UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences Chimiques

Institut Charles Sadron, CNRS

Thèse présentée par :

Eline LAURENT

soutenue le : **05 novembre 2020**

pour obtenir le grade de : **Docteure de l'Université de Strasbourg**

Discipline/Spécialité : Chimie des Polymères

Synthèse et optimisation de séquences de poly(phosphodiester)s à haute capacité de stockage d'information et à lecture facilitée

THÈSE dirigée par :

M. Jean-François LUTZ	Directeur de recherche au CNRS, ICS, Strasbourg

RAPPORTEURS :

M. Mathias DESTARAC	Professeur, IMRCP, Toulouse
Mme Sophie GUILLAUME	Directrice de Recherche au CNRS, ISCR, Rennes

AUTRES MEMBRES DU JURY :

Mme Laurence CHARLES	Professeure, ICR, Marseille
M. Marc-André DELSUC	Directeur de Recherche au CNRS, IGBMC, Strasbourg
M. Bertrand DONNIO	Directeur de Recherche au CNRS, IPCMS, Strasbourg

Remerciements

Pendant ces trois années de thèse de nombreuses personnes ont été impliquées aussi bien scientifiquement qu'humainement dans ce projet et je tiens à les remercier pour leur soutient, leur aide, leurs encouragements et encore leurs conseils. Sans elles cette thèse ne se serait pas déroulée de façon aussi intéressante et épanouissante.

Pour commencer, je tiens à remercier les membres de mon jury Mme. Sophie Guillaume, Mme. Laurence Charles, M. Mathias Destarac, M. Bertrand Donnio et M. Marc-André Delsuc pour avoir accepté d'évaluer mon travail.

Je souhaite remercier l'Agence Nationale de la Recherche qui a financé ce projet de thèse (ANR 00111001), ainsi que le LabEX CSC (Grant Numbers. ANR-16-CE29-0004-01 et ANR-16-CE29-0004-02), et le CNRS.

Je souhaite également remercier Jean-François Lutz de m'avoir engagée et de m'avoir fait confiance pour mener à bien ce projet de thèse. Travailler au sein de son équipe reconnue pour être pionnière dans la chimie de précision macromoléculaire a été épanouissant et très intéressant. Travailler sur un sujet à la pointe de l'innovation et en plein essor est très valorisant, je remercie donc Jean-François pour m'avoir permis de faire partie de son équipe. L'expertise que j'ai pu acquérir dans la chimie macromoléculaire de précision à ses côtés ne pourra m'être qu'utile dans ma future carrière. Nos échanges et ses remarques toujours constructives m'ont permis de m'améliorer, de ce fait le travail en est ressorti valorisé, notamment par l'obtention du prix du meilleur poster lors du congrès EPF.

Je veux également remercier toutes les personnes impliquées dans le projet ANR : l'équipe de Laurence Charles de l'Institut de Chimie Radicalaire de Marseille avec notamment Jean-Arthur, Salomé et bien entendue Laurence qui ont toujours été là pour répondre à mes questions sur les analyses de spectrométrie de masse. Également de l'IRC de Marseille, je tiens à remercier l'équipe de Didier Gigmes et plus particulièrement Kévin avec qui nous avons travaillé sur l'incorporation des nouvelles molécules espaceurs. Ce fut agréable de pouvoir discuter de nos petits problèmes de thésards ! Je souhaite également remercier Marc-André Delsuc qui est le dernier partenaire impliqué dans ce projet et qui m'a beaucoup aidé à comprendre comment fonctionne un code et comment le compresser. Grâce à lui je suis devenue une experte en 0 et 1 !

Pour continuer, je reviens au sein de l'équipe dans laquelle j'ai effectué ma thèse pour remercier tous les membres de l'équipe de Chimie Macromoléculaire de Précision. Tout d'abord, je tiens à remercier Laurence Oswald qui m'a permis d'évoluer en toute sécurité dans le laboratoire. Grâce à son aide, nous avons pu synthétiser très rapidement de nombreux monomères qui m'ont été essentiels. Travailler à ses côtés dans le laboratoire pendant ces trois années m'a permis d'acquérir des techniques et réflexes qui me seront toujours très utiles, ses conseils ont tous été bons à prendre !

Je tiens également à remercier les deux post-doctorants qui m'ont suivi tout au long de ces trois ans : Aziz et Tathagata et qui ont toujours été à l'écoute des soucis que je pouvais rencontrer. Leur expertise sur les synthétiseurs m'a été très bénéfique ainsi que leur force pour ouvrir les bouteilles lors du remplissage d'acétonitrile !
Ensuite je tiens à remercier l'ancienne génération de doctorants : Chloé, Guillaume, Niklas, Gianni, Denise et Benoit qui m'ont très bien accueilli au sein du « Precious Members Club ». Chloé a été là au commencement de ma thèse pour me montrer la bonne direction à prendre dès le début en me

donnant tous ses trucs et astuces pour collaborer avec mes futurs collègues. J'ai eu la chance de la retrouver un an plus tard lorsqu'elle a été engagée dans une autre équipe de l'institut, nos petites pauses potins étaient un vrai plaisir ! Guillaume a été d'une grande aide pour comprendre toutes les subtilités des demandes de l'école doctorale, ainsi que pour comprendre toutes les fonctionnalités de chaque logiciel que j'ai pu utiliser. Il a toujours su m'aider même une fois parti ! Avec lui nous avons aussi fait de grands projets artistiques que ce soit lors de la confection des chapeaux de diplômés ou lors de la création de Claude le bonhomme de neige en gobelet ! Niklas et Gianni ont été mes collègues de bureau pendant ma première année, grâce à eux ma connaissance en gros mots italien et allemand s'est grandement améliorée. Avec eux, l'ambiance dans le bureau était toujours très bonne et joyeuse. Denise, ma copine grecque, avec elle nous nous sommes parlé dans toutes les langues : français, anglais, espagnol voir un peu de grec et cela au sein d'une seule et même phrase. Notre façon de communiquer restera à jamais ma préférée. Grâce à elle j'ai pu découvrir les merveilles d'Athènes et des îles des cyclades, ainsi que leurs spécialités culinaires qui ont été un plaisir pour mon palais ! Et je sais que si j'ai besoin d'un avis sur un plat grec, je n'ai qu'à l'appeler elle sera toujours ravie de me conseiller !

Je tiens également à remercier Roza et Evgeniia avec qui j'ai pu travailler et partager. Leurs conseils que ce soit sur le domaine du travail ou culinaire et touristique ont toujours été les meilleurs. Et je sais que si un jour j'ai envie de partir les voir en Pologne ou en Russie je serai accueillie à bras ouverts.
Tous ensembles nous avons vécu des moments de rires et de joie que ce soit au travail, en pauses café ou en vacances. Et même n'étant plus avec moi au laboratoire ils continuent tous à me donner leurs conseils de loin, une vraie amitié est née et j'espère qu'elle durera longtemps.

Pendant la période qui suivie, de nouvelles personnes ont partagés mon quotidien au laboratoire : Clothilde, Duncan et Antoine. Nous avons pu partager de bons moments ensembles. Clothilde a été d'une grande aide au laboratoire grâce à son expérience entant que post-doctorante. Ça a été un plaisir d'aider Duncan et Antoine dans la poursuite de leurs études et ainsi de pouvoir les retrouver en tant que doctorants dans des équipes voisines.

J'ai également eu la chance de travailler avec deux étudiants en stage sur mon sujet. Bien que Marie ne fût en stage que trois semaines, elle a fourni un très bon travail qui m'a permis d'avancer sur des points importants de ma thèse. De plus, sa gentillesse et sa bienveillance lui ont permis de s'intégrer très rapidement au sein de l'équipe, elle a compris que les gâteaux étaient notre péché mignon … Quant à Anton, son court stage à temps partiel n'était pas forcément le mieux pour s'intégrer mais il a fourni un travail satisfaisant qui a permis d'acquérir des résultats préliminaires importants pour la continuation d'un projet.

Finalement, la nouvelle génération de doctorants, post-doctorants et stagiaires est arrivée, cette fois-ci c'est moi qui les ai accueillis au sein de l'équipe et j'espère qu'ils en sont contents. La nouvelle règle d'amener un goûter à son arrivée a été instaurée, Seydina, Ian, Laurie, Mattia, Sevtlana, Elisa, Nadia et Maria ont tous joués le jeu. Ce qui nous a permis de nous retrouver régulièrement pendant nos pauses café autour de plats typiques de chaque région du monde ! Même si malheureusement la pandémie du Covid-19 nous a obligé à arrêter nos petites réunions festives… Mon nouveau compagnon de bureau Ian est à la hauteur de mes attentes et c'est agréable de pouvoir à nouveau rigoler avec un autre doctorant et de martyriser les stagiaires en leur demandant de nous faire des gâteaux !

Je voudrais continuer en restant au sein de l'institut et remercier mes collègues un peu plus lointains, Thiebault et Sébastien avec qui mes pauses café étaient souvent axées sur nos dernières confections de produits maison : gel douche, shampoing ou encore lessive n'ont plus aucun secret pour nous !
Je remercie aussi Mélanie Legros qui a effectué les analyses de SEC et avec qui j'ai beaucoup échangé pour comprendre mes résultats .
Je souhaite également remercier toutes les personnes de l'administration avec qui j'ai pu échanger que ce soit pour l'envoi d'un de mes nombreux paquets pour Marseille avec surtout Jean-Marc de l'accueil ou des questions plus personnelles avec notamment Magali et Odile qui m'ont bien rassuré pour mon opération de la thyroïde. Et plus généralement je souhaite remercier toutes les personnes avec qui j'ai pu échanger pendant ces trois années au sein de l'institut.
Je souhaite également remercier toutes les personnes que j'ai pu côtoyer au sein du Bureau des Jeunes Chercheurs. C'était agréable d'organiser avec vous les événements sociaux de l'institut comme le découpage des citrouilles à Halloween ou la décoration de sapin vivant à Noël.
Pendant mes trois années de thèse j'ai donc pu rencontrer énormément de personnes qui m'ont permis d'avoir un quotidien au laboratoire tout à fait plaisant. Sans eux ma thèse au sein de l'ICS aurait été beaucoup plus triste.

Je veux également remercier toutes les personnes extérieures à l'institut avec qui j'ai pu partager mes doutes et craintes mais également et surtout avec qui j'ai pu me changer les idées et reprendre des forces pour revenir en pleine forme au travail.
Tout d'abord je souhaite remercier mon copain de pause Paul, avec qui j'ai pris beaucoup de pauses digestion après le repas de midi. Pendant ces temps-là nous avons pu partager tous nos petits soucis quotidiens de thésard et nous remotiver l'un l'autre. Grâce à lui j'ai pu trouver les réponses à mes nombreuses questions de chimie organique mais également augmenter considérablement mes scores à Mario Kart, mes connaissances en chansons françaises et mes abonnements aux comptes instagram indispensables pour rire tous les jours.
Ensuite je veux remercier tous mes amis de Strasbourg : Sophie, Justine, Eva, Geoffrey, Adrien, Francis, Cécile, Hugo, Maëva et Vincent. Se retrouver autour d'une bière ou d'un gin tonic avec eux est toujours un plaisir ! Si en plus ça se passe au Grincheux ou au Garde-fou je ne peux rien demander de mieux ! Mention spéciale pour Adrien et Geoffrey qui m'ont poussé à me lancer dans une nouvelle carrière d'influenceuse instagram !
C'est important aussi pour moi de remercier Marion qui est partie s'exiler à l'autre bout de la France. Mon ancienne colocataire, ma copine de voyage, ma Toupinette, même à 1038 km de distance nous restons toujours connectées. C'est grâce à elle que je sais manier une colonne de chromatographie ce qui m'a bien était utile tout au long de ma thèse. Se retrouver est toujours un vrai plaisir, son écoute et son aide m'ont été d'un bénéfice énorme pendant ces trois ans et le resteront à vie.
Je tiens aussi à remercier mes amis de CPI avec qui je suis resté en contact pendant toutes ces années. Tous ces week-ends retrouvailles sont vraiment des moments superbes remplis de rires et de joie. Spéciale dédicace à Eléonore, avec qui on se suit depuis le lycée. Même si nos chemins se sont un peu séparés au moment de rentrer en école d'ingénieur nous avons gardé le même cap qui nous mène cette année à devenir docteure. Nous avons toujours réussi à nous retrouver pour partager des moments inoubliables, avec le dernier en date, notre congrès en Crète suivi de nos petites vacances à la découverte des plus belles plages de l'île ainsi que la recherche du meilleur raki ! Yamas !

Ma famille a également été très présente pour m'épauler pendant ces trois années. Avec en premier lieu, mes parents que je souhaite remercier. Mon père pour avoir toujours été là pour répondre à mes

questions et me donner son avis entant que scientifique et surtout pour m'avoir poussé à donner le meilleur de moi-même. J'espère l'avoir rendu fière et pour ma part je suis fière d'atteindre enfin le même niveau d'étude que lui et de représenter à ses côtés les docteurs de la famille Laurent. Ma mère pour avoir toujours été à l'écoute de tous mes soucis qu'ils soient logistiques avec des questions administratives ou plus futiles avec des questions sur comment organiser mes prochains week-ends et aussi mes questionnements sur mon travail même si parfois je pouvais la perdre avec les termes techniques. Elle a toujours su me remotiver dans les moments de doutes et toujours me faire voir les bons côtés. Merci à tous les deux pour m'avoir soutenu tout au long de ces trois ans.
Je remercie également mon oncle Franck, ma tante Isabelle et mes cousins Fabien et Luc pour m'avoir soutenu pendant ces trois années de thèse. Je sais qu'ils feront tout leur possible pour venir me soutenir le jour de ma soutenance et ça me touche énormément.
Mes grands-mères sont aussi très importantes pour moi et je tiens à les remercier d'avoir partagé beaucoup de coups de téléphone passés sur mon chemin de retour du travail. C'était un plaisir de leur raconter les journées que je venais de passer et elles ont toujours essayé de comprendre un maximum ce que je faisais même si elles ne baignent pas du tout dans l'univers scientifique. Leur expliquer ma thèse et les expériences que je fais est un vrai défi de vulgarisation et elles m'ont ainsi permis de m'améliorer au cours de ces trois ans en étant toujours à l'écoute.

Pour finir je remercie Yvan, qui a été là pour moi tout au long de mes trois ans de thèse. Après notre rencontre à Montpellier, le retour à Strasbourg pour débuter ma thèse a été difficile mais la difficulté ne nous faisant pas peur nous avons réussi à braver la distance. Même si le chemin entre Strasbourg et Montpellier est long, nous avons tenu le coup et les retrouvailles ont toujours values les heures passées dans le train. Nos week-ends et voyages ont été très importants pour moi, et m'ont permis de me vider la tête et de me ressourcer à ses côtés pour toujours être au plus haut de ma forme. Il a toujours su me booster et m'épauler au maximum pour que je puisse donner le meilleur de moi-même. Il m'a également montré le bon chemin en réalisant une très belle thèse, il a été le meilleur exemple à suivre. Je le remercie énormément pour tous ces moments passés à côté de lui et pour tous ceux qui nous reste encore à vivre.

Table des matières

Remerciements 1
Liste des abréviations 11
Liste des figures 14
Liste des tableaux 18
Introduction générale 21
Chapitre I : Stockage d'information à l'échelle moléculaire 29
1. Introduction 31
2. Le stockage naturel d'information à l'échelle moléculaire dans les systèmes biologiques 31
2.1. L'acide désoxyribonucléique (ADN) 31
2.1.1. Structure de l'ADN 31
2.1.2. Réplication : la biosynthèse de l'ADN 33
2.1.3. Dénaturation de l'ADN 34
2.2. L'acide ribonucléique (ARN) 34
2.2.1. Structure de l'ARN 34
2.2.2. Biosynthèse de l'ARN 34
2.3. Les protéines 35
2.3.1. Structure des protéines 35
2.3.2. Biosynthèse des protéines 36
2.3.3. La synthèse chimique des protéines 36
3. Synthèse chimique d'oligonucléotides 38
3.1. Historique de l'optimisation de la synthèse de l'ADN 38
3.1.1. Synthèse du premier dinucléotide 38
3.1.2. Méthode phosphodiester 38
3.1.3. Développement de la synthèse sur support solide 40
3.1.4. Méthode phosphotriester 41
3.1.5. Méthode phosphite triester 42
3.1.6. Méthode phosphoramidite 43
3.2. Développement de la chimie de la phosphoramidite 44
3.2.1. Description des synthétiseurs d'ADN 44
3.2.2. Description du cycle itératif 46
3.2.3. Supports solides utilisés 50
3.2.4. Groupements protecteurs 50

3.3. Stockage de données sur l'ADN 52
3.3.1. Principe : comment stocker de l'information ? 52
3.3.2. Automatisation du stockage de données 54
4. Stockage de données sur des macromolécules synthétiques 55
4.1. Nécessité d'avoir un système bien contrôlé 55
4.1.1. Polymérisation par étape 55
4.1.2. Polymérisation en chaine 56
4.1.3. Synthèses multi-étapes 57
4.2. Diversité des espèces chimiques possibles pour le stockage d'information sur des polymères synthétiques 59
4.2.1. Les oligo(triazoles amide)s 60
4.2.2. Les oligo(alcoxyamines amide)s 60
4.2.3. Les oligo(alcoxyamines phosphodiester)s 60
4.2.4. Cryptage de données 61
4.2.5. Les polyuréthanes 62
4.2.6. Les poly(N-substitués uréthane)s 64
4.2.7. Réactions multi-composants : codage avec des petites molécules 64
4.2.8. Les séquences multi-fonctionelles 65
4.2.9. Les poly(succinimide thioether)s linéaires et dendritiques 65
4.2.10. Les polyesters 66
4.2.11. Les peptoïdes 66
4.3. Utilisation de la chimie de la phosphoramidite pour le stockage d'information sur des polymères à séquences définies 67
4.3.1. Phase d'écriture : développement de l'alphabet classique 67
4.3.2. Phase de la lecture : analyse par spectrométrie de masse 69
4.3.3. Développement de différents alphabets binaires 69
4.3.4. Augmentation de la densité de stockage via un stockage spatial 76
5. Conclusion et perspectives 77
Chapitre II : Développement d'alphabets moléculaires augmentés 79
1. Introduction 81
2. Structure des poly(phosphodiester)s utilisés pour coder de l'information 82
2.1. Synthèse utilisant la chimie de la phosphoramidite 82
2.1.1. Synthèse des monomères 83
2.1.2. Synthèse automatisée 84
2.1.3. Lecture facilitée par spectrométrie de masse 86

3. Conception d'alphabets augmentés 88
3.1. Alphabet avec codage sur la chaîne principale 89
3.1.1. Synthèse des monomères linéaires 89
3.1.2. Séquences tests avec les monomères linéaires 89
3.1.3. Séquences codant de l'information avec les monomères linéaires 92
3.2. Alphabet à chaînes pendantes 95
3.2.1. Première génération de l'alphabet à 4 symboles 95
3.2.2. Deuxième génération de l'alphabet à 4 symboles 98
3.2.3. Création de l'alphabet à 8 symboles 100
3.2.4. Evaluation de l'efficacité de couplage des monomères encombrés 105
4. Utilisation des alphabets augmentés pour stocker des images 113
4.1. Séquences codant des images avec l'alphabet à 4 symboles 113
4.2. Séquences codant des images avec l'alphabet à 8 symboles 117
5. Conclusion et perspectives 124
Chapitre III : Optimisation de l'espaceur contenant une liaison alcoxyamine 127
1. Introduction 129
2. Réactions parasites observées avec l'utilisation de l'espaceur classique 130
3. Développement d'espaceurs optimisés 133
3.1. Première stratégie : délocalisation du radical 133
3.2. Seconde stratégie : rigidifier le squelette de l'espaceur 136
3.2.1. Espaceur RISC 136
3.2.2. Espaceur RISCOP 141
4. Synthèse de polymères incorporant des images 146
4.1. Analyse et lecture manuelle des séquences 148
4.1.1. Analyse par chromatographie en phase liquide à haute performance 148
4.1.2. Lecture par spectrométrie de masse 150
4.2. Lecture automatisée des séquences avec le logiciel MS-DECODER 151
5. Amélioration de l'espaceur afin de lui conférer de nouvelles propriétés 153
5.1. Elaboration d'espaceurs photosensibles 153
5.2. Elaboration d'espaceur-marqueur 157
6. Conclusion et perspectives 159
Chapitre IV : Synthèse de polymères à très haute capacité de stockage 163
1. Introduction 165
2. Choix des marqueurs de masse 166

3. Utilisation d'outils de compression et de codes de correction pour augmenter la capacité de stockage 169
3.1. Compression de l'information 170
3.1.1. Algorithmes avec perte dits « Lossy » 170
3.1.2. Algorithmes sans perte dits « Lossless » 171
3.2. Contrôle des erreurs 174
4. Elaboration d'une séquence codant une grande quantité d'information 174
4.1. Utilisation de l'espaceur E1 176
4.2. Utilisation de l'espaceur RISCOP 179
5. Conclusion et perspectives 182
Conclusion générale 185
Partie expérimentale 191
1. Techniques d'analyses et matériels utilisés 193
1.1. Résonnance magnétique nucléaire (RMN) 193
1.2. Spectrométrie de masse 193
1.3. Spectrophotomètre UV-Vis 193
1.4. Chromatographie d'exclusion stérique 193
1.5. Equipement du laboratoire 194
1.5.1. HPLC 194
1.5.2. Lyophilisateur 194
1.5.3. Synthétiseur d'ADN 194
1.5.4. Chromatographies 194
2. Réactifs et solvants 194
3. Partie expérimentale du chapitre II 196
3.1. Synthèses des molécules présentées dans le chapitre II 196
3.1.1. Synthèse des monomères linéaires 196
3.1.2. Synthèse des monomères à chaînes latérales 196
3.1.3. Synthèse de la molécule espaceur 200
3.2. Caractérisation des molécules présentées dans le chapitre II 202
3.2.1. Caractérisation des monomères linéaires 202
3.2.2. Caractérisation des monomères à chaîne latérales 209
3.3. Caractérisation des monomères par spectrométrie de masse 229
3.4. Synthèse des polymères 229
3.4.1. Préparation des solutions de monomères 229
3.4.2. Description du langage utilisé 230

3.4.3. Description des protocoles utilisés sur les synthétiseurs d'ADN 230
3.4.4. Clivage des polymères 233
3.5. Analyses des polymères par spectrométrie de masse 234
3.5.1. Méthode de séquençage : 234
3.5.2. Polymères obtenus avec l'alphabet linéaire 236
3.5.3. Polymères obtenus avec l'alphabet à 4 symboles de première génération 243
3.5.4. Polymères obtenus avec l'alphabet à 4 symboles de seconde génération 248
3.5.5. Polymères obtenus avec l'alphabet à 8 symboles 251
3.5.6. Séquences codant des images avec l'alphabet à 4 symboles 254
3.5.7. Séquences codant des images avec l'alphabet à 8 symboles 258
4. Partie expérimentale du chapitre III 267
4.1. Analyse de polymères par spectrométrie de masse 267
4.1.1. Séquences impliquant l'espaceur ROSC 267
4.1.2. Séquences impliquant l'espaceur RISC 270
4.1.3. Séquences impliquant l'espaceur RISCOP 271
4.1.4. Séquences impliquant l'espaceur PRISC 280
4.1.5. Séquences impliquant l'espaceur NISC 282
4.1.6. Séquence impliquant l'espaceur EM-Br 284
5. Partie expérimentale du chapitre IV 285
Références bibliographiques 289
Liste des publications 297
Liste des présentations 297

Liste des abréviations

A

A	Adénine
ADN	Acide désoxyribonucléique
ADMET	Acyclic diene metathesis
anh.	Anhydre
ARGET	Activator regenerated by electron transfer
ARN	Acide ribonucléique
ARNm	ARN messager
ARNt	ARN de transfert
ASCII	American Standard code for Information Interchange
ATRP	Atom Transfer Radical Polymerization

B

Br	Brome
Bz	Benzoyle

C

C	Cytosine
CCM	Chromatographie sur couche mince
CE	2-cyanoéthyle
CID	Collision-induced dissociation
CNRS	Centre National de la Recherche Scientifique
CPG	Controlled Pore Glass
CuAAC	Copper-catalysed azide–alkyne cycloadditions

D

dA	Deoxyadenosine
dC	Deoxycytidine
DCC	*N,N'*-dicyclohéxylcarbodiimide
DCM	Dichlorométhane
dG	Deoxyguanine
DIPEA	*N,N*-diisopropyléthylamine
DMT	4,4'-diméthoxytrityl
DMT-Cl	Chlorure de 4,4'-diméthoxytrityl
DP	Degré de Polymérisation
dT	Deoxyrhymidine

E

ESI	Ionisation par électronébuliseur (Electrospray Ionization)
EtOAc	Acétate d'éthyle

F

F	Fluor

G

G	Guanine

H

HRMS	Spectrométrie de masse haute résolution
HPLC	Chromatographie en phase Liquide à Haute Performance

I

I	Iode
I_2	Iodine
ICAR	Initiators for continous activator regeneration

L

LbL	Layer by Layer (couche par couche)
lcaa	Longue-Chaine alkyle amine
LCST	Température de solubilisation critique basse

M

MALDI-TOF	Matrix Assisted Laser Desorption Ionization - Time of Flight
MeCN	Acétonitrile
MeOH	Méthanol
MMT	Monométhoxytrityle
$MgSO_4$	Sulfate de magnésium
MS	Spectrométrie de masse
MS/MS	Spectrométrie de masse en tandem
m/z	Ratio masse sur charge

N

$NaHCO_3$	Bicarbonate de sodium
Na_2SO_4	Sulfate de sodium anhydre
NEt_3	Triéthylamine
NMP	Nitroxide-Mediated polymérisation

P

P-3CR	Réaction de Passerini trois composants
PCR	Polymerase Chain reaction

PU	Poly(uréthane)

R

RAFT	Reversible addition-fragmentation chain tranfer
RMN	Résonance Magnétique Nucléaire
Rmq	Remarque
ROMP	Ring-opening metathesis polymerization
ROP	Ring-opening polymerization

S

SARA	Supplemental Activator and Reducing Agent
SEC	Chromatographie d'Exclusion Strérique
SPPS	Synthèse Peptidique en phase Solide

T

T	Thymine
Ta	Température ambiante
TCA	Acide trichloracétique
TEA	Triéthylammonium
TEMPO	(2,2,6,6-tétraméthylpipéridin-1-yl)oxy
THF	Tétrahydrofurane
TIPS	Triméthylsilyle
Tr	Trityle
TsOH	Acide *p*-toluène-sulfonique

U

U	Uracile
U-4C	Réaction Ugi quatre composants
UV-Vis	Ultraviolet/visible

Liste des figures

Figure 1: Schéma des structures des monomères menant à des oligonucléotides (à gauche) avec B représentant les bases de l'ADN (Adénine, Guanine, Thymine ou Cytosine) ou à des poly(phosphodiester)s synthétiques (à droite) avec $R_1 = R_2 = H$: monomère 0 et $R_1 = R_2 = CH_3$: monomère 1. 25
Figure 2: En haut, schéma du design d'une séquence contenant un espaceur (en rouge) et des marqueurs (en vert). En bas, exemple de structure d'une chaîne de poly(phosphodiester)s contenant un espaceur et un marqueur de masse. 26
Figure 3: Plan général des études menées. En haut, séquence type de poly(phosphodiester)s codant de l'information. Bleu : Amélioration de l'alphabet (chapitre II). Rouge : Amélioration de l'espaceur (chapitre III). Vert : amélioration des marqueurs moléculaires (chapitre IV). 26
Figure 4: Spectre RMN ^1H ($CDCl_3$) du monomère codant pour (01) 213
Figure 5: Spectre RMN ^1H ($CDCl_3$) du monomère codant pour (01) 213

Figure I. 1 : (a) Représentation schématique de la structure hélicoïdale de l'ADN. Les brins rouge et bleu représentent la chaine phosphate et les lignes horizontales représentent les liaisons hydrogènes entre les paires de bases. (b) Représentation schématique de la formule chimique d'une simple chaîne. (c) Représentation des deux paires de bases de l'ADN et de leurs liaisons hydrogènes. 32
Figure I. 2: Schéma de la réplication semi-conservative de l'ADN. 33
Figure I. 3 : Représentation des quatre structures des protéines. (a) Structure primaire : structure d'un acide aminé. (b) Structure secondaire : Conformation spécifique des chaînes polypeptides (gauche : alpha hélice, droite : feuillet beta). (c) Structure tertiaire : Conformation dans l'espace (exemple de l'insuline). (d) Structure quaternaire : Conformation dans l'espace d'une protéine formée de plusieurs sous unités (exemple de l'hémoglobine A humaine). 35
Figure I. 4: Protocole générale de la méthode de synthèse peptidique sur support solide. (a) Ajout des réactifs en excès en solution. (b) Couplage. (c) Filtration et lavages avec les solvants adéquats. 37
Figure I. 5: Réaction suivie par Todd pour obtenir le premier di-nucléotide synthétique. i) Couplage, 2,6-lutidine ii) Hydrolyse. 38
Figure I. 6: Mécanisme de la méthode phosphodiester utilisée pour la synthèse d'oligonucléotides. 39
Figure I. 7: Supports solides pour la synthèse d'oligonucléotides. (a) Support cytidine. (b) Support mono-méthoxytrityle (MMT). (c) Support 3'-ester. 41
Figure I. 8: Synthèse d'oligonucléotides avec la méthode phosphotriester. 41
Figure I. 9: Synthèse d'oligonucléotides avec la méthode phosphite triester. 43
Figure I. 10: Schéma du nucléotide substitué par la phosphoramidite N,N-dimétylamine (gauche) et le nucléotide subtitué par la phosphoramidite N,N-diisopropylamine (droite). 44
Figure I. 11: Photographie du premier synthétiseur d'ADN, modèle 280. 45
Figure I. 12: Photographie du premier synthétiseur d'ADN 380A mis en place dans le laboratoire de Caruthers en 1982. Tirée de la littérature.[73] 45
Figure I. 13: Photographie du synthétiseur Applied Biosystems, modèle 3900. 45
Figure I. 14: Schéma récapitulatif des différentes étapes du cycle itératif de la chimie de la phosphoramidite. 46
Figure I. 15: Schéma de l'étape de déprotection de l'alcool. 47
Figure I. 16 : Schéma de l'étape de couplage de la phosphoramidite. 47
Figure I. 17: Schéma de l'étape d'oxydation. 48
Figure I. 18: Schéma de l'étape de capping. 48
Figure I. 19: Structure du support solide lcaa standard. 50
Figure I. 20: Classification générale des techniques de polymérisation permettant d'obtenir des polymères synthétiques à séquences contrôlées. (adapté de la littérature[93] suggérant la terminologie des synthèses de polymères). 55

Figure I. 21: Comparaison des différents systèmes macromoléculaires et de leur capacité de stockage théorique dans une unité répétitive. Rouge : Information codée via l'ADN ; Bleu : Macromolécules à séquence contrôlées basée sur l'encodage binaire (gauche[25], milieu[117], droite[118]) ; Vert : Macromolécules à séquences contrôlées via une réaction multicomposants (gauche[119], droite[120]) Violet: Séquences contrôlées multifonctionnelles.[121] ; Orange : Macromolécules à séquences contrôlées sur des dendrimères.[122] 59
Figure I. 22: Structures moléculaires et masses molaires des monomères alcoxyamines phosphodiesters codants. 61
Figure I. 23: Structure d'une séquence codées de poly(N-substitués uréthane)s. Avec R = Me (00), Et (01), Pr (10), Bu (11). 64
Figure I. 24: Schéma expliquant la création des monomères utilisés pour synthétiser des polymères numériques. 68
Figure I. 25: Représentation schématique du séquençage en spectrométrie de masse d'un polymère numérique contenant 4 octets d'information. Séquence codée avec 2 monomères différents. Adaptée de la littérature.[25] 72

Figure II. 1 : Concept des alphabets augmentés codant sur 4 symboles (orange) ou 8 symboles (violet). 81
Figure II. 2 : Description des trois étapes permettant l'ajout d'un monomère sur la séquence. Le demi-cercle rose correspond à la chaîne en croissance attachée au support solide. La boule bleue correspond à la partie centrale de chaque monomère ajouté qui peut être n'importe quelle espèce chimique. 83
Figure II. 3: Structure des monomères de phosphoramidite contenant le groupement réactif phosphoramidite (orange) et le groupement protecteur DMT (violet). Gauche : monomère codant pour le bit **0**, droite : monomère codant pour le bit **1**. 83
Figure II. 4 : Schéma réactionnel de la synthèse des monomères de phosphoramidite. La boule bleue correspond à la partie centrale de chaque monomère qui peut être n'importe quelle espèce chimique. . 84
Figure II. 5 : Photographie des synthétiseurs d'ADN de l'équipe. (Gauche) Expédite, (Droite) ABI. 85
Figure II. 6 : Schéma de la procédure de clivage. 86
Figure II. 7 : Structure moléculaire des marqueurs de masse et de l'espaceur utilisés. Les groupements protecteurs des amines des bases sont colorés en rouge. 87
Figure II. 8 : Structure schématique d'une séquence contenant 5 blocs. Les blocs sont représentés par des rectangles orange, les espaceurs par des rectangles rouges et les marqueurs par des boules vertes. 88
Figure II. 9 : Structure des monomères synthétisés pour l'alphabet linéaire. 89
Figure II. 10 : Schéma de la séquence **L1**. (a) Spectre ESI-MS en mode négatif de L1. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux annotés en gris à des fragments formés en source. (b) Spectre MS/MS du polymère L1 avec $[M - 8H]^{8-}$ à m/z 552,0. (c) Schéma des blocs obtenus en MS/MS analysés en pseudo-MS^3 93
Figure II. 11 : Spectre ESI-MS en mode négatif de **L2**. Les pics annotés en gris correspondent aux fragments formés en source attendus pour l'oligomère recherché, ceux annotés en noirs correspondent à d'autres fragments formés en source. 94
Figure II. 12 : Structure des monomères de la première génération de l'alphabet à 4 symboles. 95
Figure II. 13 : (En haut) Schéma de la structure du polymère **P1**. Spectre ESI-MS en mode négatif de P1. L'oligomère recherché est observé sous les formes $[M-2H]^{2-}$ à m/z 917,2, $[M-3H]^{5-}$ à m/z 611,2 et $[M-4H]^{4-}$ à m/z 458,1. Les pics au pied de l'ion m/z 917,2 correspondent à des échanges H/Na et H/K. Inset : spectre ESI-MS/MS de l'ion $[M-4H]^{4-}$ à m/z 458,1 montrant la formation des 2 fragments attendus : le bloc 1 doublement déprotoné à m/z 368,1 et le bloc 2 doublement déprotoné à m/z 548,1. 97
Figure II. 14 : Structure des monomères de la seconde génération de l'alphabet à 4 symboles. 98
Figure II. 15 : Analyse de spectrométrie de masse du polymère **P4**. (En haut) Schéma de la structure globale du polymère. (a) Spectre MS montrant les différents états de charges du polymère en vert et de l'impureté en rouge. (b) Schéma de fragmentation en MS/MS. (c) Spectre de masse en tandem issu de l'ion

précurseur $[M-12H]^{12-}$ à m/z 584,4. (d) Blocs issus du spectre MS/MS, puis analyse MS^3 pour retrouver le texte encodé. 100
Figure II. 16 : Structures moléculaires des monomères synthétisés. Orange : monomères utilisés pour l'alphabet à 4 symboles (les monomères codent des dyades : 2 bits par unité). Violet : monomères utilisés pour l'alphabet à 8 symboles (les monomères codent des triades : 3 bits par unité). 101
Figure II. 17 : Spectres de masse obtenus pour les polymères **P5** (a) et **P6** (b). 103
Figure II. 18 : Spectre de ESI-MS en mode négatif des pentamères modèles M2.M2.M2.M2.M2-T (en haut), M6.M6.M6.M6.M6-T (à gauche) et M7.M7.M7.M7.M7-T (à droite). Les états de charges sont annotés entre parenthèses. Les diamants désignent des formes ioniques détectées après des échanges H/Na ou H/K. §: bruit de fond. # Amas d'acide trichloroacétique. 106
Figure II. 19 : Spectre de ESI-MS en mode négatif du pentamère modèle M4.M4.M4.M4.M4-T (en haut) et M5.M5.M5.M5.M5-T (en bas). Les états de charges sont annotés entre parenthèses. Les signaux en gris correspondent aux oligomères de plus faible DP. Les diamants désignent des formes ioniques détectées après des échanges H/Na ou H/K. # Amas d'acide trichloroacétique. 107
Figure II. 20 : Spectre de ESI-MS en mode négatif du pentamère modèle M8.M8.M8.M8.M8-T avec les états de charges annotés entre parenthèses. Signaux en gris correspondent aux oligomères de plus faible DP. # Amas d'acide trichloroacétique. 108
Figure II. 21 : Analyse de spectrométrie de masse du polymère **P8**. (En haut) Schéma de la structure globale du polymère. (a) Spectre MS montrant les différents états de charges du polymère. Les étoiles grises sont des fragments formés en source. (b) Spectre de masse de déconvolution. (c) Spectre de masse en tandem issu de l'ion précurseur $[M-15H]^{15-}$ à m/z 585,6. (d) Blocs issus du spectre MS/MS, puis analyse MS^3 pour retrouver l'image encodée. 115
Figure II. 22 : Séquençage du polymère **P8.** (a) Schéma de la fragmentation obtenue après l'analyse MS/MS de l'ion précurseur $[M-15H]^{15-}$ à m/z 585,6. Spectres Pseudo-MS^3 et la couverture associée des séquences pour (b) le premier bloc à m/z 471,1, (c) le second bloc à m/z 630,8, (d) le troisième bloc à m/z 625,5, (e) le quatrième bloc à m/z 617,5, et (f) le cinquième à m/z 581,8. Les monomères déprotonés sont désignés par un astérisque et les pics en gris sont des sous-produits de fragmentation. 116
Figure II. 23 : Schéma des fragments obtenus pour chaque monomère lors de l'analyse de pseudo-MS^3. 117
Figure II. 24 : Analyse de spectrométrie de masse du polymère **P11**. (En haut) Schéma de la structure globale du polymère P11. (a) Spectre MS montrant les différents états de charges du polymère. Les annotations grises sont des fragments formés en source. (b) Schéma de fragmentation en MS/MS. (c) Spectre de masse en tandem issu de l'ion précurseur $[M-12H]^{12-}$ à m/z 570.1. (d) Blocs issus du spectre MS/MS, puis analyse MS^3pour retrouver l'image encodée. 120
Figure II. 25 : Spectre de masse du polymère **P14** avec des temps de clivage différents. En haut : schéma de la structure globale du polymère ; (a) clivage en 30min, (b) clivage en 60 min. 123

Figure III. 1: (a) Structure de l'espaceur classique **E1**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur classique incorporé dans une séquence suivie de la fragmentation de la liaison alcoxyamine. 131
Figure III. 2: Structure des fragments primaires pour n'importe quel bloc retrouvé en analyse MS/MS après l'homolyse de la liaison alcoxyamine. 131
Figure III. 3: Mécanisme proposé pour la perte d'un radical de 100,1 Da pour tous les segments comportant un radical sur le carbone central de la terminaison ω. 132
Figure III. 4: Mécanisme proposé pour la perte du fragment à 225,1 Da avec la production en parallèle d'un ion à m/z 224,1. 133
Figure III. 5: Mécanisme proposé pour la perte combinée de 224,1 Da et la perte d'une base. 133
Figure III. 6: (a) Structure de l'espaceur **ROSC**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur ROSC incorporé dans une séquence suivie de la fragmentation de la liaison alcoxyamine. 134
Figure III. 7: Séquençage du polymère $\mathbf{P_{ROSC}1}$. (a) Spectre MS/MS de l'ion $[M - 6H]^{6-}$ à m/z 503,9 (b) Spectre pseudo-MS^3 du bloc 1 à m/z 407,4. Les pics annotés en gris correspondent à des fragments internes,

parmi lesquels les monomères sont désignés par #. Les ions annotés en rose, rouge et violet sont issus de réactions induites par le radical de l'espaceur ROSC. 135
Figure III. 8: Description des réactions secondaires observées pour la séquence $P_{ROSC}1$. 136
Figure III. 9: (a) Structure de l'espaceur **RISC**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur RISC incorporé dans une séquence suivie du schéma de la fragmentation de la liaison alcoxyamine. 136
Figure III. 10: Histogrammes obtenus par le synthétiseur Expédite lors de la synthèse des séquences de $P_{RISC}1$ (gauche) et $P_{RISC}2$ (droite). 138
Figure III. 11: Photographies de la solution obtenue après clivage (gauche) et du liquide visqueux obtenu après lyophilisation (droite). 138
Figure III. 12: (a) Spectres obtenus après l'analyse MS/MS des impuretés détectées dans l'échantillon $\mathbf{P_{RISC}1}$. (b) Réactions secondaires suggérées lors du clivage. 139
Figure III. 13: Séquençage de la séquence $\mathbf{P_{RISC}2}$. (a) Spectre MS/MS de l'ion $[M - 9H]^{9-}$ montrant l'homolyse des liaisons NO-C. (b) Spectre pseudo-MS^3 du bloc 1 à m/z 412,0 et tableau de couverture du bloc 1. (c) Spectre pseudo-MS^3 du bloc 2 à m/z 594,3 et tableau de couverture du bloc 2. (d) Spectre pseudo-MS^3 du bloc 3 à m/z 525,7 et tableau de couverture du bloc 3. Les pics annotés en gris correspondent à des fragments internes, parmi lesquels les monomères sont désignés par #. 140
Figure III. 14: (a) Structure de l'espaceur **RISCOP**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur RISCOP incorporé dans une séquence suivie du schéma de la fragmentation de la liaison alcoxyamine. 141
Figure III. 15: Séquençage de la séquence $\mathbf{P_{RISCOP}2}$. (a) Spectre ESI-MS en mode négatif. (b)Spectre MS/MS de l'ion $[M - 6H]^{6-}$ montrant l'homolyse des liaisons C–ON. (c) Spectre pseudo-MS^3 du bloc 1 à m/z 700,1 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 2 à m/z 623,1 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 3 à m/z 513,1 et couverture de la séquence. Les pics annotés en gris correspondent au monomères déprotonés. Les ions annotés avec un rond gris correspondent aux fragments libérés en source. Les pics annotés par un astérisque correspondent à des échanges H/Na ou H/K. 142
Figure III. 16: Séquençage de la séquence $\mathbf{P_{RISCOP}3}$. (a) Spectre ESI-MS en mode négatif. Séquence détectée sous les différents états de charge 6– à 15– (en vert) ainsi que fragments générés en source pour homolyse des différentes liaisons alcoxyamine : bloc 1 en rouge, bloc 2 en rose, bloc 3 en orange et bloc 4 en violet. (b) Spectre pseudo-MS^3 du bloc 2 comprenant le marqueur dA à m/z 599,1 et couverture de la séquence. Les pics annotés en gris sont des fragments libérés en source. * : échanges H/Na. 144
Figure III. 17: Spectres obtenus après les analyses HPLC des séquences P_R1 (Main en verte), P_R2 (Souris en bleu clair), P_R3 (Fichier en bleu), P_R4 (Cigogne en violet), P_R5 (Bretzel en rose) et P_R6 (Space invader en orange). 149
Figure III. 18: Spectre ESI-MS en mode négatif des séquences P_R2 à P_R6. Les oligomères recherchés sont représentés en vert. Les échantillons P_R2 et P_R4 contiennent également une petite impureté à +149 Da et -137,8 Da respectivement. Les annotations en grises sont des fragments formés en source. 150
Figure III. 19: Analyse de spectrométrie de masse du polymère P_R1. (En haut) Schéma de la structure globale du polymère P_R1. (a) Spectre MS montrant les différents états de charges du polymère. Les annotations grises sont des fragments formés en source. (b) Spectre de masse en tandem issu de l'ion précurseur $[M-15H]^{15-}$ à m/z 587,8, ainsi que le schéma de fragmentation. (c) Blocs issus du spectre MS/MS, puis analyse MS^3 pour retrouver l'image encodée. 151
Figure III. 20: Résultats obtenus par le MS-DECODER lors de l'analyse de la séquence $\mathbf{P_R5}$ codant pour l'image d'un bretzel. 153
Figure III. 21: (a) Structure de l'espaceur PRISC. (b) Structure de l'espaceur NISC. En rouge les liaisons NO-C fragiles. 154
Figure III. 22: (En haut) Spectre ESI-MS en mode négatif de $\mathbf{P_{PRISC}1}$ en vert. (En bas) Spectre ESI-MS en mode négatif de $\mathbf{P_{NISC}1}$ en vert. Les impuretés sont annotées en couleurs. Les pics annotés par un rond gris correspondent à des fragments formés en source. # Amas de sel. 155
Figure III. 23: Séquençage de $P_{PRISC}1$ et des impuretés (a) Spectre MS/MS de l'ion $[M - 5H]^{5-}$ à m/z 613,9 et schéma d'homolyse de l'espaceur. (b) Spectre MS/MS de l'ion $[A - 3H]^{3-}$ à m/z 693,2 (c) Spectre MS/MS

de l'ion $[B - 3H]^{3-}$ à m/z 676,2 et (d) Spectre MS/MS de l'ion $[C - 3H]^{3-}$ à m/z 660,2. Dans chaque schéma de dissociation, la valeur encadrée correspond à la masse du bloc codant tandis que la valeur non encadrée correspond au segment libéré. 156
Figure III. 24: Schéma des substitutions nucléophile par la méthylamine et l'ammonique des séquences (a) $P_{PRISC}1$ (impureté C) avec un synthon X à 30 Da et (b) $P_{NISC}2$ avec un synthon à 16 Da (visible en partie expérimentale). 156
Figure III. 25: Structure de l'espaceur-marqueur **EM-Br**. 157
Figure III. 26: Spectre ESI-MS en mode négatif de $\mathbf{P_{EM\text{-}Br}1}$. Les pics annotés par un astérisque correspondent à différents échanges H/Na et H/K. Les fragments formés en source sont en gris. 158
Figure III. 27: Séquençage du polymère $P_{ROSC}2$. (a) Spectre ESI-MS en mode négatif. Les ions annotés en noir sont des espèces simplement chargées. Les pics annotés en rouge correspondent à une impureté de +217,6 Da. (b) Spectre MS/MS de l'ion $[M - 9H]^{9-}$ à m/z 507,9 (c) Schéma d'homolyse des liaisons NO-C. (c) Spectre pseudo-MS^3 du bloc 1 à m/z 407,4 (d) Spectre pseudo-MS^3 du bloc 2 à m/z 589,8. (e) Spectre pseudo-MS^3 du bloc 3 à m/z 525,7. Les pics annotés en gris correspondent à des fragments internes, parmi lesquels les monomères sont désignés par #. Les ions annotés en rose sont issus de réactions induites par le radical de l'espaceur ROSC. 269
Figure III. 28: Séquençage de $P_{PRISC}1$ et des impuretés. Spectre MS/MS de l'ion $[M - 5H]^{5-}$ à m/z 613,9 et schéma d'homolyse de l'espaceur. 280

Figure IV. 1: Schéma du concept permettant la synthèse de séquence codant des données comprimées. 166
Figure IV. 2: Structure des molécules choisies comme marqueurs de masse. 168
Figure IV. 3: Structures des molécules dF_U et dF_C. 168
Figure IV. 4: Image pixélisée d'une fiole en 13*12 utilisée comme exemple pour expliquer la compression Huffman. 172
Figure IV. 5: (Gauche) Portrait du couple Lavoisier exposé au Metropolitain Museum of Art à New York, avec le zoom sur le visage d'Antoine Laurent Lavoisier. (Droite) Image pixélisée de Lavoisier. 175
Figure IV. 6: Spectre ESI-MS obtenu pour la séquence codant l'image de Lavoisier compressée. Les différentes couleurs des pics correspondent aux couleurs attribuées à chaque bloc de la séquence schématisée en inset. 177
Figure IV. 7: Spectre ESI-MS obtenu pour la séquence codant une partie de l'image de Lavoisier compressée contenant que 9 blocs (Lav1). Les différentes couleurs des pics correspondent aux couleurs attribuées à chaque bloc de la séquence schématisée en inset. 179
Figure IV. 8: Spectre ESI-MS obtenu pour la séquence codant l'image entière de Lavoisier compressée avec l'espaceur RISCOP (Lav4). Les différentes couleurs des pics correspondent aux couleurs attribuées à chaque bloc de la séquence schématisée en inset. 180
Figure IV. 9: Spectres obtenus après les analyses HPLC des séquences Lav3 (rouge) et Lav4 (bleu). 181
Figure IV. 10: Courbes obtenues après l'analyse SEC de l'échantillon Lav3. La courbe en noire donne les populations de l'échantillon et la courbe en orange donne les masses molaires calculées. 182

Figure C. 1: Plan général des études qui ont été menées au cours de cette thèse. En haut, séquence type de poly(phosphodiester)s codant de l'information. Bleu : Amélioration de l'alphabet (chapitre II). Rouge : Amélioration de l'espaceur (chapitre III). Vert : Amélioration des marqueurs moléculaires (chapitre IV). 188

Liste des tableaux

Tableau I. 1: Enchaînement des bases nucléiques formant des codons et leur traduction en acides aminés (seuls 20 des acides aminés correspondent à un codon). 36

Tableau I. 2: Liste des groupements protecteurs classiques utilisés pour la chimie de la phosphoramidite. 51
Tableau I. 3: Tableau donnant les temps de dérive des différentes dyades et permettant de déchiffrer la clé de codage.[129] 62
Tableau I. 4: Liste des alphabets composés de monomère de phosphoramidite. 70

Tableau II. 1: Liste des séquences tests effectuées avec les monomères C3-C10 pour l'alphabet linéaire. 90
Tableau II. 2 : Description des séquences plus longues codées en dyades avec l'alphabet linéaire. 92
Tableau II. 3 : Description des polymères codant du texte avec l'alphabet de première génération. 96
Tableau II. 4 : Description des polymères codant de longues séquences synthétisées avec l'alphabet à 4 symboles de seconde génération. 99
Tableau II. 5: Description des premiers essais de synthèse avec l'alphabet à 8 symboles 102
Tableau II. 6 : Table du langage SXBIT. 104
Tableau II. 7 : Description du polymère codant du texte avec l'alphabet triade et le langage sixbit. 104
Tableau II. 8: Caractérisation des oligomères modèles. 105
Tableau II. 9: Efficacité de couplage et rendement obtenus pour les monomères M6 à M8 pour les séquences d'octamères suivis dans le temps. 110
Tableau II. 10 : Efficacité de couplage et rendement obtenus pour le monomère M8 avec les différents protocoles. 112
Tableau II. 11 : Description des polymères numériques codant une image synthétisée avec l'alphabet à 4 symboles. 114
Tableau II. 12 : Rendement des synthèses de polymères codant des images avec l'alphabet à 4 symboles. 114
Tableau II. 13 : Description des sous-séquences synthétisées avec les protocoles a, b et c, menant au polymère **P11**. 118
Tableau II. 14 : Description des polymères codant des images avec l'alphabet à 8 symboles. 121
Tableau II. 15 : Rendement des synthèses de polymères codant des images en triade. 122

Tableau III. 1: Description des polymères synthétisés avec l'espaceur ROSC. 134
Tableau III. 2: Description des séquences tests synthétisées avec l'espaceur RISC. 137
Tableau III. 3: Description des séquences tests synthétisées avec l'espaceur RISCOP. 141
Tableau III. 4: Description de la longue séquence tests synthétisées avec l'espaceur RISCOP. 143
Tableau III. 5: Valeur de m/z attendues pour des segments triplements chargés généré pendant une analyse CID de séquences de poly(phosphodiester)s préparées avec l'espaceur RISCOP. 145
Tableau III. 6: Description des séquences codant des images synthétisées avec l'espaceur RISCOP noté ici « ER ». 147
Tableau III. 7 : Rendement des synthèses de polymères codant des images avec l'espaceur RISCOP. 148
Tableau III. 8: Valeur de m/z attendues pour des segments triplements chargés généré pendant une analyse CID de séquences de poly(phosphodiester)s préparées avec l'espaceur RISCOP et l'alphabet augmenté à 4 symboles. 152
Tableau III. 9: Description des séquences tests synthétisées avec les espaceurs photosensibles PRISC et NISC. 154
Tableau III. 10: Tableau récapitulatif des structures des espaceurs étudiés dans ce chapitre lorsqu'ils sont incorporés dans une chaîne de poly(phosphodiester)s. 159

Tableau IV. 1: Liste des molécules choisies convenant aux critères pour être des marqueurs de masse. 167
Tableau IV. 2: Table de compression Huffman pour l'image de la fiole. 172
Tableau IV. 3: Description des polymères numériques codant une image compressée 178
Tableau IV. 4: Description des polymères numériques codant une image compressée avec l'espaceur RISCOP 179

Introduction générale

Le stockage efficace des données est un des enjeux majeurs de notre époque. Les recherches effectuées à ce sujet sont nombreuses et beaucoup de domaines d'expertises peuvent apporter des solutions. Le travail effectué au cours de cette thèse propose une réponse chimique à ce problème. La solution envisagée est de présenter un stockage d'information à l'échelle moléculaire via la synthèse de polymères codant des données numériques. Cette solution s'inscrivant dans le domaine des sciences chimique, pourrait permettre de faire évoluer les outils de stockage actuellement utilisés.

Avec l'émergence des nouvelles technologies comme les smartphones, les objets connectés ou encore la prolifération des réseaux sociaux, notre demande de stockage de données ne cesse d'augmenter au fil des années. Rien qu'en 2018, 33 zettaoctets d'informations avaient été créés, et d'après plusieurs experts les 175 zettaoctets pourraient être atteints d'ici 2025.[1]

Cette incessante augmentation d'informations générées nécessite de trouver de nouveaux moyens de stockage. Les dispositifs traditionnels comme les disques optiques, les disques durs ou les clés USB ont atteint leurs limites, il faut donc trouver une nouvelle solution pour ne rien perdre. Depuis plusieurs années le « cloud » (ou l'informatique dans les nuages en français) semble être une solution. Via ce cloud les informations générées sont enregistrées par Internet et sont stockées dans des centres de données ou data center en anglais. Ces lieux regroupent les systèmes informatiques d'une ou plusieurs entreprise(s), tels que les ordinateurs centraux, les serveurs ou encore les baies de stockage. On peut donc imaginer plus simplement qu'il s'agit d'un grand nombre de disques durs performants qui vont stocker toutes les informations transmises dans le data center.

Le stockage est alors effectué via des supports électromagnétiques comme les disques durs. Ce dernier est composé d'un ensemble de plateaux circulaires coaxiaux qui sont recouverts d'un matériau magnétique permettant l'enregistrement des données. Une tête de lecture-écriture permet l'écriture et la lecture des informations sur le dispositif. Le tout est enfermé dans une coque étanche qui isole l'appareil de la poussière. Pour encoder l'information, le langage binaire est utilisé car c'est avec lui que les ordinateurs fonctionnent. Il s'agira donc de messages contenant une succession de bits 0 et 1. Lors de l'écriture, la tête de lecture-écriture génère un champ magnétique positif ou négatif qui permet de polariser une petite zone de la surface du disque. Lors de la lecture, en passant sur ces zones polarisées il y a un changement de polarité qui induit un courant dans la tête et qui est ensuite convertie en flux de bits compréhensibles par l'ordinateur.

Les entreprises ayant recours à ce genre de centre de données mettent donc à disposition des comptes sécurisés aux utilisateurs qui peuvent alors télécharger leurs données dans ces centres pour les stocker. Un des exemples les plus connus sont les clouds mis à disposition par Google avec ses « google drive » par exemple. De plus, la plupart des utilisateurs de cloud ne s'en rende même pas compte, car c'est via une connexion préinstallée sur leurs téléphones qu'ils y ont accès. Ainsi, les utilisateurs d'Iphone ont leur photos et vidéos transférées automatiquement sur « Icloud » et donc dans un centre de données d'Apple.

Ces centres de données restent peu connus du grand public et sont généralement construits hors des grandes villes, étant donné l'espace important nécessaire pour leur construction. Généralement ils occupent environ 10 km² mais il en existe des bien plus grands. A ce jour China Telecom, une des sociétés d'opérateurs télécoms la plus grosse au monde, détient le plus grand centre de données au monde. Situé en Chine, ce centre occupe environ 25 km² et consomme énormément d'énergie pour son fonctionnement. En 2018, la consommation de ces centres de données en Chine représentait

2,35% de la consommation totale d'électricité du pays.[2] Cette même année en moyenne 200 térawattheures avaient été consommés pour le fonctionnement global de ces centres de données mondiaux. Ce qui représente 1% de la consommation mondiale d'électricité.[3] Le fonctionnement de ces centres nécessite un apport énergétique pour alimenter les serveurs, mais également et surtout pour les refroidir. En effet, les systèmes informatiques ont besoin d'une chaleur homogène d'environ 20 °C pour un fonctionnement optimal. Ils ont donc besoin d'un système de climatisation pour contrebalancer la chaleur générée par effet Joules. Bien entendu, ce refroidissement a un impact environnemental et contribue directement à 0,3% des émissions totales de carbone.[3, 4]

Des solutions existent pour diminuer la consommation énergétique des centres de données.[5] Cette diminution peut être atteinte en utilisant des technologies plus efficientes qui vont augmenter l'efficacité des centres tout en diminuant leur consommation énergétique. On peut citer par exemple les systèmes de refroidissement utilisant l'air extérieur plutôt qu'un système de climatisation traditionnel. L'air extérieur a toujours besoin d'être filtré et d'avoir une hygroscopie contrôlée mais aucun système de refroidissement n'est utilisé. Toutefois, la localisation des centres peut poser un problème si la température extérieure est trop importante.

Une autre solution pour réduire la consommation d'énergie est de remplacer les systèmes actuels par le stockage de données à l'échelle moléculaire. Plusieurs technologies semblent pouvoir répondre à cette demande.

Des chercheurs de l'université de Manchester ont prouvé qu'il est possible de stocker de l'information sur des aimants monomoléculaires.[6] Ces derniers peuvent servir pour créer un disque dur moléculaire. En effet, en les utilisant à une basse température (-213°C), ces molécules agissent tels des aimants qui sont capables de stocker de l'information à très petite échelle. Natterer et *al*. ont montré qu'il est même possible de lire et écrire des données sur un atome.[7] Cette prouesse est possible en utilisant un atome d'Holmium sur de l'oxyde de magnésium. Configuré ainsi, l'atome possède deux états magnétiques stables, lisible grâce à l'utilisation de la magnétorésistance à effet tunnel.[8] Ils ont pu changer ces états magnétiques en appliquant une grande puissance sur l'atome. Ils ont ainsi montré qu'il est possible d'écrire les 4 états magnétiques 00, 01, 10 et 11, prouvant que le stockage d'information à l'échelle atomique est bien possible.

Les recherches sur le stockage sur des polymères synthétiques se développent également depuis la fin des années 1980. Ce dernier permet de réduire considérablement la place nécessaire pour stocker une large quantité de données. Les exemples les plus connus se trouvent dans le domaine des biopolymères. La nature compte plusieurs exemples de macromolécules contenant de l'information. Que ce soit l'ADN ou l'ARN, ces séquences sont toutes deux porteuses de données. L'ADN stocke toute l'information génétique des êtres vivants et joue donc un rôle d'archive, alors que l'ARN va plutôt accomplir un rôle catalytique. Mais qu'importe leur emploi, elles stockent toutes les deux de l'information à l'échelle moléculaire. Le contrôle de leur séquence est donc très important pour permettre un stockage sans erreur et pour transmettre les bonnes informations.

De nombreux chercheurs ont donc voulu rapidement copier les capacités de ces biopolymères. Depuis la découverte de la structure de l'ADN en 1953, fabriquer de l'ADN synthétique est devenu un domaine à la pointe de la recherche.[9] La synthèse de brins d'ADN étant maîtrisée depuis les années 1980, de nombreuses recherches se portent maintenant sur comment optimiser ce support d'information pour le stockage.[10] Tout comme dans la nature, ce sont les quatre bases de l'ADN qui sont utilisées pour

stocker l'information. Mais pour que l'information soit compréhensible de nos dispositifs informatiques actuels, il faut que celle-ci soit codée en bits. Il faut donc traduire les quatre bases A, C, G et T (Adénine, Cytosine, Guanine et Thymine) en 0 et 1. Pour se faire chaque base nucléique va alors coder pour 2 bits : les dyades 00, 01, 10 et 11. Avec ce procédé, des études ont montré qu'il est possible de coder plusieurs kilooctets de données sur des macromolécules d'ADN.[10-13] Récemment ce type de séquences codant de l'information a été utilisé dans un processus permettant de coder l'information, la stocker et la lire grâce à un seul et même appareil.[14] Ce développement est nécessaire pour espérer un jour pouvoir utiliser le codage à l'échelle moléculaire en dehors des laboratoires et pourquoi pas dans les centres de données à la place des disques durs.

Malgré leur grand attrait, les séquences de biomacromolécules codantes ne sont pas les seules intéressantes pour ce genre d'application. En effet, des polymères abiotiques pourraient être plus optimaux pour des applications non biologiques. Tout en gardant les systèmes biologiques comme modèle, de nombreux systèmes non naturels codant de l'information sont apparus au cours des dernières années.[15] L'idée reste la même, il suffit de choisir un monomère codant pour le bit 0 et un autre codant pour le bit 1 et synthétiser des copolymères contenant ainsi de l'information binaire.[4] A la différence de l'ADN synthétique qui n'a le choix qu'entre quatre monomères différents (*i.e.* les quatre bases nucléiques) une multitude de monomères peut être choisie dans le cas de ces nouveaux polymères dits polymères numériques. Pour permettre un stockage parfait et sans erreur des données contenus sur ces polymères, le plus important est que la séquence soit parfaitement définie et contrôlée. Il faudra donc utiliser une des méthodes de polymérisation contrôlées. Celles-ci sont généralement classées en trois grandes catégories (i) polymérisation par étapes (ii) polymérisation en chaîne et (iii) synthèses multi-étapes.[16] Les deux premières catégories ne permettent pas d'accéder à un contrôle parfait de la séquence, les indices de polymolécularité généralement obtenus sont trop larges pour espérer utiliser ce type de polymérisation dans une application comme le stockage d'information à l'échelle moléculaire.[17]

Les synthèses multi-étapes peuvent, en revanche, permettre un contrôle absolu de la structure des séquences synthétisées. En effet, la synthèse itérative peut aboutir à des séquences définies au monomère près, car chaque unité est attachée une à une à la chaîne en croissance.[16] Ce concept s'est popularisé avec son utilisation par Merrifield à partir de 1963 pour effectuer des synthèses sur support solide de polypeptides.[18] Ce procédé, développé à l'origine pour les acides aminés, peut facilement s'adapter à la synthèse de polymères non naturels, à condition de bien choisir les co-monomères. Il faut qu'ils réagissent seulement entre eux et non sur eux-mêmes pour éviter les homopolymérisations. Dans ces conditions, ce procédé permet facilement de synthétiser des polymères numériques étant donné qu'il est facile de bien contrôler l'ordre d'ajout de chaque monomère 0 ou 1 pour créer le message binaire.

La synthèse sur support solide est donc le processus qui semble être le plus facile à mettre en œuvre pour obtenir des séquences parfaitement contrôlées. Il reste à choisir les monomères codants pour savoir quel type de chimie est la plus efficace pour être utilisée sur support solide. Comme mentionné précédemment, la chimie offre un vaste nombre de monomères possibles et différents types de séquences non naturelles ont donc été testées au cours de ces dernières années.[19-22]

Les monomères de phosphoramidite formant des séquences de poly(phosphodiester)s sont particulièrement intéressants car ils peuvent être impliqués dans une synthèse itérative sur support solide et automatisée. Ces types de monomères sont utilisés dans la synthèse d'ADN synthétique. A la

base, ils sont donc composés d'un sucre et d'une des bases nucléiques (un nucléoside), protégé d'un côté par un groupement protecteur 4-4'-diméthoxytrityle (DMT) et fonctionnalisé de l'autre côté par une phosphoramidite. Cette structure est donnée sur la Figure 1. Pour être impliqué dans le cycle itératif le nucléoside n'est pas nécessaire et peut être remplacé par n'importe quelle espèce chimique (boule bleue sur la figure). Dans un premier temps, notre équipe de chimie macromoléculaire de précision localisée à l'Institut Charles Sadron de Strasbourg, a choisi d'utiliser une simple chaîne propyle linéaire pour le monomère 0 ; pour le monomère 1 la même chaîne propyle est fonctionnalisée par un méthyle en position centrale.[23] Le groupement protecteur DMT et la phosphoramidite sont toujours présents. Avec ces deux nouveaux monomères et la synthèse itérative impliquant la chimie de la phosphoramidite automatisée, des séquences monodisperses contenant de l'information ont pu être synthétisées.[23, 24]

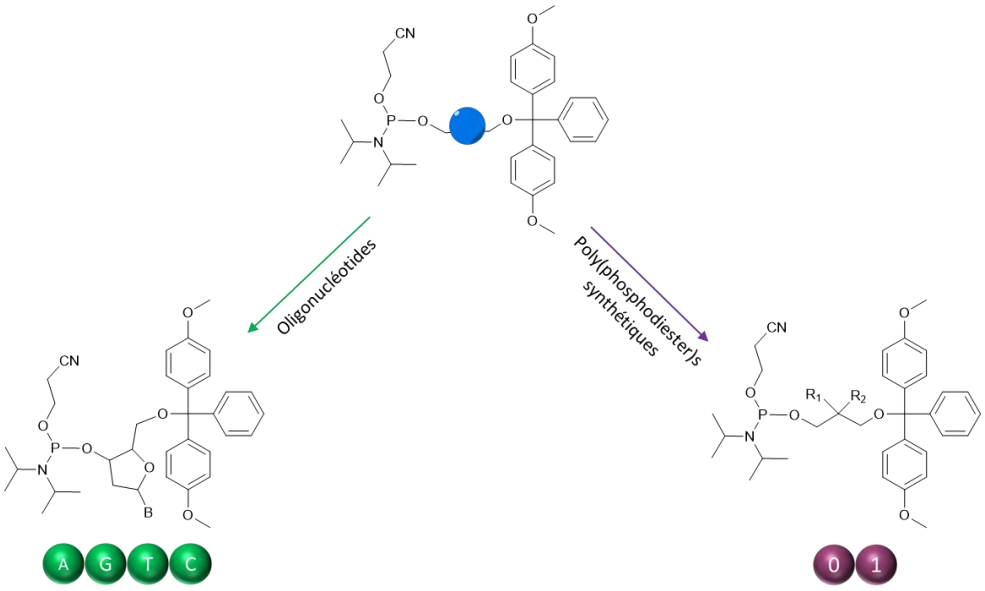


Figure 1: Schéma des structures des monomères menant à des oligonucléotides (à gauche) avec B représentant les bases de l'ADN (Adénine, Guanine, Thymine ou Cytosine) ou à des poly(phosphodiester)s synthétiques (à droite) avec $R_1 = R_2 = H$: monomère 0 et $R_1 = R_2 = CH_3$: monomère 1.

La synthèse et donc le stockage de données, est assez facile à mettre en place mais la récupération des données durant la phase de lecture reste encore très difficile avec ces séquences. La lecture se fait via une analyse de spectrométrie de masse, mais les spectres obtenus sont lourds à déchiffrer, voire indéchiffrables. A partir de 2017, des améliorations sont apportées par notre équipe grâce à l'ajout d'une molécule qui est utilisée comme espaceur et qui va permettre de faciliter la lecture en créant une fragmentation inter-octet.[25]

Cet espaceur, une alcoxyamine, est l'élément le plus fragile de la séquence (représenté en rouge sur la Figure 2). Lors de l'analyse en spectrométrie de masse il va donc être clivé en premier (schématisé par un éclair rouge). Après son clivage, la macromolécule sera découpée en plusieurs segments (généralement des octets) qui seront reconnaissables grâce à des marqueurs de masse différents présents au bout de chaque octet (représentés en vert sur la Figure 2). Grâce à ces marqueurs, la place dans la séquence de chaque octet sera également connue. Ensuite chaque octet sera fragmenté de nouveau, le clivage s'effectuant cette fois entre chaque monomère. La lecture par spectrométrie de masse de chaque fragment va ainsi permettre de recouvrer les bits encodés sur chaque fragment. En

remettant ensemble toutes les données obtenues pour chaque partie, on va alors pouvoir obtenir le message initialement encodé.

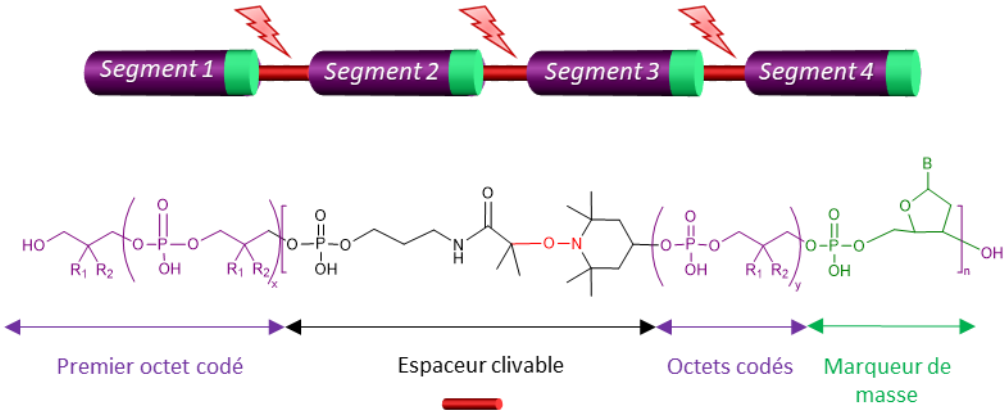


Figure 2: En haut, schéma du design d'une séquence contenant un espaceur (en rouge) et des marqueurs (en vert). En bas, exemple de structure d'une chaîne de poly(phosphodiester)s contenant un espaceur et un marqueur de masse.

Ce nouveau système permet donc de stocker et de lire des messages encodés à l'échelle moléculaire. Néanmoins, des améliorations sont encore possibles. Pour le moment, la lecture des spectres de spectrométrie de masse permettant de recouvrer l'information, s'effectue de manière manuelle. Cette lecture peut être facilitée pour diminuer le temps nécessaire au décodage, et peut même être totalement automatisée. Il est aussi possible d'augmenter la densité de stockage. Pour le moment, seuls des oligomères codants quelques octets d'information ont pu être synthétisés. Il est possible d'atteindre des capacités de stockage plus importantes en utilisant des alphabets augmentés, ainsi que des éléments informatiques permettant de compresser les données.

Pour ce faire, trois axes de recherches peuvent être explorés : (i) l'alphabet, donc les monomères utilisés, (ii) l'espaceur utilisé et (iii) les marqueurs de masse utilisés. Dans le cadre de cette thèse l'amélioration de ce type de séquence a été étudiée. Les trois axes précédemment mentionnés ont été explorés et sont résumés sur la Figure 3.

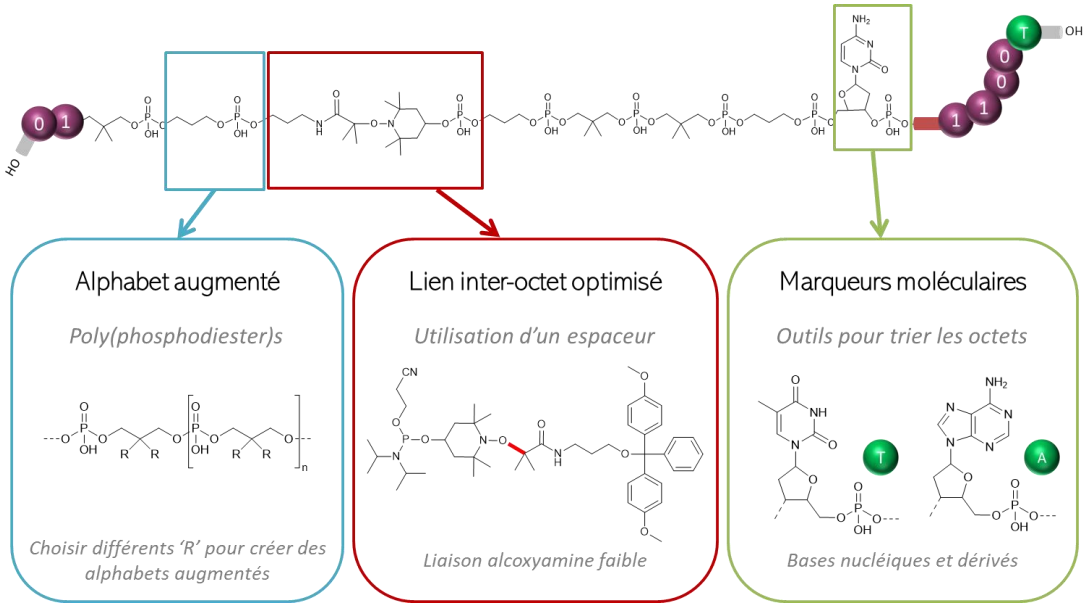


Figure 3: Plan général des études menées. En haut, séquence type de poly(phosphodiester)s codant de l'information. Bleu : Amélioration de l'alphabet (chapitre II). Rouge : Amélioration de l'espaceur (chapitre III). Vert : amélioration des marqueurs moléculaires (chapitre IV).

Dans un premier temps, le **chapitre I** est une introduction sur l'état de l'art du stockage à l'échelle moléculaire. Tout d'abord, les systèmes biologiques codant de l'information sont détaillés ainsi que l'historique de leur synthèse chimique. En particulier, des explications sur la chimie de la phosphoramidite sont données, suivis par des exemples de stockage d'information sur des biopolymères. Les polymères non naturels sont ensuite décrits et leurs applications en tant que polymères numériques est détaillées en deux temps. Des exemples de polymères synthétiques codant de l'information sont donnés, pour finalement se concentrer sur les séquences de poly(phosphodiester)s utilisant la chimie de la phosphoramidites. Par la suite, 3 chapitres expérimentaux sont développés.

Le **chapitre II** décrit le développement de nouveaux alphabets et leur utilisation dans des séquences codées de poly(phosphodiester)s. Plusieurs alphabets augmentés ont été étudiés pour mener à l'élaboration de deux alphabets complexes : le premier est un alphabet à 4 symboles et le deuxième à 8 symboles. Grâce à ce développement, la densité de stockage a été améliorée et il est maintenant possible de stocker une quantité d'information plus importante dans la même taille de polymères. Plusieurs séquences codant du texte mais également des images ont été synthétisées. L'information contenue a été recouvrée parfaitement via l'analyse par spectrométrie de masse.

Le **chapitre III** est consacré au développement d'une fragmentation inter-octet. Il est montré que des réactions secondaires peuvent se produire avec l'espaceur actuellement utilisé. En effet, des réactions parasites se produisent après la création du radical lors de l'analyse de spectrométrie de masse. De nouveaux espaceurs plus stables et ayant un radical moins accessible sont donc testés. Plusieurs séquences tests impliquants ces nouveaux espaceurs sont synthétisées pour confirmer ou non leur possible utilisation dans des séquences plus compliquées de poly(phosphodiester)s.

Le **chapitre IV** expose les améliorations possibles grâce aux marqueurs de masse. Dans ce chapitre, une description du choix des marqueurs est donnée. Leur utilisation dans des séquences ayant une plus grande densité de stockage est détaillée. Pour atteindre de telle capacité de stockage des éléments de compression et de vérification d'erreur ont été utilisés. Ces éléments supplémentaires ne codent certes pas de messages, mais augmentent la robustesse du code.

Cette thèse s'inscrit dans le projet ARN « 00111001 » et a été effectuée au sein de l'équipe de Chimie Macromoléculaire de Précision à l'Institut Charles Sadron de l'Université de Strasbourg sous la direction du Dr. Jean-François Lutz. Au cours de cette thèse plusieurs laboratoires ont été impliqués. Ainsi toutes les analyses de spectrométrie de masse ont été effectuées par l'équipe de la professeure Laurence Charles à l'Institut de Chimie Radicalaire à l'Université d'Aix-Marseille. Les espaceurs ont été fournis par Kévin Launey, doctorant effectuant sa thèse au sein de l'Institut de Chimie Radicalaire de l'Université d'Aix-Marseille sous la direction du Dr. Didier Gigmes. Marc-André Delsuc, directeur de recherche au CNRS rattaché à l'institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) de l'Université de Strasbourg a également été impliqué dans l'utilisation des outils informatiques permettant d'augmenter la capacité de stockage des séquences.

Chapitre I

Stockage d’information à l’échelle moléculaire

1. Introduction

Depuis toujours les êtres humains ont eu besoin de stocker les connaissances qu'ils génèrent. Le papyrus et les parchemins ont été des outils indispensables pour transmettre les savoirs des sociétés passées. Par la suite, l'imprimerie permit de démocratiser la transmission d'information ainsi que de la stocker de façon plus efficace. Depuis, les technologies de stockage ont bien évolué. Jusque dans les années 1960 le stockage d'information était essentiellement fait sur des supports physiques tels que le ruban perforé, mais les supports magnétiques comme la bande magnétique ou la disquette l'ont remplacé à cette époque.[26] De nos jours, ce stockage se fait grâce à des supports d'information électroniques ou électromagnétiques comme les disques durs ou les clés USB. C'est en 1956 que IBM présente le tout premier disque dur. Il a une capacité de stockage de 5 Mo, ce qui représente environ la taille d'un fichier MP3. Depuis la capacité de stockage n'a cessé d'augmenter. Hitachi dépasse le Giga en 1982 et de nos jours les 16 Téra ont été atteints par Seagate.

Cette augmentation phénoménale de capacité a pu être possible grâce à la découverte en 1988 de la magnétorésistance géante par deux équipes, celle d'Albert Fert[27] et de Peter Grünberg.[28] Ils reçoivent le prix Nobel de physique en 2007 pour leurs travaux. Déjà en 1989, une équipe d'IBM s'intéresse à cette découverte qui permet de détecter des champs magnétiques très faibles et donc de lire des données à petite échelle. L'entreprise utilise la magnétorésistance géante pour réaliser des nouveaux capteurs de champ magnétique, ce qui leur permet de mettre au point un nouveau type de tête de lecture utilisé dans les disques durs des ordinateurs. Ce nouveau dispositif est commercialisé en 1997.

Dans ce contexte, de plus en plus d'équipements de stockage, tels que les disques durs magnétiques qui sont basés sur le cryptage des données à une échelle microscopique, montrent leurs limites. Beaucoup de recherches sont menées pour pallier ce problème. Une solution proposée qui semble être prometteuse est le stockage des données à l'échelle moléculaire. Cette solution permettrait de gagner de l'espace, de l'énergie et voire même d'augmenter la densité de stockage. Comme il a été montré dans **l'introduction générale**, les systèmes biologiques sont de bons exemples sur lesquels les scientifiques se sont basés pour créer des systèmes synthétiques de stockage d'information. Ci-après sont détaillées leurs démarches, dans un premier temps via la synthèse d'ADN synthétiques puis par la synthèse de séquences codantes non naturelles.

2. Le stockage naturel d'information à l'échelle moléculaire dans les systèmes biologiques

2.1. L'acide désoxyribonucléique (ADN)

2.1.1. Structure de l'ADN

C'est dans la nature que l'on trouve l'exemple le plus courant de stockage d'information sur un polymère biologique. En effet, l'acide désoxyribonucléique permet de stocker toute l'information génétique des systèmes vivants et est présent dans toutes les cellules du monde vivant. C'est la clé pour comprendre le développement, la reproduction ou encore le fonctionnement de toutes les espèces vivantes. Cette biomacromolécule est formée d'une chaîne principale faite d'unités répétitives de sucres et de

phosphates. Sur chaque sucre se trouve une base nucléique ; l'ensemble forme alors un nucléoside. Les bases nucléiques sont au nombre de quatre et permettent le codage de l'information génétique grâce à leur enchaînement défini. On les appelle la cytosine (C) et la thymine (T) (les pyrimidines) et la guanine (G) et l'adénine (A) (les purines). L'ADN est une des biomacromolécules ayant une séquence contrôlée avec la plus haute capacité de stockage.[9]

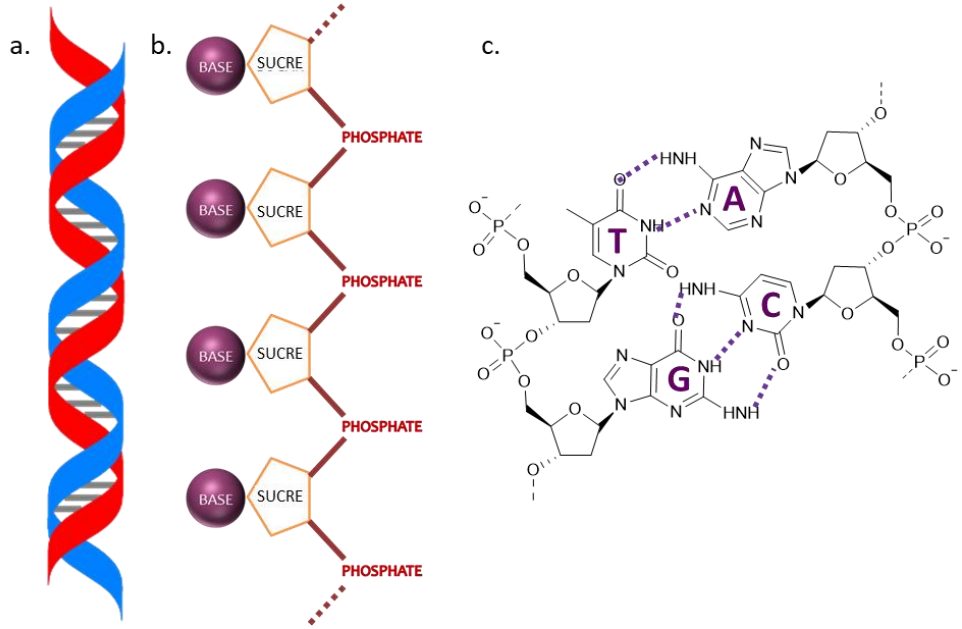


Figure I. 1 : (a) Représentation schématique de la structure hélicoïdale de l'ADN. Les brins rouge et bleu représentent la chaine phosphate et les lignes horizontales représentent les liaisons hydrogènes entre les paires de bases. (b) Représentation schématique de la formule chimique d'une simple chaîne. (c) Représentation des deux paires de bases de l'ADN et de leurs liaisons hydrogènes.

Les acides nucléiques sont isolés pour la première fois en 1869 par Friedrich Miescher, un scientifique suisse. Il découvre dans le noyau de cellules un composé riche en phosphate qu'il appelle la nucléine.[29] Il faut attendre 1944 pour que la structure de l'ADN soit vue comme vecteur d'information génétique avec les travaux d'O.T. Avery et *al*.[30] Leurs travaux, montrant le premier spectre de diffraction au rayon X de l'ADN, sont la base sur laquelle Maurice Wilkins et Rosalind Franklin travaillent.[31] Les travaux de ces derniers ont finalement permis à Watson et Crick de décrire en 1953 la structure de l'ADN. Ils montrent la relation entre sa structure primaire et sa structure secondaire qui génère l'organisation des chaînes dans l'espace.[9] Ils montrent que la structure de l'ADN est une double hélice, composée de deux brins de polynucléotides hélicoïdaux qui s'enroulent l'un autour de l'autre. Grâce à cette découverte ils obtiennent le prix Nobel en Physiologie ou Médecine en 1962 avec M. Wilkins. Cette structure hélicoïdale est due aux liaisons hydrogènes entre les paires de bases : l'adénine avec la thymine sont reliées par deux liaisons et la guanine avec la cytosine en ont trois comme le montre la Figure I. 1.c. De par ces interactions, les bases se trouvent alors vers l'intérieur de la double hélice et les groupements sucres et phosphates sont repoussés vers l'extérieur.[9] Grâce à ces interactions la structure en double hélice est assez forte. Cette robustesse augmente avec le nombre d'interactions. Il est donc logique de voir que les deux brins sont

complémentaires. Néanmoins ces liaisons hydrogènes restent plus faibles que des liaisons covalentes et sont donc réversibles, ce qui peut être un réel avantage pour plusieurs mécanismes biologiques.

2.1.2. Réplication : la biosynthèse de l'ADN

La synthèse de nouveaux brins d'ADN, se fait lors de la réplication.[32] Ce processus permet d'obtenir un nouveau brin d'ADN à l'identique de façon fiable et rapide. La réplication est dite semi-conservative, puisqu'elle est faite à partir d'un brin d'ADN déjà existant que l'on appelle le brin parent et qui va servir de matrice pour la réplication. Le brin nouvellement formé est ainsi composé d'un brin parental et d'un nouveau brin.[32] Les origines de réplications sont un enchaînement spécifique de nucléotides qui sont reconnus par les protéines de réplication. A partir de ces origines, les enzymes séparent les deux chaînes créant ainsi une fourche. L'ADN polymérase se fixe alors sur cette fourche en catalysant la formation des nouvelles liaisons nucléotidiques. Les nouveaux nucléotides créés sont donc les complémentaires de ceux présents sur le brin parental. La fiabilité de la reproduction est permise grâce à la propriété de relecture dont dispose l'ADN polymérase. En effet, celle-ci vérifie que le dernier nucléotide inséré est bien le bon et si ce n'est pas le cas elle peut revenir pour le corriger. Pour ce faire, une hydrolyse du mauvais nucléotide a lieu qui permet de l'éliminer et le bon nucléotide peut alors le remplacer.

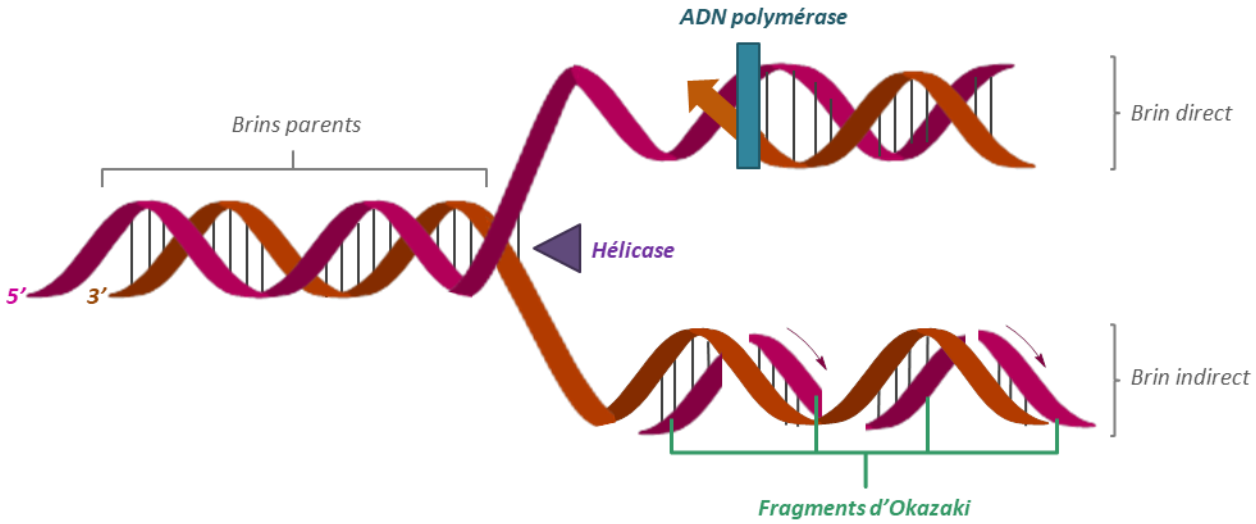


Figure I. 2: Schéma de la réplication semi-conservative de l'ADN.

Cette réplication se fait toujours dans le même sens : depuis l'extrémité 5' qui est le phosphate relié au cinquième carbone du sucre final vers le bout 3' qui est l'hydroxyle relié au troisième carbone du sucre de la chaîne. Ce processus implique donc qu'il y a un brin direct et un brin indirect. Le brin direct est celui qui est créé en suivant le même sens que la direction de l'ADN polymérase donc de façon continue. A l'inverse, le brin indirect est créé dans le sens inverse donc de façon discontinue. A cause de cette discontinuité dans la synthèse, le brin indirect est en fait formé de plusieurs petits fragments qui sont assemblés à la fin. On parle de fragments d'Okazaki, du nom de leur découvreur.[33] La dernière étape est alors l'assemblage des fragments qui se fait par un autre complexe enzymatique appelé l'ADN ligase. Toutes ces étapes sont représentées en Figure I. 2.

2.1.3. Dénaturation de l'ADN

Une des propriétés importantes que l'on trouve chez l'ADN est la capacité de réversibilité des liaisons hydrogènes formées entre ses paires. En effet, ces liaisons n'étant pas aussi fortes que des liaisons covalentes, il est possible de les faire et les défaire, comme on a pu le voir lors de la biosynthèse de nouveaux brins. Le processus de dénaturation ou fonte de l'ADN est donc le passage du complexe hélicoïdal formé de deux brins, à deux brins seuls en rompant les liaisons hydrogènes.[34] Pour mettre en place ce processus, il faut apporter de l'énergie au complexe. Dans les cellules cette énergie est apportée par la protéine hélicase. Au laboratoire, une augmentation de la température permet d'initier ce phénomène. Il est donc possible de mesurer une température de fusion T_m, pour laquelle la moitié des liaisons hydrogènes sont rompues. A contrario, lorsque la température diminue, le phénomène inverse est observé et on retrouve la structure hélicoïdale formée des deux brins d'ADN grâce aux liaisons hydrogènes reformées. Ce phénomène s'appelle la renaturation.

2.2. L'acide ribonucléique (ARN)

2.2.1. Structure de l'ARN

L'acide ribonucléique est lui aussi un polymère permettant de stocker une grande quantité d'information, mais il n'est pas utilisé pour une propriété d'archivage comme peut l'être l'ADN. Il est plutôt utilisé comme un vecteur permettant d'exporter l'information en dehors du noyau.[35] Sa structure ressemble beaucoup à celle de l'ADN, et est donc également constitué d'une succession de nucléotides, composée donc d'un phosphate, d'un sucre (ici le ribose) et d'une base nucléique.[36] Concernant les quatre bases nucléiques, elles ne sont pas toutes les mêmes que pour l'ADN. En effet, l'adénine, la guanine et la cytosine sont également présentes mais la thymine est remplacée par l'uracile. Les propriétés d'appariement avec l'adénine restent inchangées. L'ARN a moins tendance à former des complexes entre séquences complémentaires, il se retrouve donc dans les cellules sous forme d'un simple brin et non d'une double-hélice comme l'ADN. Des structures tridimensionnelles se forment tout de même, mais les appariements se font entre les bases d'une même chaîne macromoléculaire. Les structures assez complexes ainsi formées vont permettent l'utilisation des chaînes d'ARN dans des tâches où la structure en double-hélice de l'ADN aurait été un obstacle.

2.2.2. Biosynthèse de l'ARN

La synthèse de l'ARN se fait à partir de l'ADN, on parle de transcription.[37] L'enzyme ARN polymérase commence la transcription en détectant une séquence spécifique sur l'ADN, appelée le promoteur. Une fois le promoteur détecté, l'enzyme sépare sur une courte distance les deux brins d'ADN, ce qui lui permet d'utiliser un des deux brins comme matrice pour commencer la transcription de celui-ci. Par la suite, elle apporte les nucléotides complémentaires nécessaires, le tout en catalysant la formation des liaisons phosphodiesters. L'enzyme répète cette opération en se déplaçant le long de la chaîne jusqu'à détecter un message indiquant la fin de la transcription. A la lecture de ce message, la nouvelle chaîne d'ARN ainsi synthétisée est détachée de la chaîne d'ADN matrice, et cette dernière retrouve sa forme initiale de double-hélice avec l'autre brin d'ADN.

2.3. *Les protéines*

2.3.1. Structure des protéines

Dans toutes les cellules vivantes se trouvent des macromolécules à séquences définies que l'on nomme les protéines, polymères essentiels à leur bon fonctionnement.[36] Ces biomacromolécules sont constituées d'une succession d'acides aminés liés par des liaisons peptidiques.[38] Ces acides aminés sont au nombre de 22, procurant ainsi une grande diversité à cette classe de macromolécules. Tous les acides aminés ont une structure commune (à l'exception de la proline) et sont composés d'un acide carboxylique, d'une amine primaire et d'une chaîne latérale. L'enchaînement spécifique des acides aminés que l'on appelle *structure primaire* (Figure I. 3.a), généré par des gènes spécifiques, induit une *structure secondaire* en donnant une conformation spécifique aux chaînes polypeptides. Par exemple, les chaînes peuvent être retrouvées sous formes d'hélice alpha ou de feuillet beta, comme montré sur la Figure I. 3.b. Les chaînes sous leurs différentes conformations peuvent avoir des interactions intramoléculaires entre les différents acides aminés et donc avoir un impact sur la structure tridimensionnelle de la protéine, et ainsi que sur les propriétés chimiques de cette dernière. On appelle alors cette structure la *structure tertiaire*, voir Figure I. 3.c. On distingue également une *structure quaternaire* lorsque les protéines sont composées de plusieurs chaînes de polypeptides qui ont ce même genre d'interaction entre elles (Figure I. 3.d).

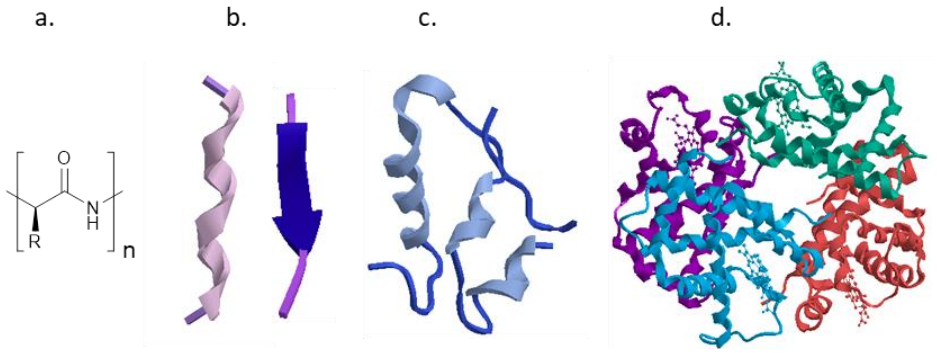


Figure I. 3 : Représentation des quatre structures des protéines. (a) Structure primaire : structure d'un acide aminé. (b) Structure secondaire : Conformation spécifique des chaînes polypeptides (gauche : alpha hélice, droite : feuillet beta). (c) Structure tertiaire : Conformation dans l'espace (exemple de l'insuline). (d) Structure quaternaire : Conformation dans l'espace d'une protéine formée de plusieurs sous unités (exemple de l'hémoglobine A humaine).

Les structures complexes de protéines confèrent de nombreuses propriétés physico-chimiques à ces macromolécules. De par cette diversité de structures on retrouve les protéines dans la plupart des processus biologiques.[35] Un exemple courant est le rôle des protéines en tant qu'enzyme, c'est-à-dire de catalyseur d'une réaction chimique. Ce rôle se retrouve dans la synthèse de l'ADN par exemple.[39] En outre, des protéines structurelles peuvent également agir sur la rigidité de composants biologiques qui, sans elles, seraient fluides. C'est par exemple le cas pour le collagène et l'élastine qui sont les composants principaux du cartilage, ou encore de la kératine composant principal des poils et ongles. Il existe également des protéines qui vont induire des forces mécaniques, comme pour la contraction musculaire.[37] Grâce à la diversité des séquences possibles due aux nombres importants des différents acides aminés, les protéines agissent sur une immense variété de fonctions dans les systèmes biologiques.

2.3.2. Biosynthèse des protéines

La biosynthèse des protéines s'effectue en plusieurs étapes impliquant plusieurs acteurs. La première étape est *la transcription*. Lors de cette étape un gène d'ADN est transcrit en une molécule d'ARN messager (ARNm), remplaçant ainsi les thymines par des uraciles.[37]

Ensuite l'ARN messager peut être traduit, on parle de *la traduction*. Lors de cette étape le brin d'ARNm se lie à un ribosome.[35] Les ribosomes permettent de traduire les codons présents dans le brin d'ARNm en un acide aminé précis. Un codon représente trois bases nucléiques, et chaque combinaison différente de codon correspond à un acide aminé spécifique, comme on peut le voir dans le Tableau I. 1. Grâce à la traduction de chaque codon, une séquence d'acides aminés spécifiques forme ainsi une protéine. C'est grâce à un autre ARN que les acides aminés sont transportés pour former la nouvelle protéine, il s'agit de l'ARN de transfert (ARNt). Il existe un ARNt spécifique pour chaque codon. Ainsi lorsque le ribosome traduit un codon de l'ARNm, l'ARNt homologue est reconnu et peut apporter le bon acide aminé. Il existe également des codons 'STOP'. Il s'agit d'un enchaînement spécifique indiquant la fin de la traduction, impliquant ainsi que l'ARNt n'apportera pas de nouvel acide aminé, terminant donc la synthèse de la chaîne peptidique.

Tableau I. 1: Enchaînement des bases nucléiques formant des codons et leur traduction en acides aminés (seuls 20 des acides aminés correspondent à un codon).

1ère base	2ème base								3ème base
	U		C		A		G		
U	UUU	Phénilalanine	UCU	Sérine	UAU	Tyrosine	UGU	Cystéine	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	Tryptophane	G
C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine	ACU	Thréonine	AAU	Aspargine	AGU	Sérine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine	AGA	Arginine	A
	AUG	Méthionine	ACG		AAG		AGG		G
G	GUU	Valine	GCU	Alanine	GAU	Acide aspartique	GGU	Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Acide glutamique	GGA		A
	GUG		GCG		GAG		GGG		G

2.3.3. La synthèse chimique des protéines

L'intérêt des protéines est tel que rapidement les chercheurs ont essayé de synthétiser en laboratoire ce genre de macromolécule. Le premier exemple date de 1903 lorsqu'Emile Fischer synthétise pour la première fois un tripeptide, le Gly-Gly-Gly.[40] Cette synthèse est réalisée en solution et jusque dans les années 60 de nombreux développements lui ont été consacrés. Mais la nécessité d'avoir des groupements protecteurs et le temps nécessaire à chaque ajout rend cette méthode fastidieuse.

En 1963, R. Bruce Merrified vient révolutionner les recherches en proposant non plus une réaction en solution, mais une réaction sur support solide. On appelle cette nouvelle méthode la synthèse peptidique sur support solide (SPPS en anglais pour *solid phase peptide synthesis*).[18] Elle consiste à faire croitre une chaîne de peptides à partir d'un support solide insoluble dans les solvants utilisés. Les étapes d'ajout de nouveaux monomères avec l'utilisation de groupements protecteurs sont toujours les mêmes, mais les étapes de purifications sont largement simplifiées. En effet, il suffit désormais de réaliser une filtration pour évacuer les excès de réactifs et les sous-produits, et non plus une étape de recristallisation après chaque ajout. La chaîne étant attachée au support solide elle ne part pas lors de la filtration, le support étant d'une taille trop importante pour passer à travers les pores. Cette nouvelle méthode est schématisée en Figure I. 4.

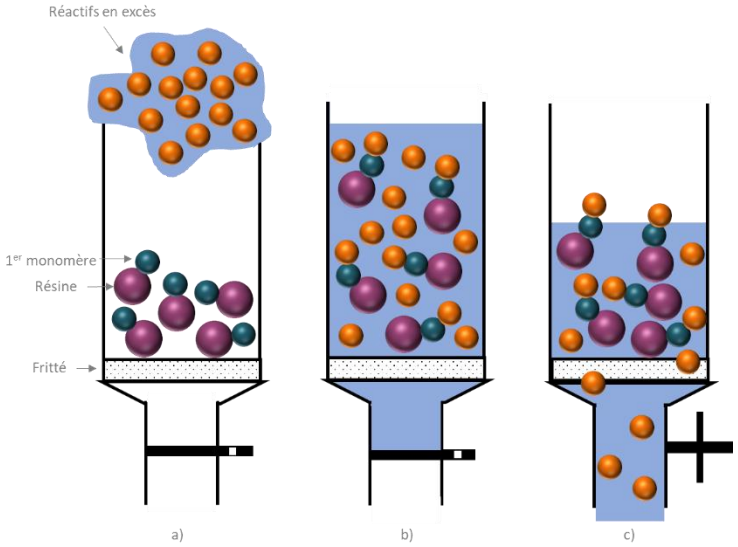


Figure I. 4: Protocole générale de la méthode de synthèse peptidique sur support solide. (a) Ajout des réactifs en excès en solution. (b) Couplage. (c) Filtration et lavages avec les solvants adéquats.

L'automatisation de cette méthode réduit encore plus le temps nécessaire pour synthétiser un polypeptide.[41] Mais de nouveaux challenges sont apparus en même temps. En effet, l'utilisation d'un support solide implique une accessibilité des sites actifs réduites et différentes selon les solvants utilisés. Cela tend à provoquer des erreurs de synthèse, notamment l'obtention de séquences possédant des monomères en moins, qui sont très difficiles à séparer de la séquence désirée. Néanmoins, les chercheurs ont continué à tester de nouvelles conditions pour améliorer cette méthode. De nos jours les instruments utilisés, les supports solides, et le choix des réactifs se sont améliorés. Les résultats obtenus sont très bons.[42, 43]

3. Synthèse chimique d'oligonucléotides

3.1. Historique de l'optimisation de la synthèse de l'ADN

En 1953, lorsque Watson et Crick décrivent la relation entre la structure primaire de l'ADN (la séquence des nucléotides) et sa structure secondaire générant l'organisation des chaînes dans l'espace, c'est le début d'une longue série de travaux vers la synthèse en laboratoire des premiers brins d'ADN synthétiques.[9]

3.1.1. Synthèse du premier dinucléotide

La synthèse de la liaison phosphodiester entre la fonction 3'-hydroxy d'un monomère et la fonction 5'-hydroxy d'un second est reconnue comme étant le meilleur moyen pour réussir à créer un polynucléotide. En 1955, Alexander R. Todd est le premier à réussir la synthèse d'un dinucléotide, un dinucléotide de dithymidine contenant la liaison 3'-5' entre les nucléotides.[44] Il montre qu'il est possible de créer la liaison phosphodiester entre deux déoxythymidines protégées, et les analyses confirment bien la structure 3'-5' de la liaison phosphodiester. A.R. Todd procède en deux étapes : premièrement une déoxythymidine avec la fonction 5'-OAc protégée et une fonction phosphorochlorodate en position 3' est condensée avec une autre déoxythymidine dont le fonction 3'-OAc est protégée, permettant d'obtenir la liaison phosphodiester protégée (Figure I. 5). Dans une seconde étape, la liaison est déprotégée conduisant ainsi à la création du premier dinucléotide synthétique.

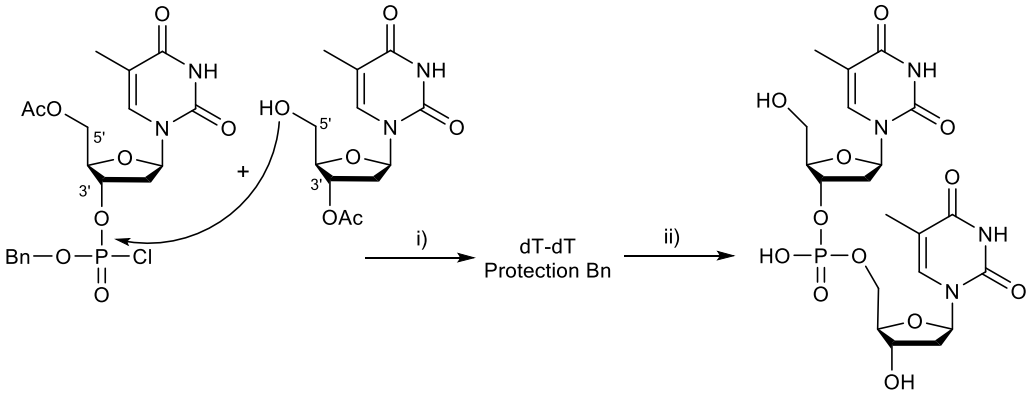


Figure I. 5: Réaction suivie par Todd pour obtenir le premier di-nucléotide synthétique. i) Couplage, 2,6-lutidine ii) Hydrolyse.

3.1.2. Méthode phosphodiester

A la suite de cette réussite, plusieurs recherches sont menées pour trouver de nouvelles voies de synthèses plus efficaces. En 1956, H. G. Khorana fait mention d'une nouvelle méthode de liaison, plus tard appelée la méthode phosphodiester. Il montre comment former un pont ester entre un groupement phosphate activé d'un premier nucléotide et un groupement hydroxyle d'un autre nucléotide. Le point clé est l'activation du groupement phosphate. Celle-ci se fait d'abord par l'acide p-toluène-sulfonique (TsOH), puis il montre que l'utilisation du N,N'-dicyclohexylcarbodiimide (DCC) permet d'atteindre de meilleurs rendements.[45, 46]Le mono-ester contenant le groupement 3'-phosphate est donc activé par le DCC. Ensuite, le groupement primaire 5'-hydroxy plus actif réagit par une attaque nucléophile, menant ainsi à la formation de la liaison phosphodiester entre le 5'-OH d'un nucléotide et le 3'-OH du suivant.

Des difficultés apparaissent lors de synthèses de plus longues chaînes, provenant de la non-protection de la liaison phosphate créée lors du précédent couplage. En effet, des branchements peuvent se former au niveau du phosphate de ce lien inter-nucléotidique. Il est donc nécessaire d'effectuer des étapes de purification longues et difficiles pour les retirer. Bien entendu, plus la longueur de la chaîne d'oligonucléotide augmente plus il y a de branchements créés, augmentant ainsi le temps de purification et la rendant de plus en plus complexe et chronophage.

Des groupements protecteurs sont alors testés pour éviter ces branchements au niveau du phosphates mais également pour d'autres sites comme l'alcool 5'. Pour ce dernier, différents groupes protecteurs sont testés dont le trityle, le 4-méthoxytrityle et le 4-4'-diméthoxytrityle (DMT).[47] Ces groupements sont facilement clivables et de plus, pour le groupement DMT, le clivage est obtenu dans conditions douces (pH 4-5) ce qui est un critère important étant donné la susceptibilité des bases nucléiques à être clivées dans des conditions plus acides. De plus, le carbocation issue du DMT a une couleur orange en solution et montre une absorbance caractéristique qui s'avère être un moyen de quantifier le rendement de l'étape de clivage.[48] L'efficacité de chaque cycle peut ainsi être calculée en mesurant l'absorbance de la solution du DMT clivé et en la comparant avec celle des étapes précédentes. C'est donc avec le DMT que les résultats les plus intéressants sont obtenus.

De plus, les amines présentes sur la cytosine et les purines sont elles aussi protégées : le *N*-benzoyle est utilisé pour protéger l'adénine, la guanine est protégé par la *N*-isobutyryle et la cytosine par un dérivé de la *N*-anisoyle.[49]

La méthode phosphodiester consiste alors à faire réagir sous activation du DCC le 5'-alcool du monomère ayant le 3'-hydroxy protégé, avec un nouveau monomère protégé par le DMT lors de la première étape. Dans une seconde étape la déprotection du DMT est effectuée dans des conditions acides douces, régénérant l'alcool primaire 5'. Le cycle comprenant l'addition du monomère (1ère étape), la déprotection du DMT (2ème étape) et les purifications (3ème étape) est effectué le nombre de fois nécessaire jusqu'à l'obtention de la séquence voulue. L'oligonucléotide final est obtenu après le clivage des groupements protecteurs dans des conditions basiques.[47, 50]

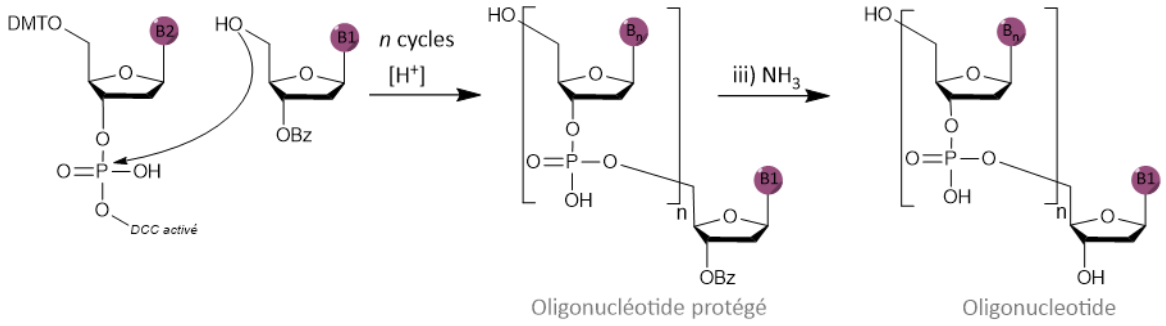


Figure I. 6: Mécanisme de la méthode phosphodiester utilisée pour la synthèse d'oligonucléotides.

Avec ces nouveaux groupements protecteurs, il est donc possible d'utiliser cette méthode phosphodiester pour synthétiser n'importe quelle séquence d'ADN ou d'ARN. Le groupe de Khorana réussit ainsi à déchiffrer le code génétique, puis parvint à la synthèse du premier gène artificiel en 1970.[49, 51-54]

3.1.3. Développement de la synthèse sur support solide

Inspiré par les travaux de Khorana sur les groupements protecteurs et par analogie aux recherches de Merrifield sur la synthèse peptidique sur support solide, Letsinger s'intéresse à la synthèse d'oligonucléotides sur support solide dans les années 1963. Son groupe montre que la synthèse d'oligonucléotides sur support solide assure une voie de synthèse plus rapide.[55, 56]

Son point de départ est l'idée que si la macromolécule est retenue à un support lors de la synthèse, la réaction est plus simple. Cette réaction suit trois étapes simples :

1. Le nucléotide est attaché de façon covalente au support.
2. Les nucléotides suivants sont ajoutés au premier nucléotide lié au support en suivant une procédure en étape.
3. La liaison covalente entre le premier nucléotide et le support est clivée et la séquence entière est retrouvée.

Cette technique simplifie énormément les étapes de purifications. Le produit est lié au support alors que les réactifs en excès et les sous-produits solubles sont séparés par une simple filtration. Pour être viable, le support doit être (i) insoluble dans les solvants utilisés et inertes face au réactifs utilisés, (ii) contenir un groupement fonctionnel sur lequel il est possible de lier le premier nucléotide et également clivable pour obtenir la séquence finale. Et finalement (iii) une structure qui n'est pas un obstacle pour la synthèse et qui permet une bonne diffusion des réactifs.

Le premier essai est effectué avec une résine polymère de styrène divinyle benzène (cf. Figure I. 7.a) comme un support solide. Un dipeptide est préparé et clivé de ce support.[55] L'équipe de Letsinger réussi par la suite à synthétiser des oligonucléotides.[57] Pour ce faire, le premier nucléotide, la 5-O-Trityle-déoxycitidine, est liée au support via un groupement amine, ce support est spécifique à la cytidine.[56] L'addition des nucléotides suivants est réalisée en deux étapes (i) une phosphorylation et (ii) la condensation du dérivé phosphoryle avec le nucléotide suivant. Concernant le clivage du support, il est fait par une hydrolyse de l'amine.

Melby et Strobach démontrent qu'il est possible de synthétiser des oligonucléotides en utilisant un copolymère de styrène divinyle benzène contenant le 4-methoxitrityle comme groupement fonctionnel utilisé pour faire le lien avec le nucléotide en formant une liaison trityle-éther, comme montré sur la Figure I. 7.b.[58] Ce type de liaison entre le support et le nucléotide est aussi utilisé par Khorana et *al.* et également par Cramer et *al.* mais en utilisant un support soluble dans la pyridine.[59, 60]

Un autre type de support également étudié est la résine polymère de styrène carboxylé qui réagit avec un nucléotide ayant sa position 5' protégée par le mono-méthoxytrityle, formant ainsi une liaison ester avec la position 3'-terminale. La structure de ce support est donnée en Figure I. 7.c.[61] Ce nouveau type de liaison tolère les conditions nécessaires à la formation d'oligonucléotides dérivées de la déoxythymidine. Ainsi un trimère, le thymidylyl-thymidylyl-thymidine, est synthétisé. La synthèse montre des rendements meilleurs que pour la synthèse effectuée avec le support solide de Melby et Strobach, mais également des meilleurs rendements que pour les synthèses faites avec le support solide bâti à partir du trityle-déoxycytidine. Il montre également de meilleurs résultats que ceux obtenus avec les supports solubles de

Cramer et ses collègues. Le système en solution de Khorana obtient toutefois toujours de meilleurs résultats.

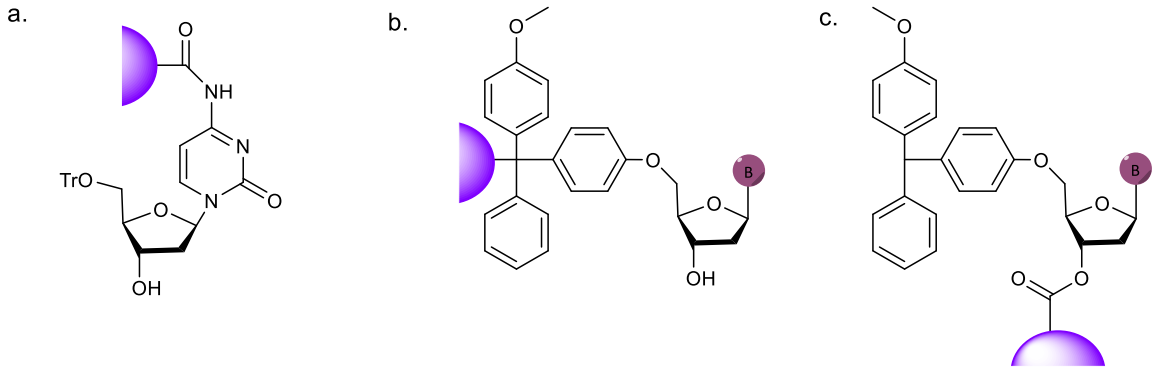


Figure I. 7: Supports solides pour la synthèse d'oligonucléotides. (a) Support cytidine. (b) Support mono-méthoxytrityle (MMT). (c) Support 3'-ester.

Grâce aux travaux réalisés sur les supports solides par Letsinger et ses collègues, il est possible de synthétiser plus facilement de courts oligomères. Néanmoins, les réactions parasites créant des hyper-branchements apparaissent toujours. Les rendements sont tout de même assez bas dans chacun des travaux présentés. De nouveaux développements sont donc nécessaires pour avoir un procédé plus efficace.

3.1.4. Méthode phosphotriester

Alors qu'il effectue ses recherches sur les supports solides, Letsinger travaille déjà sur les problèmes liés aux réactions secondaires. Eviter les hyper-branchements est la clé pour avoir une réaction plus rapide et plus efficace. Pour ce faire, la jonction inter-nucléotides doit être protégée. Letzinger utilise donc le phosphate 2-cyanoéthyle activé par le DCC. Il montre que l'utilisation de ce phosphate permet d'éviter des réactions parasites.[56, 57] Ces expériences sont réalisées en utilisant la première génération de support solide, celui issu de la cytidine. La procédure d'addition se divise en trois étapes, décrites en Figure I. 8.

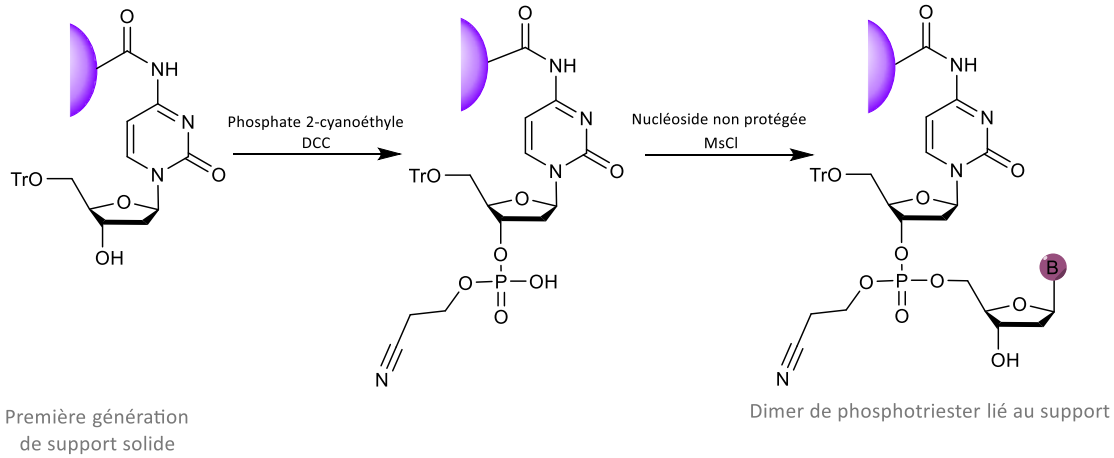


Figure I. 8: Synthèse d'oligonucléotides avec la méthode phosphotriester.

Premièrement (i) le support-cytidine réagit avec le phosphate 2-cyanoéthyle activé par le DCC, puis (ii) le phosphodiester ainsi formé est activé par le chlorure de mesitylènesulfonyle pour avoir un meilleur groupe partant comparé à la fonction hydroxyle. Finalement, (iii) la procédure est achevée par l'ajout du nucléotide voulu, créant un phosphotriester non réactif. La méthode est appelée « méthode phosphotriester » pour cette raison. Les trois étapes sont répétées le nombre de fois nécessaire pour obtenir la séquence désirée.

Le groupement protecteur 2-cyanoéthyle est retiré par un traitement avec une base faible et la séquence est clivée du support par une base plus forte. Les premiers résultats ne sont pas très probants et ne montrent pas de très bons rendements, mais Letsinger continue ses recherches et prouve dans la fin des années 1960 qu'il est possible d'augmenter ces rendements et même de synthétiser des séquences d'oligonucléotides d'ADN et d'ARN. En effet, il montre d'abord que la méthode phosphotriester peut s'étendre à tous les désoxyribonucléosides, puis il développe des groupements protecteurs pour chaque nucléoside, pour finalement montrer qu'il est possible de synthétiser des oligoribonucléotides.[62-64]

3.1.5. Méthode phosphite triester

Malgré toutes les recherches effectuées pour développer la méthode phosphotriester, cette procédure reste tout de même peu efficace pour la synthèse de longues chaînes d'oligonucléotides. C'est pourquoi de nouvelles approches sont étudiées dans le but de trouver un nouveau type de liaison inter-nucléotides menant à de meilleurs rendements. Les liens phosphites, obtenus par des réactions rapides avec de bons rendements, semblent être des intermédiaires intéressants.[65, 66] Le groupe de Letsinger décrit de nouvelles conditions menant à des meilleurs rendements ; il montre que la phosphorodichloridite a une réactivité significative dans le tétrahydrofurane (THF) à basse température vis-à-vis des alcools et qu'une simple étape d'oxydation permet d'obtenir le phosphotriester à partir de cette phosphite dans l'iode et l'eau.[65] En comparaison, les temps de réaction avec les phosphorochloridates nécessitent plusieurs heures, alors qu'avec ces nouvelles conditions la réaction est achevée en quelques minutes. En utilisant cette méthode, il réussit à synthétiser un penta-nucléotide avec un bon rendement (69%) en utilisant le trichloréthyle phosphotriesters et le phosphite correspondant.[66]

Quelques années plus tard, Caruthers (un ancien étudiant de Letsinger) publie un article démontrant qu'il est possible d'avoir encore de meilleurs résultats en changeant le support solide sur lequel la réaction se fait.[67] Il utilise comme support une colonne HPLC de silice (aussi appelé verre poreux ou « controlled pore glass » (CPG) en anglais). Celle-ci n'absorbe pas les synthons et les autres réactifs, contrairement aux supports organiques. Le premier nucléotide (habituellement la thymine) est lié au support via une liaison ester entre le 3'-OH du nucléotide et la fonction acide carboxylique du support. Pour tester ce support, il synthétise une séquence test de neuf thymidines. La synthèse, décrite en Figure I. 9 est réalisée en quatre étapes :

i) Création de la liaison phosphite entre les nucléotides.
ii) Oxydation du phosphite en phosphate.
iii) Capping : Le phénylisocyanate dans le THF et la 2,6 lutine sont utilisés pour réagir avec le groupement hydroxyle d'un nucléoside non phosphorylé.
iv) Lavage.

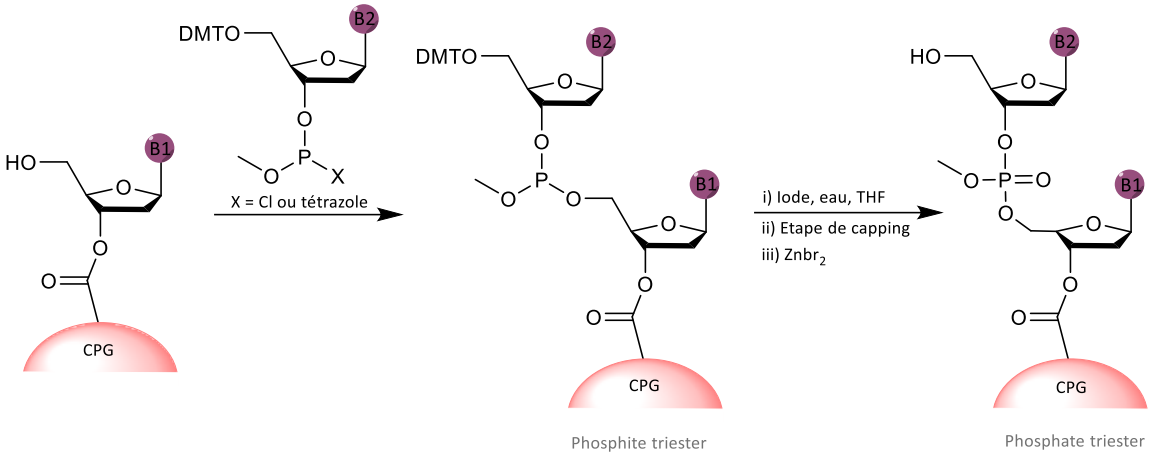


Figure I. 9: Synthèse d'oligonucléotides avec la méthode phosphite triester.

Au total, l'addition d'un nucléotide prend environ quatre heures et les rendements montent jusqu'à 90%. Il est possible de synthétiser des oligonucléotides en utilisant chaque nucléotide différent et un simple groupement méthyle est utilisé comme groupement protecteur pour le phosphate.

Des résultats encore meilleurs sont obtenus par la suite en changeant d'intermédiaire. L'intermédiaire phosphite déoxynucléoside activé est fonctionnalisé par le tétrazole à la place d'une phosphochloridite.[68] Ce changement permet d'obtenir des rendements autour de 95% en un temps de réaction d'environ 60 minutes. A la fin, le groupement protecteur méthyle est enlevé en utilisant l'oxyde de thiophène de triéthylammonium dans le dioxane, et la séquence est clivée du support par une hydrolyse de la liaison ester en utilisant de l'hydroxyde d'ammonium concentré.

De plus, ce procédé peut être automatisé. Un appareillage composé d'une colonne en verre où le support en silice fonctionnalisé par la première thymidine est utilisé. Le tout est relié à une mini pompe et une boucle d'injection. En utilisant cette méthode automatisée, plusieurs séquences contenant les quatre bases nucléiques sont synthétisées. Néanmoins, cette méthode a des limites concernant l'instabilité des intermédiaires vis-à-vis de l'oxydation à l'air et de l'hydrolyse.

3.1.6. Méthode phosphoramidite

Des intermédiaires plus stables sont donc nécessaires pour vraiment pouvoir étendre cette méthode. Après leur découverte des phosphites tétrazoles, Caruthers et son équipe continuent à chercher de meilleurs intermédiaires phosphites et découvre que la phosphoramidite *N,N*-di-méthylamine est une bonne candidate. En effet, ce type de phosphoramidite est stable à l'oxydation par l'air et à l'hydrolyse à température ambiante. De plus, elle est facilement synthétisable et peut être stockée sous forme de poudre (voir Figure I. 10).[69] Pour pouvoir réagir avec un nucléotide le monomère de phosphoramidite doit être activé et des tests sont donc effectués pour déterminer quel activateur montre les meilleurs résultats. Il en est ressorti que le 1*H*-tétrazole est l'acide faible qui convient le mieux. Sous ces conditions, ils réussissent à créer la liaison inter-nucléotidique avec les quatre nucléosides différents. Pour avoir de meilleurs résultats encore, ils ont combiné leurs dernières découvertes avec l'utilisation des supports solides CPG.

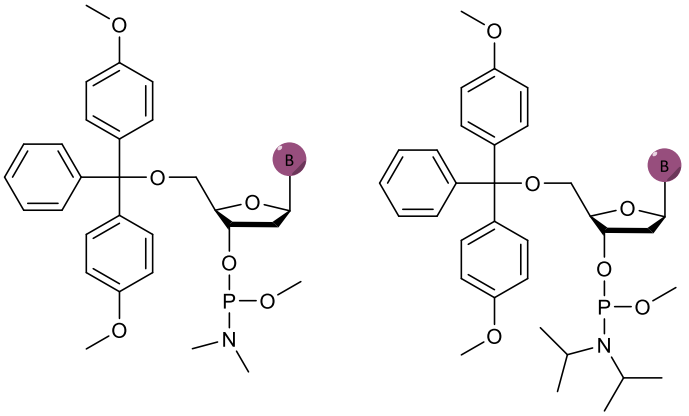


Figure I. 10: Schéma du nucléotide substitué par la phosphoramidite N,N-dimétylamine (gauche) et le nucléotide subtitué par la phosphoramidite N,N-diisopropylamine (droite).

Peu de changements ont été effectués depuis cette découverte, le principal fut la substitution du groupe *N,N*-di-méthylamine par le groupe *N,N*-di-isopropylamine (cf. Figure I. 10).[70, 71] Cette substitution aide lors des purifications des monomères et leur procure une meilleure stabilité dans le temps. Elle aide également pour l'automatisation de la procédure. Grâce à ces changements Adams *et. al* montrent qu'il est possible de synthétiser deux séquences d'ADN complémentaires de 51 monomères.[71]

3.2. Développement de la chimie de la phosphoramidite

Avec cette nouvelle méthode est née la chimie de la phosphoramidite. Cette chimie est toujours utilisée de nos jours pour effectuer des synthèses d'oligonucléotides. Grâce à sa facilité de mise en œuvre, des synthétiseurs d'ADN utilisant cette chimie sont développés. Ces robots peuvent être utilisés par des non-chimistes sans connaissance approfondie de la méthode qui est appliquée à l'intérieur, et tous les réactifs sont commerciaux.

3.2.1. Description des synthétiseurs d'ADN

La création de synthétiseur d'ADN arrive peu de temps après la découverte de la méthode phosphoramidite par Caruthers. Le premier synthétiseur d'oligonucléotide est créé en 1980 par l'entreprise Vega Biotechnologies.[72] Cet appareil peut synthétiser des branches d'ADN contenant jusqu'à 15 nucléotides en une journée. L'instrument contient deux parties : une partie chimique, permettant de synthétiser l'ADN en utilisant la chimie sur phase solide ; et une partie informatique, comprenant un ordinateur contrôlant la réaction et permettant de programmer la synthèse des séquences voulues. Cette machine est exposée au « National Museum of American History », et sa photographie est montrée en Figure I. 11.

Figure I. 11: Photographie du premier synthétiseur d'ADN, modèle 280.

Au même moment, Caruthers est contacté par de nombreuses entreprises voulant utiliser sa nouvelle façon de synthétiser de l'ADN. Il se laisse convaincre par l'une d'entre elles et collabore avec le biologiste Leroy Hood pour développer l'entreprise « Applied Biosystems ».[73] En 1982, un ingénieur de l'équipe design le premier synthétiseur d'ADN, nommé le 380A. Celui-ci est commercialisé dès 1983.

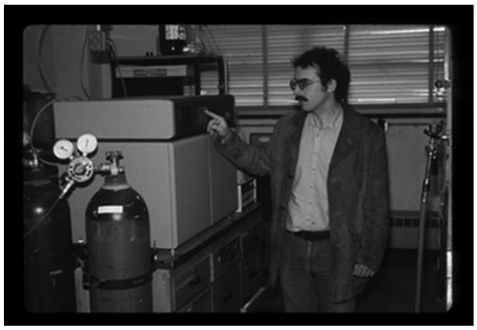


Figure I. 12: Photographie du premier synthétiseur d'ADN 380A mis en place dans le laboratoire de Caruthers en 1982. Tirée de la littérature.[73]

A cette époque, plusieurs entreprises créent leur propre synthétiseur d'ADN. Mais il s'est avéré que seule l'entreprise Applied Biosystems, dirigée par Caruthers et Hood, continua à développer de nouveaux modèles. En 1999, elle produit le synthétiseur 3900, contenant 48 positions. La photographie de ce modèle est donnée en Figure I. 13.

Figure I. 13: Photographie du synthétiseur Applied Biosystems, modèle 3900.

De nos jours, de nouvelles entreprises recommencent à manufacturer des synthétiseurs d'ADN. C'est le cas de Twist Bioscience et Agilent. Avec ces nouveaux instruments, il est possible de produire 240 000 segments uniques d'ADN préparés sur des micropuces. Agilent produit l'équivalent du génome humain (6 milliards de couplages) chaque jour grâce à ces machines fonctionnant 7 jours sur 7 et 24h sur 24.[73, 74]

3.2.2. Description du cycle itératif

Comme mentionné précédemment, la chimie de la phosphoramidite est très répandue pour l'utilisation dans des synthétiseurs d'ADN. Cet appareil est composé de plusieurs réservoirs pour les réactifs nécessaires, notamment pour les monomères de phosphoramidite. L'appareil commence la synthèse en prélevant la quantité optimale de réactifs et de monomères pour effectuer la synthèse. Tout est contrôlé par un logiciel où il est nécessaire d'indiquer la séquence voulue. Ces informations sont ensuite transmises au synthétiseur qui procède à la synthèse. Cette dernière est décrite comme étant une procédure faite de cycles itératifs effectués sur un support solide et suivant quatre étapes, qui sont schématisées sur la Figure I. 14 suivante. Elles sont également détaillées dans les paragraphes qui suivent.

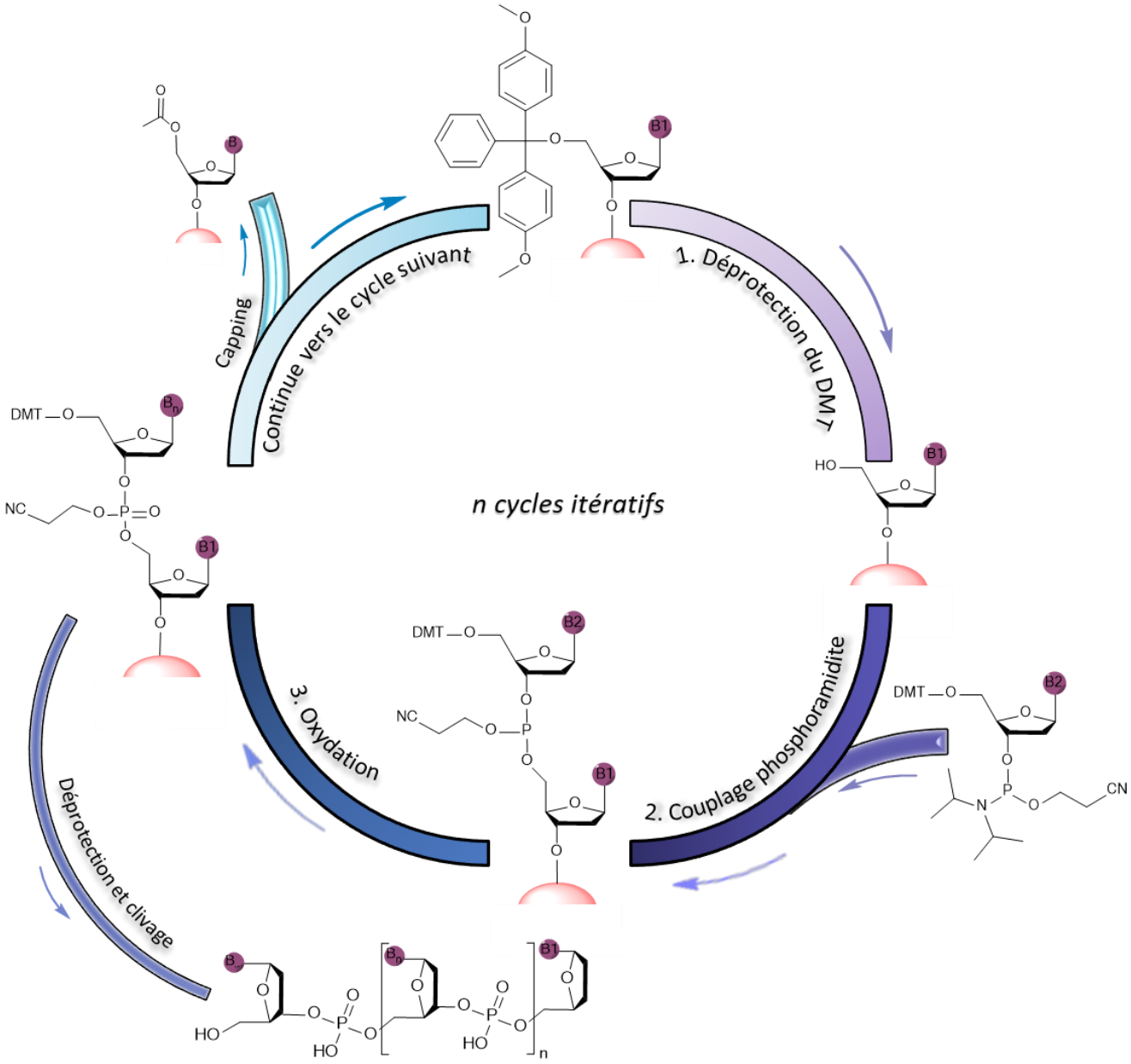


Figure I. 14: Schéma récapitulatif des différentes étapes du cycle itératif de la chimie de la phosphoramidite.

3.2.2.1. Etape de déprotection de l'alcool

Au début de la synthèse, l'alcool 5' de bout de chaîne du nucléoside est protégé par le groupement protecteur DMT. La première étape consiste donc à libérer cet alcool protégé par le DMT. Pour ce faire, l'acide trichloro-acétique est utilisé dans du dichlorométhane, comme présenté sur la Figure I. 15. Après la déprotection, le groupement 5'-hydroxy libéré peut réagir lors de la seconde étape.

Le carbocation du DMT a une couleur orange qui lui apporte une forte absorption de la lumière aux alentours de la longueur d'onde à 500 nm (cf. section 3.1.1). Ainsi, cette détection UV peut être utilisée pour suivre l'efficacité de couplage de chaque étape. Moins d'une minute est nécessaire pour effectuer la déprotection et les lavages.

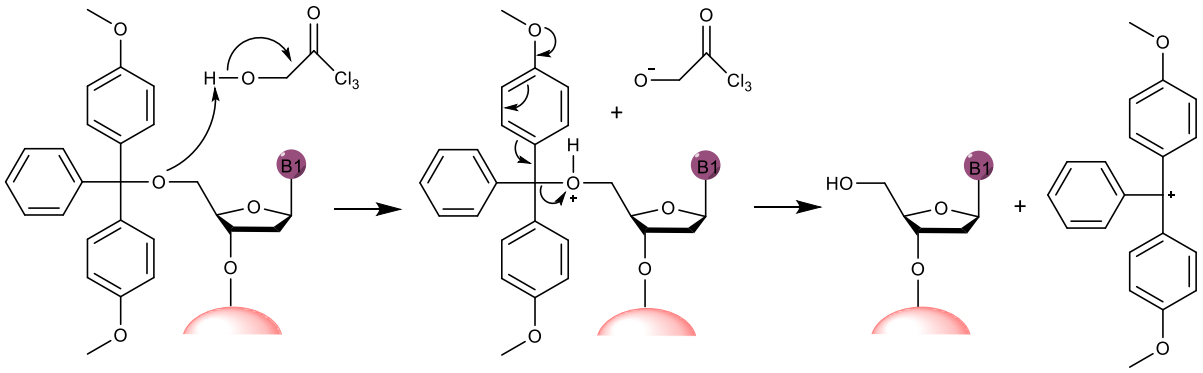


Figure I. 15: Schéma de l'étape de déprotection de l'alcool.

3.2.2.2. Etape du couplage du monomère de phosphoramidite

La chaîne en expansion est toujours reliée au support solide, à sa fin se trouve la fonction 5'-hydroxy finale. Lors de cette deuxième étape, celle-ci réagit avec le nouveau nucléotide qui est ajouté sous la forme d'un monomère de phosphoramidite activé par le 1*H*-tétrazole, créant ainsi la liaison phosphite triester. Cette activation du nouveau monomère est faite par le 5-(éthylthio)-1*H*-tétrazole, un catalyseur qui vient protoner le groupement di-isopropylamine. Le schéma réactionnel est donné en Figure I. 16. Le monomère de phosphoramidite ainsi activé est injecté en large excès dans le but d'atteindre un rendement très important. Le temps nécessaire à la réaction est d'environ 2 minutes.

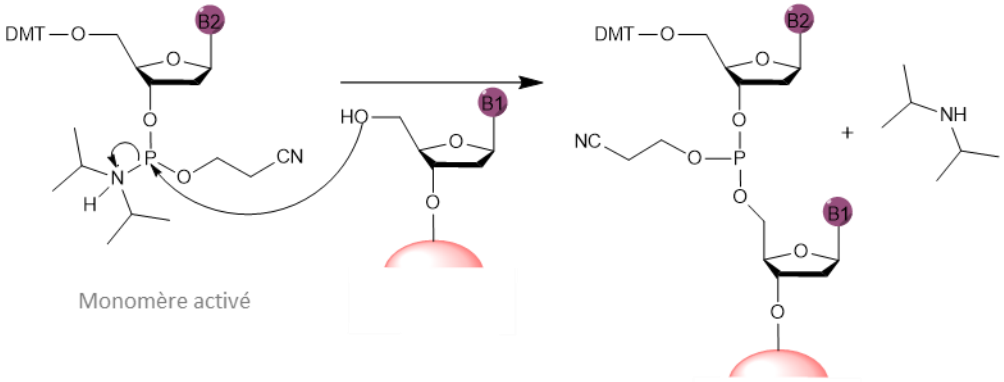


Figure I. 16 : Schéma de l'étape de couplage de la phosphoramidite.

3.2.2.3. Etape d'oxydation

Le phosphite triester formé durant la seconde étape est converti en un phosphotriester plus stable lors de l'étape d'oxydation. Cette conversion est effectuée par de l'iode en présence d'eau et de pyridine (Figure I. 17). La solution oxydante est délivrée à travers la colonne à une grande vitesse, rendant l'étape très rapide. En effet, l'oxydation et les lavages qui suivent ne prennent que 30 secondes environ.

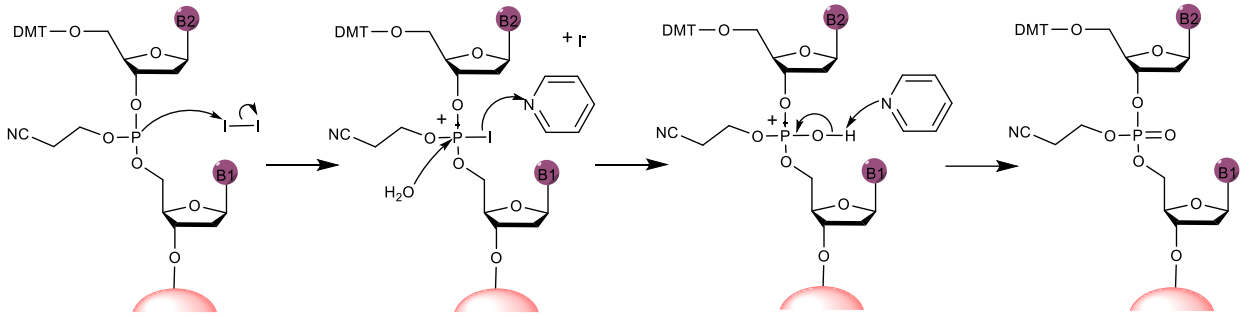


Figure I. 17: Schéma de l'étape d'oxydation.

3.2.2.4. Etape de capping

Le couplage du nouveau monomère de phosphoramidite est quasi-quantitatif. Une efficacité de 100% est en outre impossible, laissant toujours quelques chaînes présentant une fonction réactive 5'-OH à leur fin. Pour éviter les séquences avec des monomères manquants, il est nécessaire de bloquer ces dernières pour les empêcher de réagir lors de la prochaine étape de couplage. L'étape de capping est donc importante pour éviter l'accumulation de séquences tronquées. L'anhydride acétique activé avec du *N*-méthylimidazole est utilisé pour effectuer cette étape. Cet intermédiaire est mené jusqu'au nucléoside lié au support solide via les conduits de fluides de l'instrument. Le mécanisme est donné est en Figure I. 18.

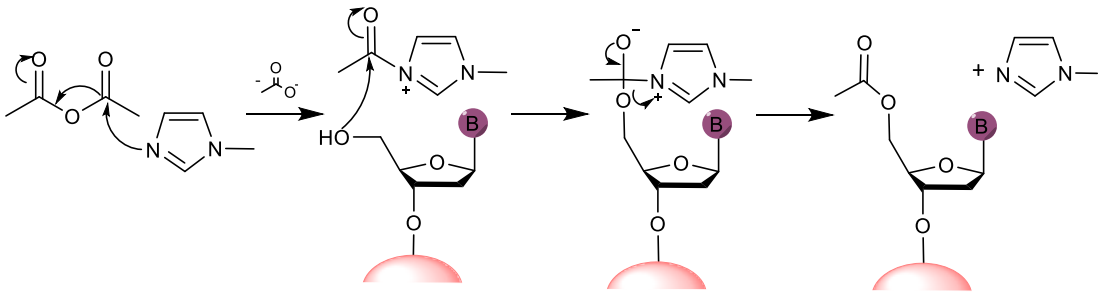


Figure I. 18: Schéma de l'étape de capping.

Après l'étape de capping, un nouveau cycle peut commencer et ainsi un nouveau monomère peut être ajouté. Ce cycle de quatre étapes est répété le nombre de fois souhaité pour obtenir la séquence désirée de poly(phosphodiester)s.

3.2.2.5. Déprotection, clivage et purification

Une fois la séquence voulue synthétisée, il est alors nécessaire de déprotéger la chaîne principale. En effet, le phosphate est toujours protégé par le groupement 2-cyanoéthyle et les bases nucléiques peuvent également être protégées, voir paragraphe 3.2.1. Ces groupements protecteurs sont détachés par l'utilisation d'une base faible, puis lavés et filtrés, alors que la séquence est toujours liée au support.

Un traitement avec une solution d'ammoniaque est généralement utilisé pour cliver l'ester succinique qui lie encore la séquence avec le support solide. Par la suite des lavages sont nécessaires pour récupérer la séquence. Finalement, la séquence obtenue est donc un oligonucléotide terminé par un 3'-OH libre.

Lorsque qu'une solution d'ammoniaque seule est utilisée, le temps de clivage est d'au moins 1h. Des études sont effectuées pour essayer de réduire ce temps de clivage. Plusieurs tests sont faits pour trouver une meilleure solution de clivage. Des solutions plus nucléophiles ainsi que des mélanges avec de l'ammoniaque sont alors essayés. Il s'est avéré que la solution de méthylamine/ammoniaque 1/1 donne les meilleurs taux de déprotection dans un temps plus court qu'une heure.[75] Cette solution est donc dorénavant utilisé pendant 30min à température ambiante pour cliver et déprotéger les séquences.

Après le clivage du support, un mélange contenant la bonne séquence mais également des séquences plus courtes est obtenu. Une purification est donc nécessaire pour pouvoir séparer les chaînes et obtenir uniquement la séquence désirée. La façon la plus simple de procéder est d'utiliser le mode « DMT-on », ce qui veut dire que le dernier monomère ajouté aura toujours sa protection DMT lorsque la séquence sera clivée du support. Grâce à cette fonction encore présente, la séquence voulue est différente de celles tronquées et une simple chromatographie permet une séparation efficace. Dû à la polarité, les séquences plus petites peuvent être séparées sur une colonne en phase inverse. Lors de cette séparation, les groupements protecteurs préalablement clivés et qui sont potentiellement encore présents dans la solution vont aussi être lavés. La fin de la séquence en DMT est moins polaire et apporte donc un mouvement plus lent à toute la séquence d'oligonucléotides. Ainsi une purification avec une HPLC préparative est possible en commençant par un solvant polaire puis en le remplaçant pour un moins polaire.

Il existe d'autres manières de purifier les séquences, notamment une méthode manuelle en utilisant un kit de purification sur phase inverse. Après le clivage et les déprotections, le mélange des oligomères de différentes tailles est manuellement disposé sur une colonne en phase inverse. Un premier lavage sépare les espèces polaires, et la séquence voulue va rester sur la colonne grâce à son bout de chaîne DMT. Une solution acide est ensuite utilisée pour cliver la fonction protectrice DMT, ce qui apporte une couleur rouge orangé à la colonne. Le DMT libre est lavé avec de l'eau puis l'oligonucléotide désiré peut être collecté. Il est alors complétement déprotégé, présentant des groupements hydroxyles en bout de chaîne.

3.2.1. Supports solides utilisés

Depuis les recherches faites par Letsinger, les supports solides ont été améliorés. Des résultats intéressants montrent l'utilisation de support solide de polymère. Des supports fortement réticulés et peu absorbants tels que le polystyrène aminométhylé macroporeux montrent leur intérêt, surtout pour leur hydrophobicité.[76] En effet, grâce à l'hydrophobicité de ce polystyrène, la phosphoramidite est moins susceptible d'être hydrolysée, les taux d'efficacité s'en trouvent donc augmentés.

Rapidement les supports inorganiques en silice montrent de nombreux avantages. Etant déjà largement utilisé pour les analyses d'HPLC, les supports en verre poreux (CPG) se sont facilement et largement démocratisés pour la chimie de la phosphoramidite. De plus, leur rigidité et donc leur incompressibilité permet une résistance au passage de solvants et réactifs sous un haut débit. Par ailleurs l'inertie de ce support permet d'accéder à un large choix de solvants, sans crainte de gonflement.[77] Ce type de support est disponible avec différentes tailles de pores. Pour la synthèse d'oligonucléotides des tailles de pore allant de 10nm à plusieurs centaines de nanomètres sont utilisés. Plus la taille des pores est grande, plus la longueur de la chaine d'oligonucléotide synthétisée efficacement peut être grande. Il a aussi été montré que les supports du type lcaa (Long Chain AlkylAmine) CPG où l'immobilisation se fait via une longue chaîne alkyle sont efficaces.[78] Ce sont ces types de support qui sont dorénavant les plus utilisés, leur structure est donnée en Figure I. 19.

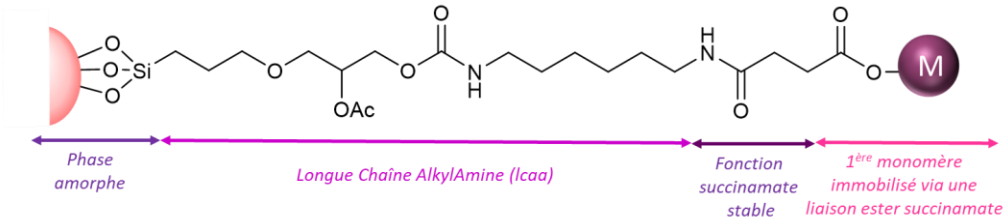
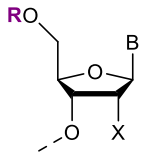
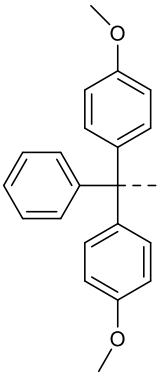
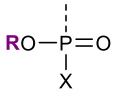
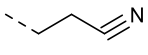


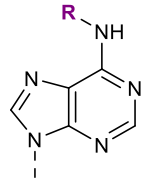
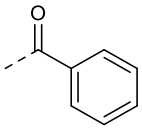
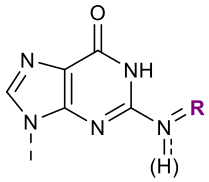
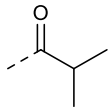
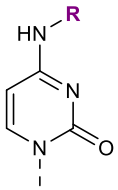
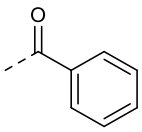
Figure I. 19: Structure du support solide lcaa standard.

3.2.2. Groupements protecteurs

Il a été prouvé que la chimie de la phosphoramidite montre les meilleurs résultats lorsqu'elle est effectuée sur un support solide et en utilisant le groupement protecteur DMT sur la fonction 5'-hydroxy. Les fonctions amines présentes sur les bases nucléiques doivent elles aussi être protégées afin d'éviter qu'elles ne réagissent avec les phosphoramidites. Les groupements protecteurs récapitulés dans le Tableau I. 2 suivant permettent une utilisation classique de la méthode de la chimie de la phosphoramidite tout en étant assez robustes. Ce genre de groupements protecteurs peut être enlevé assez rapidement en utilisant une solution aqueuse d'ammoniaque/méthylamine.[79]

Tableau I. 2: Liste des groupements protecteurs classiques utilisés pour la chimie de la phosphoramidite.

Fonction protégées	Groupements protecteurs	Commentaires
 Groupement 5'-hydroxy	**R** = Groupement DMT 	Le groupement protecteur DMT peut facilement être clivé via l'utilisation d'une solution d'acide faible telle qu'une solution à 3% de TCA dans le dichlorométhane. Ce clivage est facilité grâce à la présence des groupes méthoxy donneurs d'électrons qui vont permettre la stabilisation du carbocation central. Ainsi, un groupement trityle aurait été trop stable sous les mêmes conditions de clivage et un groupement 4,4',4''-trimethotrityle (TMT) aurait lui été trop instable. De plus, la forte absorption du cation DMT permet d'évaluer le rendement de chaque ajout lors de la synthèse.[80]
 Phosphite (Phosphate)	**R** = Groupement cyanoéthyle (CE) 	Le groupement 2-cyanoéthyle protégeant le phosphate a été identifié comme étant stable pour la chimie de la phosphoramidite. Il est clivé lors du clivage du support. Le groupe partant acrylonitrile est volatile.[57]
	R = Groupement benzoyle	Les amines peuvent être protégées via des groupements benzoyles (Bz).[48] Ce groupement

 Amine de l'adénine		protecteur est enlevé lors du clivage du support avec la solution d'ammonique/méthylamine.
 Amine de la guanine	**R** = Groupement isobutyryle 	L'amine de la guanine peut être protégée par le groupement isobutyryle.[77]
 Amine de la cytosine	**R** = Groupement benzoyle  **R** = Groupement acétyle 	Dans un premier temps, l'amine de la cytosine était protégée avec un groupement benzoyle[48], mais il s'est avéré qu'une modification pouvait arriver lorsque le clivage se fait en présence de méthylamine.[79] L'utilisation du groupement protecteur acétyle permet d'éviter ces modifications. En effet, en augmentant l'électropositivité du carbone de la fonction carbonyle, la déprotection se fait plus rapidement et donc élimine totalement les risques d'attaques sur la pyrimidine. [79]

3.3. Stockage de données sur l'ADN

Depuis le développement de la chimie de la phosphoramidite de nombreuses recherches sont faites en utilisant cette méthode. Il est dorénavant possible de produire des brins d'ADN synthétiques. Ceux-ci peuvent reproduire le code génétique mais il est aussi possible de coder n'importe quelle information sur ces brins synthétiques.

3.3.1. Principe : comment stocker de l'information ?

Le stockage d'information à l'échelle moléculaire semble être une voie prometteuse.[13] L'obsolescence des moyens de stockage actuels n'est plus à prouver, aucun dispositif actuel n'a une espérance de vie beaucoup plus longue que dix ans. Le stockage à l'échelle moléculaire via la synthèse d'ADN a un avantage

énorme sur ce point, car gardé dans de bonnes conditions d'humidité et d'exposition à la lumière, l'ADN peut être conservé pendant des siècles voire des millénaires.[11, 81]

La capacité de stockage sur l'ADN est environ six fois plus importante que celle obtenue avec les dispositifs actuels.[10, 14, 15] La densité d'information stockable est donc clairement augmentée par rapport à celles possible d'atteindre avec les traditionnels disques durs. Pour la comparaison, il faudrait seulement quelques grammes d'ADN pour stocker un zettaoctet d'informations alors qu'il faudrait 1kg d'alliage de cobalt pour stocker la même quantité d'information. La cytosine, la guanine, l'adénine et la thymine sont les quatre bases de l'ADN permettant de stocker toute l'information génétique d'un être vivant. Grâce à un codage effectué sur ces quatre bases, la densité d'information qu'il est possible de stocker est énorme sachant qu'il est donc possible de stocker 2 bits par nucléotide. Théoriquement 455 millions de téraoctets peuvent être stockés par gramme d'ADN. [11] De plus, des techniques de lectures ont déjà été développées et sont facilitées par la propriété de réplication de l'ADN en utilisant l'amplification en chaîne par polymérase, PCR (pour Polymerase Chain Reaction en anglais).[82] Il est alors pertinent d'explorer l'utilisation de l'ADN comme support de stockage pour les informations générés par les êtres humains de nos jours.

Dans cette optique, l'idée de stocker à l'échelle moléculaire est démontrée pour la première fois en 1986 via l'encodage d'une image de 35 bits. Pour ce faire Davis a recourt à une étape *in vivo*.[83] En 1999, Clelland et *al.* démontrent qu'il est possible de stocker de l'information sur des brins d'ADN sans avoir recourt à une étape *in vivo,* c'est-à-dire sans stocker le message sur des cellules vivantes.[84] Ce n'est qu'en 2012 que d'autres équipes réitèrent l'expérience. Church et Goldman montrent indépendamment qu'il est dorénavant possible de stocker des centaines de kilooctets d'informations, poussant au développement rapide des techniques d'écriture et de lecture de l'ADN synthétique.

Ainsi, Church décrit qu'il a encodé un livre entier sur des séquences d'ADN, ce qui correspond à 650 kilooctets d'informations et Goldman montre qu'il a synthétisé un logiciel informatique de 630 kilooctets. [11, 12] Récemment, 200 mégaoctets d'informations incluant un vidéoclip de haute définition et 100 livres ont été stockés sur de l'ADN contenant plus de 1.5 milliards de paires de bases nucléiques.[85] En 2019, ce sont les travaux d'Organick et *al.* qui montrent la plus haute densité de stockage encore jamais obtenue avec un total de 33 kilooctet d'information stockées en utilisant un faible nombre de copies des séquences.[14] En effet, trente-cinq dossiers sont encodés et stockés sur seulement 13,4 millions d'oligonucléotides d'ADN de 150 nucléotides, ce qui correspond environ à 36 copies pour une première partie des fichiers (32 kilooctects) et 80 copies pour les 1,3 ko restants. En comparaison Church utilisa 3000 copies pour stocker les 650 kilooctets d'informations ce qui diminue sa densité de stockage finale.[11] Ainsi Church a une densité par nucléotide de 0,6 bit, alors qu'Organick réussi à atteindre les 0,81 bit.

En outre, il est encore difficile de produire des séquences plus longues que 150 à 200 nucléotides, un trop grand nombre d'erreurs apparaissant au-delà. Pour atteindre un stockage plus important, différentes méthodes sont utilisées, l'emploi de micropuces étant la plus courante.[86] La synthèse des séquences d'oligonucléotides est réalisée sur des surfaces avec des délimitations spatiales pour chaque polymère. L'activation des synthèses s'effectue en utilisant une chimie activée par la lumière. Des surfaces contenant plusieurs séquences d'oligonucléotides sont ainsi accessibles, chaque séquence contenant une partie du message codé. L'ensemble de ces dernières code une densité d'information assez importante.

L'utilisation de cellules vivantes pour le stockage d'information reste moins performante que la méthode *in vitro* au niveau de la densité de stockage. De nombreuses recherches sont pourtant toujours en cours sur cette méthode. Par exemple, en 2017, Church montra qu'il est possible de coder une courte animation GIF en utilisant une population de bactéries vivantes et la méthodologie CRISPR-Cas.[87]

3.3.2. Automatisation du stockage de données

Le stockage d'information à l'échelle moléculaire sur des oligonucléotides d'ADN montre depuis plusieurs années son efficacité. Néanmoins, malgré l'automatisation des synthèses et le développement des lectures de séquençage, beaucoup d'étapes intermédiaires nécessitent encore une présence humaine pour être exécutées. Ce n'est qu'en 2019 qu'une équipe développe un dispositif automatisé du début à la fin, permettant de stocker de l'information sur de l'ADN. Des chercheurs de Microsoft et de l'Université de Washington créent ce premier système automatique qui est une avancée primordiale pour permettre de translater les recherches dans ce domaine du laboratoire jusque dans les centres de données commerciaux.[88]

Cet appareil encode l'information sur des séquences d'ADN, traduisant les bits 0 et 1 en A, C, G et T. Il effectue également une correction d'erreurs pour donner une plus grande robustesse aux séquences. Celles-ci sont alors synthétisées sur un synthétiseur. Elles sont ensuite regroupées pour un stockage liquide, puis lues via un séquenceur à nanopores. En utilisant leur prototype, l'équipe de Ceze a réussi à stocker un message de 5 octets, le mot « HELLO ». 21 heures ont été nécessaires pour l'encodage et la lecture du message. Sur une quantité de 1mg d'ADN synthétisé, une seule séquence correcte a pu être décodée. De nombreuses améliorations sont à venir qui vont réduire le temps nécessaire pour réussir le process et augmenter le rendement.

Des développements sont également en cours sur l'étape de stockage liquide puis lecture. Il est montré que cette étape peut être accomplie via une amplification en chaîne par polymérase (PCR) de façon aléatoire. La PCR permet d'amplifier les séquences d'oligonucléotides synthétisées ce qui permet une lecture facilitée, réduisant cependant la densité d'information stockée. Néanmoins, il a récemment été montré qu'un système d'une densité de 17 exaoctets par gramme pouvait être atteinte grâce à l'utilisation de seulement 10 copies par séquence stockées.[89] Cette densité est quasiment deux fois plus élevée que celle obtenue via les « fontaines d'ADN » qui permettent d'atteindre une densité de stockage de 215 pétaoctets.[90]

4. Stockage de données sur des macromolécules synthétiques

En suivant l'exemple du stockage de données sur de l'ADN, des recherches sont effectuées pour stocker de l'information sur des macromolécules synthétiques. Pour que le stockage soit possible, il est nécessaire d'avoir des séquences de polymères parfaitement définies, aucune erreur de synthèse n'est permise. En effet, si une erreur est commise lors de la synthèse, la séquence obtenue ne code plus l'information désirée. Pour avoir un codage optimal, il faut donc utiliser une chimie permettant de synthétiser des polymères parfaitement monodisperses.

4.1. Nécessité d'avoir un système bien contrôlé

Il y a tout juste 100 ans, Hermann Staudinger décrit pour la première fois en 1920 une réaction de polymérisation. [91] Il est le premier à présenter la formation des molécules à haut poids moléculaire comme étant la succession de plusieurs petites molécules liées de façon covalentes les unes aux autres. C'est en 1922, qu'il nomme ce nouveau concept les « macromolécules » couvrant ainsi les polymères synthétiques et naturels. Aujourd'hui les macromolécules issues de la chimie des polymères sont présentes partout, que ce soit dans les emballages alimentaires, dans l'automobile, les fibres textiles…[92] L'architecture de celles-ci sont de plus en plus précises pour convenir aux demandes des technologies modernes. Plusieurs voies de synthèses existent pour atteindre des polymères plus ou moins bien contrôlés.[93] Les trois principales que sont la polymérisation en étape, en chaîne et la synthèse multi-étapes sont décrites ci-après.

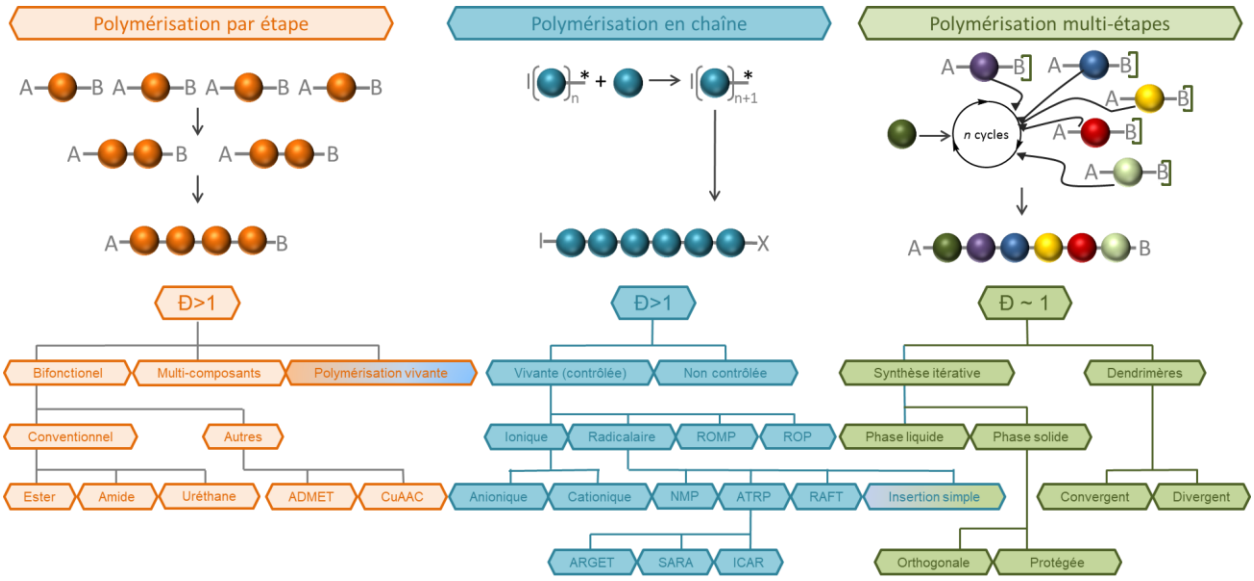


Figure I. 20: Classification générale des techniques de polymérisation permettant d'obtenir des polymères synthétiques à séquences contrôlées. (adapté de la littérature[93] suggérant la terminologie des synthèses de polymères).

4.1.1. Polymérisation par étape

La polymérisation par étape, se fait par une succession de réactions de condensation entre des monomères bi ou polyfonctionnels (partie orange de la Figure I. 20). Lorsque des monomères bi-fonctionnels (AB) sont utilisés, cette condensation aboutit en un nouveau dimère portant les mêmes

fonctions en bout de chaîne, qui peut lui-même réagir avec un nouveau monomère ou avec un autre dimère ou oligomère créé. Ainsi dans ce type de polymérisation la chaîne de polymère croît grâce à l'ajout de nouveaux monomères ou oligomères. Lors de ces condensations il y a généralement l'élimination d'une petite molécule, souvent une molécule d'eau, mais cela varie en fonction des monomères utilisés.

Carothers et Flory sont les deux pionniers dans le développement de ce type de polymérisation.[17, 94] Leurs travaux respectifs ont permis une meilleure compréhension de cette polymérisation et depuis elle est développée grandement pour l'industrie, avec notamment la polymérisation du PET pour l'obtention de bouteilles en plastiques.[95] Mais l'intérêt de cette voie pour synthétiser des polymères à séquence contrôlées est limité, car le contrôle n'est pas total et la polymolécularité est souvent aux alentours de Đ ≈ 2.

4.1.2. Polymérisation en chaine

Des réactions de polymérisation en chaine sont déjà observées dès le début du XIXème siècle.[96] Un exemple est celui de l'apothicaire allemand Eduard Simon qui en 1839 montre qu'il a pu isoler une huile qu'il appelle le « Styrol ».[97, 98] Or, quelques mois plus tard il remarque que cette huile s'est solidifiée et ressemble désormais à une masse gélatineuse et transparente qui n'est plus soluble ni dans l'alcool ni dans l'éther. Il pense que cette transformation est due à la combinaison des actions de l'air, de la chaleur et de la lumière ; il appelle ce nouveau produit le Styroloxyd" (oxyde de styrol). On sait maintenant que ce qu'il observe est en fait la polymérisation en chaine du styrène.

En 1953, Flory donne un descriptif complet de ce type de polymérisation.[99] Il s'agit d'une synthèse mettant en jeu des réactions entre un ou des monomère(s) et un ou des site(s) actif(s). Une fois que le site actif réagit celui-ci est régénéré pour pouvoir être impliqué dans une nouvelle étape. L'élément clé de la polymérisation en chaine est donc l'activation de ce site actif, sans activation les monomères ne peuvent pas réagir entre eux. Ensuite la phase de propagation a lieu, augmentant la taille du polymère et finalement la chaine se termine lors de la phase de terminaison.

Dans certain cas, aucune réaction de transfert ni de terminaison a lieu, on va alors parler de polymérisation vivante, tel que la polymérisation anionique vivante. Avec ce type de polymérisation il est possible d'obtenir des indices de polymolécularité assez faibles aux alentours de Đ < 1,1.[100]

Il existe également les réactions de polymérisations radicalaires contrôlées, celles-ci sont plus complexes que les réactions anioniques vivantes. Lors des réactions de polymérisations radicalaires contrôlées il existe des réactions de terminaison mais il est tout de même possible d'obtenir des architectures contrôlées grâce à l'utilisation de techniques spécifiques. Les exemples les plus courant sont les réactions connues sous les acronymes ***NMP*** de l'anglais Nitroxide-Mediated Polymerization (Modulation de la réactivité des radicaux propagateurs par des contre-radicaux nitroxydes),[101] ***ATRP*** de l'anglais Atom-Transfer Radical Polymerization (Polymérisation radicalaire par transfert d'atomes),[102-104] et ***RAFT*** de l'anglais Reversible-Addition Fragmentation Chain-Transfer Polymerization (Polymérisation radicalaire contrôlée par transfert de chaîne réversible par addition-fragmentation).[105-109]

Ces types de réactions sont efficaces pour obtenir des architectures bien contrôlées mais elles ne permettent pas d'accéder à un contrôle total de la séquence. Une autre voie de synthèse doit être utilisée pour obtenir un contrôle de la séquence au monomère près.

4.1.3. Synthèses multi-étapes

Grâce aux recherches effectuées sur le contrôle des polymérisations via les synthèses en étapes ou en chaînes il est possible d'atteindre des indices de polymolécularité corrects mais toujours pas une isomolécularité et un contrôle parfait de la séquence de monomère. Pour réussir à synthétiser une séquence totalement définie il faut utiliser des synthèses multi-étapes comme les synthèses itératives ou encore la synthèse de dendrimères. Ce genre de synthèse permet d'obtenir un contrôle absolu de la séquence synthétisée ce qui est nécessaire lorsque l'on veut stocker de l'information sur ces macromolécules.

4.1.3.1. Synthèse itérative linéaire

Les synthèses itératives s'effectuent via une succession de plusieurs cycles comportant un certain nombre d'étapes. Chaque cycle permet d'ajouter un nouveau synthon et chaque étape du cycle permet l'ajout efficace de ce synthon. La synthèse itérative via la chimie de la phosphoramidite décrite en section 3.2 est un exemple de ce type de synthèse.

Il s'agit donc d'une réaction entre deux fonctions A et B. Un monomère ayant ces deux fonctions est utilisé, néanmoins la fonction A est protégée. Cela permet que la fonction B ne réagisse qu'avec la fonction A de la chaîne en croissance et non avec les fonctions A qui viennent d'être ajoutées au système lors de l'ajout des nouveaux. Lors de la première étape du cycle la fonction B réagit avec la fonction A de la chaîne en croissance. Une séquence comprenant un nouveau monomère est obtenue et celle-ci est terminée par une fonction A protégée. Dans la deuxième étape, la fonction A est déprotégé ce qui permet de la rendre active pour le prochain cycle.[110]

Ce genre de synthèse peut s'effectuer en solution ou en phase solide. La synthèse sur phase solide est décrite en section 3.1.3. Concernant la synthèse en phase soluble, de nombreux travaux montrent qu'il est possible de synthétiser de nombreux type de polymères tels que les polypeptides ou des polymères périodiques.[111-113] Dans tous les cas, lorsque la synthèse itérative est effectuée en solution il est nécessaire d'effectuer à chaque étape du cycle (ajout d'un monomère ou déprotection) une purification du produit ce qui est chronophage.

4.1.3.2. Synthèse de dendrimères

En 1985, le terme dendrimère est pour la première fois utilisé par l'équipe de D. A. Tomalia chez Dow Chemical suite à la synthèse d'une macromolécule qui leur fait penser à un arbre (dendron en grec).[114] Très vite après cette découverte de nombreuses recherches portent sur la synthèse de ce type de macromolécule et son optimisation.

Ces dendrimères sont composés d'un noyau qu'on appelle aussi le cœur et duquel partent les branches appelées dendrons. Ces macromolécules peuvent être obtenues de deux manières différentes, soit via

une synthèse divergente : du cœur vers les extrémités ou via une synthèse convergente : des extrémités vers le cœur.

Le désavantage majeur de la synthèse divergente est que le nombre de réactions à effectuer à chaque étape est de plus en plus grand ce qui augmente l'encombrement. Cela résulte en la présence d'un plus grand nombre de défauts dans la structure que lorsqu'une synthèse par voie convergente est utilisée.[114]

Une structure sphérique est obtenue pour la plupart des dendrimères synthétisés, ce qui permet d'avoir des propriétés physiques bien différentes comparées à celles que l'on peut avoir avec des polymères linéaires.[115] De plus, ce type de macromolécules comportent beaucoup plus de bout de chaînes qu'un simple polymère linéaire il est donc facilement imaginable de fonctionnaliser ces bouts de chaînes apportant ainsi des propriétés intéressantes au dendrimère. On peut par exemple penser à fonctionnaliser les bouts de chaînes avec des sites actifs qui pourront être utilisés dans le domaine médicinal.[115] En biologie, ce type de macromolécule est aussi très intéressant. Il est, par exemple, possible de synthétiser des peptides dendritiques qui peuvent être utilisés comme antigènes.[116]

4.2. Diversité des espèces chimiques possibles pour le stockage d'information sur des polymères synthétiques

Les biopolymères sont de bons candidats pour le stockage d'information, leurs méthodes de synthèse, de réplication et de lecture sont développées depuis plusieurs dizaines d'années. Néanmoins, leur structure moléculaire n'est pas forcément optimale pour des applications non biologiques. Les polymères synthétiques semblent donc être une bonne alternative et apparaissent comme la clé pour développer de nouvelles méthodes pour stocker à l'échelle moléculaire.[22, 110]

La Figure I. 21 présentée ci-après, introduit les différentes familles de macromolécules qui permettent de coder de l'information.

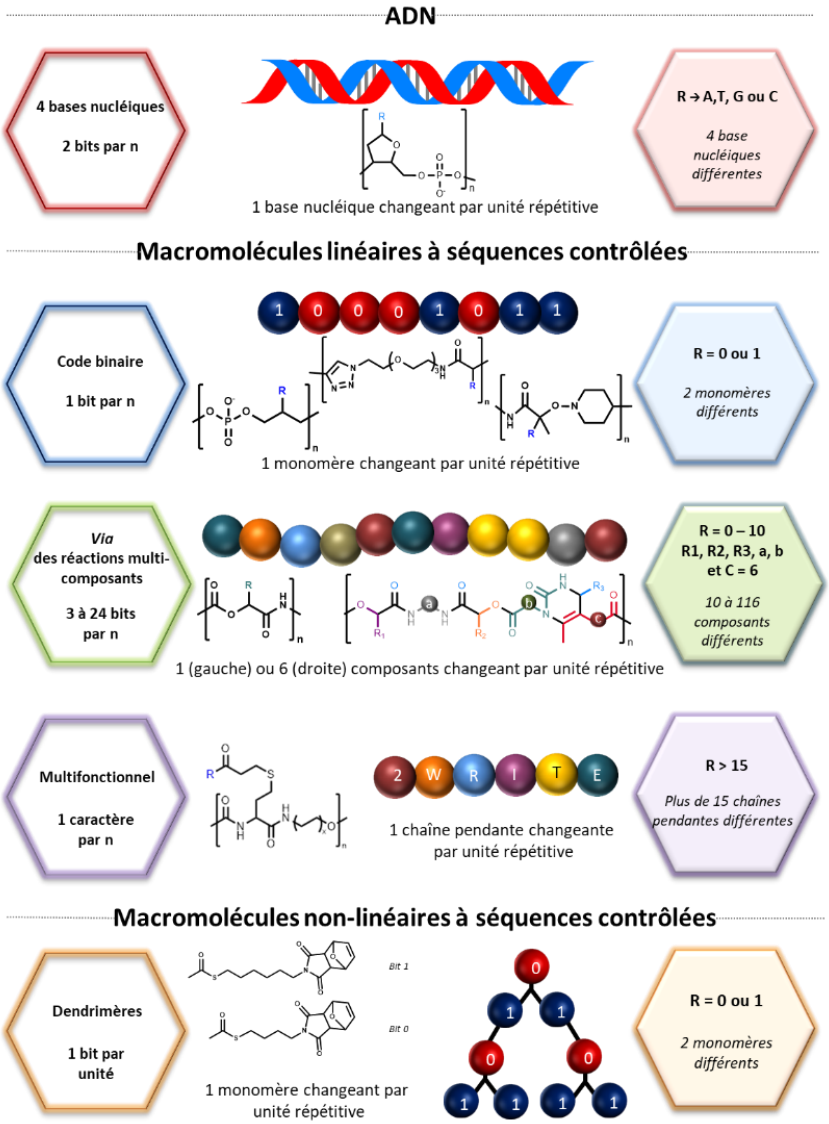


Figure I. 21: Comparaison des différents systèmes macromoléculaires et de leur capacité de stockage théorique dans une unité répétitive. Rouge : Information codée via l'ADN ; Bleu : Macromolécules à séquence contrôlées basée sur l'encodage binaire (gauche[25], milieu[117], droite[118]) ; Vert : Macromolécules à séquences contrôlées via une réaction multicomposants (gauche[119], droite[120]) Violet: Séquences contrôlées multifonctionnelles.[121] ; Orange : Macromolécules à séquences contrôlées sur des dendrimères.[122]

Elles sont présentées dans les paragraphes suivants. En rouge, les séquences de biopolymères comme l'ADN. En bleu, les séquences contrôlées encodées de manière binaires seront la base des travaux effectués au cours de cette thèse. En vert et violet sont résumés les autres exemples de macromolécules linéaires codant de l'information. Et enfin en orange, l'exemple des dendrimères démontrant que le stockage d'information peut également se faire sur des architectures non-linéaires.

4.2.1. Les oligo(triazoles amide)s

Les séquence d'oligo(triazoles amide)s sont le premier exemple de polymère non naturel numérique décrit par Lutz et *al.* en 2014.[20] Ces séquences sont basées sur le code binaire (0, 1) et préparées via l'approche (AB+CD) sur un support solide. Ce protocole repose sur deux réactions de couplages itératives chimioselectives (i) une amidification et (ii) une cycloaddition alcyne-azoture catalysée par le cuivre sur une résine de Wang non modifiée. Pour la synthèse de ces séquences deux type de composants AB (A = acide, B = alcyne) sont utilisés : l'acide 4-pentynoïque codant pour (0) et l'acide 2-méthyle-4-pentynoïque codant pour (1). Concernant le composant CD il s'agit du 1-amino-11-azido-3,6,9-trioxaundécane (C = amine et D = azoture).

Une fois les séquences désirées obtenues, les macromolécules sont clivées du support Wang en condition acide. L'information encodée est ensuite recouvrée via des analyses de MALDI-TOF, de SEC et de RMN qui montrent des chaînes monodisperse, prouvant le contrôle parfait de la structure primaire. Cependant, le séquençage par spectrométrie de masse en tandem montre des états de charges multiples de certains fragments rendant compliquée la lecture même pour des séquences simples et courtes.[123]

4.2.2. Les oligo(alcoxyamines amide)s

Les problèmes de séquençage rencontrés avec les oligo(triazoles amide)s sont contournés en utilisant des alcoxyamines à la place des triazoles.[21, 124] En effet, en 2015 Lutz et *al.* montrent qu'en utilisant une liaison alcoxyamine labile le séquençage en spectrométrie de masse en tandem est largement facilité.

Une stratégie itérative est employée pour synthétiser ce genre de macromolécule. Elle met en jeu deux réactions chimiosélectives : (i) la réaction d'une amine primaire avec un acide anhydride et (ii) la réaction de couplage radicalaire d'un radical d'un carbone central avec un nitroxyde. Deux différents anhydrides sont employés pour encoder l'information, l'un codant pour le bit 0 et l'autre pour le bit 1, le nitroxyde est utilisé comme un espaceur. Les synthèses sont réalisées sur un support solide. Une fois la séquence désirée obtenue, celle-ci est clivée du support avec une solution acide. Les résultats de SEC, RMN et ESI-MS montrent une grande pureté des macromolécules obtenues.

L'introduction de la liaison alcoxyamine labile dans la structure de la macromolécule permet un séquençage rapide et facile car à faible énergie d'ionisation, il n'y a que le clivage homolytique de la liaison alcoxyamine qui est obtenu. Cette fragmentation sélective clivant qu'une seule liaison par unité codante permet d'obtenir un spectre de fragmentation extrêmement facile à lire.

4.2.3. Les oligo(alcoxyamines phosphodiester)s

Les avantages obtenus grâce à la présence de la liaison alcoxyamine dans les monomères précédemment décrits, ont conduit l'équipe de Lutz à la synthèse de nouveaux monomères contenant également cette

liaison labile. Ici, il ne s'agit plus de séquences d'oligo(alcoxyamines amide)s mais de séquences contenant des liaisons phosphodiesters menant ainsi à des oligo(alcoxyamines phosphodiester)s. Quatre monomères sont synthétisés et codent chacun 2 bits : les dyades 00, 01, 10 et 11.[125, 126] Dans chaque monomère il y a 2 parties codantes. La première se situé au niveau de la partie phosphoramidite et la deuxième sur le nitroxyde, leur structure est donnée en Figure I. 22. Une lecture facile par spectrométrie de masse est toujours obtenue avec ce type de monomère.[127]

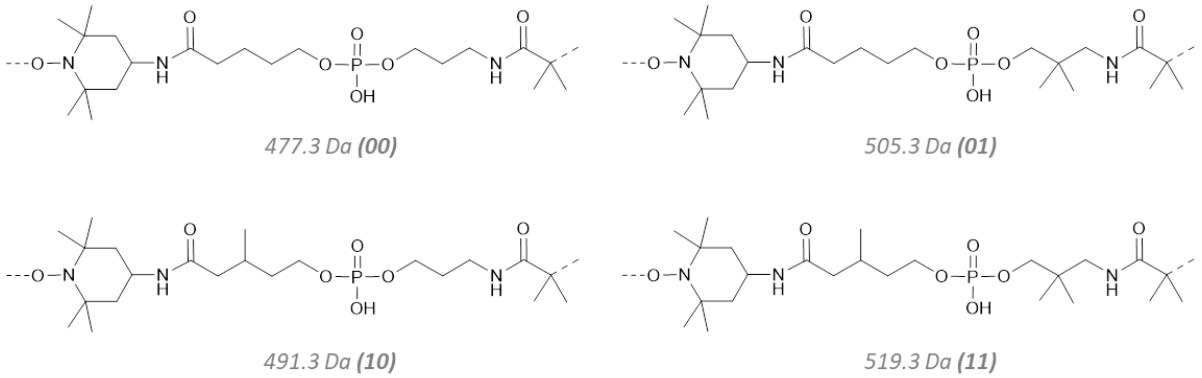


Figure I. 22: Structures moléculaires et masses molaires des monomères alcoxyamines phosphodiesters codants.

Grâce à la facilité de lecture conservée, celle-ci a pu être automatisée en utilisant le logiciel MS-DECODER. Un algorithme spécifique aux séquences est créé et permet la détection des dyades dans les séquences d'oligo(alcoxyamine phosphodiester)s. Le décodage de la séquence est alors effectué en quelques millisecondes.

L'algorithme « MS-DECODER » qui permet un décryptage rapide et explicite de polymères à séquences codées, a été développé par l'équipe de Carapito.[128] Dans leur premier essai 84 séquences codées de plusieurs types de polymères comme les polyuréthanes qui sont décrits ci-après (paragraphe 4.2.5) ont été testées et dans chacun des cas le logiciel a réussi à les décoder très rapidement (de l'ordre de quelques millisecondes) et sans erreur. Maintenant de plus en plus de type de séquences sont déchiffrables grâce au MS-DECODER.

4.2.4. Cryptage de données

Le stockage de donnée à l'échelle moléculaire peut également être utilisé pour crypter des données pour une potentielle application dans les messages moléculaires secrets. De ce fait, une lecture trop simplifiée peut être un facteur limitant pour cette application.

Dans cette optique, la structure décrite dans le paragraphe 4.2.3 est optimisée pour permettre de crypter les informations codées sur ces séquences.[129] La différence se joue sur les nitroxydes codants, qui sont dorénavant au nombre de quatre : il y a donc deux paires de nitroxyde, les nitroxydes de la même paire ont la même masse mais des conformations différentes, par contre les deux paires ont bien une masse différente. Pour former des dyades, une phosphoramidite (a ou b) est utilisée avec un des quatre nitroxyde, il est donc possible de former huit dyades différentes mais seulement quatre ont une masse différente, mais toutes auront une conformation qui varie.

La lecture de ces nouveaux polymères ne peut donc plus s'effectuer seulement via une analyse de spectrométrie de masse en tandem car les nitroxydes ont la même masse. Il faut ajouter une analyse de spectrométrie de mobilité ionique qui sera différentes dues aux conformations différentes. Cette analyse supplémentaire permet d'identifier une clé permettant de décoder en MS/MS le message secret.

Plusieurs étapes sont donc nécessaires pour réussir à décoder le message crypté sur la séquence. Tout d'abord, il est nécessaire que la personne cryptant le code et celle le décodant se mettent d'accord sur le procédé de décryptage. Il faut que les règles de codage soient connues au préalable, c'est-à-dire la façon dont le message binaire 0/1 est défini. Ensuite, il faut analyser un lot de polymère standard sous les bonnes conditions MS/MS. Cela va permettre au lecteur d'établir ses propres références pour les analyses et qui lui permettra par la suite de trouver la clé cachée dans l'espèce nitroxyde. Ces références peuvent être rassemblées dans un tableau, dont un exemple est donné en Tableau I. 3 et qui permettra d'avoir les données de références nécessaires pour pouvoir identifier la clé. Une fois la clé trouvée, il est facile de retrouver le bon code.

Tableau I. 3: Tableau donnant les temps de dérive des différentes dyades et permettant de déchiffrer la clé de codage.[129]

Ion interne	Dyades	Temps de dérive (t_D) (ms)	Δ t_D (ms)
m/z 476.3	**0'$_b$0$_a$**	7.06	0.28
	0''$_b$0$_a$	6.78	
m/z 490.3	**1'$_b$0$_a$**	7.27	0.30
	1''$_b$0$_a$	6.97	
m/z 504.3	**0'$_b$1$_a$**	7.60	0.28
	0''$_b$1$_a$	7.32	
m/z 518.3	**1'$_b$1$_a$**	7.70	0.21
	1''$_b$1$_a$	7.49	

En effet, la séquence contrôlée d'oligo(alcoxyamine phosphodiester)s contenant l'information secrète va dans un premier temps être analysée en MS/MS pour caractériser la « séquence primaire », ensuite cette analyse est répétée mais en augmentant l'énergie d'activation et en effectuant à la suite une analyse de mobilité ionique. Les résultats obtenus pour cette seconde analyse sont alors comparés aux références, ce qui permet de déterminer quel alphabet a été employé. Ensuite, sachant les conditions d'encryptage il est facile de retrouver le message secret original.

4.2.5. Les polyuréthanes

Les polyuréthanes sont des polymères largement répandus dans plusieurs industries comme pour les revêtements, les adhésifs, l'automobile, l'isolation etc…[130] Pour ce genre d'applications une polymérisation en étape est utilisée et est largement suffisante pour obtenir les propriétés recherchées. Mais pour que ce genre de polymère soit utilisé dans le domaine du stockage d'information à l'échelle moléculaire, il faut que leur structure et la distribution de leur poids moléculaire soient beaucoup plus

précis. Des recherches dans cet esprit-ci commencent depuis plusieurs années, la plupart utilise des groupements protecteurs pour atteindre un meilleur contrôle.[131]

En 2016, la première synthèse de séquences codées de polyuréthanes sans groupement protecteur est décrite par le groupe de Lutz.[22] En plus de développer une nouvelle manière de synthétiser ce genre de macromolécules, ils montrent qu'il est également possible de stocker de l'information dans ces séquences. Le même modèle de différenciation entre les bits 0 et 1 est suivit. Elle se fait via la présence (bit 1) ou non (bit 0) d'un groupement méthyle sur les chaînes pendante d'amino-alcools. Plusieurs séquences tests sont réalisées et les analyses (SEC, HRMS et RMN) montrent l'uniformité des séquences. La fragmentation sélective des liaison carbamates C-O obtenues en mode négatif de l'analyse de spectrométrie de masse en tandem permet d'acquérir un spectre facile à interpréter. Tout comme dans les cas précédents, la différentiation des séquences isobariques se fait via la différence de masse entre les bits 0 et 1.

Grâce à leur lecture très facile ce genre de séquence trouve une efficacité en tant que code-barre moléculaire. En effet, ces séquences peuvent être implémentées au sein de différents matériaux pour permettre la lutte anti-contrefaçon. Il a notamment été montré qu'elles peuvent être incorporées dans des plastiques de commodité tels que des films de polystyrène ou des matériaux imprimés en 3D via une photopolymérisation à base de méthacrylate.[22] L'incorporation est un succès et l'analyse MS permet de décoder la séquence binaire. Pour ce faire, une petite partie du matériau est coupé et immergé dans un solvant qui est un non-solvant pour la matrice mais qui peut extraire une quantité assez importante de la séquence codée pour procéder à l'analyse en ESI-MS.

Il a aussi été étudié l'incorporation de ces code-barres dans des matériaux plus complexes, comme des implants intra-oculaires.[132] Le marqueur moléculaire est ajouté soit *in situ* directement pendant la réaction formant le réseau réticulé de méthacrylate ou alors après la réaction en utilisant une lentille déjà faite et en la gonflant et dégonflant dans du THF. Les deux voies sont satisfaisantes et n'altèrent pas la biocompatibilité ni la transparence des lentilles intra-oculaires. A nouveau, une analyse en MS/MS confirme que la séquence peut être recouvrée après l'incorporation.

Ces macromolécules ont également été utilisées comme des marqueurs *in vivo* pour l'identification d'implants.[133] Ces implants sont des films d'alcool polyvinylique (PVA) réticulés. Ils sont testés dans l'abdomen de rats où ils sont implantés en intramusculaire et en sous-cutané. Les résultats montrent que les implants ne sont pas dangereux pour les rats car les parties des rats exposés n'ont pas exhibé de réactions anormales.

Il est également montré que des séquences de polyuréthanes numériquement encodées peuvent être cristallisées.[134] Ces séquences forment des systèmes cristallins orthorhombique à base centrée via des interactions entre liaisons hydrogènes. Il est montré que le volume occupé par une unité d'information basique est de l'ordre de 148-188 $Å^3$. Ce volume est trois fois plus petit que le volume occupé par un nucléotide dans la double hélice de l'ADN.[31]

4.2.6. Les poly(N-substitués uréthane)s

Les polyuréthanes montrent donc une grande diversité d'applications. Néanmoins, la cristallisation des oligomères et leur faible solubilité dans certains solvants peut s'avérer être une limite pour certains usages.[134] Pour pallier ces limites, une stratégie alternative est développée par le groupe de Lutz en 2019. Cette dernière met en jeu des séquences codées d'oligo(N-substitués uréthane)s dont la structure type est donnée en Figure I. 23.[135]

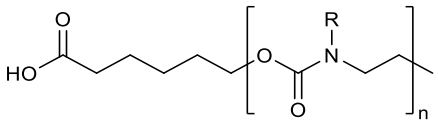


Figure I. 23: Structure d'une séquence codées de poly(N-substitués uréthane)s. Avec R = Me (00), Et (01), Pr (10), Bu (11).

Avec ce type de séquence, une meilleure solubilité est obtenue dans les solvants commun comme le THF, l'acétonitrile ou encore l'acétate d'éthyle. De plus, grâce au substituant de la fonction amine, il est possible de coder de l'information sur ces oligomères. Il suffit d'utiliser des substituant ayant une différence de masse suffisamment importante pour que le décodage de l'information puisse se faire en analyse de spectrométrie de masse. Ainsi, quatre substituants sont utilisés : un méthyle, un éthyle, un propyle et un butyle. Chaque monomère code alors pour 2 bits, à savoir Me (00), Et (01), Pr (10) et Bu (11). Des séquences allant jusqu'à 28 unités sont synthétisées.

Les analyses de spectrométrie de masse permettant de décoder les messages implémentés dans les séquences s'avèrent être très facile.[136] La fragmentation se fait au niveau des liaisons CH_2-O séparant les différents monomères. Les ions formés lors de l'analyse MS/MS sont facilement interprétable grâce à la terminaison α méthylée. Cette dernière, permet d'obtenir une couverture totale de la séquence, car aucun ion non désiré ne vient parasiter le spectre. En outre, ces séquences peuvent être analysées automatiquement grâce à l'algorithme MS-DECODER. Cet algorithme déchiffre les séquences en 100 ms.

4.2.7. Réactions multi-composants : codage avec des petites molécules

Les réactions multi-composants peuvent également servir à construire rapidement des macromolécules à séquence contrôlée. Cette stratégie consiste à faire réagir plusieurs petites molécules qui sont ajoutées à la chaîne en croissance en une seule étape. Ce type de réaction est bien connu depuis plusieurs années en chimie organique mais commence seulement à se démocratiser en chimie des polymères et plus précisément dans le domaine de polymères à séquences contrôlées.

Meier et ses collègues ont montré que la réaction 3 composants de Passerini (P-3CR) peut être utilisée dans cette voie.[119, 137, 138] Cette dernière est une réaction entre un aldéhyde ou une cétone avec un acide carboxylique et un isocyanure, formant un α-acyloxy carboxamide. Grâce à cette chimie, ils peuvent synthétiser un décamère portant dix chaînes latérales différentes en quantité importante (plus de 2 grammes), avec un rendement total de 44% après les 19 étapes nécessaires. De plus, via un groupement latéral une autométathèse peut être conduite et ainsi permettre la formation d'une séquence contrôlée de 20-mers. Toutes les séquences sont caractérisées en RMN, SEC et ESI-MS et les analyses montrent une haute pureté qui confirme la monodispersité des séquences.[139]

La combinaison de deux réactions multi-composants peut être une bonne solution pour atteindre un stockage d'information encore plus important. Dans cette optique, de nouveaux réactifs sont synthétisés via la réaction de Passerini combinée avec la réaction de Biginelli. Cette combinaison permet d'augmenter la variété de monomères possibles et d'augmenter la variété structurale obtenue.[120] Six différents composants par unités répétitives vont servir à encoder l'information. Grâce à ce design spécifique les oligomères obtenus avec cette approche ont une densité allant jusqu'à 24 bits par unités répétitives. Une analyse de spectrométrie de masse est utilisée pour lire ces séquences et retrouver l'information encodée.

A ce jour, l'équipe de Rosenstein a produit la quantité la plus importante de petites molécules codant de l'information.[140] Ils ont utilisé l'approche multi-composant avec la réaction de Ugi quatre composants de façon automatisée. Ils ont réussi à encoder 1,8 millions de bits dans ces petites molécules.

4.2.8. Les séquences multi-fonctionelles

La capacité de stockage des oligomères à séquences contrôlés basés sur la synthèse multi-fonctionnelle d'oligo(amide-uréthane)s en utilisant la chimie des thiolactones (Tla) est également explorée. [141-144] Il en résulte des séquences hautement pures et faites grâce à un système automatisé. L'équipe de Du Prez, montre qu'il est possible d'encoder de l'information digitale comme des QR codes dans de tels oligomères.[121] De plus, la lecture de ce type de séquence est facilitée grâce à la fragmentation contrôlée des liaisons uréthanes. Le stockage est effectué sur 71 oligomères différents qui sont analysés par un algorithme développé par l'équipe, nommé « Chemreader ». Celui-ci facilite la lecture de ces séquences en décodant l'information stockée dans les séquences de manière rapide, efficace et automatisée sur un ordinateur standard.

4.2.9. Les poly(succinimide thioether)s linéaires et dendritiques

La chimie click mène à des séquences de poly(succinimide thioether)s via le couplage de Michael de thiol-maléimide. Le groupe de Zhang montre en 2017, qu'une longue séquence de 128 unités peut être synthétisée en utilisant ce procédé.[145] Deux ans plus tard, ils montrent que de l'information peut être écrite dans de tels polymères et utilisant deux monomères l'un codant pour le bit 0, l'autre pour le bit 1.[146] Le monomère codant pour le bit 1 est un dithiosuccinimide qui peut être équipé de différents groupes fonctionnels, augmentant ainsi les possibilités de codage. L'information encodée est déchiffrée par une analyse de spectrométrie de masse via le clivage sélectif des liaisons S-C contenues exclusivement dans le monomère codant le bit 1. Cette fragmentation spécifique permet d'obtenir des spectres facilement interprétables. L'information encodée dans les séquences est donc facilement décryptée.

Les exemples précédents traitent tous de polymères synthétiques linéaires, mais le groupe de Zhang prouve qu'il est également possible d'utiliser des polymères fortement branchés comme des dendrimères pour encoder de l'information.[122] Son groupe utilise à nouveau les séquence de poly(succinimide tioether)s pour créer ce genre de structure. En se basant sur une organisation d'arbre binaire utilisé en informatique, ils synthétisent des dendrimères avec des structures semblables. La stratégie divergente est utilisée pour assembler les deux différents monomères codant pour les bits 0 et 1, créant ainsi un code binaire. Le décodage du message encrypté est conduit via une analyse de spectrométrie de masse en tandem (MS/MS). La fragmentation se fait au niveau de la liaison succinimide thioéther qui a une faible énergie de dissociation qui se trouve dans le design du sous monomère. Cette fragmentation de la liaison

S-C permet d'obtenir un spectre facilement lisible mais elle est non sélective et donc arrive de façon aléatoire produisant un nombre important de dendrimères possibles. Pour récupérer le bon dendrimère, ils introduisent un cryptage spécifique. C'est une règle que seul le dendrimère recherché suit. Grâce à ce nouveau paramètre, ils prouvent qu'il est possible d'encoder et de décoder des dendrimères numériquement.

4.2.10. Les polyesters

Les séquences de polyesters montrent des densités de stockage très importantes. En 2020, l'équipe de Kim décrit la synthèse d'une longue séquence de poly(phenyllactic-co-lactic acid) contenant de l'information.[147] Quatre dyades sont utilisées pour coder les données. Le mot « SEQUENCE » contenant 64 bits (8 bits par lettre) est encodé dans une chaîne unique de polymère. Ce dernier est retrouvé grâce à une seule analyse de spectrométrie de masse en tandem de MALDI-TOF. Une séquence plus longue est également synthétisée (128 unités) mais cette fois-ci l'analyse en spectrométrie de masse se réalise en plusieurs étapes. Ces polymères s'effectuent via une synthèse convergente nécessitant moins d'étapes que si la séquence avait été réalisée sur support solide. De plus, aucun réactif n'est utilisé en excès, toutes les quantités sont stœchiométriques. En outre, la capacité de stockage de telles séquences est 50% plus importante que la capacité de l'ADN. Les séquences codées de polyesters pourraient alors être des alternatives très intéressantes pour le codage d'information à l'échelle moléculaire.

4.2.11. Les peptoïdes

Les peptoïdes sont une nouvelle famille de molécules synthétisées pour la première fois en 1992 par le groupe de Zuckermann.[148] Il s'agit d'oligoamides artificiels, apparentés aux peptides mais dont les chaînes latérales sont portées par les atomes d'azote des amides et non plus par les carbones α. De nombreuses applications ont été découvertes ces dernières années pour ces nouveaux polymères biomimétiques.[149, 150] Les séquences contrôlées des peptoïdes apparaissent de plus en plus. Le contrôle de la séquence permet, par exemple, d'avoir un effet sur les propriétés thermiques et de cristallisation.[151, 152] La séquence des monomères peut également influencer le comportement en solution des polymères en affectant leur température de solubilisation critique basse (LCST).[150]

En 2020, l'équipe de Scott utilise des séquences binaires de peptoïdes pour coder de l'information.[153] Les groupements codant l'information sont situés sur les chaînes latérales des peptoïdes. Il s'agit d'une amine pour le bit 1 et d'un aldéhyde pour le bit 0.

4.3. Utilisation de la chimie de la phosphoramidite pour le stockage d'information sur des polymères à séquences définies

La chimie de la phosphoramidite montre beaucoup d'avantage pour sa facilité de synthèse, son automatisation et sa lecture qui est en plein développement. Ces avantages poussent les chercheurs à développer toujours plus les connaissances sur cette chimie. Grâce à ces recherches, il existe de nombreux exemples de séquences codées de poly(phosphodiester)s synthétisées par la chimie de la phosphoramidite de façon automatisée. Pour synthétiser des séquences codées de poly(phosphodiester)s il est important de faire attention à deux points essentiels. Dans un premier temps, il faut bien évidemment développer une synthèse menant à des séquences parfaitement contrôlées. On parle de la phase d'écriture. Dans un deuxième temps, il faut également penser à la phase de lecture. Cette phase permet de décoder et lire le message inscrit dans le polymère. Il faut que la structure de la séquence soit optimale pour permettre une lecture facilitée et recouvrer l'information.

4.3.1. Phase d'écriture : développement de l'alphabet classique

En prenant exemple sur l'ADN qui code sur quatre bases, il est possible de stocker de l'information sur des séquences de macromolécules non-naturelles. Cela est possible si la séquence est composée d'au moins deux comonomères, l'un codant pour 0 et l'autre pour 1. Ainsi, un message binaire peut être stocké à l'échelle moléculaire dans un polymère synthétique monodisperse. Parmi les différents types de polymères numériques, les séquences abiotiques de poly(phosphodiester)s préparées par la chimie itérative de la phosphoramidite, sont des structures prometteuses pour des applications de stockage d'information.[24, 154]

La Figure I. 24 explique la création des monomères utilisés pour synthétiser des polymères numériques. Pour utiliser la chimie de la phosphoramidite il est nécessaire d'avoir un monomère pouvant être impliqué dans le cycle itératif. Pour cela, il faut qu'il contienne une fonction phosphoramidite et une fonction alcool protégée par un groupement DMT. Lorsque des d'oligonucléotides sont synthétisés, on retrouve un nucléoside composé d'un sucre et d'une base nucléique entre la fonction phosphoramidite et le groupement DMT, partie verte de la Figure I. 24. Dans le cas des séquences non biologiques, le nucléoside peut être remplacé par n'importe quelle espèce chimique. Ainsi, dans les premiers travaux de notre équipe, des séquences sont effectuées en utilisant des monomères contenant une chaîne propyle entre les deux fonctions, en bleu sur la figure ci-dessous. Si cette chaîne est laissée telle quelle, elle code pour le bit 0, si elle contient deux méthyles sur la position centrale, le monomère obtenu code pour le bit 1. D'autres monomères peuvent être imaginés en choisissant d'autres R_1 et R_2.

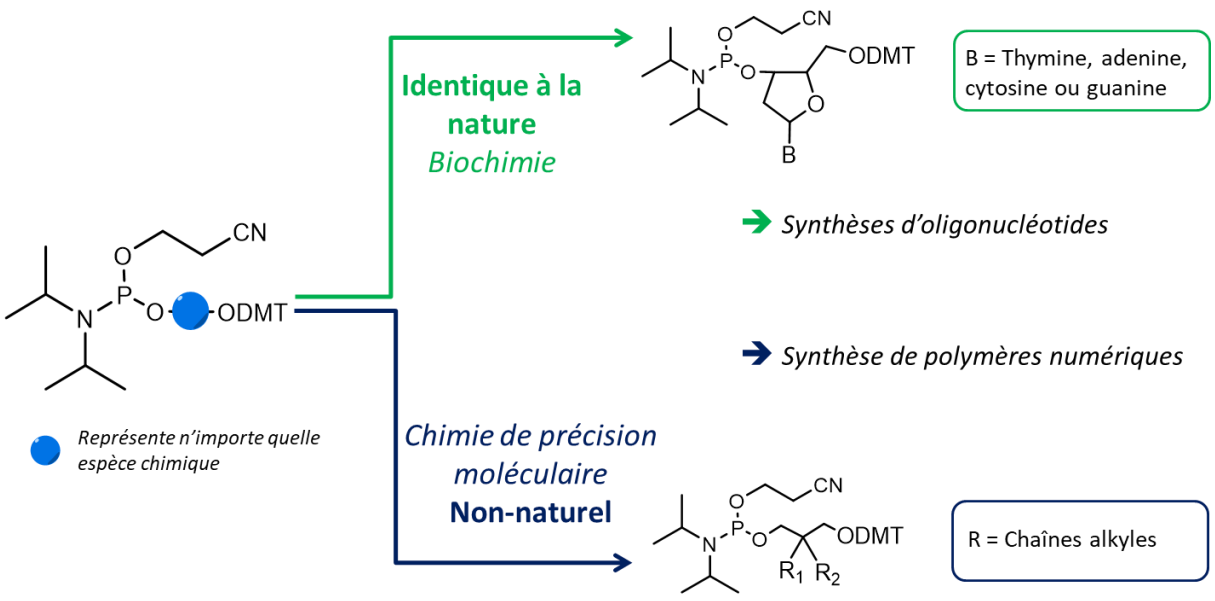


Figure I. 24: Schéma expliquant la création des monomères utilisés pour synthétiser des polymères numériques.

En 2015, le groupe de Lutz montre que de longues chaînes de polymères encodées peuvent être facilement synthétisées grâce à la chimie de la phosphoramidite, en utilisant les deux monomères cités précédemment.[23] Les séquences sont synthétisées avec un synthétiseur d'ADN, qui permet d'obtenir un polymère monodisperse contenant jusqu'à 104 monomères codants.[24] Toutes les analyses effectuées indiquent que les séquences synthétisées sont monodisperses avec une architecture contrôlée (ESI-HRMS, MALDI-HRMS, RMN et SEC).

4.3.1.1. Choix du langage de codage

Le choix du langage de codage s'est porté sur le langage ASCII « American Standard Code for Information Interchange » en anglais (Code américain normalisé pour l'échange d'information). Publié pour la première fois en 1963, ce code donne les lignes directrices à suivre pour le codage des appareils électroniques.[155] Avant cette normalisation chaque machine avait son propre langage qui était bien souvent incompatible avec celui des autres machines. A l'origine l'ASCII définit 128 caractères chacun codés en binaire, donc composé de 0 et de 1 et allant de 0000000 et 1111111. Chaque caractère est alors composé de 7 bits ; sept chiffres indiquant un 0 ou un 1. Mais depuis les années 1970 les ordinateurs travaillent le plus souvent sous des codages impliquant un multiple de huit, un huitième bit a donc été ajouté au langage ASCII, il s'agit d'un 0 en début de code. Ce huitième bit peut également servir à étendre les caractères codés. En effet, à la base l'ASCII est un langage optimisé pour la langue anglaise et permet de coder les chiffres arabes, les lettres latines en minuscules et en majuscules, la ponctuation et les opérateurs mathématiques. Mais par exemple, les lettres avec accent sont inexistantes. Ce huitième bit permet donc de les coder ainsi que d'autres caractères comme le « ç » par exemple, augmentant le nombre de caractère disponible à 256. Le codage ASCII est de nos jours très répandu et facile à utiliser grâce aux tables récapitulant le codage de chaque symbole (donnée en **partie expérimentale**). C'est donc pour cette raison que ce langage est choisi pour le stockage d'information à l'échelle moléculaire.

4.3.2. Phase de la lecture : analyse par spectrométrie de masse

L'alphabet classique se base sur deux comonomères de masse différenciée via la simple présence ou non de deux méthyles en position centrale de la chaîne propyle. Cette différence est optimale pour procéder à des analyses de spectrométrie de masse pour recouvrer l'ordre des monomères dans les séquences.[125] Les séquences sont composées d'une succession d'unités codantes séparées par des liaisons phosphates. Lors des analyses de spectrométrie de masse ESI, ces séquences sont facilement ionisables en mode négatif. Le clivage s'effectue au niveau de toutes les liaisons phosphates lors de l'activation induite par collisions.

Ceci permet d'obtenir une couverture complète des séquences lorsqu'elles sont composées de moins de 50 unités codantes. Lorsque les chaînes sont plus longues, l'interprétation des données est très fastidieuse.[125] Pour des séquences comprenant une vingtaine de monomères, la couverture de la séquence commence déjà à ne plus être complète. Les fragments obtenus avec des états de charges hauts ont tendance à avoir une dissociation supplémentaire. Cela augmente le nombre de produits secondaires observés, rendant le spectre plus difficile à interpréter. Pour les séquences plus grandes, les mêmes types de fragmentations secondaires sont observés. A cause de ces fragmentations non voulues, les spectres deviennent trop compliqués pour retrouver sans doute les séquences analysées.

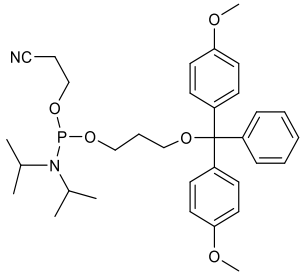
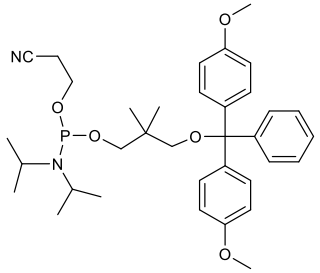
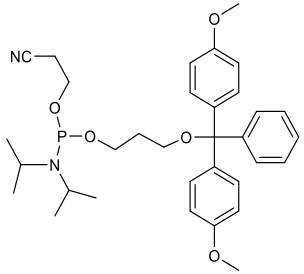
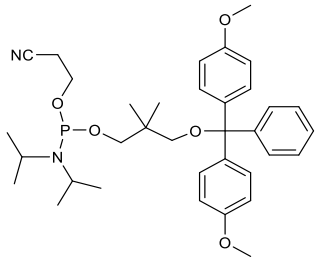
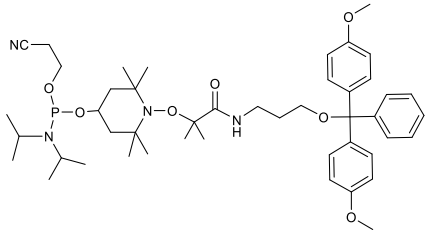
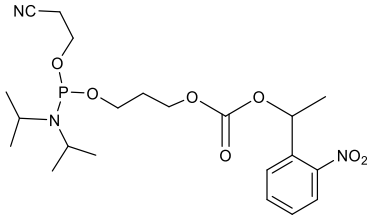
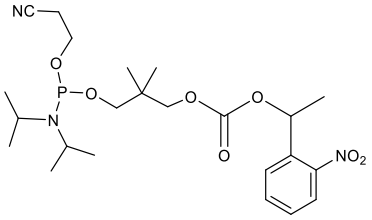
Une optimisation de la structure de la séquence de poly(phosphodiester)s est nécessaire pour espérer pouvoir analyser de longues séquences.

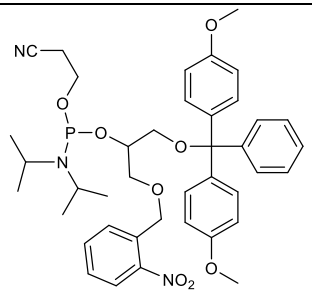
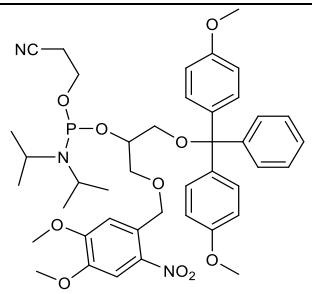
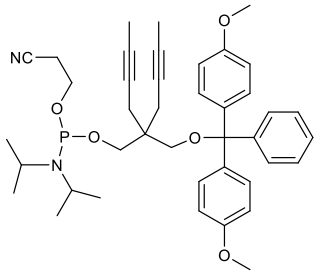
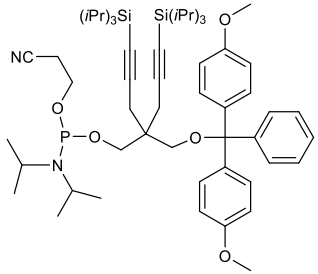
4.3.3. Développement de différents alphabets binaires

L'alphabet classique présenté n'est pas le seul qu'il est possible d'utiliser pour former des séquences de poly(phosphodiester)s codant de l'information. De nombreux développement sont possibles et peuvent apporter d'importants avantages, augmentant l'efficacité de ces séquences. Plusieurs exemples sont montrés ci-après. Le premier alphabet complexifié créé permet de simplifier la lecture des séquences de poly(phosphodiester)s. Dans cet alphabet, les monomères classiques peuvent être utilisés, mais un nouvel élément est introduit dans la séquence qui facilite la lecture lors de l'analyse MS/MS. Les autres développements jouent sur les monomères. On montre qu'il est possible d'ajouter des fonctions photoclivables qui vont, soit apporter un changement au niveau de la synthèse, soit un changement lors des analyses de spectrométrie de masse. Ces autres alphabets plus complexes sont présentés dans le

Tableau I. 4 , puis détaillés dans les paragraphes suivants.

Tableau I. 4: Liste des alphabets composés de monomère de phosphoramidite.

Nom de l'alphabet	Monomère codant pour 0	Monomère codant pour 1	Commentaires	Références
Alphabet classique			Séquence de 104 monomères synthétisée Difficile à déchiffrer en spectrométrie de masse	Paragraphe 4.3.1 23, 24
Alphabet classique avec aide à la lecture			Ajout d'un espaceur alcoxyamine qui facilite la lecture en MS/MS 	Paragraphe 4.3.3.1 25
Alphabet photocontrôlé			Utilisation du procédé itératif de chimie de la phosphoramidite photocontrôlé par le groupement NPPOC.	Paragraphe 4.3.3.2 156

Alphabet photomodifiable	 	Possibilité d'effacer, de révéler ou de changer le message encodé	Paragraphe 4.3.3.3 157
Alphabet chemo-modifiable	 	Réactions post-polymérisations possibles : cycloaddition alcyne-azoture catalysée par le cuivre (CuAAC)	Paragraphe 4.3.3.4 154

4.3.3.1. *Optimisation du design des séquences en ajoutant une fonction alcoxyamine labile*

Une optimisation de l'alphabet classique est nécessaire pour espérer pouvoir analyser de longues séquences via des analyses de spectrométrie de masse. Pour continuer à utiliser les monomères de l'alphabet classique comme monomères codants, il faut ajouter un nouveau synthon qui permet de faciliter la phase de lecture.

L'analyse des séquences d'oligo(alcoxyamines amide)s décrites dans le paragraphe 4.2.2, montre que la liaison NO-C présente dans ces structures, nécessite une très faible énergie pour se dissocier.[124, 158, 159] Ce type de liaison semble être une option intéressante pour introduire des points faibles dans le squelette des chaînes de poly(phosphodiester)s, rendant ainsi la lecture en MS/MS plus facile.

Une première structure est créée en ajoutant ces synthons alcoxyamines fragiles tous les 8 monomères codants qui vont se cliver en premier lors de l'analyse en spectrométrie de masse.[25] Après ce clivage, les fragments obtenus seront composés de petits segments d'oligo(phosphodiester)s de 8 unités qui seront facilement analysables. Ainsi, comme le montre la Figure I. 25 de façon schématique, la liaison NO-C (en rouge sur le schéma) se fragmente en premier lors de la première étape d'activation (schématisé par un éclair en jaune). Cette fragmentation spécifique permet d'obtenir un spectre de MS/MS avec chacun des octets intacts.

Par la suite, chaque octet est soumis à une seconde activation en peudo-MS^3(représentée par un éclair rose) qui fragmente au niveau des phosphates de chaque octet. Ces petites séquences composées de 8 unités sont simples à déchiffrer. Cette étape permet de recouvrer chaque monomère employé pour encoder la séquence et leur place précise. La signification de chaque octet et donc de la séquence entière est alors retrouvée.

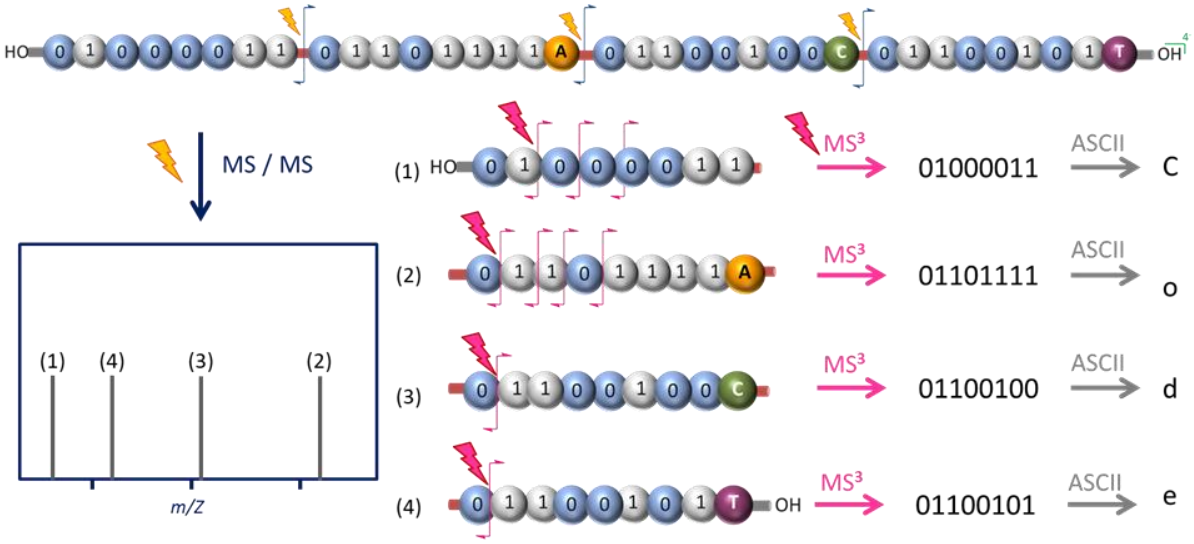


Figure I. 25: Représentation schématique du séquençage en spectrométrie de masse d'un polymère numérique contenant 4 octets d'information. Séquence codée avec 2 monomères différents. Adaptée de la littérature.[25]

Grâce à l'espaceur, l'analyse permettant de retrouver le code est beaucoup plus simple. L'analyse de séquençage en MS^3 est effectuée sur des séquences de tailles assez faibles avec des états de charges plus faibles. Les fragmentations supplémentaires des ions à haut état de charge observées sur les séquences plus longues, ne se produisent pas sur ces courtes séquences.

L'espaceur permet donc de faciliter la lecture en spectrométrie de masse, mais pour retrouver l'ordre de la séquence il est absolument nécessaire d'identifier chaque octet pour retrouver sa place dans la séquence. Ainsi, chaque octet doit contenir un marqueur le distinguant en lui procurant une masse différente. De nombreuses molécules pourraient être utilisées pour servir de marqueur, les plus simples d'usage sont les bases nucléiques naturelles A, T, G et C qui sont commerciales et peu chères. Des dérivés non naturels (appelés I et F) de ces bases peuvent également être utilisés, ils sont également commerciaux.

Pour choisir les marqueurs il faut tout de même tenir compte de certains paramètres importants : la masse molaire d'un marqueur ne doit pas être un multiple de la différence de masse entre les monomères 0 et 1. Ici, le 0 et le 1 se différencient par la présence ou non de deux méthyles, leur différence de masse est donc de 28 daltons. La masse molaire des marqueurs ne doit donc pas être un multiple de 28. Il faut également que la différence de masse molaire entre deux marqueurs ne soit pas un multiple de 28. De plus, cette différence ne doit pas être plus petite que 3 Da, car des espèces triplement chargées sont étudiées en MS^3.

Dans ce travail, une séquence contenant huit octets est synthétisée, donc huit types de marqueurs sont nécessaires. Pour le premier octet, aucun marqueur n'est utilisé ce qui permet également de le distinguer lors de l'analyse de spectrométrie de masse. L'ordre des marqueurs dans la séquence est défini arbitrairement au préalable et suit toujours le schéma :

- 1er octet : pas de marqueur
- 2ème octet : la base nucléique F
- 3ème octet : la base nucléique I
- 4ème octet : la base nucléique B
- 5ème octet : la base nucléique G
- 6ème octet : la base nucléique A
- 7ème octet : la base nucléique C
- 8ème octet : la base nucléique T

La lecture s'effectue dans le sens inverse de la synthèse, car lors de celle-ci c'est le dernier octet qui est synthétisé en premier. C'est donc toujours une thymine qui est reliée au support solide lors de la synthèse de ce type de polymère.

Grâce à cette nouvelle structure, de longues chaînes de poly(phosphodiester)s sont synthétisables. Ce design sera utilisé et optimisé davantage pour permettre de synthétiser des séquences à haute capacité de stockage.

4.3.3.2. Séquences photocontrôlées

Un deuxième alphabet complexifié est montré dans ce paragraphe. Ici, la structure des monomères classiques est changée mais la façon de coder l'information reste basée sur la différence de masse des deux comonomères. Cette différence est toujours obtenue par la présence ou non de deux méthyles en position centrale de la chaîne propyle. La différence s'effectue lors de la synthèse des séquences. La chimie de la phosphoramidite est classiquement effectuée sur un support solide et la fonction OH des monomères est protégée par le groupement DMT. Toutefois, il est montré que d'autres conditions peuvent être suivies pour synthétiser des séquences de poly(phosphodiester)s sur des micropuces notamment. Par exemple, le groupement protecteur des OH peut être changé pour un groupement photoclivable. La réaction est alors contrôlée par la lumière. En 2017, König et al. montrent que ce procédé peut s'employer pour synthétiser des polymères numériques.[156] Le groupement protecteur photoclivable utilisé est le NPPOC : 2-(2-nitrophenyl)propoxycarbonyl (structure montrée dans le

Tableau I. 4). Les séquences sont effectuées en solution. Des séquences monodisperses sont obtenues à chaque fois, les analyses MS et MS/MS l'affirmant. Cette nouvelle méthode montre que la synthèse de poly(phosphodiester)s numériques peut être effectuée en utilisant le procédé de la chimie de la phosphoramidite modifié.

4.3.3.3. Contrôle du design de la séquence via un stimulus physique

Il est décrit précédemment que la lumière peut servir à enclencher la réaction de déprotection du OH des monomères de phosphoramidite. Cette utilisation de la lumière n'est pas la seule possible. Des stimuli externes peuvent avoir un impact direct sur la structure de la séquence. Par exemple, König et *al.* montrent qu'il est possible de synthétiser des macromolécules encodées ayant des chaînes pendantes photo-contrôlables. Grâce à ces nouveaux motifs introduits dans la séquence il est alors possible d'effacer, de révéler ou de changer le message initialement contenu dans la macromolécule. Il est possible de faire ce genre de séquence en utilisant une séquence de poly(phosphodiester)s.[157] Quatre monomères différents sont utilisés, deux contenant des motifs photo-clivables et deux autres inertes face à la lumière.

La première stratégie est de montrer qu'il est possible d'*effacer de l'information* contenue dans la séquence du polymère. Dans un premier temps, un message binaire a été codé avec les deux monomères photo-clivables (les deux monomères ont des masses différentes, la séquence peut donc être retrouvée via une analyse de spectrométrie de masse en tandem). A la suite de la lecture de la séquence, la macromolécule est exposée à la lumière et de ce fait les motifs sensibles à la lumière sont clivés et une séquence non-codée contenant que la chaîne principale est recouvrée.

Dans un second temps, il est aussi démontré qu'il est possible de *révéler un message* caché dans une séquence. Cette fois-ci, un monomère sensible à la lumière et un autre inerte sont utilisés pour construire le message binaire. Mais cette séquence est impossible à déchiffrer en MS/MS car ces deux monomères ont la même masse molaire. Il faut alors exposer cette séquence à la lumière pour permettre le clivage des monomères sensibles à la lumière ce qui provoquera alors un changement de masse molaire de ce monomère et qui permettra la lecture en MS/MS.

Une troisième et plus complexe stratégie consiste à *changer le message* contenu dans la séquence. Pour cela, trois monomères sont utilisés : (a) un monomère sensible à la lumière, (b) un inerte ayant la même masse molaire et (c) un second monomère inerte contenant une chaîne latérale avec un groupement OH terminal. Les deux premiers monomères (a) et (b) sont utilisés pour coder le bit 1 et le dernier (c) code pour le bit 0. Lorsque cette séquence est exposée à la lumière, les bits 1 sensibles sont clivés et deviennent alors un bit 0, car il ne reste alors que la même chaîne pendante terminant par un OH. Le message contenu sur la séquence a donc changé et contient maintenant plus de bits 0 qu'à l'origine, ce qui change totalement le sens de l'information codée.

Il est donc montré que des stimuli externes peuvent avoir un réel impact sur la séquence et sa signification. Ce genre de propriété peut bien sûr avoir une grande utilité dans le stockage d'information qu'elle soit cryptée ou non mais également dans d'autres domaines comme la lutte anti-contrefaçon.

4.3.3.4. Contrôle du design de la séquence via une modification chimique post-polymérisation

Dans le paragraphe précédent, il est décrit que le design d'une séquence de polymère codée peut être modifié via un stimulus externe. Il est également possible de jouer sur le design de la séquence en effectuant des modifications chimiques.[154] Ces modifications ont lieu après que la séquence soit synthétisée. Les séquences initiales sont des séquences de poly(phosphodiester)s contenant les bits 0 et 1 portant des alcynes terminales ou des alcynes protégées avec le groupement TIPS. Il est ensuite choisi d'utiliser une cycloaddition alcyne-azoture catalysée par le cuivre (CuAAC) pour effectuer la modification post-polymérisation de ces précurseurs. Les alcynes codants pour le bit 0 réagissent dans un premier temps avec un azoture organique R^1N_3 pour obtenir un triazole substitué. Ensuite, le groupement protecteur TIPS des bits 1 est enlevé et les alcynes ainsi obtenus peuvent réagir avec un autre azoture R^2N_3 lors d'une seconde réaction CuAAC.

Cette approche de modification post-polymérisation s'avère être universelle pour tout type de séquence codées de poly(phosphodiester)s. Elle peut être utile pour permettre une lecture de la séquence via une autre technique que celle pour laquelle la séquence a été conçue. Par exemple, le séquençage via les nanopores qui est une technique analytique non destructive requière un design spécifique.[160] Les motifs des chaînes latérales doivent être facilement lisible plutôt que facilement clivable comme pour les analyses par spectrométrie de masse. Dans cette approche, les séquences sont soumises à un courant ionique et passent à travers des nanopores de diamètres définis. Les interactions entre les pores et les monomères vont provoquer une variation du courant qui est enregistrée. Plus les monomères provoquent une variation différente, plus il est facile de déterminer leur nature via la technique des nanopores. Il est alors facile de remonter à la séquence et à l'information qu'elle contient.

Les différents alphabets décrits dans les paragraphes ci-dessus, démontrent que la chimie de la phosphoramidite permet de synthétiser des séquences de poly(phosphodiester)s complexes. La lecture de ces polymères est effectuée via les analyses de spectrométrie de masse et montre qu'elles sont synthétisées de façon monodisperse. Le codage d'information est donc possible sur ces chaînes de polymère uniques. Pour augmenter davantage la densité de stockage, il est possible de jouer sur l'organisation spatiale des chaînes. Ce procédé est décrit dans le paragraphe suivant.

4.3.4. Augmentation de la densité de stockage via un stockage spatial

Dans les exemples cités, il est montré que de l'information peut être stockée sur des chaînes de polymères uniques. Ainsi, seuls des informations simples comme des mots peuvent être stockées sur une macromolécule. Pour augmenter la capacité de stockage (*i.e.* stocker des images ou des vidéos), il faut repenser l'organisation spatiale des oligomères, comme cela est fait avec les micropuces pour la synthèse d'oligonucléotides.[16, 86]

4.3.4.1. Assemblage couche par couche

Dans cette optique, un nouveau concept permettant d'augmenter la capacité de stockage grâce à l'assemblage non-covalent a été présentée par Szweda et *al.*[161] Une bibliothèque de 16 polyanions numériquement encodés est utilisée pour créer une structure couche par couche (Layer by Layer : LbL en anglais). Ainsi, des nano films contenant des strates de codes numériques sont fabriqués. Les polyanions sont des polyphosphodiesters préparés via le process automatisé utilisant la chimie de la phosphoramidite. Chaque strate est composée d'une séquence de 10 octets codant un texte en langage ASCII donc 80 monomères codants par strate. En combinant toutes les strates il est possible d'écrire une phrase entière qui peut être stockée dans un film multicouche.

4.3.4.2. Séquences hybrides de polymères synthétiques et d'ADN

L'ADN rencontre parfois des limites dans certaines applications. En effet, les simples interactions entre les quatre bases (A-T et G-C) peuvent rendre difficile la création de structures complexes sans utiliser un nombre énorme de séquences d'ADN. Des recherches sont faites pour essayer de rendre possible la création de structures complexes comprenant de l'ADN.

Le groupe de Sleiman montre que l'utilisation de séquences contrôlées de polymères amphiphiles appariés avec des cages d'ADN permet d'accéder à un grand nombre de nouvelles structures et d'applications.[162-164] En effet, différents assemblages sont accessibles en fonction de la taille de la chaîne de polymère utilisée mais également en fonction de l'ordre de la séquence utilisée.[163] Lorsque les polymères ont entre 5 et 9 répétitions du motif hydrophobe, deux cages peuvent interagir entre elles via ces polymères et former une structure appelée « handshake » où les deux cages se font faces et sont « attachées » via les chaînes de polymères. Lorsque les séquences sont plus longues et que tous les motifs hydrophobes sont placés en bout de chaînes, il est possible de créer des structures de « cages d'ADN en cercle ». Dans ce dernier cas, on retrouve par exemple, trois cages d'ADN structurées en forme de cercle, reliées entre elles par leurs bouts de chaînes hydrophobes. Sleiman démontre l'importance d'avoir une séquence bien contrôlée car l'ordre des monomères à une grande influence dans les structures finales obtenues. Ce genre de séquences hybrides pourrait aussi être une alternative pour stocker de l'information à la fois sur les brins d'ADN et également sur les séquences contrôlées.

Mondal et *al*. montrent que ce genre de séquence peut être réalisé à partir de poly(phosphodiester)s bio-hybrides.[165] Ces séquences contrôlées géantes sont obtenues en synthétisant des macromolécules bio-hybrides contenant un long segment non-naturel connecté à un brin d'ADN à ces extrémités. Pour former des structures linéaires, les brins d'ADN synthétisés aux extrémités des séquences sont choisis pour être complémentaires entre eux. Ainsi, en mixant et chauffant à 80°C les différents oligomères, une

hybridation des séquences d'ADN complémentaires se produit. Ceci mène alors à la création d'une séquence géante bio-hybride de 442 monomères : quatre blocs de 88 monomères non naturels connectés par des doubles brins d'ADN de 15 paires. Une étoile à 4 bras est aussi synthétisée comme preuve de concept montrant que des structures non linéaires peuvent être synthétisées. Ce genre de structure non linéaire est possible grâce à une hybridation de l'ADN bien contrôlé. De plus, tous les segments naturels sont des séquences de poly(phosphodiester)s pouvant coder de l'information. Ce genre de structure, montre à nouveau que le stockage de données peut s'effectuer sur des structures variées : de la chaîne unique de polymère linéaire, jusqu'à la structure en étoile d'une macromolécule bio-hybride.

5. Conclusion et perspectives

Ce premier chapitre a permis de faire un état de l'art du stockage d'information à l'échelle moléculaire. La recherche sur les polymères à séquences contrôlées permettant ce stockage de données est devenue un domaine d'étude majeur. Ainsi, les procédés chimiques permettant de synthétiser de telles séquences, se développent depuis plusieurs années et sont de plus en plus efficaces. Leur utilisation dans le monde industriel semble proche.

Toutes ces technologies innovantes sont inspirées des systèmes biologiques trouvés dans la nature. L'ADN et les protéines sont des exemples à suivre pour stocker efficacement de l'information à l'échelle moléculaire. Depuis leur découverte, l'Homme tente de reproduire au laboratoire ces macromolécules biologiques. La synthèse de l'ADN synthétique, optimisé depuis les années 1980, est une grande voie de recherche pour le stockage d'information à l'échelle moléculaire. Pour l'instant, aucune technologie ne permet d'accéder à des séquences aussi grandes que celles trouvées dans la nature. Par exemple, il est encore impossible de synthétiser et même de concevoir, une macromolécule synthétique aussi grande que le chromosome 1 qui contient deux brins d'ADN hybrides comprenant chacun une séquence contrôlée de 248 956 422 monomères.[166] Toutefois, les systèmes synthétiques se basant sur la structure de l'ADN sont prometteurs. De nombreuses recherches montrent qu'il est possible de stocker efficacement des données sur de l'ADN synthétique. Ces procédés artificiels semblent même pouvoir s'exporter en dehors des laboratoires et il est fortement possible de les voir apparaitre dans le commerce dans quelques années.

Ces systèmes biologiques ne sont pas les seuls en plein essor. Les systèmes non naturels trouvent leur place dans de nombreuses applications grâce à la diversité de séquences possibles de faire. Dans le domaine du stockage d'information, les séquences contrôlées de poly(phosphodiester)s semblent être particulièrement intéressantes. De nombreux systèmes utilisant la chimie de la phosphoramidite existent et sont capables de stocker une quantité de données importante.

De très longues chaînes uniques de poly(phosphodiester)s codant de l'information n'ont, cependant, pas encore été synthétisées. Au cours de cette thèse, des développements du design des séquences ont été effectués. Ces derniers ont permis d'optimiser la synthèse de longs polymères et aussi de faciliter leur lecture jusqu'à effectuer un déchiffrage automatique du message codé.

Chapitre II

Développement d'alphabets moléculaires augmentés

1. Introduction

Dans le **chapitre I**, les séquences contrôlées codant de l'information ont été exposées. Ces séquences sont pour la plupart des oligomères ayant une faible capacité de stockage. L'augmentation de la densité de stockage est un point clé pour améliorer ces séquences. Pour cela, deux paramètres majeurs sont ajustables : (i) l'alphabet utilisé et (ii) la compression de l'information. L'alphabet utilisé a un impact direct sur le design de la séquence car il joue sur les monomères utilisés. La compression d'information, quant à elle, change la façon de coder les messages en bits, mais n'a aucun impact sur l'alphabet en lui-même. Le premier paramètre s'intéressant au développement de nouveaux alphabets est expliqué dans ce **chapitre II**. La compression d'information sera illustrée dans le **chapitre IV**.

Le choix de l'alphabet est très important car il joue directement sur la densité de stockage de chaque monomère. Jusqu'à présent, seulement des alphabets à 2 symboles ont été utilisés. Dans ce chapitre, des alphabets augmentés sont étudiés. Ces nouveaux alphabets contiennent 4 ou 8 symboles. Ainsi, il est possible de stocker 2 ou 3 bits par monomère. L'alphabet à 4 symboles code les dyades (00), (01), (10) et (11). Et l'alphabet à 8 symboles code les triades (000), (001), (010), (011), (100), (101), (110) et (111). Ces nouveaux alphabets sont schématisés sur la Figure II. 1

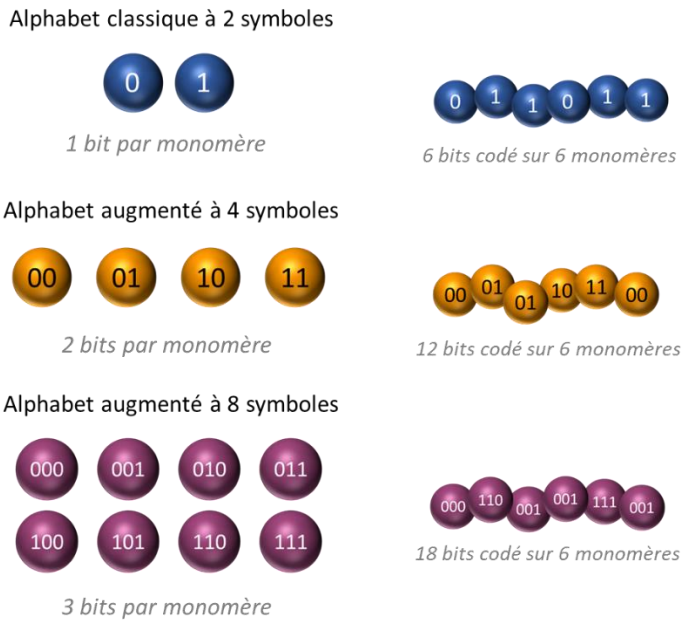


Figure II. 1 : Concept des alphabets augmentés codant sur 4 symboles (orange) ou 8 symboles (violet).

Grâce à ces nouveaux monomères codant individuellement plus de bits, il est possible de stocker plus d'information sur une même longueur de chaîne. En utilisant un alphabet classique, ne comprenant que 2 symboles qui code chacun pour 1 bit, il est possible de coder un message de 6 bits sur une chaîne comprenant 6 monomères. Cependant, en utilisant l'alphabet augmenté à 4 symboles, il sera possible de stocker un message de 12 bits sur cette chaîne de 6 monomères (2 bits/ monomères) et même 18 bits en

utilisant l'alphabet augmenté à 8 symboles (3 bits/ monomères). Cette évolution de l'alphabet permet d'augmenter considérablement la capacité de stockage des séquences.

Le **chapitre II** présente la mise en place de ces nouveaux alphabets. Dans un premier temps, il est nécessaire de sélectionner les meilleurs monomères. La chimie de la phosphoramidite automatisée est utilisée pour synthétiser les séquences, il faut donc choisir des monomères qui sont adaptables à cette chimie.

Dans une première partie du chapitre, les monomères testés seront présentés. Leurs synthèses et leur optimisation pour la chimie de la phosphoramidite seront détaillées. Ensuite, les séquences codant de l'information seront exposées. L'encodage des données, la synthèse et la récupération des informations seront exposés en détails. La phase de lecture permettant de recouvrer les informations encodées est effectuée via des analyses de spectrométrie de masse (MS). Ces analyses seront également montrées dans ce chapitre pour les séquences les plus intéressantes, les autres sont données en **partie expérimentale**.

2. Structure des poly(phosphodiester)s utilisés pour coder de l'information

Les séquences de poly(phosphodiester)s ont montré leur efficacité pour stocker de l'information au cours de différents projets effectués dans l'équipe.[23-25, 154, 156, 157] Le choix d'un alphabet permettant de synthétiser ce genre de séquence s'est donc présenté comme une évidence. Le développement d'un nouvel alphabet nécessite une bonne compréhension de la synthèse des polymères et des monomères qui le compose. Dans un premier temps, la synthèse des poly(phosphodiester)s utilisés dans ce chapitre est donc décrite en détails.

2.1. Synthèse utilisant la chimie de la phosphoramidite

Les polymères synthétisés sont tous issus d'une synthèse itérative automatisée sur support solide utilisant la chimie de la phosphoramidite, menant ainsi à des poly(phosphodiester)s.[23, 162, 164, 167] La méthode générale suivie implique quatre étapes : (i) la déprotection, (ii) le couplage, (iii) l'oxydation et (iv) le capping (expliquées en détails dans le **chapitre I**), qui sont répétées à chaque cycle jusqu'à l'obtention du polymère désiré. Pour utiliser cette méthode, il est nécessaire d'utiliser des monomères spécifiques contenant un groupement réactif phosphoramidite et un groupement 4-4'-diméthoxytrityle (DMT) qui protège l'alcool terminal. Les synthèses sont effectuées sur des supports solides en verre poreux (CPG) ou sur des résines polystyrènes en fonction du synthétiseur utilisé. Les quatre étapes pour attacher un monomère consistent en :

i. Détritylation du support (déprotection)
ii. Réaction du groupement phosphoramidite avec un groupe OH libre du support (couplage)
iii. Transformation du phosphite obtenu en phosphate (oxydation)
iv. Blocage des groupements OH n'ayant pas réagi (capping)

Les 3 premières étapes sont décrites sur la Figure II. 2 suivante.

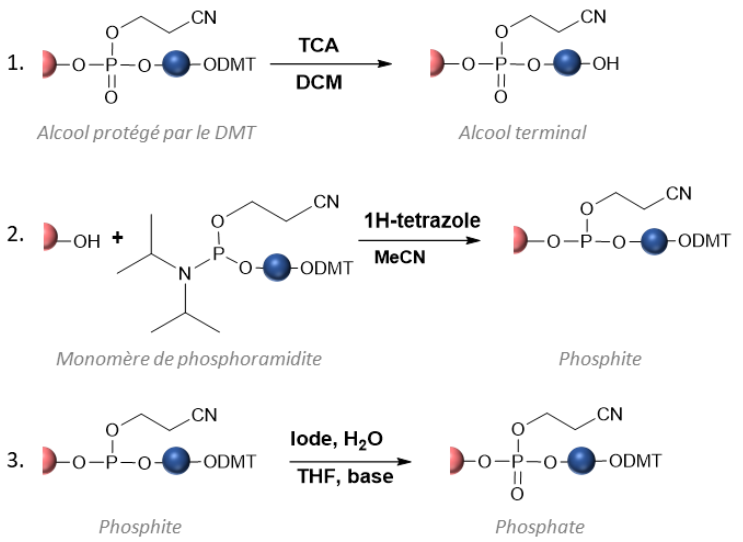


Figure II. 2 : Description des trois étapes permettant l'ajout d'un monomère sur la séquence. Le demi-cercle rose correspond à la chaîne en croissance attachée au support solide. La boule bleue correspond à la partie centrale de chaque monomère ajouté qui peut être n'importe quelle espèce chimique.

2.1.1. Synthèse des monomères

Les travaux précédents de l'équipe ont déjà porté sur la synthèse de poly(phosphodiester)s. Ainsi, deux monomères de phosphoramidite sont déjà connus et sont utilisés pour synthétiser des séquences contenant de l'information. La synthèse des nouveaux monomères des alphabets augmentés s'est basée sur celle de ces deux monomères déjà connus.

Les monomères exposés en Figure II. 3 sont les monomères dits classiques. Ils ont montré leur efficacité pour synthétiser des séquences codant de l'information. Ces monomères peuvent être utilisés dans la chimie de la phosphoramidite grâce à la présence de la fonction phosphoramidite (en orange) et au groupement protecteur DMT (en violet).[23, 24] Ils permettent de coder un message binaire en utilisant l'un pour le bit **0** (sans groupement méthyle) et l'autre pour le bit **1** (deux groupements méthyles). Une densité de stockage de 1 bit par monomère est possible lorsqu'ils sont utilisés ensemble. (Cf. chapitre I)

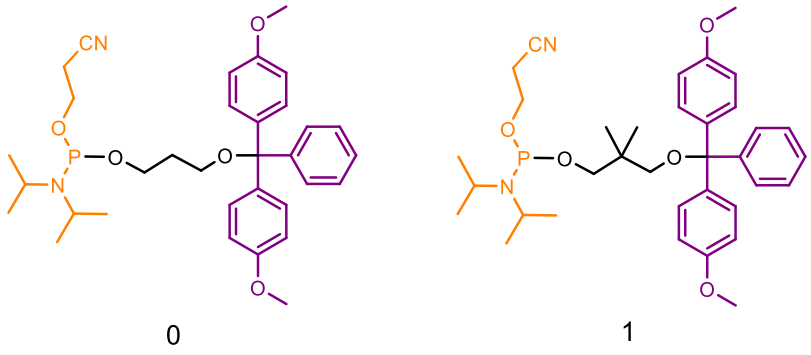


*Figure II. 3: Structure des monomères de phosphoramidite contenant le groupement réactif phosphoramidite (orange) et le groupement protecteur DMT (violet). Gauche : monomère codant pour le bit **0**, droite : monomère codant pour le bit **1**.*

Ces deux monomères sont synthétisés en suivant la procédure précédemment décrite lors de leur première utilisation.[23] Les nouveaux monomères sont synthétisés en suivant la même procédure. Leur préparation repose sur une synthèse en deux étapes à partir de diols commerciaux, montrée en Figure II. 4. Dans la première étape, les diols commerciaux sont mono-protégés avec le groupement DMT, mais les rendements obtenus ne sont généralement que de 50% environ à cause du sous-produit di-protégé également obtenu. La deuxième étape consiste à fonctionnaliser l'alcool encore libre avec une phosphoramidite (la 2-cyanoéthyle-*N,N*-diisopropylchlorophosphoramidite). Cette étape, effectuée sous atmosphère inerte, est quasi quantitative.

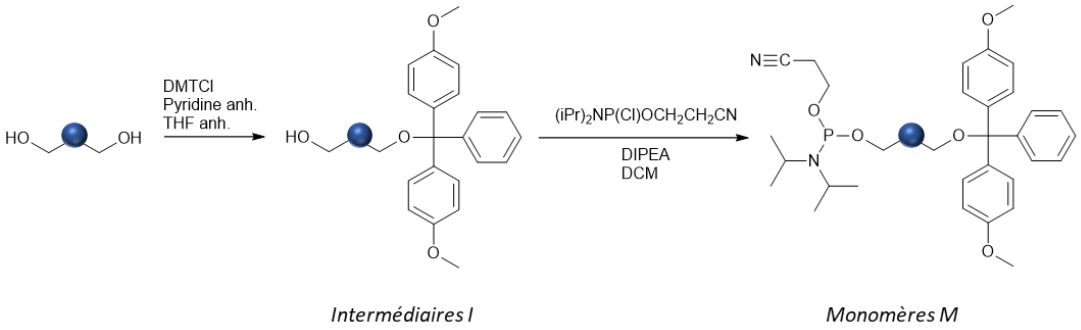


Figure II. 4 : Schéma réactionnel de la synthèse des monomères de phosphoramidite. La boule bleue correspond à la partie centrale de chaque monomère qui peut être n'importe quelle espèce chimique.

Les nouveaux monomères impliqués dans la formation des alphabets augmentés sont tous synthétisés de cette façon. Le diol commercial utilisé lors de la première étape détermine les fonctions codantes du monomère. Tous les monomères (M), ainsi que les intermédiaires (I) sont analysés en RMN ^{1}H, RMN ^{13}C et en spectrométrie de masse en ionisation par électronébuliseur (ESI-HRMS). Les conditions expérimentales et les analyses sont détaillées dans la **partie expérimentale**.

2.1.2. Synthèse automatisée

Les monomères ainsi synthétisés peuvent être utilisés dans l'élaboration de séquences de poly(phosphodiester)s. Toutes les séquences sont synthétisées sous argon dans des conditions parfaitement anhydres via une méthode automatisée en phase solide sur un synthétiseur Expedite (modèle Perseptive BioSystem 8900) ou un synthétiseur ABI (modèle Applied BioSystem 3900). La photographie des deux appareils est donnée en avec à gauche l'Expedite et à droite l'ABI. Les deux synthétiseurs diffèrent essentiellement sur la manière dont les réactifs sont amenés jusqu'au support solide.

Figure II. 5 : Photographie des synthétiseurs d'ADN de l'équipe. (Gauche) Expédite, (Droite) ABI.

Dans le cas du synthétiseur Expedite, les réactifs sont poussés jusqu'au support solide en un nombre de pulsations et un temps défini. Pour le synthétiseur ABI, les réactifs sont aspirés à travers le support, permettant une diminution des quantités engagées ainsi qu'une diminution des temps de réaction.

Deux autres paramètres sont également à prendre en compte dans le choix du synthétiseur : le suivi UV des étapes de déprotection ainsi que le nombre de séquences différentes synthétisables en parallèle. Le synthétiseur Expedite permet la synthèse de deux séquences en parallèle et un suivi UV des étapes de déprotection est fait automatiquement, donnant ainsi directement une idée du rendement de chaque étape, ainsi que le rendement global. En outre, si le rendement d'une étape calculé via la quantité de DMT déprotégée est trop faible, l'appareil arrêtera automatiquement la synthèse. Concernant le synthétiseur ABI, il n'y a pas de suivi UV mais 24 séquences peuvent être faites en même temps. Les synthétiseurs ne seront donc pas utilisés pour les mêmes raisons. L'Expedite permet d'avoir un suivi direct des synthèses, ce qui est utile lors de la synthèse de séquences avec de nouveaux monomères. Mais l'appareil ABI permet des synthèses plus rapides et moins coûteuses en réactifs. Cet appareil est donc utilisé pour la synthèse des séquences lorsque les conditions sont optimisées.

Avec les deux synthétiseurs, la première étape est de préparer les solutions nécessaires à la synthèse. Tous les monomères sont solubilisés dans de l'acétonitrile anhydre sous argon. Des solutions à 100 mM sont utilisées avec l'Expedite et des solutions à 50 mM pour l'ABI. Celles-ci sont placées sur le synthétiseur ainsi que tous les réactifs nécessaires (réactifs d'activation, solution oxydante, solutions de capping, solution de clivage du DMT). Ensuite, chaque solution est utilisée pour rincer les lignes afin d'être sûr que toutes les traces d'anciens réactifs ou d'acétonitrile présent à cette position au préalable, soit bien évacuées. Le support solide (1 µmol) est placé sur le synthétiseur, le monomère accroché au support est toujours une thymine (dT). La séquence voulue est écrite sur le logiciel. Elle est écrite dans le sens de la lecture, qui est le sens contraire de la synthèse. Ainsi, la thymine liée au support est le dernier monomère de la séquence lorsque celle-ci est lue et écrite mais le premier monomère à être inclus dans le cycle de synthèse avec la première déprotection. La synthèse est ensuite effectuée avec le mode « DMT-on », cela signifie que le dernier groupement protecteur DMT ne sera pas clivé et sera gardé en bout de chaîne.

Une fois la synthèse terminée, le support solide est enlevé du synthétiseur et la séquence protégée par le DMT est clivée du support. Une solution d'ammoniaque (28%) et de méthylamine (1/1 v/v) est introduite

dans le support, deux seringues à chaque bout servent à effectuer quelques allers-retours de la solution dans le support (voir Figure II. 6) puis celui-ci est agité dans la solution pendant 30 min à température ambiante.

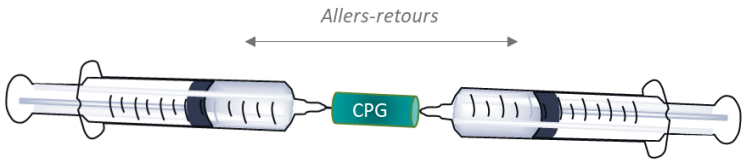


Figure II. 6 : Schéma de la procédure de clivage.

Ensuite, un kit de purification (glen-pak DNA purification cartridge) est utilisé pour effectuer la procédure finale de clivage en suivant les mêmes étapes que celles décrites au préalable.[24] Cette procédure permet de séparer les structures désirées terminant par un groupement DMT de celles tronquées et désactivées par les réactions de capping. Enfin, les groupements DMT sont clivés avec une solution acide et lavés avec de l'eau. La séquence désirée est maintenant isolée et peut être récupérée grâce à la solution d'élution. La procédure complète est décrite en **partie expérimentale**. Après lyophilisation, les polymères sont récupérés généralement sous forme de poudre blanche et peuvent être analysés.

2.1.3. Lecture facilitée par spectrométrie de masse

Les polymères synthétiques sont principalement décryptés grâce à une analyse de spectrométrie de masse en tandem (MS/MS).[168, 169] La différence de masse molaire entre les monomères est donc un outil utile pour effectuer cette analyse et recouvrer les informations encodées. Néanmoins, ce n'est pas suffisant car les spectres MS/MS des séquences standards de poly(phosphodiester)s sont compliqués à interpréter si ces séquences sont plus longues que quelques octets.

Une optimisation du design de la séquence permet toutefois de faciliter l'étape de séquençage. Cette optimisation repose sur l'incorporation d'un nouvel élément que l'on nomme élément espaceur. Cette nouvelle unité n'est pas codante mais permet une lecture par spectrométrie de masse plus facile, en privilégiant un premier chemin de fragmentation. Comme cela est expliqué dans le **chapitre I**, cette entité contient une liaison alcoxyamine qui est l'élément le plus fragile de la séquence. La structure de l'espaceur E est donnée en Figure II. 7, c'est cet espaceur qui a été utilisé lors de la synthèse des polymères présentés dans ce **chapitre II**.

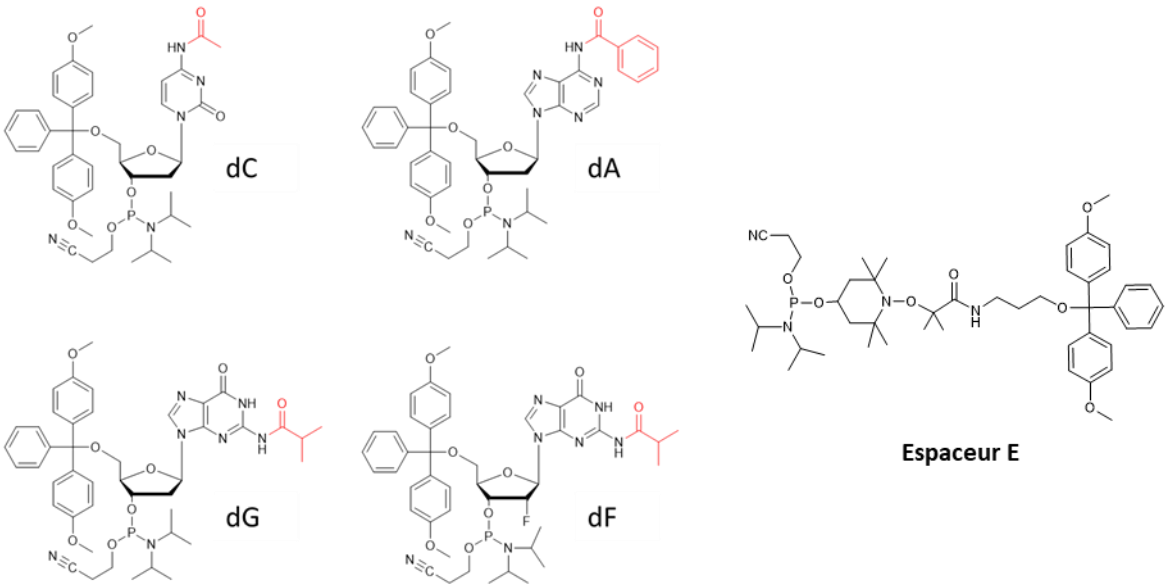


Figure II. 7 : Structure moléculaire des marqueurs de masse et de l'espaceur utilisés. Les groupements protecteurs des amines des bases sont colorés en rouge.

Cette liaison fragile est essentielle, car c'est elle qui se fragmente en premier lors de l'analyse MS/MS. En plaçant cet espaceur à des endroits stratégiques de la séquence, il est possible de prédire le premier chemin de fragmentation et ainsi d'obtenir un spectre facilement interprétable. Ici, il est choisi de placer l'espaceur tous les 8 monomères. Après la première fragmentation de la liaison alcoxyamine, des blocs de 8 unités sont donc obtenus. Ces petits blocs sont ensuite analysés un à un lors d'une analyse en pseudo-MS^3. La procédure suivie lors de cette analyse de pseudo-MS^3 est détaillée dans la partie expérimentale. Cette fois-ci, les liaisons phosphates liant les monomères codants sont clivées. Grâce à la petite taille des blocs, l'analyse est beaucoup plus simple et toutes les données sont interprétables. Une fois les informations de chaque bloc recouvrées, toutes les données sont rassemblées et le message entier est décodé.

Le décodage des données est possible seulement si tous les blocs sont lus dans le bon ordre. Pour cela des marqueurs de masse sont présents au bout de tous les blocs, ce qui permet de les reconnaitre des uns des autres. En connaissant l'ordre précis dans lequel les marqueurs sont insérés dans la séquence, il est possible de remettre les blocs dans le bon ordre. Ces entités sont généralement des nucléotides commerciaux dont leur structure est décrite en Figure II. 7. Ainsi on retrouve le nucléotide dC pour la cytosine, dA pour l'adénine, dG pour la guanine et dF pour la guanine substituée avec un groupement fluor sur le sucre. Leurs masses molaires sont différentes ce qui permet d'ajouter un poids moléculaire unique en plus au bloc. Ainsi même des blocs isobariques ont une masse bien différente grâce à la présence du marqueur de masse. Il est alors facile de les différencier lors de l'analyse. Un exemple plus précis est donné ci-dessous avec la Figure II. 8.

Figure II. 8 : Structure schématique d'une séquence contenant 5 blocs. Les blocs sont représentés par des rectangles orange, les espaceurs par des rectangles rouges et les marqueurs par des boules vertes.

Chaque marqueur a une place prédéfinie dans la séquence que le décodeur connaît, lui permettant ainsi de remettre dans le bon ordre chaque bloc analysé pour finalement retrouver la bonne information encodée. La Figure II. 8 donne un exemple de structure schématique d'une séquence contenant 5 blocs (schématisé par des rectangles orange). Lors de l'analyse MS/MS, les espaceurs (rectangles rouges) se clivent, libérant ainsi les différents blocs. Une fois les blocs libérés, les marqueurs de masse (boules vertes) servent à retrouver l'ordre initial de la séquence. Dans l'exemple ci-dessus, chaque marqueur est assigné à un bloc précis : marqueur dG au bloc 2, marqueur dA au bloc 3, marqueur dC au bloc 4, marqueur dT au bloc 5 et aucun marqueur pour le bloc 1. Cet ordre sera toujours suivi dans une séquence contenant 5 blocs.

Si une séquence de longueur *n* est synthétisée, on aura toujours le bloc *n* marqué par l'acide nucléique dT, le bloc *n-1* par dC, le bloc *n-2* par dA et le bloc *n-3* par dG. Quant au bloc 1, il ne comprend jamais de marqueur, ce qui permet également de le différencier lors de l'analyse. Les blocs supplémentaires *n-5*, *n-6* … seront marqués par de nouveaux marqueurs. Il est important que l'ordre de ces nouveaux marqueurs soit également connu par le décodeur.

Ce type de design est utilisé en combinaison des alphabets augmentés permettant ainsi de synthétiser de longues séquences codantes et facilement lisibles en spectrométrie de masse en tandem. La mise en place des alphabets augmentés est décrite dans la section suivante.

3. Conception d'alphabets augmentés

La première section de ce **chapitre II**, permet de comprendre quel type de monomères peut être utilisé pour synthétiser des séquences de poly(phosphodiester)s. Le choix le plus important reste à faire : quels diols commerciaux utiliser pour synthétiser les monomères ? De nombreuses possibilités existent et semblent convenir d'un point de vue théorique. La section qui suit détaille les différents types de monomères qui ont été testés durant cette thèse et qui ont mené à la création des alphabets augmentés à 4 et 8 symboles. Ces alphabets ont permis la synthèse de plusieurs polymères numériques à haute capacité de stockage.

De nombreuses possibilités s'offrent pour la construction des différents monomères. Il a été décidé de construire ces nouveaux alphabets en incorporant des chaînes alkyles dans la structure des monomères. Deux manières de faire sont possibles :

1) Il est possible de construire des monomères à partir du squelette propyle des monomères classiques et d'ajouter des chaînes alkyles latérales de plus en plus longues sur le carbone central.
2) Il est également possible d'augmenter la taille de la chaîne carbonée principale de chaque monomère.

3.1. *Alphabet avec codage sur la chaîne principale*

La deuxième possibilité mentionnée ci-dessus mène à l'élaboration d'un alphabet contenant des monomères linéaires. Elle est exposée dans cette partie.

3.1.1. *Synthèse des monomères linéaires*

Ce nouvel alphabet augmenté comprend 8 monomères contenant des chaînes principales linéaires de plus en plus longues. Ils sont nommés C3 à C10 en fonction du nombre de carbones qu'ils contiennent dans leur chaîne principale. Ainsi le monomère C3, contient 3 carbones dans son squelette et le monomère C10 en contient 10. Ils sont tous présentés dans la Figure II. 9.

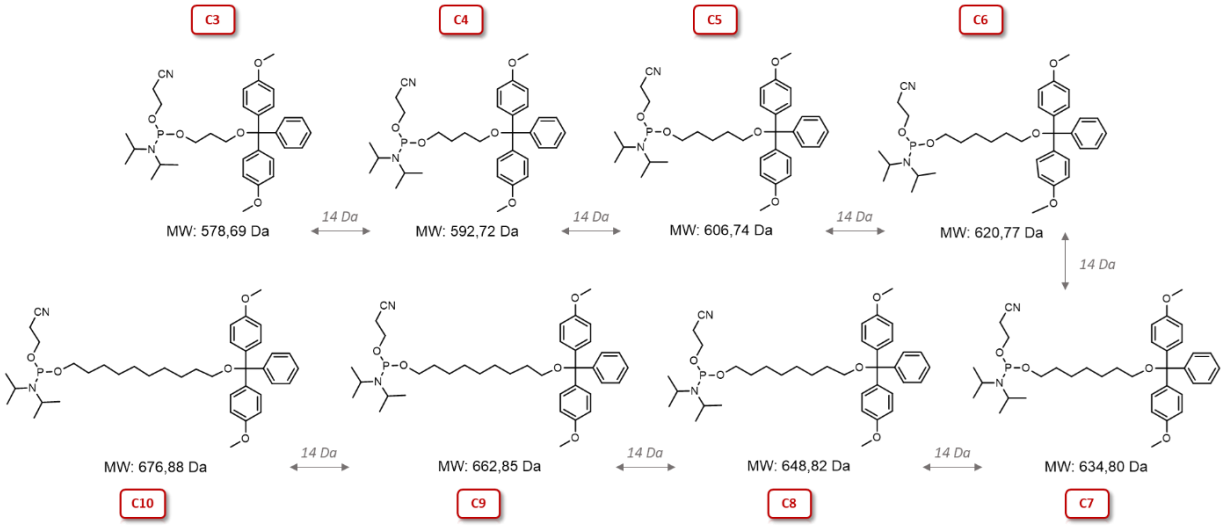


Figure II. 9 : Structure des monomères synthétisés pour l'alphabet linéaire.

Cet alphabet est composé de huit monomères différents, chacun ayant une masse molaire différente avec un incrément de 14 Da, qui correspond à la masse molaire du groupement CH_2 ajouté dans la chaîne principale. Grâce à ces huit monomères il est possible de coder des triades en utilisant tous les monomères, chaque monomère code donc pour 3 bits. Il est également possible de n'utiliser que quatre des huit monomères pour coder des dyades avec 2 bits par monomère.

Il est important de remarquer que le monomère C3 est un des monomères classiques précédemment utilisé. Sa synthèse et son efficacité en tant que monomère de phosphoramidite sont donc déjà connus.

3.1.2. *Séquences tests avec les monomères linéaires*

Les monomères ont été synthétisés en suivant la procédure montrée en paragraphe 2.1.1 (caractérisations par RMN données en **partie expérimentale**). Des premiers polymères ont été synthétisés avec ces monomères. Il s'agit de séquences simples permettant de vérifier si les monomères ont de bonnes efficacités de couplage. Deux séquences tests ont été effectuées par monomère :

1. C_x·C_x·C_x·C_x-E-C_x·C_x·C_x·C_x·A-E-C_x·C_x·C_x·C_x-T
2. C_x·C_3·C_x·C_3-E-C_x·C_3·C_x·C_3-T

La première séquence test est composée de 3 blocs de quatre fois le monomère testé avec le deuxième bloc labélisé par le marqueur dA. Tous les blocs sont séparés par un espaceur nommé E. Il s'agit de l'espaceur classique, dont la structure est montrée sur la Figure II. 7. La deuxième séquence test est composée de 2 blocs de quatre monomères avec dans chaque bloc le monomère C3. Le monomère étant déjà connu, il est utilisé comme standard pour vérifier que les nouveaux monomères peuvent se coupler avec un autre monomère et pas seulement sur eux-mêmes. Toutes ces séquences sont listées dans le Tableau II. 1 suivant, les résultats MS y sont résumés.

Tableau II. 1: Liste des séquences tests effectuées avec les monomères C3-C10 pour l'alphabet linéaire.

Monomère testé	**Séquence de monomères[a]**	***M[b] (Da)***	***Impuretés***
C4	C_4·C_4·C_4·C_4-E-C_4·C_4·C_4·C_4·A-E-C_4·C_4·C_4·C_4-T *Rmq : Trouvé en faible abondance*	3135.8180	+ 226 Da
	C_4·C_3·C_4·C_4-E-C_4·C_3·C_4·C_3-T *Rmq : Trouvé en faible abondance*	1780.4104	+231 Da + 226 Da
C5	C_5·C_5·C_5·C_5-E-C_5·C_5·C_5·C_5·A-E-C_5·C_5·C_5·C_5-T	3304.0058	+302 Da : Groupement DMT non déprotégé
	C_5·C_3·C_5·C_3-E-C_5·C_3·C_5·C_3-T	1836.4730	+302 Da : Groupement DMT non déprotégé
C6	C_6·C_6·C_6·C_6-E-C_6·C_6·C_6·C_6·A-E-C_6·C_6·C_6·C_6-T	3472.1936	+192 Da : Ajout d'un espaceur et perte de C6 -180 Da : Perte de C6
	C_6·C_3·C_6·C_3-E-C_6·C_3·C_6·C_3-T	1892.5356	+302 Da : Groupement DMT non déprotégé
C7	C_7·C_7·C_7·C_7-E-C_7·C_7·C_7·C_7·A-E-C_7·C_7·C_7·C_7-T	3640.3814	+302 Da : Groupement DMT non déprotégé
	C_7·C_3·C_7·C_3-E-C_7·C_3·C_7·C_3-T	1948.5982	+56 Da
C8	C_8·C_8·C_8·C_8-E-C_8·C_8·C_8·C_8·A-E-C_8·C_8·C_8·C_8-T	3808.5692	+302 Da : Groupement DMT non déprotégé
	C_8·C_3·C_8·C_3-E-C_8·C_3·C_8·C_3-T *Rmq : Trouvé en faible abondance*	2004.6608	-378 Da : Perte de l'espaceur - 516 Da : Perte de l'espaceur et d'un C3 -586 Da : Perte de l'espaceur et d'un C8

C9	C_9·C_9·C_9·C_9-E-C_9·C_9·C_9·C_9·A-E-C_9·C_9·C_9·C_9-T *Rmq : Trouvé en faible abondance*	3976.7570	+302 Da : Groupement DMT non déprotégé
	C_9·C_3·C_9·C_3-E-C_9·C_3·C_9·C_3-T *Rmq : Pas détectée*	2060.7234	-76 Da : Perte de l'espaceur et groupement DMT non déprotégé (-378 Da +302 Da = 76Da) - 378 Da : Perte de l'espaceur
C10	C_{10}·C_{10}·C_{10}·C_{10}-E-C_{10}·C_{10}·C_{10}·C_{10}·A-E-C_{10}·C_{10}·C_{10}·C_{10}-T	4144.9448	+302 Da : Groupement DMT non déprotégé
	C_{10}·C_3·C_{10}·C_3-E-C_{10}·C_3·C_{10}·C_3-T	2116.7860	+302 Da : Groupement DMT non déprotégé

Chaque séquence a été identifiée lors de l'analyse en spectrométrie de masse, excepté pour la séquence impliquant le monomère C9 avec le monomère C3. Pour les monomères C5, C6, C7 et C10, les ions détectés en majorité sont ceux correspondant à la séquence synthétisée, les impuretés sont mineures. Concernant les séquences pour les monomères C4, C8 et C9, lorsqu'elles ont été retrouvées, elles le sont avec une intensité semblable aux impuretés voire une intensité plus faible. Les impuretés sont majoritaires dans ces échantillons.

Dans la plupart des cas, une impureté à +302 Da a été détectée dans le premier bloc. Celui-ci est le dernier bloc à être synthétisé et comprend donc encore le groupement protecteur DMT lorsque la synthèse est finie. Ce groupement est normalement clivé lors de la purification. Mais il arrive parfois que cette purification ne soit pas efficace à 100%, ce qui est souvent le signe que la solution de clivage est trop ancienne. Lorsque le clivage du groupement DMT n'est pas effectué sur toutes les séquences de l'échantillon, on retrouve cette impureté à +302 Da. Cette différence de masse correspond exactement à la masse du groupement DMT. Si on omet cette impureté correspondant à la séquence recherchée mais pas totalement déprotégée, on peut remarquer que la plupart des séquences ont été retrouvées sans aucune autre impureté. C'est le cas pour les deux séquences obtenues pour les monomères C5 et C10, ainsi que pour la séquence synthétisée avec les monomères C6 et C3, et les homopolymérisations de C7 et C8. En tout 8 séquences sur les 14 synthétisées ont été retrouvées avec cette impureté mineure.

Néanmoins, il semblerait que trois monomères aient une moins bonne réactivité car les séquences n'ont été détectées qu'en faible quantité voire pas détectées du tout. Il s'agit des monomères C4, C8 et C9. Dans les cas des monomères C8 et C9, la première séquence ne comprenant que le monomère testé a été retrouvée mais l'impureté à +302 (groupement DMT non déprotégé) l'est également. Dans le cas des séquences comprenant également le monomère C3, d'autres impuretés ont été détectées avec à chaque fois la perte de l'espaceur. Une optimisation de l'étape de couplage semble donc nécessaire pour utiliser couramment ces monomères.

3.1.3. Séquences codant de l'information avec les monomères linéaires

Les résultats obtenus pour les séquences tests n'encouragent pas à effectuer une synthèse comprenant les 8 monomères différents. Dans un premier temps, les polymères codants de l'information ont été donc synthétisés avec un alphabet à 4 symboles en utilisant les monomères C3, C4, C5 et C6. Ainsi chaque monomère code pour 2 bits, on dit que ce sont des dyades (cf. Figure II. 1). Deux séquences ont été synthétisées. Elles sont nommées L1 et L2 et sont décrites dans le Tableau II. 2.

Tableau II. 2 : Description des séquences plus longues codées en dyades avec l'alphabet linéaire.

Signification		Séquence binaire [a]	Séquence de monomères [b]	*C* [c]	*DP* [d]	*M* [e] *(Da)*	*Impuretés*
L1	Texte ATGC	01000001.01010100. 01000111.01000011	C_4·C_3·C_3·C_4-E-C_4·C_4·C_4·C_3·A-E-C_4·C_3·C_4·C_6·G-E-C_4·C_3·C_3·C_6-T	32	22	4423,1266	+226 Da +70 Da (dG non déprotégé)
L2	Image Fiole	01111110.00100100. 00100100.00100100. 00100100.00100100. 01000010.10000001. 10000001.01111110	C_4·C_6·C_6·C_5 C_3·C_5·C_4·C_3-E-C_3·C_5·C_4·C_3·C_3·C_5·C_4·C_3·G-E-C_3·C_5·C_4·C_3·C_3·C_5·C_4·C_3·A-E-C_4·C_3·C_3·C_5·C_5·C_3·C_3·C_4·C-E-C_5·C_3·C_3·C_4·C_4·C_6·C_6·C_5-T	80	48	8794,9999	Fragments formés en source

***Rmq** : Pas détectée*

[a] Les séquences binaires ont été séparées en bloc de huit bits pour plus de clarté. Chaque bloc correspond donc à un octet. [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. Les points sont insérés entre chaque monomère et les tirets autour de l'espaceur ainsi qu'avant la thymine venant du support solide. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation incluant les monomères, l'espaceur et les marqueurs de masse. [e] Masse monoisotopique.

La séquence L1 a été retrouvée lors de l'analyse de spectrométrie de masse. Le spectre de masse enregistré en mode négatif montre l'oligomère ciblé sous les états de charges 4- à 9- (en vert, Figure II. 10). L'analyse de spectrométrie de masse en tandem confirme cette attribution et l'analyse en pseudo-MS^3 montre une couverture de chaque séquence permettant de recouvrer l'information encodée. Deux impuretés de plus faible abondance ont cependant également été détectées à +226 Da et +70 Da. Elles sont montrées respectivement en rouge et orange sur la Figure II. 10.

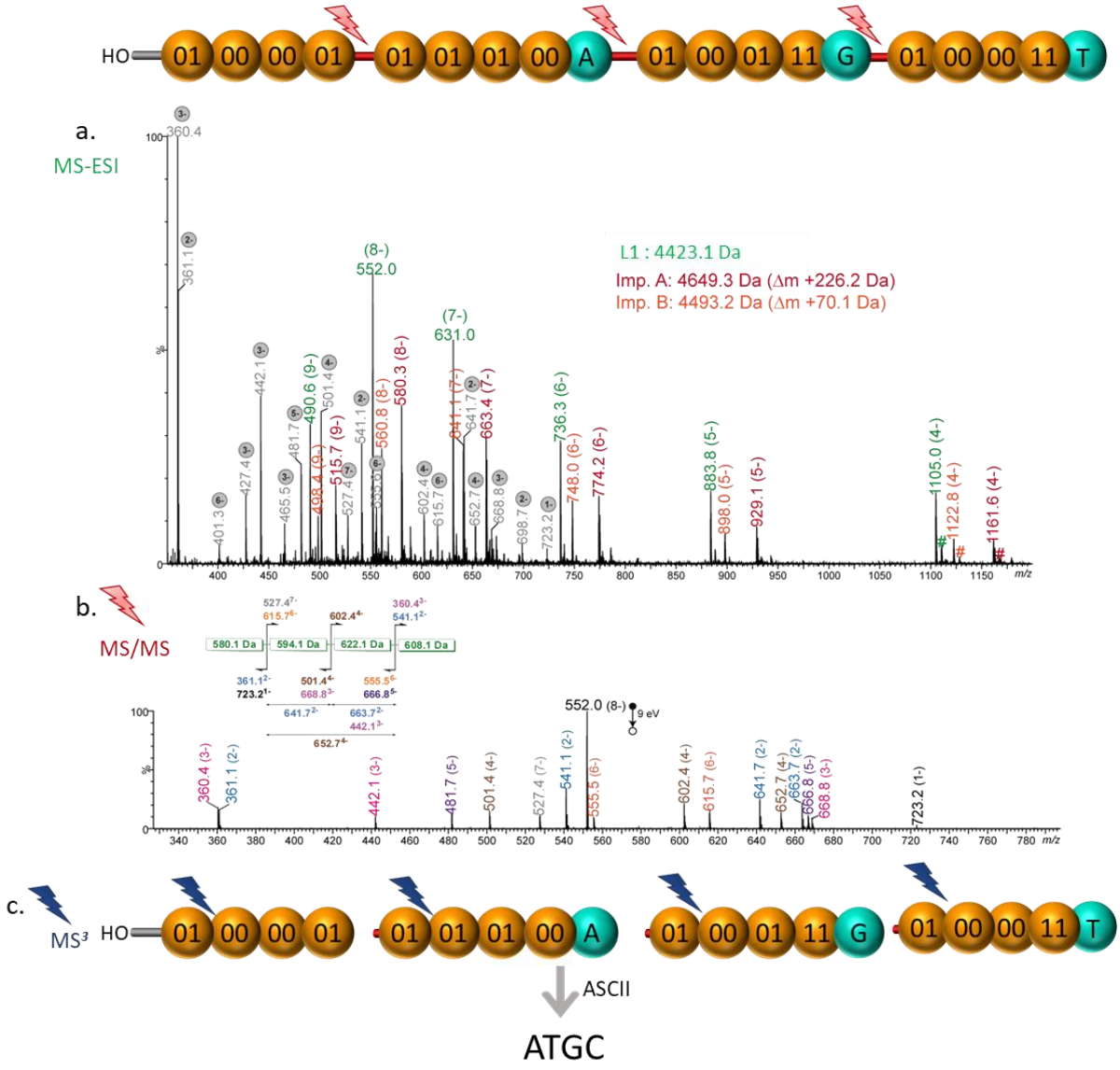


*Figure II. 10 : Schéma de la séquence **L1**. (a) Spectre ESI-MS en mode négatif de L1. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux annotés en gris à des fragments formés en source. (b) Spectre MS/MS du polymère L1 avec $[M - 8H]^{8-}$ à m/z 552,0. (c) Schéma des blocs obtenus en MS/MS analysés en pseudo-MS^3.*

Grâce à l'analyse MS/MS de l'impureté orange à +70 Da, il a été déterminé que cette dernière se trouve dans le bloc #3, celui contenant le marqueur dG. Il est remarqué que ce marqueur peut parfois être mal déprotégé. La différence de masse détectée dans l'impureté correspond au groupement isobutyryle protecteur de l'amine du cycle de la base nucléique dG qui est toujours présent. Cette impureté est facilement lavable en déprotégeant plus longtemps l'échantillon. Concernant la deuxième impureté, celle-ci ne correspond à aucune impureté connue. Elle a déjà été détectée lors des séquences tests effectuées avec le monomère C4. L'analyse MS/MS montre que cette impureté est localisée sur le bloc #2, contenant trois fois le monomère C4 à la suite. Il se pourrait que ce type d'enchaînement induise des réactions parasites responsables de l'apparition de nouveaux pics sur le spectre.

Quant à la séquence L2, elle n'a pas été pas détectée lors de l'analyse de spectrométrie de masse. Les ions détectés sont de faible état de charge et correspondent à des fragments formés en sources attendus ou non, comme on peut le voir sur la Figure II. 11. Ce résultat suggère que les polymères multi-blocs construits à partir des monomères linéaires sont trop fragiles pour survivre aux conditions d'ionisation. Ce résultat semble être en accord avec ceux obtenus pour les oligomères plus courts. Lors de l'analyse MS/MS de ces séquences plus courtes, des fragments formés en source sont observés systématiquement avec une grande intensité.

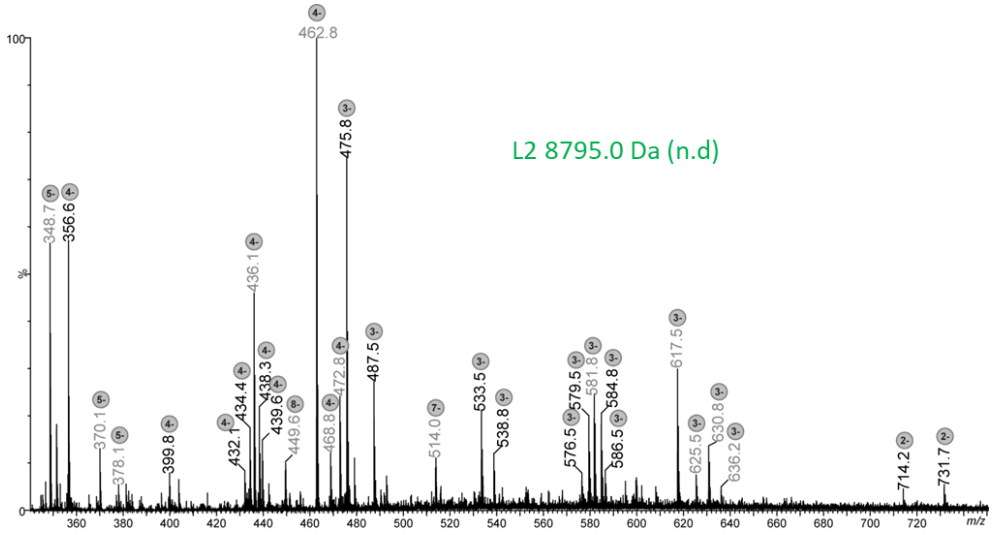


*Figure II. 11 : Spectre ESI-MS en mode négatif de **L2**. Les pics annotés en gris correspondent aux fragments formés en source attendus pour l'oligomère recherché, ceux annotés en noirs correspondent à d'autres fragments formés en source.*

Cette fragilité peut être due au fait que les différents blocs du polymère n'ont pas la même longueur. En effet, les monomères impliqués dans la séquence n'ont pas le même nombre de carbone dans leur squelette ce qui va induire des longueurs de blocs différentes en fonction de quel monomère est utilisé. Par exemple pour le polymère L2, le bloc #1 comprend une chaîne carbonée de 36 carbones pour 8 monomères, alors que le deuxième bloc comprend une chaîne carbonée de seulement 30 carbones pour 8 monomères. Cette différence peut avoir un impact lors de l'analyse de spectrométrie de masse. En effet, les états de charge ne se répartissent pas uniformément sur chaque bloc ce qui peut poser un problème lors de l'analyse en spectrométrie de masse en tandem.

Étant donné les résultats obtenus, il a été décidé de stopper l'élaboration d'un nouvel alphabet à partir des monomères linéaires. La fragilité en spectrométrie de masse des séquences obtenues est trop importante pour une application comme monomère codant de l'information via un design comprenant un fragment clivable inter-bloc. Toutefois, ce genre d'alphabet pourrait être utilisé dans d'autres applications ne nécessitant pas une lecture par une analyse de spectrométrie de masse. On peut par exemple penser à une application les mettant en jeu dans des séquences décryptées par une analyse via les nanopores.

3.2. *Alphabet à chaînes pendantes*

Les résultats obtenus avec les monomères contenant une chaîne carbonée de plus en plus longue n'ont pas été encourageants. Un autre type d'alphabet a été testé et semble montrer de meilleurs résultats. Il s'agit d'un alphabet basé sur le monomère classique contenant une chaîne propyle entre les fonctions phosphoramidite et DMT. Pour créer les nouveaux monomères, des chaînes alkyles de plus en plus longues sont ajoutés sur le carbone central de la chaîne carbonée. Ceci mène alors vers un alphabet contenant des monomères ayant des chaînes alkyles pendantes de plus en plus longues. Dans un premier temps, il a fallu déterminer quels diols commerciaux pouvaient être utilisés pour créer ce genre d'alphabet. Un premier groupe de quatre monomères a été testé, il est présenté dans le paragraphe suivant.

3.2.1. Première génération de l'alphabet à 4 symboles

Pour commencer, les monomères classiques ont été réutilisés pour créer ce nouvel alphabet. Deux autres monomères ont été recherchés pour créer un alphabet à 4 symboles. Plusieurs diols commerciaux conviennent, l'unique règle à respecter pour le moment est de trouver des diols ayant une masse suffisamment différente de ceux déjà utilisés. Pour l'instant, la différence de masse des monomères classiques est de 28 Da. Cette différence de masse peut être plus petite si nécessaire, avec un minimum de 3 Da de différence car des espèces triplement chargées sont analysées. Toutefois, une aussi petite différence de masse n'a pas été atteinte, car les monomères testés comprennent tous des chaînes alkyles plus ou moins longues. Elles diffèrent donc généralement d'au moins 14 Da, ce qui correspond au CH_2 ajouté dans la chaîne. Ainsi, un monomère contenant un seul méthyle sur le carbone central a été ajouté, de même qu'un monomère fonctionnalisé par un groupement isobutyle. Ces quatre monomères sont nommés M1 (aucune chaîne alkyle pendante), M2 (groupement méthyle), M3 (groupement di-méthyle) et M4-*iso* (groupement isobutyle) et sont présentés sur la Figure II. 12.

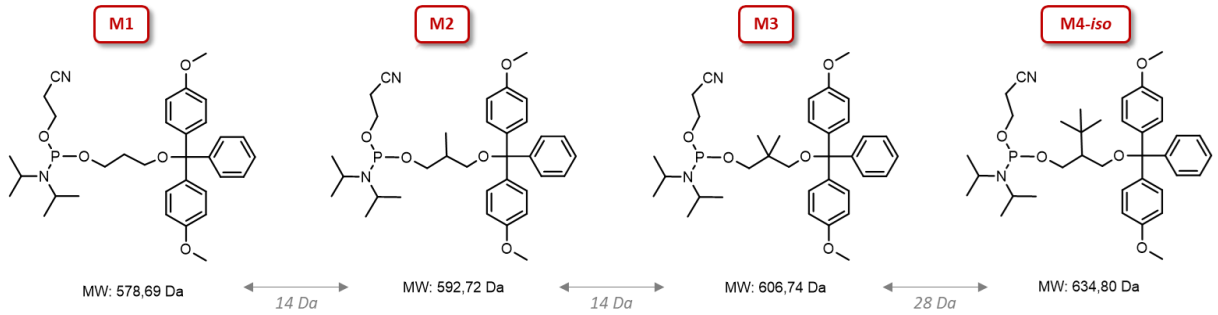


Figure II. 12 : Structure des monomères de la première génération de l'alphabet à 4 symboles.

Des séquences tests ont été menées en utilisant cet alphabet. Tout d'abord une petite séquence impliquant les quatre monomères mais ne codant aucun caractère a été synthétisée, suivi d'une séquence impliquant tous les monomères et l'espaceur. Lors de l'analyse par spectrométrie de masse, une impureté a été détectée pour la séquence avec l'espaceur. Chacune des séquences synthétisées a également bien été retrouvée sur le spectre.

Avec la validation de séquences tests, des séquences codant de l'information ont été réalisées. Les monomères M1, M2, M3 et M4-*iso* code respectivement pour les bits 00, 01, 10 et 11.

La première séquence code les initiales EL (séquence **P1**). Elle comporte deux blocs de 4 monomères et un espaceur. Le langage ASCII a été utilisé pour coder du texte (table donnée en **partie expérimentale**). Ce langage est expliqué en détails dans le **chapitre I**, et code une lettre sur 8 bits. Avec 4 monomères de dyade, il est donc possible de coder une lettre. Ainsi, chaque bloc code pour une lettre.

Ensuite, les lettres « ATGC » ont également été encodées sur une séquence de 4 blocs : une lettre par bloc, la séquence est nommée **P2**. Cette séquence étant plus longue, les marqueurs de masse dA et dC sont également utilisés. La description des séquences synthétisées est donnée dans le Tableau II. 3.

L'analyse de spectrométrie de masse de ces séquences montre qu'elles ont été retrouvées. Concernant la séquence P1, elle a été retrouvée de façon monodisperse donc l'information a pu être recouvrée sans problème. Les spectres de masse et de MS/MS sont donnés en Figure II. 13, et permettent de mettre en évidence le polymère P1 à différents états de charge en vert, ainsi que deux pics bleus correspondant à chaque bloc lors de l'analyse en tandem.

Tableau II. 3 : Description des polymères codant du texte avec l'alphabet de première génération.

Texte		**Séquence binaire** [a]	**Séquence de monomères** [b]	***C*** [c]	***DP*** [d]	***M*** [e] ***(Da)***	***Impuretés***
P1	EL	01000101. 01001100	$M_2 \cdot M_1 \cdot M_2 \cdot M_2$-E- $M_2 \cdot M_1 \cdot M_{4\text{-iso}} \cdot M_1$-T	16	10	1836.4730	/
P2	ATGC	01000001. 01010100. 01000111. 01000011	$M_2 \cdot M_1 \cdot M_1 \cdot M_2$-E- $M_2 \cdot M_2 \cdot M_2 \cdot M_1$·A-E- $M_2 \cdot M_1 \cdot M_2 \cdot M_{4\text{-iso}}$·C-E- $M_2 \cdot M_1 \cdot M_1 \cdot M_{4\text{-iso}}$-T	32	22	4411.1517	-289 Da : dC -427 Da : dC et M_1

[a] Les séquences binaires ont été séparées en bloc de huit bits pour plus de clarté. Chaque bloc correspond donc à un octet (i.e. une lettre). [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation. [e] Masse monoisotopique.

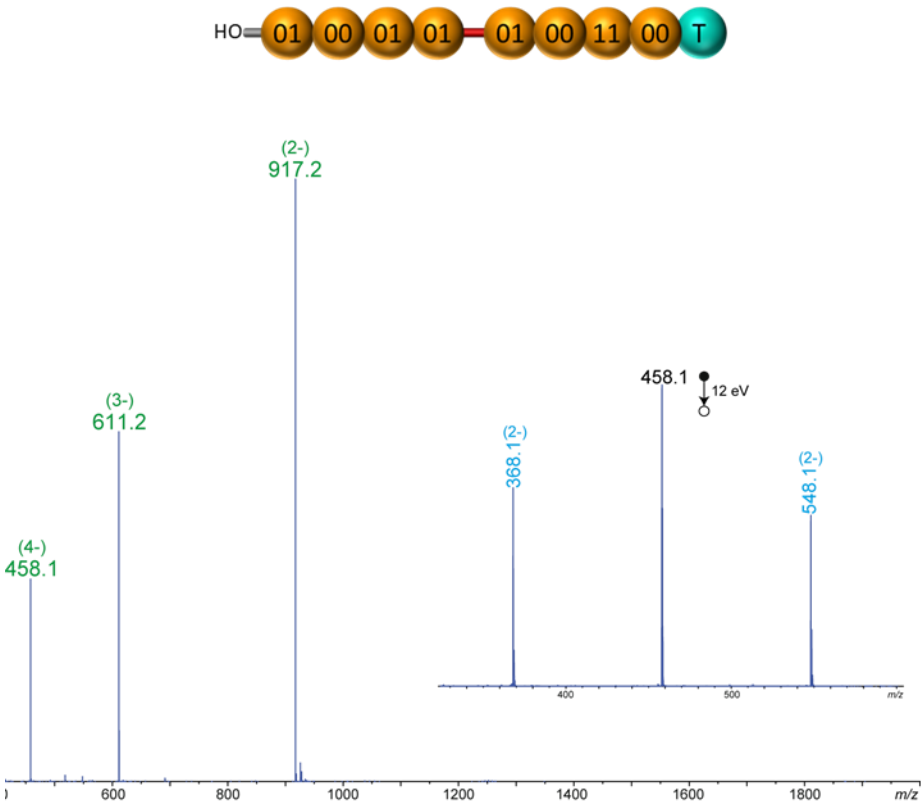


*Figure II. 13 : (En haut) Schéma de la structure du polymère **P1**. Spectre ESI-MS en mode négatif de P1. L'oligomère recherché est observé sous les formes $[M-2H]^{2-}$ à m/z 917,2, $[M-3H]^{5-}$ à m/z 611,2 et $[M-4H]^{4-}$ à m/z 458,1. Les pics au pied de l'ion m/z 917,2 correspondent à des échanges H/Na et H/K. Inset : spectre ESI-MS/MS de l'ion $[M-4H]^{4-}$ à m/z 458,1 montrant la formation des 2 fragments attendus : le bloc 1 doublement déprotoné à m/z 368,1 et le bloc 2 doublement déprotoné à m/z 548,1.*

Pour la séquence P2, deux impuretés ont également été détectées, même si la séquence désirée est majoritaire. Ces deux impuretés sont des défauts de masse de -289 Da et -427 Da, qui pourraient correspondre au manque du marqueur dC (-289 Da), et au manque du marqueur dC ainsi que du monomère M1 pour la deuxième impureté (-427 Da).

Deux séquences plus grandes ont également été synthétisées : la première est composée de 4 blocs de 8 monomères, donc 32 monomères codant pour 64 bits. La deuxième contient 7 blocs dont 6 contenants 12 monomères et un bloc contenant 6 monomères ; le tout codant pour 78 bits. Ces séquences n'ont pas été détectées lors de l'analyse de spectrométrie de masse. Toutefois, même si les macromolécules n'ont pas été retrouvées, les impuretés détectées pour la deuxième longue séquence ont été analysées. Elles apportent des éléments de compréhension importants pour optimiser la synthèse de ces longues séquences. Des signaux correspondant seulement au bloc 1 et seulement au bloc 2 ont été retrouvés. Il se peut que des réactions secondaires détériorent le polymère. Deux hypothèses sont alors avancées:

1) Les macromolécules attendues ne sont pas stables aux états de charge qu'elles doivent adopter pour être détectées, et se décomposent spontanément dans la source.
2) Le transfert des ions depuis la source vers l'analyseur est trop énergétique dans le spectromètre de masse utilisé.

Lors de l'analyse initiale le spectromètre de masse Synapt G2, QTOF a été utilisé. Cet appareil est plus résolutif et donc plus adapté pour un séquençage en pseudo-MS^3. Néanmoins, cet instrument possède une interface de transfert réputée pour être énergétique. L'échantillon a donc été analysé une nouvelle fois avec un appareil moins résolutif mais plus doux en termes de transfert des ions, le QTOF (Qstar Elite). Cette nouvelle analyse permet de vérifier si la deuxième hypothèse avancée est la bonne. Le nouveau spectre montre les mêmes signaux avec en plus une levée de ligne de base sur la gamme m/z 500-950, ce qui est typique de la présence de mélange complexe de macromolécules hautement chargées.

L'analyse en spectrométrie de masse nous montre donc que le problème vient de la synthèse de la séquence et non de l'analyse en elle-même, nous confortant sur son utilisation pour les prochaines séquences. La synthèse doit être optimisée, mais la phase de lecture effectuée en spectrométrie de masse peut continuer à être utiliser avec ce genre de macromolécule.

Ces résultats montrent que l'alphabet de première génération n'est pas optimal pour la synthèse de longs polymères. Un changement est nécessaire pour espérer atteindre des capacités de stockage plus importantes.

3.2.2. Deuxième génération de l'alphabet à 4 symboles

Un autre alphabet a donc été testé pour tenter d'achever la synthèse contrôlée de séquences plus longues. Cette fois l'alphabet est composé des mêmes monomères M1, M2 et M3 mais le monomère M4 est changé. La synthèse du monomère M4-*iso* de première génération pose des problèmes lors de sa synthèse. Les purifications sur colonnes de chromatographie sont difficiles, certainement à cause du groupement encombrant isobutyle. Un autre monomère de même masse molaire a alors été utilisé pour le remplacer, il est montré sur la Figure II. 14. Le groupement di-éthyle sur la position centrale de la chaîne carbonée est moins encombrant, la synthèse de ce monomère est alors facilitée. De par sa fonction latérale di-éthyle, ce nouveau monomère est nommé M4-*diet.*

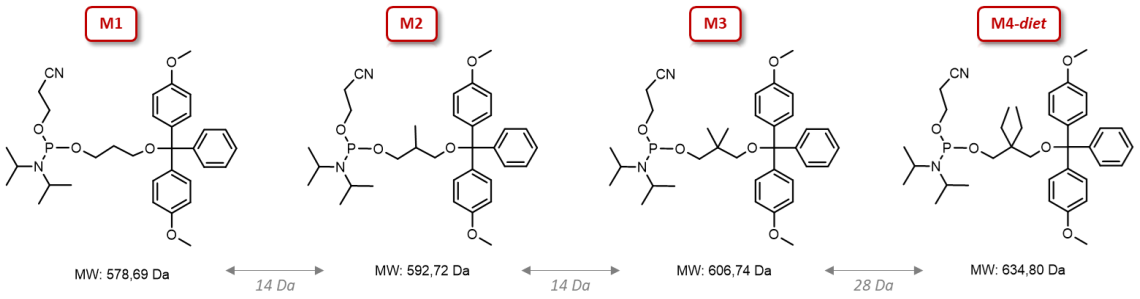


Figure II. 14 : Structure des monomères de la seconde génération de l'alphabet à 4 symboles.

Cet alphabet a été utilisé pour synthétiser deux séquences, elles sont détaillées dans le Tableau II. 4. La première séquence code pour « ATGC » (**P3**) et peut être comparée à la séquence P2 synthétisée avec l'alphabet de première génération. La deuxième séquence plus longue (**P4**) code pour le mot « CHIMIE ». Cette dernière est composée de 6 blocs de 4 monomères chacun codant ainsi pour un total de 48 bits. Pour synthétiser une séquence aussi longue, il faut utiliser 4 marqueurs de masse en plus de la thymine du support : dA, dC, dG et dF.

Les deux séquences ont été retrouvées en majorité lors de l'analyse MS, malgré une petite impureté dans chacune. L'impureté de P3 à +183, n'a pas pu être attribuée. Néanmoins, en comparaison avec la séquence P2 qui comprenait deux impuretés, cette synthèse semble plus efficace. Par ailleurs, l'impureté de P4 à -307 Da a pu être attribuée et correspond au manque du marqueur de masse dF dans le bloc 5. Son analyse en spectrométrie de masse est montrée en Figure II. 15.

Tableau II. 4 : Description des polymères codant de longues séquences synthétisées avec l'alphabet à 4 symboles de seconde génération.

Texte		Séquence binaire [a]	Séquence de monomères [b]	*C* [c]	*DP* [d]	*M* [e] *(Da)*	*Impuretés*
P3	ATGC	01000001.01010100. 01000111.01000011	M_2·M_1·M_1·M_2-E-M_2·M_2·M_2·M_1·A-E-M_2·M_1·M_2·$M_{4\text{-diet}}$·G-E-M_2·M_1·M_1·$M_{4\text{-diet}}$-T	32	22	4411.1517	+183 Da
P4	CHIMIE	01000011.01001000. 01001001.01001101. 01001001.01000101	M_2·M_1·M_1·$M_{4\text{-diet}}$-E-M_2·M_1·M_3·M_1·A-E-M_2·M_1·M_3·M_2·C-E-M_2·M_1·$M_{4\text{-diet}}$·M_2·G-E-M_2·M_1·M_3·M_2·F-E-M_2·M_1·M_2·M_2-T	48	34	7033.8316	-307 Da : dF

[a] Les séquences binaires ont été séparées en bloc de huit bits pour plus de clarté. [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation. [e] Masse monoisotopique.

Le spectre (a) de la Figure II. 15 montre les résultats obtenus lors de l'analyse du polymère P4, les pics verts correspondant au polymère recherché et les pics rouges correspondant à l'impureté. Le spectre (c) montre l'analyse de spectrométrie de masse en tandem obtenue à la suite de la sélection de l'ion précurseur à $[M\text{-}12H]^{12-}$. A cette étape tous les espaceurs sont clivés et chaque bloc est alors libéré. Ensuite, tous les blocs sont analysés un par un en analyse pseudo-MS^3. Puis, grâce au marqueur de masse il est alors possible de réordonner les différents blocs et de retrouver le message encodé.

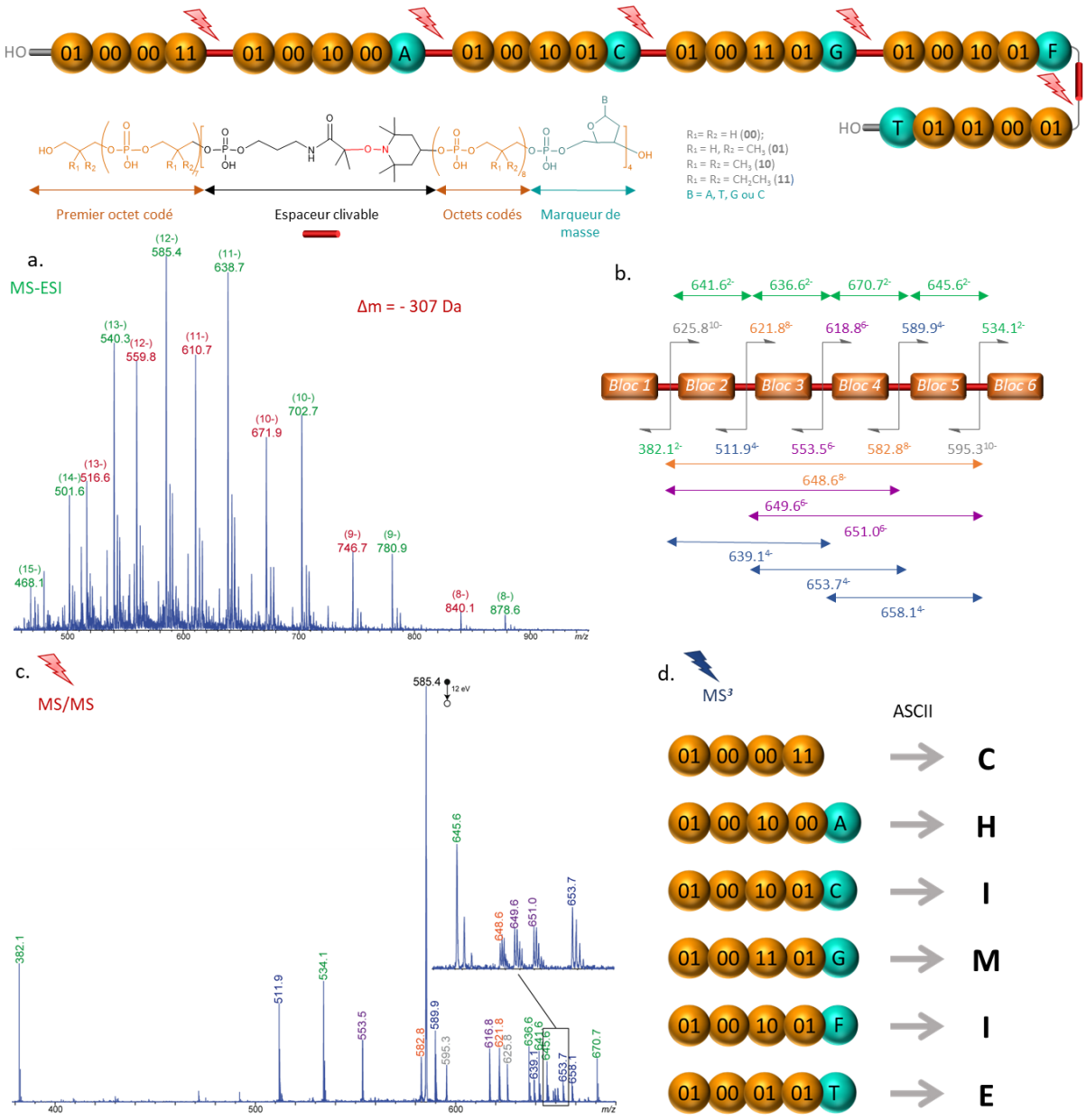


*Figure II. 15 : Analyse de spectrométrie de masse du polymère **P4**. (En haut) Schéma de la structure globale du polymère. (a) Spectre MS montrant les différents états de charges du polymère en vert et de l'impureté en rouge. (b) Schéma de fragmentation en MS/MS. (c) Spectre de masse en tandem issu de l'ion précurseur $[M-12H]^{12-}$ à m/z 584,4. (d) Blocs issus du spectre MS/MS, puis analyse MS^3 pour retrouver le texte encodé.*

L'alphabet de seconde génération donne de meilleurs résultats que celui de première génération. Le remplacement du monomère M4-*iso* par le monomère M4-*diet* est un succès. L'alphabet de première génération a été écarté des recherches et n'a plus été utilisé pour synthétiser d'autres séquences.

3.2.3. Création de l'alphabet à 8 symboles

Les résultats obtenus avec la deuxième génération de l'alphabet à 4 symboles sont encourageants. Pour continuer sur cette voie, il a été décidé de créer un alphabet à 8 symboles se basant sur ce dernier. Il est

alors nécessaire de trouver quatre nouveaux monomères compatibles avec les quatre monomères déjà utilisés. Des diols comprenant de plus longues chaînes alkyles sont recherchés. L'alphabet final est alors composé de huit monomères ayant une différence de 14 ou 28 daltons entre eux. Cette différence de masse vient de l'ajout d'un ou deux groupement(s) CH_2 sur la position centrale du squelette de la molécule. Trois monomères sont ajoutés après le monomère M4 de l'alphabet de seconde génération. Un nouveau monomère ayant une masse intermédiaire est ajouté entre M3 et M4 (14 Da de différence entre chaque monomère). La liste complète des nouveaux monomères est donnée en Figure II. 16. Ces huit monomères composent l'alphabet à 8 symboles et codent chacun pour 3 bits, ils sont présentés en violet. Les quatre premiers monomères, présentés en orange, quant à eux, composent l'alphabet à 4 symboles de troisième génération.

Par une convention arbitraire suivie dans l'équipe, le monomère le plus léger code pour le bit **0** (ou **00** / **000** ici) et le plus lourd pour le bit **1** (ou **11** /**111** ici). Ainsi, pour l'alphabet à 4 symboles, quatre monomères sont impliqués M1, M2, M3 et M4 codant pour **00**, **01**, **10** et **11** respectivement, avec M1 le monomère le plus léger et M4 le monomère le plus lourd. De même, pour l'alphabet à 8 symboles, les huit monomères M1, M2, M3, M4, M5, M6, M7 et M8 utilisés codent pour **000**, **001**, **010**, **011**, **100**, **101**, **110** et **111** respectivement avec M1 le monomère ayant la plus petite masse molaire et M8 ayant la plus haute.

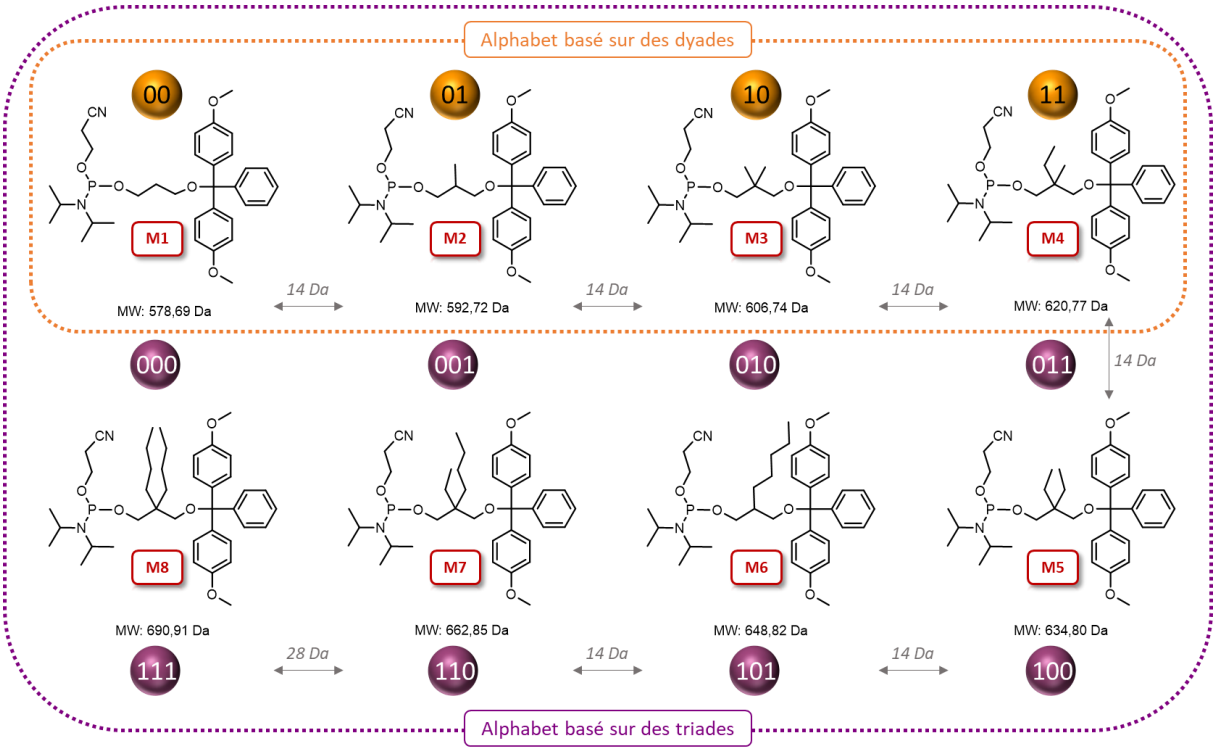


Figure II. 16 : Structures moléculaires des monomères synthétisés. Orange : monomères utilisés pour l'alphabet à 4 symboles (les monomères codent des dyades : 2 bits par unité). Violet : monomères utilisés pour l'alphabet à 8 symboles (les monomères codent des triades : 3 bits par unité).

3.2.3.1. Premiers essais de synthèses avec l'alphabet à 8 symboles

Les premières séquences effectuées avec l'alphabet triade codent pour du texte en langage ASCII. Deux séquences codant « ATGC » ont été synthétisées (voir Tableau II. 5).

Tableau II. 5: Description des premiers essais de synthèse avec l'alphabet à 8 symboles

Texte		Séquence binaire [a]	Séquence de monomères [b]	*C* [c]	*DP* [d]	*M* [e] *(Da)*	*Impuretés*
P5	ATGC	01000001. 01010100. 01000111. 01000011.*1*	$M_3 \cdot M_1 \cdot M_3 \cdot M_6$-E- $M_3 \cdot M_2 \cdot M_1 \cdot M_8$·G-E- $M_3 \cdot M_1 \cdot M_8$-T	32+1	15	3266.0864	-208 Da : M_6 -250 Da : M_8 -458 Da : M_6 + M_8 -708 Da : M_6 + 2x M_8
P6	ATGC	*0*.01000001. 01010100. 01000111. 01000011	$M_2 \cdot M_1 \cdot M_2 \cdot M_3$-E- $M_6 \cdot M_1 \cdot M_5 \cdot M_4$·A-E- $M_6 \cdot M_1 \cdot M_4$-T	32+1	15	3165.9976	+216 Da

[a] Les séquences binaires ont été séparées en bloc de huit bits pour plus de clarté. [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation. [e] Masse monoisotopique.

Le langage ASCII code une lettre sur 8 bits, ici il faudra donc 32 bits pour coder les quatre lettres ATGC. Or, avec l'alphabet triade, un monomère code pour 3 bits. Il faudra donc ajouter un bit non codant en début ou fin de message pour coder un mot en langage ASCII. Pour le polymère **P5**, le bit **1** a été ajouté en fin de séquence. L'analyse de spectrométrie montre le polymère recherché en très faible abondance, et quatre impuretés ont été détectées en plus grande quantité comme le montre la Figure II. 17.a. Chaque impureté correspond au signal d'une séquence où il manque un ou plusieurs monomères. Il s'agit des monomères M6 et M8. Le monomère M7 n'étant pas impliqué dans la séquence, ces deux monomères sont les plus encombrés de la séquence car ils comprennent une chaîne pentyle (M6) et un groupement di-butyle (M8). Leur réactivité a besoin d'être étudiée pour vérifier l'efficacité des couplages.

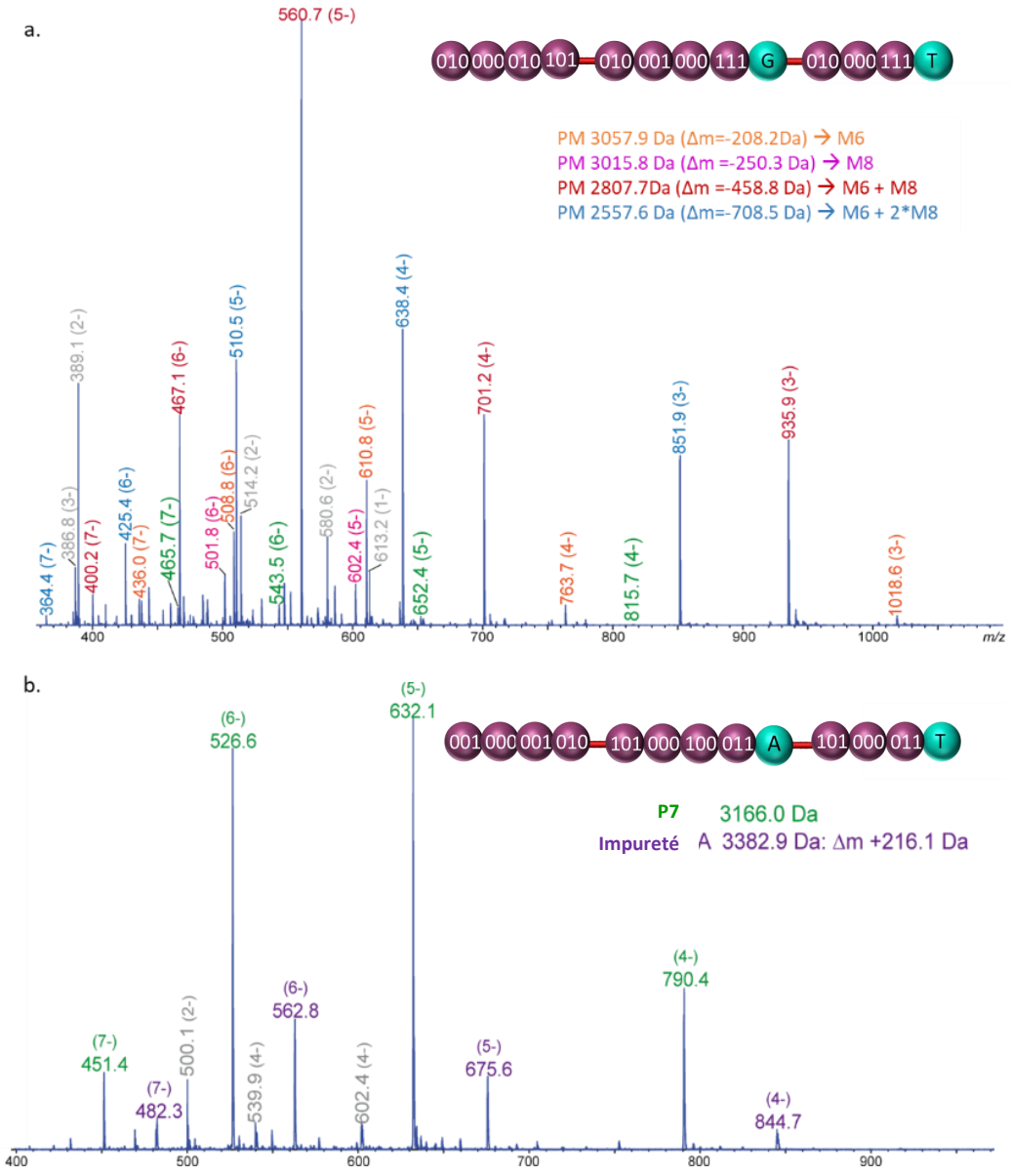


*Figure II. 17 : Spectres de masse obtenus pour les polymères **P5** (a) et **P6** (b).*

Un autre test a également été effectué avec le polymère **P6**, ici le bit **0** non codant est ajouté en début de séquence. Une impureté mineure (non attribuée) a été détectée lors de l'analyse de cette séquence, mais la séquence a été retrouvée en majorité. Les différents états de charges de la séquence sont annotés en verts sur la Figure II. 17.b.

L'ajout d'un bit non codant en début ou fin de séquence n'est pas viable pour une application telle que le stockage d'information. Un autre langage est donc utilisé pour remplacer le langage ASCII qui ne convient pas pour les monomères codant des triades. Il s'agit du langage SIXBIT, qui, comme son nom l'indique, code sur 6 bits et qui est donc plus adapté aux monomères triades (voir table en Tableau II. 6).

Tableau II. 6 : Table du langage SXBIT.

2nd bits / 1ers bits	0 000	0 001	0 010	0 011	0 100	0 101	0 110	0 111	1 000	1 001	1 010	1 011	1 100	1 101	1 110	1 111
00	SP	!	"	#	$	%	&	‘	(	)	*	+	,	-	.	/
01	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
10	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
11	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_

A nouveau, une séquence codant les lettres ATGC a été synthétisée (**P7** présentée en Tableau II. 7). L'analyse montre le polymère recherché en très faible abondance et sept impuretés ont aussi été détectées. La plus abondante correspond à une séquence dont le monomère M8 manque. Encore une fois, l'utilisation des monomères les plus encombrés semble mener à des synthèses de macromolécules polymoléculaires.

Tableau II. 7 : Description du polymère codant du texte avec l'alphabet triade et le langage sixbit.

Texte		Séquence binaire [a]	Séquence de monomères [b]	*C* [c]	*DP* [d]	*M* [e] *(Da)*	*Impuretés*
P7	ATGC (sixbit)	100001.110100. 100111.100011	$M_5 \cdot M_2 \cdot M_7 \cdot M_5$-E-$M_5 \cdot M_8 \cdot M_5 \cdot M_4$-T	24	10	2200.8799	7

[a] *La séquence binaire a été séparée en blocs de six bits pour plus de clarté. Chaque bloc correspond donc à une lettre.* [b] *Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté.* [c] *Capacité de stockage en bit/chaîne.* [d] *Degré de polymérisation.* [e] *Masse monoisotopique.*

Il semble donc que les monomères impliqués dans une même taille de séquence aient une grande importance. En effet, en comparant les polymères P5 et P6 qui ont la même taille (DP 15), on observe des résultats différents. Cette différence semble venir des monomères impliqués dans les séquences, et plus particulièrement des monomères encombrés. Dans le cas du polymère P6, il n'y a que deux fois un monomère encombré (M6), alors que pour le polymère P5, il y a trois monomères encombrés (M6 et deux fois M8). Il apparaît que la synthèse des séquences comprenant plus de monomères encombrés est plus difficile. Elle mène vers des échantillons non monodisperse, présentant plus d'impuretés. De plus, même dans des séquences plus petites comme P7 à DP 10, pour lesquelles on pourrait croire que les synthèses sont plus faciles, on observe des impuretés. A nouveau, il semble que ces impuretés ont pour origine les monomères encombrés (M6 à M8).

Pour pouvoir synthétiser des séquences comportant plusieurs monomères encombrés, il faut dans un premier temps améliorer les conditions de couplage de ces monomères. Une évaluation de l'efficacité de couplage des monomères encombrés est donc nécessaire. Cette évaluation est présentée dans le paragraphe suivant.

3.2.4. Evaluation de l'efficacité de couplage des monomères encombrés

3.2.4.1. Séquences tests de pentamères évaluées en spectrométrie de masse

Des séquences tests ont d'abord été effectuées pour les monomères M2 et M4 à M8 pour juger leur efficacité avec la chimie de la phosphoramidite automatisée. M1 et M3 étant déjà largement utilisés dans l'équipe, il n'était pas nécessaire d'effectuer ces tests avec eux. Une série de pentamères a été réalisée en utilisant la procédure classique, puis chaque oligomère a été analysé en ESI-HRMS. Les résultats sont regroupés dans le Tableau II. 8 et tous les spectres obtenus sont donnés dans les figures qui suivent.

Tableau II. 8: Caractérisation des oligomères modèles.

Séquence	*M* [a] [Da]	m/z_{th} [b]	m/z [c]
M_2·M_2·M_2·M_2·M_2-T	1002.2095	500.0975	500.0971
M_4·M_4·M_4·M_4·M_4-T	1142.3660	570.1757	570.1768
M_5·M_5·M_5·M_5·M_5-T	1212.4433	605.2149	605.2152
M_6·M_6·M_6·M_6·M_6-T	1282.5225	640.2540	640.2559
M_7·M_7·M_7·M_7·M_7-T	1352.6007	675.2931	675.2925
M_8·M_8·M_8·M_8·M_8-T	1492.7573	745.3714	745.3725

[a] Masse mono-isotopique. [b] m/z théorique. [c] m/z mesuré en mode négatif lors d'une ESI-HRMS comme des espèces $[M–2H]^{2–}$.

Les résultats pour M2, et M4 à M7, montrent majoritairement les oligomères recherchés. Les séquences pour M2, M6 et M7 ont été retrouvées de façon monodisperse comme le montre les spectres présentés en Figure II. 18. Concernant les monomères M4 et M5, les résultats montrent principalement les pentamères recherchés, mais également des oligomères avec un degré de polymérisation plus petit, témoins d'un couplage raté. (Cf. Figure II. 19).

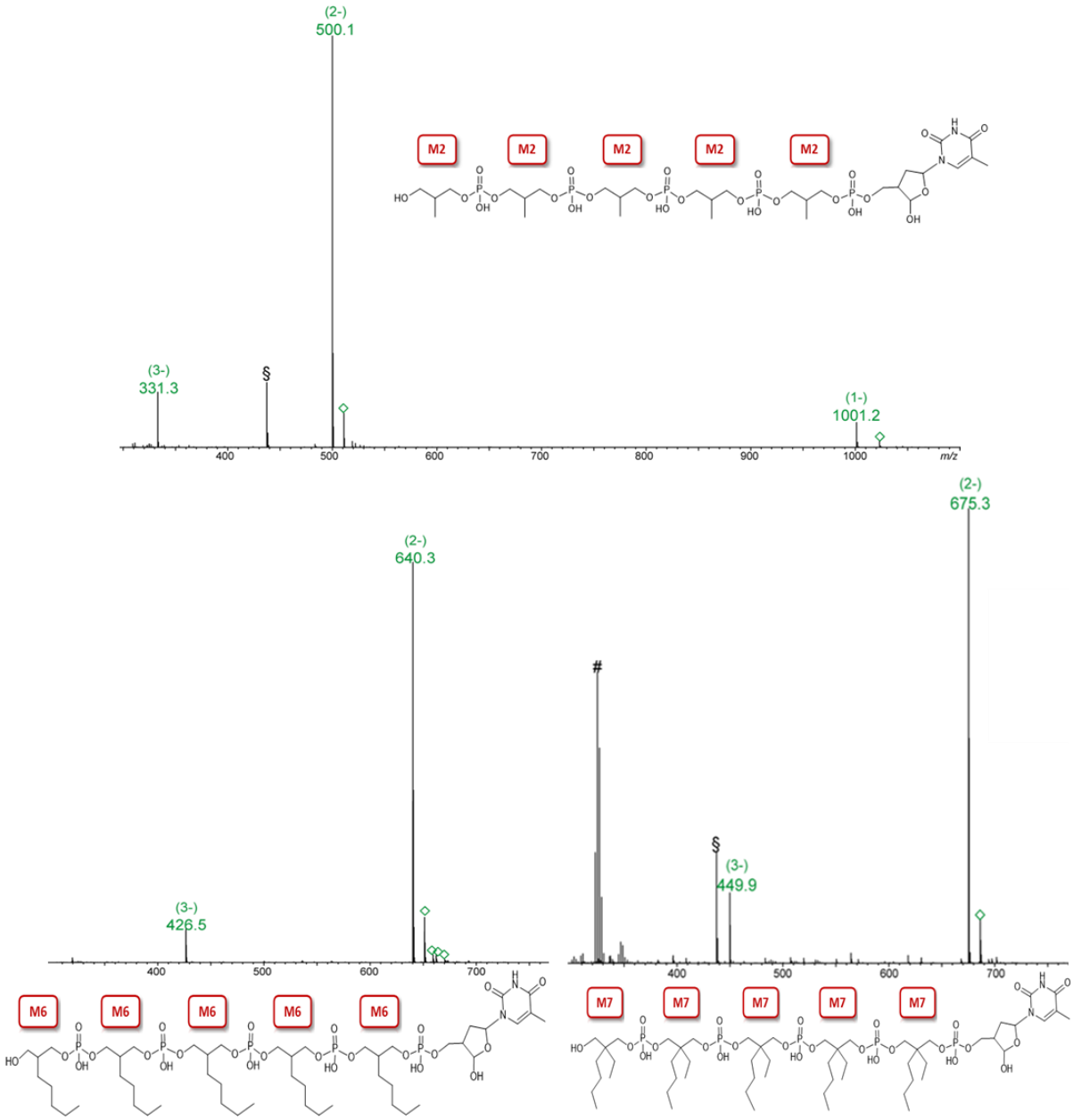


Figure II. 18 : Spectre de ESI-MS en mode négatif des pentamères modèles M2.M2.M2.M2.M2-T (en haut), M6.M6.M6.M6.M6-T (à gauche) et M7.M7.M7.M7.M7-T (à droite). Les états de charges sont annotés entre parenthèses. Les diamants désignent des formes ioniques détectées après des échanges H/Na ou H/K. §: bruit de fond. # Amas d'acide trichloroacétique.

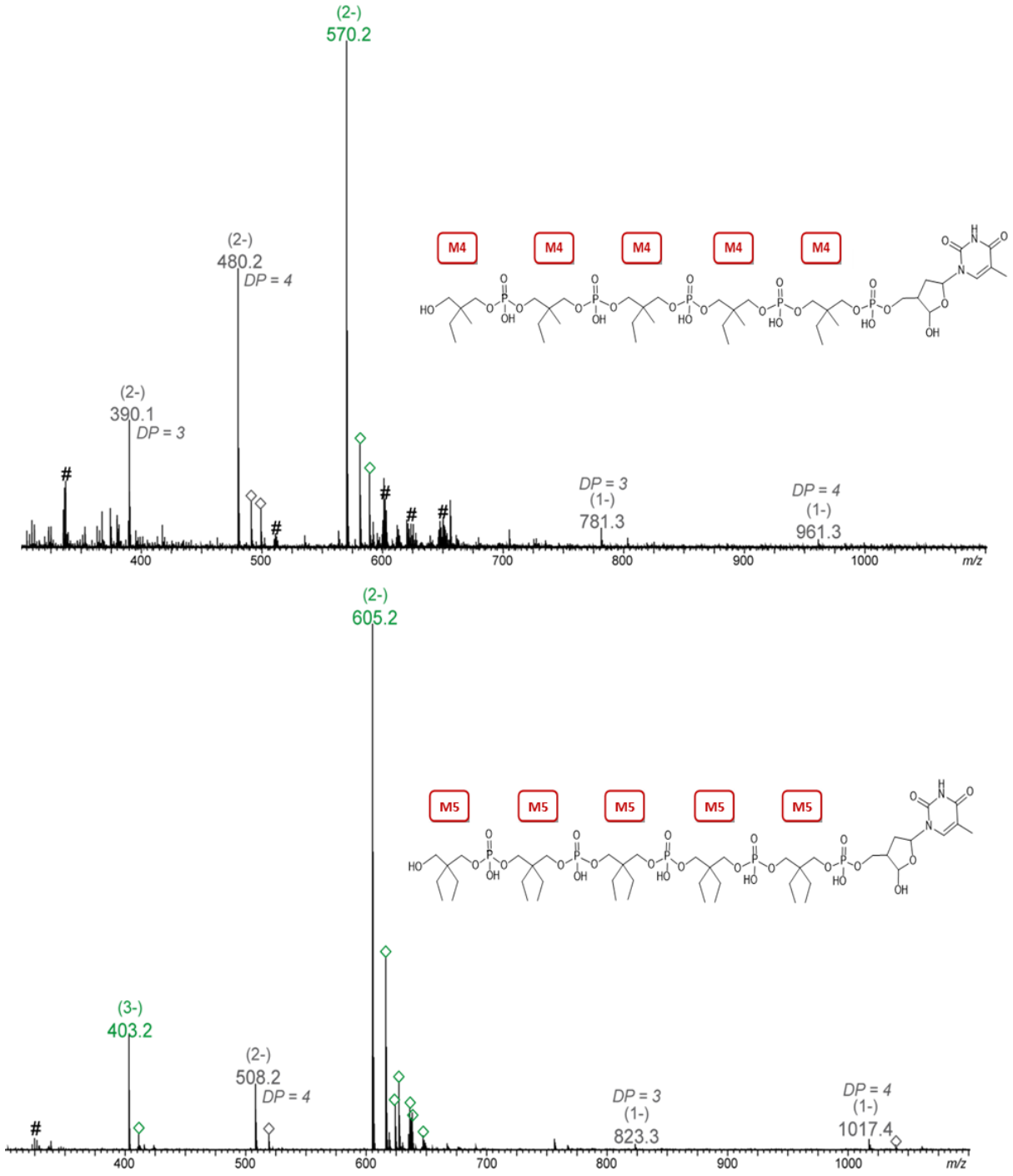


Figure II. 19 : Spectre de ESI-MS en mode négatif du pentamère modèle M4.M4.M4.M4.M4-T (en haut) et M5.M5.M5.M5.M5-T (en bas). Les états de charges sont annotés entre parenthèses. Les signaux en gris correspondent aux oligomères de plus faible DP. Les diamants désignent des formes ioniques détectées après des échanges H/Na ou H/K. # Amas d'acide trichloroacétique.

Concernant le monomère M8, qui est le monomère le plus encombré à cause de ses deux longues chaînes butyles latérales, les résultats sont moins bons. La structure recherchée a été retrouvée mais en faible

quantité et plusieurs impuretés qui sont des oligomères ayant un DP plus faible sont présents en plus grande quantité (visible sur la Figure II. 20).

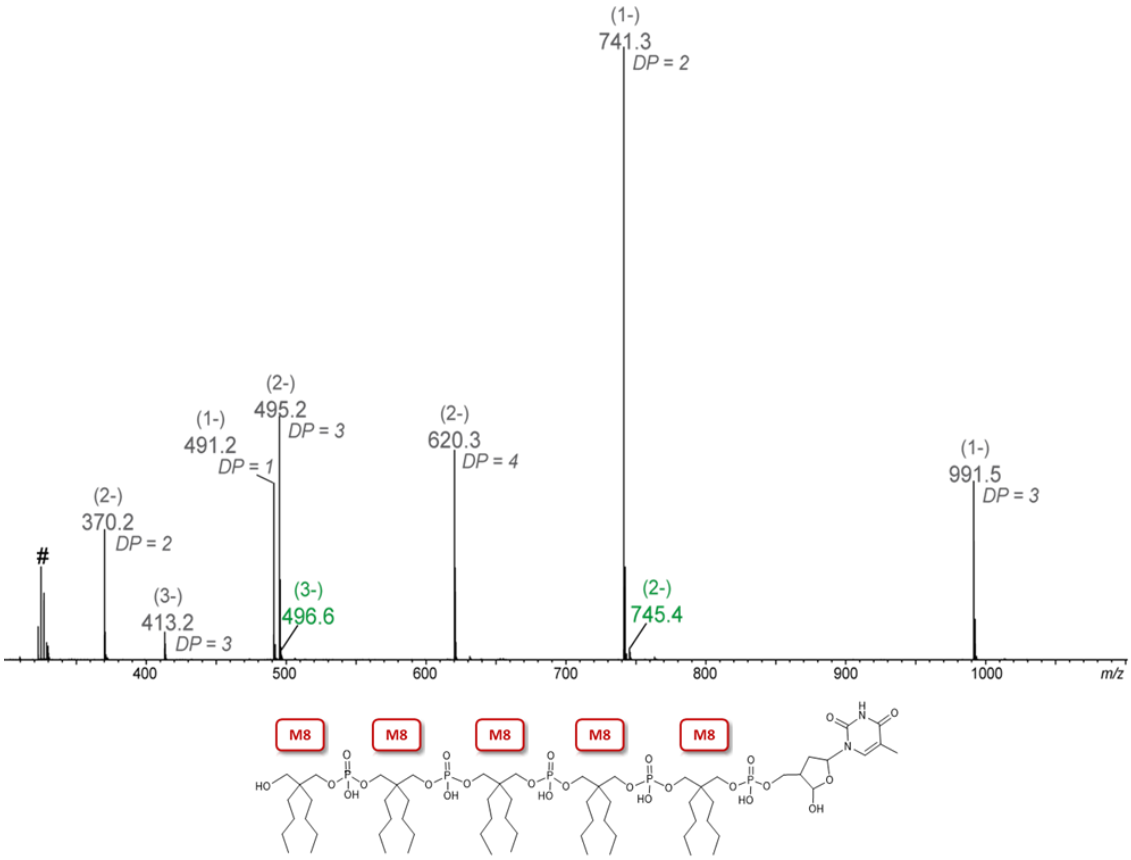


Figure II. 20 : Spectre de ESI-MS en mode négatif du pentamère modèle M8.M8.M8.M8.M8-T avec les états de charges annotés entre parenthèses. Signaux en gris correspondent aux oligomères de plus faible DP. # Amas d'acide trichloroacétique.

Les résultats obtenus grâce à l'analyse des pentamères tests nous confirment que le monomère M8 a une efficacité de couplage moins bonne que les autres monomères. Tous les autres monomères ont été retrouvés en majorité, alors que M8 a été retrouvé en faible quantité. Il a cependant été décidé de continuer d'effectuer des essais avec le monomère M8. Ce monomère est issu d'un diol commercial peu cher qui suit le design des autres monomères utilisés qui comprennent des chaînes pendantes. Trouver un autre monomère efficace et ayant la bonne différence de masse molaire aurait été possible mais aurait également pu être très chronophage. Il a donc été décidé d'effectuer, dans un premier temps, des essais d'améliorations de l'efficacité du monomère M8.

Des améliorations peuvent être apportées, mais une étude plus approfondie sur l'efficacité de couplage des monomères est dans un premier temps mené.

3.2.4.2. Etude de l'efficacité de couplage par spectroscopie UV

Il est connu que l'encombrement stérique peut influencer l'efficacité de couplage des monomères de phosphoramidite.[80] Par exemple, les monomères d'ARN ont besoin d'un temps de couplage plus

important que les monomères d'ADN à cause de l'encombrement causé par les groupements protecteurs du ribose.[170] L'efficacité de couplage des monomères les plus encombrés a donc été étudiée. Pour cela, lors de la synthèse de séquences d'homopolymères (octamères), le relargage du DMT a été analysé. A chaque cycle itératif, à la suite de l'ajout d'un nouveau monomère il y a une étape de déprotection du groupement DMT. Le carbocation ainsi formé, a une forte absorption de la lumière à 504 nm. Sa quantité peut être analysée à chaque cycle. L'efficacité de couplage peut donc être analysée. Plus l'absorbance est grande, plus il y a de carbocations DMT+ relâchés, donc plus le nombre de nouveaux monomères ajoutés est important, augmentant ainsi l'efficacité du couplage. Ainsi, durant la synthèse, l'espèce DMT déprotégée est collectée à chaque étape et analysée en spectroscopie UV.

L'efficacité est estimée en comparant la troisième fraction à la fraction antépénultième, comme décrit au préalable.[24] La troisième fraction est analysée car il s'agit de la première fraction pour laquelle un monomère est couplé à un autre monomère. En effet, la première fraction correspond à la déprotection du support et la deuxième correspond au couplage d'un monomère avec le support. Les fractions sont collectées dans des fioles jaugées de 5 mL, complétées à 5 mL avec du dichlorométhane. Un volume de 250 µL est transféré dans une deuxième fiole jaugée de 5 mL, cette dernière étant complétée cette fois avec une solution à 3% d'acide trichloracétique dans le dichlorométhane (v/v). L'absorbance est mesurée à 504 nm avec du dichlorométhane comme solvant de référence. L'absorbance est généralement inférieure à 1 pour des synthèses à une échelle de 1 µmol.

Ces analyses ont été effectuées pour les trois monomères les plus encombrés : M6 avec son groupement pentyle, M7 avec ses chaînes pendantes butyle-éthyle et finalement le plus encombré M8 ayant deux chaînes butyles pendantes. De plus, un suivi dans le temps a également été effectué. Les solutions des monomères ont été réalisées le jour 1. Elles sont restées en place sur l'appareil à température ambiante jusqu'au jour 9. Une analyse à jour 4 ou 5 puis une autre à jour 8 ou 9 ont été réalisées. Les résultats sont rassemblés dans le Tableau II. 9 et montrent les absorbances mesurées, l'efficacité de couplage, le rendement global et la concentration des solutions qu'il est possible de calculer d'après les valeurs obtenues.

L'efficacité de couplage est obtenue avec le calcul suivant : $EC = \left(\frac{A_{n-1}}{A_3}\right)^{\frac{1}{n-3}}$ et le rendement global est obtenu avec le calcul $Rdt = (EC)^n$

Où n est le nombre total d'étapes effectuées pour avoir la séquence désirée. Dans le cas ci-après, des octamères sont synthétisés, n est donc égale à 8. A : absorbance mesurée.

Le calcul de l'efficacité de couplage permet de déterminer si le couplage des monomères sur une même séquence s'effectue toujours de la même manière. Il permet de vérifier si l'antépénultième monomère ajouté, est couplé aussi efficacement que le troisième. Si la quantité de couplages réussis diminue lorsque le nombre d'unité ajouté croît, cela signifie que l'efficacité de couplage n'est pas constante lors de la synthèse d'une séquence.

Lorsque l'efficacité est notée à 1 cela signifie qu'il n'y a pas de différences notables entre les deux valeurs d'absorbances mesurées, et donc que le taux de DMT déprotégé est le même pour toute la séquence. On

peut alors en conclure que tous les couplages ont pu se faire de la même manière. L'efficacité de couplage est donc constante lors de la synthèse de la séquence.

A contrario, plus l'efficacité s'éloigne de 1 plus le nombre de couplages réussis diminue. Le rendement global reflète l'efficacité de tous les couplages de la séquence. Lorsque l'efficacité des couplages n'est pas bonne, le rendement est diminué.

La concentration est obtenue grâce à la loi de Bert-Lambert $A = \varepsilon . l . c$

Avec ε le coefficient d'absorption molaire (76000 L·mol^{-1}·cm^{-1}), l la longueur de la cuve (1 cm) et A l'absorbance mesurée.

Le calcul de la concentration des fractions de DMT collectées permet de déterminer si la quantité de DMT déprotégée est toujours la même au cours du temps. En effet, il se peut que l'efficacité de couplage soit constante au cours du temps, donc que les monomères se couplent toujours bien entre eux, mais que le nombre de couplage (*i.e*. la concentration de couplage) soit diminué dans le temps. Si la concentration diminue au cours du temps, cela voudrait dire que la quantité finale de séquences obtenues est plus faible, qu'il y a donc moins de séquences synthétisées.

Tableau II. 9: Efficacité de couplage et rendement obtenus pour les monomères M6 à M8 pour les séquences d'octamères suivis dans le temps.

Monomère	Jour	A_3	A_{n-1}	Concentration (mol/L)	Efficacité de couplage	Rendement (%)
M6	1	0.846	0.798	1,11^{E}-5	0.988	91.0%
	4	0.787	0.7976	1,04^{E}-5	1	100%
	8	0.667	0.515	8,70^{E}-6	0.917	50,2%
M7	1	0.812	0.837	9,60^{E}-6	1	100%
	5	0.834	0.808	1,10^{E}-5	0.993	95.1%
	9	0.629	0.638	8,3^{E}-6	1	100%
M8	1	0.709	0.701	9,31^{E}-6	0.997	97.9%
	4	0.725	0.677	1,06^{E}-5	0.986	89.6%
	9	0.722	0.683	8,99^{E}-6	0.988	91.8%

On voit que pour le monomère M6 l'efficacité de couplage est relativement bonne jusqu'au 4ème jour, mais au 8ème jour l'efficacité chute et le rendement de réaction diminue drastiquement (il passe de 91% à 50%). Pour le monomère M7, l'efficacité et le rendement sont constants et bons tout au long de l'étude (avec un rendement de réaction de plus de 95% à chaque fois). Concernant le monomère M8, l'efficacité semble moins bonne que pour les deux autres monomères. Une diminution de l'efficacité de couplage se fait voir dès le jour 4 (rendement aux alentours de 90% dès le 4ème jour).

La concentration quant à elle est plutôt constante, aux alentours de de 1^{E}-5 mol/L les deux premiers jours. Mais on remarque qu'elle diminue lors du dernier jour. Il semblerait qu'au cours du temps, la quantité totale de DMT déprotégé diminue. Lors des deux premiers prélèvements, la quantité totale de DMT relarguée (donc le nombre de couplage réussi) est plus grande que celle du dernier jour. Il semblerait donc que le nombre total de couplages réussis soit plus important avec une solution récente. Autrement dit, avec des solutions fraichement préparées, la quantité totale de couplages réussis est plus grande.

D'après ces observations, il peut être conclu que l'efficacité de couplage semble diminuer lorsque les solutions vieillissent (efficacité et rendement diminués pour M6 et M8 au dernier jour de prélèvement). Ceci est corroboré avec les calculs de concentration qui montrent une diminution au dernier jour de collecte. On peut imaginer que l'efficacité de couplage est impactée par la diminution de la concentration. Il y a moins de monomères disponibles pour effectuer les couplages, ceux-ci se font donc moins efficacement.

Cette étude nous montre donc qu'il faut être très vigilent lors de l'utilisation de ces nouveaux monomères. Une fois en solution, leur efficacité semble se dégrader au cours du temps. Il a donc été décidé de toujours utiliser des solutions fraîchement préparées avec les monomères gardés au réfrigérateur au préalable. Pour être sûr d'obtenir les meilleurs résultats possibles, cette règle est appliquée pour les huit monomères utilisés.

De plus pour espérer obtenir une meilleure efficacité dès le premier jour d'utilisation, des améliorations des protocoles de synthèse ont été recherchés. Ces différents protocoles sont présentés dans le paragraphe suivant.

3.2.4.3. Mise de place de différents protocoles pour augmenter l'efficacité de couplage des monomères

Différents protocoles ont été explorés pour essayer d'augmenter l'efficacité de couplage pour le monomère le plus encombré M8. Par exemple, le temps de couplage a été augmenté et passe de 1,36 min à 5 min, les étapes de couplage/oxydation ont été doublées et la concentration initiale des solutions de monomères a été augmentée. Les différents protocoles testés sont expliqués ci-dessous plus en détails.

<u>Protocole standard :</u>

Dans le protocole standard, l'étape de couplage prend 96 secondes et peut être décrite comme la succession de quatre phases :

a. Rincer les lignes avec les solvants (vitesse maximale)
b. Activer le monomère avec une solution de 1*H*-tetrazole 5-thioéthyle (8 s)
c. Faire réagir le monomère activé avec la séquence liée au support (88 s)
d. Rincer les lignes avec les solvants (vitesse maximale)

<u>Protocole modifié **a** :</u>

Dans cette méthode l'étape de couplage est allongée est peut se décrire comme :

a. Rincer les lignes avec les solvants (vitesse maximale)
b. Activer le monomère avec une solution de 1*H*-tetrazole 5-thioéthyle (75 s)
c. Faire réagir le monomère activé avec la séquence liée au support (225 s)
d. Rincer les lignes avec les solvants (vitesse maximale)

Protocole modifié **b** :

Cette méthode modifiée inclut six étapes : (i) déprotection, (ii) couplage, (iii) oxydation, (iv) couplage, (v) oxydation et (vi) capping. Les étapes de couplage et oxydation sont doublées pour obtenir une meilleure efficacité de couplage. La durée de chaque étape de couplage est de 96 secondes comme pour le protocole standard.

Protocole modifié **c** :

Dans cette méthode, la concentration des monomères choisis est augmentée par un facteur de 1,2 (e.g. 120 mM au lieu de 100 mM pour les synthèses sur le système Expédite).

Des séquences tests pour le monomère M8 ont été effectuées pour déterminer si les protocoles **a** et **b** permettent d'augmenter l'efficacité de couplage de ce monomère. Ces séquences sont suivies en spectroscopie UV comme pour les séquences préalablement décrites. Trois séquences d'octo-homopolymères M8 ont été synthétisées à la suite : une en utilisant la méthode standard, la deuxième avec la méthode issue du protocole **a** (temps de couplage allongé) et la dernière avec le protocole **b** (étapes de couplage/oxydation doublées). L'analyse de spectroscopie UV permet de déterminer l'absorbance des solutions collectées et montre que l'efficacité de couplage est meilleure lorsque les protocoles **a** et **b** sont utilisés (voir Tableau II. 10).

Tableau II. 10 : Efficacité de couplage et rendement obtenus pour le monomère M8 avec les différents protocoles.

Méthode	A_3	A_7	Efficacité de couplage	Rendement (%)
Standard	0.725	0.677	0.986	89.6%
Protocole **a**	0.759	0.791	1	100%
Protocole **b**	0.814	0.846	1	100%

Ces tests, effectués sur des séquences modèles d'homopolymère, montrent que la réactivité des monomères les plus encombrés peut être améliorée grâce à des nouveaux protocoles de synthèse. Néanmoins, ces tests n'ont été effectués que sur des homopolymères, et la réactivité des monomères peut être différente lors de couplage avec d'autres monomères.

L'utilisation de l'alphabet augmenté à 8 symboles semble être possible avec la mise en place des protocoles d'amélioration des étapes de couplages. Il semble alors possible de synthétiser des séquences codant de l'information avec cet alphabet. Il a été choisi de coder des images dans ces séquences de poly(phosphodiester)s. Dans un premier temps, la validation du concept a été effectuée en utilisant l'alphabet augmenté à 4 symboles. Cet alphabet est utilisé dans un premier temps, car il ne nécessite pas la mise en place des protocoles d'amélioration. Il est donc plus simple de vérifier avec lui que des images

peuvent bien être encodées sur des chaînes uniques de poly(phosphodiester)s. Les résultats obtenus lors de la synthèse de séquences codant des images sont donnés dans la section suivante.

4. Utilisation des alphabets augmentés pour stocker des images

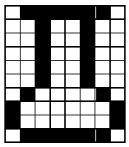
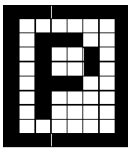
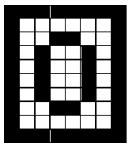
L'information codée sur des séquences de polymères ne doit pas forcément être du texte, des images ou des QR codes peuvent également facilement être codés sur une macromolécule.[121, 171] Ici, il est montré que de petites images pixélisées peuvent être encodées à l'échelle moléculaire. Ainsi, des images en noir et blanc ont été créées dans une grille de 8 colonnes et 10 lignes. Ces images sont converties en séquence binaires en utilisant la convention noir = 1 et blanc = 0. Ce flux de bits peut ensuite être traduit en séquence codée en utilisant les alphabets augmentés montrés dans la Figure II. 16.

4.1. Séquences codant des images avec l'alphabet à 4 symboles

Les premières séquences effectuées avec l'alphabet de troisième génération ont été très prometteuses et ont menées à la synthèse de séquences parfaitement contrôlées.

Le Tableau II. 11 montre les séquences de polymères qui ont été synthétisées pour coder des images avec l'alphabet dyade. Trois images différentes ont été synthétisées avec succès, chacune dans une seule chaîne de polymère. Il s'agit : d'une fiole (polymère **P8**) et des deux symboles atomiques du phosphore (**P9**) et de l'oxygène (**P10**). En tout huit monomères de phosphoramidite ont été nécessaires : les quatre monomères **M1-M4** de l'alphabet à quatre symboles, un espaceur **E** facilitant la lecture, et trois marqueurs de masse **dC**, **dA** et **dG**. Le design précédemment décrit a été suivi.[25] L'espaceur a été inclut dans la chaîne après une série de 8 monomères codants. Chacun des 8 monomères engagés code pour 2 bits, chaque bloc contient donc 16 bits d'information. La synthèse s'effectue dans le sens inverse de lecture de l'information. La synthèse commence par l'extrémité T (élément venant du support solide), alors que la séquence est lue en commençant par l'autre extrémité.[25]

Tableau II. 11 : Description des polymères numériques codant une image synthétisée avec l'alphabet à 4 symboles.

	Image	Séquence binaire [a]	Séquence de monomères [b]	*C* [c]	*DP* [d]	*M* [e] *(Da)*
P8		01111110·00100100· 00100100.00100100· 00100100·00100100. 01000010·10000001· 10000001.01111110	M_2·M_4·M_4·M_3·M_1·M_3·M_2·M_1-*E*- M_1·M_3·M_2 M_1·M_1·M_3·M_2·M_1·G-*E*- M_1·M_3·M_2·M_1·M_1·M_3·M_2·M_1·A-*E*- M_2·M_1·M_1·M_3·M_3·M_1·M_1·M_2·C-*E*- M_3·M_1·M_1·M_2·M_2·M_4·M_4·M_3-T	80	48	8794.9999
P9		11111111·10000001· 10111001.10100101· 10100101·10111001. 10100001·10100001· 10000001.11111111	M_4·M_4·M_4·M_4·M_3·M_1·M_1·M_2-*E*- M_3·M_4·M_3 M_2·M_3·M_3·M_2·M_2·G-*E*- M_3·M_3·M_2·M_2·M_3·M_4·M_3·M_2·A-*E*- M_3·M_3·M_1·M_2·M_3·M_3·M_1·M_2·C-*E*- M_3·M_1·M_1·M_2·M_4·M_4·M_4·M_4-T	80	48	9159.4068
P10		11111111·10000001· 10011001.10100101· 10100101·10100101. 10100101·10011001· 10000001.11111111	M_4·M_4·M_4·M_4·M_3·M_1·M_1·M_2-*E*- M_3·M_2·M_3 M_2·M_3·M_3·M_2·M_2·G-*E*- M_3·M_3·M_2·M_2·M_3·M_3·M_2·M_2·A-*E*- M_3·M_3·M_2·M_2·M_3·M_2·M_3·M_2·C-*E*- M_3·M_1·M_1·M_2·M_4·M_4·M_4·M_4-T	80	48	9131.3755

[a] Les séquences binaires ont été séparées en bloc de huit bits pour plus de clarté. Chaque bloc correspond donc à une ligne de l'image. [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation. [e] Masse monoisotopique.

L'espaceur a été placé tous les 8 monomères. La taille des blocs obtenus après la fragmentation lors de l'analyse MS/MS est alors favorable pour l'analyse par spectrométrie de masse, car les fragments sont chargés de façon homogène.[25, 125] Lors de la synthèse, l'incorporation de l'espaceur **E** est directement suivie de l'addition d'un marqueur de masse en suivant l'ordre dC, dA, dG. L'espaceur du premier bloc n'est suivi d'aucun marqueur, ce qui permet également de le distinguer des autres blocs.

Après la synthèse, les polymères numériques ont été clivés du support et purifiés. Le Tableau II. 12 donne les rendements de synthèse obtenus. Ces rendements sont obtenus en calculant la masse théorique en fonction du loading de support (*i.e*. quantité de chaînes qu'il est possible de synthétiser sur ce support). Ici le loading est à 1 µmol, la masse théorique sera donc égale au loading multiplié par la masse molaire du polymère. Le rendement peut donc être calculé en divisant la masse du polymère obtenue (généralement une poudre) par cette masse théorique.

Tableau II. 12 : Rendement des synthèses de polymères codant des images avec l'alphabet à 4 symboles.

	m (mg)	Rendement (%)
P8	7.1	81
P9	7.1	78
P10	6.8	74

Après la purification, les polymères ont été analysés en spectrométrie de masse (cf. Figure II. 21). Dans tous les cas, la séquence désirée a été identifiée comme l'espèce dominante par spectrométrie de masse ESI en mode négatif.

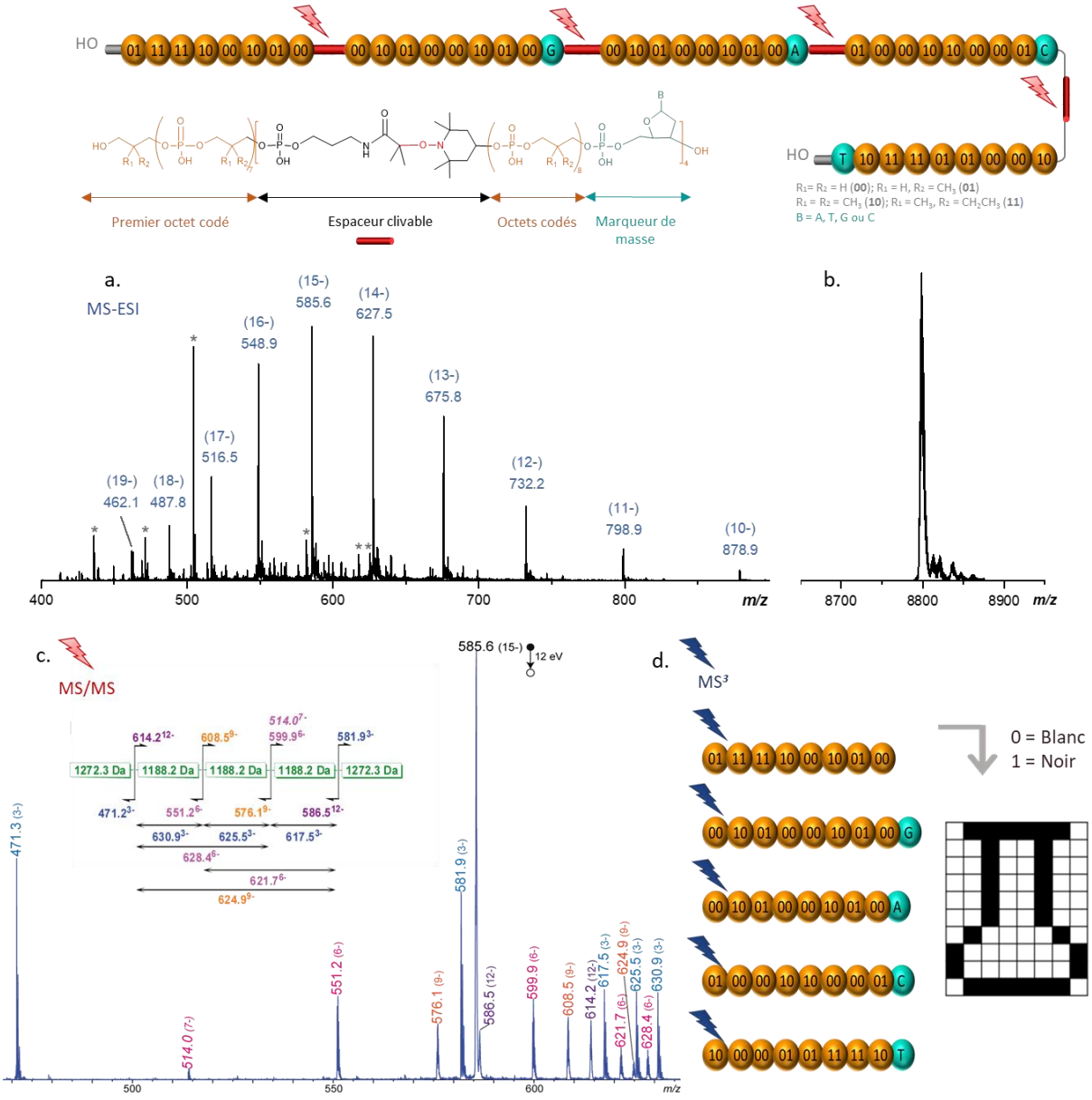


*Figure II. 21 : Analyse de spectrométrie de masse du polymère **P8**. (En haut) Schéma de la structure globale du polymère. (a) Spectre MS montrant les différents états de charges du polymère. Les étoiles grises sont des fragments formés en source. (b) Spectre de masse de déconvolution. (c) Spectre de masse en tandem issu de l'ion précurseur [M-15H]$^{15-}$ à m/z 585,6. (d) Blocs issus du spectre MS/MS, puis analyse MS^3 pour retrouver l'image encodée.*

La Figure II. 21 montre l'analyse en spectrométrie de masse du polymère P8 avec les spectres MS et MS/MS. Le signal obtenu en MS peut facilement être exploité pour séquencer le polymère en utilisant la procédure pseudo-MS^3 préalablement décrite.[25]

Lors de la lecture du polymère, ce dernier est analysé dans un premier temps en ESI-MS en mode négatif. Si une seule espèce est détectée, la synthèse s'est bien déroulée et la séquence obtenue est monodisperse, comme on peut le voir sur le spectre (a) de la Figure II. 21. Une fois cette analyse MS effectuée, il est possible de choisir un ion précurseur (ici l'ion $[M-15H]^{15-}$) qui est transféré et analysé en MS/MS. La dissociation par collision induite (CID) est utilisée pour produire la fragmentation programmée de la liaison faible alcoxyamine, (*i.e*. l'espaceur **E**), schématisée par un éclair rouge sur le schéma. Ces conditions MS/MS mènent à la formation d'une liste des différents blocs libérés, qui sont identifiés grâce aux marqueurs de masse. Dans l'exemple du polymère P8, les fragments contiennent 8 monomères codant donc 16 bits d'information.

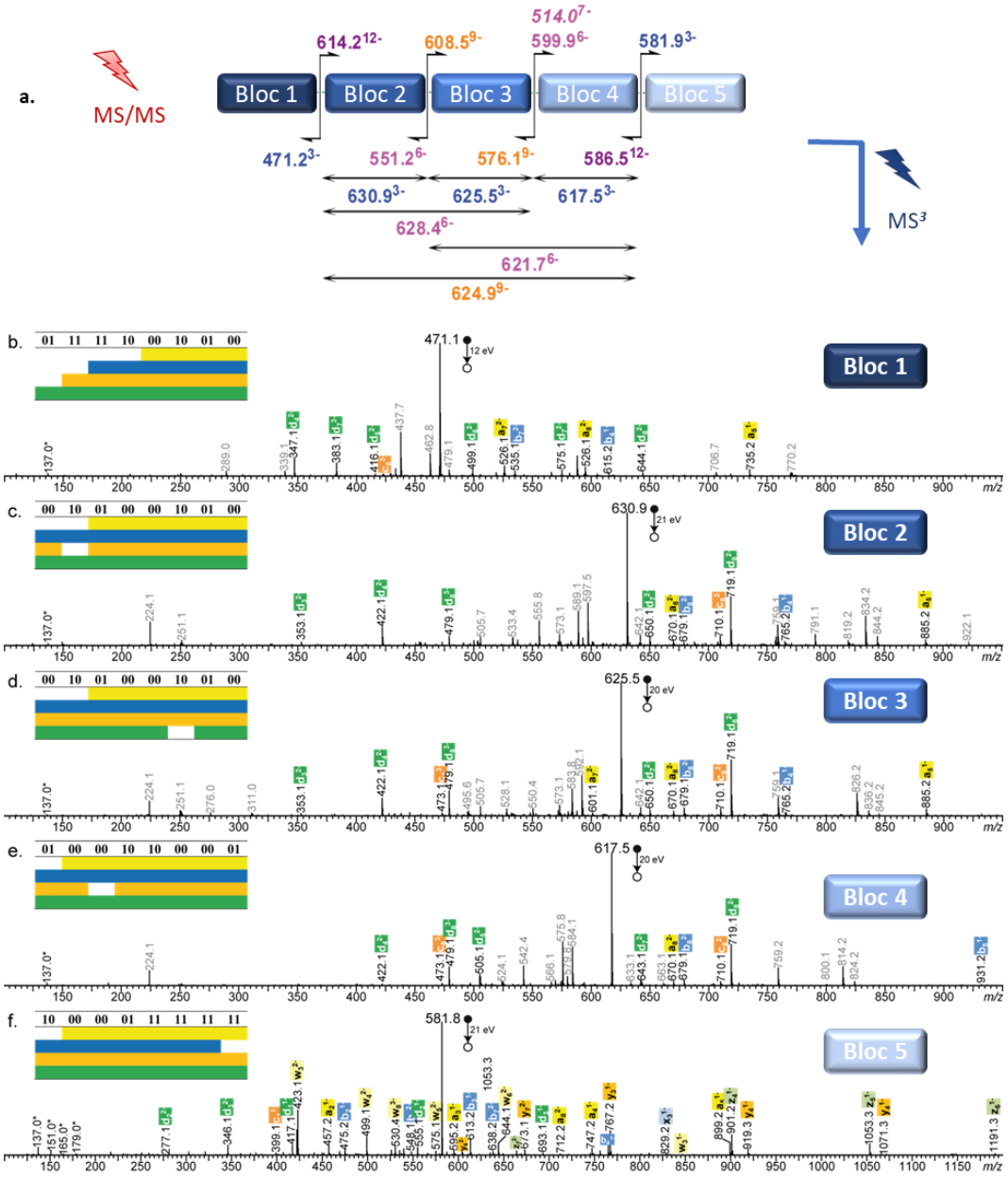


*Figure II. 22 : Séquençage du polymère **P8.** (a) Schéma de la fragmentation obtenue après l'analyse MS/MS de l'ion précurseur $[M-15H]^{15-}$ à m/z 585,6. Spectres Pseudo-MS^3 et la couverture associée des séquences*

pour (b) le premier bloc à m/z 471,1, (c) le second bloc à m/z 630,8, (d) le troisième bloc à m/z 625,5, (e) le quatrième bloc à m/z 617,5, et (f) le cinquième à m/z 581,8. Les monomères déprotonés sont désignés par un astérisque et les pics en gris sont des sous-produits de fragmentation.

Par la suite, les blocs subissent chacun une analyse CID (schématisée par un éclair bleu sur la Figure II. 21 et Figure II. 22) qui décompose chaque bloc et permet de retrouver chaque monomère (conditions pseudo-MS^3, expliquées plus en détails dans la **partie expérimentale**). La Figure II. 22 montre les spectres de pseudo-MS^3 de chaque bloc, ainsi que la couverture associée des séquences. La couverture est représentée par un tableau représentant les monomères dans l'ordre de chaque bloc et la présence ou non des fragments associés à chaque monomère. Les couleurs jaune, bleue, orange et verte correspondent aux différents types de fragments obtenus pour chaque monomère en condition de pseudo-MS^3, montrés sur la Figure II. 23. Les cases du tableau de couverture seront remplies par les couleurs correspondantes au fur et à mesure que les fragments sont retrouvés. Les ions notés a, b, c et d contiennent la terminaison α (terminaison vers la gauche du bloc) et les ions w, x, y et z contiennent la terminaison ω (terminaison vers la droite du bloc).

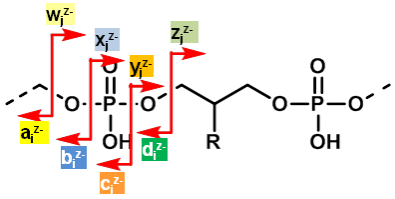


Figure II. 23 : Schéma des fragments obtenus pour chaque monomère lors de l'analyse de pseudo-MS^3.

En combinant toutes les informations trouvées, il est possible de remettre l'information dans le bon ordre en utilisant les marqueurs de masse. Les bonnes images peuvent ainsi être dessinées.

4.2. Séquences codant des images avec l'alphabet à 8 symboles

Des polymères numériques codant des images ont également été codés avec l'alphabet à 8 symboles. D'après les premiers résultats obtenus lors de la synthèse de séquences préliminaires et les résultats des tests d'efficacité, il est montré que les monomères les plus encombrés peuvent avoir des problèmes de réactivité. Ces problèmes peuvent s'accroître lorsque la longueur de la chaîne augmente, il faut donc être de plus en plus prudent pour avoir des efficacités de couplage performantes.

Comme discuté auparavant, les monomères encombrés stériquement peuvent avoir de faibles efficacités de couplage lorsque le protocole standard de la chimie de la phosphoramidite est utilisé. Les protocoles expliqués dans la section 3.2.4.3 ont donc été testés lors des synthèses des premières images avec l'alphabet à 8 symboles. Ceci permet d'avoir un avis sur l'efficacité des différents protocoles lorsqu'ils sont utilisés sur des séquences impliquant différents monomères et non plus que sur des séquences d'homopolymères.

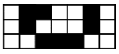
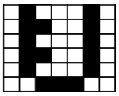
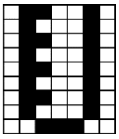
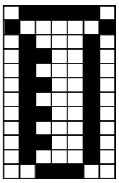
Une image d'un tube à essai d'une taille de 12x7 codant pour 84 bits a été codée en plusieurs étapes. La séquence codant toute l'image contient 4 blocs de 7 monomères. Dans un premier temps, la synthèse de la séquence complète a été interrompue après la formation du premier, puis du deuxième et enfin du troisième bloc. Menant ainsi à trois sous-séquences distinctes codant pour 1 bloc, 2 blocs et 3 blocs. Pour rappel, la synthèse s'effectue dans le sens inverse de la lecture, la séquence comprenant qu'un bloc est

donc le bloc #4, celle comprenant 2 blocs sont les blocs #4 et #3, et finalement celle comprenant 3 blocs sont les blocs #4, #3 et #2. Chaque sous-séquence a été synthétisée avec les trois protocoles :

a) Temps de couplage plus long pour les monomères M5 à M8 (protocole a).
b) Double étape de couplage/oxydation pour les monomères M5 à M8 (protocole b).
c) Augmentation de la concentration des solutions des monomères M4 à M8 (protocole c).

Les neuf sous-séquences ainsi synthétisées ont été analysées en spectrométrie de masse. La description des séquences est donnée dans le Tableau II. 13. Les résultats montrent que le protocole **a** est le moins efficace pour chacune des sous-séquences. En effet, avec ces conditions, des défauts importants sont déjà présents dès la sous-séquence codant pour bloc #4. Bien que le protocole **b** permette d'obtenir des séquences contrôlées pour les deux premières sous-séquences, des défauts sont détectés lorsque la séquence est plus longue d'un bloc et atteint la taille de 3 blocs. Seul le protocole **c** permet la synthèse de chaque sous-séquence de façon monodisperse.

*Tableau II. 13 : Description des sous-séquences synthétisées avec les protocoles a, b et c, menant au polymère **P11**.*

Image	Séquence binaire [a]	Séquence de monomères [b]	*M [c] (Da)*	*Protocole*	*Impuretés*
	0110010· 0100010· 0011100	M_4·M_2·M_2·M_1·M_5·M_4·M_5-T	1432.3981	a	-56 Da
				b	*1 très faible*
				c	*1 très faible*
	0100010·0110010· 0100010·0110010· 0100010·0011100	M_3·M_2·M_2·M_5·M_5·M_5·M_3·C-*E*- M_4·M_2·M_2·M_1·M_5·M_4·M_5-T	3317.9755	a	-154 Da + 148 Da
				b	/
				c	/
	0110010·0100010· 0110010·0100010· 0110010·0100010· 0110010·0100010· 0011100	M_4·M_2·M_2·M_1·M_5·M_7·M_3·A-*E*- M_3·M_2·M_2·M_5·M_5·M_5·M_3·C-*E*- M_4·M_2·M_2·M_1·M_5·M_4·M_5-T	5215.5548	a	- 418 Da -238 Da + 64 Da
				b	-238 Da + 64 Da
				c	/
P11 	0111110·1000001· 0100010·0110010· 0100010·0110010· 0100010·0110010· 0100010·0110010· 0100010·0011100	M_4·M_8·M_3·M_1·M_3·M_5·M_3-*E*- M_4·M_2·M_2·M_1·M_5·M_7·M_3·A-*E*- M_3·M_2·M_2·M_5·M_5·M_5·M_3·C-*E*- M_4·M_2·M_2·M_1·M_5·M_4·M_5-T		c	/

[a] Les séquences binaires ont été séparées en bloc de sept bits pour plus de clarté. Chaque bloc correspond donc à une ligne de l'image. [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Masse monoisotopique.

Il est possible de plus détailler les défauts détectés dans les séquences. Pour la première sous-séquence, une impureté de -56 Da a été détectée dans l'échantillon synthétisé avec le protocole **a** (une trace de la même impureté a été détectée dans l'échantillon synthétisé avec les deux autres protocoles **b** et **c**). Cette différence de masse peut être expliquée par un échange de monomère : les monomères M2 (152 Da) ont été remplacés par des unités de masse 166 Da, ce qui correspond à la masse de M3. Les monomères M5 (194 Da) ont été remplacés par des unités de masse 152 Da, donc M2. En prenant en compte ces changements, il a été possible d'obtenir un spectre de masse en pseudo-MS^3 et une couverture de séquence cohérente.

Pour la deuxième sous-séquence les impuretés trouvées dans l'échantillon synthétisé avec le protocole **a** ont été détectées avec des défauts à -154 Da (-98 Da dans le $2^{ème}$ bloc synthétisé et -56 Da dans le premier) et la deuxième à + 148 Da (+204 Da dans le $2^{ème}$ bloc synthétisé et -56 Da dans le premier). La première impureté est issue des mêmes changements que pour la première sous-séquence. Les monomères M2 ont été remplacés par les monomères M3 et les monomères M5 ont été remplacés par les monomères M2.

Lorsque la séquence atteint la taille de 3 blocs, des impuretés ont été détectées dans les échantillons synthétisés avec les protocoles **a** et **b**. L'impureté la plus abondante de ces deux sous-séquences est la même et est issue des remplacements qui ont pu être mis en évidence via l'analyse MS/MS. Il s'agit des remplacements des monomères M2 par M3, M5 par M2 et M7 par M2.

Ces changements ne semblent pas être une erreur lors de la mise en place des solutions des monomères sur le synthétiseur, car les séquences voulues ont également été détectées dans ces échantillons. Il se peut néanmoins, que la description des protocoles comprenne des erreurs. Le synthétiseur ne prélèverait alors pas uniquement la solution demandée mais également une autre solution. Cela mènerait à ce genre d'impureté où des monomères non voulus sont incorporés à la place des bons monomères. Il se peut également que les lignes ne soient pas bien lavées entre chaque ajout de monomère (fait automatiquement). Des traces des monomères non voulus seraient alors encore présentes dans ces lignes et pourraient également être utilisées lors d'un couplage non désiré. Les étapes de lavages peuvent être allongées pour éviter cette option.

Ces résultats montrent que le protocole **c** mène vers les séquences les mieux définies, ce protocole a donc été choisi pour synthétiser les séquences codant des images. La séquence complète du tube à essai a alors été effectuée (séquence **P11**), les résultats obtenus montrent une séquence monodisperse (cf. Figure II. 24). L'image peut être reconstruite après l'analyse.

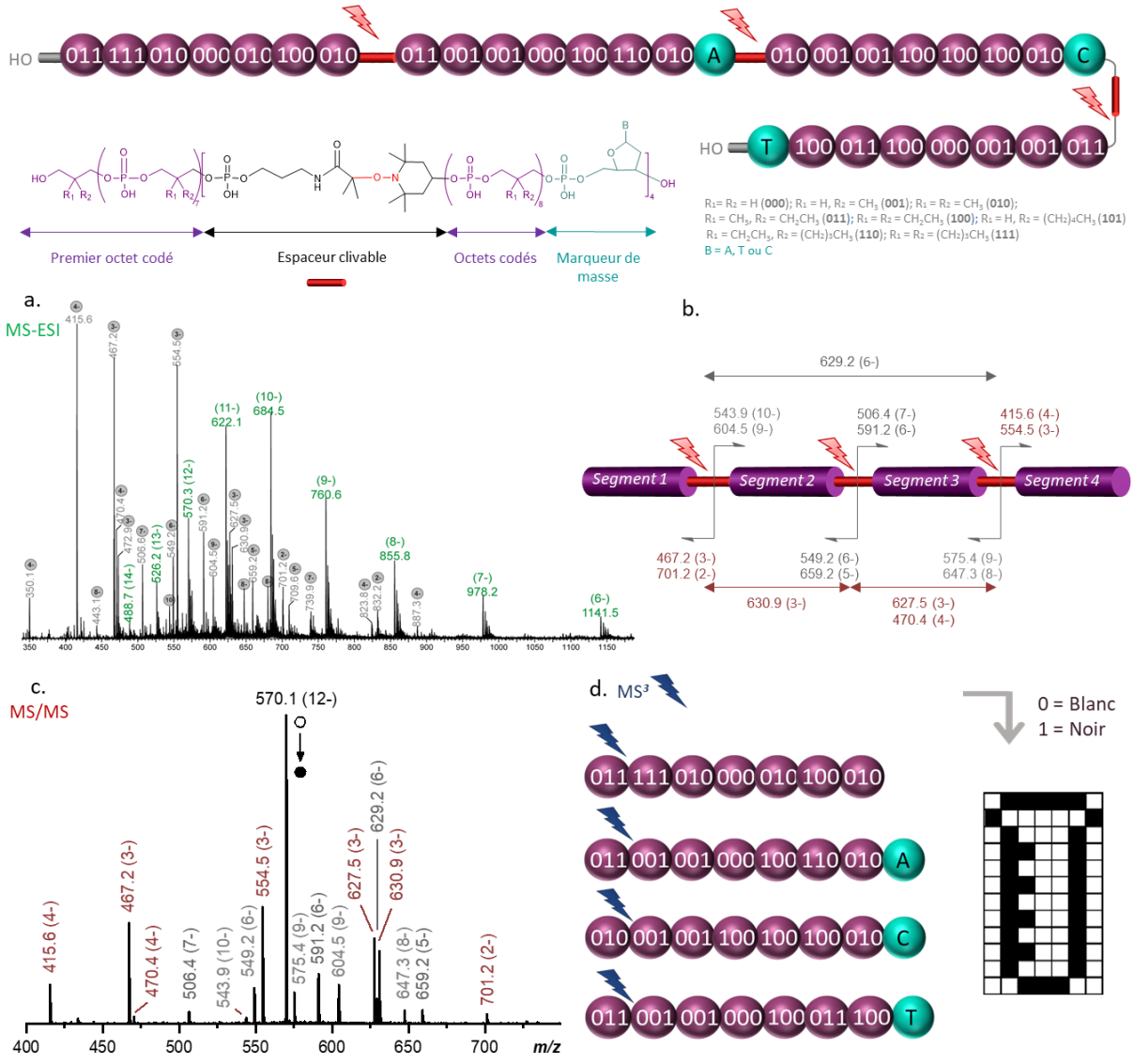
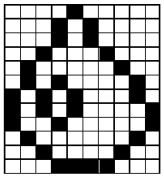
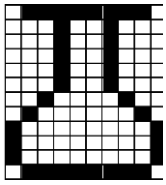
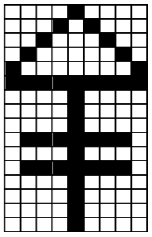


*Figure II. 24 : Analyse de spectrométrie de masse du polymère **P11**. (En haut) Schéma de la structure globale du polymère P11. (a) Spectre MS montrant les différents états de charges du polymère. Les annotations grises sont des fragments formés en source. (b) Schéma de fragmentation en MS/MS. (c) Spectre de masse en tandem issu de l'ion précurseur $[M-12H]^{12-}$ à m/z 570.1. (d) Blocs issus du spectre MS/MS, puis analyse MS^3pour retrouver l'image encodée.*

L'optimisation des paramètres de synthèse mène donc à la synthèse de séquences monodisperses. Il a alors été décidé d'utiliser le second synthétiseur ABI à la place du synthétiseur Expédite. Comme mentionné au début du chapitre (section 2.1.2, Synthèse automatisée), le synthétiseur Expédite est utilisé lorsque les conditions doivent encore être améliorées. Le synthétiseur ABI, peut être utilisé lorsque les conditions sont optimisées. A présent, ce synthétiseur ABI peut alors être utilisé.

Trois images utilisant 13 monomères différents : les huit monomères, un espaceur, et quatre marqueurs de masse, ont été synthétisées. Il s'agit d'une goutte (**P12**), d'une fiole (**P13**) et du symbole alchimique du phosphore (**P14**). Chacune a une taille différente, P12 et P13 : 12x10 = 120 bits et P14 : 16x9 = 144 bits.

Tableau II. 14 : Description des polymères codant des images avec l'alphabet à 8 symboles.

	Image	Séquence binaire [a]	Séquence de monomère[b]	*C* [c]	*DP* [d]	*M* [c] *[Da]*
P12		0000100000·0001010000· 0001010000·0010001000· 0100000100·0101000010· 1010100010·1010100001· 1001000001·0100000010· 0010000100·0001111000	M_1·M_3·M_1·M_1·M_3·M_5·M_1·M_2-*E*- M_3·M_1·M_2·M_1·M_5·M_2·M_1·M_2.G-*E*- M_1·M_6·M_1·M_3·M_6·M_3·M_2·M_3·A-*E*- M_6·M_1·M_4·M_2·M_1·M_2·M_3·M_1·C-*E*- M_2·M_1·M_5·M_2·M_1·M_2·M_8·M_1-T	120	48	9075.3129
P13		0111111110.0001001000. 0001001000.0001001000. 0001001000.0001001000. 0010000100.0100000010. 1000000001.1000000001. 1000000001.0111111110	M_4·M_8·M_8·M_1·M_3·M_3·M_1·M_2-*E*- M_2·M_1·M_1·M_5·M_5·M_1·M_3·M_3.G-*E*- M_1·M_2·M_2·M_1·M_2·M_1·M_3·M_2·A-*E*- M_1·M_1·M_6·M_1·M_1·M_2·M_5·M_1·C-*E*- M_1·M_7·M_1·M_1·M_3·M_8·M_8·M_7-T	120	48	9369.6415
P14		000010000·000101000 001000100·010000010· 100000001·111111111· 000010000·000010000· 000010000·011111110· 000010000·011111110· 000010000·000010000· 000010000·000010000	M_1·M_3·M_1·M_1·M_6·M_1·M_2·M_1-*E*- M_5·M_3·M_1·M_3.M_5·M_1·M_2·M_8·F-*E*- M_8·M_8·M_1·M_3·M_1·M_1·M_3·M_1·G-*E*- M_1·M_3·M_1·M_4·M_8·M_7·M_1·M_3·A-*E*- M_1·M_4·M_8·M_7·M_1·M_3·M_1·M_1·C-*E*- M_3·M_1·M_1·M_3·M_1·M_1·M_3·M_1-T	144	58	11367.1299

[a] Les séquences binaires ont été séparées en bloc correspond à une ligne de l'image. [b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation. [e] Masse monoisotopique.

Les résultats de spectrométrie de masse montrent que les séquences visées ont été obtenues de façon majoritaire dans tous les cas. Ces dernières ont toutes été décryptées via le processus d'analyse en pseudo-MS^3 précédemment décrit.

Les séquences subissent donc dans un premier temps une analyse CID et sont décomposées en segments prévus via le clivage de l'espaceur. Ensuite, tous les segments sont séquencés individuellement et l'information contenue dans les séquences peut être reconstruite en suivant l'ordre dicté par les marqueurs de masse. Ainsi, les séquences contenant 84 (**P11**), 120 (**P12** et **P13**) et 144 (**P14**) bits d'information ont été recouvrées de façon compréhensible via l'analyse de spectrométrie de masse en pseudo-MS^3. Le rendement de chaque synthèse a également été calculé comme expliqué dans le paragraphe 4.1. Ils sont donnés dans le Tableau II. 15.

Tableau II. 15 : Rendement des synthèses de polymères codant des images en triade.

	m (mg)	Rendement (%)
P11	4.9	71
P12	7.5	83
P13	8,2	87,5
P14	9.7	85

A notre connaissance, une chaîne comprenant 144 bits d'information est la capacité de stockage la plus importante décrite à ce jour pour une chaîne de polymère synthétique.

Les séquences P11 et P12 ont été retrouvées de manière totalement monodisperse. La Figure II. 24 montre les résultats d'analyse obtenus pour le polymère P11.

Néanmoins les séquences P13 et P14 montrent également des impuretés. Concernant P13, deux impuretés ont été détectées à +70 Da et à +302 Da. Grâce à l'analyse MS/MS, l'emplacement des impuretés est déterminé.

L'impureté à +70 Da est localisée dans le bloc 2 (le 4ème bloc à être synthétisé) : $M_2{\cdot}M_1{\cdot}M_1{\cdot}M_5{\cdot}M_5{\cdot}M_1{\cdot}M_3{\cdot}M_3$.G-*E*. Dans ce bloc le marqueur de masse dG est impliqué, lui aussi contient un groupement protecteur qu'il faut également déprotéger. Cette déprotection a lieu lors de la purification des séquences. Si elle n'est pas totale, elle mène à la formation de l'impureté détectée. En effet, dans ce cas la séquence non déprotégée est alors aussi détectée. Le groupement protecteur de la base nucléique guanine est un isobutyryle dont la masse est exactement 70 Da. L'ajout de la masse du groupement protecteur à la séquence voulue correspond exactement à l'impureté détectée dans l'échantillon. En combinant cette information avec le fait que l'impureté soit bien localisée dans le bloc comprenant ce groupement protecteur, on peut considérer que le groupement protecteur n'a pas été clivé convenablement lors de la purification.

Concernant l'impureté +302 Da, celle-ci est détectée dans le bloc 1. Ce bloc est le dernier à être synthétisé avant la purification, il contient donc encore un groupement protecteur DMT en bout de chaîne. La masse du groupement est exactement 302 Da et correspond donc à l'impureté détectée.

Concernant l'impureté du polymère P14, elle est détectée à +70 Da et correspond à deux isomères qui portent l'excès de charge sur deux blocs distincts : le bloc 2 ou le bloc 3 : #2 : $M_5{\cdot}M_3{\cdot}M_1{\cdot}M_3.M_5{\cdot}M_1{\cdot}M_2{\cdot}M_8$·F-*E* et #3 : $M_8{\cdot}M_8{\cdot}M_1{\cdot}M_3{\cdot}M_1{\cdot}M_1{\cdot}M_3{\cdot}M_1$·G-*E*. Ces deux blocs contiennent les marqueurs dG et dF, les deux ont la même structure avec la présence d'une molécule de fluor en plus pour dF. Ils ont donc le même groupement protecteur isobutyryle ayant une masse de +70 Da qui correspond à l'impureté détectée.

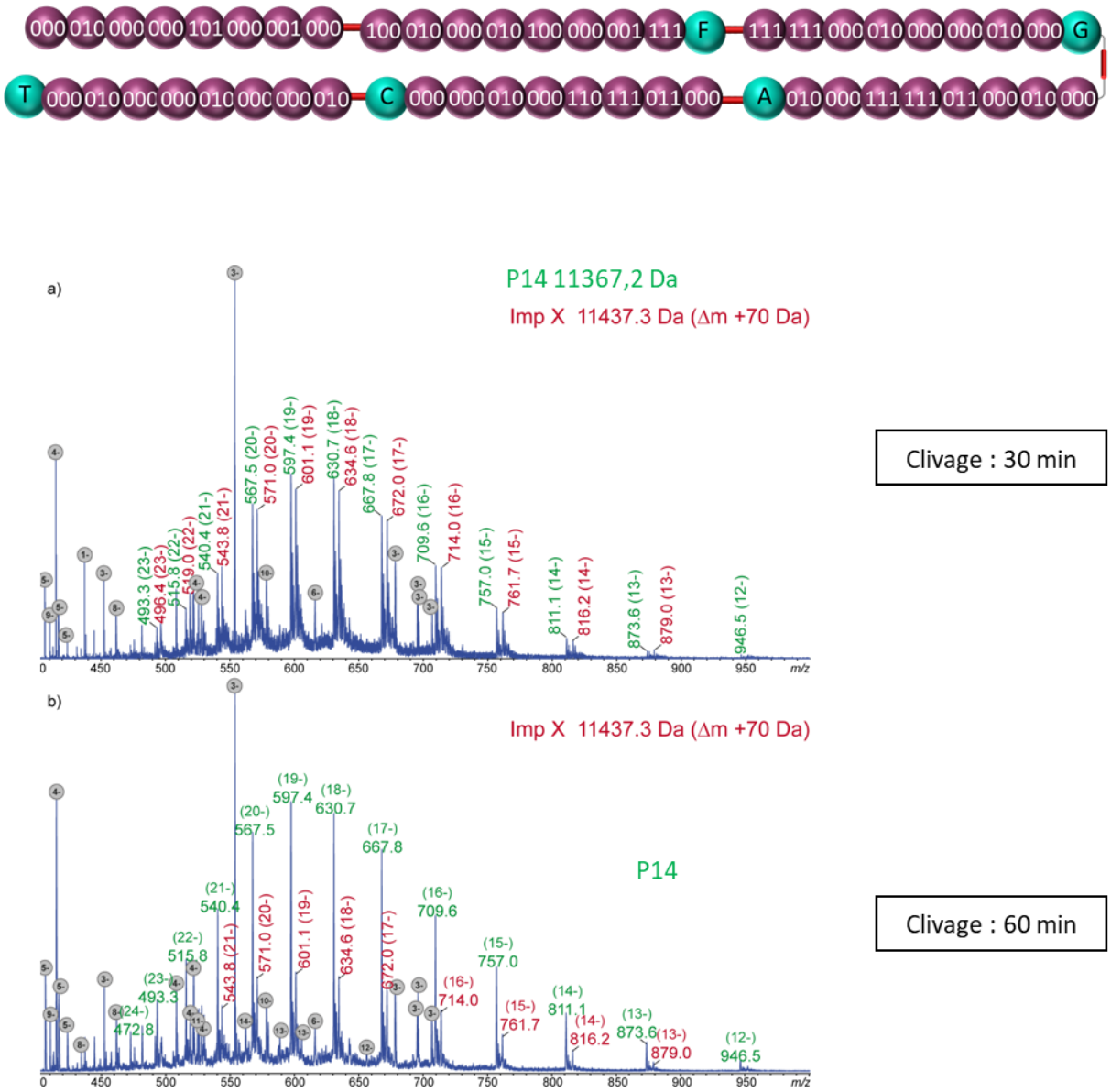


*Figure II. 25 : Spectre de masse du polymère **P14** avec des temps de clivage différents. En haut : schéma de la structure globale du polymère ; (a) clivage en 30min, (b) clivage en 60 min.*

Ce type d'impureté peut être enlevé en effectuant à nouveau la purification des séquences collectées. Elles montrent qu'il faut être particulièrement précautionneux avec les solutions de clivage utilisées. La solution de clivage du DMT est une solution acide de TCA à 3%, celle-ci doit être changée fréquemment pour éviter ce genre de problème. De plus, cela montre qu'il faut faire attention lors de l'utilisation des groupements dG et dF contentant le groupement protecteur isobutyryle. Un temps de purification plus long semble nécessaire lorsqu'ils sont utilisés. Ceci est validé grâce à une étude menée sur le temps de purification : deux séquences de **P14** ont été synthétisées en parallèle dans les mêmes conditions, en utilisant deux positions différentes sur le synthétiseur ABI. Ces deux séquences ont ensuite été purifiées avec des temps de clivage différents : 30 min pour l'une (conditions usuelles) et 60 min pour l'autre, les deux à température ambiante. Les deux séquences purifiées sont ensuite analysées en spectrométrie de

masse. Les résultats présentés en Figure II. 25, montrent que la quantité d'impureté de +70 Da diminue par rapport à la séquence désirée lorsque le temps de clivage est plus important.

Ces analyses de spectrométrie de masse apportent à chaque fois des informations utiles quant à la bonne utilisation des monomères de phosphoramidite. Premièrement, elles permettent de prendre garde à la longueur de vie des solutions de marqueurs de masse et des monomères utilisés. Ces solutions doivent toujours être préparées le plus récemment possible pour éviter d'avoir des séquences avec un monomère manquant. La plupart des impuretés simples des séquences viennent de couplages ratés lorsque la solution utilisée est trop ancienne. Par exemple, pour la séquence P4 la seule impureté détectée est une séquence où il manque le marqueur dF. Il est facile d'obtenir une séquence monodisperse en réitérant la synthèse de la séquence avec de nouvelles solutions fraichement préparées. Néanmoins, cette manière de procéder peut être très coûteuse en monomère car il faut préparer à chaque fois suffisamment de solution pour ne pas risquer d'avoir une solution vide au cours de la synthèse d'une séquence, mais ne pas faire un trop grand excès pour ne pas perdre trop de solution. Bien entendu, il est possible de récupérer les solutions trop vieilles pour évaporer le solvant et récupérer le reste de monomère. Cette procédure est très chronophage pour finalement récupérer une quantité assez faible de monomère. Il faut rassembler plusieurs anciennes solutions pour espérer récupérer une quantité suffisante de matière réutilisable.

Deuxièmement, les analyses de spectrométrie de masse permettent aussi de constater qu'un temps de déprotection plus long est nécessaire pour déprotéger complètement tous les groupements protecteurs des marqueurs de masse. Elle nous met également en garde sur l'étape de déprotection du groupement DMT. Ce clivage n'est parfois pas effectué sur toutes les séquences. Certaines d'entre elles gardent le groupement DMT et sont alors détectées comme des impuretés lors de l'analyse de spectrométrie de masse. En utilisant une solution de clivage fraichement préparée, celui-ci est toujours efficace et aucune séquence non déprotégée n'est recouvrée.

5. Conclusion et perspectives

Des séquences de poly(phosphodiester)s à haute capacité de stockage sont décrites dans ce **chapitre II**. Ces macromolécules contenant de l'information sont préparées en utilisant des alphabets augmentés contenant 4 ou 8 monomères phosphoramidites. L'avantage majeur de ce design est de permettre de combiner deux atouts : une longueur de chaîne importante et une haute densité de stockage. En effet, l'utilisation de la chimie de la phosphoramidite automatisée permet la synthèse de polymères relativement longs (jusqu'à 58 monomères) et l'utilisation d'alphabets augmentés permet d'insérer plus d'information dans un polymère donné (2 bits ou 3 bits par monomère). Les deux alphabets permettent la synthèse de macromolécules numériques uniformes, même si les monomères les plus encombrés de l'alphabet codant sur des triades requièrent une optimisation du protocole de couplage pour être incorporés avec de hauts rendements dans les chaînes encodées. Un espaceur clivable inter-segments et des marqueurs de masse ont été inclus dans les polymères pour permettre d'induire une fragmentation programmée de la chaîne lors du séquençage en ESI pseudo-MS^3. En conséquence, tous les échantillons synthétisés ont été décodés facilement en spectrométrie de masse.

En prenant en compte toutes les informations récoltées au fur et à mesure des synthèses, il a été possible de synthétiser des polymères numériques à haute capacité de stockage. Non seulement du texte, mais également de petites images pixellisées en noir et blanc ont été stockées dans des chaînes uniques de polymère, et ont été décryptées. Par exemple, une image contenant 144 pixels a été inclue dans une séquence de 58 monomères de poly(phosphodiester)s (séquence **P14**). Une telle capacité de stockage est à notre connaissance la plus haute jamais décrite pour un copolymère synthétique.

Le design présenté ici n'est, en outre pas une limite. Il est en effet possible d'augmenter d'avantage la capacité de stockage d'une simple chaîne en augmentant la taille de la séquence synthétisée jusqu'à au moins 100 monomères. Il est par ailleurs possible d'étendre encore l'alphabet utilisé, en passant à un alphabet codant sur 16 symboles ou encore en créant un alphabet comprenant 27 symboles, facilitant l'encodage de texte en français car basé sur les lettres de notre alphabet (26 symboles pour les lettres de l'alphabet et 1 symbole pour les espaces entres les mots). Bien sûr, il est aussi possible d'utiliser des algorithmes de compression sans perte comme l'encodage « arithmetic » ou le codage de Huffman, cette stratégie est explorée dans le **chapitre IV**.

Il est également possible d'améliorer encore l'analyse de spectrométrie de masse. La facilité de l'analyse repose essentiellement sur l'efficacité de l'espaceur placé dans la séquence. Cet espaceur peut être amélioré et peut rendre encore plus facile la lecture par spectrométrie de masse. L'amélioration de l'espaceur permettrait en outre de rendre automatique l'analyse de spectres des masse obtenus après la fragmentation en pseudo-MS^3. Le logiciel MS-DECODER, récemment développer par une équipe de l'université de Strasbourg,[128] pourrait être utilisé pour effectuer le décodage automatique de séquences numériques de poly(phosphodiester)s. Cet aspect est développé dans le **chapitre III**.

Somme toute, ce chapitre démontre que des fichiers de données complets (textes ou images) de tailles significatives peuvent être stockés dans une macromolécule unique.

Chapitre III

Optimisation de l'espaceur contenant une liaison alcoxyamine

1. Introduction

Les **chapitres I** et **II** ont montrés que les séquences de poly(phosphodiester)s contenant un motif espaceur alcoxyamine peuvent être décodées facilement. L'espaceur incorporé dans la macromolécule est l'élément clé pour faciliter la lecture. Sans cet élément il est possible de synthétiser de longues chaînes de poly(phosphodiester)s, mais leur lecture est difficile voire impossible en spectrométrie de masse. L'incorporation de l'espaceur est donc nécessaire pour recouvrir de façon efficace l'information codée sur ces macromolécules.

Le choix de cette molécule repose sur une règle générale : elle doit comporter une liaison faible nécessitant moins d'énergie pour être cassée dans les conditions de dissociation induite par collision (CID) que les liaisons connectant les monomères codant de l'information. Dans le cas des chaînes de poly(phosphodiester)s, l'espaceur choisi comprend une liaison alcoxyamine qui est plus facile à cliver que les liaisons phosphates qui relient les monomères codant l'information. Cependant, ce concept n'est probablement pas limité au poly(phosphodiester)s et pourrait s'appliquer à d'autres types de polymères numériques.

Lorsque ce type de motif espaceur est utilisé, la lecture effectuée par l'analyse de spectrométrie de masse se déroule en trois étapes (schéma explicatif disponible dans le chapitre I et des détails sont donnés en **partie expérimentale**). Dans un premier temps, la séquence est analysée en spectrométrie de masse en ionisation par électronébuliseur ESI-MS en mode négatif. Le spectre obtenu montre les différents états de charges de la macromolécule recherchée. Dans la deuxième étape, un ion précurseur choisi est transféré et analysé en MS/MS. Lors de cette étape, la liaison faible de l'espaceur est clivée, libérant ainsi différents blocs intacts qui sont identifiés grâce aux marqueurs de masse incorporés dans la séquence à la fin de chaque bloc (*i.e.* juste avant l'espaceur). Dans la troisième étape, un ion précurseur de chaque bloc est transféré et analysé en pseudo-MS^3. Ils sont alors soumis un à un à une seconde analyse CID qui fragmente les liaisons phosphates.

Lors des premières réactions de fragmentation en MS/MS, il n'y a que la liaison faible de l'espaceur qui est clivée. Ici, il s'agit de la liaison NO-C de l'alcoxyamine. Lors de la troisième étape où la seconde activation par CID a lieu (pseudo-MS^3), il y a deux types de dissociations. La première comprend tous les clivages des liaisons du squelette de la chaîne de poly(phosphodiester)s. Les fragments détectés sont ceux contentant les terminaisons α (*i.e.* a_i^{z-}, b_i^{z-}, c_i^{z-}, d_i^{z-}) et/ou ω (*i.e.* w_i^{z-}, x_i^{z-}, y_i^{z-}, z_i^{z-}) qui permettent de reconstruire la séquence à partir de la gauche (α) ou de la droite (ω) du bloc. Le deuxième type de dissociation est dû à la réactivité du radical du carbone central sur la terminaison ω de tous les blocs excepté le dernier. Ceci va générer de nombreux ions parasites qui ne sont pas utiles pour reconstruire la séquence. Ces nouveaux ions générés en grande abondance rendent plus difficile la lecture du spectre de pseudo-MS^3.

Des pistes de recherche ont été étudiées afin d'éviter la formation d'un radical trop réactif, permettant ainsi de réduire la formation de ces ions parasites. Ces dernières seront exposées dans ce **chapitre III**, avec dans une première partie une explication détaillée des réactions secondaires. Par la suite, on s'intéressera à la stratégie suivie pour essayer de les éliminer.

L'élimination de ces réactions parasites permettrait d'utiliser le logiciel MS-DECODER qui décrypte automatiquement les messages codés dans les polymères numériques.[128] Cet algorithme a été développé par Christine Carapito et Alexandre Burel de l'équipe de Spectrométrie de Masse BioOrganique (LSMBO) de Strasbourg. Son objectif est de décoder automatiquement les spectres obtenus lors des analyses en pseudo-MS^3, permettant ainsi de recouvrer le message codé de façon automatique.

Le design de la séquence peut être amélioré grâce à l'espaceur. Ceci peut passer par l'ajout sur l'espaceur de nouvelles fonctions apportant de nouvelles propriétés à la séquence entière. Ainsi, il sera montré, dans une dernière partie, qu'il est possible de synthétiser des séquences contenant des fragments photosensibles ou des éléments pouvant être utilisés comme marqueur de masse.

Ces recherches ont été effectuées en collaboration avec Kévin Launay, doctorant effectuant sa thèse au sein de l'Institut de Chimie Radicalaire de l'Université d'Aix-Marseille sous la direction du Dr. Didier Gigmes. Les molécules des différents espaceurs testés ont été synthétisées par ses soins. J'ai effectué leur validation pour une incorporation dans des séquences de poly(phosphodiester)s, dans un premier temps avec de courtes séquences tests, puis dans des séquences plus conséquentes contenant une plus grande quantité d'information.

2. Réactions parasites observées avec l'utilisation de l'espaceur classique

L'espaceur dit classique est une molécule contenant une liaison alcoxyamine NO-C basée sur le 2,2,6,6-tétraméthyl-1-pipéridinyl)oxydanyl (TEMPO). Cette molécule a été utilisée dans plusieurs travaux de l'équipe, notamment dans les travaux présentés dans le **chapitre II**.[25, 172] Cet espaceur est nommé **E1** dans la suite du chapitre.

Lorsque l'espaceur classique E1 est utilisé, la liaison alcoxyamine NO-C (représentée en rouge sur la structure donnée en Figure III. 1) est la liaison la plus fragile de la séquence. C'est donc celle-ci qui va se cliver en premier lors de l'activation par l'analyse CID. A la suite de la fragmentation, (vaguelette rouge sur le schéma (b) de la Figure III. 1) les segments correspondant aux différents blocs ont à leurs extrémités les éléments correspondant à la partie gauche ou droite de l'espaceur comme montré sur la Figure III. 1.c. La partie droite de l'espaceur (comprenant le motif TEMPO) est localisé en α des segments correspondant aux différents blocs et la partie gauche est la terminaison ω des segments.

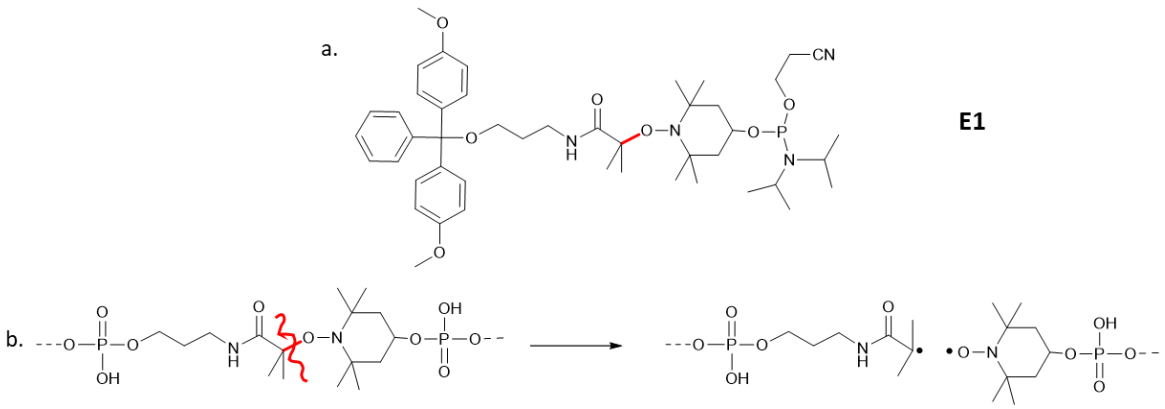


*Figure III. 1: (a) Structure de l'espaceur classique **E1**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur classique incorporé dans une séquence suivie de la fragmentation de la liaison alcoxyamine.*

Tous les blocs excepté le dernier ont donc une extrémité ω se terminant par un radical sur le carbone central. Le dernier bloc se termine par le marqueur dT issus du support. L'extrémité α est composée soit du groupement hydroxyle pour le premier bloc ou alors de l'autre partie de l'espaceur. Dans les séquences composées de blocs de huit monomères, les blocs sont généralement triplement chargés, un schéma des segments obtenus après la fragmentation est donné en Figure III. 2.

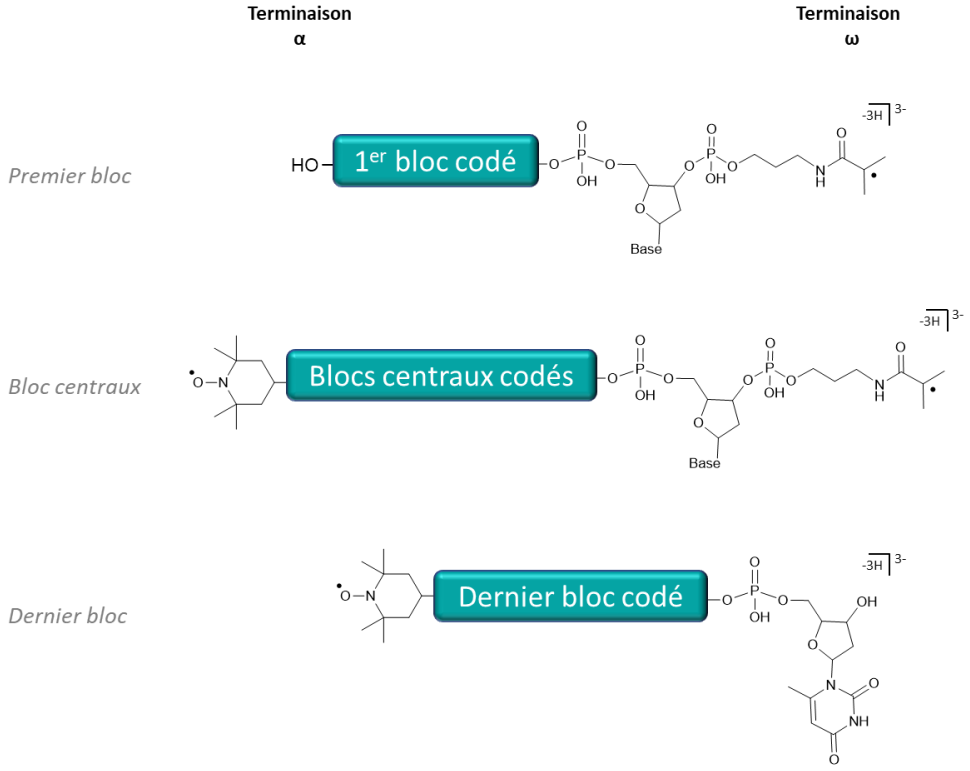


Figure III. 2: Structure des fragments primaires pour n'importe quel bloc retrouvé en analyse MS/MS après l'homolyse de la liaison alcoxyamine.

A cause de la présence du radical carbone en ω, de nombreuses réactions parasites ont lieu. Les nouveaux ions ainsi formés encombrent l'analyse de pseudo-MS^3 et compromettent la couverture totale de la séquence.

Pour éviter la synthèse des multiples impuretés détectées lors de l'analyse de spectrométrie de masse en pseudo-MS^3, leur formation a dans un premier temps été analysée. L'équipe de Laurence Charles de l'université d'Aix-Marseille a ensuite proposé un mécanisme de formation des ions.

Seules les réactions intramoléculaires en phase gazeuse ont été considérées, car la formation de ces impuretés se produit uniquement lors de l'analyse de spectrométrie de masse en tandem. Ces impuretés sont généralement issues de faibles pertes de masses. Les plus courantes sont une perte de 100,1 Da et 225,1 Da. La perte de masse correspondant aux différentes bases de l'ADN est également détectée. Ces bases sont un des éléments constituant les nucléosides utilisés comme marqueurs de masse (sucre + base = nucléoside).

Parmi ces réactions, certaines sont caractérisées par l'élimination d'espèces radicalaires issues du précurseur activé. Le mécanisme proposé en Figure III. 3 montre comment il est possible d'expliquer la perte de masse de 100,1 Da observée lors des analyses.

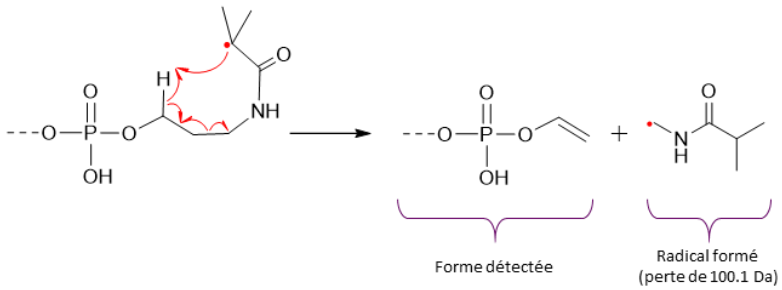


Figure III. 3: Mécanisme proposé pour la perte d'un radical de 100,1 Da pour tous les segments comportant un radical sur le carbone central de la terminaison ω.

D'autres réactions sont observées avec, cette fois, une perte neutre de 225,1 Da sur le segment normalement observé, avec en parallèle l'apparition d'un pic à m/z 224,1. Ce pic correspond à l'ion formé à partir d'un transfert de H·, menant à la perte du fragment indiqué en Figure III. 4.

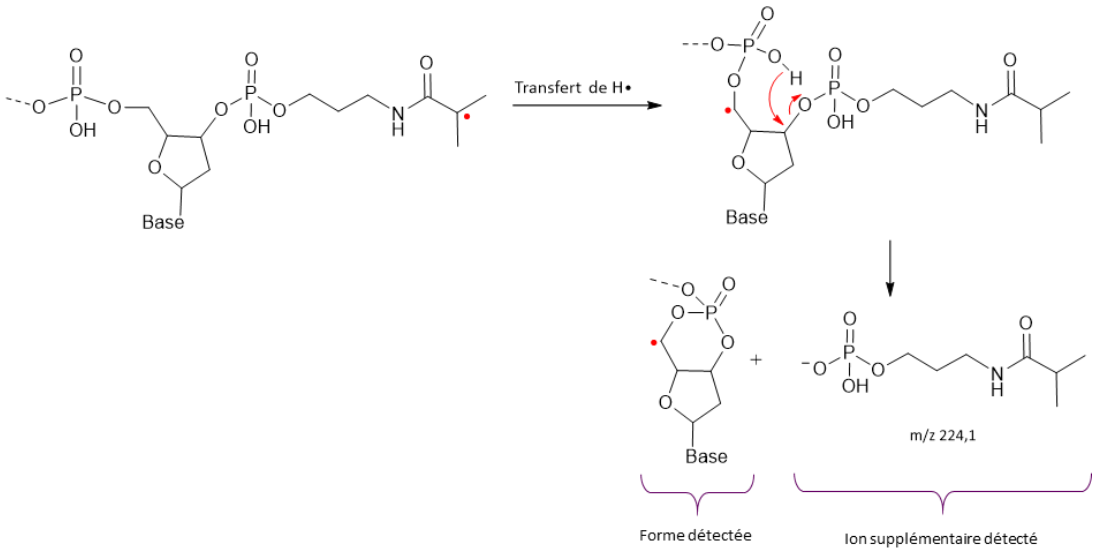


Figure III. 4: Mécanisme proposé pour la perte du fragment à 225,1 Da avec la production en parallèle d'un ion à m/z 224,1.

Il se peut que ce type de réaction secondaire soit accompagné d'une perte de la base. Lorsque cela se produit, il n'y a pas de transfert de H• et donc pas de production d'un ion supplémentaire à m/z 224,1. Il y aura la perte d'un radical de 224,1 Da et la perte de la base comme montré sur la Figure III. 5.

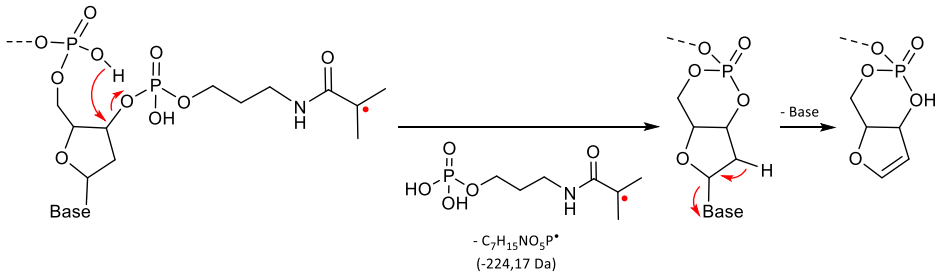


Figure III. 5: Mécanisme proposé pour la perte combinée de 224,1 Da et la perte d'une base.

Toutes les réactions secondaires analysées demandent une bonne accessibilité du radical pour pouvoir se produire. Deux stratégies sont donc testées pour remédier à ce problème. Dans un premier temps, la délocalisation du radical alkyle est testée afin de diminuer l'apparition des réactions secondaires. La seconde stratégie mise en place, vise à rendre la chaîne moins flexible afin de diminuer l'accessibilité du radical. Ces deux voies d'optimisation seront discutées dans le paragraphe suivant.

3. Développement d'espaceurs optimisés

3.1. Première stratégie : délocalisation du radical

La première stratégie testée est la délocalisation du radical alkyl. Pour se faire un nouveau design est élaboré. Il est décidé d'ajouter un groupement phényle permettant la délocalisation du radical par effet mésomère.

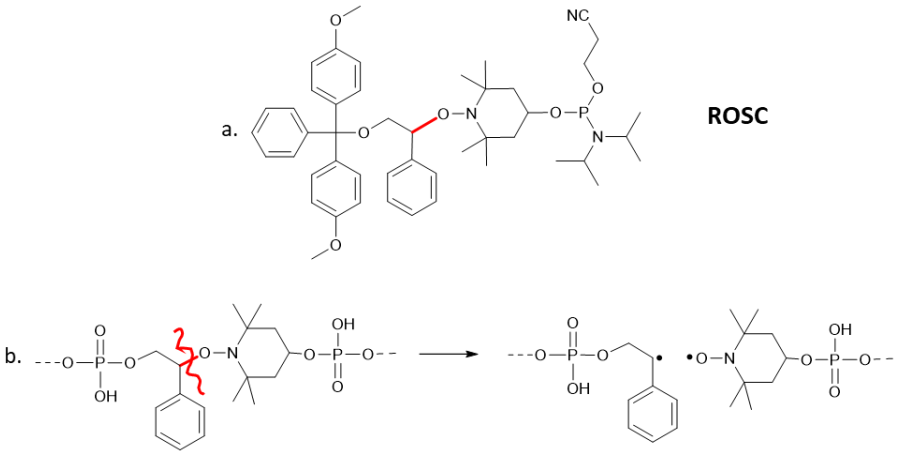


*Figure III. 6: (a) Structure de l'espaceur **ROSC**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur ROSC incorporé dans une séquence suivie de la fragmentation de la liaison alcoxyamine.*

La première molécule testée comprend donc une liaison alcoxyamine avec un groupement phényle latéral par rapport au squelette de la molécule. La Figure III. 6 présente la structure de cet espaceur **ROSC** seule (a), et sa structure lorsqu'il est incorporé dans une chaîne de poly(phosphodiester)s suivie par le schéma de la fragmentation (b). Il a été nommé ROSC pour « Ring OutSide the Chain ».

La synthèse de poly(phosphodiester)s incorporant l'espaceur ROSC a été effectuée. Deux séquences tests (ne codant pas de l'information) ont été synthétisées : la première n'implique que les monomères M1 et M3, et la deuxième met en jeu également un marqueur de masse dA. Ces molécules sont présentées dans le chapitre II. Aucun problème n'a été détecté lors de la synthèse de ces deux séquences. Elles ont été retrouvées de façon quasi monodisperse ; elles sont décrites dans le Tableau III. 1. La première séquence ne contient aucune impureté et l'analyse de la séquence $P_{ROSC}2$ montre que le polymère a été retrouvé de façon quasi monodisperse sous les états de charge -8 et -9, avec néanmoins une faible impureté à +217,6 Da (voir **partie expérimentale**).

L'analyse en spectrométrie de masse en tandem puis en pseudo-MS^3 de ces deux séquences permet de juger si ce nouvel espaceur ROSC est plus efficace que l'espaceur E1. Elles permettent donc de vérifier si les réactions secondaires, induites par le radical formé après l'homolyse de la liaison alcoxyamine, sont évitées.

Tableau III. 1: Description des polymères synthétisés avec l'espaceur ROSC.

Séquence	Séquence de monomères	*M [a] (Da)*	*Impuretés*
$P_{ROSC}1$	$M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1$-ROSC-$M_2{\cdot}M_2{\cdot}M_2{\cdot}M_2{\cdot}M_2{\cdot}M_2{\cdot}M_2{\cdot}M_2$-T	3029,6267	/
$P_{ROSC}2$	$M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1$-ROSC-$M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1$A-ROSC-$M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1{\cdot}M_1$-T	4605,6856	+217,6 Da

[a] *Masse monoisotopique.*

Les résultats obtenus lors de l'analyse complète de la séquence **P_{ROSC}1** montrent que l'homolyse de l'espaceur ROSC lors de l'analyse MS/MS mène à deux sous-segments complémentaires triplement chargés, comme montré en Figure III. 7.a. L'activation de l'ion à m/z 407,4 contenant le premier octet permet d'obtenir le spectre de pseudo-MS^3 montré en Figure III. 7.b. La séquence peut être retrouvée grâce aux fragments contentant les terminaisons α. Néanmoins, le produit majoritaire obtenu est un ion doublement chargé à m/z 504,0 (annoté en rouge), sa formation résulte d'une réaction secondaire qui n'a pas été évitée par à la présence du groupement phényle. De plus, deux autres réactions secondaires sont également détectées : (i) l'élimination du styrène à partir du radical terminal mène à l'espèce triplement chargée à m/z 372,7 (en rose) et (ii) l'oxydation du précurseur est également observée via la détection du fragment doublement chargé à m/z 611,1 (en violet).

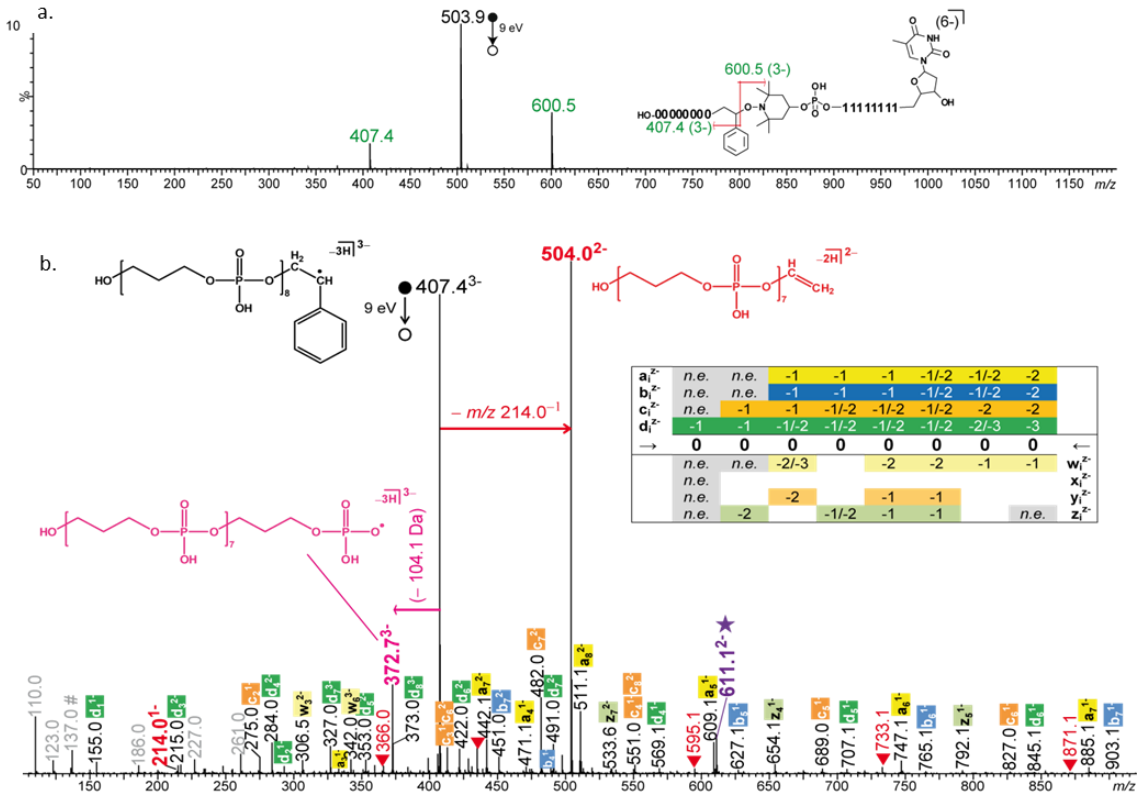


*Figure III. 7: Séquençage du polymère **P_{ROSC}1**. (a) Spectre MS/MS de l'ion $[M - 6H]^{6-}$ à* m/z *503,9 (b) Spectre pseudo-MS^3 du bloc 1 à* m/z *407,4. Les pics annotés en gris correspondent à des fragments internes, parmi lesquels les monomères sont désignés par #. Les ions annotés en rose, rouge et violet sont issus de réactions induites par le radical de l'espaceur ROSC.*

A cause de ces différentes réactions secondaires (décrites en Figure III. 8) de nombreux fragments ω sont absents comme on peut le voir sur le tableau de couverture de séquence montré en Figure III. 7.b.

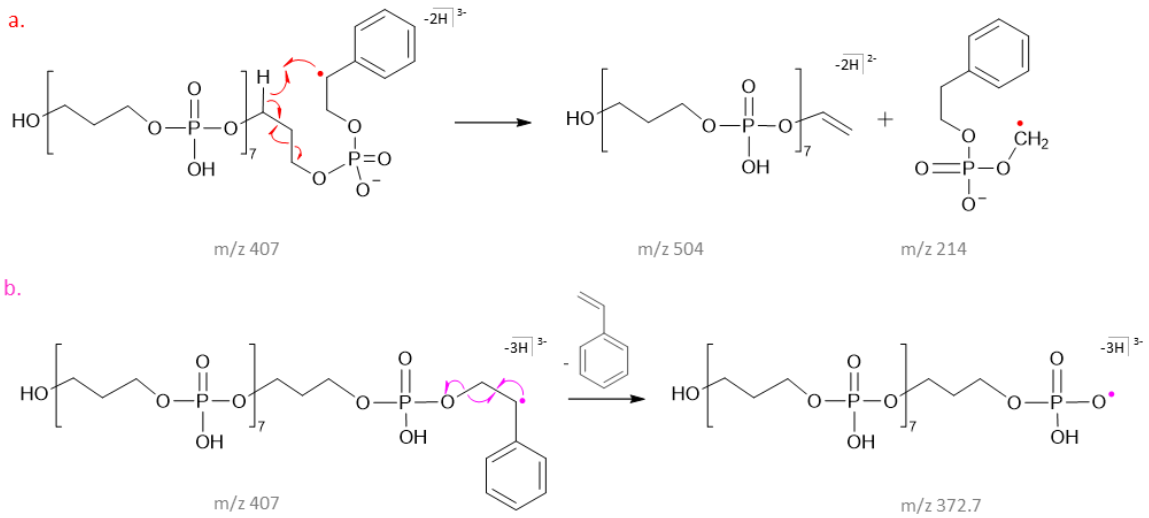


Figure III. 8: Description des réactions secondaires observées pour la séquence $P_{ROSC}1$.

La réactivité des radicaux alkyles générés après le clivage de l'espaceur ROSC ne diminue donc pas par rapport à celle observée avec l'espaceur E1. La couverture des séquences des différents blocs est affectée, empêchant notamment la formation de plusieurs fragments w, x, y et z. L'analyse du séquençage du polymère **$P_{ROSC}2$** montre les mêmes types de réactions secondaires induites par le radical. L'emploi du nouvel espaceur ROSC n'implique aucune amélioration notable vis-à-vis de l'espaceur E1, il a donc été écarté des recherches et la seconde stratégie est employée.

3.2. Seconde stratégie : rigidifier le squelette de l'espaceur

3.2.1. Espaceur RISC

Dans cette seconde stratégie l'objectif est de rendre plus rigide la molécule. Ceci permettrait d'éviter les réactions parasites qui se produisent grâce à la flexibilité de la molécule. Dans cet aspect, un nouvel espaceur nommé **RISC** pour « Ring InSide the Chain » a été synthétisé. Cette nouvelle molécule implique également un groupement phényl, mais directement compris dans le squelette de la molécule, sa structure est donnée en Figure III. 9.

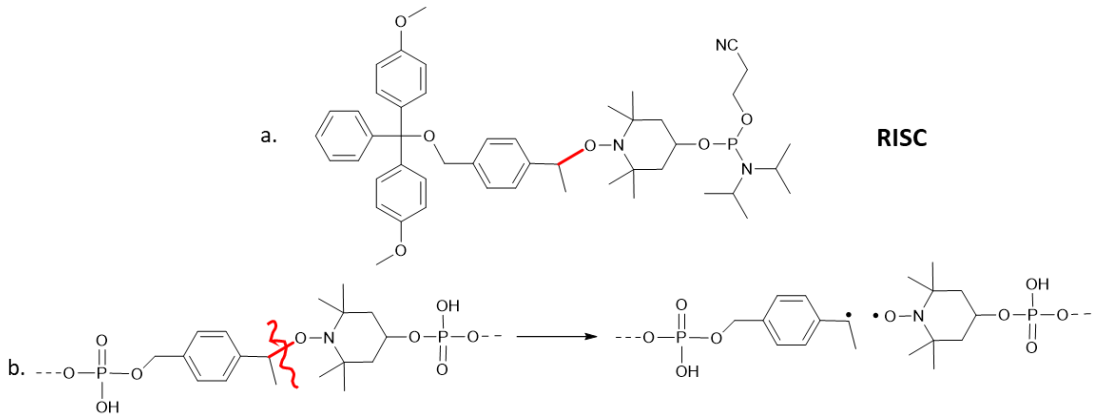


*Figure III. 9: (a) Structure de l'espaceur **RISC**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur RISC incorporé dans une séquence suivie du schéma de la fragmentation de la liaison alcoxyamine.*

Ce nouveau design permet de garder l'effet de délocalisation du radical grâce au groupement phényle mais augmente aussi la rigidité de la chaîne principale. Des séquences tests ont à nouveau été effectuées : une comprenant uniquement les monomère M1 et M3 et une autre impliquant aussi le marqueur de masse dA, comme indiqué dans le Tableau III. 2.

Tableau III. 2: Description des séquences tests synthétisées avec l'espaceur RISC.

Séquence	Séquence de monomères	*M [a] (Da)*	*Impuretés*
$P_{RISC}1$	M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-RISC-M_3·M_3·M_3·M_3·M_3·M_3·M_3·M_3-T	3034.6423	• Manque de M1 • 4 impuretés issus de réactions parasites lors du clivage.
$P_{RISC}2$	M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-RISC-M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1·A-RISC-M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-T	4605.685	3 impuretés issus de réactions parasites lors du clivage.

[a] *Masse monoisotopique.*

Ces séquences ont été effectuées sur le synthétiseur Expedite qui permet de suivre les rendements de couplage en direct via une analyse UV des solutions de DMT déprotégées. Les rendements sont donnés sous forme d'histogramme récapitulant l'efficacité de chaque déprotection, ces derniers sont montrés en Figure III. 10. Ces graphiques montrent l'efficacité de chaque couplage en la représentant par une barre sur l'histogramme. La première n'est pas à prendre en compte car elle correspondre à la déprotection du support. Ici, les barres numérotés 7 correspondent au monomère M3, les barres 5 au monomère M1, les barres T (excepté la première) à l'espaceur RISC et la barre A au marqueur dA.

Ces histogrammes ne sont pas assez précis pour donner les vraies efficacités de couplage de chaque monomère, mais ils donnent une idée globale de la qualité de la synthèse. Tant qu'aucune barre de l'histogramme n'atteint pas un niveau très bas, on peut estimer que la synthèse se déroule correctement. Pour être plus précis il faudrait effectuer des analyses UV manuelles en collectant les solutions des étapes de déprotections (procédure détaillée dans le chapitre II, paragraphe 3.2.4.2). Ici on remarque que les rendements de chaque étape sont bons, même s'ils semblent décroitre au fur et à mesure des couplages aucun n'est drastiquement plus faible que les autres. Cette diminution du niveau des barres de l'histogramme est référencé dans le manuel d'utilisation car elle est couramment observée.[173] De surcroît, l'efficacité de l'espaceur RISC donné par les barres notés T, semble être dans la moyenne des autres efficacités des monomères de la séquence. L'incorporation d'espaceur RISC dans la séquence semble donc s'effectuer correctement.

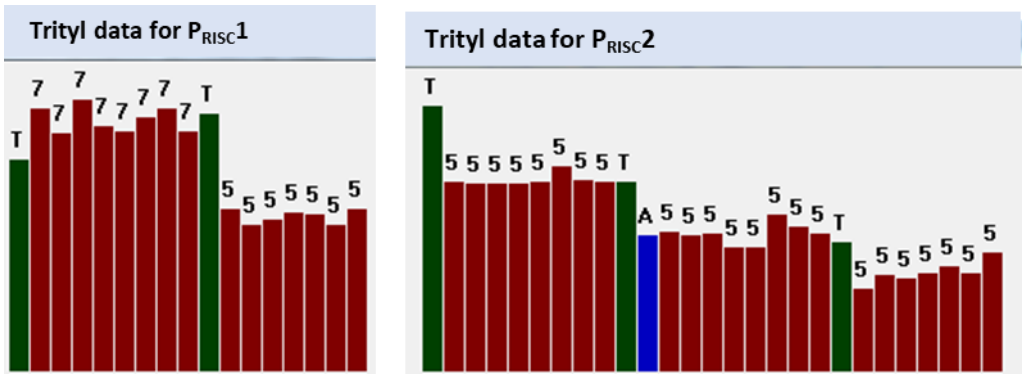


Figure III. 10: Histogrammes obtenus par le synthétiseur Expédite lors de la synthèse des séquences de $P_{RISC}1$ (gauche) et $P_{RISC}2$ (droite).

Toutefois, des problèmes ont été observés lors du clivage des séquences. L'ajout de la solution de clivage dans le support solide a provoqué une coloration jaune orangé. Après la purification complète la solution obtenue est également jaunâtre alors qu'en temps normal elle est transparente. De même, après lyophilisation ce n'est pas une poudre blanche qui est retrouvée, mais plutôt un liquide visqueux orange-marron, comme cela peut être vu sur la Figure III. 11. Cette apparition de couleur peut être le signe d'un clivage anticipé de la liaison alcoxyamine de l'espaceur RISC, pouvant alors libérer un radical TEMPO coloré.

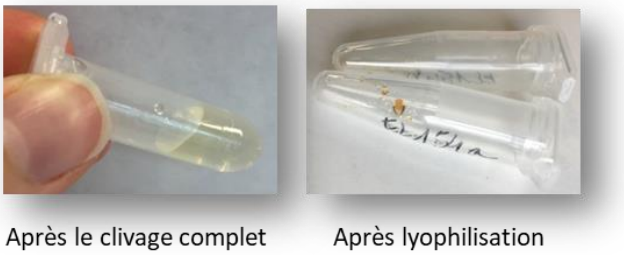


Figure III. 11: Photographies de la solution obtenue après clivage (gauche) et du liquide visqueux obtenu après lyophilisation (droite).

Le polymère recherché a été détecté par spectrométrie de masse. Mais de nombreuses impuretés ont également été détectées en plus large excès. Les impuretés majeures détectées dans les deux échantillons correspondent à des espèces contenant le dernier bloc ou le dernier et l'avant-dernier bloc pour la séquence plus longue **$P_{RISC}2$**, avec des groupements α similaires. La structure de certaines impuretés peut être mise en évidence lors de l'analyse MS/MS. Ces impuretés peuvent s'expliquer par l'apparition de réactions secondaires lors de l'étape de clivage avec la solution de $MeNH_2$: NH_3 (v : v / 1 : 1), montrées en Figure III. 12. L'hypothèse est qu'une substitution nucléophile en position benzylique avec la solution de clivage d'ammoniaque et méthylamine pourrait être à l'origine de ces réactions secondaires.

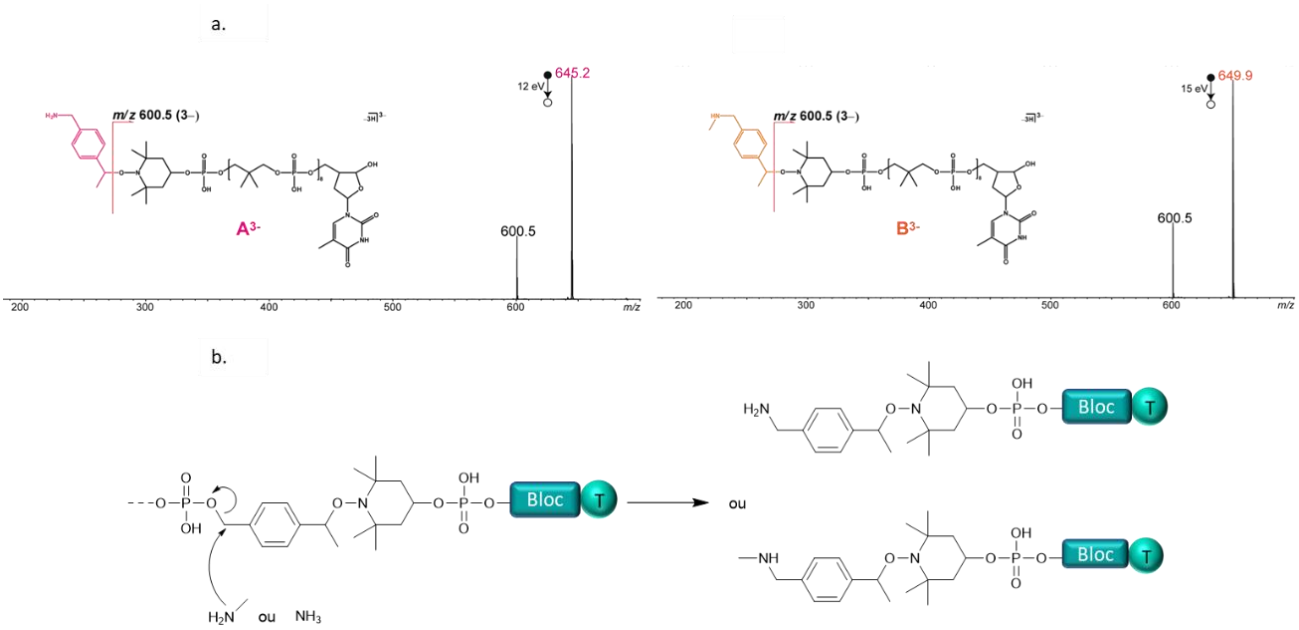


Figure III. 12: (a) Spectres obtenus après l'analyse MS/MS des impuretés détectées dans l'échantillon ***PRISC1****. (b) Réactions secondaires suggérées lors du clivage.*

Ces deux formes d'impuretés sont détectées dans les deux échantillons ainsi que d'autres avec des groupements α de plus grande masse. Malgré l'existence de ces réactions parasites, il est tout de même possible d'identifier la séquence désirée et d'effectuer les analyses de spectrométrie de masse en tandem et de pseudo-MS^3 qui valident cette attribution. Lors de ces analyses aucune réaction parasite due à la forte réactivité des radicaux alkyles générés après l'homolyse de la liaison alcoxyamine de l'espaceur n'est observée. En conséquence, la couverture des séquences en pseudo-MS^3 est totale et tous les fragments de chaque monomère ont été retrouvés (voir Figure III. 13).

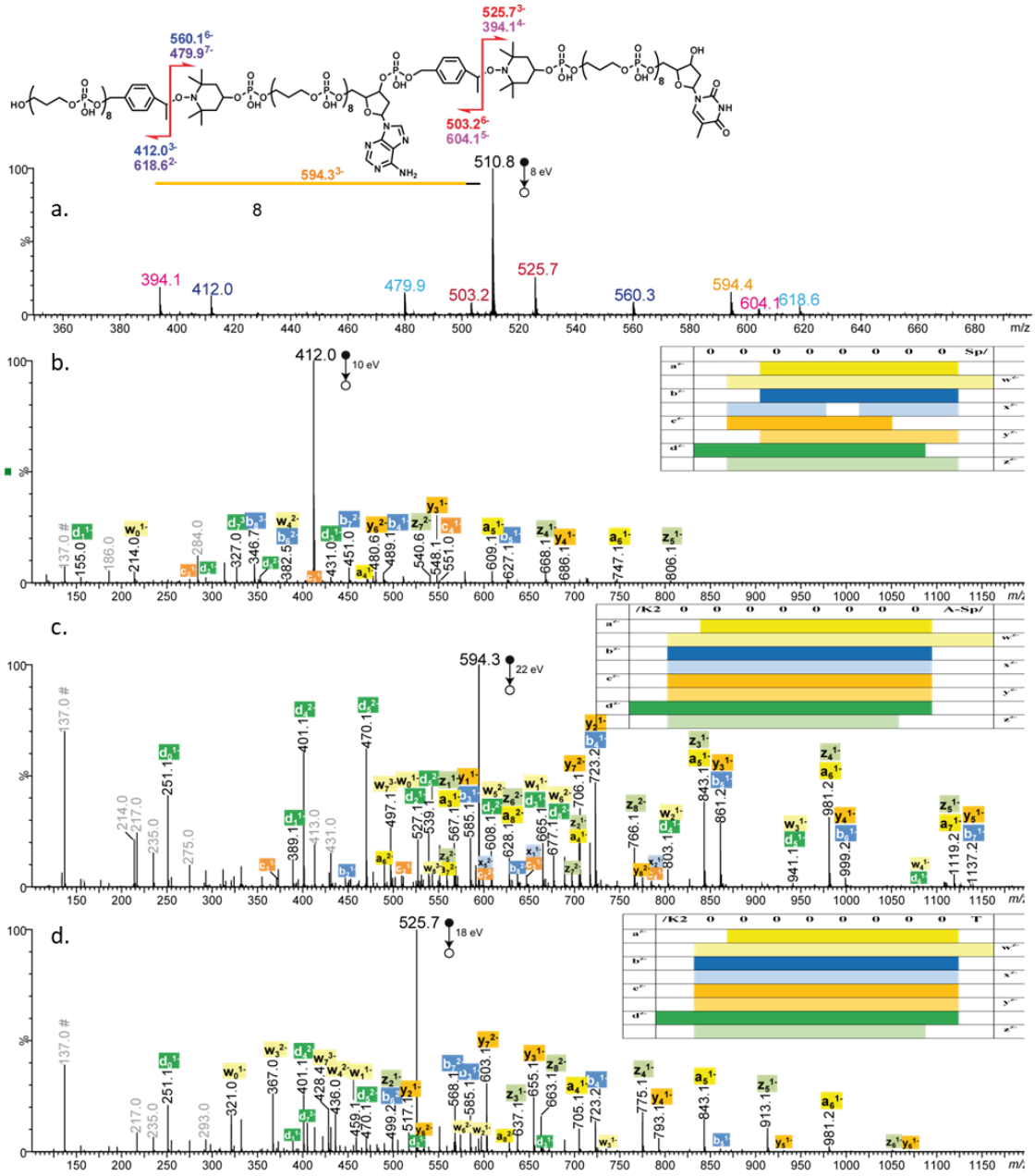


Figure III. 13: Séquençage de la séquence ***PRISC2****. (a) Spectre MS/MS de l'ion [M – 9H]$^{9-}$ montrant l'homolyse des liaisons NO-C. (b) Spectre pseudo-MS^3 du bloc 1 à m/z 412,0 et tableau de couverture du bloc 1. (c) Spectre pseudo-MS^3 du bloc 2 à m/z 594,3 et tableau de couverture du bloc 2. (d) Spectre pseudo-MS^3 du bloc 3 à m/z 525,7 et tableau de couverture du bloc 3. Les pics annotés en gris correspondent à des fragments internes, parmi lesquels les monomères sont désignés par #.*

Cette couverture totale obtenue en analyse pseudo-MS^3 encourageant à optimiser la synthèse de cet espaceur. De nouvelles conditions de clivage plus douces ont donc été testées afin d'éviter les réactions parasites. Il a donc été tenté de cliver les solutions avec juste une solution d'ammoniaque pendant soit 30 min, soit pendant 1h, toujours à température ambiante. Ces essais n'ont pas été concluant, la coloration apparaît toujours lors du clivage des séquences. Les analyses en spectrométrie de masse

montrent toujours les mêmes types d'impuretés comprenant le (ou les) dernier(s) bloc(s) intact(s) mais avec un premier bloc plus court formé par des réactions parasites.

3.2.2. Espaceur RISCOP

L'espaceur RISC semble prévenir des réactions secondaires détectées lors de l'analyse en spectrométrie de masse grâce à la production d'un radical alkyle moins réactif qu'auparavant. Cependant, des problèmes apparaissent lors de la synthèse macromoléculaire. La purification des séquences ne permet pas d'obtenir des polymères monodisperses et crée de nouvelles réactions secondaires qu'il est difficile d'éviter. Une optimisation du design de l'espaceur RISC a donc été imaginée. Ce design optimisé garde les avantages de l'espaceur RISC lors de l'analyse de spectrométrie de masse, mais permet également d'éviter les réactions secondaires lors de la synthèse. La structure de l'espaceur optimisé comprend ainsi un groupement benzyle plutôt qu'un groupement phényle dans son squelette. L'allongement de la chaîne carbonée permet d'éviter la position benzylique réactive. Cette modification a pour but d'éviter la substitution nucléophile avec la solution de clivage en cette position. Cette réflexion a mené à la synthèse de l'espaceur optimisé, nommé **RISCOP** (pour RISC Optimisé) dont la structure est donnée en Figure III. 14.

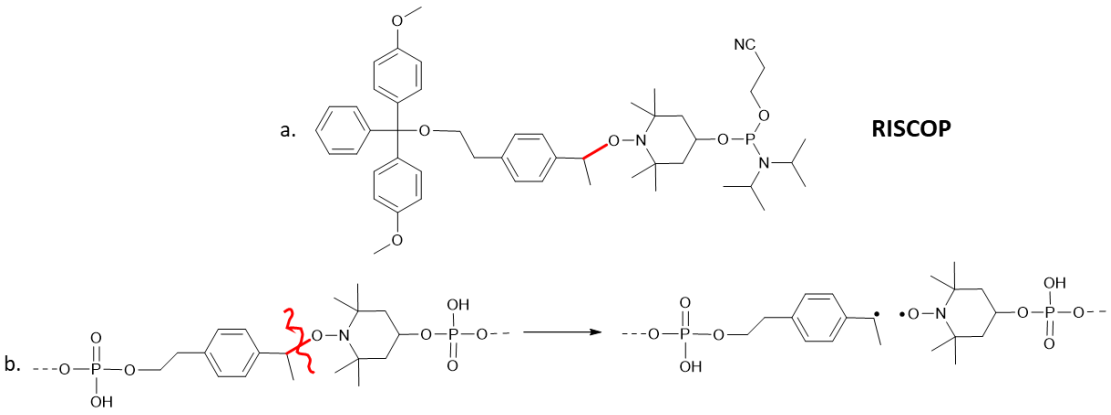


*Figure III. 14: (a) Structure de l'espaceur **RISCOP**. En rouge liaison NO-C fragile. (b) Structure de l'espaceur RISCOP incorporé dans une séquence suivie du schéma de la fragmentation de la liaison alcoxyamine.*

Plusieurs séquences tests ont été effectuées pour vérifier si la synthèse et la purification des échantillons contenant l'espaceur RISCOP se déroulent sans l'apparition de réactions secondaires. Ces dernières seraient alors visibles par la coloration de la solution obtenue après le clivage ainsi que la coloration de l'échantillon après sa lyophilisation. Les séquences synthétisées sont rassemblées dans le Tableau III. 3 suivant.

Tableau III. 3: Description des séquences tests synthétisées avec l'espaceur RISCOP.

Séquence	Séquence de monomères	*M [a] (Da)*	*Impuretés*
$P_{RISCOP}1$	M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-RISCOP-M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-T	2833,4076	-138 Da : M_1
$P_{RISCOP}2$	M_1·M_1·M_1·M_1-RISCOP- M_1·M_1·M_1·M_1·A-RISCOP-M_1·M_1·M_1·M_1-T	2833,4076	-138 Da : M_1

[a] *Masse monoisotopique.*

Lors de la purification des séquences, après leur synthèse aucune coloration n'est apparue, la solution était transparente. Le produit récupéré après lyophilisation n'avait aucune coloration. L'analyse de spectrométrie de masse confirme que la séquence qui a été majoritairement détectée correspond bien au polymère synthétisé avec l'espaceur RISCOP. La seule impureté mineure détectée dans les deux échantillons correspond à un couplage raté, avec le manque d'une unité M1. Le spectre MS/MS des ions précurseurs ainsi que le séquençage pseudo-MS^3 des deux séquences tests permettent de valider les attributions. De plus, le radical carboné généré comme groupement ω du bloc 1 après homolyse de l'alcoxyamine de l'espaceur RISCOP n'induit plus aucune des réactions indésirables qui altèrent usuellement la couverture de la séquence de ce bloc en pseudo-MS^3. L'analyse complète du polymère **P_{RISCOP}2** est donnée en Figure III. 15.

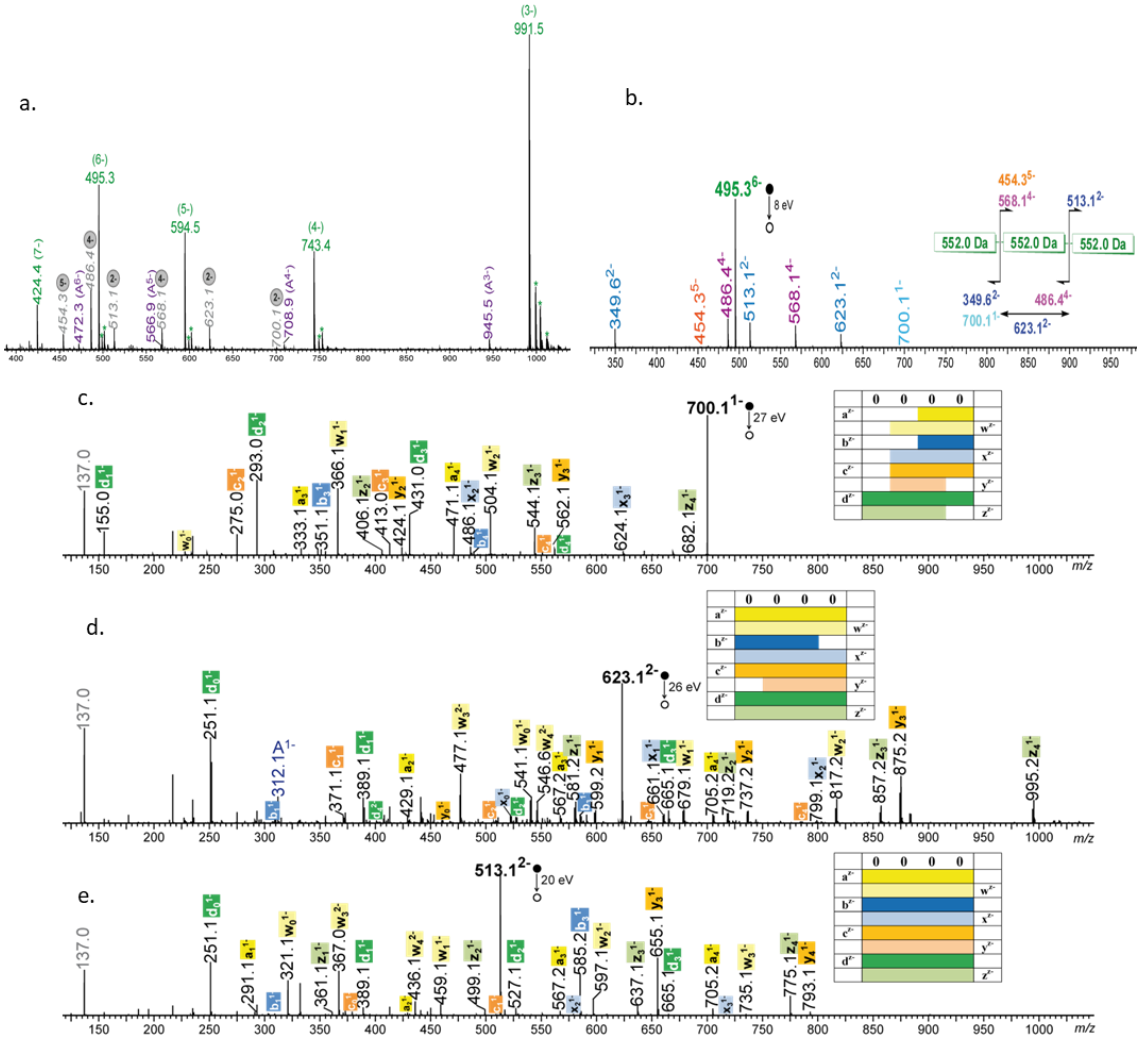


*Figure III. 15: Séquençage de la séquence **P_{RISCOP}2**. (a) Spectre ESI-MS en mode négatif. (b)Spectre MS/MS de l'ion [M – 6H]$^{6-}$ montrant l'homolyse des liaisons C–ON. (c) Spectre pseudo-MS^3 du bloc 1 à m/z 700,1 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 2 à m/z 623,1 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 3 à m/z 513,1 et couverture de la séquence. Les pics annotés en gris correspondent au monomères déprotonés. Les ions annotés avec un rond gris correspondent aux fragments libérés en source. Les pics annotés par un astérisque correspondent à des échanges H/Na ou H/K.*

Les spectres de pseudo-MS^3, donnés en Figure III. 15, c, d et e, ne montrent aucune détection de réactions parasites et la couverture des séquences est facilement assurée quelle que soit la localisation initiale des blocs. Cependant, l'analyse du bloc central montre systématiquement des ions supplémentaires à +1 m/z pour les fragments a, b, c, d et à -1 m/z pour les fragments w, x, y, z. Ces ions sont le signe de réactions de transfert de H·. Deux hypothèses sont à vérifier pour déterminer l'origine de ces réactions : (i) elles peuvent être dues à la taille des blocs de la séquence. Ici, la séquence synthétisée comprend trois blocs de quatre monomères chacun menant à une détection des états de charge 2-. Cette taille limitée et le faible état de charge peuvent mener aux réactions de transferts observées. (ii) Elles peuvent aussi avoir pour origine la combinaison du marqueur dA et de l'espaceur RISCOP.

Une séquence test a été étudiée pour trancher entre ces deux hypothèses. Il s'agit une séquence de 4 blocs comprenant 8 monomères chacun avec les marqueurs dT, dC et dA. Si aucune réaction de transfert n'est détectée, l'hypothèse des blocs de faibles tailles sera retenue et les prochaines séquences synthétisées avec l'espaceur RISCOP devront toujours comprendre des blocs d'au moins 8 unités. En revanche, si des réactions de transfert sont détectées dans le deuxième bloc comprenant le marqueur dA, cela indiquera que les réactions de transfert sont causées par la combinaison du marqueur dA et de l'espaceur RISCOP et donc le marqueur dA ne sera pas utilisé dans les prochaines synthèses impliquant l'espaceur RISCOP.

Néanmoins, ces réactions de transfert de H· ne sont pas un problème lors d'un séquençage manuel et la séquence peut être retrouvée. Ces réactions de transfert peuvent cependant, devenir gênantes lors d'un séquençage automatique. Des développements ont été fait et il est maintenant possible d'appliquer l'algorithme du MS-DECODER au séquençage des chaînes de poly(phosphodiester)s numériques. Le logiciel MS-DECODER permet de retrouver automatiquement, en quelques secondes, l'information contenue dans les macromolécules.[128] Cette analyse automatique était impossible jusqu'ici à cause des réactions parasites se produisant en pseudo-MS^3. Mais les avancements obtenus avec l'espaceur RISCOP semblent prometteurs et si les réactions de transferts de H· sont écartées il sera alors possible de tester de décoder automatiquement des séquences comprenant le RISCOP.

La séquence **P_{RISCOP}3**, comprenant 4 blocs de huit monomères M1 avec les marqueurs dA, dC et dT a donc été synthétisée (Tableau III. 4).

Tableau III. 4: Description de la longue séquence tests synthétisées avec l'espaceur RISCOP.

Séquence	Séquence de monomères	*M[a] (Da)*	*Impuretés*
P_{RISCOP}3	M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-RISCOP- M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1·A-RISCOP- M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1·C-RISCOP- M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-T	6410,0150 Da	/

[a] *Masse monoisotopique.*

A nouveau la synthèse et la purification se sont bien déroulées et lors de l'analyse de spectrométrie de masse aucune impureté n'a été détectée, comme il est possible de voir sur le spectre obtenu et montré en Figure III. 16 .

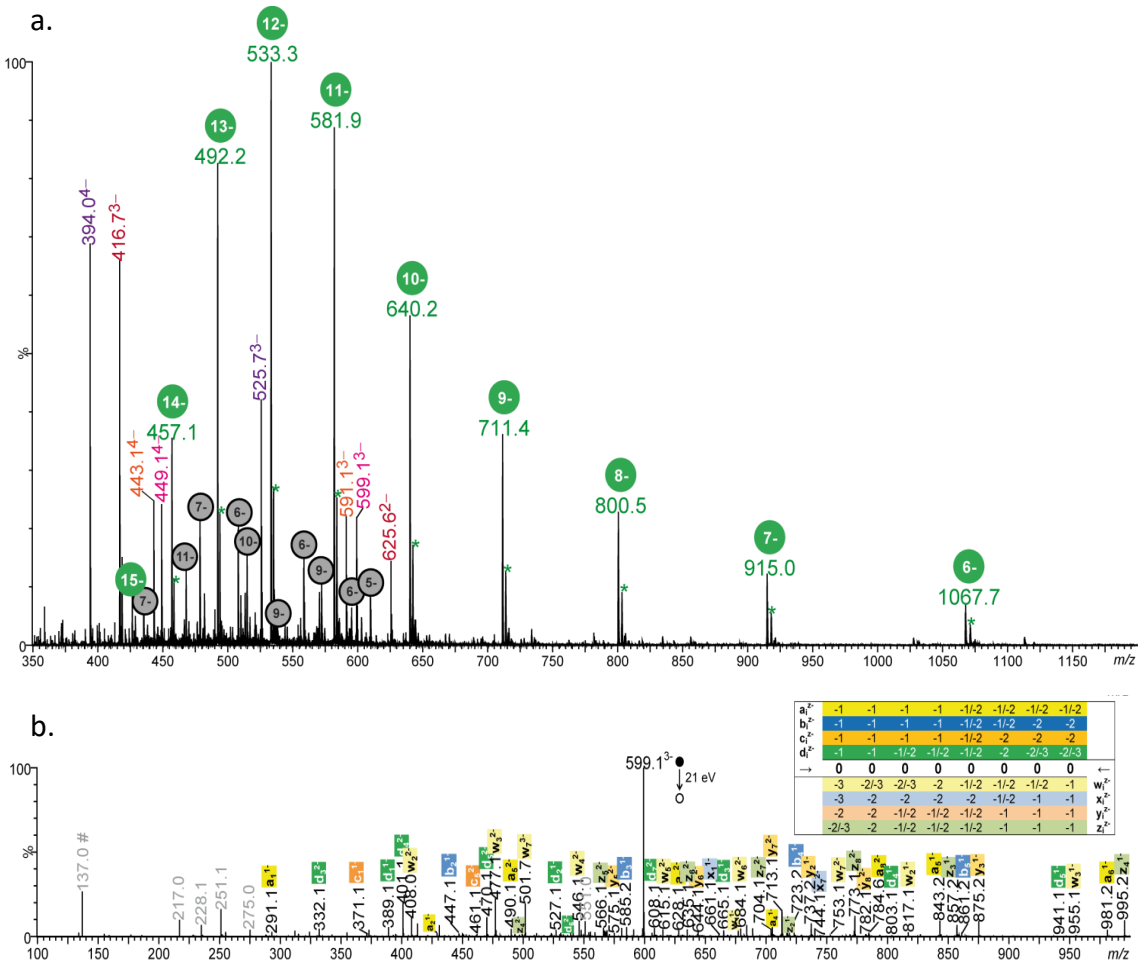


Figure III. 16: Séquençage de la séquence ***PRISCOP3****. (a) Spectre ESI-MS en mode négatif. Séquence détectée sous les différents états de charge 6– à 15– (en vert) ainsi que fragments générés en source pour homolyse des différentes liaisons alcoxyamine : bloc 1 en rouge, bloc 2 en rose, bloc 3 en orange et bloc 4 en violet. (b) Spectre pseudo-MS³ du bloc 2 comprenant le marqueur dA à m/z 599,1 et couverture de la séquence. Les pics annotés en gris sont des fragments libérés en source. * : échanges H/Na.*

Lors de l'analyse MS la séquence a été détectée sous les états de charge 6- à 15- (en vert sur la Figure III. 16.a). Les autres ions détectés correspondent aux fragments générés en sources. Ceux indiqués en couleur rouge, rose, orange et violet correspondent au clivage des liaisons alcoxyamines menant à la détection de fragments pour chaque bloc. La Figure III. 16.b montre l'analyse en pseudo-MS³ du bloc 2 comprenant le marqueur dA. Ce spectre ne montre aucune réaction de transfert détectée lors de l'analyse de la séquence PRISCOP2. L'hypothèse impliquant des réactions de transferts causées par la combinaison marqueur dA et l'espaceur RISCOP n'est donc pas vérifiée et ce marqueur peut toujours être utilisé avec ce nouvel espaceur.

En outre, aucune réaction de transfert n'est détectée dans aucun des blocs. Les réactions parasites précédemment observées, peuvent être attribuées à la longueur des blocs qui étaient trop courts. Désormais des blocs comprenant au moins 8 monomères seront synthétisés lors de l'utilisation de l'espaceur RISCOP.

3.2.2.1. Séquençage automatique avec le logiciel MS-DECODER

Grâce aux très bons résultats obtenus avec la dernière séquence test $P_{RISCOP}3$, il est possible de tenter la détection automatique de la séquence via l'algorithme MS-DECODER.

La construction de cet algorithme se base sur le fait que ces séquences de poly(phosphodiester)s sont divisées en différents sous-segments séparés par les espaceurs. En connaissant le nombre de bits compris dans chaque bloc, il est possible de lister un nombre raisonnable de combinaisons potentielles ainsi que leur masse théorique. Lorsque l'alphabet classique est utilisé, on retrouve 1 bit par monomère. Chaque bloc étant composé de 8 monomères, il y a donc 8 bits par blocs. Ainsi, il y a 2^8 soit 256 combinaisons possibles qui vont de 00000000 à 11111111.

L'analyse automatique se déroule en deux temps. Tout d'abord, le manipulateur doit indiquer la valeur de l'ion précurseur qui a été utilisé pour effectuer l'analyse en pseudo-MS^3, ainsi qu'une tolérance sur les valeurs de m/z détectées (à + / – 0,1 par exemple). Cet ion correspond à l'ion détecté lors de l'analyse MS/MS qui permet de déterminer le bloc duquel il est issu. Avec la valeur de cet ion, l'algorithme peut retrouver sans aucun doute la position dans la séquence du bloc et sa composition. Pour cela, les valeurs théoriques de chaque composition possible (*i.e* nombre de monomères 0 et de monomère 1 dans le bloc) avec tous les marqueurs sont calculés. Pour l'alphabet classique à 2 monomères, on obtient le Tableau III. 5. Il est alors possible de retrouver la valeur de l'ion précurseur dans le tableau et de savoir son rang ainsi que sa composition. Par exemple, avec un ion précurseur à m/z 644 (en orange dans le tableau), on retrouve qu'il est localisé dans le bloc n-2 contenant le marqueur dA et qu'il contient 3 bits 0 et 5 bits 1. L'algorithme garde en mémoire la position de ce bloc, qu'il connaît en fonction du marqueur qu'il détecte. L'ordre dans lequel sont utilisés les marqueurs dans la séquence est implémentée dans l'algorithme, il peut alors remettre les blocs dans le bon ordre tout seul. A cette étape, on ne connaît pas encore l'ordre dans lequel les bits sont disposés.

Tableau III. 5: Valeur de m/z attendues pour des segments triplements chargés généré pendant une analyse CID de séquences de poly(phosphodiester)s préparées avec l'espaceur RISCOP.

Composition		Rang 1	Rang (n – 2)	Rang (n – 1)	Rang n
0	**1**	"pas de marqueur"	dA	dC	dT
8	0	415.0487	597.4311	589.4273	525.7412
7	1	424.3925	606.7749	598.7711	535.0849
6	2	433.7362	616.1186	608.1149	544.4287
5	3	443.0800	625.4624	617.4586	553.7725
4	4	452.4238	634.8062	626.8024	563.1162
3	5	461.7675	644.1499	636.1462	572.4600
2	6	471.1113	653.4937	645.4899	581.8038
1	7	480.4551	662.8375	654.8337	591.1475
0	8	489.7988	672.1812	664.1775	600.4913

Par la suite, pour déterminer l'ordre des bits de la composition trouvée, il faut calculer tous les m/z théoriques de chaque série d'ions produits (*i.e.* a_i^{z-}, b_i^{z-}, c_i^{z-}, d_i^{z}, w_i^{z-}, x_i^{z-}, y_i^{z-}, z_i^{z-}) pour chaque état de charge de 1 à 3. Cette étape permet de générer une palette de valeurs de m/z unique qui sera possible de comparer avec les valeurs de m/z enregistrées lors de l'analyse. Cette étape permet donc de trouver les concordances et ainsi de reconstruire la séquence dans le bon ordre. L'algorithme lit donc le spectre, et détermine quelle combinaison obtient le plus de fragments détectés avec une intensité globale la plus haute (somme des intensités de tous les pics). Une fois cette combinaison obtenue, il la garde en mémoire jusqu'à ce que tous les blocs (*i.e*. tous les spectres de pseudo-MS^3) soient analysés. Lorsque tous les blocs sont analysés, il peut alors remettre dans l'ordre l'information en fonction des marqueurs de masse, dont il connaît l'ordre.

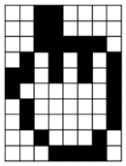
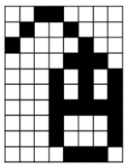
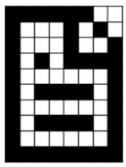
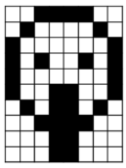
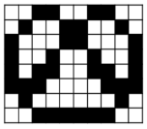
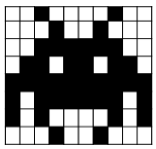
Grâce à cet algorithme, le temps d'analyse ne dépend pas de la complexité des données. En effet, l'analyse de chaque polymère devrait prendre environ le même laps de temps, car après avoir généré les valeurs de m/z théoriques, l'algorithme n'a qu'à lire une seule fois les spectres expérimentaux pour trouver les concordances avec les valeurs théoriques.

Les données obtenues lors de l'analyse de la séquence $P_{RISCOP}3$ sont donc implémentées dans l'algorithme développé pour les séquences de poly(phosphodiester)s utilisant l'alphabet classique. Le séquençage et la recomposition de la séquence initiale est effectuée parfaitement en 50 à 100 ms.

4. Synthèse de polymères incorporant des images

A la suite des résultats encourageant obtenus avec la synthèse de la séquence $P_{RISCOP}3$, des séquences plus complexes et codant des images ont été synthétisées en utilisant l'espaceur RISCOP. L'alphabet augmenté présenté dans le **chapitre II** comportant les quatre monomères M1 à M4 a été utilisé, formant ainsi des séquences composées de dyades. Six nouvelles petites images comprenant 80 pixels ont été synthétisées et sont rassemblées dans le Tableau III. 6. Pour simplifier la liste des séquences de monomères donné dans le tableau, l'espaceur RISCOP est noté « ER ». Ainsi une main (**P_R1**), une souris d'ordinateur (**P_R2**), un fichier (**P_R3**), une tête de cigogne (**P_R4**), un bretzel (**P_R5**) et un space invader (**P_R6**) ont été pixélisées.

Tableau III. 6: Description des séquences codant des images synthétisées avec l'espaceur RISCOP noté ici « ER ».

	Image	Séquence binaire [a]	Séquence de monomères [b]	*C* [c]	*DP* [d]	*M* [e] *(Da)*
P_R1		00110000.00110000. 00111110.01010101. 11010101.10000001. 01000001.01000001. 00100010.00111110.	M_1·M_4·M_1·M_1·M_1·M_4·M_1·M_1-*ER*- M_1·M_4·M_4·M_3·M_2·M_2·M_2·M_2·G-*ER*- M_4·M_2·M_2·M_2·M_3·M_1.M_1·M_2·A-*ER*- M_2·M_1·M_1·M_2·M_2·M_1·M_1·M_2·C-*ER*- M_1·M_3·M_1·M_3·M_1·M_4·M_4·M_3-T	80	48	8828.9923
P_R2		00110000.01001000. 10000100.00001110. 00010101.00010101. 00011111.00010001. 00010001.00001110.	M_1·M_4·M_1·M_1·M_2·M_1·M_3·M_1-*ER*- M_3·M_1·M_2·M_1·M_1·M_1·M_4·M_3·G-*ER*- M_1·M_2·M_2·M_2·M_1·M_2·M_2·M_2·A-*ER*- M_1·M_2·M_4·M_4·M_1·M_2·M_1·M_2·C-*ER*- M_1·M_2·M_1·M_2·M_1·M_1·M_4·M_3-T	80	48	8730.8827
P_R3		11111100.10001010. 10001001.10101111. 10000001.10111101. 10000001.10111101. 10000001.11111111.	M_4·M_4·M_4·M_1·M_3·M_1·M_3·M_3-*ER*- M_3·M_1·M_3·M_2·M_3·M_3·M_4·M_4·G-*ER*- M_3·M_1·M_1·M_2·M_3·M_4·M_4·M_2·A-*ER*- M_3·M_1·M_1·M_2·M_3·M_4·M_4·M_2·C-*ER*- M_3·M_1·M_1·M_2·M_4·M_4·M_4·M_4-T	80	48	9193.3992
P_R4		00111100.01000010. 10000001.10100101. 10000001.10011001. 01011010.00111100. 00011000.00011000.	M_1·M_4·M_4·M_1·M_2·M_1·M_1·M_3-*ER*- M_3·M_1·M_1·M_2·M_3·M_3·M_2·M_2·G-*ER*- M_3·M_1·M_1·M_2·M_3·M_2·M_3·M_2·A-*ER*- M_2·M_2·M_3·M_3·M_1·M_4·M_4·M_1·C-*ER*- M_1·M_2·M_3·M_1·M_1·M_2·M_3·M_1-T	80	48	8857.0236
P_R5		00110011.00010011. 00101000.11000110. 01001001.10110011. 01111000.01110100. 00001000.11111100.	M_1·M_4·M_1·M_4·M_1·M_2·M_1·M_4-*ER*- M_1·M_3·M_3·M_1·M_4·M_1·M_2·M_3·G-*ER*- M_2·M_1·M_3·M_2·M_3·M_4·M_1·M_4·A-*ER*- M_2·M_4·M_3·M_4·M_2·M_4·M_2·M_4·C-*ER*- M_1·M_1·M_3·M_1·M_4·M_4·M_4·M_1-T	80	48	8983.1644
P_R6		00100001.00000100. 10000011.11110001. 10110110.11111111. 11101111.11011010. 00010100.01001000.	M_1·M_3·M_1·M_2·M_1·M_1·M_2·M_1-*ER*- M_3·M_1·M_1·M_4·M_4·M_4·M_1·M_2·G-*ER*- M_3·M_4·M_2·M_3·M_4·M_4·M_4·M_4·A-*ER*- M_4·M_3·M_4·M_4·M_4·M_2·M_3·M_3·C-*ER*- M_1·M_2·M_2·M_1·M_2·M_1·M_3·M_1-T	80	48	9067.2583

[a] Les séquences binaires ont été séparées en bloc de huit bits pour plus de clarté.[b] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [c] Capacité de stockage en bit/chaîne. [d] Degré de polymérisation incluant les monomères, l'espaceur et les marqueurs de masse. [e] Masse monoisotopique.

Comme pour la séquence test précédente, ces nouvelles séquences codant des images sont composées de blocs de huit monomères codants. Les séquences ont été clivées pendant 45 min chacune puis

lyophilisées et récupérées sous formes d'une poudre blanche à la suite de la procédure de purification. Les rendements obtenus pour chaque séquence sont donnés Tableau III. 7.

Tableau III. 7 : Rendement des synthèses de polymères codant des images avec l'espaceur RISCOP.

	m (mg)	Rendement (%)
P_R1	7,3	83
P_R2	6,7	77
P_R3	6,4	70
P_R4	6,6	74,5
P_R5	6,7	75
P_R6	6,6	73

4.1. Analyse et lecture manuelle des séquences

4.1.1. Analyse par chromatographie en phase liquide à haute performance

Les séquences ont été analysées en chromatographie en phase liquide à haute performance (HPLC). Les spectres obtenus sont rassemblés en Figure III. 17. Ces analyses montrent pour chaque séquence un pic principal à environ 11min. Ces séquences ont toutes la même taille, à savoir 48 unités. Il est donc normal de retrouver un temps de rétention semblable pour toutes les séquences. Les pics principaux ont tous des épaulements. Celui éluant après le pic principal est certainement issus de chaînes non déprotégées. Il se peut que lors de la purification le groupement DMT final et/ou les groupes protecteurs de bases nucléiques ne soi(en)t pas bien clivé(s). Si une seule chaîne non-déprotégée est présente dans l'échantillon, elle suffit à créer ce petit épaulement.[174] Les analyses sont faites avec une colonne échangeuse d'ions, les interactions entre la chaîne du polymère et la colonne sont donc différentes si les groupements protecteurs sont encore présents. La différence d'interaction provoque alors ces petits épaulements que l'on remarque sur les spectres.

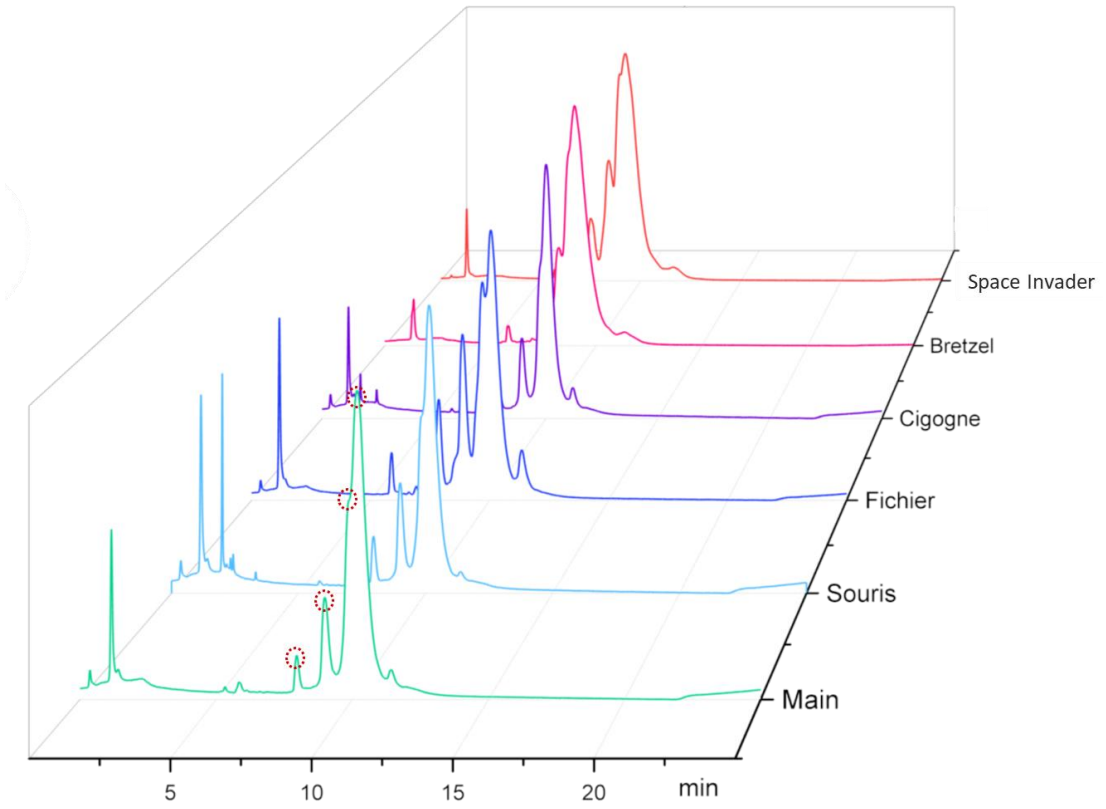


Figure III. 17: Spectres obtenus après les analyses HPLC des séquences P_R1 (Main en verte), P_R2 (Souris en bleu clair), P_R3 (Fichier en bleu), P_R4 (Cigogne en violet), P_R5 (Bretzel en rose) et P_R6 (Space invader en orange).

Les épaulements et les autres petits pics visibles sur les spectres avant le pic principal, peuvent venir de chaînes plus courtes ou de bases libres qui n'ont pas été lavées pendant la purification. Ce type d'impureté est plus visible en HPLC, car la détection s'effectue seulement via les bases nucléiques présentes dans la séquence. Lors de l'analyse, l'absorbance est mesurée à 269 nm ce qui correspond au maximum d'absorption des bases nucléiques. Dans les séquences analysées, seulement quatre bases nucléiques sont présentes (dT, dC, dA et dG). De ce fait, les impuretés contenant même juste une base nucléique (chaîne courte ou base libre), peuvent être suffisantes pour être détecté. Ce signal peut être assez intense en comparaison à la séquence recherchée si le nombre de base compris dans l'impureté se rapproche de celui de la séquence. Ainsi, les épaulements et signaux visibles sur les spectres peuvent être surestimés.

Il semble avoir 3 impuretés pour chaque séquence. Elles sont entourées en pointillés rouges pour la première courbe obtenue pour la séquence P_R1. La séquence désirée contient 4 marqueurs de blocs (dT, dC, dA et dG), les trois impuretés pourraient alors correspondre à des sous-séquences contenant des blocs en moins, donc une base nucléique en moins à chaque impureté. L'épaulement correspondrait donc à une sous-séquence comprenant 3 marqueurs donc 3 blocs, puis les pics comprendraient à une sous-séquence avec 2 marqueurs donc 2 blocs puis finalement plus qu'à 1 marqueur donc un seul bloc. Ces impuretés peuvent correspondre à des clivages partiels des espaceurs lors de l'analyse HPLC qui provoquerait la

détection de sous-séquences. Le nombre de pic sur le spectre HPLC correspond alors au nombre total de bloc contenant un marqueur dans la séquence.

Les analyses de spectrométrie de masse apportent une preuve en plus que la séquence recherchée est bien la séquence principale de l'échantillon.

4.1.2. Lecture par spectrométrie de masse

Les analyses de spectrométrie de masse montrent que dans tous les cas les séquences ont été retrouvées, les spectres des analyses ESI-MS sont donnés en Figure III. 18. L'analyse des séquences P_R1, P_R3, P_R5 et P_R6 codant respectivement pour l'image d'une main, d'un fichier, d'un bretzel et d'un space invader montre que les échantillons sont monodisperses. Les signaux détectés correspondent uniquement aux séquences désirées.

De très faibles signaux ne correspondant pas à la séquence recherchée peuvent être détectés avec une très faible intensité. De ce fait, ils se trouvent dans le bruit de fond et ne sont pas analysés. Les épaulements détectés en HPLC, peuvent ainsi se retrouver en bruit de fond lors de l'analyse en spectrométrie de masse car la détection n'est pas la même. En HPLC, c'est l'absorption de bases nucléiques qui permet la détection d'une espèce, ce qui n'est pas le cas en MS. Une impureté visible en HPLC peut devenir très faible lors de l'analyse de spectrométrie de masse. C'est pour cela que l'on ne retrouve pas d'impuretés dans les séquences P_R1, P_R3, P_R5 et P_R6 alors que l'analyse d'HPLC semblait indiquer la présence de certaines.

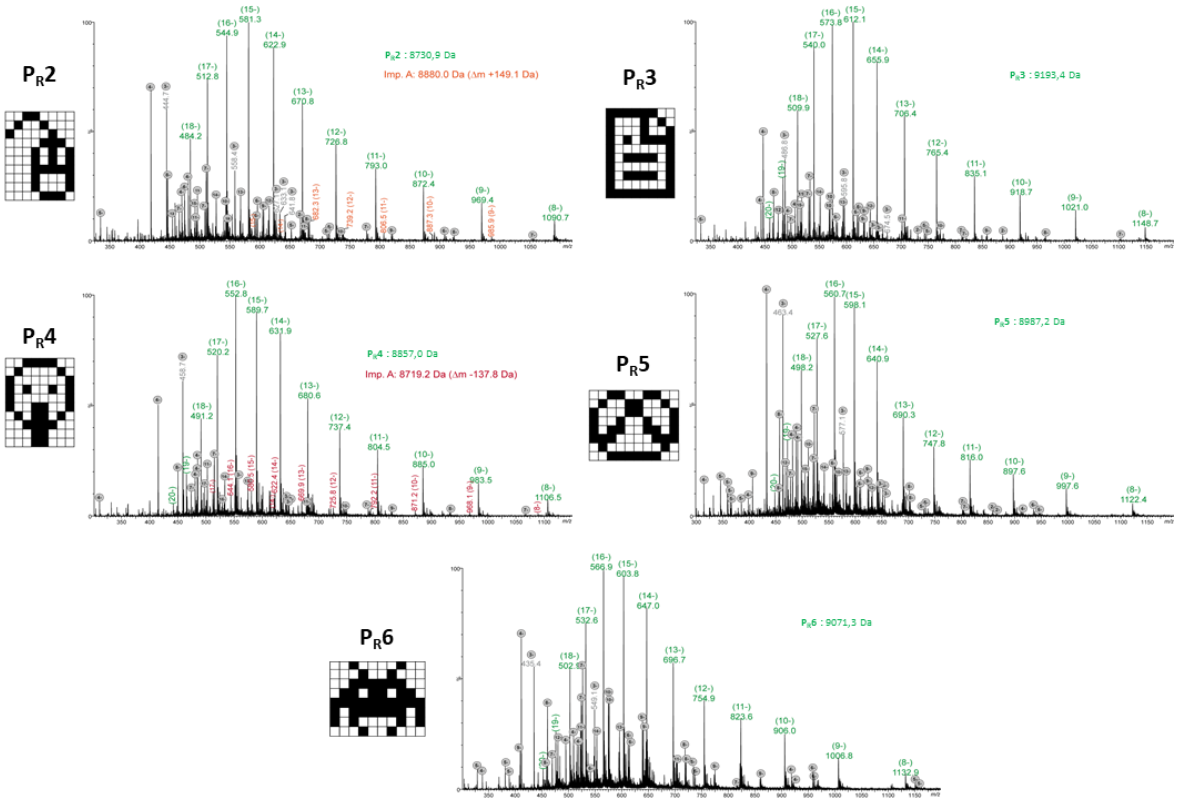


Figure III. 18: Spectre ESI-MS en mode négatif des séquences P_R2 à P_R6. Les oligomères recherchés sont représentés en vert. Les échantillons P_R2 et P_R4 contiennent également une petite impureté à +149 Da et -137,8 Da respectivement. Les annotations en grises sont des fragments formés en source.

L'analyse complète de la séquence P_R1 est donnée en Figure III. 19. Concernant l'analyse des séquences P_R2 et P_R4, une impureté mineure est détectée dans chaque échantillon. L'impureté de l'échantillon P_R4 correspond à la perte d'un monomère M1. Concernant celle de l'échantillon P_R2 il s'agit d'une différence de masse Δ = +149 Da. Cette impureté étant présente en faible quantité, aucune analyse supplémentaire n'a pu être effectuée. Il se peut qu'elle soit issue d'une mauvaise déprotection des marqueurs de masse.

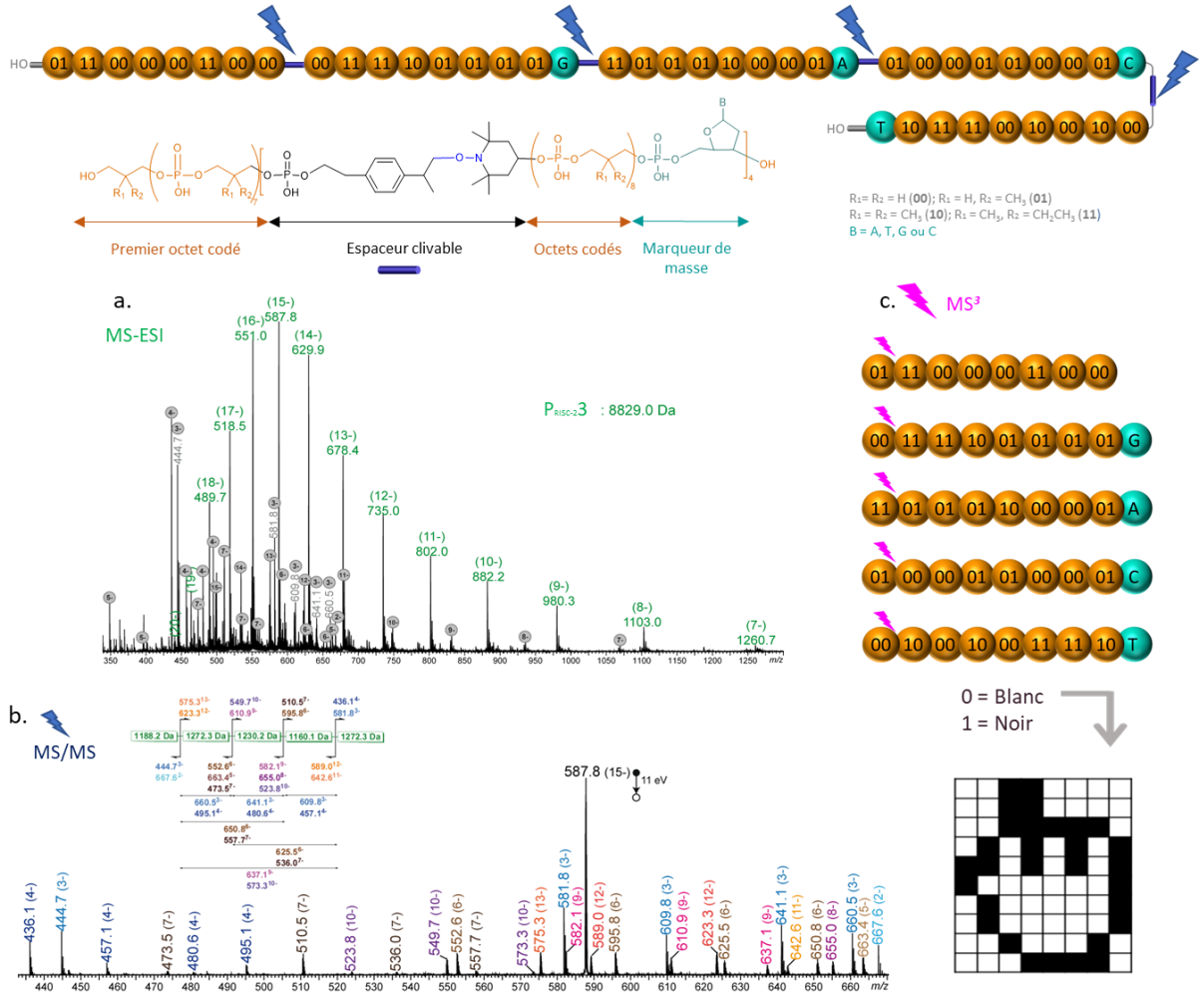


Figure III. 19: Analyse de spectrométrie de masse du polymère P_R1. (En haut) Schéma de la structure globale du polymère P_R1. (a) Spectre MS montrant les différents états de charges du polymère. Les annotations grises sont des fragments formés en source. (b) Spectre de masse en tandem issu de l'ion précurseur $[M-15H]^{15-}$ à m/z 587,8, ainsi que le schéma de fragmentation. (c) Blocs issus du spectre MS/MS, puis analyse MS^3 pour retrouver l'image encodée.

4.2. Lecture automatisée des séquences avec le logiciel MS-DECODER

L'algorithme MS-DECODER a permis d'obtenir un séquençage automatique de la séquence test présentée dans le paragraphe précédent. Cet algorithme peut également être utilisé pour des séquences plus complexes comme les séquences P_R1 à P_R6 codant de petites images. Néanmoins, une nouvelle version de l'algorithme doit être écrite pour qu'elle soit compatible avec l'alphabet à 4 symboles qui est utilisé pour synthétiser ces séquences.

A nouveau aucune nouvelle analyse de spectrométrie de masse n'est nécessaire pour obtenir cette lecture automatisée. Les mêmes fichiers obtenus lors de l'analyse manuelle en pseudo-MS^3 sont utilisés par l'algorithme MS-DECODER. Il effectue alors la lecture automatiquement et reconstitue également automatiquement les images à partir des données qu'il a recouvré.

La lecture des spectres s'effectue d'une façon proche de celle utilisée pour les séquences codées avec l'alphabet à 2 symboles. La grosse différence porte sur le nombre des combinaisons possibles. En effet, cette fois-ci les blocs comportent toujours 8 monomères, mais ils codent pour 16 bits et non plus 8. Il y a donc 2^{16} soit 65 536 combinaisons possibles pour organiser les bits dans un bloc. Le calcul des m/z théoriques de chaque série d'ions produits (*i.e.* a_i^{z-}, b_i^{z-}, c_i^{z-}, d_i^{z}, w_i^{z-}, x_i^{z-}, y_i^{z-}, z_i^{z-}) pour chaque état de charge de 1 à 3, prend donc plus de temps.

Tableau III. 8: Valeur de m/z attendues pour des segments triplements chargés généré pendant une analyse CID de séquences de poly(phosphodiester)s préparées avec l'espaceur RISCOP et l'alphabet augmenté à 4 symboles.

Composition				Rang 1	Rang (n-3)	Rang (n-2)	Rang (n-1)	Rang n
00	**01**	**10**	**11**	Pas de marqueur	dG	dA	dC	dT
8	0	0	0	416.7	604.4	599.1	591.1	525.7
6	0	1	1					
5	2	0	1					
5	1	2	0	440.1	627.8	622.5	614.5	549.1
4	3	1	0					
3	5	0	0					

En outre, avec l'alphabet classique ne comprenant que deux monomères, la valeur de l'ion précurseur ne correspondait qu'à une seule composition possible. Avec l'alphabet augmenté à 4 symboles, la valeur m/z de l'ion précurseur peut correspondre à plusieurs compositions des 4 monomères. Par exemple, le Tableau III. 8 donne les compositions possibles pour des valeurs de m/z à 416,7 et à 440,1. L'ion à m/z 416,7 correspond qu'a une seule composition, mais l'ion à m/z 440,1 peut correspondre à 5 compositions différentes. Ainsi, lorsque les compositions possibles sont nombreuses, le temps de calcul est forcément impacté. La lecture des spectres prend donc un peu plus de temps, mais toutes les séquences testées sont tout de même retrouvées en quelques secondes.

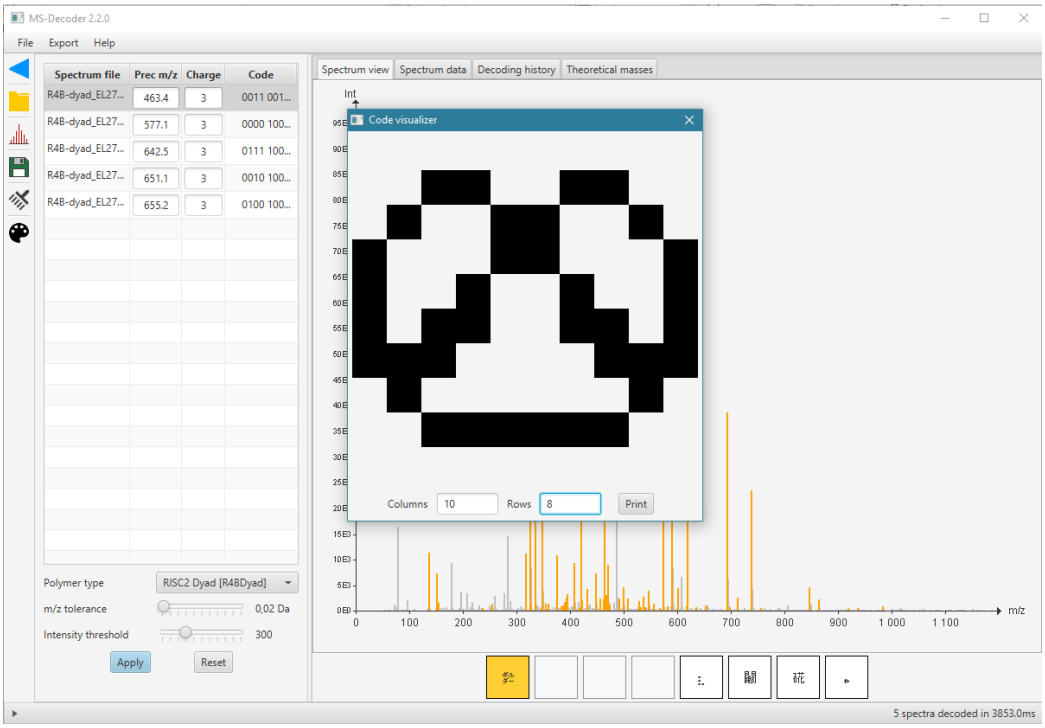


Figure III. 20: Résultats obtenus par le MS-DECODER lors de l'analyse de la séquence ***PR5*** *codant pour l'image d'un bretzel.*

La Figure III. 20 montre une capture d'écran des résultats obtenus avec MS-DECODER lors de l'analyse de la séquence PR5 codant pour l'image d'un bretzel. Sur la gauche de l'écran, on retrouve l'endroit où la valeur de l'ion précurseur pour chaque bloc est notée ainsi que la tolérance en m/z (ici 0,02). A la fin de l'analyse, on voit apparaître au milieu de l'écran l'image du bretzel qui est généré automatiquement par l'algorithme. Il suffit de lui donner la taille de l'image. En bas à droite, on retrouve le temps total de l'analyse qui est de 3853 ms pour la séquence PR5.

Il serait possible d'étendre l'algorithme pour l'utiliser avec l'alphabet augmenté à 8 symboles, mais à nouveau le nombre de composition possibles pour des blocs à 8 unités triades seraient considérablement augmenté. Il passerait de 65 536 combinaisons à 2^{24} soit 16 777 216 possibilités. Le temps de calcul serait alors beaucoup plus long.

5. Amélioration de l'espaceur afin de lui conférer de nouvelles propriétés

En parallèle des optimisations effectuées pour chercher un espaceur permettant une meilleure couverture de séquence lors de l'analyse de pseudo-MS^3, des recherches ont été menées pour apporter de nouvelles propriétés à l'espaceur. Deux points essentiels ont été étudiés : (i) l'apport de la photosensibilité et (ii) l'élaboration d'un espaceur-marqueur.

5.1. Elaboration d'espaceurs photosensibles

Des alcoxyamines photosensibles ont été synthétisées afin de produire des macromolécules contenant de l'information photo-clivables. Ces nouvelles alcoxyamines sont basées sur le design de l'espaceur RISC et

comprennent donc un groupement phényle inséré dans la chaîne principale. Deux différents groupements photosensibles ont été testés. Dans un premier temps un groupement nitro a été substitué en *ortho* du noyau aromatique. Il est nommé PRISC pour « Photosensible Ring InSide the Chain ». Une deuxième molécule photosensible a été synthétisée, elle contient un groupement naphtalène dans la chaîne principale de la molécule. Elle est nommée NISC pour « Naphtalène InSide the Chain ». La structure des deux molécules photosensibles est visible sur la Figure III. 21.

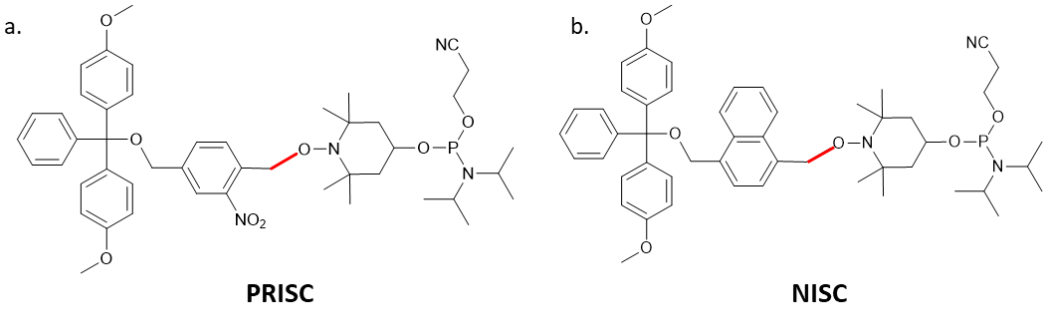


Figure III. 21: (a) Structure de l'espaceur PRISC. (b) Structure de l'espaceur NISC. En rouge les liaisons NO-C fragiles.

Ces espaceurs photosensibles PRISC et NISC ont été incorporées dans différentes séquences tests permettant de juger leurs efficacités. Ces essais sont listés dans le Tableau III. 9 suivant.

Tableau III. 9: Description des séquences tests synthétisées avec les espaceurs photosensibles PRISC et NISC.

Séquence	Séquence de monomères	*M[a] (Da)*	*Impuretés*
$P_{PRISC}1$	M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-PRISC-M_3·M_3·M_3·M_3·M_3·M_3·M_3·M_3-T	3074,6118	4
$P_{PRISC}2$	M_1·M_1·M_1·M_1-PRISC-M_1·M_1·M_1·M_1·A-PRISC- M_1·M_1·M_1·M_1-T	3011,5261	6
$P_{NISC}1$	M_1·M_1·M_1·M_1·M_1·M_1·M_1·M_1-NISC-M_3·M_3·M_3·M_3·M_3·M_3·M_3·M_3-T	3079,6423	5
$P_{NISC}2$	M_1·M_1·M_1·M_1-NISC-M_1·M_1·M_1·M_1·A-NISC- M_1·M_1·M_1·M_1-T	3021,5873	3
	Rmq : Pas détectée		

[a] *Masse monoisotopique.*

Trois séquences sur les quatre synthétisées ont pu être retrouvées lors de l'analyse par spectrométrie de masse. Les séquences contenant l'espaceur PRISC ont été retrouvées avec une abondance plus importante que celles effectuées avec l'espaceur NISC.

En effet, la séquence ne comprenant que deux blocs ($P_{NISC}1$) est bien retrouvée lors de l'analyse en spectrométrie de masse mais en très faible quantité, visible en vert sur un zoom de la Figure III. 22. A cause de cette trop faible abondance l'attribution n'a pas pu être confirmée par l'analyse MS/MS ni

pseudo-MS^3. Concernant la deuxième séquence test comprenant l'espaceur NISC, elle n'a pas été retrouvée lors du séquençage.

Les deux séquences synthétisées avec l'espaceur PRISC ont, quant à elle, été retrouvées. Leur attribution a pu être confirmée par une analyse MS/MS et une couverture totale est observée en pseudo-MS^3. Les analyses MS des séquences **$P_{PRISC}1$** et **$P_{NISC}1$** sont montrées en Figure III. 22.

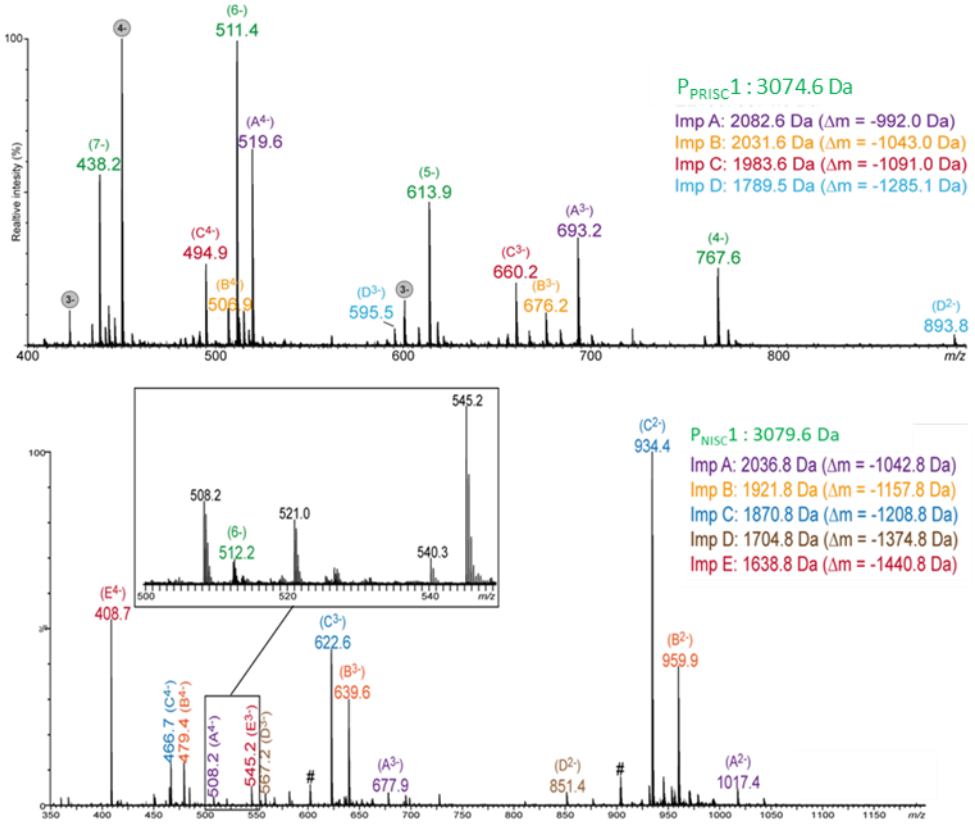


*Figure III. 22: (En haut) Spectre ESI-MS en mode négatif de **$P_{PRISC}1$** en vert. (En bas) Spectre ESI-MS en mode négatif de **$P_{NISC}1$** en vert. Les impuretés sont annotées en couleurs. Les pics annotés par un rond gris correspondent à des fragments formés en source. # Amas de sel.*

Plusieurs impuretés sont tout de même détectées dans tous les échantillons, comme on peut le voir les spectres MS des séquences $P_{PRISC}1$ et $P_{NISC}1$ donnés en Figure III. 22. Ces impuretés semblent être du même type dans tous les échantillons, quel que soit l'espaceur photosensible utilisé. Elles peuvent être interprétées comme des séquences tronquées ayant le second bloc intact. A la place du premier bloc, la partie à gauche de la liaison C-ON de l'espaceur est intacte et on trouve un synthon noté X de masse comprise entre 16 et 129 Da en fonction des séquences analysées. Par exemple pour la séquence $P_{PRISC}1$ les trois impuretés principales A, B et C peuvent s'expliquer ainsi :

- Impureté A : 278.1 Da = 149 + 129 (en violet)
- Impureté B : 227.1 Da = 149 + 78 (en orange),
- Impureté C : 179.1 Da = 149 + 30 (en rouge),

où 149 Da est la masse du segment PRISC (si intact) à gauche de la liaison alcoxyamine de l'espaceur PRISC. Les spectres MS/MS de ces trois impuretés sont donnés en Figure III. 23.

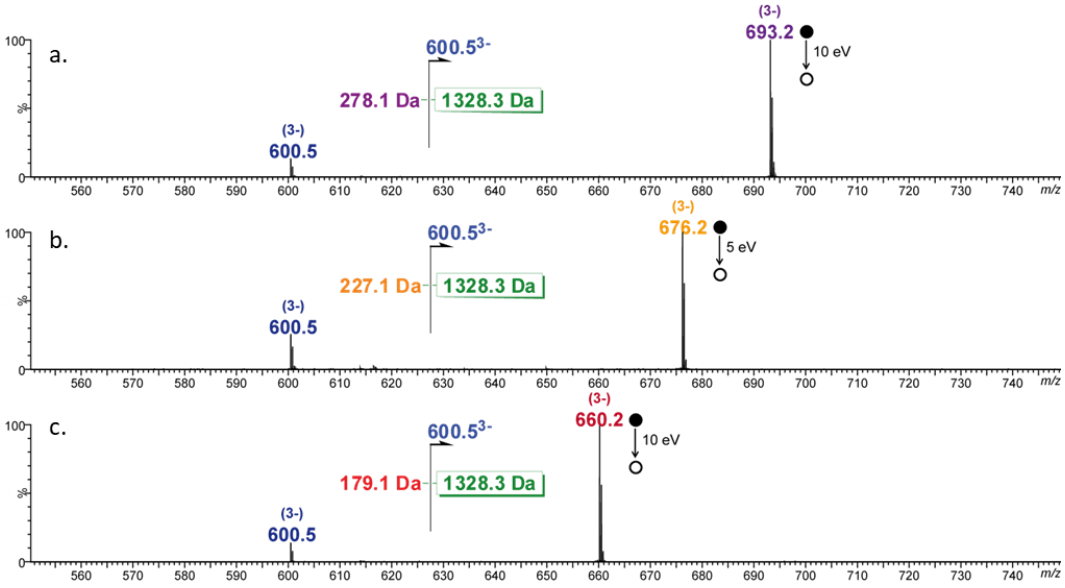


Figure III. 23: Séquençage de $P_{PRISC}1$ et des impuretés (a) Spectre MS/MS de l'ion $[M - 5H]^{5-}$ à m/z 613,9 et schéma d'homolyse de l'espaceur. (b) Spectre MS/MS de l'ion $[A - 3H]^{3-}$ à m/z *693,2 (c) Spectre MS/MS de l'ion $[B - 3H]^{3-}$ à* m/z *676,2 et (d) Spectre MS/MS de l'ion $[C - 3H]^{3-}$ à* m/z *660,2. Dans chaque schéma de dissociation, la valeur encadrée correspond à la masse du bloc codant tandis que la valeur non encadrée correspond au segment libéré.*

Lorsque les impuretés détectées ont un synthon X de 16 ou 30 Da (comme l'impureté C de la séquence $P_{PRISC}1$), elles peuvent s'expliquée par la même interprétation. Elles semblent issues de la substitution nucléophile par l'ammoniaque ou la méthylamine du carbone benzylique lors du clivage. Lorsque le synthon X est retrouvé à une masse de 16 Da la substitution s'est faite par l'ammoniaque, à 30 Da il s'agit de la substitution par la méthylamine. Les deux cas de figure sont montrés sur la Figure III. 24.

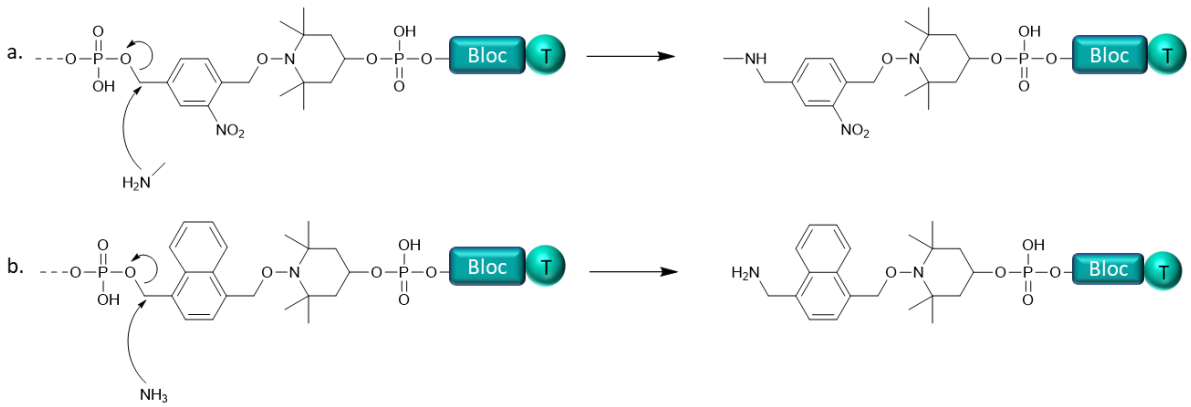


Figure III. 24: Schéma des substitutions nucléophile par la méthylamine et l'ammonique des séquences (a) $P_{PRISC}1$ (impureté C) avec un synthon X à 30 Da et (b) $P_{NISC}2$ avec un synthon à 16 Da (visible en partie expérimentale).

Ces quatre séquences tests nous montrent que la synthèse avec les espaceurs photosensibles PRISC et NISC semble compliquée. Les impuretés détectées montrent que des réactions secondaires semblent apparaitre lors du clivage comme pour l'espaceur RISC. Les séquences synthétisées comprennent bien l'espaceur mais le bloc suivant ne correspond pas à celui désiré. La synthèse de ce type d'espaceur a donc été abandonnée. Plus généralement, tous les espaceurs de type RISC avec la position benzylique très réactive, ne conviennent pas pour être incorporés dans une séquence de poly(phosphodiester)s numérique. Cependant, il semble possible de synthétiser des dérivés photosensibles à partir de l'espaceur RISCOP qui paraît remédier aux problèmes rencontrés avec les autres espaceurs.

5.2. Elaboration d'espaceur-marqueur

Une optimisation intéressante pour les espaceurs est de leur donner également la fonction de marqueur de bloc. En effet, en utilisant un espaceur-marqueur il ne serait plus nécessaire d'utiliser les nucléotides commerciaux comme marqueur de masse. L'ordre des différents blocs serait réparti en fonction du marqueur présent sur l'espaceur et non plus en fonction de monomères marqueurs ajoutés en plus dans la séquence. Cela permettrait d'avoir des séquences impliquant uniquement des monomères synthétiques.

Un premier essai est effectué en utilisant la structure de l'espaceur ROSC contenant un groupement phényl en position latérale de la chaîne principale. Le noyau aromatique de cet espaceur a été bromé en position *para*. Ce nouvel élément dans la molécule permet de lui conférer une masse spécifique qui pourra être reconnue spécifiquement lors d'une analyse en spectrométrie de masse. Cette nouvelle molécule est nommée l'espaceur-marqueur **EM-Br**, sa structure est donnée en Figure III. 25.

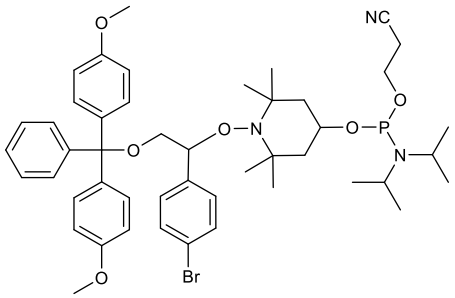


*Figure III. 25: Structure de l'espaceur-marqueur **EM-Br**.*

Une seule séquence a été effectuée avec l'espaceur-marqueur EM-Br. La séquence $\mathbf{P_{EM\text{-}Br}1}$ est composée de deux blocs comprenant huit monomères chacun (bloc #1 : 8 x M1 et bloc #2 : 8 x M3) séparés par l'espaceur EM-Br. Celle-ci nous permet de conclure que la synthèse de séquences comprenant cet élément est possible. En effet, comme pour son homologue espaceur ROSC, la synthèse s'est bien effectuée et la séquence a été recouvrée de façon monodisperse comme le montre la Figure III. 26.

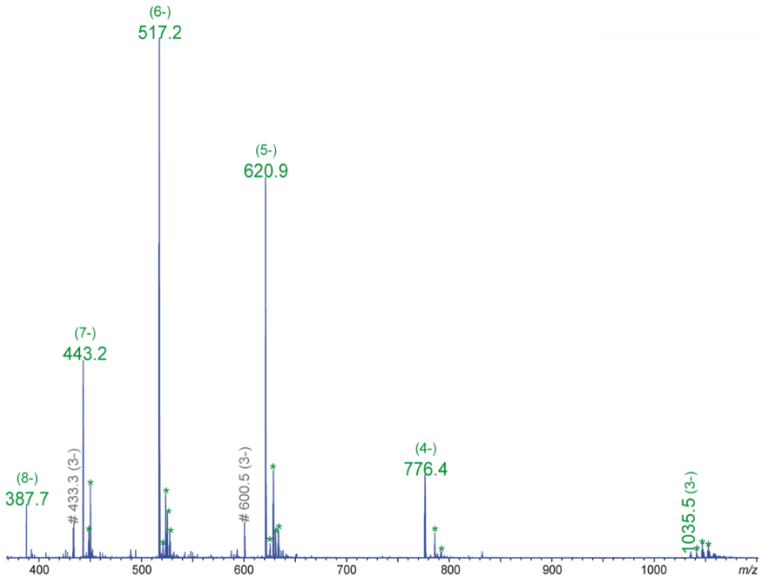


Figure III. 26: Spectre ESI-MS en mode négatif de ***$P_{EM\text{-}Br}1$****. Les pics annotés par un astérisque correspondent à différents échanges H/Na et H/K. Les fragments formés en source sont en gris.*

Néanmoins, aucune amélioration n'a été remarquée lors de la fragmentation en pseudo-MS^3. Etant donné que cette nouvelle molécule est basée sur la structure de l'espaceur ROSC, il est normal de n'avoir aucun progrès comme constaté lors des essais avec ce dernier.

Toutefois, cet espaceur-marqueur peut être utilisé dans les séquences de poly(phosphodiester)s comme l'élément le plus fragile de la séquence grâce à la présence de la liaison alcoxyamine. Mais également comme l'élément qui permet de marquer le bloc dans lequel il est incorporé, grâce à la présence de l'élément brome. Cette amélioration permet l'économie d'un couplage et permet également d'éviter de possibles réactions secondaires qui peuvent se produire avec les bases nucléiques lors de la création du radical alkyle réactif. Pour l'instant il n'y a que cet espaceur-marqueur qui a pu être synthétisé et incorporé dans une séquence de poly(phosphodiester)s. Il est possible d'imaginer d'autres espaceurs-marqueurs qui contiendraient d'autres atomes comme l'iode ou le fluor sur les espaceurs précédemment présentés.

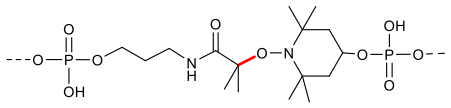
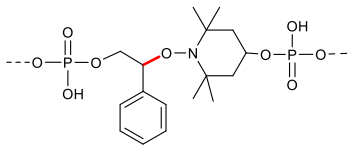
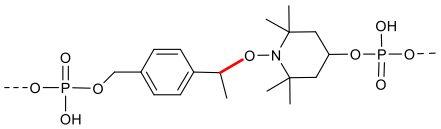
6. Conclusion et perspectives

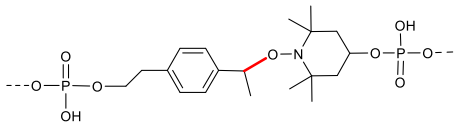
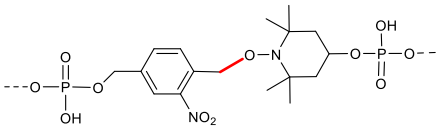
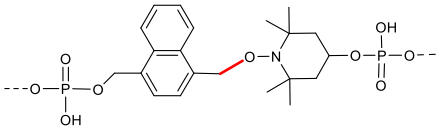
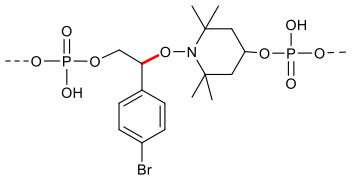
Il a été montré dans ce **chapitre III** que l'élément espaceur utilisé dans le design de séquence de poly(phosphodiester)s présenté dans le **chapitre II** peut être optimisé. Cette optimisation est nécessaire pour obtenir une meilleure analyse de spectrométrie de masse en pseudo-MS^3. En effet, lors du clivage de la liaison alcoxyamine contenue dans l'espaceur, un radical alkyle très réactif est créé et produit de nombreuses réactions secondaires qui parasitent le spectre obtenu en rendant beaucoup plus difficile sa lecture. Ceci empêche l'automatisation de cette étape. En effet, l'élimination des réactions secondaires observées en pseudo-MS^3 peut permettre d'automatiser tout le séquençage des séquences de poly(phosphodiester)s numériques. L'information encodée serait alors recouvrée en quelques secondes.

Plusieurs pistes ont été étudiées pour perfectionner cette molécule. La structure globale de la molécule reste la même : il est nécessaire d'avoir le groupement protecteur DMT et le groupement phosphoramidite pour être utilisé avec la chimie de la phosphoramidite. Une liaison alcoxyamine doit être présente pour être l'élément le plus fragile de la séquence et permettre une fragmentation contrôlée lors de l'analyse en spectrométrie de masse en tandem. Mais il est possible de jouer sur le reste du squelette de la molécule. Le point de départ est de rendre moins accessible le radical créé lors de la première fragmentation en analyse MS/MS.

Plusieurs espaceurs avec des structures différentes mais contenant tous cette liaison alcoxyamine faible, ont été incorporés dans la synthèse de poly(phosphodiester)s. Le Tableau III. 10 récapitule les structures des molécules étudiées au cours de ce **chapitre III**, lorsqu'elles sont comprises dans une chaîne de poly(phosphodiester)s.

Tableau III. 10: Tableau récapitulatif des structures des espaceurs étudiés dans ce chapitre lorsqu'ils sont incorporés dans une chaîne de poly(phosphodiester)s.

Nom de l'espaceur	Structure	Commentaire
Espaceur classique **E1**		*Après la fragmentation, obtention d'un radical sur le carbone central très réactif qui provoque des réactions parasites.*
Espaceur **ROSC**		*Après la fragmentation, aucune amélioration n'est observée.*
Espaceur **RISC**		*Réactions secondaires lors de la synthèse, mais de nettes améliorations sont apportés lors de l'analyse en pseudo-MS^3.*

Espaceur **RISCOP**		*Améliorations apportées : plus aucune réaction parasite n'est détectée.*
Espaceurs Photosensibles **PRISC**		*Réactions secondaires lors de la synthèse.*
NISC		
Espaceur-marqueur **EM-Br**		*Permet de jouer le rôle d'espaceur et de marqueur de masse.*

Dans un premier temps, un groupement phényle est ajouté en position latérale (**ROSC**), mais celui-ci ne permet aucune amélioration lors de l'analyse en pseudo-MS^3, malgré une bonne efficacité de couplage lors de la synthèse. Dans un deuxième temps, le groupement phényle est ajouté directement dans la chaîne principale de l'espaceur (**RISC**). Cet espaceur donne des résultats significativement meilleurs lors de l'analyse en pseudo-MS^3, mais des problèmes lors de la synthèse des séquences sont à déplorer. Malgré différents tests d'amélioration des conditions de purification, des réactions secondaires se produisent à chaque fois. Un autre design se basant également sur l'ajout d'un groupement phényle dans la chaîne principale est développé. Cette fois-ci, la fonction benzylique très réactive est supprimée en allongeant d'un carbone la chaîne principale entre le phosphate et le cycle (**RISCOP**). Cette optimisation semble être la bonne solution car les spectres de pseudo-MS^3 obtenus sont nettement améliorés par rapport aux précédents. De plus, aucune réaction secondaire n'apparait lors de la synthèse des séquences.

Ce design semble donc être le bon pour permettre d'améliorer les conditions de séquençage en pseudo-MS^3. Grâce à ce nouvel espaceur il est possible de coder de l'information sur des séquences de poly(phosphodiester)s et de retrouver les données grâce à une analyse automatisée faite avec le logiciel MS-DECODER. En combinant ce nouvel espaceur et les alphabets augmentés présentés dans le **chapitre II**, il est donc possible de coder des séquences à haute capacité de stockage et de les décoder en quelques secondes de façon automatisée.[175] Il sera également possible d'atteindre des capacités de stockage encore plus haute avec toujours une lecture automatisée en combinant l'utilisation de l'espaceur RISCOP

avec des outils de compressions d'information. Cette nouvelle voie de recherche est explorée dans le **chapitre IV**.

D'autres améliorations sont également possibles et jouent sur les propriétés de l'espaceur. Une première voie est de rendre l'espaceur photosensible pour permettre une première fragmentation via un stimulus externe de la lumière et non plus une fragmentation par ionisation. Deux espaceurs ont été testés (**PRISC** et **NISC**) mais leurs designs se basent sur la structure de l'espaceur RISC qui montrent des défaillances lors de la synthèse. Par conséquent, les mêmes problèmes de synthèses sont remarqués. Le même design de molécules photosensibles pourrait être envisagée avec pour base l'espaceur RISCOP.

Une seconde voie de recherche est l'élaboration d'espaceur-marqueur (**EM-Br**). Ce type de molécule permettrait d'économiser des étapes de couplage en éliminant les nucléotides utilisés pour ordonner les différents blocs. Un espaceur-marqueur a pu être synthétisé et montre une bonne efficacité de couplage lors de la synthèse de poly(phosphodiester)s mais aucune amélioration vis-à-vis des analyses de spectrométrie de masse n'est observée. Néanmoins, ce design permet de s'affranchir des nucléotides commerciaux et permet d'obtenir une séquence totalement synthétique déchiffrable. Ce type de molécule espaceur-marqueur pourrait être testé sur d'autres espaceurs, notamment sur l'espaceur RISCOP qui permet d'obtenir de meilleurs résultats lors de l'analyse en pseudo-MS^3.

En résumé, une amélioration de l'espaceur alcoxyamine classique a été possible. Le nouveau design de la molécule **RISCOP** donne des résultats très intéressants en permettant notamment de synthétiser des séquences stockant des images. De surcroit, toutes ces séquences peuvent maintenant être analysées de façon automatique en utilisant le logiciel MS-DECODER. Ce nouvel espaceur peut donc être utilisé comme espaceur principal pour la synthèse de prochaines séquences.

Des séquences à plus haute capacité de stockage peuvent ainsi être envisagées avec l'utilisation de cet espaceur. Mais pour synthétiser de grandes séquences de poly(phosphodiester)s contenant des éléments espaceurs clivables, il faut également développer les marqueurs de masse. Sans ces éléments, le décodage de telles séquences ne peut s'effectuer correctement. Une étude a donc été menée (**chapitre IV**) pour développer ces éléments et ainsi pouvoir synthétiser des séquences à très haute capacité de stockage.

Chapitre IV

Synthèse de polymères à très haute capacité de stockage

1. Introduction

Dans les **chapitres II** et **III** il a été montré que les séquences de poly(phosphodiester)s codant de l'information peuvent être améliorées grâce à l'extension de l'alphabet moléculaire et à l'optimisation de l'espaceur alcoxyamine facilitant la lecture. Grâce à ces améliorations il peut maintenant être envisager de créer des chaînes à haute capacité de stockage. Pour atteindre de plus hautes densités de stockage il faut synthétiser des séquences plus longues. Pour l'ADN synthétique, la limite se trouve entre 115 et 200 nucléotides ajoutées et synthétisées sur une chaîne unique.[13] Il semble possible d'atteindre une longueur de chaîne semblable avec des séquences de poly(phosphodiester)s synthétiques. Toutefois des outils aidant à la lecture doivent être utilisés pour pouvoir recouvrer l'information encodée.

Ces outils sont les marqueurs de masse qui sont utilisés pour ordonner les séquences lors de l'analyse de spectrométrie de masse pendant la phase de lecture. Lors des premiers essais de lecture en analyse MS des séquences de poly(phosphodiester)s, les spectres obtenus sont difficiles à interpréter. La mise en place de la fragmentation inter-segments permet une analyse plus facile menant jusqu'au décryptage des données encodées. Pour faciliter encore la lecture, la longueur optimale des segments est de 8 monomères. Pour synthétiser de longues séquences contenant plusieurs blocs de 8 monomères, il faut augmenter l'ajout d'espaceurs et de marqueurs moléculaires entre chacun de ces blocs. Le choix de bons marqueurs de masse est donc primordial. En effet, s'ils ne répondent pas à certains critères précis, la réorganisation des segments post analyse MS/MS n'est pas possible. Il est alors impossible de récupérer l'information initialement encodée.

Dans ce **chapitre IV**, nous allons voir dans un premier temps comment le choix de nouveaux marqueurs de masse s'effectue. Bien entendu, ce choix repose également sur le choix de l'alphabet utilisé pour encoder l'information. Ici, l'alphabet à 8 symboles codant des triades sur les monomères M1 à M8 est utilisé (alphabet exposé sur la Figure II. 16 du chapitre II). Le choix des nouveaux marqueurs s'est porté sur des marqueurs commerciaux dérivés des bases nucléiques déjà utilisées auparavant.

L'utilisation d'une plus grande quantité de marqueurs permet d'atteindre une densité de stockage encore plus importante. Pour améliorer davantage la capacité de stockage de ces longues chaînes de polymères, il est aussi possible d'utiliser des éléments informatiques permettant de comprimer l'information encodée. Ceci pourrait permettre la synthèse de longues chaînes de poly(phosphodiester)s contenant une très haute densité d'information. Le schéma du concept est exposé en Figure IV. 1. Il montre qu'une image peut être facilement pixellisée et transformée en flux de bits qui pourra, à son tour, être compressé et traduit en la suite de monomères qui formeront une séquence de poly(phosphodiester)s numériques à très haute capacité de stockage. Cette séquence sera alors formée de plusieurs blocs contenant de l'information compressée (en violet sur la Figure IV. 1), avec un marqueur de masse en bout de bloc (représenté en vert). Plus il y aura de blocs dans la séquence, plus le nombre de marqueur nécessaire sera augmenté.

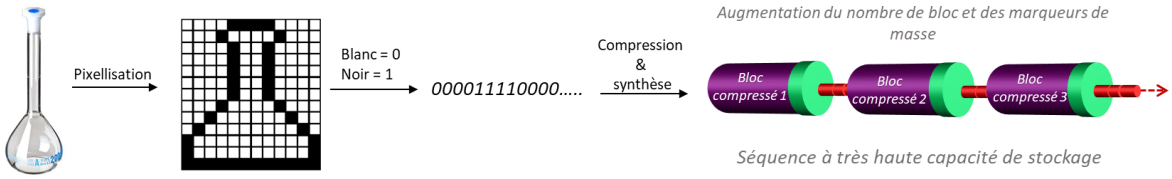


Figure IV. 1: Schéma du concept permettant la synthèse de séquence codant des données comprimées.

Pour donner plus de robustesse à la séquence, des éléments de correction d'erreurs sont également ajoutés. La création de séquences comprimées a été effectuée avec la collaboration de Marc-André Delsuc, directeur de recherche au CNRS rattaché à l'institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) de l'Université de Strasbourg.

La synthèse de longues chaînes de poly(phosphodiester)s codant de l'information comprimée est détaillée dans la suite du **chapitre IV**. Il est décidé de coder à nouveau une image. Cette fois-ci, un portrait est encodé et comprimé en flux de bits puis celui-ci est synthétisé en utilisant la chimie de la phosphoramidite automatisée. L'échantillon récupéré est alors analysé en spectrométrie de masse. Les spectres qui ne sont pas présentés dans ce chapitre, sont disponibles dans la **partie expérimentale**. La méthode de compression des images encodées est détaillée, ainsi que la méthode permettant la correction d'erreurs.

2. Choix des marqueurs de masse

Les marqueurs de masse utilisés dans les séquences de poly(phosphodiester)s doivent répondre à certains critères.[25] Tout d'abord ils doivent pouvoir être utilisés en tant que monomères pour la chimie de phosphoramidite automatisée, ils doivent donc contenir une fonction phosphoramidite et un groupement protecteur DMT. Ces deux fonctions permettent leur utilisation dans le cycle itératif. Ensuite, le critère le plus important est leur masse molaire. Celle-ci va permettre la différenciation des différents segments à la suite de la fragmentation lors de l'analyse de spectrométrie de masse en tandem. Leur masse molaire doit répondre à des caractéristiques bien précises qui sont mentionnées dans le **chapitre I** lors de l'explication de la mise en place du nouveau design des séquences comprenant ces éléments. Les critères auxquels doivent répondre les marqueurs de masse sont basés sur les monomères impliqués dans la séquence et leur masse molaire. Pour ces nouvelles séquences comprenant les huit monomères M1 à M8 les critères sont plus contraignants que lorsqu'il n'y a que deux monomères utilisés. Les trois paramètres à respectés sont :

1) La masse molaire d'un marqueur ne doit pas être un multiple de 14.
2) La différence de masse molaire entre deux marqueurs ne doit pas être un multiple de 14.
3) La différence de masse molaire entre deux marqueurs doit être supérieur à 3.

Les deux premiers critères à respecter impliquent le chiffre 14 et ses multiples. Cette valeur vient de la différence de masse molaire entre les monomères choisis. En effet avec l'alphabet à 8 symboles, la plus petite différence de masse entre deux monomères est de 14 daltons. Il faut donc que la masse molaire d'un marqueur ne soit pas un multiple de 14 pour permettre de le reconnaitre sans aucun doute. De plus, il faut également que la différence de masse entre deux marqueurs réponde à ce critère : elle ne devra pas être un multiple de 14 pour pouvoir reconnaitre sans ambiguïté un marqueur de masse d'un monomère. Concernant le dernier critère, il s'explique par le fait que des espèces

triplements chargées sont analysées en pseudo-MS^3, il faut donc une différence de masse plus grande que 3 Da pour être sûr de bien différencier 2 segments qui seront triplement chargés.

En respectant ces trois critères, une liste de marqueurs de masse potentiellement utilisables en chimie de la phosphoramidite automatisée a été établie. Les marqueurs composés des bases nucléiques, utilisés usuellement : dA, dC, dG et dT sont gardés ainsi que leur dérivés dI, dB et dF, qui sont renommés dI_U, dB_U et dF_G en fonction de la molécule dont est issus ce dérivé et de la fonction ajoutée. Par exemple la molécule dF_G est issue de la base nucléique dG avec un fluor en plus. D'autres marqueurs dérivés des bases nucléiques ont alors été recherchés. De nombreux dérivés commerciaux existent, néanmoins il faut qu'ils répondent aux critères mentionnés ci-dessus. Plus d'une trentaine de molécules commerciales ont été examinées. La liste suivante donne les onze molécules retenues :

Tableau IV. 1: Liste des molécules choisies convenant aux critères pour être des marqueurs de masse.

Nom	M [a] (g/mol)	m [b] (g/mol)	Multiple de 14 ?[c]	Δ [d]	Δ multiple de 14 ?[c]
dC	833,93	289,18	20,66		
dP	738,82	298	21,29	8,82	0,63
dT	744,83	304,1	21,72	6,1	0,44
dF_U	748,79	307,16	21,94	4,06	0,22
dA	857,95	313,21	22,37	6,05	0,43
dG	839,92	329,21	23,52	10	0,71
dE	777,86	338,23	24,16	9,02	0,64
dF_G	857,91	347,19	24,80	8,96	0,64
dBr_U	809,69	369,07	26,36	21,88	1,56
dBr_G	903,9	408,1	29,15	5,74	0,41
dI_U	856,69	416,07	29,72	7,97	0,57

[a]Masse molaire des molécules avec les groupements protecteur. [b]Masse molaire des molécules lorsqu'elles sont incorporées dans la chaîne de poly(phosphodiester)s, masse utilisée pour les calculs de différences. [c]Permet de vérifier qu'une masse molaire d'une molécule ou de la différence n'est pas divisible par 14. [d]Différence des masses molaires de deux molécules consécutives.

Pour chaque molécule, sa masse molaire avec les groupements protecteurs est donnée mais celle-ci ne peut pas être utilisée pour vérifier si les molécules conviennent aux critères d'éligibilité. En effet, il faut comparer les différentes molécules en fonction de leur masse molaire sans groupements protecteurs. Il faut donc utiliser leur structure lorsqu'elles sont incorporées dans une chaîne de poly(phosphodiester)s (notée « m » dans le Tableau IV. 1). La masse molaire de la molécule incorporée dans la séquence permet alors de vérifier si cette dernière convient. On vérifie ainsi qu'elle n'est pas un multiple de 14, en divisant sa valeur par 14 (4ème colonne du tableau), que la différence entre deux molécules notée Δ, n'est également pas divisible par 14 (dernière colonne du tableau) et que cette différence est bien supérieure à 3. Les structures de tous les marqueurs retenus sont données en Figure IV. 2.

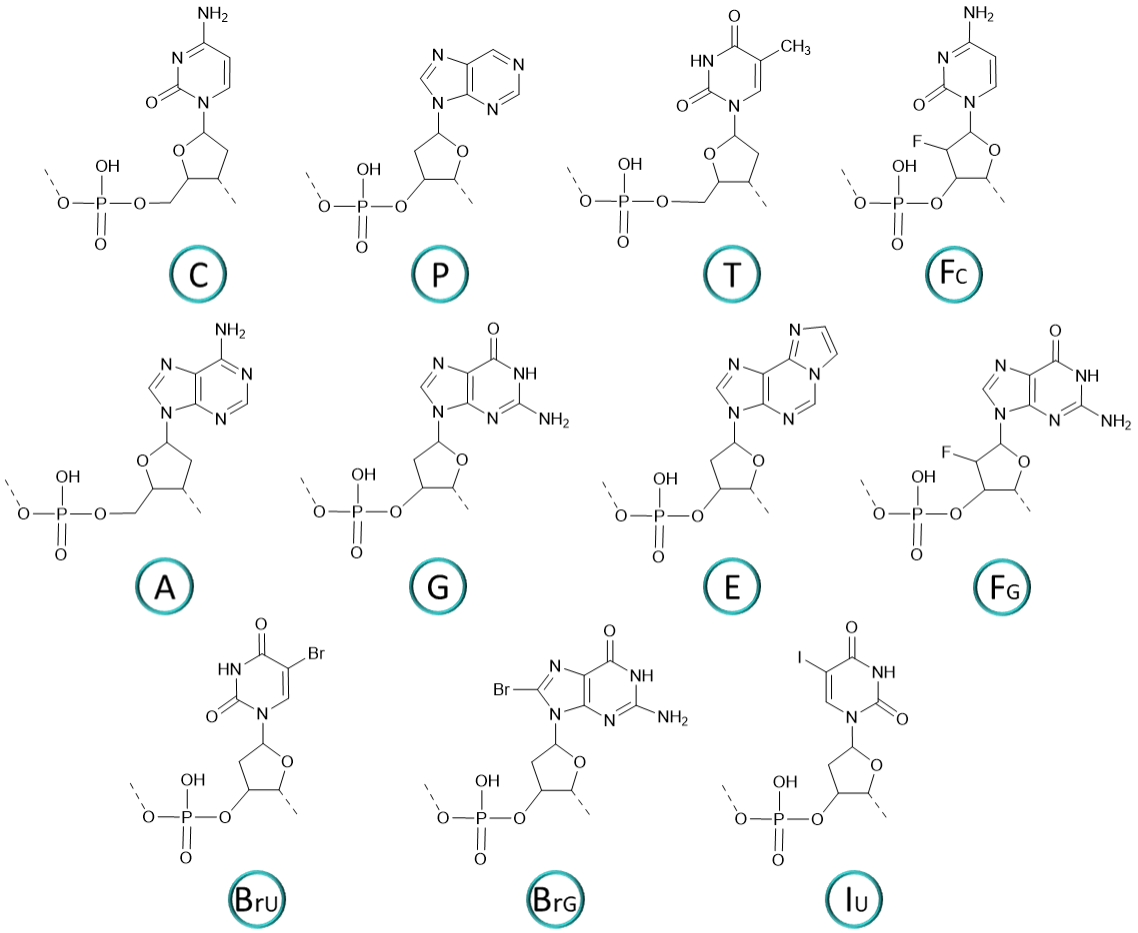


Figure IV. 2: Structure des molécules choisies comme marqueurs de masse.

Bien entendu d'autres structures existent mais ne répondent pas toujours aux critères demandés. Par exemple, c'est le cas pour la molécule dF_U qui est un dérivé de la base nucléique dU avec un fluor en plus. En effet, dF_U a la même masse que la molécule dF_C lorsqu'elle est incorporée dans une chaîne de poly(phosphodiester)s. Leurs structures sont montrées en Figure IV. 3.

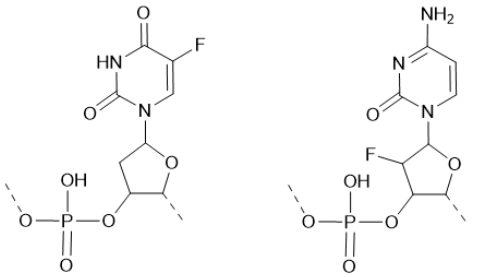


Figure IV. 3: Structures des molécules dF_U et dF_C.

Les deux molécules conviennent mais elles ne peuvent pas être utilisées en même temps car leur masse molaire est la même. Il est donc impossible de les différencier lors d'une analyse en spectrométrie de masse. L'ordre originel des différents blocs libérés ne pourrait pas être retrouvé.

Avec cette nouvelle liste de marqueurs de masse possibles contenant 11 molécules différentes il est possible de concevoir des séquences ayant jusqu'à 12 blocs différents, le premier bloc ne comprenant pas de marqueur de masse. Quatre nouveaux marqueurs sont donc ajoutés à la liste déjà existante, permettant ainsi d'augmenter la taille des séquences possibles. Un design comprenant des blocs de huit monomères est à nouveau choisi. Cela permet d'atteindre au minimum une longueur de chaîne de 12 blocs de 8 monomères chacun donnant une séquence de 96 monomères codants. L'alphabet à 8 symboles est utilisé donc chaque monomère code pour 3 bits, ce qui permet d'atteindre une capacité de stockage de 288 bits. Cette capacité de stockage est déjà le double de celle obtenue avec la chaîne la plus longue synthétisée dans le **chapitre II** (P14 avec 144 bits). Mais, en plus de cela, l'information encodée peut également être compressée pour permettre d'augmenter encore plus la capacité de stockage d'une chaîne unique de polymère. Différents outils peuvent être utilisés pour réduire le message encodé tout en gardant une robustesse qui permettra de ne perdre aucune information. Ces outils sont présentés dans le paragraphe suivant.

3. Utilisation d'outils de compression et de codes de correction pour augmenter la capacité de stockage

Plusieurs outils sont utilisés pour augmenter de la capacité de stockage d'une chaîne de polymère. Il est possible de les répartir en deux catégories : (i) la compression de l'information et (ii) le contrôle des erreurs. La première catégorie permet de réduire la taille du message à encoder. La deuxième permet de ne perdre aucune information tout en contrôlant qu'aucune chaîne erronée est transmise comme étant le bon message. Ces deux éléments changent totalement le flux de bits encodés mais l'information contenue est la même qu'initialement. Néanmoins, les messages ne sont pas autonomes, il faut mettre en place des conventions pour les comprendre.

Il y a 2 types de conventions : (i) une externe qui détermine les règles globales que l'encrypteur et le décodeur connaissent et (ii) une convention interne, qui est directement implémenté dans le polymère codé.

Il faut trouver un juste milieu entre ces deux conventions pour que les informations données dans la convention interne ne soient pas trop importantes, ce qui prendrait beaucoup de place. Mais il faut également ne pas donner toutes les informations dans la convention externe, car le codage perdrait de son intérêt.

La convention interne, se traduit par l'utilisation d'un entête dans le codage du polymère. Dans cet en tête on retrouve les informations non contenues dans la convention externe. Par exemple, pour le codage d'une image, la convention interne peut comprendre un entête avec :

- Le type d'information encodé : une image dans ce cas (mais peut aussi dire que c'est un texte, une vidéo, un QR code …).
- La forme de code utilisé, c'est-à-dire si une compression est utilisée il faudra dire laquelle.

Cet entête ajouté pour coder la convention interne directement dans la chaîne de polymère ajoute des bits non codants mais nécessaires pour la compréhension du code. En effet, les informations de l'entête sont transformées en une suite de bits qui correspondent à un message que le lecteur comprend car leur décryptage fait partie des conventions externes. On peut imaginer que le premier

monomère de la séquence code pour le type d'information encodé, ainsi avec l'utilisation de l'alphabet à 8 symboles, il y a 8 types d'information qu'il est possible d'encoder, par exemple on peut avoir :

> Monomère M1 code pour les bits 000, correspond à une image.
> Monomère M2 code pour les bits 001, correspond à un texte.
> Monomère M3 code pour les bits 010, correspond à une vidéo… Et ainsi de suite.

Le deuxième monomère de la séquence donne la forme du code utilisé, ainsi comme pour le premier monomère, il y a 8 moyens d'exprimer la façon de coder choisie :

> Monomère M1 code pour les bits 000, correspond à un code sans outils de compression.
> Monomère M2 code pour les bits 001, correspond à un code avec outils de compression « Huffmann ».
> Monomère M3 code pour les bits 010, correspond à un code avec outils de compression « Arithmetic coding » … Et ainsi de suite.

La convention externe peut comprendre :

- La taille de l'image (22*20 par exemple)
- L'alphabet utilisé (à 2, 4 ou 8 symboles)
- L'ordre des marqueurs de masse qui permet de recouvrir l'information
- Les outils permettant de décrypter l'entête du polymère
- Les outils pour décompresser le flux de bits obtenu et obtenir le message initialement encodé

Grâce à l'ajout de ces éléments il serait alors possible d'obtenir de longues chaînes de polymères codant une quantité importante d'information. Il est important de remarquer que ces conventions existent pour n'importe quel type d'information à transmettre. Même pour de très simples informations comme un message écrit à la main, la convention externe est que le lecteur est capable de comprendre l'alphabet avec lequel le message est écrit. Si cette convention n'est pas remplie, le message ne pourra pas être transmis correctement. Une exception connue est le code génétique, qui fonctionne très bien seul sans avoir besoin de conventions externes produites par les êtres humains pour être compris par les systèmes biologiques. Mais toutes les transmissions d'information effectuées par les êtres humains ont besoin de conventions pour être comprises sans erreurs.

3.1. Compression de l'information

La compression de données est une opération informatique permettant de réduire la taille d'un message formée par des bits. Ainsi une suite de bits α est transformée en une suite de bits β plus courte mais qui va pouvoir tout de même restituer les mêmes informations ou des informations voisines. Cette restitution est possible grâce à l'utilisation d'un algorithme de décompression, qui utilise les méthodes inverses de celles utilisées pour compresser le message. Il existe deux types de méthodes de compression : (i) algorithme de compression avec perte et (ii) sans pertes. L'algorithme sans perte restitue le message initial dans sa totalité sans aucune perte, celui avec perte restitue un message plus ou moins voisin de l'original.

3.1.1. Algorithmes avec perte dits « Lossy »

Dans ce genre de méthode avec perte, l'information originale encodée n'est pas retrouvée dans sa totalité. Ce genre d'algorithme s'appuie sur les connaissances extérieures des décrypteurs. Par

exemple, il peut être utilisé pour encoder un texte mais en enlevant toutes les voyelles, cela permet de gagner énormément d'espace de stockage et avec un peu de travail annexe le texte est tout de même lisible. Néanmoins, le décodeur doit avoir des capacités plus développées pour recouvrer le code que s'il est codé sans perte. En effet, si un livre écrit en français est encodé et comprimé en enlevant les voyelles, il n'y aura qu'un décodeur francophone qui pourra réussir à décoder le message.

Généralement cette technique avec perte est employée pour compresser des images ou des sons. Par exemple tous les fichiers JPEG sont des compressions avec perte. Pour des fichiers de ce type, l'idée est qu'un simple sous-ensemble du fichier permet au lecteur de percevoir la même information. Pour des images, l'œil ne perçoit pas forcément tous les détails il sera alors possible de réduire la quantité de données transmises sans que l'œil humain ne s'aperçoive de la différence.

3.1.2. Algorithmes sans perte dits « Lossless »

Les algorithmes sans pertes, comme leur nom l'indique, transmettent les données initialement encodées sans aucune perte. Le message peut être retrouvé dans sa totalité, l'information est juste réécrite de façon plus concise, autrement dit elle est compactée.

Malheureusement il n'existe pas d'algorithme général permettant de comprimer les données sans perte. Une méthode compressant efficacement un type de fichier peut également amplifier un autre type de fichier. Ces méthodes de compression dépendent du fichier à comprimer et leur efficacité se base sur l'entropie de l'information à comprimer. Cette entropie reflète la quantité d'information contenue dans le message pouvant mener à un taux de compression maximal.

L'entropie d'une image en noir et blanc est minimale si l'image est toute blanche ou toute noire ou si les pixels sont totalement mis au hasard et qu'il n'y a pas de répétitions. L'entropie est maximale si tous les pixels sont équiprobables, il y a autant de pixels blancs que de pixels noirs donc autant de 0 que de 1. La compression dépend de la façon dont est codée l'information ainsi que l'information elle-même.

Pour concevoir un stockage de donnée comprimée sur une chaîne de polymère, un algorithme sans perte doit être utilisé car aucune perte d'information n'est possible pour recouvrir le message dans sa totalité. En effet, la perte d'éléments codants de l'information changerait complétement le message. Les données retrouvées ne correspondraient alors pas à celles initialement transmises. Concernant le choix de la méthode sans perte à faire, il n'y a pas de règles générales facilitant le choix. Le plus simple est de tester différentes méthodes pour utiliser celle qui finalement donne le meilleur résultat de compression. Chaque méthode suppose une certaine régularité dans le signal, son utilisation optimale dépend donc de l'adéquation signal/méthode. Dans le cas où le signal est parfaitement aléatoire, il est alors incompressible quelle que soit la méthode. Néanmoins, il est rare d'avoir un message complétement irrégulier. Dans un texte en français il y a toujours plus de a, e et s que de k, w et y (c'est pour cela que ces dernières comptent pour plus de points au scrabble). Ainsi, en codant les lettres courantes sur peu de bits et les lettres plus rares sur un nombre de bit plus important, le message est compressé. Ceci est le principe de codes dit « entropiques » qui sont utilisés pour les méthodes Huffman ou Arithmetic. Ils sont expliqués plus en détails dans les paragraphes suivant.

3.1.2.1. Méthode Huffman

La méthode de compression Huffman est l'une des plus courante et peut se rapprocher de la méthode utilisée pour le code Morse.[176] Dans ce type de méthode, les symboles de la source sont représentés

par un code à longueur variable. Ainsi on code les symboles qui sont les plus fréquents par des codons courts et on code les symboles sources qui sont plus rares sur des codons plus longs. C'est pour cela qu'en morse le symbole source « e » qui est très fréquent, est codé par simplement un point, le plus court de tous les codons utilisé dans ce langage.

Le nouveau code est donc déterminé à partir d'une estimation de probabilité d'apparition des symboles de source. Dans un premier temps, il faut donc compter combien de fois chaque code source apparait. Prenons un exemple pour faciliter l'explication : l'image en 13 * 12 d'une fiole, donnée en Figure IV. 4. Cette dernière est composée de pixels noirs et blancs, qui sont traduits en bits avec la règle noir = 1 et blanc = 0. Le flux de bits obtenu peut alors être synthétisé avec l'alphabet à 8 symboles puis compressée avec la méthode Huffman.

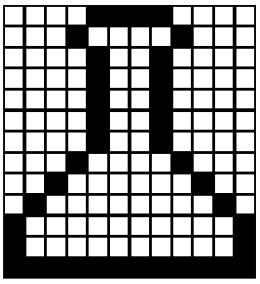


*Figure IV. 4: Image pixélisée d'une fiole en 13*12 utilisée comme exemple pour expliquer la compression Huffman.*

Pour utiliser cette méthode, il faut dans un premier temps compter le nombre de fois où apparaissent les symboles sources. Ces symboles sources correspondent aux triades codées par les monomères employés. Il s'agit donc des symboles sources 000, 001, 010, 011, 100, 101, 110 et 111. En fonction du nombre de fois qu'ils apparaissent ils ont un code comprimé plus ou moins grand. Le Tableau IV. 2 recense les nombres d'apparition de chaque symbole source ainsi que le code comprimé qui leur est attribué avec la compression Huffman.

Tableau IV. 2: Table de compression Huffman pour l'image de la fiole.

Symbole source	Nombre d'apparition	Code comprimé
000	24	0
001	5	1110
010	12	10
011	1	111100
100	5	1101
101	0	1111110
110	1	111101
111	4	11101

Dans cet exemple, on voit que le symbole source '000' revient le plus souvent (24 fois) il a donc le nouveau code comprimé le plus petit de la série. A contrario, le code source '110' qui n'apparait qu'une seule fois a comme nouveau code le '111101'. Finalement, la fiole compressée fait maintenant 120 bits au lieu de 156 bits au préalable.

Le défaut de ce type de codage est que l'algorithme doit dans un premier temps lire tout le fichier avant de pouvoir le comprimer. En effet, pour connaitre la fréquence de chaque symbole source, ce dernier doit connaitre chaque caractère utilisé dans le fichier et leur fréquence pour ainsi leur conférer un nouveau code optimal. De plus, pour pouvoir décompresser ce type de données il faudra que le lecteur connaisse la table qui a servie à compresser les informations. C'est-à-dire connaître l'attribution de chaque symbole source à son nouveau code. Cette table peut être ajoutée devant le fichier ce qui diminue la compression, ou elle peut être donnée dans les conventions externe pour chaque séquence. A cause de ces défauts ce type de codage n'est pas universel et le codage en grande dimension n'est pas possible et reste purement théorique.

3.1.2.2. Méthode universelle

Codage Elias

Pour rémédier à ces difficultés, d'autres systèmes de codage sans pertes sont développés. Ces nouveaux systèmes permettent de coder n'importe quel message source sans connaitre les statistiques de chaque symbole source formant le message. Ils sont donc dits « codages universels » car tout message pourrait être codé efficacement avec ces nouvelles méthodes. De nos jours, le codage arithmétique est l'un des plus répandu et suit ces nouvelles règles d'encodage universel. Il est une extension du codage d'Elias qui a été développé dans les années 1960.[177] Néanmoins, le code décrit de cette façon est impraticable car il demande une précision trop importante. Des modifications ont été effectuées pour améliorer ces inconvénients. En 1976, les travaux de J. Rissanen et R. Pasco montrent les premiers schémas de codage arithmétiques pratiques.[178, 179] La méthode de codage arithmétique est développée dans le paragraphe suivant.

Méthode de codage arithmétique

Le codage arithmétique est également basé sur l'entropie du message à compresser, tout comme le codage de Huffman. Ainsi, il associe aux symboles sources les plus courant les codes comprimés les plus petits. C'est donc un code à longueur variable, les symboles sources ayant tous la même taille ne sont pas codés par des codons de même taille en fonction de leur fréquence dans le message.

L'avantage majeur par rapport au codage de Huffman, est que le codage arithmétique peut produire des codes vides. Grâce à ces codes vides, la compression est meilleure car pour le codage de Huffman le plus petit code était de au moins 1 bit. Grâce au codage arithmétique cette lacune peut être comblée. Ainsi, l'un des plus grands avantages du codage arithmétique est que les symboles sources peuvent être codé sur un nombre non-entier de bits. Avec ce codage, l'algorithme ne code pas chaque symbole par un caractère spécifique mais par une chaîne de caractères plus ou moins longue. De plus, c'est un codage adaptatif : il ne suppose pas que la distribution de probabilité est connue dès le départ, mais qu'il est possible de l'estimer au fur et à mesure.

Grâce au codage arithmétique il est donc possible de compresser n'importe quel type d'information. Le message ainsi compressé permet un gain de place considérable et donc dans notre cas permet de coder la même information sur une chaîne de polymère plus courte ou alors de coder encore plus d'information sur la même longueur de chaîne et ainsi augmenter abondamment la densité de stockage. Néanmoins, lorsque ce type d'outils est utilisé aucune erreur de synthèse n'est permise. En effet, si la séquence est incorrecte, les données qu'elle contient sont illisibles car il ne sera pas possible de décompresser le bon message. Pour empêcher la transmission de séquences erronées en phase de lecture, des outils existent et apportent de la robustesse à la chaîne de polymère synthétisée. Il s'agit

d'ajouter des éléments permettant le contrôle des erreurs. Leur emploi est détaillé dans le paragraphe suivant.

3.2. Contrôle des erreurs

L'utilisation d'outils de compression peut rendre le message plus fragile, il est donc nécessaire de rendre les messages suffisamment robustes pour permettre de retrouver les bonnes informations si des erreurs de synthèse se produisent. Pour ce faire, des outils permettant de contrôler les erreurs sont utilisés, on parle d'un contrôle par redondance, car des données sont ajoutées à la fin du message pour détecter des défauts.

Les outils les plus simples sont les outils appelés les sommes de contrôle (ou « checksum » en anglais). Cette somme de contrôle est un nombre ajouté à la fin du message à transmettre qui permet au lecteur de vérifier que le message reçu est bien celui envoyé. Il existe différentes sommes de contrôle comme le bit de parité, le contrôle par redondance longitudinale ou encore le contrôle sur les chiffres.

Lors de l'élaboration des séquences nous avons choisi d'utiliser le bit de parité. Le principe du bit de parité est de faire la somme d'un groupement de bits, si le résultat est pair le bit de parité ajouté à la fin de ce groupement est le bit 0, si le résultat est impair le bit de parité est le bit 1. Cette méthode de contrôle permet de transmettre des informations partielles sur un segment de taille et d'emplacement connus. En effet, le bit de parité permet de signaler si la somme sur la série de bit est bonne et donc dans ce cas, d'indiquer qu'aucune faute n'est détectée. Néanmoins si le bit de parité montre qu'il y a une erreur dans la série de bit, il ne précise pas quel bit est faux. De plus, si plus d'une erreur se produit il se peut également que le bit de parité n'indique aucune erreur alors qu'il se pourrait qu'il y ait plusieurs erreurs qui se compensent. Ce contrôle permet toutefois d'augmenter la fiabilité du message transmis.

4. Elaboration d'une séquence codant une grande quantité d'information

Grâce à l'élaboration de la liste des nouveaux marqueurs, des outils de compression et de contrôle d'erreurs il est possible dorénavant de mettre en place la synthèse de séquence à haute capacité de stockage codant de l'information comprimée.

Pour ce faire, une taille de séquence à atteindre et le type d'information à coder est déterminé dans un premier temps. Il est choisi ici de coder une image. Celle-ci est plus facilement codée qu'un texte lorsque l'alphabet à huit symboles est utilisé. Il n'est pas nécessaire d'ajouter de bit non codant, qui est nécessaire lorsque qu'un texte ASCII est codé avec cet alphabet (comme expliqué dans le paragraphe 3.2.2 du chapitre II). Un polymère contenant 88 unités codées a été étudié. Cette séquence permet de coder une image comprenant 264 pixels sans outils de compression.

Un premier essai a été effectué en pixélisant le portrait d'Antoine Laurent Lavoisier, visible en Figure IV. 5. Le portrait de Lavoisier sur lequel la pixélisation est effectuée est donné à gauche de la Figure IV. 5. Il est issu du tableau Jacques-Louis David représentant Lavoisier et sa femme. L'image de droite montre le résultat obtenu lorsque la résolution du portrait est diminuée jusqu'à obtenir une image pixélisée d'une taille de 22*20 soit 440 pixels. Les formes du visage de Lavoisier sont encore visibles avec notamment ces yeux et son nez en noir sur la partie du milieu vers la gauche. Son jabot est également reconnaissable en bas de l'image.

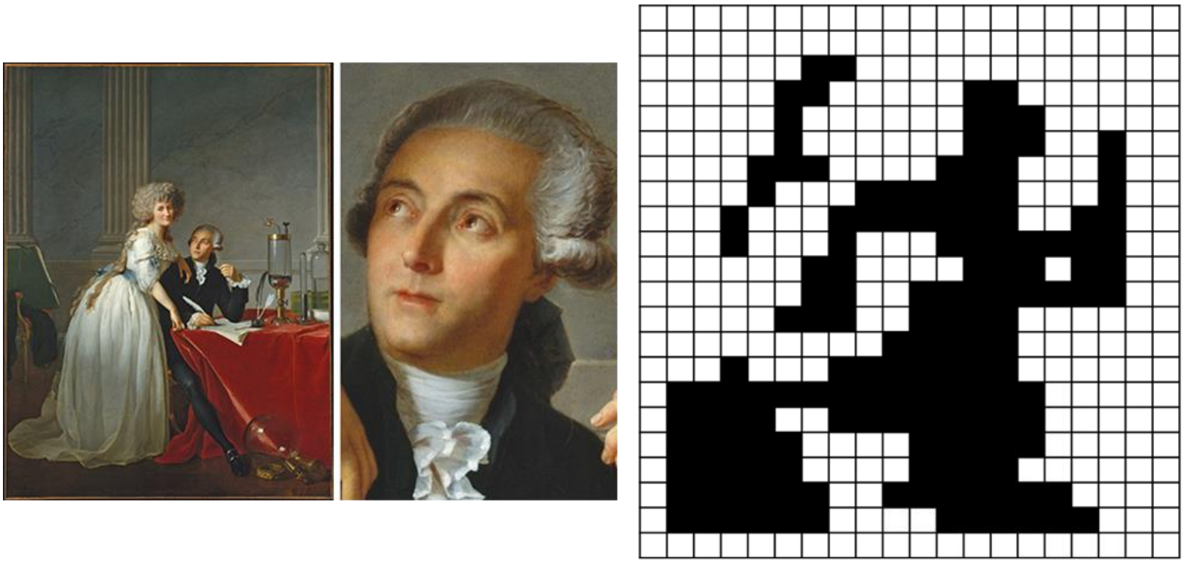


Figure IV. 5: (Gauche) Portrait du couple Lavoisier exposé au Metropolitain Museum of Art à New York, avec le zoom sur le visage d'Antoine Laurent Lavoisier. (Droite) Image pixélisée de Lavoisier.

Cette image pixélisée peut ensuite être convertie en flux de bits, donc en une suite de 440 bits 0 et 1, comme indiqué ci-dessous :

*[0000000000000000000000000000000000000000000011000000000000000000110000011000000000000100
0000111000000000000100000011100100000001110000111000100000010001111110001000001000111111001
1000001000100011111110000000011000011101100000000110011111111000000011100111100000000000000
0001111100000000001000111111100000000111111111111100000011110011111111000000111110000111111
0000000111110000111100000000111111001111111000000111111000011111000000000000000000000000]*

Un bit de parité est ajouté à la fin de la séquence des 440 bits, ici la somme des bits de la séquence donne 155, ce chiffre étant impair le bit de parité sera donc un 1. En utilisant la méthode de compression arithmétique il est possible de réduire ce flux de 441 bits à 264 bits soit une compression de 40%. Le nouveau code est donné ci-après.

*[0000000011101100110011111110100110111111110011000110011001110110100101100111101000100101 1
1110110011110110100100101011010000010110001100111010000101111110100011101000011111100100
10111100000111100011111000111101000101110100011111110011111101100011101101101100110111 1]*

Ce flux de bits comprimé peut se traduire en une suite de 88 monomères M1 à M8, avec un espaceur E tous les huit monomères, précédé par un des marqueurs préalablement définis.

[M1.M1.M2.M7.M7.M4.M2.M8-E-
M8.M3.M4.M4.M8.M7.M4.M1.P-E-
M7.M4.M2.M7.M7.M5.M6.M5.F_C-E-
M8.M6.M1.M5.M6.M8.M6.M5.Br_G-E-
M8.M6.M6.M2.M2.M3.M7.M5.F_G-E-
M1.M6.M5.M4.M2.M7.M5.M2.I_U-E-
M4.M8.M6.M1.M8.M3.M1.M8.Br_U-E-
M8.M5.M5.M6.M8.M1.M2.M8.G-E-
M1.M8.M7.M2.M8.M3.M2.M4.A-E-
M6.M1.M8.M8.M5.M8.M8.M4.C-E-
M1.M8.M4.M4.M4.M2.M6.M8-T]

Lors de l'élaboration de cette première séquence, il a été choisi de ne pas mettre d'entête définissant les conventions internes. Elles sont toutes connues en convention externe. Ainsi, pour décrypter le message le lecteur doit savoir que :

1) Il s'agit d'une image d'une taille 22*20 pixels
2) L'alphabet utilisé est l'alphabet à 8 symboles et l'ordre des marqueurs est : pas de marqueur, dP, dF_C, dBr_G, dF_G, dI_U, dBr_U, dG, dA, dC et dT
3) Cette image est compressée en utilisant la méthode arithmétique, la table de décryptage est également transmise
4) Il y a un bit de parité en fin de séquence décompressée

4.1. Utilisation de l'espaceur E1

Cette séquence a donc été synthétisée en utilisant l'espaceur E1 dit classique, dont la structure est donnée dans **le chapitre I**. Cette synthèse est effectuée sur le synthétiseur ABI en appliquant le protocole c : la concentration des solutions des monomères M4 à M8 est augmentée à 60 mM au lieu de 50 mM. Deux séquences ont été effectuées en parallèle et diffèrent l'une de l'autre par le temps de clivage. A la fin de la synthèse, les séquences ont été purifiées en suivant la méthode habituelle mais en augmentant le temps de clivage à 1h pour la première et 1h30 pour la deuxième pour permettre de déprotéger toutes les fonctions amines des marqueurs dG et dF_G. Finalement, après lyophilisation des solutions obtenues, une poudre blanche a été obtenue avec 53% de rendement pour la première et 36% de rendement pour la deuxième. Elles ont été toutes les deux analysées en spectrométrie de masse et aucune différence n'est notable entre les deux échantillons, il a donc été décidé de garder un temps de clivage de 1h. Le spectre de masse obtenu est donné en Figure IV. 6.

Aucun signal attribuable à la séquence intacte n'a pu être détecté. Le spectre montre essentiellement des signaux libérés en sources attribuables à chaque octet à différents états de charge. Une couleur est donnée à chaque bloc (montré sur le schéma de la séquence en Figure IV. 6). Les pics correspondant aux différents blocs sont annotés dans la même couleur. Une fragmentation en source semble se produire et cette analyse ne permet pas de fournir une preuve formelle que la séquence intacte est bien synthétisée. Même si la détection de chaque bloc tend à nous faire penser que le polymère est bien synthétisé, il n'est pas observable. Cela peut s'expliquer par une taille trop importante de la séquence.

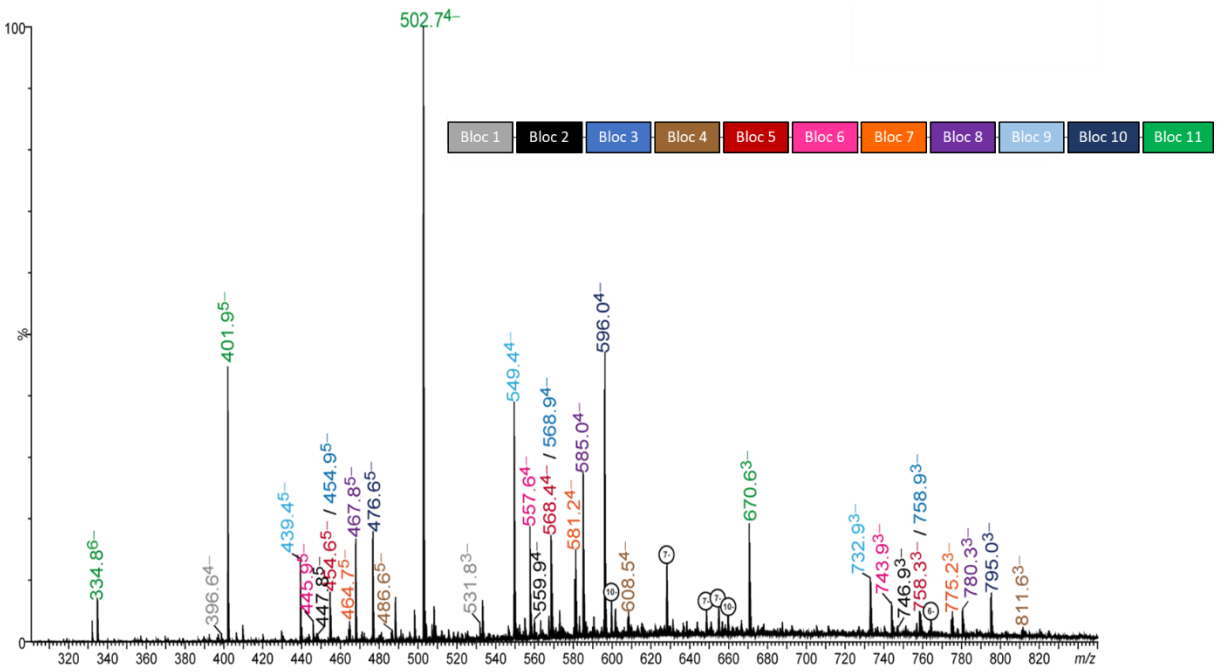


Figure IV. 6: Spectre ESI-MS obtenu pour la séquence codant l'image de Lavoisier compressée. Les différentes couleurs des pics correspondent aux couleurs attribuées à chaque bloc de la séquence schématisée en inset.

De plus, lors de cette analyse une nouvelle règle à respecter pour les marqueurs de masse a été identifiée. Il s'est avéré que les blocs 3 et 5 n'avaient pas une différence de masse suffisante pour permettre leur différenciation. La masse du bloc 3 est 2279,8423 Da et celle du bloc 5 est 2277,8015 Da, il y a donc moins de 3 Da de différence entre les deux blocs. Or, les règles décrites aux préalables ont bien été respectées mais avec le problème soulevé ici il s'est avéré qu'il faut affiner les règles pour que ce problème soit éliminé.

La nouvelle règle à respecter est que la différence de masse entre deux marqueurs doit être supérieure à un multiple de 14 **+/- 3**. Autrement dit, l'écart de masse entre -la différence de masse entre 2 blocs sans marqueur- et -la différence de masse entre les 2 marqueurs utilisés pour ces blocs- doit être plus grande que 3 Da et pas un multiple de 14.

Ici, entre le bloc 3 et 5 sans marqueurs la différence de masse est de 42 Da, or les 2 marqueurs utilisés F_C (307g/mol) et F_G (347g/mol) ont une différence de 40 Da. Donc en combinant les 2 la différence globale n'est que de 2 Da, ce qui n'est pas suffisant pour les différencier. La paire F_G/F_C ne respecte donc pas la nouvelle règle. Il s'est avéré qu'il en est de même pour la paire A/G avec une différence de masse de 16 < (1*14 - 3). Néanmoins, la composition des blocs pour lesquels les marqueurs dA et dG ont été utilisés nous a permis de passer outre ce problème. Dans le cas de la séquence codant pour l'image de Lavoisier, il a été décidé d'échanger la position des marqueurs dF_C et dP pour permettre d'avoir une différence supérieure à 3 Da entre chaque bloc de la séquence. Cependant, pour avoir des marqueurs universels il faudrait enlever de la liste un marqueur de chaque couple qui pose un problème : dA ou dG et dF_C ou dF_G.

La séquence a été synthétisée à nouveau en échangeant les marqueurs dP et dF_C. Une séquence plus petite avec 2 blocs en moins a également été synthétisée. Cette séquence plus courte a été synthétisée pour déterminer si une séquence de ce type contenant 89 monomères peut être détectée intacte lors de l'analyse de spectrométrie de masse. Les 2 séquences ont été purifiées et retrouvées avec un

rendement de 67 % pour la séquence tronquée (Lav1) et un rendement de 39% pour la séquence complète (Lav2). La description des deux séquences est donnée dans le Tableau IV. 3 suivant.

Tableau IV. 3: Description des polymères numériques codant une image compressée

Nom	Séquence de monomères [a]	*DP* [b]	*M* [c] *(Da)*
Lav1	M_7·M_4·M_2·M_7·M_7·M_5·M_6·M_5·P-*E1*- M_8·M_6·M_1·M_5·M_6·M_8·M_6·M_5·Br_G-*E1*- M_8·M_6·M_6·M_2·M_2·M_3·M_7·M_5·F_G-*E1*- M_1·M_6·M_5·M_4·M_2·M_7·M_5·M_2·I- *E1*- M_4·M_8·M_6·M_1.M_8·M_3·M_1·M_8·Br-*E1*- M_8·M_5·M_5·M_6·M_8·M_1·M_2·M_8·G-*E1*- M_1·M_8·M_7·M_2·M_8·M_3·M_2·M_4·A-*E1*- M_6·M_1·M_8·M_8·M_5·M_8·M_8·M_4·C-*E1*- M_1·M_8·M_4·M_4·M_4·M_2·M_6·M_8-T	89	20271.6120
Lav2	M_1·M_1·M_2·M_7·M_7·M_4·M_2·M_8-*E1*- M_8·M_3·M_4·M_4.M_8·M_7·M_4·M_1·F_C-*E1*- M_7·M_4·M_2·M_7·M_7·M_5·M_6·M_5·P-*E1*- M_8·M_6·M_1·M_5·M_6·M_8·M_6·M_5·Br_G-*E1*- M_8·M_6·M_6·M_2·M_2·M_3·M_7·M_5·F_G-*E1*- M_1·M_6·M_5·M_4·M_2·M_7·M_5·M_2·I- *E1*- M_4·M_8·M_6·M_1.M_8·M_3·M_1·M_8·Br-*E1*- M_8·M_5·M_5·M_6·M_8·M_1·M_2·M_8·G-*E1*- M_1·M_8·M_7·M_2·M_8·M_3·M_2·M_4·A-*E1*- M_6·M_1·M_8·M_8·M_5·M_8·M_8·M_4·C-*E1*- M_1·M_8·M_4·M_4·M_4·M_2·M_6·M_8-T	108	24358,0781

[a] Les points et les tirets n'ont pas de valeur chimique et sont insérés pour plus de clarté. [b] Degré de polymérisation incluant les monomères, l'espaceur et les marqueurs de masse. [c] Masse monoisotopique.

La séquence Lav1 ne contenant que 9 blocs sur les 11 de la séquence entière a été retrouvée intacte mais en très faible quantité comme montré sur la Figure IV. 7. Les signaux ont une trop faible intensité pour permettre d'effectuer une analyse MS/MS qui permettrait de valider la composition de la séquence. A nouveaux, tous les blocs ont été libérés en source et sont directement visible dès la première analyse. Chaque bloc est bien retrouvé grâce aux marqueurs de masse qui permettent cette fois de différencier sans aucun doute chaque bloc. La détection de la séquence entière nous montre toutefois que la synthèse d'une séquence comprenant 9 blocs est bien possible. La synthèse de longues séquences est donc bien possible, mais ces dernières semblent fragiles car même lorsqu'elles sont détectées, elles le sont en très faible quantité. Il apparait qu'une grande partie de la séquence synthétisée soit fragmentée avant même de pouvoir être détectée lors de l'analyse MS.

L'analyse de la séquence complète Lav2 ne permet pas de retrouver la séquence intacte. Comme précédemment, il n'y a que les blocs libérés en sources qui sont retrouvés, aucun signal attribuable au polymère intact n'est visible. Aucune impureté ne vient cependant parasiter le spectre, ceci est le signe que la synthèse s'est bien déroulée et qu'aucune réaction secondaire notable se produit durant la synthèse. Des impuretés auraient sinon été détectées sur le spectre.

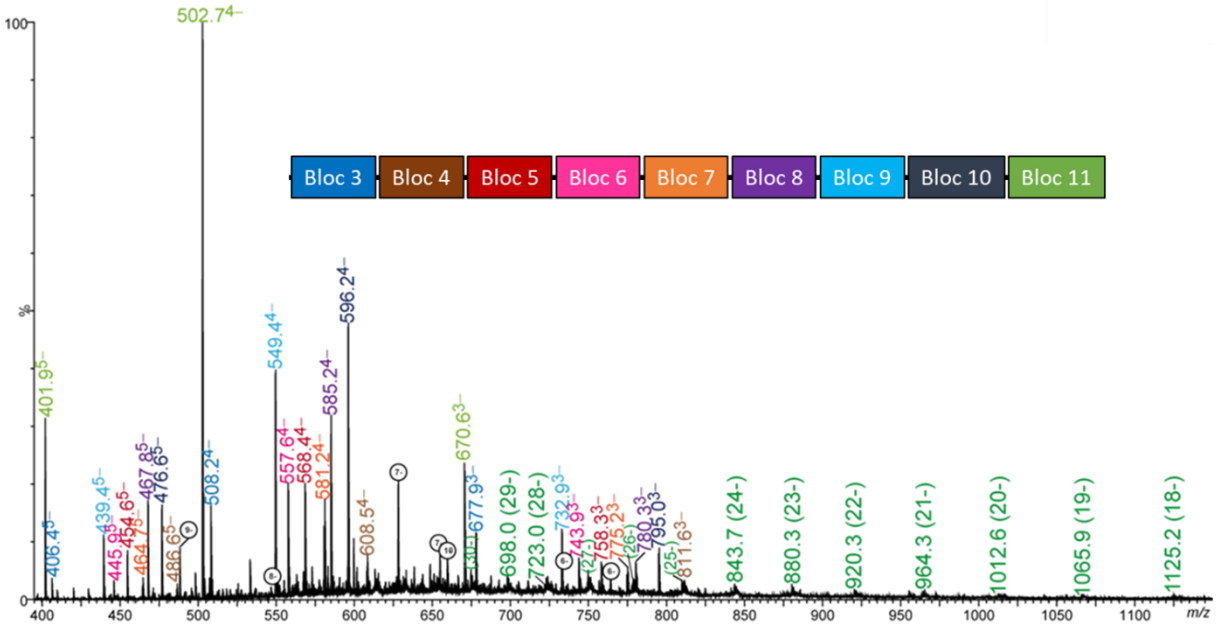


Figure IV. 7: Spectre ESI-MS obtenu pour la séquence codant une partie de l'image de Lavoisier compressée contenant que 9 blocs (Lav1). Les différentes couleurs des pics correspondent aux couleurs attribuées à chaque bloc de la séquence schématisée en inset.

D'après ces deux analyses, il semblerait que la synthèse de grandes séquences telles que celles-ci, soit possible. Toutefois leur fragilité ne permet pas de les détecter lors de l'analyse par spectrométrie de masse. Cette fragilité semble venir de l'espaceur car les éléments retrouvés lors de l'analyse MS correspondent aux différents blocs qui sont normalement détectés lors de l'analyse MS/MS après la fragmentation de cet espaceur. Il semble donc que l'espaceur soit trop fragile pour résister aux conditions d'analyse.

4.2. Utilisation de l'espaceur RISCOP

Une nouvelle série de deux séquences a été effectuée, mais cette fois en utilisant l'espaceur RISCOP décrit dans le **chapitre III** (paragraphe 3.2.2) et non plus l'espaceur E1. Cet espaceur montrant de meilleurs résultats lors des analyses pourrait permettre d'obtenir des signaux attribuables à la séquence intacte lors de la lecture de cette grande séquence. A nouveau, une première séquence plus courte ne contenant que 9 blocs a été synthétisée en parallèle de la séquence complète. Ces deux séquences ont été clivées pendant 1h et une poudre blanche a été isolée après lyophilisation avec un rendement de 49% pour la séquence Lav3 comprenant 9 blocs et 44 % pour la séquence complète Lav4. Leur description est donnée dans le Tableau IV. 4.

Tableau IV. 4: Description des polymères numériques codant une image compressée avec l'espaceur RISCOP

Nom	Nombre de bloc	*DP* [a]	*M* [b] *(Da)*
Lav3	9	89	20311,5696
Lav4	11 (séquence entière)	108	24408,0244

[a] Degré de polymérisation incluant les monomères, l'espaceur et les marqueurs de masse. [b]Masse monoisotopique.

L'analyse de spectrométrie de masse ne montre pas d'amélioration avec l'utilisation de l'espaceur RISCOP. Les signaux correspondant aux séquences entières n'ont pas pu être détectés. Mais, de nouveau, des signaux libérés en sources correspondant à chaque bloc sont trouvés. Le spectre de la séquence entière Lav4 est donné en Figure IV. 8.

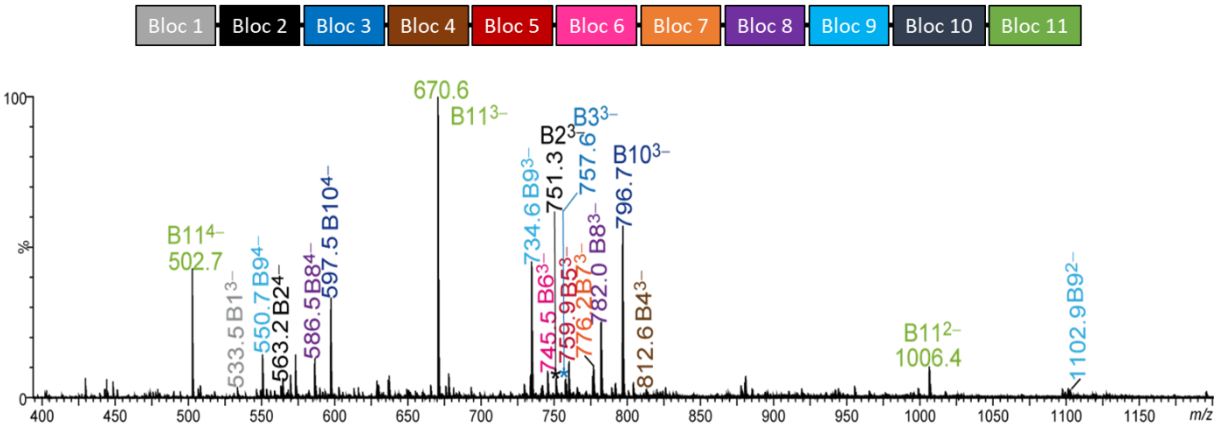


Figure IV. 8: Spectre ESI-MS obtenu pour la séquence codant l'image entière de Lavoisier compressée avec l'espaceur RISCOP (Lav4). Les différentes couleurs des pics correspondent aux couleurs attribuées à chaque bloc de la séquence schématisée en inset.

L'analyse pseudo-MS^3 a tout de même été effectuée sur chaque bloc libéré en source. Celle-ci permet de confirmer la composition de tous les octets. Ainsi, il est possible de retrouver le flux de bits comprimés de chaque bloc. Avec les marqueurs de masse, la position d'origine des octets est retrouvée et le flux de bits entier est remis dans le bon ordre. Sachant que ce flux a été comprimé avec le codage arithmétique, il est maintenant facile de retrouver le message non compressé. De plus, grâce à l'utilisation de l'espaceur RISCOP, la couverture de chaque bloc est totale (données en **partie expérimentale**).

Il semble donc que la séquence entière ait bien été bien synthétisée mais que le polymère se fragmente très facilement au niveau de l'espaceur. Cette fragmentation peut se produire lors du transport avant l'analyse de spectrométrie de masse qui s'effectue dans les laboratoires de l'université d'Aix-Marseille. L'analyse MS en elle-même peut également les produire. Des analyses d'HPLC et de chromatographie d'exclusion stérique (SEC) ont alors été effectuée pour mesurer la pureté et la polymolécularité du polymère intact.

Les analyse d'HPLC ont été effectuées après la lyophilisation des échantillons. Les résultats obtenus, présenté en Figure IV. 9, montrent un échantillon contenant plusieurs populations, dont aucune n'est majoritaire. Cette analyse ne nous permet pas d'affirmer que la séquence a bien été synthétisée. Les différents pics observés peuvent correspondre aux différents blocs que l'on observe lors de l'analyse MS. Le nombre de pic présent sur l'amas éluant entre 8 et 15 min de la séquence Lav3 semble corroborer cette hypothèse. En effet, on peut compter 8 pics principaux (entourés en pointillés verts sur la Figure IV. 9), qui pourraient correspondre aux 8 blocs de la séquence Lav3 contenant un marqueur de masse (Le premier n'en comprenant pas).

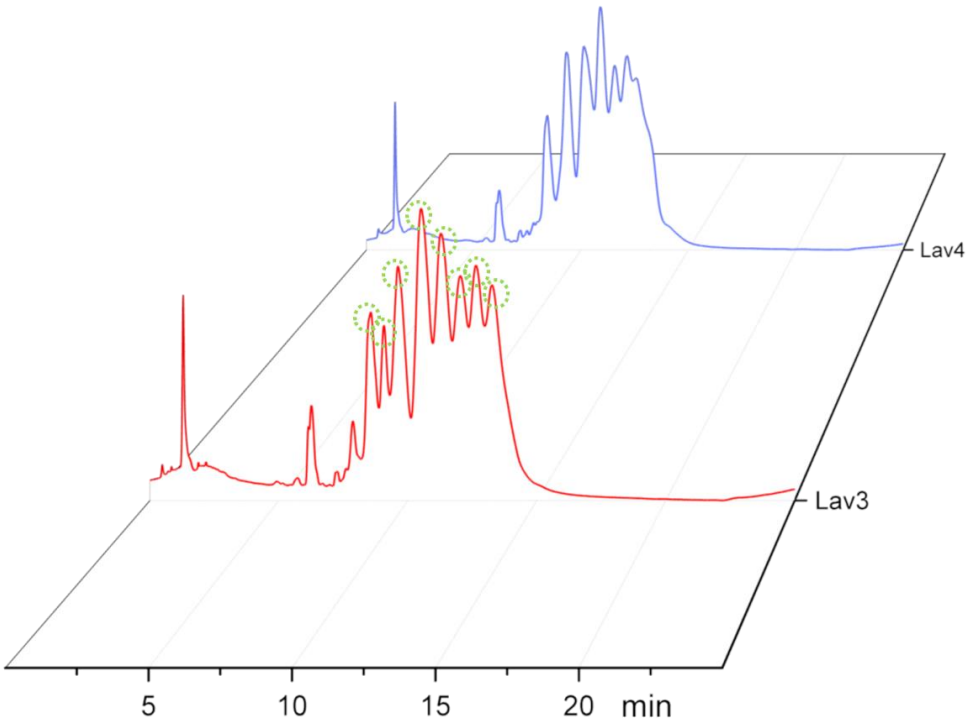


Figure IV. 9: Spectres obtenus après les analyses HPLC des séquences Lav3 (rouge) et Lav4 (bleu).

Une analyse par SEC a également été effectuée. Les résultats obtenus pour l'échantillon Lav3 sont données en Figure IV. 10. Le calcul de masse molaire est effectué lors de l'analyse SEC grâce à un couplage avec la diffusion de lumière. Pour cela, le paramètre d'incrément d'indice de réfraction (dn/dc) est essentiel. Il n'a pas été possible de calculer ce paramètre pour ces séquences. Néanmoins, ce calcul avait été effectué auparavant dans l'équipe, pour les séquences de poly(phosphodiester)s contenant les monomères M1 et M3 (structure donnée dans le **chapitre II**). Le paramètre dn/dc obtenu était de 0,107 mL/g,[23] cette valeur a donc été utilisée pour l'analyse de la séquence Lav3. Il est fort probable qu'il ne soit pas correct car les monomères impliqués dans la structure de la séquence Lav3 sont différents de ceux utilisés lors de la mesure. Les résultats de masse molaires obtenues sont montrés sur la courbe orange de la Figure IV. 10. La masse molaire moyenne calculée est 12 000 g/mol alors que la masse théorique est de 24 408 g/mol. Il se peut que le calcul de cette valeur soit faussé par le manque du bon paramètre d'incrément d'indice de réfraction. Le paramètre dn/dc de l'ADN (0,17) a également été utilisé pour effectuer les calculs, mais les résultats ne se rapprochent pas des valeurs théoriques avec une masse molaire moyenne calculée à 7500 g/mol. Des analyses sont prévues pour mesurer l'incrément d'indice de réfraction propre à ses séquences, permettant ainsi d'avoir des résultats plus justes. Dans tous les cas l'analyse montre un chromatogramme polymodal (trimodal) avec un indice de polymolécularité à 1,254.

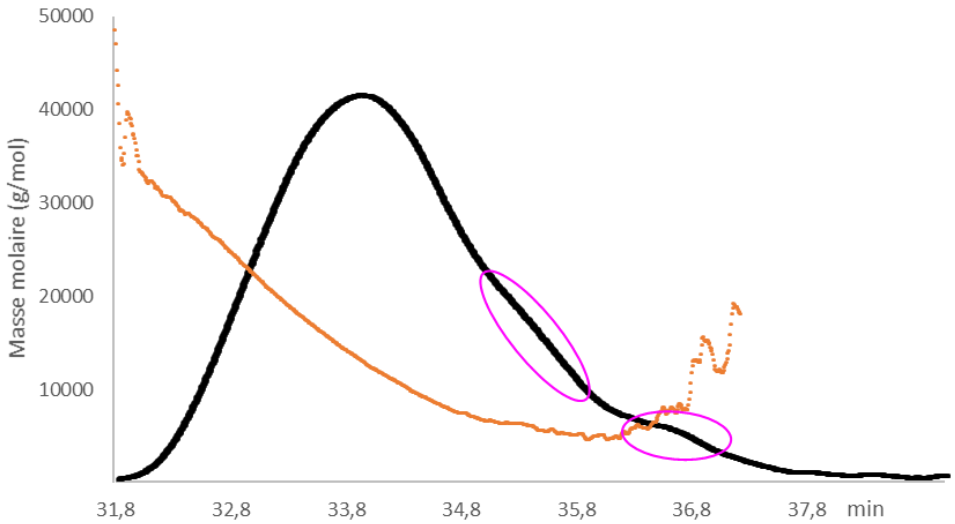


Figure IV. 10: Courbes obtenues après l'analyse SEC de l'échantillon Lav3. La courbe en noire donne les populations de l'échantillon et la courbe en orange donne les masses molaires calculées.

En effet, le profil du chromatogramme noir de la Figure IV. 10, nous indique que l'échantillon ne contient pas une seule population étant donné que le pic est assez large et qu'on observe plusieurs épaulements qui sont encerclés en rose. Il apparaît qu'il a au moins trois populations différentes dans l'échantillon. Toutefois, la masse au pic principal est calculée à 16 000 g/mol, ce qui montre qu'il y a une population ayant une masse molaire importante et qui pourrait correspondre au polymère recherché. La masse au pic des deux autres populations est calculée à 7500 g/mol pour le deuxième pic et à 5000 g/mol pour le troisième.

L'analyse de SEC ne permet pas pour l'instant de confirmer qu'une séquence monodisperse correspondant à la séquence recherchée a bien été synthétisée, car les paramètres de calculs ne sont pas adaptés à la structure de cette séquence. Toutefois, elle montre une population majoritaire qui pourrait correspondre à la séquence désirée. Ces résultats préliminaires sont intéressants, et encouragent à poursuivre la synthèse de longue chaîne de poly(phosphodiester)s à très haute capacité de stockage. D'autres séquences vont être prochainement synthétisées et analysées le plus rapidement possible en gardant les échantillons au congélateur directement après la lyophilisation.

5. Conclusion et perspectives

Les études menées dans le **chapitre IV** montrent qu'il est encore possible d'augmenter la densité de stockage dans une chaîne de polymère unique. Nous avons vu que l'emploi de marqueurs de masse permet d'élaborer de très grandes séquences telles qu'une séquence de poly(phosphodiester)s ayant une capacité de stockage de 440 bits. Cette nouvelle séquence code plus de 3 fois plus de bits que celle présentée dans le **chapitre II** et synthétisée sans outils de compression.

Quel que soit l'espaceur utilisé, la synthèse de ces longues séquences semble possible. Une séquence de 9 blocs contenant donc 216 bits a été synthétisé, même si cette séquence ne code qu'une information partielle elle démontre qu'il est possible de synthétiser de longue séquence de poly(phosphodiester)s. Néanmoins, ces séquences sont très fragiles, et il est difficile de les analyser. Les trois techniques d'analyses utilisées (MS, HPLC et SEC) montrent des échantillons contenant plusieurs populations. Pour l'instant, ces analyses ne permettent pas de vérifier systématiquement

que la séquence entière est retrouvée après l'étape de purification. La synthèse de nouvelles séquences va être effectuée. Ces dernières pourront alors être analysées en chromatographie d'exclusion stérique en utilisant leur paramètre d'incrément d'indice de réfraction qui serra mesurer prochainement.

Les marqueurs de masse sont un des éléments essentiels pour réussir à encoder ces longues chaînes de polymères. Sans eux il serait impossible de lire dans le bon ordre l'information initialement encodée. Nous avons vu qu'il est primordial de bien choisir ces marqueurs pour permettre de retrouver le message de manière inéquivoque. Pour l'instant, les marqueurs utilisés sont tous commerciaux, mais il est possible d'en synthétiser au laboratoire en se basant sur l'exemple donné dans le **chapitre III**, où il est montré qu'il est possible de synthétiser un espaceur-marqueur. Ce type d'espaceur-marqueur permettrait d'avoir une séquence composée entièrement d'éléments synthétiques et de se séparer complétement des éléments contenus également dans des séquences naturelles. Des messages de mêmes longueurs seraient alors encodés sur des chaînes de polymères encore plus courtes, la densité de stockage serait encore augmentée.

Pour atteindre des densités de stockage toujours plus hautes, il a également été montré que l'utilisation d'outils informatiques est indispensable. Ces outils permettent d'obtenir un message comprimé à 40%, ainsi 441 bits sont encodés sur seulement 108 monomères. Sans une compression du message initial, il est impossible de synthétiser une séquence de 441 monomères, beaucoup trop longue pour être synthétisée dans les conditions utilisées. Ces outils sont donc essentiels pour permettre d'encoder toujours plus d'information sur une chaîne unique de polymère.

Ce **chapitre IV** nous a donc permis de constater que l'utilisation simultanée d'une grande quantité de marqueurs de masse et d'outils informatiques est primordiale pour atteindre une très haute capacité de stockage sur une chaîne unique de poly(phosphodiester).

Conclusion générale

Au cours de cette thèse, la synthèse de de poly(phosphodiester)s numériques a été étudiée. Ces polymères sont utilisés pour stocker de l'information à l'échelle moléculaire. De nombreux systèmes existent pour stocker des données sur des chaînes de polymères, mais les séquences de poly(phosphodiester)s sont pour l'instant les seules permettant d'atteindre de hautes capacités de stockage.

Leurs synthèses se basent sur la chimie de la phosphoramidite qui est une stratégie très efficace et facile à mettre en œuvre car automatisée. Au cours de la synthèse un message binaire est encodé sur la séquence. Les monomères forment ainsi une succession de bits 0 et 1 qui codent de l'information. A la suite de leurs synthèses, les séquences numériques sont analysées par spectrométrie de masse. Cette analyse permet de retrouver l'information encodée sur ces macromolécules. Un design spécifique a toutefois dû être créé pour faciliter cette analyse, qui s'avère compliquée lorsque les séquences sont longues. Un espaceur a ainsi été ajouté tous les 8 monomères. Cette nouvelle molécule est l'élément le plus fragile de la séquence et permet une fragmentation contrôlée lors de l'analyse de spectrométrie de masse en tandem. En effet lors de cette analyse, le clivage de l'espaceur permet la libération de petits blocs d'oligo(phosphodiester)s qui sont plus facilement déchiffrables en pseudo-MS^3. Les données comprises dans chaque bloc peuvent ainsi être recouvrées. Un marqueur moléculaire, présent en bout de bloc, permet de les remettre dans le bon ordre. Le message initialement encodé est alors déchiffré.

Ce design a permis de synthétiser des chaînes de poly(phosphodiester)s contenant de l'information. Il peut toutefois être optimisé pour améliorer la capacité de stockage et la lecture de ces polymères numériques. Cette optimisation a été réalisée au cours de cette thèse. Différents points d'améliorations ont été effectués, ces derniers se sont répartis en trois catégories représentées sur la Figure C. 1. Dans un premier temps, l'alphabet moléculaire a été amélioré. Des alphabets augmentés ont ainsi été créés. Dans un second temps, le lien inter-octet a été perfectionné grâce à la synthèse d'un nouvel espaceur s'incorporant dans des séquences de poly(phosphodiester)s numériques dont la lecture peut être automatisée. Finalement, une étude sur les marqueurs moléculaires a été effectuée, ce qui a permis d'augmenter leur nombre. Grâce à cette quantité augmenté de marqueurs moléculaires, il a alors été possible de synthétiser des séquences à très haute capacité de stockage. De surcroît, l'utilisation d'outils informatiques a permis d'augmenter encore plus la densité de stockage d'une chaîne unique de poly(phosphodiester)s. Ces trois axes de recherches ont été développés dans les différents chapitres de cette thèse.

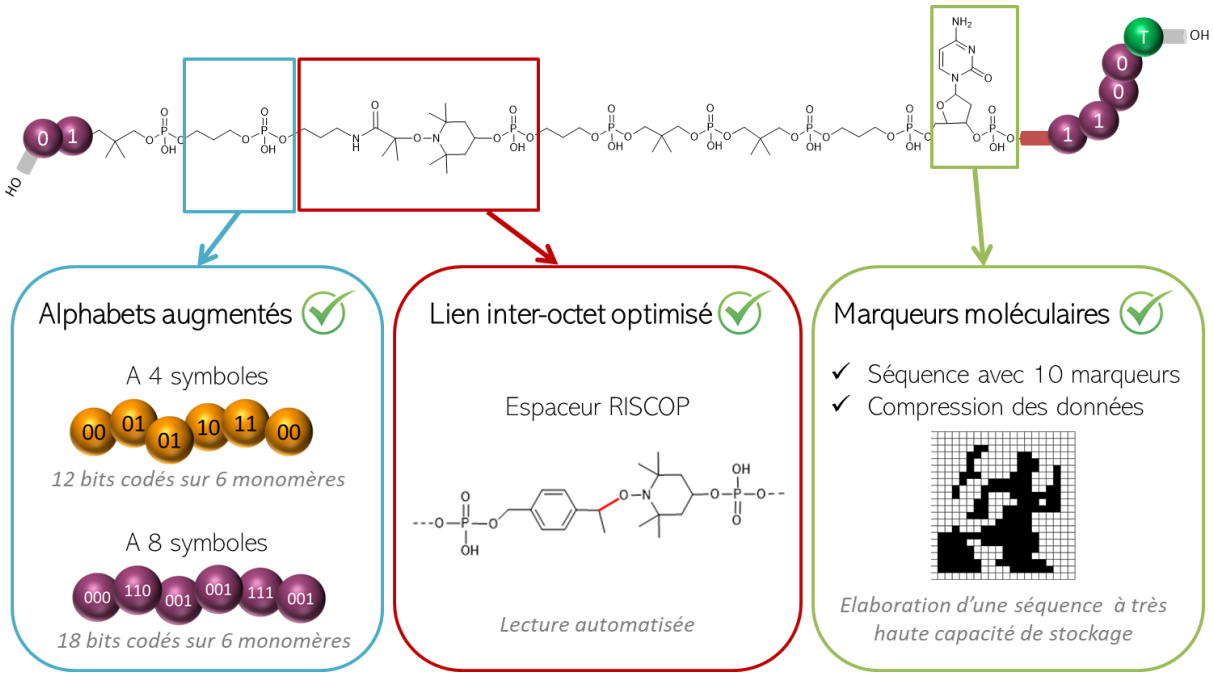


Figure C. 1: Plan général des études qui ont été menées au cours de cette thèse. En haut, séquence type de poly(phosphodiester)s codant de l'information. Bleu : Amélioration de l'alphabet (chapitre II). Rouge : Amélioration de l'espaceur (chapitre III). Vert : Amélioration des marqueurs moléculaires (chapitre IV).

Ainsi, l'objectif du **chapitre II** était de développer un nouvel alphabet plus efficace permettant de coder plus de bits sur un même monomère. Il a été montré qu'au lieu d'utiliser seulement deux monomères codant chacun pour un bit (0 ou 1), il était plus efficace d'utiliser des systèmes composés de 4 ou 8 monomères encodant pour 2 et 3 bits respectivement. En effet, en utilisant l'alphabet augmenté contenant 4 symboles, les bits sont codés sur des dyades (*i.e.* 2 bits/monomère) et donc pour encoder une lettre il ne faudra plus que 4 monomères et non plus 8 comme auparavant. Avec le système augmenté à 8 symboles, chaque monomère code pour une triade (*i.e.* 3 bits/monomère), le nombre nécessaire de monomère pour encoder la même information est donc encore réduite par rapport aux deux autres systèmes. Grâce à la mise en place de ces nouveaux alphabets il a été possible d'encoder de nombreuses données comme du texte et des images. L'encodage d'une image de 16 lignes sur 9 colonnes donnant un total de 144 bits a permis la synthèse de la plus longue chaîne codante de poly(phosphodiester)s. L'alphabet comprenant 8 monomères, n'est pas une limite, il est encore possible d'étendre l'alphabet utilisé, et des systèmes à 16 monomères codants pourraient être envisagés.

Le **chapitre III** était consacré à l'incorporation de nouveaux espaceurs dans des chaînes de poly(phosphodiester)s numériques. L'espaceur classique E1 qui avait déjà été utilisé auparavant provoque des réactions parasites lors de l'analyse de spectrométrie de masse. La synthèse de séquences comprenant des nouveaux espaceurs a alors été effectuée. Lorsque leurs synthèses étaient possibles, ces séquences ont été analysées en spectrométrie de masse pour vérifier qu'aucune réaction secondaire n'est détectée. Les études menées ont conduit à la validation d'un nouvel espaceur nommé RISCOP. L'incorporation de ce dernier dans une chaîne de poly(phosphodiester)s est possible et il permet d'éviter toutes les réactions parasites observées lors de l'analyse. Grâce à ce nouvel espaceur RISCOP, des analyses automatisées des spectres de masse ont été effectuées. Le logiciel MS-DECODER utilisé a permis de retrouver les informations encodées en quelques secondes.

Le **chapitre IV** décrivait la synthèse de poly(phosphodiester)s numériques ayant une très haute capacité de stockage. Pour atteindre une grande taille de chaîne, qui soit toujours déchiffrable lors de l'analyse de spectrométrie de masse, il faut utiliser un nombre augmenté de marqueurs masses. Ces derniers permettent de structurer la séquence en blocs de 8 monomères qui sont faciles à déchiffrer lors de l'analyse. L'étude menée sur ces marqueurs moléculaires a permis d'en sélectionner plusieurs qui ont été utilisé pour synthétisés une longue séquence. Pour atteindre une densité de stockage encore plus grande, des outils informatiques ont été utilisés pour compresser le message. Cette procédure a été employée pour encoder une grande image de 440 bits sur une chaîne de polymère contenant au total 108 unités. Une compression de 40% a été possible grâce à l'utilisation de l'algorithme « arithmetic coding ». Dix marqueurs de blocs différents ont été nécessaires pour élaborer le design de cette séquence.

La lecture des très longues séquences de poly(phosphodiester)s semble néanmoins plus compliquée en spectrométrie de masse. Ces polymères sont trop fragiles pour pouvoir recouvrer un signal correspondant à leur chaîne entière lors de l'analyse de spectrométrie de masse. Tous les petits blocs sont libérés en sources dès la première analyse MS. Ces blocs permettent de retrouver l'information encodée en les analysant en pseudo-MS^3, mais une preuve formelle que la séquence entière a bien été synthétisée n'est pas obtenue grâce à cette analyse. D'autres analyses comme la HPLC et la SEC donnent pour l'instant des résultats similaires. Une étude approfondie des analyses SEC est néanmoins possible, ce qui permettrait d'avoir des résultats moins approximatifs. La combinaison de différentes techniques d'analyses semble pour le moment nécessaire pour la lecture des très longues séquences de poly(phosphodiester)s.

Au cours des différents chapitres il a été montré à quel point l'utilisation de la chimie de la phosphoramidite est efficace pour synthétiser de longues chaines codantes de poly(phosphodiester)s. A ce jour et à notre connaissance, c'est avec ce type de séquence qu'il est possible d'avoir la plus grande densité de stockage sur une chaîne unique. Et ceci en utilisant ou non des outils de compression. En effet, sans outils de compression un polymère contenant 144 bits est encodé, et avec l'utilisation d'outils de compression la capacité de stockage monte jusqu'à 440 bits codants. Ainsi, les chaînes de poly(phosphodiester)s décrites dans cette thèse sont, à notre connaissance, les séquences numériques qui permettent de stocker le plus d'information et ayant une lecture automatisée facilitée grâce à l'élément espaceur clivable implémenté dans leurs structures.

Pour imaginer une utilisation viable dans l'avenir de ces séquences, il faudrait automatiser encore plus la totalité de la procédure. Pour l'instant, l'étape de synthèse est automatisée grâce à l'utilisation de synthétiseur d'ADN et l'étape de lecture des spectres l'est également via l'utilisation de l'algorithme MS-DECODER. Néanmoins, une personne est tout de même nécessaire pour suivre la synthèse, purifier les séquences et pour transmettre les échantillons aux différentes personnes impliquées dans le processus de lecture.

Tout d'abord pour la phase de synthèse, des améliorations simples peuvent être imaginées pour s'affranchir du besoin humain. Pour l'instant, les synthétiseurs utilisés sont des synthétiseurs d'ADN, le nombre de places pour les monomères est limité à 9 au maximum. Or, pour la synthèse d'un polymère synthétisé avec l'alphabet à 8 symboles et contenant 4 blocs par exemple, 11 solutions différentes sont déjà nécessaires (8 pour les monomères codants, 1 pour l'espaceur, et 2 pour les marqueurs des blocs centraux). On pourrait alors imaginer un nouveau synthétiseur avec beaucoup plus de places disponibles. Ainsi, en ayant à disposition le bon nombre de position correspondant aux

nombres de monomères nécessaires pour la synthèse totale d'un polymère il ne serait plus nécessaire d'effectuer les synthèses en plusieurs étapes et l'intervention humaine ne se ferait qu'en début de synthèse pour écrire le message voulu sur le logiciel. Pour l'instant toutes les longues séquences impliquant plus de 9 monomères différents ont été faites en plusieurs fois. Ceci est chronophage et peut augmenter le risque d'erreurs de manipulation.

Ensuite pour optimiser encore plus le processus, il faudrait imaginer une machine complète comportant une partie synthèse et une partie lecture. Un tel appareil a déjà été fabriqué pour la synthèse d'ADN synthétique.[88] Ainsi avec ce genre d'instrument, on peut imaginer que l'expérimentateur n'a qu'à entrer le texte ou l'image qu'il souhaite encoder à l'échelle moléculaire dans le logiciel prévu à cet effet, puis 5 étapes seraient suivies:

1) Un programme changerait le message en flux de bits.
2) Le flux de bits serait traduit par le même programme en la suite de monomères correspondant puis les espaceur et marqueurs seraient ajoutés tous les 8 monomères.
3) La série de monomères serait envoyée au synthétiseur et la synthèse pourrait commencer.
4) Les séquences seraient purifiées.
5) Les séquences purifiées seraient analysées par un spectromètre de masse et analysées par le logiciel MS-DECODER qui permettrait de recouvrer le message initial.

Pour imaginer une utilisation industrielle, l'automatisation complète du procédé est indispensable. Même si certaines étapes comme la purification et l'analyse des séquences sont encore à automatisées complètement, l'automatisation complète pourrait être atteinte un jour.

Somme toute, au cours de cette thèse de nombreux avancements ont été faits concernant la synthèse de séquence de poly(phosphodiester)s à haute capacité de stockage et leur lecture par spectrométrie de masse a largement pu être simplifiée jusqu'à être automatisée. Ces nouveaux accomplissements laissent la voie pour de nouvelles perspectives dans le domaine des polymères à haute capacité de stockage. En conséquence, les séquences de poly(phosphodiester)s contenant de l'information seront très certainement de bonnes candidates pour une utilisation encore plus complexe dans le domaine du stockage d'information à l'échelle moléculaire. Même si ce domaine n'est pas le seul dans lequel ces séquences peuvent être efficaces. Elles pourraient tout à fait être utilisées comme codes-barres moléculaires ou comme séquences cryptées pour servir dans la lutte anti-contrefaçon.

Partie expérimentale

1. Techniques d'analyses et matériels utilisés

1.1. Résonnance magnétique nucléaire (RMN)

Tous les spectres RMN ont été enregistrés sur un spectromètre Bruker Avance 400 MHz équipé d'un aimant Ultrashield. Les déplacements chimiques sont reportés en parties par million (ppm) par rapport au signal du solvant résiduel (RMN 1H, $CDCl_3$: δ = 7.26 ppm ; RMN ^{13}C, $CDCl_3$: δ = 77.16 ppm ; RMN ^{1}H, $(CD_3)_2CO$: δ = 2.05 ppm ; RMN ^{13}C, $(CD_3)_2CO$: δ = 29.84 ppm.) Les spectres de RMN ^{1}H sont enregistrés à 400.13 MHz, RMN ^{13}C à 100.62 MHz et ceux de RMN ^{31}P à 161.96 MHz. Les solvants utilisés : chloroforme deutéré ($CDCl_3$, 99.8%) et acétone deutéré ($(CD_3)_2CO$, 99.8%) sont achetés chez Aldrich.

1.2. Spectrométrie de masse

Le groupe de Laurence Charles de l'université d'Aix-Marseille a effectué toutes les analyses de spectrométrie de masse.

Tous les échantillons sont dissous dans une solution de H_2O/MeCN (50/50, v/v) contenant 0,1% d'acide formique. Les solutions ainsi faites sont ensuite diluées dans une solution d'acétate d'ammonium dans le méthanol (3 mM) et introduites à un débit de 10 µL.min^{-1} dans la source d'ionisation par électronébuliseur (ESI) en mode négatif (tension capillaire : -2,27 kV) sous un flux de gaz désolvant (N_2) à 100 L.h^{-1} chauffé à 35°C. La tension appliquée est -10 V ou -20 V pour les expériences en MS et MS^2, mais celle-ci est ajustée entre -30 V et -50 V pour induire la fragmentation en source lors des analyses en MS^3. La dissociation induite par collision (CID) est réalisée dans un piège ionique (ion trap) utilisant l'argon comme le gaz collisionnant après la sélection du premier (en MS^2) ou du second (pseudo-MS^3) ion précurseur. Les analyses des données sont faites via le programme MassLynx 4.1 fourni par Waters.

1.3. Spectrophotomètre UV-Vis

Les mesures d'absorbances sont faites sur un spectrophotomètre UV-Vis Perkin Elmer Lambda 25 en utilisant le logiciel UV WinLab.

1.4. Chromatographie d'exclusion stérique

L'installation comprend :

- 1 ensemble DIONEX, série Ultimate 3000 (dégazeur, pompe, passeur d'échantillons)
- 4 colonnes Shodex OH-pak 30 cm en série (802.5HQ, 804HQ, 806HQ, 807HQ) et une précolonne (gamme de séparation : 500 à 10 000 000 g/mol)
- 1 réfractomètre différentiel OPTILAB rEX (Wyatt Techn.)
- 1 détecteur diffusion de lumière multi-angles DAWN HELEOS II (Wyatt Techn.)

Le débit utilisé est de 0,5 mL/min et le solvant est 60% eau (qualité Millipore) + 40% acétonitrile (qualité HPLC) + 0,1 M $NaNO_3$.

Les échantillons sont mis en solution dans 60% H_20 et 40% acétonitrile avec 0,1 M de $NaNO_3$ pendant 12h, puis ils sont filtrés sur un filtre Dynagard (Ester de cellulose) de 0,2 µm.

1.5. Equipement du laboratoire

1.5.1. HPLC

Les analyses d'HPLC sont faites avec l'appareil Agilent 1220 Infinity II LC avec un passeur automatique et ayant un détecteur UV à une seule longueur d'onde. La colonne utilisée est une colonne échangeuse d'ion DIONEX, DNA Pac PA 100, 4x250 mm. Les chromatogrammes sont enregistrés à λ=260nm. Conditions expérimentales : phase A :10 % MeCN, 20% 2M NH3 dans l'eau, phase B : 2.5M NaCl dans l'eau. Gradient d'élution : 0-3 min de 100% à 95% A, 3-20 min de 95% à 70% A, 20-25 min de 70% à 100% A. Débit : 1 $mL \cdot min^{-1}$.

1.5.2. Lyophilisateur

Le lyophilisateur utilisé est un FreeZone 2,5 de chez Labconco.

1.5.3. Synthétiseur d'ADN

Deux synthétiseurs d'ADN sont utilisés :

- Le synthétiseur Expedite, modèle Perseptive BioSystem 8900
- Le synthétiseur ABI, modèle Applied BioSystem 3900

Qu'importe le synthétiseur utilisé, les réactifs nécessaires sont tous commerciaux et utilisés sans aucun traitement supplémentaire. Tous les réactifs (commerciaux et les phosphoramidites synthétisées) sont placés sur le synthétiseur avec une asséchant dans tous les réservoirs. Avant de commencer une synthèse, les lignes sont lavées à l'acétonitrile puis purgées deux fois avec le réactif. Pour le synthétiseur ABI, à chaque changement de réactif, il est nécessaire de refaire une calibration. Ensuite le support solide est connecté au synthétiseur et le cycle itératif automatique débute.

1.5.4. Chromatographies

Chromatographie sur couche mince (CCM) : Les chromatographies sur couche mince sont effectuées sur des plaques de silice (gel de silice MERCK 60F 254, épaisseur : 0.25mm). Les révélations sont faites avec une solution d'acide phosphomolybdique dans l'éthanol ou une solution de molybdate d'ammonium cérique dans l'eau.

Chromatographie liquide sur colonne : Les purifications par chromatographie sur colonne sont réalisées sur gel de silice en éluant sous pression d'air (Sigma, taille des pores 60 Å, 230 - 400 mesh).

2. Réactifs et solvants

Acétate d'éthyle (EtOAc, Carlo Erba), acide acétique (AcOH, Fisher Chemical, 99.7%), acide chlorhydrique (HCl, Sigma-Aldrich, 37%), 2-butyle-2-éthyle-1,3-propanediol (>98.0%, TCI), 2-tert-butylepropane-1,3-diol (98%, Alfa Aesar), 2,2-di-n-butyle-1,3-propanediol (96%, Alfa Aesar), chlorure de 4,4'-dimethoxytrityl (DMTr-Cl, ChemGenes), chlorure de sodium (pour solution de saumure, ESCO), 2-cyanoéthyle-N,N-diisopropylchlorophosphoramidite (97 %, ABCR), cyclohexane (Carlo Erba), dichlorométhane anhydre (99.8 %, anh. DCM, cont. amylene comme stabiliseur, Sigma-Aldrich), N,N-diisopropyléthylamine (>99.0

%, DIPEA, TCI), 2-éthyle-2-méthyle-1,3-propandiol (>97%, TCI), 2,2-diéthyle-1,3-propanediol (>98.0%, TCI), hexane (Carlo Erba), hydroxyde de sodium (NaOH, VWR, 99%), hydroxyde de potassium (KOH, VWR, 87%), méthanol (MeOH, Carlo Erba, 99.9%), 2-méthyl-1,3-propanediol (99%, TCI), n-pentane (Carlo Erba, 95%), 2-n-pentylpropane-1,3-diol (97%, Alfa Aesar), pyridine anhydre (99.8 %, anh. pyr, in Sure/Seal™, Sigma-Aldrich), sulfate de sodium anhydre (99.6 %, VWR), et triéthylamine (97 %, Et3N, Acros Organics) ont été utilisé sans aucune purification supplémentaire. Le THF anhydre a été obtenu en utilisant une station pour sécher les solvants GT S100. Le bromure de cuivre (I) (CuBr, Sigma-Aldrich, 98%) a été lavé à l'acide acétique glacé dans le but d'éliminer toutes les espèces solubles oxydées, filtré, lavé à l'éthanol, et séché.

Toutes les réactions sensibles à l'air ont été menées sous argon. Les monomères **M1**, **M3** et l'espaceur **E**, présentés dans le **chapitre II**, ont été préparés comme décrit précédemment et stockés à -18°C.[23]

Réactifs pour la synthèse phosphoramidite automatisée : acétonitrile anhydre (phosphoramidite diluent & dry washings, ChemGenes), acétonitrile (≥99.9 %, lavages, Roth), réactifs d'activation (0.25 M 5-éthylthio tétrazole dans MeCN, ChemGenes), Cap A (anhydride acétique /pyridine/THF, ChemGenes), Cap B (10 % *N*-méthylimidazole dans THF, ChemGenes), réactif de clivage du DMT (3 w% TCA dans DCM, Roth), dessiccant (petit, 10 - 15 mL, ChemGenes), dT-CPG 1000 (1 µmol comme support solide, taille de pores 1000 Å, Glen Research), dT-support polystyrène(1 µmol pour ABI) cartouche de purification ADN glen-pak (10 nmole - 1.0 µmole, Glen Reasearch) et solution oxydante (0.02 M iode/pyridine/H_2O/THF, ChemGenes) ont été utilisés sans aucune purification supplémentaire. Les nucléotides dA-CE phosphoramidite, Ac-dC-CE phosphoramidite, dG-CE phosphoramidite, et 2'-F-G-CE phosphoramidite (Glen Research) ont également été utilisés sans purification et stockés au congélateur à -18°C.

3. Partie expérimentale du chapitre II

3.1. Synthèses des molécules présentées dans le chapitre II

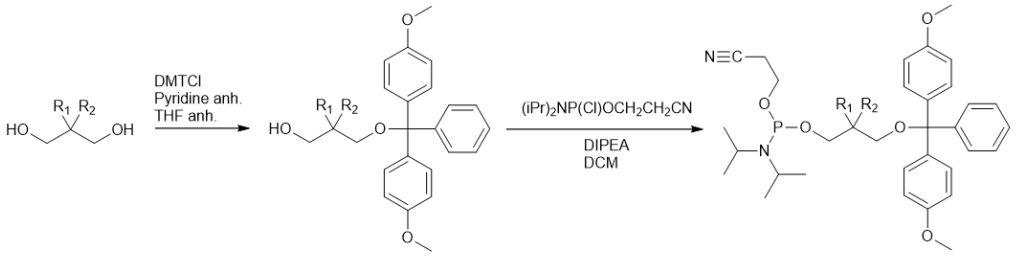
3.1.1. Synthèse des monomères linéaires

Tous les monomères présentés dans le chapitre II, sont synthétisés en suivant la même procédure, adaptée de celle préalablement décrite et utilisée dans le groupe.[23]

La synthèse des monomères linéaires a été fait avec l'aide de Laurence Oswald, ingénieure d'étude travaillant dans l'équipe. Elle a effectué les synthèses en suivant la même procédure que celle utilisée pour les monomères à chaînes latérales qui est détaillée dans le paragraphe suivant.

3.1.2. Synthèse des monomères à chaînes latérales

A nouveau, ces monomères sont synthétisés en utilisant la procédure utilisée dans le groupe. Le schéma de synthèse est donné en Figure SI-II. 1.



Intermédiaires I

(**I1**) $R_1 = R_2 = H$
(**I2**) $R_1 = H$, $R_2 = CH_3$
(**I3**) $R_1 = R_2 = CH_3$
(**I4-*iso***) $R_1 = H$, $R_2 = C(CH_3)_3$
(**I4**) $R_1 = CH_2CH_3$, $R_2 = CH_3$
(**I5**) $R_1 = R_2 = CH_2CH_3$
(**I6**) $R_1 = H$, $R_2 = (CH_2)_4CH_3$
(**I7**) $R_1 = CH_2CH_3$, $R_2 = (CH_2)_3CH_3$
(**I8**) $R_1 = R_2 = (CH_2)_3CH_3$

Monomères M

(**M1**) $R_1 = R_2 = H$
(**M2**) $R_1 = H$, $R_2 = CH_3$
(**M3**) $R_1 = R_2 = CH_3$
(**M4-*iso***) $R_1 = H$, $R_2 = C(CH_3)_3$
(**M4**) $R_1 = CH_2CH_3$, $R_2 = CH_3$
(**M5**) $R_1 = R_2 = CH_2CH_3$
(**M6**) $R_1 = H$, $R_2 = (CH_2)_4CH_3$
(**M7**) $R_1 = CH_2CH_3$, $R_2 = (CH_2)_3CH_3$
(**M8**) $R_1 = R_2 = (CH_2)_3CH_3$

Figure SI-II. 1 : Synthèse des monomères M1 à M8 et des intermédiaires I1 à I8.

<u>Première étape :</u> Protection de l'alcool par le groupement diméthoxytrityle (DMT).

Dans un ballon monocol de 250 mL, le diol choisi (2 g, 1 eq.) est introduit. Un bouchon à jupe en caoutchouc ferme le ballon, et 3 cycles vide-argon permettent d'assurer une atmosphère inerte. La pyridine anhydre et du THF anhydre sont introduits à travers le bouchon grâce à une seringue, préalablement conditionnée sous argon. La poudre de chlorure de 4,4'-dimethoxytrityl (DMTCl) (1,1 eq.) est ajoutée rapidement en ouvrant le bouchon pour réagir avec le diol. Le DMTCl est ajouté en trois proportions égales en attendant une heure entre chaque ajout. Après les trois additions, le mélange est

agité toute la nuit. Ensuite, du méthanol est ajouté pour stopper la réaction et le mélange est concentré à l'évaporateur rotatif. L'huile visqueuse obtenue est alors dissous dans l'acétate d'éthyle puis le tout est transvasé dans une ampoule à décanter et est lavé avec une solution saturée de bicarbonate de sodium ($NaHCO_3$). La phase aqueuse est alors lavée avec de l'acétate d'éthyle (deux fois 20mL), les phases organiques sont rassemblées et lavées à l'eau (une fois 20 mL), puis à la saumure (une fois 20 mL) et finalement séchée sur du sulfate de sodium anhydre (Na_2SO_4) et évaporé à l'évaporateur rotatif. Le produit brut est purifié sur colonne de silice en utilisant un éluant généralement au cyclohexane/acétate d'éthyle (60/40, mais peut différer en fonction du diol utilisé) avec 1 % de triéthyleamine. Les composés sont isolés sous la forme d'une poudre blanche voire d'une huile très visqueuse en fonction des substituants présent sur le diol de départ, avec un rendement entre 30 et 60 % en fonction du diol utilisé. CCM (Cyclohexane : Acétate d'éthyle : Triéthyleamine, 60 :40 :1 v/v/v).

Deuxième étape : Ajout de la phosphoramidite

L'huile (1 eq. 3 g) obtenue lors de l'étape 1 est mise sous atmosphère inerte grâce à trois cycles vide-argon. Ensuite celle-ci est dissoute dans le dichlorométhane anhydre avec de la *N,N*-diisopropyléthylamine (DIPEA) (4 eq.). Le mélange est refroidi à 0 degré dans un bain d'eau glacée. La 2-cyanoethyl-*N,N*-diisopropylchlorophosphoramidite (1 eq.) est solubilisée dans quelques millilitres de DCM anhydre et est ensuite ajouté goutte à goutte sous une agitation continue. Le mélange est agité pour 1 heure à température ambiante, puis il est concentré à l'évaporateur rotatif. Le produit brut est alors dissous dans l'acétate d'éthyle puis directement purifié sur colonne de silice en éluant avec du cyclohexane et de l'acétate d'éthyle avec 1 % de triéthylamine (50-50). Une huile transparente plus ou moins visqueuse est obtenue avec un rendement autour de 95%. CCM (Cyclohexane : acétate d'éthyle : triéthyleamine, 50 :50 :1 v/v/v).
Le Tableau SI-II. 1 suivant donne les valeurs exactes utilisées pour la synthèse des monomères M1, M2, M3, M4, M5, M6, M7, M8 et M4-*iso*.

Tableau SI-II. 1 : Liste des composés synthétisé pour le chapitre II.

Composé	Diol utilisé				DMTCl				Diol monoprotégé DMT			Phosphoramidite			
	Nom	Masse (g)	Quantité de matière (mmol)	Eq.	Masse (g)	Quantité de matière (mmol)	Eq.	Rdt (%)	Masse (g)	Quantité de matière (mmol)	Eq.	Masse (g)	Quantité de matière (mmol)	Eq.	Rdt (%)
M1	1,3-Propanediol, 99%	2	26,3	1	9,8	28,9	1,1	58	3	7,9	1	2	8,7	1,1	95
M2	2-Methyl-1,3-propanediol, 99%	2	22,2	1	8,3	24,4	1,1	55	3	7,6	1	1,99	8,4	1,1	98
M3	2,2-Dimethyl-1,3-propanediol, 99%	2	19,2	1	7,2	21,1	1,1	61	3	7,4	1	1,9	8,1	1,1	86
M4	2-Ethyl-2-methyl-1,3-propanediol, > 97%	2	16,9	1	6,3	18,6	1,1	60	3	7,1	1	1,85	7,8	1,1	84
M5	2,2-Diethyl-1,3-propanediol, >98,0%	2	15,1	1	5,6	16,6	1,1	59	3	6,9	1	1,8	7,6	1,1	91,5
M6	2-n-Pentylpropane-1,3-diol, 97%	2	13,7	1	5	15,0	1,1	67	3	6,7	1	1,75	7,3	1,1	94,5

M7	2-Butyl-2-ethyl-1,3-propanediol, >98,0%	2	12,5	1	4,6	13,7	1,1	66	3	6,5	1	1,7	7,1	1,1	77
M8	2,2-Di-n-butyl-1,3-propanediol, 96%	2	10,6	1	3,9	11,7	1,1	54	3	6,1	1	1,6	6,7	1,1	86
M4-*iso*	2-tert-Butylpropane-1,3-diol, 98%	2	15,1	1	5,6	16,6	1,1	49	3	6,9	1	1,8	7,6	1,1	88

3.1.3. Synthèse de la molécule espaceur

La molécule espaceur utilisée dans le chapitre II a été synthétisée en suivant le protocole indiqué dans le travaux effectués au préalable dans l'équipe.[25]

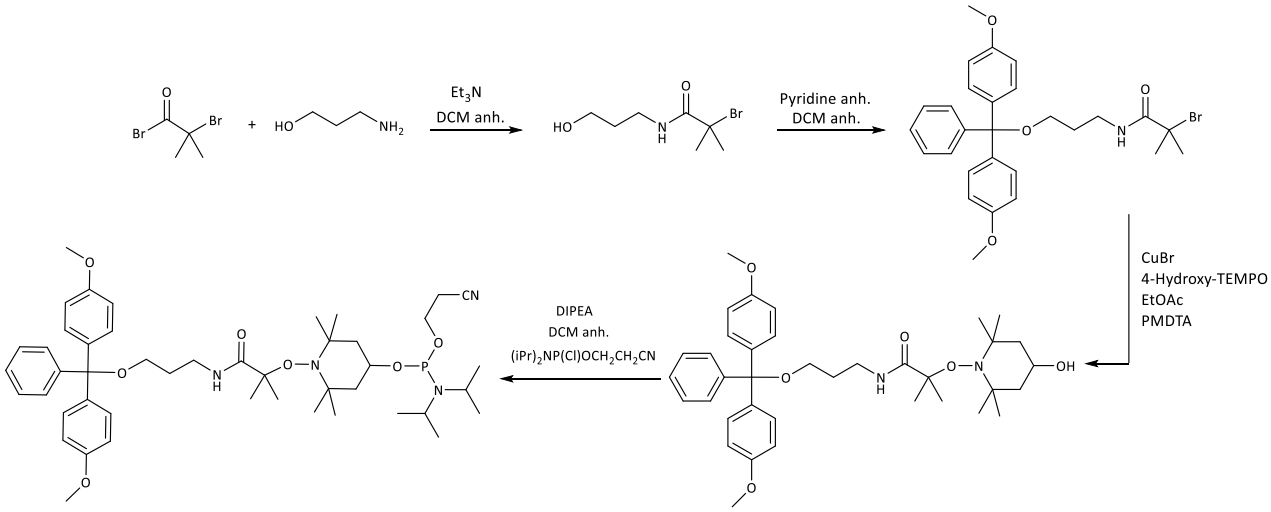


Figure SI-II. 2 : Schéma réactionnel de la molécule espaceur.

Cette synthèse se fait en quatre étapes décrites ainsi :

Etape 1 :

Dans un ballon monocol, le 3-Amino-1-propanol (3.75 g, 50 mmol) et la triéthylamine (10.1 g, 100 mmol) sont mélangés dans 80 mL de dichlorométhane anhydre. La solution est ensuite refroidie dans un bain d'eau glacée pour permettre l'ajout goutte à goutte d'une solution de bromure de 2-bromoisobutyryle (22.9 g, 100 mmol) solubilisée dans 40 mL de DCM anhydre. La réaction est mélangée pendant 1h30 à température ambiante. Celle-ci est ensuite extraite avec de l'eau (2x50 mL) puis lavée avec une solution saturée de $NaHCO_3$ et finalement évaporée à l'évaporateur rotatif. Ensuite, une solution contenant 5 mL de triéthylamine dans 40 mL de méthanol est ajoutée et le tout est laissé sous agitation pendant 1 h. La solution est concentrée puis finalement purifiée sur colonne de silice en utilisant un éluant à l'acétate d'éthyle. (Rf = 0,49) Une huile incolore est récupérée avec environ 80% de rendement.

Etape 2 :

L'huile obtenue lors de l'étape 1 (0.7 g, 3.13 mmol) est coévaporée avec de la pyridine anhydre pour enlever les dernières traces d'eau. La réaction est effectuée sous atmosphère inerte ce qui est assurée par trois cycles de vide/argon. L'huile est ensuite solubilisée dans 4 mL de pyridine anhydre et 6 mL de DCM anhydre. Une solution de chlorure de 4,4'-diméthoxytrityle (1.0714 g, 3.16 mmol) dans du DCM anhydre est ensuite ajouté à la seringue. Le mélange est agité toute la nuit. Ensuite, 2.5 mL de méthanol sont ajoutés pour stopper la réaction et le tout est évaporé à l'évaporateur rotatif. Le produit obtenu est dissous dans de l'acétate d'éthyle et lavé avec une solution saturée de $NaHCO_3$. L'acétate d'éthyle est utilisé pour extraire la phase aqueuse et ensuite les phases organiques combinées sont lavées avec de l'eau et finalement séchées sur du Na_2SO_4 anhydre et concentrées. Le produit final est purifié sur colonne de silice en utilisant un éluant acétate d'éthyle / cyclohéxane 1/4 avec 1% de triéthylamine. (Rf= 0,26) Un solide blanc est récupéré avec un rendement de 80%.

Etape 3 :

Dans un ballon monocol et sous argon, le solide récupéré (0.7 g, 1.33 mmol), du CuBr (0.38 g, 2.6 mmol) et de l'hydroxy-TEMPO (0.22 g, 1.3 mmol) sont solubilisés dans de l'acétate d'éthyle. Ensuite la N,N,N',N'',N''-pentaméthyle-diéthyl-ène-triamine (PMDTA) (0.46 g, 2.6 mmol) est ajouté goutte à goutte et le mélange est agité toute la nuit. La solution verte obtenue est filtrée et le gâteau du filtre est lavé avec 100 mL d'acétate d'éthyle. Ensuite, le filtrat récupéré est lavé à l'eau (1x20 mL), avec une solution saturée de $NaHCO_3$ (2x20 mL) puis une solution de saumure (1x20 mL). Les phases organiques combinées sont séchées sur du Na_2SO_4 anhydre et concentrées. Le produit obtenu est finalement purifié sur colonne de silice en utilisant un éluant acétate d'éthyle / cyclohéxane 1/4 avec 1% de triéthylamine. (Rf = 0,26). Une poudre blanche est récupérée avec un rendement de 80%.

Etape 4 :

La poudre blanche obtenue en étape 3 (0.150 g, 0.24 mmol) est engagée dans trois cycle vide/argon pour permettre une réaction sous atmosphère inerte. Ensuite 10 mL de DCM anhydre et 0.25 mL de DIPEA (1.4 mmol) sont ajoutés successivement sous argon et la solution est refroidie dans un bain d'eau glacée. La 2-cyanoéthyle-*N,N*-diisopropylchlorophosphoramidite (0.063 g, 0.26 mmol) est solubilisé dans du DCM anhydre et cette solution est ajoutée goutte à goutte au mélange. Celui-ci est remis à température ambiante et mélangé pendant 1h. La solution est ensuite concentrée à l'évaporateur rotatif puis diluée dans l'acétate d'éthyle et directement purifiée sur colonne de silice en utilisant un éluant acétate d'éthyle / cyclohéxane 1/1 avec 1% de triéthylamine. (Rf = 0,74). Une poudre blanche est récupérée avec un rendement de 95%.

3.2. Caractérisation des molécules présentées dans le chapitre II

Tous les monomères sont caractérisés par RMN ^{1}H, RMN ^{13}C et par RMN ^{31}P. Toutes les caractérisations sont décrites ci-après.

3.2.1. Caractérisation des monomères linéaires

- **<u>Monomère C4</u>**

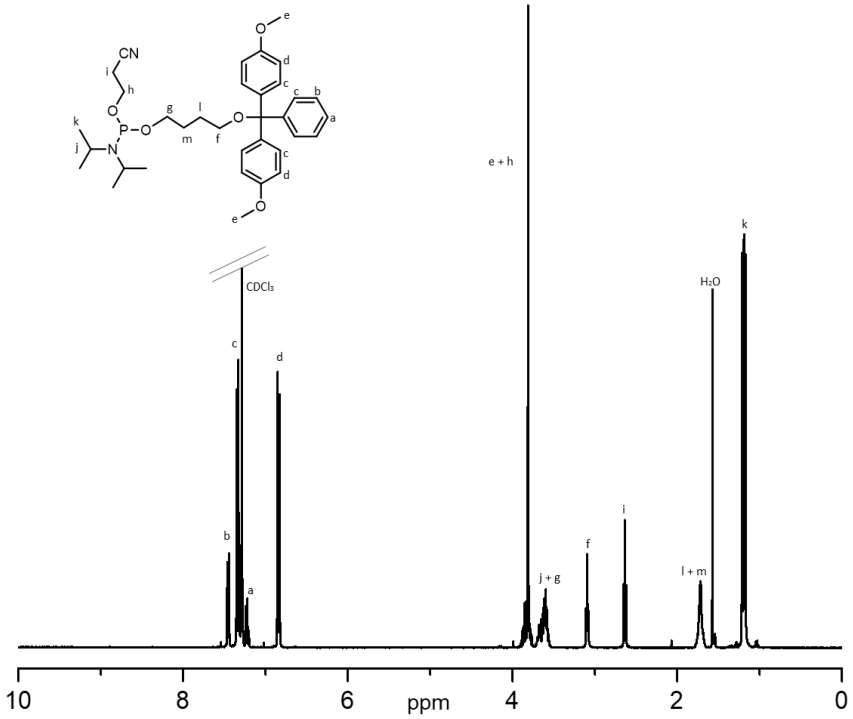


Figure SI-II. 3 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C4.

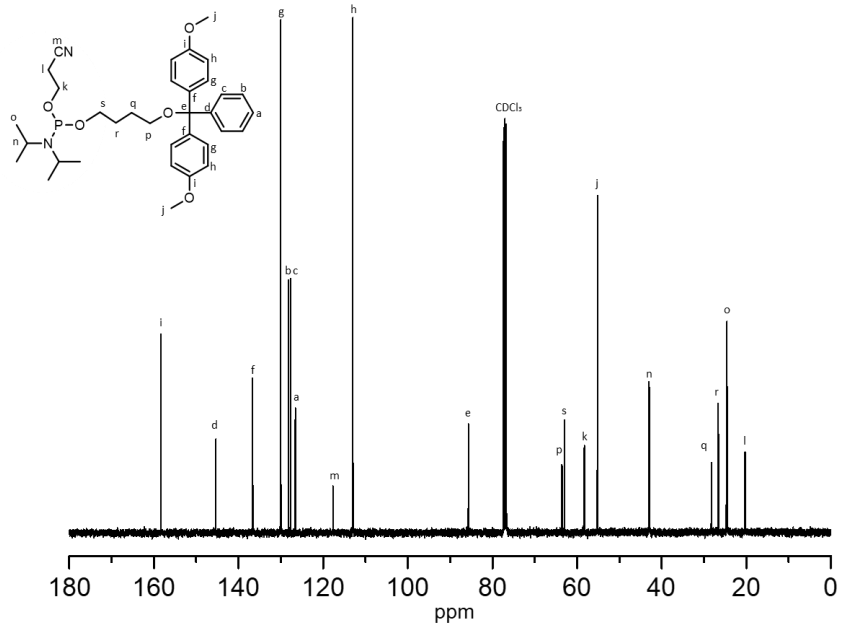


Figure SI-II. 4 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C4.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.28.

- **Monomère C5**

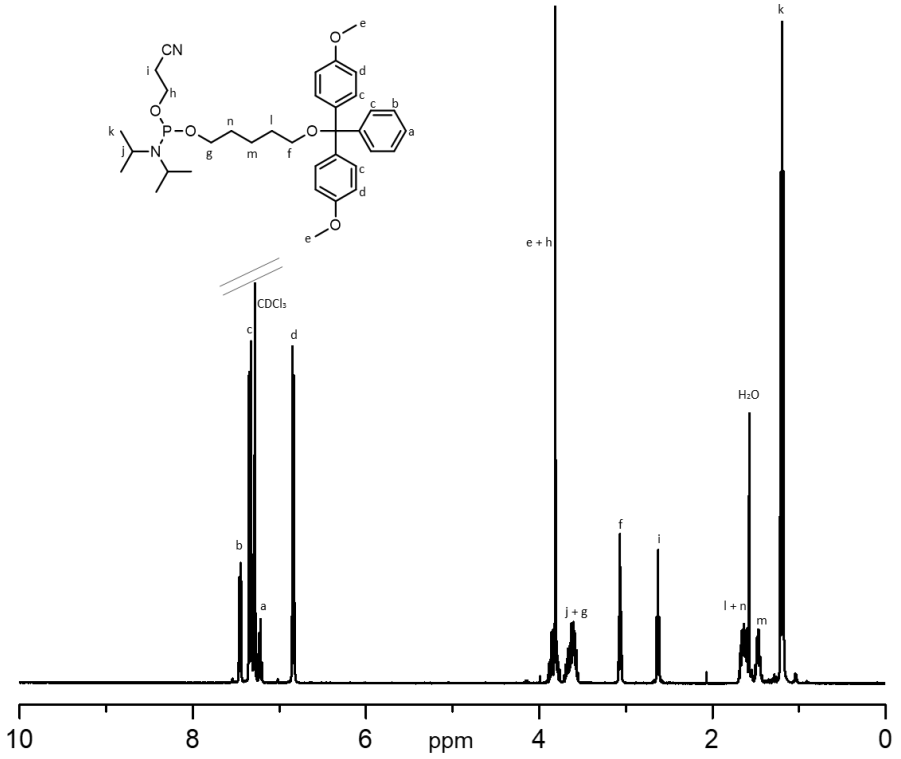


Figure SI-II. 5 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C5.

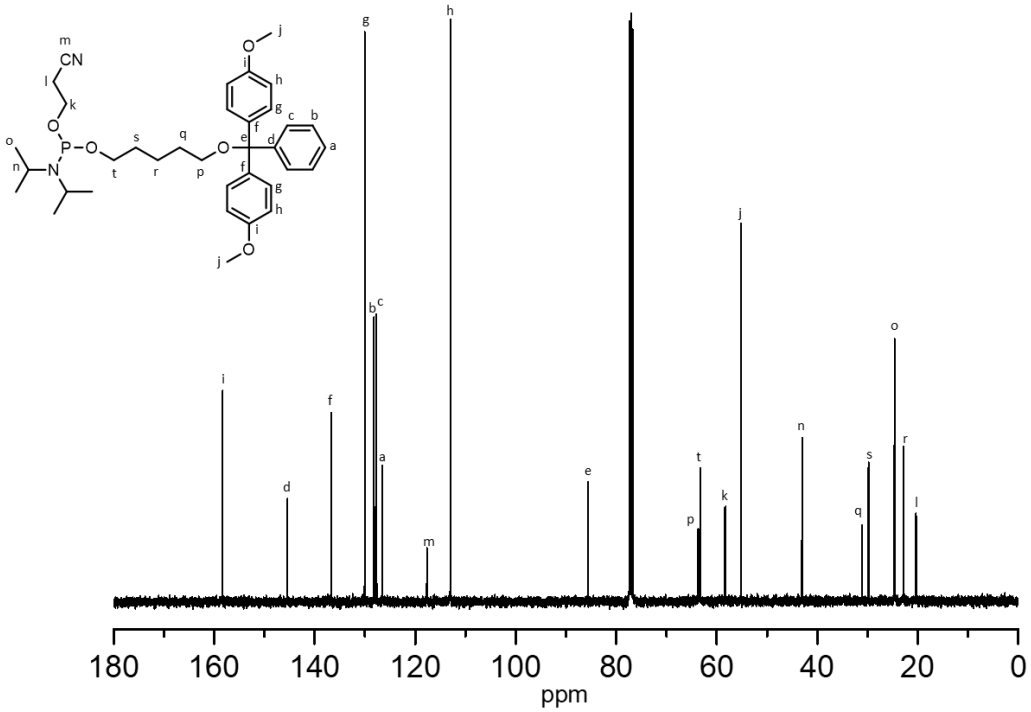


Figure SI-II. 6 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C5.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.28.

- **Monomère C6**

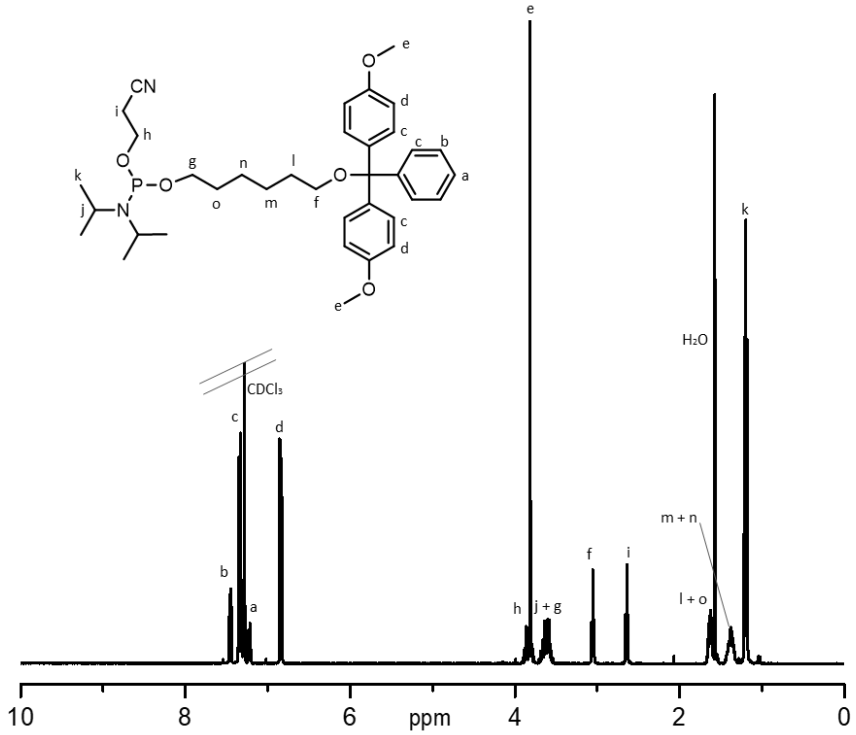


Figure SI-II. 7 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C6.

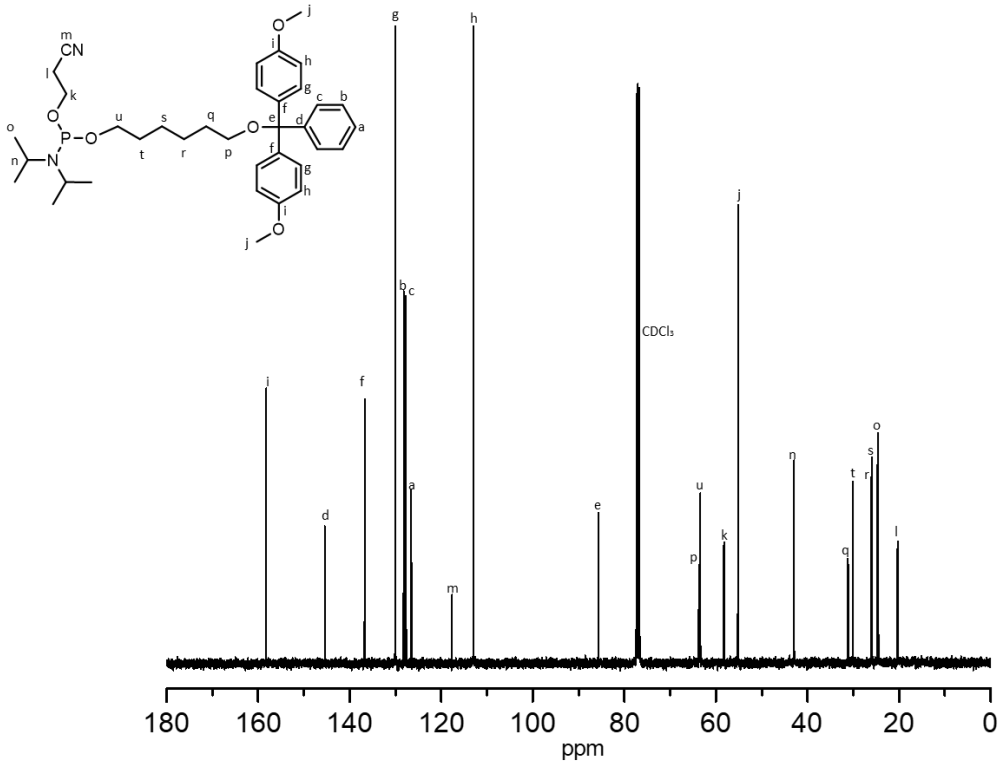


Figure SI-II. 8 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C6.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.25.

- **Monomère C7**

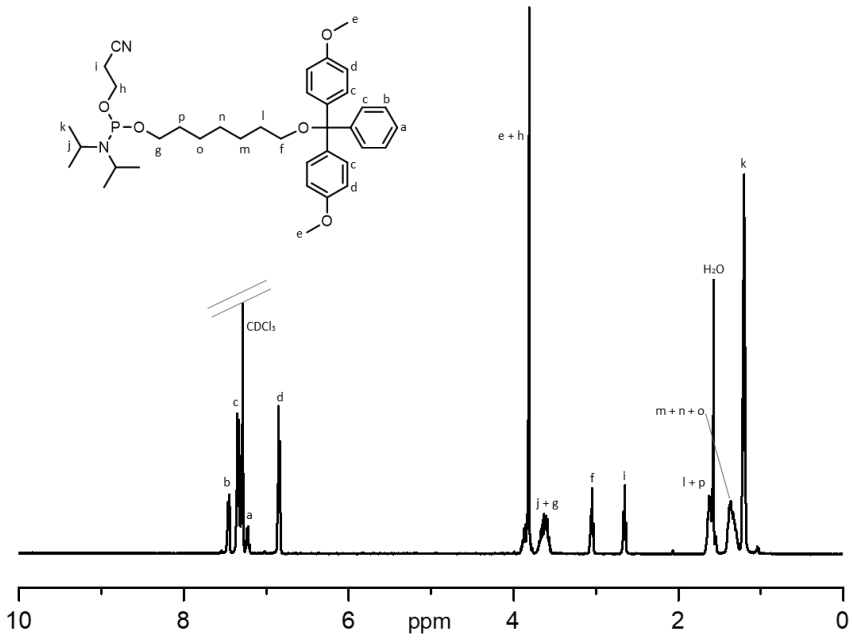


Figure SI-II. 9 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C7.

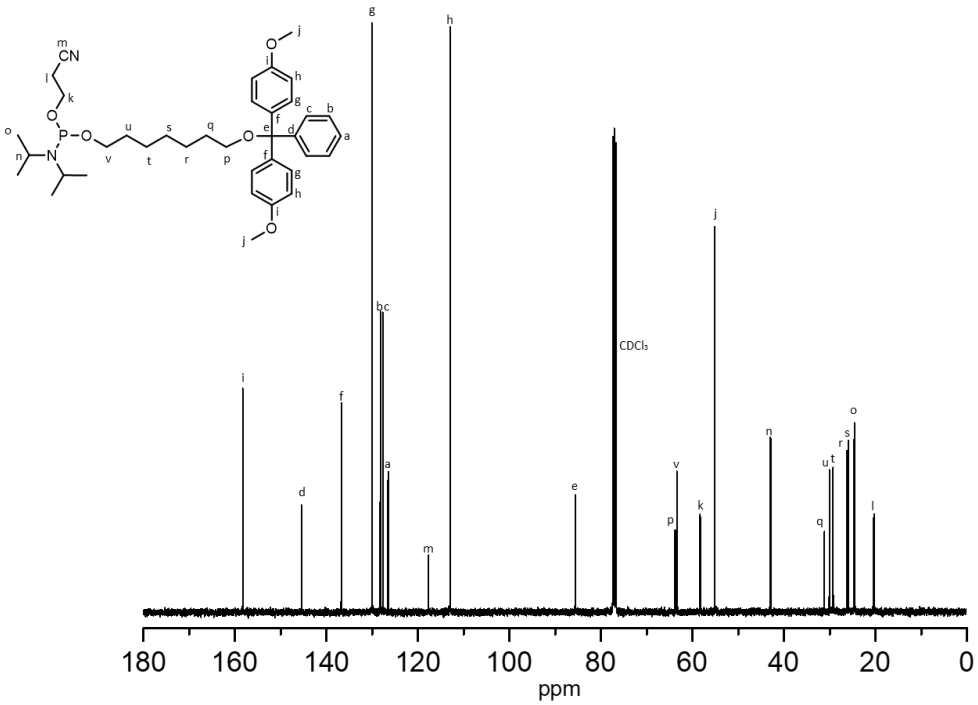


Figure SI-II. 10 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C7.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.23.

- **Monomère C8**

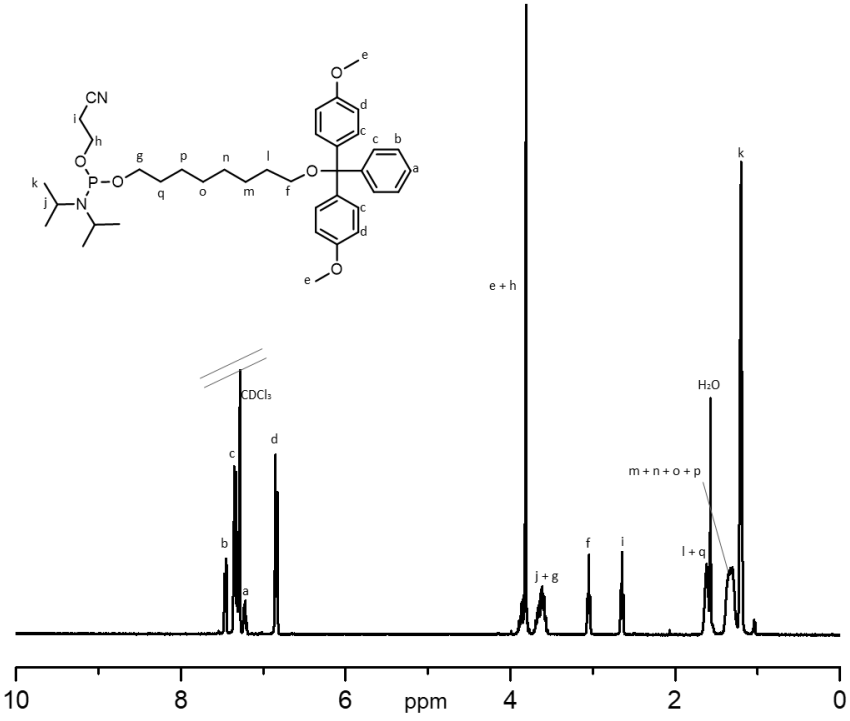


Figure SI-II. 11 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C8.

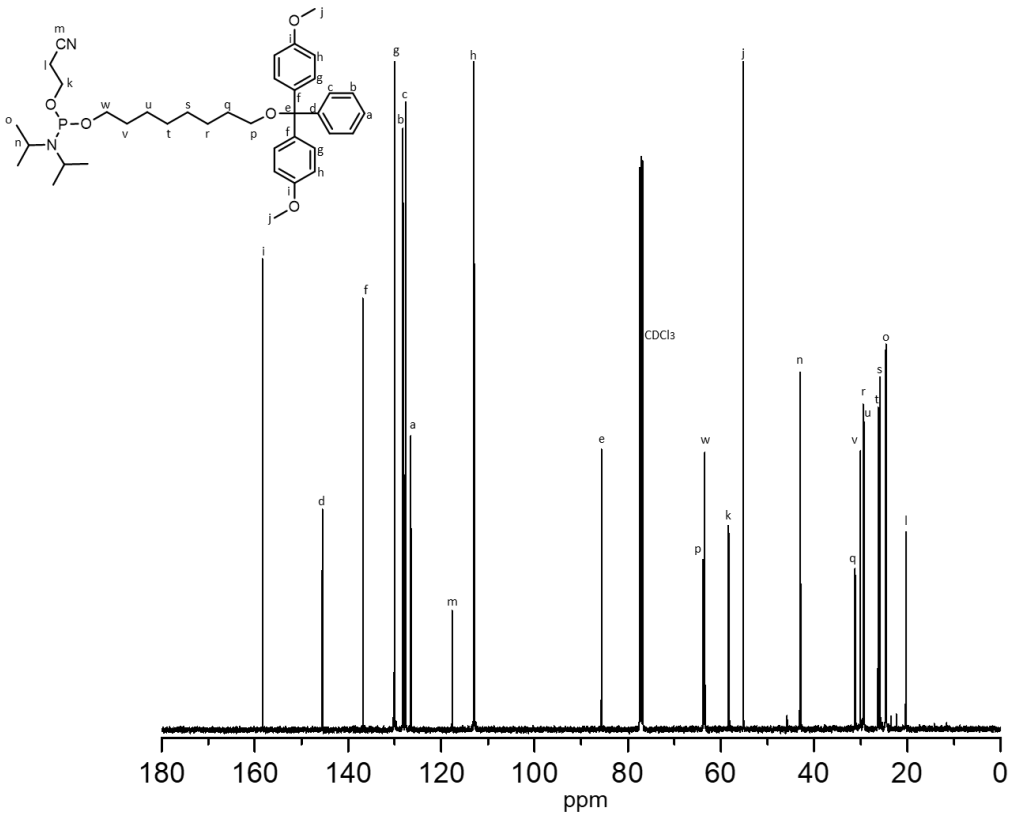


Figure SI-II. 12 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C9.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.25.

- **Monomère C9**

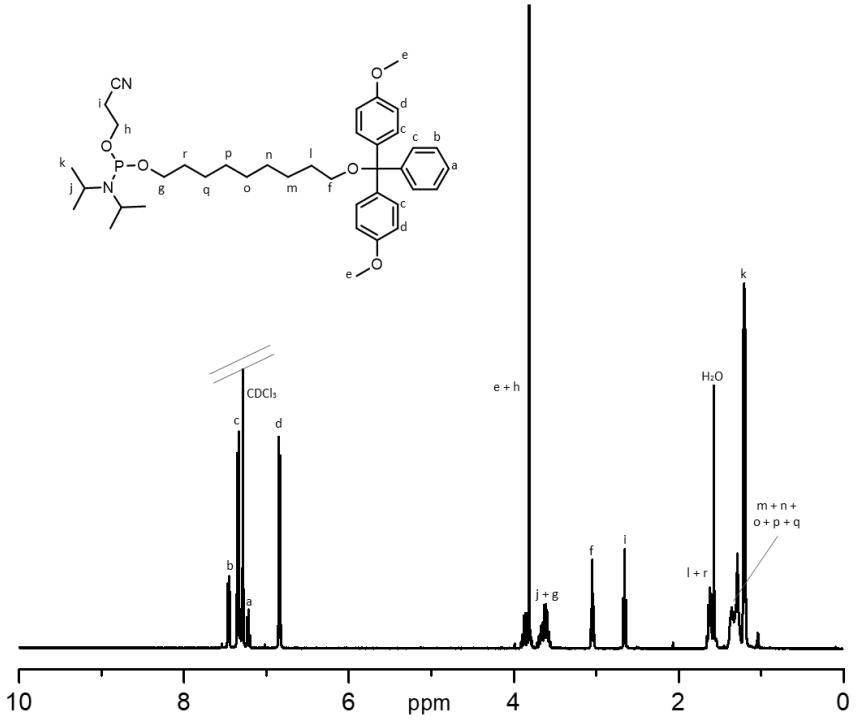


Figure SI-II. 13 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C9.

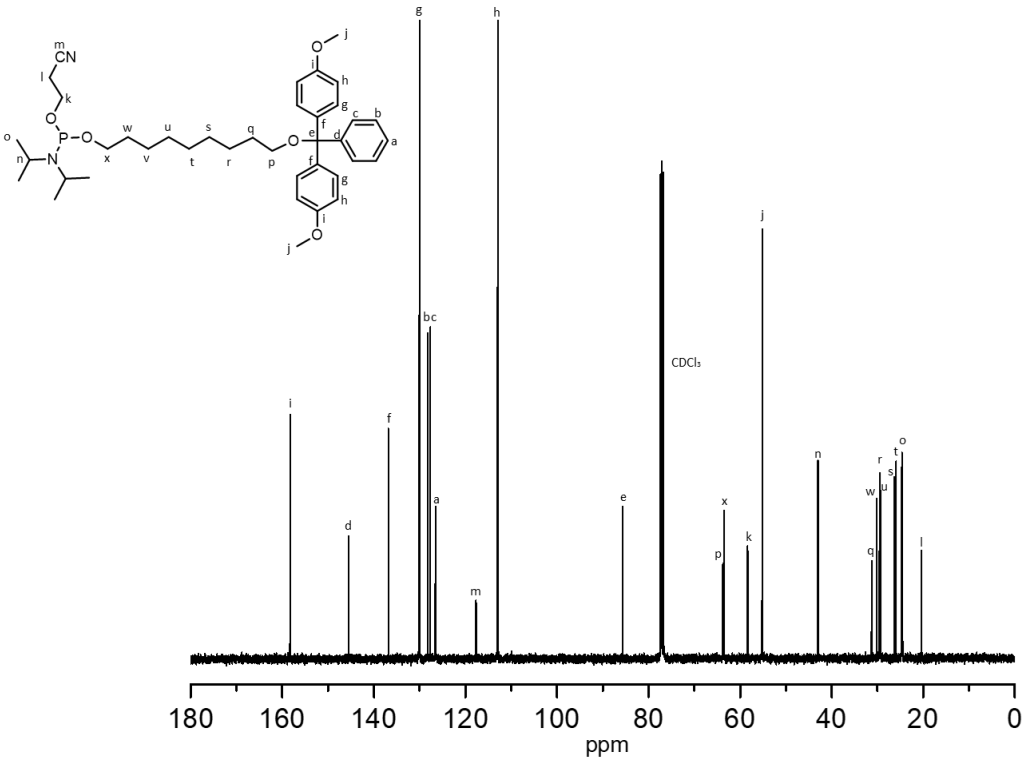


Figure SI-II. 14 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C9.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.22.

- **Monomère C10**

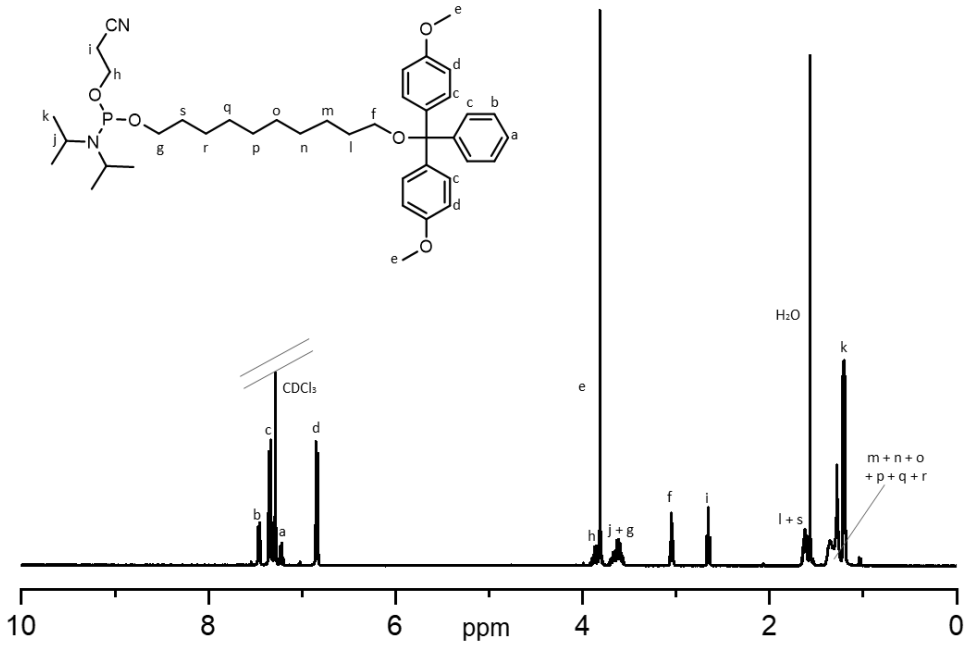


Figure SI-II. 15 : Spectre RMN ^{1}H ($CDCl_3$) du monomère C10.

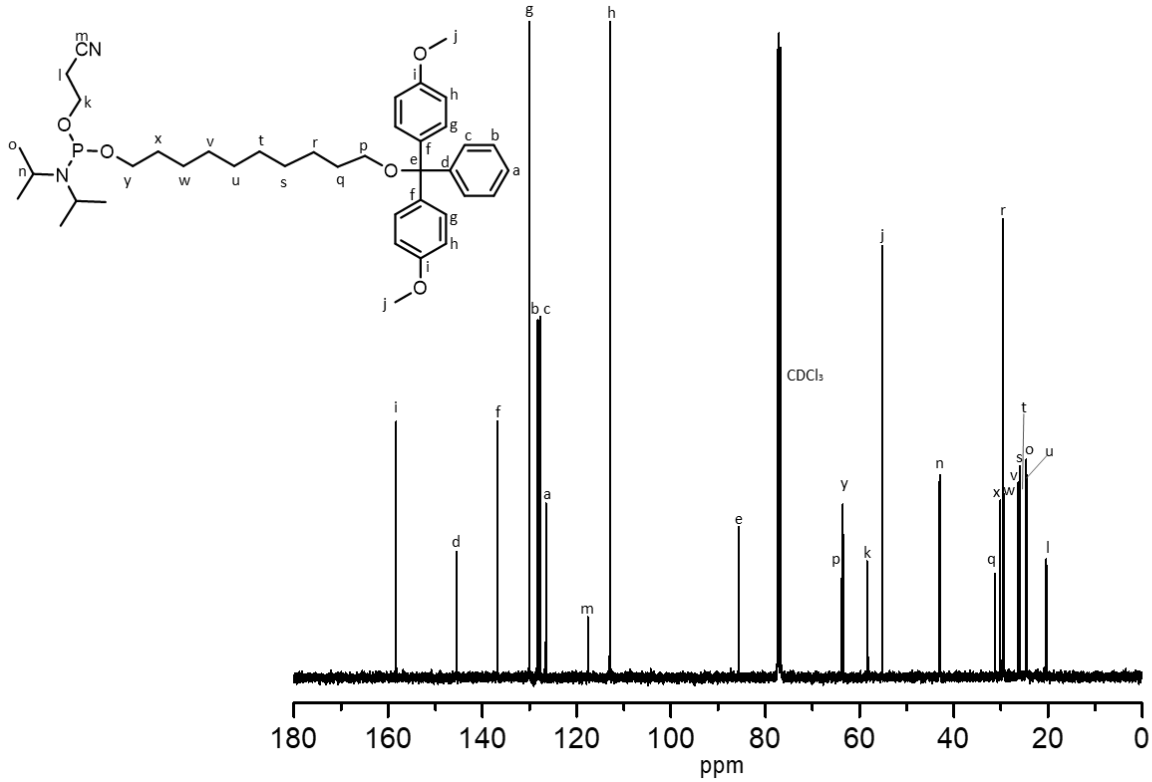


Figure SI-II. 16 : Spectre RMN ^{13}C ($CDCl_3$) du monomère C10.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.25.

3.2.2. Caractérisation des monomères à chaîne latérales

- **Monomère M1**

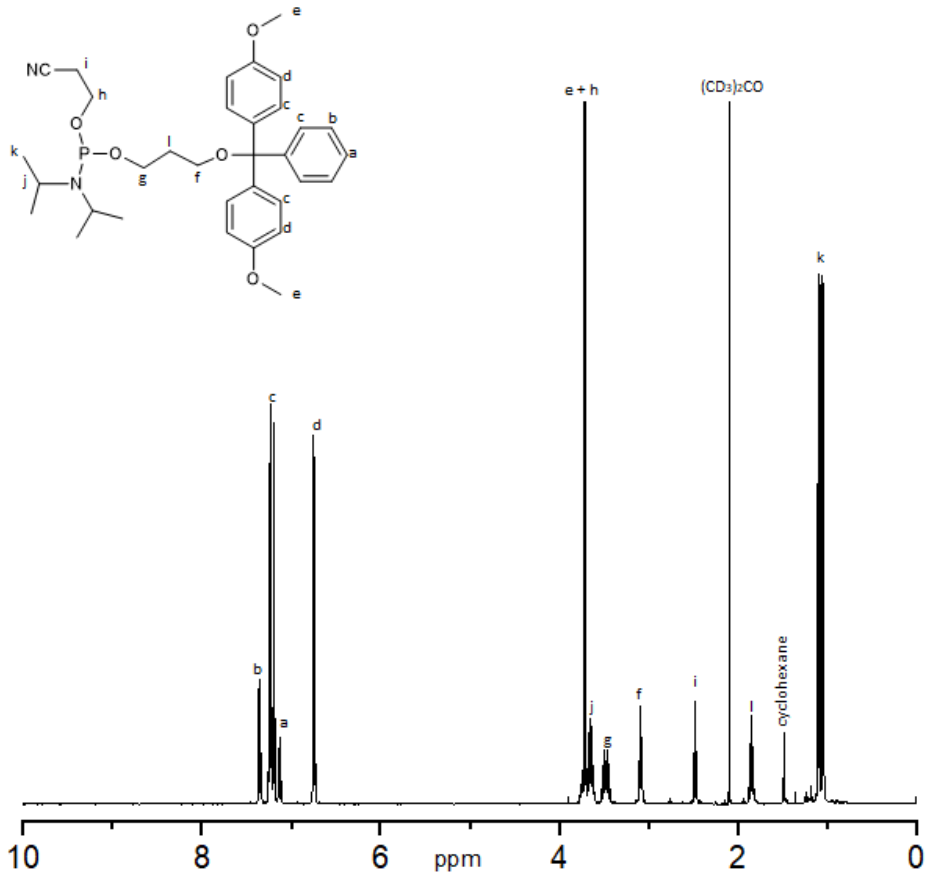


Figure SI-II. 17 : Spectre RMN ^{1}H ($(CD_3)_2CO$) du monomère M1.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.30 (d, J = 8.13 Hz, 2H, b), 7.16 (m, 6H, c), 7.07 (m, 1H, a), 6.69 (d, J =10.17 Hz, 4H, d), 3.66(s, 6H, e), 3.64-3.73 (m, 2H, h), 3.55-3.63 (m, 2H, j), 3.38-3.46 (m, 2H, g), 3.02-3.05(m, 2H, f), 2.43 (m, 2H, i), 1.79 (m, 1H, l), 1.04-1.00 (dd, 12H, k).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.4, 145.3, 136.5, 130.04, 128.19, 127.68, 126.6, 118.9, 112.9, 85.81, 60.19, 58.4, 58.2, 55.2, 43.06, 31.9, 24.6, 20.27.

RMN ^{31}P ($CDCl_3$, δ, ppm): 147.7

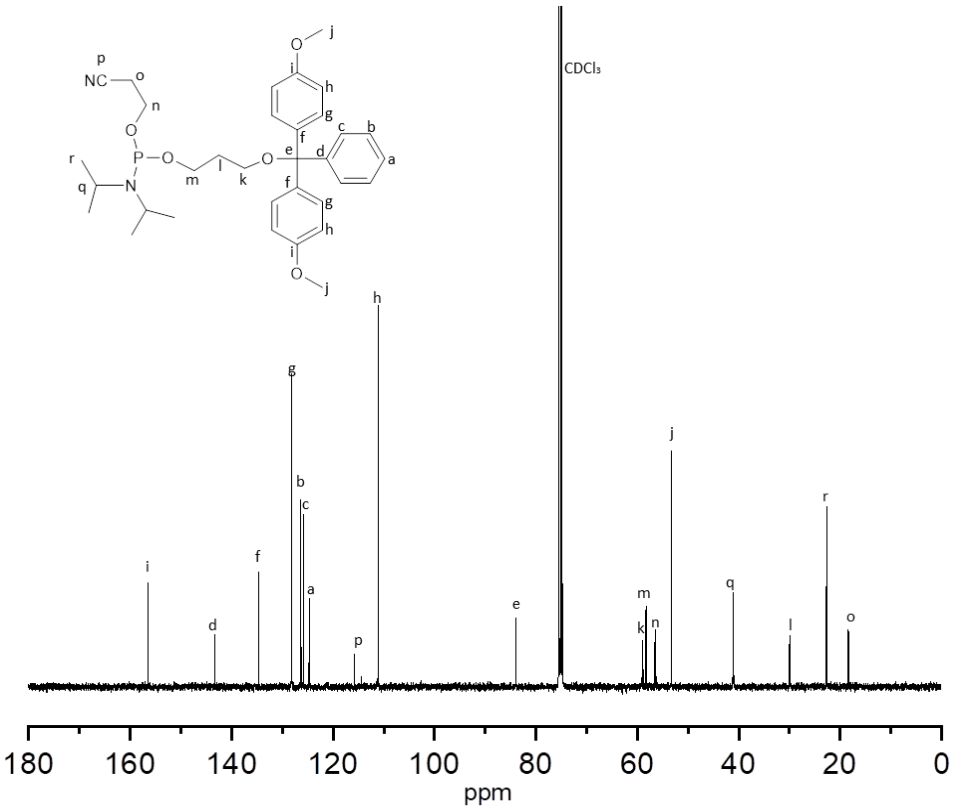


Figure SI-II. 18 : Spectre RMN ^{13}C ($CDCl_3$) du monomère M1.

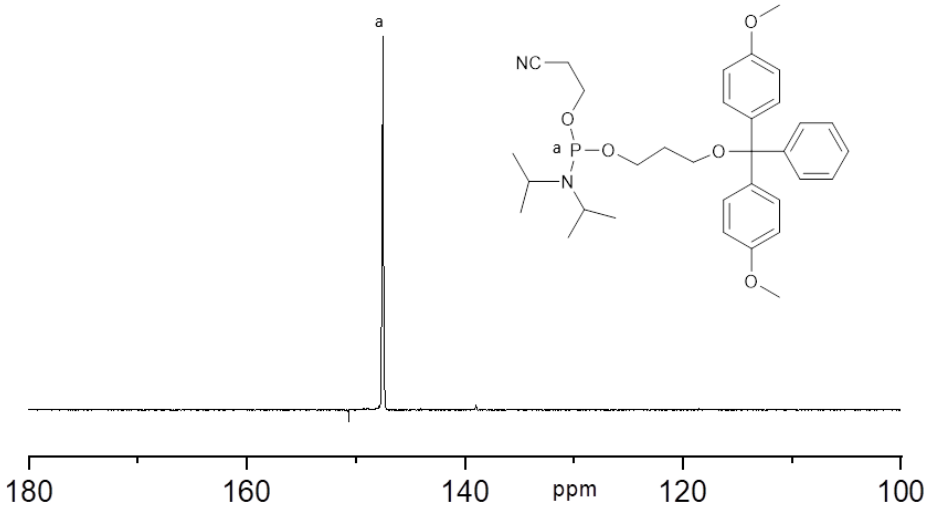


Figure SI-II. 19: Spectre RMN ^{31}P ($CDCl_3$) du monomère M1.

- **Monomère M2**

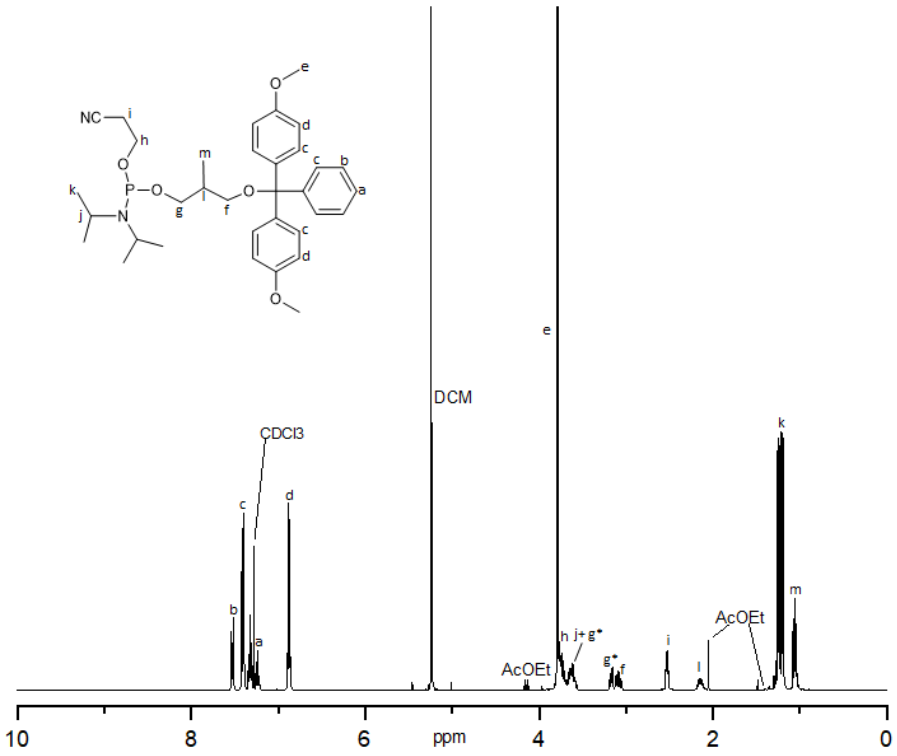


Figure SI-II. 20 : Spectre RMN ^{1}H ($CDCl_3$) du monomère M2.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.52 (d, J = 8.9 Hz, 2H, b), 7.32 (m, 6H, c), 7.23 (m, 1H, a), 6.87 (d, J =9.22 Hz, 4H, d), 3.79(s, 6H, e), 3.70-3.77 (m, 2H, h), 3.62 (m, 3H, h et un proton de g), 3.14-3.19 (m, 1H, deuxième proton de g), 3.05-3.13(m, 2H, f), 2.53 (m, 2H, i), 2.15 (m, 1H, l), 1.19-1.25 (dd, 12H, k), 1.06 (t, 3H, m).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.49,145.52, 136.57,130.16, 128.27, 127.71, 126.62, 117.74, 113.03, 85.68, 66.13, 65.27, 58.46, 53.60, 43.15, 35.89, 24.56, 20.36, 14.57.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.35

Rf_{I2} = 0,31 et Rf_{M2} = 0,69

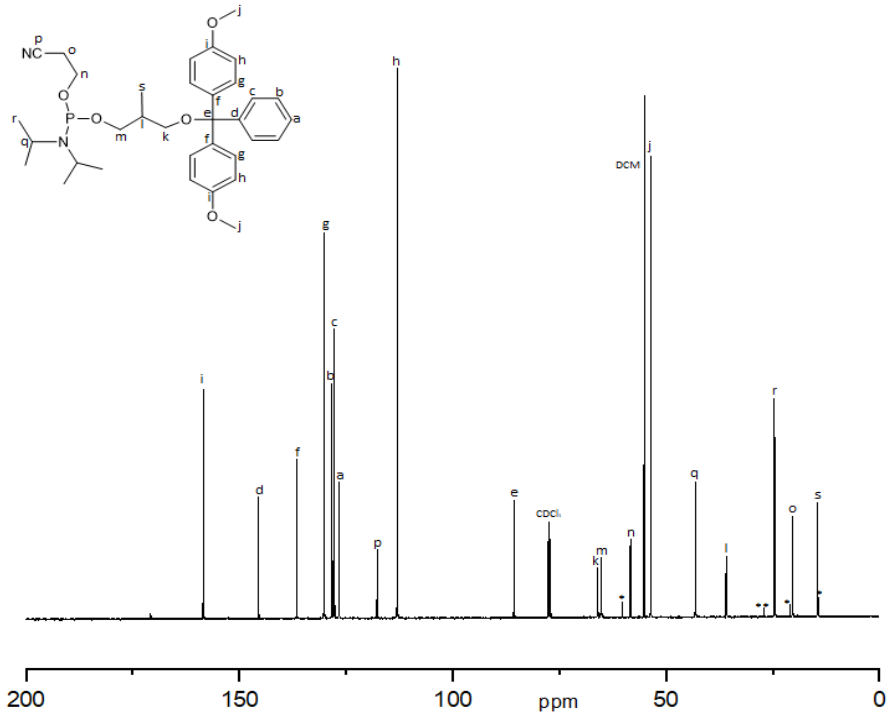


Figure SI-II. 21 : Spectre RMN ^{1}H ($CDCl_3$) du monomère M2

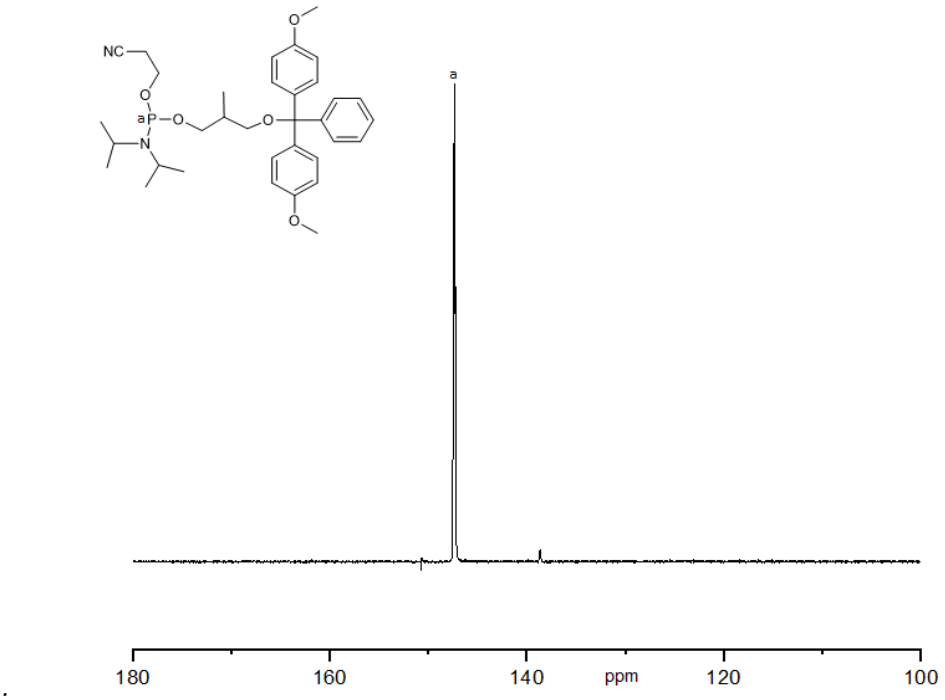


Figure SI-II. 22 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M2.

- **<u>Monomère M3 :</u>**

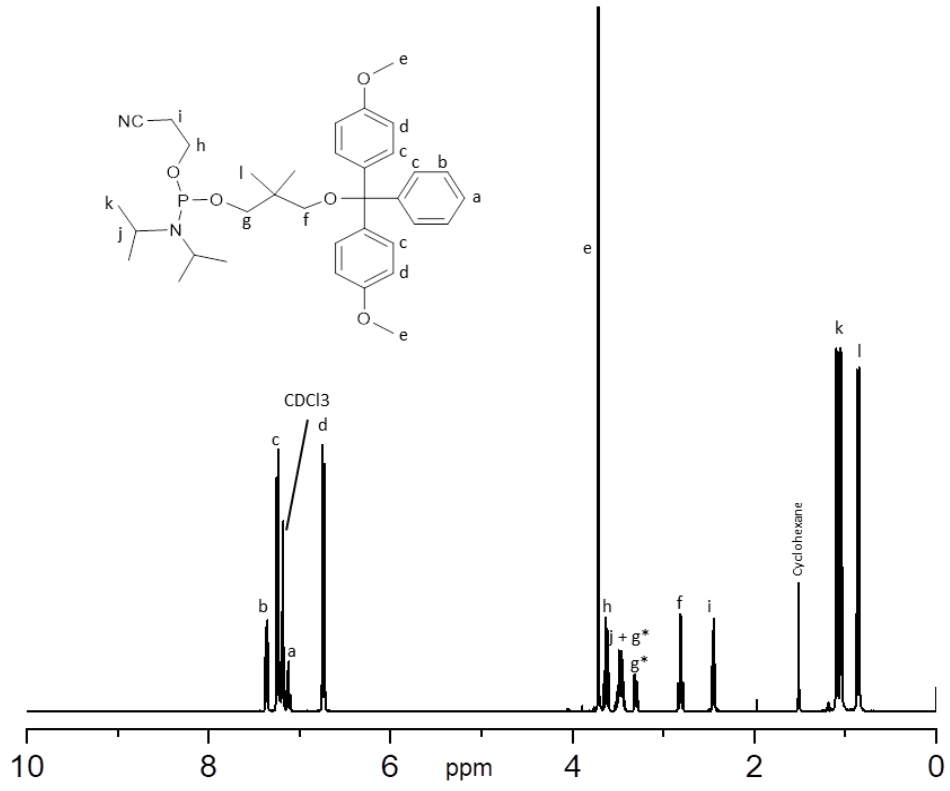


Figure SI-II. 23 : Spectre RMN ^{1}H ($CDCl_3$) du monomère M3.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : : 7.30 (d, J = 8.13 Hz, 2H, b), 7.16 (m, 6H, c), 7.07 (m, 1H, a), 6.69 (d, J =10.17 Hz, 4H, d), 3.66(s, 6H, e), 3.64-3.73 (m, 2H, h), 3.55-3.63 (m, 3H, j et un proton de g), 3.46 (m, 1H, deuxième proton de g), 3.02-3.05(m, 2H, f), 2.43 (m, 2H, i), 1.79 (m, 1H, l), 1.04-1.00 (dd, 12H, k).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.4, 145.3, 136.5, 130.04, 128.19, 127.68, 126.6, 118.9, 112.9, 85.81, 60.19, 58.4, 58.2, 55.2, 43.06, 31.9, 24.6, 20.27.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.7

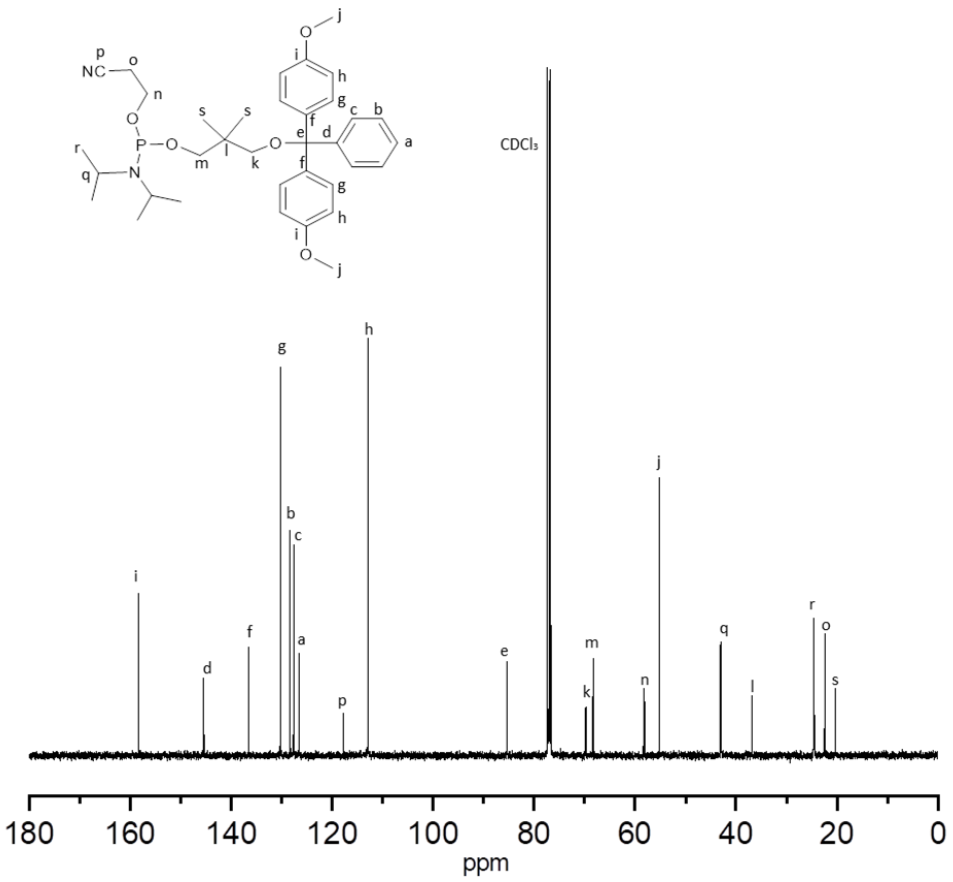


Figure SI-II. 24 : Spectre RMN ^{13}C ($CDCl_3$) du monomère M3.

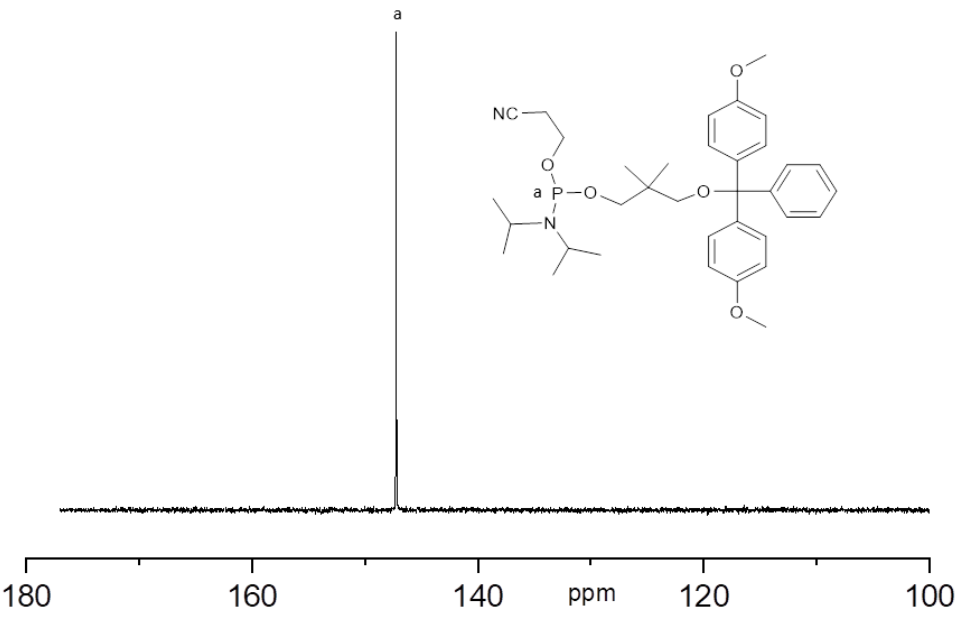


Figure SI-II. 25 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M3.

- **Monomère M4 :**

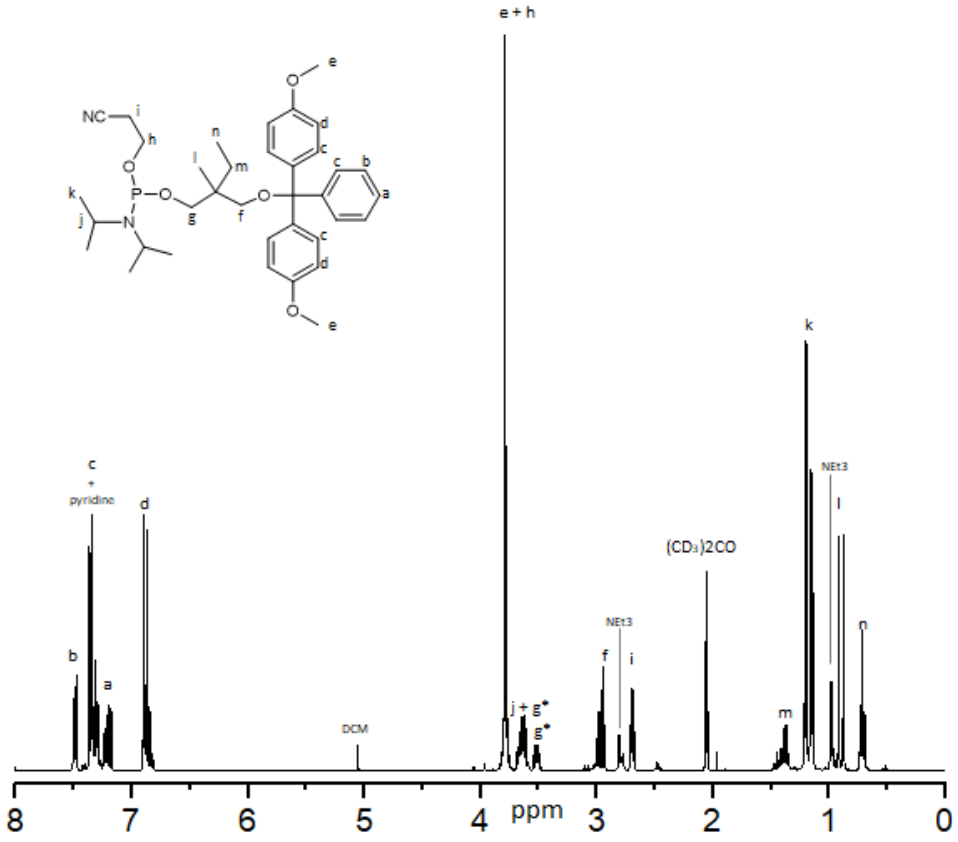


Figure SI-II. 26 : Spectre RMN ^{1}H ($(CD_3)_2CO$) du monomère M4.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.48 (d, J = 8.21 Hz, 2H, b), 7.32 (m, 6H, c), 7.20 (m, 1H, a), 6.88 (m,4H, d), 3.78(s, 6H, e), 3.74-3.77 (m, 2H, h), 3.63 (m, 3H, j et un proton de g), 3.48-3.53 (m, 1H, deuxième proton de g), 2.92-3.0(m, 2H, f), 2.69 (m, 2H, i), 1.34-1.42 (m, 2H, m), 1.13-1.20 (dd, 12H, k), 086 (d, J=15.97Hz, 3H, l), 0.71(td, 3H, n).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.58, 145.74, 136.44, 129.85, 128.21, 127.57, 126.48, 112.85, 112.2, 85.43, 66.91, 66.56, 65.95, 54.87, 54.58, 39.29, 26.82, 26.64, 18.92, 18.60, 7.11.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.15

Rf_{I4} = 0,70 et Rf_{M4} = 0,83

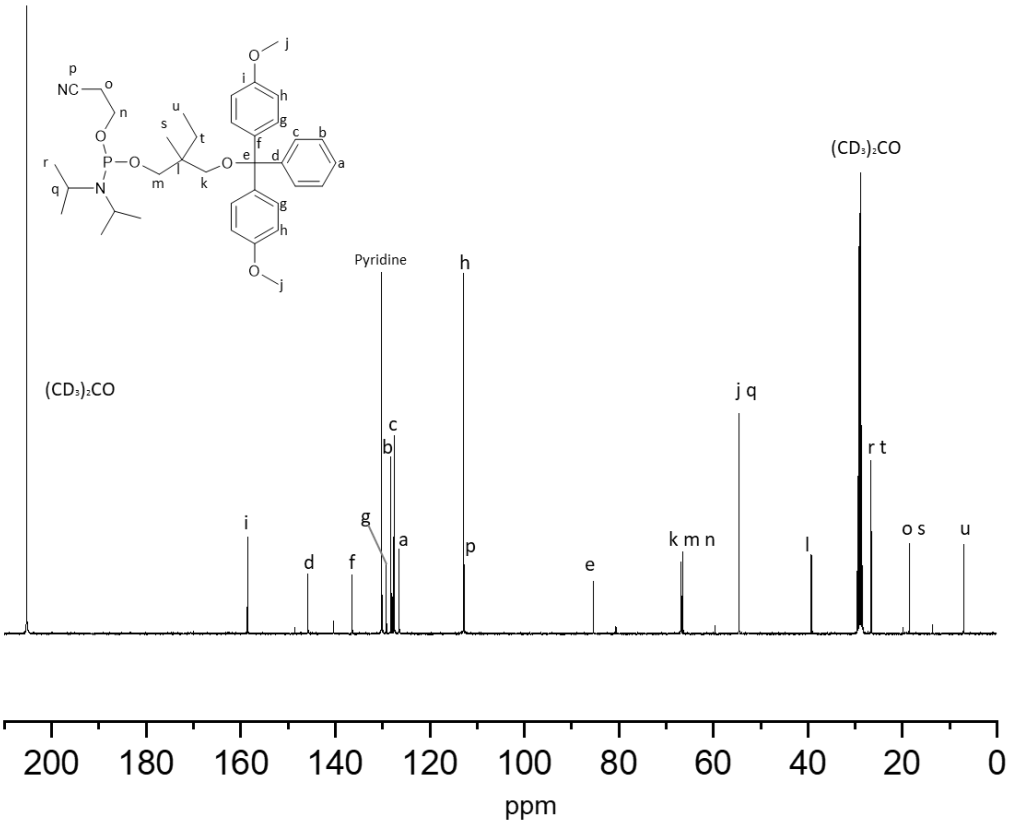


Figure SI-II. 27 : Spectre RMN ^{13}C ($(CD_3)_2CO$) du monomère M4.

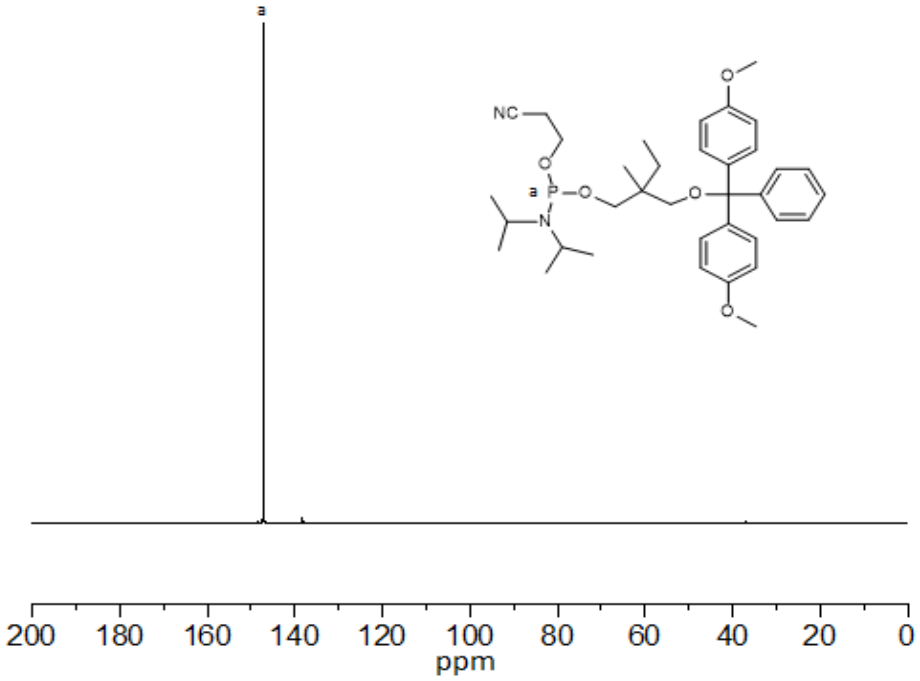


Figure SI-II. 28 : Spectre RMN ^{31}P ($(CD_3)_2CO$) du monomère M4.

- **Monomère M4-*iso* :**

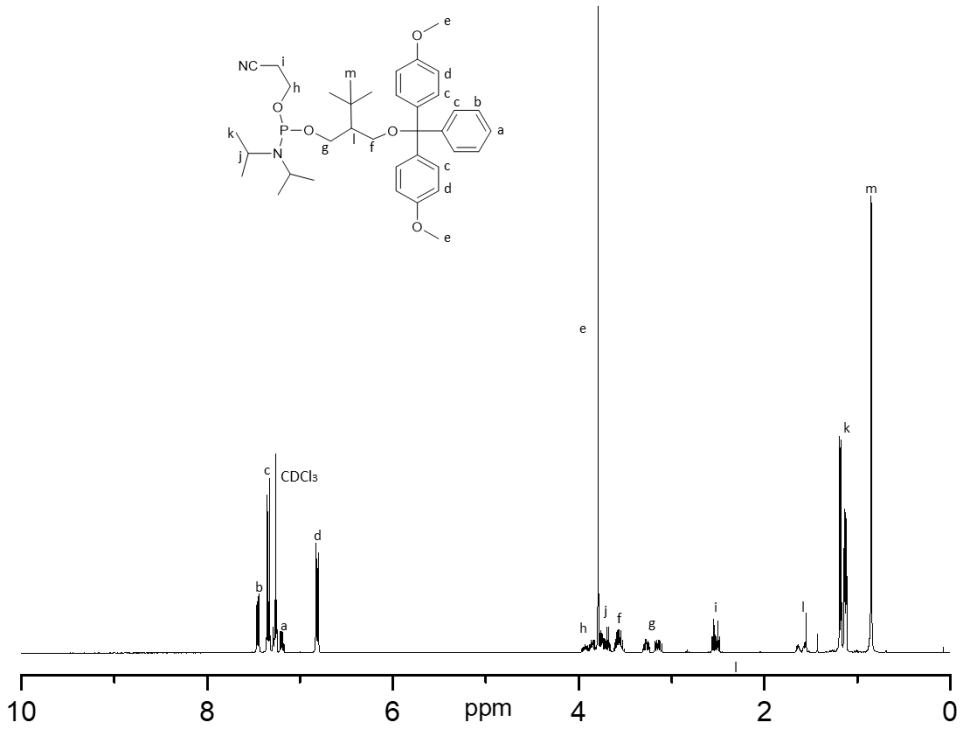


Figure SI-II. 29 : Spectre RMN ^{1}H ($CDCl_3$) du monomère M4-iso.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.45 (d, J = 7.13 Hz, 2H, b), 7.34 (m, 6H, c), 7.20 (m, 1H, a), 6.81 (m,4H, d), 3.82-3.97 (m, 2H, h), 3.79(s, 6H, e), 3.66-3.76 (m, 2H, j), 3.11-3.31 (m, 2H, g), 2.52 (m, 2H, i), 1.60 (m, 1H, l), 1.13-1.18 (dd, 12H, k), 0.85 (s, 9H, m).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.31, 145.43, 136.73, 130.20, 128.38, 127.64, 126.55, 117.70, 112.92, 86.11, 60.96, 60.92, 58.26, 55.16, 49.80, 42.98, 32.07, 28.81, 24.73, 20.59.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 150.68

$Rf_{I4\text{-}iso}$ = 0,42 et $Rf_{M4\text{-}iso}$ = 0,86

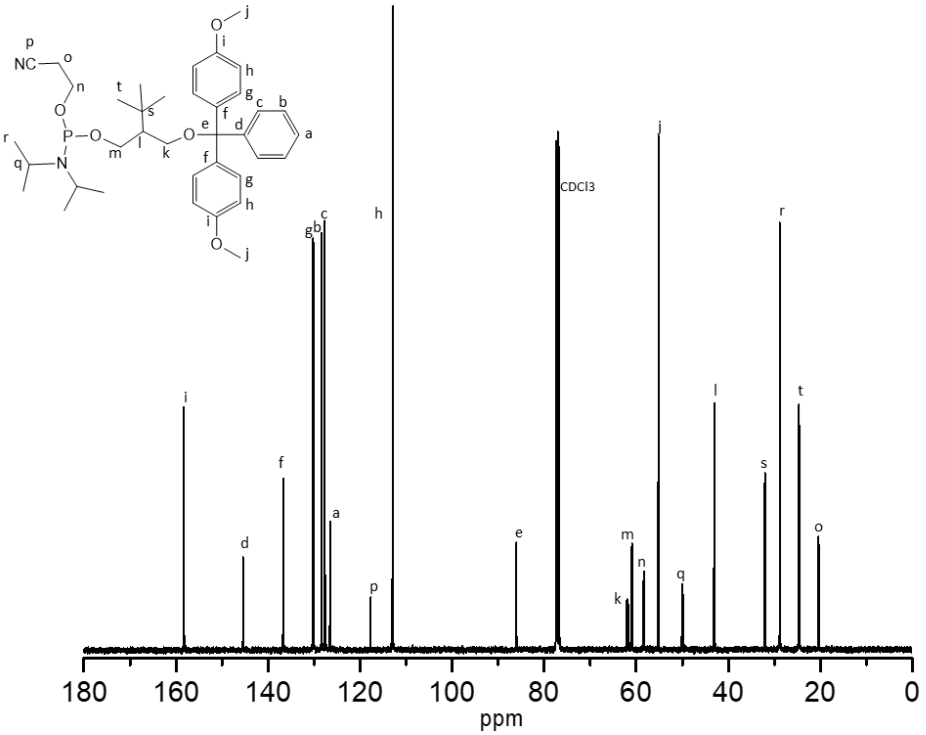


Figure SI-II. 30 : Spectre RMN ^{13}C ($CDCl_3$) du monomère M4-iso.

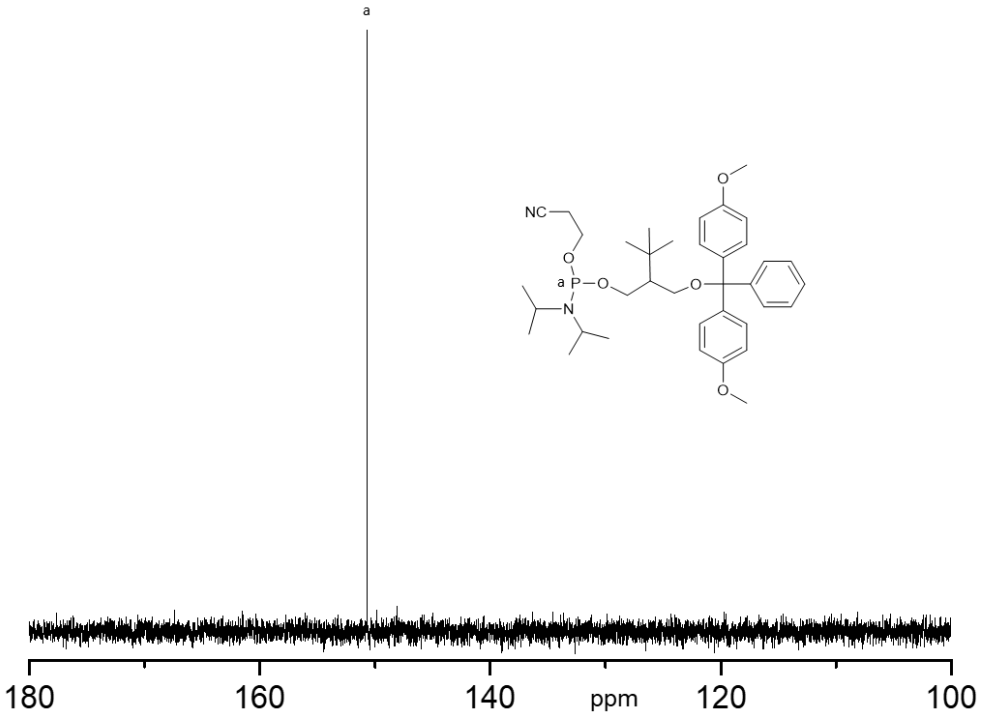


Figure SI-II. 31 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M4-iso.

- **Monomère M5 :**

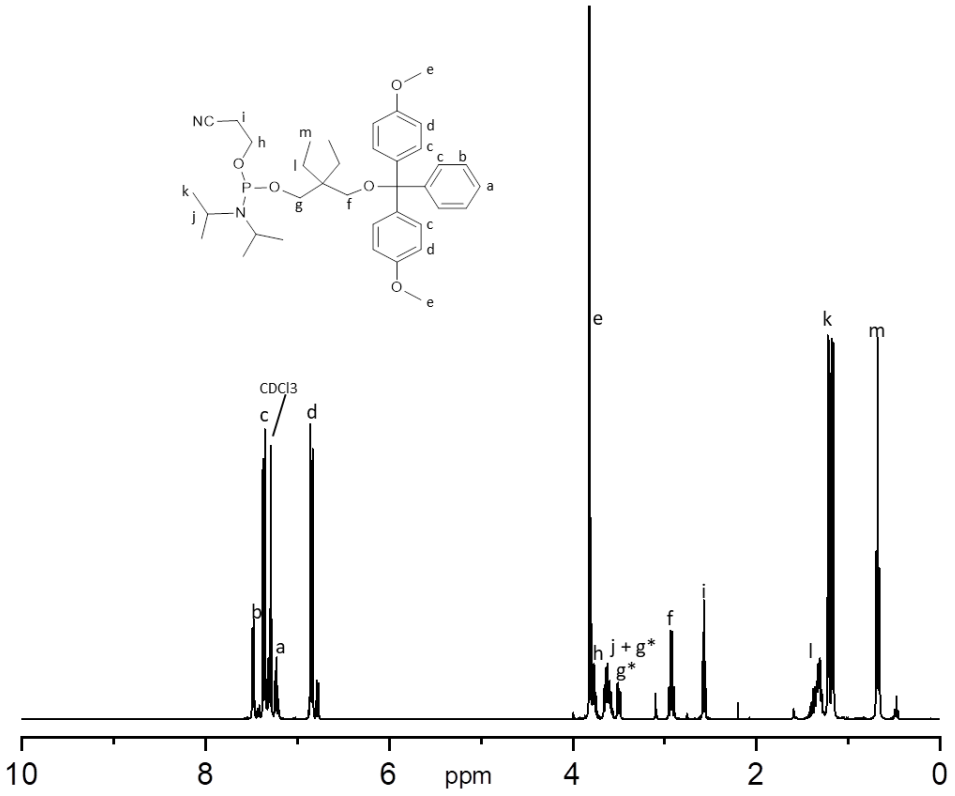


Figure SI-II. 32 : Spectre RMN ^{1}H ($(CD_3)_2CO$) du monomère M5.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.48 (d, J = 7.44 Hz, 2H, b), 7.32 (m, 6H, c), 7.23 (m, 1H, a), 6.84 (m,4H, d), 3.81(s, 6H, e), 3.74-3.80 (m, 2H, h), 3.56-3.66 (m, 3H, j et un proton de g), 3.48-3.52 (m, 1H, deuxième proton de g), 2.90-2.96(m, 2H, f), 2.57 (m, 2H, i), 1.29-1.42 (m, 4H, l), 1.17-1.21 (dd, 12H, k), 0.68 (t, 6H, m).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.30, 145.39, 136.51, 130.30, 128.42, 127.57, 126.51, 117.67, 112.84, 85.30, 66.12, 64.00, 60.38, 58.33, 55.17, 39.29, 26.82, 26.64, 18.92, 18.60, 7.11.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.07.

Rf_{I5} = 0,65 et Rf_{M5} = 0,79

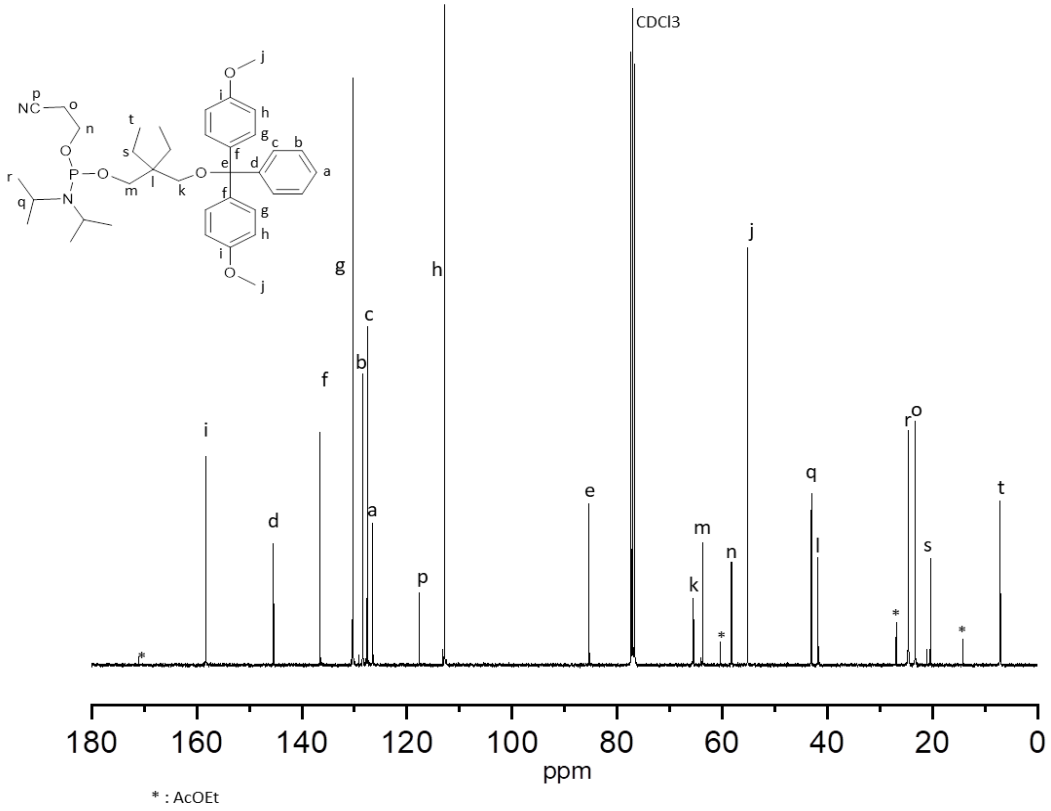


Figure SI-II. 33 : Spectre RMN ^{13}C ($(CD_3)_2CO$) du monomère M5.

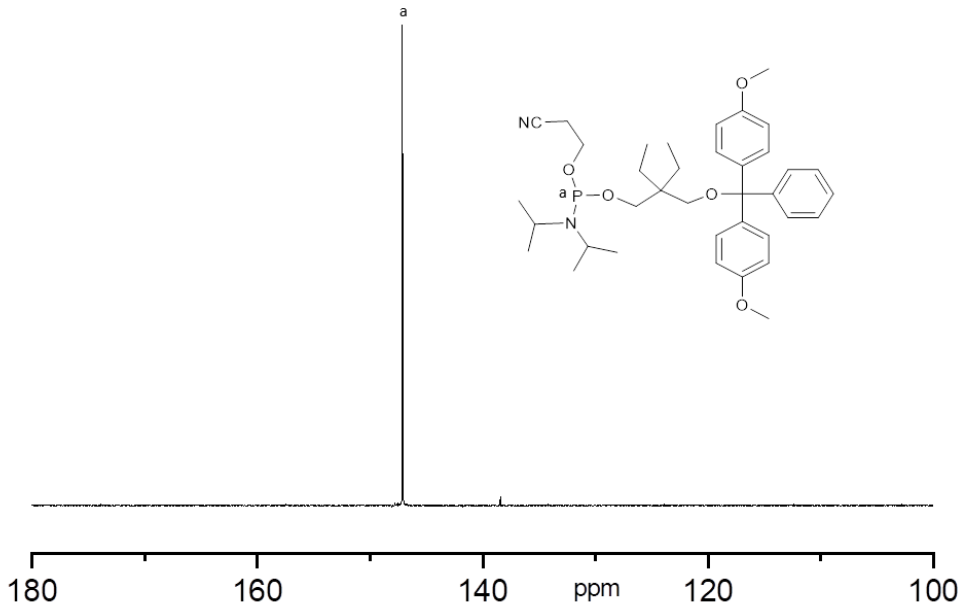


Figure SI-II. 34 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M5.

- **Monomère M6 :**

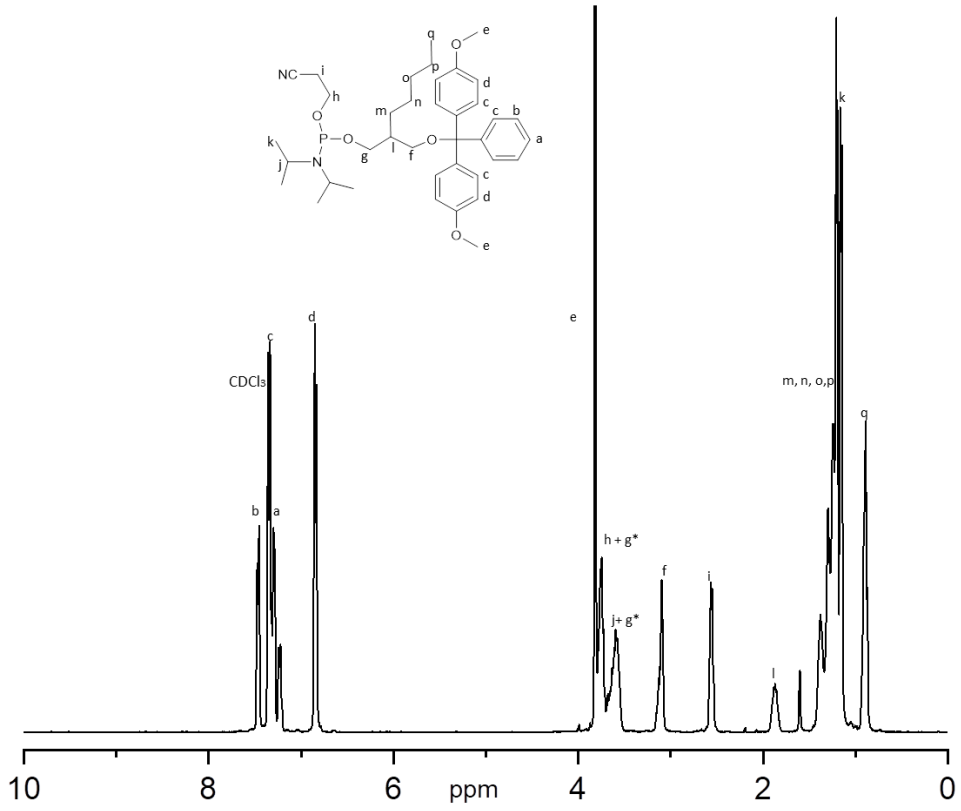


Figure SI-II. 35 : Spectre RMN ^{1}H ($CDCl_3$) du monomère M6.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.46 (d, J = 7.84 Hz, 2H, b), 7.34 (m, 6H, c), 7.23 (m, 1H, a), 6.84 (m,4H, d), 3.81(s, 6H, e), 3.70-3.79 (m, 3H, h et un proton de g), 3.52-3.70 (m, 3H, j et un proton de g), 3.06-3.16 (m, 2H, f), 2.55 (m, 2H, i), 1.82-1.93 (m, 1H, l), 1.29-1.42(m,8H, m,n,o et p), 1.15-1.20 (dd, 12H, k), 0.89 (m, 6H, q).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.33, 145.37, 136.59, 130.14, 128.30, 127.65, 126.57, 117.66, 112.93, 85.60, 64.29, 64.3, 58.22, 55.19, 43.08, 40.78, 32.12, 28.39, 26.58, 25.12, 22.60, 20.37, 14.08.

RMN ^{31}P ($CDCl_3$, δ, ppm): 147.30.

Rf_{I6} = 0,43 et Rf_{M6} = 0,86

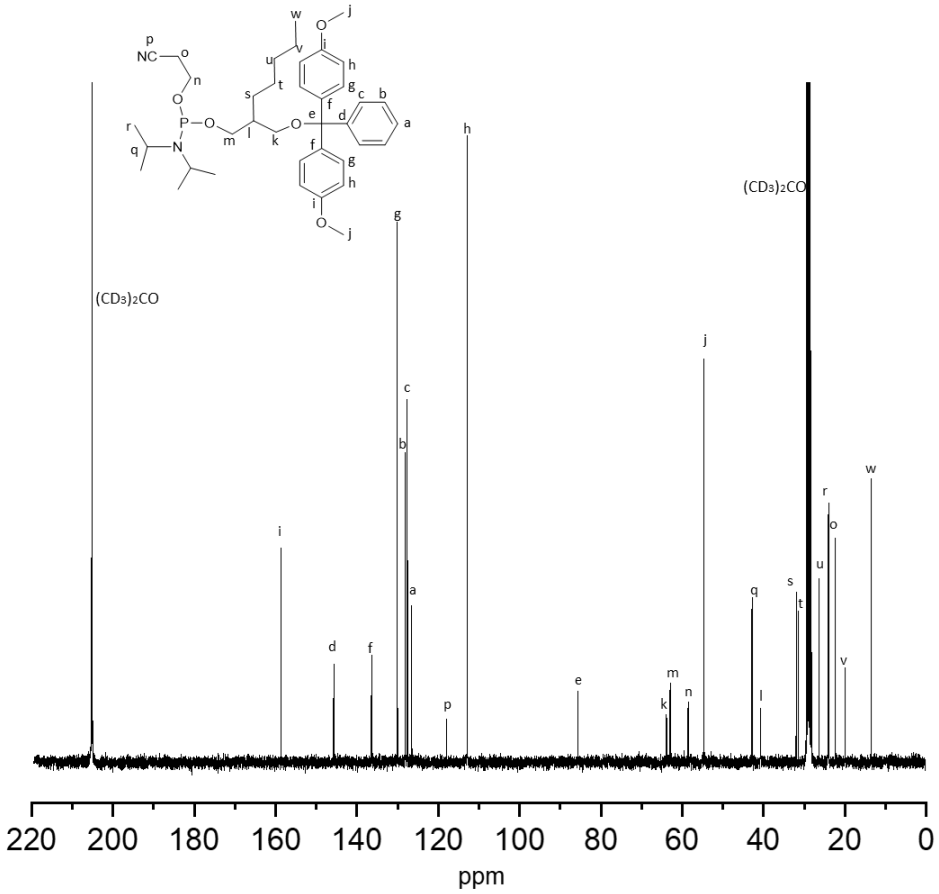


Figure SI-II. 36 : Spectre RMN ^{13}C ($(CD_3)_2CO$) du monomère M6.

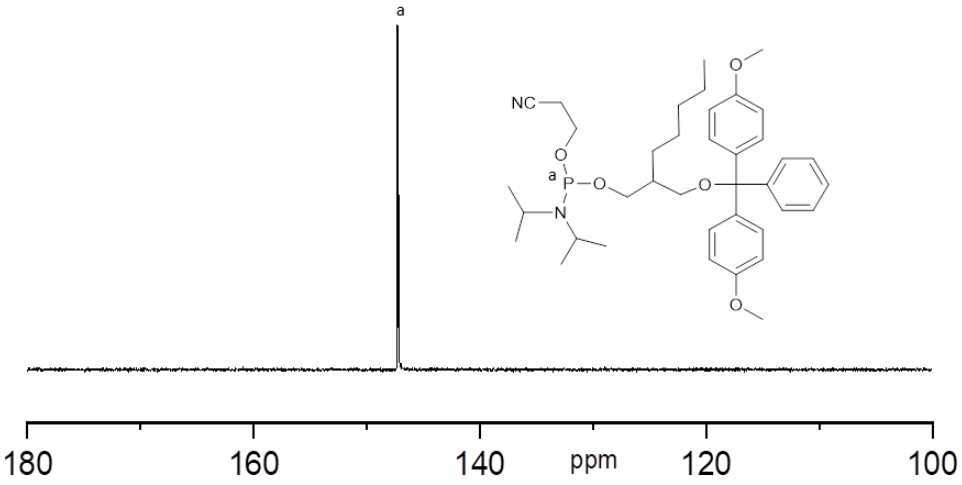


Figure SI-II. 37 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M6.

- **Monomère M7 :**

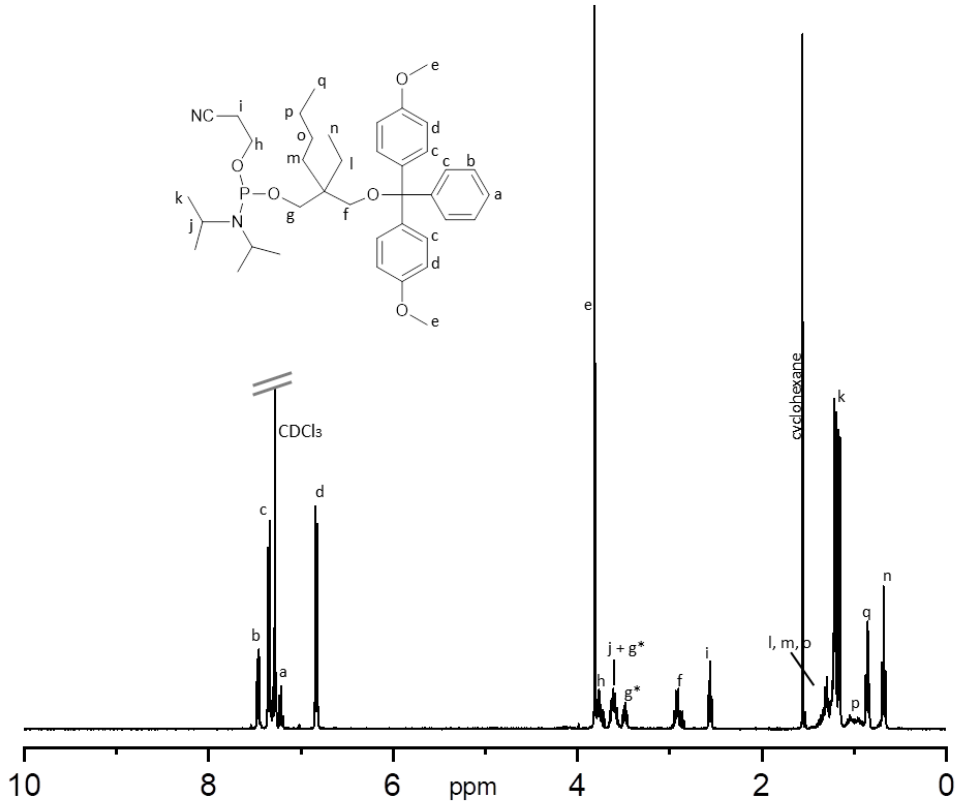


Figure SI-II. 38 : Spectre RMN ^{1}H ($CDCl_3$) du monomère M7.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.47 (d, J = 7.07 Hz, 2H, b), 7.32 (m, 6H, c), 7.22 (m, 1H, a), 6.83 (m,4H, d), 3.81(s, 6H, e), 3.73-3.80 (m, 2H, h), 3.56-3.66 (m, 3H, j et un proton de g),3.46-3.52 (m, 1H, un proton de g), 2.89-2.96 (m, 2H, f), 2.56 (t, 2H, i), 1.24-1.38(m,6H, l,m et o), 1.17-1.215 (dd, 12H, k), 0.9-1.12(m, 2H, p), 0.86 (td, 3H, q), 0.68 (t, 3H, n).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.29, 145.38, 136.52, 130.30, 128.43, 127.57, 126.52, 117.68, 112.84, 85.32, 63.87, 63.70, 60.39, 55.19, 43.75, 41.66, 30.65, 24.67, 23.73, 23.74, 23.61, 23.62, 20.43, 14.11, 7.16.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.14.

Rf_{I7} = 0,39 et Rf_{M7} = 0,85

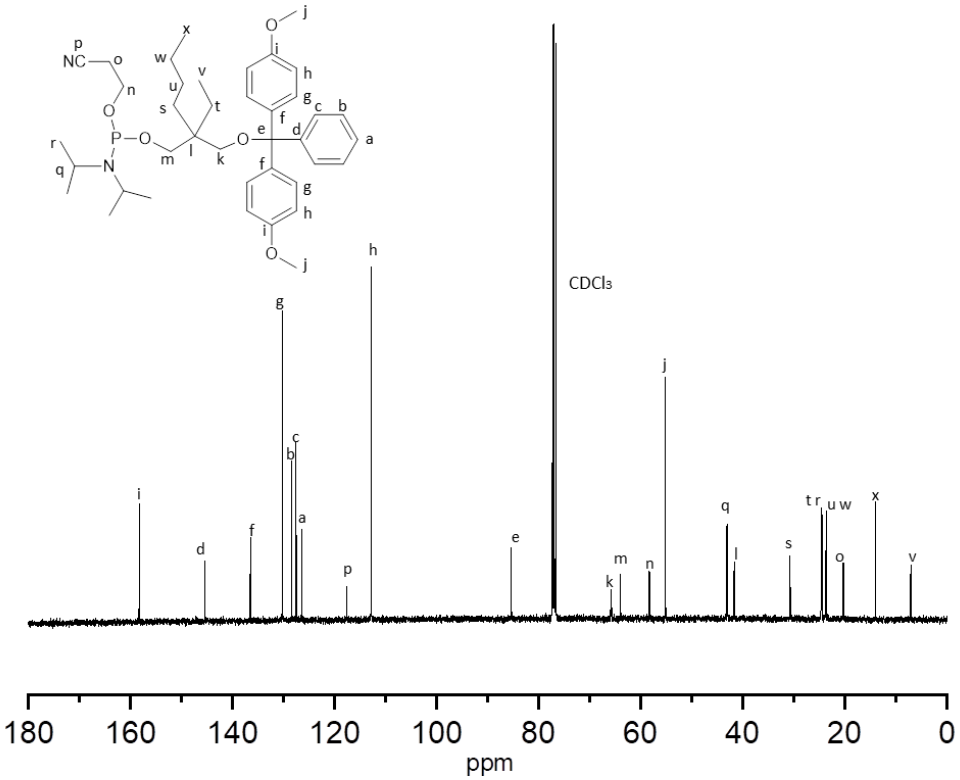


Figure SI-II. 39 : Spectre RMN ^{13}C ($CDCl_3$) du monomère M7.

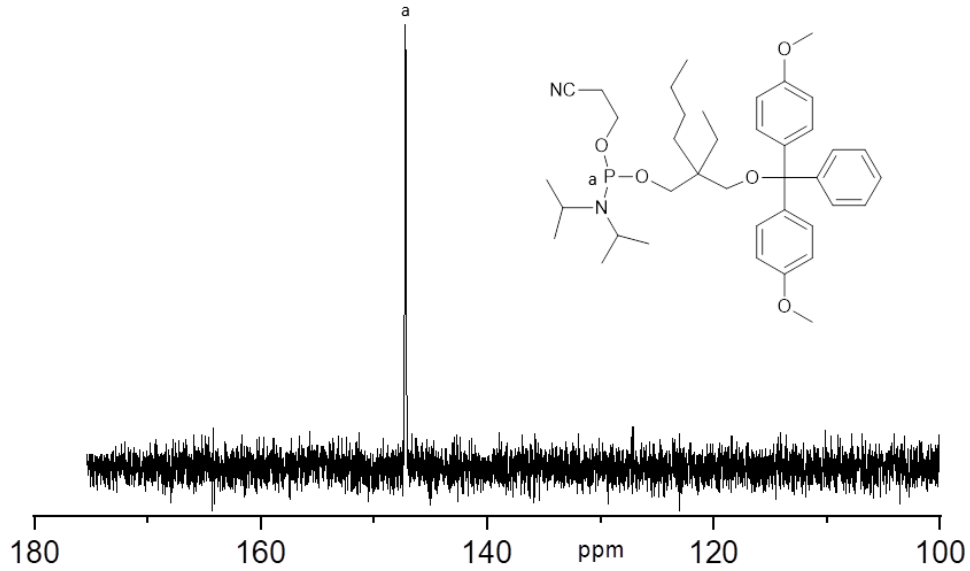


Figure SI-II. 40 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M7.

- **Monomère M8 :**

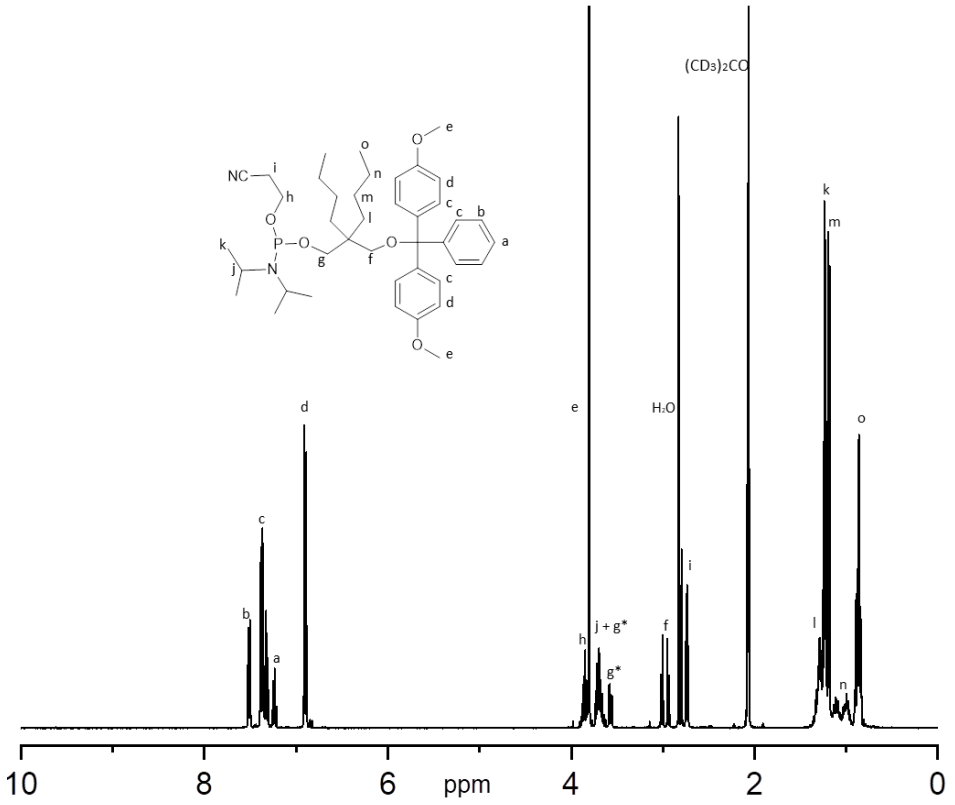


Figure SI-II. 41 : Spectre RMN ^{1}H ($(CD_3)_2CO$) du monomère M8.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.47 (d, J = 7.79 Hz, 2H, b), 7.32 (m, 6H, c), 7.22 (m, 1H, a), 6.83 (m,4H, d), 3.81(s, 6H, e), 3.82-3.85 (m, 2H, h), 3.57-3.66 (m, 3H, j et un proton de g),3.48-3.52 (m, 1H, un proton de g), 2.89-2.96 (m, 2H, f), 2.57 (t, 2H, i), 1.24-1.30(m,4H, l), 1.17-1.22 (m, 16H, k et m), 0.9-1.17(m, 2H, n), 0.86 (td, 6H, n).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 158.30, 145.39, 136.51, 130.30, 128.42, 127.57, 126.51, 117.67, 112.84, 85.30, 66.12, 64.00, 60.38, 55.17, 43.12, 41.55, 31.11, 26.92, 24.71, 23.60, 21.03, 14.11.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 147.77.

Rf_{I8} = 0,41 Rf_{M8} = 0,77

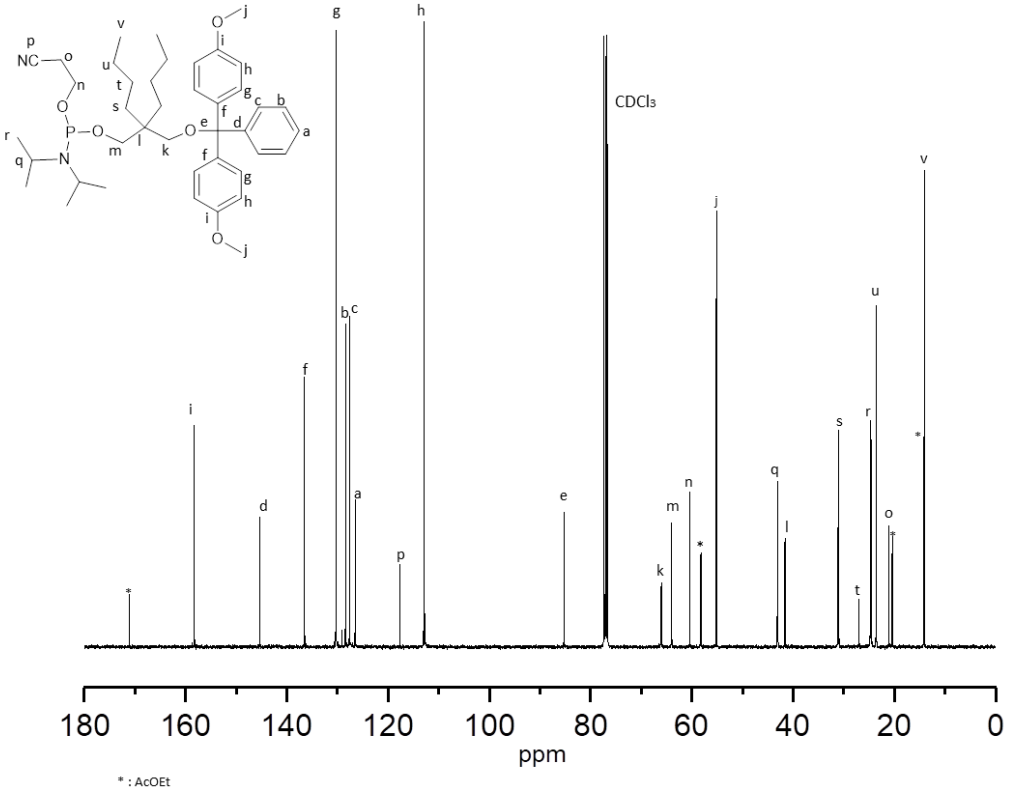


Figure SI-II. 42 : Spectre RMN ^{13}C ($CDCl_3$) du monomère M8.

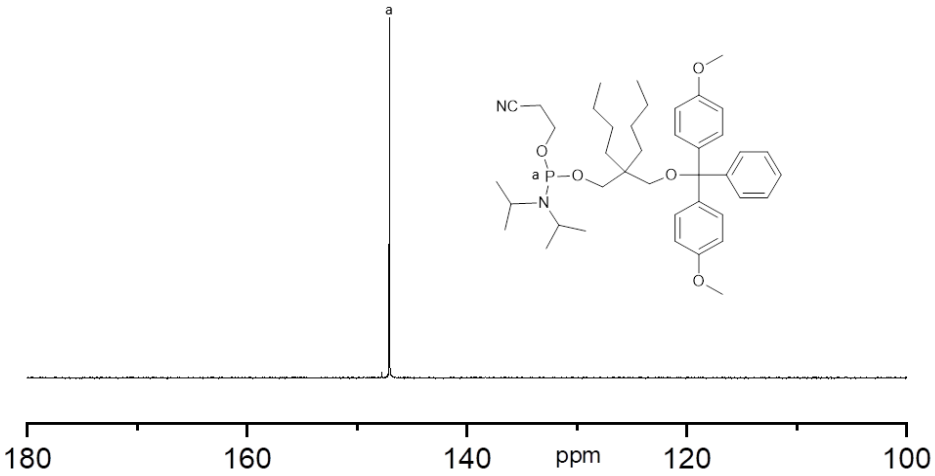


Figure SI-II. 43 : Spectre RMN ^{31}P ($CDCl_3$) du monomère M8.

- **Espaceur :**

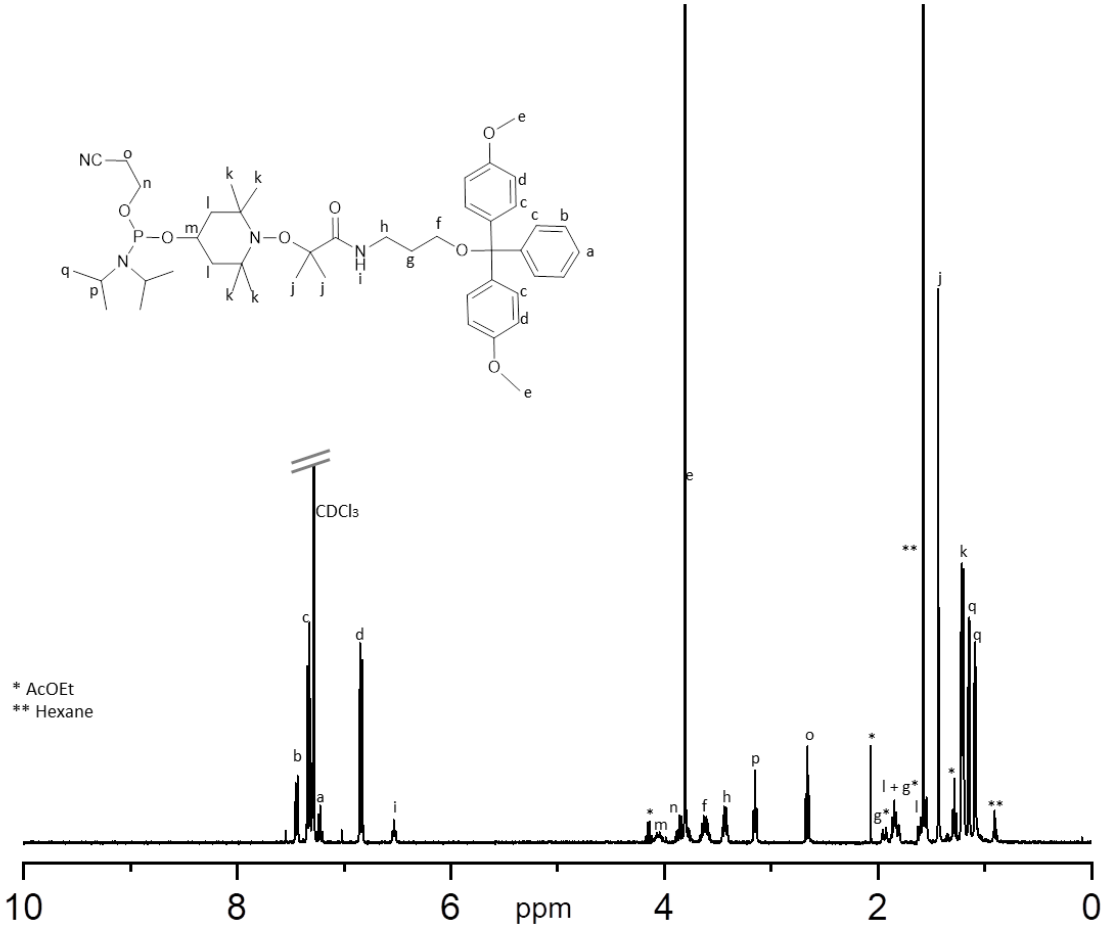


Figure SI-II. 44 : Spectre RMN ^{1}H ($CDCl_3$) du monomère espaceur E.

RMN ^{1}H (400 Mhz, $CDCl_3$, δ, ppm) : 7.44 (d, J = 8.54 Hz, 2H, b), 7.30 (m, 6H, c), 7.22 (m, 1H, a), 6,84 (d, J=9.15 Hz, 4H, d), 6,53 (m, 1H, i), 4,06 (m, 1H, m), 3.85 (m, 1H, n), 3.80 (s, 6H, e), 3.62 (m, 2H, f), 3.43 (m, 2H, h), 3.15 (m, 2H, p), 2.66 (m, 2H, o), 1.94 (m, 1H, un proton de g), 1.85 (m, 3H, deux protons de l et un proton de g), 1.59 (m, 2H, deux protons de l), 1.43 (s, 6H, j), 1.21 (dd, 12H, k), 1.15 (d, J= 3.68Hz, 6H, q), 1.09 (d, J= 2.52Hz, 6H, q).

RMN ^{13}C (75 MHz, $CDCl_3$, δ, ppm) : 176.9, 158.3, 145.05, 136.3, 129.9, 128.1, 127.8, 126.7, 117.7, 113.1, 85.9, 83.2, 65.2, 61.12, 60.1, 58.3, 55.2, 47.8, 43.1, 33.4, 29.9, 24.7, 24.5, 21.71, 20.4 ppm.

RMN ^{31}P ($CDCl_3$, δ, ppm) : 145.39 ppm.

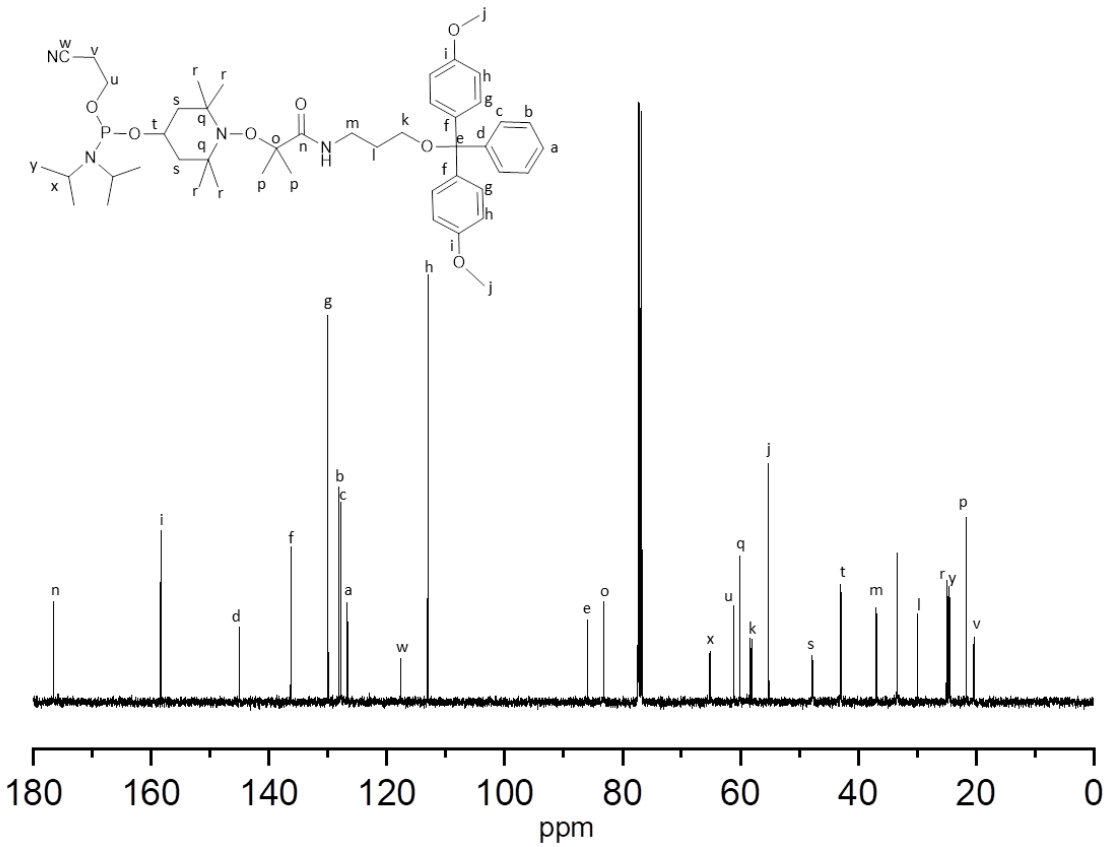


Figure SI-II. 45 : Spectre RMN ^{13}C ($CDCl_3$) du monomère espaceur E.

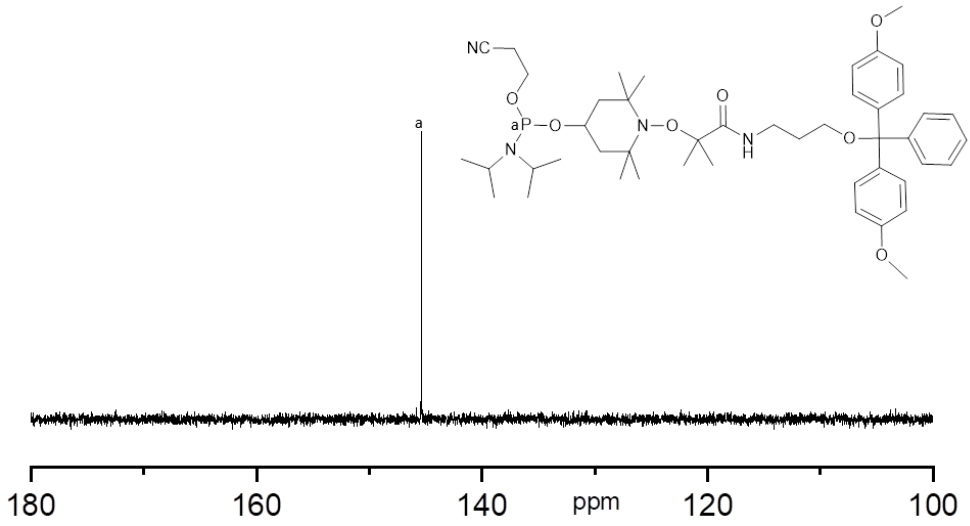


Figure SI-II. 46 : Spectre RMN ^{31}P ($CDCl_3$) de l'espaceur E.

3.3. Caractérisation des monomères par spectrométrie de masse

L'analyse de spectrométrie de masse est effectuée par une ionisation par électronébuliseur (ESI-MS). Tous les monomères de l'alphabet augmenté à 8 symboles sont analysés et les résultats sont donnés ci-dessous :

Tableau SI-II. 2 : Caractérisation des monomères M2 et M4 à M8 et les intermédiaires respectifs en spectrométrie de masse.

Monomères	Forme protonée		m/Z théorique	m/Z calculée
I2	$[M+Na]^+$	$C_{24}H_{28}O_4Na^+$	415.1880	415.1881
M2	$[M+H]^+$	$C_{34}H_{46}N_2O_5P^+$	593.3139	593.3141
I4	$[M+Na]^+$	$C_{27}H_{32}O_4Na^+$	443.2193	443.2192
M4	$[M+H]^+$	$C_{36}H_{50}N_2O_5P^+$	621.3452	621.3454
I5	$[M+Na]^+$	$C_{28}H_{34}O_4Na^+$	457.2349	457.2347
M5	$[M+H]^+$	$C_{37}H_{52}N_2O_5P^+$	635.3608	635.3611
I6	$[M+Na]^+$	$C_{29}H_{36}O_4Na^+$	471.2506	471.2513
M6	$[M+H]^+$	$C_{38}H_{54}N_2O_5P^+$	649.3765	649.3770
I7	$[M+Na]^+$	$C_{30}H_{38}O_4Na^+$	485.2662	485.2662
M7	$[M+H]^+$	$C_{39}H_{56}N_2O_5P^+$	663.3921	663.3920
I8	$[M+Na]^+$	$C_{32}H_{42}O_4Na^+$	513.2975	513.2981
M8	$[M+H]^+$	$C_{41}H_{60}N_2O_5P^+$	691.4234	691.4237

3.4. Synthèse des polymères

3.4.1. Préparation des solutions de monomères

Les polymères P1 à P14 ont été synthétisés avec la chimie de la phosphoramidite automatisée. Deux synthétiseurs ont été utilisés : le synthétiseur Expedite et le synthétiseur ABI. Les solutions de monomères nécessaires ont été préparées dans de l'acétonitrile anhydre à 100mM pour l'Expedite et à 50mM pour l'ABI. Généralement des solutions de 10 mL étaient préparées, la quantité nécessaire de monomère est donnée en Tableau SI-II. 3.

Tableau SI-II. 3 : Masse de monomère nécessaire pour la préparation de 10 mL de solution à 100 mM (Expédite) et à 50 mM (ABI).

Monomère	M1	M2	M3	M4	M5	M6	M7	M8	Espaceur
m_Expedite (g)	0,563	0,577	0,591	0,605	0,619	0,633	0,647	0,675	0,820
m_ABI (g)	0,282	0,289	0,296	0,303	0,310	0,317	0,324	0,338	0,410

3.4.2. Description du langage utilisé

Le langage ASCII a principalement été utilisé pour coder du texte sur les séquences d'oligomères présentées. Sa table est donnée en Tableau SI-II. 4.

Tableau SI-II. 4 : Table ASCII des lettres de l'alphabet en majuscule et minuscule.

Caractère	Code binaire	Caractère	Code binaire
A	01000001	a	01100001
B	01000010	b	01100010
C	01000011	c	01100011
D	01000100	d	01100100
E	01000101	e	01100101
F	01000110	f	01100110
G	01000111	g	01100111
H	01001000	h	01101000
I	01001001	i	01101001
J	01001010	j	01101010
K	01001011	k	01101011
L	01001100	l	01101100
M	01001101	m	01101101
N	01001110	n	01101110
O	01001111	o	01101111
P	01010000	p	01110000
Q	01010001	q	01110001
R	01010010	r	01110010
S	01010011	s	01110011
T	01010100	t	01110100
U	01010101	u	01110101
V	01010110	v	01110110
W	01010111	w	01110111
X	01011000	x	01111000
Y	01011001	y	01111001
Z	01011010	z	01111010

3.4.3. Description des protocoles utilisés sur les synthétiseurs d'ADN

3.4.3.1. Synthétiseur Expedite

Les commandes suivies par le synthétiseur Expedite sont données ci-après et montre un cycle entier du protocole standard pour l'ajout d'un monomère. Les commentaires pour chaque ligne sont donnés entre guillemet ou après un astérisque. Le premier numéro de chaque ligne correspond à la position du réactif (ou à la commande pour la détection UV), le second définie le nombre de pulsations nécessaires pour délivrer le réactif à la colonne et le troisième donne le temps dans lequel ces pulsations doivent être effectuées. Dans les cas où un « 0 » est noté, la pompe fonctionne à la vitesse maximale. L'exemple suivant est pour le couplage d'un monomère en position 8 en utilisant le protocole standard.

Protocole standard

```
$Deblocking
144     /*Index Fract. Coll.    */ NA       1     0     "Event out ON"
0        /*Default              */ WAIT     0     1.5   "Wait"
141     /*Trityl Mon. On/Off    */ NA       1     1     "START data collection"
16      /*Dblk                  */ PULSE    10    0     "Dblk to column"
16      /*Dblk                  */ PULSE    50    49    "Deblock"
38      /*Diverted Wsh A        */ PULSE    80    0     "Flush system with Wsh A"
141     /*Trityl Mon. On/Off    */ NA       0     1     "STOP data collection"
38      /*Diverted Wsh A        */ PULSE    80    0     "Flush system with Wsh A"
144     /*Index Fract. Coll.    */ NA       2     0     "Event out OFF"

$Coupling
1       /*Wsh                   */ PULSE    5     0     "Flush system with Wsh"
2       /*Act                   */ PULSE    5     0     "Flush system with Act"
25      /*8 + Act               */ PULSE    6     0     "Monomer + Act to column"
25      /*8 + Act               */ PULSE    1     8     "Couple monomer"
2       /*Act                   */ PULSE    4     32    "Couple monomer"
1       /*Wsh                   */ PULSE    7     56    "Couple monomer"
1       /*Wsh                   */ PULSE    8     0     "Flush system with Wsh"
$Capping
12      /*Wsh A                 */ PULSE    20    0     "Flush system with Wsh A"
13      /*Caps                  */ PULSE    8     0     "Caps to column"
12      /*Wsh A                 */ PULSE    6     90    "Cap"
12      /*Wsh A                 */ PULSE    14    0     "Flush system with Wsh A"

$Oxidizing
15      /*Ox                    */ PULSE    15    0     "Ox to column"
12      /*Wsh A                 */ PULSE    60    0     "Flush system with Wsh A"

$Capping
13      /*Caps                  */ PULSE    7     0     "Caps to column"
12      /*Wsh A                 */ PULSE    6     90    "Cap"
12      /*Wsh A                 */ PULSE    30    0     "End of cycle wash"
```

Protocole a : Etape de couplage allongée

Le protocole **a** est décrit ci-dessous, à nouveau pour un monomère placé en position 8. Les étapes différentes par rapport au protocole standard sont mises en couleur.

```
$Deblocking
144     /*Index Fract. Coll.    */ NA       1     0     "Event out ON"
0        /*Default              */ WAIT     0     1.5   "Wait"
141     /*Trityl Mon. On/Off    */ NA       1     1     "START data collection"
16      /*Dblk                  */ PULSE    10    0     "Dblk to column"
16      /*Dblk                  */ PULSE    50    49    "Deblock"
38      /*Diverted Wsh A        */ PULSE    80    0     "Flush system with Wsh A"
141     /*Trityl Mon. On/Off    */ NA       0     1     "STOP data collection"
38      /*Diverted Wsh A        */ PULSE    80    0     "Flush system with Wsh A"
144     /*Index Fract. Coll.    */ NA       2     0     "Event out OFF"

$Coupling
1       /*Wsh                   */ PULSE    15    0     "Flush system with Wsh"
2       /*Act                   */ PULSE    5     0     "Flush system with Act"
25      /*8 + Act               */ PULSE    6     0     "Monomer + Act to column"
```

```
25      /*8 + Act           */ PULSE    2       75      "Couple monomer"
2       /*Act               */ PULSE    3       75      "Couple monomer"
1       /*Wsh               */ PULSE    7       150     "Couple monomer"
1       /*Wsh               */ PULSE    8       0       "Flush system with Wsh"
$Capping
12      /*Wsh A             */ PULSE    20      0       "Flush system with Wsh A"
13      /*Caps              */ PULSE    8       0       "Caps to column"
12      /*Wsh A             */ PULSE    6       90      "Cap"
12      /*Wsh A             */ PULSE    14      0       "Flush system with Wsh A"

$Oxidizing
15      /*Ox                */ PULSE    15      0       "Ox to column"
12      /*Wsh A             */ PULSE    60      0       "Flush system with Wsh A"

$Capping
13      /*Caps              */ PULSE    7       0       "Caps to column"
12      /*Wsh A             */ PULSE    6       90      "Cap"
12      /*Wsh A             */ PULSE    30      0       "End of cycle wash"
```

Protocole b : Etape de couplage et oxydation doublées

Le protocole **b** est décrit ci-dessous, à nouveau pour un monomère placé en position 8. Les étapes ajoutées par rapport au protocole standard sont mises en couleur.

```
$Deblocking
144     /*Index Fract. Coll.   */ NA       1       0       "Event out ON"
0       /*Default             */ WAIT     0       1.5     "Wait"
141     /*Trityl Mon. On/Off  */ NA       1       1       "START data collection"
16      /*Dblk                */ PULSE    10      0       "Dblk to column"
16      /*Dblk                */ PULSE    50      49      "Deblock"
38      /*Diverted Wsh A      */ PULSE    80      0       "Flush system with Wsh A"
141     /*Trityl Mon. On/Off  */ NA       0       1       "STOP data collection"
38      /*Diverted Wsh A      */ PULSE    80      0       "Flush system with Wsh A"
144     /*Index Fract. Coll.  */ NA       2       0       "Event out OFF"

$Coupling
1       /*Wsh               */ PULSE    5       0       "Flush system with Wsh"
2       /*Act               */ PULSE    5       0       "Flush system with Act"
25      /*8 + Act           */ PULSE    6       0       "Monomer + Act to column"
25      /*8 + Act           */ PULSE    1       8       "Couple monomer"
2       /*Act               */ PULSE    4       32      "Couple monomer"
1       /*Wsh               */ PULSE    7       56      "Couple monomer"
1       /*Wsh               */ PULSE    8       0       "Flush system with Wsh"

$Oxidizing
15      /*Ox                */ PULSE    15      0       "Ox to column"
12      /*Wsh A             */ PULSE    60      0       "Flush system with Wsh A"

$Coupling
1       /*Wsh               */ PULSE    5       0       "Flush system with Wsh"
2       /*Act               */ PULSE    5       0       "Flush system with Act"
25      /*8 + Act           */ PULSE    6       0       "Monomer + Act to column"
25      /*8 + Act           */ PULSE    1       8       "Couple monomer"
2       /*Act               */ PULSE    4       32      "Couple monomer"
1       /*Wsh               */ PULSE    7       56      "Couple monomer"
```

```
1       /*Wsh              */ PULSE    8       0       "Flush system with Wsh"

$Oxidizing
15      /*Ox               */ PULSE    15      0       "Ox to column"
12      /*Wsh A            */ PULSE    60      0       "Flush system with Wsh A"

$Capping
12      /*Wsh A            */ PULSE    20      0       "Flush system with Wsh A"
13      /*Caps             */ PULSE    8       0       "Caps to column"
12      /*Wsh A            */ PULSE    6       90      "Cap"
12      /*Wsh A            */ PULSE    14      0       "Flush system with Wsh A"
13      /*Caps             */ PULSE    7       0       "Caps to column"
12      /*Wsh A            */ PULSE    6       90      "Cap"
12      /*Wsh A            */ PULSE    30      0       "End of cycle wash"
```

3.4.3.2. Synthétiseur ABI

Les commandes suivies par le synthétiseur ABI sont données ci-après et montre un cycle entier du protocole standard pour l'ajout d'un monomère.

Tableau SI-II. 5 : Commandes suivies par le synthétiseur ABI.

PREPROCESSING	Purges	# Iterations	Wait sec.	Command Code	Param1 Chemical	Param2 Volume
PREWASH1		1	0	DISP	ACN	250
	LONG_PURGE					
START_LOOP						
DETRITYLATION		3	0	DISP	DEBLOCK	180
	LONG_PURGE					
TCA_WASH		1	0	DISP	ACN	250
	LONG_PURGE					
COUPLING		3	0	DISP	ACTIVATOR	115
				DISP	AMIDITE	75
	LONG_PURGE					
CAPPING		1	0	DISP	CAPB	80
				DISP	CAPA	80
	LONG_PURGE					
OXIDATION		1	0	DISP	OXIDIZER	150
	LONG_PURGE					
OX_WASH		1	0	DISP	ACN	250
	LONG_PURGE					
CAPPING		1	0	DISP	CAPB	80
				DISP	CAPA	80
	LONG_PURGE					
TRITYLOFF		3	0	DISP	DEBLOCK	180
	LONG_PURGE					
FINAL_FLUSH		4	0	DISP	ACN	250
	LONG_PURGE					
DRY_SUPPORT		1	0	DISP	ACN	180
	DRY_BEADS					
END_PROGRAM						

3.4.4. Clivage des polymères

La procédure de clivage des polymères est toujours la même. Une fois la synthèse automatisée effectuée, le support solide est enlevé du synthétiseur puis la solution de clivage (généralement NH_3/ NH_2CH_3, 1/1) est introduite dans le support et le tout est agité pendant 30 min à température ambiante. Ensuite une cartouche

de purification (colonne de chromatographie en phase inverse, DNA purification kit) est utilisée pour finir la procédure de clivage tel que :

1) Préparation de la cartouche

La cartouche est rincée avec 0,5 mL d'acétonitrile (MeCN) et 1 mL d'une solution tampon aqueuse à 2 M d'acétate de triéthylammonium.

2) Préparation de la solution de polymère à purifiée

Après le clivage du support la solution obtenue est mélangée avec 1 mL d'une solution de NaCl (aq. 10 m%) qui est ensuite appliquée sur la cartouche de purification.

3) Purification du polymère sur la cartouche

La cartouche est lavée avec plusieurs solutions :

a) 1 mL d'une solution salée de lavage (0,1g NaCl/mL dans une solution à 95 : 5, H_2O : MeCN).

Permet de laver les séquences tronquées qui ont été capées lors de la synthèse. En effet, lors de la synthèse automatique une étape de capping est effectuée à chaque cycle pour neutraliser le groupement terminal OH des séquences pour lesquelles le dernier couplage n'a pas pu se faire. Ces séquences sont tout de même présentes dans la solution récupérée après le clivage du support, il faut donc les lavées. Pour ce faire, cette solution salée est appliquée. Les séquences désirées vont donc avoir un bout de chaine se terminant par un groupement DMT. Ce groupement est suffisamment apolaire pour établir une interaction assez importante avec la phase statique pour accrocher la séquence entière. Les séquences capées plus petites qui ne contiennent pas de groupement apolaire DMT seront alors lavées avec la phase mobile.

b) 2 mL d'une solution acide (2 m% TCA dans H_2O).

Permet de cliver le dernier groupement DMT présent en bout de chaine de chaque séquence.

c) 2 mL d'eau

Permet de laver l'acide et les groupement DMT clivés.

d) 1 mL de solution d'élution (0,5 v% NH3 (28% dans H_2O) dans une soliton à 1 : 1, H_2O : MeCN).

La solution est ensuite collectée en plusieurs fractions. Les premières gouttes de contiennent généralement pas le polymère désiré, néanmoins toutes les fractions montrant une absorption UV sont lyophilisées et analysées.

3.5. Analyses des polymères par spectrométrie de masse

3.5.1. Méthode de séquençage :

La méthodologie pseudo-MS^3 développée pour recouvrer la séquence entière à partir de segments tronqués du polymère numérique repose sur trois étapes expérimentale principale, en utilisant le mode ESI négatif.

Etape 1 : En mode MS, la distribution des états de charge du polymère indique le nombre de ses segments. En effet, il a été remarqué qu'en moyenne des blocs contenant 8 unités codantes sont observés avec 3 charges. Ainsi, l'état de charge z préférentiel est le plus souvent un multiple de 3, duquel il est facile de déterminer le nombre n de segment comme n=z/3. Par exemple pour un polymère dont l'abondance maximale est obtenue pour un état de charge à z=15 indique qu'il y a n=z/3 = 5 segments. Cet ion est donc choisi pour l'étape 2.

Etape 2 : Une analyse CID est effectuée pour l'ion précurseur à l'état de charge z=3n : car les charges négatives sont distribuées uniformément, ceci diminue la complexité des données obtenues en MS/MS. Cette première étape de CID permet de cliver les liaisons alcoxyamines et génère des fragments contenant 1 à n-1 segments avec un état de charge préférentiel qui sera un multiple de 3. Par conséquent les segments simples sont à un état de charge 3- et leur valeur de m/z dépend de leur composition en monomères codants et de leur marqueur de masse. Etant donné que les marqueurs ont été sélectionnés pour fournir une signature de masse unique, chaque fragment 3- peut être assigné uniquement en se basant sur valeur m/z du marqueur. A cette étape il n'est toujours pas possible de déterminer la composition exacte des comonomères. Néanmoins, l'indentification des segments via les marqueurs permet de déterminer leur localisation initiale dans la séquence d'origine. Ainsi pour le polymère à z=15, on va pouvoir retrouver 5 fragments simples à un état de charge 3-. Ces 5 fragments seront catégorisés via le marqueur qu'ils contiennent qui leur donne une masse unique, néanmoins la composition exacte de chaque segment ne peut pas être déterminé à ce niveau.

Etape 3 : La tension appliquée est augmentée pour induire la production de fragments en source. Les segments triplement chargés sont à nouveau analysés en CID. Ainsi les spectres de pseudo-MS ^{3} obtenus sont analysés en se basant sur les règles de dissociation établies pour les poly(phosphodiester)s pour recouvrer chaque segment de la séquence.[25] Ces règles peuvent être rapidement décrite : le clivage des liaisons phosphates fait augmenter la présence des fragments contenant les bouts de chaîne α , fragments nommés : a_i^{z-}, b_i^{z-}, c_i^{z-}, d_i^{z-} ou fait augmenter la présence des fragments contenant les bouts de chaîne ω : w_i^{z-}, x_i^{z-}, y_i^{z-}, z_i^{z-} (cf. Figure SI-II. 47).

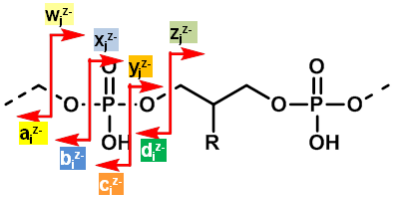


Figure SI-II. 47 : Schéma des fragments obtenus pour chaque monomère lors de l'analyse de pseudo-MS^3.

A cause de la réactivité du radical centré sur le carbone se trouvant dans le bout de chaîne ω, tous les ions contenant un seul segment (excepté le dernier) sont immédiatement dissociés à nouveau, compromettant ainsi la couverture complète de la séquence. Cependant, en considérant seulement les fragments contenant le bout de chaîne α, il est possible de recouvrer la séquence entièrement. Les ions issus de la fragmentation d_i^{z-} sont généralement tous recouvrés et comme ils correspondent au monomère en entier on va s'attendre pour le premier fragment d_1^{1-} le m/z correspondant à chaque monomère impliqué dans la séquence. Ainsi pour une séquence codée avec l'alphabet dyade, le fragment d_1^{1-} est attendu à *m/z* 155.0, *m/z* 169.0, *m/z* 183.0 ou *m/z* 197.1 en fonction du premier monomère codant du segment M1, M2, M3 ou M4 respectivement. En fonction de la valeur du pic trouvé, le début de la séquence sera connu. Ensuite, les autres membres de la série d_i^{z-} sont recherchés en ajoutant la masse de chaque unité différente (138.0 Da pour M1, 152.0 Da M2, 166.0 Da pour M3, 180.1 Da pour M4) au m/z trouvé au préalable. Les autres fragments α sont attendus à des m/z tels que $c_i^{z-} = d_i^{z-} - 18$ Da, $b_i^{z-} = c_i^{z-} - 62$ Da, et $a_i^{z-} = b_i^{z-} - 18$ Da. Avec toutes les informations récoltées il sera alors possible de retrouver et confirmer le bon ordre des monomères dans chaque segment.

Finalement, tous les segments sont ordonnés en fonction des règles définies par le système des marqueurs révélant ainsi la séquence complète du polymère.

3.5.2. Polymères obtenus avec l'alphabet linéaire

- **Séquences linéaires tests**
  - **C4**

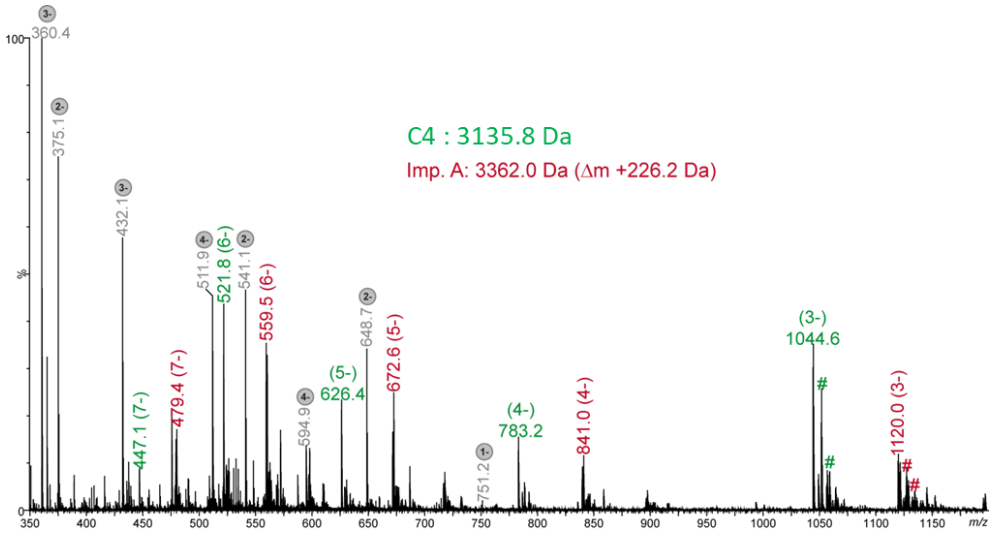


Figure SI-II. 48 : Spectre ESI-MS en mode négatif de la séquence test C4. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux annotés en gris à des fragments formés en source.

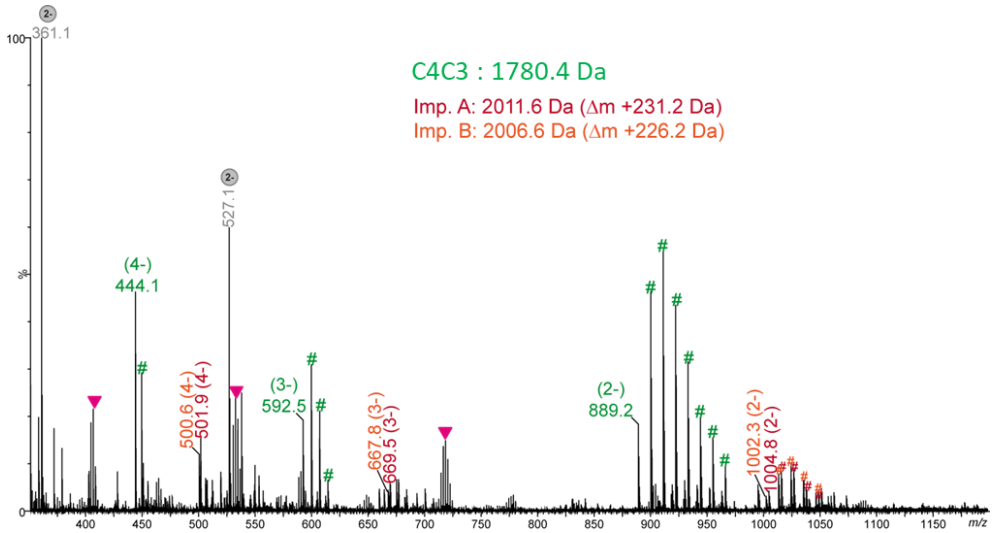


Figure SI-II. 49 : Spectre ESI-MS en mode négatif de la séquence test C4C3. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux annotés en gris à des fragments formés en source. Triangles rose correspondent à des agrégats d'acide trichloroacétique.

- **C5**

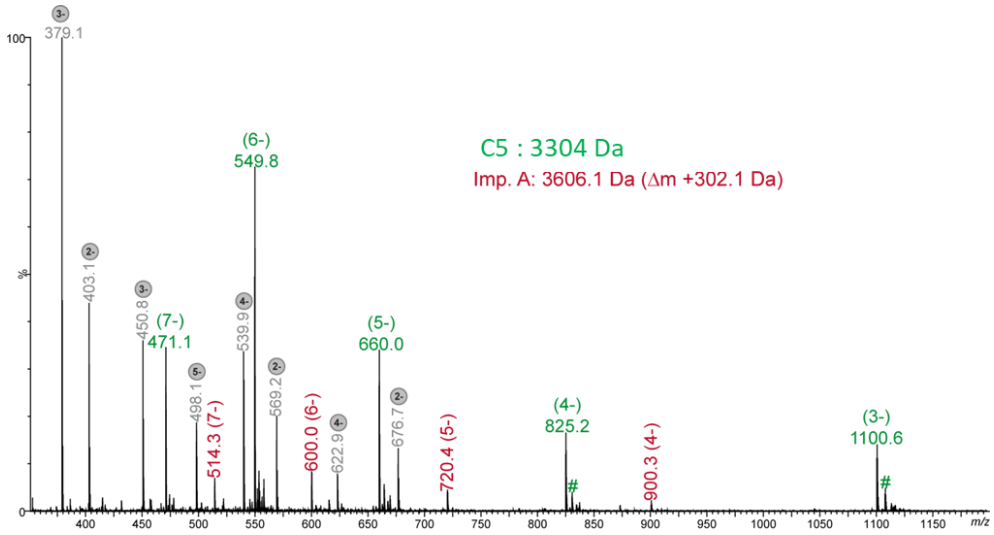


Figure SI-II. 50 : Spectre ESI-MS en mode négatif de la séquence test C5. Les pics annotés en gris correspondent à des fragments formés en source.

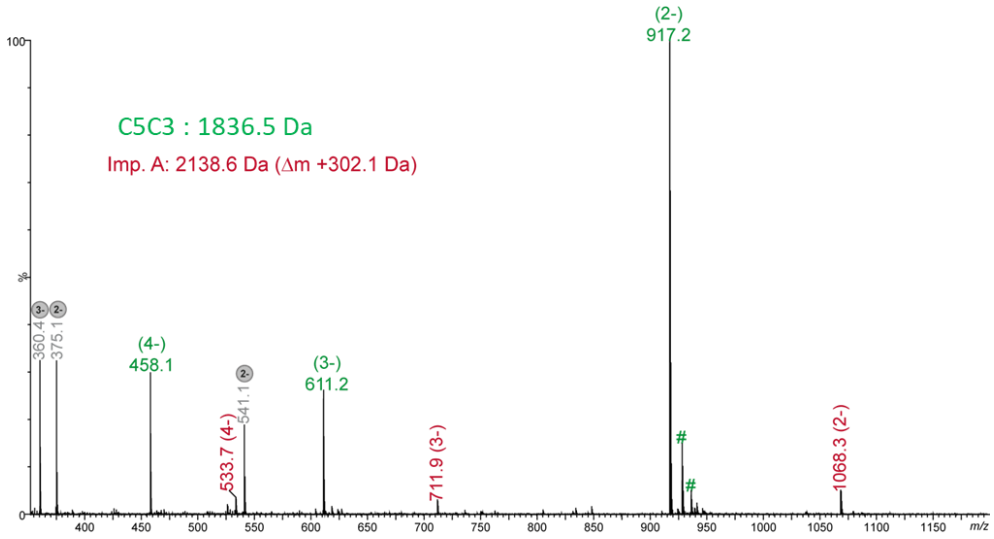


Figure SI-II. 51 : Spectre ESI-MS en mode négatif de la séquence test C5C3. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

- **C6**

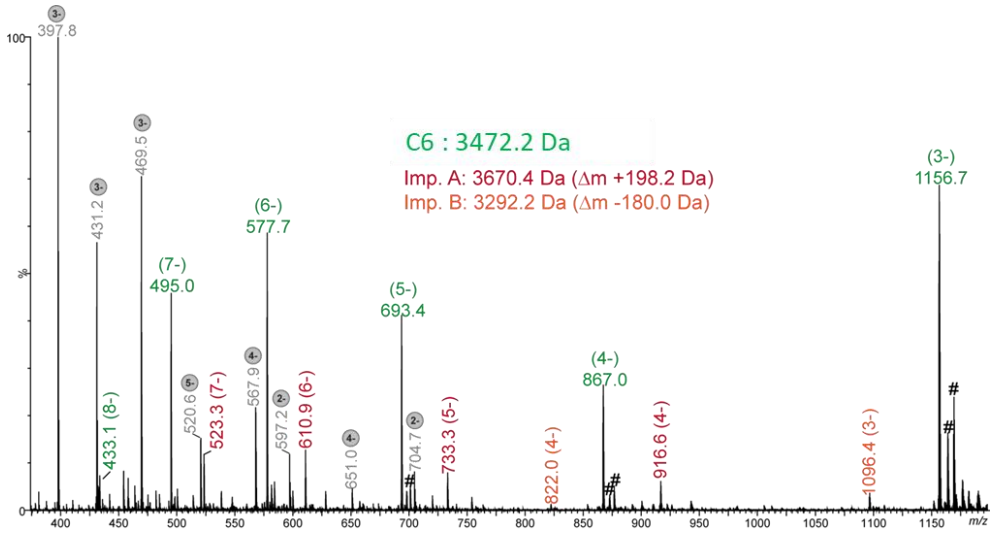


Figure SI-II. 52 : Spectre ESI-MS en mode négatif de la séquence test C6. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

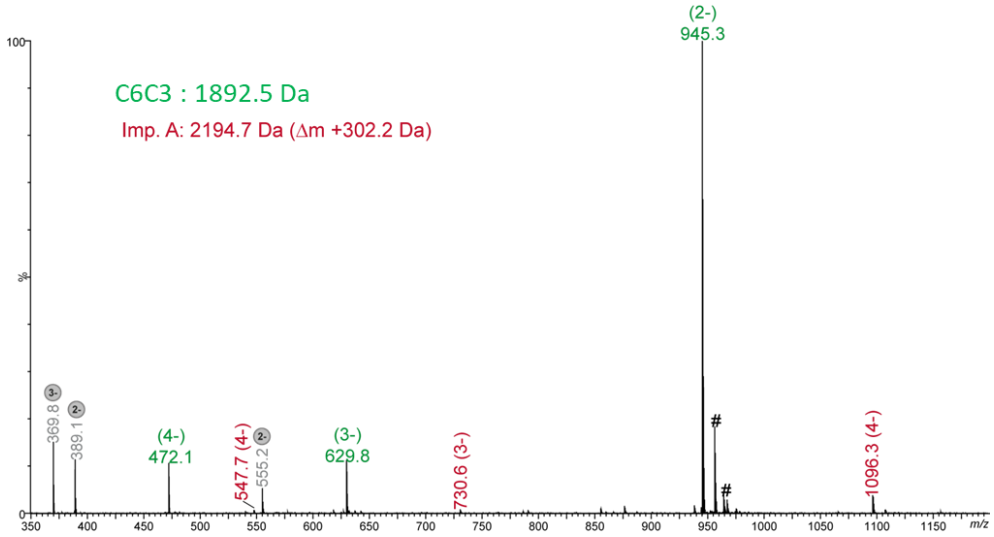


Figure SI-II. 53 : Spectre ESI-MS en mode négatif de la séquence test C6C3. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

- **C7**

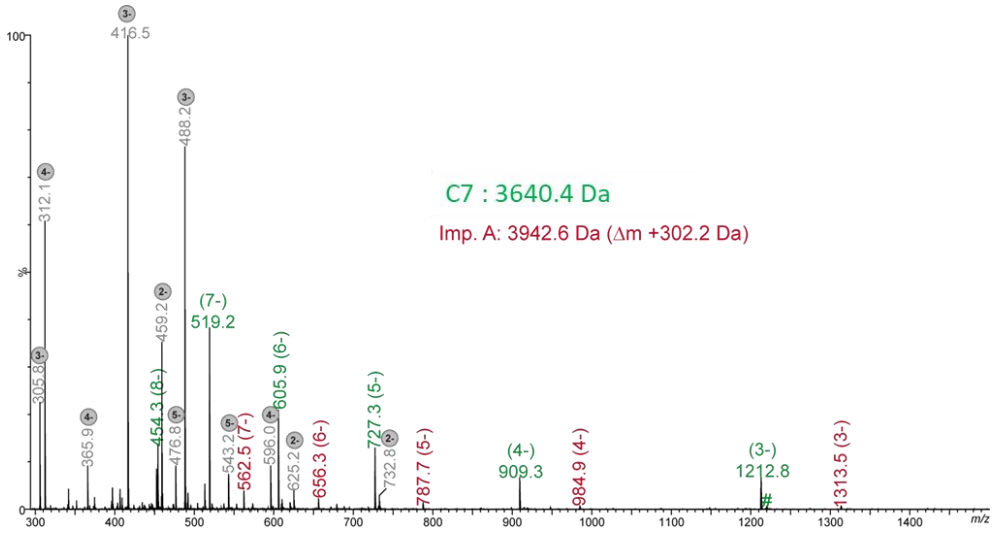


Figure SI-II. 54 : Spectre ESI-MS en mode négatif de la séquence test C7. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

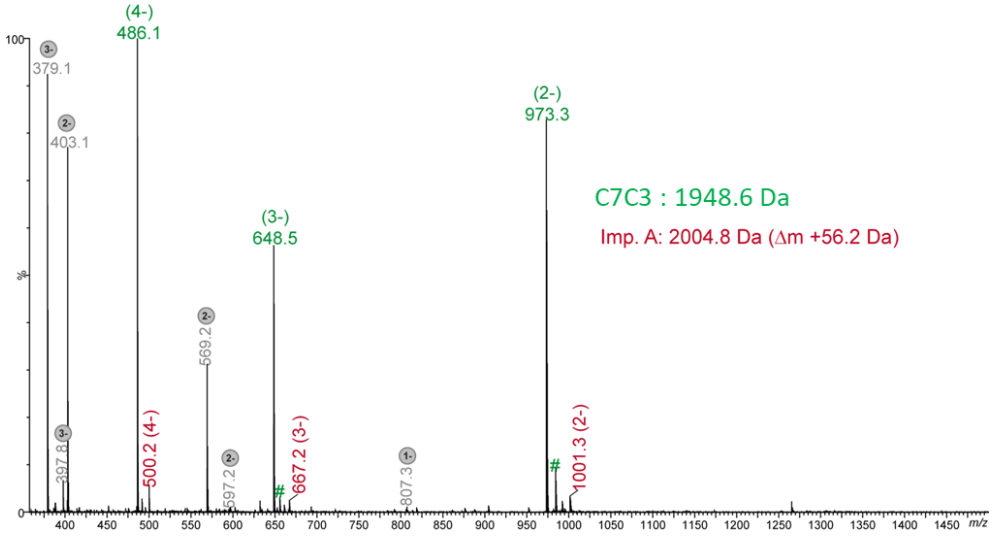


Figure SI-II. 55 : Spectre ESI-MS en mode négatif de la séquence test C7C3. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

- **C8**

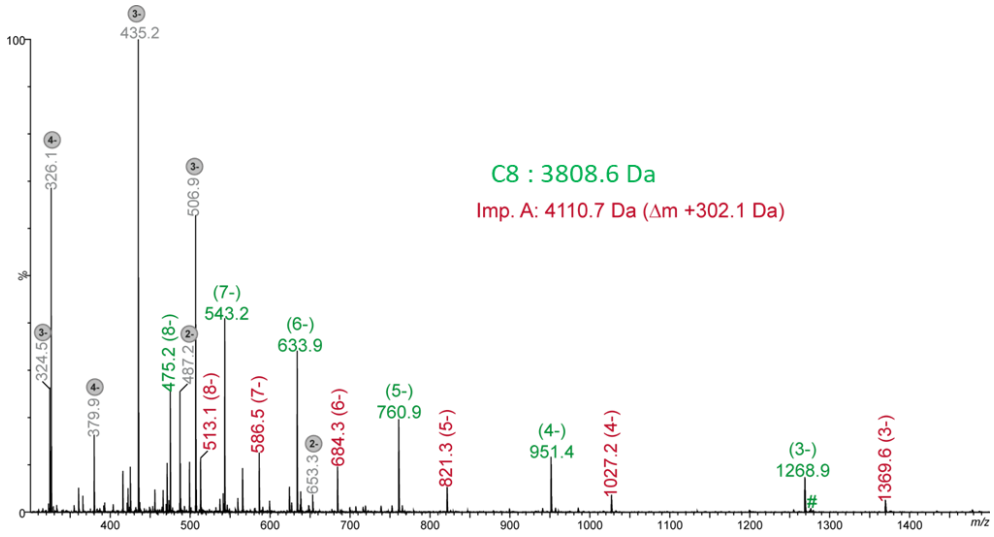


Figure SI-II. 56 : Spectre ESI-MS en mode négatif de la séquence test C8. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

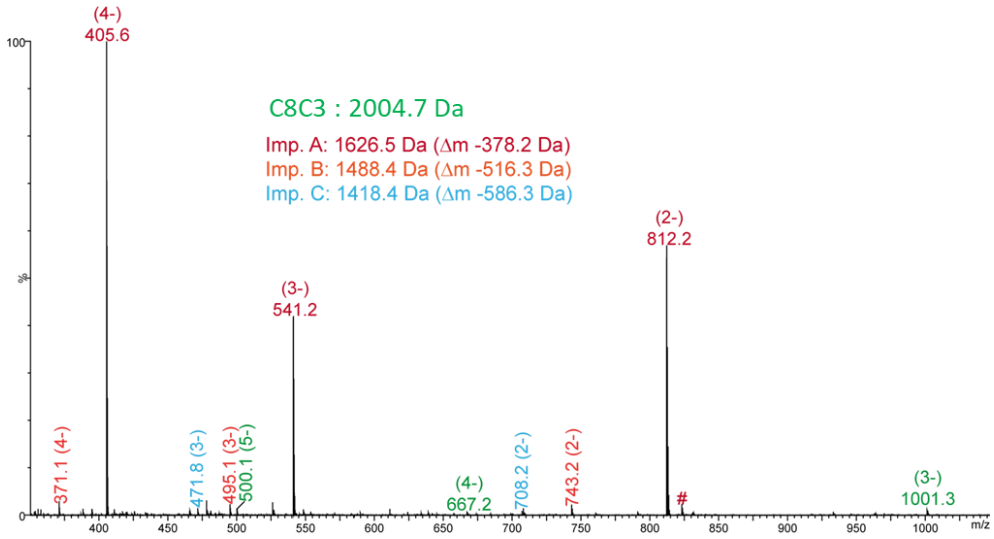


Figure SI-II. 57 : Spectre ESI-MS en mode négatif de la séquence test C8C3. Les pics annotés par # correspondent à des échanges H/Na ou H/K.

- **C9**

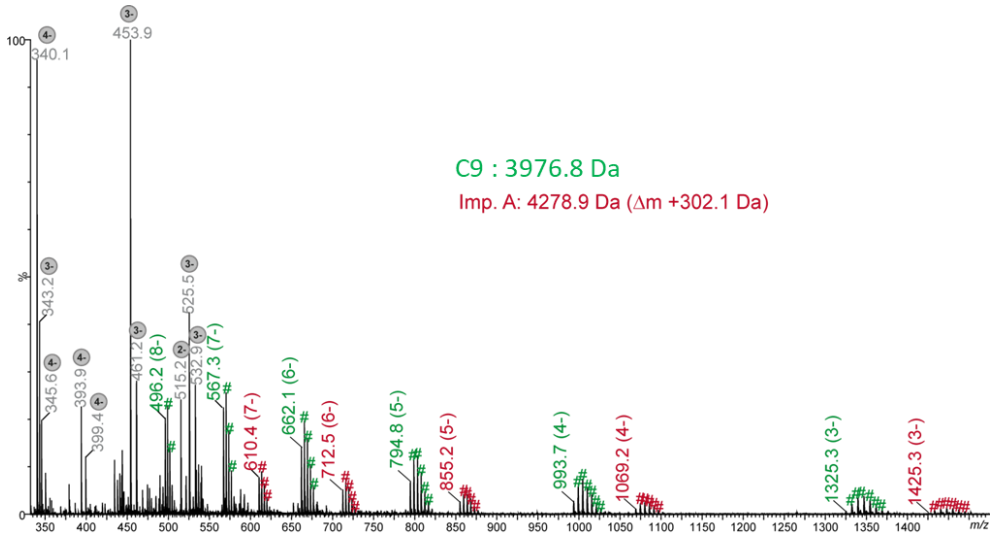


Figure SI-II. 58 : Spectre ESI-MS en mode négatif de la séquence test C9. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

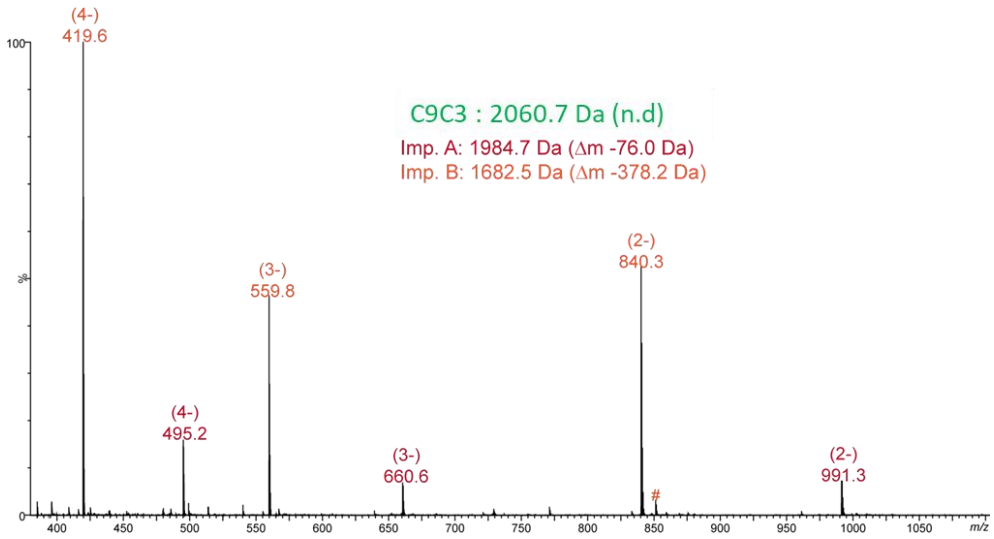


Figure SI-II. 59 : Spectre ESI-MS en mode négatif de la séquence test C9C3 (non détectée). Les pics annotés par # correspondent à des échanges H/Na ou H/K.

- **C10**

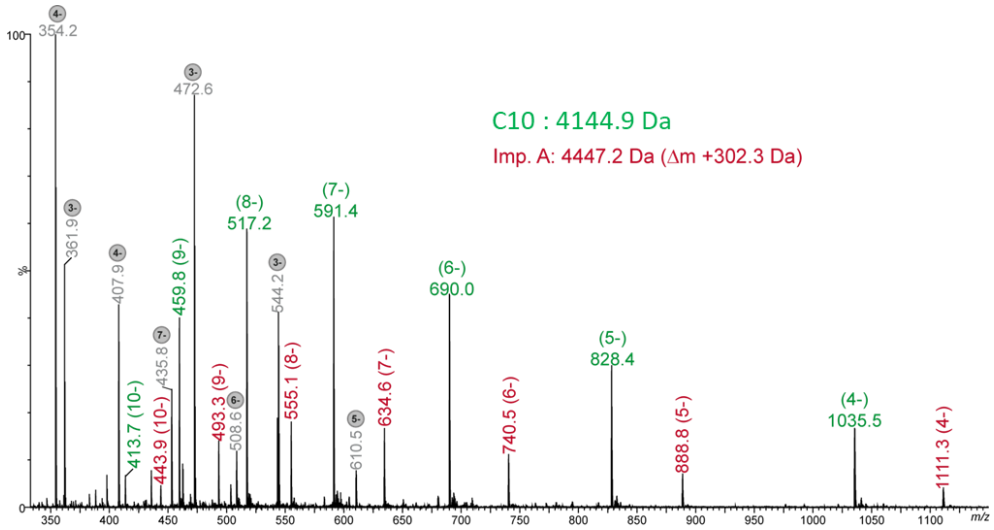


Figure SI-II. 60 : Spectre ESI-MS en mode négatif de la séquence test C10. Les pics annotés en gris correspondent à des fragments formés en source.

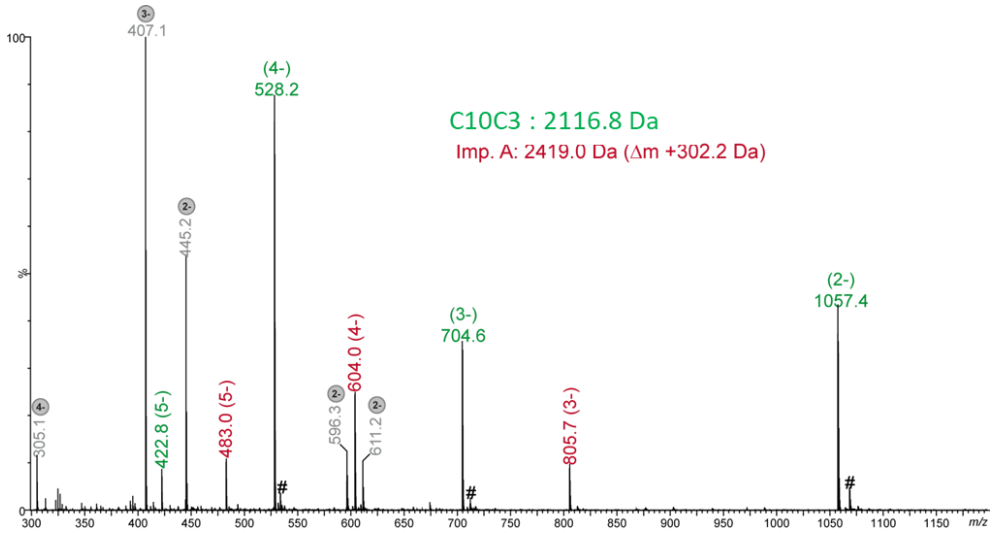


Figure SI-II. 61 : Spectre ESI-MS en mode négatif de la séquence test C10C3. Les pics annotés par # correspondent à des échanges H/Na ou H/K et ceux en gris correspondent à des fragments formés en source.

3.5.3. Polymères obtenus avec l'alphabet à 4 symboles de première génération

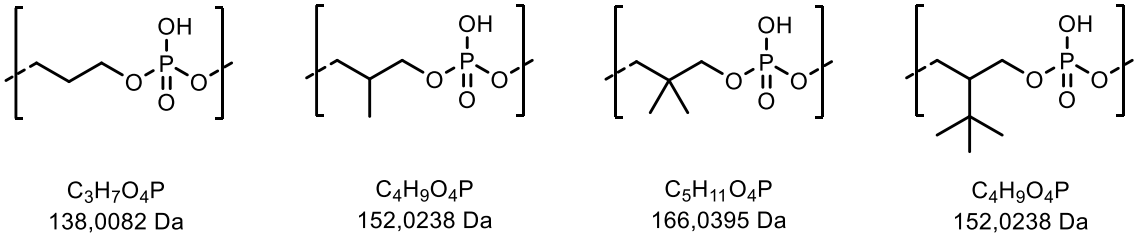


Figure SI-II. 62 : Structure des monomères de l'alphabet à 4 symboles première génération lorsqu'ils sont incorporés dans un polymère.

- **Séquences tests**

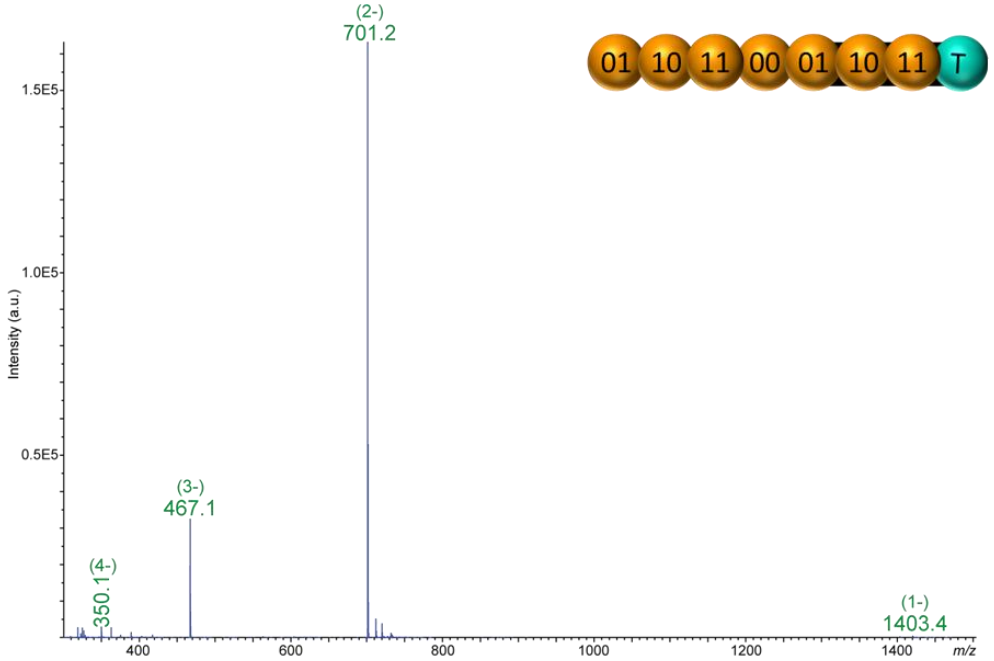


Figure SI-II. 63 : Spectre ESI-MS en mode négatif de l'échantillon contenant la séquence test à 1404,3668 Da. L'oligomère recherché est observé sous les formes $[M-H]^-$ à m/z 1403,4, $[M-2H]^{2-}$ à m/z 701,2, $[M-3H]^{3-}$ à m/z 467,1 et $[M-4H]^{4-}$ à m/z 350,1. Les pics au pied de l'ion m/z 701,2 correspondent à des échanges H/Na et H/K.

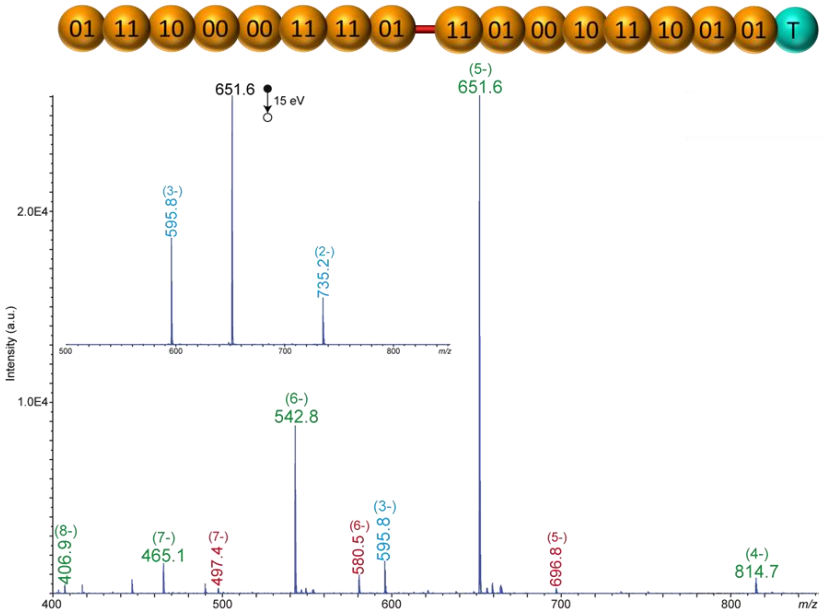


Figure SI-II. 64 : Spectre ESI-MS en mode négatif de l'échantillon contenant le séquence test impliquant un espaceur. L'oligomère recherché est observé sous les formes $[M-4H]^{4-}$ à m/z 814,7, $[M-5H]^{5-}$ à m/z 651,6, $[M-6H]^{6-}$ à m/z 542,8, $[M-7H]^{7-}$ à m/z 465,1 et $[M-8H]^{8-}$ à m/z 406,9. Les pics au pied des ions m/z 651,6 et m/z 543,8 correspondent à des échanges H/Na et H/K. Le pic annoté en bleu correspond à un fragment interne (octet 2). Inset : spectre ESI-MS/MS de l'ion $[M-5H]^{5-}$ à m/z 651,6. Les pics annotés en rouge correspondent aux 3 états de charge d'une molécule de masse 3489,1 Da (Δm = +227 Da).

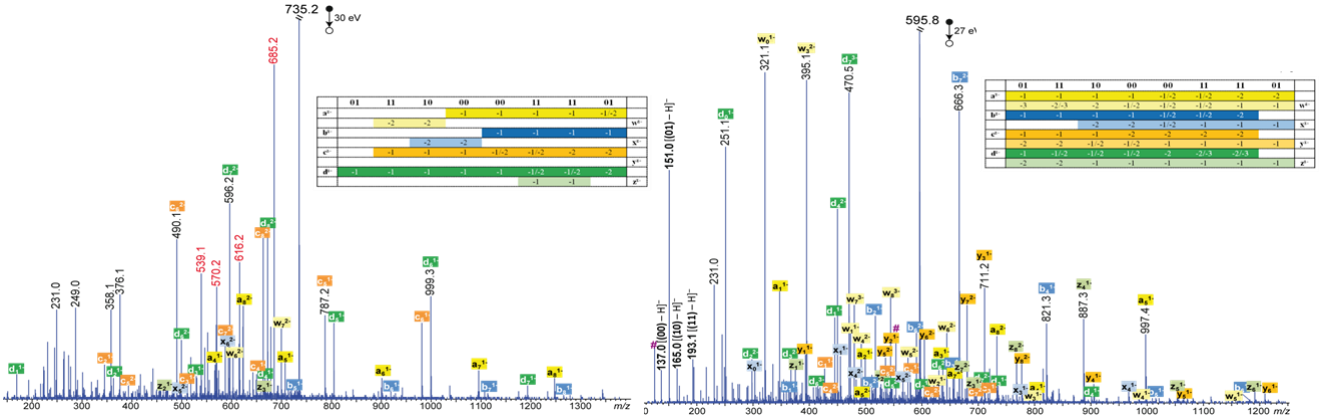


Figure SI-II. 65 : Analyse de pseudo-MS^3 de la séquence test comprenant un espaceur. (Gauche) Spectre pseudo-MS^3du bloc 1 à m/z 735,2 et couverture de la séquence. (Droite) Spectre pseudo-MS^3du bloc 2 à m/z 595,8 et couverture de la séquence. (Les pics annotés par # sont issus de la perte du marqueur T.

- **P1**

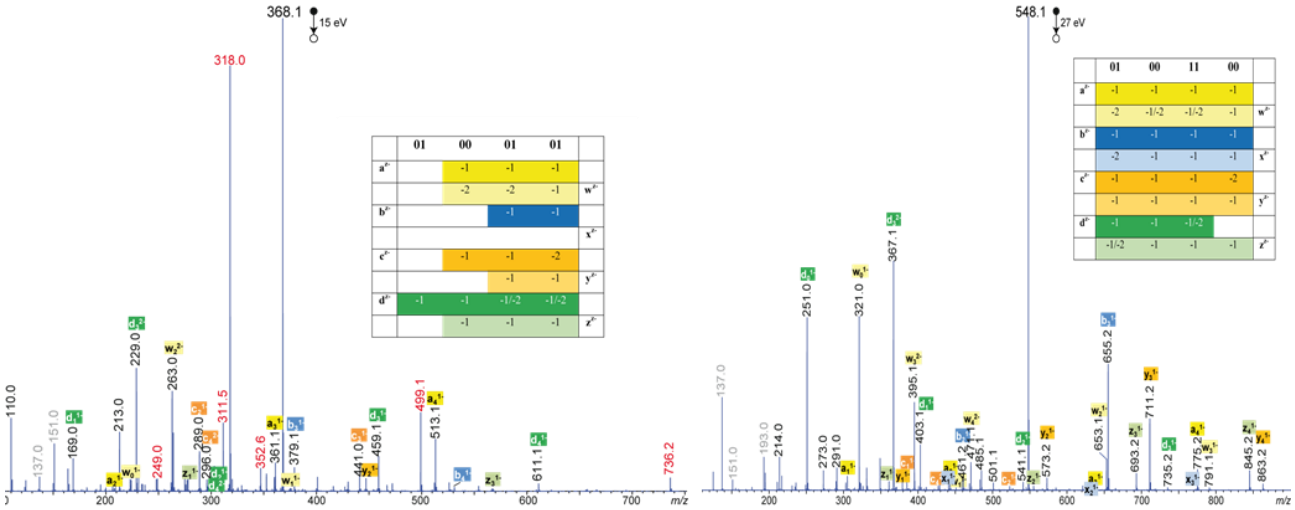


Figure SI-II. 66 : Analyse de pseudo-MS³ de la séquence P1. (Gauche) Spectre pseudo-MS³du bloc 1 à m/z 368,1 et couverture de la séquence. (Droite) Spectre pseudo-MS³du bloc 2 à m/z 548,1 et couverture de la séquence.

- **P2**

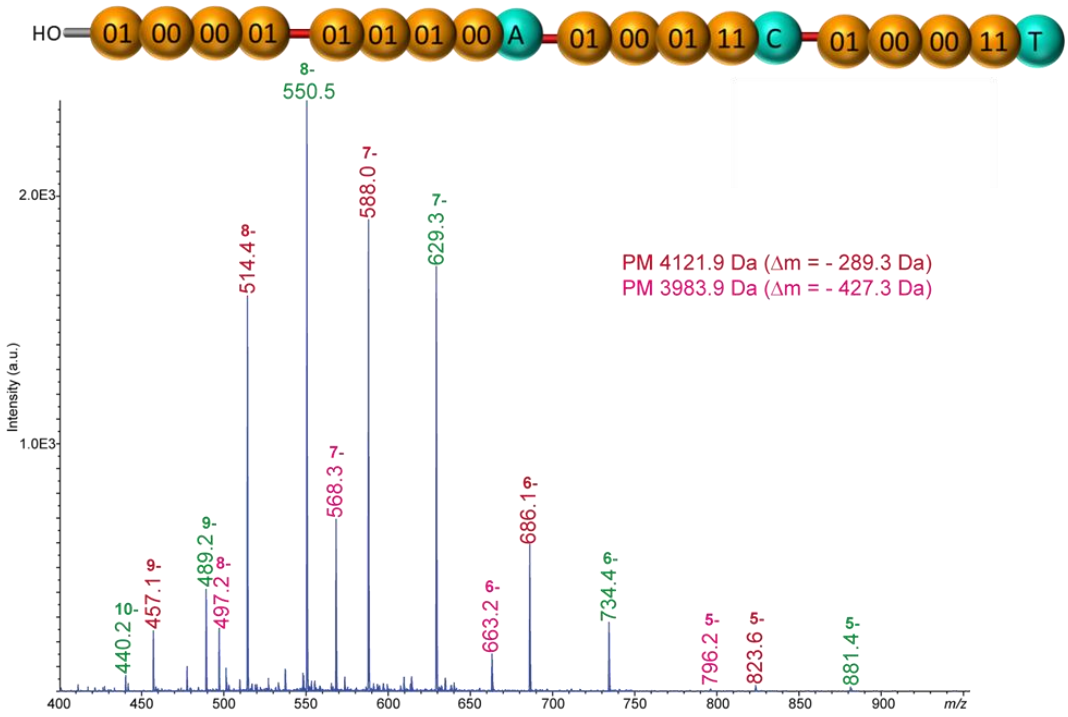


Figure SI-II. 67 : Spectre ESI-MS en mode négatif de P2. Le polymère recherché est détecté sous les états de charges -5 à -10 (en vert). Les pics annotés en rouge correspondent aux états de charge d'une molécule de masse 4121,9 Da (Δm = -289,3 Da), ceux en rose correspondent aux états de charge d'une molécule de masse 3983,9 Da (Δm = -427,3 Da).

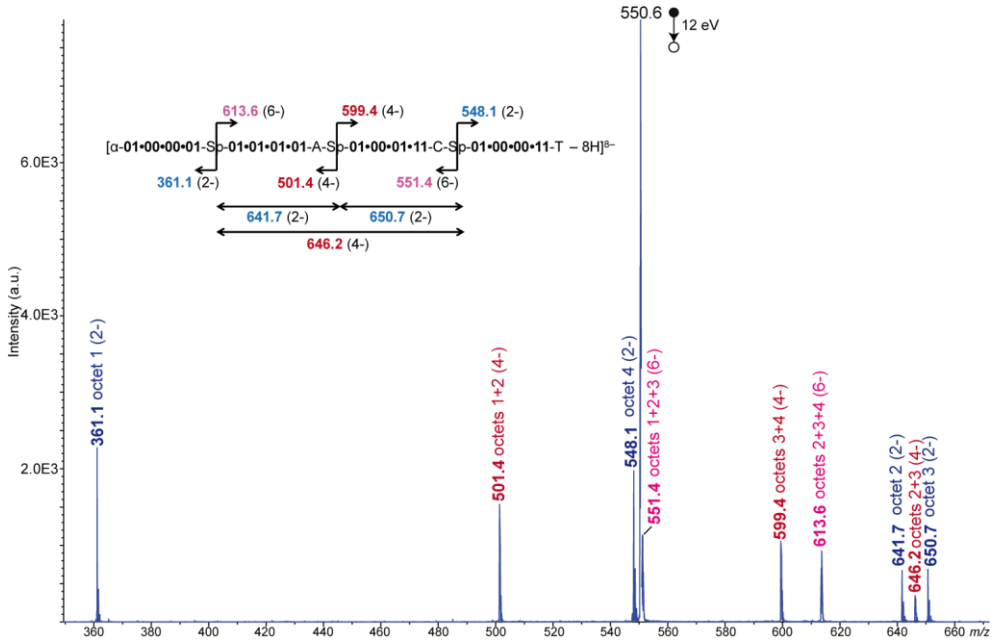


Figure SI-II. 68 : Spectre ESI-MS/MS de l'ion $[M - 8H]^{-}$ à m/z 550,7 dans P2.

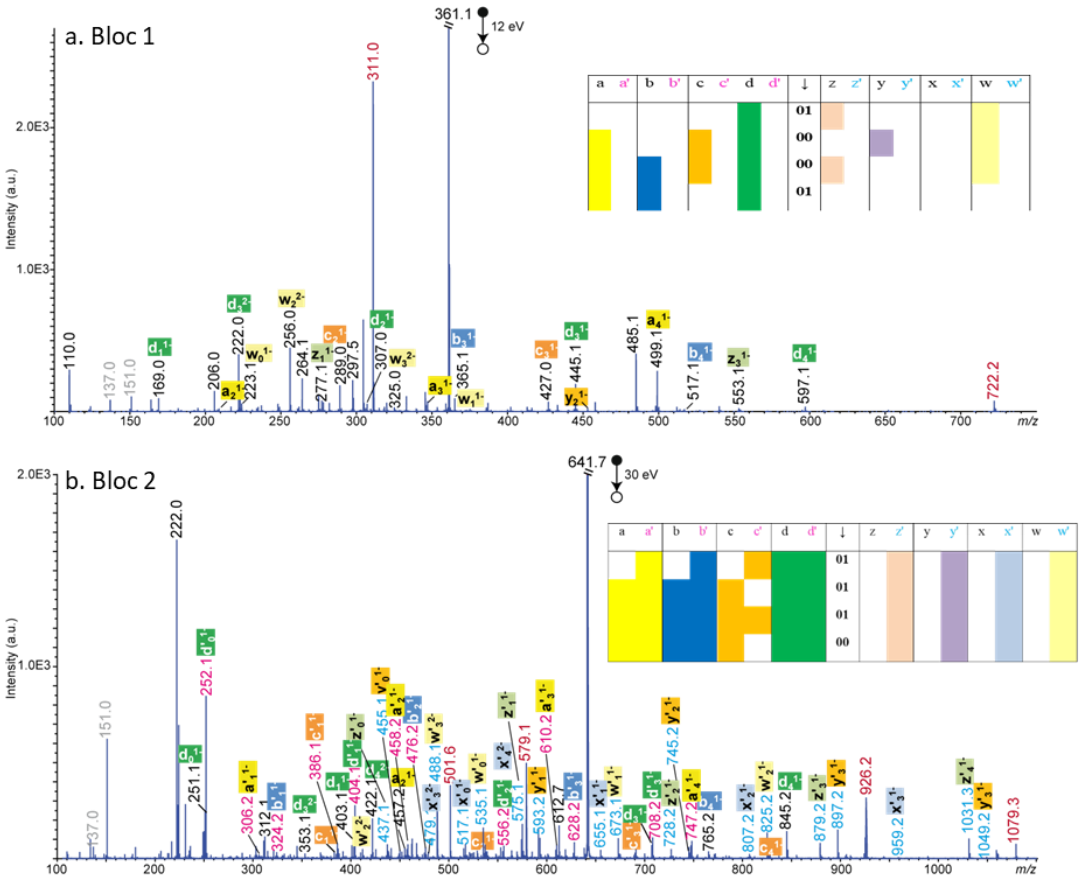


Figure SI-II. 69 : Analyse de pseudo-MS³ de la séquence P2. (a) Spectre pseudo-MS³du bloc 1 à m/z 361,1 et couverture de la séquence. (b) Spectre pseudo-MS³du bloc 2 à m/z 641,7 et couverture de la séquence. En gris : monomères déprotonés. En rouge : fragments issus d'autres dissociations que les ruptures de liaisons phosphate.

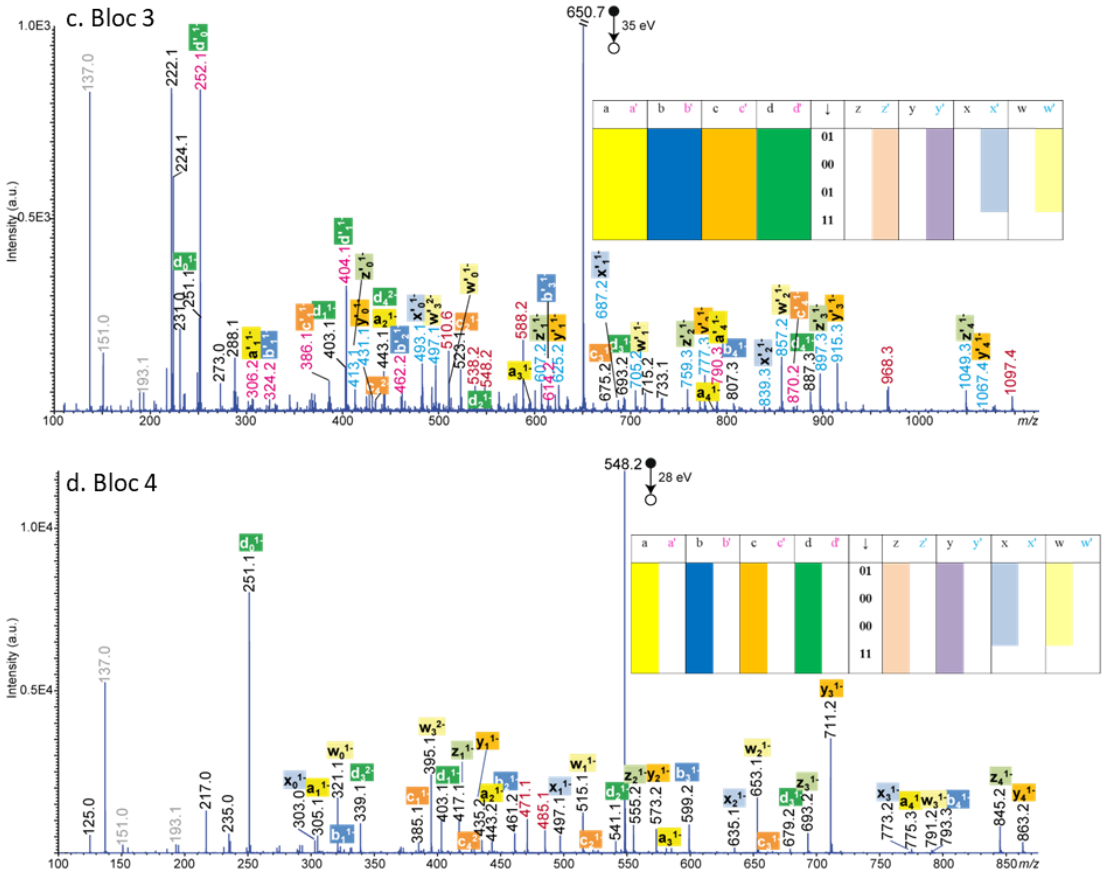


Figure SI-II. 70 : Suite de l'analyse de pseudo-MS³ de la séquence P2. (c) Spectre pseudo-MS³du bloc 3 à m/z 650,7 et couverture de la séquence. (d) Spectre pseudo-MS³du bloc 2 à m/z 548,2 et couverture de la séquence. En gris : monomères déprotonés. En rouge : fragments issus d'autres dissociations que les ruptures de liaisons phosphate.

- **Séquence plus longue non détectée**

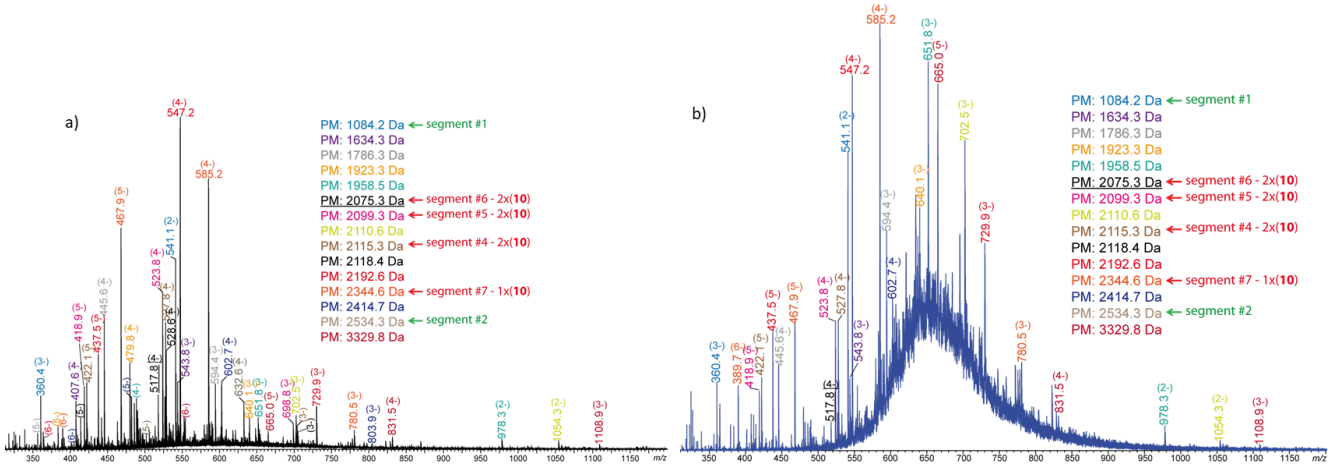


Figure SI-II. 71 : (a) Spectre ESI-MS en mode négatif de l'échantillon enregistré avec le spectromètre Synapt G2. Il ne montre aucun signal présentant des états de charge nécessaires à la détection de la macromolécule (typiquement entre 15 et 35). Inset : Liste des impuretés détectées et leur possible interprétation. (b) Spectre ESI-MS en mode négatif enregistré avec le spectromètre QStar Elite.

3.5.4. Polymères obtenus avec l'alphabet à 4 symboles de seconde génération

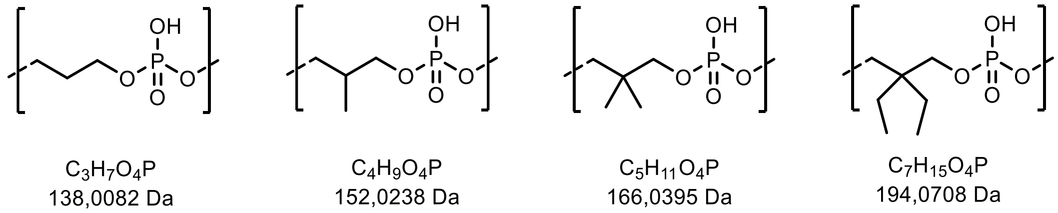


Figure SI-II. 72 : Structure des monomères de l'alphabet à 4 symboles seconde génération lorsqu'ils sont incorporés dans un polymère.

- **P3**

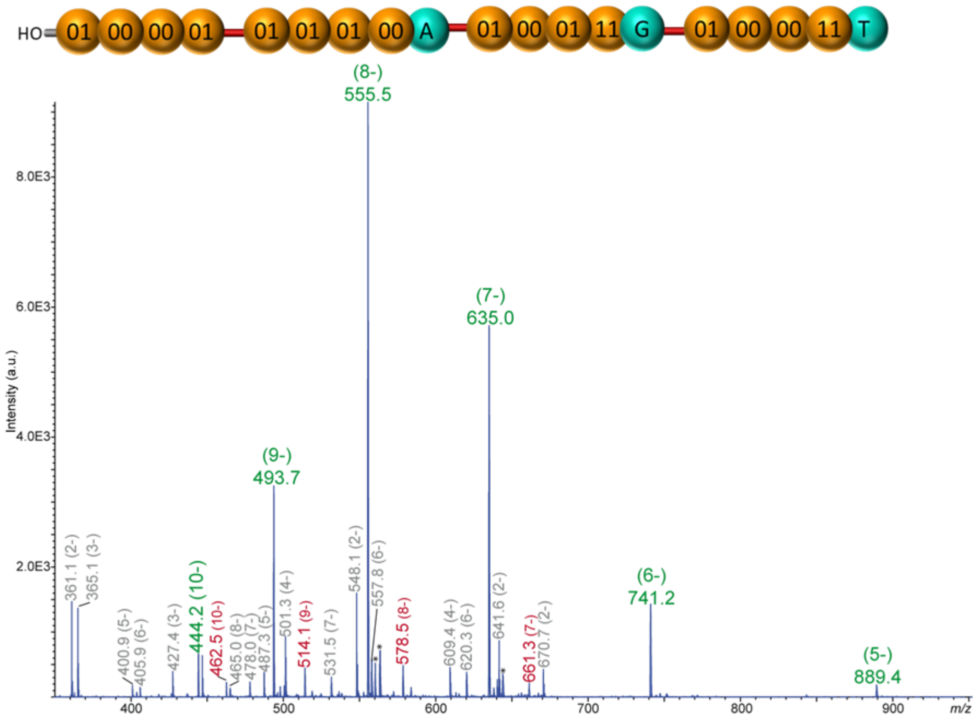


Figure SI-II. 73 : Spectre ESI-MS en mode négatif de P3. L'oligophosphate est observé sous les états de charge -5 à -10 (en vert) ainsi que des fragments formés en source après différents clivages de liaisons alcoxyamine (en gris). Une impureté mineure est également observée avec un excès de masse de 183,8 Da (en rouge).

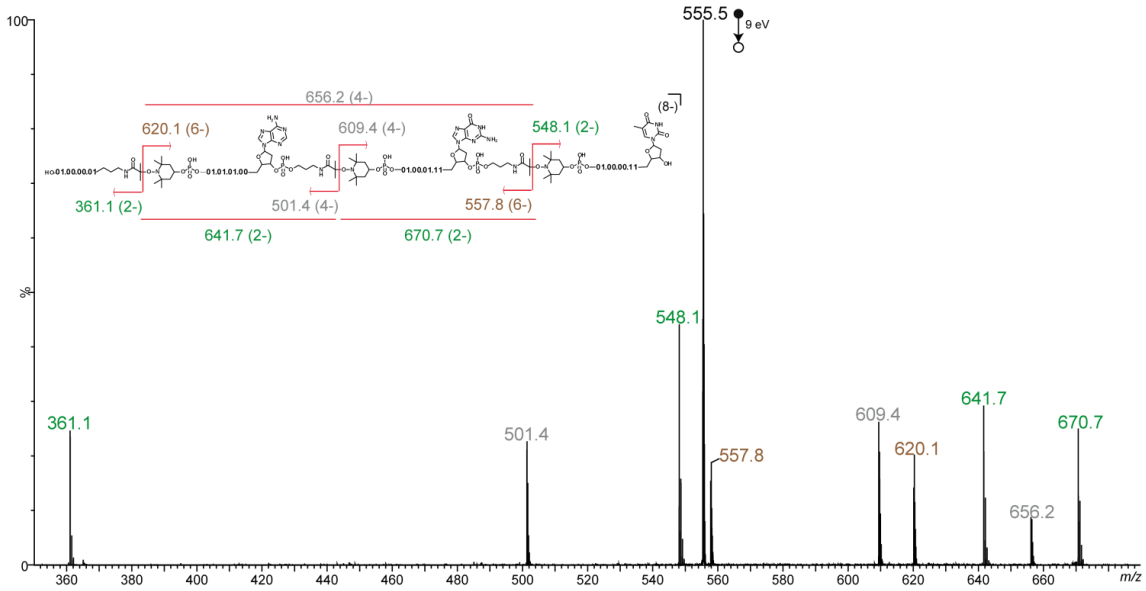


Figure SI-II. 74 : Spectre ESI-MS/MS de l'ion $[M–8H]^{8–}$ à m/z 555,5 et schéma de fragmentation.

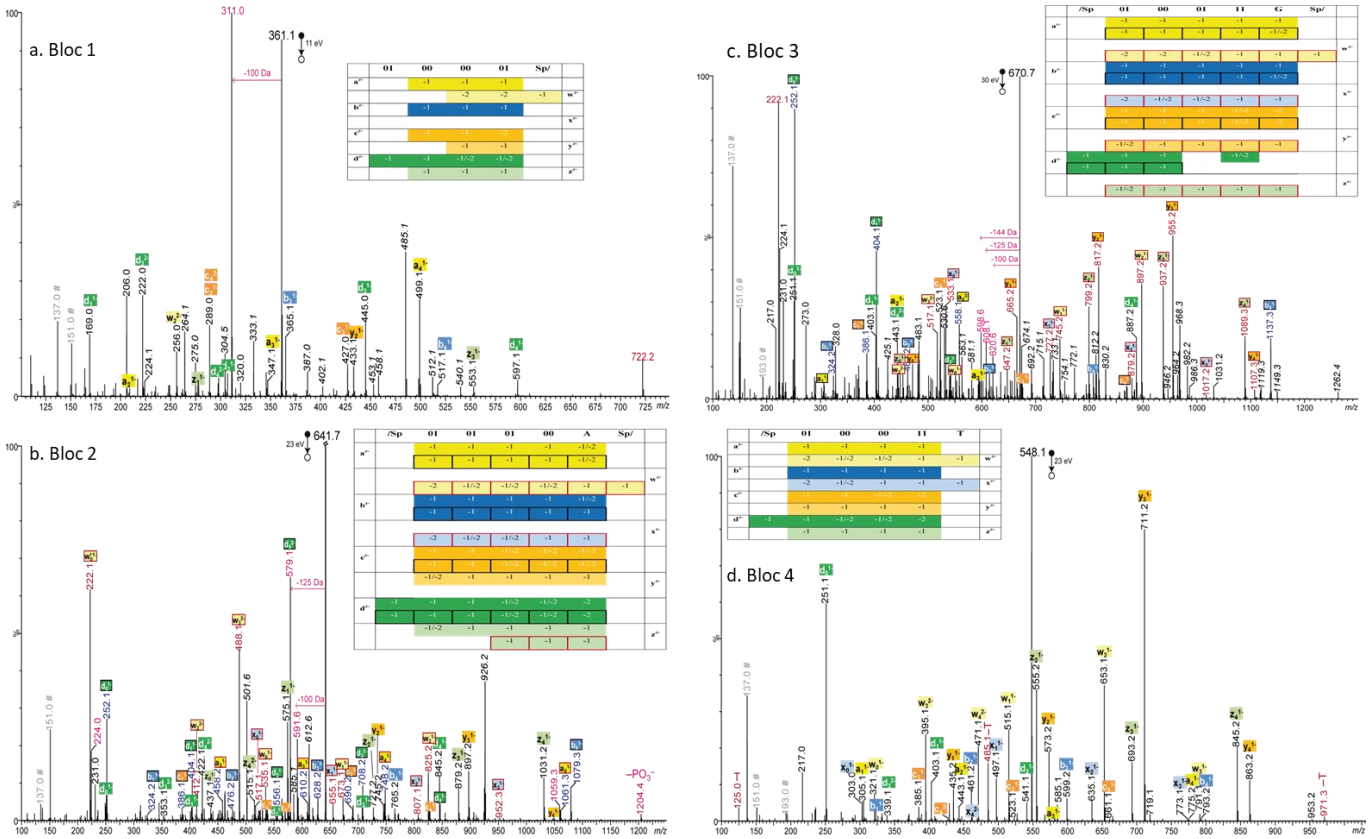


Figure SI-II. 75 : Analyse de pseudo-MS³ de la séquence P3. (a) Spectre pseudo-MS³du bloc 1 à m/z 361,1 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 2 à m/z 641,7 et couverture de la séquence. (c) Spectre pseudo-MS³du bloc 3 à m/z 670,7 et couverture de la séquence. (d) Spectre pseudo-MS³du bloc 4 à m/z 548,1 et couverture de la séquence. Les ions encadrés en rouge sont des fragments obtenus après abstraction d'un H• et les ions encadrés en bleu sont des fragments obtenus après gain d'un H•. Fragments non attribués en italique. En gris : monomères déprotonés. En rouge : fragments issus d'autres dissociations que les ruptures de liaisons phosphate.

- **P4**

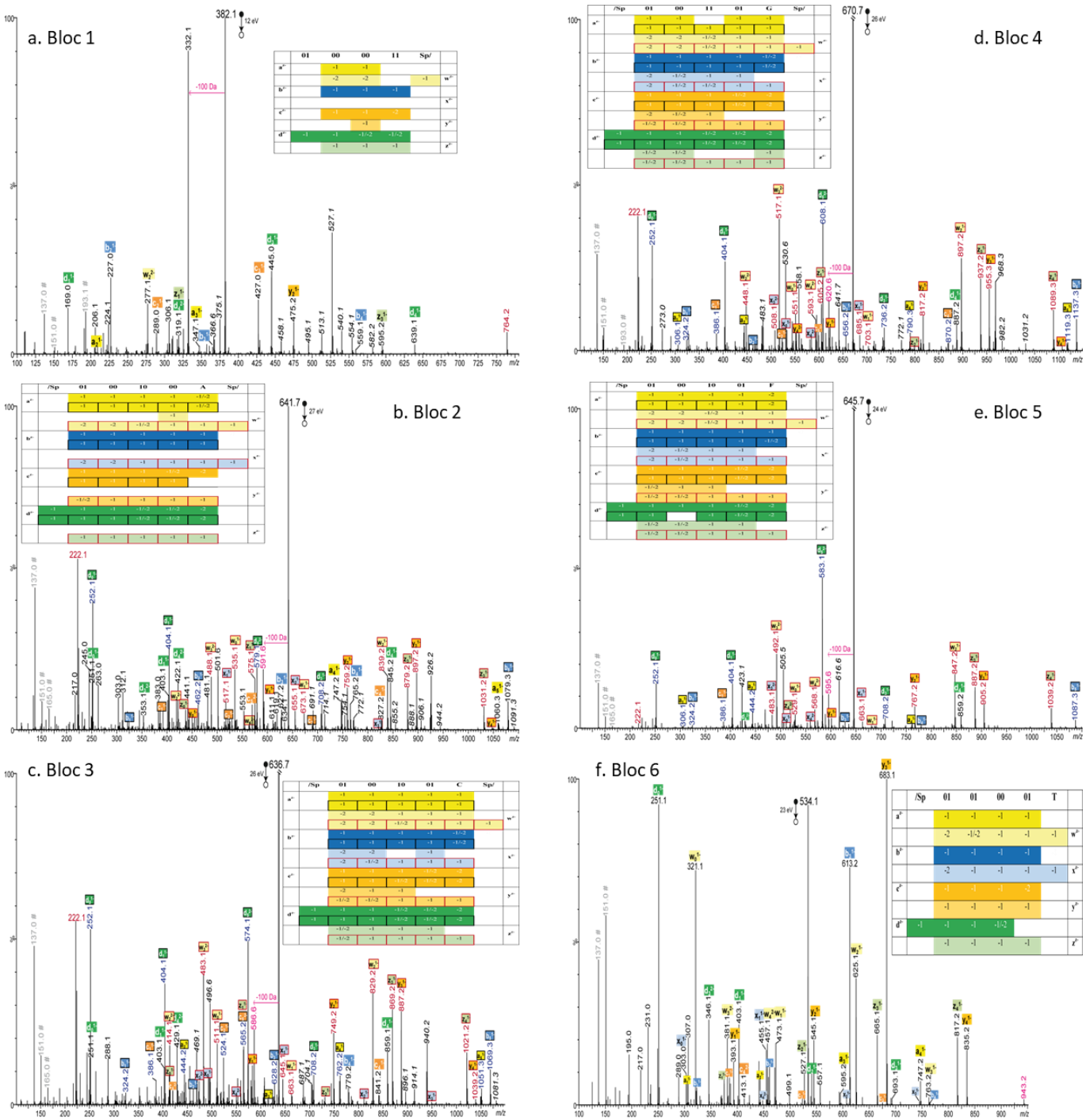


Figure SI-II. 76 : Analyse de pseudo-MS³ de la séquence P4. (a) Spectre pseudo-MS³du bloc 1 à m/z 382,1 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 2 à m/z 641,7 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 3 à m/z 636,7 et couverture de la séquence. (d) Spectre pseudo-MS³ du bloc 4 à m/z 670,7 et couverture de la séquence. (e) Spectre pseudo-MS³du bloc 5 à m/z 645,7 et couverture de la séquence. (f) Spectre pseudo-MS³ du bloc 6 à m/z 534,2 et couverture de la séquence. Les ions encadrés en rouge sont des fragments obtenus après abstraction d'un H• et les ions encadrés en bleu sont des fragments obtenus après gain d'un H•. Fragments non attribués en italique. En gris : monomères déprotonés. En rouge : fragments issus d'autres dissociations que les ruptures de liaisons phosphate

.

3.5.5. Polymères obtenus avec l'alphabet à 8 symboles

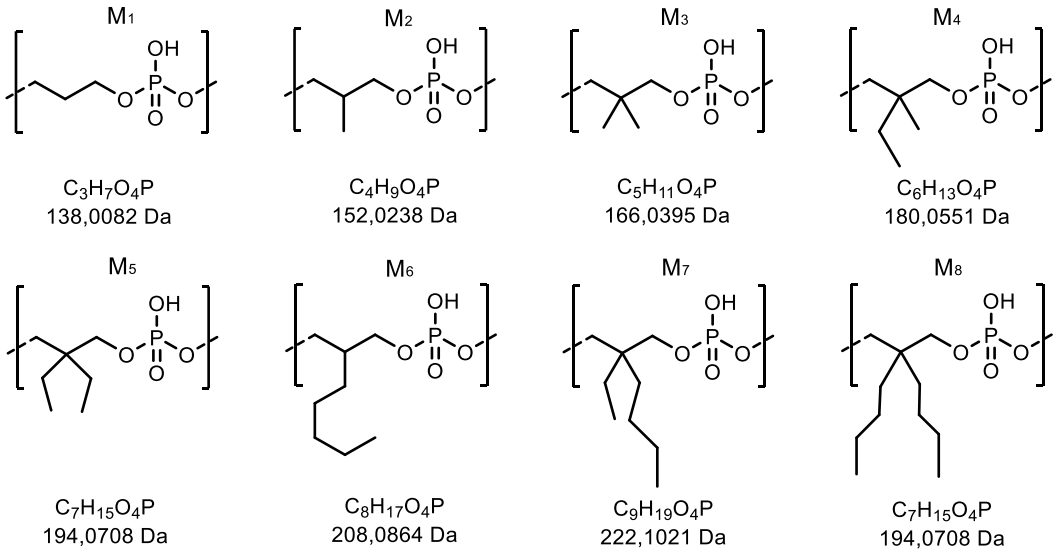


Figure SI-II. 77 : Structure des monomères codant du texte avec l'alphabet à 8 symboles lorsqu'ils sont incorporés dans une séquence.

- **P5**

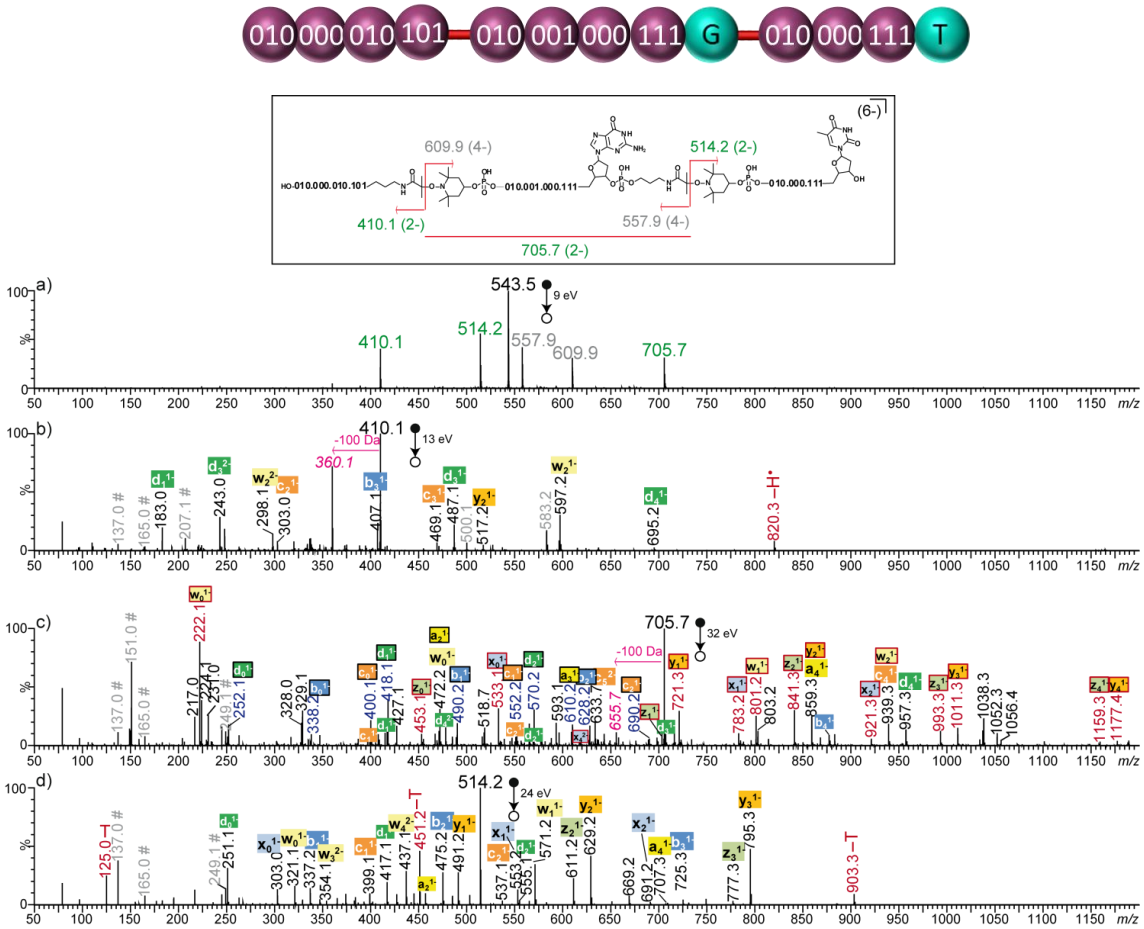
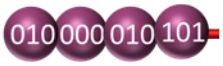


Figure SI-II. 78 : Séquençage de P5. (a) Spectre ESI-MS/MS de l'ion $[M–6H]^{6–}$ à m/z 543,5 et schéma de fragmentation. (b) Spectre pseudo-MS^3 du bloc 1 à m/z 410,1. Dans ce spectre, on note l'élimination d'un radical H• à partir de l'ion précurseur pour engendrer l'ion m/z 820,3. (c) Spectre pseudo-MS^3 du bloc 2 à m/z 705,7. Les ions encadrés en rouge sont des fragments obtenus après abstraction d'un H• et les ions encadrés en bleu sont des fragments obtenus après gain d'un H•. (d) Spectre pseudo-MS^3 du bloc 3 à m/z 514,2. Les fragments secondaires (dont les monomères déprotonés) sont annotés en gris.

a) Bloc 1 

	010	000	010	101	E/	
a^{z-}		-1				
			-1/-2	-1	-1	w^{z-}
b^{z-}		-1	-1	-1		
						x^{z-}
c^{z-}		-1	-1	-2		
			-1			y^{z-}
d^{z-}	-1		-1/-2	-1/-2		
		-1		-1		z^{z-}

b) Bloc 2 

	/E	010	001	000	111	G	E/	
a^{z-}			-1	-1	-1	-1		
			-1	-1		-1		
				-1		-1		w^{z-}
		-2	-2	-1/-2	-1	-1	-1	
b^{z-}		-1	-1	-1	-1	-1		
		-1	-1	-1	-1	-1/-2		
								x^{z-}
		-2	-2	-1	-1	-1		
c^{z-}		-1	-1	-1/	-1/-2	-2		
		-1	-1	-1	-1	-2		
						-1		y^{z-}
		-1/-2	-1	-1	-1	-1		
d^{z-}	-1	-1	-1	-1	-1/-2	-2		
	-1	-1	-1			-2		
								z^{z-}
		-1	-1	-1	-1	-1		

c) Bloc 3 

	/E	010	000	111	T	
a^{z-}			-1	-1		
		-2	-1/-2	-1/-2	-1	w^{z-}
b^{z-}		-1	-1	-1		
		-2	-1	-1	-1	x^{z-}
c^{z-}		-1	-1	-1/-2		
		-1	-1	-1		y^{z-}
d^{z-}	-1	-1	-1/-2	-1/-2		
		-1	-1	-1		z^{z-}

Figure SI-II. 79 : Tableaux de couverture des différents blocs de P5. (a) bloc 1, (b) bloc 2 et (c) bloc 3.

- **P6**

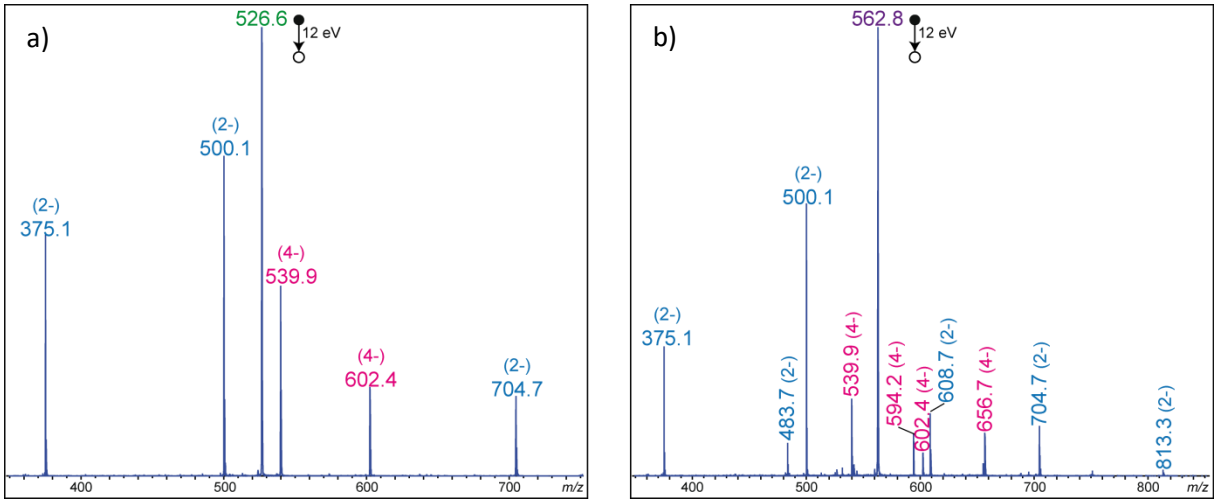


Figure SI-II. 80 : Spectre MS/MS a) du polymère P6 et b) de l'impureté détectée.

- **P7**

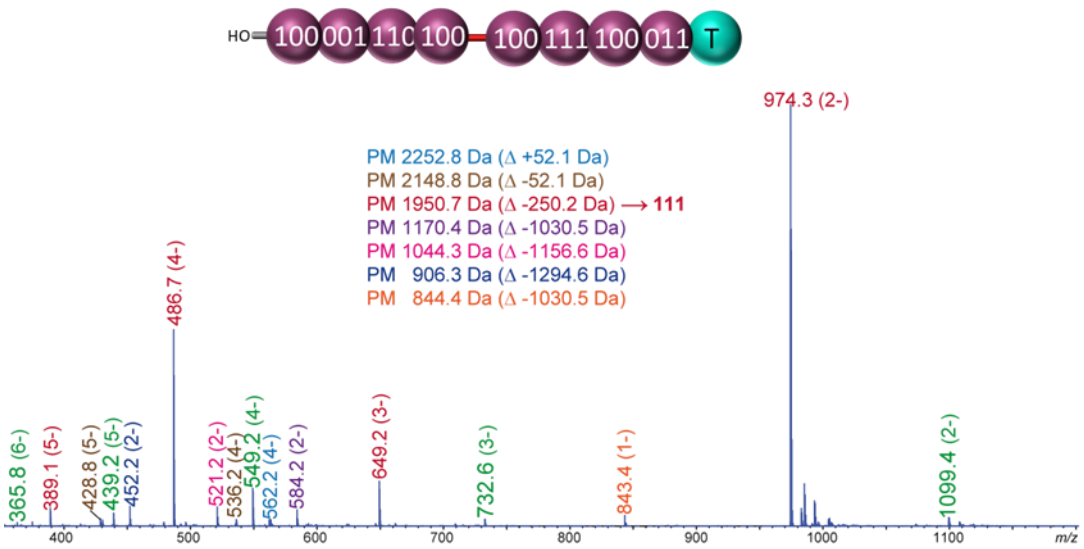


Figure SI-II. 81 : Spectre ESI-MS en mode négatif de P7 montrant l'oligophosphate recherché en très faible abondance sous les états de charge -2 à -6 (en vert). Inset : masse des impuretés principales et défaut de masse par rapport à l'oligomère ciblé.

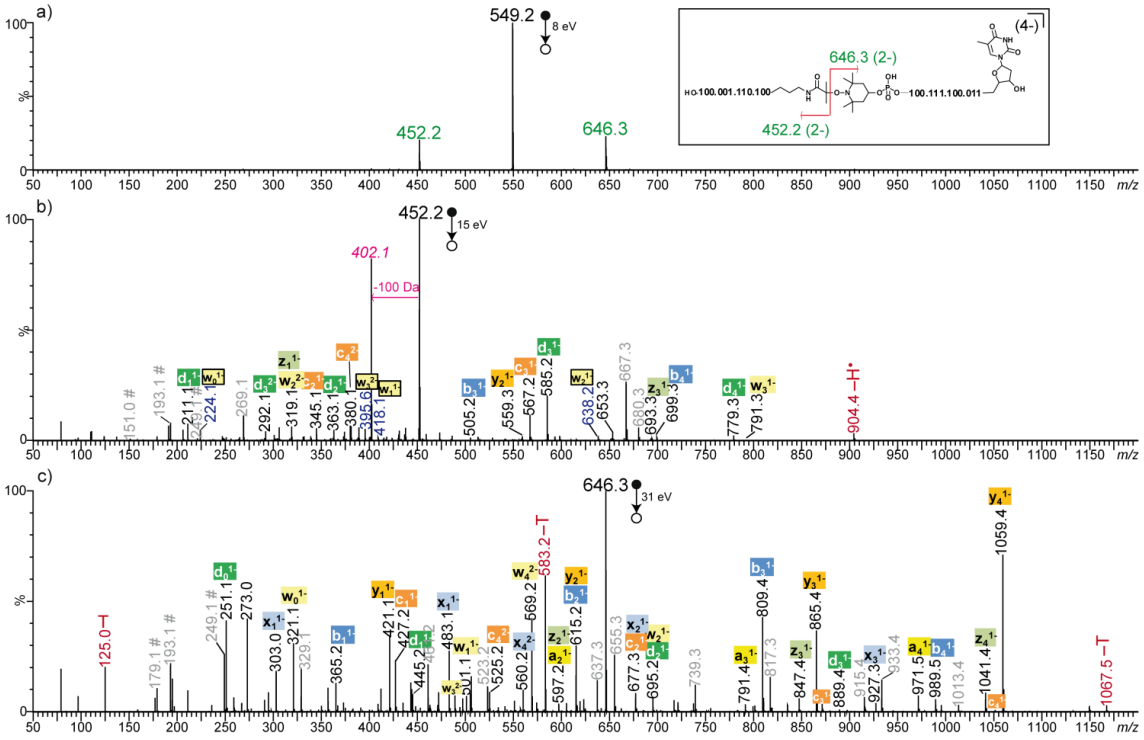


Figure SI-II. 82 : Séquençage de l'oligophosphate détecté comme étant P7. (a) Spectre ESI-MS/MS de l'ion $[M–4H]^{4–}$ à m/z 549,2 et schéma de fragmentation. (b) Spectre pseudo-MS^3 du bloc 1 à m/z 452,2. Les ions encadrés en bleu sont des fragments obtenus après gain d'un H•. (c) Spectre pseudo-MS^3 du bloc 2 à m/z 643,2 Les fragments secondaires (dont les monomères déprotonés, #) sont annotés en gris.

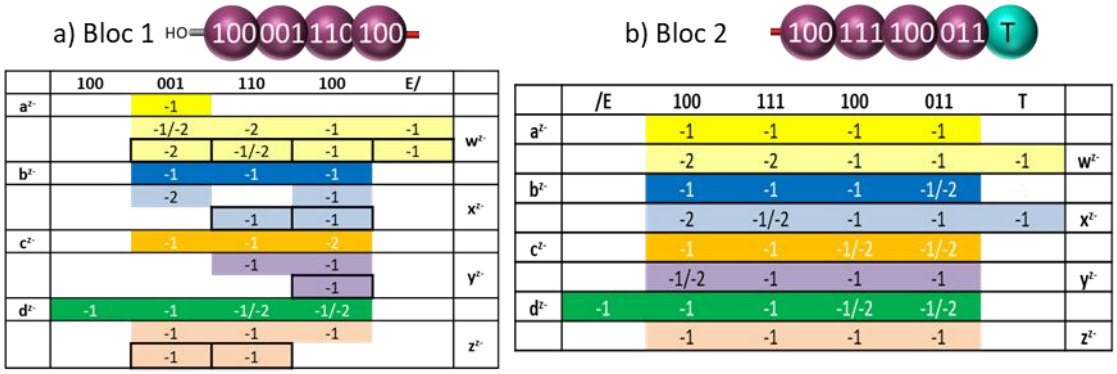


Figure SI-II. 83 : Tableau de couverture de P7 (a) bloc 1, (b) bloc 2.

3.5.6. Séquences codant des images avec l'alphabet à 4 symboles

- **P9**

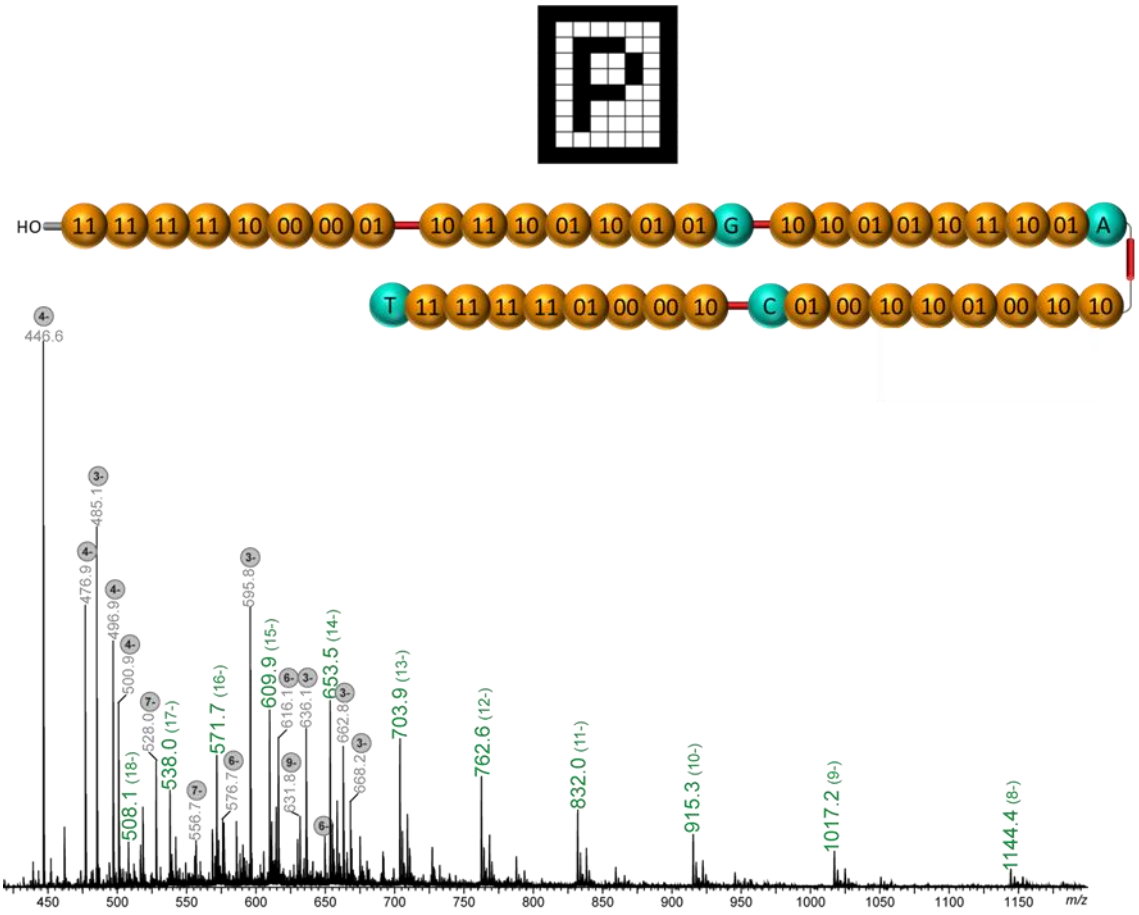


Figure SI-II. 84 : Spectre ESI-MS en mode négatif de P9 détecté sous les états de charge 8- à 19- (en vert). Les pics annotés en gris correspondent à des fragments formés en source. Les pics non annotés sont issus d'échanges H/Na.

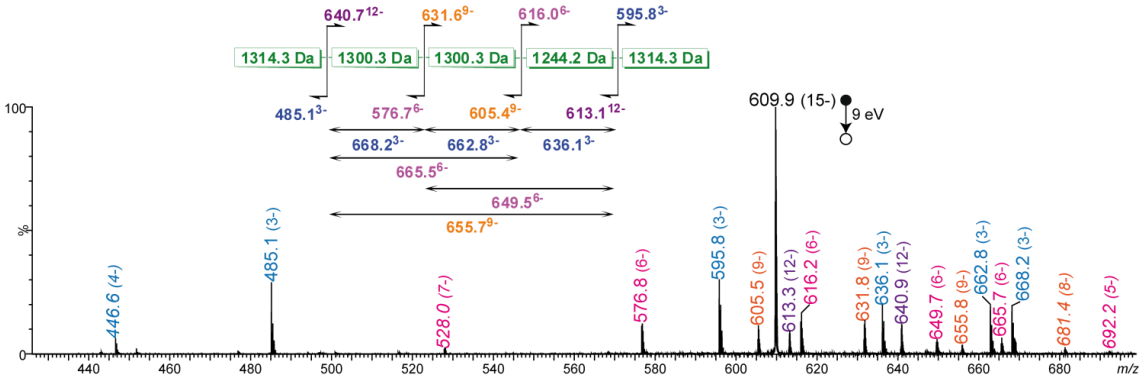


Figure SI-II. 85 : Spectre MS/MS de l'ion m/z 609,9 de P9. Les ions annotés en italique sont des fragments observés sous d'autres états de charge que 3 charges par octet. Inset : schéma de la fragmentation.

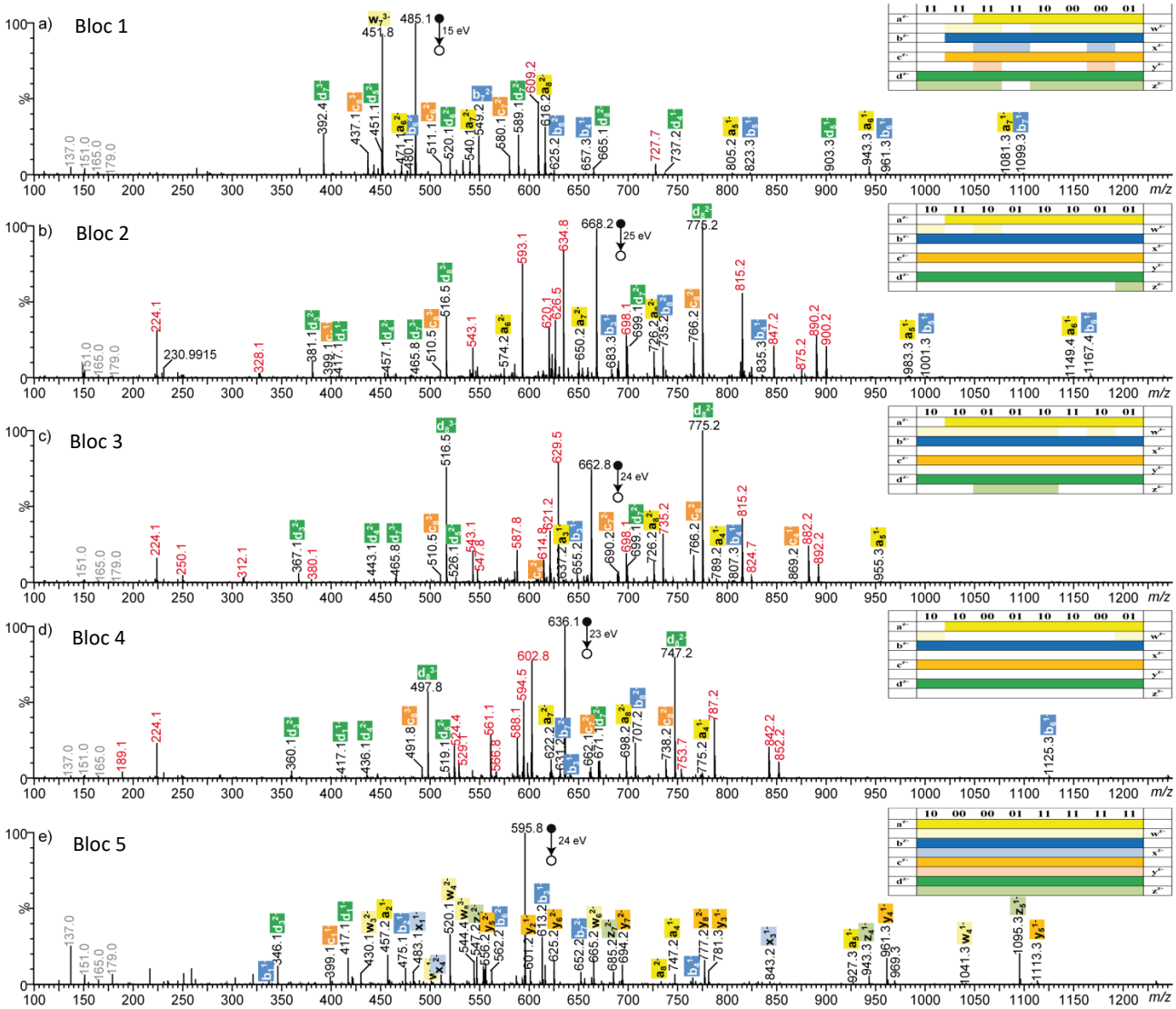


*Figure SI-II. 86 : Séquençage du polymère P9. (a) Spectre pseudo-MS³ du bloc 1 à M/z 485 et couverture de la séquence (b) Spectre pseudo-MS³ du bloc 2 à m/z 668,2 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 3 à m/z 661,8 et couverture de la séquence. Les monomères déprotonés, [**00** – H]⁻ à m/z 137,0, [**01** – H]⁻ à m/z 151,0, [**10** – H]⁻ à m/z 165.0 et [**11** – H]⁻ à m/z 179,0 sont annotés en gris. Les ions annotés en rouge sont des fragments issus de réactions induites par le radical carboné.*

- **P10**

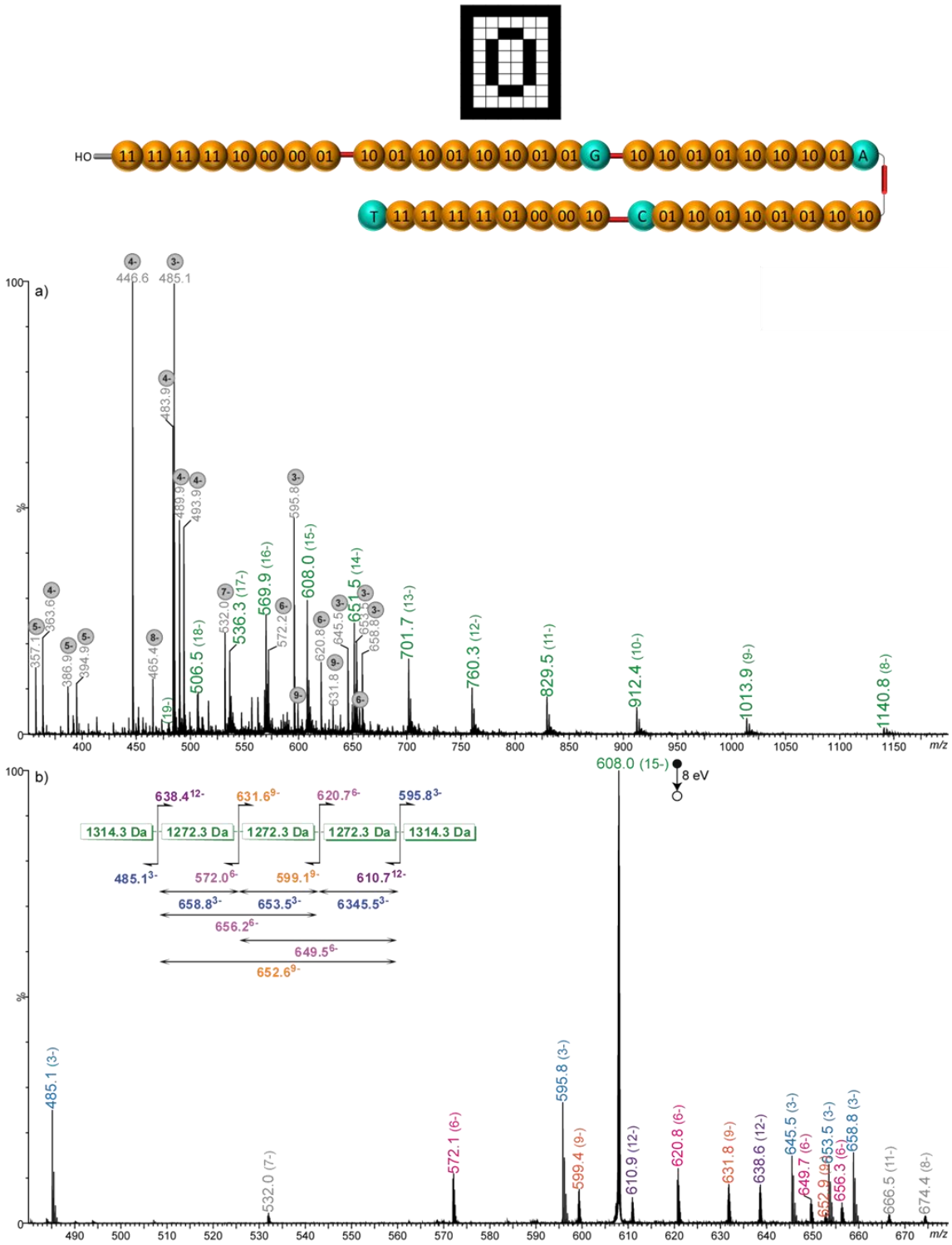


Figure SI-II. 87 : (a) Spectre ESI-MS en mode négatif de P10 observé sous les états de charge 8- à 19- (vert). Les pics annotés en gris correspondent à des fragments formés en source. (b) Spectre MS/MS de l'ion m/z 608,0 (15–). Inset : Schéma de la fragmentation.

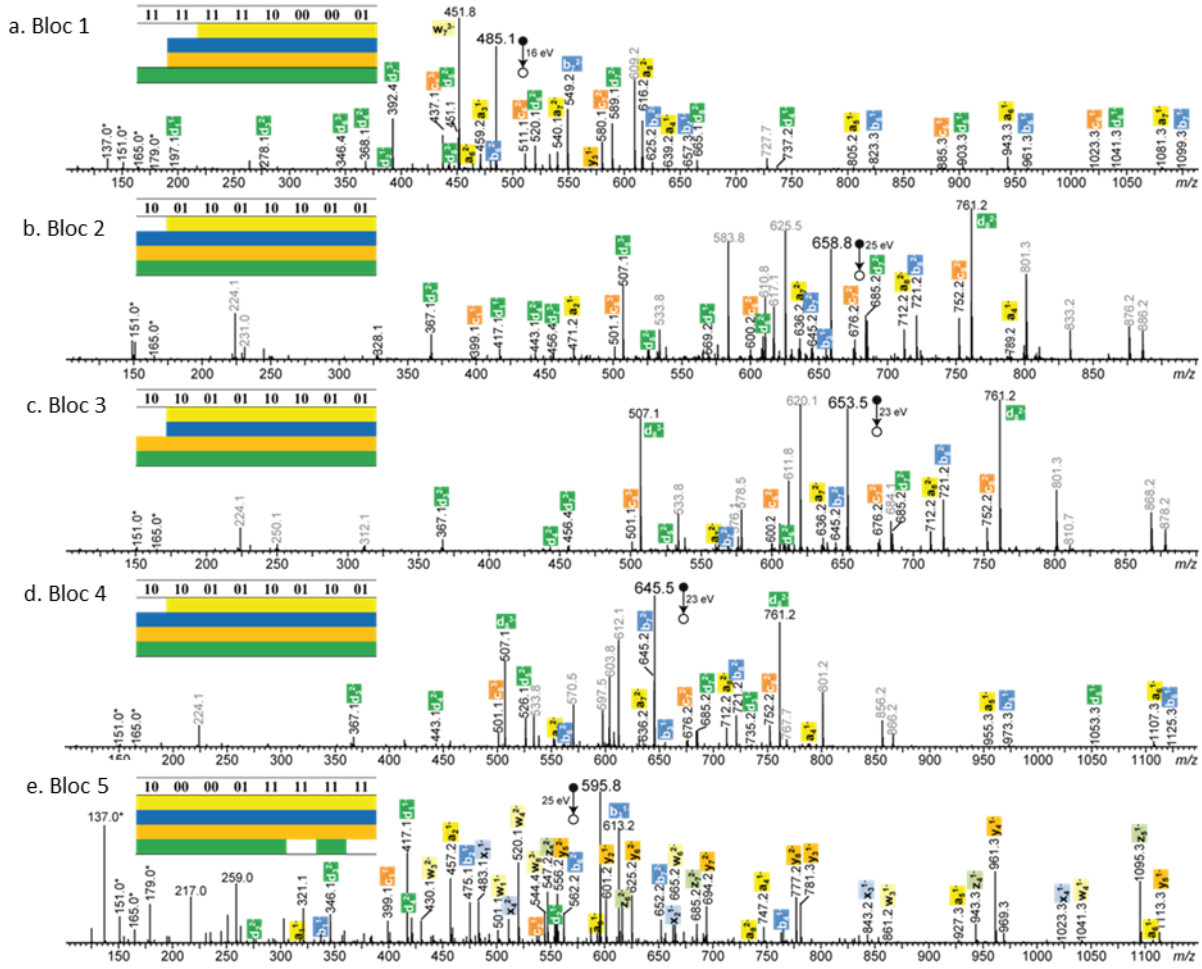


*Figure SI-II. 88 : Séquençage du polymère P10. (a) Spectre pseudo-MS³ du bloc 1 à m/z 485,1 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 2 à m/z 658,8 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 3 à m/z 653,5 et couverture de la séquence. (d) Spectre pseudo-MS³ du bloc 4 à m/z 645,5 et couverture de la séquence. (e) Spectre pseudo-MS³ du bloc 5 à m/z 595,8 et couverture de la séquence. Les monomères déprotonés, [**00** – H]⁻ à m/z 137,0, [**01** – H]⁻ à m/z 151,0, [**10** – H]⁻ à m/z 165,0 et [**11** – H]⁻ à m/z 179,0 sont annotés en gris. Les ions annotés en rouge sont des fragments issus de réactions induites par le radical carboné.*

3.5.7. Séquences codant des images avec l'alphabet à 8 symboles

- **Sous séquences de P11 permettant l'optimisation des conditions de synthèse**

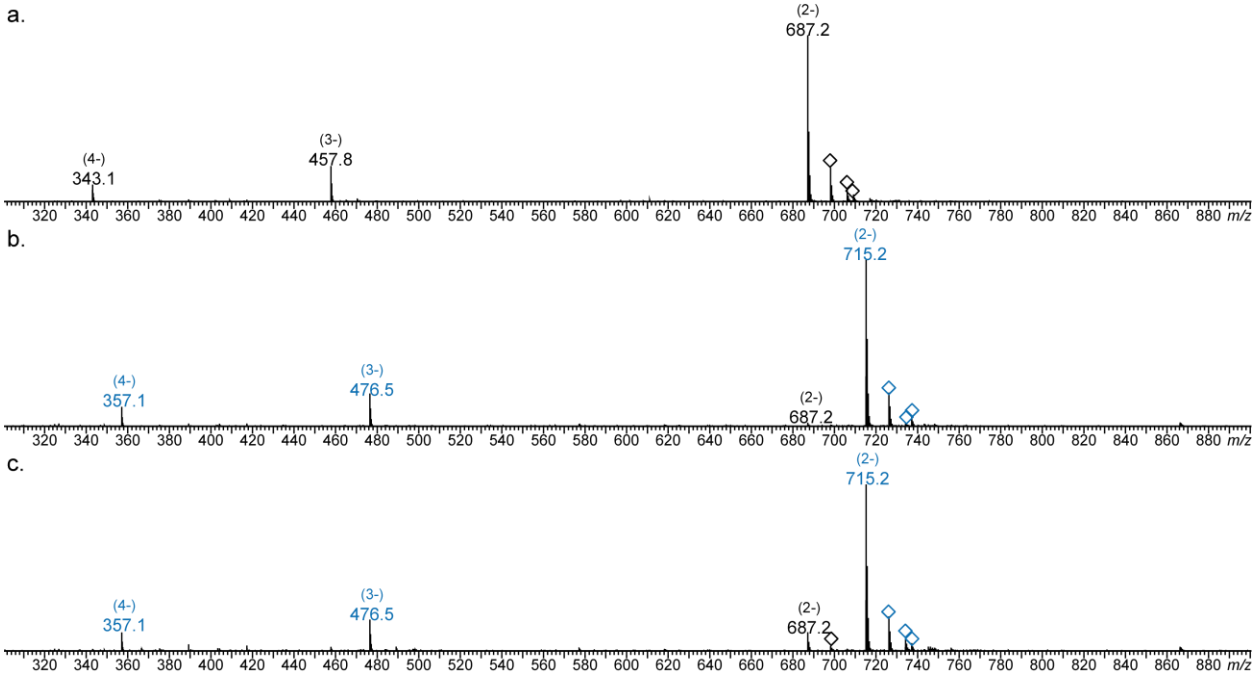


Figure SI-II. 89 : Spectres ESI-MS en mode négatif du premier segment synthétisé de P11 : $M_4{\cdot}M_2{\cdot}M_2{\cdot}M_1{\cdot}M_5{\cdot}M_4{\cdot}M_5$-T (1432,4 Da) préparé avec (a) le protocole a, (b) le protocole b et (c) le protocole c. Les fragments du polymère intact sont annotés en bleu avec l'état de charge entre parenthèse. En noir sont annotés les structures défectueuses. Les diamants désignent les fragments issus des échanges H/Na et/ou H/K.

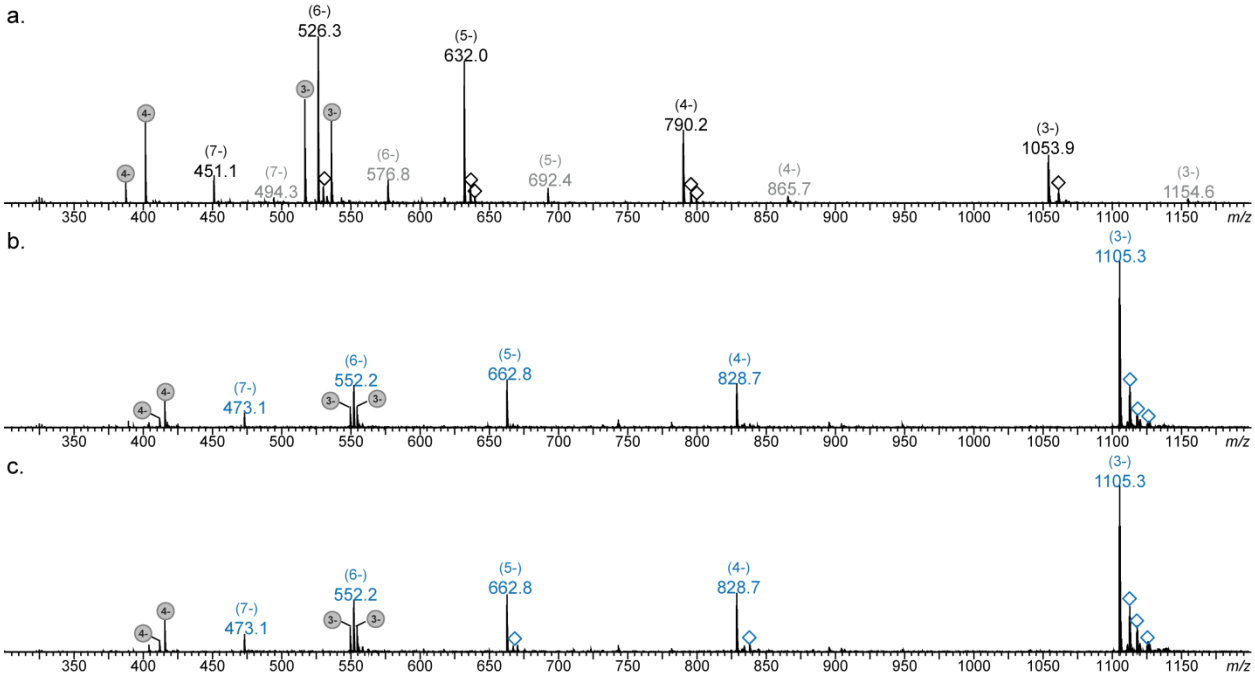


Figure SI-II. 90 : Spectres ESI-MS en mode négatif du premier et deuxième segment synthétisés de P11 : $M_3{\cdot}M_2{\cdot}M_2{\cdot}M_5{\cdot}M_5{\cdot}M_5{\cdot}M_3$-C-L-$M_4{\cdot}M_2{\cdot}M_2{\cdot}M_1{\cdot}M_5{\cdot}M_4{\cdot}M_5$-T (3318,0 Da), préparé avec (a) le protocole a, (b) le protocole b et (c) le protocole c. Les fragments du polymère intact sont annotés en bleu avec l'état de charge entre parenthèse. En noir et gris sont annotés les structures défectueuses. Les diamants désignent les fragments issus des échanges H/Na et/ou H/K. Les pics annotés dans une rond gris correspondent à des fragments formés en source.

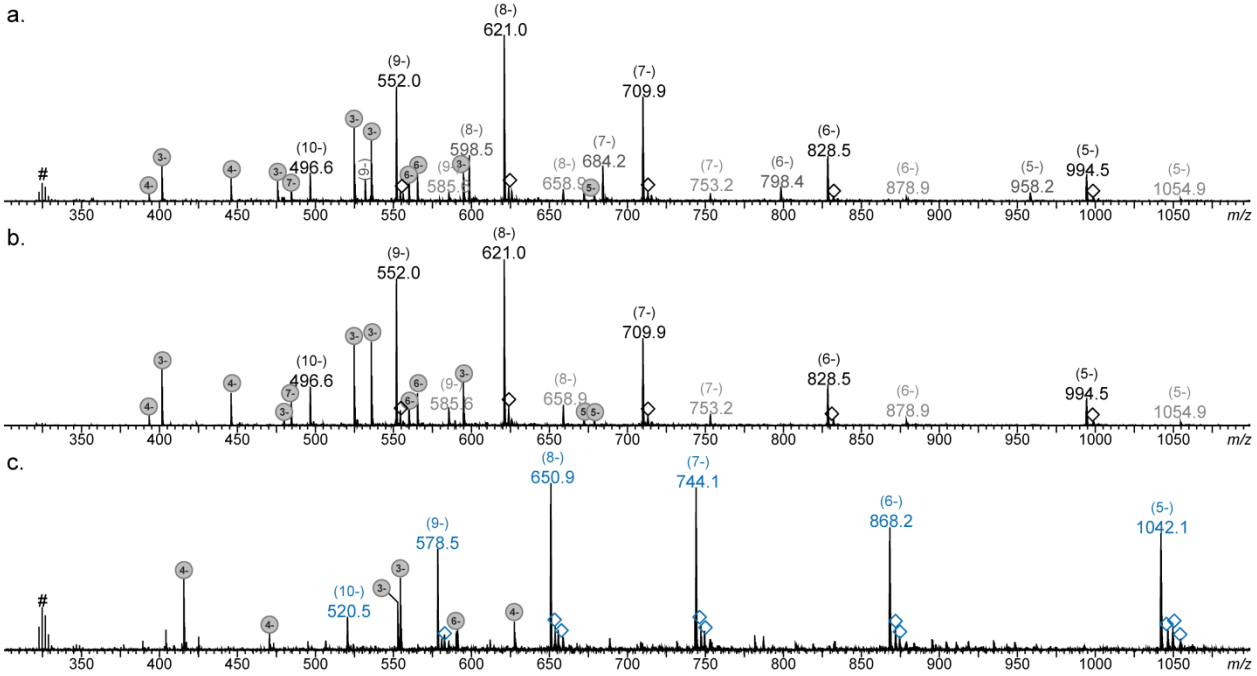


Figure SI-II. 91 : Spectres ESI-MS en mode négatif des trois premiers segments synthétisés de P11 : M₄·M₂·M₂·M₁·M₅·M₇·M₃-A-L-M₃·M₂·M₂·M₅·M₅·M₅·M₃-C-L-M₄·M₂·M₂·M₁·M₅·M₄·M₅-T (5213,5485 Da), préparé avec (a) le protocole a, (b) le protocole b et (c) le protocole c. Les fragments du polymère intact sont annotés en bleu avec l'état de charge entre parenthèse. En noir et gris sont annotés les structures défectueuses. Les diamants désignent les fragments issus des échanges H/Na et/ou H/K. Les pics annotés dans une rond gris correspondent à des fragments formés en source. # Amas d'acide trichlocoacétique.

- **P11**

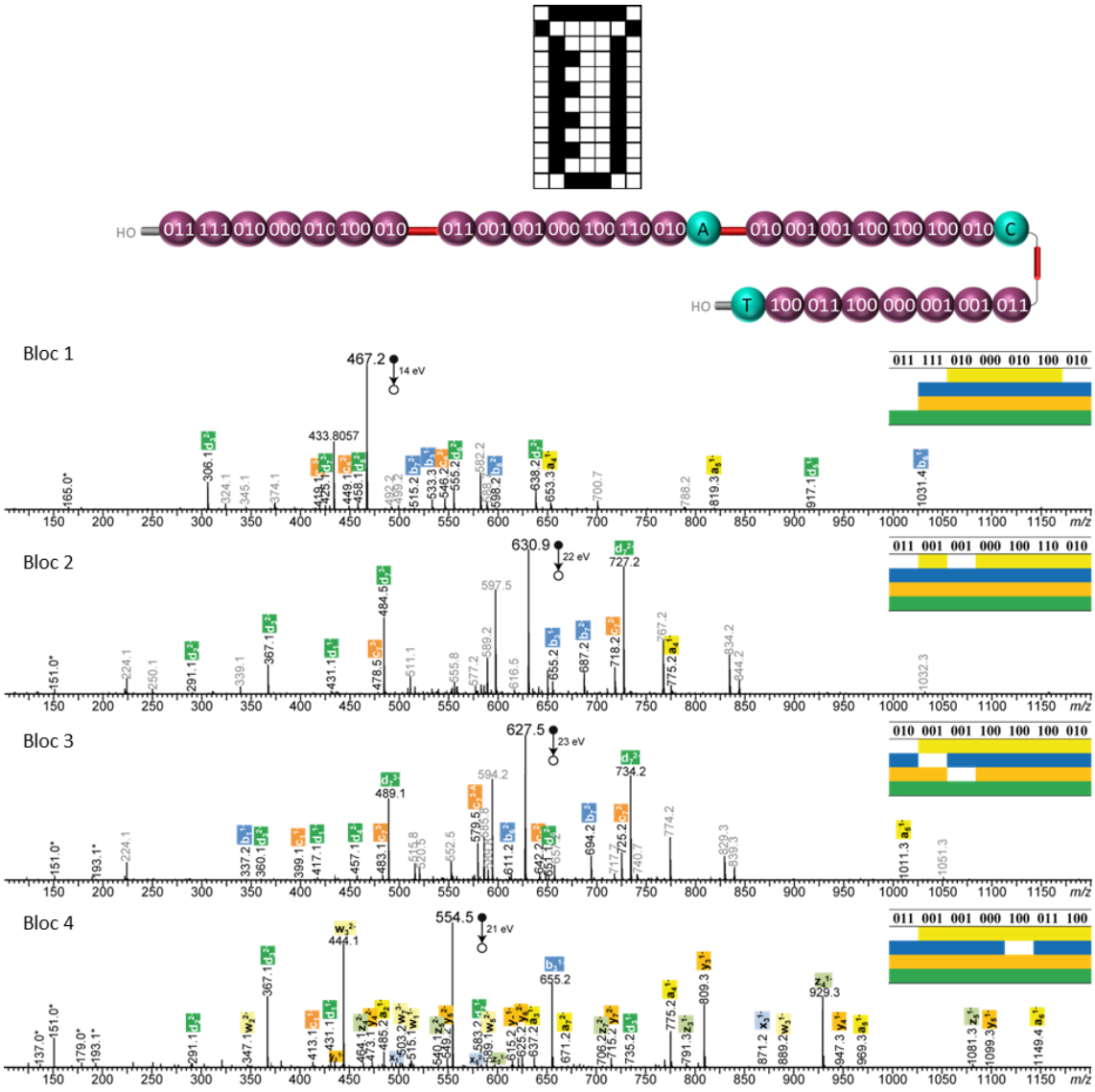


*Figure SI-II. 92 : Séquençage du polymère P11. (a) Spectre pseudo-MS³ du bloc 1 à m/z 467,2 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 2 à m/z 630,9 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 3 à m/z 627,5 et couverture de la séquence. (d) Spectre pseudo-MS³ du bloc 4 à m/z 554,5 et couverture de la séquence. Les monomères déprotonés sont annotés avec *. Les ions annotés en gris sont des fragments issus de réactions induites par le radical carboné.*

- **P12**

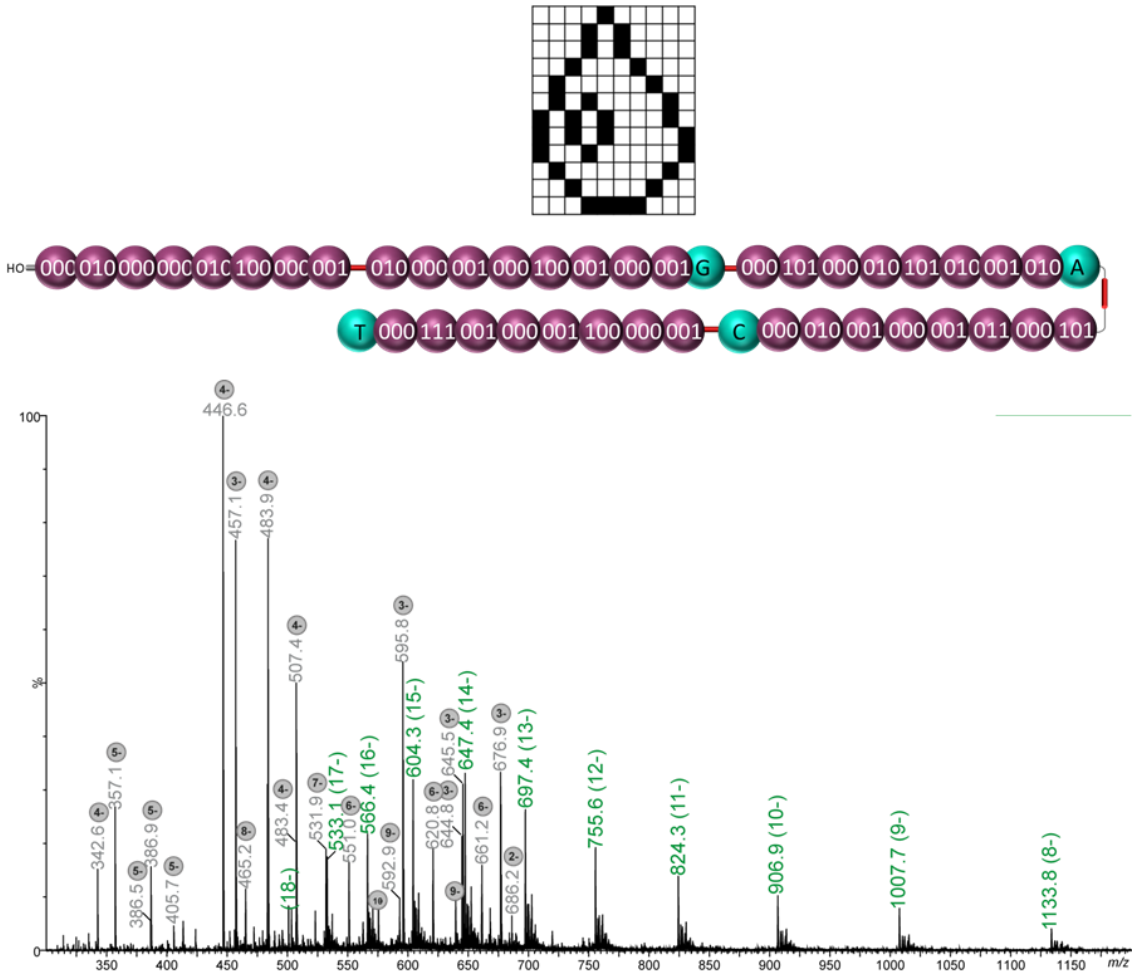


Figure SI-II. 93 : Spectre ESI-MS en mode négatif de P12. Le polymère est détecté exclusivement sous les états de charge 6- à 14- (vert). Les pics annotés en gris correspondent à des fragments formés en source.

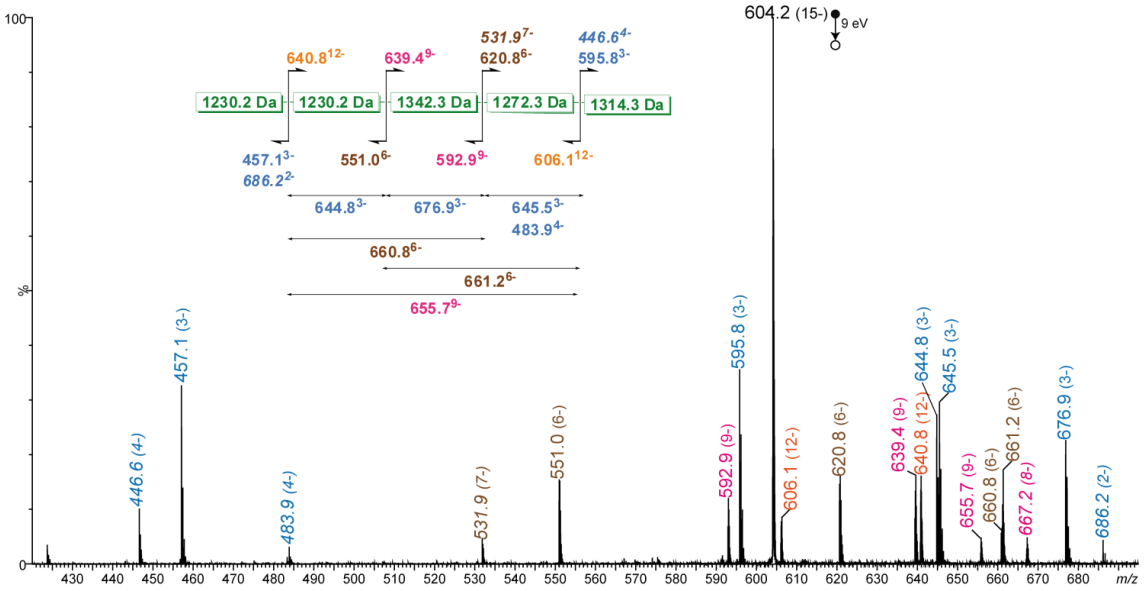


Figure SI-II. 94 : Spectre MS/MS de l'ion $[M - 15H]^{15-}$ à m/z 604,3 du polymère P12.

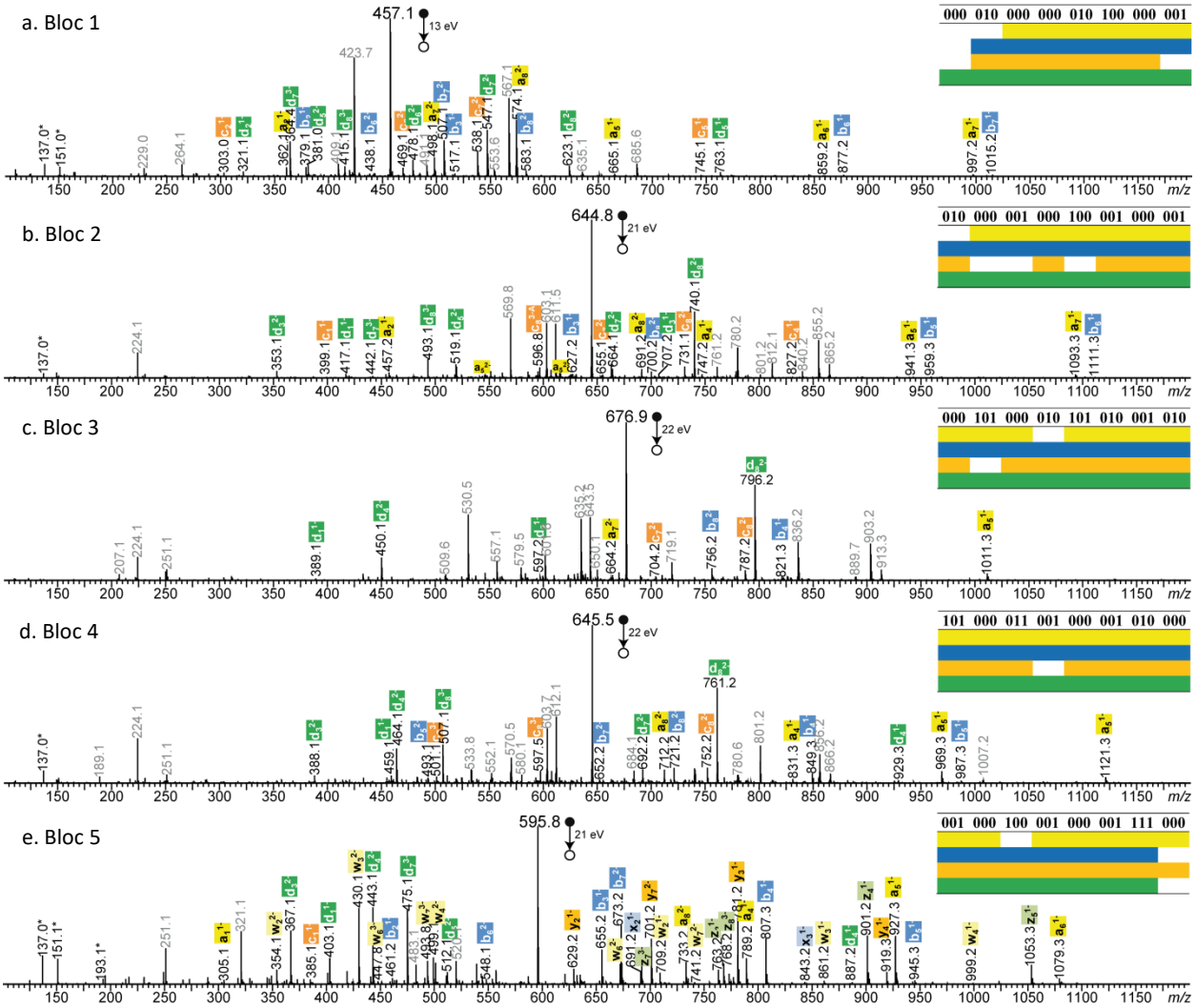


*Figure SI-II. 95 : Séquençage du polymère P12. (a) Spectre pseudo-MS³ du bloc 1 à m/z 457,1 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 2 à m/z 644,8 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 3 à m/z 676,9 et couverture de la séquence. (d) Spectre pseudo-MS³ du bloc 4 à m/z 645,5 et couverture de la séquence. (e) Spectre pseudo-MS³ du bloc 5 à m/z 595,8 et couverture de la séquence. Les monomères déprotonés sont annotés avec *. Les ions annotés en gris sont des fragments issus de réactions induites par le radical carboné.*

- **P13**

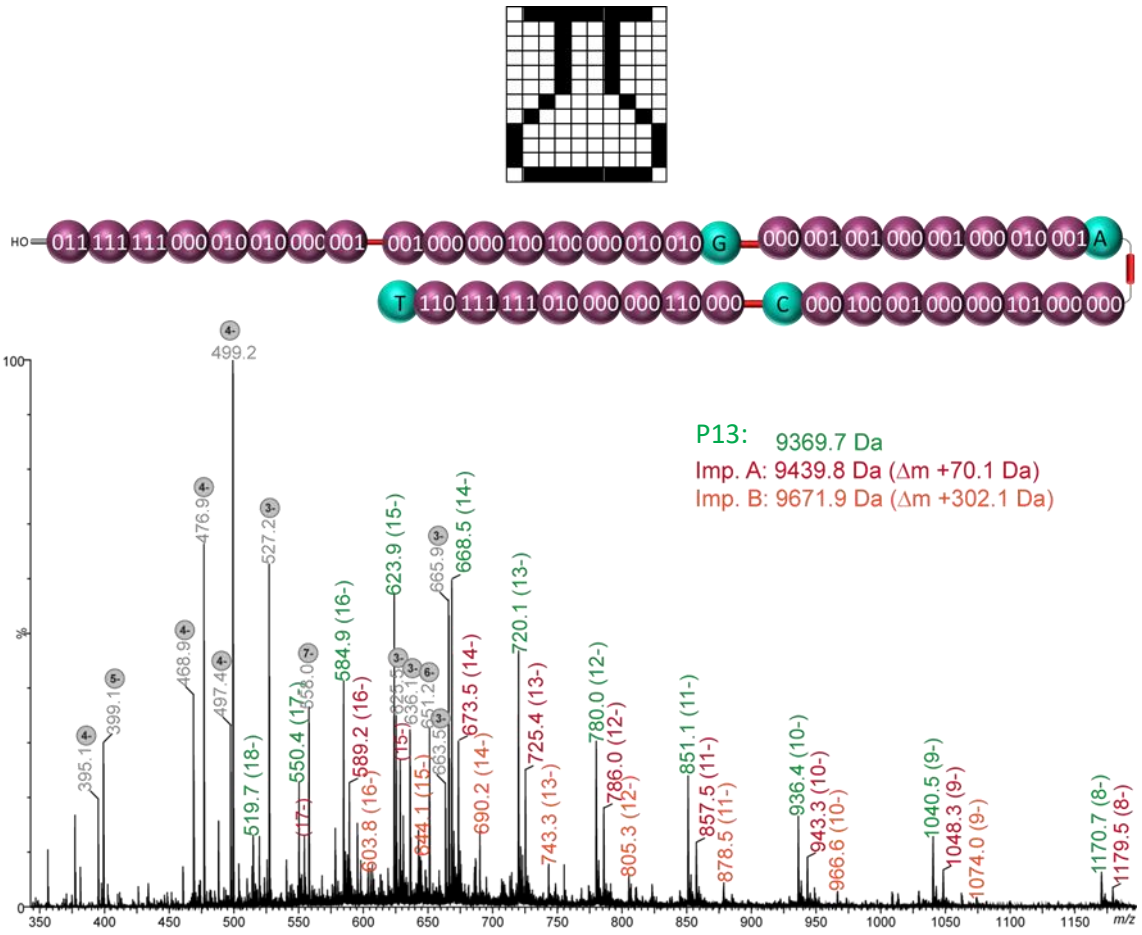


Figure SI-II. 96 : Spectre ESI-MS en mode négatif de P13 (pics verts). En rouge, impureté A de Δm + 70,1 Da et en orange impureté B de Δm + 302,1 Da. Les pics annotés en gris correspondent à des fragments formés en source.

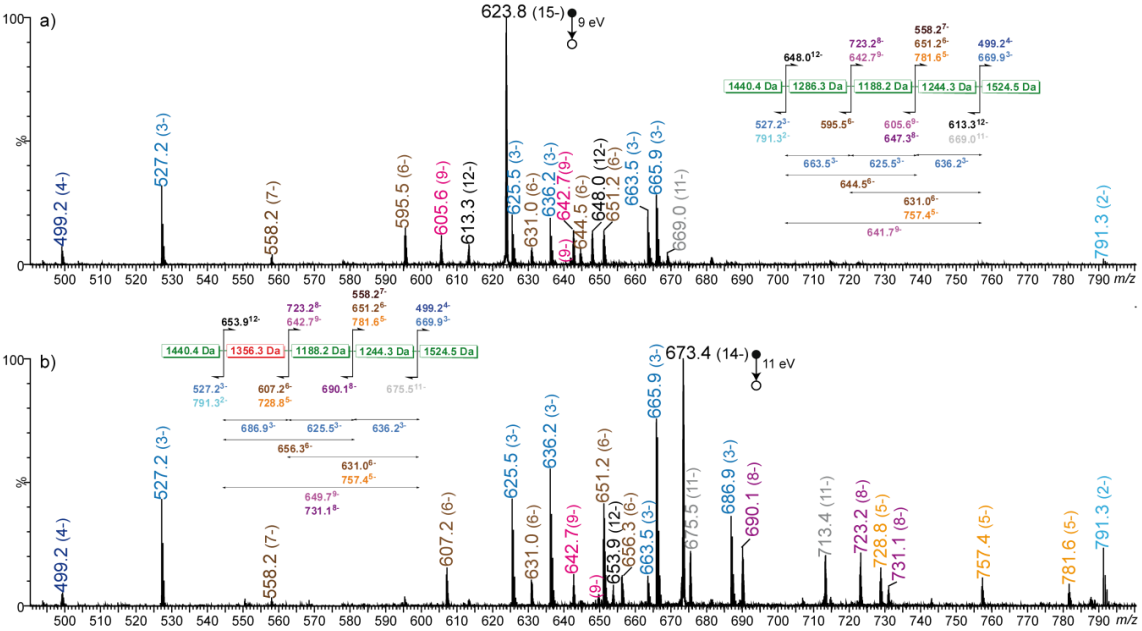


Figure SI-II. 97 : Spectre MS/MS des ions (a) $[M - 15H]^{15-}$ à m/z 623,8 correspondant à P13, et (b) $[A - 14H]^{14-}$ à m/z 673,4 (impureté annotée en rouge) détectés dans l'échantillon.

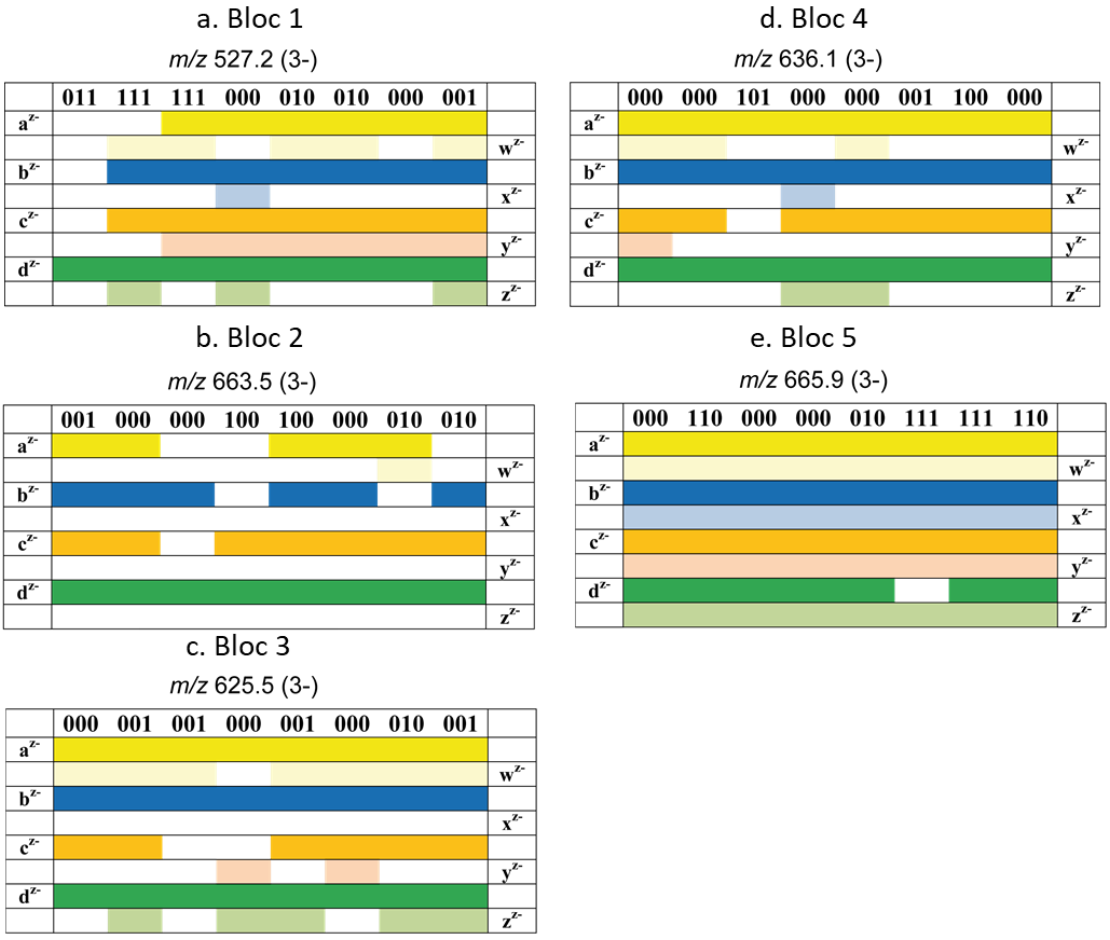


Figure SI-II. 98 : Couverture de séquence des 5 blocs de P13. (a) bloc 1, (b) bloc 2, (c) bloc 3, (d) bloc 4 et (e) bloc 5.

- **P14**

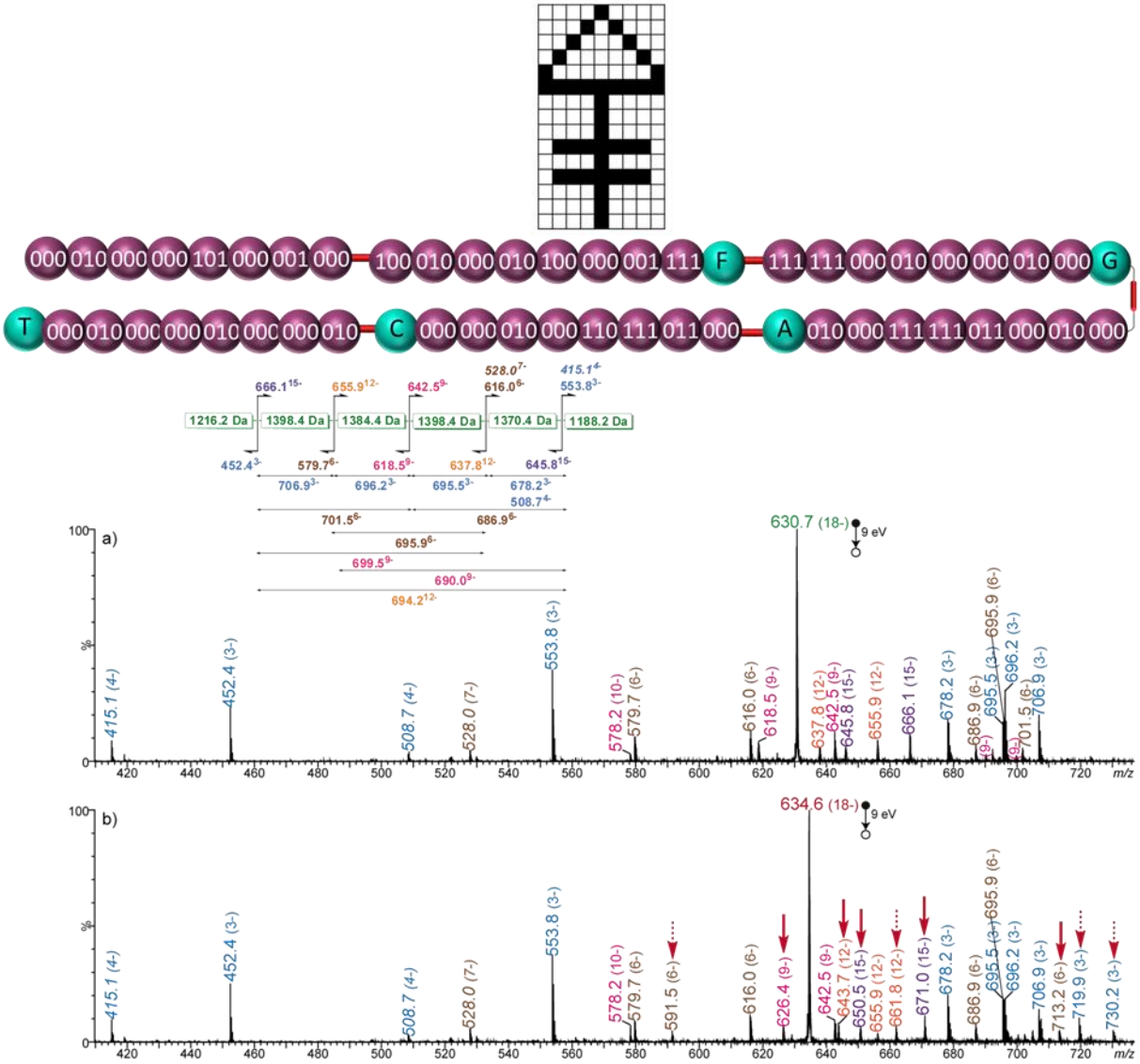


Figure SI-II. 99 : Spectre MS/MS des ions (a) $[M - 18H]^{18-}$ à m/z 630,7 de P14 et (b) $[X - 18H]^{18-}$ à m/z 634,6 de l'impureté détectés dans l'échantillon P14.

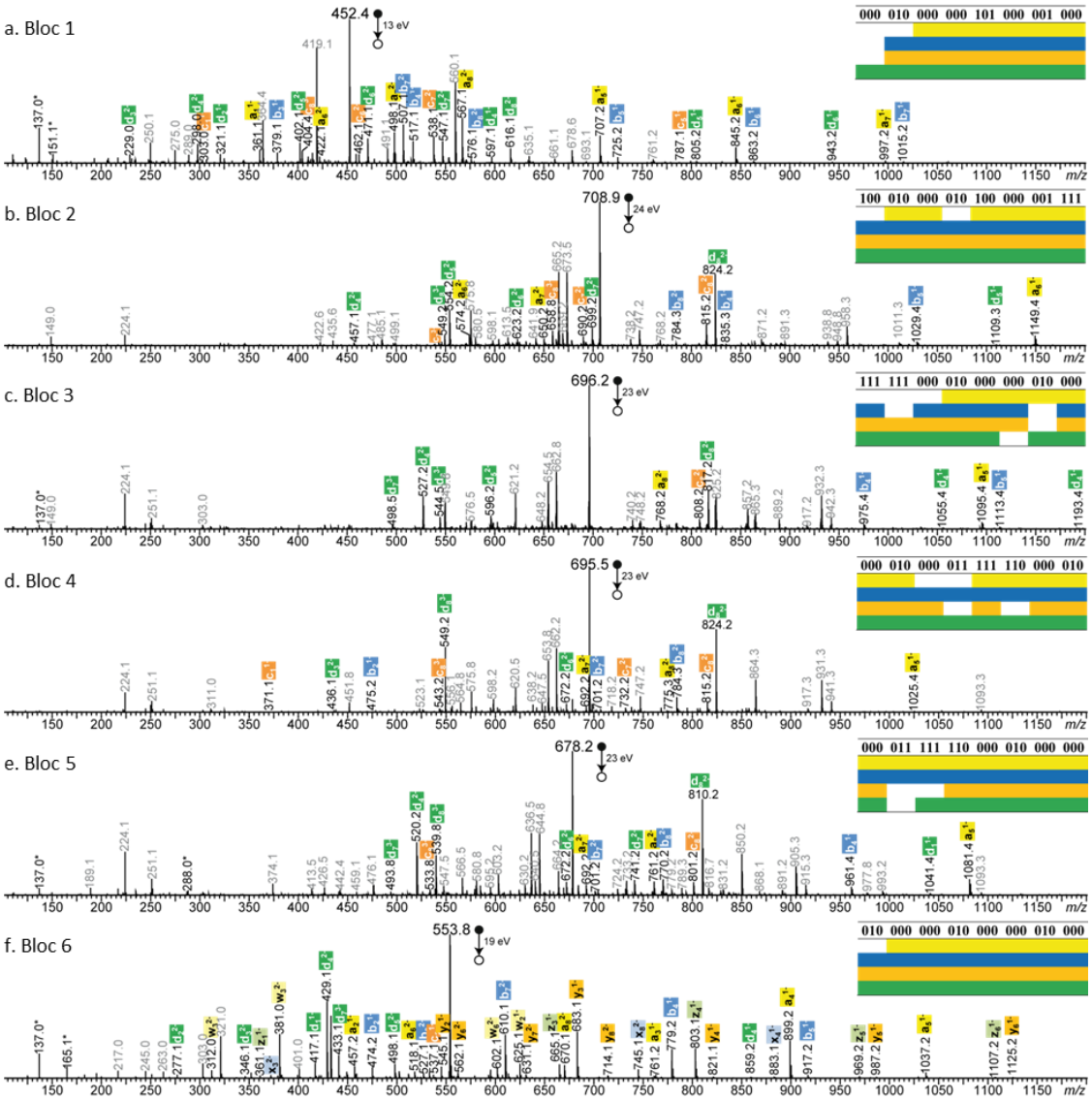


*Figure SI-II. 100 : Séquençage du polymère P14. (a) Spectre pseudo-MS3 du bloc 1 à m/z 452,4 et couverture de la séquence. (b) Spectre pseudo-MS3 du bloc 2 à m/z 706,9 et couverture de la séquence. (c) Spectre pseudo-MS3 du bloc 3 à m/z 696,2 et couverture de la séquence. (d) Spectre pseudo-MS3 du bloc 4 à m/z 695,6 et couverture de la séquence. (e) Spectre pseudo-MS3 du bloc 5 à m/z 678,2 et couverture de la séquence. (f) Spectre pseudo-MS3 du bloc 6 à m/z 553(8 et couverture de la séquence. Les monomères déprotonés sont annotés avec *. Les ions annotés en gris sont des fragments issus de réactions induites par le radical carboné.*

4. Partie expérimentale du chapitre III

4.1. Analyse de polymères par spectrométrie de masse

4.1.1. Séquences impliquant l'espaceur ROSC

- **$P_{ROSC}1$**

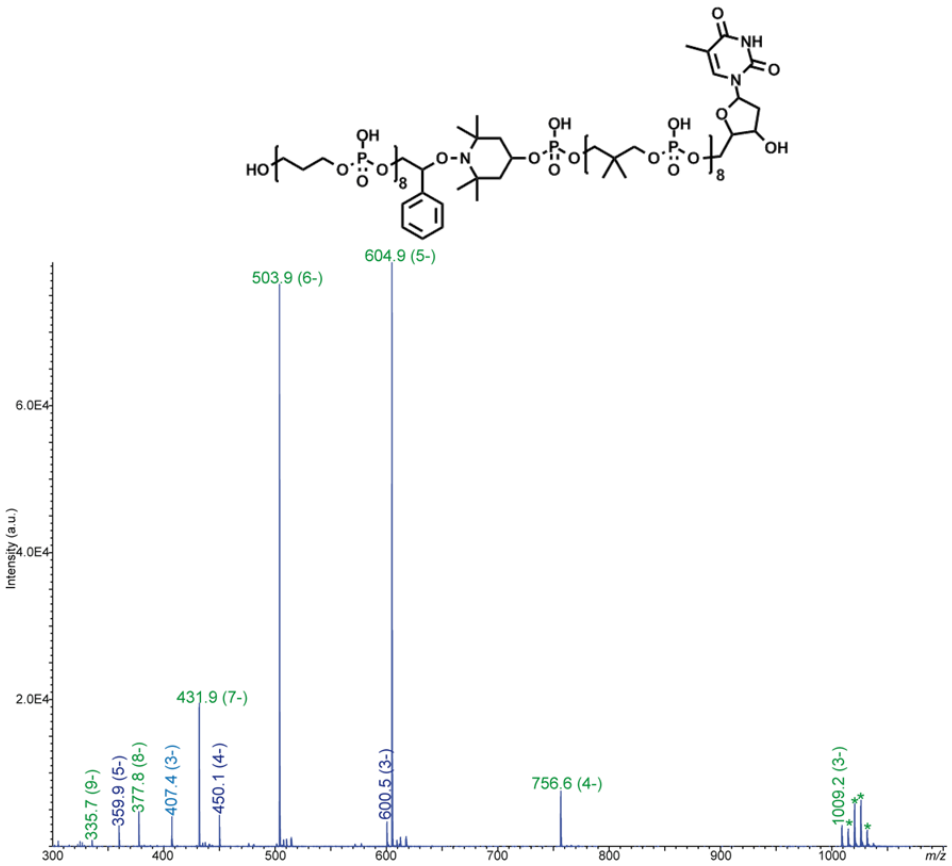


Figure SI-III. 1 : Spectre ESI-MS en mode négatif de $P_{ROSC}1$ (pics verts). Les pics annotés en bleu sont issus de l'homolyse de la liaison alcoxyamine, en cyan pour le premier bloc et en bleu marine pour le second bloc. Les ions annotés par un astérisque sont issus d'échanges H/Na et/ou H/K.

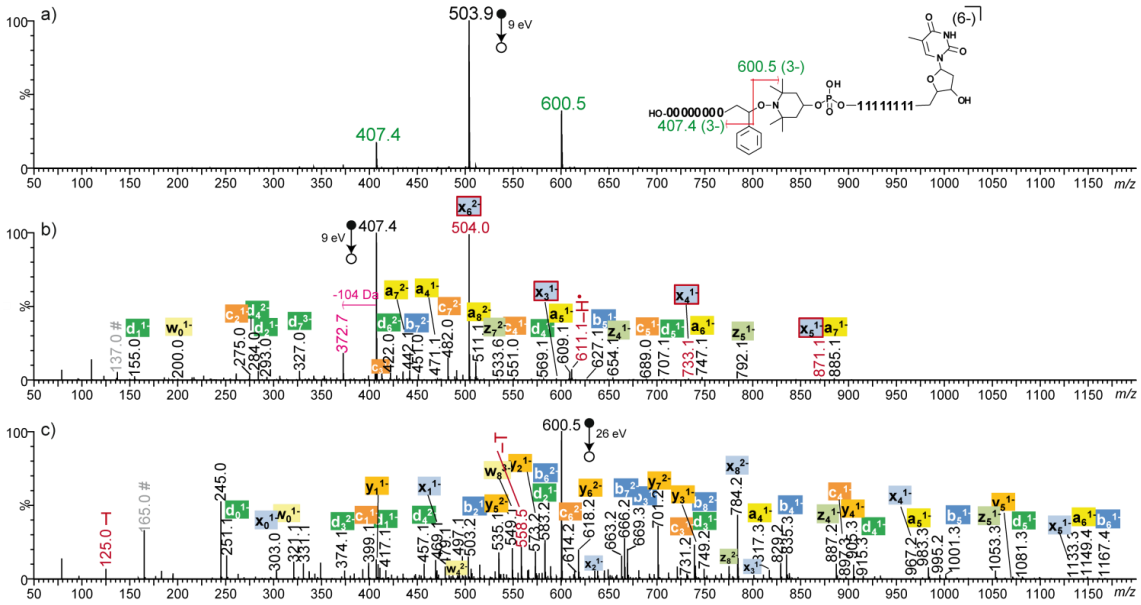


Figure SI-III. 2 : Séquençage de $P_{ROSC}1$. (a) Spectre ESI-MS/MS de l'ion $[M–6H]^{6–}$ à m/z 503,9 et schéma de fragmentation. (b) Spectre pseudo-MS^3 du bloc 1 à m/z 407,4. L'ion à ms/Z 611,1 provient de l'élimination d'un H• à partir de l'ion précurseur. Les ions encadrés en rouge sont des fragments obtenus après abstraction d'un H•. (c) Spectre pseudo-MS^3 du bloc 2 à m/z 600,5. Les fragments secondaires (dont les monomères déprotonés, #) sont annotés en gris.

	0	0	0	0	0	0	0	0	RO/	
a^{2-}			-1	-1	-1	-1/-2	-2	-2		
					-2				-1	w^{2-}
			-2	-2		-1	-1	-1		
b^{2-}			-1	-1	-1	-1	-1/-2			
										x^{2-}
			-2	-1/-2	-1/-2	-1	-1	-1		
c^{2-}		-1	-1	-1	-1/-2	-2	-2			
		-1	-1	-1	-1					
			-2	-2	-1	-1				y^{2-}
d^{2-}	-1	-1	-1	-1/-2	-1/-2	-2	-2/-3			
		-2	-2	-1/-2	-1	-1		-1		z^{2-}

	/SC	1	1	1	1	1	1	1	1	T	
a^{2-}			-1	-1	-1	-1/-2	-1/-2	-2	-2/-3		
		-3	-3	-2	-2	-1/-2	-1/-2	-1/-2	-1	-1	w^{2-}
b^{2-}		-1	-1	-1	-1/-2	-1/-2	-1/-2	-2	-3		
		-3	-2/-3	-2	-1/-2	-1/-2	-1/-2	-1	-1	-1	x^{2-}
c^{2-}	-1	-1	-1	-1/-2	-1/-2	-1/-2	-2	-2/-3	-3		
		-2/-3	-2	-2	-1/-2	-1/-2	-1/-2	-1	-1		y^{2-}
d^{2-}	-1	-1	-1/-2	-1/-2	-1/-2	-1/-2	-2	-2	-2		
		-2/-3	-2	-2	-1/-2	-1	-1	-1	-1		z^{2-}

Figure SI-III. 3 : Couverture des deux blocs de la séquence $P_{ROSC}1$. (Gauche) Séquence du premier bloc avec M1 codant pour le bit 0. Les ions encadrés en rouge sont des fragments obtenus après l'abstraction d'un H•. (Droite) Séquence du second bloc avec M2 codant pour le bit 1.

- **$P_{ROSC}2$**

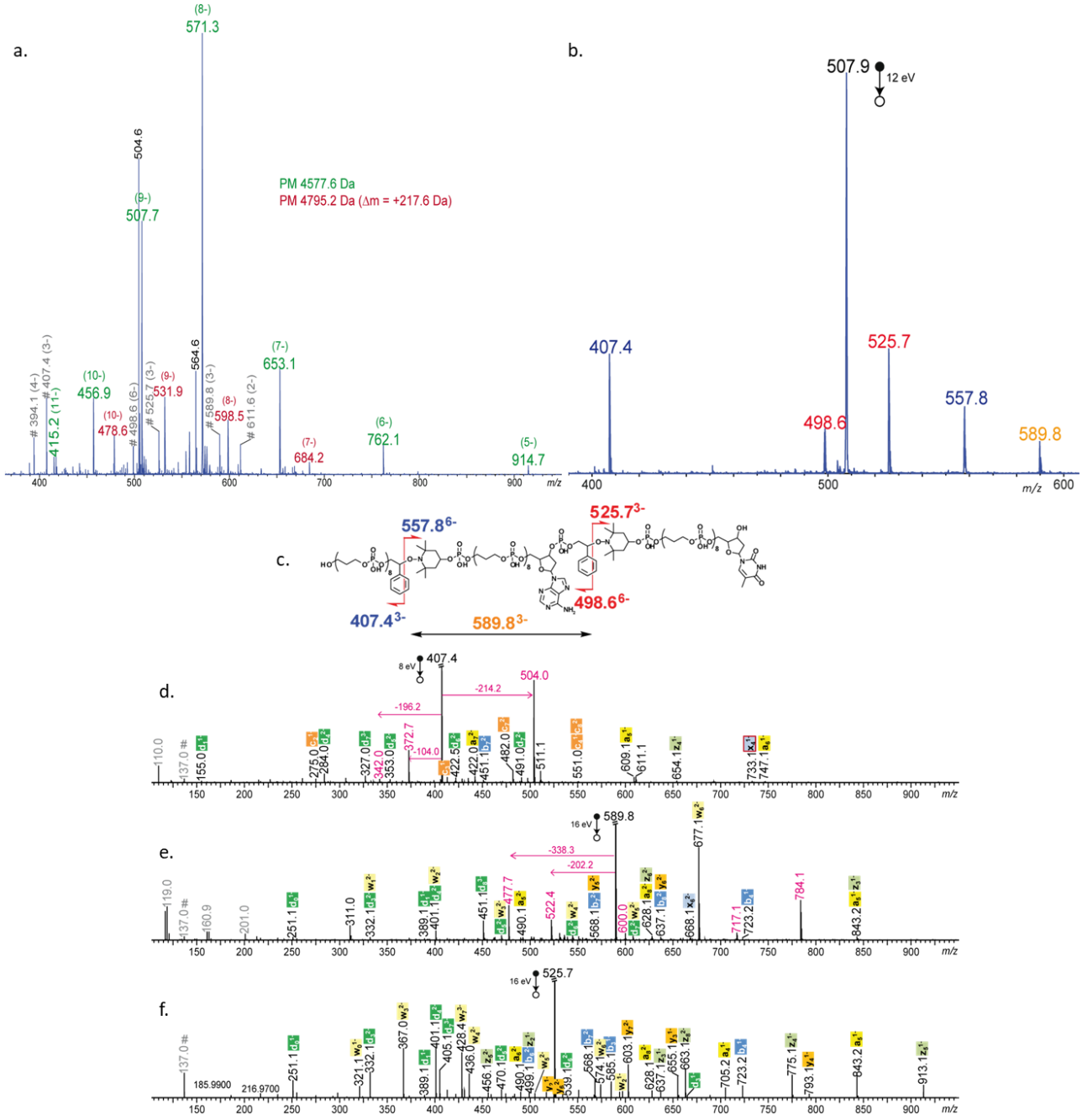


Figure III. 27: Séquençage du polymère $P_{ROSC}2$. (a) Spectre ESI-MS en mode négatif. Les ions annotés en noir sont des espèces simplement chargées. Les pics annotés en rouge correspondent à une impureté de +217,6 Da. (b) Spectre MS/MS de l'ion $[M - 9H]^{9-}$ à m/z *507,9 (c) Schéma d'homolyse des liaisons NO-C. (c) Spectre pseudo-MS^3 du bloc 1 à* m/z *407,4 (d) Spectre pseudo-MS^3 du bloc 2 à* m/z *589,8. (e) Spectre pseudo-MS^3 du bloc 3 à* m/z *525,7. Les pics annotés en gris correspondent à des fragments internes, parmi lesquels les monomères sont désignés par #. Les ions annotés en rose sont issus de réactions induites par le radical de l'espaceur ROSC.*

4.1.2. Séquences impliquant l'espaceur RISC

- **$P_{RISC}1$**

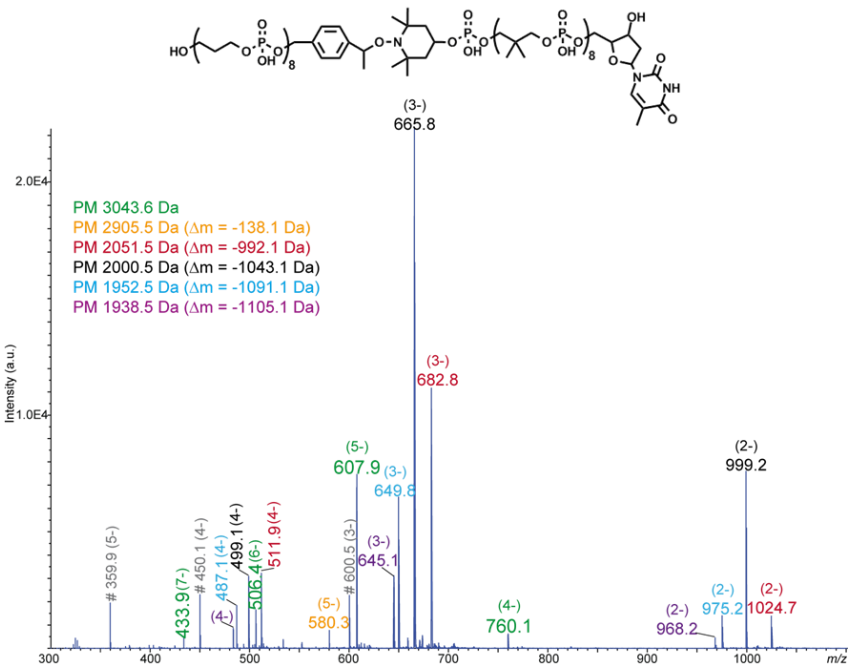


Figure SI-III. 4 : Spectre ESI-MS en mode négatif de $P_{RISC}1$ (pics verts). Les pics en orange, rouge, noir, bleu et violet correspondent à des impuretés dont la perte de masse est donnée en inset. Les pics annotés en gris correspondent au second octet libéré en source.

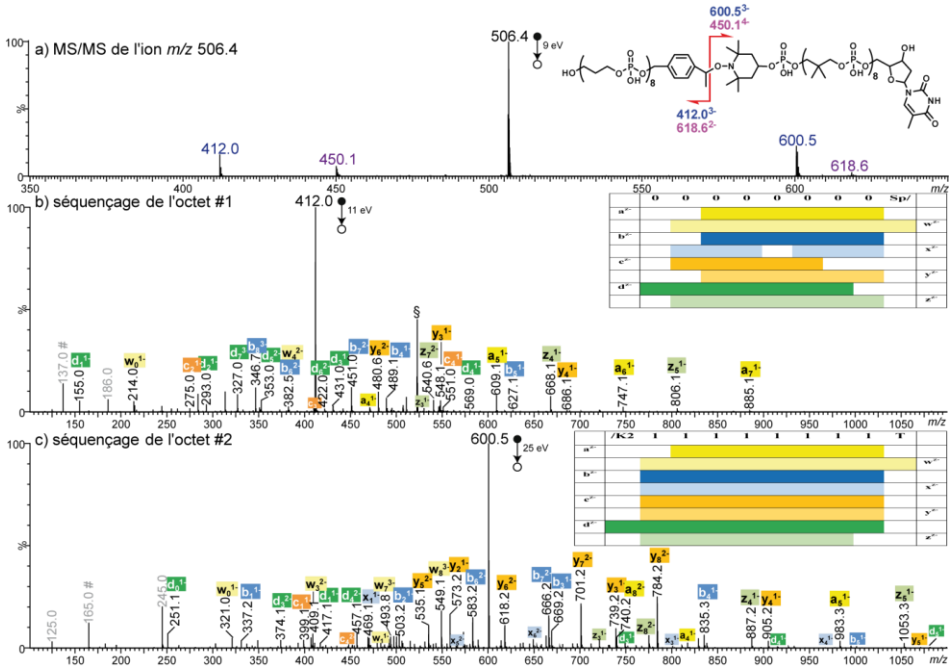


Figure SI-III. 5 : Séquence de $P_{RISC}1$. (a) Spectre MS/MS de l'ion $[M - 6H]^{6-}$ avec l'homolyse de la liaison alcoxyamine. (b) Spectre pseudo-MS^3 du bloc 1 à m/z 412,0 (3–) (c) Spectre pseudo-MS^3 du bloc 2 à m/z 600,5. Les pics annotés en gris correspondent à des fragments internes, parmi lesquels les monomères sont désignés par #.

- **$P_{RISC}2$**

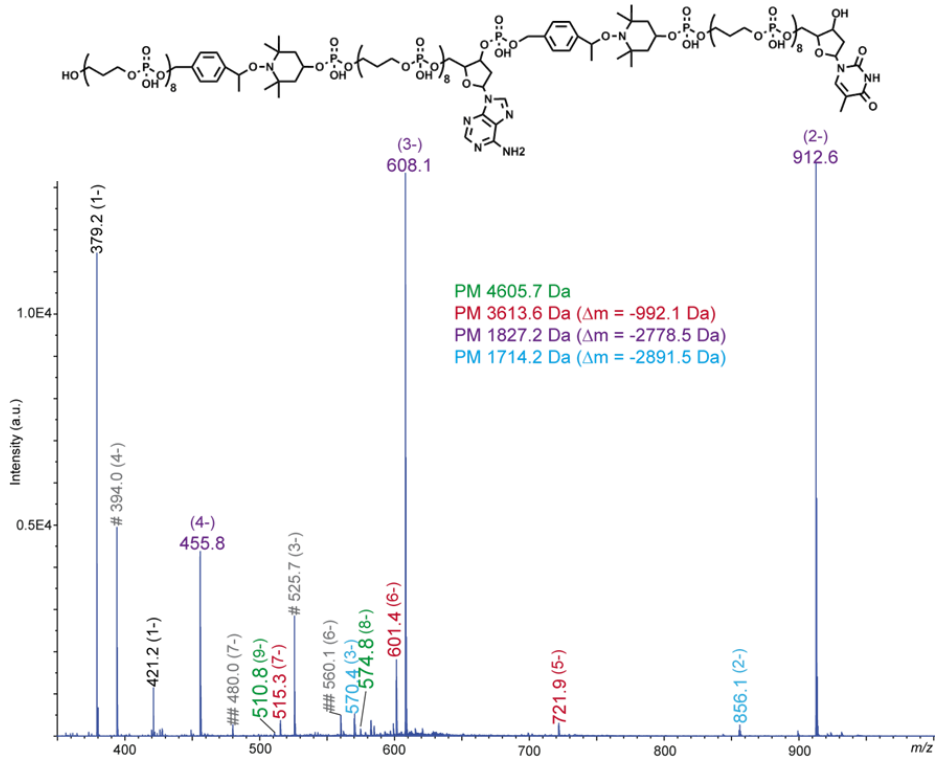


Figure SI-III. 6 : Spectre ESI-MS en mode négatif de $P_{RISC}2$ (pics verts). Les pics annotés en rouge, violet et bleu sont des impuretés détectées, donc la perte de masse est donnée en inset. Les pics annotés en gris correspondent à des fragments formés en source tels que l'octet 3 (#) ou les octets 2 et 3 (##).

4.1.3. Séquences impliquant l'espaceur RISCOP

- **$P_{RISCOP}1$**

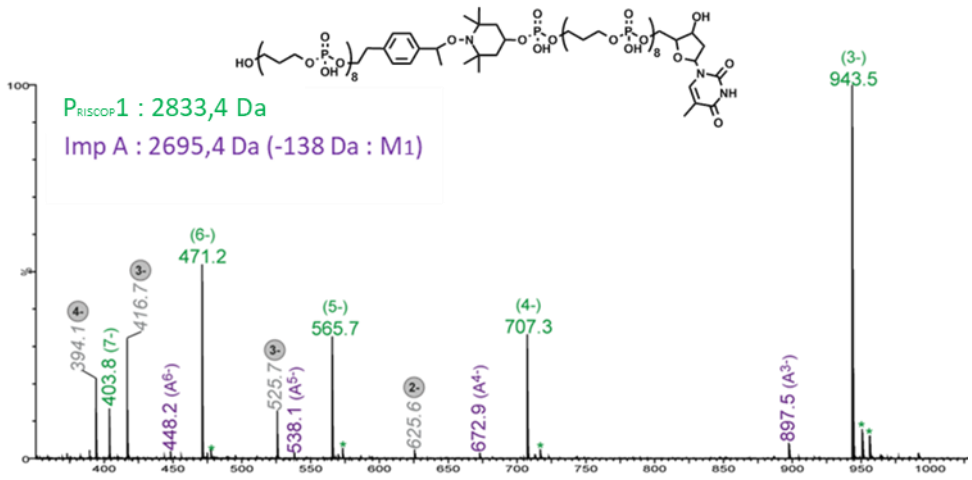


Figure SI-III. 7 : Spectre ESI-MS en mode négatif de $P_{RISCOP}1$ (pics en verts). Les ions annotés avec un rond gris correspondent aux fragments libérés en source. Les pics en violet correspondent à l'impureté violette. Les pics annotés par un astérisque correspondent à des échanges H/Na ou H/K.

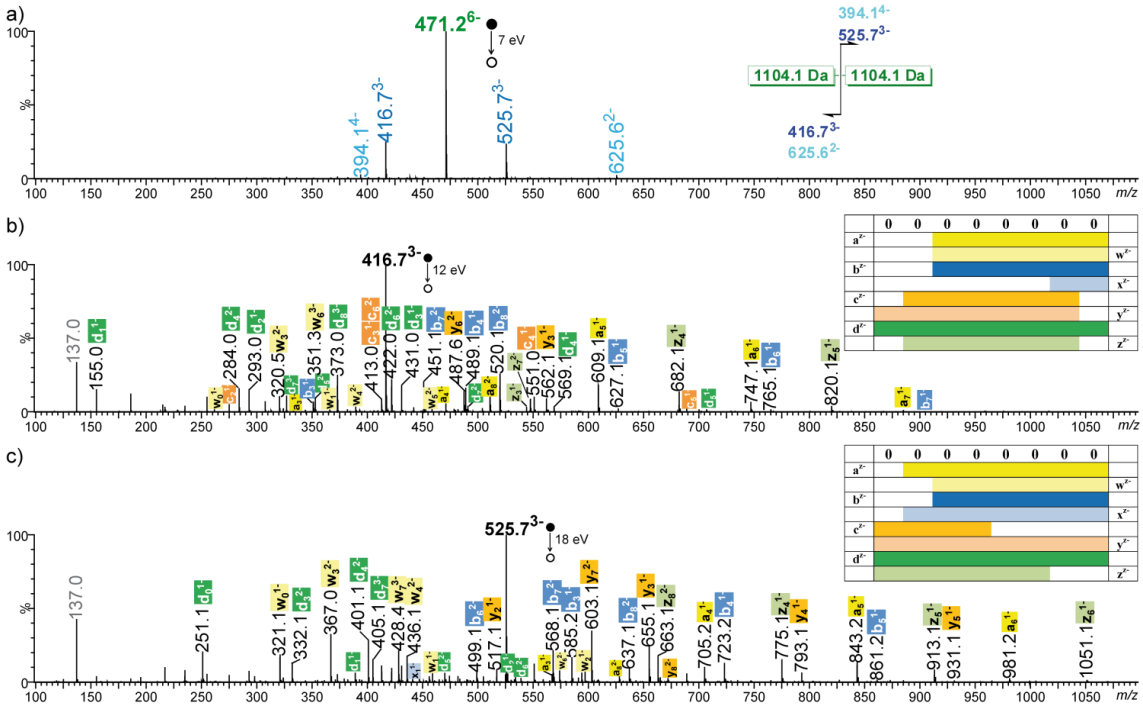


Figure SI-III. 8 : Séquençage de PRISCOP1. (a) Spectre MS/MS de l'ion $[M - 6H]^{6-}$ à m/z *471,2 et schéma d'homolyse de la liaison alcoxyamine. (b) Spectres pseudo-MS^3 du bloc 1 à* m/z *416,7 et couverture de la séquence. (c) Spectres pseudo-MS^3 du bloc 2 à* m/z *525,7. Les monomères déprotonés, $[\mathbf{0} - H]^-$ à* m/z *137,0 pour les deux blocs, sont annotés en gris.*

- **PRISCOP3**

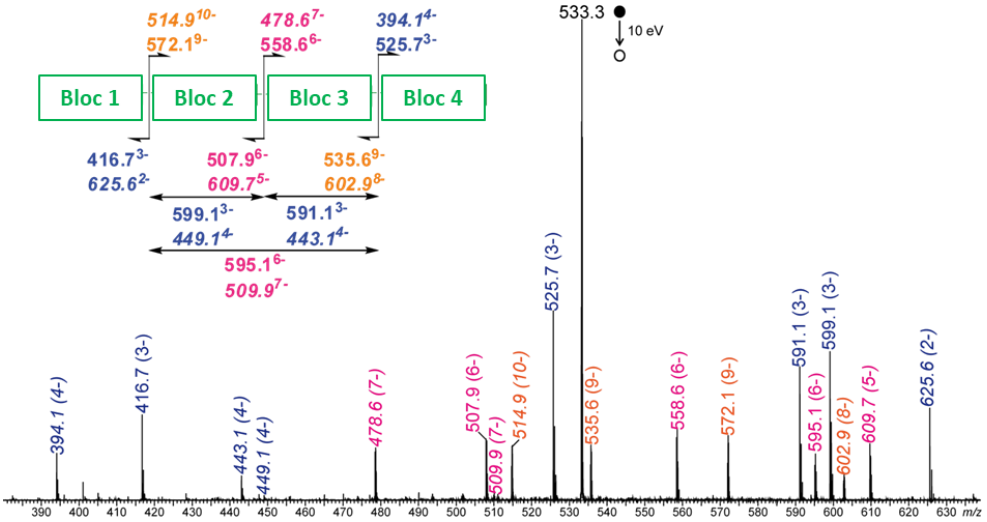


Figure SI-III. 9 : Séquençage de PRISCOP3. Spectre MS/MS de l'ion $[M - 12H]^{12-}$ à m/z *533,3 et schéma d'homolyse des liaisons alcoxyamines.*

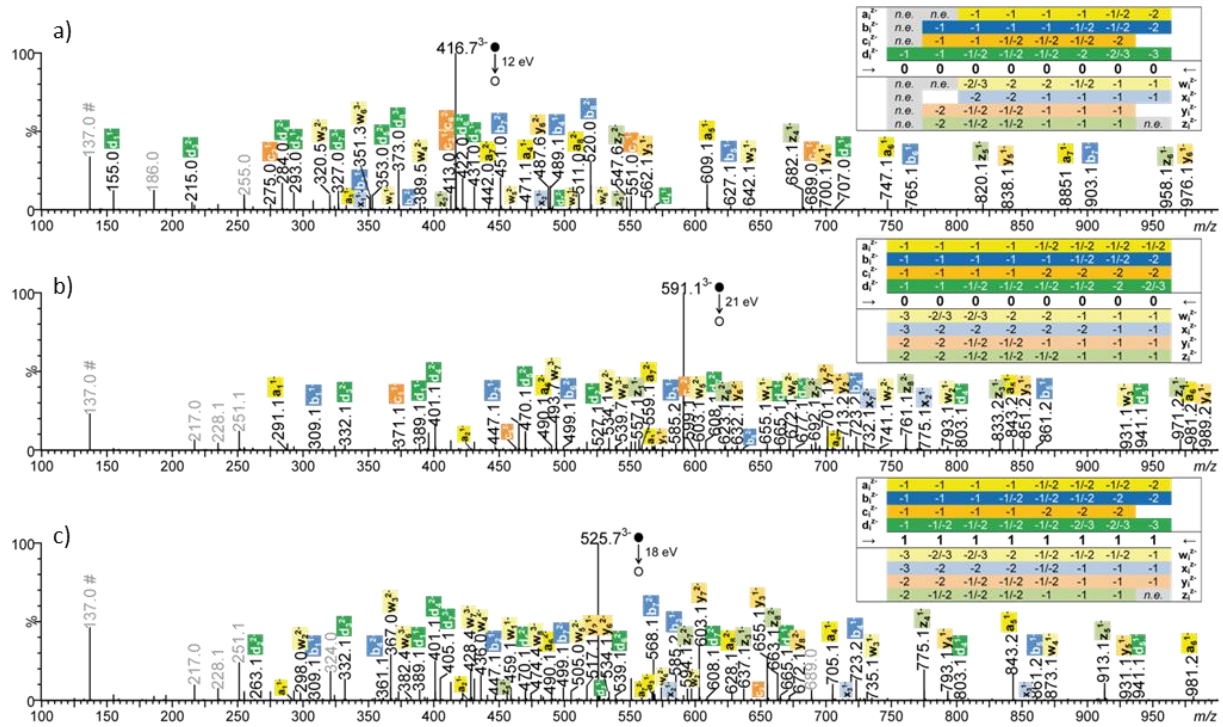


Figure SI-III. 10 : Séquençage en pseudo-MS³de $P_{RISCOP}3$. (a) Spectre pseudo-MS³ du bloc 1 à m/z *416,7 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 3 à* m/z *591,1 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 4 à* m/z *525,7 et couverture de la séquence. Les monomères déprotonés, [**0** – H]⁻ à* m/z *137,0 pour les trois blocs, sont annotés en gris.*

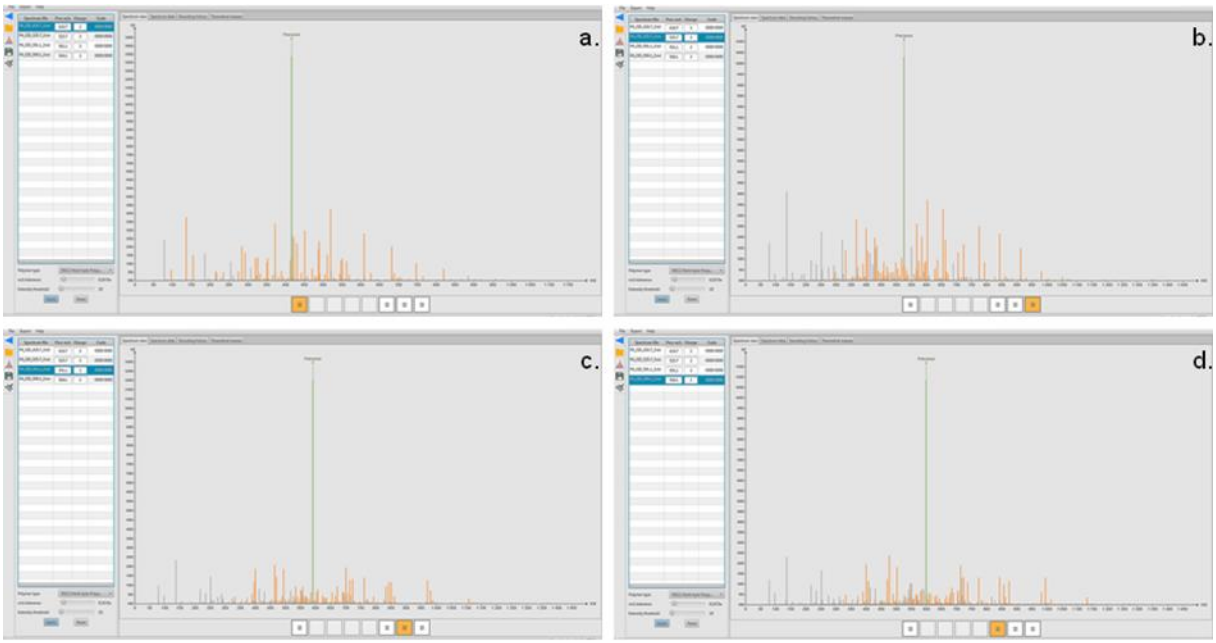


Figure SI-III. 11 : Séquençage automatique de $P_{RISCOP}3$ avec MS-DECODER pour les 4 blocs à (a) m/z 416,7, (b) m/z 525,7, (c) m/z 591,1 et (d) m/z 599,1.

- **P_R1**

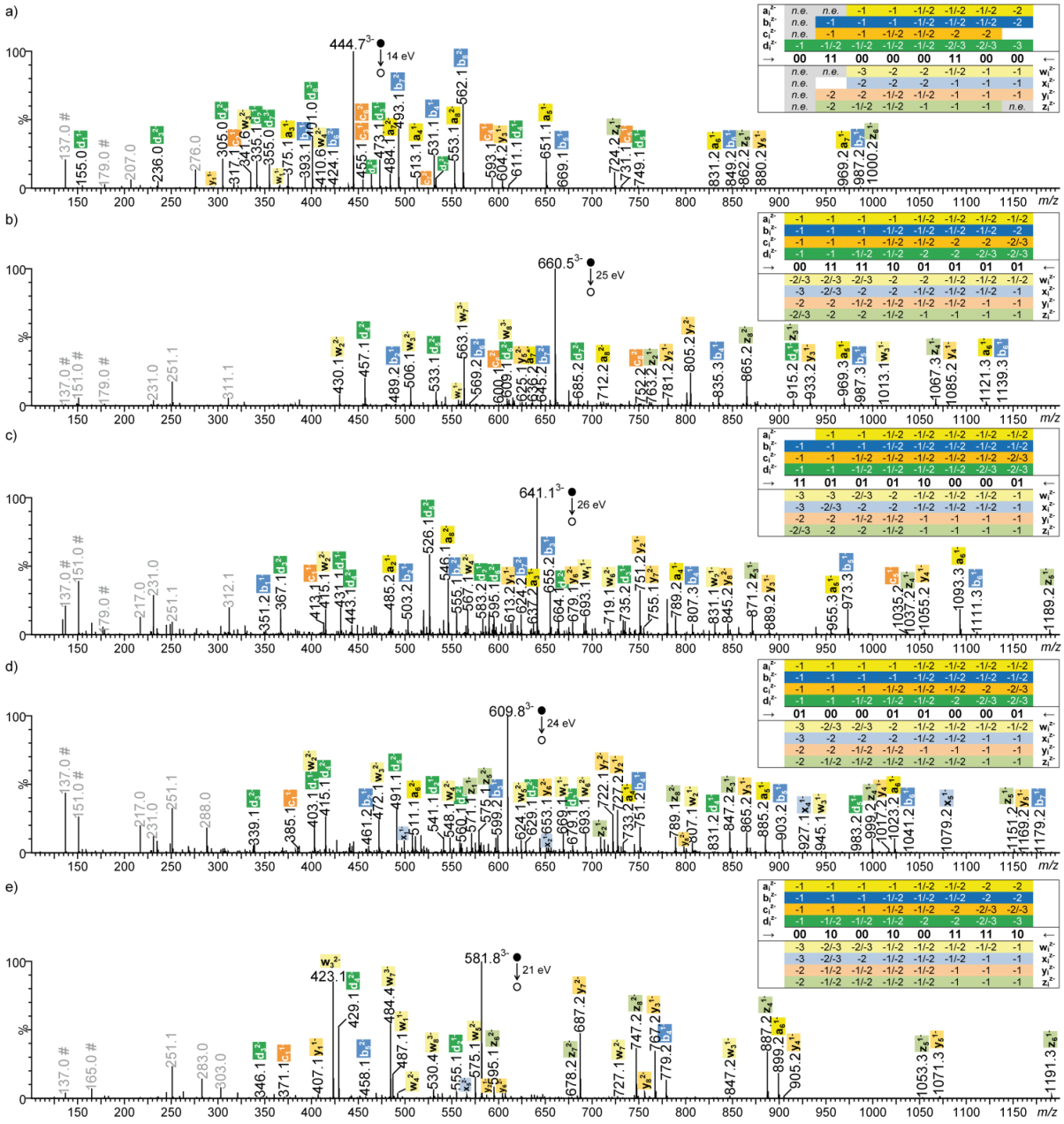


Figure SI-III. 12 : Séquençage en pseudo-MS³de P_R1. (a) Spectre pseudo-MS³ du bloc 1 à m/z *447,7 et couverture de la séquence. (b) Spectre pseudo-MS³ du bloc 2 à* m/z *660,5 et couverture de la séquence. (c) Spectre pseudo-MS³ du bloc 3 à* m/z *641,1 et couverture de la séquence. (d) Spectre pseudo-MS³ du bloc 4 à* m/z *609,8 et couverture de la séquence. (e) Spectre pseudo-MS³ du bloc 5 à* m/z *581,8 et couverture de la séquence. Les monomères déprotonés sont annotés en gris.*

- **P_R2**

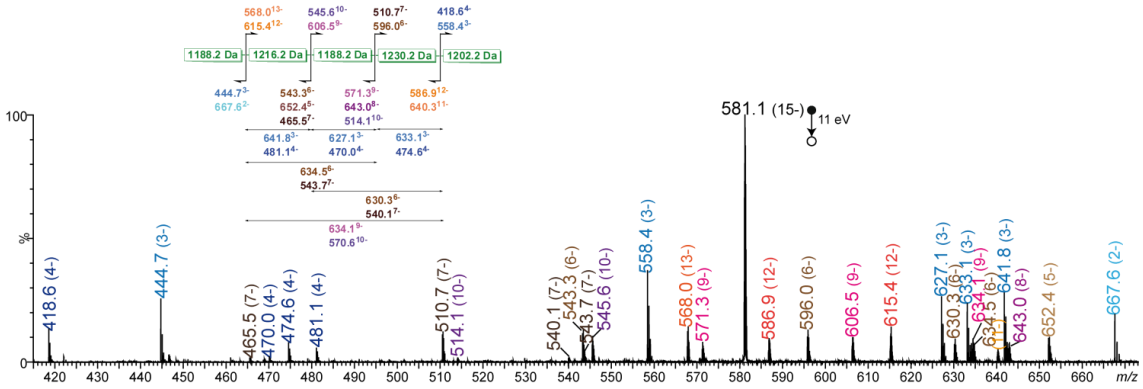


Figure SI-III. 13 : Spectre MS/MS de l'ion $[M - 15H]^{15-}$ à m/z 581,1 de P_R2.

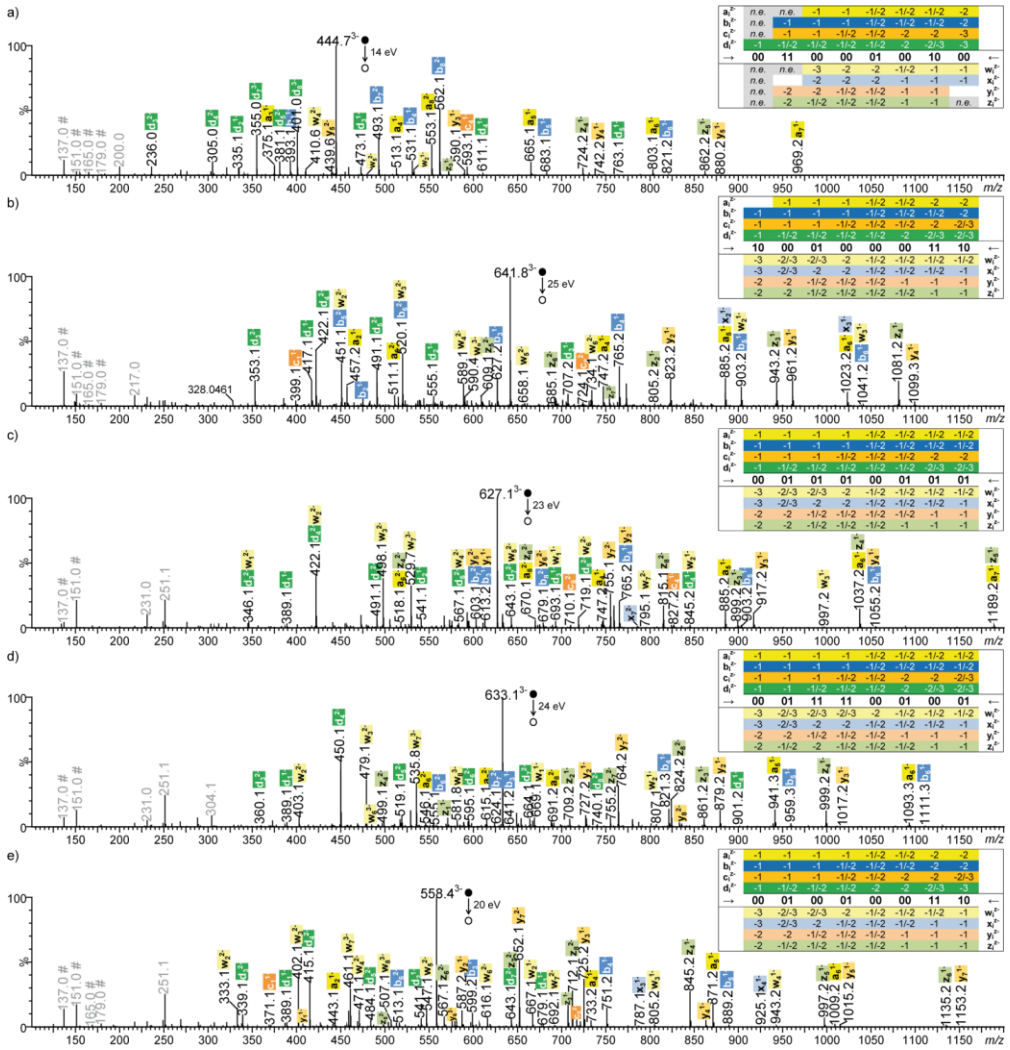


Figure SI-III. 14 : Séquençage en pseudo-MS^3de P_R2. (a) Spectre pseudo-MS^3 du bloc 1 à m/z 447,7 et couverture de la séquence. (b) Spectre pseudo-MS^3 du bloc 2 à m/z 641,8 et couverture de la séquence. (c) Spectre pseudo-MS^3 du bloc 3 à m/z 627,1 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 4 à m/z 633,1 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 5 à m/z 558,4 et couverture de la séquence. Les monomères déprotonés sont annotés en gris.

- **P_R3**

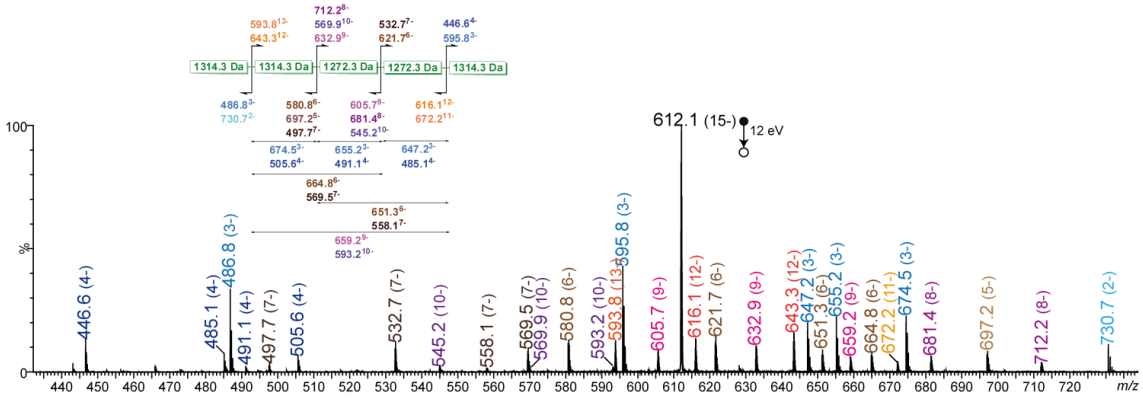


Figure SI-III. 15 : Spectre MS/MS de l'ion $[M - 15H]^{15-}$ à m/z 612,1 du polymère P_R3.

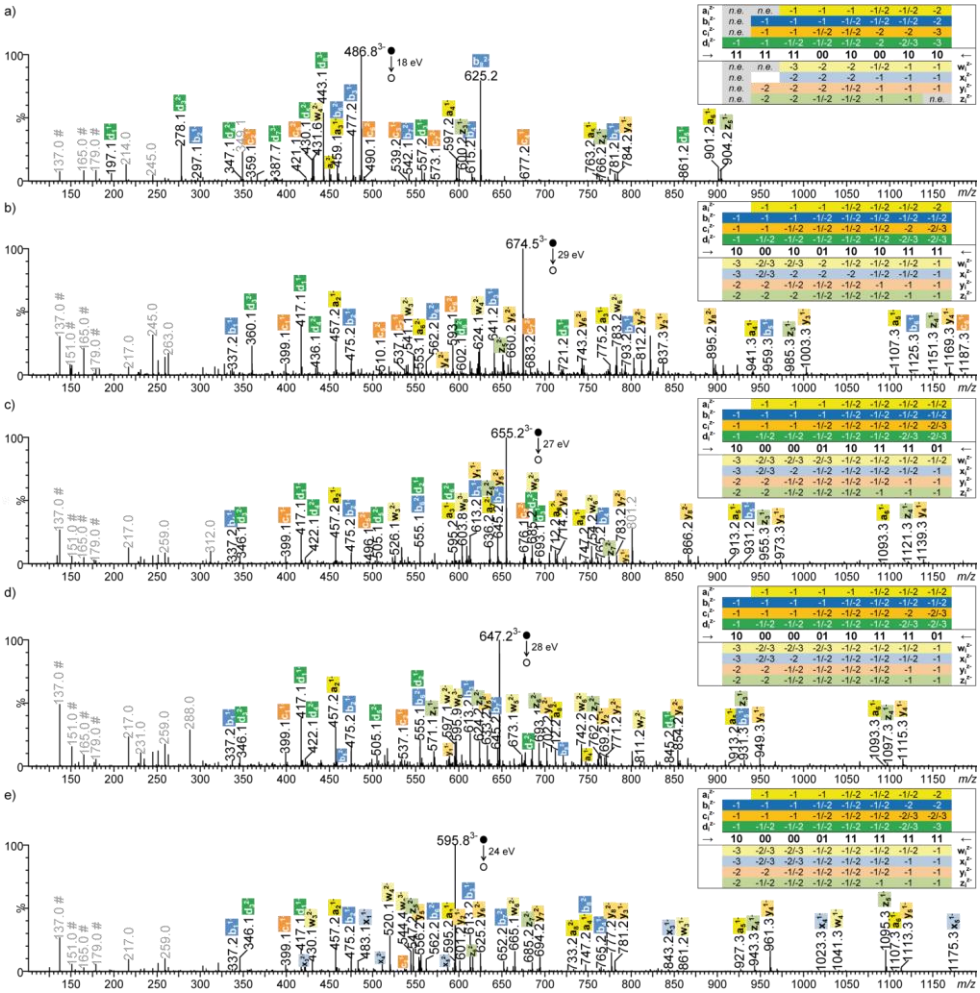


Figure SI-III. 16 : Séquençage en pseudo-MS^3de P_R3. (a) Spectre pseudo-MS^3 du bloc 1 à m/z *486,8 et couverture de la séquence. (b) Spectre pseudo-MS^3 du bloc 2 à* m/z *674,5 et couverture de la séquence. (c) Spectre pseudo-MS^3 du bloc 3 à* m/z *655,2 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 4 à* m/z *647,2 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 5 à* m/z *595,8 et couverture de la séquence. Les monomères déprotonés sont annotés en gris.*

- **P_R4**

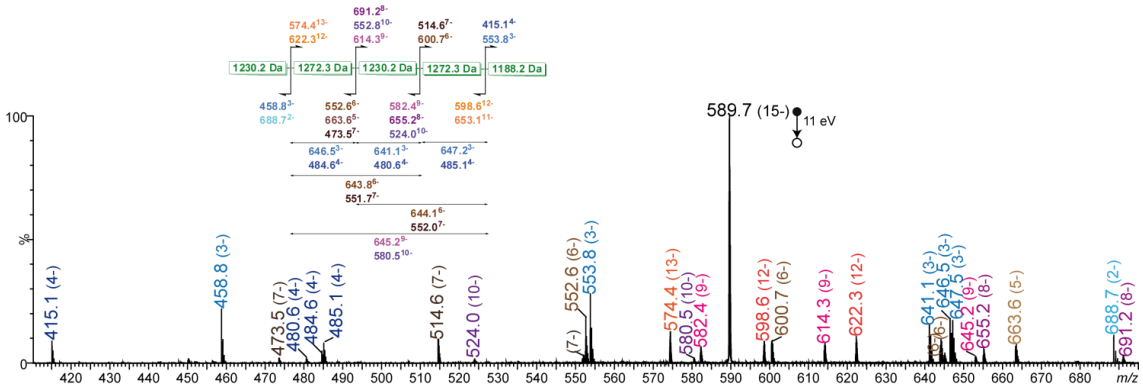


Figure SI-III. 17 : Spectre MS/MS de l'ion $[M - 15H]^{15-}$ à m/z 589,7 du polymère P_R4.

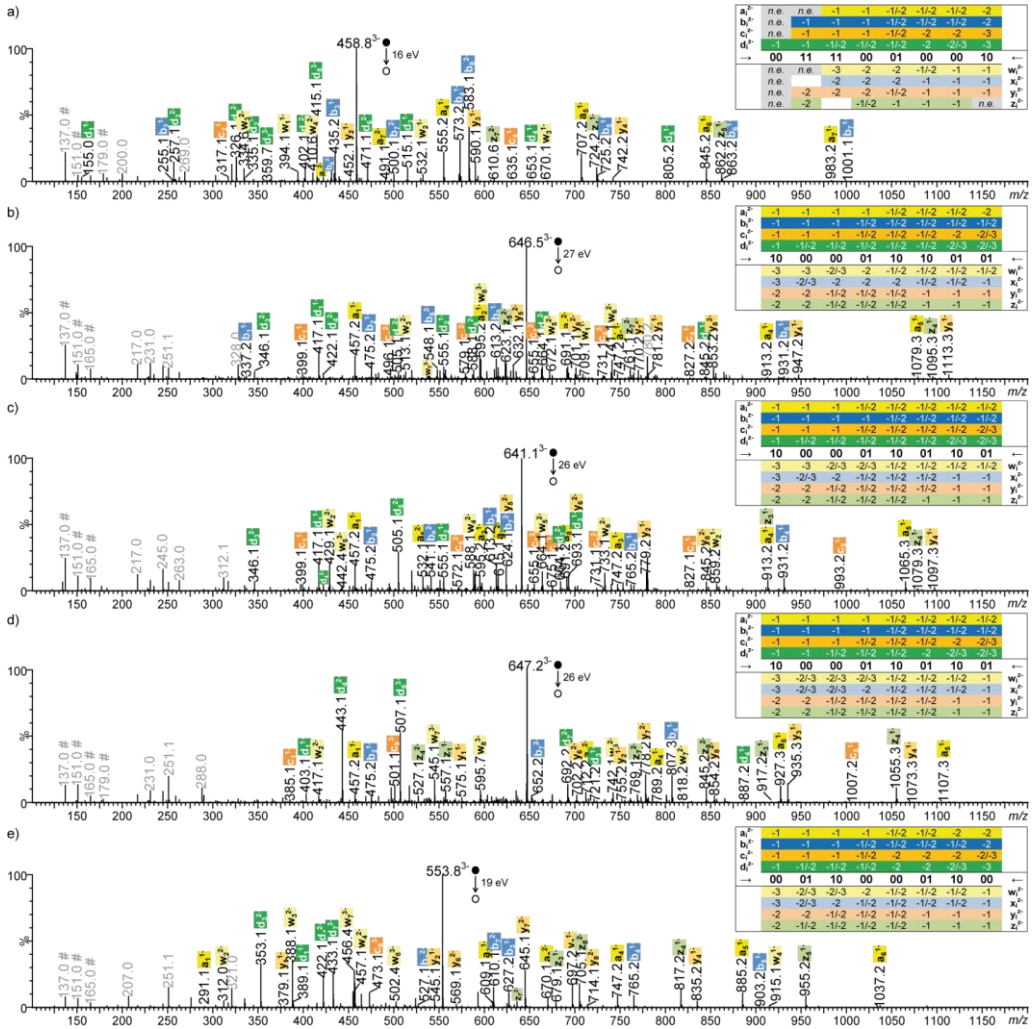


Figure SI-III. 18 : Séquençage en pseudo-MS^3de P_R4. (a) Spectre pseudo-MS^3 du bloc 1 à m/z *458,8 et couverture de la séquence. (b) Spectre pseudo-MS^3 du bloc 2 à* m/z *646,5 et couverture de la séquence. (c) Spectre pseudo-MS^3 du bloc 3 à* m/z *641,1 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 4 à* m/z *647,2 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 5 à* m/z *553,8 et couverture de la séquence. Les monomères déprotonés sont annotés en gris.*

- **P_R5**

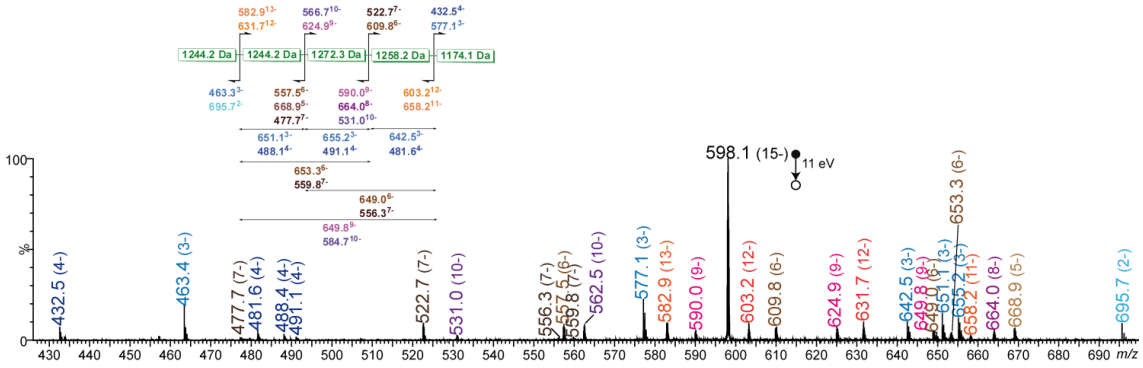


Figure SI-III. 19 : Spectre MS/MS de l'ion $[M - 15H]^{15-}$ à m/z 598,1 du polymère P_R5.

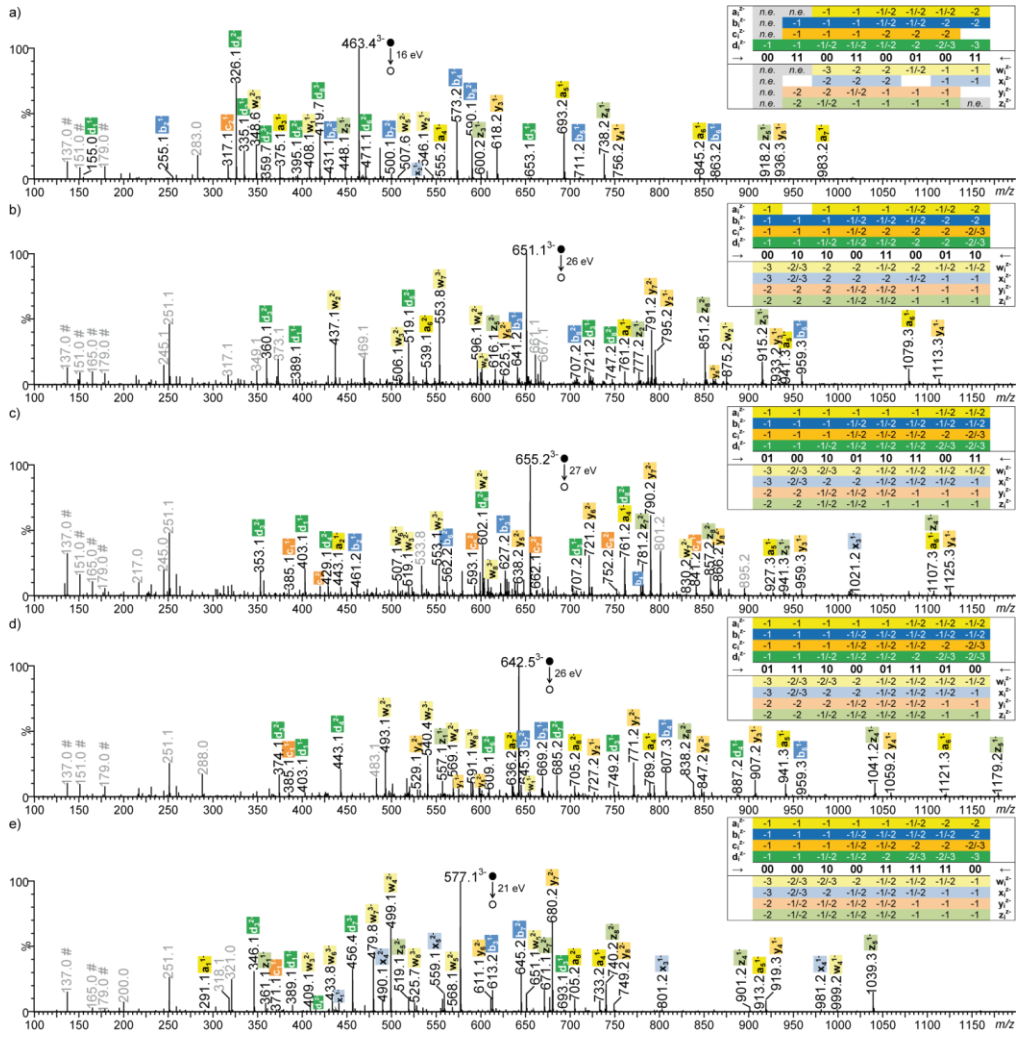


Figure SI-III. 20 : Séquençage en pseudo-MS^3de P_R5. (a) Spectre pseudo-MS^3 du bloc 1 à m/z 463,4 et couverture de la séquence. (b) Spectre pseudo-MS^3 du bloc 2 à m/z 651,1 et couverture de la séquence. (c) Spectre pseudo-MS^3 du bloc 3 à m/z 655,2 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 4 à m/z 642,5 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 5 à m/z 577,1 et couverture de la séquence. Les monomères déprotonés sont annotés en gris.

- **P_R6**

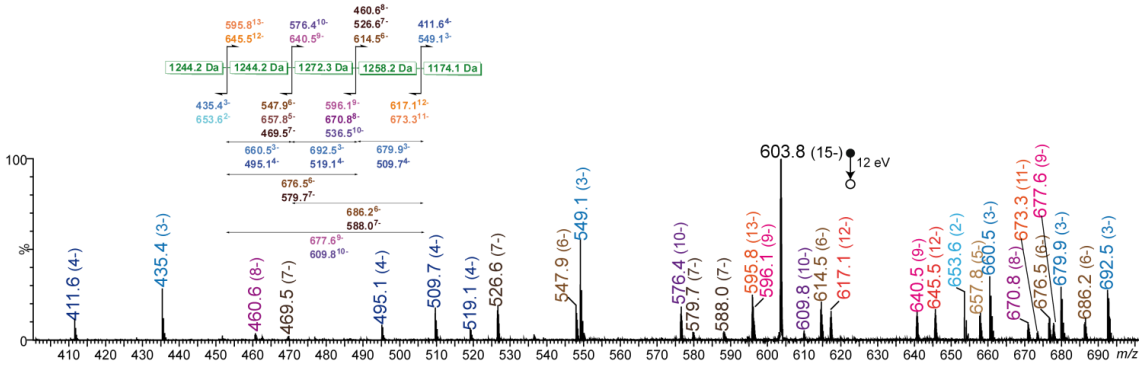


Figure SI-III. 21 : Spectre MS/MS de l'ion $[M - 15H]^{15-}$ à m/z 603,8 du polymère P_R6.

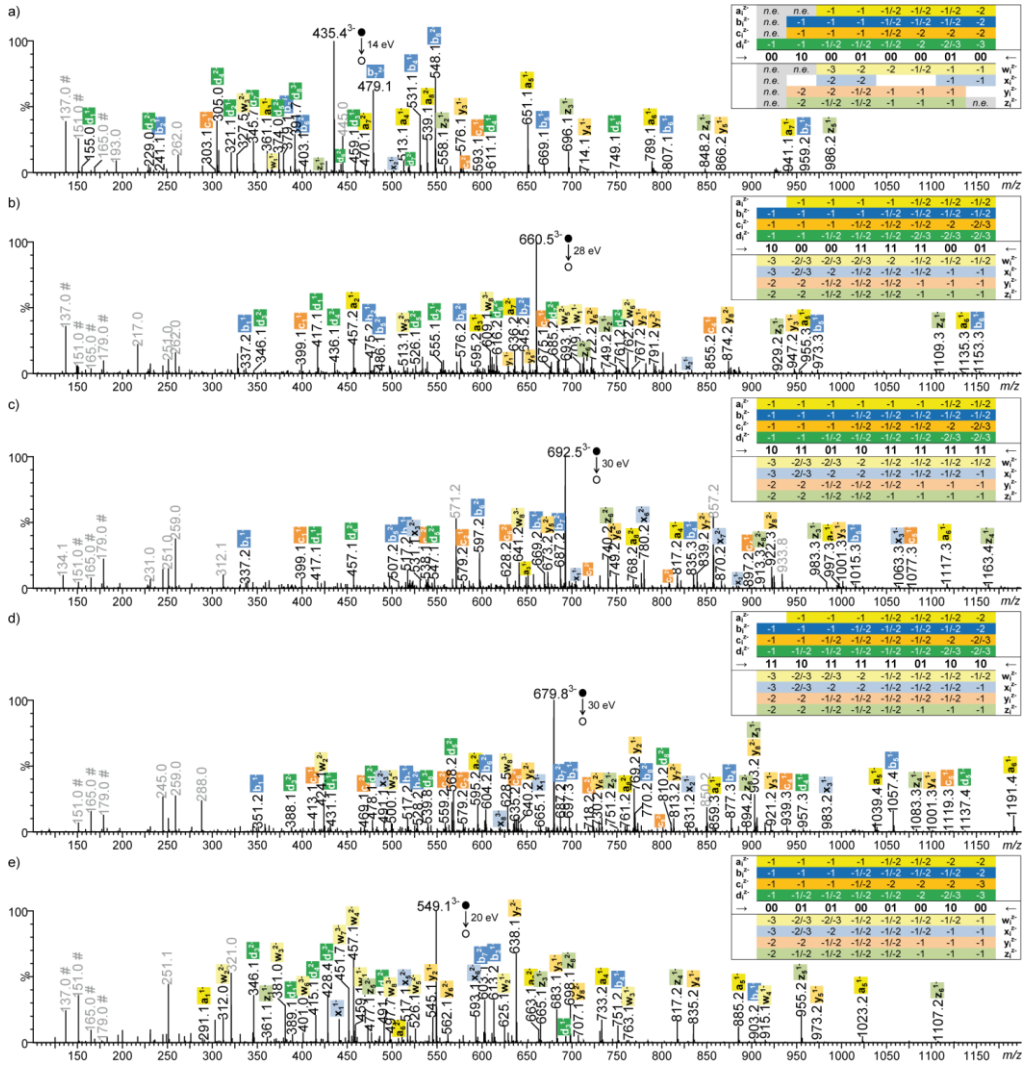


Figure SI-III. 22 : Séquençage en pseudo-MS^3de P_R6. (a) Spectre pseudo-MS^3 du bloc 1 à m/z 435,4 et couverture de la séquence. (b) Spectre pseudo-MS^3 du bloc 2 à m/z 660,5 et couverture de la séquence. (c) Spectre pseudo-MS^3 du bloc 3 à m/z 692,5 et couverture de la séquence. (d) Spectre pseudo-MS^3 du bloc 4 à m/z 679,8 et couverture de la séquence. (e) Spectre pseudo-MS^3 du bloc 5 à m/z 549,1 et couverture de la séquence. Les monomères déprotonés sont annotés en gris.

4.1.4. Séquences impliquant l'espaceur PRISC

- **$P_{PRISC}1$**

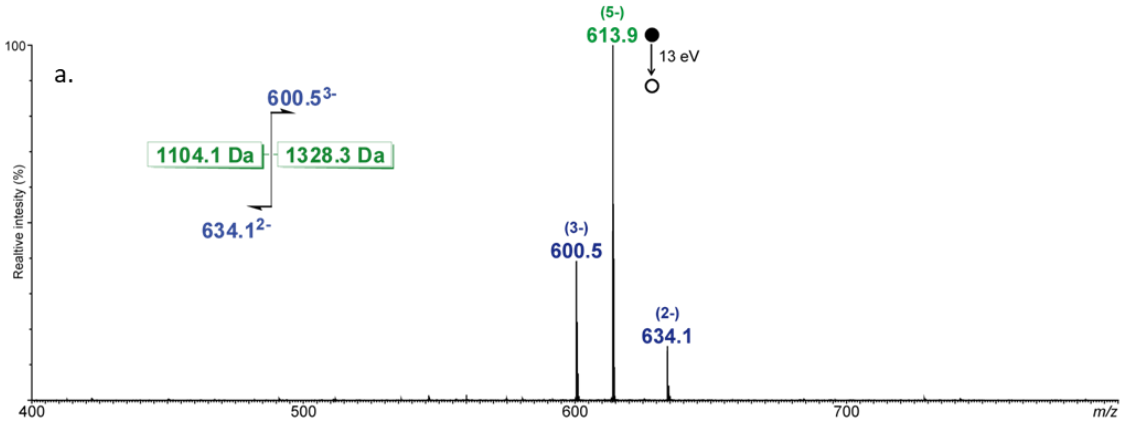


Figure III. 28: Séquençage de $P_{PRISC}1$ et des impuretés. Spectre MS/MS de l'ion $[M - 5H]^{5-}$ à m/z 613,9 et schéma d'homolyse de l'espaceur.

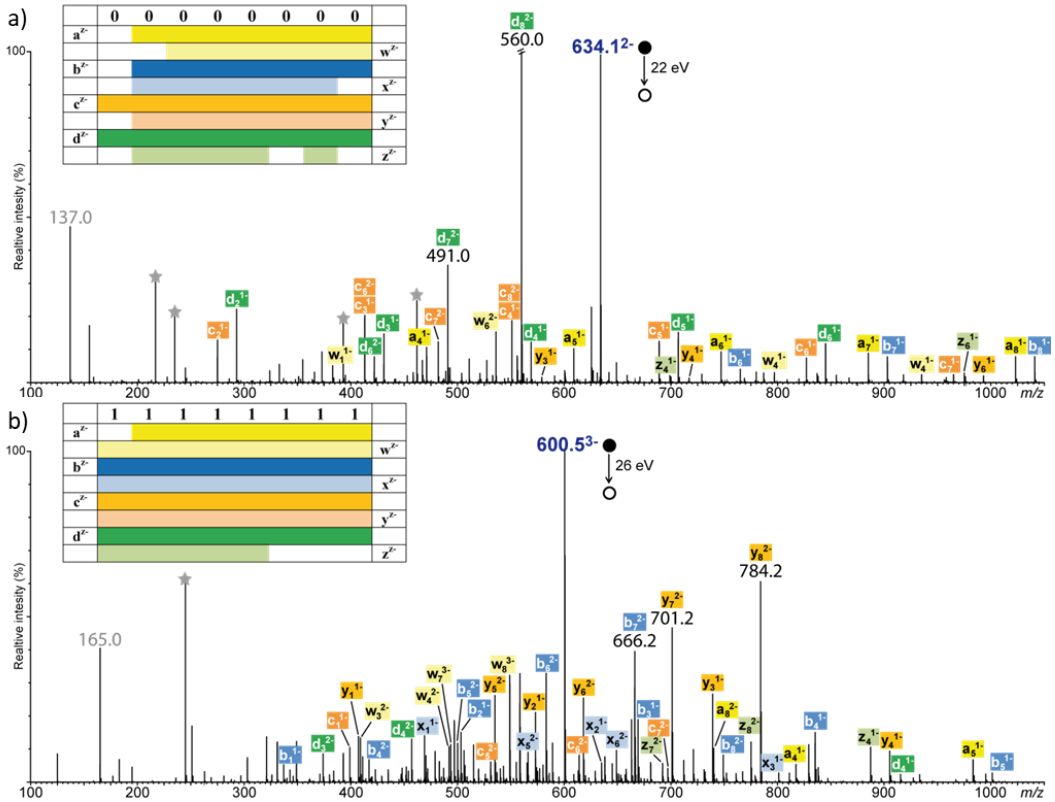


Figure SI-III. 23 : Séquençage de $P_{PRISC}1$. (a) Spectre pseudo-MS^3 du bloc 1 à m/z 634,1. (b) Spectre pseudo-MS^3 du bloc 2 à m/z 600,5. Les monomères déprotonés, $[\mathbf{0} - H]^-$ à m/z 137,0 pour l'octet 1 et $[\mathbf{1} - H]^-$ à m/z 165,0 pour l'octet 2, sont annotés en gris. Les pics désignés par des étoiles correspondent à des fragments secondaires.

- **$P_{PRISC}2$**

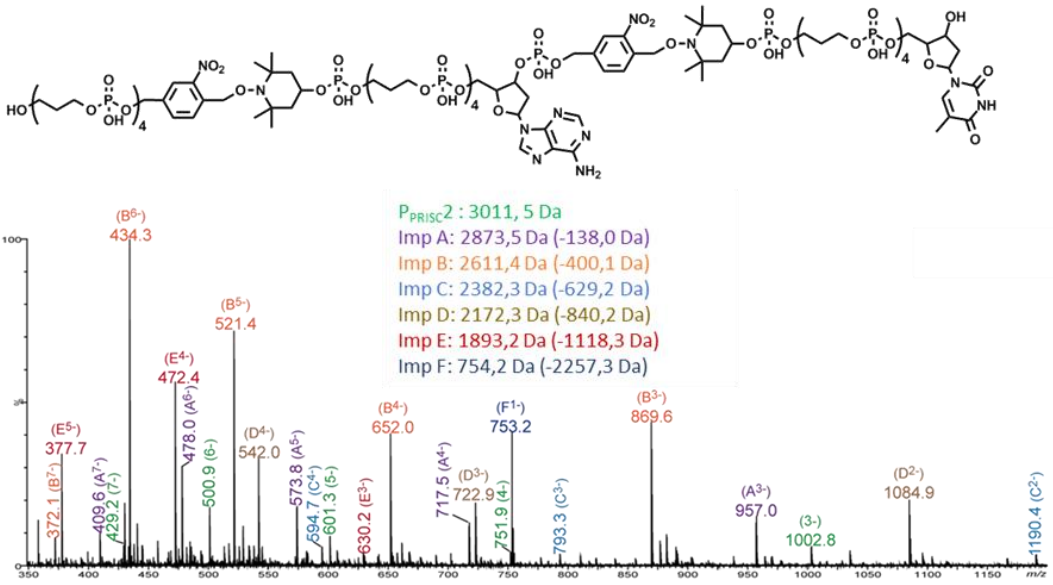


Figure SI-III. 24 : Spectre ESI-MS en mode négatif de $P_{PRISC}2$ (pics en verts). Les pics en couleurs autres que vert correspondent aux impuretés A à F listées.

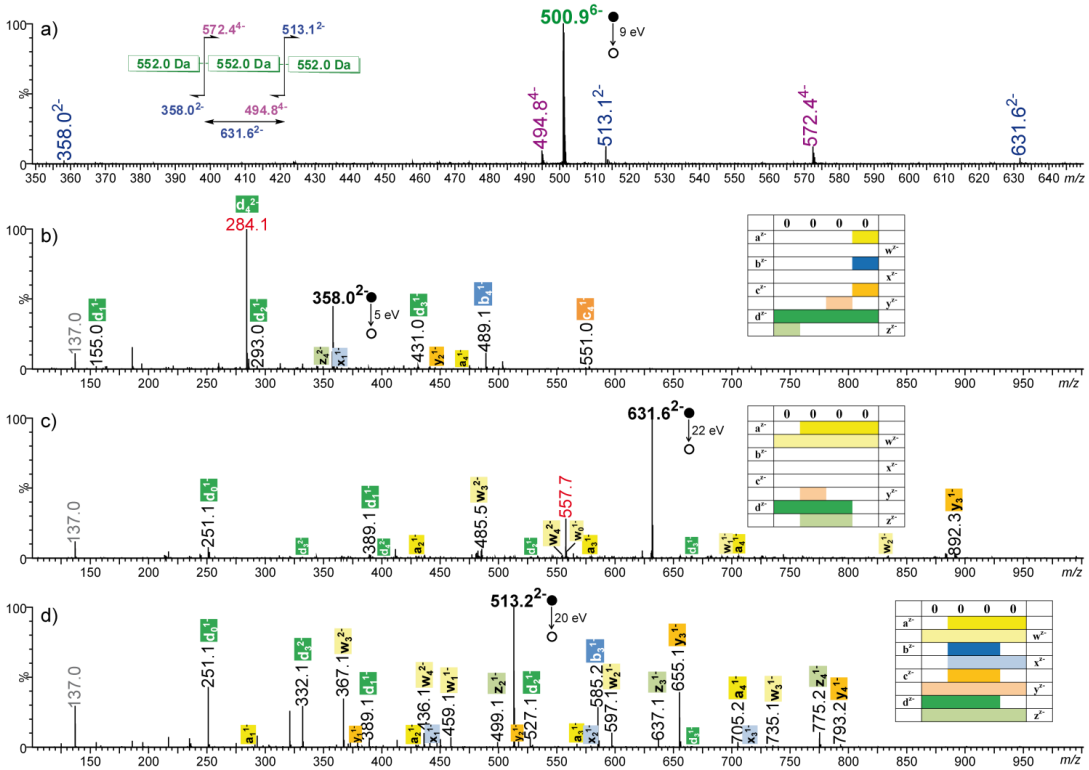


Figure SI-III. 25 : Séquençage de $P_{PRISC}2$. (a) Spectre MS/MS de l'ion $[M - 6H]^{6-}$ à m/z 500,9 et schéma d'homolyse de la liaison alcoxyamine. (b) Spectres pseudo-MS^3 du bloc 1 à m/z 358,0. (c) Spectres pseudo-MS^3 du bloc 2 à m/z 631,6. (d) Spectres pseudo-MS^3 du bloc 3 à m/z 513,2. Le monomère déprotoné, $[\mathbf{0} - H]^-$ à m/z 137,0, est annoté en gris.

4.1.5. Séquences impliquant l'espaceur NISC

- **$P_{NISC}1$**

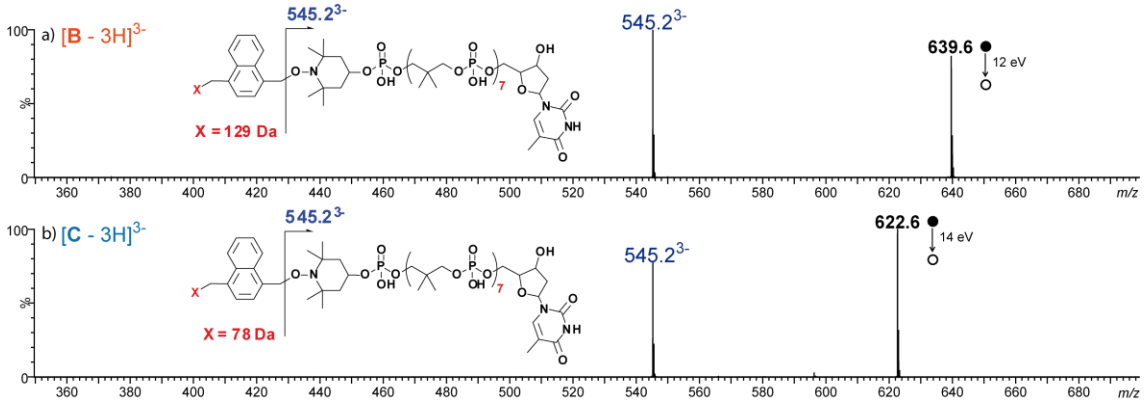


*Figure SI-III. 26 : Spectres ESI-MS/MS des impuretés majeurs détectées dans l'échantillon de $P_{NISC}2$. (a) [**B** – 3H]$^{3-}$ à m/z 639,6 et (b) [**D** – 3H]$^{3-}$ à m/z 622,6.*

- **$P_{NISC}2$**

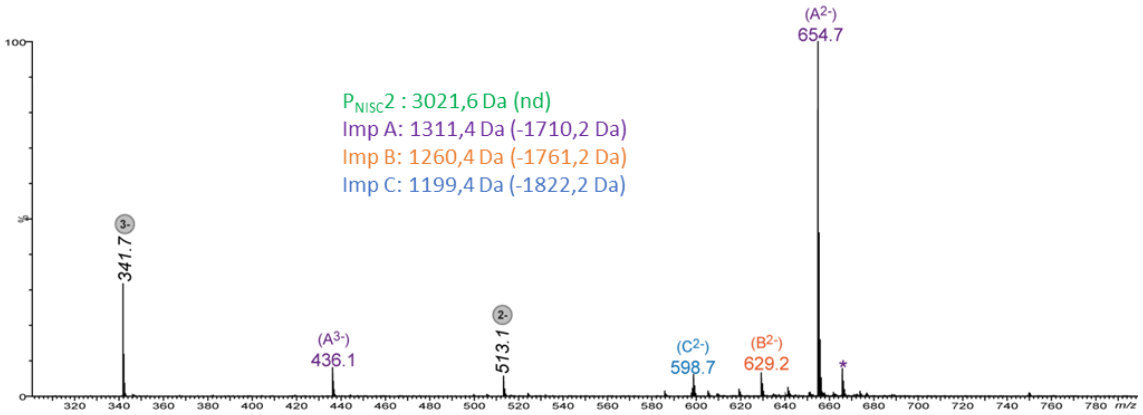


Figure SI-III. 27 : Spectre ESI-MS en mode négatif enregistré pour l'échantillon de $P_{NISC}2$. Les impuretés 1, B et C sont annotées en violet, orange et bleu. Les ions annotés avec un rond gris correspondent au bloc #3 libérés en source. Les pics annotés par un astérisque correspondent à des échanges H/Na.

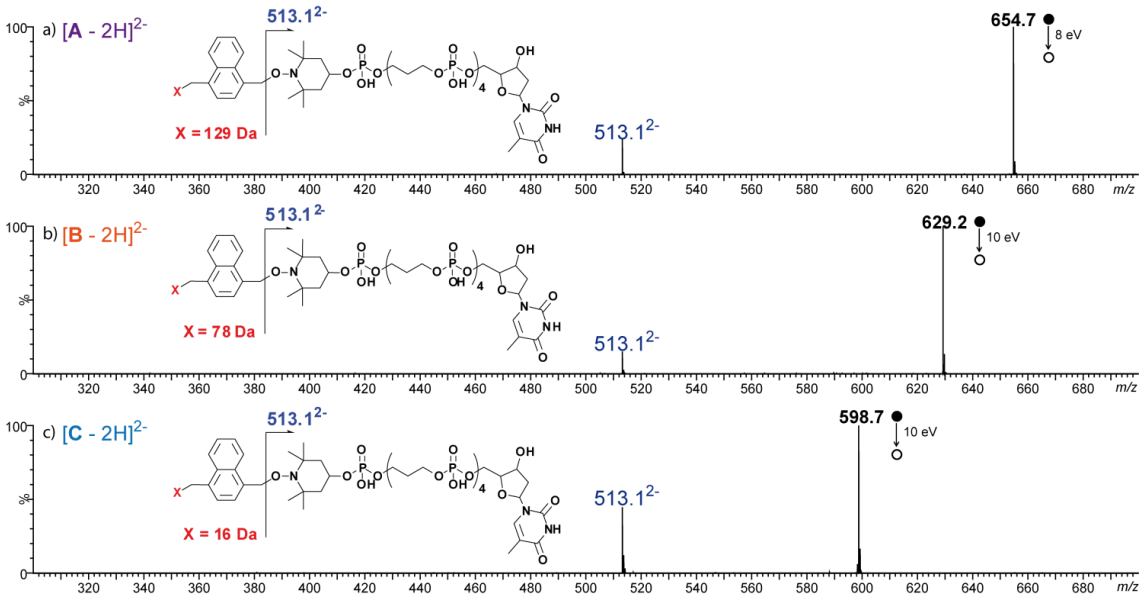


*Figure SI-III. 28: Spectres ESI-MS/MS des impuretés détectées dans l'échantillon de $P_{NISC}2$. (a) [**A** – 2H]$^{2-}$ à m/z 654,7, (b) [**B** – 2H]$^{2-}$ à m/z 629,2 et (c) [**C** – 2H]$^{2-}$ à m/z 598,7.*

4.1.6. Séquence impliquant l'espaceur EM-Br

- **$P_{EM\text{-}Br}1$**

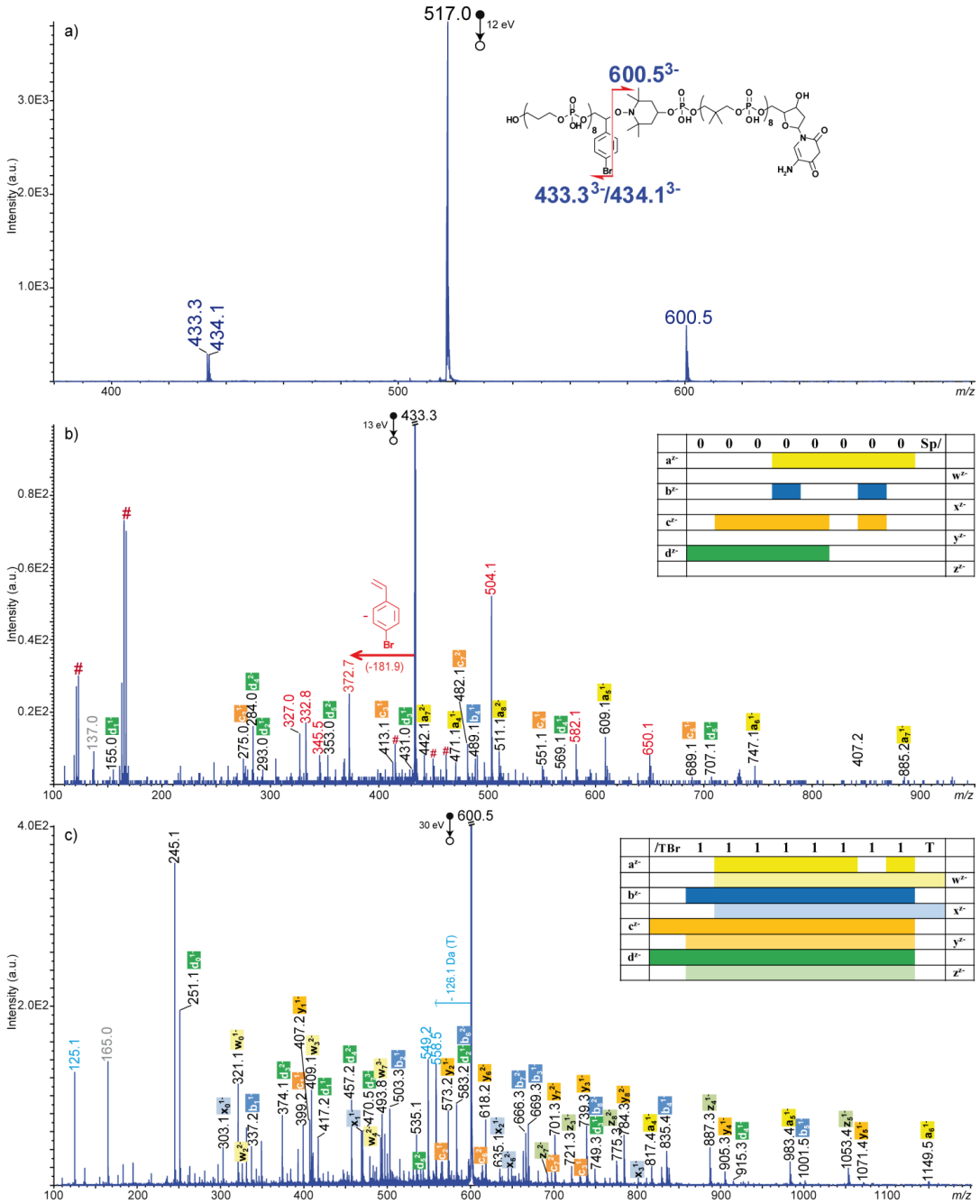


Figure SI-III. 29 : Séquençage de $P_{EM\text{-}Br}1$. (a) Spectre MS/MS de l'ion $[M - 6H]^{6-}$ à m/z 517,0 et schéma d'homolyse des liaisons C–ON. (b) Spectre pseudo-MS^3 du bloc 1 m/z 433,3. Les pics annotés en rouge sont des fragments secondaires après perte de l'espaceur EM-Br et ceux annotés par # résultent d'interférences en source. Le tableau de couverture de séquence montre qu'aucun fragment contenant le radical styrényl n'est observé. (c) Spectre pseudo-MS^3 du bloc 2 à m/z 600,5. Les ions annotés en cyan sont liés à la perte de la base T. Les tableaux en inset montrent les couvertures de chaque bloc. Dans les spectres (b) et (c) les pics annotés en gris correspondent aux monomères déprotonés.

5. Partie expérimentale du chapitre IV

- **Lav2**

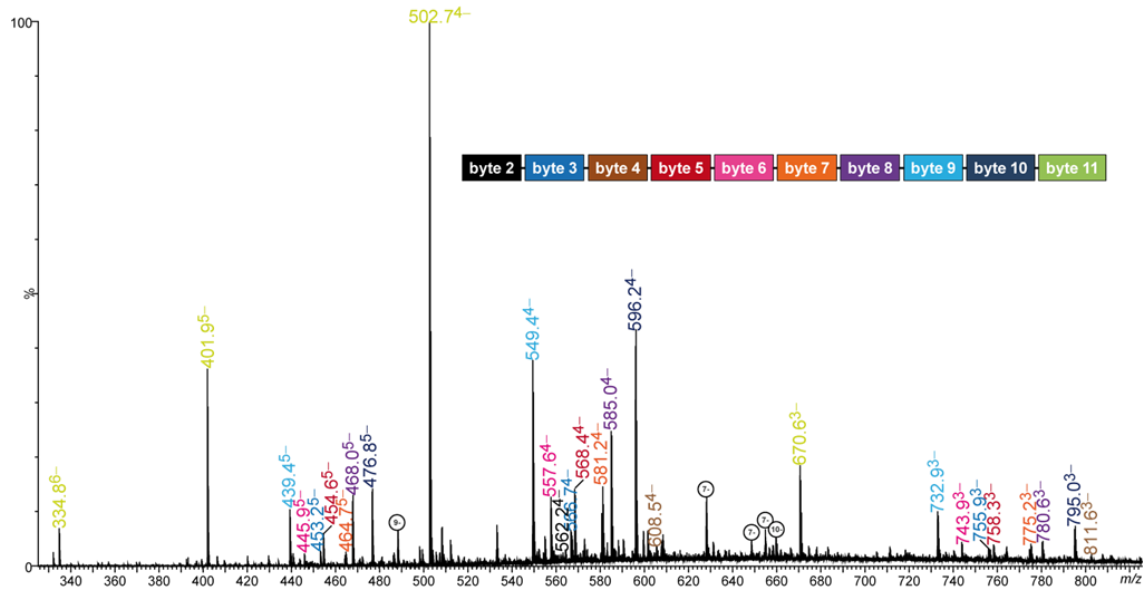


Figure SI-IV. 1 : Spectre ESI-MS en mode négatif de la séquence Lav2. Les signaux correspondant aux blocs libérés en sources sont annotés de la même couleur que sur le schéma de la séquence.

- **Lav3**

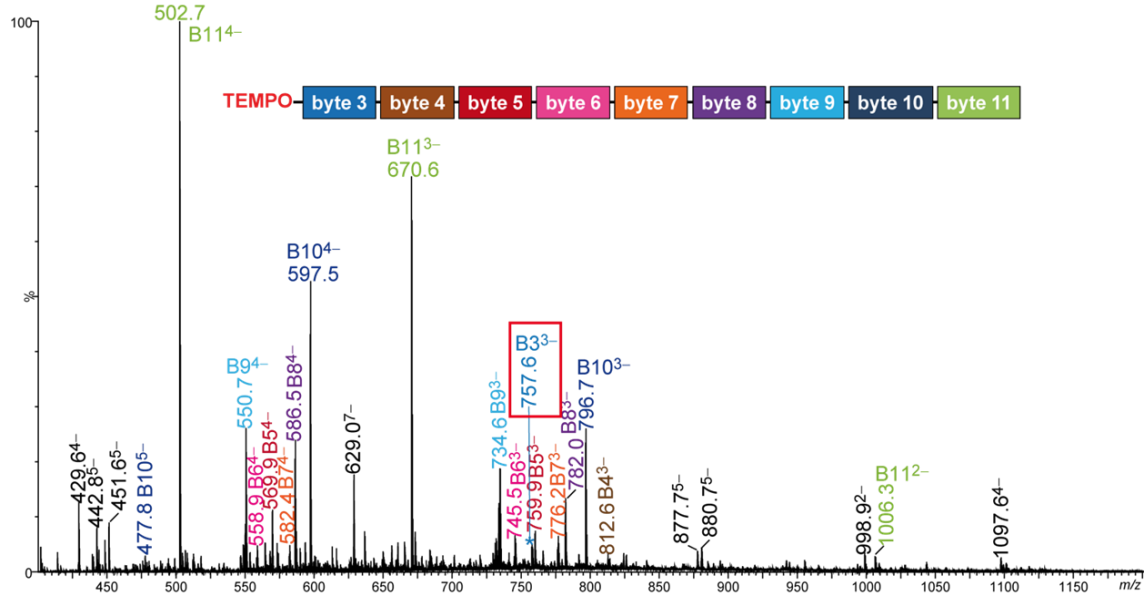


Figure SI-IV. 2 : Spectre ESI-MS en mode négatif de la séquence Lav3. Les signaux correspondant aux blocs libérés en sources sont annotés de la même couleur que sur le schéma de la séquence.

- **Lav4**

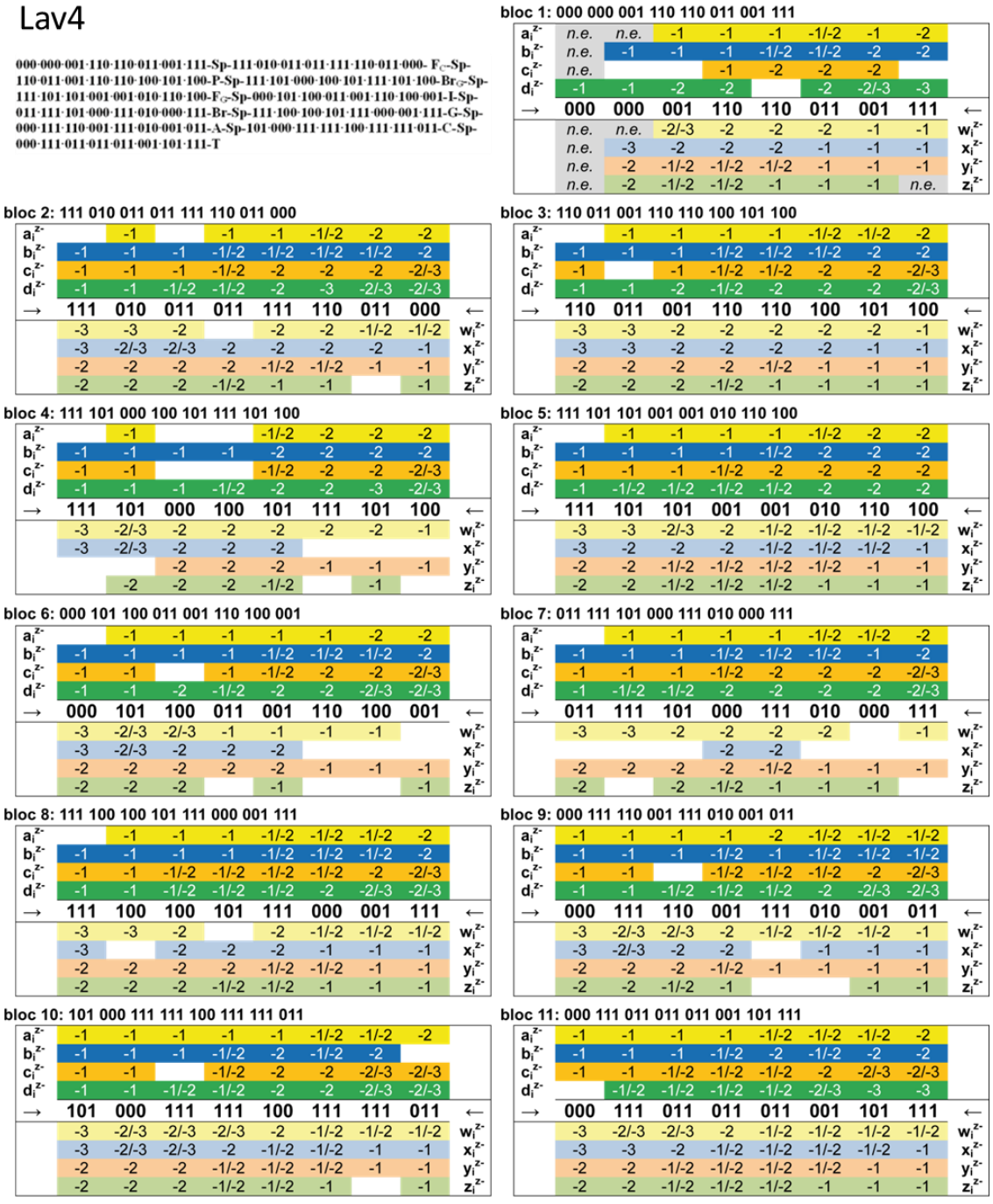


Figure SI-IV. 3 : Tableaux des couvertures obtenus pour chaque bloc de la séquence Lav4.

Références bibliographiques

1. D. Reinsel, J. Gantz and J. Rydning, *The Digitization of the World - From Edge to Core*, Report US44413318, 2018.
2. Greenpeace, *Powering the Cloud: How China's Internet Industry Can Shift to Renewable Energy*, China Water Risk 2019.
3. N. Jones, *Nature*, 2018, **561**, 163-167.
4. H. Colquhoun and J.-F. Lutz, *Nature chemistry*, 2014, **6**, 455-456.
5. J. Judge, J. Pouchet, A. Ekbote and S. Dixit, *ASHRAE Journal*, 2008, 14-26.
6. C. A. P. Goodwin, F. Ortu, D. Reta, N. F. Chilton and D. P. Mills, *Nature*, 2017, **548**, 439-442.
7. F. D. Natterer, K. Yang, W. Paul, P. Willke, T. Choi, T. Greber, A. J. Heinrich and C. P. Lutz, *Nature*, 2017, **543**, 226-228.
8. M. Julliere, *Physics Letters A*, 1975, **54**, 225-226.
9. J. D. Watson and F. H. Crick, *Nature*, 1953, **171**, 964-967.
10. V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church and W. L. Hughes, *Nature materials*, 2016, **15**, 366-370.
11. G. M. Church, Y. Gao, S. Kosuri and C. T. Clelland, *Science* 2012, **337**, 1628-1628.
12. N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos and E. Birney, *Nature*, 2013, **494**, 77-80.
13. L. Ceze, J. Nivala and K. Strauss, *Nature Reviews Genetics*, 2019, **20**, 456-466.
14. L. Organick, S. D. Ang, Y. J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H. Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze and K. Strauss, *Nature biotechnology*, 2018, **36**, 242-248.
15. M. G. T. A. Rutten, F. W. Vaandrager, J. A. A. W. Elemans and R. J. M. Nolte, *Nature Reviews Chemistry*, 2018, **2**, 365-381.
16. J.-F. Lutz, J.-M. Lehn, E. W. Meijer and K. Matyjaszewski, *Nature Reviews Materials*, 2016, **1**, 16024.
17. W. H. Carothers, *Chemical Reviews* 1931, **8**.
18. R. B. Merrifield, *Journal of the American Chemical Society*, 1963, **85**, 2149-2154.
19. G. Fiers, D. Chouikhi, L. Oswald, A. Al Ouahabi, D. Chan-Seng, L. Charles and J.-F. Lutz, *Chemistry – A European Journal*, 2016, **22**, 17945-17948.
20. T. T. Trinh, L. Oswald, D. Chan-Seng and J. F. Lutz, *Macromolecular Rapid Communications*, 2014, **35**, 141-145.
21. R. K. Roy, C. Laure, D. Fischer-Krauser, L. Charles and J.-F. Lutz, *Chemical Communications*, 2015, **51**, 15677-15680.
22. Ufuk S. Gunay, Benoît E. Petit, D. Karamessini, A. Al Ouahabi, J.-A. Amalian, C. Chendo, M. Bouquey, D. Gigmes, L. Charles and J.-F. Lutz, *Chem*, 2016, **1**, 114-126.
23. A. Al Ouahabi, L. Charles and J.-F. Lutz, *Journal of the American Chemical Society*, 2015, **137**, 5629-5635.
24. A. Al Ouahabi, M. Kotera, L. Charles and J.-F. Lutz, *ACS Macro Letters*, 2015, **4**, 1077-1080.

25. A. Al Ouahabi, J.-A. Amalian, L. Charles and J.-F. Lutz, *Nature Communications*, 2017, **8**, 967-967.
26. IBM, The Floppy Disk, https://www.ibm.com/ibm/history/ibm100/us/en/icons/floppy/, (accessed 08 31, 2020).
27. M. N. Baibich, J. M. Broto, A. Fert, F. Nguyen Van Dau, F. Petroff, P. Etienne, G. Creuzet, A. Friederich and J. Chazelas, *Physical Review Letters*, 1988, **61**, 2472-2475.
28. G. Binasch, P. Grünberg, F. Saurenbach and W. Zinn, *Physical Review B*, 1989, **39**, 4828-4830.
29. J. James, *Journal of Histochemistry & Cytochemistry*, 1970, **18**, 217-219.
30. O. T. Avery , C. M. MacLeod and M. McCarty *The Journal of Experimental Medicine*, 1944, **79**, 137-158.
31. M. H. F. Wilkins, A. R. Stokes and H. R. Wilson, *Nature*, 1953, **171**, 738-740.
32. M. Meselson and F. W. Stahl, *Proceedings of the National Academy of Sciences*, 1958, **44**, 671.
33. R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto and A. Sugino, *Proc Natl Acad Sci U S A*, 1968, **59**, 598-605.
34. G. M. Blackburn and M. J. Gait, *Nucleic acids in chemistry and biology*, Oxford, 1990.
35. S. Brenner, F. Jacob and M. Meselson, *Nature*, 1961.
36. D. L. Nelson and M. M. Cox, *Principles Of Biochemistry*, Lehninger, 5th edn., 2008.
37. J. M. Berg, J. L. Tymoczko, G. J. Gatto and L. Stryer, *Biochemistry* W. H. Freeman, 8th revised edn., 2015.
38. E. Buxbaum, *Fundamentals of Protein Structure and Function*, Springer, 2015.
39. D. Maiorano, M. Lutzmann and M. Méchali, *Current Opinion in Cell Biology*, 2006, **18**, 130-136.
40. E. Fischer, *Berichte der Deutschen Chemischen Gesellschaft*, 1903, **36**, 2982-2992.
41. R. B. Merrifield, *Science*, 1965, **150**, 178-185.
42. R. B. Merrifield, *Science*, 1986, **232**, 341-347.
43. M. Amblard, J. A. Fehrentz, J. Martinez and G. Subra, *Molecular biotechnology*, 2006, **33**, 239-254.
44. A. M. Michelson and A. R. Todd, *Journal of the Chemical Society (Resumed)*, 1955, DOI: 10.1039/jr9550002632, 2632-2638.
45. H. G. Khorana, G. Tener, J. Moffatt and E. Pol, *Chemistry & Industry* 1956, 1523.
46. H. G. Khorana, W. Razzell, P. Gilham, G. Tener and E. Pol, *Journal of the American Chemical Society*, 1957, **79**, 1002-1003.
47. M. Smith, D. H. Rammler, I. H. Goldberg and H. G. Khorana, *Journal of the American Chemical Society*, 1962, **84**, 430-440.
48. H. Schaller, G. Weimann, B. Lerch and H. G. Khorana, *Journal of the American Chemical Society*, 1963, **85**, 3821-3827.
49. M. H. Caruthers, *Resonance*, 2012, 1143-1156.
50. Y. Lapidot and H. G. Khorana, *Journal of the American Chemical Society*, 1963, **85**, 3857-3862.
51. D. Söll, E. Ohtsuka, D. S. Jones, R. Lohrmann, H. Hayatsu, S. Nishimura and H. G. Khorana, *Proceedings of the National Academy of Sciences*, 1965, **54**, 1378.

52. K. L. Agarwal, H. Büchi, M. H. Caruthers, N. Gupta, H. G. Khorana, K. Kleppe, A. Kumar, E. Ohtsuka, U. L. Rajbhandary, J. H. Van De Sande, V. Sgaramella, H. Weber and T. Yamada, *Nature*, 1970, **227**, 27-34.
53. H. G. Khorana, *Resonance*, 2012, **17**, 1174-1197.
54. H. G. Khorana, *Science*, 1979, **203**, 614.
55. R. L. Letsinger and M. J. Kornet, *Journal of the American Chemical Society*, 1963, **85**, 3045-3046.
56. R. L. Letsinger and V. Mahadevan, *Journal of the American Chemical Society*, 1966, **88**, 5319-5324.
57. R. L. Letsinger and V. Mahadevan, *Journal of the American Chemical Society*, 1965, **87**, 3526-3527.
58. L. R. Melby and D. R. Strobach, *Journal of the American Chemical Society*, 1967, **89**, 450-453.
59. H. Hayatsu and H. G. Khorana, *Journal of the American Chemical Society*, 1967, **89**, 3880-3887.
60. F. Cramer, R. Helbig, H. Hettler, K. H. Scheit and H. Seliger, *Angewandte Chemie International Edition*, 1966, **5**, 601-601.
61. R. L. Letsinger, M. H. Caruthers and D. M. Jerina, *Biochemistry*, 1967, **5**, 1379-1388.
62. R. L. Letsinger, K. K. Ogilvie and P. S. Miller, *Journal of the American Chemical Society*, 1969, **91**, 3360-3365.
63. R. L. Letsinger and P. S. Miller, *Journal of the American Chemical Society*, 1969, **91**, 3356-3359.
64. G. W. Grams and R. L. Letsinger, *The Journal of Organic Chemistry*, 1970, **35**, 868-870.
65. R. L. Letsinger, J. L. Finnan, G. A. Heavner and W. B. Lunsford, *Journal of the American Chemical Society*, 1975, **97**, 3278-3279.
66. R. L. Letsinger and W. B. Lunsford, *Journal of the American Chemical Society*, 1976, **98**, 3655-3661.
67. M. D. Matteucci and M. H. Caruthers, *Tetrahedron Letters*, 1980, **21**, 719-722.
68. M. D. Matteucci and M. H. Caruthers, *Journal of the American Chemical Society*, 1981 **103**, 3185-3191.
69. S. L. Beaucage and M. H. Caruthers, *Tetrahedron Letters*, 1981, **22**, 1859-1862.
70. L. J. McBride and M. H. Caruthers, *Tetrahedron Letters*, 1983, **24**, 245-248.
71. S. P. Adams, K. S. Kavka, E. J. Wykes, S. B. Holder and G. R. Galluppi, *Journal of the American Chemical Society*, 1983, **105**, 661-663.
72. J. A. Menosky, *The Washington Post*, 1981, **C1**.
73. M. H. Caruthers, *Journal of Biological Chemistry*, 2013, **288**, 1420-1427.
74. E. M. LeProust, B. J. Peck, K. Spirin, H. B. McCuen, B. Moore, E. Namsaraev and M. H. Caruthers, *Nucleic Acids Research*, 2010, **38**, 2522-2540.
75. M. P. Reddy, N. B. Hanna and F. Farooqui, *Tetrahedron Letters*, 1994, **35**, 4311-4314.
76. C. McCollum and A. Andrus, *Tetrahedron Letters*, 1991, **32**, 4069-4072.
77. H. Köster, A. Stumpe and A. Wolter, *Tetrahedron Letters*, 1983, **24**, 747-750.
78. S. L. Beaucage, *Tetrahedron Letters*, 1984, **25**, 375-378.
79. M. P. Reddy, N. B. Hanna and F. Farooqui, *Nucleosides and Nucleotides*, 1997, **16**, 1589-1598.

80. S. L. Beaucage and R. P. Iyer, *Tetrahedron*, 1992, **48**, 2223-2311.
81. R. N. Grass, R. Heckel, M. Puddu, D. Paunescu and W. J. Stark, *Angewandte Chemie International Edition*, 2015, **54**, 2552-2555.
82. R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich and N. Arnheim, *Science*, 1985, **230**, 1350.
83. J. Davis, *Art Journal*, 1996, **55**, 70-74.
84. C. T. Clelland, V. Risca and C. Bancroft, *Nature*, 1999, **399**, 533-534.
85. M. Peplow, *ACS Central Science*, 2016, **2**, 874-877.
86. S. Kosuri and G. M. Church, *Nature Methods*, 2014, **11**, 499-507.
87. S. L. Shipman, J. Nivala, J. D. Macklis and G. M. Church, *Nature*, 2017, **547**, 345-349.
88. C. N. Takahashi, B. H. Nguyen, K. Strauss and L. Ceze, *Scientific Reports*, 2019, **9**, 4998.
89. L. Organick, Y.-J. Chen, S. Dumas Ang, R. Lopez, X. Liu, K. Strauss and L. Ceze, *Nature Communications*, 2020, **11**, 616.
90. Y. Erlich and D. Zielinski, *Science*, 2017, **355**, 950-954.
91. H. Staudinger, *Berichte der Deutschen Chemischen Gesellschaft* 1920, **53B**, 1073-1085.
92. PlasticsEurope, *Plastics–The Facts 2018: An Analysis of European Plastics Production, Demand and Waste Data*, 2018.
93. J.-F. Lutz, *Macromolecular Rapid Communications*, 2017, **38**, 1700582.
94. P. J. Flory, *Chemical Reviews*, 1946, **39**.
95. I. group, *PET Bottle Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2019-2024*, 2019.
96. S. C. Rasmussen, *Ambix*, 2018, **65**, 356-372.
97. E. Simon, *Annalen der Pharmacie*, 1839, **31**, 265–277.
98. M. Bonastre, *Journal de pharmacie et des sciences accessoires*, 1831, **17**, 338–350.
99. P. J. Flory, *Principles of polymer chemistry*, Cornell University Press, 1953.
100. M. Szwarc, *Nature*, 1956, **178**, 1168-1169.
101. J. Nicolas, Y. Guillaneuf, C. Lefay, D. Bertin, D. Gigmes and B. Charleux, *Progress in Polymer Science*, 2013, **38**, 63-235.
102. J.-S. Wang, D. Greszta and K. Matyjaszewski, *Polymeric Materials Science and Engineering*, 1995, **73**, 416-417.
103. M. Sawamoto, M. Kato, M. Kamigaito and T. Higashimura, *Polymer Preprints (American Chemical Society, Division of Polymer Chemistry)*, 1995, **36**, 539-540.
104. X. Pan, M. Fantin, F. Yuan and K. Matyjaszewski, *Chemical Society Reviews*, 2018, **47**, 5457-5490.
105. J. Chiefari, R. T. A. Mayadunne, C. L. Moad, G. Moad, E. Rizzardo, A. Postma and S. H. Thang, *Macromolecules*, 2003, **36**, 2273-2283.
106. D. J. Keddie, G. Moad, E. Rizzardo and S. H. Thang, *Macromolecules*, 2012, **45**, 5321-5342.
107. Y. K. Chong, J. Krstina, T. P. T. Le, G. Moad, A. Postma, E. Rizzardo and S. H. Thang, *Macromolecules*, 2003, **36**, 2256-2272.
108. J. Chiefari, Y. K. Chong, F. Ercole, J. Krstina, J. Jeffery, T. P. T. Le, R. T. A. Mayadunne, G. F. Meijs, C. L. Moad, G. Moad, E. Rizzardo and S. H. Thang, *Macromolecules*, 1998, **31**, 5559-5562.
109. P. Delduc, C. Tailhan and S. Z. Zard, *Journal of the Chemical Society, Chemical Communications*, 1988, DOI: 10.1039/C39880000308, 308-310.

110. J. F. Lutz, M. Ouchi, D. R. Liu and M. Sawamoto, *Science*, 2013, **341**, 1238149.
111. N. Badi and J.-F. Lutz, *Chemical Society Reviews*, 2009, **38**, 3383-3390.
112. S. Binauld, D. Damiron, L. A. Connal, C. J. Hawker and E. Drockenmuller, *Macromolecular Rapid Communications*, 2011, **32**.
113. J. C. Barnes, D. J. C. Ehrlich, A. X. Gao, F. A. Leibfarth, Y. Jiang, E. Zhou, T. F. Jamison and J. A. Johnson, *Nature chemistry*, 2015, **7**, 810.
114. D. A. Tomalia, H. Baker, J. Dewald, M. Hall, G. Kallos, S. Martin, J. Roeck, J. Ryder and P. Smith, *Polymer Journal*, 1985, **17**, 117-132.
115. A. W. Bosman, H. M. Janssen and E. W. Meijer, *Chemical Reviews*, 1999, **99**, 1665-1688.
116. K. Sadler and J. P. Tam, *Reviews in Molecular Biotechnology*, 2002, **90**, 195-229.
117. J.-A. Amalian, T. T. Trinh, J.-F. Lutz and L. Charles, *Analytical Chemistry*, 2016, **88**, 3715-3722.
118. C. Laure, D. Karamessini, O. Milenkovic, L. Charles and J.-F. Lutz, *Angewandte Chemie International Edition*, 2016, **55**, 10722-10725.
119. S. Oelmann, S. C. Solleder and M. A. R. Meier, *Polymer Chemistry*, 2016, **7**, 1857-1860.
120. A. C. Boukis and M. A. R. Meier, *European Polymer Journal*, 2018, **104**, 32-38.
121. S. Martens, A. Landuyt, P. Espeel, B. Devreese, P. Dawyndt and F. Du Prez, *Nature Communications*, 2018, **9**, 4451-4451.
122. Z. Huang, Q. Shi, J. Guo, F. Meng, Y. Zhang, Y. Lu, Z. Qian, X. Li, N. Zhou, Z. Zhang and X. Zhu, *Nature Communications*, 2019, **10**, 1918.
123. L. Charles, G. Cavallo, V. Monnier, L. Oswald, R. Szweda and J.-F. Lutz, *Journal of The American Society for Mass Spectrometry*, 2017, **28**, 1149-1159.
124. R. K. Roy, A. Meszynska, C. Laure, L. Charles, C. Verchin and J.-F. Lutz, *Nature Communications*, 2015, **6**, 7237.
125. J. A. Amalian, A. Al Ouahabi, G. Cavallo, N. F. König, S. Poyer, J.-F. Lutz and L. Charles, *Journal of Mass Spectrometry*, 2017, **52**, 788-798.
126. G. Cavallo, S. Poyer, J.-A. Amalian, F. Dufour, A. Burel, C. Carapito, L. Charles and J.-F. Lutz, *Angewandte Chemie International Edition*, 2018, **57**, 6266-6269.
127. G. Cavallo, A. Al Ouahabi, L. Oswald, L. Charles and J.-F. Lutz, *Journal of the American Chemical Society*, 2016, **138**, 9417-9420.
128. A. Burel, C. Carapito, J.-F. Lutz and L. Charles, *Macromolecules*, 2017, **50**, 8290-8296.
129. J.-A. Amalian, G. Cavallo, A. Al Ouahabi, J.-F. Lutz and L. Charles, *Analytical Chemistry*, 2019, **91**, 7266-7272.
130. S. L. David Randall, *The Polyurethanes Book*, Wiley, 2002.
131. P. Wender, J. Rothbard, T. Jessop, E. Kreider and B. Wylie, *Journal of the American Chemical Society*, 2002, **124**, 13382-13383.
132. D. Karamessini, B. E. Petit, M. Bouquey, L. Charles and J.-F. Lutz, *Advanced Functional Materials*, 2017, **27**, 1604595.
133. D. Karamessini, T. Simon-Yarza, S. Poyer, E. Konishcheva, L. Charles, D. Letourneur and J.-F. Lutz, *Angewandte Chemie International Edition*, 2018, **57**, 10574-10578.
134. B. É. Petit, B. Lotz and J.-F. Lutz, *ACS Macro Letters*, 2019, **8**, 779-782.
135. T. Mondal, V. Greff, B. É. Petit, L. Charles and J.-F. Lutz, *ACS Macro Letters*, 2019, **8**, 1002-1005.

136. L. Charles, T. Mondal, V. Greff, M. Razzini, V. Monnier, A. Burel, C. Carapito and J.-F. Lutz, *Rapid Communications in Mass Spectrometry*, 2020, **34**, e8815.
137. S. C. Solleder, D. Zengel, K. S. Wetzel and M. A. R. Meier, *Angewandte Chemie International Edition*, 2016, **55**, 1204-1207.
138. A. Llevot, A. C. Boukis, S. Oelmann, K. Wetzel and M. A. R. Meier, *Topics in Current Chemistry*, 2017, **375**, 66-66.
139. X. Guo, K. S. Wetzel, S. C. Solleder, S. Spann, M. A. R. Meier, M. Wilhelm, B. Luy and G. Guthausen, *Macromolecular Chemistry and Physics*, 2019, **220**, 1900155.
140. C. E. Arcadia, E. Kennedy, J. Geiser, A. Dombroski, K. Oakley, S. L. Chen, L. Sprague, M. Ozmen, J. Sello, P. M. Weber, S. Reda, C. Rose, E. Kim, B. M. Rubenstein and J. K. Rosenstein, *Nature Communications*, 2020, **11**, 691.
141. S. Martens, J. Van den Begin, A. Madder, F. E. Du Prez and P. Espeel, *Journal of the American Chemical Society*, 2016, **138**, 14182-14185.
142. P. Espeel, L. L. G. Carrette, K. Bury, S. Capenberghs, J. C. Martins, F. E. Du Prez and A. Madder, *Angewandte Chemie International Edition*, 2013, **52**, 13261-13264.
143. J. O. Holloway, S. Aksakal, F. E. Du Prez and C. R. Becer, *Macromolecular Rapid Communications*, 2017, **38**, 1700500-1700500.
144. J. O. Holloway, C. Mertens, F. E. Du Prez and N. Badi, *Macromolecular Rapid Communications*, 2019, **40**, e1800685.
145. Z. Huang, J. Zhao, Z. Wang, F. Meng, K. Ding, X. Pan, N. Zhou, X. Li, Z. Zhang and X. Zhu, *Angewandte Chemie International Edition*, 2017, **56**, 13612-13617.
146. K. Ding, Y. Zhang, Z. Huang, B. Liu, Q. Shi, L. Hu, N. Zhou, Z. Zhang and X. Zhu, *European Polymer Journal*, 2019, **119**, 421-425.
147. J. M. Lee, M. B. Koo, S. W. Lee, H. Lee, J. Kwon, Y. H. Shim, S. Y. Kim and K. T. Kim, *Nature Communications*, 2020, **11**, 56.
148. R. J. Simon, R. S. Kania, R. N. Zuckermann, V. D. Huebner, D. A. Jewell, S. Banville, S. Ng, L. Wang, S. Rosenberg and C. K. Marlowe, *Proceedings of the National Academy of Sciences*, 1992, **89**, 9367.
149. T. Szekely, C. Caumes, O. Roy, S. Faure and C. Taillefumier, *Comptes Rendus Chimie*, 2013, **16**, 318-330.
150. S. Wang, Y. Tao, J. Wang, Y. Tao and X. Wang, *Chemical Science*, 2019, **10**, 1531-1538.
151. A. M. Rosales, H. K. Murnen, R. N. Zuckermann and R. A. Segalman, *Macromolecules*, 2010, **43**, 5627-5636.
152. J. Sun, X. Jiang, R. Lund, K. H. Downing, N. P. Balsara and R. N. Zuckermann, *Proceedings of the National Academy of Sciences*, 2016, **113**, 3954-3959.
153. S. C. Leguizamon and T. F. Scott, *Nature Communications*, 2020, **11**, 784.
154. N. F. König, A. Al Ouahabi, S. Poyer, L. Charles and J.-F. Lutz, *Angewandte Chemie International Edition*, 2017, **56**, 7297-7301.
155. S. Gorn, R. W. Bemer and J. Green, *Communucations of the ACM*, 1963, **6**, 422–426.
156. N. F. König, S. Telitel, S. Poyer, L. Charles and J. F. Lutz, *Macromolecular Rapid Communications*, 2017, **38**.
157. N. F. König, A. Al Ouahabi, L. Oswald, R. Szweda, L. Charles and J.-F. Lutz, *Nature Communications*, 2019, **10**, 3774.
158. L. Charles, C. Laure, J.-F. Lutz and R. K. Roy, *Macromolecules*, 2015, **48**, 4319-4328.

159. L. Charles, C. Laure, J.-F. Lutz and R. K. Roy, *Rapid Communications in Mass Spectrometry*, 2016, **30**, 22-28.
160. J. J. Kasianowicz, J. W. F. Robertson, E. R. Chan, J. E. Reiner and V. M. Stanford, *Annual Review of Analytical Chemistry*, 2008, **1**, 737-766.
161. R. Szweda, M. Tschopp, O. Felix, G. Decher and J.-F. Lutz, *Angewandte Chemie International Edition*, 2018, **130**, 16043-16047.
162. T. G. W. Edwardson, K. M. M. Carneiro, C. J. Serpell and H. F. Sleiman, *Angewandte Chemie International Edition*, 2014, **53**, 4567-4571.
163. P. Chidchob, T. G. W. Edwardson, C. J. Serpell and H. F. Sleiman, *Journal of the American Chemical Society*, 2016, **138**, 4416-4425.
164. D. de Rochambeau, Y. Sun, M. Barlog, H. S. Bazzi and H. F. Sleiman, *The Journal of Organic Chemistry*, 2018, **83**, 9774-9786.
165. T. Mondal, M. Nerantzaki, Y. Cong, S. S. Sheiko and J. F. Lutz, unpublished work.
166. Genome Reference Consortium, https://www.ncbi.nlm.nih.gov/grc/human/data, (accessed 08 28 2020).
167. M. Vybornyi, Y. Vyborna and R. Häner, *Chemical Society Reviews*, 2019, **48**, 4347-4360.
168. H. Mutlu and J.-F. Lutz, *Angewandte Chemie International Edition*, 2014, **53**, 13010-13019.
169. C. Wesdemiotis, *Angewandte Chemie International Edition*, 2017, **56**, 1452-1464.
170. M. H. Caruthers, *Biochemical Society Transaction*, 2011, **39**, 575-580.
171. C. Mayer, G. R. McInroy, P. Murat, P. Van Delft and S. Balasubramanian, *Angewandte Chemie International Edition*, 2016, **55**, 11144-11148.
172. E. Laurent, J.-A. Amalian, M. Parmentier, L. Oswald, A. Al Ouahabi, F. Dufour, K. Launay, J.-L. Clément, D. Gigmes, M.-A. Delsuc, L. Charles and J.-F. Lutz, *Macromolecules*, 2020, DOI: 10.1021/acs.macromol.0c00666.
173. *Expedite 8900 - Nucleic Acid Synthesis System - User's Guide*, Applied Biosystems, Inc, 2001.
174. A. Andrus and R. G. Kuimelis, *Current Protocols in Nucleic Acid Chemistry*, 2000, **1**, 10.15.11-10.15.13.
175. K. Launay, J.-A. Amalian, E. Laurent, L. Oswald, A. A. Ouahabi, A. Burel, F. Dufour, C. Carapito, J.-L. Clément, J.-F. Lutz, L. Charles and D. Gigmes, unpublished work.
176. D. A. Huffman, *Proceedings of the IRE*, 1952, **40**, 1098-1101.
177. N. Abramson, *Information theory and coding*, McGraw-Hill, 1963.
178. J. J. Rissanen, *IBM Journal of Research and Development*, 1976, **20**, 198-203.
179. R. Pasco, *IEEE Transactions on Information Theory*, 1977, **23**, 548-548.

Liste des publications

High-capacity Digital Polymers: Storing Images in Single Molecules
E. Laurent, J.-A. Amalian, M. Parmentier, L. Oswald, A. Al Ouahabi, F. Dufour, K. Launay, J.-L. Clément, D. Gigmes, M.-A. Delsuc, L. Charles and J.-F. Lutz, 53, 10, 4022−4029, *Macromolecules*, 2020, DOI: 10.1021/acs.macromol.0c00666.

Precise alkoxyamine-design enables automated MS/MS sequencing of digital poly(phosphodiester)s
K. Launay, J.-A. Amalian, E. Laurent, L. Oswald, A. A. Ouahabi, A. Burel, F. Dufour, C. Carapito, J.-L. Clément, J.-F. Lutz, L. Charles and D. Gigmes, Travaux non-publiés, *Angewandte Chemie International Edition, 2020.*

Synthetic Polymers with finely-regulated Monomer Sequences: Properties and Emerging Applications
R. Szweda, E. Laurent, and J.-F. Lutz*, Travaux non-publié, Chapitre d'un livre dans "Macromolecular Engineering: From Precise Synthesis to Macroscopic Materials and Applications, Second Edition", 2020.

Liste des présentations

47ème Journée des Etudes de Polymères (JEPO 2019)
Optimal sequence-coded polyphosphodiesters
Eline Laurent, Abdelaziz Al Ouahabi, Kevin Launay, Jean-Louis-Clément, Didier Gigmes, Jean-Arthur Amalian, Laurence Charles and Jean-François Lutz
29 sept. – 4 oct. 2019, Aérocampus Aquitaine, Latresne, France (présentation orale)

European Polymer Congress (EPF 2019)
Design Of Optimal Sequence-Coded Polyphosphodiesters For Digital Applications
Eline Laurent, Abdelaziz Al Ouahabi, Kevin Launay, Jean-Louis-Clément, Didier Gigmes, Jean-Arthur Amalian, Laurence Charles and Jean-François Lutz
9 - 14 Juin 2019, Hersonissos Heraklion, Crète, Grèce (poster)

Journée des Doctorants de l'Ecole Doctorale des Sciences Chimiques
Design of optimal sequence-coded polyphosphodiesters for digital applications
E. Laurent, A. Al Ouahabi, K. Launay, J.-L. Clément, D. Gigmes, J.-A. Amalian, M.-A. Delsuc, L. Charles, J.-F. Lutz
23 Novembre 2018, Strasbourg, France (présentation orale)

Min-Symposium on Complex Molecular Systems towards Adaptive Materials
Design of optimal sequence-coded polyphosphodiesters for digital applications
E. Laurent, A. Al Ouahabi, J.-A. Amalian, L. Charles, J.-F. Lutz
21 Juin 2018, Strasbourg, France (poster)

Journée des Thésards de l'ICS

Synthèse de polymère codés

E. Laurent, A. Al Ouahabi, J.-A. Amalian, L. Charles, J.-F. Lutz

11 Juin 2018, Strasbourg, France (présentation orale)

Résumé

Cette thèse porte sur la synthèse de poly(phosphodiester)s numériques. Ces polymères non-naturels sont créés à l'aide d'un alphabet de monomères permettant d'écrire un message binaire à l'échelle moléculaire. Ces polymères sont lus par une analyse de spectrométrie de masse de type peuso-MS^3. Une optimisation du design de ces séquences a été effectuée. Des polymères à haute capacité de stockage ont ainsi été synthétisés.

Tout d'abord, des alphabets augmentés compatibles avec la chimie de la phosphoramidite ont été créés. Ils contiennent 4 et 8 symboles et permettent respectivement de coder 2 et 3 bits/monomère, ce qui augmente considérablement la densité de stockage d'un monomère par rapport aux systèmes décrits au préalable. De plus, la lecture des données encodées a été facilitée grâce à l'utilisation d'un espaceur optimisé dont la structure a été élaborée avec nos collaborateurs. Le design comprenant un motif benzyle en chaîne principale a été validé par à la synthèse de séquences modèles et de séquence plus complexes codant de l'information. Finalement, l'utilisation simultanée des alphabets augmentés et de l'espaceur optimisé a permis la synthèse de longues séquences numériques. Pour augmenter encore plus la capacité de stockage de ces séquences, des outils informatiques ont été utilisés pour compresser le message. Pour permettre la lecture de cette longue séquence, l'utilisation de dix marqueurs moléculaires a été nécessaire. Ils ont pour but d'aider lors de l'analyse par spectrométrie de masse. Une étude a permis de trouver lesquels étaient optimaux pour la synthèse de ces longs polymères numériques. Grâce à l'utilisation du design optimisé développé durant cette thèse, il a été possible de synthétiser et de décrypter un poly(phosphodiester) numérique contenant 441 bits.

Mots clés : Polymères à séquence contrôlées, Macromolécules contenant de l'information, Chimie de la phosphoramidite automatisée, Haute capacité de stockage.

Résumé en anglais

This thesis focused on the synthesis of digital poly(phosphodiester)s. These non-natural polymers are built thanks to a monomeric alphabet which enables to write a binary message at the molecular level. These polymers are read by a pseudo-MS^3 mass spectrometry analysis. An optimisation of the sequence's design has been performed. High storage capacity polymers have thus been synthesized.

First, extended alphabets which are compatible with the phosphoramidite chemistry were developed. They are composed of 4 or 8 symbols and they code respectively 2 and 3 bits/monomer. These new alphabets extended drastically the density of storage of a unique monomer in comparison with the previously described systems. Secondly, the reading of the encoding data was facilitated thanks to the use of an optimized spacer. Its structure was designed with our co-workers and involve a benzyl moiety in the main chain. It was validated by the synthesis of model sequences and more complex sequences which stored information. Finally, the simultaneous use of extended alphabets and the optimized spacer lead to the synthesis of long digital sequences. To further increase the storage capacity of these sequences, computational tools were used to compress the message. To achieve the sequence's reading, ten molecular tags were required. They are used to help during the mass spectrometry analysis. A study enabled to find which tags were ideal for the synthesis of long digital polymers. The optimized design developed during this Ph.D., enabled the synthesis and the deciphering of a digital poly(phosphodiester) containing 441 bits.

Key words: Sequence-controlled polymers, Information-containing macromolecules, Automated phosphoramidite chemistry, High capacity of storage.