Institut de Génétique et de Biologie Moléculaire et Cellulaire

Département de Biologie et de Génomique Structurale

Laboratoire de Bioinformatique et de Génomique Intégratives

Genomic Clinical Synergy

Thèse présentée pour obtenir le grade de Docteur de l'Université de Strasbourg

Discipline : Science de la vie
Spécialité : Bioinformatique
par David KIEFFER

Études bioinformatiques et statistiques des mécanismes de l'infidélité de la transcription.

Soutenue publiquement le 28 octobre 2009

Membres du jury

Directeur :	POCH Olivier, Directeur de Recherche CNRS
Rapporteur interne :	KIEFFER Bruno, Professeur à l'Université de Strasbourg
Rapporteur externe :	MURUA Alejandro, Professeur agrégé à l'Université de Montréal
Rapporteur externe :	FRIEDERICH Evelyne, Professeur à l'Université du Luxembourg
Membre invité :	BIHAIN Bernard, Directeur de Recherche, Genclis.
Membre invité :	WICKER Nicolas, Maître de conférences à l'Université de Strasbourg

Remerciements

Tout d'abord je tiens à remercier Bruno Kieffer, Alejandro Murua et Evelyne Friederich pour avoir accepté de juger ce travail de thèse.

Je remercie Bernard Bihain pour m'avoir permis de réaliser ce travail de thèse avec l'équipe de bioinformatique de Genclis, ainsi que Dino Moras pour m'avoir accepté au sein du département de Biologie et Génomique Structurale.

Je ne pourrais aller plus loin sans exprimer mon infinie gratitude à Olivier Poch mon cher directeur de thèse. Je ne suis pas prêt d'oublier ces nuits blanches par connexion interposée. Je n'oublierai pas la chance que j'ai eu de connaître le seul homme capable de dire : « J'étais tranquillement en train de rédiger le rapport vers 2h du matin, j'ai cligné des yeux et je ne sais pas ce qui c'est passé, il était 6h ! Je m'excuse.». Tu es un exemple et le moteur du laboratoire car on ne peut décemment se contenter du minimum à tes côtés.
Je n'oublie pas que je dois ces trois années de folles découvertes à Nicolas Wicker. Sacré Nicolas ! Tu as réussi à t'en sortir pour ton premier encadrement de thésard. Tu sais ce qu'on dit, c'est toujours la première fois…. Bref, nous avons appris ensemble pendant ces trois années et je te remercie pour ton soutien moral et ta bonne humeur que tu as partagée pendant les meilleurs et les pires moments.
Un merci tout particulier à Laurent Bianchetti avec qui j'ai partagé de folles aventures au pays du SAGE. Merci aussi pour ton aide en anglais. J'ai beaucoup appris en termes de gestion du travail (et du stress) grâce à toi.

Je n'aurais pu réaliser ce travail sans mes chers collègues de Nancy. Un grand merci aux filles qui ont partagée leur bureau avec moi, ainsi que leurs passionnantes discussions de filles. J'ai beaucoup appris sur la vie à vos cotés. Merci à toute l'équipe de bioinfo de Genclis. Merci à toi Stéphanie pour tes mots de soutien qui m'ont beaucoup aidé pour ces derniers mois. Je te souhaite bon courage pour la suite de ta vie professionnelle. Merci Pascal pour ta disponibilité. J'espère que tu finiras par obtenir du câble électrique et réseaux pour pouvoir installer les PC sans que l'on risque de les débrancher à chaque fois que l'on se lève. Tes talents de Mc Gyver te seront, je pense, utiles dans ton rôle de papa. Valentin, merci pour ton aide sur les stats, tous mes vœux de bonheur dans ta nouvelle vie d'homme marié. Olivier

C., sans toi l'ambiance n'aurait certainement pas été la même. Ne change pas. Tu es comme un souffle d'air comique dans un monde qui a perdu son humour. L'équipe de bioinfo Genclis c'est avant tout Marie. Ta voix langoureuse va manquer aux Strasbourgeois. Tu pourras leur faire plaisir en les rappelant de temps en temps. Bon courage pour ton boulot de maman Je n'oublie pas aussi tous ceux qui sont passés dans cette équipe, Walter le dieu du Perl, Dahlia l'informaticienne de caractère avec qui le courant est tout de suite passé, Manu qui m'a encadré pendant mes premiers mois de thèse, Nicolas le statisticien toujours souriant, et Philippe bien sûr ! Toujours prêt à écouter les autres, sans juger. Le partenaire clope idéal. Bonne chance dans ton nouveau job.
Merci à tous ceux qui ont rendu plus qu'agréables mes passages à Nancy, et avec qui je me suis bien marré pendant les repas.

Cette thèse a été l'occasion de rencontrer des personnes qui m'ont profondément marqué. Tout d'abord mes chers amis qui me répondaient après « salut la plate-forme ! » le traditionnel « salut l'labo ! ». Laurent bien sûr, toujours pro. Sophie à javatiser son linux par uimaJEE et netbeanisation (oui ce n'est pas simple à comprendre). Tu es maintenant une pointure dans ton domaine. Mais ce ne sont pas tes chevilles qui enflent… Merci Véronique pour l'optimisme que tu m'as fait partager. Merci Stéphanie pour ton sourire matinal, ça fait toujours chaud au cœur. Merci Tao pour ta gentillesse et tes news de la communauté chinoise. Et bien sûr Fred. Historien, photographe, illustrateur, professeur de taïchi, maître en épée chinoise, gourou du thé d'avant 9 heures et aussi expert en bioinformatique. Qui a dit que l'on ne pouvait pas avoir une vie après le labo ? J'ai bien aimé croiser le fer avec toi et j'espère en avoir à nouveau l'occasion.

Ce laboratoire ne s'en sortirait jamais sans nos secrétaires de chocs Laetitia G. et Anne. Merci Laetitia de m'avoir guidé de tes mains expertes dans les méandres de l'administration.

Et puis il y a le LBGI. Sacrée bande de copains toujours prêts à dire une bêtise entre deux portes ! Luc, combien de fois m'as-tu surpris par une vanne quand je m'y attendais le moins! Dao merci pour le site ! J'y ai laissé ma marque ! Ne t'inquiète pas pour ta thèse tu es dans un bon labo! Laetitia P, merci pour tes visites quotidiennes au bureau et tes mots d'encouragements. Un grand merci à toi Raymond pour ta disponibilité. Toujours là quand on a besoin d'un coup de main. Merci à toi Odile pour m'avoir fait découvrir les joies de la génomique. Merci Julie pour ton petit bonjour tous les matins et pour toujours avoir répondu présent quand j'avais besoin d'aide. Parmi les personnes du labo avec qui j'ai le plus

travaillé, il y a bien sûr Wolfgang. Grand Manitou de R et des réunions de labo. Ça a été un plaisir de travailler avec toi. Nicodème, merci pour ton rire si communicatif. Ca met une sacrée ambiance. Un merci particulier à Yann pour ta précieuse aide pour corriger mes fautes d'orthographes. Et alors Jean c'est maintenant que tu reviens ? Tu pars avant le début de ma thèse et tu es de retour quand je la fini ! Sans toi et tes ARP, je ne me serais jamais fais la main sur les alignements. Merci, car ça m'a bien aidé par la suite. Hoan, toujours en train de jouer avec ton petit Bird. Merci de m'avoir laissé y toucher un peu. Et bien sûr, il y a tous les thésards avec qui j'ai commencé cette thèse. Florence, qui a subi cette rédaction de thèse en même temps que moi. Ça fait du bien quand ça s'arrête. Enfin tu vas l'avoir cette thèse ! Radouene et Yannick les deux compères. Toujours à la recherche du meilleur *design* pour un futur site. Je n'aurais jamais assez de mots pour dire à quel point Nicolas G et Laurent Philippe ont contribué au maintien de ma santé mentale durant ces trois années. A coup de X-wars, de Magic, et de trop rare sorties, merci d'avoir réussi à me traîner jusqu'au bout. J'espère que les nouveaux thésards auront la même ambiance que nous. Alors va falloir bosser Benjamin. T'es toujours sûr de faire une thèse ?

Et puis, il y a ceux qui ne sont plus dans le labo. Anne F, qui maintenant est sur l'espla… Merci d'avoir toujours été dispo dans les couloirs et au coin café. Guillaume, le maitre javaScript ! Quel choc quand tu nous as quitté ! Merci pour la décoration du bureau. Merci Manu pour tes conseils sur le roller. Mais je n'ai pas eu une fois l'occasion d'essayer pendant ces trois ans !
Et puis il y a les Post-Doc qui ont ramené un peu de leur pays et de leur soleil chez nous, Francisco, Gioia, Valentin. Un merci à Emeline pour ces discussions matinales et pour avoir partagé son amour du ski avec moi et qui maintenant est à 100% chez les structuralistes.

Je n'oublie pas les fondateurs de notre bureau. Ben et Christophe. Heureusement que vous étiez là pour mettre de l'ambiance quand j'étais encore dans ce qui servait de bibliothèque aux structuralistes et que l'on entendait les mouches voler.
Et puis il y a les potes de fac, Ivo, Alex, Fred W, Nath, Manu, Claude et Agathe. Merci d'avoir gardé le contact. J'espère que l'on continuera à se voir ou à s'écrire à défaut.
Le labo c'est aussi d'innombrables stagiaires. Merci à tous pour ce que vous avez apporté, et surtout Fabrice pour X wars, Némo pour son bon esprit, Enzo pour m'avoir fait reprendre goût au magic et bien sûr Sophie, que j'ai torturée pour lui faire tester diverses façons d'afficher un alignement multiple et sans qui beaucoup de choses n'auraient pas avancées.

Un grand merci à Myriam et Fred Bertrand qui m'ont aidé à y voir plus clair dans les statistiques. Bonne chance à vous. Je vous souhaite tout le bonheur possible.

Durant cette thèse, il n'y a pas que le boulot, mais aussi ceux qui m'ont soutenu dans la vie de tous les jours. Je n'aurai jamais réussi sans toi Christelle. Merci pour ton soutien. On commençait à ne plus y croire, mais ça y est ! Comme dirait Olivier : « C'est fini ! ». Merci Thomas, Céline et Céline pour avoir soutenu Christelle pour qu'elle puisse me soutenir.
Merci aux parents de Christelle, toujours prêts à me changer les idées. Vous vous investissez vraiment beaucoup et je ne l'oublie pas. Merci à mes parents pour m'avoir permis d'atteindre ce niveau d'étude. Et merci à toute ma famille pour ses multiples encouragements, et pour toujours avoir cru en moi. Merci à mon petit Moka pour ses léchouilles et je finirai par une petite pensée pour Pistache.

Liste des abréviations

ADN	Acide DésoxyriboNucléique
ADNc	ADN complémentaire
API	*Application Programming Interface*
ARN	Acide RiboNucléique
ARNm	ARN messager
ARNr	ARN ribosomique
ARNt	ARN transfert
BIPS	Plate-forme de BioInformatique de Strasbourg
BIRD	*Biological Integration and Retrieval Data*
BIRD-QL	*BIRD Query language*
BLAST	*Basic Local Alignment Search Tool*
CD	*Compact Disc*
CGH	*Comparative Genomic Hybridization*
ChIP	*Chromatin ImmunoPrecipitation*
CPSF	*Cleavage/Polyadenylation Specificity Factor*
CTD	*Carboxy Terminal Domain*
DAVID	*Database for Annotation, Visualization and Integrated Discovery*
DDBJ	*DNA Data Bank of Japan*
DTD	*Document Type Definition*
DVD	*Digital Versatile Disc*
EJC	*Exon Junction Complex*
EMBL	*European Molecular Biology Laboratory*
EST	*Expressed Sequenced Tag*
FDR	*False Discovery Rate*
Genclis	*Genomic Clinical Synergy*
GEO	*Gene Expression Omnibus*
GGR	*Global Genomic Repair*
GOLD	*Genomes OnLine Database*
GPL	*Geo Platform*
GSE	*Geo SErie*
GSM	*Geo SaMple*
HapMap	*Haplotype Map*
HGP	*Human Genome Project*
HTC	*High Throughput cDNA*
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*
ICARUS	*Interpreter of Commands And RecUrsive Syntax*
IGBMC	Institut de Génétique et de Biologie Moléculaire et Cellulaire
ILP	*Inductive Logic Programming*
ISCSI	*Internet Small Computer System Interface*
IT	Infidélité de Transcription
JRI	*Java R Interface*
JMACS	*Java for Multiple Alignment of Complete Sequences*
KDD	*Knowledge Discovery in Database*
LANL	*Los Alamos National Laboratory*
LBE	*Location Based Estimator*
LBGI	Laboratoire de Bioinformatique et Génomique Intégratives
MACS	*Multiple Alignment of Complete Sequences*

MACSIMS	*Multiple Alignment of Complete Sequences Information Management System*
MAO	*Multiple Alignment Ontology*
MCM	*Mini Chromosome Maintenance complex*
NCBI	*National Center for Biotechnology Information*
NFS	*Network File System*
NGD	*No-Go mRNA Decay*
NMD	*Nonsense-Mediated mRNA Decay*
NSD	*NonStop mRNA Decay*
OBO	*Open Biomedical Ontologies*
PAP	Poly (A) Polymérase
PC	*Personal Computer*
PCNA	*Proliferating Cell Nuclear Antigen*
PCR	Réaction de Polymérisation en Chaîne
PHP	*PHP: Hypertext Preprocessor*
RAID	*Redundant Array of Independent Disks*
REB	Réparation par Excision de Base
REN	Réparation par Excision de Nucléotides
ReNaBi	Réseau National des plates-formes Bioinformatique
RFC	*Replication Factor C*
RIO	Réunion Inter-Organismes
RMA	Réparation des MésAppariements
RPA	*Replication Protein A*
RPL	*Ribosomal Protein Large subunit*
RPS	*Ribosomal Protein Small subunit*
SAGE	*Serial Analysis of Gene Expression*
SNP	*Single Nucleotide Polymorphism*
SQL	*Structured query language*
SRS	*Sequence Retrieval System*
TCP	*Transmission Control Protocol*
TCR	*Transcription Coupled Repair*
TE	Tags Expérimentaux
TFIID	*Transcription Factor II D*
TFIIH	*Transcription Factor II H*
TFIIS	*Transcription Factor II S*
TRC	Transcrit de protéine Ribosomale Cytoplasmique
TV	Tag Virtuel
TVA	Tag Virtuel Amont
UCSC	*University of California Santa Cruz*
UTR	*UnTranslated Region*
XML	*eXtensible Markup Language*

Table des matières

Table des figures

Table des tableaux

Avant-propos

J'ai effectué ma thèse dans le cadre d'une collaboration entre le LBGI (Laboratoire de Bioinformatique et Génomique Intégratives) à l'IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire) de Strasbourg et l'entreprise Genclis (Genomic Clinical Synergy) établie à Vandoeuvre les Nancy. L'équipe de Bioinformatique de Genclis a démontré, sur la base des EST (*Expressed Sequence Tag*) que les ARNm provenant de tissus cancéreux humains sont statistiquement plus hétérogènes en séquence que ceux provenant de tissus sains (Brulliard et al., 2007). Ce résultat a permis de proposer le concept d'Infidélité de la Transcription (IT), phénomène qui interviendrait dans les cellules humaines saines, mais qui serait fortement augmenté dans les cancers. C'est dans ce contexte que Genclis et l'IGBMC ont mis en place une collaboration, pour essayer de caractériser et de comprendre, par des approches bioinformatique et statistique, les aspects fondamentaux du phénomène de l'IT. Par delà la compréhension des mécanismes mis en jeu lors de l'IT, une telle connaissance a pour objectif de diagnostiquer les patients atteints en prédisant de manière de plus en plus fiable et précoce les molécules aberrantes générées par l'IT dans les cancers.

Durant ma thèse, j'ai participé à la caractérisation de l'hétérogénéité des ARNm dans les cancers humains par l'étude d'une source de données complémentaires des EST, les SAGE (*Serial Analysis of Gene Expression*). J'ai également réalisé une étude longitudinale des modifications observées sur les EST afin d'établir des règles caractérisant les mécanismes aboutissant à cette hétérogénéité. Sur le plan informatique et statistique, ceci m'a permis de mettre au point des protocoles informatiques et statistiques originaux qui pourront être réutilisés dans d'autres travaux de recherches. Sur le plan de la compréhension des mécanismes liés à l'IT, ceci m'a permis de caractériser de nouveaux gènes sensibles à l'IT et de mettre en évidence un phénomène de couplage des modifications introduites par l'IT au sein des ARNm. Ce phénomène de couplage sera discuté au regard des nombreux mécanismes connus intervenant dans le maintien de l'intégrité du message génétique au sein des cellules d'eucaryotes.

Le manuscrit s'articule autour d'une introduction qui présente succinctement les connaissances qui ont été nécessaires pour réaliser ce travail de thèse et qui recouvrent des notions d'informatique et de bioinformatique, des notions biologiques sur les mécanismes impliqués dans le maintien de l'information génétique, des notions sur la nature et l'origine

des cancers ainsi que des informations concernant les différents objets biologiques utilisés dans mon étude, à savoir : les EST et les données SAGE. Face à l'étendue des sujets abordés, j'ai été amené à ne présenter dans cette partie introductive qu'un survol incomplet des différents domaines. Enfin, au sein des descriptions des différents mécanismes de maintien de l'intégrité génétique, j'ai isolé, lorsque cela était possible, les tailles des régions nucléotidiques impliquées dans chaque mécanisme pour faciliter l'interprétation des résultats (voir plus bas). La deuxième partie concerne les matériels et méthodes utilisés qui englobent les banques de données et outils bioinformatiques utilisés durant cette thèse ainsi qu'une rapide présentation des outils et concepts statistiques nécessaires à la compréhension des résultats obtenus et des développements réalisés. La troisième partie comprend la présentation des résultats qui distingue d'une part, les résultats liés à l'analyse des données SAGE qui m'ont permis de caractériser de nouveaux gènes et d'estimer l'importance des modifications présentes dans les ARNm provenant de tissus cancéreux et d'autre part, les travaux statistiques réalisés sur la base de données d'EST qui ont permis de révéler l'existence de modifications couplées au sein des messagers. L'étude des distances entre modifications couplées au sein des transcrits et des génomes nous a permis d'émettre des hypothèses sur l'origine des modifications et sur les types de mécanismes de maintien de l'intégrité de l'information qui pourrait être perturbés dans les cancers humains.

Introduction

1 Contexte bioinformatique et statistique

Le terme de bioinformatique désigne la mise au point et l'utilisation de méthodes informatiques pour collecter, organiser et analyser les données biologiques. Il existe trois grands aspects de la bioinformatique. Tout d'abord, l'aspect le plus mathématique qui consiste à créer des algorithmes et des modèles mathématiques pour intégrer les données existantes et découvrir de nouvelles propriétés permettant de mieux comprendre leurs interactions et faire des prédictions *in silico*. Le second, plus informatique, consiste en l'élaboration de programmes pour optimiser la gestion, l'accès et les traitements des données, en implémentant les nouveaux modèles et les nouveaux algorithmes. Le dernier aspect, appelé aussi bioanalyse, est centré sur l'utilisation des outils bioinformatiques pour analyser les données biologiques. Ces trois approches sont complémentaires et le bioinformaticien qui doit être à même de saisir et manipuler des notions de mathématiques, informatique et biologie est donc, par définition, pluridisciplinaire. Dans ce contexte, la force d'un laboratoire de bioinformatique est souvent de réunir des chercheurs de ces trois disciplines pour apporter un regard complémentaire sur la biologie.

La bioinformatique doit son développement tout d'abord à la diffusion d'ordinateurs de plus en plus puissants capables de stocker et de traiter un nombre croissant de données biologiques. Ensuite, elle doit son essor à l'accumulation des connaissances sur les molécules et agents biologiques tels, l'ADN, l'ARN et les protéines. Notamment, grâce au développement du séquençage, qui permet au bioinformaticien de traiter ces molécules comme une séquence de caractères correspondant, soit à des nucléotides dans le cas des ADN et ARN, soit à des acides aminés dans le cas des protéines. Ce sont les molécules de bases de la transmission de l'information génétique dans la cellule dont le dogme central a été énoncé dans les années 50 par Crick (CRICK, 1958). La figure ci-dessous illustre comment nous considérons aujourd'hui la propagation du message génétique.

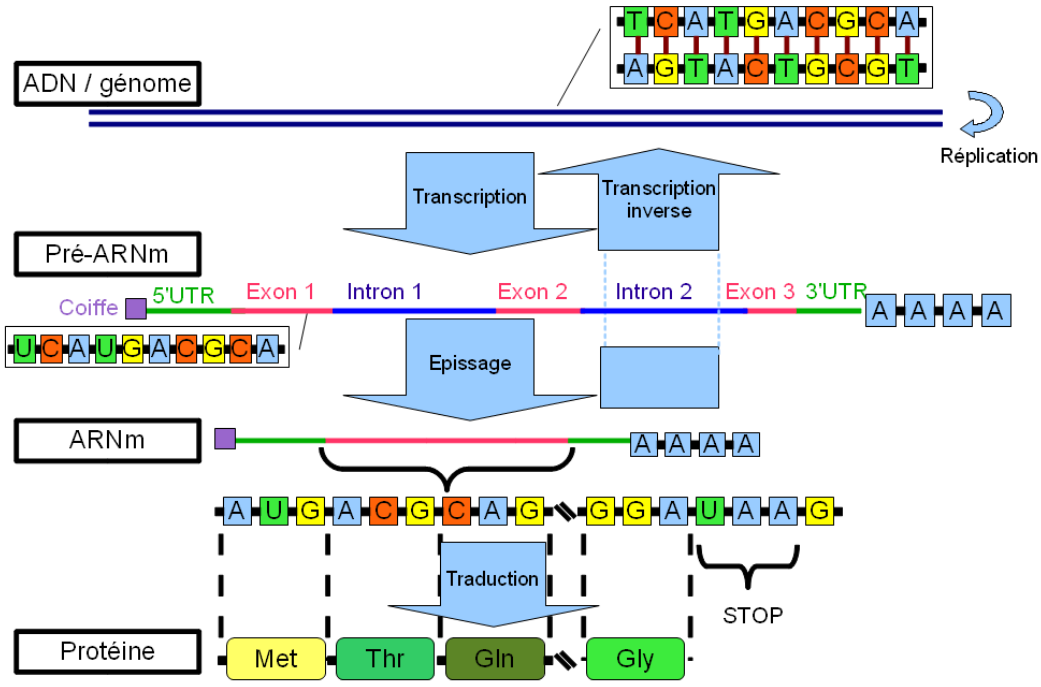


Figure 1 : Représentation du transfert de l'information génétique de l'ADN à la protéine chez les eucaryotes.
L'ADN est le support de l'information génétique, codé par les bases A, T, G et C. Il se transmet aux cellules filles grâce à sa réplication avant la division cellulaire. La transcription va générer des molécules d'ARN qui seront une copie de l'information de l'ADN. Dans le cas des ARN messagers (ARNm), il y aura une étape d'épissage qui consiste à enlever les fragments nommés introns. Ces ARNm seront ensuite décodés pendant la traduction par « codon » de trois nucléotides afin d'assembler les acides aminés qui constitueront la protéine. Il existe également un mécanisme de reverse transcription permettant de synthétiser un ADN à partir d'un ARN qui peut aboutir à la réintégration de la molécule d'ADN formée dans le génome.

La bioinformatique est une discipline dynamique qui évolue en même temps que les connaissances biologiques et les nouvelles technologies informatiques ou biotechnologiques. Le Tableau 1 présente les principaux évènements qui ont influencé la bioinformatique.

Tableau 1 : Chronologie non exhaustive des grands événements influençant la bioinformatique. En noir sont indiqués les grands évènements de biologie et de biochimie, en bleu, ceux de l'informatique et en vert, ceux de la bioinformatique. (Tableau construit à partir de http://bio.cc/Bioinformatics/history_of_bioinformatics.html et de http://villemin.gerard.free.fr/Wwwgvmm/Histoire/Informat.htm)

Date	Auteur	Evènement
1944	Avery	Démontre que l'ADN est le support de l'information génétique (Avery, MacLeod, & McCarty, 1944)
1953	Watson, Crick	Etablissement de la structure en double hélice de l'ADN (Watson & Crick, 1974)
1953	IBM 650	Premier ordinateur commercial

1953	Sanger	Détermination de la séquence des chaines A et B de l'insuline (F. Sanger & E. O. P. Thompson, 1953a, 1953b)
1956	Tijo, Levan	Démontre que le nombre de paires de chromosomes chez l'homme est 23 (Tjio JH, 1956)
1956	Anfinsen	La structure tridimensionnelle d'une protéine est fonction de sa séquence (ANFINSEN & REDFIELD, 1956)
1956	IBM	Commercialisation des premiers disques durs
1958	Crick	Enonciation du Dogme Central de la biologie moléculaire (CRICK, 1958)
1962	Nirenberg, Matthaei	Déchiffrage du code génétique (MATTHAEI, JONES, MARTIN, & NIRENBERG, 1962)
1965	Monod, Jacob, Wolf	Découverte des mécanismes de la régulation génétique impliqués dans le dogme central
1965	Dayoff	Premier atlas de séquences et structures de protéines (Dayhoff, 1965)
1967	Fitch	Construction d'arbres phylogénétiques (Fitch & Margoliash, 1967)
1969	ARPANET	Premières interconnexions universitaires
1970	Needlman, Wunsh	Algorithme d'alignement global optimal entre deux séquences de protéines (Needleman & Wunsch, 1970)
1974	Crerf, Kahn	Développement du concept d'Internet et du protocole TCP
1974	Chou	Algorithme de prédiction des structures secondaires des protéines
1977	Sanger	Méthode de séquençage de séquences nucléiques (approche enzymatique) (F Sanger, Nicklen, & A R Coulson, 1977)
1977	Maxam, Gilbert	Méthode de séquençage des séquences nucléiques (approche chimique) (Maxam & W. Gilbert, 1977)
1978	Sanger	Séquençage du génome du bactériophage phi174 (5386 pb)(F Sanger et al., 1977)
1980	EMBL	Création d'une banque européenne de séquences nucléiques
1981	Smith, Waterman	Algorithme d'alignement local optimal entre deux séquences
1981	IBM	Premier ordinateur sous le nom de *Personal Computer* (PC).
1981	Anderson	Séquençage du génome mitochondrial humain (Anderson et al., 1981)
1982	Genbank	Création de la banque américaine de séquences nucléiques
1983	Mullis	Invention de la Réaction de Polymérisation en Chaine (PCR)(Mullis, 1994)
1984	Gouy	ACNUC, logiciel d'interrogation de banques de séquences(Gouy, C. Gautier, Attimonelli, Lanave, & di Paola, 1985)
1985	Lipman, Pearson	FASTA, programme de recherche de similarité dans les banques de données(D J Lipman & Pearson, 1985)
1985	Sonny, Philips	Création d'un nouveau support numérique, le Compact Disc (CD)
1986	Swiss-Prot	Création de la banque de séquences protéiques
1986	DDBJ	Création de la banque japonaise de séquences nucléiques
1986	Roderick	Apparition du terme "*genomic*"
1987	Applied Biosystems	Commercialisation du premier séquenceur automatisé
1987	McKusic	Première carte génétique du génome humain (McKusick & Ruddle, 1987)

1987	Kulesh	Apparition de la technologie des puces à ADN (Kulesh, Clive, Zarlenga, & Greene, 1987)
1988	HUGO	Coordonne le décryptage mondial du génome humain (National Research Council (U.S.)., 1988)
1989	Internet	Internet succède à ARPANET
1990	Berners-Lee	Publication du premier document HTML
1990	HGP	Initiation du « *Human Genome Project* » (HGP), visant à décrypter l'intégralité du génome humain
1990	Altschul	BLAST, programme de recherche de séquences par similarité (Altschul, W. Gish, W. Miller, E. W. Myers, & D J Lipman, 1990)
1991	Adams	Premier séquençage à grande échelle d'ADNc (EST) (Adams et al., 1991)
1991	Roberts	GRAIL : programme de localisation de gènes (L. Roberts, 1991)
1992		Séquençage du chromosome III de *Saccharomyces cerevisiae*
1993	Cohen	Première carte physique du génome humain
1993	Boguski	dbEST: banque de données internationale d'EST
1993	Etzold	SRS: logiciel d'interrogation de banques
1995	Fleischmann	Séquençage du premier organisme vivant, *Haemophilus influenza* (R. D. Fleischmann et al., 1995)
1995	Velculescu	Apparition de la technologie du "*Serial Analysis of Gene Expression*" (SAGE) (V. E. Velculescu, L. Zhang, B. Vogelstein, & K. W. Kinzler, 1995)
1995	DVD Forum	Création d'un support numérique de stockage de haute capacité, le DVD.
1996	Walsh	Séquençage du premier génome eucaryote, *Saccharomyces cerevisiae* (Walsh & B. Barrell, 1996)
1996	Affymetrix	Commercialisation de la première puce à ADN
1998	Pinkel	Adaptation de la technologie de "*Comparative Genomic Hybridization*" (CGH) sur puce à ADN (Pinkel et al., 1998)
1998	W3C	Création du format XML
2000	Adams	Séquençage du génome de *Drosophila melanogaster* (Adams et al., 2000)
2000	Ren	Adaptation de la technique de "*Chromatin ImmunoPrecipitation*" sur puce à ADN (Chip on chip) (Ren et al., 2000)
2000	Ashburner	Création de la banque d'annotation Gene Ontology (Ashburner et al., 2000)
2001	Lander	Séquence préliminaire du génome humain par HGP (Lander et al., 2001)
2001	Venter	Séquence préliminaire du génome humain par Celera Genomics (J. C. Venter et al., 2001)
2001	Fred Hutchinson Cancer Research Center	Lancement du projet BioConductor une librairie pour le logiciel de statistique R, pour l'analyse des expériences à haut débit (Gentleman et al., 2004)
2001	Pruitt	Création de la banque RefSeq, une banque non redondante de séquences (K D Pruitt & D R Maglott, 2001)
2001	Ensembl	Genome browser Ensembl (http://www.ensembl.org/Homo_sapiens/index.html)
2001	NCBI	Genome browser du NCBI (www.ncbi.nlm.nih.gov/mapview)
2001	AMD et INTEL	Premier processeur cadencé à 1GhZ pour PC

2002	Ron Edgar	Création de la base de données *Gene Expression Omnibus* au NCBI pour les données d'expression à haut débit.
2002	Kent	BLAT pour la recherche de séquences génomiques (W. James Kent, 2002)
2002	UCSC	Publication du "*UCSC genome browser database*" permettant une vue annotée du génome sur un navigateur internet.
2005	AMD et INTEL	1ers processeurs double-cœur pour PC
2006	Sony	Création d'un support numérique de stockage à haute définition, le Blu-ray.
2006	Roche-454	Lancement d'un séquenceur à haut débit pouvant traiter 20 Mbase par run avec une lecture de 100 bases
2007	Illumina	Lancement du séquenceur SOLEXA pouvant traiter 1Gbase par run avec une lecture de 32 bases
2008	Applied BioSystem	Lancement du séquenceur SOLID à ligation de base pouvant traiter 20 Gbase par run.

Le principe de l'hérédité était déjà connu depuis le début du 20ème siècle grâce aux travaux de Gregor Mendel, le fondateur de ce qui s'appellera la génétique. Cette théorie a été d'ailleurs le point de départ des statistiques en biologie qui ont initialement servi à argumenter cette théorie, puis ont été élargies dès les années 1920 à l'étude de la génétique des populations et de leur évolution génétique (R.A. Fisher, 1930). Mais, à cette époque, les molécules responsables de l'hérédité étaient encore inconnues. Ce n'est qu'au milieu du 20ème siècle qu'il a été prouvé que l'ADN est le porteur de l'information génétique. Il est à la base de la transmission de l'information génétique aux protéines. La communauté scientifique s'est donc lancée dans le décryptage du génome dans l'espoir de comprendre comment cette molécule d'ADN est capable de diriger le développement d'un organisme vivant. Depuis la mise au point des premières techniques de séquençage de protéines (F. Sanger & E. O. P. Thompson, 1953a) et des acides nucléiques (F Sanger et al., 1977), jusqu'aux récentes techniques de séquençage à haut débit, de nombreux projets de séquençage de génomes entiers ont été lancés. Ainsi, à ce jour, 5553 projets de génomes sont répertoriés sur le site GOLD (http://www.genomesonline.org/gold.cgi) dont 1071 sont complets représentant 65 archées, 890 bactéries et 116 eucaryotes. Le séquençage devenant de plus en plus rapide et fiable, le nombre de projets de séquençages ne cesse d'augmenter (Figure 2). Aujourd'hui, il est question de métagénome, désignant l'ensemble de tous les génomes des espèces présentes dans un environnement particulier.

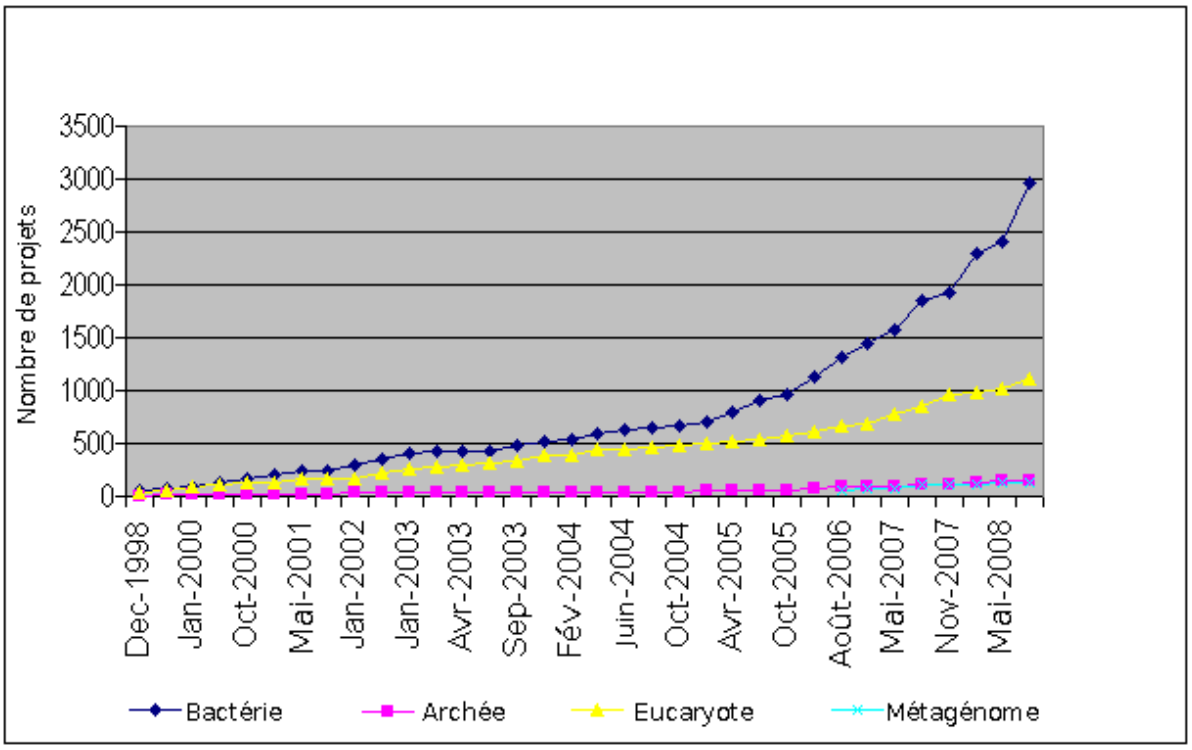


Figure 2 : Evolution du nombre de projets de séquençage de génomes disponibles sur le site GOLD en fonction du groupe phylogénétique.

Il faut tout de même préciser que le séquençage seul ne suffit pas à obtenir un génome complet. Il est nécessaire de réaliser un travail d'assemblage pour disposer correctement les séquences les unes par rapport aux autres. Les améliorations constantes des techniques d'assemblage illustrent bien le lien entre experts en bioinformatique (en l'occurrence, algorithmiciens) et augmentation de la puissance de calcul des ordinateurs.

Nous sommes actuellement dans ce que nous appelons l'ère post-génomique, où nous pouvons nous baser sur la connaissance et l'accès aux séquences de nombreux génomes. Cependant, il reste encore un travail titanesque d'annotation de ces génomes (comme l'illustre de façon humoristique la Figure 3) : «Où sont et que sont les gènes? Où sont les séquences codantes pour les protéines? ». Ces questions peuvent être abordées, grâce à la transcriptomique et à la protéomique qui étudient respectivement les ARN et les protéines produits au sein des différents niveaux d'organisation du vivant (cellule, tissu, organe, organisme ou environnement) et qui se sont développées en parallèle de la génomique. A titre d'exemple, il est à noter que des mises à jour du génome humain, obtenues en 2001, sont régulièrement publiées, la dernière en date étant la version 37.1 de février 2009. Pour mettre à disposition ces connaissances, plusieurs instituts ont développé une interface web (*Genome Browser*) présentant le génome et ces annotations, notamment, l'UCSC (University of

California, Santa Cruz, http://genome.ucsc.edu), le NCBI (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?_taxid=9606) et l'EBI (http://www.ensembl.org/Homo_sapiens/index.html).

Il faut toutefois préciser que nous ne parlons pas ici d'un génome universel, mais bien d'un génome de référence. Ainsi, chez l'homme, de nombreux projets ont porté sur l'étude des particularités génétiques qui différencient chaque individu et de leurs liens avec des particularités phénotypiques ou vis-à-vis des prédispositions à développer telle ou telle maladie. C'est dans le cadre de l'étude de ces particularités que le projet HapMap (Haplotype Map) ("The International HapMap Project," 2003) a été initié en 2002 pour obtenir la description des variations génétiques les plus fréquentes. Les variations génétiques peuvent impliquer plusieurs nucléotides à la suite dans l'ADN mais, dans environ 90% des cas, elles sont ponctuelles (Francis S. Collins, Brooks, & Chakravarti, 1998) et sont appelées SNP (Single Nucleotide Polymorphism). En fait, 93% des gènes humains contiendraient au moins un SNP (Chakravarti, 2001).

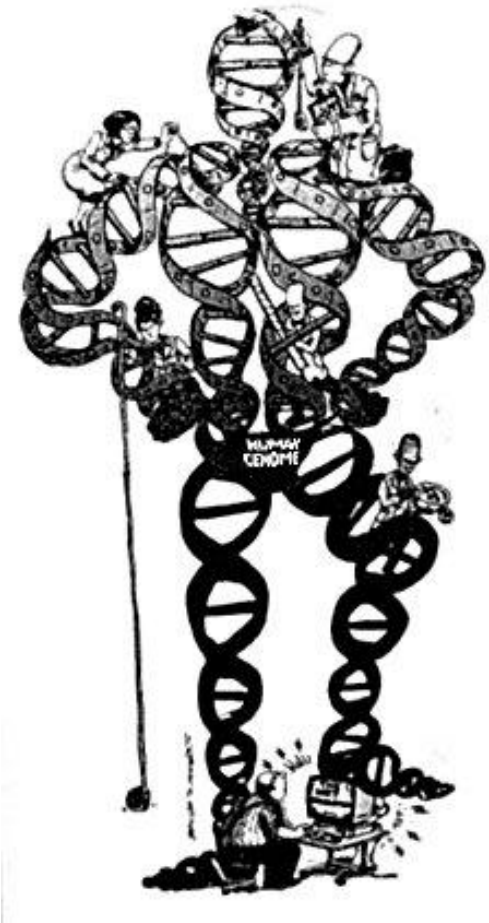


Figure 3 : Biologistes cherchant à comprendre le génome humain et ordinateur ayant du mal à gérer le « poids » des données.
Source : http://www2c.ac-lille.fr/bts-lettres/images\clonage.jpg)

Comme nous l'avons vu, la biologie utilise de plus en plus de technologies dites à haut débit permettant de tester des milliers de gènes en une seule expérimentation (voir chapitre 4). Ainsi, la nouvelle génération des séquenceurs permet d'aller encore plus loin dans l'analyse du génome et du transcriptome en fournissant en une seule expérience plusieurs centaines de millions de nucléotides (Sultan et al., 2008). Ces technologies ne sont possibles que grâce à l'informatique et à la bioinformatique qui prennent en charge la collecte, l'organisation et l'analyse des données générées. La recherche d'algorithmes et l'élaboration d'outils informatiques de plus en plus performants sont devenues un véritable enjeu pour obtenir des analyses de qualité croissante. En effet, de nombreuses méthodes visent à améliorer le rapport signal/bruit en éliminant ou, à tout le moins, en diminuant les multiples « bruits de fond » inhérents aux technologies à haut débit. L'analyse n'est donc possible que grâce aux nombreuses approches statistiques développées pour traiter des données de plus en plus complexes. Face à cette demande, en 2001, démarre le projet Bioconductor (Gentleman et al., 2004) qui est une librairie de statistiques pour le logiciel R (Team, 2008) regroupant les développements d'outils statistiques appliqués à la biologie et notamment, aux expériences à haut débit. Ainsi, aujourd'hui, une analyse efficiente du vivant implique non seulement, une grande masse de données mais aussi, des données de très grande qualité intégrant des informations sur les traitements réalisés et les taux d'erreurs probables.

La compilation de telles données venant de plusieurs domaines d'expertises comme la génétique, la transcriptomique, la protéomique, la métabolomique et bien d'autres, permettrait d'avoir une vue globale du fonctionnement de la cellule. Regrouper cette masse de connaissances biologiques est maintenant le grand défi de la bioinformatique (Snoep, Bruggeman, Olivier, & Westerhoff) et pour atteindre cet objectif, l'intégration des avancées venant des domaines de l'informatique, des mathématiques et des statistiques est devenue indispensable.

2 Biologie : Les mécanismes de conservation de l'information génétique

La conservation de l'intégrité de l'information génétique est essentielle pour la survie d'un organisme. Pour que la synthèse des protéines ne soit pas altérée au cours du temps, toute une série de mécanismes de contrôle et de complexes de réparation de l'ADN et de l'ARN sont actifs dans la cellule. Les principaux mécanismes intervenant dans le transfert et le maintien de l'information génétique (Chapitre 1.2 et 1.3) ainsi que les conséquences d'erreurs intervenant lors des ces processus (Chapitre 1.3) seront présentés succinctement. Nous décrirons tout d'abord, la structure de la molécule d'ADN dans la cellule, puis nous présenterons les grandes étapes du maintien de l'information génétique de l'ADN durant la réplication, la transcription et la traduction.

2.1 La molécule d'ADN

L'ADN est un assemblage de deux brins de nucléotides composés d'une base azotée (adénine, thymine, cytosine ou guanine), d'un sucre (désoxyribose) et d'un phosphate. Les deux brins forment une structure en double hélice et sont orientés d'une extrémité 5' phosphate vers une extrémité 3' OH. Ils sont appariés de façon antiparallèle (l'extrémité 5' de l'un en face de l'extrémité 3' de l'autre) par complémentarité de base (A avec T, G avec C). Chez les eucaryotes, ces deux brins s'enroulent autour des histones et forment un complexe histones/ADN, appelé nucléosome, qui est une structure dynamique permettant à l'ADN de passer d'un état décondensé (chromatine) à un état condensé (chromosome) (Figure 4). La dynamique de la chromatine permet non seulement la compaction/décompaction chromosomique nécessaire à la division cellulaire mais joue également un rôle dans le contrôle de l'expression des gènes en modifiant l'accessibilité de l'ADN aux facteurs de transcription.

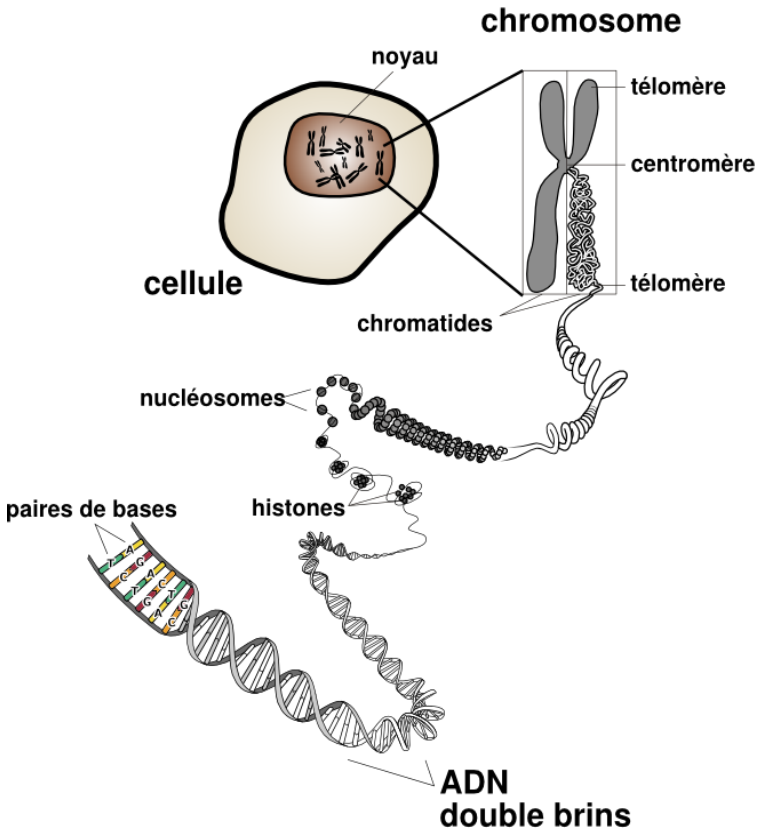


Figure 4 : Compaction de l'ADN en chromosome.
L'ADN double brins s'enroule autour des histones en nucléosomes. L'ensemble des nucléosomes formes la chromatine qui va se condenser pour former un chromosome (source : http://fr.wikipedia.org/wiki/Fichier:Chromosome_fr.svg) .

2.2 La réplication de l'ADN

La réplication de l'ADN est un processus semi-conservatif qui, à partir de chaque brin d'ADN, va synthétiser un brin par complémentarité des bases (Figure 5). Il en résultera deux nouvelles molécules d'ADN ayant chacune un brin père qui aura servi de base et un brin fils néo-synthétisé.

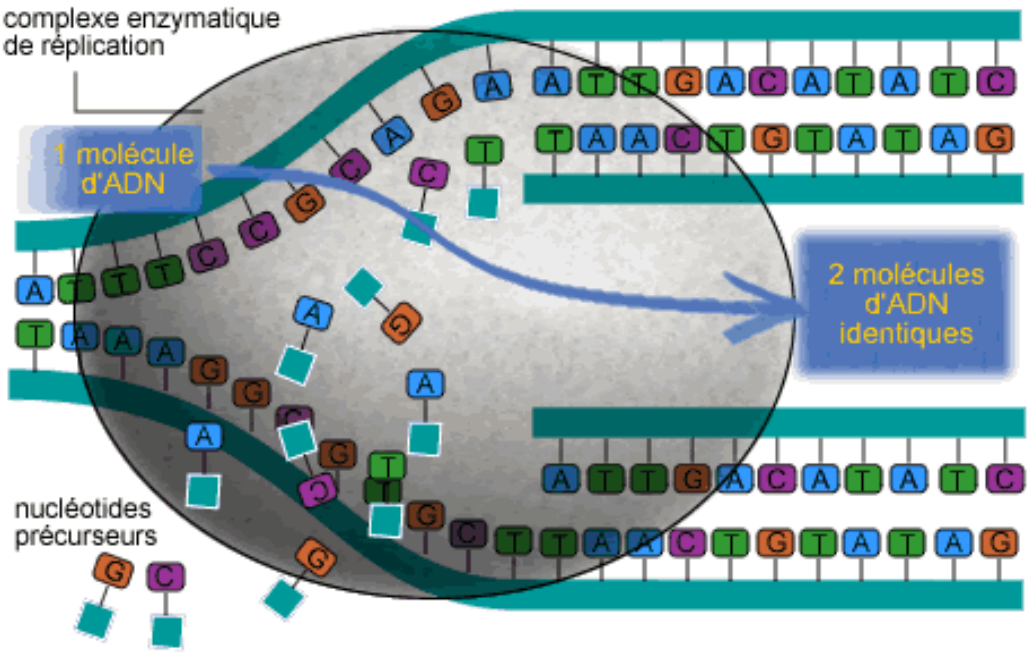


Figure 5: La réplication de l'ADN : un processus semi-conservatif.
A partir d'une molécule d'ADN parent, le complexe de réplication crée deux molécules d'ADN identiques en séquence. Chacune des molécules est constituée d'un brin « père » et d'un brin « fils » néo-synthétisé (source : http://fr.wikipedia.org/wiki/Fichier:RéplicationdelADN.png) .

La réplication débute par la formation d'un réplicon où se mettront en place deux fourches de réplication qui vont se déplacer le long du brin d'ADN en sens opposés (Figure 6). Il y a plusieurs réplicons à la fois par chromosome pour accélérer ce processus. L'origine de la formation des réplicons et leurs contrôles sont encore peu connus.

Figure 6 : La réplication bidirectionnelle.
Un réplicon et ses deux fourches de réplication qui parcourent le chromosome en sens opposé.

Cette organisation de l'ADN implique une machinerie de réplication complexe pouvant se déplacer dans une telle structure. Les Figure 5 et Figure 7 schématisent le mode de réplication de l'ADN.

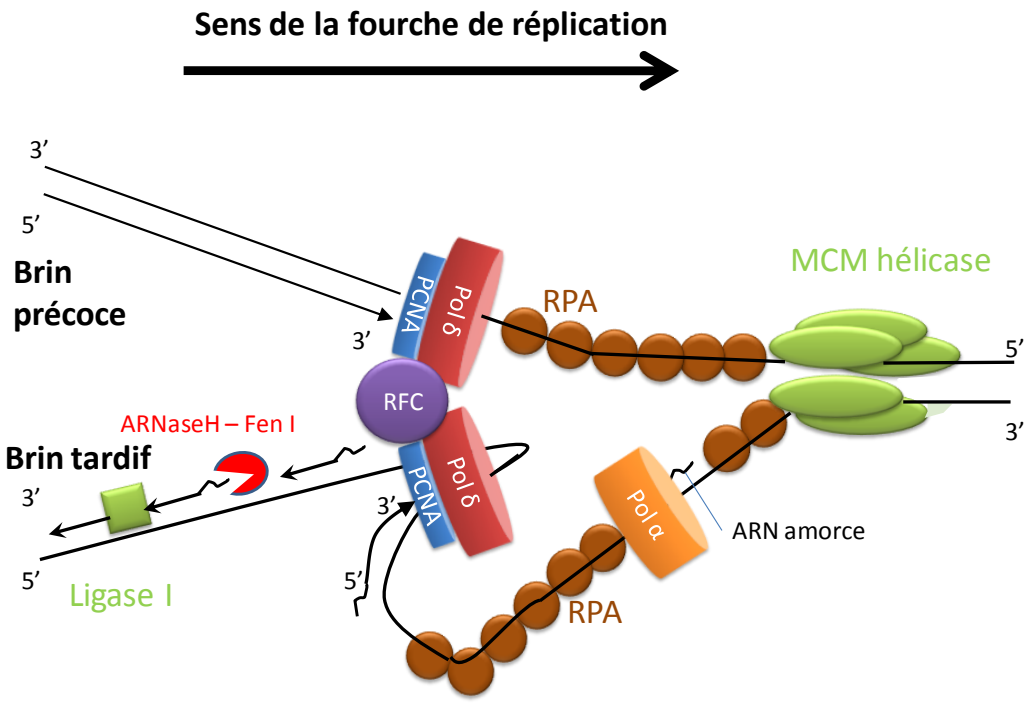


Figure 7 : Schéma de la fourche de réplication chez les eucaryotes.
La réplication est assurée par l'holoenzyme composée de deux polymérases δ, avec les PCNA associées et du complexe RFC. Une polymérase δ synthétisera le brin précoce en continu tandis que l'autre aura besoin des amorces produites par la polymérase α pour synthétiser les fragments d'Okazaki. L'ARNase H digérera les amorces, pour permettre à la ligase d'assembler le brin tardif (adapté de (Helmut Pospiech, 2002)).

Le complexe MCM (*Mini Chromosome Maintenance complex*) ouvre la double hélice d'ADN grâce à son activité hélicase. La RPA (*Replication Protein A*) va se fixer à l'ADN simple brin pour stabiliser sa structure. Plusieurs types de polymérases permettent ensuite d'ajouter des nucléotides à l'extrémité 3' du nouveau brin. Des polymérases α vont synthétiser des ARN amorces sur les deux brins. Ce type de polymérase reste peu de temps (processivité faible) et sera rapidement remplacée par la polymérase δ qui a besoin d'une amorce double brin pour se fixer. Stabilisée par la PCNA (*Proliferating Cell Nuclear Antigen*), cette dernière aura une forte processivité et pourra allonger le nouveau brin d'ADN. La polymérase δ a une activité exonucléasique 3' → 5', c'est-à-dire la capacité de corriger automatiquement une éventuelle erreur. En fait, le complexe de réplication met en jeu deux polymérases δ reliées par le complexe RFC (*Replication Factor C*), qui vont parcourir conjointement l'ADN. La synthèse d'un brin par les polymérases se réalise toujours par ajout d'un nucléotide à l'extrémité 3' du brin néo-synthétisé. En conséquence, il y aura deux systèmes d'élongation du nouveau brin. Le brin précoce est le brin où la polymérase δ

peut réaliser directement, de façon continue, la synthèse dans le même sens que le déplacement de la fourche de réplication. Sur le brin tardif, la polymérase α va régulièrement construire un ARN amorce. Le brin tardif va former une boucle et permettre à la polymérase de synthétiser un nouveau fragment d'ADN depuis les amorces pour former un fragment d'Okazaki. Il sera synthétisé sur **une centaine de nucléotides** (Weil, 2001) jusqu'à rencontrer le fragment d'Okazaki précédent, l'amorce d'ARN ayant été digérée par l'ARNase H. L'extrémité 3' du nouveau fragment sera reliée à l'extrémité 5' du fragment d'Okazaki précédent par la ligase I. D'autres protéines impliquées dans le contrôle de la stabilité de l'ADN et tout un ensemble d'ADN polymérases différentes sont également recrutées selon le contexte rencontré lors de l'avancement de la fourche de réplication. Pour contrôler la stabilité de l'ADN nous pouvons citer les topoisomérases, qui sont des protéines permettant de baisser les tensions dans la structure de l'ADN dues à son « désenroulement » par l'hélicase. Comme autre polymérase, nous pouvons citer la polymérase ε qui interviendra lors de processus de réparations ou pour terminer la réplication aux bords des chromosomes (les télomères).

Non seulement ce système de réplication est très fidèle (moins d'une erreur pour dix millions de nucléotides) mais de plus, toute une batterie de machineries de réparations (décrit au chapitre 2.4) est mise en jeu durant cette étape pour assurer l'intégrité génétique abaissant ainsi le taux d'erreur final à moins d'un nucléotide pour dix milliards.

2.3 La transcription des ARNm

La transcription est le mécanisme permettant de produire les molécules d'ARN à partir de l'ADN. Il existe de nombreux types d'ARN différents dont les plus connus sont les ARNr (ARN ribosomique) éléments constitutifs du ribosome, les ARNt (ARN de transfert) qui serviront au recrutement des acides aminés durant la traduction et les ARNm (ARN messager) qui codent pour les protéines. Dans ce chapitre, nous nous intéresserons particulièrement aux ARNm. La transcription des ARNm s'effectue en plusieurs étapes. L'initiation de la transcription, l'élongation, la terminaison et enfin, la maturation. Chez les eucaryotes, seuls les ARNm maturés pourront être transportés hors du noyau vers le cytoplasme de la cellule pour permettre la production des protéines.

2.3.1 L'initiation

L'initiation de la transcription n'est possible que si l'état de la chromatine autorise l'accessibilité des facteurs de transcriptions aux régions promotrices (B. Li, M. Carey, & Workman, 2007) et s'effectue grâce à des séquences promotrices dont nous ne citerons ici que la boite TATA, située à une trentaine de nucléotides du début de la transcription. Le facteur de transcription TFIID va reconnaître la boite TATA et recruter la machinerie de transcription. Deux hypothèses coexistent : i) le modèle séquentiel, où les facteurs sont recrutés séquentiellement et ii) le modèle de l'holoenzyme où l'ARN polymérase, située dans un complexe multi-protéique préexistant, est recrutée (Figure 8).

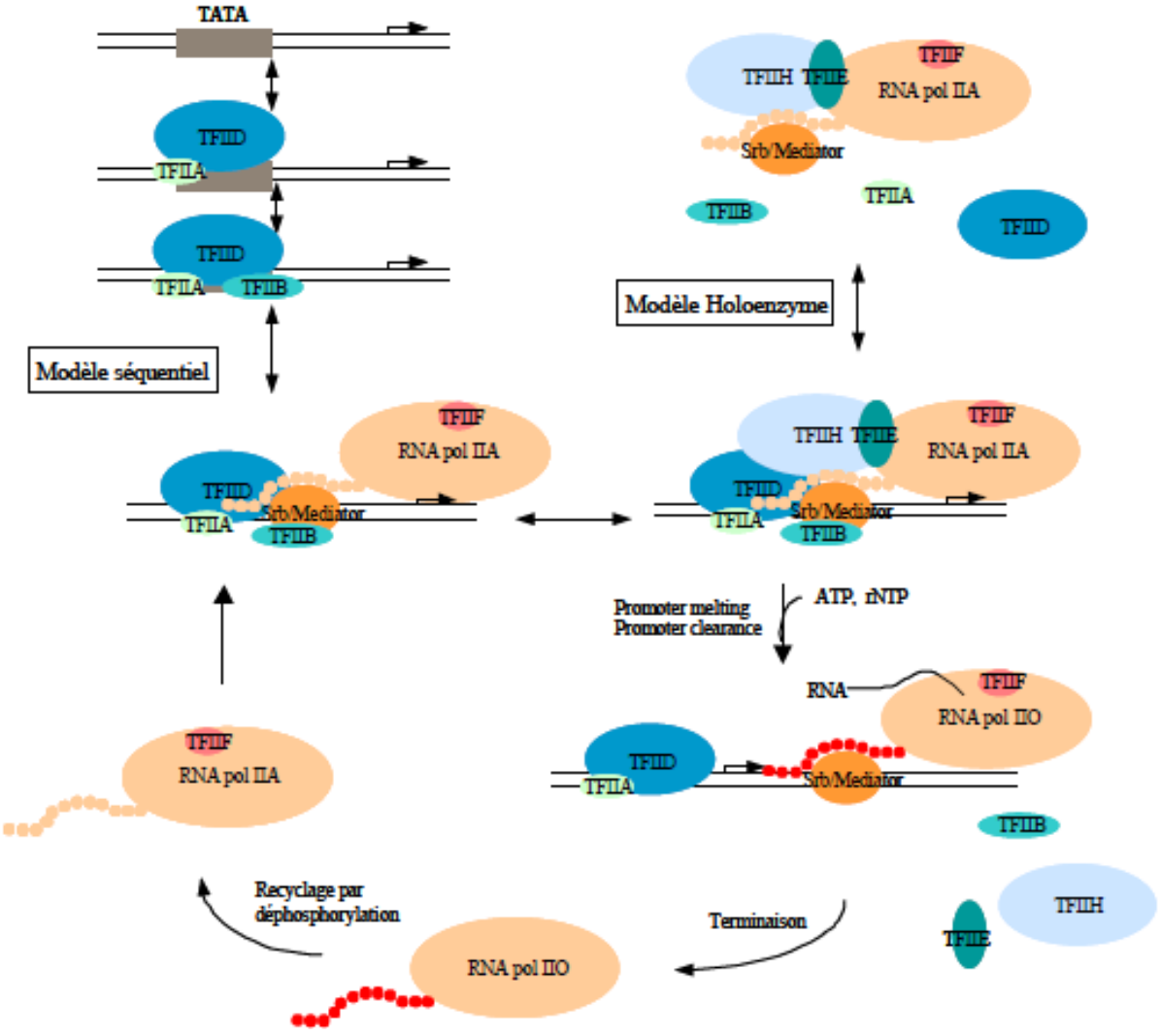


Figure 8 : Deux modèles pour l'assemblage du complexe d'initiation de la transcription. (Uhring, 2004)

Une fois le complexe d'initiation de la transcription recruté, le domaine CTD (*Carboxy Terminal Domain*) de l'ARN polymérase II sera phosphorylé par TFIIH entraînant la libération de la polymérase pour commencer l'élongation du brin d'ARN (Wade & Struhl, 2008).

2.3.2 L'élongation

Le complexe d'élongation, composé de 12 sous-unités, synthétise le brin d'ARN sur le même principe que le brin continu lors de la réplication (Svejstrup, 2007)(Armache, Kettenberger, & Cramer, 2005). Sur le plan chimique, une différence majeure est liée à la nature des nucléotides intégrés dont le sucre sera un ribose et au fait que le nucléotide uracile sera incorporé à la place de la thymine en complémentarité de l'adénine (Figure 9).

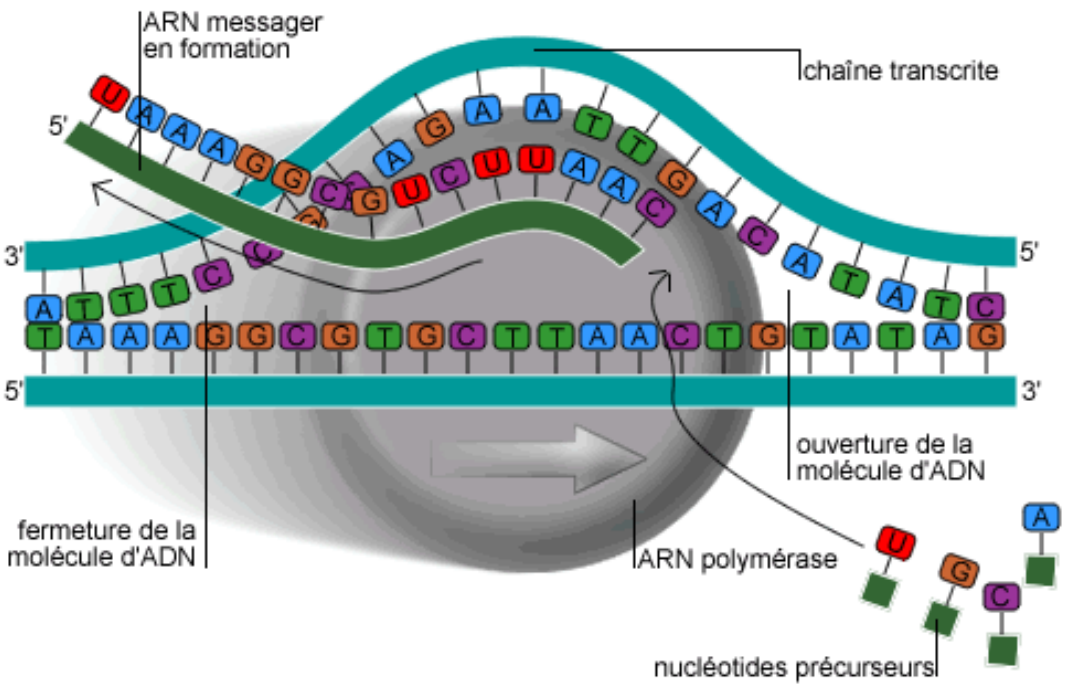


Figure 9 : Production d'une molécule d'ARN par une ARN polymérase.
La chaine transcrite (ou brin matrice) servira à la polymérase pour assembler les ribonucléotides A, U, G, C en brin d'ARN en suivant le principe de complémentarité des bases. La séquence de l'ARN résultant sera une copie de l'autre brin (brin codant), à l'exception des T qui seront remplacés par des U. (source : http://fr.wikipedia.org/wiki/Fichier:TranscriptionARN.png)

Dès le début de l'élongation, l'extrémité 5' de l'ARN néo-synthétisé sera « coiffée » pour le protéger de la dégradation (chapitre 2.3.4). La polymérase II possède également une activité de contrôle en assurant la relecture des ARN néo-synthétisés (Zenkin, Yuzenkova, & Severinov, 2006). Cette relecture peut être engagée directement par la polymérase (Walmacq et al., 2009) ou avec l'aide du facteur de transcription TFIIS (Koyama, Ito, Nakanishi, Kawamura, & Sekimizu, 2003).

2.3.3 Terminaison

Plusieurs hypothèses existent pour expliquer la terminaison de la transcription. Les modèles prédominants sont ceux de « l'anti-terminateur » et du « torpédo » (Buratowski, 2005). Le modèle de l'anti-terminateur a pour hypothèse qu'un facteur de transcription va stabiliser la phosphorylation du domaine CTD de la polymérase II pour maintenir l'élongation. Lorsque ce facteur est modifié ou quitte le complexe de transcription, le CTD se dé-phosphoryle, déstabilisant le complexe de transcription. Le second modèle postule que lorsque le brin d'ARN a passé le signal de polyadénylation (chapitre 2.3.4), la maturation de l'ARN se produit alors que la polymérase continue l'élongation. Le brin d'ARN sera coupé lors de ce processus pour être maturé et quitter le noyau. Le reste du brin en élongation par la polymérase ne sera donc plus protégé par la coiffe et sera dégradé depuis son extrémité 5' par une exonucléase, plus rapidement qu'il ne sera allongé. L'exonucléase finit par rejoindre la polymérase et par la dissocier de l'ADN. La régulation de la terminaison de la transcription semble être contrôlée par les séquences de polyadénylation ainsi que par la distance au promoteur (Jenks, O'Rourke, & Reines, 2008).

2.3.4 Maturation

L'ARNm subit plusieurs modifications post-transcriptionnelles.

- La coiffe est ajoutée pendant l'élongation de la transcription pour stabiliser l'ARN en formation. Elle est constituée d'un nucléotide modifié à l'extrémité 5' de l'ARNm.

- La polyadénylation, qui consiste en l'addition d'une queue poly (A) (environ 200 nucléotides chez les eucaryotes supérieurs) à l'extrémité 3' de l'ARNm. Le complexe

CPSF (*Cleavage/Polyadenylation Specificity Factor*) va couper le brin d'ARN après le motif AAUAAA du site de polyadénylation situé une quinzaine de nucléotides en amont d'un site enrichi en U/G. La PAP (*Poly(A)Polymérase*) va interagir avec CPSF et l'ARN pour synthétiser la queue polyA depuis le site de polyadénylation.

- L'épissage, mécanisme propre aux eucaryotes, est une étape qui consiste à enlever les introns de la séquence d'ARNm. Un intron se caractérise par une séquence UAUAAC appelée « boite de branchement » située au milieu de l'intron qui va former une boucle avec une de ces extrémités. Ce processus est catalysé par un complexe nucléoprotéique, le spliceosome, constitué d'ARN et de protéines. Ce complexe va libérer les introns sous forme d'ARN « en lasso » qui seront dégradés dans la cellule (Figure 10). Il peut y avoir plusieurs introns par ARNm pouvant produire des épissages alternatifs.

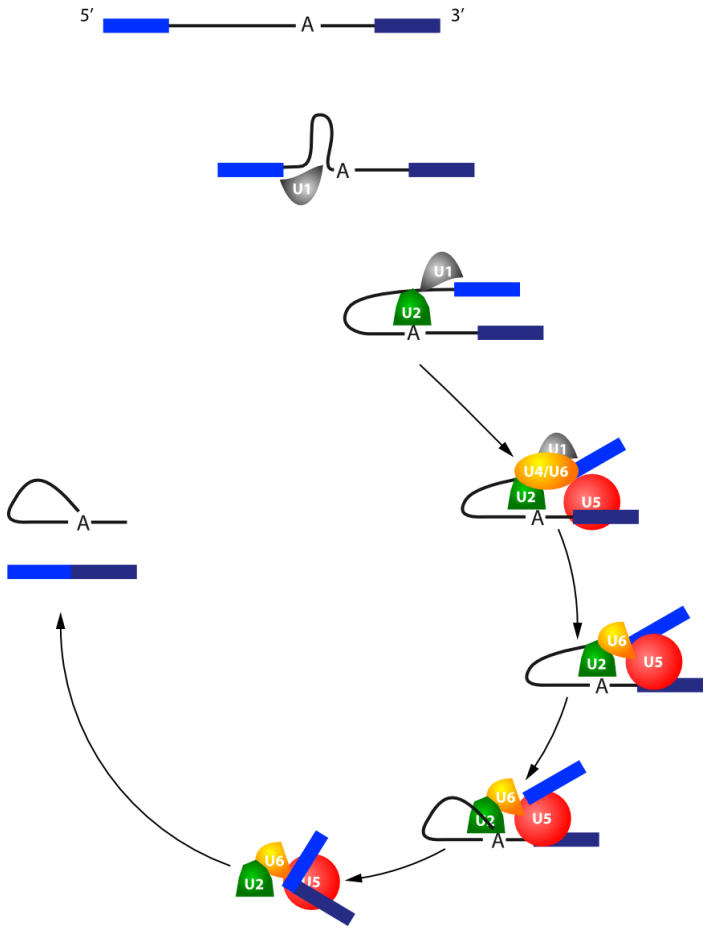


Figure 10: Séquence d'événements aboutissant à la constitution du spliceosome.
La sous-unité U1 s'associe à la jonction 5' de l'intron. U2 s'associe à la boite de branchement. U4 associée à U6 rapproche U1 et U2, réalisant ainsi un pont entre la jonction 5' de l'intron et la boite de branchement. U5 s'associe à son tour et rapproche les bords 3' et 5' des exons à suturer. U4 et U1 quittent le complexe. Le 2'-OH du A de la boite de branchement coupe la jonction 5' de l'intron. Le 3'-OH du nucléotide en 3' de l'exon en amont coupe l'autre jonction. L'ARNm épissé et l'intron en "lasso" sont libérés. Source (http://fr.wikipedia.org/wiki/Fichier:Splicing.svg).

- L'édition d'ARNm est un mécanisme qui, comme son nom l'indique, modifie des nucléotides de la séquence. Ces mécanismes sont différents selon l'organisme et le type cellulaire (pour revue voir (Aphasizhev, 2007)) Le premier démontré est l'insertion d'une séquence poly-U dans certains transcrits chez les Trypanosomes. Nous pouvons par exemple, citer le changement de la séquence de l'ARNm par la désamination de certains C en U, ou encore de A en I dans le cytoplasme de certaines lignées cellulaires des mammifères.

2.4 La réparation de l'ADN

Le maintien de l'intégrité de l'information génétique portée par l'ADN est assuré par de nombreuses machineries de réparation au sein de la cellule. Ces machineries sont regroupées en quatre grandes familles (Hanawalt & Spivak, 2008) :

- La réparation directe de lésion comme la photolyase qui enlève les dimères de thymine, ou la méthyltransférase qui enlève les groupements méthyl en position O-6 de la guanine.

- La réparation par excision où l'erreur est éliminée en se servant du brin non endommagé.
    - La réparation par excision de nucléotide (REN) qui supprime un ou plusieurs nucléotides du brin présentant une erreur.
    - La réparation par excision de base (REB) qui enlève une base altérée ou une base inappropriée comme l'uracile dans l'ADN.
    - La réparation des erreurs d'appariements (RMA) qui enlève du brin d'ADN, les bases qui ne sont pas appariées au brin complémentaire.

- La réparation par recombinaison lorsqu'il y a une cassure sur un brin de l'ADN.

- La synthèse des trans-lésions lorsqu'il y a des lésions en opposition sur les deux brins de l'ADN.

Ces mécanismes peuvent être activés soit, directement, par la structure que provoque la lésion elle-même, notamment lors des mécanismes de réparation directe, soit par les complexes rencontrant les bases altérées (Ljungman, 2005). La réparation peut donc être induite par la machinerie de réplication ou être couplée à la transcription.

Un exemple de mécanisme couplé à la réplication est la réparation par recombinaison. Si un brin porte une base altérée, la réplication peut entraîner une brèche du brin synthétisé, qui sera réparée par recombinaison (Figure 11).

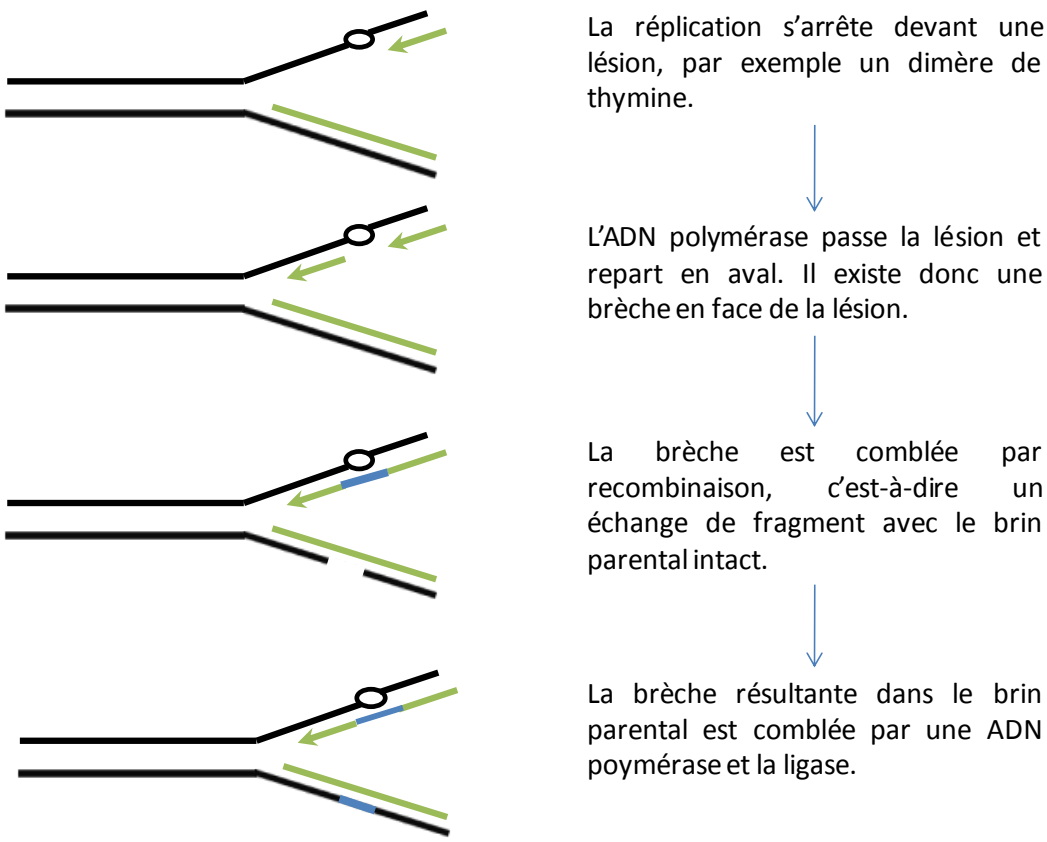


Figure 11: Réparation des brèches par recombinaison.
Lorsqu'une brèche se crée sur un brin de l'ADN, lors de la réplication, elle peut être réparée par recombinaison avec le brin parental matrice de l'autre brin synthétisé (adapté de (Cooper, 1999)).

Un autre exemple couplé à la réplication est la réparation des erreurs d'appariements (Iyer, Pluciennik, Burdett, & Modrich, 2006). Dans les cas où, malgré l'activité de relecture de l'ADN polymérase, le brin néo-synthétisé comporte un mésappariement, sa structure sera reconnue et entraînera le recrutement d'une endonucléase qui va couper le nouveau brin en

amont du mésappariement, créant une brèche qui sera par la suite comblée par une ADN polymérase et la ligase.

La réparation peut également être initiée lors de la transcription par deux mécanismes de réparation qui sont la réparation par excision de nucléotides (Figure 12) et par excision de bases (Figure 13).

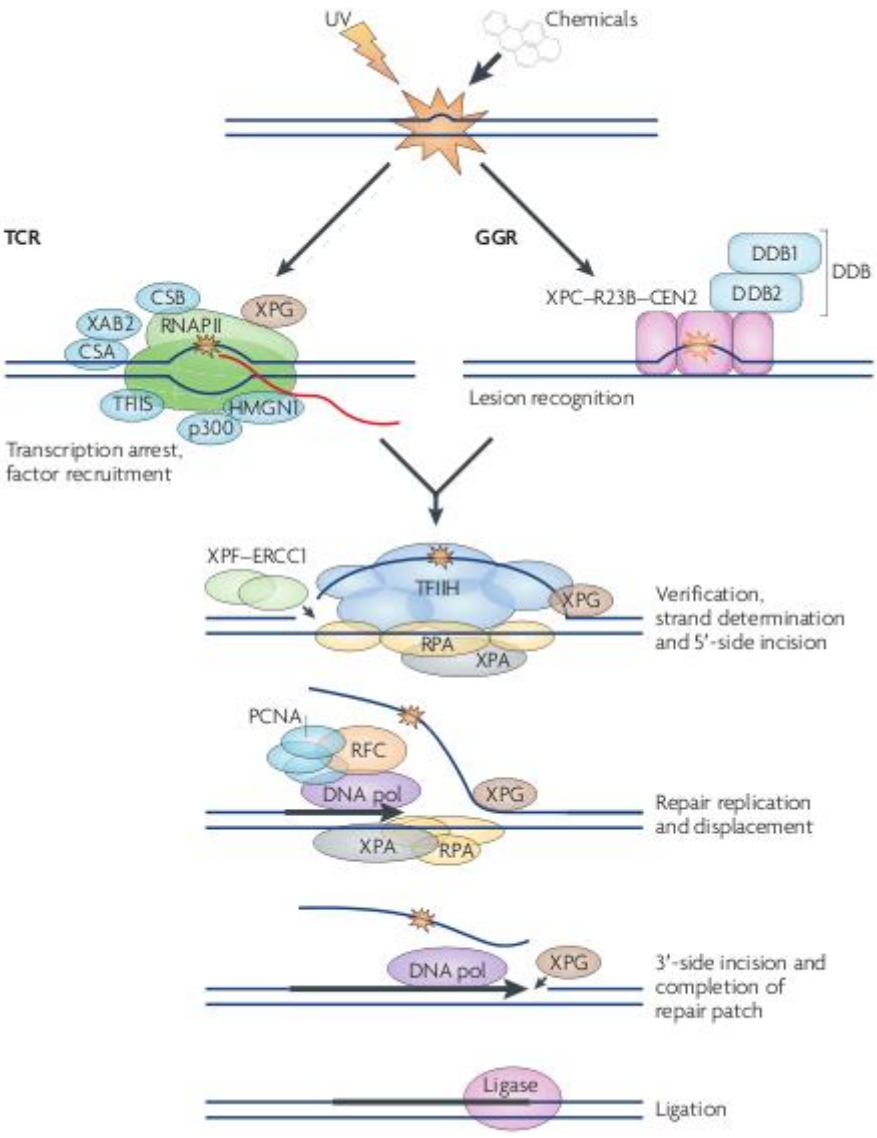


Figure 12 : Modèle de réparation par excision de nucléotides.
Extrait de (Hanawalt & Spivak, 2008). La machinerie de réparation peut être recrutée, soit après un arrêt de la machinerie de transcription dans la voie TCR (Transcription Coupled Repair), soit par un complexe de reconnaissance de la lésion dans la voie GGR (Global Genomic Repair). La machinerie de réparation est composée notamment de TFIIH qui va permettre l'action des endonucléases pour ouvrir le brin avec la lésion du coté 5'. L'ouverture du brin permettra une nouvelle synthèse par la polymérase.

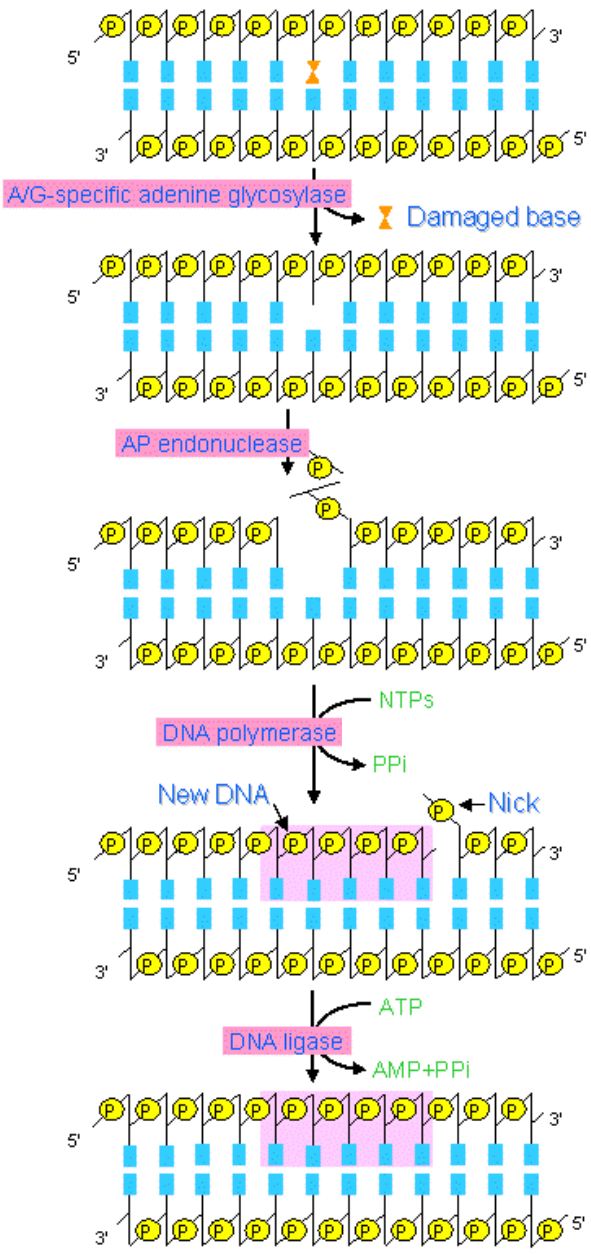


Figure 13 : Modèle de réparation par excision de base.
Une ADN glycosylase spécifique du type de base rompt la liaison sucre-base, laissant un site AP (apurique ou apyrimidique). Ce site sera reconnu par une endonucléase qui va ouvrir le brin du côté 5', permettant une nouvelle synthèse du brin par l'ADN polymérase. Tiré de (Svejstrup, 2002)

Pour réparer la lésion, tous ces mécanismes se basent sur le brin non endommagé. Lorsqu'il y a des lésions sur chacun des deux brins, ou une coupure double brin, les mécanismes de réparation impliqués restaurent la continuité de la molécule d'ADN de façon biochimique sans tenir compte de la séquence. Ce type de réparation crée donc une mutation dans la séquence génomique.

2.5 De l'ARN à la protéine

L'ARNm peut être représenté en 3 parties : la région 5'UTR (*UnTranslated Region*) qui débute par la coiffe ; la région codante délimitée par le codon d'initiation en 5' et par le codon stop en 3' ; la région 3' UTR se terminant par la queue polyadénylée.
L'ARNm sera traduit en protéine grâce au ribosome. Chaque codon de l'ARNm composé de trois nucléotides correspond à un acide aminé particulier. Cette correspondance est le code génétique (Figure 14). Plusieurs codons peuvent coder pour le même acide aminé, mais selon l'organelle ou l'organisme, il existe des différences entre les codes génétiques utilisés. Trois codons (UAA, UAG et UGA) qui ne codent pas pour des acides aminés et qui constituent les codons stop entraînant l'arrêt de la traduction.

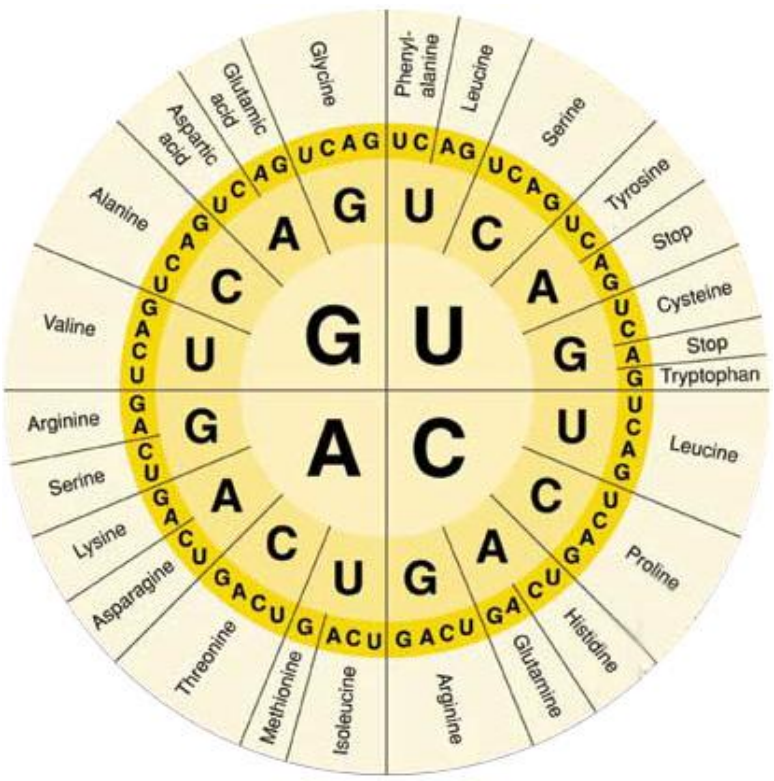


Figure 14: Le code génétique.
La première lettre du codon est au centre et la dernière en périphérie.

Le ribosome est un complexe de deux sous-unités ribonucléoprotéiques. Chez l'homme, la grande sous-unité 60S est composée de 49 protéines et des ARNr 28S, 5,8S, et 5S et la petite sous-unité 40S est composée de 33 protéines ribosomales et de l'ARNr 18S. Les deux sous-unités se fixent avec l'ARNt initiateur sur l'ARNm au niveau du codon d'initiation de la transcription qui est le codon AUG. La séquence d'ARN sera lue dans son cadre de lecture, c'est-à-dire de codon en codon de façon non chevauchante. Chaque codon de l'ARNm permettra de recruter, au sein du ribosome, l'ARNt porteur de l'acide aminé qui sera ajouté à la chaine peptidique. Chaque ARNt sera recruté spécifiquement grâce à son anticodon qui correspond à la séquence complémentaire du codon de l'ARNm. Quand le ribosome atteint le codon stop, les protéines de terminaison de la traduction ERF-1 et ERF-2 dissocieront le ribosome et libéreront la chaine peptidique synthétisée.

2.6 Contrôle de la qualité des ARNm

Nous avons vu précédemment que les mécanismes de réplication et de transcription, grâce à leur capacité de relecture, commettent peu d'erreurs. Nous avons également vu que toute une batterie de voies de réparation empêche l'apparition d'une mutation dans le génome. Cependant, la probabilité d'obtenir un ARNm portant un nucléotide erroné n'est pas nulle. Un tel événement aboutit à une matrice produisant une protéine erronée. Pour minimiser l'impact des erreurs pouvant se produire sur les ARNm, la cellule possède différents mécanismes de contrôle. Les trois mécanismes les plus connus sont:

- Le NMD (*Nonsense-Mediated mRNA Decay*) qui élimine les ARNm ayant un codon stop prématuré (Figure 15).

- Le NSD (*NonStop mRNA Decay*) qui élimine les ARNm n'ayant pas de codon stop (Figure 16).

- Le NGD (*No-Go mRNA Dec*ay) qui élimine les ARNm qui ont une structure secondaire bloquant les ribosomes (Figure 17).

Ces mécanismes assurent la dégradation des ARNm produisant des protéines aberrantes dès le premier cycle de traduction. Ces ARNm sont donc dégradés presque instantanément dès leur sortie du noyau (Isken & Maquat, 2007).

Le mécanisme NMD dégrade les ARNm qui auront subi une modification codant pour un codon stop ou une modification qui modifie l'épissage en introduisant un codon stop avant l'exon suivant. Ce mécanisme permet à la cellule de minimiser la production de protéines tronquées non fonctionnelles.

Si une modification fait apparaître un site précoce de polyadénylation avant un codon stop ou bien si une modification entraîne un épissage qui résulterait en un ARNm sans codon stop, le mécanisme de NSD sera recruté. Ce mécanisme permet de libérer les ribosomes ayant participé à la traduction de l'ARNm aberrant et d'éviter une accumulation de peptides enrichis en poly-lysine (la lysine est codée par AAA).

Le mécanisme de NGD permet de dégrader les ARNm qui provoquent un « blocage » des ribosomes en ayant par exemple, une structure secondaire infranchissable, en étant en interaction forte avec une autre molécule ou parce que sa séquence implique de nombreux codons rares, ralentissant l'avancé du ribosome.

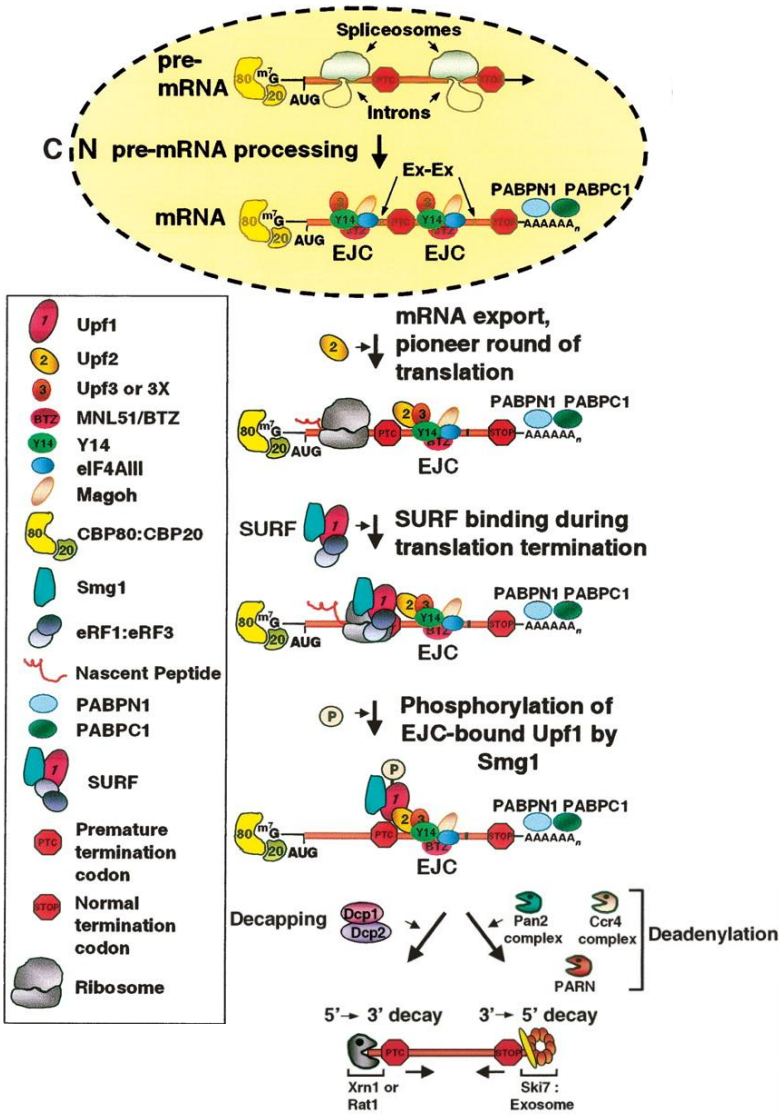


Figure 15 : Modèle du NMD.
Après maturation, l'ARNm porte sur ses jonctions exons-exons un complexe protéique dérivant du spliceosome EJC (*Exon Junction Complex*). Ce complexe recrute dans le noyau Upf3 qui, après transfert de l'ARNm dans le cytoplasme, recrutera à son tour Upf2. Ce complexe sera dissocié durant la traduction par le premier passage du ribosome. Si le ribosome s'arrête sur un codon stop, alors qu'il existe encore un complexe EJC avec Upf3 et Upf2 en aval (50ène de nucléotides) alors le complexe SURF, formé entre autres d'Upf1, sera recruté sur l'EJC et Upf1 sera phosphorylé. Cette phosphorylation va libérer le ribosome. La déphosphorylation d'Upf1 permettra de recruter soit des facteurs qui détruiront la coiffe (decapping), soit des complexes protéiques digérant la queue poly-A (Pan2 et Ccr4). L'ARNm « mis à nu » sera sans protection face aux exonucléases (Xrn1, Rat1, exosome) qui le digéreront (Isken & Maquat, 2007).

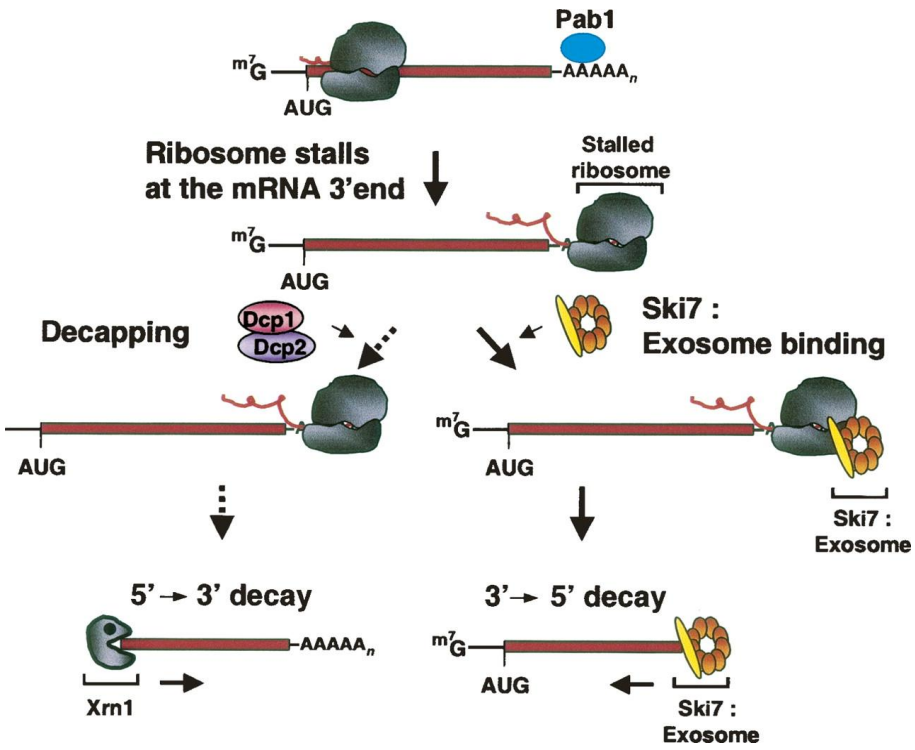


Figure 16 : Modèle du NSD.
Lorsque le ribosome atteint la fin de l'ARNm, la protéine Ski7 de l'exosome, pourra se lier à son site vide en mimant les facteurs de fin de traduction. Ski7 permet donc de dissocier le ribosome et de dégrader sans déadénylation préalable l'ARNm. Il peut également y avoir un recrutement par le ribosome des facteurs digérant la coiffe de l'ARNm pour permettre à Xrn1 de le digérer (Isken & Maquat, 2007).

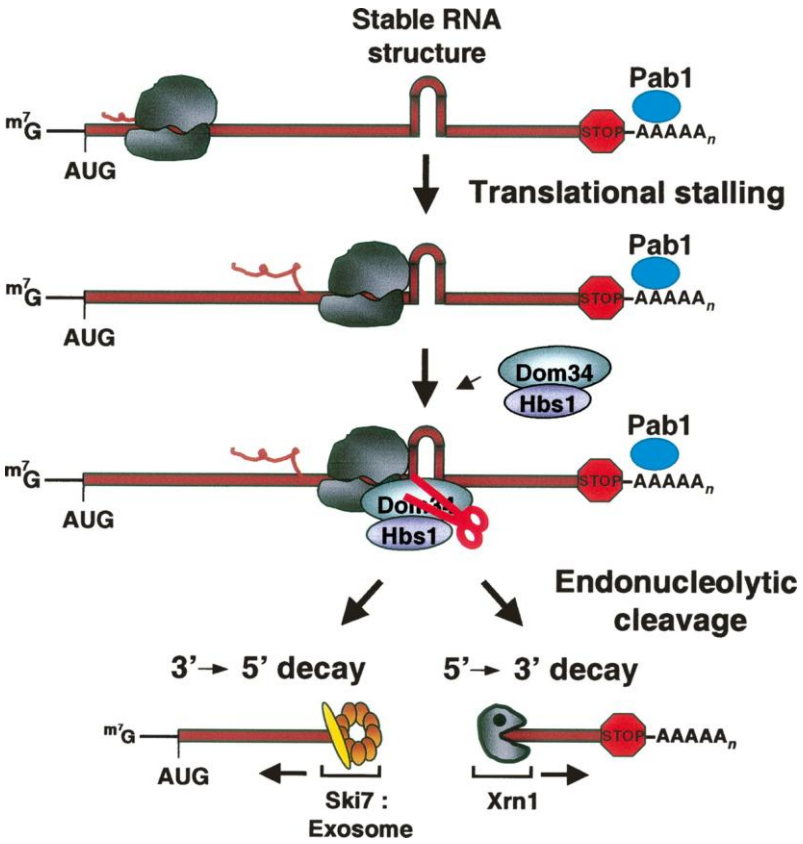


Figure 17 : Modèle du NGD.
Lorsque le ribosome se retrouve bloqué sur l'ARNm, par exemple lorsque la séquence forme une structure infranchissable, le facteur Dom34 est recruté (sûrement par mimétisme d'un ARNt), avec Hbs1 qui a une activité endonucléasique. Une fois l'ARNm coupé par Hbs1, il sera digéré par les voies de dégradation, l'exosome et Xrn1 (Isken & Maquat, 2007).

2.7 Impact des modifications de l'ARNm sur les protéines

Lorsque tous les systèmes que nous venons de décrire ne parviennent pas à empêcher l'expression d'un ARNm modifié, une protéine aberrante pourra être produite. Par modification on sous-entend soit, la substitution d'un nucléotide par un autre, soit, la perte (délétion) ou le gain (insertion) d'un ou plusieurs nucléotides. Une modification dans la séquence d'un ARN pré-messager ou messager peut occasionner plusieurs effets différents selon sa localisation. Une modification dans les sites responsables de l'épissage comme le site de branchement ou les bords des exons, peut entraîner la perte d'un exon ou le gain d'un intron dans la séquence codante. Cette modification peut entraîner une suppression ou un ajout de domaine à la protéine, mais peut aussi changer de manière dramatique la protéine, si cette modification affecte le cadre de lecture. Une modification dans la séquence codante peut entraîner plusieurs types d'effets sur la protéine (Figure 18).

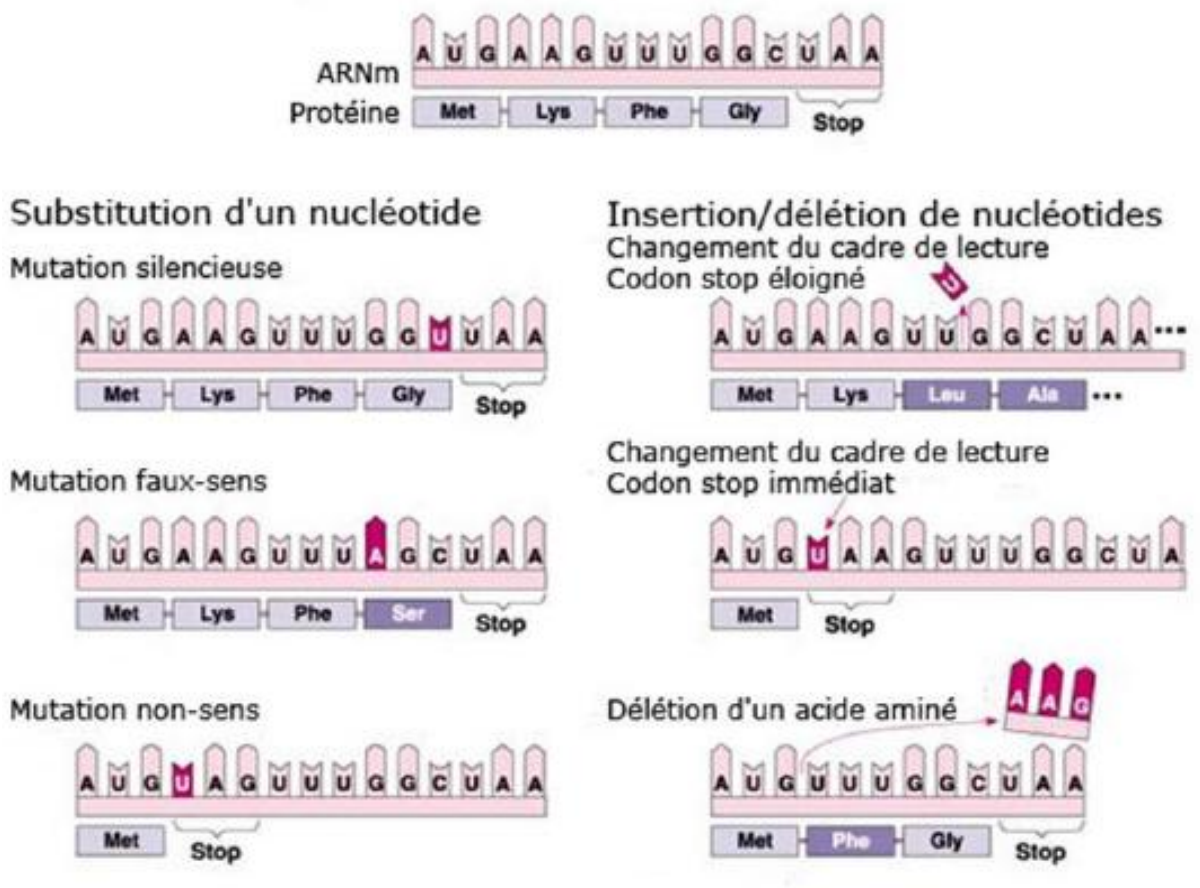


Figure 18 : Exemples d'effets d'une modification de l'ARNm sur la protéine résultante. (adapté de http://fajerpc.magnet.fsu.edu/Education/2010/Lectures/26_DNA_Transcription.htm)

Dans le cas d'une substitution, il est possible qu'elle soit « silencieuse » si elle ne change pas l'acide aminé. En effet, le troisième nucléotide de l'anticodon a souvent une importance mineure sur le choix de l'acide aminé (Figure 14). Si la modification est non silencieuse, alors on pourra avoir une modification de l'acide aminé dans la protéine (mutation faux sens). L'impact de cette modification est très différent selon l'importance de l'acide aminé modifié et son remplaçant. Si ce dernier présente des propriétés physico-chimiques très différentes de l'acide aminé initial, cela peut induire une conformation moins stable de la protéine. Si l'acide aminé modifié est impliqué directement dans un site catalytique, la protéine peut devenir non fonctionnelle. Une modification peut introduire un codon stop prématuré (mutation non sens) ou supprimer un codon stop et aboutir à une protéine tronquée ou plus longue respectivement.

Dans le cas d'une insertion ou d'une délétion de nucléotides, il y aura un décalage du cadre de lecture et la synthèse d'une protéine dont la séquence, en aval de la position de la

modification, sera complètement différente de celle d'origine. Il faut nuancer ce dernier point car il est possible de retrouver le cadre de lecture si le décalage est un multiple de 3. Par exemple, une perte d'un acide aminé dans la protéine sera observée s'il y a délétion d'un codon entier.

Enfin, la modification dans la séquence codante peut potentiellement changer sa maturation, si la modification fait apparaitre un nouveau site d'épissage, ou un nouveau site de polyadénylation.

Le maintien de l'intégrité du message génétique est donc primordial pour le bon fonctionnement cellulaire. Par exemple, la modification de protéines régulatrices de l'expression des gènes de prolifération cellulaire peut être à l'origine de cancers. Le cancer est souvent décrit comme une maladie due à une perte de l'intégrité génétique et à une accumulation de mutations.

3 Cancer : Un seul mot pour des maladies complexes

Le terme cancer recouvre un ensemble de maladies ayant en commun une augmentation anormale de la prolifération cellulaire au sein d'un tissu de l'organisme. Certaines cellules, peuvent faire des invasions, d'autres peuvent également quitter leur lieu de production et former des métastases. Ces cellules ne répondent pas aux facteurs de contrôle de la prolifération et n'entrent pas en apoptose ou mort cellulaire, conférant un pouvoir de division infinie. Les cellules cancéreuses peuvent également influencer leur environnement en sécrétant des facteurs activant certains processus biologiques, par exemple l'angiogénèse, c'est-à-dire la création de capillaires sanguins permettant de « nourrir » la tumeur.

Face aux multiples types de cancers existants, plusieurs hypothèses non-exclusives ont été élaborées pour expliquer l'origine d'un cancer.

La première hypothèse propose qu'au sein d'une tumeur cancéreuse, toutes les cellules cancéreuses dérivent d'un même clone qui a accumulé différentes mutations au fil des générations cellulaires (les cellules précancéreuses). Les cellules filles acquérant des

mutations favorisant leur prolifération deviennent majoritaires jusqu'à obtention de cellules cancéreuses qui vont former une tumeur (Loeb, Bielas, & Beckman, 2008).

La seconde hypothèse suggère que seules quelques cellules sont cancéreuses dans la tumeur, les cellules souches cancéreuses et que les autres sont des cellules saines subissant l'effet des cellules cancéreuses (Maitland & Anne Collins, 2005). Sous cette hypothèse les cellules souches seraient suffisantes pour initier et maintenir un cancer

La troisième hypothèse est que des mutations apparaissent dans les cellules de la matrice cellulaire (comme le tissu conjonctif ou la lame basale), impliquant une dérégulation de la différenciation et la prolifération des cellules dépendant de cette matrice en périphérie des organes (comme les cellules épithéliales) (Cunha, Cooke, & Kurita, 2004; Tsuruta et al., 2008).

Toutes ces hypothèses peuvent expliquer la récidive des cancers. Dans la première hypothèse, une logique d'évolution de la population cellulaire est considérée. Les cellules ayant acquis les mutations les plus favorables à la prolifération formeront un premier cancer. Si ce cancer est soigné par ablation ou destruction des cellules, alors une autre population de cellules cancéreuses dérivant des cellules précancéreuses ayant développé d'autres mutations en parallèle pourra à son tour former une tumeur. Dans la seconde, nous aurons beau soigner le cancer, il n'y aura pas de guérison tant que les cellules souches qui provoquent la maladie ne seront pas atteintes. Enfin, dans la troisième hypothèse, il est inutile d'agir sur les cellules cancéreuses car, tant que la communication avec la matrice sera défectueuse, il y aura production de cellules cancéreuses. La seule alternative est l'ablation de l'organe entier (comme pour la prostate) ou de la région exhibant une production de cellules cancéreuses (comme pour le cancer du colon). Il est aussi possible que l'adhésion des cellules cancéreuses à la matrice soit modifiée, ces cellules vont alors pouvoir migrer pour former des métastases. Dans ce dernier cas, l'ablation de l'organe ne soignera pas le cancer.

Des soins aussi extrêmes pourraient être évités si le cancer était détecté dans ses stades les plus précoces. C'est pour cela qu'il est important de trouver des marqueurs spécifiques précoces des cancers. Dans cette optique, de nombreuses expériences de transcriptomique ont été réalisées dans le but de caractériser l'expression des gènes au sein des tissus cancéreux (Boon et al., 2002; Chung et al., 2002; Fujii et al., 2002; Greenman et al., 2007; Hanahan & Weinberg, 2000; E. M. Reis et al., 2005; Wood et al., 2007).

4 La transcriptomique

La transcriptomique représente l'analyse à haut débit de l'expression des gènes par mesure de la quantité d'ARN. Dans de nombreuses approches à haut débit, les ARN sont caractérisés par leur séquence. Ainsi, l'évolution des techniques de transcriptomique est directement dépendante des progrès réalisés par les techniques de séquençage et d'assemblage. Le séquençage est réalisé à partir de banques d'ADN complémentaire (ADNc). L'une des techniques de base de la transcriptomique est l'analyse des EST (*Expressed Sequence Tag*) que nous présenterons au chapitre 4.1. Les séquences répertoriées par cette technique ont permis de prédire des séquences hybridant spécifiquement les ARN que l'on désire étudier, ce qui a abouti à la mise au point d'autres techniques de transcriptomique telles que, les puces à ADN, basées sur l'hybridation des ARN de la cellule sur des fragments de séquences complémentaires ou encore, la technique du SAGE (*Serial Analysis Gene Expression*) présentée au chapitre 4.2, dont l'analyse nécessite de connaître les séquences d'ARNm.

4.1 EST (Expressed Sequence Tag)

La technique des EST permet d'analyser les gènes exprimés dans un tissu et/ou de positionner le gène dans le génome en comparant sa séquence à celui de la séquence génomique. L'EST est un fragment de séquence réalisé à partir d'un ADNc (ADN complémentaire) issu d'un ARNm (Figure 19).

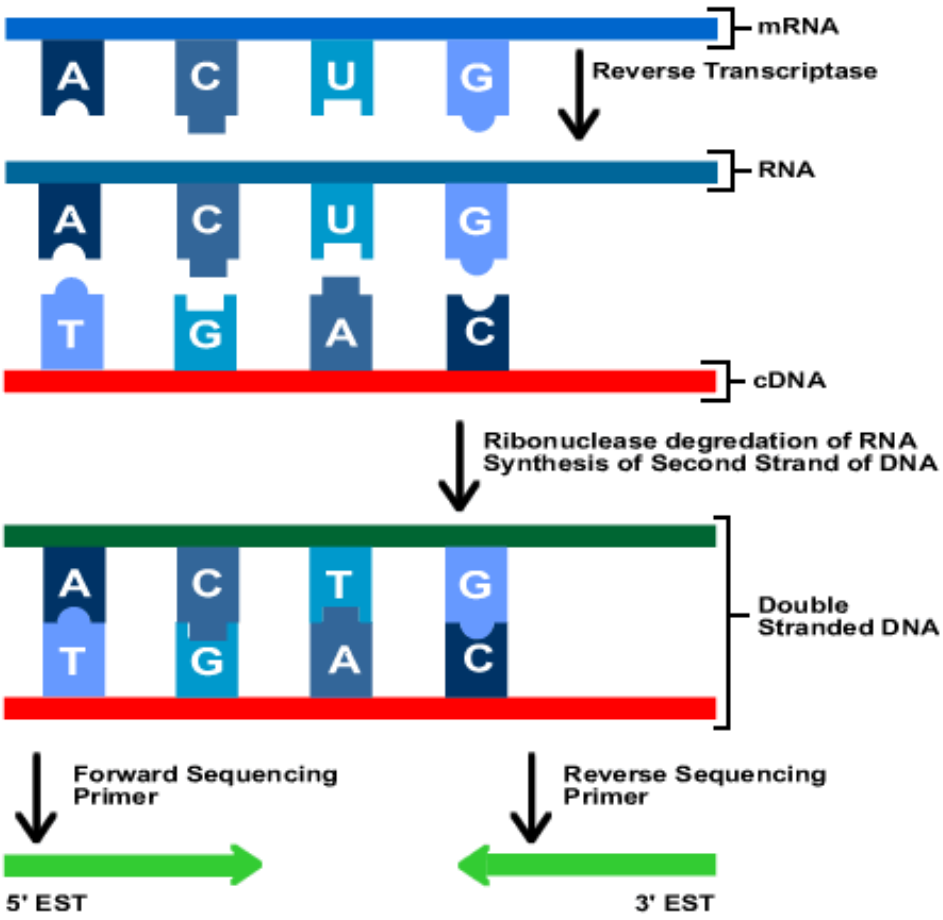


Figure 19 : Résumé de l'obtention des EST (*Expressed Sequence Tag*).

Les séquences d'EST sont des marqueurs de l'ARNm dont ils sont issus. Le décompte des EST spécifiques de chaque gène et provenant d'une cellule ou d'un tissu permet d'estimer les proportions d'ARNm présents au sein d'une cellule ou d'un tissu.

Les séquences d'EST sont référencées et disponibles au NCBI depuis 1993 dans la banque dbEST (http://www.ncbi.nlm.nih.gov/dbEST). Elle contient aujourd'hui presque 63 millions d'EST (septembre 2009) dont plus de 8 millions provenant de l'homme. La banque UniGene (http://www.ncbi.nlm.nih.gov/unigene) regroupe les EST correspondant au même gène.

4.2 SAGE (Serial Analysis of Gene Expression)

Le SAGE (V. E. Velculescu et al., 1995) est une technique de séquençage qui, comme pour les EST, permet de mesurer l'expression des gènes dans une cellule ou un tissu. Cette

technique est originale car elle effectue une amplification des ARNm en début de protocole. Cette importante augmentation du matériel rend le SAGE environ six fois plus sensible que les EST pour détecter un ARNm rare (M. Sun et al., 2004).

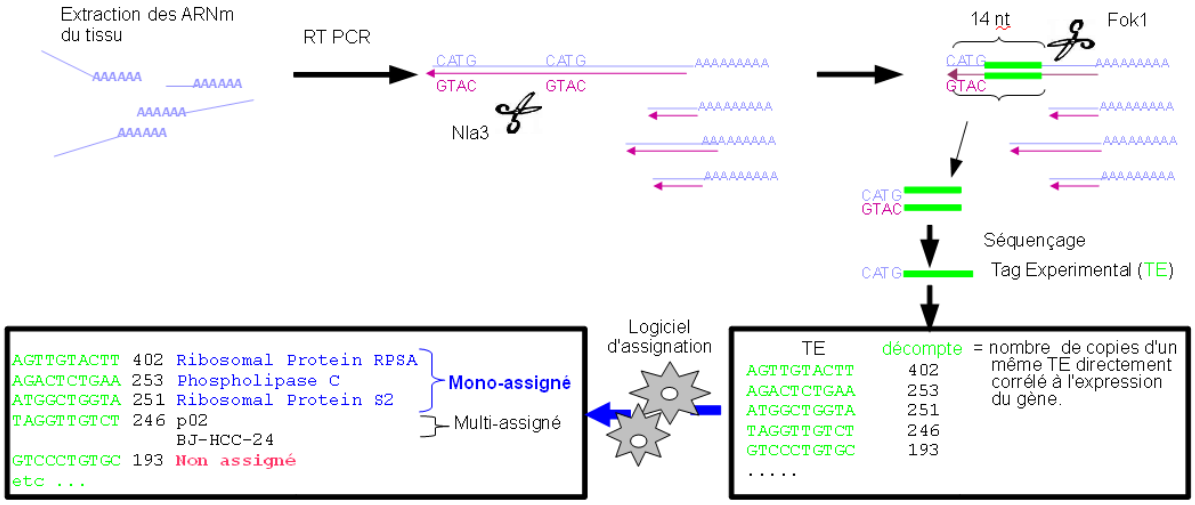


Figure 20 : Résumé de la technique de SAGE Nla3.

Les ARNm du tissu sont extraits, puis subissent une RT-PCR, c'est-à-dire une transcription inverse qui fera une copie de l'ARNm en ADNc, puis une étape de PCR (*Polymérase Chain Reaction*) qui va multiplier les copies d'ADNc. Ces ADNc seront capturés sur des billes par leur queue polyadénylée. Les ADNc vont subir une première digestion par l'enzyme Nla3 qui va reconnaitre tous les sites CATG. Après élution des fragments digérés, une seconde digestion de l'ADN sera effectuée par l'enzyme Fok1, qui va couper 14 nucléotides en incluant le site CATG. Les fragments élués seront séquencés et décomptés. Enfin, un logiciel assignera les séquences ou Tags Expérimentaux (TE) aux gènes correspondants.

Les méthodes de transcriptomique présentées rapidement utilisent donc la connaissance des séquences des gènes pour pouvoir analyser leur expression et ont notamment été utilisées pour étudier les différences d'expression des gènes au sein de tissus sains et cancéreux. Ces études se basent sur le principe que la séquence de l'ARN exprimé par le gène sera identique dans tous les types de tissus. Cependant, dans une étude préliminaire réalisée par des membres de la société Genclis (Genomique Clinical Synergy) et de l'IGBMC (Brulliard et al., 2007), il a été démontré que, pour 17 gènes fortement exprimés, les séquences des EST provenant de tissus cancéreux montraient, de façon significative, une plus grande variabilité

de séquence. Ce résultat remet en question les analyses réalisées en transcriptomique qui n'ont jamais pris en compte l'existence d'une diversité accrue des ARNm dans les tissus cancéreux.

5 Les EST significativement plus hétérogènes en Cancer

Nous présenterons rapidement ici les principaux résultats publiés (Brulliard et al., 2007). Ces résultats s'appuient sur l'analyse de plus de 2 millions d'EST provenant de tissus cancéreux comparés à 2 millions d'EST de tissus sains, en prenant systématiquement comme référence, l'ARNm de qualité présent dans la banque RefSeq. Les études prélimianires et complémentaires ont permis d'isoler plus de 700 gènes qui présentent des positions ayant une augmentation significative de modifications (mutations, insertions ou délétions) dans les EST provenant des tissus cancéreux par rapport aux EST provenant de tissus sains. A l'inverse, seuls 300 gènes environs ont révélé une augmentation significative de positions modifiées supérieures dans les tissus sains par rapport aux tissus cancéreux. Pour des raisons de confidentialité et afin de préserver un avantage face aux concurrents potentiels, dans le cadre de la publication (Brulliard et al., 2007), seuls 17 gènes ont été présentés. Ces gènes ont été sélectionnés pour leur forte expression aboutissant à un grand nombre de séquences d'EST sensiblement équivalent dans les 2 populations (Normal et Cancer). La disponibilité de nombreuses séquences permettait des études statistiques robustes qui ne pouvaient prêter le flanc à des critiques basées sur des effectifs très divergents. Pour chaque position de l'ARNm considéré, les taux de modifications observées entre tissus sains et cancéreux ont été comparés (protocole décrit au chapitre 10.2). Cette étude a permis de mettre en évidence pour chacun des gènes une liste de positions ayant une proportion de modifications significativement différentes (Figure 21).

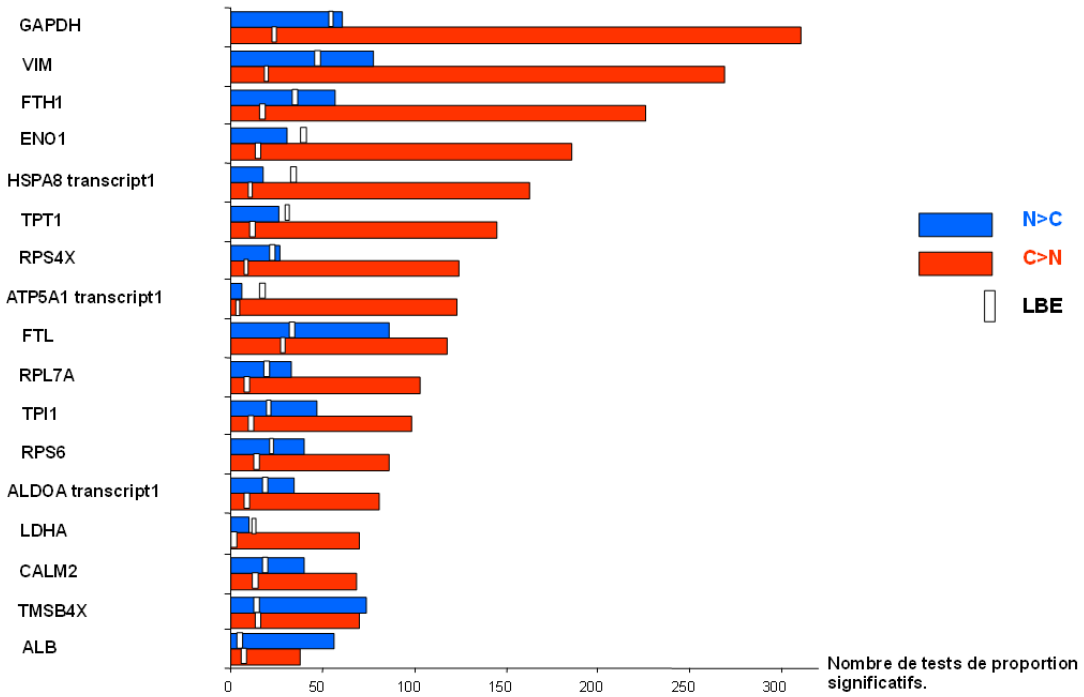


Figure 21 : Nombre de modifications statistiquement significatives sur les EST originaires de tissus sains (N) et cancéreux (C) pour 17 gènes ayant un grand nombre d'EST référencés.
Le nombre de modifications ayant une proportion statistiquement supérieure dans les EST issus de tissus sains que dans les EST issus de tissus cancéreux (N>C), et inversement (C>N) sont représentés par les barres bleues et rouges respectivement. Le nombre de faux positifs estimé par le LBE (*Location-Based Estimator*) est représenté par un trait blanc sur chaque barre (Brulliard et al., 2007).

Cette étude préliminaire a été réalisée en prenant en compte uniquement les substitutions de nucléotides et non, les insertions et délétions. Sur les 17 gènes étudiés, 15 ont plus de positions présentant des taux de modifications significativement augmentés dans les EST provenant de tissus cancéreux.

Cette étude a été réalisée avec des critères de stringence stricts. En effet, comme nous l'avons décrit précédemment, de nombreux éléments peuvent introduire des modifications dans les ARNm. Ainsi, lors de l'analyse des séquences d'EST, nous savons que nous observons le résultat final de toute une série de processus biologiques, allant de la réplication du génome à l'obtention d'un ARNm mature, qui peuvent tous générer des altérations de la séquence finale. De plus, la séquence de l'EST est obtenue à partir d'une technologie qui peut incorporer des erreurs dans la séquence. Cependant, le fait de comparer des EST provenant de

tissus sains et cancéreux est un moyen d'éliminer le biais dû aux erreurs de séquençage, car il est du même ordre de grandeur dans les deux groupes, les enzymes utilisées ne sachant pas, *à priori*, d'où proviennent les ARNm et les fragments d'ADNc en résultant. De plus, pour réaliser des tests statistiques robustes, un soin particulier a été apporté pour assembler, pour chaque gène de l'étude, un nombre équivalent d'EST de qualité (au moins 90% d'identité avec la séquence de référence et une longueur supérieure à 100 nucléotides). Dans ce contexte, l'hétérogénéité des EST plus importante en Cancer n'est pas un artefact provenant de l'expérimentation ou de l'échantillon.

Pour expliquer ce phénomène, l'hypothèse d'une augmentation de l'IT en Cancer a été mise en avant, car le nombre de modifications vues sur les EST est 3 à 4 fois plus grand que le nombre de mutations somatiques observés habituellement lors de l'analyse des séquences génomiques provenant de cancers du sein et du colon. Cette étude a également démontré qu'il y a plus souvent des substitutions observées pour les nucléotides A et T que pour G et C, et que lorsqu'il y a substitution, la base de remplacement est dans 73% des cas identique à la base directement en amont ou en aval. L'ensemble de ces observations implique que le phénomène de substitution n'est pas aléatoire, mais qu'il doit être lié à un, ou à des, mécanisme lié à la transcription.

6 A la recherche de marqueurs de cancers

Le fait qu'il existe un mécanisme augmentant l'hétérogénéité des ARNm en Cancers, ouvre de nouvelles perspectives. Prédire des ARNm modifiés qui se retrouveraient amplifiés dans les cellules cancéreuses et en déduire les protéines correspondantes, permettraient de mettre au point des marqueurs pour la détection des cancers et, éventuellement, pour suivre leur évolution. Pour cela, il faut réussir à mettre au point des méthodes rapides et robustes pour identifier ces ARNm et comprendre les mécanismes misent en jeu.

Pour aborder ces mécanismes, nous avons essayé de définir des règles à partir des observations des ARNm provenant d'expériences de transcriptomique. Nous avons mis en place d'un côté, une étude sur l'hétérogénéité des ARNm en tissus sains et cancéreux au travers des expériences de SAGE et de l'autre, une étude des liaisons qui pouvaient exister entre deux modifications consécutives sur un ARNm. Cette dernière étude permettant, à long

terme, de prédire les types et séquences d'ARNm sensibles et/ou les protéines caractéristiques d'un type de cancer.

Matériel et Méthodes

L'analyse de données dans l'univers du haut-débit nécessite des ressources informatiques appropriées. Le traitement massif de plusieurs milliers de séquences réclame de grandes capacités de stockage et de calcul ainsi que des banques de données spécialisées.

Dans ce chapitre, seront présentées tout d'abord, les ressources informatiques disponibles au laboratoire, puis les différentes banques de données biologiques utilisées pendant ce travail de thèse et enfin, les outils disponibles pour les interroger. Suivra une présentation des techniques de programmation choisies ainsi que les outils bioinformatiques. Cette partie se conclura par la description des méthodes bioinformatiques qui ont été utilisées pour récupérer en haut débit les données mettant en évidence les modifications observées dans les ARNm, ainsi que les techniques statistiques qui ont servi à valider ce travail.

7 La plate-forme de bioinformatique de Strasbourg : BIPS

Le BIPS (*BioInformatics Platform of Strasbourg)* est la plate-forme à haut-débit pour la génomique comparative et structurale (http://bips.u-strasbg.fr). Elle a pour mission de maintenir les ressources bioinformatiques de l'IGBMC, aussi bien au niveau des logiciels qu'au niveau des banques de données biologiques. La plate-forme met également un point d'honneur à la formation des biologistes dans l'utilisation des logiciels de bioinformatique et à la transmission des connaissances aux autres bioinformaticiens. Enfin, elle met à disposition son expertise pour les différents projets scientifiques, aussi bien pour les équipes de l'IGBMC que pour de grands projets internationaux. C'est donc naturellement que la plate-forme de bioinformatique est en constante interaction avec le LBGI, pour participer aux projets de recherche, apportant d'une part, sa connaissance des outils existants et d'autre part, en valorisant les outils développés au sein du laboratoire.

Cette plate-forme appartient au Réseau National des plates-formes Bioinformatique (ReNaBi). Elle a été identifiée plate-forme nationale RIO (Réunion Inter-Organismes) en 2003 et fait partie du Génopôle Grand-Est « du Gène au Médicament ». En 2005, elle est devenue la première plate-forme française de bioinformatique à être certifiée norme ISO9001:2000.

Les ressources mises en place par BIPS utilisent un puissant serveur de calculs : le serveur « Star » qui est une grappe composée de huit nœuds (« star1 » à « star8 ») d'architecture x64 Optéron™. C'est-à-dire, huit serveurs indépendants interconnectés, permettant l'accès aux mêmes logiciels et aux mêmes données. Les calculs qui prendraient plusieurs jours sur un ordinateur personnel, peuvent grandement être accélérés en les parallélisant sur les différents nœuds de ce serveur. Deux nœuds sont des *Sun Fire V40z* quadri-processeurs dotés de 32 Go de mémoire utilisant le système d'exploitation Solaris 10. Ces 2 nœuds sont responsables de la gestion de la grappe et du partage des disques de stockage via le protocole NFS (Network File System). Les Stars 3 à 8, exploités pour les calculs, sont sous *Red Hat Entreprise Linux ES 5* et disposent de 16 Go de mémoire au minimum (

Tableau 2). Le serveur de disques connecté à la grappe est un *Sun Fire V480* sous *Solaris 9* et dispose d'une capacité de 12 To de stockage RAID5 (*Redundant Array of Independent Disks*).

Tableau 2 : Spécification des serveurs constituant la grappe « Star ».

<table>
<tr><th>Serveur Star</th><th>Type de serveur</th><th>Mémoire (Go)</th><th>Architecture</th><th>Système d'exploitation</th><th>Utilisation</th></tr>
<tr><td>1</td><td rowspan="5">Sun Fire V40z</td><td rowspan="2">32</td><td rowspan="4">quadri-processeurs</td><td>Solaris 10</td><td>Gestion de la grappe et partage des disques</td></tr>
<tr><td>2</td><td rowspan="4">Red Hat Enterprise Linux ES 5</td><td rowspan="6">Calculs intensifs</td></tr>
<tr><td>3</td><td rowspan="3">16</td></tr>
<tr><td>4</td></tr>
<tr><td>5</td><td rowspan="2">bi-proces. quadri-cœurs</td></tr>
<tr><td>6</td><td>Supermicro H8DMU+</td><td>32</td></tr>
<tr><td>7</td><td rowspan="2">Sun Fire X4100 M2</td><td rowspan="2">16</td><td rowspan="2">bi-process. bi-cœurs</td></tr>
<tr><td>8</td></tr>
</table>

7.1 Alnitak, l'étoile du LBGI

Pour héberger les différents sites web du LBGI, le laboratoire à mis en place un serveur dédié, le serveur Alnitak (http://alnitak.igbmc.fr). Il s'agit d'un *Sun Fire X4200 M2*, biprocesseurs bi-cœurs, doté de 8 Go de mémoire fonctionnant sous système d'exploitation *Ubuntu 8.04 LTS*. Alnitak accède directement par protocole iSCSI (*Internet Small Computer*

System Interface) à un serveur de disques dédié disposant de 2 To de stockage RAID1. Alnitak abrite les différents sites web du laboratoire et les différentes bases de données relationnelles du LBGI utilisées par les sites.

7.2 Banques de données biologiques

7.2.1 Genbank

GenBank (Kim D. Pruitt, Tatusova, Klimke, & Donna R. Maglott, 2009) a été mise en place en 1979 au LANL (Los Alamos National Laboratory), à Los Alamos (Etats-Unis). Depuis 1992, elle est maintenue au NCBI (National Center for Biotechnology Information), à Bethesda (États-Unis). Elle fait partie de l'International Nucleotide Sequence Database (http://www.insdc.org) qui regroupe également l'EMBL Nucleotide Sequence Database (Cochrane et al., 2009) et la DDBJ (*DNA Data Bank of Japan*)(Tateno & Gojobori, 1997). Ces 3 banques constituent les plus grands centres de dépôt et de consultation de séquences nucléotidiques du monde. Le libre accès à la soumission de séquence a permis une augmentation quasi-exponentielle du nombre de séquences disponibles (Figure 22). La collaboration accrue entre les 3 banques se traduit par des échanges de séquences quotidiens assurant une synchronisation entre les 3 centres de dépôts.

Cependant, il en résulte une grande redondance de séquences dans la banque. La banque de données RefSeq (K D Pruitt & D R Maglott, 2001; Kim D. Pruitt, Tatusova, & Donna R. Maglott, 2007) a été initiée depuis 2000 par le NCBI pour éliminer cette redondance et fournir les séquences de meilleures qualités possibles c'est-à-dire, complètes et sans erreur. Genbank est accessible via le site du NCBI (http://www.ncbi.nlm.nih.gov/) et elle est également mise à disposition par BIPS.

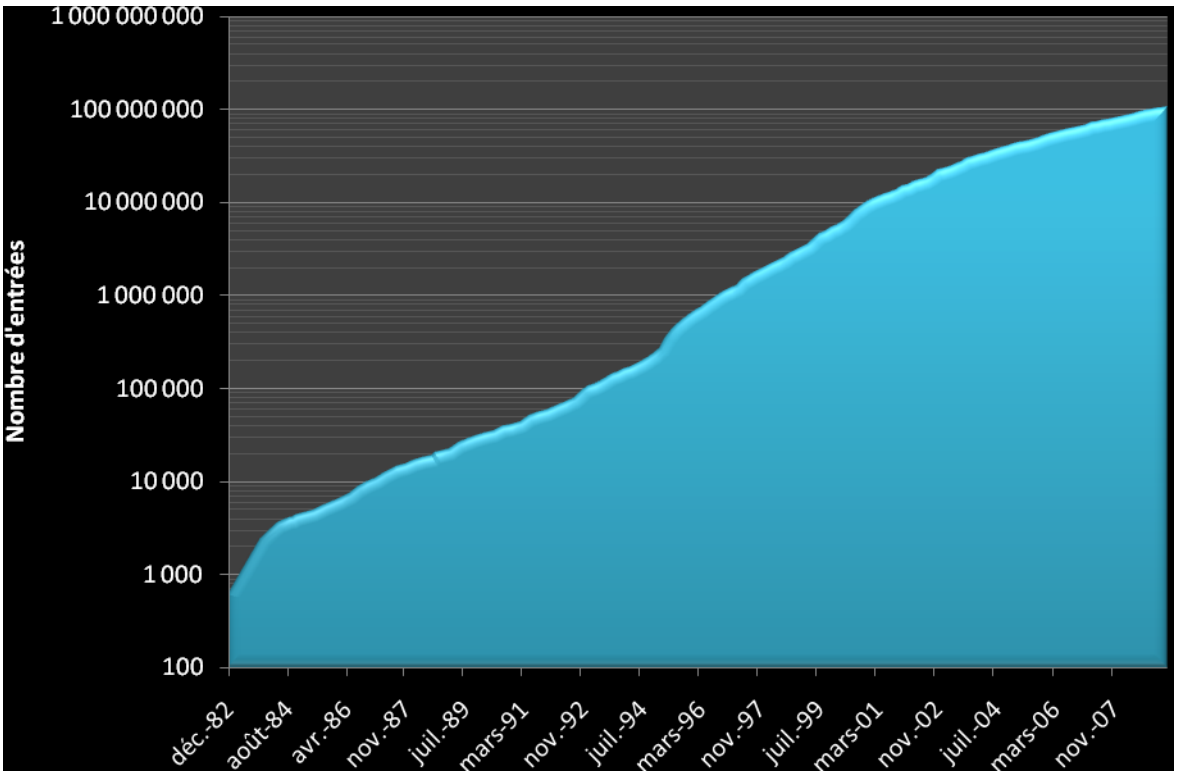


Figure 22 : Évolution du nombre d'entrées dans GenBank de Décembre 1982 à Juin 2009. Données extraites de ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (GenBank version 172).

7.2.2 Gene Expression Omnibus (GEO)

L'étude des niveaux d'expression des gènes par des méthodes à haut débit a été particulièrement utilisée ces dernières années. Les techniques comme les puces à hybridation ou le SAGE (Serial Analysis of Gene Expression) peuvent mesurer en une seule fois l'expression de dizaines de milliers de gènes. Dans les années 2000, a été initiée la banque publique Gene Expression Omnibus (GEO) hébergée au NCBI (Barrett et al. 2009). Elle contient les données des expériences à haut débit de transcriptomique mises à disposition par la communauté scientifique. Nous pouvons ainsi trouver entre autres, les expériences de puces à ARN, de puces à ADN ou d'expériences SAGE. Avec l'avènement des séquenceurs à haut débit de seconde génération, comme le SOLiD ou l'Illumina Solexa, de nouvelles expériences de transcriptomique basées sur le séquençage massif d'ADNc font maintenant leur apparition et sont également répertoriées dans GEO.

L'architecture de la base de données GEO, a été optimisée pour l'enregistrement, le stockage et la recherche des jeux hétérogènes de données à haut débit disponibles. La structure se veut

assez flexible pour s'adapter à l'évolution des technologies. Chaque expérience est référencée dans une fiche GSM (pour *Geo SaMple*) décrivant la source biologique, le protocole expérimental, le sujet et les données avec chaque mesure. Les fichiers GSM sont classés de deux façons :

- GSE (pour *Geo SErie*) définissant le jeu d'échantillons considérés dans une étude et décrivant l'objectif d'une étude et son protocole qui illustre le contexte expérimental aboutissant au GSM.

- GPL (pour *Geo Platform*) qui spécifie un type précis d'expérience à haut débit sur un organisme donnée. Nous utiliserons dans ce travail GPL4, la plate-forme regroupant les SAGE réalisés avec les enzymes Fok1 (TE de 10 nucléotides) et Nla3 sur l'humain.

En septembre 2009, GEO contenait 6 343 plate-formes GPL différentes, concernant plus d'une centaine d'organismes et 13 467 séries d'expériences regroupant 345 454 résultats d'expériences d'expression. Cette masse de données peut être explorée efficacement, questionnée et visualisée en utilisant le site web (http://www.ncbi.nlm.nih.gov/geo/) (Figure 23).

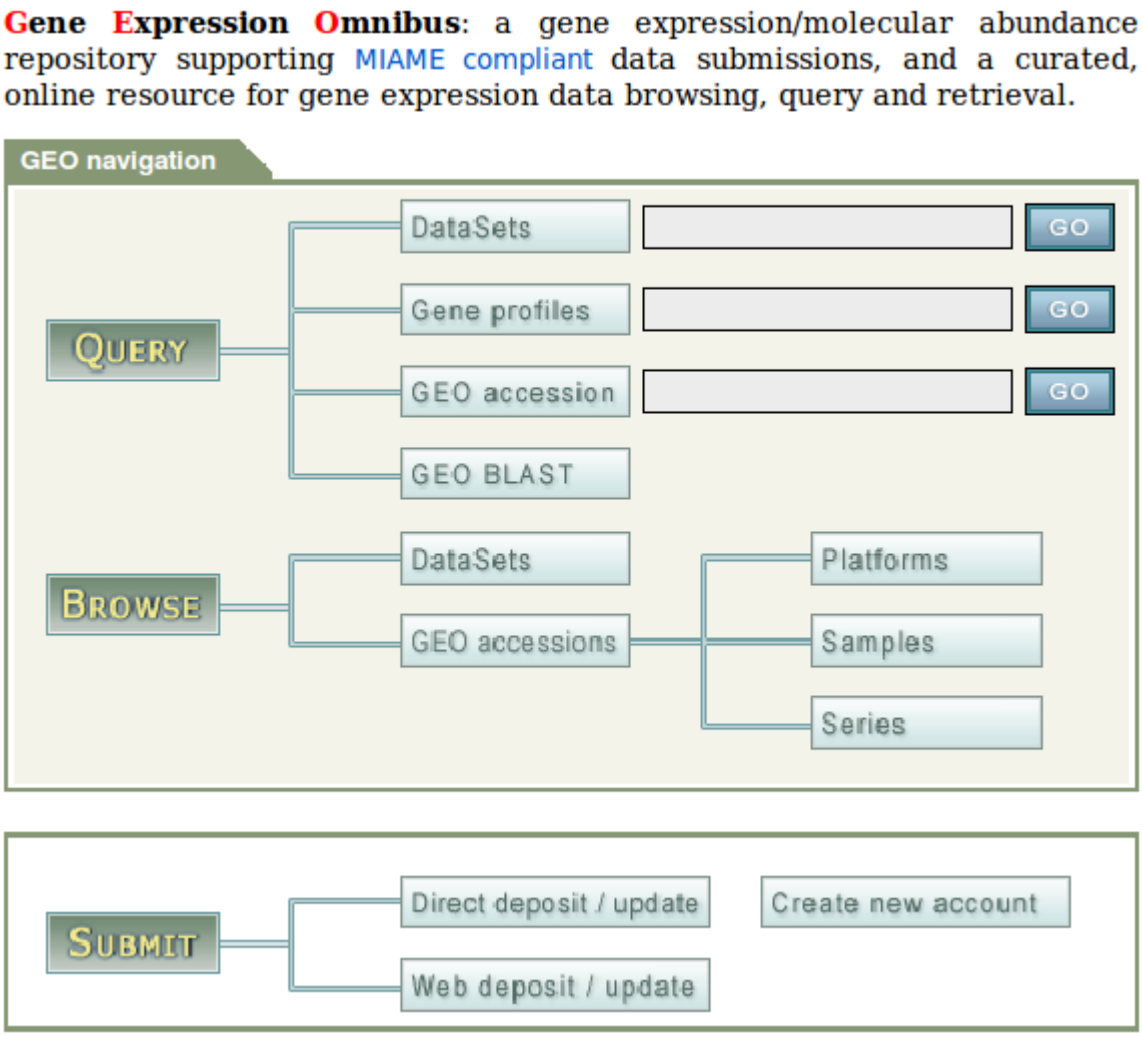


Figure 23 : Capture d'écran du site de GEO. (http://www.ncbi.nlm.nih.gov/geo/). A partir de l'interface web, il est possible de rechercher une expérience de transcriptomique soit directement, depuis les formulaires de recherche (Query), soit, par navigation (Browse). Le site propose aussi un outil de soumission des résultats des expériences d'un utilisateur.

7.3 Interrogation des banques

7.3.1 BLAST

Les séquences nucléiques des ARN étudiés durant ma thèse ont été recherchées dans les banques grâce à la suite logicielle BLAST (Altschul et al., 1997) fournie par le NCBI (D. L. Wheeler et al., 2007) et dédiée à la recherche de séquences similaires. BLAST fournit des alignements locaux deux à deux des régions les mieux conservées entre la séquence requête

et celles des banques BLAST. Pour chaque alignement, nous obtenons un score et une *E-value* (ou *expect*) qui représente la significativité de l'alignement par rapport à des alignements de même longueur générés de façon aléatoire. Plus la *E-value* est proche de 0 et plus l'alignement est statistiquement significatif. En pratique, on considère qu'une séquence associée à une *E-value* inférieure à 0,001 présente un lien significatif avec la séquence requête.

7.3.2 SRS

SRS pour *Sequence Retrivial System* (Etzold, Ulyanov, & Argos, 1996) est un système d'interrogation qui indexe les banques biologiques pour pouvoir faire des requêtes rapides sur les clés indexées. SRS a été développé à l'EMBL puis, par LION BioScience et maintenant, par Biowisdom (http://www.biowisdom.com/).

SRS est installé et géré par le BIPS dans sa version 8.3. Il est interrogeable via l'interface web (Figure 24) et en ligne de commande. L'interrogation en ligne de commande peut s'effectuer par le programme « getz », par un script « ICARUS » (*Interpreter of Commands And RecUrsive Syntax*) ou *via* la librairie Java « *SRS WSObjects* ». L'utilisation d'un script ICARUS ou de la librairie Java permet des requêtes plus complexes et plus rapides.

Figure 24 : Interface web de recherche de SRS 8.3 sur le serveur BIPS.

7.3.3 BIRD

BIRD (*Biological Integration and Retrieval Data*) (Nguyen H et al., 2008) est un système d'interrogation des banques biologiques basé sur le système IBM WebSphere (http://www.ibm.com/). WebSphere contient, entre autres, le serveur de données IBM DB2 et le serveur de fédération de bases de données IBM WebSphere Federation Server. BIRD permet d'interroger différentes banques de données biologiques, avec différents formats, distantes ou locales, comme si elles n'en faisaient qu'une, virtuelle. A terme, BIRD permettra l'extraction automatique de connaissances dans des bases de données hétérogènes (KDD pour *Knowledge Discovery in Database*). Ces développements sont en cours au laboratoire en utilisant différentes techniques et outils d'extraction de connaissances.

BIRD est interrogeable soit, par son interface web (http://decrypthon.u-strasbg.fr/birdweb/query.do/) soit, directement par son service web, par l'intermédiaire de n'importe quel langage supportant le protocole HTTP (*HyperText Transfert Protocol*). Cette dernière possibilité est particulièrement utile lorsqu'un programme à besoin de créer et traiter de nombreuses requêtes BIRD à la volée. La récupération des données passe par un langage de requête propre à BIRD appelé BIRD-QL (*BIRD Query Language*). Ce langage dérivé du langage d'interrogation de bases de données SQL (*Structured Query Language*) est présenté dans un format proche du format EMBL et permet de formuler facilement des requêtes élaborées en masquant la complexité de l'architecture sous-jacente (Figure 25).

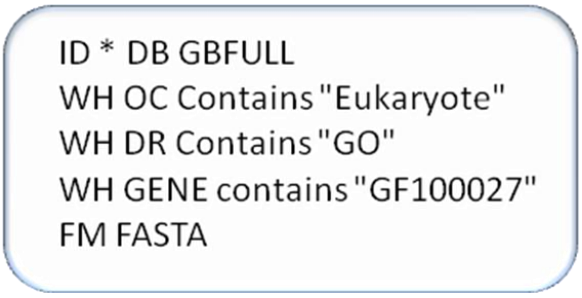


Figure 25 : Exemple de requête BIRD-QL.
Demande afin de récupérer, sous format Fasta, dans tous les enregistrements GENBANK, les séquences eucaryotiques ayant une annotation GeneOntology contenant une référence au gène GF100027.

8 Outils de programmation

8.1 Le langage Java

Java est un langage orienté-objet, c'est-à-dire un langage qui permet de manipuler des structures de données, composées d'attributs et de méthodes. Ces méthodes modélisent des concepts concrets et les interactions qui unissent les différentes structures de données. Sa création a été initiée en 1990 par une équipe de Sun microsystems (http://fr.sun.com/) pour palier aux lacunes du langage C++, également un langage orienté objet, jugé trop complexe pour le développeur. En effet, le langage C++ demande une grande maîtrise pour gérer la mémoire et la gestion de tâches en parallèle.

Java se veut un langage orienté-objet simple à utiliser, robuste et sûr, qui propose une gestion automatique de la mémoire en implémentant un mécanisme de ramasse-miettes (*garbage collector*) et permet aisément de réaliser des applications multitâches. Mais son originalité réside dans son indépendance vis-à-vis de la machine employée pour l'exécution des programmes Java. Cette indépendance est assurée par la machine virtuelle Java (*Java Virtual Machine*, JVM) qui fournit un environnement d'exécution émulé par la machine hôte, tout en conservant des performances proches des langages compilés tels que C++ (Cf. banc d'essai à l'adresse http://shootout.alioth.debian.org/u32q/java.php).

Depuis 1996, date de la première version distribuée de Java, ce langage n'a cessé d'évoluer grâce à l'engouement des développeurs qui ont réalisé de nombreuses librairies additionnelles. Java est particulièrement connu pour sa facilité d'utilisation sur des pages internet avec la technologie embarquée des Applets et son exhaustivité d'outils et de librairies dédiés à la réalisation d'applications client–serveur (spécifications Java EE, *Enterprise Edition*). La version 1.6 de Java a été utilisée sur le serveur de calcul Star.

8.2 Éditeur intégré de développement Netbeans

Netbeans est un environnement de développement intégré mis à disposition, dès 2000, par Sun microsystems pour créer des applications Java. Outre un ensemble complet de modules dédié à la conception d'applications Java, il possède de nombreux modules supplémentaires permettant, par exemple, de créer visuellement des interfaces graphiques ou de programmer à l'aide d'autres langages (comme C ou PHP). Les modules qui ont particulièrement été utilisés dans ce travail de thèse sont : i) l'éditeur XML, pour manipuler des alignements au format MACSIMS (voir chapitre 9.1) ii) l'interface de gestion de versions, qui m'a permis de conserver l'historique de tous les développements informatiques que j'ai réalisé et iii) le *profiler* qui permet de mesurer les performances d'une application en comptabilisant : le nombre d'appels à chaque méthode de l'application considérée, le temps de calcul nécessaire et la quantité de mémoire utilisée.

8.3 R, un projet de calculs statistiques

R (Team, 2008) est un logiciel de statistiques qui implémente le langage de programmation S (Chambers, 1993). Ce langage de programmation a été développé dans les années 70 par AT&T Bell Labs. Le but premier était de proposer un langage suffisamment flexible pour supporter les activités de recherche du département statistique de Bell Labs. S permit aussi, par la suite, de réaliser les tâches répétitives aux moyens de scripts. Il est d'usage de parler du langage R (au lieu de S), lorsque l'on désigne le langage utilisé dans le logiciel R. Ce langage se veut flexible, c'est-à-dire que l'utilisateur peut implémenter ses propres techniques statistiques en complément de celles déjà fournies.

R est un logiciel libre depuis 1995, ce qui permet aux utilisateurs experts dans leur domaine de contribuer à améliorer le logiciel. Ainsi, R possède de nombreuses librairies additionnelles développées par les utilisateurs. Ces librairies sont référencées sur le dépôt du projet R (http://cran.r-project.org/web/packages/). Nous pouvons citer le «*package epicalc* » qui permet des études d'épidémiologie et qui a servi dans la réalisation des graphiques des populations de Tags SAGE dans cette étude, ainsi que le «*package rJava* » qui permet d'utiliser les applications Java directement depuis R. Il est à noter qu'il existe également une librairie Java : JRI (*Java R Interface*), qui permet d'utiliser du code R directement dans un programme Java.

R est aujourd'hui incontournable dans les laboratoires de recherche et de nombreux projets, tel que *RReportGenerator* (Raffelsberger et al., 2008), sont construits autour de ce langage pour faciliter le traitement de données à haut débit.

R est installé en version 2.9.1 sur les machines Star.

9 Outils bioinformatiques

9.1 Format XML MACSIMS

MACSIMS (*Multiple Alignment of Complete Sequences Information Management System*) (J. D. Thompson et al., 2006) est un système de gestion de l'information basé sur l'ontologie de l'alignement multiple, MAO (*Multiple Alignment Ontology*) (J. D. Thompson et al., 2005). MAO est répertorié sur le dépôt d'ontologies biologiques OBO (*Open Biomedical Ontologies*, http://www.obofoundry.org/). MACSIMS permet une annotation automatique des MACS (*Multiple Alignment of Complete Sequences*, alignement multiple de séquences complètes) par l'intégration et l'organisation de différents types de données disponibles dans le cadre d'un alignement multiple. Les informations caractérisant les séquences présentes dans un MACS sont collectées depuis des banques de données publiques ou calculées à partir de programmes propres ou externes. Ainsi, MACIMS annote les alignements avec des informations concernant les domaines structuraux, les données taxonomiques, les sites actifs ou de liaison, les segments transmembranaires....

Le format de stockage de MACSIMS est un fichier XML (*Extensible Markup Language*, http://www.w3.org/XML/) qui reprend la structure définie par MAO. Les spécifications de format auxquelles j'ai participé, sont décrites dans un fichier DTD (*Document Type Definition*), (http://www-bio3d-igbmc.u-strasbg.fr/Spine/public/xml/macsim.dtd, http://www-bio3d-igbmc.u-strasbg.fr/Spine/public/xml/DeKieffer/macsimDocumentation.2.1.html). Ce format permet de stocker toutes les informations concernant les séquences et les alignements de façon structurée afin d'assurer un accès aisé depuis n'importe quel programme. Un exemple de fichier est présenté en Annexe B.

9.2 Sagettarius : Un logiciel d'assignation de Tags SAGE basé sur des banques de Tags virtuels de qualité

Comme nous l'avons vu dans l'Introduction (chapitre 4.2), la technique SAGE, appliquée essentiellement à ce jour chez l'homme et chez la souris, permet d'obtenir rapidement de nombreuses séquences de 10 nucléotides appelées Tags Expérimentaux (TE). Classiquement, pour rechercher le gène d'origine d'une séquence, il est courant de faire une recherche de type BLAST dans une banque de séquences lorsque l'on dispose d'une séquence de longueur suffisante. Mais cette démarche, dans le cas de l'assignation des TE, pose un problème car 10 nucléotides ne suffisent pas à discriminer de façon non-ambigüe l'ARNm et donc le gène, correspondant au TE. Dès lors, nous devons faire appel à ce qui fait la spécificité d'un TE provenant d'une expérience SAGE, à savoir, sa localisation sur le transcrit, c'est-à-dire, une séquence de 10 nucléotides en 3' du site de restriction Nla3 (ou Sau3a) le plus proche de la séquence de poly-adénylation. La stratégie est donc d'utiliser les banques de données de séquences nucléotidiques contenant l'ensemble des séquences d'ARNm d'un organisme (en l'occurrence l'homme ou la souris), afin de prédire l'ensemble des séquences de Tags qui résulteraient d'une expérience SAGE. C'est ce qui est nommé les Tags Virtuels (TV). Il suffit pour cela d'extraire la séquence des 10 nucléotides suivant le site Nla3 (ou Sau3a) le plus proche de la queue poly(A) des ARNm connus (Figure 26) et de rassembler l'ensemble de ces TV dans une banque de données.

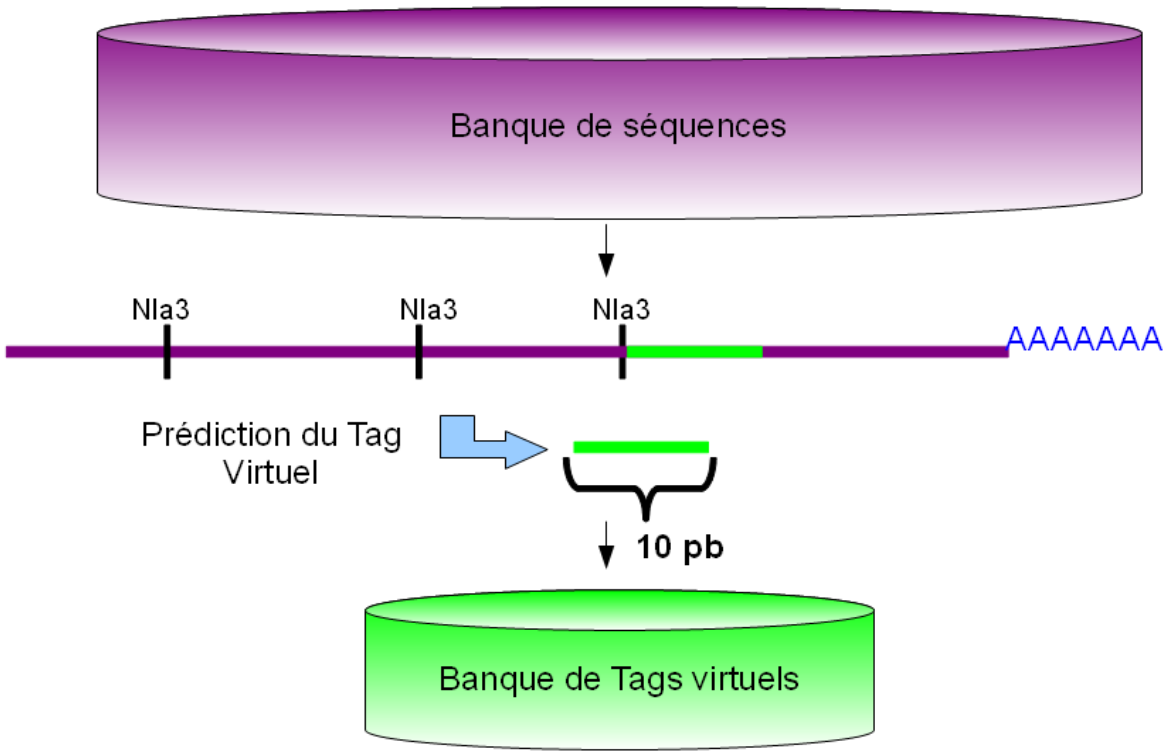


Figure 26 : Construction d'une banque de Tags virtuels depuis les séquences connues.

Dès lors, les logiciels d'assignation des TE (Tags Expérimentaux) comprennent une, ou plusieurs, banques de TV afin de comparer la séquence de chaque TE aux séquences de TV disponibles dans les banques et d'assigner ainsi à chaque TE, son transcrit et son gène. Nous distinguerons ainsi trois types de TE : i) les Tags mono-assignés, assignés à un seul et unique gène, ii) les Tags multi-assignés, assignés à plusieurs gènes et iii) les Tags non-assignés.

Dans ce contexte, le logiciel Sagettarius (Bianchetti, Y. Wu, Guerin, Frederic Plewniak, & Poch, 2007) (http://bips.u-strasbg.fr/Sage_docs/Sagettarius.php) développé au BIPS offre la possibilité d'assigner les TE en utilisant différentes banques de TV construites à partir de séquences dont la qualité a été prise en compte. Ainsi, l'assignation des TE par Sagettarius permet de distinguer les assignations sûres des assignations plus douteuses. Sagettarius distingue quatre banques de TV, allant des séquences d'ADNc de meilleure qualité aux moins valides, qui correspondent : (i) aux séquences de tous les transcrits (variants documentés inclus) résultant des 80 protéines ribosomiques cytoplasmiques, les TRC (Transcrit de protéine Ribosomale Cytoplasmique) et extraits de RefSeq et de publications, (ii) aux séquences d'ADNc documentés, (iii) aux séquences d'expériences HTC (*High Throughput cDNA*), (iv) aux EST. Lors de l'assignation, Sagettarius prend en entrée le tableau du

décompte des TE et assigne ces derniers par étapes (Figure 6), en précisant la banque source. De cette façon, l'utilisateur à une notion de la fiabilité de l'assignation.

Sagettarius privilégie la qualité de l'assignation à la quantité, contrairement aux autres logiciels, tels que SageMap (Lash et al., 2000) qui, pour construire les banques de TV, considére tous les sites Nla3 de l'ensemble des ADNc et EST disponibles dans les banques nucléotidiques. L'approche utilisée dans SageMap aboutit à un faible taux de TE non-assignés, mais entraîne un grand nombre de TE multi-assignés, difficilement exploitables.

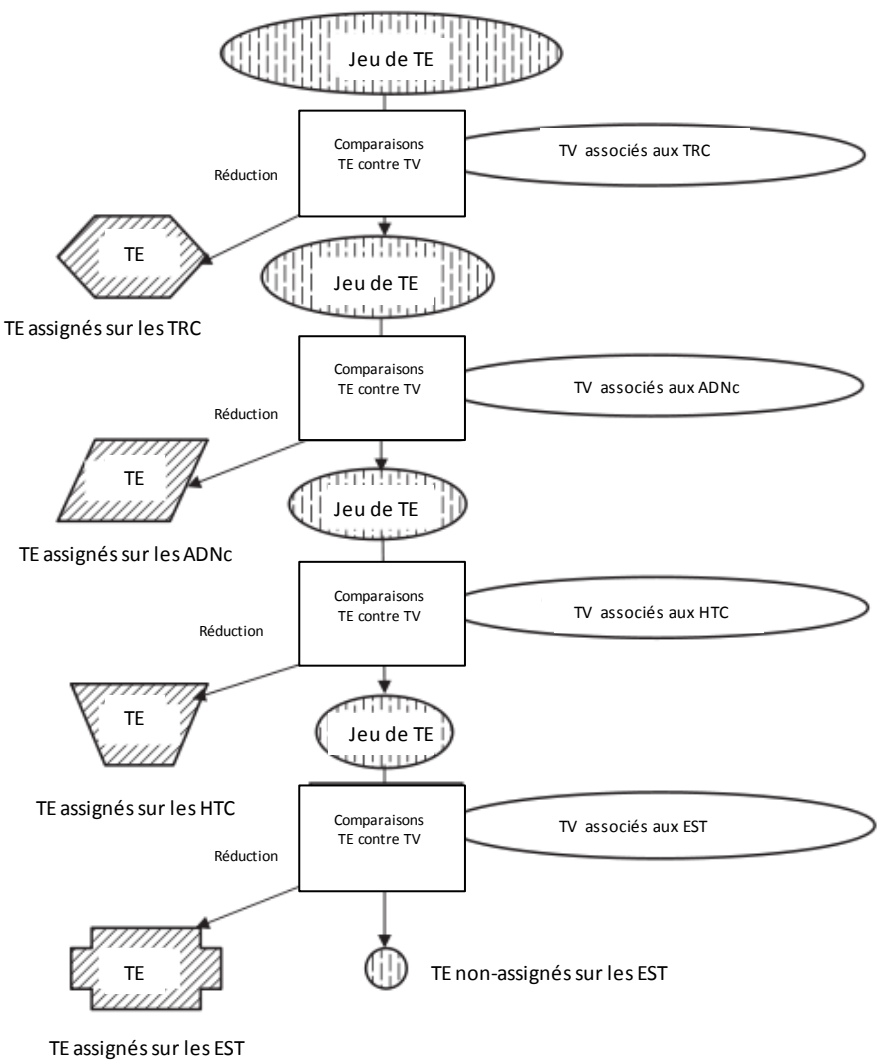


Figure 27 : Protocole d'assignation de Sagettarius.
Les Tags Expérimentaux (TE) sont comparés successivement à chaque banque de Tags Virtuels (TV). L'assignation se fait de la banque la plus sûre qui est celle des Transcrits des protéines Ribosomales Cytoplasmiques (TRC), à la plus bruitée basée sur les séquences d'EST. Le jeu de TE est réduit à chaque étape. Les TE restants sont annotés comme non-assignés.

9.3 DAVID, logiciel d'annotation fonctionnelle

DAVID (*Database for Annotation, Visualization and Integrated Discovery*) (Dennis et al., 2003) regroupe un ensemble d'outils Web destinés à l'annotation fonctionnelle d'ensemble de gènes à l'aide de sa propre banque de données, *DAVID knowledgebase*. Cette dernière intègre les identifiants de gènes ou de protéines de plusieurs espèces, ainsi que leurs annotations, à partir d'une grande variété de banques de données publiques.

DAVID analyse des listes de gènes fournies par l'utilisateur et est disponible à l'adresse http://david.abcc.ncifcrf.gov/. Il comprend l'outil de conversion d'identifiants, l'outil de classification fonctionnelle de gènes (regroupement de gènes ayant une annotation fonctionnelle similaire) et l'outil d'annotation fonctionnelle (analyse de l'enrichissement en catégories fonctionnelles, cartographie sur les voies métaboliques, résumé d'annotation sous forme de graphiques…).

Nous nous sommes servis de DAVID pour identifier nos gènes d'intérêt à partir des numéros d'accés Genbank puis, pour analyser les enrichissements fonctionnels.

10 Méthodes d'étude des modifications de séquences liées au cancer

Ce chapitre décrit les concepts et les méthodes utilisés pour définir et extraire les données utilisées durant ma thèse pour étudier les modifications (mutations, insertions ou délétions) des séquences de transcrits liées aux cancers.

10.1 Etude des modifications aux travers des données provenant d'expériences SAGE

La Figure 28 présente les modifications de séquences de transcrits qui peuvent influencer les résultats d'un SAGE. Trois types de modifications ont été distinguées :

1) La modification intervient dans les 10 nucléotides situés immédiatement en aval du site Nla3 entraînant la création d'un Tag Expérimental (TE) dont la séquence est différente du Tag Virtuel lui correspondant dans la banque. Le TE produit par une telle modification peut être, soit comptabilisé comme un nouveau TE non-assigné, soit être ajouté au décompte des TE provenant d'un autre transcrit.

2) La modification introduit un site Nla3 en 3' du TE théorique. Le TE produit peut être, soit comptabilisé comme un nouveau TE non-assigné, soit être ajouté au décompte des TE d'un autre transcrit.

3) La modification intervient dans l'un des 4 nucléotides du site Nla3. Cet événement entraîne la création d'un nouveau TE correspondant aux 10 nucléotides situés en 3' du premier site Nla3 localisés en amont du site modifié. Là encore, le TE produit peut être, soit comptabilisé comme un TE non-assigné, soit être ajouté au décompte des TE d'un autre transcrit.

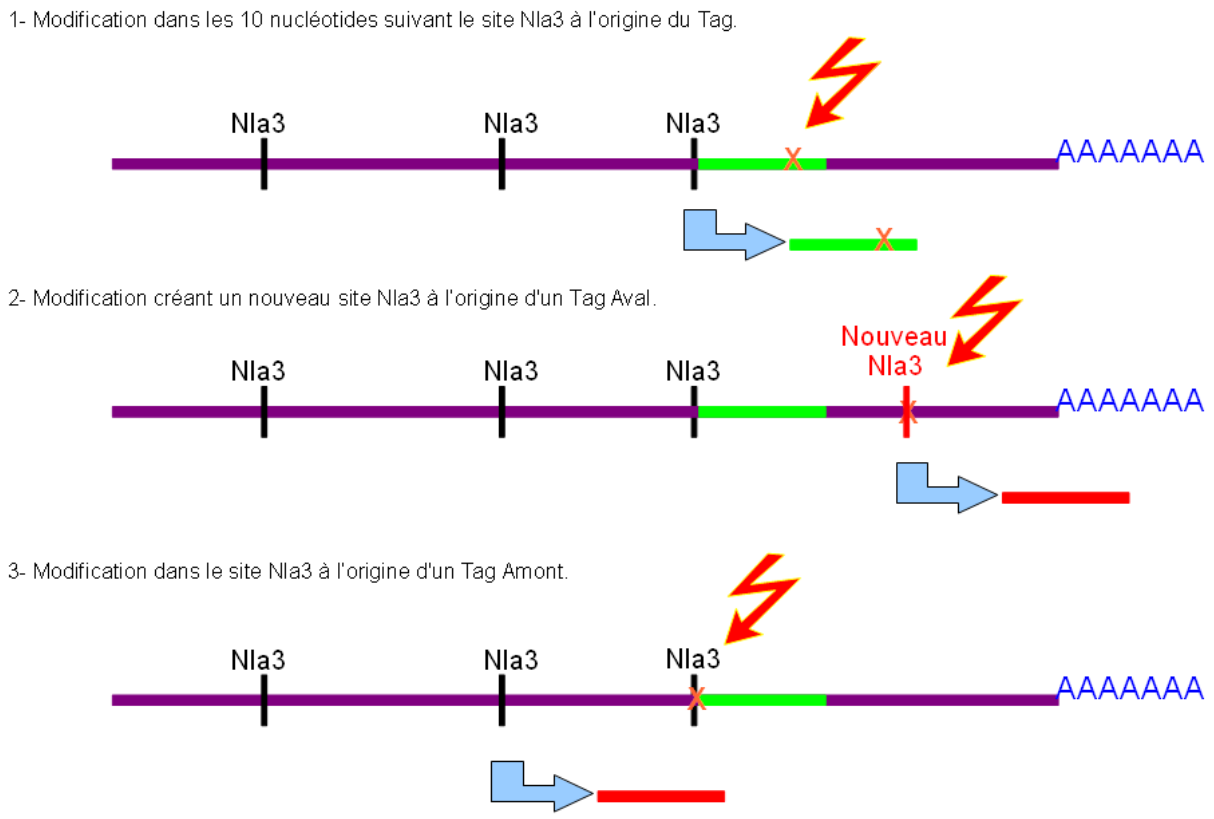


Figure 28 Modifications de la séquence d'un transcrit influençant les résultats obtenus par la méthode SAGE.

Dans le cadre de la comparaison des données SAGE provenant de tissus sains et cancéreux, seules les modifications du troisième type ont été considérées. En effet, ces modifications entraînent la création d'un Tag Expérimental situé immédiatement en amont du TE canonique. A la différence des séquences multiples pouvant résulter des modifications de types 1 et 2, les séquences découlant des modifications de type 3 sont uniques et donc, prédictibles de façon non-ambigüe. Dès lors, nous avons créé au sein de Sagettarius, une banque de Tags Virtuels Amont (TVA) correspondant aux séquences de 10 nucléotides situés en 3' des sites Nla3 localisés en amont du site Nla3 canonique. Pour limiter les sources d'erreurs, cette banque de TVA a été prédite uniquement à partir des transcrits de bonne qualité, à savoir les TRC et les ADNc validés et documentés. De plus, pour assurer la qualité de la banque de TVA générés, un protocole de sélection a été développé (Figure 29) afin de définir la liste des transcrits possédant une paire TV/TVA unique. Ce protocole a entraîné l'élimination des transcrits ne possédant pas de site Nla3 en amont ou ne possédant pas de TVA unique, c'est-à-dire, un TVA ne correspondant à aucun TV.

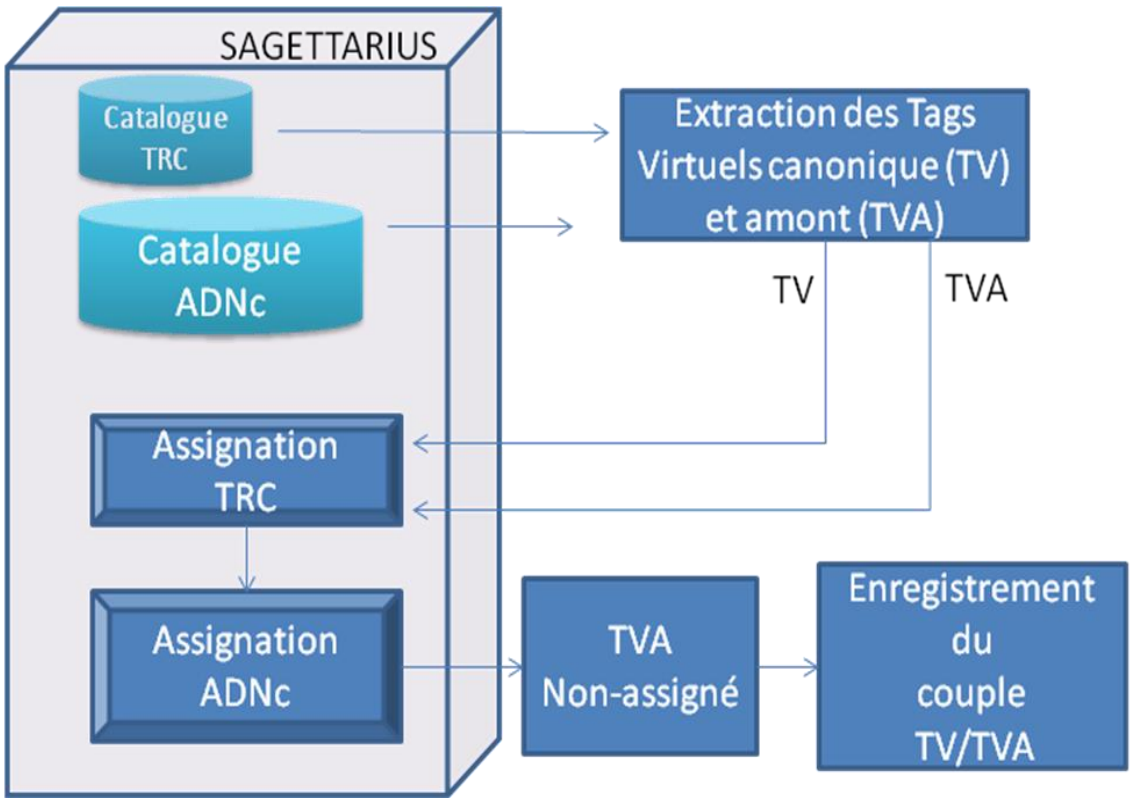


Figure 29 : Protocole d'extraction des couples de Tags Virtuels canoniques (TV) et des Tags Virtuels Amont (TVA) depuis les séquences d'ADNc et des transcrits de protéines ribosomales cytoplasmiques (TRC).

A partir des séquences d'ADNc de Sagettarius, et en respectant la définition stricte des Tags SAGE (tag à partir du site Nla3 le plus en 3'), nous avons pu extraire 6537 séquences possédant un Tag Amont. En éliminant tous les TVA assignés sur la banque de TV d'ADNc de Sagettarius, ce protocole à permis de déterminer 4778 couples TV/TVA uniques.

Pour éviter la redondance de l'information et optimiser la rapidité d'accès, nous avons utilisé deux tables dans notre base de données (Figure 30). Ainsi, notre base de données « SAGE » contient une table : « Sagemapping » où sont enregistrées, les informations sur les transcrits, et une seconde table « virtualtag » où sont enregistrés, les couples TV/TVA. Cette base de données a été intégrée dans le serveur BIRD.

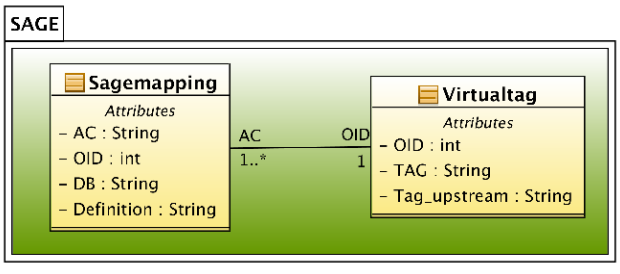


Figure 30 : Structure de la base de données relationnelle des TV et TVA.

Chaque couple de Tags est associé à un unique « OID ». Donc, les couples de Tags sont enregistrés une seule et unique fois. A chaque ADNc, indexé par un unique AC, sera attribué l'OID correspondant. Ceci nous permettra donc de retrouver pour un couple de Tags, tous les ADNc possibles associés au même OID.

10.2 Positions significativement plus modifiées sur les EST

En comparant pour une séquence d'ARNm de qualité, répertoriée dans la base de données RefSeq, les séquences des EST provenant de tissus sains à ceux provenant de tissus cancéreux, nous pouvons déduire la liste de positions sur la séquence de référence, où nous avons significativement plus de modifications en Cancer ou en tissus sains. Pour cela, les EST sont tout d'abord alignés sur la séquence de référence correspondante en utilisant l'algorithme MegaBLAST (Z Zhang, S. Schwartz, L Wagner, & W Miller, 2000) qui est optimisé pour l'alignement de séquences fortement identiques. En effet, nous cherchons des EST de plus de 100 nucléotides ayant plus de 90% d'identité avec la séquence de référence. Nous obtenons donc une série d'alignements deux à deux entre un EST et la séquence de référence. Pour chaque groupe d'EST et pour chaque position de la séquence de référence, le nombre de modifications observées sur les EST est comptabilisé (Figure 31). Les nombres de modifications observés sont comparés à chaque position par un test de χ^2 (voir Chapitre 10.5.5.2) afin de vérifier si le nombre de modifications est significativement influencé par son tissu d'origine. Il faut encore ajouter que chaque type de modifications est traité séparément. Ce protocole est exécuté trois fois en considérant tout d'abord uniquement les substitutions, puis les délétions et enfin, les insertions.

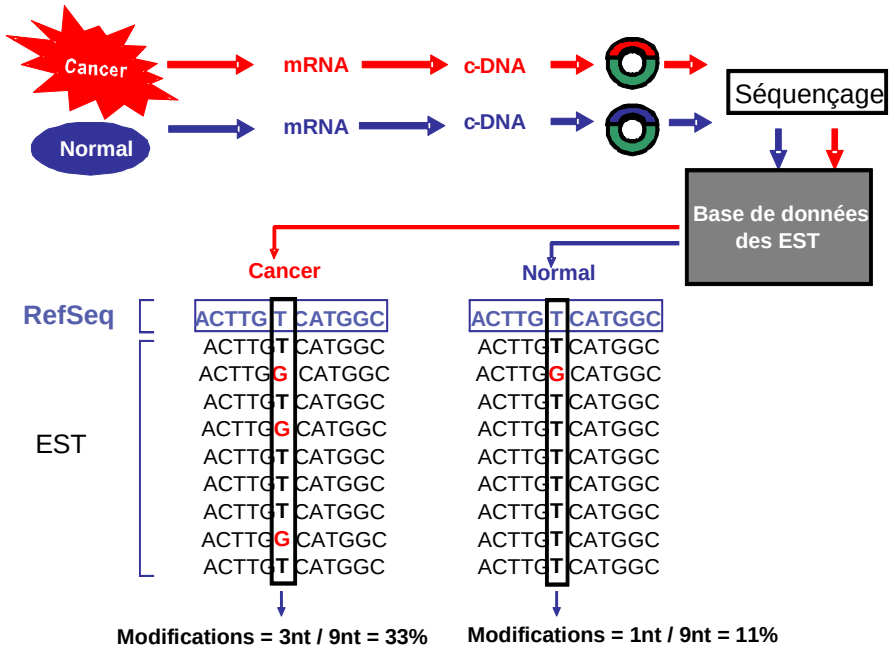


Figure 31 : Décompte du nombre de modifications par position observées sur des EST provenant de tissus cancéreux (Cancer) et de tissus sains (Normal).

Afin de minimiser les sources d'erreurs, des critères très stricts ont été définis pour retenir les EST présentant deux modifications. Les paires de séquences EST-séquence de référence alignées par MegaBLAST sont considérées comme significatives si l'EST fait plus de 100 résidus de long et est aligné avec plus de 90% d'identité. L'EST doit être aligné de façon continue sur la séquence de référence (absence de « fragment » aligné), ceci sur au moins 70% de sa longueur (Figure 32). Comme les EST sont réputés de mauvaise qualité aux extrémités 5' et 3', les bords des alignements (arbitrairement 50 nucléotides) ne sont pas pris en compte.

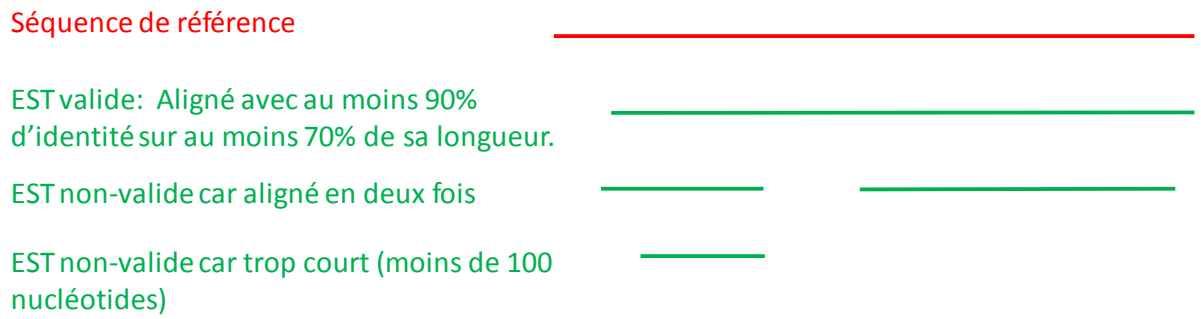


Figure 32 : Sélection des EST à considérer pour comptabiliser le nombre de nucléotides modifiés ou non à une position donnée.

10.3 Méthode d'extraction des positions modifiées en co-occurrence sur les EST Cancer et Normal

Nous considérons trois types de modifications: substitution, insertion et délétion. Pour chaque séquence de référence, il est possible d'avoir trois listes de positions (une pour chaque type de modification) significativement plus fréquentes en Cancer ou en sain. En considérant deux évènements de modifications sur deux positions différentes, nous pourrons avoir 6 listes de positions avec des modifications liées : deux substitutions (sub_sub), deux délétions (gap_gap), deux insertions (ins_ins), une substitution et un gap (sub_gap), une substitution et une insertion (sub_ins) et enfin, une délétion et une insertion (gap_ins).

Pour chacune de ses 6 listes, nous avons décompté, à partir des alignements MegaBLAST des EST, le nombre de fois, pour un couple de positions données, où deux modifications se produisent : sur le même EST, uniquement sur une des positions considérées ou sur aucune des deux positions. Ce décompte nous permettra de calculer si les modifications a ces deux positions sont liées ou non (voir chapitre 10.5.4,

Tableau 6). Ces fichiers de sortie MegaBLAST contiennent un grand nombre d'alignements d'EST (plusieurs milliers). Traiter une masse aussi importante de données nécessite une stratégie efficace de programmation. Pour améliorer les performances d'analyse des fichiers MegaBLAST, une procédure multitâche a été implémentée. Tous les tableaux de contingences pour une séquence de référence sont remplis en parallèle au fur et à mesure que les alignements d'EST sont lus.

En ce qui concerne le traitement des délétions et des insertions dans les régions de faible complexité, nous ne pouvons nous prononcer sur leur localisation exacte car le programme MegaBLAST place les insertions/délétions de façon heuristique. Arbitrairement, tous ces évènements ont été regroupés en 3'de la région de faible complexité (Figure 33).

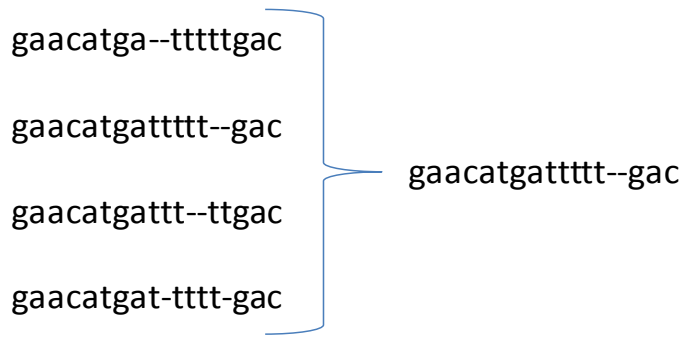
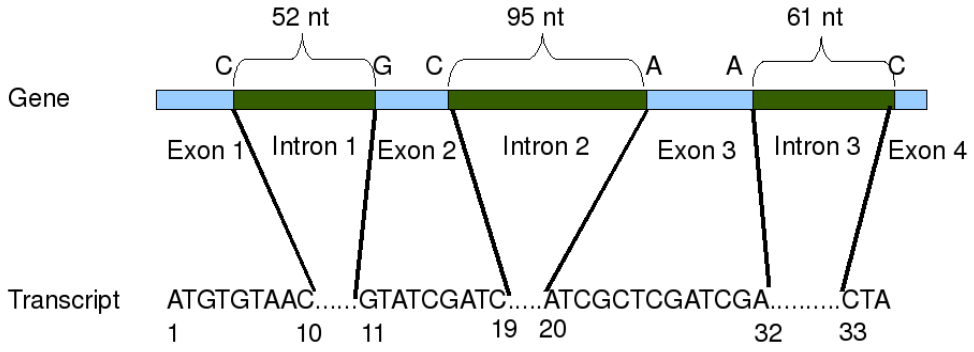


Figure 33 : Déplacement des gaps.
Une insertion/délétion (indel) est indiquée par un gap (-). Lorsqu'une indel (dans la figure un gap) est dans une région de faible complexité (dans la figure, les nucléotides répétés, ttttt), le gap est placé par le programme MegaBLAST de façon heuristique (partie gauche de la figure). Pour homogénéiser le résultat, tous les gaps sont arbitrairement regroupés en 3' de la région de répétition de nucléotides.

10.4 Obtention des distances génomiques

Pour connaître la distance génomique entre deux nucléotides observés sur un ARNm, il faut savoir si ces deux positions sont séparées par aucun, un ou plusieurs introns, ainsi que la longueur de ces introns (Figure 34).



Output

N° intron	position	length
1	10	52
2	19	95
3	32	61

Figure 34 : Définition des introns sur le transcrit.
Un intron est caractérisé, par son numéro, la position du dernier nucléotide de l'exon qui le précède et par le nombre de nucléotides qui le compose.

Pour chaque séquence d'ARNm de référence, les coordonnées chromosomiques des exons ont été extraites des fiches chromosomiques de référence de Genbank. Ces positions nous permettent de calculer les positions des introns sur le transcrit ainsi que leur longueur (Équation 1).

$$[position\_intron_1] = [fin\_exon_1] - [début\_exon_1] + 1$$
$$[position\_intron_n] = [fin\_exon_n] - [début\_exon_n] + 1 + [position\_intron_{n-1}]$$
$$[taille\_intron_n] = [début\_exon_{n+1}] - [fin\_exon_n] - 1$$

Équation 1 : Calcul des positions des introns sur le transcrit ainsi que leur taille à partir des positions génomiques des exons.

10.5 Méthodes statistiques utilisées

Après avoir présenté les méthodes de bioinformatique qui ont permis de collecter les données, ce chapitre décrira les méthodes statistiques qui ont été utilisées pour valider les

observations. Tout d'abord, une rapide introduction aux méthodes statistiques initiera le néophyte au vocabulaire nécessaire à la compréhension de ces méthodes, puis chaque technique statistique employée dans ce travail sera décrite en précisant dans quel contexte. Tous les tests utilisés sont disponibles dans la librairie de base du logiciel R.

10.5.1 Vocabulaire minimum à connaître en statistique

- **La population** : La population est le groupe réel que l'on désire étudier. Par exemple : tous les hommes entre 30 et 40 ans vivant en France.

- **L'échantillon** : C'est l'ensemble des individus réellement observés qui va nous permettre de généraliser des conclusions à la population globale. Une des grandes difficultés en statistique est de définir si l'échantillon étudié est représentatif de la population afin d'éviter de faire, à partir de quelques exceptions, une fausse généralité. Ceci est surtout le travail des experts du domaine auquel les statistiques sont appliquées, experts qui connaissent intimement les données et sont responsables de la création de l'échantillon.

- **Une variable qualitative** : C'est un groupe dans lequel on classe les individus de l'échantillon. Par exemple le sexe (homme, femme) ou encore la couleur des cheveux (blond, brun…).

- **Une variable quantitative** : C'est une mesure faite sur un individu.

- **Une variable quantitative discrète** : C'est une mesure qui ne peut prendre que certaines valeurs. Par exemple, le nombre de dents d'un individu ne peut être qu'un entier compris entre 0 et 32.

- **Une variable quantitative continue** : C'est une mesure qui ne peut prendre que des valeurs décimales, par exemple la taille d'un individu. Le nombre de mesures possibles est considéré comme infini, bien qu'en réalité il soit limité par la précision de l'appareil.

- **Un modèle** : Un ensemble d'équations mathématiques qui décrivent un objet en termes de variables aléatoires. Tous les tests statistiques sont basés sur un modèle permettant de décrire une population de façon mathématique.

- **Une statistique** : Une fonction de variables aléatoires. Ces variables sont typiquement des observations sur un échantillon.

- **Un test statistique** : C'est une démarche consistant à confronter deux hypothèses : H0, l'hypothèse nulle, et H1, l'hypothèse alternative. L'hypothèse H0, correspond à un modèle pour lequel on sait qu'une statistique S suit une certaine loi. Cette variable aléatoire S peut être observée ou calculée à partir d'observations, on obtient ainsi une valeur x qui sera confrontée à la loi de S. Si x est « conforme » à la loi de S, l'hypothèse H0 est acceptée, sinon elle est rejetée.

- **p-valeur** : La probabilité d'avoir |S|>x (cf. test statistique)

- **Le risque de première espèce α :** C'est la probabilité de rejeter à tort l'hypothèse nulle lorsqu'elle est vraie. Classiquement, on fixe cette valeur α à 5%. Si la p-valeur du test est inférieure au risque de première espèce, alors on dira que le test est significatif en faveur de l'hypothèse H1.

- **Le risque de seconde espèce β** : C'est la possibilité d'accepter à tort l'hypothèse nulle lorsqu'elle est fausse. La valeur 1-β est appelée la « puissance du test » (Tableau 3).

Tableau 3 : Les risques d'un test statistique.

Décision \Vérité	H0	H1
H0	1-α	B
H1	α	1 – β

10.5.2 Statistiques descriptives

La première démarche statistique consiste à caractériser un échantillon en décrivant ses propriétés. Voici la liste des propriétés les plus couramment utilisées :

- **Le minimum et le maximum** : ils nous informent sur les extrêmes de l'échantillon.

- **Quartile** : Le premier, second et troisième quartiles correspondent respectivement à 25%, 50%, 75% des valeurs observées par ordre croissant.

- **La médiane :** C'est la valeur du second quartile.

- **La moyenne** : Elle donne le point d'équilibre des valeurs.

Si la moyenne et la médiane sont proches, alors on a une population équilibrée. Le nombre d'individus étant supérieur à la moyenne compense de façon équilibrée ceux qui lui sont inférieurs.

Il est également informatif de calculer la variance qui illustre la variabilité des données. En fait, sa racine carrée positive, c'est-à-dire l'écart-type, est plus souvent utilisée car elle est plus facile à interpréter. Ainsi, dans une population où les individus seront regroupés autour de la moyenne, l'écart type sera faible. Au contraire, si la disparité est importante, l'écart type sera élevé.

La "boite à moustaches" est une façon pratique de représenter les statistiques descriptives (Figure 35).

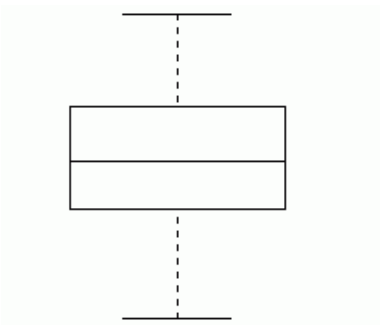


Figure 35 : Exemple de "boite à moustaches".
Le rectangle contient le deuxième et troisième quartile et est coupé par la médiane. Les "moustaches" vont du minimum au maximum du premier quartile et du minimum du quatrième quartile au maximum.

10.5.3 Quelques lois de probabilité

Les statistiques descriptives constituent la première démarche de l'étude d'un échantillon. La seconde est d'essayer de modéliser la population, soit pour étudier certaines propriétés à travers des tests statistiques, soit pour permettre de faire des prédictions sur son évolution. Les lois de probabilités sont des outils de base pour construire les modèles statistiques. Nous présenterons dans ce chapitre, quelques unes de ces lois.

10.5.3.1 La Loi binomiale

La loi binomiale est la loi de probabilité qui décrit une suite d'essais ne pouvant avoir que deux valeurs (réussite ou échec). Elle a été utilisée dans ce travail de thèse comme base pour créer un test sur les Tags SAGE en considérant cette expérience comme une suite de tirages où l'on recherche un Tag particulier (Figure 36).

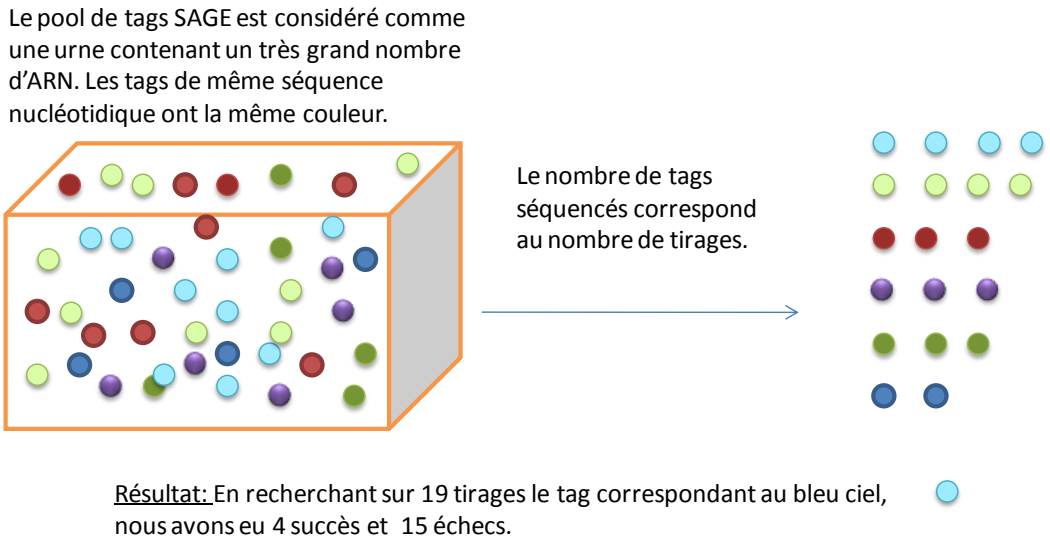


Figure 36 : Analogie d'une expérience SAGE avec une loi Binomiale.
Le nombre d'ARN dans la cellule est assez grand pour que le tirage soit considéré avec remise.

La loi binomiale de paramètres *n* et *p* décrit la probabilité d'avoir au bout de *n* tirages aléatoires avec remise, un nombre de succès *k*, avec *p*, la probabilité que le tirage soit un succès et q=1-p la probabilité que le tirage soit un échec.

$$p(k) = P(X = k) = \binom{n}{k} p^k q^{n-k}$$

Avec

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

10.5.3.2 La Loi normale

La loi normale est la loi de probabilité la plus souvent utilisée pour créer les modèles statistiques. Lors de l'étude de phénomènes naturels, cette distribution est très souvent retrouvée. Ainsi, lorsque l'on étudie des échantillons, il est d'usage de vérifier si les données suivent une loi normale avec un test, appelé test de normalité. Si c'est le cas, nous avons à notre disposition toute une batterie de tests statistiques pour réaliser notre étude. La densité de probabilité de la loi normale dessine une courbe dite courbe en cloche ou courbe de Gauss ou encore une Gaussienne (Figure 37).

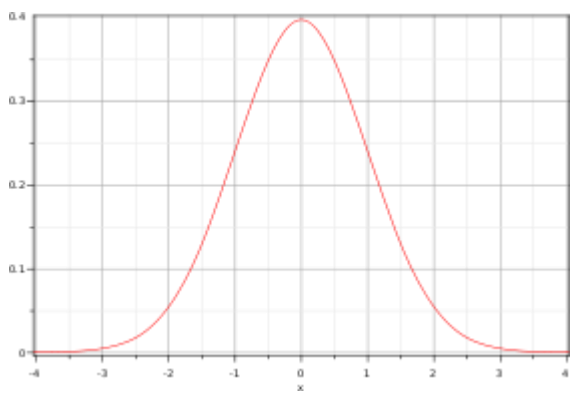


Figure 37 : Densité de probabilité d'une loi normale.
Ici la loi est centrée réduite, c'est-à-dire qu'elle est symétrique par rapport à 0 et que l'écart-type est égal à 1.

10.5.3.3 La Loi hypergéométrique

La Loi hypergéométrique est la loi de probabilité qui décrit un tirage sans remise d'un ensemble d'objets appartenant à seulement deux catégories possibles. Nous considérons une urne contenant un nombre « A » de boules. Parmi ces boules, nous en avons « pA » boules gagnantes et «1-pA » qui sont perdantes. Nous tirons simultanément « n » boules et notons « K » le nombre de boules gagnantes. La probabilité d'obtenir K boules gagnantes est donnée par la formule :

$$p(k) = \frac{C_{pA}^{k} C_{1\text{-}pA}^{n-k}}{C_{A}^{n}}$$

Elle a été utilisée dans ce travail de thèse pour modéliser les substitutions de nucléotides au hasard dans les alignements d'EST afin de mettre en évidence des zones particulièrement riches en modifications par rapport à un modèle théorique. Les constructions des alignements multiples ont été réalisées sur la base des alignements deux à deux fournis par MegaBLAST grâce à la librairie Java JMACS que j'ai développée et qui est décrite au chapitre 12. Nous avons choisi la loi hypergéométrique car il est important, quand nous regardons une fenêtre de l'alignement, que le tirage des nucléotides soit simultané et non avec remise, pour ne considérer qu'une seule fois chaque nucléotide de l'alignement de départ.

Le principe du test est que nous considérons une fenêtre glissante de taille variable de paramètres (c, S) avec « c » la première colonne de la fenêtre, et « S » le nombre de colonnes recouvertes par la fenêtre. A la fin du processus, cette fenêtre recouvre toutes les séquences de l'alignement (Figure 38).

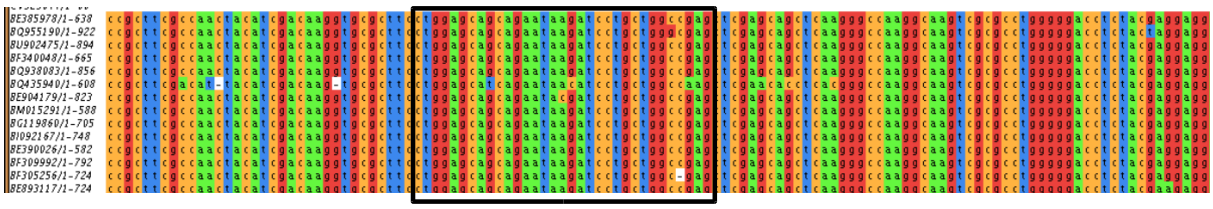


Figure 38 : Une fenêtre glissante sur l'alignement multiple de séquence.

L'alignement contient A nucléotides dont pA correspondent à des substitutions. Nous avons « n » nucléotides observés dans la fenêtre de l'alignement. Parmi ces « n » nucléotides, « X » nucléotides sont modifiés par rapport à la séquence de référence. Nous nous demandons quelle est la chance de trouver au moins « X » nucléotides modifiées dans cette fenêtre par rapport au modèle aléatoire décrit par la loi hypergéométrique. Si cette probabilité est faible, alors c'est que nous avons un enrichissement anormal de modifications. En utilisant la fonction de R pour la loi hypergéométrique, nous avons calculé $p(k \geq X)$, avec k le nombre de nucléotides modifiés attendus pour chaque fenêtres possibles, c'est-à-dire pour chaque colonne « c » et pour chaque taille « S ».

Cette méthode générant un grand nombre de tests, nous risquons d'accumuler beaucoup de faux positifs. Nous avons donc décidé de fixer le taux de faux positif à 1% et sélectionné le seuil α de significativité de la p-valeur en conséquence avec la méthode du FDR (*False Discovery Rate*, voir chapitre 10.5.6).

Nous avons ensuite sélectionné les fenêtres non chevauchantes avec les p-valeurs les plus faibles inférieures au seuil α.

10.5.4 Mesurer la liaison entre deux variables quantitatives

Il est souvent utile lorsque l'on étudie un échantillon, de savoir si deux paramètres observés évoluent en même temps ou non. Ceci peut être approximé par le coefficient de corrélation. Nous avons utilisé le coefficient de corrélation pour montrer l'intensité de la liaison entre le nombre de Tags séquencés en SAGE et le nombre de types de Tags différents rencontrés. Nous présenterons ici deux façons de calculer le coefficient de corrélation.

- Le coefficient de corrélation de Pearson est utilisé lorsque les deux variables sont gaussiennes.

- Le coefficient de corrélation de Spearman est utilisé lorsque les deux variables ne sont pas gaussiennes. Le calcul se base sur les rangs des observations.

Dans les deux cas, ce coefficient a une valeur entre -1 et 1. Plus sa valeur sera éloignée de 0 et plus les deux facteurs seront corrélés. Le signe négatif indique que les facteurs évoluent dans des sens opposés.

Lorsque nous avons une corrélation sur des variables gaussiennes, on peut modéliser cette corrélation par une régression linéaire. Le modèle linéaire implémenté dans le logiciel R permet non seulement de calculer cette droite, mais aussi de calculer un intervalle de confiance, mettant en évidence les observations qui ne semblent pas suivre le modèle. Un exemple de script R pour obtenir ce résultat est fourni en Annexe C.

10.5.5 Comparer 2 échantillons

Une des grandes questions en biologie est : « Mes deux échantillons sont-ils différents ? ». Cette question qui paraît simple est souvent le début d'une longue étude. La première difficulté est de définir mathématiquement « différent » ! Différent par rapport à quels critères ? Le biologiste a souvent bien du mal à préciser sa question, surtout si elle s'inscrit dans une phase de recherche durant laquelle on ne sait *à priori* pas ce qui pourrait changer parmi les nombreuses mesures caractérisant les échantillons. Dans ce cas, le biologiste ne peut poser une question précise sur l'influence d'un facteur et mettre au point un protocole d'analyse robuste pour pouvoir tester la différence sur ce facteur. Dans le cadre des études de bioinformatique sur un ensemble d'expériences publiées dans les banques (telles GEO), il est simplement impossible de trouver des échantillons d'expériences correspondant parfaitement à la question que l'on veut aborder. Le statisticien se voit donc contraint d'effectuer toute une batterie de tests couramment utilisés, sur toutes les variables possibles, laissant le soin au biologiste de jouer à « *Jeopardy* » une fois que le statisticien lui a donné une réponse purement mathématique.

Nous allons ici décrire la suite de tests qui ont permis de trouver quelles différences nous pouvions observer entre les données provenant de tissus sains et cancéreux. Nous décrirons aussi les tests permettant de vérifier que deux évènements sont statistiquement liés.

10.5.5.1.1 Moyennes des deux échantillons similaires

Le test de Student (ou test t) permet de comparer les moyennes de deux échantillons gaussiens. Il permet de savoir si les mesures sur les deux échantillons sont statistiquement différentes ou non. Le test de Student est un test paramétrique, qui fait intervenir les paramètres d'une loi gaussienne. Ceci est directement lié à la construction des modèles qui servent au test. Ainsi, pour pouvoir le réaliser, il faut admettre que les individus des échantillons sont indépendants et que leurs distributions suivent une loi normale. Si le test de normalité n'est pas concluant, on utilisera plutôt un test de Mann-Withney.

10.5.5.1.2 Ordre de grandeur des mesures de deux échantillons similaire

Le test de Mann-Whitney (ou Wilcoxon) est un test de rang, c'est-à-dire qu'il ne va pas comparer directement les valeurs. Ce test à l'avantage d'être non-paramétrique et de pouvoir être appliqué dans tous les cas. Nous avons utilisé ce test pour chercher s'il y avait moins de TRC différents détectés dans les expériences SAGE réalisées sur des tissus cancéreux que sur celles provenant de tissus sains. Il faut savoir que les tests non-paramétriques ont une puissance plus faible.

10.5.5.1.3 Variances des échantillons identiques

Le test de variance ou test F va tester si les variances sont égales. La variance va nous renseigner sur l'hétérogénéité des mesures dans un échantillon. Si le test des variances est positif, alors cela signifie que les deux échantillons ont le même degré d'hétérogénéité. Ce test est paramétrique et ne peut s'appliquer qu'aux mesures suivant une loi normale. Il existe un autre test, le test de Levene qui permet d'avoir un test équivalent, mais de façon non-paramétrique. Nous avons utilisé le test de Levene pour vérifier que l'hétérogénéité de la détection des TRC était plus grande dans les données SAGE provenant des tissus cancéreux que dans celles provenant des tissus sains.

10.5.5.2 Liens entre deux variables qualitatives

Pour tester si deux variables sont liées, le statisticien dispose des tests d'indépendances. L'hypothèse nulle H0 du test est : « Les deux variables sont indépendantes » (d'où le nom du test) contre l'hypothèse alternative H1 : « Les variables sont dépendantes ».
Dans le cadre de l'étude des données SAGE, nous avons testé H0 : « Le nombre de TE non-assignés est indépendant de l'origine tissulaire Normal/Cancer ». Ceci correspond à réaliser le test d'indépendance sur un tableau de contingence (Tableau 4).

Tableau 4: Tableau de contingence sur les variables TE assignés/non-assignés et Normal/Cancer.

Assignation\ Phénotype	Normal	Cancer
TE assignés	Nombre total de TE assignés Expériences SAGE Normal.	Nombre total de TE assignés Expériences SAGE Cancer.
TE non-assignés	Nombre total de TE non-assignés Expériences SAGE Normal.	Nombre total de TE non-assignés Expériences SAGE Cancer.

Si la p-valeur du test est inférieure à 5%, alors nous acceptons H1. Le nombre de TE non-assignés est donc influencé par l'origine du tissu. Nous avons utilisé cette stratégie pour étudier l'effet du cancer sur la détection des Tags Amont (Tableau 5), ainsi que pour vérifier s'il y avait un lien entre deux modifications de nucléotides consécutives sur la base des EST (

Tableau 6).

Pour faire le test, nous demandons un effectif total d'au moins 70, pour ne pas réaliser un test sur un échantillon trop petit pour être représentatif.
Historiquement, le test de dépendance utilisé était celui du χ^2 car il est facile à calculer, mais il s'agit d'un test asymptotique, autrement dit, dont l'exactitude théorique n'est vraie que pour un nombre infini d'observations. Aujourd'hui, avec la disponibilité des ordinateurs, on privilégie le test de Fisher exact, un test d'indépendance fournissant, contrairement au test du χ^2, une p-valeur exacte.

Tableau 5: Tableau de contingence sur l'assignation des TE sur les critères TV/TVA et Normal/Cancer.

assignation\Phénotype	Normal	Cancer
TVA	Nombre total de Tags Amont observés dans les expériences SAGE Normal.	Nombre total de Tags Amont observés dans les expériences SAGE Cancer.
TV	Nombre total de Tags canoniques Observés dans les expériences SAGE Normal	Nombre total de Tags canoniques observés dans les expériences SAGE Cancer

Tableau 6: Tableau de contingence sur les variables : nucléotide modifié ou non à la première position et nucléotide modifié ou non à la seconde position sur la séquence.

Position 1 \ Position 2	Nucléotide modifié	Nucléotide non modifié
Nucléotide modifié	Nombre de séquences ayant un nucléotide modifié à la fois en position 1 et 2.	Nombre de séquences ayant un nucléotide modifié uniquement en position 1.
Nucléotide non modifié	Nombre de séquences ayant un nucléotide modifié uniquement en position 2.	Nombre de séquences n'ayant aucun nucléotide modifié en position 1 et 2.

10.5.6 Estimation du nombre de faux positifs sur des tests multiples

Lorsque nous effectuons une série de tests statistiques, nous accumulons à chaque test un risque de faux positifs dû au risque de première espèce. On peut l'estimer avec le LBE (*Location Based Estimator*) (Dalmasso, Broët, & Moreau, 2005) qui est en réalité une borne supérieure du nombre de faux positifs. Cette fonction est disponible dans R avec la librairie Bio-conductor (http://www.bioconductor.org/).

Pour pouvoir contrôler le taux de faux positifs, nous pouvons utiliser les méthodes de type FDR (*False Discovery Rate*)(Benjamini Y, & Hochberg Y., 1995),(Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001)., pas de date). Ces méthodes FDR permettent de choisir le seuil α de significativité à utiliser pour chaque test en fonction du taux de faux positifs maximum souhaité. Nous avons utilisé le FDR de Benjamini et Hochberg.

Résultats

Dans cette partie, nous présenterons tout d'abord les différents outils bioinformatiques et statistiques développés durant ma thèse dans le but d'une part, de pouvoir comparer les données SAGE et d'autre part, de manipuler les alignements multiples de séquences. Ensuite, nous présenterons les résultats de notre étude comparative portant sur l'hétérogénéité des ARNm provenant de tissus sains et cancéreux par l'utilisation des données SAGE. Nous enchaînerons ensuite sur les études approchant le mécanisme de l'hétérogénéité des ARNm au sein des tissus cancéreux. Ces études ont été effectuées sur la base des EST. Nous avons étudié la liaison possible entre deux modifications consécutives sur un ARNm en analysant plusieurs millions de séquences d'EST. D'autre part, nous avons recherché dans les transcrits issus de tissus cancéreux les régions enrichies en modifications.

11 Un nouveau modèle statistique pour comparer les niveaux d'expression des gènes sur la base des données SAGE

Le SAGE permet de comparer en une seule expérience le niveau d'expression de milliers de gènes (voir chapitre 4.2). Cependant, la résolution des données dépend du nombre de Tags séquencés. Plus il y aura de Tags séquencés et plus le nombre de transcrits de gènes différents sera élevé. Par exemple, il faut séquencer plus de 100 000 Tags pour espérer détecter au moins 70 des 80 transcrits de protéines ribosomales cytoplasmiques (Bianchetti et al., 2007). Cet effet prend une importance particulière lorsque nous désirons comparer plusieurs expériences SAGE qui ont des nombres de Tags séquencés différents.

Dans notre étude, dans un premier temps, nous désirons savoir si un gène est surexprimé dans les expériences SAGE réalisées sur les tissus sains ou, au contraire, dans les SAGE issus de tissus cancéreux. La méthode couramment utilisée est de réaliser un test d'indépendance (voir chapitre 10.5.5.2) entre les deux types de SAGE en se basant sur les décomptes du TE du gène d'intérêt. Cette méthode ne prend pas en compte la variation du nombre de Tags séquencés entre les expériences. C'est pour combler ce manque que nous avons mis au point un nouveau test qui prend en compte ce paramètre. Pour cela, nous avons réalisé une modélisation statistique de nos expériences SAGE. La technique du SAGE peut s'assimiler à une expérience suivant une loi binomiale (voir chapitre 10.5.3.1). Pour savoir si notre gène est plus exprimé dans les SAGE issus de tissus sains ou de tissus cancéreux, nous posons

deux hypothèses : l'hypothèse nulle où le type de tissu (sain ou cancéreux) n'a pas d'influence sur l'expression du gène et son hypothèse alternative, où le type de tissu a une influence. Résoudre ce test, revient à considérer un modèle pour chaque hypothèse et à choisir le plus probable. La probabilité d'un modèle « M » basée sur des observations « x » est classiquement représentée par la formule de Bayes :

$$P(M|x) = \frac{P(x|M)P(M)}{P(x)}$$

Avec M le modèle et x un vecteur contenant les variables observées. P(M|x) est la probabilité d'avoir le modèle en ayant les observations x et P (x|M) est la probabilité d'avoir les observations x avec le modèle M. P(M) et p(x) sont respectivement les probabilités d'observer ce modèle et ces observations parmi tous les cas possibles. Dans notre cas, x est un vecteur contenant les décomptes observés en SAGE du TE du gène d'intérêt.

Comme nous n'avons pas d'*a priori* sur le choix du modèle, nous considérons les deux modèles équiprobables. Donc si P(M) est égale pour les deux modèles, seule P(x|M) diffère. Donc comparer P(M|x) pour chacun des modèles nous amène à calculer ce que l'on nomme le facteur de Bayes :

$$\frac{P(x|M_1)}{P(x|M_2)}$$

En pratique, on utilise comme score le logarithme de ce facteur. Si ce score est supérieur à zéro, alors le modèle 1 est le plus probable. Le passage au logarithme permet d'une part de pouvoir simplifier les équations dans la plupart des cas et d'autre part, d'éviter le dépassement de capacité de calcul d'un ordinateur lorsque ce rapport nécessite le calcul de factoriels.

Après cette explication sur notre stratégie de comparaison de modèles, nous allons brièvement présenter les équations de modélisation.

Pour chaque expérience SAGE « i », avec un nombre de Tags séquencés « n_i », la probabilité d'avoir le compte du Tag « xi » sera donc donnée par une loi binomiale :

$$P(X = x_i) = C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i - x_i}$$

où p est la probabilité d'apparition du Tag considéré. A travers cette approche, nous prenons bien en compte le nombre de Tags séquencés n_i de chaque SAGE. Lorsque l'on considère le modèle 1, où la probabilité p est indépendante de l'origine saine ou cancéreuse des tissus, la probabilité d'avoir le modèle M_1 pour l'ensemble des SAGE étudiés sera :

$$P(x|M_1) = \int_0^1 \prod_{i=1}^{n} C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i - x_i} dp$$

Dans le modèle 2, la probabilité d'apparition de l'ET sera différente dans les SAGE issus de tissus sains ou cancéreux. Donc, si on considère un Tag que l'on retrouve dans « n » expériences SAGE, « m » dans des tissus sains et donc « n – m » dans des tissus cancéreux, nous obtiendrons l'équation suivante :

$$P(x|M_2) = \int_0^1 \prod_{i=1}^{m} C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i - x_i} dp \int_0^1 \prod_{i=m+1}^{n} C_{n_i}^{x_i} p^{x_i} (1-p)^{n_i - x_i} dp$$

Le développement de ces équations pour calculer le score du test à partir de ces deux probabilités est fourni en Annexe A (pour les spécialistes désirant connaitre l'implémentation exacte du test).

Nous pouvons remarquer que le facteur de Bayes des deux modèles revient à une loi hypergéométrique à un coefficient multiplicatif près. Cela s'observe en calculant l'intégrale :

$$\begin{aligned} \int_0^1 p^x (1-p)^{n-x} dp &= \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \\ &= \frac{1}{(n+1)} \times \binom{n}{x}^{-1} \end{aligned}$$

Alors le facteur de Bayes devient :

$$\text{constante} \times \frac{\binom{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} x_i} \binom{\sum_{i=m+1}^{n} n_i}{\sum_{i=m+1}^{n} x_i}}{\binom{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} x_i}}$$

Ce test a été implémenté sur les serveurs de calculs du laboratoire en langage Java. Nous l'avons utilisé pour comparer le niveau d'expression d'un transcrit entre un jeu d'expériences SAGE provenant de tissus sains et un autre provenant de tissus cancéreux, en utilisant pour la

variable x_i le décompte des Tags. Si le modèle 2 est choisi, on estimera que le gène sera plus exprimé dans un des deux groupes.

Ce test a été utilisé sur les données SAGE, mais il peut également être utilisé sur toutes les expériences que l'on peut assimiler à une loi binomiale et pour lesquelles on désire comparer deux groupes.

12 JMACS, une librairie de manipulation d'alignements multiples

L'alignement multiple de séquences est un outil privilégié des bioinformaticiens. Contrairement aux alignements deux à deux du programme MegaBLAST, dans un alignement multiple, nous pouvons comparer en une fois une grande quantité de séquences entre elles. L'alignement multiple est utilisé dans de nombreux cas tel que, la validation d'une prédiction de gène, l'annotation fonctionnelle d'un gène, la réalisation d'analyses phylogénétiques ou la mise en évidence de motifs ou de résidus importants d'une protéine que se soit en repérant les conservations ou les mutations.

Un alignement multiple ne se résume pas à un empilement de lettres représentant les acides aminés ou les nucléotides, mais contient un grand nombre d'informations. Dans un alignement multiple de séquences, nous pouvons identifier et visualiser des domaines fonctionnels ou de conservation, des régions de structure connue et bien d'autres informations sur les séquences. Un exemple de fichier contenant ce type d'informations est fourni par le programme MACSIMS (J. D. Thompson et al., 2006) qui annote automatiquement les alignements multiples de séquences complètes de protéines comprenant aussi bien des informations de structure que de taxonomie (voir chapitre 9.1) Dans ce cadre, nous avons réalisé la librairie JMACS (*Java for Multiple Alignment of Complete Sequence*) en Java qui permet d'accéder et de manipuler, de façon plus efficace, l'ensemble des informations disponibles. Il est à noter qu'il existait déjà dans la librairie Java pour la bioinformatique, BioJava (Holland et al., 2008) une implémentation d'un alignement multiple de séquences (http://www.biojava.org/docs/api/org/biojava/ bio/alignment/package-summary.html). Cependant, comme nous le verrons par la suite, cette implémentation

présentait de nombreuses limites qui m'ont amené à créer une librairie susceptible de traiter rapidement et efficacement des alignements multiples contenant de nombreuses séquences.

12.1 Une implémentation en Java basée sur MAO et MACSIMS

Nous nous sommes basés sur l'ontologie MAO (*Multiple Alignment Ontologie*) (J. D. Thompson et al., 2005) pour construire notre librairie. L'ontologie est une façon de modéliser les concepts d'un domaine de connaissance. Elle est organisée en classes représentant les concepts de base. Chaque classe possède des attributs décrivant ses caractéristiques. Les classes sont reliées entre elles par des relations basiques telles que « est » que l'on nomme héritage ou encore « contient » que l'on nomme agrégation. Nous avons commencé par retranscrire les classes de MAO, leurs attributs et leurs relations (Figure 39) sous forme de codes sources Java. Ces codes sources qui définissent tout ce qu'un objet est capable de faire que se soit pour un traitement de données ou pour un accès à un autre objet sont appelés, en langage de programmation orienté objet, des interfaces. Les interfaces sont donc la traduction directe de l'ontologie en code source.

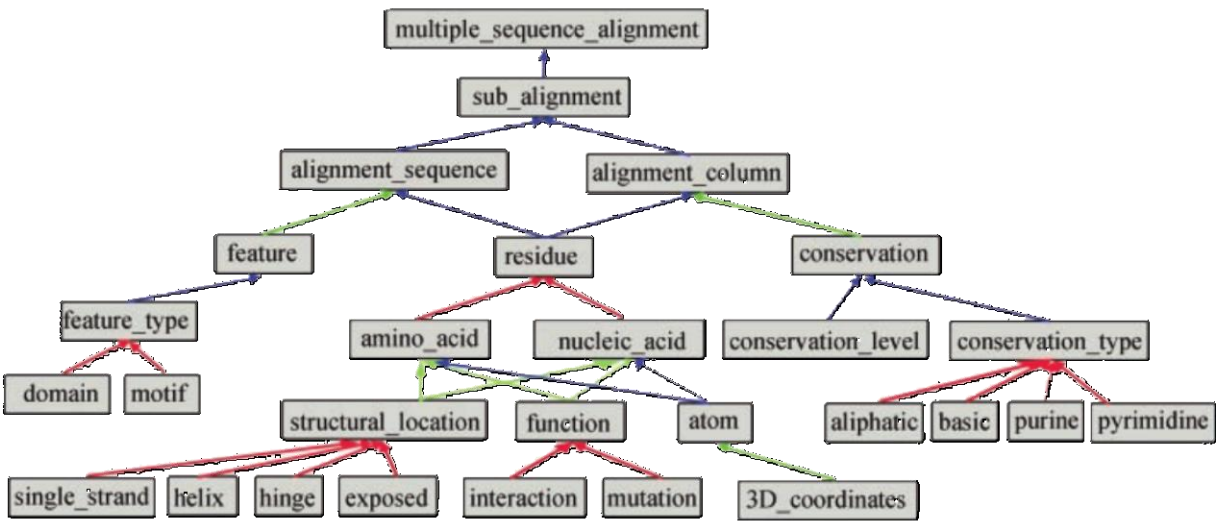


Figure 39 : Représentation de MAO.
Les flèches vertes désignent des attributs (*feature* est un attribut de *alignment_sequence*), les flèches bleues une agrégation (*sub_alignment* est contenu dans *multiple_sequence_alignment*) et les flèches rouges un héritage (*amino_acid* est un *residue*).

En complément de l'ontologie des alignements, la hiérarchie des annotations utilisée par le programme MACSIMS a été également ajoutée. L'implémentation de la librairie a été optimisée pour privilégier la vitesse de lecture et d'édition des données de séquences biologiques pour les traitements à haut débit.

12.2 Description de la librairie

La librairie est téléchargeable depuis le dépôt subversion : http://alnitak.u-strasbg.fr/svn/dkieffer/public/JMACS/. Elle a été structurée en neuf paquets regroupant les fonctions essentielles (Figure 40) suivantes: i) la création et la modification d'un alignement, ii) la communication automatisé entre les objets de la librairie, iii) la lecture et l'écriture des alignements sous différents formats, iv) la visualisation d'un alignement, v) les interfaces de MAO, vi) la création et la modification des séquences, vii) l'accès à toutes les annotations sur les séquences, viii) les constantes que sont les types de résidus, ix) une boite à outils pour travailler sur les alignements. Nous allons passer succinctement cette librairie en revue.

Packages	
org.igbmc.jmacs.alignment	Implementation of Alignment based from Macsim and MAO
org.igbmc.jmacs.alignment.event	Event System of API
org.igbmc.jmacs.alignment.files	To read and write the multiple sequence alignments
org.igbmc.jmacs.graphics	Package for display sequences and alignments
org.igbmc.jmacs.mao	This Package define the interfaces based from MAO
org.igbmc.jmacs.sequence	Define implementation of alignment sequence
org.igbmc.jmacs.sequence.infos	Define Object for the sequence informations
org.igbmc.jmacs.sequence.residues	Package defining Residues.
org.igbmc.jmacs.util	Util Objects with static methods to obtain and convert data in JMACS

Figure 40: Aperçu des paquets de la librairie JMACS.

12.2.1 Paquet « MAO »

JMACS fournit une série d'interfaces Java respectant les spécifications de l'ontologie des alignements multiple MAO (Figure 41). A ce jour, la librairie JMACS ne fournit pas encore tous les objets décrits dans MAO. Les objets tels que les structures secondaires (*helix*) ou les types de conservations (*conservation-type*) sont indisponibles. Le développement a été très vite orienté vers la rapidité d'accès aux données de séquences, telles que le type de nucléotide

(a,t,g,c…) à une position donnée ou les annotations sur la séquence : les *Features*. Ces interfaces ont pour objectif d'être une base pour tous les programmes Java, respectant l'ontologie des alignements. Ceci permettra la réutilisation des codes entre les différents projets de la communauté bioinformatique.

Interface Summary	
AlignmentColumn	A basic interface based on MAO for alignment column
AlignmentSequence	sequence definition by MAO
AminoAcid	A aminoacid representation
Atom	A atom representation
Conservation	A conservation for columns in alignment.
Feature	a MAO feature representation
MultipleSequenceAlignment	A MAO multiple alignment representation
NucleicAcid	A nucleic acid representation.
Residue	A representation of Residue.

Figure 41 : Interfaces Java représentant les spécifications de MAO.

12.2.2 Paquets « *residues* », « *sequence* », « *infos* » et « *alignment* ».

Les informations sur une séquence, telles que le nom de la séquence, la définition, le numéro d'accession dans une banque, mais aussi les domaines de protéines Pfam (Finn, Griffiths-Jones, & Alex Bateman, 2003) ou encore les interactions entre résidus sont toutes des annotations répertoriées dans un fichier MACSIMS. J'ai représenté ce système comme une « surcouche » des classes de MAO. Les 3 classes principales de cette nouvelle couche sont l'interface « *MacsimsSequence* » qui permet l'accès à un objet « *MacsimsSeqInfo* » contenant entre autres, le nom, la définition de la séquence et une liste de « *MacsimsFeatures* », les annotations localisées sur la séquence telles que des motifs structuraux, Pfam, la taxonomie, mais aussi, toute autre annotation que l'utilisateur estime utile d'ajouter. Ce sont également des interfaces qui héritent des interfaces MAO présentées précédemment, c'est-à-dire possédant les attributs pour l'annotation (Figure 42). A ce stade, nous avons présenté uniquement des interfaces, c'est-à-dire les définitions des objets. Comme nous l'avons déjà dit, faire l'effort de passer par ces interfaces, garantit de travailler avec un standard (ici, le modèle de l'ontologie MAO) et de permettre aux programmeurs d'avoir un modèle strict pour les développements informatiques. En effet, programmer à partir d'un standard facilite la relecture du code par un tiers, la maintenance des programmes et la transmission des savoir-faire.

Nous allons maintenant rapidement présenter les objets dérivant de ces interfaces, proposés dans JMACS. Ces objets sont optimisés pour pouvoir le plus rapidement possible changer leurs attributs. Nous aurions aussi pu imaginer orienter le développement vers un aspect favorisant l'espace mémoire utilisé ou la récupération optimisée des données d'alignements via les banques de données en ligne par exemple, mais la motivation première de ce travail était de créer un outil permettant d'analyser en un temps record plusieurs milliers de séquences biologiques.

Figure 42 : Arbre des interfaces de la librairie JMACS.
Les interfaces MACSIMS héritent des interfaces MAO et ajoutent de nouveaux attributs.

Tout d'abord, l'objet « *Sequence* ». C'est l'objet central de la librairie. Une séquence biologique ayant forcément un type (ADN, ARN ou protéine), il n'a pas été permis de pouvoir directement utiliser l'objet *Sequence* pour ne pas donner l'opportunité de créer une séquence sans type. Lorsque nous parlons des séquences JMACS, nous parlons en fait des objets *DNASequence*, *RNASequence* et ProteinSequence dont la principale différence est le type de séquence biologique représenté. Depuis ces séquences, il est possible d'accéder ou de

créer si besoin, le tableau des « *Features* » dont chacun possèdera, les positions de début et de fin de l'annotation, ainsi qu'un descriptif de type de *feature* considéré. Un rapide descriptif des « *Features* » est disponible à cette adresse : http://decrypthon.igbmc.fr/sm2ph/html/help.html#macsims_feat.

Les séquences JMACS ont l'avantage de faire la distinction entre 2 localisations : la position réelle du résidu dans la séquence et la colonne de l'alignement où se situe le résidu (Figure 43). Ainsi, un *gap* n'aura qu'une seule position sur la séquence, mais pourra correspondre à plusieurs colonnes. Une table d'association pour chaque séquence permet de faire facilement et rapidement la conversion entre ces deux types de localisation. Ceci est bien utile, lorsque nous avons sur nos séquences, des positions d'étude (dans le cas ici, des positions statistiquement plus hétérogènes en Cancer (ou en Normal)) et que nous voulons traiter l'analyse de ces séquences, en s'assurant de ne pas regarder une grande délétion. L'index des colonnes est défini depuis la colonne 0 car c'est le standard informatique. La position réelle peut être définie de multiples façons selon le contexte. Par défaut, une séquence commence à la position 1. Elle peut être donnée par exemple, par les logiciels BLAST si nous considérons des fragments. La liste des positions peut être inversée si nous avons un ADN complémentaire par exemple, qui est représenté de 3' vers 5'. La position 1 est souvent arbitrairement donnée sur le premier nucléotide du premier codon traduit de l'ARNm, la séquence en amont contenant alors des numéros de position négatifs.

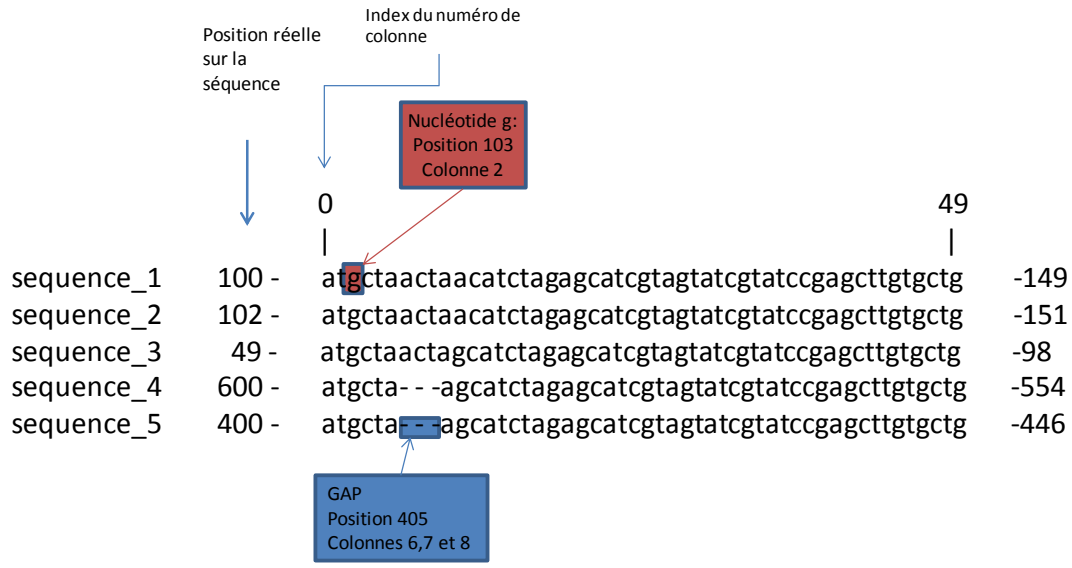


Figure 43 : Illustration des deux systèmes de localisation sur un alignement multiple de séquences.

Dans le premier système, la séquence numéro 1 a un « g » en position 103 et la séquence 5 a un gap en position 405 de longueur 3. Dans le second système, on dira que la séquence 1 a un « g » en colonne 2 et que la séquence 5 a un gap en colonne 6,7 et 8.

12.2.3 Paquet « *files* »

Il existe de nombreux formats de fichier pour stocker les alignements multiples. Développer des programmes pour lire chaque format de la façon la plus efficace possible est un travail courant. Pour éviter la redondance du code et permettre une maintenance facile, le traitement des fichiers d'alignements multiples a été séparé en deux tâches distinctes.

- La première consiste à créer un moteur de prise en charge des données du fichier ou un « *DocumentHandler* », qui sera unique pour tous les formats de fichier d'alignement multiple. Son rôle est de créer les objets Java au fur et à mesure qu'il reçoit les informations contenues dans le fichier.

- La seconde consiste en une série d'analyseurs de fichier ou « *Parsers* », spécifiques de chaque format, qui auront pour tâche de lire le fichier et d'envoyer les données au *DocumentHandler*.

Ainsi, pour pouvoir traiter un nouveau format, il suffit de coder uniquement la partie d'analyse du fichier, la création de l'objet Java représentant l'alignement étant la même pour tous. A ce jour, la librairie peut gérer les formats fasta, msf et MACSIMS.

12.2.4 Paquet « *graphics* »

Ce paquet contient les outils de base pour construire une interface graphique de visualisation d'un alignement multiple. Tout d'abord, il y a les classes de rendus ou «*renderer* » qui vont servir à créer une image de l'alignement. Plusieurs façons de créer l'image sont disponibles, soit par coloration de résidus (Figure 44), soit par coloration de certaines annotations (Figure 46). Ensuite, des objets sont mis à disposition pour afficher l'image de l'alignement, soit par une vue classique de la séquence avec les noms des séquences et les numéros des colonnes, soit par un aperçu global de tout l'alignement (Figure 45 et Figure 47). Cette dernière possibilité permet rapidement de voir les résidus et les domaines conservés, ainsi que les

zones d'insertion/délétion de fragment. Ces données permettent de distinguer les différentes familles de séquences et de visualiser des parties fonctionnelles ou conservées de la séquence.

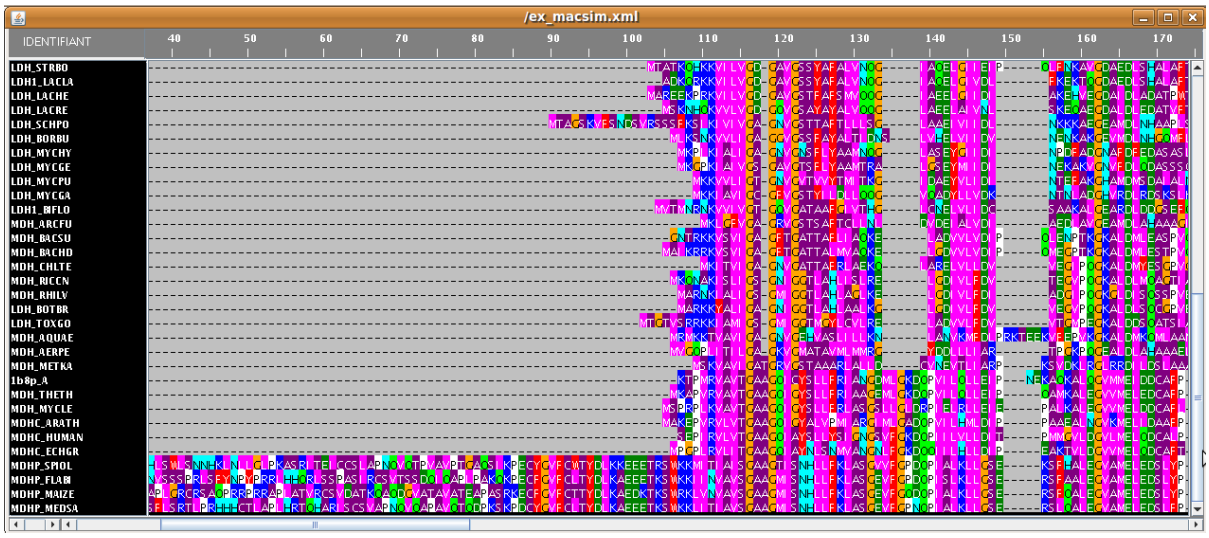


Figure 44 : Exemple d'un alignement de séquences protéiques avec coloration par résidu.

Figure 45 : Une vue d'ensemble d'un alignement multiple de séquences protéiques.

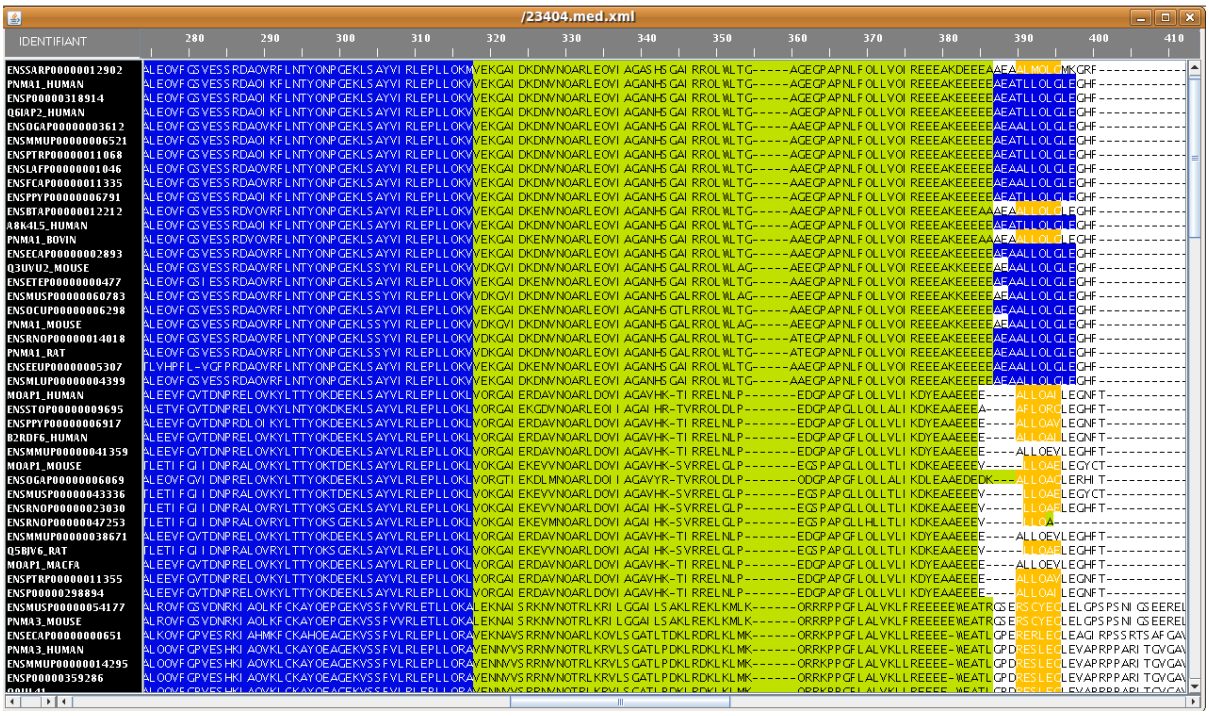


Figure 46 : Alignement multiple de séquences protéiques colorées par *Features*.
Dans cette exemple, nous avons en bleu les *Pfam-A* (domaine protéique), en jaune les *Blocks* (séquence conservée) et en orange les *phyloblocks* (séquence conservée par phylogénie).

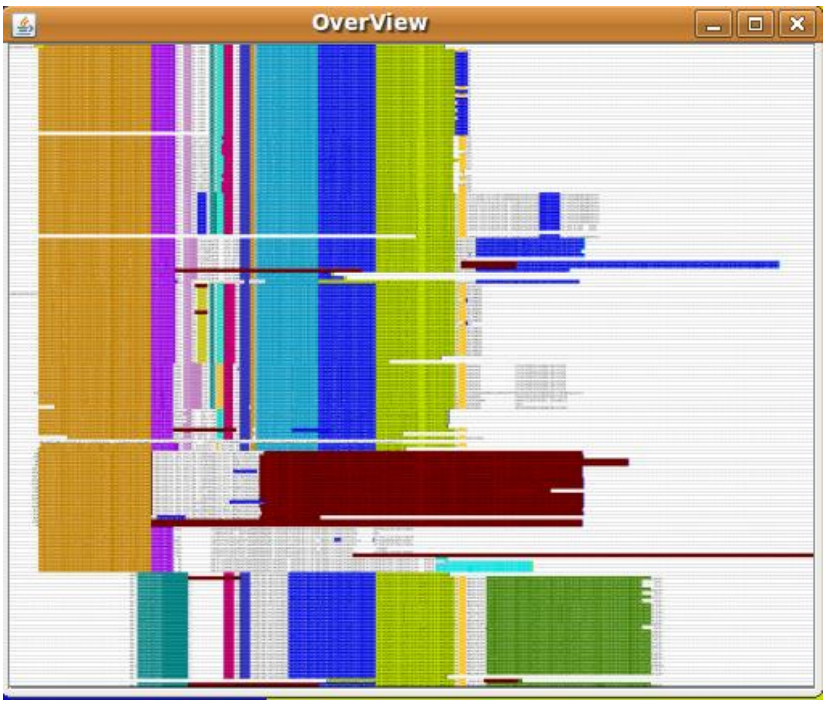


Figure 47 : Vue d'ensemble d'un alignement multiple de séquences protéiques colorées par *Features*.

12.2.5 Paquet « *util* »

Ce paquet contient tout d'abord des classes facilitant l'obtention des objets alignement à partir de fichiers, ainsi qu'une classe pour sauvegarder les alignements dans le format choisi : MACSIMS, msf ou fasta. Ce paquet contient également des classes pour réaliser des analyses et des éditions sur les alignements telles que, calculer un pourcentage d'identité entre deux séquences, trier les séquences d'un alignement multiple selon leurs groupes d'appartenance ou encore, obtenir la séquence complémentaire d'une séquence nucléique.

12.3 Comparaison avec une autre implémentation d'alignement multiple de séquence

La librairie BioJava propose une implémentation d'un objet représentant un alignement multiple. Cette implémentation considère un alignement multiple uniquement comme une liste de séquences et non pas comme un objet avec des attributs dynamiques et des annotations. Nous avons testé la vitesse de traitement d'un alignement par la librairie BioJava et par la librairie JMACS. Dans un même programme Java, nous avons réalisé avec chaque implémentation une tâche courante consistant à ajouter une insertion dans un alignement. Cette tâche a été effectuée 100 fois de suite (Figure 48). Avec BioJava, le temps de traitement est de 440ms, tandis qu'il n'est que de 30ms avec JMACS. Dans le monde du haut débit, le gain d'un facteur 14 permet généralement passer d'un traitement prenant plusieurs jours à quelques heures.

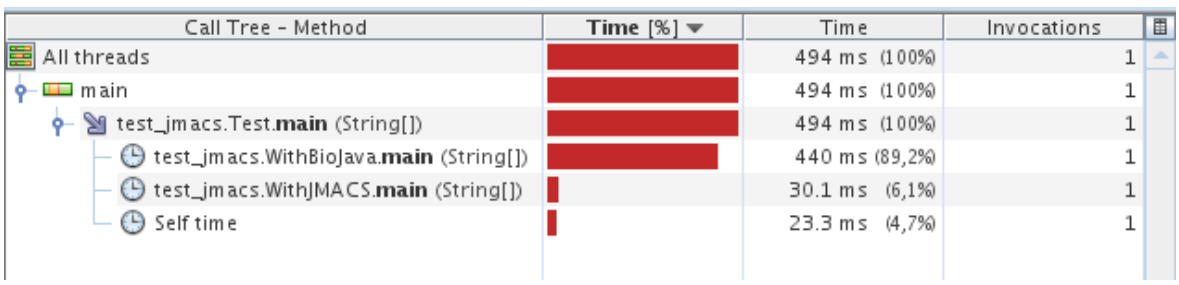


Figure 48 : Comparaison des vitesses d'exécutions entre BioJava et JMACS.
Mesure du temps d'exécution nécessaire pour insérer 2 résidus dans un alignement de 100 séquences et 1000 colonnes en utilisant soit la librairie BioJava (440ms), soit la librairie JMACS (30,1ms).

Il faut noter que BioJava est avant tout une librairie dédiée à l'extraction des informations de fichiers biologiques. Dans cette optique, les objets sont implémentés comme étant « immutables », c'est-à-dire que l'on ne peut pas les modifier. Pour pouvoir modifier un alignement, il faut donc en faire une copie comportant les modifications. Au contraire, la librairie JMACS considère l'alignement et les séquences comme des objets dynamiques, ce qui permet de les modifier directement et d'augmenter la rapidité du traitement.

La librairie JMACS fournit des outils pour créer, acquérir et modifier les données d'un alignement multiple de séquences. Grâce à l'implémentation de l'ontologie des alignements multiple MAO, il est facile d'accéder aux données à tous les niveaux, de l'alignement jusqu'au résidu. Dans JMACS, l'implémentation ajoute des méthodes simples pour accéder aux multiples annotations présentes dans les fichiers MACSIMS. Non seulement JMACS permet, en quelque lignes de code, de faire des manipulations et des analyses sur les alignements grâce à une panoplie de routines préprogrammées (comme la visualisation, des calculs d'identité entre deux séquences, l'insertion ou la délétion de colonne ou de séquence dans l'alignement), mais permet également de traiter les données de façon très rapide, ce qui est essentiel pour le traitement à haut débit.

Grâce aux avantages de cette librairie, j'ai pu aisément « empiler » les EST des alignements deux à deux obtenus par MegaBLAST dans un unique alignement multiple de séquences. Ceci m'a permis notamment de visualiser facilement les séquences mais surtout, de pouvoir réaliser des décomptes (par exemple, compter le nombre de nucléotides substitués) de façon optimisée, ce qui est d'autant plus utile que, pour certains gènes, quelques milliers d'EST leurs sont associés.

13 Etude statistique de l'hétérogénéité des ARNm dans les tissus cancéreux au travers des données SAGE

13.1 Collecte des données SAGE

Les expériences SAGE Nla3 d'origine humaine ont été classées manuellement, sur la base des descriptions contenues dans les GSM, entre expériences réalisées sur tissus sains et celles effectuées sur tissus cancéreux. Les expériences réalisées sur des tissus non-cancéreux qui ne peuvent être qualifiés de « sains » (définitions non fournies ou ambigües, autres maladies, traitements physico-chimiques…) n'ont pas été retenues. Lorsque nous parlons de Cancer, nous parlons en fait de tous les types de cancers observés dans tous tissus. Dans tous les résultats présentés par la suite, nous référencerons nos deux groupes par « Normal » et « Cancer ». Le résultat de ce traitement est montré dans la Figure 49.

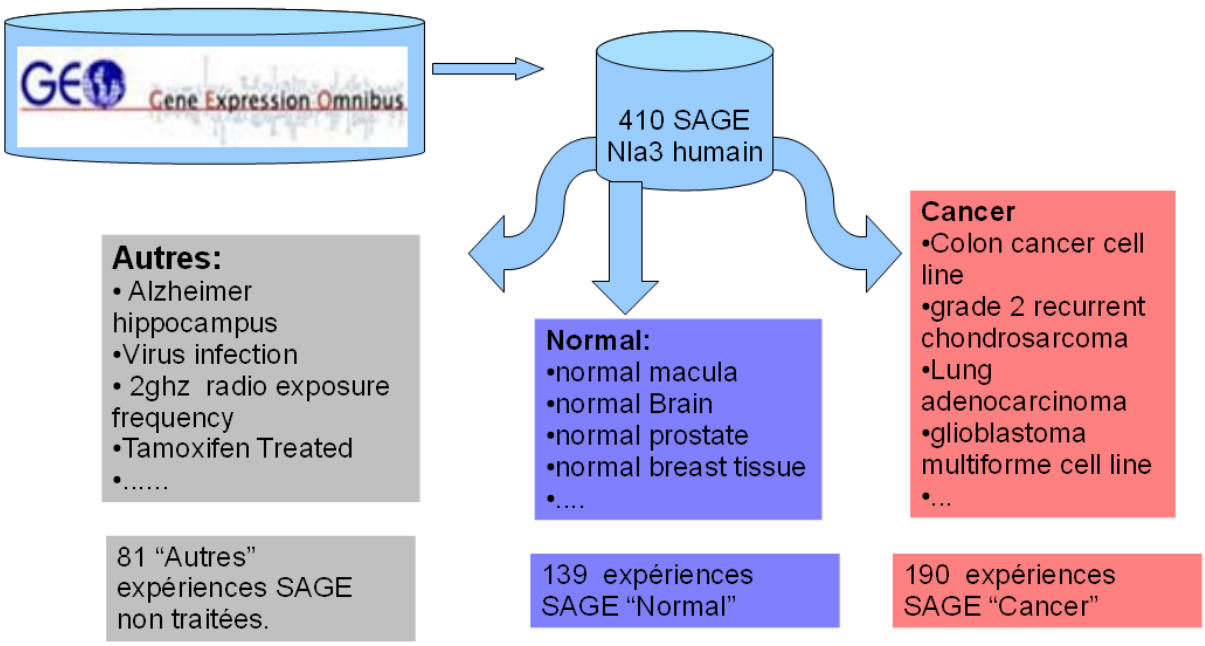


Figure 49 : Collecte des expériences SAGE provenant de tissus cancéreux (Cancer) et ceux provenant de tissus sains (Normal).
Les expériences provenant de maladies autres que le cancer, ou ayant subi un traitement physique ou chimique (autres), ont été éliminées de l'étude.

Nous avons extrait 190 expériences SAGE issues de tissus cancéreux, 139 de tissus sains et 81 de tissus présentant diverses maladies ou divers traitements chimiques ou physiques qui ont été exclues de notre étude. Nous avons donc obtenu deux échantillons d'expériences SAGE (Normal et Cancer) de tailles suffisantes pour permettre une étude statistique.

13.2 Diversité des origines tissulaires des expériences SAGE

Notre étude a été réalisée de façon transversale à tous types de tissus Normal et Cancer (Figure 50 et Figure 51). Il y a un peu plus d'expériences SAGE Cancer (190 contre 139), notre échantillon Normal couvre une plus grande variété de tissus que l'échantillon Cancer. Les origines tissulaires observées en Cancer sont directement liées aux organes touchés par la maladie et nous avons quelques tissus communs entre les deux groupes (colon, estomac, ovaire, peau, prostate, cerveau, moelle osseuse et poumon).

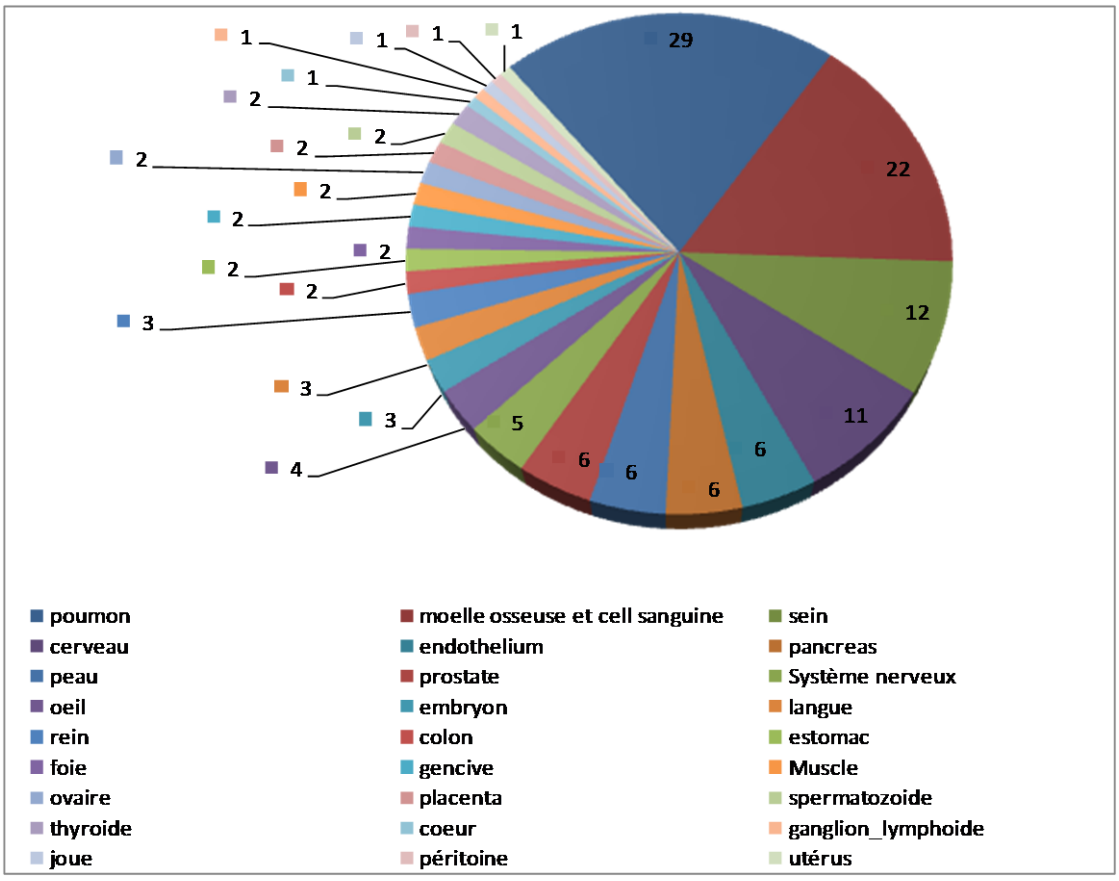


Figure 50 : Origines tissulaires des SAGE Normal provenant de la plate-forme GEO GPL4. Les 27 portions représentent le nombre d'expériences SAGE issues d'un même organe.

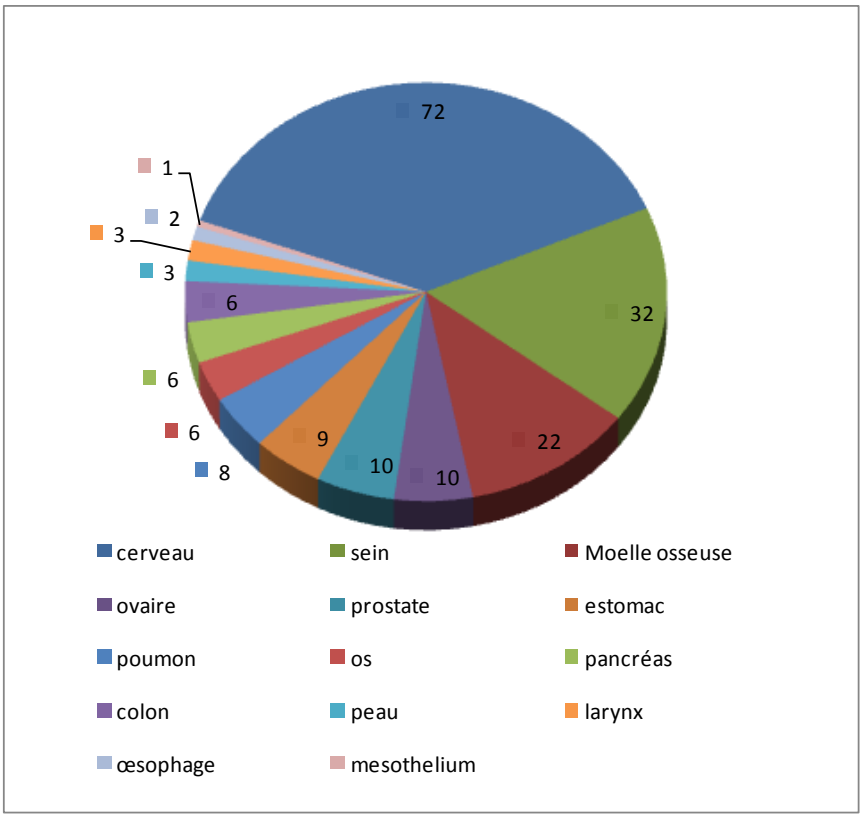


Figure 51 : Origines tissulaires des données SAGE Cancer provenant de la plate-forme GEO GPL4.
Les 14 portions représentent le nombre d'expériences SAGE issues d'un même organe.

Il y a aussi une disparité dans les proportions de tissus entre les deux échantillons. En Normal, les tissus les plus représentés sont originaires du poumon avec 29 expériences SAGE puis viennent, la moelle osseuse et les cellules sanguines avec 22 expériences, le sein (12) et le cerveau (11). Pour réaliser une étude stricte, il aurait fallu avoir à notre disposition des expériences en grand nombre et en même proportion entre Normal et Cancer venant des mêmes tissus, afin de pouvoir correctement prendre en compte le facteur tissulaire dans notre étude. Faute d'effectifs, nous assumons que, dans le cadre d'une étude transversale des modifications observables dans tous les gènes comparant tissus sains et cancéreux, l'origine tissulaire a peu, ou pas, d'impact significatif. Cela est bien sûr arbitraire et pourra être discuté à chaque étape de nos analyses.

13.3 Diversité des types de cancer

Pour une même origine tissulaire, il peut y avoir des cancers de différents types. Par exemple, pour les cancers du cerveau, nous avons, dans les expériences SAGE, des astrocytoma, des ependymoma, des medulloblastoma, des oligodendroglioma, des ependymoblastoma, des glioblastoma et des neuroblastoma. Pour notre approche, nous ne prenons pas non plus en considération les types de cancer et cela également, pour des raisons de manque d'effectif.

13.4 Diversité de la qualité des expériences SAGE

La sensibilité de la technique SAGE pour détecter des transcrits peu exprimés est directement liée au nombre de Tags séquencés. Il est donc important de prendre en compte cet aspect. Le Tableau 7 présente quelques statistiques descriptives de nos échantillons et la Figure 52, une représentation de la distribution des expériences selon leur nombre de Tags séquencés. Les médianes et les moyennes sont relativement proches, mais l'échantillon Normal présente quelques expériences avec un nombre de Tags séquencés beaucoup plus important que ce qui est observé en Cancer.

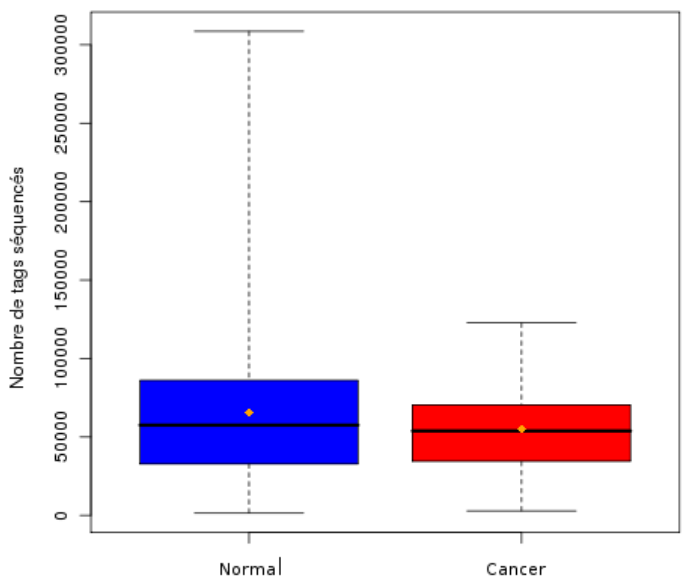


Figure 52 : Représentation en "boite à moustaches" de la dispersion du nombre de Tags séquencés des SAGE Normal et Cancer.
La moyenne est reportée sur les boites par un losange orange.

Tableau 7: Statistiques descriptives du nombre de Tags séquencés dans les échantillons Cancer et Normal.

	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
normal	1430	32814	57523	65543	86112	308589
cancer	2881	34568	53802	55048	70350	122690

Pour éviter d'introduire un biais trop fort dû aux différences entre Cancer et Normal, nous avons arbitrairement décidé d'éliminer les expériences SAGE du groupe Normal ayant un nombre de Tags séquencés supérieur à 125 000 qui sont absentes dans les expériences du groupe Cancer (Figure 53). 15 expériences SAGE ont ainsi été éliminées de l'étude.

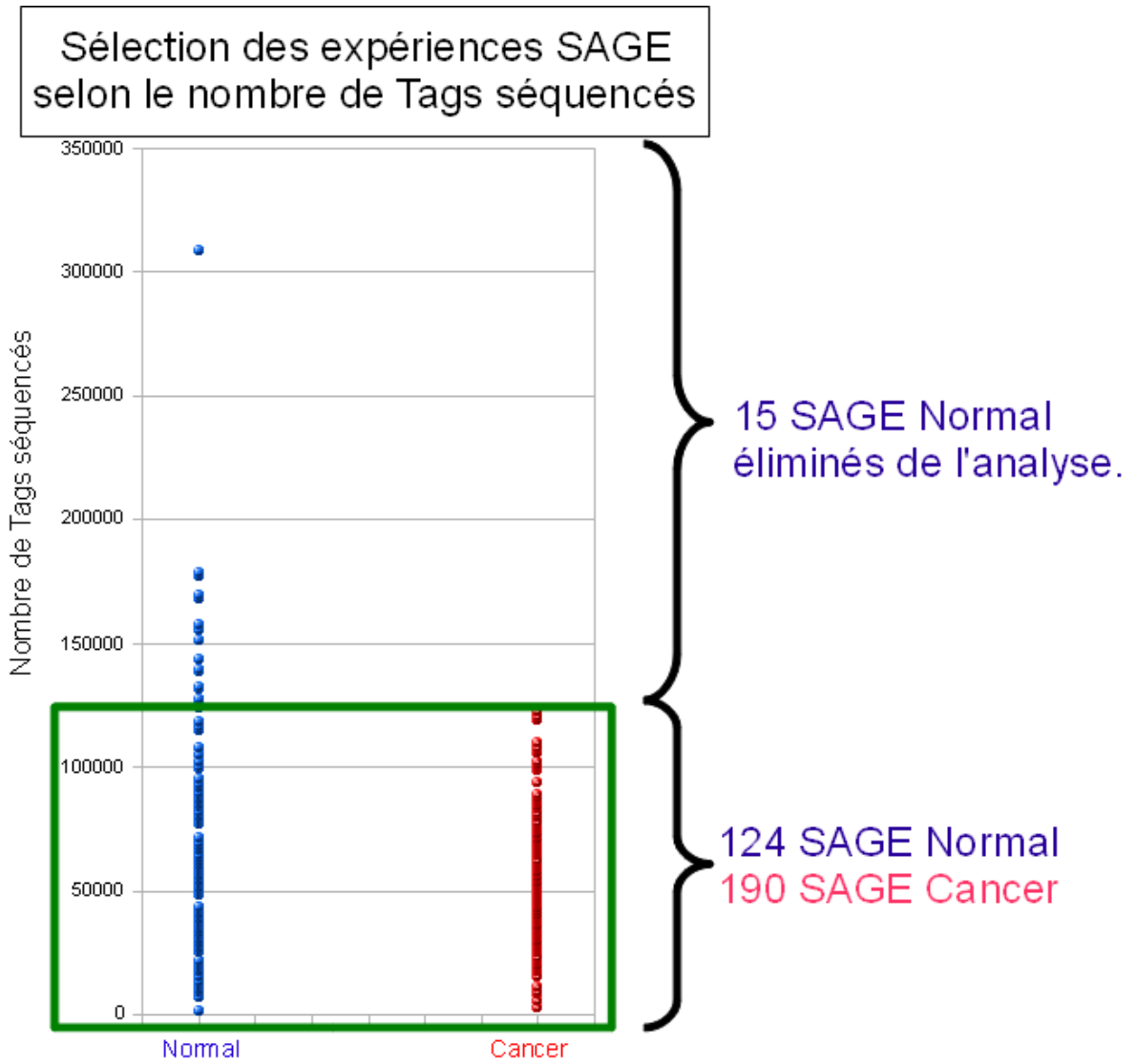


Figure 53 : Sélection des expériences SAGE à analyser selon le nombre de Tags séquencés

L'hétérogénéité des données disponibles nous a contraint à essayer d'être le plus stringent possible dans notre étude, tout en maintenant un effectif raisonnable pour une analyse statistique. Cet équilibre entre la quantité et la qualité des données est souvent un point délicat dans les analyses à haut débit qui demande d'assumer des choix arbitraires de seuil.
C'est donc sur la base d'échantillons de 124 expériences SAGE Normal et 190 Cancer que nous avons réalisé nos études statistiques.

Analyse du facteur Cancer sur le nombre de TE assignés et non-assignés.

Nous avons effectué l'assignation des jeux de TE de chaque expérience SAGE en utilisant le logiciel Sagettarius. Nous avons choisi de considérer uniquement les assignations utilisant les banques TRC et ADNc de TV qui sont les banques les plus fiables (voir Chapitre 9.2). Dans le même souci de rigueur, dans cette analyse, nous avons éliminé les décomptes de TE égaux à 1. En effet, il est accepté qu'il peut y avoir des erreurs de séquençage pouvant mener à des TE de 10 nucléotides portant une erreur. Ce taux est cependant estimé trop faible, pour produire plus d'une fois le même TE erroné. Ne pouvant faire la différence entre un TE, avec un décompte de 1 juste et un TE erroné, ces TE n'ont pas été assimilés dans nos calculs. Nous avons observé 152 658 séquences différentes de TE, toutes expériences confondues (Figure 54). Nous avons ici une estimation du nombre d'ARNm différents qui existent dans nos cellules. Par rapport aux 4 905 TV uniques de notre sous-banque Sagettarius correspondant à 4905 transcrits de qualité (TRC et ADNc), nous avons « mono-assigné » 4 414 séquences de TE en considérant l'ensemble des expériences retenues (124 « Normal» et 190 « Cancer »). 3 862 TE sont communs aux deux types, 260 TE uniquement retrouvés dans les SAGE Normal et 292 TE uniquement dans les SAGE Cancer. Il y a donc déjà à ce niveau des ARNm qui sont spécifiquement détectés, soit dans les SAGE Normal, soit dans les SAGE Cancer. La détection d'un ARNm en SAGE est directement corrélée à son niveau d'expression. Ces 260 TE non détectés en SAGE Cancer peuvent correspondre à des gènes dont l'expression a été fortement diminuée par le cancer, au point de ne plus pouvoir les détecter et inversement, les 292 TE Cancer spécifiques peuvent correspondre à des gènes dont l'expression est très faible en condition normale et/ou largement augmentée dans un cancer. D'autre part, 148 244 TE n'ont pu être assignés. Ce nombre donne un aperçu du nombre d'ARNm qui n'ont pas encore

été séquencés, annotés et vérifiés. Parmi ceux-ci, 55 718 TE sont communs aux deux types, 32 520 sont uniquement retrouvés en Normal et 60 006 uniquement en Cancer. On peut noter que le nombre de TE différents non-assignés spécifiques de l'échantillon Cancer est presque le double de celui spécifique de l'échantillon Normal. Nous avons ici un premier signe d'une hétérogénéité de transcrits beaucoup plus grande en Cancer. Cela peut être lié à la diversité des données Cancer qui pourraient présenter des profils d'expression très différents. Cependant, comme nous l'avons décrit précédemment, pour des raisons statistiques, il n'est pas possible de traiter chaque cancer séparément.

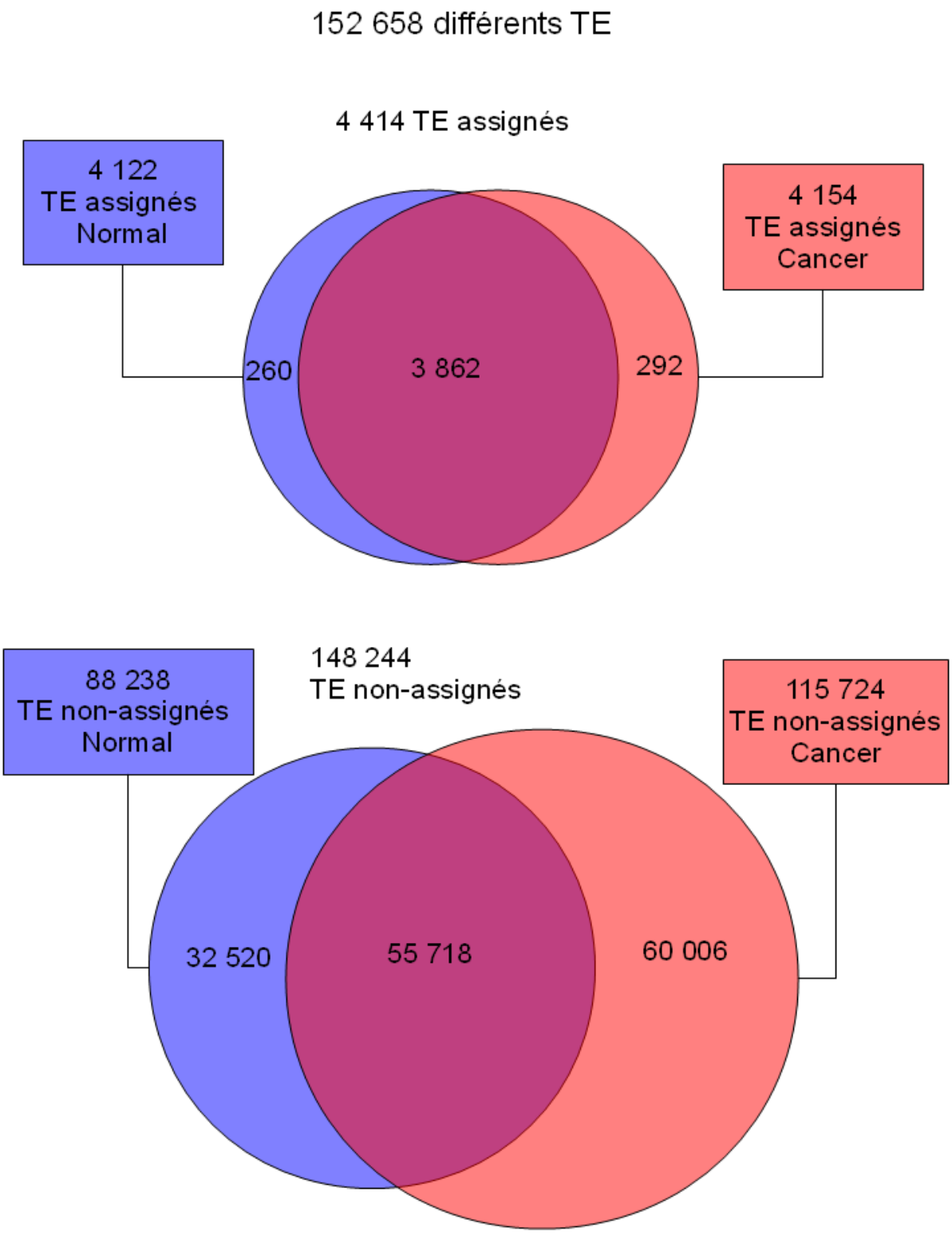


Figure 54 : Résultat de l'assignation des TE des SAGE Normal et Cancer par Sagettarius. Seules les assignations sur la banque des TRC et ADNc de SAGETTARIUS ont été considérées, les autres étant comptabilisées comme non-assignées.

Après cette approche globale, nous avons poursuivi nos investigations en recherchant les expériences SAGE montrant un nombre de TE assignés ou non-assignés anormal. Nous

avons analysé séparément deux données complémentaires. La première est le nombre de séquences différentes observées au sein des TE assignés ou non, qui va nous permettre d'approcher la diversité des ARNm produits dans la cellule. La seconde est la proportion de TE non-assignés, en travaillant sur les décomptes des TE, qui va nous permettre d'analyser le niveau d'expression relatif de ces Tags non-assignés (Figure 55).

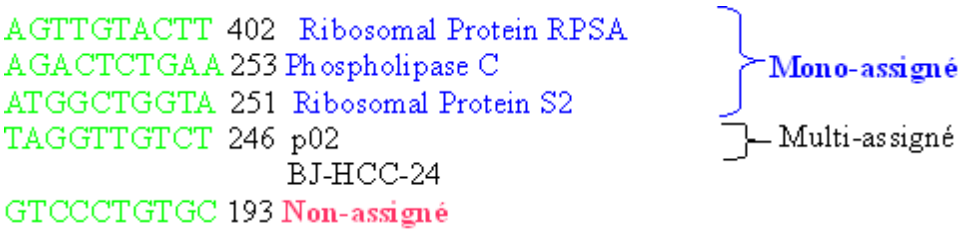


Figure 55 : Exemple de résultat d'assignation de TE.
Sur cette exemple, nous avons 4 types de TE assignés et 1 type de TE non-assigné. Ils représentent un total de 1345 (402 + 253 + 251 + 246) TE assignés et 193 TE non-assignés.

Pour définir les SAGE avec un nombre de séquences de TE assignés et non-assignés anormal, nous avons utilisé un modèle linéaire reliant le nombre de Tags séquencés et le nombre de séquences différentes de TE assignés ou non-assignés, et estimer son intervalle de confiance (voir chapitre 10.5.4). Cette démarche nous a permis de distinguer graphiquement 13 expériences SAGE avec un nombre de séquences de TE assignés plus faible. Après examen, tous sont des SAGE Cancers (Figure 56). Trois expériences SAGE ont au contraire un nombre de séquences de TE assignés plus élevé, un Normal et deux Cancer. De la même façon, nous avons observé 12 SAGE avec un nombre de séquences de TE non-assignés étonnamment élevé dont deux SAGE Normaux (Figure 57). Nous avons vérifié les tissus d'origines de ces SAGE. Ces SAGE proviennent majoritairement de cellules de cartilage et de cellules souches de moelle osseuse. Ces résultats mettent en évidence un effet du cancer observable.

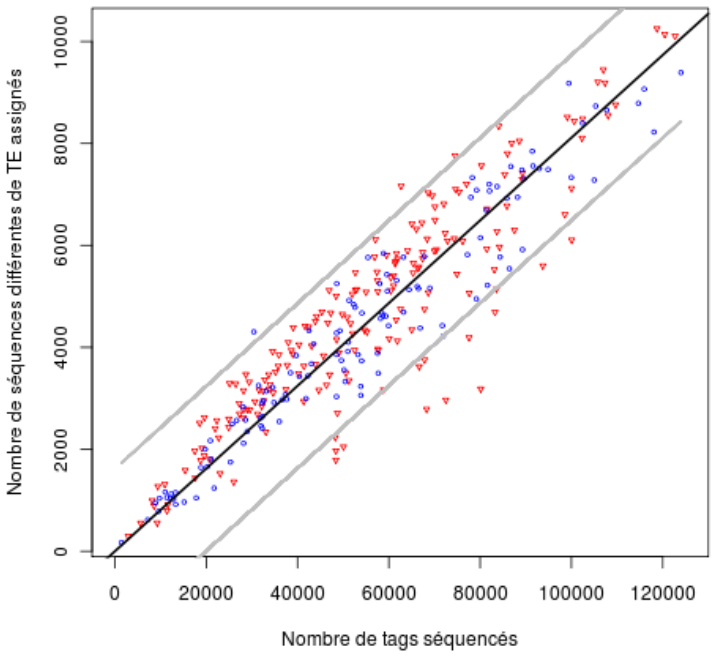


Figure 56 : Nombre de séquences différentes de TE assignés en fonction du nombre de Tags séquencés.
Les SAGE Normal sont représentés par les cercles bleus et les Cancer par des triangles rouges.

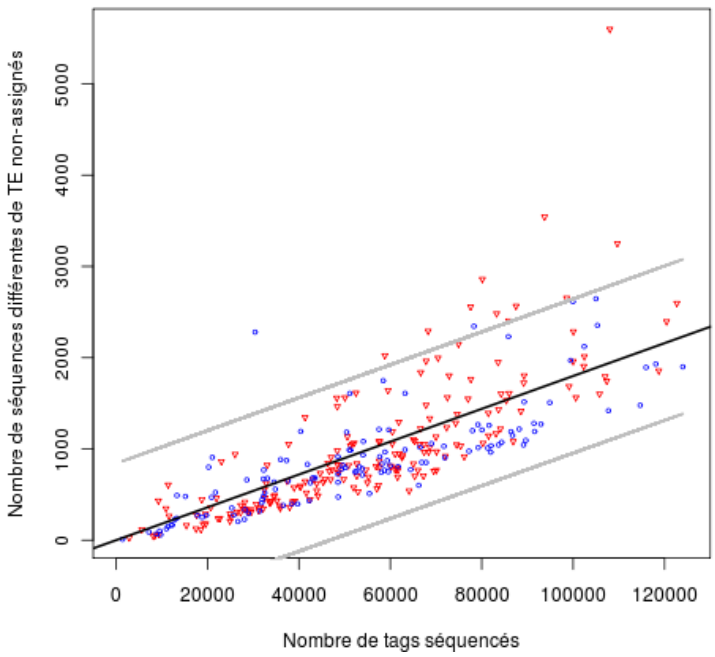


Figure 57 : Nombre de séquences différentes de TE non-assignés en fonction du nombre de Tags séquencés.
Les SAGE Normal sont représentés par les cercles bleus et les Cancer par des triangles rouges.

Nous avons ensuite calculé la proportion de TE non-assignés observés pour chaque SAGE (Figure 58). La moyenne des TE non-assignés est d'environ 10% (10,3% en Normal et 11,5% en Cancer). Nous voyons par exemple que 20 expériences (16 Cancers et 4 Normal) ont une

proportion de TE non-assignés supérieure au double de cette moyenne. Parmi les 16 expériences SAGE Cancer, 6 ont plus de 30% de TE non-assignés. Parmi les tissus considérés, nous retrouvons également des cultures cellulaires de cartilage et de cellules souches de moelle osseuse, mais aussi des cancers de la prostate et de la peau. Nous avons testé si le nombre de TE non-assignés était dépendant du groupe de SAGE (Normal ou Cancer), en utilisant le test de Fisher exact (voir 10.5.5.2) et nous avons observé que le nombre de TE non-assignés était significativement plus grand dans les SAGE Cancer avec une p-valeur inférieure à 10^{-16}.

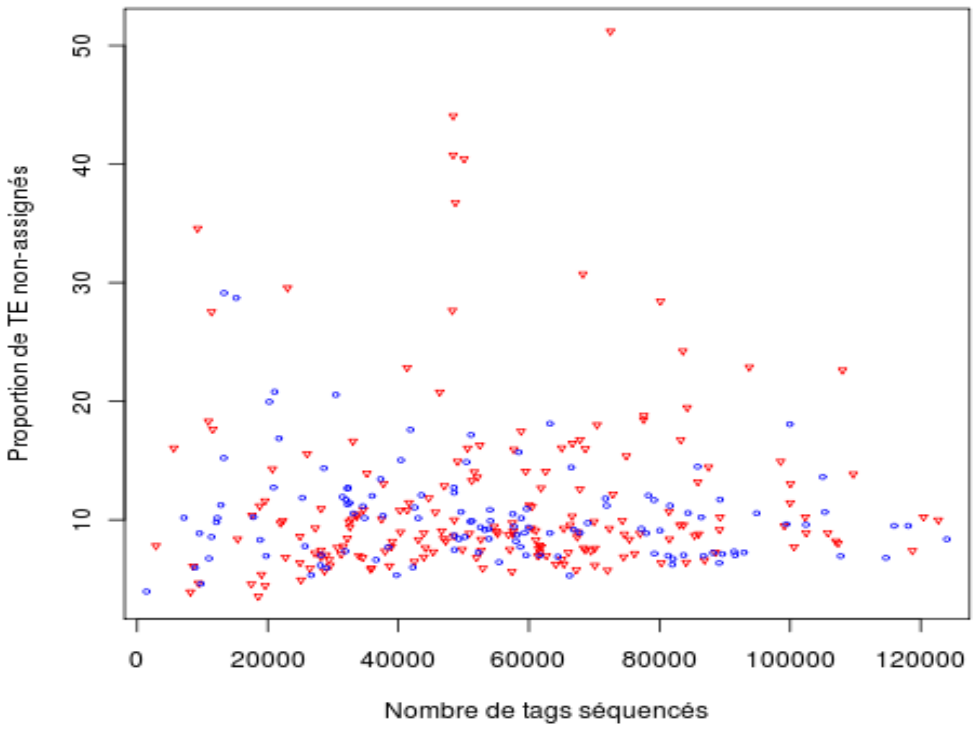


Figure 58 : Proportion de TE non-assignés selon le nombre de Tags séquencés.
Les SAGE Normal sont représentés par les cercles bleus et les Cancer par des triangles rouges.

13.5 Impact du facteur Cancer sur la détection des TRC

Nous avons vu que, dans le groupe Cancer, un grand nombre de TE non-assignés apparaissent, signe d'une hétérogénéité accrue des transcrits au sein des tissus cancéreux. Nous allons maintenant étudier si cette population de TE non-assignés peut empêcher la

détection des produits d'un gène. Pour cela, nous avons utilisé la banque de Tags Virtuels (TV) des transcrits des TRC du programme Sagettarius (voir chapitre 9.2). Nous savons que les transcrits de ces protéines essentielles sont abondamment exprimés dans tous les tissus. Nous connaissons également tous les variants référencés possibles. Les TRC sont donc le meilleur jeu de test possible pour faire une étude transversale à tous les SAGE. Nous avons, pour chaque SAGE Cancer et Normal, comptabilisé le nombre de séquences de TE correspondant aux TV des TRC, ainsi que le nombre de Tags séquencés.

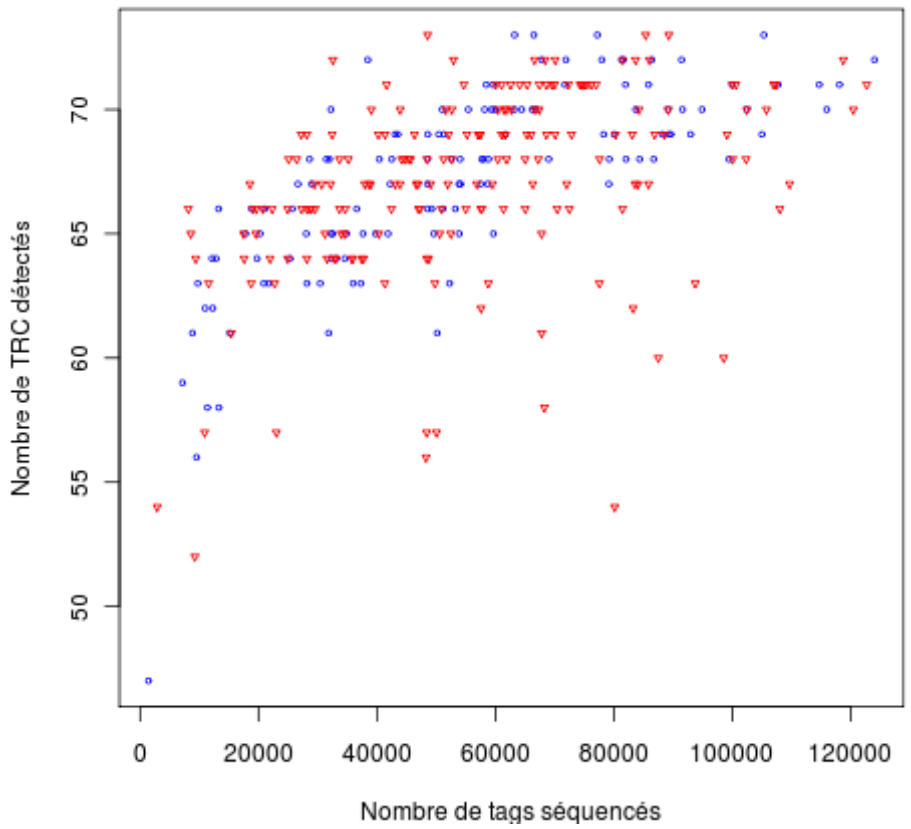


Figure 59 : Nombre de TRC (Transcrit de protéine Ribosomale Cytoplasmique) détectés en fonction du nombre de Tags séquencés.
Les SAGE Normal sont représentés par les cercles bleus et les Cancer par des triangles rouges.

Nous avons comparé la corrélation entre le nombre de Tags séquencés et le nombre de TRC détectés. Pour l'échantillon Normal, nous avons un coefficient de corrélation de Spearman (voir chapitre 10.5.4) de 0,79, donc une corrélation moyennement forte. Pour l'échantillon Cancer, nous obtenons un coefficient de 0,52, soit une corrélation très moyenne. Ces coefficients nous permettent de conclure que le nombre de TRC détectés est plus corrélé au nombre de Tags séquencés dans les tissus sains que dans les tissus cancéreux. On observe que pour le même nombre de Tags séquencés, on peut en Cancer avoir un nombre de TRC détéctés anormalement faible. Ceci peut être dû, soit à une baisse d'expression de certains TRC en Cancer, soit à une augmentation du taux de modifications dans les Tags.

13.6 Analyse des proportions de Tags Amont pour les gènes de protéines ribosomales.

On considère ici qu'un TRC peut être détecté dans une expérience SAGE non seulement par son Tag canonique, mais aussi par son Tag Amont si le site Nla3 du Tag est modifié. Dès lors, nous nous intéressons à l'ensemble des Tags produits pour un gène considéré et à la proportion de Tags canoniques et Tags Amont. Dans ce but, nous avons extrait pour chaque TRC, les Tags Virtuels Amont (TVA) associés au TV canoniques en vérifiant que les TVA ne correspondaient à aucun TV connu (voir chapitre 10.1). Le TVA ne correspondant pas non plus à un variant d'épissage du gène, l'origine d'un TE assigné à un TVA est donc clairement attribuable à une mutation du site Nla3. Nous avons ainsi obtenu 52 couples TV/TVA spécifiques des TRC. Les TE assignés à un de ces couples TV et TVA spécifiques d'un TRC ont été comptabilisés dans les expériences SAGE Normal d'une part, et Cancer d'autre part. Nous avons testé l'influence de l'origine tissulaire du SAGE sur la quantité de TVA observés pour chaque TRC en utilisant le test de Fisher exact (voir chapitre 10.5.5.2). Ce test a été effectué uniquement pour les TRC ayant un nombre de TV et TVA observés suffisant pour réaliser un test statistique. En effet, certains TVA ne sont pas observés et pour certain TRC, le nombre de TV et TVA est trop faible pour pouvoir réaliser une étude statistique. 50 TRC ont ainsi pu être testés et nous avons pu mettre en évidence 20 TRC présentant une quantité de TVA significativement dépendante de l'origine tissulaire (Figure 60).

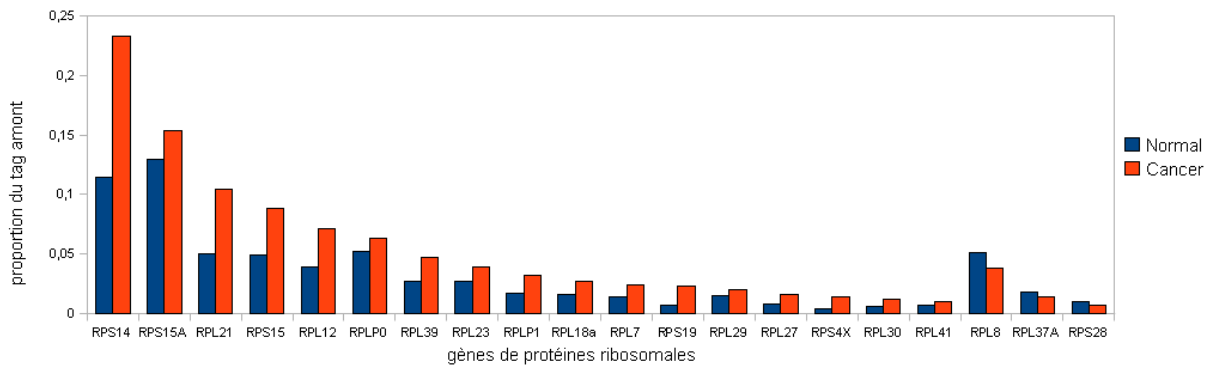


Figure 60 : Gènes de protéines ribosomales cytoplasmiques dont la détection en SAGE présente un taux de Tags Amont significativement différent en Cancer.
Nous considérons qu'un TRC produit un ensemble de Tags dont certains seront des Tags canoniques et d'autres des Tags Amont. Les barres représentent la proportion de Tags Amont observés, pour chacun des TRC, pour les SAGE Normal (en bleu) et Cancer (en orange).

Pour 17 gènes, la proportion de Tags Amont est supérieure en Cancer. Par exemple, les Tags Amont représentent 25% des Tags attribuables aux transcrits codant pour la protéine RPS14 dans les SAGE issus des tissus cancéreux contre 11% au sein des tissus sains. Selon le gène considéré, la différence des proportions des Tags Amont observés en Normal et en Cancer est plus ou moins grande. Elles peuvent être plus que doublées, comme dans le cas de RPS14 (11 et 25%), ou rester proches comme pour RPS15a (12 et 15%). Pour 3 gènes, la proportion de Tags Amont est légèrement plus grande en Normal. On observe une tendance à avoir une différence des proportions assez faible lorsque la proportion de Tags Amont en Normal est supérieure à celle observée en Cancer. Obtenir 17 gènes sur 50 ayant une proportion de Tags Amont plus élevée en Cancer démontre que l'une des sources possibles de l'hétérogénéité des données observées en SAGE Cancer provient de l'augmentation du nombre d'ARNm dont la séquence a été modifiée.

Nous nous permettons de tirer cette conclusion, car nous savons que les TRC sont fortement exprimés dans tous les tissus et que cet effet ne peut pas être expliqué par l'existence de variants alternatifs du gène dont nous avons tenu compte dans notre étude. Nous n'avons pu être aussi précis pour les autres gènes dont on ne connaît pas tous les variants d'épissage et qui peuvent présenter des profils d'expression très différents d'un tissu à l'autres. Cependant, malgré ces limites, nous avons utilisé ce même protocole sur l'ensemble des séquences de la banque d'ADNc de Sagettarius afin d'identifier les gènes présentant ce profil.

13.7 Etude des Tags Amont pour l'ensemble des ADNc

La banque d'ADNc de Sagettarius, contient les séquences de meilleur qualité pour effectuer l'assignation de TE après celle des TRC. En utilisant notre protocole pour identifer les couples TV/TVA uniques, nous avons obtenu 4778 couples TV/TVA extraits sur l'ensemble des ADNc de la banque Sagettarius (voir chapitre 10.5.5.2). Nous avons pu tester l'importance du cancer sur l'apparition du Tag Amont pour 2 120 couples qui présentaient des effectifs suffisants. Les résultats des tests ont mis en évidence 372 couples présentant une proportion de Tags Amont sensiblement augmentée en Cancer et 145 couples présentant une proportion de Tags Amont diminuée en Cancer. Comme pour les TRC, nous voyons des gènes présentant une augmentation des Tag Amont en Cancer et d'autres, moins nombreux,

présentant une diminution des Tags Amont en Cancer. L'« effet cancer » observé sur les Tags Amont des TRC semble bien pouvoir être étendu à tous les gènes!

A partir de ces résultats, nous avons quantifié le taux de croissance de la proportion de Tags Amont, observée sur les SAGE Cancer par rapport à la proportion observée sur les SAGE Normal (Équation 2), lorsque cette proportion était significativement augmentée en Cancer, et inversement lorsque cette proportion était significativement augmenté en Normal (Figure 61).

$$Croissance_{Cancer} = \frac{\left(\frac{TVA}{TV+TVA}\right)_{Cancer} - \left(\frac{TVA}{TV+TVA}\right)_{Normal}}{\left(\frac{TVA}{TV+TVA}\right)_{Normal}}$$

Équation 2 : Taux de croissance de la proportion de Tags Amont en Cancer par rapport au Normal.
Il suffit d'échanger les indices (Cancer/Normal) pour obtenir le taux de croissance du Tags Amont en Normal par rapport au Cancer.

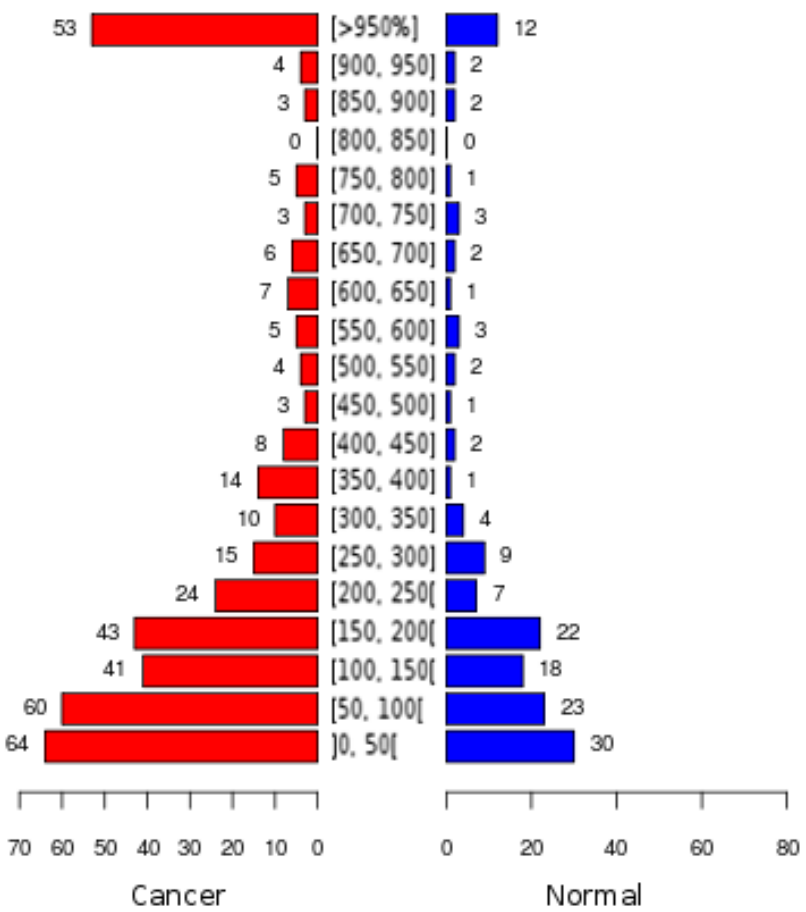


Figure 61: Nombre de gènes ayant un nombre de Tags Amont significativement différents. Gènes présentant une augmentation en Cancer (rouge) ou une augmentation en Normal (bleu) classé selon le taux d'augmentation du Tag Amont.

Dans la suite de l'analyse, nous avons décidé de ne retenir que les 248 gènes présentant une proportion de Tags Amont plus que doublée dans les SAGE Cancer (soit plus de 100% d'augmentation), afin d'essayer de caractériser les gènes les plus affectés dans les tissus cancéreux. Ces gènes sont répertoriés dans l'Annexe D

13.8 Analyse des gènes présentant une augmentation importante des Tags Amont.

Nous avons utilisé le programme DAVID (voir chapitre 9.3) pour vérifier s'il y avait, dans ces 248 gènes, un enrichissement en fonctions cellulaires particulières, mais aucun enrichissement n'est détecté. Ces gènes très modifiés dans les tissus cancéreux ne le sont donc pas à cause d'un lien fonctionnel.

Nous avons également comparé, entre SAGE Normal et Cancer, les niveaux d'expression des 248 gènes (voir chapitre 11) sur la base de la méthode classique (à savoir le décompte des Tags canoniques). Cette analyse a montré que, pour 184 gènes, aucune expression différentielle n'est observée, par contre, 64 gènes semblent sous-exprimés en Cancer et un gène est sur-exprimé en Cancer. Cependant, au-delà du constat que la population considérée semble clairement biaisée, nous ne pouvons objectivement tirer des conclusions robustes sur les sous-expressions, voire sur les expressions similaires, observées. En effet, dans la mesure où les 248 gènes ont été sélectionnés sur la base de taux de Tags Amont anormalement élévés en Cancer, nous ne pouvons exclure que, dans ces gènes particulièrement sensibles, d'autres modifications interviennent à d'autres positions (cf Figure 28) perturbant le décompte des transcrits observés et induisant, en Cancer, une sous-estimation systématique du nombre de Tags assignés.

14 Etude des dépendances entre les positions présentant un taux de modifications supérieur en cancer

Pour aborder les mécanismes menant à l'augmentation de l'hétérogénéité des ARNm dans les cancers, nous avons tout d'abord vérifié si les positions significativement plus modifiées en Cancer ou en Normal étaient indépendantes ou liées. Cette étude a été réalisée sur l'ensemble des EST (environ 2 millions en Cancer et 2 millions en Normal) traités par la société Genclis. Le protocole permettant d'obtenir les positions significativement modifiées est décrit dans le chapitre 10.2. Ainsi, lorsque nous ne considérons qu'un seul type de modification, par exemple les substitutions, nous avons pu observer plusieurs milliers de positions avec un taux de substitutions significativement plus élevé en Cancer qui impliquent environ 1600 transcrits c'est-à-dire un nombre de gènes légérement inférieur car les différents produits d'un gène, incluant tous les variants d'épissage sont pris en compte dans nos études. De la même façon, pour les insertions ou délétions, plusieurs milliers de positions ont été identifiées impliquant environ plusieurs centaines de gènes. Les mêmes analyses effectuées sur les données provenant de tissus sains ont également permis d'identifier des positions et des gènes présentant des taux de modifications significativement élévés. Pour des raisons de confidentialité, les positions, leur environnement qui joue un rôle dans l'émergence de modifications ainsi que les gènes identifiés ne seront pas décrits dans ce manuscrit et nous ne présenterons que la partie des résultats portant sur la mise en évidence de positions modifiées liées dans les tissus cancéreux.

14.1 Résultats des tests d'indépendance

Pour déterminer si des modifications à deux positions sont liées, nous avons compté, pour chaque couple de positions significativement plus modifiées au sein de chaque trancrit, combien d'EST portaient les deux modifications en même temps, combien d'EST n'avaient qu'une modification à une seule position et combien n'en portaient pas. Grâce à ces observations, nous avons testé si les modifications aux deux positions étaient liées ou

indépendantes (test décrit au chapitre 10.5.5). Ce test a été réalisé pour tous les types de modifications (substitution, insertion et délétion) et pour tous les couples possibles de positions ayant un taux de modification significatif (protocole bioinformatique détaillé au chapitre 10.3). Nous définissons un couple de positions comme lié quand le test d'indépendance réalisé est négatif (on rejette l'indépendance) et ce, quel que soit le type de modifications considérées (substitution liée à une substitution, substitution liée à une insertion…). Les résultats en Normal et Cancer sont présentés respectivement dans le Tableau 8 et le Tableau 9.

Nous avons tout d'abord sélectionné les couples de positions ayant des modifications assez fréquentes pour pouvoir être observés en même temps aux deux positions (partie rouge des tableaux). La première observation est que le nombre de couples possibles est plus élevé en Normal qu'en Cancer (1ère colonne). En effet, les positions significativement plus modifiées en Normal qu'en Cancer sont concentrées sur un nombre restreint de gènes, ce qui augmente le nombre de couples possibles sur un même gène.

Lorsque nous regardons le nombre de couples vérifiant les critères stringents que nous avons retenus pour réaliser un test statistique robuste (pour mémoire : au moins 70 EST alignés sur les deux positions, et un effectif théorique strictement supérieur à 1, chapitre 10.5.5.2), le nombre de couples de positions testables est alors beaucoup plus faible en Normal (de 244 à 2188) qu'en Cancer (de 3874 à 105294) (2nde colonne). Ces valeurs indiquent clairement que la plupart des modifications observées dans les tissus sains correspondent à des événements uniques et que les doubles modifications sont presqu'exclusivement observables en Cancer. Cependant, on peut noter que, pour les deux types de tissus, ce sont les couples de modifications impliquant 2 insertions conjointes qui sont nettement moins fréquentes.

Lorsque l'on applique le test de Fisher à ces couples de positions (partie verte des tableaux), on constate alors que les couples liés (4ième colonne) sont largement supérieurs aux nombres de faux positifs attendus (5ième colonne). Les proportions de couples de positions liées varient de 34 à 86% dans l'échantillon Normal et de 26 à 66% pour le Cancer.

.

Tableau 8 : EST Normal

	Nombre de couples possibles	Nombre de couples testables	% de couples testables	Nombre de couples liés	Nombre de faux positifs attendu (LBE) Fisher	% de couples liés	Nombre de séquences de rérérence avec un couple testable	Nombre de séquence de référence avec un couple lié
SUB_VS_SUB	465064	2188	0,47%	1048	92,05	48%	87	59
GAP_VS_GAP	205786	1562	0,76%	1346	7,98	86%	58	58
INS_VS_INS	236554	214	0,09%	92	11,11	43%	11	10
SUB_VS_GAP	588186	1729	0,29%	815	53,16	47%	81	48
SUB_VS_INS	554411	470	0,08%	163	23,98	35%	28	17
GAP_VS_INS	328957	350	0,11%	119	13,24	34%	21	17

Tableau 9 : EST Cancer

	Nombre de couples possibles	Nombre de couples testables	% de couples testables	Nombre de couples liés	Nombre de faux positifs attendu (LBE) Fisher	% de couples liés	Nombre de séquences de rérérence avec un couple testable	Nombre de séquence de référence avec un couple lié
SUB_VS_SUB	283701	50095	17,66%	24093	1566,81	48%	317	263
GAP_VS_GAP	151799	57802	38,08%	38261	1183,02	66%	470	447
INS_VS_INS	40977	3874	9,45%	2115	88,60	55%	53	50
SUB_VS_GAP	386912	105294	27,21%	40709	4062,51	39%	420	307
SUB_VS_INS	176511	24817	14,06%	10373	755,90	42%	110	90
GAP_VS_INS	118287	27999	23,67%	7391	1130,67	26%	115	96

Tests d'indépendance réalisés sur les modifications observées entre deux positions significativement plus modifiées dans les EST Normal (Tableau 8) et Cancer (Tableau 9). Pour chaque couple de modifications possibles, entre substitution (SUB), délétion (GAP) et insertion (INS), est reporté (en rouge) le nombre de couples de positions traitées, le nombre de couples de positions ayant assez d'effectifs pour réaliser le test de Fisher exact, ainsi que le pourcentage que cela représente. Dans les colonnes suivantes, (en vert) le nombre de couples liés d'après le test est reporté, ainsi que le nombre de faux positifs maximum attendu, et le pourcentage de couples liés obtenus. Dans les deux dernières colonnes, (en bleu) est reporté le nombre de séquences de référence qui ont été effectivement testées et le nombre de séquences de référence ayant au moins un couple lié.

Sur le plan statistique, ces fortes valeurs indiquent que les phénomènes sont dépendants et que, pour environ 50% des couples, ces deux modifications auront une tendance à se produire en même temps. Sur le plan biologique, le fait que nous observions des pourcentages de couples liés sensiblement équivalents pour les tissus Normal et Cancer et ce, quelque soit les couples de modifications considérés, nous inclinent à penser que les mécanismes responsables de ces modifications sont sans doute équivalents dans les tissus sains et cancéreux.

Dans une dernière étape (partie bleue des tableaux), nous avons cherché à vérifier que les double-modifications liées n'étaient pas uniquement dues à un nombre restreint de transcrits. Pour ce faire, nous avons caractérisé les nombres de transcrits de références différents ayant au moins un couple de positions testable (colonne 7 et 8). Les valeurs observées montrent clairement que, dans l'échantillon Cancer, les positions liées se répartissent sur plusieurs dizaines à centaines de gènes. Cependant, dans le cas de l'échantillon Normal, seuls quelques dizaines de gènes semblent présentés des doubles événements. Dès lors, pour éviter d'introduire un biais lié à une sous-population de gènes particulière, pour la suite de notre étude, nous nous sommes focalisés sur les données Cancer qui présentent un effectif beaucoup plus robuste statistiquement.

14.2 Analyse du nombre et type de modifications liées par séquence de référence

En cumulant tous les types de modifications liées dans l'échantillon Cancer, nous obtenons un total de 509 transcrits distincts (incluant les séquences de référence et les variants) qui possédent au moins un couple de positions liées. L'analyse des transcrits a montré que les 6 couples de modifications sont possibles sur une même séquence de référence. Nous nous sommes intéressés aux séquences de référence présentant plusieurs couples de positions liées (Figure 62). On constate qu'un nombre important de transcrits (460) présentent au moins deux couples de positions liés et que dans un cas correspondant à un transcrit particulièrement long, on observe jusqu'à plus de 10 000 positions liées. Ces données sont particulièrement intéressantes car elles nous ont permis d'isoler des gènes hyper-sensibles aux modificationx dans les tissus cancéreux. Ces gènes représentent des marqueurs de choix pour la Société Genclis.

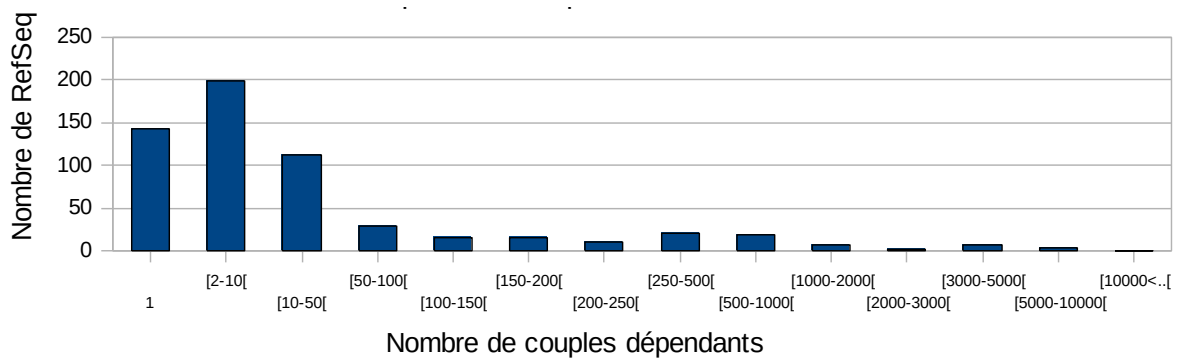


Figure 62 : Nombre de séquences de référence par nombre de couples de positions liées possibles.

14.3 Analyse des distances génomiques séparant deux positions liées

Pour discriminer les mécanismes mis en jeu pouvant impliquer ces modifications liées, nous avons calculé les distances les séparant sur l'ARNm.

Nous avons, quel que soit le type de modifications liées, des distances observées sur les transcrits très proches (Tableau 10), variant, en moyenne, de 93 nucléotides pour deux substitutions liées à 162 nucléotides lorsque deux délétions sont présentes conjointement.

	Type de distances	minimum	1st Quartile	Médiane	3rd Quartile	Maximum	Moyenne	Écart-type
Sub_VS_Sub	transcrit	1	26	60	115	895	93,14	107,38
	gène	1	28	78	299	18354	468,2	1175,14
Gap_VS_Gap	transcrit	1	41	96	221	752	162	166,9
	gène	1	46	182	894	50686	893,3	1997,13
Ins_VS_Ins	transcrit	1	25	61	151	733	118,9	141,45
	gène	1	26	80,5	622,2	19626	501,5	1087,54
Sub_VS_Gap	transcrit	1	31	72	151	819	125,1	144,13
	gène	1	33,75	101	491	58891	627,71	1603,54
Sub_VS_Ins	transcrit	0	35	80	180	895	146,1	170,42
	gène	0	39	149	1077	17804	736,7	1283,99
Gap_VS_Ins	transcrit	1	55	120	265	756	119	173,27
	gène	1	67	283,5	903	1089	23639	1850,03

Tableau 10 : Statistiques descriptives des distances séparant deux positions liées sur l'ARNm et sur le gène (distances génomiques).

Les écarts-types calculés sont très élevés, ce qui indique que les valeurs sont très dispersées et que les moyennes ne sont pas représentatives des distances. Par contre, sur l'analyse des valeurs des médianes, nous observons que, dans 50% des cas, les distances se regroupent autour d'une centaine de nucléotides pour les ARNm (de 60 à 120 nucléotides).
Ces observations portent à penser que les modifications liées sont dues à un phénomène agissant à courte distance.

Cependant, si l'on considère un mécanisme affectant la transcription ou la réplication, nous devons considérer les distances avant épissage, sur le gène. Ainsi, pour chaque couple de positions liées, nous avons calculé la distance génomique qui les sépare en tenant compte des introns (voir chapitre 10.4). Les valeurs des médianes sont alors comprises entre 78 et 283. Le fait que les distances observées sur le génome soient du même ordre de grandeur, indique que les mécanismes impliqués dans l'émergence d'une double modification sont des mécanismes à courte distance, d'environ une centaine de nucléotides

Une distance de l'ordre de 100 nucléotides correspond à une distance fréquemment observée dans divers complexes biologiques impliqués dans des processus liés aux acides nucléiques, que ce soit pour les complexes liés à l'ADN polymérase et impliqués dans le processus de réplication ou pour les complexes liés à l'ARN polymérase et aux processus de transcription ou de réparation.

14.4 Des régions enrichies en mutations dans les EST Cancer.

Pour continuer la caractérisation de nos modifications en Cancer, nous avons étendu le principe de « positions » significativement plus modifiées en Cancer, à celui de fragment, afin de savoir si nous avions des zones où les mutations se concentraient. Pour cela, pour chaque segment possible sur la séquence de référence (depuis chaque position et de toute les tailles possibles), nous allons considérer la fenêtre correspondante sur l'alignement des séquences d'EST et comparer le nombre de modifications observées à celui obtenu pour un modèle aléatoire. Ce modèle aléatoire basé sur une loi hypergéométrique qui est décrite au chapitre 10.5.3.3. Les alignements deux à deux du MegaBLAST contenant les EST issus des échantillons Cancer, qui ont servi de base à la détection des positions plus substituées en Cancer ont été « empilés » en conservant l'alignement sur la séquence de référence grâce à la

librairie JMACS (voir chapitre 12) pour obtenir un alignement multiple de séquences. Le nombre important de décomptes des modifications de nucléotides dans l'alignement, pour chaque segment considéré, a été grandement simplifié en utilisant JMACS fournissant des fonctions intégrées qui ont permis de convertir les bornes des segments indexées sur la séquence de référence, en fenêtres indexées sur les colonnes de l'alignement. Cette étude a été réalisée à ce jour sur chacun des 17 gènes où a été démontrée pour la première fois l'augmentation de l'hétérogénéité des ARNm en Cancer (Brulliard et al., 2007) afin de tester notre démarche sur un premier jeux de données contenant beaucoup d'EST.

Pour l'ensemble des 17 gènes, nous obtenons des segments particulièrement modifiés d'une longueur comprise entre 50 et 200 nucléotides. Toutes les positions qui avaient été annotées comme significativement plus substituées en Cancer sont comprises dans ces segments. Cette distance est du même ordre que nos dépendances et corrobore l'hypothèse de régions « sensibles » où se produirait une accumulation de modifications dans la séquence. Ce résultat change la vision « ponctuelle » des modifications de la séquence des trancrits au sein des tissus cancéreux qui a été suivie durant les différents travaux présentés. Les positions mises en évidence comme étant plus substituées en Cancer, ne sont que les plus souvent affectées parmi toutes les autres d'une région particulièrement sensible. Ce sont dans ces régions que nous pourrons comprendre les caractéristiques qui les rendent si sensibles aux modifications et avec les quelles nous pourront mettre au point un modèle prédictif des séquences alternatives après modifications.

Discussion et perspectives

Ce projet de thèse a été initié suite à la collaboration entre la Société Genclis et le LBGI. Cette collaboration s'est constituée autour de l'observation que les séquences des EST issus de tissus cancéreux étaient plus hétérogènes que les EST issus de tissus sains. Cette hétérogénéité se caractérisait par la présence de nombreuses modifications (mutations, insertions ou délétions) au sein des séquences « cancéreuses ». Ce constat a soulevé un nombre important d'interrogations sur les mécanismes pouvant mener à un tel résultat. Une fois éliminées, par des méthodes statistiques robustes, les hypothèses de modifications uniquement imputables aux erreurs de séquences ou à des phénomènes d'origine exclusivement génomique, l'hypothèse avancée face au fort taux de modifications observées a été l'existence d'un dérèglement d'une ou de plusieurs activités liées au maintien de l'intégrité du message génétique lors de la transcription. L'Infidélité de Transcription est le terme générique utilisé pour identifier l'ensemble de ces mécanismes intervenant durant la transcription et qui pourraient aboutir à un transcrit modifié par rapport à la séquence attendue du gène associé.

Au regard de nos travaux, cet ensemble de mécanismes semble exister aussi bien dans les cellules saines que dans les cellules cancéreuses. Cependant, dans les cancers, l'ampleur du phénomène est multipliée par des facteurs allant d'au moins 3 à plus de 10. Par delà les multiples conséquences biologiques, de tels taux de modifications ont des effets directs sur les études à haut débit, tels le SAGE ou les puces à ADN, qui s'appuient toutes sur le postulat que les ARNm provenant de tissus sains ou cancéreux sont identiques en séquence ou, à tout le moins, que les perturbations introduites dans les séquences sont de faible amplitude.

Durant cette thèse, nous avons entrepris d'estimer et d'étudier l'impact des modifications générées au travers des données provenant des expériences SAGE, car cette technique très sensible, basée sur le séquençage direct (comme les EST), nous permettait potentiellement d'observer les expressions de tous les gènes présents dans une cellule. Cette étude a nécessité de mettre au point des méthodes statistiques et informatiques originales pouvant traiter les données hétérogènes résultant des différentes expériences SAGE. Nous avons été limité par l'hétérogénéité liée à la qualité des données SAGE et aux nombreux types de tissus et types de cancers différents, ce qui nous a contraint à accepter de nombreuses approximations. Cependant, les résultats obtenus suggèrent non seulement, que le nombre d'ARNm ayant une séquence modifiée est bien plus élevé dans les tissus cancéreux que dans des tissus sains mais aussi, que cette hétérogénéité peut biaiser significativement les analyses visant à comparer les données SAGE provenant de ces deux tissus. Ce résultat majeur implique que les expériences

SAGE comparant les deux types de tissus doivent être sérieusement réévalués. Une telle réévaluation pourrait s'appuyer sur la comparaison, pour chaque gène, des taux des Tags Expérimentaux Amont uniques et spécifiques présents dans les tissus sains et cancéreux. Cependant, il est clair que ces estimations d'événements impliquant la modification du seul site de restriction canonique, ne constituent qu'une grossière approximation de l'ensemble des modifications possibles, et que de plus amples études, s'appuyant sur un séquençage massif des ARNm provenant des différents tissus, sont nécessaires pour obtenir une réelle estimation des taux de modification affectant les transcrits d'une cellule cancéreuse.

Comme l'ont montré les analyses réalisées sur la base des séquences d'EST (Brulliard et al., 2007), nos études s'appuyant sur l'analyse des données SAGE ont montré que tous les gènes ne présentaient pas des taux de modifications équivalents. Ainsi, certains gènes présentent de fortes augmentations des rapports Tags Amont/Tags canoniques dans les tissus cancéreux par rapport aux tissus sains et, pour un nombre restreint de gènes, nous avons pu observer une diminution de ce rapport. Ceci a permis d'identifier plus de 200 gènes où l'effet 'Cancer' est particulièrement visible sur les transcrits, gènes qui représentent autant de nouvelles cibles pour l'étude des cancers. Pouvoir prédire quel gène est particulièrement sensible, savoir quelle séquence aura l'ARNm modifié, à quel taux et donc, par extension, quelles séquences protéiques en résulteront constituent autant de nouveaux défis. C'est autour de ces aspects que la Société Genclis a porté son effort pour aboutir à la mise au point de méthodes de diagnostics précoces des cancers qui mettent à profit l'émergence, au sein des cellules cancéreuses, de peptides atypiques résultant des modifications, notamment ceux résultant des sauts de cadre générés par les modifications.

Pour aborder l'étude des mécanismes impliqués dans l'émergence des nombreuses modifications observées au sein des transcrits provenant de cellules cancéreuses, nous avons cherché à mieux caractériser ces modifications sur les séquences d'EST en nous appuyant sur des outils bioinformatiques et statistiques. Ces études ont nécessité le traitement d'alignements de plusieurs milliers d'EST et la mise au point de librairies informatiques optimisant l'exploitation et l'analyse de ces données. Nous avons ainsi démontré qu'une modification nucléotidique n'est pas un évènement isolé, mais s'inscrit plutôt dans un contexte favorisant l'émergence de plusieurs modifications sur une région d'une centaine de bases environ. Ce résultat majeur suggère que les modifications introduites préférentiellement au sein des cellules cancéreuses, ne sont pas fortuites, mais pourraient découler de

mécanismes précis impliquant une centaine de nucléotides qui seraient fortement perturbés dans les cellules cancéreuses.

Cette distance d'une centaine de nucléotides à été observée aussi bien lors de nos études portant sur les couplages statistiques de positions modifiées isolées que lors de l'étude des régions présentant de forts taux de modifications. De plus, la comparaison entre les distances observées sur les transcrits et celles déduites des génomes a révélé une grande similitude entre les deux types de valeurs, ce qui semble militer pour des mécanismes introduisant des modifications à courtes distances. Cependant, nous sommes conscients que ces résultats sont très préliminaires et que cette hypothèse, pour être étayée, nécessite des études beaucoup plus robustes et surtout, de nombreuses confirmations expérimentales.

Il est d'autant plus difficile d'attribuer l'origine des modifications observées à un processus biologique précis que de nombreux mécanismes impliqués dans le maintien de l'information génétique mettent en jeu des distances d'une centaine de nucléotides. C'est le cas des distances séparant deux fragments d'Okasaki durant la réplication ainsi que celles qui sépareraient deux complexes consécutifs lors de la transcription. De même, les derniers travaux suggèrent que le nombre de nucléotides contenus à l'intérieur du complexe transcriptionnel, ce que l'on appelle « la bulle de transcription », et approximativement égal à une soixantaine de nucléotides. En se plaçant au niveau de cette « bulle », de multiples facteurs pourraient expliquer les modifications observées. L'ARNm modifié pourrait acquérir durant son élongation une structure secondaire déstabilisant partiellement les complexes d'élongation, d'épissage ou de réparation. L'ARN polymérase pourrait être bloquée sur la matrice d'ADN, que se soit par les structures précédemment évoquées ou par l'introduction de bases erronées, ce qui favoriserait l'introduction d'une seconde modification en aval ou en amont. Il est également probable que les taux de modifications soient influencés par les niveaux d'expression d'un gène, ce qui pourrait expliquer les sensibités différentes observées selon les gènes.

La plus grande limite pour aborder les mécanismes responsables des modifications observées est que nous ne disposons que de la séquence de l'ARNm qui cumule toutes les modifications éventuellement introduites antérieurement, depuis la duplication du brin d'ADN durant la réplication jusqu'aux nombreux processus de maturation des transcrits. Dès lors, l'obtention en parallèle des séquences d'ARNm et des régions génomiques correspondantes, incluant les introns et exons, de chaque tissu ou cellule représente une étape incontournable pour l'étude

des mécanismes mis en jeu. Ce type d'approche, bien que coûteuse, est désormais tout à fait envisageable grâce aux progrès récents des techniques de séquençage à haut débit. Cependant, malgré toutes ces limitations, il est clair que nos travaux indiquent que, dans les tissus cancéreux, certains facteurs et mécanismes indispensables au maintien de l'intégrité de l'information génétique sont fortement perturbés entraînant une accumulation de mutations dans des zones particulièrement sensibles.

Enfin, nous avons également montré que l'accumulation de modifications sur l'ARNm existe aussi dans les tissus sains, mais en plus faible amplitude. Cette observation confirme les résultats initiaux obtenus sur la base de 17 gènes et conforte l'hypothèse d'un phénomène présent dans les cellules saines et amplifié dans les cancers. Un des éléments militant pour cette hypothèse découle de nos travaux sur l'analyse des modifications liées qui a révélé que, bien qu'ils soient moins fréquents dans les cellules saines, les différents types d'événements doubles (substitutions/substitutions, substitutions/insertions, etc) suivent une distribution similaire dans les tissus sains et cancéreux. Cette surprenante observation incline à penser que ce sont les mêmes mécanismes qui sont mis en jeu dans les deux types de tissus, mécanismes sans doute en lien direct avec le contrôle de l'intégrité de l'information génétique. Dans l'état actuel de nos observations et de nos connaissances, il est difficile de comprendre l'origine d'un tel phénomène et surtout, de lui attribuer un rôle éventuel au sein des cellules saines. En effet, de nombreux travaux montrent, qu'au sein de la cellule, de nombreux mécanismes concourent à l'élimination des erreurs et dans ce contexte, nos observations pourraient ne refléter que les limites de ces mécanismes au sein des cellules saines, limites et de leurs dérèglements au sein des cellules cancéreuses. Cependant, une autre hypothèse, plus iconoclaste, pourrait expliquer nos observations, hypothèse faisant appel à l'existence d'un mécanisme basal, autorisant, voire introduisant, des modifications à différents niveaux du cycle transcriptionnel. Ce genre de mécanismes est déjà connu pour augmenter le nombre d'anticorps différents dans nos cellules, où des modifications dans l'ARN et dans l'ADN vont être conservées dans les cellules souches des anticorps (Steele, 2009). Un tel mécanisme pourrait offrir de nombreux avantages à la cellule en constituant une sorte de test opportun des mécanismes de réparation et un indicateur 'naturel' de l'existence de problèmes au sein de ces mécanismes comme cela est observé dans les cancers ou dans les ultimes étapes de la vie cellulaire. Cette hypothèse, hautement spéculative au regard de nos résultats, s'inscrit clairement dans la vision moderne de la biologie où processus et réseaux biologiques sont fortement intriqués afin d'assurer un contrôle mutuel qui s'appuie sur de multiples signaux et mécanismes souvent convergents mais aussi, parfois, contradictoires.

Annexe A

Implémentation du modèle pour SAGE :

Modèle 1

$$
\begin{aligned}
P(x|M_1) &= \left(\prod_{i=1}^{n} C_{n_i}^{x_i}\right) \frac{\Gamma\left(1+\sum_{i=1}^{n} x_i\right)\Gamma\left(1+\sum_{i=1}^{n}(n_i-x_i)\right)}{\Gamma\left(2+\sum_{i=1}^{n} n_i\right)} \\
\ln P(x|M_1) &= \ln\left[\left(\prod_{i=1}^{n} C_{n_i}^{x_i}\right) \frac{\Gamma\left(1+\sum_{i=1}^{n} x_i\right)\Gamma\left(1+\sum_{i=1}^{n}(n_i-x_i)\right)}{\Gamma\left(2+\sum_{i=1}^{n} n_i\right)}\right] \\
&= \ln\left[\left(\prod_{i=1}^{n} C_{n_i}^{x_i}\right)\right] + \ln\left[\Gamma\left(1+\sum_{i=1}^{n} x_i\right)\right] + \ln\left[\Gamma\left(1+\sum_{i=1}^{n}(n_i-x_i)\right)\right] \\
&\quad - \ln\left[\Gamma\left(2+\sum_{i=1}^{n} n_i\right)\right]
\end{aligned}
$$

Pour chacun des quatre groupes:

1)

$$
\ln\left(\prod_{i=1}^{n} C_{n_i}^{x_i}\right) = \sum_{i=1}^{n} \ln C_{n_i}^{x_i} = \sum_{i=1}^{n} \ln \frac{n_i!}{x_i!(n_i-x_i)!} = \sum_{i=1}^{n}\left[\sum_{a=1}^{n_i} \ln a - \sum_{b=1}^{x_i} \ln b - \sum_{c=1}^{n_i-x_i} \ln c\right]
$$

2)

$$
\ln\left[\Gamma\left(1+\sum_{i=1}^{n} x_i\right)\right] = \ln\left[\left(\sum_{i=1}^{n} x_i\right)!\right] = \sum_{a=1}^{\sum_{i=1}^{n} x_i} \ln a
$$

3)

$$
\ln\left[\Gamma\left(1+\sum_{i=1}^{n} n_i - x_i\right)\right] = \ln\left[\left(\sum_{i=1}^{n}(n_i-x_i)\right)!\right] = \sum_{b=1}^{\sum_{i=1}^{n} ni-x_i} \ln b
$$

4)

$$
\ln\left[\Gamma\left(2+\sum_{i=1}^{n} n_i\right)\right] = \ln\left[\left(1+\sum_{i=1}^{n} n_i\right)!\right] = \sum_{c=1}^{1+\sum_{i=1}^{n} n_i} \ln c
$$

Modèle 2

$$
\begin{aligned}
P(x|M_2) &= \left(\prod_{i=1}^{n} C_{n_i}^{x_i}\right) \frac{\Gamma\left(1+\sum_{i=1}^{m} x_i\right)\Gamma\left(1+\sum_{i=1}^{m}(n_i-x_i)\right)}{\Gamma\left(2+\sum_{i=1}^{m} n_i\right)} \\
&\quad * \frac{\Gamma\left(1+\sum_{i=m+1}^{n} x_i\right)\Gamma\left(1+\sum_{i=m+1}^{n}(n_i-x_i)\right)}{\Gamma\left(2+\sum_{i=m+1}^{n} n_i\right)} \\
\ln P(x|M_2) &= \ln\left[\left(\prod_{i=1}^{n} C_{n_i}^{x_i}\right)\right] + \ln\left[\Gamma\left(1+\sum_{i=1}^{m} x_i\right)\right] + \ln\left[\Gamma\left(1+\sum_{i=1}^{m}(n_i-x_i)\right)\right] \\
&\quad + \ln\left[\Gamma\left(1+\sum_{i=m+1}^{n} x_i\right)\right] + \ln\left[\Gamma\left(1+\sum_{i=m+1}^{n}(n_i-x_i)\right)\right] \\
&\quad - \ln\left[\Gamma\left(2+\sum_{i=1}^{m} n_i\right)\right] - \ln\left[\Gamma\left(2+\sum_{i=m+1}^{n} n_i\right)\right]
\end{aligned}
$$

Pour chacun des sept groupes:

1) identique au modèle 1.

2)

$$\ln\left[\Gamma\left(1+\sum_{i=1}^{m} x_i\right)\right] = \ln\left[\left(\sum_{i=1}^{m} x_i\right)!\right] = \sum_{d=1}^{\sum_{i=1}^{m} x_i} \ln d$$

3)

$$\ln\left[\Gamma\left(1+\sum_{i=1}^{m}(n_i-x_i)\right)\right] = \ln\left[\left(\sum_{i=1}^{m}(n_i-x_i)\right)!\right] = \sum_{f=1}^{\sum_{i=1}^{m}(n_i-x_i)} \ln f$$

4)

$$\ln\left[\Gamma\left(1+\sum_{i=m+1}^{n} x_i\right)\right] = \ln\left[\left(\sum_{i=m+1}^{n} x_i\right)!\right] = \sum_{g=1}^{\sum_{i=m+1}^{n} x_i} \ln g$$

5)

$$\ln\left[\Gamma\left(1+\sum_{i=m+1}^{n}(n_i-x_i)\right)\right] = \ln\left[\left(\sum_{i=m+1}^{n}(n_i-x_i)\right)!\right] = \sum_{h=1}^{\sum_{i=m+1}^{n}(n_i-x_i)} \ln h$$

6)

$$\ln\left[\Gamma\left(2+\sum_{i=1}^{m} n_i\right)\right] = \ln\left[\left(1+\sum_{i=1}^{m} n_i\right)!\right] = \sum_{k=1}^{1+\sum_{i=1}^{m} n_i} \ln k$$

7)

$$\ln\left[\Gamma\left(2+\sum_{i=1}^{m} n_i\right)\right] = \ln\left[\left(1+\sum_{i=m+1}^{n} n_i\right)!\right] = \sum_{l=1}^{1+\sum_{i=m+1}^{n} n_i} \ln l$$

Chaque groupe est calculé en parallèle. Les groupes sont ensuite ajouté (ou soustrait) pour obtenir le score.

Annexe B

Exemple du contenu d'un fichier d'alignement multiple MACSIMS.xml

```
<macsim>

 <alignment>
 <aln-name>ref3/test/1ldg_ref3</aln-name>

 <sequence seq-type="Protein">
        <seq-name>1ldg_</seq-name>

        <seq-info>
                <accession>1ldg_</accession>
                <nid>1ldg_</nid>
                <ec>0.0.0.0</ec>
                <group>0</group>

                <ftable>

                  <fitem>
                        <ftype>STRUCT</ftype>
                        <fstart>5</fstart>
                        <fstop>9</fstop>
                        <fcolor>0</fcolor>
                        <fscore>0.00</fscore>
                        <fnote>STRAND</fnote>
                  </fitem>

                  <fitem>
                        <ftype>STRUCT</ftype>
                        <fstart>13</fstart>
                        <fstop>24</fstop>
                        <fcolor>3</fcolor>
                        <fscore>0.00</fscore>
                        <fnote>HELIX</fnote>
                  </fitem>
                </ftable>
                <length>492</length>
                <weight>100</weight>
        </seq-info>

        <seq-data>
                ------------------------------------------------------------------------------------------------
                apkakivlvgs-gmiggvmatlivqkn------lgdvvlfdi-------vknmphgkaldtshtnvm---snckvsgs------
                ntyddlagsdvvivtagftkew--------nrldllplnnkimieigghikknc---afiivvtnpvdvmvqllhqhsg-
                vpknkiiglggvldtsrlkyyisqklnvcprdvn-ahivgahgnkmvllkryitv------efinnklis--------daele-
                aifdrtvntaleivnlh--aspyvapaaaiiemaesylkd--lkkvlicstlleg----qyg---hsdifggtpvvlgangveqv-
                ielqlnseekakfdeaiaetkrmkala-----------------------
        </seq-data>
 </sequence>
 <sequence seq-type="Protein">
.......

 </sequence>
 <column-score>
 <colsco-name>coreblock</colsco-name>
 <colsco-owner>0</colsco-owner>
 <colsco-type>int</colsco-type>
```

<colsco-data>

-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -
1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -
1 -1 -1 -1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 0 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1
1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 0 0 0 0 0 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
</colsco-data>
</column-score>
</alignment>
</macsim>

Annexe C

Exemple de construction d'un modèle linéaire avec R.

Soit nos données stocké dans la table « data », avec le nombre de différent type de TE assigné dans sa colonne « mapped » et le nombre de Tags séquencés dans sa colonne « Sequenced_ET ».
Voici le code affichant un modèle linéaire entre les variable « séquences de TE » et « nombre de Tags séquencés ».

```
%affiche les points
plot(data$mapped ~ data$Sequenced_ET, ylab = "Nombre de séquences différentes de TE assignés", xlab ="Nombre de Tags séquencés")
%Calcul le modèle linéaire en stipulant que pour 0 Tags séquencés nous avanons 0 TE assigné
regressionLinMapped <- lm(formula = data$mapped ~ data$Sequenced_ET +0)
%Prédit les points extrèmes à partir du modèle linéaire.
XmappedPredict <- predict(regressionLinMapped,interval="prediction", level =0.95)
%affiche la droite de regression
abline(regressionLinMapped,lwd=2)
%affiche la droite de confiance inférieur
lines(data$Sequenced_ET,XmappedPredict[,"lwr"],lwd = 2)
%affiche la droite de confiance supérieur
lines(data$Sequenced_ET,XmappedPredict[,"upr"],lwd = 2")
```

Annexe D

Liste des gènes dont la proportion de Tag Amont détecté est augmentée en cancer.

GENBANK_ACCESSION	Gene Name
AF113124	FASCICULATION AND ELONGATION PROTEIN ZETA 2 (ZYGIN II)
AF193047	HLA-B ASSOCIATED TRANSCRIPT 5
AB061546	SIGNAL RECOGNITION PARTICLE 14KDA (HOMOLOGOUS ALU RNA BINDING PROTEIN)
AY848700	MYELOID LEUKEMIA FACTOR 1
AF340151	SH3 DOMAIN BINDING GLUTAMIC ACID-RICH PROTEIN LIKE 2
AY232654	GRANZYME B (GRANZYME 2, CYTOTOXIC T-LYMPHOCYTE-ASSOCIATED SERINE ESTERASE 1)
AF015524	CHEMOKINE (C-C MOTIF) RECEPTOR-LIKE 2
AY729650	INTERSEX-LIKE (DROSOPHILA)
AF133732	FOUR AND A HALF LIM DOMAINS 3
AB032427	TRANSIENT RECEPTOR POTENTIAL CATION CHANNEL, SUBFAMILY V, MEMBER 4
AF285596	HYPOTHETICAL PROTEIN FLJ20416
AF328684	C-TYPE LECTIN DOMAIN FAMILY 4, MEMBER A
AF109146	C-TYPE LECTIN DOMAIN FAMILY 4, MEMBER A
U02683	NUCLEAR RESPIRATORY FACTOR 1
AF240633	MYOZENIN 1
AF331521	SOLUTE CARRIER FAMILY 26, MEMBER 7
AY358884	SLIT HOMOLOG 3 (DROSOPHILA)
AB007191	C-MYC BINDING PROTEIN
AY082014	RELAXIN-LIKE FACTOR B
AY359050	GASTROKINE 1
AL832452	CARDIOMYOPATHY ASSOCIATED 3
AY358978	KSP37 PROTEIN
AF217982	CDK5 REGULATORY SUBUNIT ASSOCIATED PROTEIN 3
AF155658	SIMILAR TO C. ELEGENS HYPOTHETICAL 55.2 KD PROTEIN F16A11.2
AY578063	NUCLEAR PROTEIN UKP68
AF050641	NADH DEHYDROGENASE (UBIQUINONE) 1 ALPHA SUBCOMPLEX, 9, 39KDA
AF124993	PEROXIREDOXIN 5
AF258583	RAB24, MEMBER RAS ONCOGENE FAMILY
AF183420	METHIONINE SULFOXIDE REDUCTASE A
AY027862	EPENDYMIN RELATED PROTEIN 1 (ZEBRAFISH)
AF332197	SINE OCULIS HOMEOBOX HOMOLOG 2 (DROSOPHILA)
AF164791	AHA1, ACTIVATOR OF HEAT SHOCK 90KDA PROTEIN ATPASE HOMOLOG 1 (YEAST)
AF458592	CHROMOSOME 3 OPEN READING FRAME 57
AF104629	ATPASE, H+ TRANSPORTING, LYSOSOMAL 34KDA, V1 SUBUNIT D
AF319553	TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY, MEMBER 19-LIKE
AF144094	MYOSIN XVA
AF096834	Y BOX BINDING PROTEIN 2

BX537381	RHODOPSIN (OPSIN 2, ROD PIGMENT) (RETINITIS PIGMENTOSA 4, AUTOSOMAL DOMINANT)
AF242517	PEPTIDOGLYCAN RECOGNITION PROTEIN 1
AF064200	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B4
AF068754	HEAT SHOCK FACTOR BINDING PROTEIN 1
AF114833	NEURITIN 1
AF364517	ERYTHROID ASSOCIATED FACTOR
AF440434	MAP-KINASE ACTIVATING DEATH DOMAIN
AF053944	AE BINDING PROTEIN 1
AY358483	HYPOTHETICAL PROTEIN MGC45438
AL832639	MYOSIN, HEAVY POLYPEPTIDE 9, NON-MUSCLE
AY358837	CD99 ANTIGEN-LIKE 2
AF269223	T-COMPLEX 11 (MOUSE)
AY827490	FILAGGRIN 2
AF245044	CHROMOSOME 6 OPEN READING FRAME 200
AY321367	DKFZP586J0917 PROTEIN
AF153608	SIN3A-ASSOCIATED PROTEIN, 18KDA
AF464935	COFACTOR OF BRCA1
AF401520	HIGH MOBILITY GROUP NUCLEOSOMAL BINDING DOMAIN 3
AY964667	IQ MOTIF CONTAINING B1
AF297872	TRANSCRIPTIONAL REGULATING FACTOR 1
AF063301	KERATOCAN
AB032255	DKFZP434H071 PROTEIN
AF152929	A KINASE (PRKA) ANCHOR PROTEIN 7
AB063321	CAMP RESPONSIVE ELEMENT BINDING PROTEIN 3-LIKE 1
AF084457	COATOMER PROTEIN COMPLEX, SUBUNIT BETA
AF332193	EXONUCLEASE NEF-SP
AY256461	INTERFERON-STIMULATED TRANSCRIPTION FACTOR 3, GAMMA 48KDA
AY358564	CYSTEINE-RICH SECRETORY PROTEIN LCCL DOMAIN CONTAINING 1
AY009106	REGULATOR OF G-PROTEIN SIGNALLING 22
AF395889	CLUSTERIN-LIKE 1 (RETINAL)
AF225419	HSCARG PROTEIN
AF083190	DNAJ (HSP40) HOMOLOG, SUBFAMILY C, MEMBER 8
AF212371	SPINSTER
AF213465	DUAL OXIDASE 1
AF435958	TETRATRICOPEPTIDE REPEAT DOMAIN 18
AF264785	HAIRY AND ENHANCER OF SPLIT 1, (DROSOPHILA)
AF316855	HYPOTHETICAL PROTEIN FLJ22795
DQ145726	ARNT-INTERACTING PROTEIN 2
AB047786	SNF1-LIKE KINASE
AF321876	DPH1 HOMOLOG (S. CEREVISIAE)
AF043472	POTASSIUM VOLTAGE-GATED CHANNEL, DELAYED-RECTIFIER, SUBFAMILY S, MEMBER 3
AB094095	NOD9 PROTEIN
U03626	ARRESTIN 3, RETINAL (X-ARRESTIN)
AF023612	THIOREDOXIN-LIKE 4A
AF023477	ADAM METALLOPEPTIDASE DOMAIN 12 (MELTRIN ALPHA)
AF506289	G PROTEIN-COUPLED RECEPTOR, FAMILY C, GROUP 5, MEMBER A

AF386504	HYPOTHETICAL PROTEIN MGC15416
AY960291	ALBUMIN
AF328731	TAO KINASE 3
AY376242	RIBOSOMAL PROTEIN L9
AF263452	EUKARYOTIC TRANSLATION INITIATION FACTOR 1B
AF172993	PALATE, LUNG AND NASAL EPITHELIUM CARCINOMA ASSOCIATED
BX640623	IMMUNOGLOBULIN HEAVY CONSTANT GAMMA 2 (G2M MARKER)
AF265206	RAN GUANINE NUCLEOTIDE RELEASE FACTOR
AF052955	ATP SYNTHASE, H+ TRANSPORTING, MITOCHONDRIAL F1 COMPLEX, EPSILON SUBUNIT
AY279380	KALLIKREIN 5
AF082569	CYCLIN D-TYPE BINDING-PROTEIN 1
AF067226	PHOSPHODIESTERASE 9A
AF214731	DEAD (ASP-GLU-ALA-ASP) BOX POLYPEPTIDE 24
AF119226	DUAL SPECIFICITY PHOSPHATASE 12
AF139540	INTRAFLAGELLAR TRANSPORT 81 HOMOLOG (CHLAMYDOMONAS)
AY013295	MYOZENIN 2
S82198	CHYMOTRYPSIN C (CALDECRIN)
AF258591	MORF4 FAMILY ASSOCIATED PROTEIN 1-LIKE 1
AF506820	PLECKSTRIN HOMOLOGY-LIKE DOMAIN, FAMILY B, MEMBER 2
AF277187	PROTEIN TYROSINE PHOSPHATASE, MITOCHONDRIAL 1
BX647064	SEVEN IN ABSENTIA HOMOLOG 1 (DROSOPHILA)
AF053070	NADH DEHYDROGENASE (UBIQUINONE) FLAVOPROTEIN 1, 51KDA
AB025432	TSC22 DOMAIN FAMILY, MEMBER 3
AF020202	UNC-13 HOMOLOG B (C. ELEGANS)
AF141332	APOLIPOPROTEIN B48 RECEPTOR
AY245432	EUKARYOTIC TRANSLATION INITIATION FACTOR 3, SUBUNIT 12
AF092441	POLY(RC) BINDING PROTEIN 4
AF414442	HYPOTHETICAL PROTEIN FLJ14303
AF120266	TETRASPANIN 15
AF328905	STROMAL INTERACTION MOLECULE 2
AF133587	RHABDOID TUMOR DELETION REGION GENE 1
AF361746	HUEL (C4ORF1)-INTERACTING PROTEIN
AF417114	CHROMOSOME 1 OPEN READING FRAME 41
AF429969	CHROMOSOME 10 OPEN READING FRAME 9
AF345910	NYD-SP14 PROTEIN
AF479813	CKLF-LIKE MARVEL TRANSMEMBRANE DOMAIN CONTAINING 3
AF333389	DEOXYRIBONUCLEASE II BETA
AY359046	DEHYDROGENASE/REDUCTASE (SDR FAMILY) MEMBER 9
AB043104	NUCLEOLAR PROTEIN FAMILY A, MEMBER 3 (H/ACA SMALL NUCLEOLAR RNPS)
AF261089	CALPAIN 2, (M/II) LARGE SUBUNIT
AF076464	PHOSDUCIN
AF076465	PHOSDUCIN
AF052124	SECRETED PHOSPHOPROTEIN 1 (OSTEOPONTIN, BONE SIALOPROTEIN I, EARLY T-LYMPHOCYTE ACTIVATION 1)
AF132947	G PROTEIN-COUPLED RECEPTOR 89A
AF099137	POTASSIUM LARGE CONDUCTANCE CALCIUM-ACTIVATED CHANNEL, SUBFAMILY M, BETA MEMBER 2

AF331796	CHROMOSOME CONDENSATION PROTEIN G
AF214680	RING FINGER PROTEIN 141
AY072034	CHROMOSOME 14 OPEN READING FRAME 8
AF275258	P ANTIGEN FAMILY, MEMBER 4 (PROSTATE ASSOCIATED)
BX538343	SERPIN PEPTIDASE INHIBITOR, CLADE B (OVALBUMIN), MEMBER 6
AF063605	LEPTIN RECEPTOR OVERLAPPING TRANSCRIPT-LIKE 1
AY358674	LEPTIN RECEPTOR OVERLAPPING TRANSCRIPT-LIKE 1
AF166331	CRYSTALLIN, BETA A2
AF087853	GROWTH ARREST AND DNA-DAMAGE-INDUCIBLE, BETA
AF288208	UDP-GLCNAC:BETAGAL BETA-1,3-N-ACETYLGLUCOSAMINYLTRANSFERASE 1
AY174127	CKLF-LIKE MARVEL TRANSMEMBRANE DOMAIN CONTAINING 1
AF060228	RETINOIC ACID RECEPTOR RESPONDER (TAZAROTENE INDUCED) 3
AF092922	RETINOIC ACID RECEPTOR RESPONDER (TAZAROTENE INDUCED) 3
AF089106	HEMATOPOIETIC STEM/PROGENITOR CELLS 176
AF140501	POLYMERASE (DNA DIRECTED) IOTA
AY074488	ARGINASE, LIVER
AF001979	CHEMOKINE (C-C MOTIF) LIGAND 21
AF112217	UBIQUINOL-CYTOCHROME C REDUCTASE COMPLEX (7.2 KD)
AY509193	HEMOGLOBIN, BETA
AF272151	CHROMOSOME 6 OPEN READING FRAME 5
AY358433	DMC
AF130736	NUCLEAR DISTRIBUTION GENE C HOMOLOG (A. NIDULANS)
AF121260	CYSTEINE AND GLYCINE-RICH PROTEIN 3 (CARDIAC LIM PROTEIN)
AF074002	LECTIN, GALACTOSIDE-BINDING, SOLUBLE, 8 (GALECTIN 8)
BX537392	PURINERGIC RECEPTOR P2Y, G-PROTEIN COUPLED, 5
AF039103	HIV-1 TAT INTERACTIVE PROTEIN 2, 30KDA
AF144745	GUANINE DEAMINASE
AF030555	ACYL-COA SYNTHETASE LONG-CHAIN FAMILY MEMBER 4
AF065388	TETRASPANIN 1
AY268104	CARBOXYLESTERASE 1 (MONOCYTE/MACROPHAGE SERINE ESTERASE 1)
AB011031	PHD FINGER PROTEIN 11
AB062396	TUMOR PROTEIN D52-LIKE 2
AY157580	TRANSMEMBRANE PROTEIN 123
AF125045	UBIQUITIN-CONJUGATING ENZYME E2D 4 (PUTATIVE)
AY359035	MESODERM INDUCTION EARLY RESPONSE 1 HOMOLOG (XENOPUS LAEVIS)
AF276808	LIM DOMAIN BINDING 3
AB046613	MYOSIN, LIGHT POLYPEPTIDE 6, ALKALI, SMOOTH MUSCLE AND NON-MUSCLE
AF210057	CHROMOSOME 3 OPEN READING FRAME 1
AY522342	HOMEOBOX CONTAINING 1
AF494509	DOWN-REGULATED IN GASTRIC CANCER GDDR
AY424284	PROGESTIN AND ADIPOQ RECEPTOR FAMILY MEMBER VI
AB033284	SOLUTE CARRIER FAMILY 12 (POTASSIUM/CHLORIDE TRANSPORTERS), MEMBER 9
AF157316	CHROMOSOME 5 OPEN READING FRAME 5
BX640689	TROPONIN T TYPE 3 (SKELETAL, FAST)
AF087873	PROTEIN KINASE (CAMP-DEPENDENT, CATALYTIC) INHIBITOR BETA
AF026851	PET112-LIKE (YEAST)
AF221842	CHROMOSOME 20 OPEN READING FRAME 14
AF331038	TBC1 DOMAIN FAMILY, MEMBER 10A

AF155660	SOLUTE CARRIER FAMILY 25, MEMBER 37
AF043498	PROSTATE STEM CELL ANTIGEN
AB030000	INTERLEUKIN 23, ALPHA SUBUNIT P19
AF305225	APOLIPOPROTEIN L, 2
BX537418	TUMOR PROTEIN P53 BINDING PROTEIN, 1
AF095287	PITUITARY TUMOR-TRANSFORMING 1
AF061736	UBIQUITIN-CONJUGATING ENZYME E2L 6
AF226053	SET AND MYND DOMAIN CONTAINING 2
AF131214	SORTING NEXIN 9
AY903446	ANKYRIN REPEAT DOMAIN 1 (CARDIAC MUSCLE)
AF528099	TRANSFORMING, ACIDIC COILED-COIL CONTAINING PROTEIN 2
AF004230	LEUKOCYTE IMMUNOGLOBULIN-LIKE RECEPTOR, SUBFAMILY B (WITH TM AND ITIM DOMAINS), MEMBER 1
AY358201	ABHYDROLASE DOMAIN CONTAINING 14A
AF272894	UBX DOMAIN CONTAINING 1
AF261072	MEDIATOR OF RNA POLYMERASE II TRANSCRIPTION, SUBUNIT 25 HOMOLOG (YEAST)
AF131802	CHROMOSOME 2 OPEN READING FRAME 25
AF026947	ALDO-KETO REDUCTASE FAMILY 7, MEMBER A2 (AFLATOXIN ALDEHYDE REDUCTASE)
AF106966	BRAIN PROTEIN I3
AF171938	NUMB HOMOLOG (DROSOPHILA)
AY359114	HYPOTHETICAL PROTEIN MGC4825
AF262024	PRAJA 1
AF073312	SEPTIN 4
AF085235	SEPTIN 1
AY461712	ATPASE TYPE 13A2
AB033823	TEKTIN 2 (TESTICULAR)
AF276423	ERBB2 INTERACTING PROTEIN
AF530581	POLYPYRIMIDINE TRACT BINDING PROTEIN 2
AF273049	ZINC FINGER PROTEIN 638
AF526533	DEVELOPMENTALLY REGULATED RNA-BINDING PROTEIN 1
AY317139	RETINOIC ACID INDUCED 14
AF279897	CDC28 PROTEIN KINASE REGULATORY SUBUNIT 1B
AY103469	CRUMBS HOMOLOG 3 (DROSOPHILA)
AL832642	CD44 ANTIGEN (INDIAN BLOOD GROUP)
AY427823	HEME BINDING PROTEIN 2
AF146019	HEPATOCELLULAR CARCINOMA ANTIGEN GENE 520
AF251039	JUMONJI DOMAIN CONTAINING 1B
AY507846	CHROMOSOME 16 OPEN READING FRAME 63
AF176006	MITOCHONDRIAL CARRIER HOMOLOG 1 (C. ELEGANS)
AF268037	CHROMOSOME 8 OPEN READING FRAME 4
AY196212	ANKYRIN REPEAT DOMAIN 23
AB038158	SERINE PROTEASE TADG12
AF054589	MYOD FAMILY INHIBITOR DOMAIN CONTAINING
AF000652	SYNDECAN BINDING PROTEIN (SYNTENIN)
AF245436	HYPOTHETICAL PROTEIN FLJ23518
AF389338	STEAROYL-COA DESATURASE 5

AF053453	TETRASPANIN 6
AF494381	FAMILY WITH SEQUENCE SIMILARITY 3, MEMBER D
AF087892	COMPLEMENT COMPONENT 1, Q SUBCOMPONENT, C CHAIN
AB098731	MUCIN 20
AY358469	PHOSPHOLIPASE INHIBITOR
AF036268	SH3-DOMAIN GRB2-LIKE 2
AF196834	TWISTED GASTRULATION HOMOLOG 1 (DROSOPHILA)
AF242521	ORNITHINE DECARBOXYLASE ANTIZYME 2
AF539739	S100 CALCIUM BINDING PROTEIN P
AF282851	HYPOTHETICAL PROTEIN FLJ20467
AF034091	NUCLEAR LOCALIZATION SIGNAL DELETED IN VELOCARDIOFACIAL SYNDROME
AB016068	ZINC METALLOPEPTIDASE (STE24 HOMOLOG, YEAST)
AF037204	RING FINGER PROTEIN 13
AB055804	PREFOLDIN SUBUNIT 5
AF164679	RING FINGER PROTEIN 7
AF277182	CHROMOSOME 18 OPEN READING FRAME 21
AY055826	LECTIN, GALACTOSIDE-BINDING, SOLUBLE, 13 (GALECTIN 13)
AF188698	SULFOTRANSFERASE FAMILY 4A, MEMBER 1
AF068651	LIM DOMAIN BINDING 2
AF334406	REGULATOR OF CHROMOSOME CONDENSATION (RCC1) AND BTB (POZ) DOMAIN CONTAINING PROTEIN 1
AY839725	
EU520261	
EF489426	
DQ679794	
EF488978	
DQ338435	
DQ132786	

Références Bibliographiques

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, *287*(5461), 2185–2195.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., et al. (1991). Complementary DNA sequencing: expressed sequence Tags and human genome project. *Science (New York, N.Y.)*, *252*(5013), 1651–1656.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1006/jmbi.1990.9999.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Anderson, S., Bankier, A. T., Barrell, B. G., Bruijn, M. H. D., Coulson, A. R., Drouin, J., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806), 457–465.

ANFINSEN, C. B., & REDFIELD, R. R. (1956). Protein structure in relation to function and biosynthesis. *Advances in Protein Chemistry*, *11*, 1-100.

Aphasizhev, R. (2007). RNA editing. *Molecular Biology*, *41*(2), 227-239. doi: 10.1134/S0026893307020057.

Armache, K., Kettenberger, H., & Cramer, P. (2005). The dynamic machinery of mRNA elongation. *Current Opinion in Structural Biology*, *15*(2), 197-203. doi: 10.1016/j.sbi.2005.03.002.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. doi: 10.1038/75556.

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES: INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J. Exp. Med.*, *79*(2), 137-158. doi: 10.1084/jem.79.2.137.

Benjamini Y,, & Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing., 57,1:289–300.

Bianchetti, L., Wu, Y., Guerin, E., Plewniak, F., & Poch, O. (2007). SAGETTARIUS: a program to reduce the number of Tags mapped to multiple transcripts and to plan SAGE sequencing stages. *Nucleic Acids Res*, *35*(18), e122. doi: 10.1093/nar/gkm648.

Boon, K., Osório, E. C., Greenhut, S. F., Schaefer, C. F., Shoemaker, J., Polyak, K., et al. (2002). An anatomy of normal and malignant gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11287–11292. doi: 10.1073/pnas.152324199.

Brulliard, M., Lorphelin, D., Collignon, O., Lorphelin, W., Thouvenot, B., Gothié, E., et al. (2007). Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis. *Proc Natl Acad Sci U S A*, *104*(18), 7522–7527. doi: 10.1073/pnas.0611076104.

Buratowski, S. (2005). Connections between mRNA 3' end processing and transcription termination. *Current Opinion in Cell Biology*, *17*(3), 257-261. doi: 10.1016/j.ceb.2005.04.003.

Chakravarti, A. (2001). Single nucleotide polymorphisms: . . .to a future of genetic medicine. *Nature*, *409*(6822), 822-823. doi: 10.1038/35057281.

Chambers, J. (1993). *Statistical models in S*. New York: Champan & Hall.

Chung, E. J., Sung, Y. K., Farooq, M., Kim, Y., Im, S., Tak, W. Y., et al. (2002). Gene

expression profile analysis in human hepatocellular carcinoma by cDNA microarray. *Mol Cells*, *14*(3), 382–387.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., et al. (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Research*, *37*(Database issue), D19-25. doi: 10.1093/nar/gkn765.

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Research*, *8*(12), 1229-1231. doi: 10.1101/gr.8.12.1229.

Cooper, G. M. (1999). *La cellule*. De Boeck Université.

CRICK, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, *12*, 138-163.

Cunha, G. R., Cooke, P. S., & Kurita, T. (2004). Role of stromal-epithelial interactions in hormonal responses. *Archives of Histology and Cytology*, *67*(5), 417-434.

Dalmasso, C., Broët, P., & Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics (Oxford, England)*, *21*(5), 660-668. doi: 10.1093/bioinformatics/bti063.

Dayhoff, M. O. (1965). Computer aids to protein sequence determination. *Journal of Theoretical Biology*, *8*(1), 97-112.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, *4*(9), R60. doi: 10.1186/gb-2003-4-9-r60.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). (pas de date). Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc., 96, 1151?1160.

Etzold, T., Ulyanov, A., & Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods in Enzymology*, *266*, 114-128.

Finn, R., Griffiths-Jones, S., & Bateman, A. (2003). Identifying protein domains with the Pfam database. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al*, *Chapter 2*, Unit 2.5. doi: 10.1002/0471250953.bi0205s01.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science (New York, N.Y.)*, *155*(760), 279-284.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, *269*(5223), 496–512.

Fujii, T., Dracheva, T., Player, A., Chacko, S., Clifford, R., Strausberg, R. L., et al. (2002). A preliminary transcriptome map of non-small cell lung cancer. *Cancer Res*, *62*(12), 3340–3346.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. doi: 10.1186/gb-2004-5-10-r80.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., & di Paola, G. (1985). ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Computer Applications in the Biosciences: CABIOS*, *1*(3), 167-172.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, *446*(7132), 153–158. doi: 10.1038/nature05610.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70.

Hanawalt, P. C., & Spivak, G. (2008). Transcription-coupled DNA repair: two decades of progress and surprises. *Nature Reviews. Molecular Cell Biology*, *9*(12), 958-970. doi: 10.1038/nrm2549.

Helmut Pospiech. (2002). The role of DNA polymerases, in particular DNA polymerase

ε in DNA repair and replication. Text.Thesis.Doctoral, . Retrouvé Septembre 8, 2009, de http://herkules.oulu.fi/isbn9514266692/.

Holland, R. C. G., Down, T. A., Pocock, M., Prlic, A., Huen, D., James, K., et al. (2008). BioJava: an open-source framework for bioinformatics. *Bioinformatics*, *24*(18), 2096-2097. doi: 10.1093/bioinformatics/btn397.

Isken, O., & Maquat, L. E. (2007). Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes & Development*, *21*(15), 1833-3856. doi: 10.1101/gad.1566807.

Iyer, R. R., Pluciennik, A., Burdett, V., & Modrich, P. L. (2006). DNA Mismatch Repair: Functions and Mechanisms. *Chemical Reviews*, *106*(2), 302-323. doi: 10.1021/cr0404794.

Jenks, M. H., O'Rourke, T. W., & Reines, D. (2008). Properties of an intergenic terminator and start site switch that regulate IMD2 transcription in yeast. *Molecular and Cellular Biology*, *28*(12), 3883-3893. doi: 10.1128/MCB.00380-08.

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. doi: 10.1101/gr.229202.

Koyama, H., Ito, T., Nakanishi, T., Kawamura, N., & Sekimizu, K. (2003). Transcription elongation factor S-II maintains transcriptional fidelity and confers oxidative stress resistance. *Genes to Cells*, *8*(10), 779-788. doi: 10.1046/j.1365-2443.2003.00677.x.

Kulesh, D. A., Clive, D. R., Zarlenga, D. S., & Greene, J. J. (1987). Identification of interferon-modulated proliferation-related cDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(23), 8453-8457.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi: 10.1038/35057062.

Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J.,

et al. (2000). SAGEmap: a public gene expression resource. *Genome Research*, *10*(7), 1051-1060.

Li, B., Carey, M., & Workman, J. L. (2007). The Role of Chromatin during Transcription. *Cell*, *128*(4), 707-719. doi: 10.1016/j.cell.2007.01.015.

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science (New York, N.Y.)*, *227*(4693), 1435-1441.

Ljungman, M. (2005). Activation of DNA damage signaling. *Mutation Research*, *577*(1-2), 203-216. doi: 10.1016/j.mrfmmm.2005.02.014.

Loeb, L. A., Bielas, J. H., & Beckman, R. A. (2008). Cancers exhibit a mutator phenotype: clinical implications. *Cancer Research*, *68*(10), 3551-3557; discussion 3557. doi: 10.1158/0008-5472.CAN-07-5835.

Maitland, N. J., & Collins, A. (2005). A tumour stem cell hypothesis for the origins of prostate cancer. *BJU International*, *96*(9), 1219-1223. doi: 10.1111/j.1464-410X.2005.05744.x.

MATTHAEI, J. H., JONES, O. W., MARTIN, R. G., & NIRENBERG, M. W. (1962). Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America*, *48*, 666-677.

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564.

McKusick, V. A., & Ruddle, F. H. (1987). Toward a complete map of the human genome. *Genomics*, *1*(2), 103-106.

Mullis, K. (1994). *The Polymerase chain reaction*. Boston: Birkhäuser.

National Research Council (U.S.). (1988). *Mapping and sequencing the human genome Committee on Mapping and Sequencing the Human Genome, Board on Basic Biology, Commission on Life Sciences, National Research Council.* Washington D.C.: National Academy Press.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*(3), 443-453.

Nguyen H, Friedrich A, Berthommier G, Poidevin L, Moulinier, Ripp R, et al. (2008). Introduction du nouveau Centre de Données Biomedicales Décrypthon. CORIA, Tregastel.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics, 20*(2), 207-211. doi: 10.1038/2524.

Pruitt, K. D., & Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research, 29*(1), 137-140.

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research, 37*(Database issue), D32–D36. doi: 10.1093/nar/gkn721.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 35*(Database issue), D61–D65. doi: 10.1093/nar/gkl842.

R.A. Fisher. (1930). *The genetical Theory of Natural Selection*. Clarendon.

Raffelsberger, W., Krause, Y., Moulinier, L., Kieffer, D., Morand, A., Brino, L., et al. (2008). RReportGenerator: automatic reports from routine statistical analysis using R. *Bioinformatics (Oxford, England), 24*(2), 276-278. doi: 10.1093/bioinformatics/btm556.

Reis, E. M., Ojopi, E. P. B., Alberto, F. L., Rahal, P., Tsukumo, F., Mancini, U. M., et al. (2005). Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. *Cancer Res, 65*(5), 1693–1699.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., et al. (2000).

Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.), 290*(5500), 2306-2309. doi: 10.1126/science.290.5500.2306.

Roberts, L. (1991). GRAIL seeks out genes buried in DNA sequence. *Science (New York, N.Y.), 254*(5033), 805.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature, 265*(5596), 687-695.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America, 74*(12), 5463-5467.

Sanger, F., & Thompson, E. O. P. (1953a). The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal, 53*(3), 353–366.

Sanger, F., & Thompson, E. O. P. (1953b). The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal, 53*(3), 366–374.

Snoep, J. L., Bruggeman, F., Olivier, B. G., & Westerhoff, H. V.Towards building the silicon cell: A modular approach. *Biosystems, 83*(2-3), 207-216. doi: 10.1016/j.biosystems.2005.07.006.

Steele, E. J. (2009). Mechanism of somatic hypermutation: Critical analysis of strand biased mutation signatures at A:T and G:C base pairs. *Molecular Immunology, 46*(3), 305-320. doi: 10.1016/j.molimm.2008.10.021.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., et al. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science, 321*(5891), 956-960. doi: 10.1126/science.1160342.

Sun, M., Zhou, G., Lee, S., Chen, J., Shi, R. Z., & Wang, S. M. (2004). SAGE is far more

sensitive than EST for detecting low-abundance transcripts. *BMC Genomics*, *5*(1), 1. doi: 10.1186/1471-2164-5-1.

Svejstrup, J. Q. (2007). Elongator complex: how many roles does it play? *Current Opinion in Cell Biology*, *19*(3), 331-336. doi: 10.1016/j.ceb.2007.04.005.

Svejstrup, J. Q. (2002). Mechanisms of transcription-coupled DNA repair. *Nat Rev Mol Cell Biol*, *3*(1), 21-29. doi: 10.1038/nrm703.

Tateno, Y., & Gojobori, T. (1997). DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Research*, *25*(1), 14-17.

Team, R. D. C. (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrouvé de http://www.R-project.org.

The International HapMap Project. (2003). *Nature*, *426*(6968), 789-796. doi: 10.1038/nature02168.

Thompson, J. D., Muller, A., Waterhouse, A., Procter, J., Barton, G. J., Plewniak, F., et al. (2006). MACSIMS : multiple alignment of complete sequences information management system. *BMC Bioinformatics*, *7*, 318. doi: 10.1186/1471-2105-7-318.

Thompson, J. D., Holbrook, S. R., Katoh, K., Koehl, P., Moras, D., Westhof, E., et al. (2005). MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. *Nucleic Acids Research*, *33*(13), 4164–4171. doi: 10.1093/nar/gki735.

Tjio JH, L. A. (1956). The chromosome number in man. *Hereditas*, *42*, 1-6.

Tsuruta, D., Kobayashi, H., Imanishi, H., Sugawara, K., Ishii, M., & Jones, J. C. R. (2008). Laminin-332-integrin interaction: a target for cancer therapy? *Current Medicinal Chemistry*, *15*(20), 1968-1975.

Uhring, M. (2004). Contribution à l'étude structure/fonction du facteur de transcription TFIIH.

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*(5235), 484–487.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–1351. doi: 10.1126/science.1058040.

Wade, J. T., & Struhl, K. (2008). The transition from transcriptional initiation to elongation. *Current Opinion in Genetics & Development*, *18*(2), 130-136. doi: 10.1016/j.gde.2007.12.008.

Walmacq, C., Kireeva, M. L., Irvin, J., Nedialkov, Y., Lubkowska, L., Malagon, F., et al. (2009). Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *The Journal of Biological Chemistry*, *284*(29), 19601-19612. doi: 10.1074/jbc.M109.006908.

Walsh, S., & Barrell, B. (1996). The Saccharomyces cerevisiae genome on the World Wide Web. *Trends in Genetics: TIG*, *12*(7), 276–277.

Watson, J. D., & Crick, F. H. (1974). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953. *Nature*, *248*(5451), 765.

Weil, J. (2001). Réplication chez les eucaryotes. Dans *Biochimie générale* (9 éd., p. 396). Dunod.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, *35*(Database issue), D5–12. doi: 10.1093/nar/gkl1031.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*, *318*(5853), 1108–1113. doi: 10.1126/science.1145720.

Zenkin, N., Yuzenkova, Y., & Severinov, K. (2006). Transcript-assisted transcriptional proofreading. *Science (New York, N.Y.)*, *313*(5786), 518-520. doi: 10.1126/science.1127422.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *7*(1-2), 203-214. doi: 10.1089/10665270050081478.

Publications

RreportGenerator : automatic reports from routine statistical analysis using R

Wolfgang RAFFELSBERGER, Yannick KRAUSE, Luc MOULINIER, David KIEFFER, Anne-Laure MORAND, Laurent BRINO and Olivier POCH

Bioinformatics, 2008*, *Volume 24, Numéro 2, pages 276-278

A new projection method for biological semantic map generation

Hoan N Nguyen*, Nicolas Wicker*, David Kieffer and Olivier Poch

Laboratoire de bioinformatique et génomique intégratives,
Institut de Génétique et de Biologie Moléculaire et Cellulaire,
67404 Illkirch Cedex, France
University of Strasbourg
*The two first authors contributed equally to the work.

ABSTRACT

Low-dimensional representation is a convenient method of obtaining a synthetic view of complex datasets and has been used in various domains for a long time. When the representation is related to words in a document, this kind of representation is also called a semantic map. The two most popular methods are self-organizing maps and generative topographic mapping. The second approach is statistically well-founded but far less computationally efficient than the first. On the other hand, a drawback of self-organizing maps is that they do not project all points, but only map nodes.

This paper presents a method of obtaining the projections for all data points complementary to the self-organizing map nodes. The idea is to project points so that their initial distances to some cluster centers are as conserved as possible. The method is tested on an oil flow dataset and then applied to a large protein sequence dataset described by keywords. It has been integrated into an interactive data browser for biological databases.

1. Introduction

Thanks to the availability of the human and other genomes and the rapid progress of biotechnologies and information technologies, numerous large biomedical datasets have been generated. Modern biomedical information thus corresponds to a high volume of heterogeneous data that doubles in size every year and that covers very different data types, including phenotypic data, genotypic data as well as standards, processes, protocols or treatments used to generate information from raw data. In this context, systemic approaches are now needed to store, analyze and compare the huge amount of relevant information.

In addition, the knowledge provided by classical query services on biological data is often unsatisfactory (e.g. a list of proteins or sequences) and there is a need for user-friendly visual representations of the data. Such a representation exists and is called a feature or semantic map. It is used to visualize "land maps" in two or three dimensions that represent, for example, the distribution (similarity and neighborhood) of protein annotations in biological databases. When query results are represented on the map, the repartition of the proteins can be easily observed, as well as their proximity to clusters labeled according to their content. In addition, it is straightforward to superpose the information obtained from additional requests. Thus, a semantic map can greatly facilitate the interpretation of results from large scale data analyses. To quote a few examples, semantic maps have already been used in fluid mechanics [1], astronomy [2], internet data mining [3-4], scientific literature mining [5] and biology [6].

Many low-dimensional methods have been devised [5, 7, 8, 9] and two of the most popular are the WEBSOM method [9] and the Generative Topographic Mapping (GTM) [1]. These two methods are briefly outlined below.

WEBSOM originates from self-organizing maps [10] which is a classification algorithm where nodes move towards cluster centers. In WEBSOM, the nodes are fixed on a two-dimensional grid and at the same time live in the space of the dataset, typically a $\mathbb{R}^p$ space. First, a point y is picked at random from the dataset. Next, the closest node w_i in $\mathbb{R}^p$ is selected and then each node w_j moves towards y according to the equation $w_j(t+1) = w_j(t) + \eta(t)h_{ij}(t)\|y - w_j(t)\|$ where $\eta(t)$ is the learning rate decreasing in time and $h_{ij}(t)$ is a neighborhood function in the two-dimensional grid. These steps are then iterated for all data points. The initialization of the p-dimensional space can be

performed randomly, but a more effective method is to select points along the two first principal axes of the dataset [4]. Finally, the dataset is used again by assigning each point to its closest node in the p-dimensional space using a Euclidean distance. Then, for each node, the number of points it has captured is taken as its density up to a given scaling factor (the size of the dataset).

The generative topographic map (GTM) [1] is a statistical method which is provably (locally) convergent and which does not require a shrinking neighborhood or a decreasing step size. It is a generative model: the data is assumed to arise by probabilistically picking points in a low-dimensional space and mapping them to the observed high-dimensional input space. The statistical model can be described in the following way:

$p(y|x_i, W, \beta) = \left(\frac{\beta}{2\pi}\right)^{p/2} exp\left\{-\frac{\beta}{2}\|W.\phi(x_i) - y\|^2\right\}$ where x_i is a two-dimensional grid node, β is a scaling parameter, $W.\phi(x_i)$ a generalized regression model, W a $p \times m$ matrix and the elements of $\phi(x)$ consist of m basic functions $\phi_j(x)$ typically equal to radially symmetric Gaussians centered on the nodes of a two-dimensional grid. The parameters W and β of the model are estimated through the expectation-maximization (EM) algorithm [11]. This model can be considered to be the probabilistic counterpart of SOM/WEBSOM. However, the WEBSOM method is quicker than GTM when large amounts of data must be dealt with, especially if the winner selection is optimized so that millions of documents and nodes can be treated [4].

An alternative choice is to follow Flexer's approach [12] which first clusters the points in the data space and then projects cluster centers using Sammon's multidimensional scaling method [13]. However this means that only a subset of points are effectively projected. In this paper, we present a complementary method that projects all points using their distances to the cluster centers.

First this new projection method is presented, then it is evaluated on a benchmark data set and compared to other methods. Finally, it is used in the results section to generate a semantic map in the context of a new integrative navigator for biological databases.

2. Method

The principle of the presented method is to project points after they have been clustered and the cluster centers have been projected onto a two-dimensional map. This is done by conserving as much as possible the original distances between the points and the cluster centers. Basically, for each point indexed by i, the two-dimensional coordinates are search such as to minimize the difference between the distances computed in the n-dimensional data space with those computed on the map.

This comes down to finding the point x_i in two dimensions minimizing the following function $E(x_i)$:

$$E(x_i) = \sum_{g=1}^{G}\left(\sum_{k=1}^{2}(x_k - c_{g,k})^2 - d_g^2\right)^2$$

with d_g denoting the distance between point i and cluster g and $c_{g,k}$ the projection of the k^{th} cluster center. The Newton-Raphson algorithm was used to minimize $E(x_i)$. At each step, $x_i^{t+1} = x_i^t - H^{-1}.\nabla E$ with H the Hessian and ∇E the gradient of E.

$$\nabla E = \begin{pmatrix}\frac{\partial E}{\partial x_1}\\ \frac{\partial E}{\partial x_2}\end{pmatrix}, H = \begin{pmatrix}\frac{\partial^2 E}{\partial x_1^2} & \frac{\partial^2 E}{\partial x_1 \partial x_2}\\ \frac{\partial^2 E}{\partial x_1 \partial x_2} & \frac{\partial^2 E}{\partial x_2^2}\end{pmatrix}$$

The optimizing function is not convex as the Hessian is not always semi-definite positive. To show this, it is sufficient to find a point X verifying $X'HX < 0$. In particular, we show that H_{11} can be negative which is also sufficient. First let us note that

$$\frac{\partial E}{\partial x_l} = \sum_{g=1}^{G} 2\left(\sum_{k=1}^{2}(x_k - c_{g,k})^2 - d_g^2\right)2(x_l - c_{g,l})$$

$$\frac{\partial^2 E}{\partial x_l^2} = 8\sum_{g=1}^{G}(x_l - c_{g,l})^2 + 4\sum_{g=1}^{G}\left(\sum_{k=1}^{2}(x_k - c_{g,k})^2 - d_g^2\right)$$

and then set

$$\begin{cases} x_1 = c_{1,1} = c_{2,1} = c_{3,1} \\ d_1 = \sqrt{(x_1 - c_{1,1})^2 + (x_2 - c_{1,2})^2} \\ d_2 = \sqrt{(x_1 - c_{2,1})^2 + (x_2 - c_{2,2})^2} \\ d_3 > \sqrt{(x_1 - c_{3,1})^2 + (x_2 - c_{3,2})^2} \end{cases}$$

Thus, $H_{11} = 4\left((x_1 - c_{3,1})^2 + (x_2 - c_{3,2})^2 - d_3^2\right) < 0$ ■

Consequently, a global optimization process was performed using different initial values. Each cluster center projection was used as an initial value and the best solution after convergence was kept.

3. Results and Discussion

3.1. Validation using the oil flow dataset

To validate the new points projection method, a previously established oil flow dataset [14] was used as a benchmark. This training dataset is available at http://www.ncrg.aston.ac.uk/GTM/ and contains 1000 points in 12 dimensions corresponding to 12 measurements on the mixture of oil, water and gas passing through a pipeline. The three phases in the pipe can belong to three different configurations corresponding to laminar, homogeneous and annular flows.

First, the dataset was clustered into 15 clusters and the cluster centers projected according to Sammon's multidimensional scaling method [13]. Then the 1000 points were projected in two dimensions using the method described above. The results are shown on figure 1, where it be seen that three different groups are rather well linearly separated. The groups obtained with the GTM and principal component analysis (PCA) methods are shown on figures 2 and 3 respectively. In order to objectively measure the quality of these results, we computed the ratio of the between-class inertia and the total inertia for each method. For our method, GTM and the PCA, we obtained a ratio of 0.83, 0.25 and 0.23 respectively, thus confirming the visual impression. Nevertheless, it should be stated that, if only separation is desired and not specifically linear separation, GTM performs better, even though it has the drawback of making the underlying grid very visible.

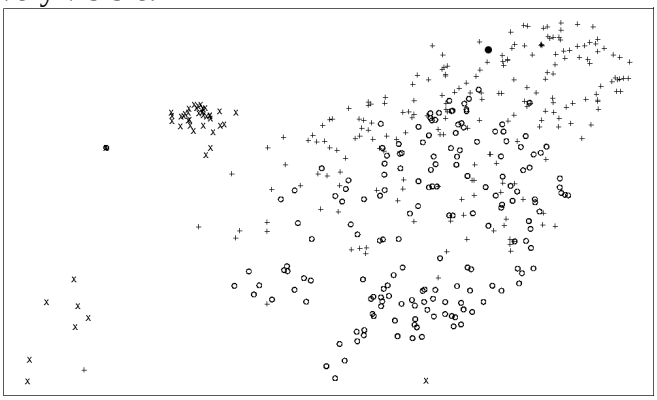


Figure 1- New projection of the dataset. Results of the presented projection on the oil flow dataset. Crosses, circles and plus-signs represent stratified, annular and homogeneous multi-phase configurations respectively. The three group separations are clearly identified.

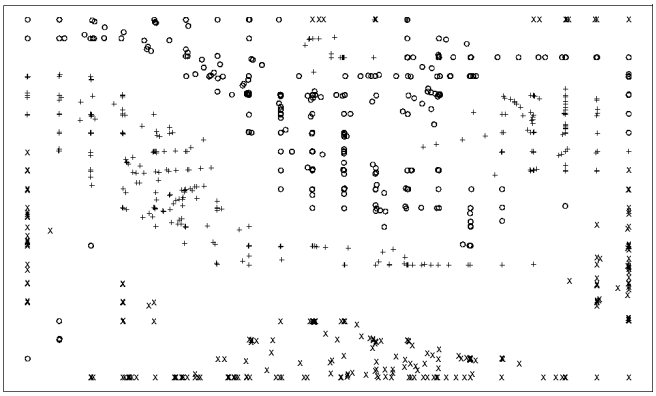


Figure 2- Oil flow dataset after GTM. After projection of the oil flow dataset using the Generative Topographic Mapping, the three group separations are clearly separated, but in a complex way that is far from linear.

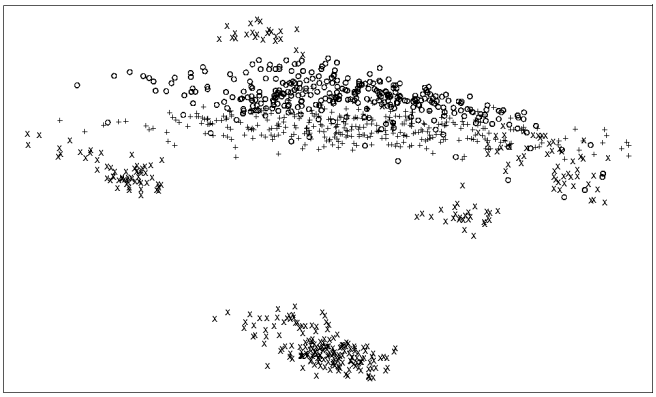


Figure 3- Oil flow dataset after PCA projection. After projection of the oil flow dataset using principal component analysis, the separation of the three groups is not clearly identified. In particular, the crosses are very scattered.

3.2. Semantic map generation for biological database

The Laboratory of Genomics and Integrative Bioinformatics (LGBI) at the IGBMC Strasbourg, has developed a new high-performance biomedical information system, called the BIRD System [15-16]. BIRD is able to integrate very quickly heterogeneous data either from the large generalist databases (sequence, structure, function and evolution, etc.) or from specialized databases dedicated to high throughput biology (transcriptomics, interactomic, etc.) in a relational database (IBM DB2). Thus, it allows to organize massive sets of biomedical data according to real world requirements. An original biological query engine, called BIRDQL, has been designed to facilitate access to the heterogeneous databases and to allow pertinent information extraction via a web server. This system has been used in the Decrypthon computing grid [17] in order to provide data to the runtime applications.

To complete the visualization and analyze functionalities of the BIRD System, the new method described above to build semantic maps was integrated in the BIRD query engine (BIRDQL). The maps can be used to explore the data using a combination of high level queries and area selections (figure 4). The method was tested by building a semantic map of the Uniprot database [18] using the keyword descriptions for each protein. After removal of redundant vectors, we obtained 60,000 vectors $z_1, \dots, z_{60000}$ in a 914-dimensional space corresponding to the 914 keywords extracted from about 6 million proteins. In the following lines, to avoid focusing on the numerical details, we will consider n proteins described by p keywords where n and p stand for 60000 and 914 respectively.

Before projecting the points, some preliminary steps were necessary:

Step 1: dimension reduction

The n proteins were described by p keywords and were thus represented by n points $z_1, \dots, z_n$ in p dimensions. As in the preprocessing step of WEBSOM [3-4], an initial dimension reduction was performed to reduce p coordinates to p^* using random projection directions. More specifically, random vectors v_j were generated on the p^*-dimensional unit-sphere and then new coordinates were obtained by computing the scalar product $y_{ij} =< v_j, z_i >$ on each document i. Thus, the n proteins were described by n points $y_1, \dots, y_n$.

Step 2: mixture models clustering

In a second step, these points were clustered using mixture models. Mixture models are a powerful method to cluster datasets of points described by coordinates. The points are assumed to be independent realizations from a mixture of several distributions. Here the mixture is only briefly described for G components $f_{\alpha_1}, \dots, f_{\alpha_G}$ with parameters $\alpha_1, \dots, \alpha_G$. A general presentation of this method and its applications can be found in [19-22]. If $\tau_1, \dots, \tau_G$ indicate the different weights of the components, the likelihood of the model for n points $y_1, \dots, y_n$ is expressed as:

$$L_M(\tau_1, \dots, \tau_G, \alpha_1, \dots, \alpha_G | y_1, \dots, y_n) = \prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g f_{\alpha_g}(y_i)$$

The estimation of the different coefficients of the mixture model is commonly performed via the EM (Expectation-Maximization) algorithm of Dempster [11]. Here, in order to simplify the estimation, a variant of the EM algorithm called CEM was used [22]. In this application G was chosen to be equal to 30.

Step 3: cluster centers projection

Once G clusters were obtained, the centers of gravity $c_1^*, \dots, c_G^*$ were computed in the p-dimensional space. Then, multidimensional scaling (MDS) [23] was applied on the cluster centers to produce two-dimensional coordinates $c_1, \dots, c_G$. MDS was used because Sammon's method [13] failed on this dataset, since it produced many points with the same coordinates.

After these three preliminary steps, the points were projected on the map using the new projection method. The density $m(x)$ for each point x of the map is given using a kernel method [24]:

$$m(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\pi |\Sigma|^{1/2}} exp\{(x - x_i)' \Sigma^{-1} (x - x_i)\}$$

Then, a color scale ranging from purple to white, with intermediary colors red, orange and yellow, was assigned to each point according to its density. The map is represented in figure 4.

This visual representation allows a global comprehension of the whole database, which is easier to understand than numerical or textual data. Some important keywords shared by many proteins are visible on this map, such as kinase, ligase and protease. At the same time, frequent keywords, such as ”complete proteome”, that are non-informative, are avoided because they are shared by several clusters. Another observation is that the density is far from being homogeneous, the map being more crowded in the bottom-left corner than elsewhere.

When using the integrated biological query engine BIRD-QL of the BIRD System via a web service or http protocol, as shown in figure 5, the selected proteins are represented on the maps by a plus sign of a given color. If different selections have been performed, different colors are used. An example is shown in figure 6, where proteins selected by a query with the keyword “apoptosis” are shown by blue plus signs. Some of these proteins were selected by the user and are surrounded by a white square. One of the proteins, DNJA3, belongs to the small cluster labeled “disease mutation” but does not possess the “disease mutation” keyword. Interestingly its deficiency implies dilated cardiomyopathy [25] (MIM608382).

There is still room for improvement in the construction of semantic maps both at the algorithmic level and at the software functionality level. The points projection is formalized as a global optimization problem and currently, it is resolved simply using different starting points with the Newton-Raphson method. However global optimization methods could also be tested [26-27]. From a practical point of view it would also be useful to determine how many clusters or nodes are necessary to achieve a good projection of the data points.

4. Conclusion

The main contribution of this work is a new computational solution to the construction of semantic maps. The idea is to project points by locating them according to cluster centers. This method can thus be coupled with other methods such as self-organizing maps or Flexer's approach.

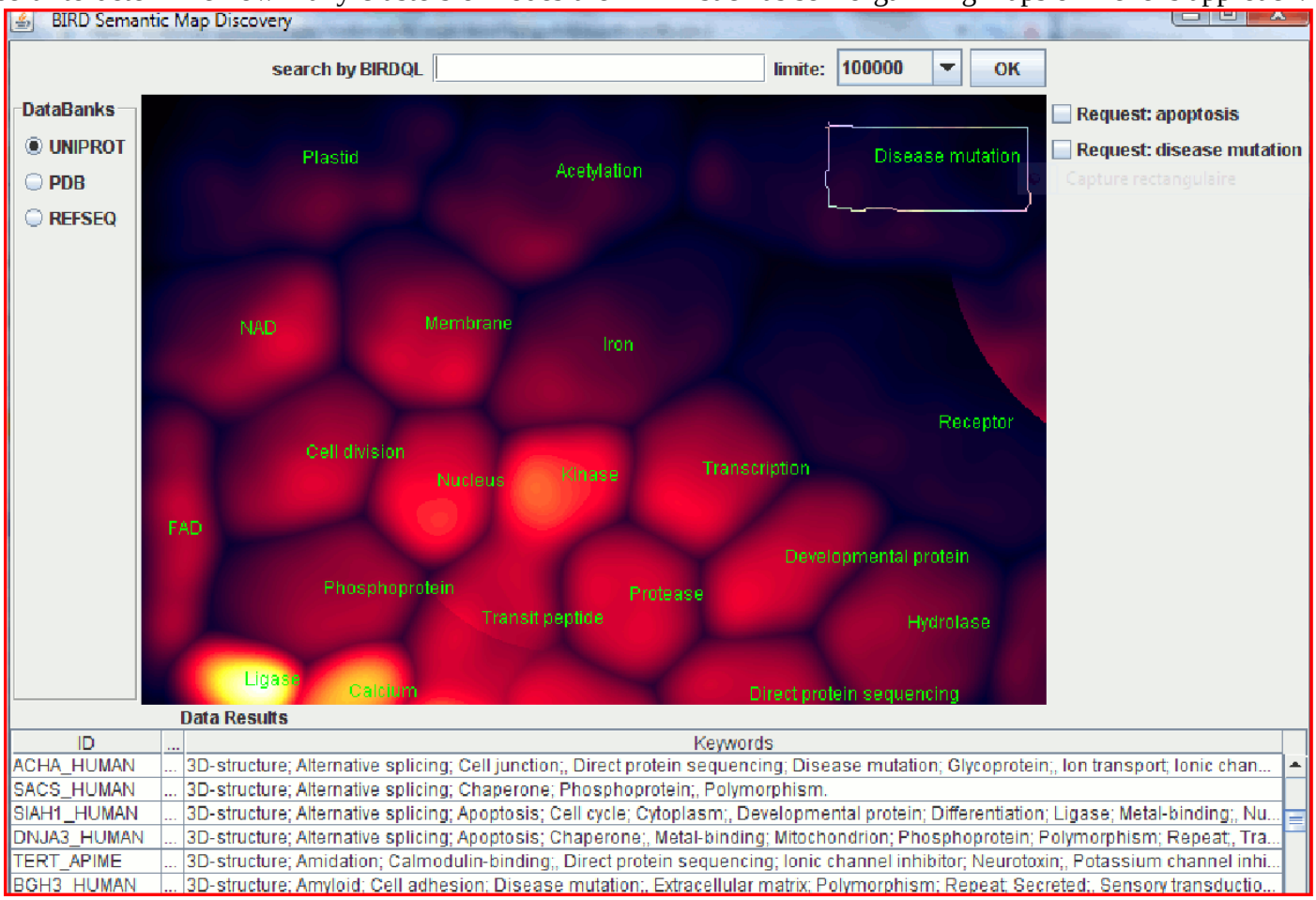


Figure 4- Semantic map with density colours and most frequent keyword labels.

Acknowledgements

This work was supported by the CNRS, the University of Strasbourg and the Décrypthon program initiated by the Association Française contre les Myopathies, IBM and the CNRS. We are grateful to all internship students who participated in this work by programming some parts of it, namely Xavier Brotel, Jérémy Némo Trouslard and Julien Cadet. The authors would like to thank Anne Friedrich, Laurent Philippe Albou and Julie Thompson for helpful suggestions.

References

[1]C.M. Bishop, M. Svens'en, and C.K.I. Williams, (1998) GTM: the generative topographic mapping, Neural Computation, **10**, 215-234.

[2]K. Lesteven, (1995) Multivariate data analysis applied to bibliographical information retrieval: SIMBAD quality control. Vistas in Astronomy, **39**, 187-193

[3]S. Kaski, (1998) Dimensionality reduction by random mapping: Fast similarity computation for clustering, Proceedings of IJCNN'98, International Joint Conference on Neural Networks, IEEE Service Center, 413-418.

[4]K. Lagus, S. Kaski, and T. Kohonen, (2004) Mining massive document collections by the WEBSOM method. Information Sciences, **163**, 135-156.

[5]C. Chen, (2005) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, Journal of the American Society for Information Science **57**, 359—377.

[6]M. Grimmelstein, and W. Urfer, W (2005) Analyzing Protein Data with the Generative Topographic Mapping Approach. Innovations in Classification, Data Science, and Information Systems, Baier, D and Wernecke, K-D, Springer Berlin Heidelberg, 585-592.
[7]P.G. Ossorio, (1966) Classification space: a multivariate procedure for automated document indexing and retrieval, Multivariate Behavioral Research **1**, 479–524.
[8]S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, (1990) Indexing by latent semantic indexing, Journal of the American Society for Information Science **41**, 391-407.
[9]T. Kohonen, (1997) Self-Organizating Maps, Springer-Verlag
[10]T. Kohonen, (1982) Analysis of a simple self-organizing process, Biological Cybernetics **44**, 135-140.
[11]A. Dempster, N. Laird, and D. Rubin, (1977) Maximum likelihood from incomplete data via the {EM} algorithm, Journal of the Royal Statistical Society, Ser. B, **39**, 249-282.
[12]A. Flexer, (1997) Limitations of self-organizing maps for vector quantization and multi-dimensional scaling, Advances in neural information processing systems **9**, 445-451.
[13]J.W. Sammon, (1969) A non-linear mapping for data structure analysis, IEEE Transactions on computers **18**, 401-409.
[14]C.M. Bishop, and G.D. James, (1993) Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. Nuclear Instruments and Methods in Physics Research Section A **327**, 580-593.
[15]H. Nguyen, G. Berthommier, A. Friedrich, L. Poidevin, R. Ripp, L. Moulinier, and O. Poch, (2008) Introduction to the new Decrypthon Data Center for biomedical data, Proc CORIA'2008 32-44.
[16]BIRDQL - Wikili, http://alnitak.u-strasbg.fr/wikili/index.php/BIRDQL
[17]Décrypthon : le grid-computing au service de la génomique et la protéomique, http://www.decrypthon.fr
[18]The UniProt Consortium (2008) The Universal Protein Resource (UniProt), Nucleic Acids Research **36**, D190--D195.
[19]D. Titterington, A. Smith, and U. Makov, (1985) Statistical Analysis of Finite Mixture Distribution, John Wiley and Sons
[20]G. McLachlan, and K. Basford, (1988) Mixture Models: Inference and Applications to Clustering, Marcel Dekker
[21]J. Banfield, and A. Raftery, (1993) Model-based Gaussian and non-Gaussian clustering, Biometrics **49**, 803-821.
[22]G. Celeux, and G. Govaert, (1992) A classification EM algorithm for clustering and two stochastic versions, Journal of Computational Statistics and Data Analysis **14**, 315-332.
[23]K.V. Mardia, J.T. Kent, and J.M. Bibby, (1979) Multivariate Analysis, Academic Press.
[24]E. Parzen, (1962) On estimation of a probability density function and mode. Annals of Mathematical Statistics **33**, 1065-1076.
[25]M. Hayashi, K. Imanaka-Yoshida, T. Yoshida, M. Wood, C. Fearns, R.J. Tatake, and J.D. Lee, (2006) A crucial role of mitochondrial Hsp40 in preventing dilated cardiomyopathy, Nature Medecine **12**, 128-132.
[26]M. Laguna, and R. Marti, (2005) Experimental testing of advanced scatter search designs for global optimization of multimodal functions, Journal of Global Optimization **33**, 235-255.
[27]A. Neumaier, O. Shcherbina, W. Huyer, and T. Vinko, (2005) A comparison of complete global optimization solvers, Mathematical Programming **103**, 335-356.

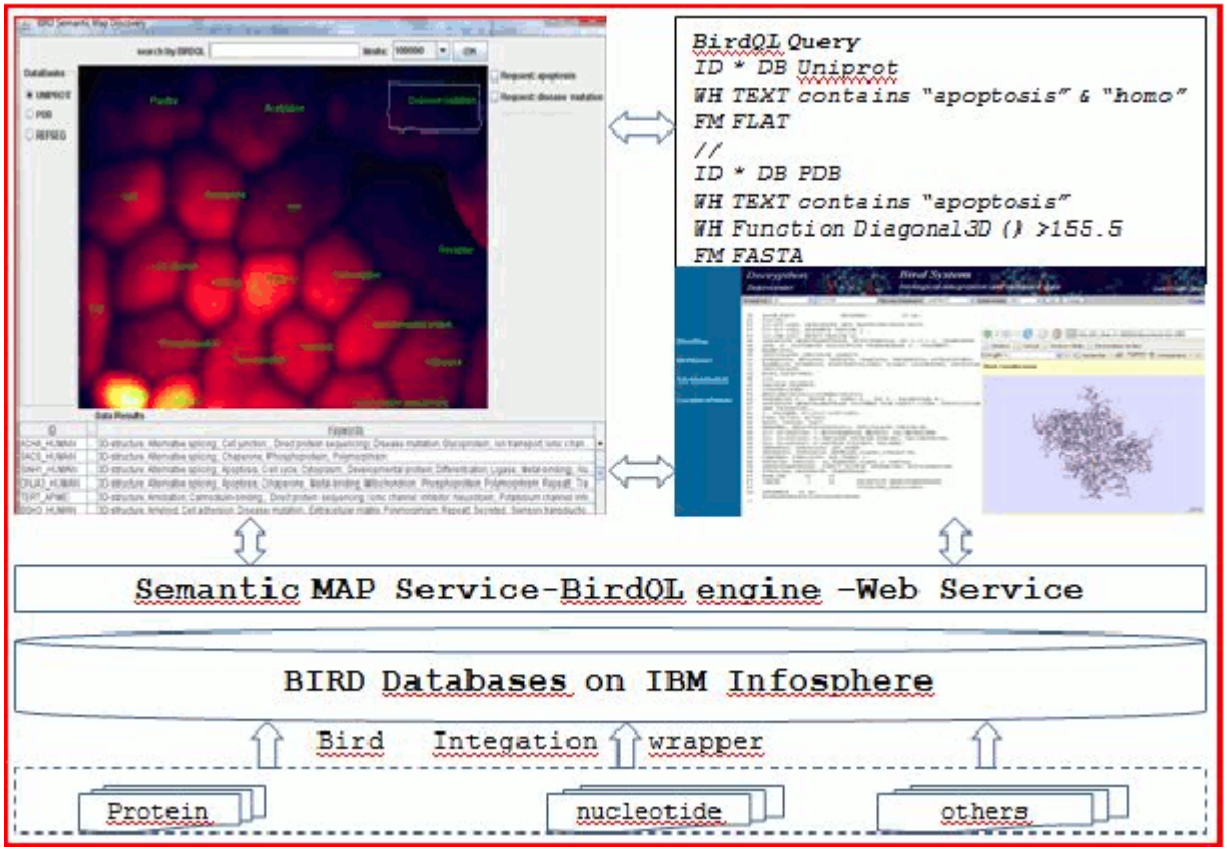


Figure 5 – The global architecture of the Semantic Map Discovery prototype coupled with the BIRD System using the BirdQL query engine.

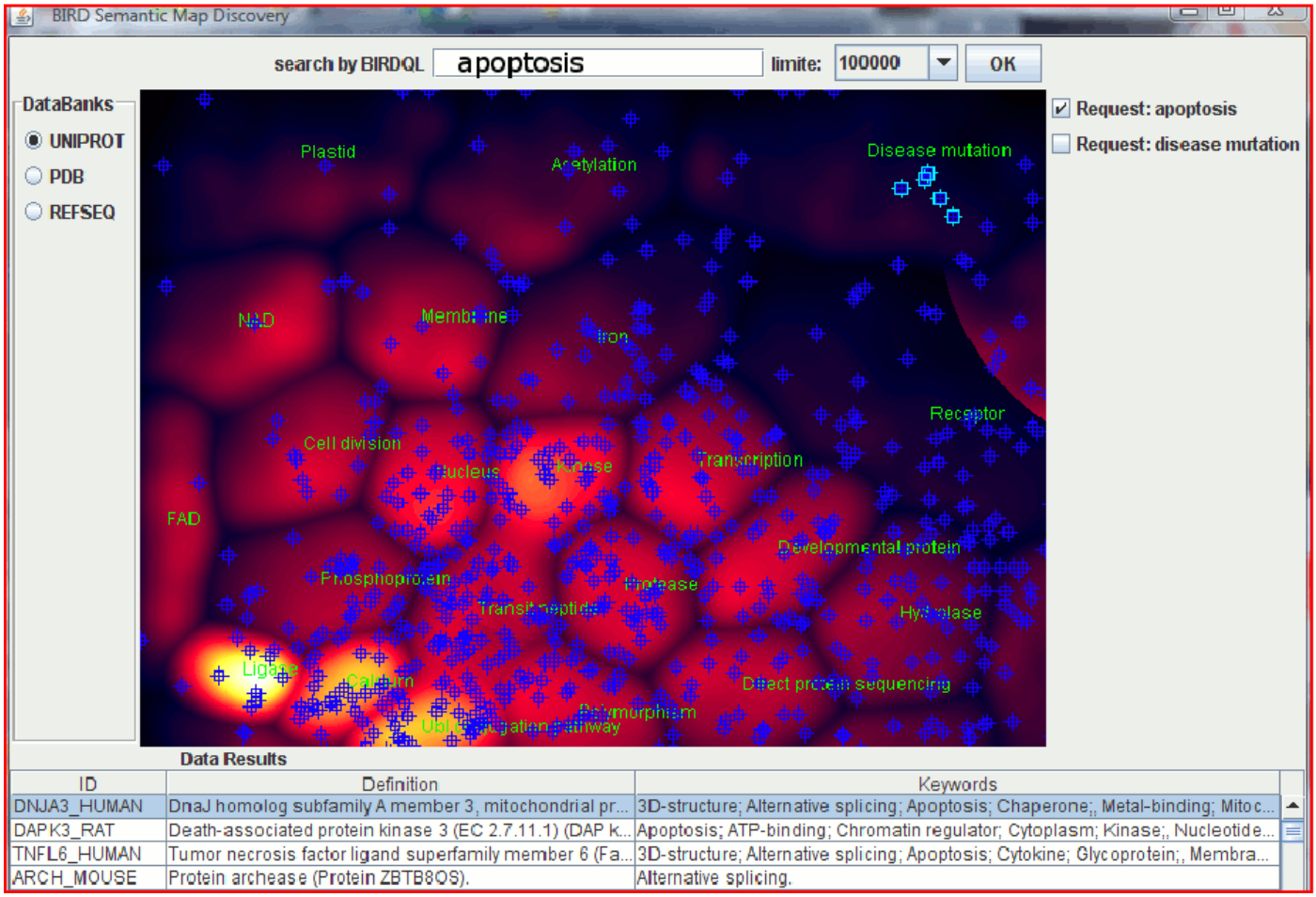


Figure 6 – Semantic map with selected proteins. The labels represent the most frequent keywords present inside the cluster points which are not shared between different clusters.