École Supérieure de Biotechnologie de Strasbourg

École Doctorale des Sciences de la vie et de la santé

Développement d'une méthode d'interprétation rapide des spectres RMN pour les protéines

THÈSE

présentée et soutenue publiquement le 11 février 2005

pour l'obtention du

Doctorat de l'Université Louis Pasteur – Strasbourg I

par

Marie-Aude Coutouly

Composition du jury

Rapporteurs :	Dr. Martin Blackledge	Rapporteur externe
	Dr. Marc-André Delsuc	Rapporteur externe
	Dr. Martial Piotto	Rapporteur interne
Examinateurs :	Dr. Michael Nilges	Examinateur
	Dr. Andrew Atkinson	Examinateur
	Pr. Bruno Kieffer	Directeur de thèse

Laboratoire de Biologie et Génomique Structurales — UMR 7104

Mis en page avec la classe thloria.

Remerciements

Je tiens à remercier le Pr B. Kieffer pour m'avoir accepté dans son équipe et pour m'avoir définitivement convaincu de planifier mon travail et par la même de m'avoir appris la nécessité d'un certain flegmatisme. Je tiens à remercier la Région Alsace et Aventis pour leur financement.

Je remercie les chercheurs du laboratoire de chimie quantique de l'ULP sans qui je n'aurais pas fait de thèse. Je remercie Marie-Pierre Reck grâce à qui j'ai atterri à l'ESBS. Et je remercie Chantal Lorentz pour m'avoir poussée et accompagnée dans ce choix.

Je remercie tout particulièrement le Dr RA Atkinson pour m'avoir guidé tout au long de ce travail de thèse. Il m'a fait partagé ses connaissances et ses idées par rapport à QUASI. Il m'a permis d'approfondir ce projet grâce à de longues discussions et des échanges de point de vue fructueux.

Je remercie Dominique Desplancq, pour avoir eu la lourde tâche de m'encadrer lors de la purification du fragment 24kDa de la sous-unité B de l'ADN-gyrase. Vu mes connaissances plus que sommaires dans ce domaine, son rôle a été bien plus complexe qu'il n'y parait.

Je remercie les docteurs Finton Sirockin et Virginie Lafont, pour leur présence lors des premières années de ma thèse. Ils m'ont aidé plus qu'ils ne le pensent, Finton par son côté informaticien et Virginie par ses connaissances en biologie moléculaire. Outre ces points scientifiques, ils m'ont apportés des moments de bonne humeur inestimables à mes yeux : les pauses cafés et les repas partagés.

Autres personnes ayant apporté un peu de gaieté au laboratoire et que je tiens à remercier : Michael Oberlin, Marc Vitorino, Nicolas Ambert, Cédric Grauffel, Thomas Hoellinger et Marc-Olivier Sercki.

Je remercie également Martina Schechner pour les footings partagés entre midi et deux et Dr Emeric Wasielewski pour m'avoir accepté dans son espace vital pendant 3 ans.

Un grand merci à mes amis du SUC Judo Anne-Thérèse Schneider, Sophie et Bruno Vonpierre, Eric Befort et tous les autres avec qui j'ai passé de nombreuses heures qui m'ont permis de me défouler et de me décontracter.

Un petit mot aussi (par ordre alphabétique) pour Anne Bodlenner, Anne Wittmeyer, Cédric Carlier, Dr Gaétan Weck, Dr Guillaume Hirsch, Gwendoline Quilici, Dr Maggy Hologne, Martine Maruani,Olivier Korkman, Serge Sartor et Vincent Gembus pour les randonnées/ballades vosgiennes du dimanche, la découverte du Vercors, les soirées de jeux et autres moments conviviaux...

Egalement Frédérique, Julie, Catherine, Séverine, Raphaelle et tous ceux que je ne

nomme pas mais à qui je pense.

Enfin je remercie les gens qui ont été là pour me soutenir à tout moment, ma famille, tout particulièrement mes parents, ainsi que Valérie.

Table des matières

Glossaire

ADPNP : adenosine 5′-adenylyl β,γ-imidodiphosphate
ALFA : ALgorithm for Fast Assignment
ANSIG : Assignment NMR Spectra by Interactive Graphics
ASCII : American Standard Code for Information Interchange
AWK : Aho Weinberger Kernighan

BMRB : BioMag Res Bank
BPTI : Bovine Pancreatic Trypsin Inhibitor

CAMRA : Computer Aided Magnetic Resonance Assignment
CLOUDS : Computed Location Of UnassigneD Spins
CSI : Chemical Shift Index
CSP : Chemical Shift Pattern

DFT : Density-Functionnal Theory

FID : Free Induction Decay
FILTER : Foc Identification via LowesT ErroR

GARANT : General Algorithm for Resonance AssignmeNT

NOE : Nuclear Overhauser Effect
NOESY : Nuclear Overhauser Effect SpectroscopY

PACES : Protein Sequential Assignment by Computer Assisted Exhaustive Search
PASTA : Protein ASsignment by Threshold Accepting
PDB : Protein Data Bank

QUASI : QUick Access to Spectral Interpretation

RDC : Residual Dipolar Couplings
REDCAT : REsidual Coupling Analysis Tool

RESCUE : RESidue prediCtion with neUral nEtworks
RMN : Résonance Magnétique Nucléaire
RMSD : Root Mean Square Deviation

SCOP : Structural Classification of Proteins

TATAPRO : Tracked Automated Assignments in PROteins
TOCSY : TOtal Correlation SpectroscopY
TROSY : Transverse Relaxation-Optimized SpectroscopY

Introduction générale

A l'heure où le génome est décrypté, il est nécessaire, afin de pouvoir en utiliser l'information, de connaître la structure tridimensionnelle des protéines encodées. Cette connaissance permet entre autres choses, la mise au point de médicaments qui vont pouvoir modifier spécifiquement l'activité de la protéine cible en s'ajustant, au plus près, à la géométrie de celle-ci. Les deux méthodes majeures utilisées pour l'élucidation des structures de protéines sont la cristallographie aux rayons X et la RMN à haute résolution en solution. La cristallographie est la méthode la plus ancienne et la plus aguerrie dans ce domaine. C'est elle qui fut, la première, capable de permettre la détermination d'une structure tridimensionnelle de protéine publiée en 1958 [Kendrew *et al.*, 1958]. Les premières structures RMN, elles, ont été publiées près de 30 ans plus tard [Wagner *et al.*, 1987]. Un des avantages de la cristallographie X est qu'il n'existe pas de taille limite à la protéine étudiée ; par contre, elle doit pouvoir être cristallisée. Cette contrainte exclut certaines protéines contenant des régions flexibles. De plus, la cristallographie piège la structure dans un minimum énergétique. La conformation peut alors dépendre des forces de "packing", c'est-à-dire des interactions protéine-protéine existant dans le cristal.

Une fois la protéine produite et cristallisée, la cristallographie permet, grâce à des logiciels bien rodés, une détermination automatique et rapide de la structure. D'autre part, la spectroscopie RMN effectue quant à elle l'analyse en solution : elle se trouve donc dans des conditions plus proches de celles rencontrées dans le milieu naturel et on peut y étudier les effets des phénomènes dynamiques. Les régions désordonnées rendent l'interprétation des spectres RMN plus complexe, mais les informations restent utilisables. Enfin la RMN permet le positionnement d'un domaine par rapport aux autres. Elle est aussi très prisée par la recherche pharmaceutique [Jahnke Widmer, 2004] car elle permet la détection et la caractérisation des interactions moléculaires. Toutefois, pour des protéines de taille importante (>40 kDa), l'encombrement spectral ainsi qu'une relaxation transverse très efficace représentent des obstacles importants pour une étude structurale par RMN.

La RMN souffre encore d'un laps de temps trop important entre l'acquisition des données et le dépôt de la structure dans la PDB (Protein Data Bank) [Berman *et al.*, 2000]. Cependant, grâce à l'évolution de la technologie des spectromètres RMN et aux méthodes de détermination de structure automatisées qui se développent, l'écart entre la RMN et la cristallographie se réduit.

L'étape d'interprétation des spectres RMN a donné lieu à de nombreuses tentatives d'automatisation. Toutefois, aucun programme ne s'est imposé. Les résultats obtenus divergent de façon significative en fonction des conditions de l'étude. Cette difficulté d'automatisation s'explique par la présence d'artefacts dans les spectres, par des listes de pseudo-résidus incomplètes à cause de recouvrements de signaux ou de mouvements internes.

De l'acquisition des spectres au dépôt de la structure dans la PDB, de nombreuses étapes doivent être franchi. La première étape consiste à différencier entre les signaux RMN et les artefacts. Cette étape est cruciale et conditionne le bon déroulement du projet de détermination de structure de protéine par RMN. Aprés cette localisation des signaux intéressants, il s'agit d'identifier chaque pic sélectionné à un atome précis de la séquence primaire : c'est l'attribution. La méthode d'attribution, dite classique, a été proposée par Wüthrich et collaborateurs dès 1982 [Billeter *et al.*, 1982, Wüthrich, 1986]. Cette méthode qui a marqué de son empreinte de nombreux programmes d'attribution comporte les 3 étapes suivantes :

- identification du type d'acide aminé (résidu) sur la base des corrélations observées entre les noyaux appartenant à un même résidu (appelé **système de spin**). Pour certains types d'acide aminé tels que les glycines, qui n'ont pas de pic C^{β}, les alanines dont la présence du groupe méthyle est facilement reconnaissable etc..., l'identification se fait de façon de non ambiguë.

- identification des relations séquentielles possibles entre les systèmes de spins en utilisant les NOE ^{1}H-^{1}H. Cette étape lève les ambiguités. Le placement unique des système de spin connectés sur les acides aminés de la séquence, établit ainsi *l'attribution spécifique à la séquence.*

- extension des attributions aux noyaux des chaînes latérales et détermination de l'attribution stéréospécifique.

Cette méthode permet l'attribution des lignes de résonance ^{1}H de tous ou presque tous les atomes d'hydrogène présents dans de petites protéines (<15 kDa). Le proton étant le noyau qui a l'abondance naturelle la plus importante des trois présents sur la chaîne principale des protéines : ^{1}H (99,98%), ^{13}C (1,11%) et ^{15}N (0,37%) il sert de point de départ aux stratégies d'attribution. Dans les faits, pour une protéine de 100 acides aminés, l'attribution peut se faire pour 400 à 700 protons non-labiles (qui ne s'échangent pas avec le solvant) et pour 110-140 protons labiles qui ne sont observables par RMN que dans certaines conditions. Afin de gérer cette difficulté, l'identification des systèmes de spins se fait en différents temps.

Tout d'abord, les motifs des protons non-labiles sont identifiés, pour chaque système de spin individuellement, dans une solution de la protéine sous sa forme native dans l'eau lourde (D_2O) en utilisant les corrélations scalaires ^{1}H-^{1}H. Dans ce milieu, les protons amides ainsi que ceux liés à un atome d'azote dans les chaînes latérales s'échangent avec le deutérium du solvant, ils ne présentent donc aucun signal dans les spectres RMN. Dans un second temps, l'étude de la protéine en solution dans H_2O, permet de compléter l'identification des systèmes de spin grâce aux corrélations scalaires avec des protons labiles. Les voisins directs de chaque système de spin sont identifiés à l'aide des NOE provenant des courtes distances entre les protons des acides aminés successifs : $d_{\alpha N}$, d_{NN} et $d_{\beta N}$. Le but de ces trois étapes est d'identifier des groupes de lignes de résonance proton qui correspondent à des segments de peptides assez longs pour être uniques dans la séquence primaire de la protéine. L'attribution spécifique avec la séquence est réalisée en identifiant (Figure : 1) les segments de peptides à la séquence d'acide aminé obtenue de façon indépendante.

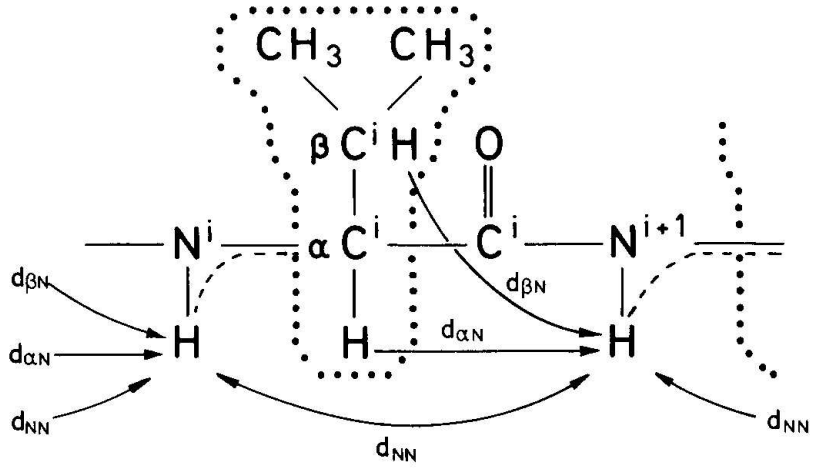

FIG. 1 – Segment polypeptidique; présentation d'un système de spin de protons non-labiles à l'intérieur des points; des connexions H^{α}-NH COSY en pointillés et des connexions liées aux NOE séquentiels représentées par des flèches.

Depuis cette première approche, la RMN a connu deux évolutions : le marquage en ^{15}N, permet d'éditer l'information spectrale proton (NOESY ou COSY) en utilisant une troisième dimension spectrale correspondant à la fréquence de l'azote 15. Cette évolution a consacré le spectre 2D ^{1}H-^{15}N HSQC [Bodenhausen Ruben, 1980] comme empreinte de la protéine. Ce spectre a la particularité de présenter une tache de corrélation par acide-aminé présent dans la séquence (excepté les acide aminés de type proline). De plus pour des

protéines allant jusqu'à 15 kDa, les spectres sont très bien résolus. Cependant, cette évolution n'a pas fondamentalement modifié le processus d'attribution [Marion *et al.*, 1989]. La seconde innovation est le marquage en ^{13}C [Ikura *et al.*, 1990]. La disponibilité de protéines marquées en ^{15}N et ^{13}C a par contre fondamentalement modifié le processus d'attribution. Les expériences RMN sont alors différentes en fonction des corrélations que l'on veut observer. Les spectres triple-résonances classiques utilisés lors de l'attribution d'une protéine sont (Figure : 1.1) l'HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HN(CA)CO et HNCO.

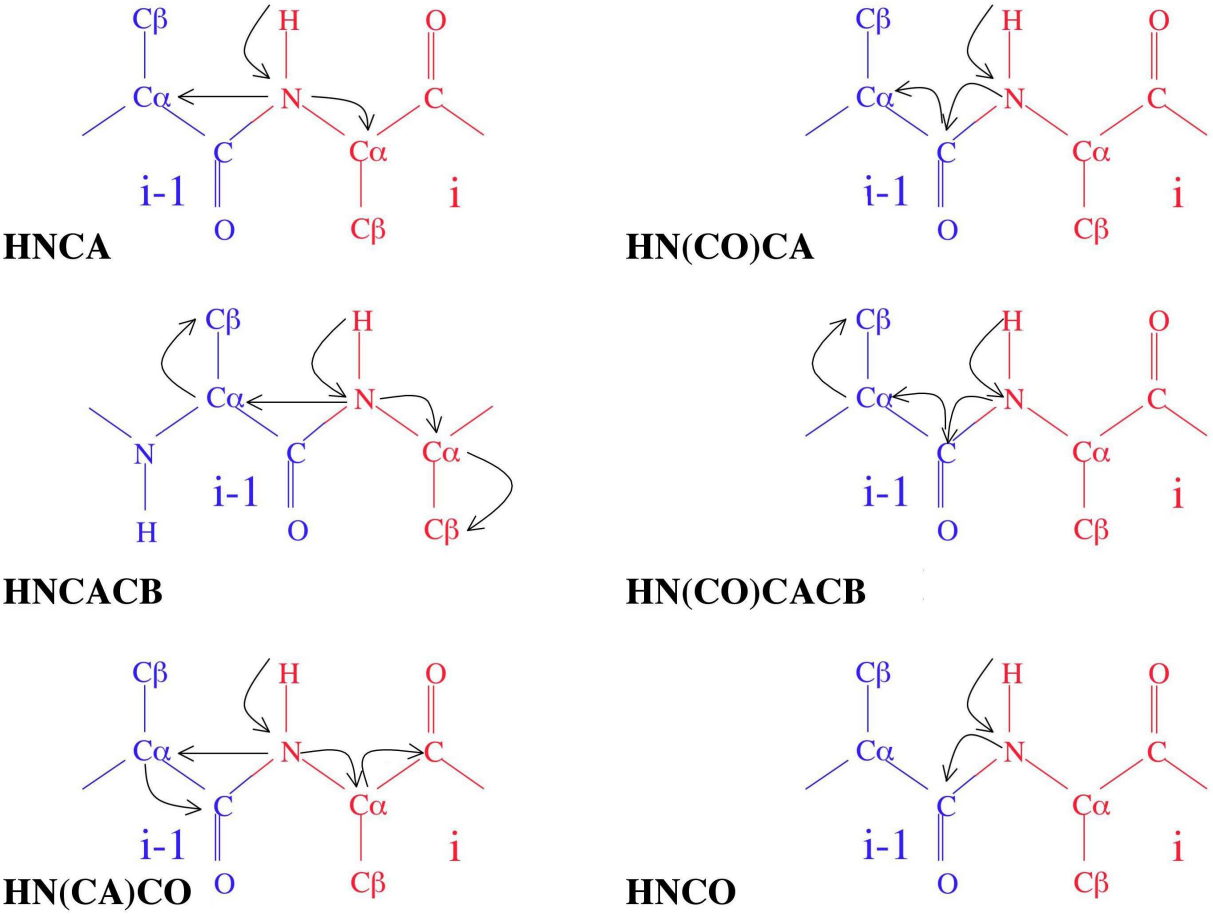

FIG. 2 – Corrélations issues des expériences tridimensionnelles hétéronucléaires RMN utilisées de façon courante pour obtenir les informations sur les couplages scalaires.

Depuis le début des études des protéines par RMN multi-dimensionnelle, de nombreuses tentatives d'automatisation du processus d'attribution ont été proposées (Figure : 3).

Des programmes ont été élaborés avec pour objectif, soit l'attribution complètement automatique (par ex : ANSIG), soit l'aide à l'attribution (par ex : Smartnotebook).

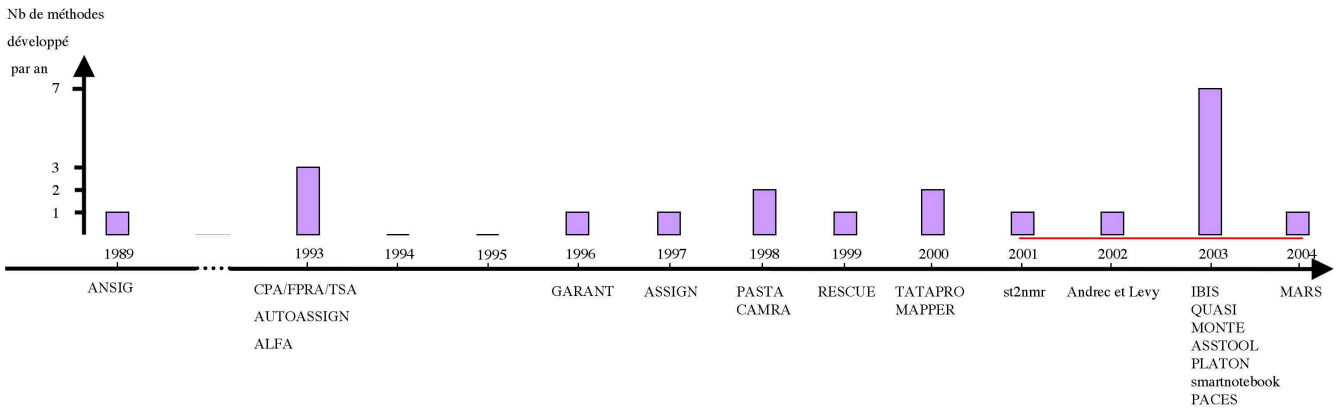

FIG. 3 – Chronologie retraçant l'apparition des programmes d'attribution séquentielle. Les années, ainsi que le nom des méthodes correspondantes, sont en abscisse. Le nombre de méthodes nouvelles par an est en ordonnée. La ligne rouge horizontale indique la période de mon travail de thèse.

Dans les deux cas, l'étape déterminante est la sélection des pics de corrélation sur les spectres. Cette sélection s'est avérée particulièrement délicate dans le cas des spectres homonucléaires du fait de l'encombrement spectral et de la présence de nombreux artefacts. L'utilisation d'échantillons marqués associés à la spectroscopie triple-résonance a permis de simplifier considérablement ce problème et a relancé le développement des méthodes d'attribution automatique. Une des voies qui est actuellement explorée consiste à coupler le calcul de la structure tridimensionnelle avec l'attribution. Dans cette logique, certains chercheurs ont proposé des méthodes de calcul qui évitent toute étape préalable d'attribution, celle-ci étant effectuée de façon implicite au cours du calcul de la structure 3D.

C'est dans cette optique que s'inscrit mon projet de thèse : le développement d'une méthode d'interprétation rapide des spectres RMN pour les protéines. Ce projet de thèse est lié à deux autres projets menés à l'Institut :

- la production de protéines recombinantes marquées aux isotopes stables en utilisant la cyanobactérie *Anabaena* sp. PCC 7120. Ce système permet d'obtenir des molécules marquées à des coûts compétitifs et permet d'envisager l'utilisation de stratégies de marquage spécifique dans le processus d'attribution.

- l'étude du fragment de la sous-unité B de l'ADN-gyrase (topoisomérase de type II) de *E. coli* était en cours au laboratoire depuis plusieurs années. Il fait l'objet de simulations de dynamique moléculaire réalisées par M. Schechner [Schechner *et al.*, 2004] afin de caractériser l'influence de la fixation des ligands sur la flexibilité de la protéine.

Cette protéine a été choisi pour valider le programme QUASI et l'étude de la dynamique a été réalisé de façon à pouvoir comparer les résultats de la simulation à des données expérimentales.

Partie I

Les stratégies d'automatisation d'études de protéines par RMN

1

Les premières tentatives d'automatisation de l'interprétation des spectres de protéine

Sommaire

Dans un premier temps, les efforts se sont portés vers des programmes d'automatisation complète.

1.1 ANSIG : Assignment NMR Spectra by Interactive Graphics

Ecrit en FORTRAN-77, ANSIG est un logiciel qui vient du groupe d'Uppsala (Suède) [Kraulis, 1989]. Il utilise des spectres 2D protons pour obtenir les liaisons chimiques et spatiales nécessaires à l'attribution séquentielle. L'objectif est d'obtenir un logiciel graphique qui permet d'analyser et d'attribuer les spectres 2D comme on le ferait à la main. De plus, il doit réaliser le référencement des opérations et la vérification de cohérence.

Les données d'entrée sont les parties réelles des spectres 2D ^{1}H traités, et éventuellement un dictionnaire des protons trouvés dans les résidus de la protéine, une description des spectres et la séquence primaire de la protéine. L'objet de base est le pic, toutes les opérations d'attribution y font référence. Ils sont extraits des spectres par un algorithme simple de sélection de pics. La liste peut être éditée ; l'utilisateur peut effacer, bouger, modifier les pics afin de corriger les résultats de la procédure. Les pics peuvent être attribués dans les deux dimensions. Une attribution est marquée par 3 entrées, le numéro du résidu, le type du résidu et le numéro du proton ; chaque valeur peut être modifiée de façon indépendante. L'évolution de l'attribution est gardée en temps réel par le programme, de plus il en vérifie la cohérence (pas de H^β pour la Gly) et les NOE qui en découlent sont mis à jour automatiquement. La valeur du déplacement chimique de chaque proton de la table d'attribution est calculée comme la valeur moyenne pondérée des coordonnées des déplacements chimiques de tous les pics impliqués dans le noyau, afin d'égaliser les petites erreurs dans les coordonnées des pics. L'écart-type des déplacements chimiques est calculé avec la valeur moyenne, afin de s'assurer que le proton n'est pas attribué plusieurs fois.
Des spectres enregistrés dans différentes conditions expérimentales sont utilisés pour résoudre les problèmes de recouvrement et de dégénérescence de pics. Habituellement, on enregistre un spectre HOHAHA et/ou une NOESY dans l'eau, et d'autres dans du D_2O. Le programme gère ceci en permettant à l'utilisateur de spécifier quels spectres sont considérés comme équivalents, la table d'attribution a alors une entrée par jeu de spectres équivalents. C'est la base d'une partie importante d'ANSIG, l'identification automatique des attributions possibles des pics NOESY. Il est primordial que ces attributions soient correctes, et ceci nécessite que chaque paire possible de protons soit étudiée pour chaque pic NOE. Dans ce programme, l'utilisateur peut sélectionner un pic sur le graphique et le programme lui indique la liste des protons qui correspondent au pic sélectionné (avec une marge d'erreur).
Ce programme est constamment amélioré et une version plus récente de ce logiciel [Helgstrand *et al.*, 2000] permet l'utilisation de spectres triple-résonances. Il est régulièrement couplé au programme AZARA [Boucher, 2002] depuis 1989. C'est essentiellement pour son aspect graphique, qui permet une analyse fine des spectres, que ce programme est utilisé.

1.2 CPA/FPRA/TSA

Cette méthode d'attribution de spectres 2D homonucléaires est basée sur la théorie des graphes, la reconnaissance de motifs dans des graphes flous et la recherche dans les arbres [Xu *et al.*, 1993]. Trois algorithmes sont utilisés :

- **CPA : Constrained Partitioning Algorithm** extrait et identifie le réseau des couplages de spins à partir d'une combinaison de spectres COSY et/ou TOCSY. Mathématiquement, un graphe topologique des spins peut se présenter sous la forme d'une matrice, d'une table de connectivité, d'un arbre binaire ou d'un ensemble d'arcs. Un pseudo-résidu correspond à un sous-graphe des pics DQF-COSY. Dans la théorie des graphes, ce sous-graphe est un ensemble d'arcs de couplage de spins. Dans le cas idéal, un sous-graphe des pics DQF-COSY correspond à un pseudo-résidu entier qui correspond à un acide aminé (Figure : 1.1), mais en réalité, du fait de problème de recouvrement des pics, on a des sous-graphes qui ne correspondent qu'à une partie du pseudo-résidu. La raison pour laquelle les topologies complètes ne peuvent être construites directement par CPA est que certains arcs ne sont pas prédits. La sortie de CPA est un jeu de pseudo-résidus sous la forme de tables adjacentes. Les informations permettant le regroupement des pics en sous-graphes sont aussi répertoriées afin que l'utilisateur puisse vérifier par lui-même.

- **FPRA : Fuzzy Pattern Recognition Algorithm** place les réseaux des couplages de spins sur des résidus spécifiques. Cet algorithme compare entre les graphes obtenus à la sortie de CPA et les graphes idéaux. Ceci se fait en deux temps, la reconnaissance de motif pour décider si on a des sous-graphes du graphe idéal d'un acide aminé donné, puis le calcul des valeurs d'appartenance pour déterminer quel sous-graphe est le plus proche du type d'acide aminé considéré (dans le cas où il y a plusieurs candidats). En théorie, les 20 acides aminés donnent 27 systèmes de couplage de spins (3 topologies pour le tryptophane, 2 pour l'asparagine, la glutamine, l'histidine, la phénylalanine et la tyrosine) ; ils forment un espace topologique flou. FPRA trouve un "bon positionnement" d'une topologie sur cet espace. De nombreuses sources d'erreurs, comme la dégénérescence ou le manque de pics COSY rendent l'attribution incertaine. Afin d'éviter de faire des erreurs, FPRA n'attribue pas le motif à un acide aminé mais donne les candidats possibles.

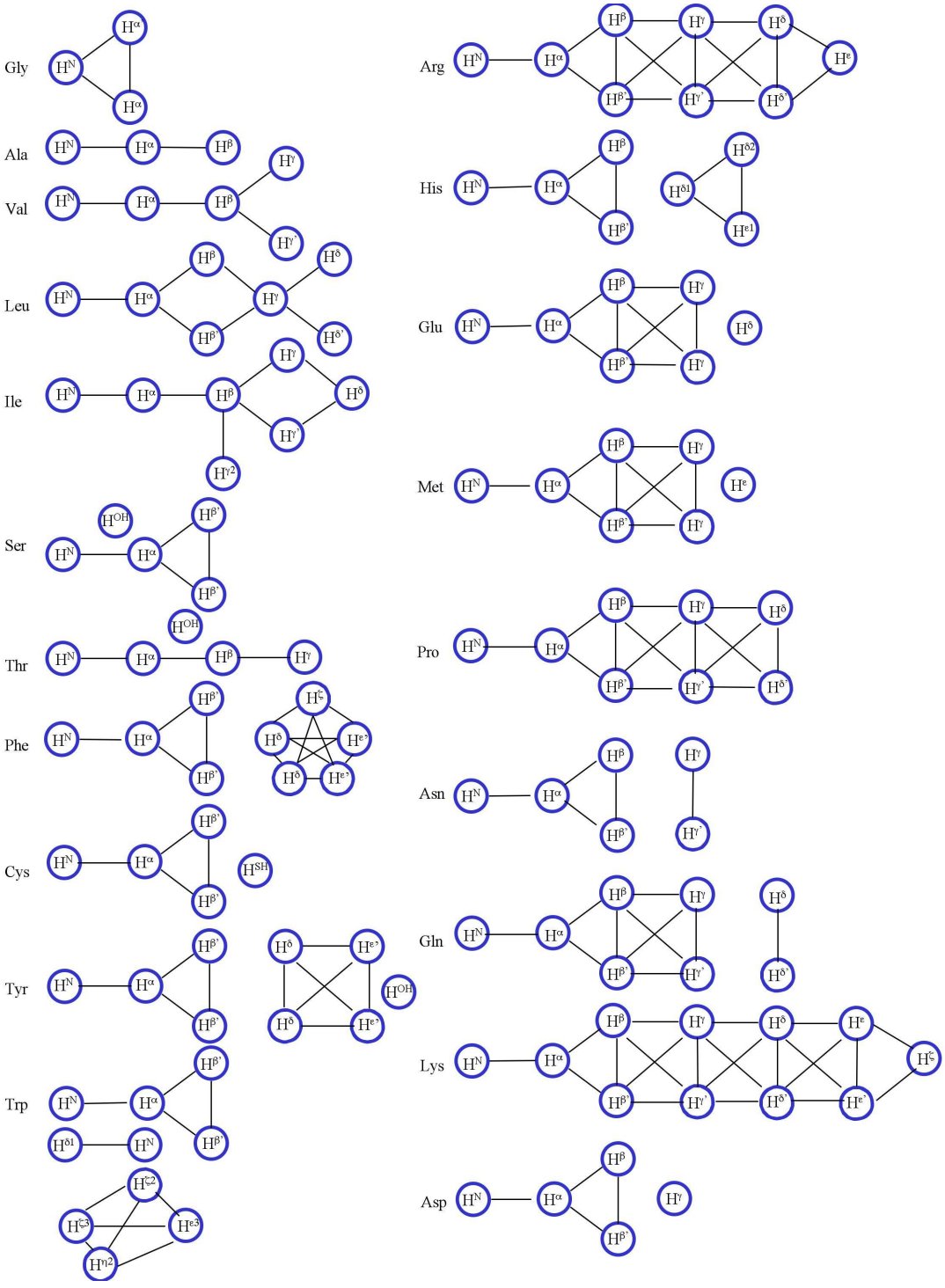

FIG. 1.1 – Les 27 systèmes de couplage de spin des 20 acides aminés.

- **TSA : Tree Search Algorithm** extrait l'attribution spécifique à la séquence, qui existe à l'intérieur du graphe, à partir des pics NOESY. L'algorithme de recherche dans les arbres est utilisé pour trouver la meilleure correspondance entre la topologie et la séquence primaire. Ceci est réalisé en cherchant le maximum de corrélations NOESY entre un candidat et la séquence primaire, et en gardant le graphe qui a le plus grand nombre d'attributions réalisées. Ces comparaisons se font en terme de supergraphes ; un pour la topologie de recherche et un pour la séquence primaire. TSA tente de trouver le maximum de connexions, donc même si il manque un pic de corrélation, deux résidus peuvent être connectés. En calculant d'abord les corrélations NOESY, pour éviter de refaire plusieurs fois le même calcul et en attribuant des fragments, comme on le ferait à la main, on diminue significativement le nombre de possibilités. Lorsque la protéine à traiter est de taille importante, la recherche du meilleur chemin peut aboutir à une explosion combinatoire. TSA permet à l'utilisateur d'attribuer la séquence fragment par fragment.

Outre BPTI, aucune protéine n'a été attribuée par ces trois algorithmes. Toutefois l'approche qui consiste à mêler attribution et théorie des graphes est intéressante. Cela permet de manipuler les déplacements chimiques et leurs relations avec une représentation graphique naturelle. La théorie des graphes apporte dans son sillage de nombreuses techniques et de nombreux outils mathématiques, toutefois leur maîtrise nécessite un apprentissage important. Ce facteur joue contre cette méthode.

1.3 AUTOASSIGN

Publié dès 1993, AUTOASSIGN [Zimmerman *et al.*, 1993, Zimmerman *et al.*, 1997] est un programme qui détermine l'attribution des résonances du squelette et du $^{13}C^{\beta}$ en utilisant des méthodes d'intelligence artificielle. La première étape d'analyse consiste à relever les correspondances les plus fortes. Seuls les déplacements chimiques pouvant être attribués sans ambiguïté sont examinés dans les échelles (CA ou CO) (Figure : 1.2). L'échelle CA regroupe les déplacements chimiques intrarésiduels et l'échelle CO les déplacements chimiques séquentiels. Puis on permet les déplacements chimiques dégénérés, cette étape affine certaines échelles non complètes et permet de lier les autres non encores liées. La troisième étape, dite d'attribution étendue des fragments, utilise les attributions et liens établis pour guider la spécification des échelles désignées de façon incomplète. Cette étape a pour but d'étendre les zones attribuées de la séquence. Dans le cas où il y a un nombre supérieur de pseudo-résidus par rapport au nombre de résidus dans la séquence primaire, les pseudo-résidus avec les intensités de pics les plus basses sont mis de côté dès le début.

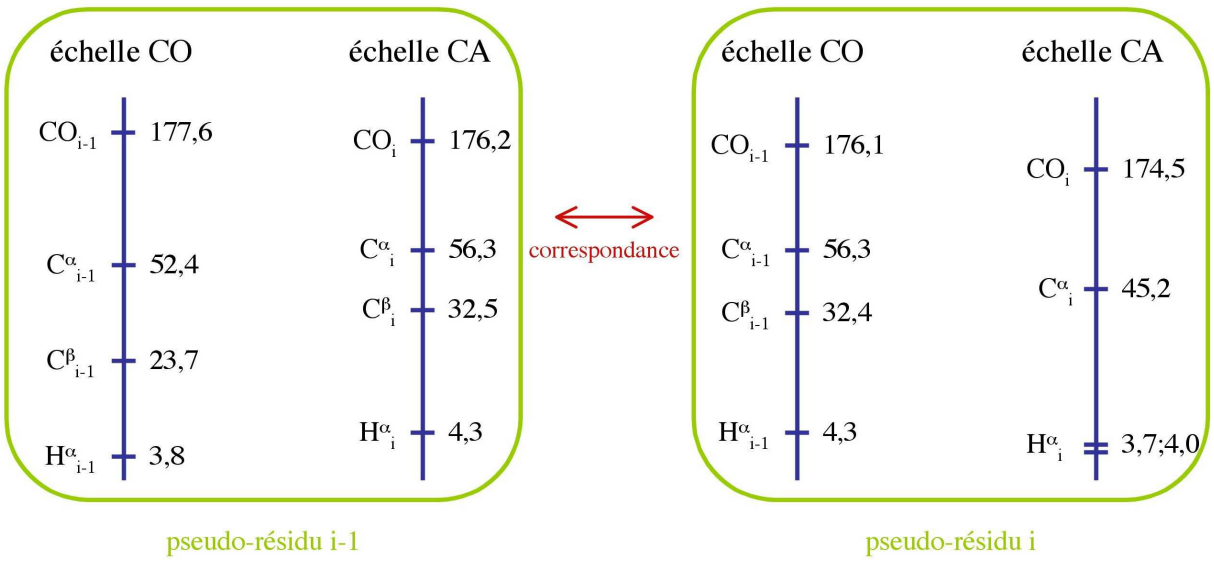

FIG. 1.2 – Echelles de comparaison utilisées dans AUTOASSIGN. L'échelle CA de gauche correspond à l'échelle CO de droite. Les deux pseudo-résidus correspondent et vont s'enchaîner.

La quatrième étape reprend ces pseudo-résidus et les remet dans les données de départ pour un nouveau cycle. L'ultime étape est l'attribution complète ; il s'agit d'examiner les déplacements chimiques et leur attribution séquentielle en quête d'éventuels désaccords, les corrigeant si possible et en concluant avec un nouveau cycle.

Ce programme est le plus souvent appliqué. Il a été utilisé pour l'attribution de plus d'une vingtaine de protéines depuis 1993.

1.4 ALFA : ALgorithm for Fast Assignment

Les données de départ sont les pseudo-résidus issus des spectres NOESY et TOCSY ; dans certains cas, il est possible d'utiliser une structure ou la connaissance de la structure secondaire comme informations supplémentaires. Il faut, de plus, une liste des pseudo-résidus classifiés en fonction du type d'acide aminé, et une liste de voisins potentiels ainsi que leur nombre de contacts inter-résiduels. L'algorithme utilisé dans ALFA [Bernstein *et al.*, 1993] est basé sur la technique de la minimisation combinatoire. L'accord entre les informations expérimentales observées sur les spectres et l'hypothèse d'attribution est décrit par une pseudo-énergie. Le principal terme d'énergie est celui lié à la topologie ($E_{topologie}$). Il dépend de la manière dont un pseudo-résidu observé correspond au résidu auquel il est attribué dans la séquence et de l'établissement de contacts inter-

résiduels dans le cas ou deux pseudo-résidus sont attribués à des positions adjacentes. D'autres termes sont liés aux contacts à longues distances que l'on peut tirer de structures 3D disponibles, aux structures secondaires et à la prise en compte que la probabilité conditionnelle que deux résidus adjacents développent des liaisons à longues distances avec deux autres résidus également adjacents est forte (surtout vrai pour les brins-β).
Le programme minimise l'énergie totale. Après une première attribution choisie de manière aléatoire, le programme sélectionne au hasard une paire de fragments. L'attribution des résidus sélectionnés dans les deux fragments sont réarrangés de façon systématique pour optimiser l'énergie totale. Si aucune attribution n'est acceptée, un pseudo-résidu pris au hasard est attribué au résidu. Les deux fragments modifiés sont replacés dans la séquence et deux autres fragments sont utilisés. Cette procédure est répétée jusqu'à ce que l'énergie ne puisse plus être minimisée. Le programme produit l'attribution optimale ainsi que les statistiques des étapes intermédiaires de l'attribution. A chaque étape, le programme utilise toutes les données d'entrée, et les attributions incorrectes sont supprimées au début de chaque cycle. L'algorithme est donc robuste face aux données erronées. ALFA a été testé sur la protéine MPI (mucous trypsin inhibitor) formée de 107 acides aminés. Le taux d'erreur dans l'attribution est de 17%. Ce programme est souvent cité parmi ceux qui comptent pour l'automatisation de l'attribution, mais il n'est jamais cité comme ayant servi à l'attribution d'une protéine.

1.5 GARANT : General Algorithm for Resonance AssignmeNT

GARANT [Bartels *et al.*, 1996] se veut une approche systématique et complètement automatique d'attribution du squelette des protéines en se basant sur les déplacements chimiques des spectres COSY, NOESY, TOCSY et éventuellement sur ceux de protéines homologues. GARANT permet de combiner l'usage des pics observés dans différents spectres NOESY et d'informations supplémentaires venant de protéines homologues (Figure : 1.3). Avec GARANT, les pics attendus sont corrélés aux pics observés et la meilleure correspondance donne l'attribution des résonances.

Dans un premier temps, la liste de pics attendus est construite à partir de la séquence primaire. La seconde étape consiste à placer ces pics sur les pics observés. Les pics liés aux couplages scalaires sont prédits à partir des connaissances de la séquence, de la structure covalente des acides aminés, et de différentes expériences RMN. La liste des pics observés est générée avec différents algorithmes qui se trouvent dans XEASY [Bartels *et al.*, 1995].

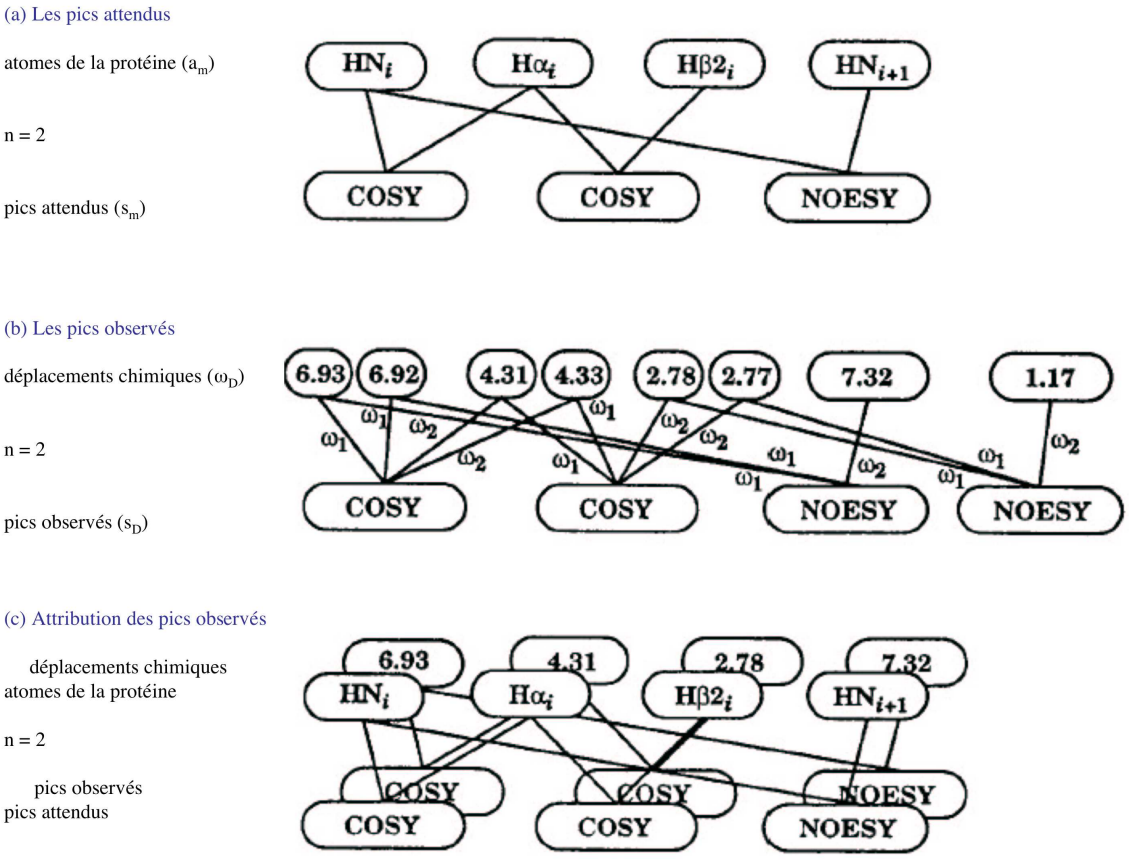

FIG. 1.3 – Représentation schématique (a) des pics attendus, (b) des pics observés, (c) de la grille utilisée pour décrire les attributions possibles dans le cas de spectres 2D homonucléaires.

A ce niveau, GARANT offre la possibilité d'utiliser des informations additionnelles issues de protéines homologues afin d'améliorer la prédiction des pics NOESY attendus. Les listes de pics sont sorties des spectres automatiquement avec XEASY. Alors que les pics observés sont caractérisés par des positions précises, les pics attendus sont donnés par des positions approximées. Les attributions obtenues par GARANT sont définies comme les positions optimales des pics attendus sur les pics observés. Pendant l'attribution, le programme évalue les attributions intermédiaires afin de guider le processus. Un algorithme génétique utilise les principes de sélection et d'héritage pour optimiser la population des solutions. Une routine d'optimisation locale est par la suite appliquée pour identifier, parmi les générations parents, les combinaisons de solutions qui permettront une solution susceptible de s'améliorer lors des générations suivantes.
Les structures 3D de protéines homologues sont utilisées pour trouver des pics NOE

issus de liaisons proton-proton de courtes distances, ils sont ajoutés à la liste des pics attendus. Des paramètres additionnels, comme une marge sur les distances proton-proton et un nombre minimal de structures homologues qui doivent contenir la courte distance, déterminent à quel moment il faut inclure un NOE entre deux atomes d'hydrogène donnés dans la liste des pics attendus.
Un autre type d'information issu de l'homologie affecte les distributions de probabilité des pics attendus. Au point de départ, les distributions sont considérées comme Normales, avec une moyenne et un écart-type, qui sont issus de librairies. Quand les déplacements chimiques de protéines homologues sont connus, les valeurs moyennes sont égales à celles des déplacements chimiques homologues et des écart-types beaucoup plus petits sont utilisés (choisis par l'utilisateur). Ainsi la connaissance de ces informations permet une description beaucoup plus fine des positions des pics attendus.

Depuis cette première version, le programme à été amélioré. Il fait partie, avec AUTOPSY [Koradi *et al.*, 1998] et PICS [Malmodin *et al.*, 2003] d'un groupe de programmes d'automatisation complète des spectres RMN triple-résonances. GARANT a été utilisé pour l'attribution d'une dizaine de protéines. Ces attributions sont partielles et complétées de façon manuelle.

Les premières versions de ces programmes ont pour but l'attribution complètement automatisée. Les attributions sont alors connues avec un pourcentage d'erreur caractéristique de l'accord global. Les mauvaises attributions ne sont pas localisées, or elles sont susceptibles d'entraîner des erreurs dans la suite de l'étude structurale. Cette démarche ne donne donc pas une réponse complètement satisfaisante à la question d'attribution rapide des résonances RMN. C'est pourquoi, depuis quelques années, les efforts se tournent vers des méthodes semi-automatiques qui permettent à l'utilisateur de suivre pas à pas les décisions prises par le programme et d'intervenir le cas échéant (IBIS, PACES...).

2

Développements récents de programme d'attribution

Sommaire

2.1 Les données d'entrée

Afin d'utiliser les programmes d'attribution, il faut définir les données qui seront utilisées comme point de départ par la méthode. Ces données représentent le premier contact de l'utilisateur avec le programme. En fonction de la complexité de ces données, le programme sera plus ou moins utilisé. En effet l'étape d'attribution reste, encore pour l'instant, un passage obligatoire que l'on désire rapide. Ainsi un programme avec comme point de départ des connaissances très complètes de la structure de la protéine, ou des spectres expérimentaux très particuliers risque d'être peu utilisé. Ceci peut certainement expliquer le fait que, comme nous le voyons dans le tableau : 2.1, la plupart des programmes utilisent des listes de pics issus de spectres triple-résonances hétéronucléaires (Tableau : 2.1).

Nous verrons également quelques cas particuliers.

Méthode	Données d'entrée	Stratégie
ASSIGN	Spectres 3D	Probabiliste
ASSTOOL	Spectres 3D	Probabiliste
AUTOASSIGN	Spectres 3D	Déterministe
GARANT	COSY-NOESY-TOCSY	Algo. génétique
IBIS	Spectres 3D	Déterministe
MAPPER	Fragments	Déterministe
MARS	Spectres 3D	Déterministe
MONTE	Spectres 3D	Probabiliste
PACES	Spectres 3D	Recherche exhaustive
PASTA	Spectres 3D	Acceptation à seuil
PLATON	Spectres 3D	Déterministe
PROCESS	δ_{ORB}, $\delta_{CAPTURE}$	Déterministe
QUASI	Spectres 3D	Déterministe
RESCUE	COSY-NOESY-TOCSY	Réseau de neurones
Smartnotebook	Spectres 3D	Déterministe
st2nmr	COSY-NOESY-TOCSY	Probabiliste
TATAPRO	Spectres 3D	Déterministe

TAB. 2.1 – Tableau récapitulatif des données d'entrée et des stratégies utilisées par les programmes d'attribution automatique. La notation "Spectres 3D" indique une partie ou la totalité des spectres 3D classiques utilisés pour l'attribution.

2.1.1 Les spectres RMN hétéronucléaires

Dans le cas général, les programmes utilisent des listes de pics issus de spectres triple-résonances hétéronucléaires. Ces listes sont créées au préalable soit de façon manuelle (avec un programme de visualisation des données), soit par un programme de traitement des données. Cette étape de sélection de signaux est particulièrement ardue car elle rencontre les problèmes liés au recouvrement des signaux. Les programmes tels que IBIS, MAPPER, MARS, MONTE, PASTA et Smartnotebook utilisent la totalité ou un sous-ensemble des spectres hétéronucléaires dits classiques : HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HN(CA)CO, HNCO.

En plus de ces spectres, ils utilisent un spectre de référence qui est souvent un spectre 2D ^{1}H-^{15}N HSQC ou quelque fois une projection 2D (^{1}H-^{15}N) d'un spectre HNCO. La liste des protons amides est extraite de ce spectre 2D. C'est à partir de cette liste numérotée arbitrairement, que les pseudo-résidus seront référencés tout au long du processus d'attribution. En effet, pour chaque paire de déplacements chimiques ^{1}H-^{15}N on trouve, dans les spectres 3D, les déplacements chimiques des atomes $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}CO issus du même résidu *i* et du résidu précédent *i-1*.

Méthode	^{15}N HSQC	$^{13}C^{\alpha}$	$^{13}C^{\beta}$	^{13}CO	HN(CO)CB	HN(CA)HA	HA(CACO)NH	HNCB	COCAH	CA(CO)NH	HA(CA)NH	CANH
ASSIGN		√	√	√		√		√	√	√		
ASSTOOL	√	√	√	√								
AUTOASSIGN	√		√	√				√		√	√	√
IBIS	√	√	√	√								
MARS	√	√	√	√								
MONTE		√		√	√			√				
PACES	√	√	√	√								
PASTA	√	√	√	√								
PLATON*	√	√	√	√								
QUASI	√	√	√	√								
Smartnotebook	√		√	HNCO								
TATAPRO			√	√		√	√					

Méthode	2D COSY	2D NOESY	2D TOCSY	3D HSQC-TOCSY	3D ^{15}N HSQC-NOESY	3D ^{13}C HSQC-NOESY	HN(CO)CACB
GARANT	√	√	√		√	√	√
st2nmr		√			√	√	
RESCUE	√	√	√	√	√		

TAB. 2.2 – Tableau récapitulatif des spectres RMN hétéronucléaires utilisés par les programmes d'attribution de ligne de résonance. Le symbole √ dans la colonne $^{13}C^{\alpha}$ indique que le programme utilise les spectres HNCA et HN(CO)CA. De même, ce symbole dans la colonne $^{13}C^{\beta}$ indique l'utilisation des spectres HNCACB et HN(CO)CACB, dans la colonne ^{13}CO, il implique les spectres HN(CA)CO et HNCO. Lorsque le nom du spectre est indiqué à l'intérieur du tableau, seul ce spectre (de la paire concernée) est utilisé.(*) Le programme PLATON représente ces informations et réalise l'attribution d'une façon originale, il sera présenté de façon détaillée dans le prochain chapitre.

Problèmes liés à l'utilisation des spectres expérimentaux

La principale difficulté rencontrée lors de cette première étape, est l'identification des signaux RMN au sein des spectres. Lorsque ces derniers sont bien dispersés, les pics sont bien discernables les uns des autres et cette étape est aisée. Cette situation se rencontre généralement pour les spectres enregistrés sur de petites protéines (qui ne sont pas les cibles de programmes d'attribution). Dans la plupart des cas, certaines zones du spectre sont très denses et la séparation des signaux se fait difficilement. Il existe des programmes entièrement dédiés à l'analyse des spectres et quelques concepteurs de programmes d'attribution utilisent les formats de ces logiciels (MARS [Jung Zweckstetter, 2004] utilise SPARKY [Goddard Kneller, 2004], IBIS [Hyberts Wagner, 2003] utilise XEASY [Bartels *et al.*, 1995]...). Toutefois, quelque soit le logiciel utilisé, l'inspection manuelle des spectres est nécessaire afin de valider les listes de pics. La plupart des programmateurs préfèrent sélectionner les signaux manuellement afin de minimiser le nombre d'erreurs transmises au programme.
Notons que devant des pics surnuméraires, les programmes ne réagissent pas tous de la même manière : certains vont les éliminer "naturellement" au cours du processus et d'autres risquent de les attribuer comme de vrais pics.

2.1.2 Les données d'entrée atypiques

Pour certains programmes d'attribution automatique, les informations nécessaires au départ ne sont pas directement les valeurs des déplacements chimiques sélectionnés sur les spectres RMN. Ces informations restent à la base de ces méthodes, mais ce ne sont pas les programmes d'attribution qui les gèrent.

Le programme PROCESS est la partie décisionnelle du pack CAMRA (Computer Aided Magnetic Resonance Assignment) [Gronwald *et al.*, 1998] (Figure : 2.1). Il utilise les prédictions de déplacements chimiques faites par le programme ORB [Gronwald *et al.*, 1997] et les déplacements chimiques issus du programme CAPTURE. Ce dernier réalise la sélection des signaux et le regroupement de ceux-ci en pseudo-résidus. Le programme ORB utilise une base de données de protéines homologues déjà attribuées. Au lieu d'utiliser les connectivités, CAMRA se base uniquement sur les correspondances entre l'observation (CAPTURE) et la prédiction (ORB).

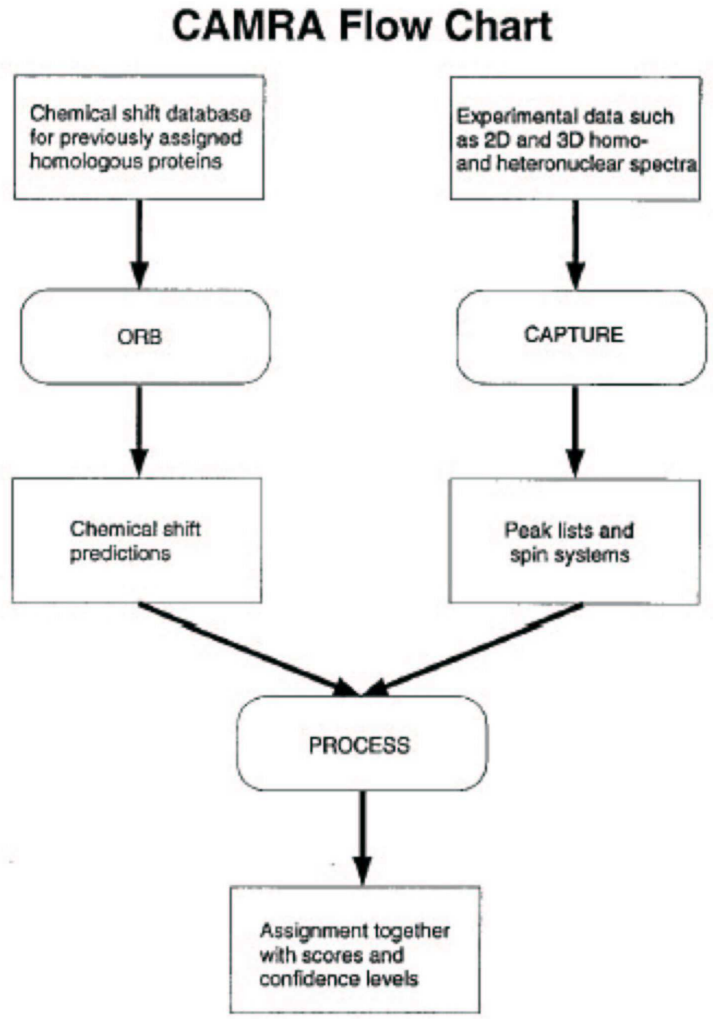

FIG. 2.1 – Diagramme du logiciel CAMRA qui décrit la suite des programmes qui le constitue.

Le programme d'attribution semi-automatique MAPPER [Güntert *et al.*, 2000] utilise lui, comme données d'entrée, des fragments de pseudo-résidus enchaînés séquentiellement. L'utilisateur (Figure : 2.2) crée ces fragments, avec l'aide d'un programme d'analyse spectrale, puis les confie au programme afin qu'il les place sur la séquence.

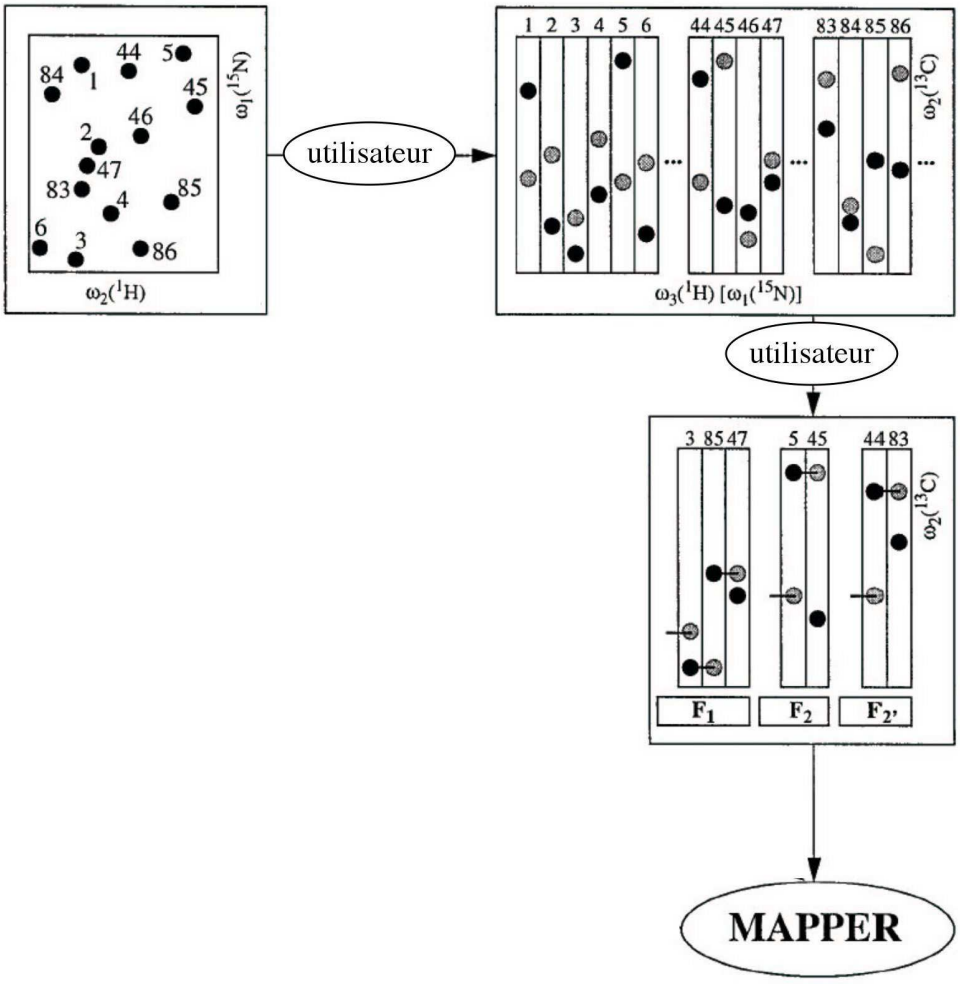

FIG. 2.2 – Représentation schématique des étapes de sélection des pics et d'enchaînement des pseudo-résidus qui précèdent l'utilisation de MAPPER.

2.2 Les stratégies

Les différentes méthodes se singularisent essentiellement par l'approche qu'elles utilisent pour connecter les pseudo-résidus puis pour placer les fragments ainsi formés sur la séquence primaire. L'attribution du squelette d'une protéine est un problème NP-Complet. On peut voir de façon intuitive les problèmes NP-Complets comme des problèmes pour lesquels la recherche d'une solution consiste à parcourir un arbre. Cet arbre de recherche contient l'ensemble des solutions possibles et une branche de cet arbre représente une éventuelle solution. Le nombre de ses branches évolue de façon exponentielle en fonction de la taille de l'arbre. Chaque noeud de l'arbre correspond à un point de choix pour une variable. La seule façon d'obtenir une solution acceptable est de parcourir l'arbre jusqu'à trouver une solution correcte, ce qui, dans le pire des cas, peut demander le parcours complet de cet arbre. Cette exploration devient très vite matériellement impossible et d'autres approches sont donc utilisées.
L'archétype de cette classe de problème est celui du *voyageur de commerce*, proposé au XIX^{ème} siècle par Sir William Rowan Hamilton (1805 - 1865), mathématicien irlandais et Thomas Penyngton Kirkman (1806 - 1895), mathématicien britannique. Il s'énonce ainsi : *Un voyageur de commerce doit, pour son travail, traverser un ensemble de N villes. Il doit toutes les traverser une fois et une seule. Le chemin choisi doit le ramener à son point de départ.* Plusieurs solutions ont été proposées pour ce type de problèmes. On peut distinguer deux catégories d'approches, les approches déterministes et les approches probabilistes.

2.2.1 Les stratégies déterministes

Les approches déterministes voient le problème de l'attribution comme évoluant de façon prédictible une fois que les données d'entrée sont fixées. Ainsi le comportement du système est complètement indépendant de la statistique. Le résultat obtenu est uniquement dépendant des données d'entrée. Parmi les approches déterministes, on trouve les algorithmes de recherche exhaustive ou bien les algorithmes du type meilleur premier (*best first search*) qui ne garde que la meilleure possibilité. Les programmes IBIS, MAPPER, MARS, PACES, PLATON, PROCESS, RESCUE, TATAPRO utilisent des stratégies déterministes. Parmi eux, la méthode d'attribution codée dans IBIS [Hyberts Wagner, 2003] est celle qui se rapproche le plus du processus de l'attribution faite à la main. Le processus d'attribution se fait à partir des données issues de XEASY. Les pseudo-résidus prennent la forme de fourche avec des branches en S (pour les pics séquentiels) ou en I (pour les pics intrarésiduels).

L'algorithme utilisé dans le programme PACES (Protein Sequential Assignment by Computer Assisted Exhaustive Search) [Coggins Zhou, 2003] fait une recherche exhaustive des pseudo-résidus pour établir les connexions séquentielles et les attributions. Les résonances résiduelles de chaque pseudo-résidu sont comparées aux résonances séquentielles de tous les autres afin de construire une table de connexions de dipeptides, pour deux systèmes i et j, une connexion est établie suivant le système d'équations 2.1.

$$\begin{aligned} |^{13}C_i^{\alpha} - ^{13}C_{j-1}^{\alpha}| &\leq \delta_{C\alpha} \\ |^{13}C_i^{\beta} - ^{13}C_{j-1}^{\beta}| &\leq \delta_{C\beta} \\ |^{13}CO_i - ^{13}CO_{j-1}| &\leq \delta_{CO} \\ |H_i^{\alpha} - H_{j-1}^{\alpha}| &\leq \delta_{H\alpha} \end{aligned} \tag{2.1}$$

où $\delta_{C\alpha}$, $\delta_{C\beta}$, δ_{CO} et $\delta_{H\alpha}$ sont des marges définies par l'utilisateur. Dans le cas de résonances manquantes, des règles ont été établies : s'il n'y a qu'un type d'atome, la connexion se fait d'après lui seul. Si il y a au moins deux types d'atomes, il faut au moins deux tests qui coïncident pour accepter la connexion. Les pseudo-résidus, avec des données manquantes, qui ne répondent pas à ces critères, ne sont pas utilisés dans ce processus. La liste complète des connectivités est stockée dans une table. Sur la base de cette table, de plus grands fragments sont assemblés. Les données sont vues comme un réseau directionnel (Figure : 2.3). Le point de départ est choisi arbitrairement, et tous les pseudo-résidus susceptibles de se trouver en N-terminal sont tracés. Les pseudo-résidus susceptibles de se trouver en C-terminal sont tracés au fur et à mesure qu'on les rencontre. Le premier pseudo-résidu de la table est sélectionné ainsi que son réseau de connexions. Chaque pseudo-résidu parcouru est marqué afin d'éviter les redondances. L'algorithme va au prochain noeud non marqué et regarde tous ces voisins. Si un pseudo-résidu se connecte à un autre qui est en bout d'un sous ensemble, les fragments se collent et la nouvelle connexion est acceptée. Dans le cas de données dégénérées, plusieurs scénari sont envisageables, PACES génère tous les fragments possibles et élimine les connexions circulaires. Avec cette approche, les fragments générés englobent toutes les connectivités potentielles.

L'identification des acides aminés est réalisée en comparant entre les déplacements chimiques observés et ceux de la BMRB. Tous les types d'acides aminés possibles pour un pseudo-résidu sont enregistrés de la même manière ; PACES ne les classifie pas.

Le placement des fragments sur la séquence se fait en glissant les fragments sur la

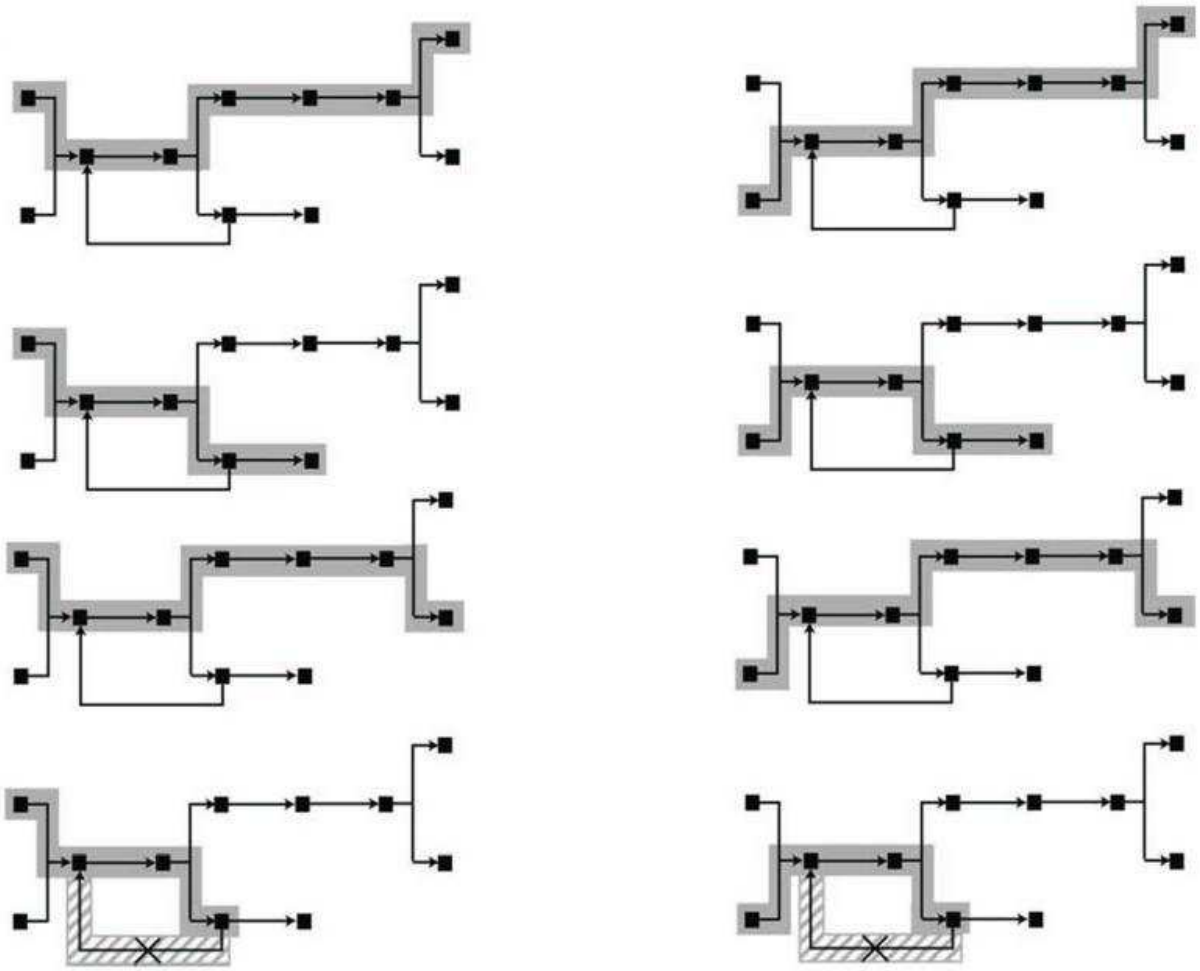

FIG. 2.3 – Exemple de réseau directionnel représentant les connexions entre les pseudo-résidus. Chaque carré est un pseudo-résidu et chaque flèche est une connexion. Les connexions grisées donnent des fragments générés, les connexions hachurées (qui sont circulaires) sont éliminées.

séquence primaire de façon circulaire afin que chaque pseudo-résidu passe sur chaque position. A chaque position, si le résidu de la séquence primaire coïncide avec un élément de la liste des types d'acide aminé possible pour le pseudo-résidu, on admet qu'il y a correspondance. Si plusieurs pseudo-résidus consécutifs s'accordent, l'algorithme identifie cela comme un fragment de correspondance. Pour le premier pseudo-résidu d'un tel fragment, le système précédent est immédiatement testé contre la séquence. Idéalement un fragment correct doit correspondre totalement à la séquence s'il est bien placé. On garde les morceaux de fragments qui ne se placent pas pour essayer de les positionner plus tard. Dans le cas où plusieurs fragments se positionnent au même endroit, on place le plus grand. Les fragments les plus longs ont les plus grandes probabilités d'être bien

placés. Une fois que ceux ci sont placés, PACES rejette les possibilités qui ne sont pas cohérentes avec les positions déjà occupées. Les fragments de 7 résidus et plus sont souvent placés de façon immédiate, mis à part les endroits où il y a conflit avec un autre gros fragment. Il peut arriver que plusieurs petits fragments se placent côte à côte : l'utilisateur peut alors examiner les liaisons et décider de les accepter ; ils seront ainsi considérés comme un gros fragment.
Ce programme propose la procédure de test la plus complète. Elle est réalisée sur 27 protéines de 76 à 723 acides aminés (10 à 80 kDa) dont trois jeux de données expérimentaux. Les tests ont tous été faits à partir des seuls déplacements chimiques des carbones. Les protéines sont réparties en 3 classes représentatives de données de moins en moins complètes.

Dans leur souci de faire une méthode robuste d'attribution complètement automatique, les concepteurs de MARS [Jung Zweckstetter, 2004] quant à eux proposent une méthode combinant l'accord local et global. Il est d'abord déterminé résidu par résidu, puis pour les fragments formés de 5 pseudo-résidus. Ces fragments sont construits en acceptant toutes les connexions qui se trouvent à l'intérieur du seuil choisi par l'utilisateur (même démarche que celle suivie dans le programme PACES). Les scores obtenus, sur la totalité de la séquence, sont triés dans l'ordre croissant. Le pseudo-résidu qui débute le fragment classé premier est pris comme point de départ pour reconstituer "en reculant" tous les fragments possibles. Puis le score est recalculé avec ces nouveaux fragments. Si le meilleur score est obtenu pour le même fragment (même constitution en terme de pseudo-résidus), l'attribution est favorisée. Ceci est répété pour tous les fragments de 5 pseudo-résidus puis pour les fragments de tailles inférieures. Les attributions faites par résidu et par fragment sont comparées et une attribution unique est conservée. Afin de minimiser l'influence de la qualité des déplacements chimiques prédits sur l'attribution, le processus d'attribution est répeté 40 fois en y introduisant progressivement du bruit.
Le programme MARS est également validé sur de nombreuses protéines : 14 représentant le même panel de taille que pour PACES. Pour 12 d'entre elles, les données sont issues de la BMRB, les deux autres, de 71 et 110 résidus sont des données issues de spectres expérimentaux.

Ces deux derniers programmes forment leurs fragments de la même manière. MARS les gèrent petit à petit et PACES les gèrent tous ensemble. Les démarches suivies par la suite sont pourtant très différentes. Alors que MARS offre une automatisation complète de l'attribution, PACES propose une aide à l'attribution. Cette nuance implique pour MARS la nécessité de remplacer l'oeil de l'utilisateur par plusieurs routines

informatiques. L'algorithme utilisé par le programme MARS est donc plus lourd et pesant.

Le programme PLATON [Labudde *et al.*, 2003] offre une démarche très différente des autres programmes. Il attribue à un jeu de déplacements chimiques intrarésiduels un type d'acide aminé avec sa structure secondaire. Il est utilisé lorsque la structure tridimensionnelle de la protéine cible est connue. Une base de données CSP (Chemical Shift Pattern) de référence est constituée à partir des déplacements chimiques des bases de données publiques de RMN. Le CSP est un vecteur de booléens qui décrit les positions relatives des déplacements chimiques par rapport à une référence. Le point de départ est la création d'un espace à N dimensions en fonction du nombre de types d'atomes disponibles dans la base de données : $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}CO, H^{α} ou tout sous-groupe. Il contient un '+' ou un '-' en fonction de la position du déplacement chimique étudié par rapport à une valeur de référence (Figure : 2.4). On obtient des CSP du type "+ + + - - - " pour un espace du type $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}CO, $^{13}C^{\alpha\prime}$, $^{13}C^{\beta\prime}$ et $^{13}CO'$. Le '+' est dans le cas ou le déplacement chimique étudié est plus grand que la référence, le '-' est ajouté dans le cas où il est égal ou inférieur à la référence.

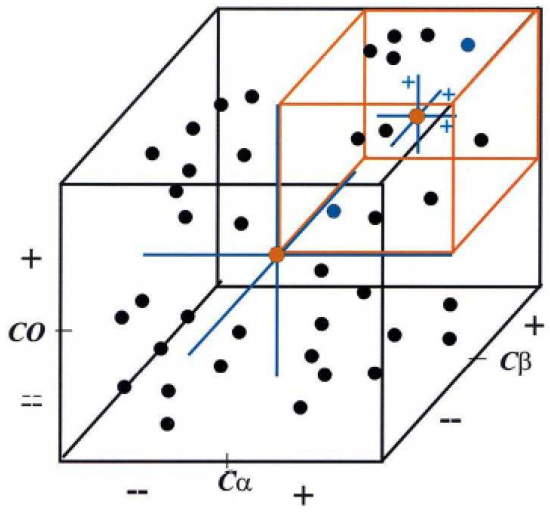

FIG. 2.4 – Illustration de la création de CSP dans un espace de déplacements chimiques à N dimensions. Ici les noyaux $^{13}C^{\alpha}$, $^{13}C^{\beta}$ et ^{13}CO sont pris en compte. Chaque point noir correspond à un acide aminé. Les points oranges sont les centres de masses qui servent de valeurs de référence pour la construction des CSP. Le cube de couleur orange est obtenu après une première subdivision de l'espace initial, les nouvelles coordonnées seront différenciées grâce à des apostrophes. [Labudde *et al.*, 2003]

Des données obtenues pour 51 protéines sont utilisées pour construire la base de données de référence CSP. Les données de 15 protéines sont sorties de la base de données du programme TALOS [Cornilescu *et al.*, 1999]. Toutes les protéines complexées ou avec un centre paramagnétique ont été exclues afin de supprimer les valeurs de déplacements chimiques atypiques. Après cette sélection, 36 autres protéines sont recrutées. Comme on se base sur des booléens, les références utilisées pour enregistrer les spectres (DSS, TSP, etc...) ne sont pas prises en compte lors de la sélection ; idem pour les types de structures secondaires qui n'ont pas été pris comme références. Toutefois, on note la représentation de 6 classes structurales (SCOP : Structural Classification of Proteins [Murzin *et al.*, 1995]) : tout-α, tout-β, α/β, protéines de surfaces et membranaires, peptides et petites protéines. Les déplacements chimiques de $^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}CO, H^{α} et les éléments de structure secondaire de 5896 acides aminés servent de données pour générer la base de données de référence de CSP.
Pour chacune des 51 protéines, les paramètres de Chou-Fasman sont calculés afin de s'assurer qu'ils reflètent bien un espace structural représentatif d'une protéine.
La base de données est étudiée en détail et un tableau à double entrée est créé (Figure : 2.5).

No.	CSP	ALA			ARG			...	Somme	Entrées	%
		B	α	β	B	α	β	...			
1	- - -	4	3	0	2	0	0	...	13	7	12%
2	- - +	0	0	0	1	0	0	...	14	8	14%
3	- + -	6	2	3	2	0	0	...	23	9	16%
4	+ - -	0	1	0	4	0	1	...	13	10	18%
5	+ - +	7	7	3	2	1	0	...	26	10	18%
6	+ + -	0	0	0	0	0	0	...	8	4	7%
7	- + +	5	0	5	4	0	0	...	32	11	19%
8	+ + +	0	0	0	1	1	0	...	10	6	11%
Somme		22	13	11	16	2	1				
Entrées		4	4	3	7	2	1				
%		50%	50%	37,5%	87,5%	25%	12,5%				

FIG. 2.5 – Exemple fictif de la présentation d'une base de données de référence CSP à 3 dimensions. La première colonne indique le numéro du CSP (de 1 à 8 (2^3)), la seconde est sa représentation sous la forme de booléens. Chaque type d'acide aminé est subdivisé en trois en fonction du type de structure secondaire : B pour la boucle, α pour l'hélice-α et β pour le brin-β. Chaque ligne expose les types d'acide aminé (ainsi que leur structure secondaire) que caractérise chaque CSP. Ici, le CSP '- - -' se retrouve 4 fois sous la forme d'une alanine dans une boucle, 3 fois sous la forme d'une alanine dans une hélice, 2 fois sous la forme d'une arginine dans une boucle...Dans la **colonne** intitulée **Somme** on retrouve le nombre total d'occurence de chaque CSP. La colonne suivante indique le nombre de types d'acides aminés caractérisés par chaque CSP. La dernière colonne indique si le CSP correspond à beaucoup de types d'acides aminés. Dans la **ligne** intitulée **Somme** on retrouve le nombre total de fois que l'acide aminé est présent dans la base de données. La ligne suivante indique le nombre de CSP qui code pour l'acide aminé. La dernière ligne indique le pourcentage de CSP correspondant à ce type d'acide aminé.

On y voit, sur une ligne, les types d'acides aminés (avec un type de structure secondaire) auxquels peut être relié un motif CSP. Dans une colonne, on trouve les types de motifs qui peuvent représenter un acide aminé dans une structure secondaire donnée. Par la suite, on utilisera le fait que les CSP des acides aminés non-attribués correspondent à une occurence élevée. Les CSP de la référence et de la protéine cible sont comparés et un tableau à deux entrées est créé avec les CSP de la protéine cible. Les colonnes du tableau sont interprétées en terme de fonctions de pénalité. Elles sont utilisées afin d'associer

chaque possibilité à une probabilité.
Ce programme n'est jamais cité comme ayant été utilisé pour l'attribution de protéines. L'approche très particulière et le choix d'une référence adéquate pourraient expliquer que cette méthode n'est pas encore utilisée.

Une démarche très différente est également suivie par le programme RESCUE [Pons Delsuc, 1999]. C'est un outil d'attribution des résonances des protons des protéines qui utilise un réseau artificiel de neurones. L'optimisation du réseau se fait à partir d'un sous-ensemble de 142 protéines représentatives de la BMRB. Cette approche permet de construire un programme s'appliquant à un grand nombre de cas, contrairement à des approches plus spécifiques privilégiées par d'autres programmes. La reconnaissance des pseudo-résidus peut s'effectuer à l'aide de deux réseaux :

- un premier réseau de neurones est créé afin de discriminer entre les 20 types d'acides aminés. On retrouve dans ce cas des difficultés pour les acides aminés qui sont proches comme Ile et Leu, Met, Glu et Gln, etc...

- un second réseau de neurones dans lequel les acides aminés sont regroupés (Tableau : 2.3), et dans un premier temps prédits comme membres du groupe. Puis une série de réseaux sont spécialisés dans la séparation des acides aminés au sein des groupes de la première étape. Ces réseaux seront référencés comme NN2.

Première étape	Deuxième étape
Ile, Leu	Ile
	Leu
Ala	Ala
Gly	Pro
Thr	Val
Lys, Arg	Lys
	Arg
Phe, Tyr, Trp, His, Asp, Asn, Cys	Phe, Tyr, Trp, His, Cys
	Asp, Asn,
Glu, Met, Gln	Glu,Gln
	Met
Ser	Ser

TAB. 2.3 – Groupes d'acides aminés utilisés par NN2.

Les réseaux de neurones utilisés sont tous des perceptrons.

Les Perceptrons

Le perceptron est le premier des réseaux de neurones. Il fut mis au point par Rosenblatt [Rosenblatt, 1957]. Le but du perceptron est d'associer des configurations (des formes) en entrée à des réponses. Le perceptron classique se compose de deux couches : la rétine et la couche de sortie qui donne la réponse correspondant à la stimulation présente en entrée. Les cellules de la première couche (Figure : 2.6) répondent en oui/non. La réponse 'oui' correspond à une valeur '1' et la réponse 'non' correspond à une valeur '0' à la sortie du neurone. L'état de sortie d'un neurone de la couche de sortie dépend de la somme pondérée des états des synapses. L'apprentissage du réseau consiste à modifier les valeurs des pondérations de façon à obtenir les valeurs de sortie souhaitées sur un jeu de valeurs d'entrée déterminé.

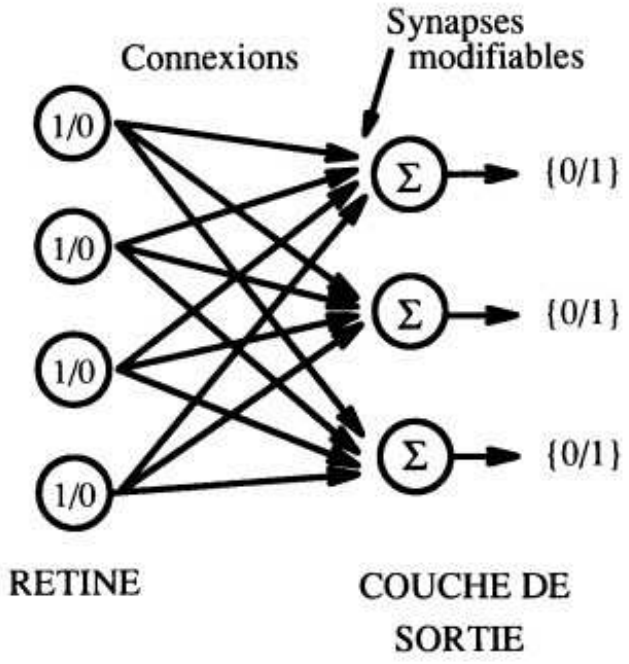

FIG. 2.6 – Représentation schématique d'un perceptron. La rétine est une couche d'unités sensorielles, fournissant des réponses modulées en fonction de l'intensité du stimulus, une couche de cellules d'association, constituant la couche de neurones formels proprement dits et une couche de cellules de décisions, qui représentent la réponse de sortie du perceptron. http ://www.isib.be/Etudiant/eln3/percep.htm

Dans le cas de RESCUE, la topologie retenue est un réseau à 3 couches : les déplacements chimiques sont présentés à la rétine, et les types d'acides aminés sont obtenus après la $3^{ème}$ couche (la couche de sortie).

Le nombre de déplacements chimiques varie d'un type d'acide aminé à l'autre et peut même varier pour un type d'acide aminé donné ; ceci n'est pas facilement gérable par un perceptron, d'où l'ajout d'une couche de logique floue [Zadeh, 1988] avant la rétine, afin d'avoir un nombre constant de déplacements chimiques à analyser. Cette couche est une grille à l'échelle des déplacements chimiques sur laquelle la position de chaque ligne spectrale est codifiée (Figure : 2.7).

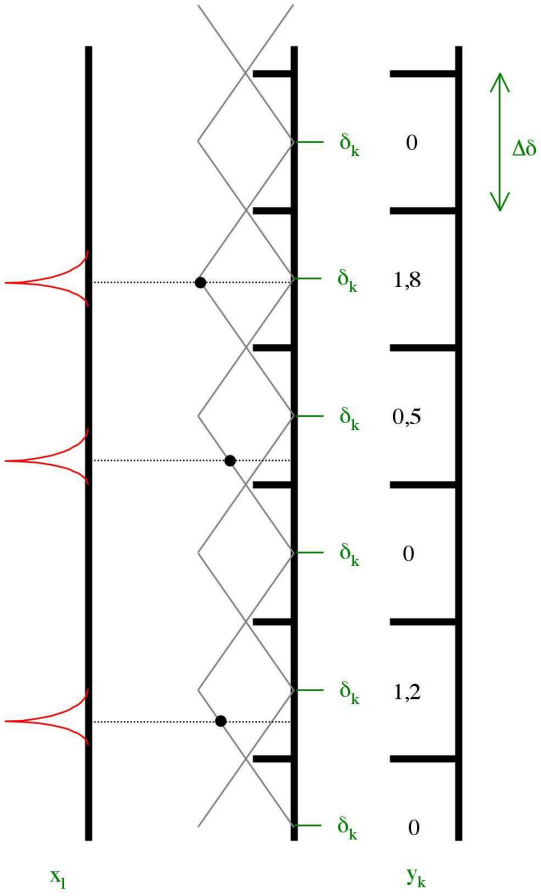

FIG. 2.7 – Représentation schématique du processus de codage des spectres RMN dans RESCUE. Les trois pics rouges représentent le spectre RMN, l'information qui en sort est les valeurs x_l des déplacements chimiques. Les lignes grises symbolisent la grille de logique floue codant les déplacements chimiques, elle échantillonne de façon régulière $\Delta\delta$ l'axe des déplacements chimiques en 6 entrées aux positions δ_k. La dernière ligne noire verticale présente les valeurs d'entrée y_k du réseau de neurones.

Dans le but d'avoir une résolution plus fine que celle de l'espacement $\Delta\delta$ entre les lignes de la grille, l'intensité y_k (Equation : 2.2) issue de la grille est proportionnelle à la distance de la ligne spectrale au centre δ_k de la division.

$$y_k = \sum_l \left(max \left(1 - \frac{||\delta_k - x_l||}{\Delta\delta}, 0 \right) \right) \tag{2.2}$$

L'introduction de logique floue permet de se rapprocher du traitement de l'information réalisé par l'oeil de l'utilisateur, cela allège les algorithmes utilisés. Cette approche rend RESCUE attractive.

2.2.2 Les stratégies probabilistes

Une approche probabiliste a pour caractéristique de ne pas donner deux fois de suite exactement la même réponse. Les diverses constantes intervenant dans un modèle paramétré ne sont pas uniquement déterminées en fonction des entrées données. Les programmes ASSIGN, ASSTOOL, MONTE et PASTA utilisent des stratégies probabilistes. L'algorithme le plus utilisé (par ASSIGN, MONTE...) pour réaliser le placement des fragments sur la séquence est le recuit-simulé. En 1983, Kirkpatrick et al. [Kirkpatrick *et al.*, 1983] proposent une méthode utilisant la *simulation Metropolis Monte Carlo* pour déterminer l'orientation la plus stable du système. Leur méthode est basée sur la procédure utilisée pour créer le verre le plus solide possible. Ce processus chauffe le verre à une température élevée pour que le verre soit liquide et que les atomes puissent bouger librement. La température est baissée lentement pour qu'à chaque pas les atomes soient assez libres pour adopter l'orientation la plus stable. Ce procédé de refroidissement est appelé le *recuit*. La méthode de Kirkpatrick et al. est connue sous le nom de *recuit simulé*. Le programme PASTA (Protein ASsignment by Threshold Accepting) [Leutner *et al.*, 1998] se démarque car il utilise une alternative au recuit simulé.

Le processus d'optimisation utilisé par PASTA est basé sur l'algorithme d'acceptation à seuil [Dueck Scheuer, 1990], une approche similaire à l'algorithme de simulation de recuit simulé. L'algorithme consiste à modifier une solution de départ (choisie au hasard) petit à petit afin d'optimiser la valeur de la fonction cible. Une modification élémentaire est acceptée si le gain de la fonction cible est supérieur à un seuil prédéterminé.

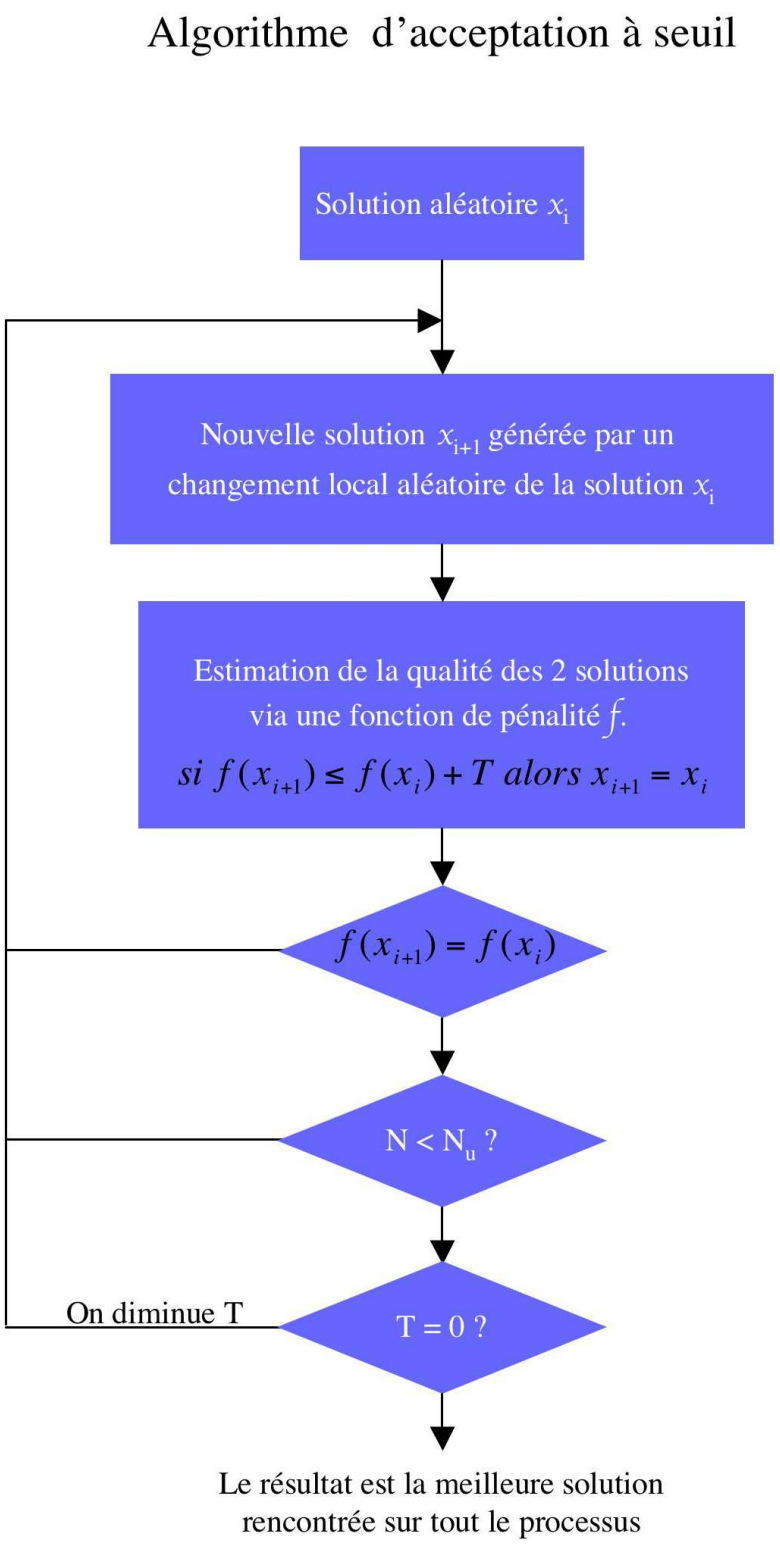

FIG. 2.8 – Schéma de l'algorithme d'acceptation à seuil. x_i et x_{i+1} sont deux conformations testées consécutivement. N est le nombre de cycles effectués et N_u est un paramètre donné par l'utilisateur qui fixe le nombre limite de cycles à effectuer.

Présenté dans la figure 2.8, l'algorithme utilise comme fonction de pénalité ou pseudo-énergie, deux composantes : une composante qui décrit l'accord entre deux résidus adjacents :

$$\mathbf{E}_{MATCH} = \mathbf{-12\ E}_{C\alpha}\mathbf{-12\ E}_{C\beta}\mathbf{-10\ E}_{H\alpha}\mathbf{-15\ E}_{CO}\mathbf{-20\ E}_{H}$$

La seconde composante est, le cas échéant, les scores additionnés issus de l'attribution étendue. Deux stratégies différentes sont utilisées pour obtenir la nouvelle solution x_{i+1} à partir de la première x_i, soit l'échange entre deux résidus choisis au hasard, soit en coupant et en recollant à un autre endroit un grand fragment. L'endroit où on recolle le fragment est choisi de façon aléatoire.
Le seuil de départ (dE) doit être assez élevé pour que presque toutes les possibilités soient acceptées au début du processus afin de s'assurer qu'aucune solution n'est écartée initialement.
Au fil du processus, la pseudo-énergie diminue et le seuil est réduit afin que le nombre de mauvaises solutions acceptées décroisse. L'algorithme converge vers des fragments correctement attribués. Il est à noter que le nombre de cycles N_u après lequel l'algorithme est considéré comme ayant convergé, doit être relativement grand. Si tel n'est pas le cas, on peut rester dans un minimum local et créer des erreurs dans l'attribution.

2.3 L'utilisation des déplacements chimiques

L'arrivée de la spectroscopie RMN tridimensionnelle renforce l'intérêt porté aux déplacements chimiques. Cette grandeur s'avère renfermer les informations de type chimique sur l'acide aminé auquel il est rattaché, mais également des informations issues de la structure secondaire de la protéine.

Cette propriété est utilisée par le programme TATAPRO [Atreya *et al.*, 2000] (Tracked Automated Assignments in PROteins). L'algorithme déterministe utilisé dans TATAPRO se base sur les distributions statistiques des déplacements chimiques répertoriés dans la BMRB.

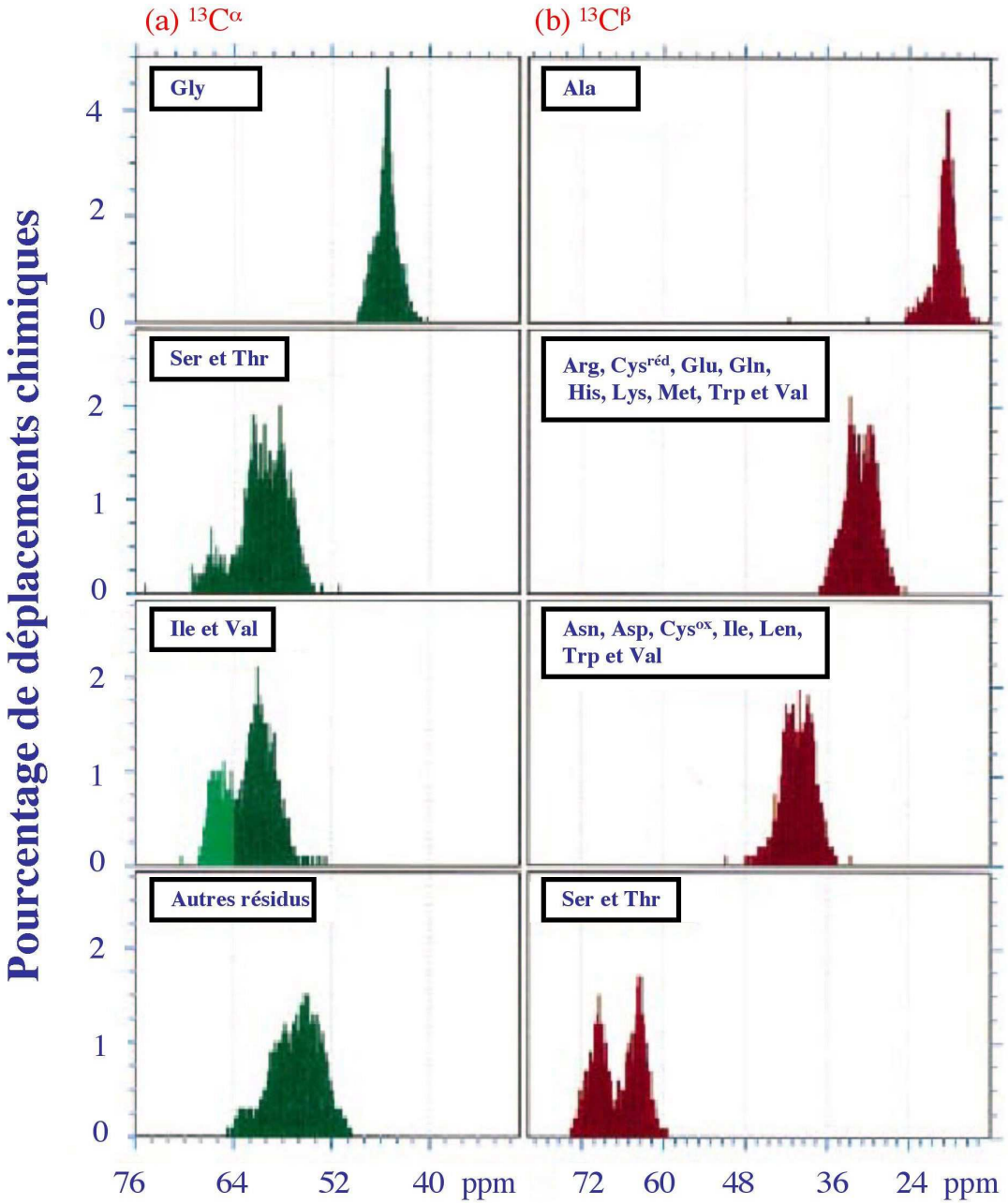

FIG. 2.9 – Distributions des déplacements chimiques des atomes de types $^{13}C^{\alpha}$ et $^{13}C^{\beta}$ dans les différents acides aminés.

Les distributions des déplacements chimiques des atomes de types $^{13}C^{\alpha}$ et $^{13}C^{\beta}$ (Figure : 2.9) montrent que l'on peut définir 6 catégories distinctes de résidus sur la base des déplacements chimiques des carbones, selon le tableau ci-dessous (Tableau : 2.4). L'analyse de la fréquence du $^{13}C^{\alpha}$ permet de discriminer 2 groupes de résidus dans 2 catégories précédemment définis et porte à 8 le nombre total de pseudo-résidus.

Propriétés du $\delta^{13}C^{\beta}$	Propriétés du $\delta^{13}C^{\alpha}$	Code	Acides aminés
-	50 ppm $< {}^{13}C^{\alpha}$	10	Gly
15 ppm $< {}^{13}C^{\beta} <$ 24 ppm		20	Ala
${}^{13}C^{\beta} >$ 58 ppm		30	Ser et Thr
24 ppm $< {}^{13}C^{\beta} <$ 36 ppm	${}^{13}C^{\alpha} <$ 64 ppm	40	Arg, Cys^{red}, Gln, Glu, His, Lys, Met, Val, Trp
	${}^{13}C^{\alpha} >$ 64 ppm	41	Val
36 ppm $< {}^{13}C^{\beta} <$ 50 ppm	${}^{13}C^{\alpha} <$ 64 ppm	50	Asp, Asn, Cys^{ox}, Ile, Leu, Phe, Tyr
	${}^{13}C^{\alpha} >$ 64 ppm	51	Ile
-	-	60	Pro

TAB. 2.4 – Regroupement des acides aminés en fonction des caractéristiques des déplacements chimiques de leur ${}^{13}C^{\alpha}$ et ${}^{13}C^{\beta}$. Dans TATAPRO II [Atreya *et al.*, 2002] les résidus Ser et Thr sont dissociés. On a alors 56 ppm $< {}^{13}C^{\beta} <$ 67 ppm pour Ser et ${}^{13}C^{\beta} >$ 67 ppm pour Thr.

Ce tableau montre que seuls les résidus Ser, Thr, Gly et Ala peuvent être définis de façon unique. Ils servent de marqueurs pour la suite du processus d'attribution spécifique à la séquence. En revanche, les acides aminés Val, Ile et Cys se retrouvent dans 2 catégories différentes.

Une fois les pseudo-résidus reconstitués, la liste les répertoriant est réorganisée dans l'ordre des groupes définis : 10, 20, 30, 40, 41, 50, 51 et 60. La même chose est faite pour la séquence primaire. Nous avions noté que, dans le tableau, certains acides aminés apparaissaient deux fois : Ile est codifié à 51, Val à 41 et les Cys sont considérées comme oxydées, code 50. La séquence primaire est ainsi transformée en table de codes à deux chiffres.

Les fragments sont reconstitués de façon séquentielle "en avançant" puis "en reculant". Le fragment est ensuite placé sur la séquence en cherchant une correspondance à chaque pseudo-résidu. En théorie, cette approche devrait donner les attributions de tout résidu différent de Pro. Dans les faits, les problèmes tels que :

- plus d'une paire ${}^{15}N_i$ et ${}^{1}H^{N}{}_i$ satisfont les conditions
- le fragment se place à différents endroits sur la séquence primaire
- au moins un résidu du fragment a un déplacement chimique atypique

sont gérés. TATAPRO n'accepte les attributions que lorsqu'elles ne sont pas ambiguës.

2.4 Les mesures de la qualité

La plupart des algorithmes d'attribution nécessitent la définition d'une fonction cible. La valeur de cette fonction permet d'estimer la validité de la solution retenue. Dans certains cas, la valeur de la fonction cible correspond à un critère statistique (MAPPER) et permet d'apprécier directement la pertinence d'une solution. Le choix de la fonction cible est primordial dans le développement d'une méthode. Il va conditionner la qualité des résultats obtenus. Si la fonction n'est pas assez spécifique, les bons résultats ne pourront pas être sélectionnés. Elle doit être utilisable, sans trop de modifications, pour la totalité des protéines étudiées.

La fonction d'énergie (Equation : 2.3) utilisée dans ASSTOOL [Reed *et al.*, 2003] est basée sur les correspondances des déplacements chimiques entre résidus adjacents et sur les fourchettes caractéristiques de déplacements chimiques pour chaque type de résidu.

$$E = \sum_{j=2}^{Nres} \sum_{L} \alpha_L f_L \left(\delta_L^j - \delta_L^{j-1}\right) + \sum_{j=1}^{Nres} \sum_{B} \beta_B g_B \left(R_B^j\right) \tag{2.3}$$

N_{res} est le nombre de résidus. La première double somme décrit à quel point les déplacements chimiques des résidus adjacents coïncident ; la seconde double somme calcule à quel point les déplacements chimiques des pseudo-résidus concordent avec ceux de l'acide aminé auxquels ils sont attribués.
La somme sur L couvre les types de noyaux présents dans la liste des pseudo-résidus. ${\delta_X}^j$ est le déplacement chimique du noyau X dans le pseudo-résidu attribué au résidu j. La forme de f_L(x) est choisie par l'utilisateur. Elle est égale à "-1" lorsque les déplacements chimiques sont proches et à "0" sinon. α_L est le terme de pondération du noyau de type L.
La somme sur B est faite sur les atomes ${}^{13}C^{\alpha}{}_i$, ${}^{13}C^{\beta}{}_i$, ${}^{13}CO_i$, ${}^{13}C^{\alpha}{}_{i-1}$, ${}^{13}C^{\beta}{}_{i-1}$ et ${}^{13}CO_{i-1}$. ${R^j}_B$ est égal à zéro si le déplacement chimique du pseudo-résidu B est proche du déplacement chimique du résidu de la séquence. Une fois passée une certaine limite, il prend la valeur de la différence par rapport à la limite. β_B est le terme de pondération du déplacement chimique de type B. La composante négative du score provient uniquement de la cohérence entre les déplacements chimiques séquentiels. La composante positive provient de la cohérence entre le pseudo-résidu et le résidu de la séquence. L'accord avec la séquence n'est pris en compte que si l'accord séquentiel est bon.
Cette fonction, basée sur la somme de différences entre valeurs de déplacements chimiques permet une évalution globale du positionnement. Par contre, elle ne permet pas d'évaluation au niveau local. Elle entraîne donc un risque important d'erreurs.

Dans son approche déterministe, IBIS utilise deux fonctions cibles.
Les distributions des déplacements chimiques sont approximées par des gaussiennes, elles sont par la suite utilisées pour calculer la probabilité $\rho\,(aa|l)$ qu'une branche (définie comme la liste des déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C^{\gamma}$, ^{13}CO) l appartienne à un acide aminé aa (Equation : 2.4).

$$\rho\,(aa|l) = \frac{0,15 \times \left(1 - \frac{|n_{exp}^{aa} - n_{obs}|}{n_{exp}^{aa}}\right) + \sum_{c} w_{atome} \times G\,(\delta_c^{aa}{}_{exp} - \delta_c^{aa}{}_{obs}, \sigma_c^{aa})}{0,15 + \sum_{c} w_{atome}}$$

$$G\,(\delta_c^{aa}{}_{exp} - \delta_c^{aa}{}_{obs}, \sigma_c^{aa}) = e^{-\left(\frac{\delta_c^{aa}{}_{exp} - \delta_c^{aa}{}_{obs}}{\sigma_c^{aa}}\right)^2} \tag{2.4}$$

Les $\delta_c^{aa}{}_{exp}$ et $\sigma_c^{aa}{}_{exp}$ sont les valeurs moyennes et les écart-types, calculés à partir du contenu de la BMRB, pour le déplacement chimique c de l'acide aminé aa. n_{obs} est le nombre de déplacements chimiques de la chaîne latérale observés dans une fourche. n_{exp}^{aa} est le nombre de déplacements chimiques de la chaîne latérale attendus dans l'acide aminé aa. L'accord entre les valeurs n_{obs} et n_{exp}^{aa} est nuancé par le facteur empirique de 0,15. Les nombres w_{atome} pondérent de façon différente chaque type d'atome. La fonction G est une fonction gaussienne classique.

La seconde fonction cible quantifie la cohérence entre un fragment et les acides aminés de la séquence (Equation : 2.5).

$$\Pi\,(i,j) = \frac{ln(n+1)}{n} \sum_{k=0}^{n-1} \frac{\rho\,(aa[position_{j+k-1}]|S) + \rho\,(aa[position_{j+k}]|I)}{2} \tag{2.5}$$

La $position_j$ est le point de départ du placement du fragment i. S et I sont les branches regroupant les déplacements chimiques séquentiels et intra-résiduels. n est le nombre de fourches dans le fragment traité. La fonction Π est construite comme la moyenne des probabilités individuelles pondérée par le terme comprenant le logarithme népérien. Ce terme diminue la moyenne. Cette approche implique le fait que l'approximation de la distribution des déplacements chimiques par une gaussienne soit toujours valable.
Ces deux fonctions sont visiblement le fruit d'une démarche du type essais/erreurs. Elle risque donc de ne pas être adéquate pour toutes les protéines cibles (taille, qualité des spectres...).

Le programme MAPPER utilise également le fait que la distribution des déplacements

chimiques puisse être approximée par une gaussienne. Après avoir fait une attribution globale, MAPPER détermine l'attribution individuelle de la façon suivante (Equation : 2.6) :

$$\chi^2(i,k) = \sum_{j=0}^{n(i)} \sum_{a \in Aj(i)} \left[\frac{\delta^a_{obsj}(i) - \delta^a_{refR(k+j)}}{\sigma^a_{refR(k+j)}} \right]^2 \tag{2.6}$$

n(i) est le nombre de résidus dans le fragment i, j est le compteur interne au fragment, $A_j(i)$ est l'ensemble des atomes du fragment i, à la position j pour lesquels les déplacements chimiques sont disponibles. δ^a_{obsj}(i) est le déplacement chimique expérimental de l'atome a du j^{ème} résidu du fragment i. $\delta^a_{refR(k+j)}$ et $\sigma\ ^a_{refR(k+j)}$ sont respectivement la valeur moyenne et l'écart-type de l'atome a du $j^{ème}$ résidu du fragment i à la position k de la séquence primaire. Ce score est comparé à la valeur critique du χ^2 ayant comme nombre de degrés de liberté le nombre de déplacements chimiques connus dans le fragment. Le taux de confiance utilisé est de 1%.

Lorsque la différence entre les valeurs de déplacements chimiques est inférieure à la valeur de l'écart-type, le terme élevé au carré est inférieur à un.

Afin de placer les fragments construits sur la séquence, MARS calcule pour chaque pseudo-résidu observé la distance entre les déplacements chimiques expérimentaux et des valeurs prédites. Les valeurs prédites sont construites comme la somme normalisée de la valeur Random-Coil et de la valeur attendue si le résidu est impliqué dans une hélice-α ou un brin-β. Cette somme est pondérée par la probabilité que le résidu soit dans une structure secondaire (issue de PSIPRED [McGuffin *et al.*, 2000]). La forme de fonction cible (Equation : 2.7) est la même que celle utilisée par MAPPER :

$$D\,(i,k) = \sum_{k=1}^{N_{CS}} \left\{ \frac{\delta(i)^{exp}_k - \delta(j)_k}{\sigma_k} \right\}^2 \tag{2.7}$$

$\delta(i)^{exp}_k$ est le déplacement chimique du noyau k du pseudo-résidu i, $\delta(j)_k$ est le déplacement chimique prédit pour le noyau k du résidu j. N_{CS} est le nombre de types de déplacements chimiques disponibles et σ_k est l'écart-type de la distribution statistique des déplacements chimiques utilisée pour calculer $\delta(j)_k$. Si le type de noyau k n'est pas disponible, $\delta(i)^{exp}_k - \delta(j)_k$ est mis à zéro. Contrairement au programme MAPPER , MARS considère chaque pseudo-résidu de façon isolée. Ils sont triés par la suite en fonction du score obtenu.

Programme faisant partie de CAMRA, PROCESS réalise la comparaison entre des déplacements chimiques prédits par le programme ORB et des déplacements chimiques expérimentaux (issus du programme CAPTURE). Il calcule un score individuel pour chaque pic observé contre chaque pic attendu. Le corps de la fonction cible est le même que celui de MARS. La différence est que, MARS garde le score sous cette forme alors que PROCESS utilise l'exponentiel de l'opposé du score(Equation : 2.8).

$$\exp -\left(\left(\frac{obs(x)-pred(x)}{\sigma(pred(x))}\right)^2 + \left(\frac{obs(y)-pred(y)}{\sigma(pred(y))}\right)^2 + \left(\frac{obs(z)-pred(z)}{\sigma(pred(z))}\right)^2\right) \tag{2.8}$$

obs(x/y/z) et *pred(x/y/z)* sont les valeurs de déplacements chimiques, en ppm, d'un pic observé ou prédit dans les dimensions x, y et z. σ est l'écart-type calculé par ORB, il sert de mesure à la précision attendue pour les prédictions. Cette approche exhaustive permet par la suite de sélectionner la meilleure possibilité. Lorsque la concordance est excellente pour tous les noyaux disponibles, le score est égal à "1", plus elle se dégrade, plus le score diminue.

Les fonctions cibles sont construites en calculant l'accord avec le pseudo-résidu précédent, en calculant l'accord avec la séquence, et/ou en calculant l'accord entre l'expérience et des valeurs de référence. Ces accords sont souvent calculés en utilisant les différences élevées au carré. Deux différences de signes opposés ne se compensent donc pas. Le type le plus répandu de fonction cible est celui qui est lié au test du χ^2 : test qui consiste à élever au carré la comparaison de la valeur expérimentale à la moyenne d'une distribution et de diviser le résultat par l'écart-type. Les trois dernières fonctions sont très proches, et très semblables à celle utilisée dans le programme QUASI. Elles présentent l'avantage de donner plus de poids à la cohérence observée au niveau de l'acide aminé.

3

Les approches alternatives

Malgré cette multiplication de méthode, aucune n'offre vraiment une approche universelle.

Dans le sillage de Malliavin et al. [Malliavin *et al.*, 1992], plusieurs approches ont pour but de contourner cette étape d'attribution en couplant l'étude des déplacements chimiques aux calculs de structure.

La méthode est basée sur l'utilisation de l'effet Overhauser nucléaire. L'acquisition de nombreux spectres NOESY (jusqu'à 50) permet de déterminer de façon précise les distances entre les protons amides du squelette de la protéine cible. Des mesures très nombreuses de distances permettent d'être d'autant plus précis sur la structure de la protéine cible. L'étude de l'évolution des pics de corrélation, observés dans une expérience NOESY non attribuée, en fonction du temps de mélange permet l'extraction des distances interatomiques avec une précision de l'ordre de $\pm$ 0,5% pour les distances jusqu'à 5 ou 6 Å. Les paramètres de relaxation et donc les distances entre protons sont dégagés d'une analyse de ces évolutions. Cette technique est appliquée [Reisdorf *et al.*, 1992] sur un lysozyme. 56 expériences NOESY ont été enregistrées et 81 courbes NOE ont été obtenues et analysées en terme de distances. Le résultat est un nuage de protons amide non-attribués. La deuxième étape consiste à modéliser la position de la chaîne principale du peptide dans le nuage de protons amides ainsi déterminé. A l'issue de cette étape, on dispose de la structure et, de façon implicite, de l'attribution des protons amides.
Cette méthode a été déclinée et explorée par plusieurs autres équipes. Elles diffèrent essentiellement sur l'algorithme choisi pour placer et analyser le nuage de protons.

L'approche DG (Distance Geometry) proposée par Oshiro et Kuntz [Oshiro Kuntz, 1993], utilise les NOESY non-attribuées pour calculer les positions tridimensionnelles de protons en appliquant les méthodes classiques de géométrie de

distance. Les relations spatiales entre ces protons peuvent être utilisées pour en déduire les attributions. Les sous-structures de domaines correspondant à des fragments isolés de structures secondaires, sont identifiées par l'analyse de graphes théoriques. A l'intérieur de chaque domaine, une sous-structure moyenne est tirée des structures données par DG et les acides aminés de la séquence sont placés sur la sous-structure. Si la géométrie des $^{13}C^{\alpha}$ se compare bien aux structures en hélices-α ou en brins-β, alors l'attribution est conservée. Après avoir testé la méthode sur des données réelles et simulées, les auteurs concluent que de bons résultats nécessitent des données NOE de haute qualité.

De son côté , l'approche *ab-initio* ANSRS se base sur le fait que toutes les informations nécessaires pour l'attribution séquentielle, et la détermination de structure, peuvent, en principe, être obtenues à partir d'un spectre NOESY multi-dimensionnel. Une structure 3D de protons non-attribués est calculée par recuit simulé sans tenir compte de structures covalentes [Kraulis, 1994]. Des attributions possibles pour chaque proton dérivent de surfaces de probabilité précompilées pour chaque atome de chaque type d'acide aminé.

Afin de valider l'utilisation de l'information structurale des spectres non-attribués pour déterminer directement la structure, Atkinson et Saudek [Atkinson Saudek, 1996], proposent une autre méthode. Ils utilisent chaque pic observé dans un spectre NOESY comme une distance entre deux protons. Entre ces protons isolés et non attribués, seules les contraintes de distances sont connues. Le problème suivant est le placement des distances dans l'espace, il est résolu par le programme XPLOR [Badger *et al.*, 1999]. L'utilisation éventuelle de spectres TOCSY et COSY permet d'affiner ce placement car ils permettent l'identification des atomes contenus dans un même acide aminé. Cette méthode permet d'obtenir un point de départ pour l'affinement de la structure.

CLOUDS [Grishaev Llinas, 2002] calcule les distributions spatiales des protons des protéines à partir des données NOESY et d'une attribution minimale. Ce programme repose sur d'abondantes contraintes de distances interprotons calculées. Ces calculs se font via une analyse de la matrice de relaxation des données NOESY. Un gaz de protons isolés et non attribués est soumis à un champ de force. La pseudo-énergie est minimisée par dynamique moléculaire, cela crée un nuage, des structures moléculaires composées uniquement de protons exempts de liaisons covalentes. Une famille de nuages, sélectionnée par FILTER (Foc Identification via LowesT ErroR), donne une densité en 3D de protons dans l'espace réel.

Pour toutes ces méthodes, le problème d'ambiguïté reste un frein majeur. Afin de contourner ce problème, une autre approche proposée par Atkinson et Saudek en 2002 [Atkinson Saudek, 2002] interprète les pics de corrélation observés dans les spectres, comme des distances entre atomes. Le spectre 2D ^{1}H-^{15}N HSQC donne une distance de 1,02 Å entre les azotes et les protons (par ex : $N_{118,00}$-$H_{8,30}$), le spectre HN(CO)CA donne des distances à 2,40 Å entre l'atome d'azote et l'atome $^{13}C^{\alpha}$ (par ex : $^{15}N_{118,00}$-$^{13}C^{\alpha}_{55,40}$)et 2,49 Å entre le proton et l'atome $^{13}C^{\alpha}$ (par ex : $^{1}H_{8,30}$-$^{13}C^{\alpha}_{55,40}$). Des traitements similaires sont appliqués aux 5 autres spectres triple-résonances classiquement utilisés pour l'attribution manuelle ainsi qu'aux spectres HNHA, HNHB, 2D ^{1}H-^{13}C HSQC. Les spectres NOESY sont également interprétés en terme de distances entre des atomes non connectés et non attribués. Cette approche a été validée sur un jeu de données artificielles de l'ubiquitine. En utilisant les données expérimentales, les problèmes de recouvrement et de signaux manquants rendent la méthode plus difficile à utiliser. Ils restent donc les principales difficultés non encore gérées par ces approches alternatives.

4

Conclusion

Les programmes d'attribution complètement automatiques ont, par rapport à des programmes d'aide à l'attribution, la difficulté supplémentaire de devoir remplacer l'oeil de l'utilisateur par des routines informatiques (MARS, AUTOASSIGN). Cette tâche est certainement aussi difficile à gérer que le problème de l'attribution. Ces programmes présentent donc des démarches beaucoup plus "lourdes" que leurs homologues qui utilisent l'utilisateur comme faisant partie du processus d'attribution.
Chaque programme d'aide à l'attribution se singularise par sa méthode, sa fonction cible... Certains d'entre eux sortent des approches traditionnelles : PLATON avec son système de CSP (Chemical Shift Pattern) qui lui permet de relier chaque pseudo-résidu à un acide aminé avec sa structure secondaire, le programme RESCUE par l'utilisation d'un réseau de neurones.
En parallèle, des approches alternatives ne nécessitant pas d'étape d'attribution, sont mises au point depuis une dizaine d'années. Elles se heurtent toujours aux problèmes liés à l'utilisation de spectres expérimentaux.
Le programme QUASI, qui fait l'objet de la prochaine partie, s'inscrit dans cette démarche de combiner à la fois l'étude du déplacement chimique et la détermination de la structure.

Bibliographie

[Atkinson Saudek, 1996] Atkinson, R. A. Saudek, V. (1996). *Dynamics and the problem of recognition in biological macromolecules* chapter The direct determination of protein structure from multidimensional NMR spectra without assignment : an evaluation of the concept. New York London : Plenum Press.

[Atkinson Saudek, 2002] Atkinson, R. A. Saudek, V. (2002). *FEBS Lett*, **510** (1-2), 1–4. *The direct determination of protein structure by NMR without assignment.*

[Atreya *et al.*, 2002] Atreya, H. S., Chary, K. V., Govil, G. (2002). *Curr Sci*, **83** (11), 1372–1376. *Automated NMR assignments of proteins for high thrughput structure determination : TATAPRO II.*

[Atreya *et al.*, 2000] Atreya, H. S., Sahu, S. C., Chary, K. V., Govil, G. (2000). *J Biomol NMR*, **17** (2), 125–136. *A tracked approach for automated NMR assignments in proteins (TATAPRO).*

[Badger *et al.*, 1999] Badger, J., Kumar, R. A., Yip, P., Szalma, S. (1999). *Proteins*, **35** (1), 25–33. *New features and enhancements in the X-PLOR computer program.*

[Bartels *et al.*, 1996] Bartels, C., Billeter, M., Güntert, P., Wüthrich, K. (1996). *J Biomol NMR*, **7**, 207–213. *Automated sequence-specific NMR assignment of homologous proteins using the program GARANT.*

[Bartels *et al.*, 1995] Bartels, C., Xia, T.-H., Güntert, P., Wüthrich, K. (1995). *J Biomol NMR*, **5**, 1–10. *The program XEASY for computer-supported NMR spectral analysis of biological macromolecules.*

[Berman *et al.*, 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000). *Nucleic Acids Res*, **28** (1), 235–242. *The Protein Data Bank.*

[Bernstein *et al.*, 1993] Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J., Holak, T. A. (1993). *J Biomol NMR*, **3** (2), 245–251. *Computer-assisted assignment of multidimensional NMR spectra of proteins : Application to 3D NOESY-HMQC and TOCSY-HMQC spectra.*

[Billeter *et al.*, 1982] Billeter, M., Braun, W., Wüthrich, K. (1982). *J Mol Biol*, **155** (3), 321–346. *Sequential resonance assignments in protein 1H nuclear magnetic resonance spectra. Computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations.*

[Bodenhausen Ruben, 1980] Bodenhausen, G. Ruben, D. (1980). *Chem Phys Lett*, **69** (1), 185–189. *Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy.*

[Boucher, 2002] Boucher, W. (2002). *AZARA program.*

[Coggins Zhou, 2003] Coggins, B. E. Zhou, P. (2003). *J Biomol NMR*, **26** (2), 93–111. *PACES : Protein sequential assignment by computer-assisted exhaustive search.*

[Cornilescu *et al.*, 1999] Cornilescu, G., Delaglio, F., Bax, A. (1999). *J Biomol NMR*, **13** (3), 289–302. *Protein backbone angle restraints from searching a database for chemical shift and sequence homology.*

[Dueck Scheuer, 1990] Dueck, G. Scheuer, T. (1990). *J Comput Phys*, **90**, 161–175. *Threshold accepting : a general purpose optimization algorithm appearing superior to simulated annealing.*

[Güntert *et al.*, 2000] Güntert, P., Salzmann, M., Braun, D., Wüthrich, K. (2000). *J Biomol NMR*, **18** (2), 129–137. *Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER.*

[Goddard Kneller, 2004] Goddard, T. D. Kneller, D. G. (2004). *Programme SPARKY.*

[Grishaev Llinas, 2002] Grishaev, A. Llinas, M. (2002). *Proc Natl Acad Sci U S A*, **99** (10), 6707–6712. *CLOUDS, a protocol for deriving a molecular proton density via NMR.*

[Gronwald *et al.*, 1997] Gronwald, W., Boyko, R., Sonnichsen, F., Wishart, D., Sykes, B. D. (1997). *J Biomol NMR*, **10** (2), 165–179. *ORB, a homology-based program for the prediction of protein NMR chemical shifts.*

[Gronwald *et al.*, 1998] Gronwald, W., Willard, L., Jellard, T., Boyko, R. F., Rajarathnam, K., Wishart, D. S., Sonnichsen, F. D., Sykes, B. D. (1998). *J Biomol NMR*, **12** (3), 395–405. *CAMRA : chemical shift based computer aided protein NMR assignments.*

[Helgstrand *et al.*, 2000] Helgstrand, M., Kraulis, P., Allard, P., Hard, T. (2000). *J Biomol NMR*, **18** (4), 329–336. *Ansig for Windows : an interactive computer program for semiautomatic assignment of protein NMR spectra.*

[Hyberts Wagner, 2003] Hyberts, S. G. Wagner, G. (2003). *J Biomol NMR*, **26** (4), 335–344. *IBIS–a tool for automated sequential assignment of protein spectra from triple resonance experiments.*

[Ikura *et al.*, 1990] Ikura, M., Kay, L. E., Bax, A. (1990). *Biochemistry*, **29** (19), 4659–4667. *A novel approach for sequential assignment of 1H, 13C, and 15N spectra of proteins : heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin.*

[Jahnke Widmer, 2004] Jahnke, W. Widmer, H. (2004). *Cell Mol Life Sci*, **61** (5), 580–599. *Protein NMR in biomedical research.*

[Jung Zweckstetter, 2004] Jung, Y.-S. Zweckstetter, M. (2004). *J Biomol NMR*, **30** (1), 11–23. *Mars – robust automatic backbone assignment of proteins.*

[Kendrew *et al.*, 1958] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., Phillips, D. C. (1958). *Nature*, **181** (4610), 662–666. *A three-dimensional model of the myoglobin molecule obtained by X-ray analysis.*

[Kirkpatrick *et al.*, 1983] Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. (1983). *Science*, **220**, 671–680. *Optimization by Simulated Annealing.*

[Koradi *et al.*, 1998] Koradi, R., Billeter, M., Engeli, M., Güntert, P., Wüthrich, K. (1998). *J Magn Reson*, **135** (2), 288–297. *Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY.*

[Kraulis, 1989] Kraulis, P. J. (1989). *J Magn Reson*, **84**, 627–633. *ANSIG : A Program for the Assignment of Protein 1H 2D NMR spectra by Interactive Graphics.*

[Kraulis, 1994] Kraulis, P. J. (1994). *J Mol Biol*, **243** (4), 696–718. *Protein three-dimensional structure determination and sequence-specific assignment of 13C and 15N-separated NOE data. A novel real-space ab initio approach.*

[Labudde *et al.*, 2003] Labudde, D., Leitner, D., Kruger, M., Oschkinat, H. (2003). *J Biomol NMR*, **25** (1), 41–53. *Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts.*

[Leutner *et al.*, 1998] Leutner, M., Gschwind, R. M., Liermann, J., Schwarz, C., Gemmecker, G., Kessler, H. (1998). *J Biomol NMR*, **11** (1), 31–43. *Automated backbone assignment of labeled proteins using the threshold accepting algorithm.*

[Malliavin *et al.*, 1992] Malliavin, T. E., Rouh, A., Delsuc, M. A., Lallemand, J. Y. (1992). *CRAS série II*, **315**, 653–659. *Approche directe de la détermination de structures moléculaires à partir de l'effet Overhauser nucléaire.*

[Malmodin *et al.*, 2003] Malmodin, D., Papavoine, C. H. M., Billeter, M. (2003). *J Biomol NMR*, **27** (1), 69–79. *Fully automated sequence-specific resonance assignments of heteronuclear protein spectra.*

[Marion *et al.*, 1989] Marion, D., Driscoll, P. C., Kay, L. E., Wingfield, P. T., Bax, A., Gronenborn, A. M., Clore, G. M. (1989). *Biochemistry*, **28** (15), 6150–6156. *Overcoming the overlap problem in the assignment of 1H NMR spectra of larger proteins by use of three-dimensional heteronuclear 1H-15N Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy : application to interleukin 1 beta.*

[McGuffin *et al.*, 2000] McGuffin, L. J., Bryson, K., Jones, D. T. (2000). *Bioinformatics*, **16** (4), 404–405. *The PSIPRED protein structure prediction server.*

[Murzin *et al.*, 1995] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C. (1995). *J Mol Biol*, **247** (4), 536–540. *SCOP : a structural classification of proteins database for the investigation of sequences and structures.*

[Oshiro Kuntz, 1993] Oshiro, C. M. Kuntz, I. D. (1993). *Biopolymers*, **33** (1), 107–115. *Application of distance geometry to the proton assignment problem.*

[Pons Delsuc, 1999] Pons, J. L. Delsuc, M. A. (1999). *J Biomol NMR*, **15** (1), 15–26. *RESCUE : an artificial neural network tool for the NMR spectral assignment of proteins.*

[Reed *et al.*, 2003] Reed, M. A. C., Hounslow, A. M., Sze, K. H., Barsukov, I. G., Hosszu, L. L. P., Clarke, A. R., Craven, C. J., Waltho, J. P. (2003). *J Mol Biol*, **330** (5), 1189–1201. *Effects of domain dissection on the folding and stability of the 43 kDa protein PGK probed by NMR.*

[Reisdorf *et al.*, 1992] Reisdorf, C., Malliavin, T. E., Delsuc, M. A. (1992). *Biochimie*, **74** (9-10), 809–813. *Accurate estimation of inter-atomic distances in large proteins by NMR.*

[Rosenblatt, 1957] Rosenblatt, F. (1957). *Technical Report 85-460-1*, . *The perceptron : A perceiving and recognizing automaton (project PARA).*

[Schechner *et al.*, 2004] Schechner, M., Sirockin, F., Stote, R. H., Dejaegere, A. P. (2004). *J Med Chem*, **47** (18), 4373–4390. *Functionality maps of the ATP binding site of DNA gyrase B : generation of a consensus model of ligand binding.*

[Wagner *et al.*, 1987] Wagner, G., Braun, W., Havel, T. F., Schaumann, T., Go, N., Wüthrich, K. (1987). *J Mol Biol*, **196** (3), 611–639. *Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN.*

[Wüthrich, 1986] Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. New-York : John Wiley and Sons.

[Xu *et al.*, 1993] Xu, J., Straus, S. K., Sanctuary, B. C., Trimble, L. (1993). *J Chem Inf Comput Sci* **33** (5), 668–682. *Automation of protein 2D proton NMR assignment by means of fuzzy mathematics and graph theory.*

[Zadeh, 1988] Zadeh, L. (1988). *IEEE Computer* **21** (4), 83–92. *Fuzzy Logic.*

[Zimmerman *et al.*, 1997] Zimmerman, D. E., Kulikowski, C. A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R., Montelione, G. T. (1997). *J Mol Biol* **269** (4), 592–610. *Automated analysis of protein NMR assignments using methods from artificial intelligence.*

[Zimmerman *et al.*, 1993] Zimmerman, D. E., Kulikowski, C. A., Montelione, G. T. (1993). *Proc Int Conf Intell Syst Mol Biol* **1**, 447–455. *A constraint reasoning system for automating sequence-specific resonance assignments from multidimensional protein NMR spectra.*

Partie II

QUASI : QUick Access to Spectral Interpretation

1

Introduction

Dans la première partie, les logiciels qui existent à l'heure actuelle pour l'attribution séquentielle des protéines par RMN ont été présentés. Certains d'entre eux existaient déjà au début du projet mené lors de cette thèse, les autres sont apparus pendant ces trois dernières années. La plupart des approches antérieures à 2002 visaient l'automatisation complète de l'attribution tel que AUTOASSIGN. QUASI s'inscrit dans une autre démarche : plus que l'automatisation, il cherche à assister l'attribution en réalisant rapidement les étapes sans ambiguïtés, en demandant l'intervention de l'utilisateur lorsque l'interprétation est ambiguë puis en évaluant le placement des pseudo-résidus sur la séquence primaire suivant plusieurs approches. QUASI est utilisable sur des protéines, au minimum, doublement marquées (^{13}C et ^{15}N). En effet, il utilise les expériences tridimensionnelles hétéronucléaires scalaires RMN présentées dans la figure 1.1.

Chacun d'entre eux apporte une information utilisée lors de la formation des pseudo-résidus. Le **pseudo-résidu** est, dans ce manuscrit, l'**ensemble des résonances liées à une corrélation NH** : $\{\mathrm{N}_i^H,\ \mathrm{H}_i,\ ^{13}\mathrm{C}_i^\alpha,\ ^{13}\mathrm{C}_i^\beta,\ ^{13}\mathrm{CO}_i,\ ^{13}\mathrm{C}_{i-1}^\alpha,\ ^{13}\mathrm{C}_{i-1}^\beta,\ ^{13}\mathrm{CO}_{i-1}\}$. Les relations existant entre ces résonances sont **intra-résiduelles** : $^{13}\mathrm{C}_i^\alpha$, $^{13}\mathrm{C}_i^\beta$ et $^{13}\mathrm{CO}_i$ et **séquentielles** : $^{13}\mathrm{C}_{i-1}^\alpha$, $^{13}\mathrm{C}_{i-1}^\beta$ et $^{13}\mathrm{CO}_{i-1}$. En enchaînant ces pseudo-résidus, on définit des **fragments**.

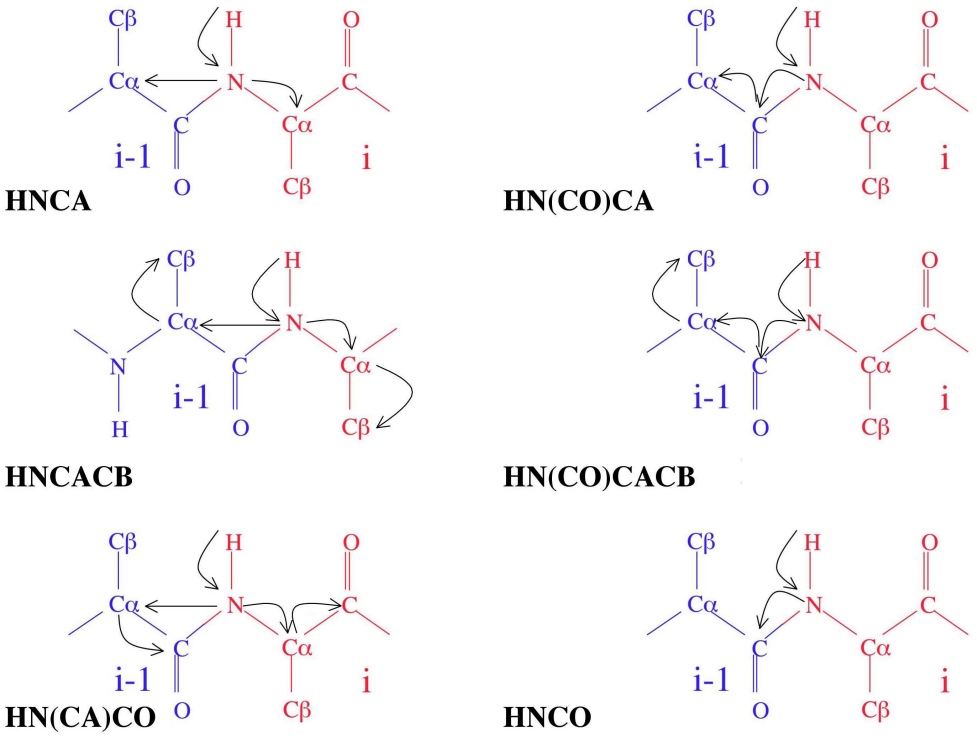

FIG. 1.1 – Expériences tridimensionnelles hétéronucléaires RMN utilisées pour obtenir les informations sur les couplages scalaires.

QUASI s'articule autour de deux sous-routines (Figure : 1.2) : la première (QUASI-1) réalise l'attribution des fréquences aux noyaux de résidus (non identifiés) et les connecte. Elle constitue des fragments. La seconde (QUASI-2) attribue les fragments à la séquence primaire. Ce placement se fait en fonction de méthodes que l'utilisateur choisit.

L'utilisation de QUASI et le comportement de chacune de ses deux sous-routines seront détaillées et illustrées par le cas de l'ubiquitine, protéine de 76 résidus qui sert souvent de plate-forme d'expérimentation lors de la mise au point des programmes informatiques pour la RMN.

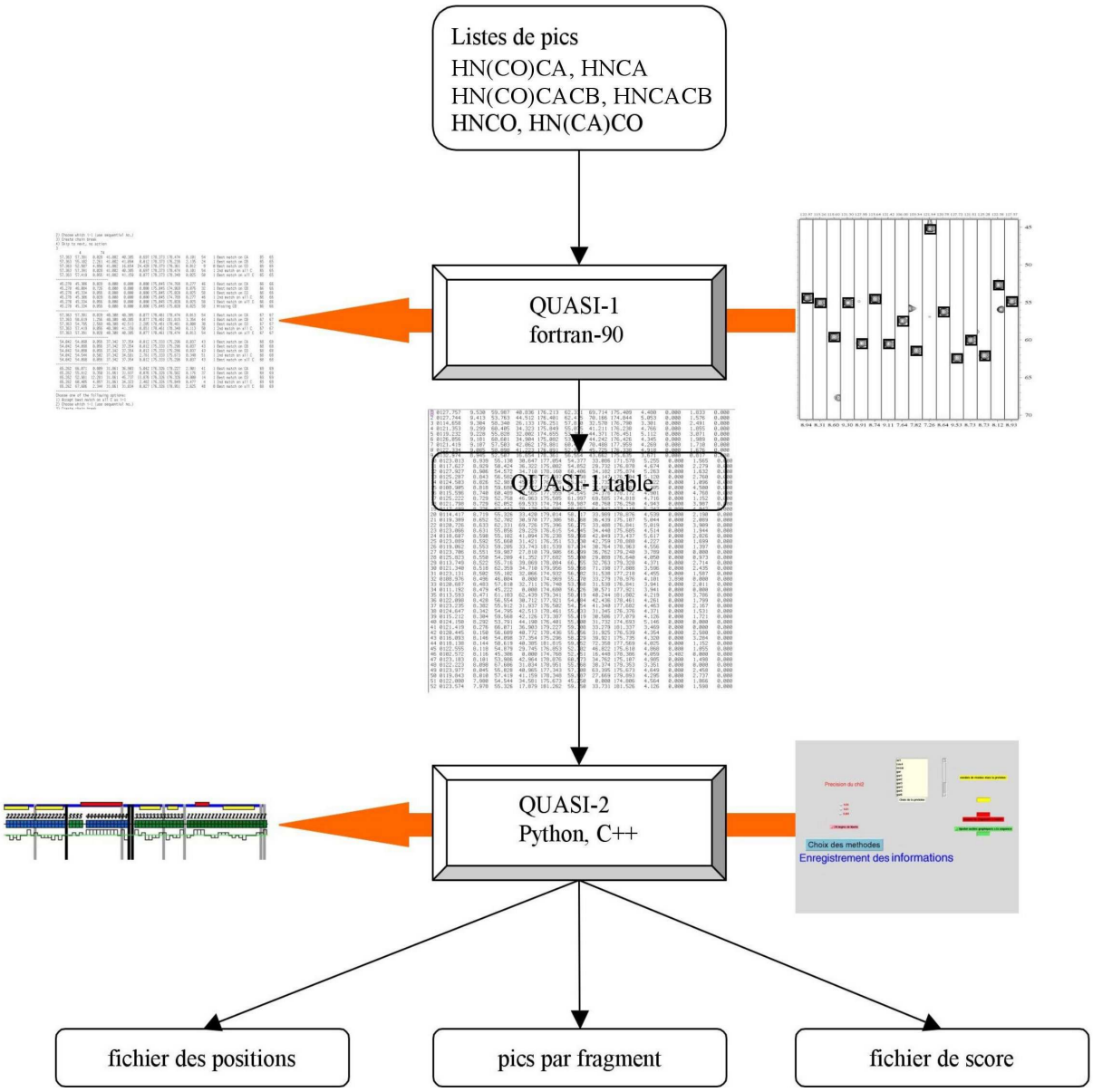

FIG. 1.2 – Diagramme schématique du programme QUASI. QUASI-1 développé en fortran 90 réalise les connexions (automatiques ou manuelles) des pseudo-résidus. QUASI-2 place les fragments sur la séquence primaire. Outre l'interface graphique qui permet une première appréciation des résultats obtenus, les données de sortie sont les fichiers récapitulant les pics de chaque fragment, les fichiers de score et les attributions faites par les méthodes choisies.

2

QUASI-1

Sommaire

2.1 Données d'entrée de QUASI-1

Le jeu d'expériences RMN, optimum, pour utiliser QUASI-1 est celui regroupant les 6 expériences triple-résonances communément utilisées : HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HNCO et HN(CA)CO. Les informations liées aux prolines ($^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}CO) n'apparaissent qu'en tant qu'atomes *i-1* associés au NH qui suit.
La sélection des pics est réalisée manuellement de façon à assurer le contrôle de l'utilisateur dès le départ du processus d'attribution. La liste des corrélations observées dana le spectre 2D ^{1}H-^{15}N HSQC est affinée en examinant le spectre HN(CO)CA qui permet de séparer les corrélations qui se recouvrent éventuellement dans l'HSQC et d'éliminer les pics n'ayant pas de pic C^{α}_{i-1} (tels que les NH_2).

L'utilisateur prépare un fichier par spectre acquis, les informations sur les déplacements chimiques δ_H, δ_N et δ_C y étant répertoriées. Les fichiers sont croisés afin d'obtenir une liste unique par type d'atome (HN(CO)CA $\rightarrow \delta_{C\alpha i-1}$, HNCA $\rightarrow \delta_{C\alpha i}$,

HN(CO)CACB $\rightarrow \delta_{C\beta i-1}$, HNCACB $\rightarrow \delta_{C\beta i}$, HNCO $\rightarrow \delta_{COi-1}$, HN(CA)CO $\rightarrow \delta_{COi}$. Afin de permettre le fonctionnement de QUASI, les déplacements chimiques des atomes N_i et H_i doivent être identiques d'un fichier à l'autre. Cela lui permet, par lecture des fichiers de pics, de former simplement les pseudo-résidus.

Une fois cette étape accomplie, QUASI-1 examine les données associées avec chaque NH pour enchaîner les pseudo-résidus et former ainsi des fragments.

2.2 Constitution des fragments

2.2.1 Méthode

Pour chaque pseudo-résidu, QUASI-1 prend les déplacements chimiques δ_{Ci-1} pour chaque type de carbone α, β, ^{13}CO et les compare avec les δ_{Ci} de tous les autres pseudo-résidus en en calculant les différences $\Delta\delta_C$. Une tolérance est fixée par type de noyau (Tableau : 2.1). Chacune d'elles correspond à l'estimation de l'erreur faite lors de la sélection des pics dans les différents spectres, elle est liée à la largeur spectrale de chaque spectre.

Noyaux	Marge en ppm
$^{13}C^{\alpha}$	0,14
$^{13}C^{\beta}$	0,2
^{13}CO	0,14

TAB. 2.1 – Tableau présentant les marges utilisées par QUASI lors de la connexion des pseudo-résidus.

Une fois toutes les comparaisons réalisées, QUASI-1 sélectionne et présente, par pseudo-résidu, plusieurs informations :

- la meilleure correspondance obtenue pour chaque type de noyau ($^{13}C^{\alpha}$, $^{13}C^{\beta}$, ^{13}CO). (**BEST match on...**)

- la seconde meilleure correspondance sur tous les noyaux (calculée comme la somme des différences des déplacements chimiques : δ_C). (**2nd match on all**)

- la meilleure correspondance sur tous les noyaux. (**BEST match on all**)

- les correspondances qui se trouvent dans les tolérances choisies mais dont certaines données sont manquantes (principalement le C_{β}). (**Missing...**)

La présentation se fait sous la forme d'un tableau de 14 colonnes (Figure : 2.1). La première présente le déplacement chimique du noyau $^{13}C^{\alpha}_{i-1}$; la seconde est le déplacement

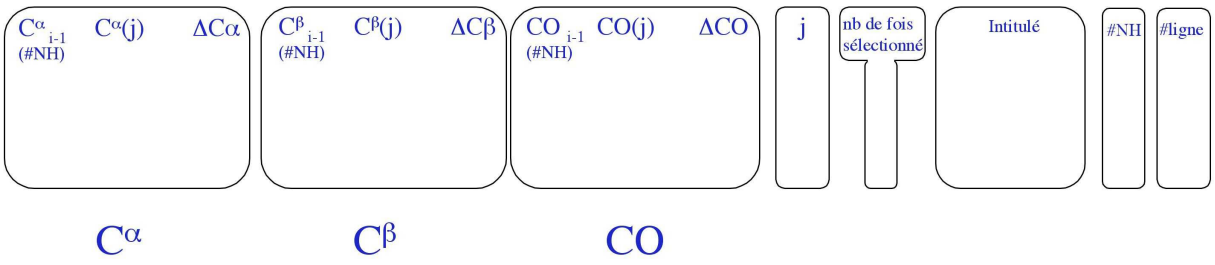

FIG. 2.1 – Détails des 14 colonnes constituant l'interface de QUASI-1.

chimique du $^{13}C^{\alpha}$ du pseudo-résidu candidat ; la troisième colonne est la différence de déplacement chimique calculée entre la première et la deuxième colonne. Les colonnes 4 à 6 regroupent les mêmes informations pour le noyau $^{13}C^{\beta}$ et celles de 7 à 9 traitent les ^{13}CO. La dixième colonne indique le numéro (j) du pseudo-résidu candidat. La $11^{ème}$ indique le nombre de fois que le pseudo-résidu candidat est sélectionné de façon automatique. Un nombre supérieur à '1' dans cette colonne indique que le pseduo-résidu candidat a été sélectionné automatiquement comme précédent plus d'une fois, l'utilisateur doit alors choisir entre les différentes possibiliés. La colonne suivante est l'intitulé de l'information disponible dans la ligne. La $13^{ème}$ est le numéro du pseudo-résidu examiné et la dernière est le numéro de la ligne correspondante dans le tableau ordonné. Cinq cas se présentent alors :

- les correspondances sont, pour chaque atome, dans l'intervalle de tolérance accepté. Il n'y a qu'un seul pseudo-résidu candidat qui réponde à ces exigences, c'est- à-dire que le candidat n'est pas sélectionné par un autre pseudo-résidu et que la seconde proposition n'est pas dans la marge d'acceptation. QUASI-1 accepte automatiquement la connexion (Figure : 2.2).

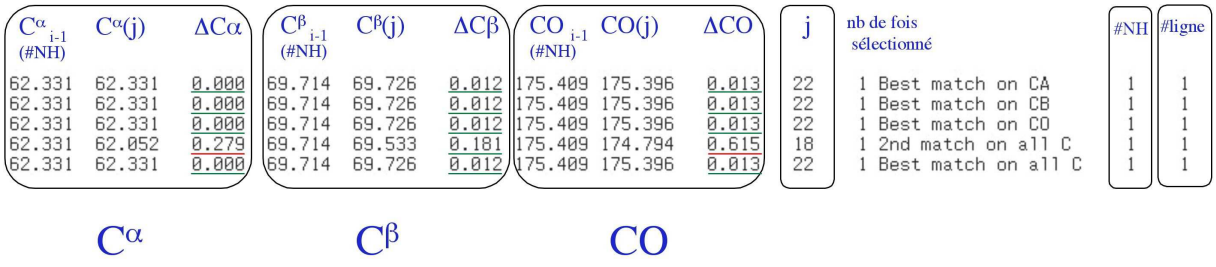

FIG. 2.2 – Présentation d'un cas où QUASI-1 accepte automatiquement la connexion. Ici la connexion se fait entre les pseudo-résidus 22 et 1.

- les deux meilleurs pseudo-résidus candidats donnent des $\Delta\delta$ dans les marges d'acceptation, QUASI-1 invite alors l'utilisateur à inspecter les spectres correspondants

(Figure : 2.3) afin de choisir éventuellement entre les deux propositions.

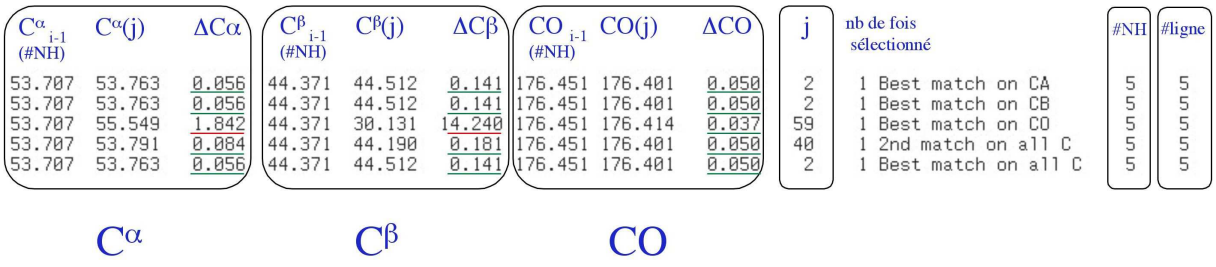

FIG. 2.3 – Les deux meilleurs pseudo-résidus candidats (les numéros 2 et 40) présentent des déplacements chimiques dans les marges d'acceptation.

- il y a au moins un noyau pour lequel la correspondance est hors de l'intervalle fixé : QUASI invite alors l'utilisateur à consulter les spectres. Ce cas peut se présenter pour plusieurs raisons : la sélection d'un des pics peut être difficile parce que le pic est très large ou qu'il y a recouvrement, dans ce cas le pic peut être sélectionné légèrement à côté du maximum et la différence de déplacements chimiques est légèrement au delà de la marge acceptée (Figure : 2.4).

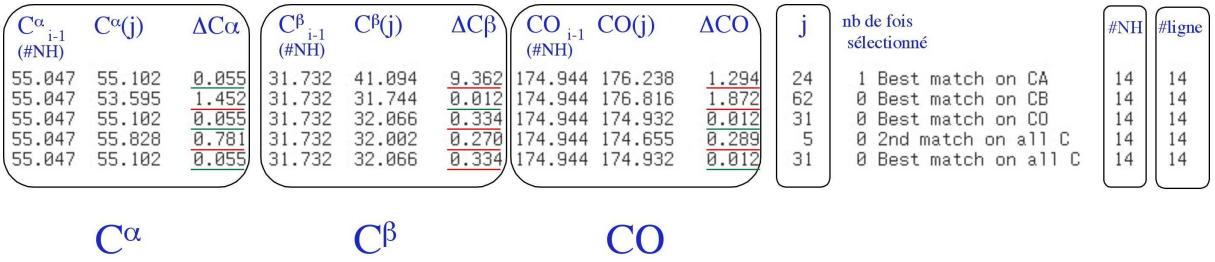

FIG. 2.4 – Dans ce cas, le pseudo-résidu 31 a un C^{β} qui n'est pas dans les marges d'acceptation ($\Delta C^{\beta} > 0,2ppm$).

L'inspection du spectre peut alors conduire à accepter la connexion. Soit il n'y a pas de NH précédent, dans le cas d'une proline ou d'un NH manquant ; après avoir inspecté les spectres l'utilisateur conclut qu'il n'y a pas de pseudo-résidu candidat acceptable (Figure : 2.5).

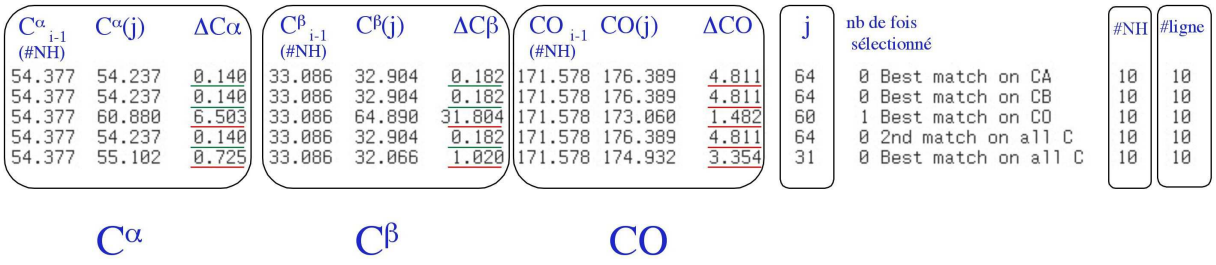

FIG. 2.5 – Aucun candidat ne se rapproche vraiment du pseudo-résidu étudié (numéro 10).

Soit, le déplacement chimique de l'un des noyaux trouve un très bon précédent mais un autre ne trouve pas de correspondant acceptable. L'inspection des spectres peut conduire à accepter une connexion avec le pseudo-résidu présenté dans la $5^{ème}$ ligne pour laquelle des données manquent (Figure : 2.6).

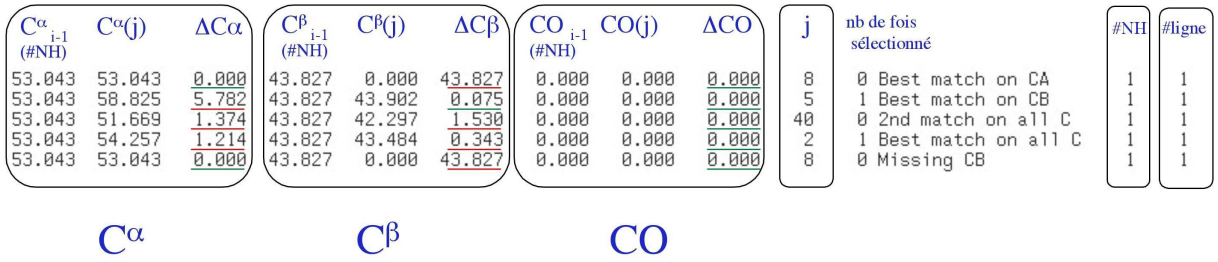

FIG. 2.6 – Pour le déplacement chimique du noyau C^{α} du pseudo-résidu 1, le meilleur précédent est le pseudo-résidu numéro 8 pour lequel le déplacement chimique du noyau $^{13}C_i^{\beta}$ n'a pas pu être identifié, par exemple, à cause de recouvrement des pics $^{13}C_{i-1}^{\beta}$ et $^{13}C_i^{\beta}$. L'inspection des spectres mène l'utilisateur à accepter le pseudo-résidu candidat (cas emprunté à l'utilisation de QUASI-1 sur l'α-actinine)

- dans le cas où la décision doit se faire juste sur un atome ($^{13}C^{\alpha}$) QUASI (ceci englobe par exemple le cas des glycines quand les données sur le déplacement chimique du ^{13}CO sont indisponibles), afin de ne pas accepter une fausse liaison, ne sélectionne aucun précédent, par contre il invite l'utilisateur à consulter les spectres (Figure : 2.7) ;

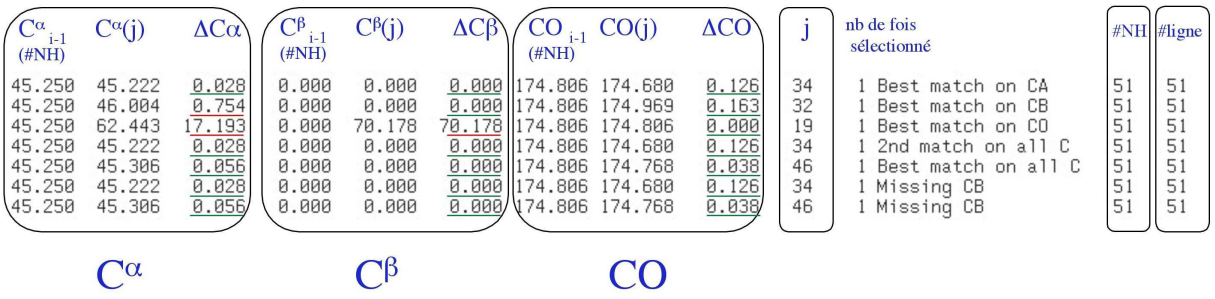

FIG. 2.7 – Le choix de connexion doit se faire sur un seul type de déplacement chimique. QUASI-1 invite l'utilisateur à inspecter les spectres.

- toutes les correspondances sont bonnes, ou proches du seuil fixé, mais le candidat en question est déjà identifié comme étant le précédent d'un autre pseudo-résidu, QUASI invite alors l'utilisateur à consulter les spectres (Figure : 2.8).

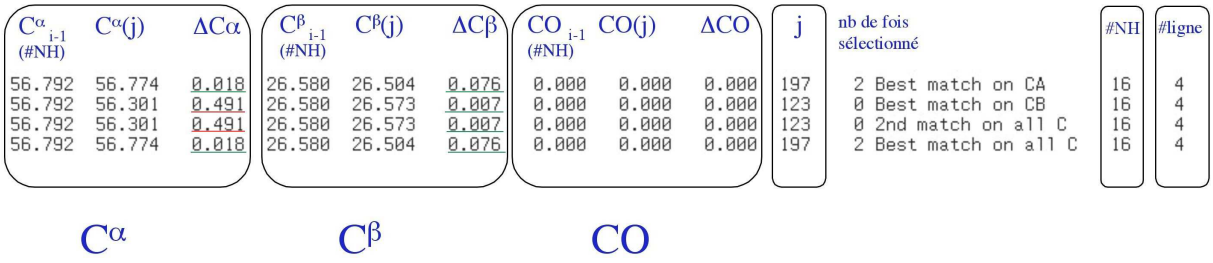

FIG. 2.8 – Le pseudo-résidu numéro 197 est identifié comme précédent potentiel par deux pseudo-résidus (cas emprunté à l'utilisation de QUASI sur le fragment 24 kDa de la sous-unité B de l'ADN-gyrase).

Dans les cas des pseudo-résidus pour lesquels aucun précédent n'est identifié, les δ_{Ci-1} sont ajoutés, sans δ_N ni δ_H à la liste des pseudo-résidus.Ces pseudo-résidus sont de la forme : { 0,000 ; 0,000 ; $^{13}C_i^{\alpha}$; $^{13}C_i^{\beta}$; $^{13}CO_i$}. Ils marquent le début d'un nouvel enchaînement.

A la fin de QUASI-1, un pseudo-résidu précédent est identifié pour chaque pseudo-résidu. L'ordre des pseudo-résidus et les données associées sont consignés dans un tableau ordonné (le fichier QUASI1.table). A partir de chaque début d'enchaînement, on peut construire et définir à rebours un fragment : enchaînement séquentiel de pseudo-résidus, non encore placés sur la séquence primaire de la protéine. Ce placement fait l'objet de la seconde partie de QUASI : QUASI-2.

2.3 Résultats avec l'ubiquitine

2.3.1 Ubiquitine

L'ubiquitine est une protéine de 76 acides aminés, qui est présente dans toutes les cellules. Elle constitue une étiquette moléculaire dans le processus de dégradation des protéines par le protéasome. Sous l'effet d'enzymes appelées E1 (activatrice), E2 (conjugase) et E3 (ligase) et en présence d'ATP, il y a biosynthèse d'une chaîne poly-ubiquitine qui se fixe sur les protéines à dégrader. La protéine ainsi poly-ubiquitinée est un substrat du protéasome 26 S qui l'hydrolyse en peptides de 3 à 25 acides aminés qui seront partiellement hydrolysés en acides aminés, lesquels pourront être réutilisés par la cellule. Les molécules d'ubiquitine libérées par dé-ubiquitination seront également réutilisées.

2.3.2 Application de QUASI-1 à la structure de l'ubiquitine

Le comportement de QUASI-1 va être détaillé sur le cas de l'ubiquitine, protéine de taille moyenne (76 résidus). Elle est très étudiée en RMN [Cavanagh *et al.*, 1996] et souvent utilisée comme plate-forme de validation pour les nouvelles méthodes d'études de protéines par RMN.

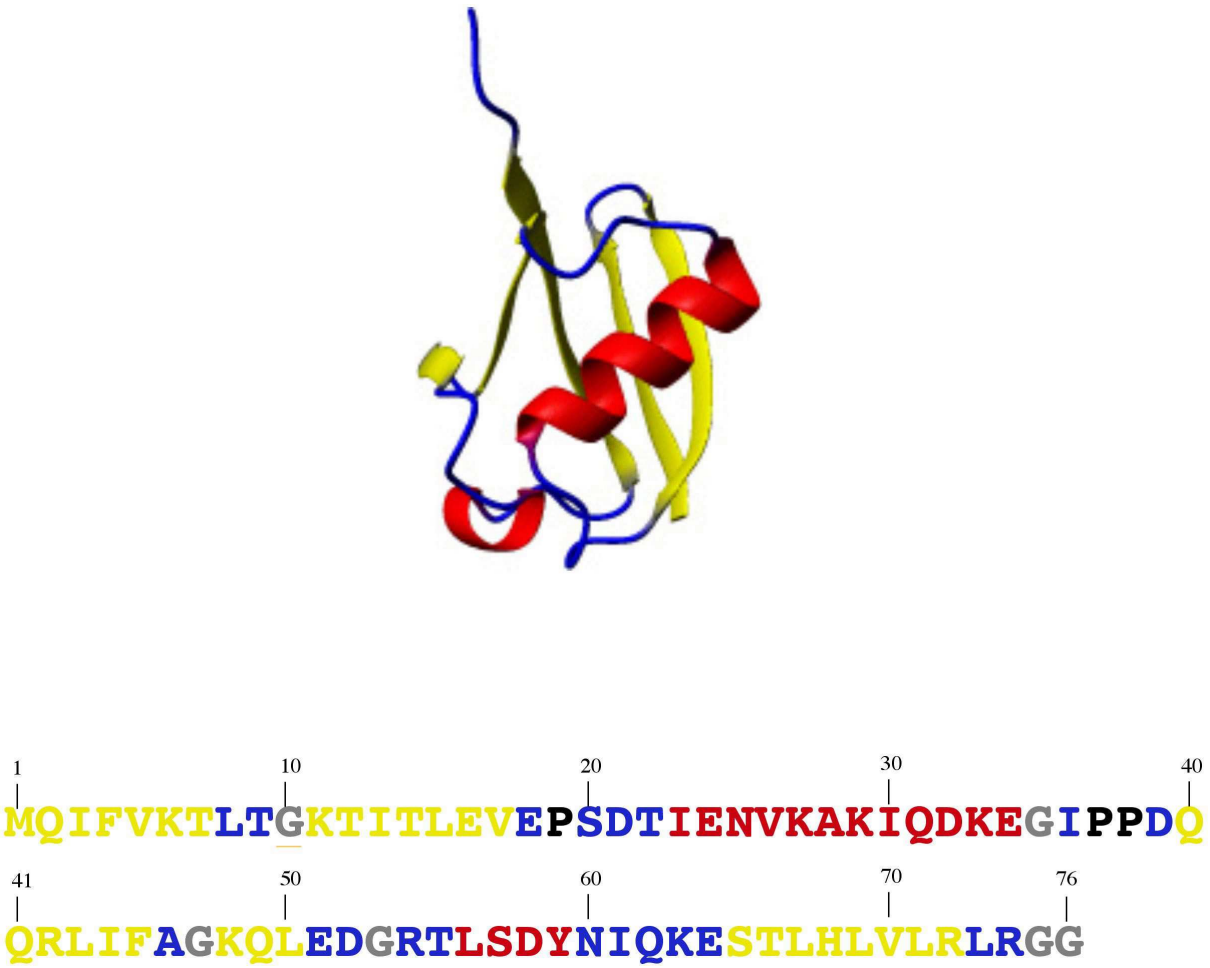

FIG. 2.9 – Représentation en ruban de l'ubiquitine. Séquence primaire de la protéine. Les prolines sont présentées en noir et les glycines en gris. Les hélices-α sont en rouge et les brins-β en jaune.

Parmi ces 76 résidus, il y a 3 prolines (Figure : 2.9) en position 19, 37 et 38 ; on s'attend donc à obtenir 3 fragments de 18, 18 et 39 résidus (1UBQ.pdb). Le jeu de spectres utilisés est : HNCA, HN(CO)CA, CBCA(CO)NH, CBCANH, HNCO et HN(CA)CO. Ils ont été acquis à UCL, Londres et sont disponibles sur le site : http ://www.biochem.ucl.ac.uk/bsm/nmr/ubq/.

Spectre	Nombre de pics
HN(CO)CA	70
HNCA	70
CBCA(CO)NH	65
CBCANH	65
HNCO	70
HN(CA)CO	70

TAB. 2.2 – Nombre de pics de chaque fichier utilisé par QUASI-1. Le nombre de corrélations sélectionnées dans l'HSQC est 92.

On a donc 70 pseudo-résidus (Tableau : 2.2) qu'il s'agit de connecter avec son précédent. Dans cette étape, l'utilisateur est invité 19 fois à intervenir et à faire un choix manuellement. Pour un cas, les deux meilleures propositions se trouvent dans la marge fixée (Figures : 2.3, 2.10A et 2.4, 2.10C). Dans 12 cas, la différence entre les C^{β} est juste au dessus du seuil de 0,2 ppm fixé par le programme. 5 autres cas sont des cas où les meilleures correspondances sont à plus de 1 ppm (Figure : 2.5 et 2.10D,E).

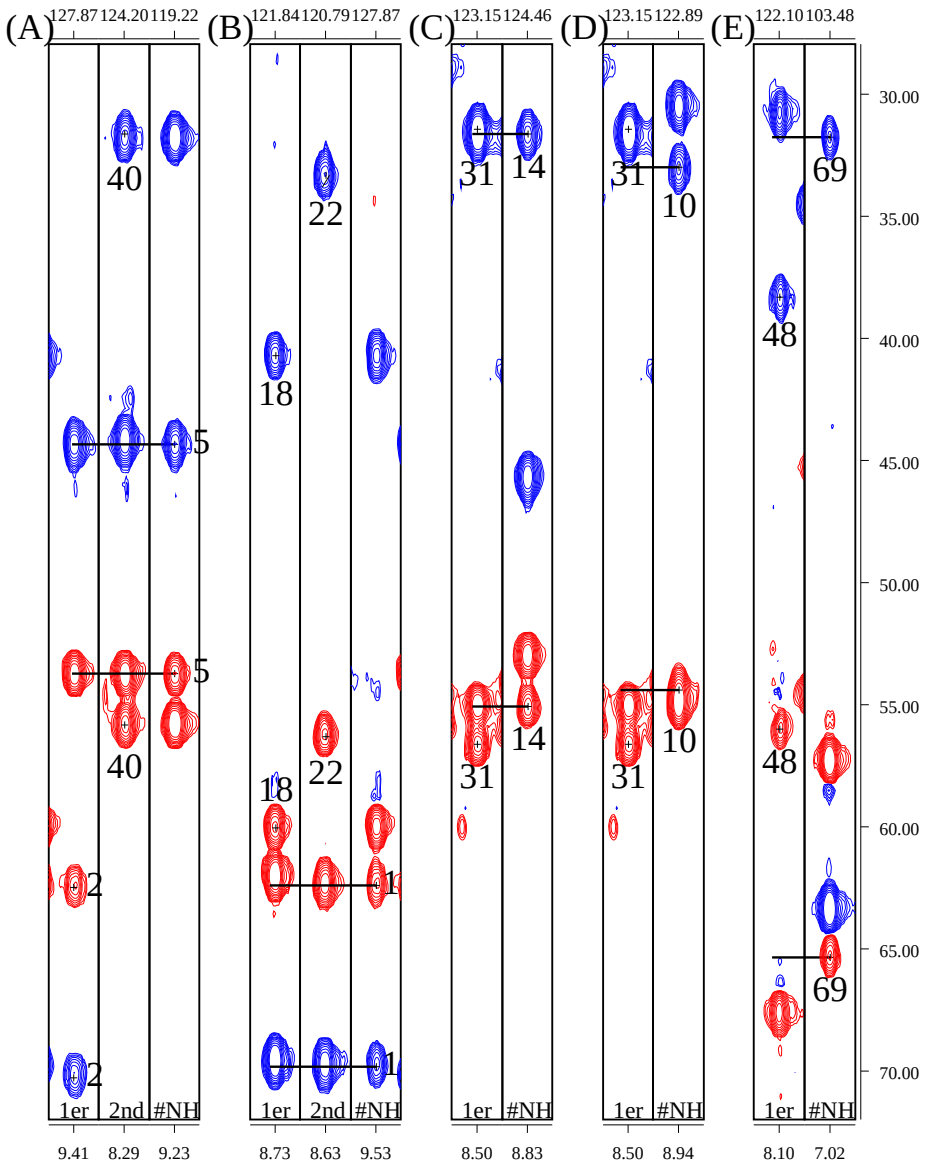

FIG. 2.10 – Spectres HNCACB de l'ubiquitine. Les pics négatifs sont représentés en bleu, les positifs en rouge. L'échelle supérieure indique le δ_{Ni}, celle de droite est le δ_C et celle du bas le δ_{Hi}. Dans le premier cas (A), les pics $^{13}C_i^{\alpha}$ des pseudo-résidus 40 et 2 coïncident très bien avec le pic $^{13}C_{i-1}^{\alpha}$ du pseudo-résidu 5. Par contre au niveau des $^{13}C^{\beta}$, le pic de pseudo-résidu 2 est plus proche. Les différences de δ des 3 noyaux ($^{13}C^{\alpha}$, $^{13}C^{\beta}$ et ^{13}CO) sont dans les marges pour les deux pseudo-résidus. L'inspection du spectre conduit à accepter le pseudo-résidu 2 comme précédent du 5 (Figure : 2.3). L'illustration (B) présente une connexion automatiquement acceptée par QUASI (Figure : 2.2). (C) Cas où la différence entre les C^{β} est légèrement plus grande que la marge fixée (0,334 ppm > 0,200 ppm) (Figure : 2.4). Le recouvrement des pics $^{13}C_i^{\beta}$ et $^{13}C_{i-1}^{\beta}$ rend la sélection des pics difficile. Dans les deux derniers cas (D) et (E), aucune proposition acceptable n'est trouvée, la meilleure possibilité est très éloignée. Le cas (D) (Figure : 2.5) ne trouve aucun précédent. Le cas (E) est représentatif d'une situation où les déplacements chimiques des $^{13}C^{\alpha}$ et $^{13}C^{\beta}$ suggèrent que le pseudo-résidu qui précède est une proline.

Dans l'ultime cas, les deux meilleures propositions se trouvent dans les marges d'acceptation, le pseudo-résidu précédent est de type glycine (identifiable par un $\delta_{C\alpha i}$ proche de 45 ppm et par l'absence de pic C_i^{β}) (Figure : 2.11).

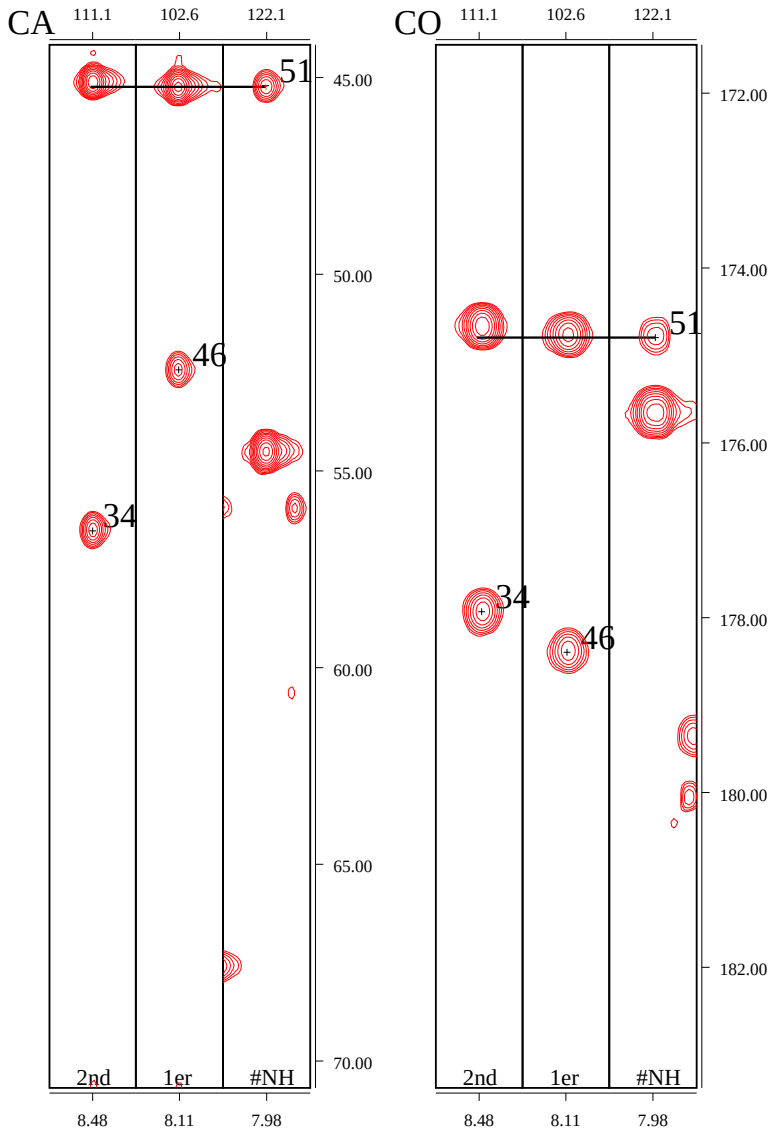

FIG. 2.11 – Présentation du cas illustré par la figure 2.7. A droite (spectre HNCA), la correspondance entre le déplacement chimique du $^{13}C_{i-1}^{\alpha}$ du pseudo-résidu 51 et les déplacements chimiques du $^{13}C_i^{\alpha}$ des pseudo-résidus 34 et 46 : le pseudo-résidu 46 semble plus proche. A gauche (spectre HN(CA)CO), les correspondances entre le déplacement chimique du $^{13}CO_{i-1}$ du pseudo-résidu 51 et les déplacements chimiques des $^{13}CO_i$ des pseudo-résidus 34 et 46 : le pseudo-résidu 46 semble être le meilleur candidat. La connexion est donc choisie entre les pseudo-résidus 46 et 51.

A la fin de cette étape, on obtient 5 fragments, deux de plus que le nombre attendu au départ, de 5, 13, 15, 18 et 24 acides aminés. Ce nombre de fragments s'explique par le fait que deux NH n'ont pas été sélectionnés sur le spectre HSQC (Figure : 2.12). Ils sont de faibles intensités et n'ont pas été sélectionnés.

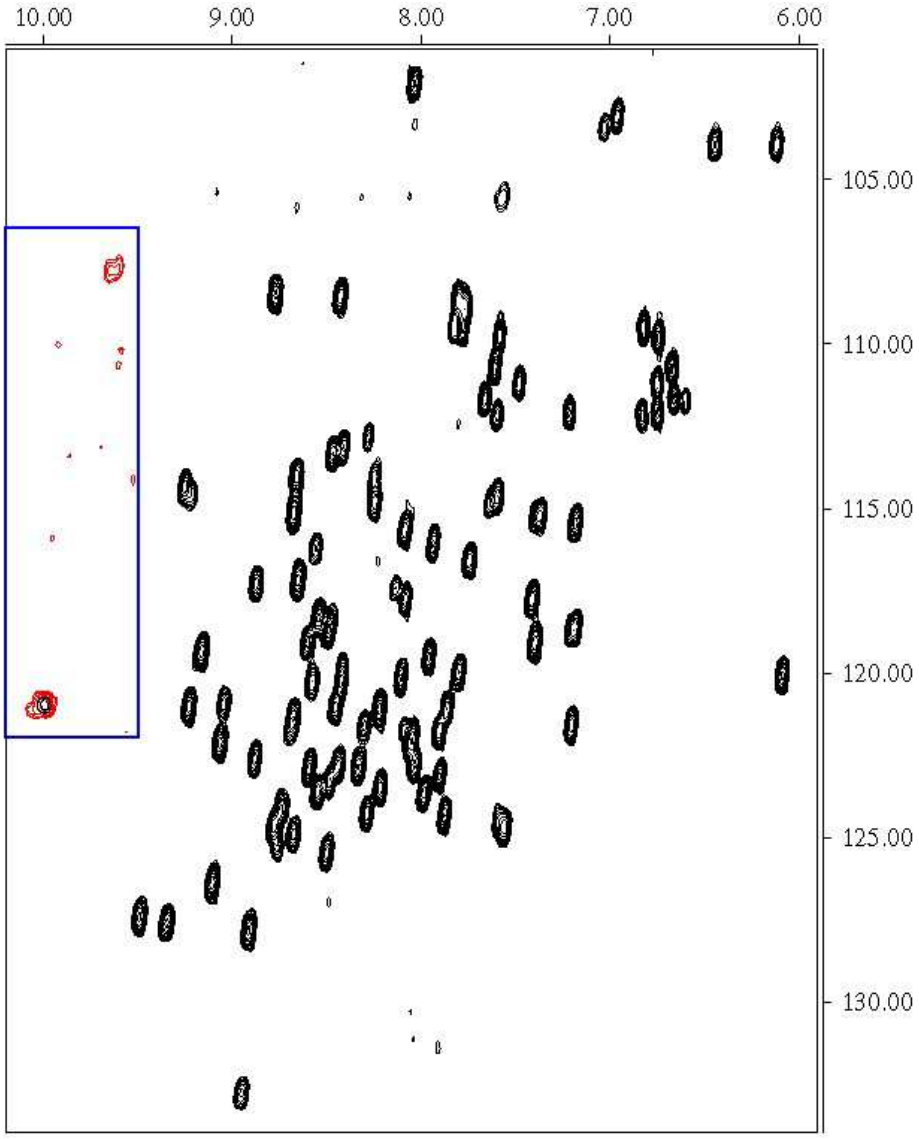

FIG. 2.12 – Spectre 2D ^{1}H-^{15}N HSQC acquis sur l'ubiquitine. En noir le niveau auquel les pics ont été sélectionnés, en rouge entouré de bleu, la zone où se trouvent les deux NH manquants.

Dans le cas de l'ubiquitine, QUASI-1 a donc accepté 51 connexions automatiquement (Figure : 2.10B), soit 73% des connexions. Ce résultat est dû à une bonne qualité et une bonne dispersion des spectres RMN. QUASI-1 écrit le tableau ordonné des pseudo-résidus

dans le fichier QUASI1.table (Figure 2.13).

```
  0.000   0.000  54.377  33.086 171.578   0.000   0.000   0.000
123.013   8.939  55.130  30.647 177.054  54.377  33.086 171.578
115.212   8.304  59.568  42.126 173.387  55.019  30.506 177.079
118.607   8.598  55.102  41.094 176.238  59.568  42.049 173.437
121.353   9.299  60.405  34.323 175.849  55.075  41.211 176.238
127.927   8.906  54.572  34.710 178.160  60.406  34.182 175.874
115.596   8.740  60.489  70.565 177.959  54.545  34.376 178.172
121.419   9.107  57.503  42.062 179.881  60.489  70.488 177.959
105.962   7.631  61.354  68.953 176.539  57.475  41.921 179.893
109.308   7.813  45.334   0.000 175.020  61.382  69.005 176.564
121.998   7.260  56.247  33.485 176.828  45.278   0.000 175.045
120.726   8.633  62.331  69.726 175.396  56.275  33.408 176.841
127.757   9.530  59.987  40.836 176.213  62.331  69.714 175.409
121.798   8.729  62.052  69.533 174.794  59.987  40.760 176.250
125.222   8.729  52.758  46.963 175.585  61.997  69.585 174.818
122.555   8.118  54.879  29.745 176.853  52.702  46.822 175.610
117.627   8.929  58.424  36.322 175.082  54.852  29.732 176.878
119.389   8.652  52.702  30.970 177.306  58.368  36.439 175.107
  0.000   0.000  66.155  32.763 179.328   0.000   0.000   0.000
113.749   8.522  55.716  39.869 178.084  66.155  32.763 179.328
116.959   7.809  55.549  30.131 176.414  55.689  39.728 178.097
118.153   7.474  56.637  31.744 177.205  55.521  30.055 176.464
123.131   8.502  55.102  32.066 174.932  56.582  31.538 177.218
124.503   8.826  52.981  45.737 176.326  55.047  31.732 174.944
122.334   9.085  58.898  41.223 176.891  52.954  45.725 176.338
125.287   8.843  56.582  43.738 175.597  58.898  41.147 176.904
132.974   8.945  52.507  16.654 178.361  56.554  43.662 175.635
102.572   8.116  45.306   0.000 174.768  52.451  16.448 178.386
122.080   7.980  54.544  34.581 175.673  45.250   0.000 174.806
123.066   8.631  55.856  29.229 176.615  54.545  34.440 175.685
125.823   8.550  54.209  41.352 177.682  55.800  29.088 176.640
123.235   8.382  55.912  31.937 176.502  54.154  41.340 177.682
120.445   8.150  56.609  40.772 178.436  55.856  31.925 176.539
  0.000   0.000  60.657  28.701 180.019   0.000   0.000   0.000
121.440   7.921  56.023  38.515 179.341  60.657  28.701 180.019
122.223   8.098  67.606  31.034 178.951  55.968  38.374 179.353
119.062   8.553  59.205  33.743 181.539  67.634  30.764 178.963
123.574   7.978  55.326  17.879 181.262  59.150  33.731 181.526
120.315   7.856  59.735  33.485 181.363  55.298  17.673 181.275
121.419   8.276  66.071  36.903 179.227  59.708  33.279 181.337
123.706   8.551  59.987  27.810 179.906  66.099  36.762 179.240
119.843   8.010  57.419  41.159 178.348  59.987  27.669 179.893
115.539   7.420  58.117  34.065 178.876  57.363  41.082 178.373
114.417   8.719  55.326  33.420 179.014  58.117  33.989 178.876
108.976   8.496  46.004   0.000 174.969  55.270  33.279 178.976
120.374   6.145  57.726  40.578 174.454  45.948   0.000 174.994
  0.000   0.000  45.055   0.000 175.848   0.000   0.000   0.000
119.445   7.456  54.237  32.904 176.389  45.055   0.000 175.848
108.905   8.818  59.680  72.177 177.544  54.210  32.699 176.401
118.138   8.144  58.619  40.385 181.815  59.652  72.358 177.569
113.593   8.471  61.103  62.439 179.341  58.619  40.244 181.802
124.605   7.928  57.391  40.385 178.474  61.048  62.492 179.328
115.865   7.249  58.228  39.933 175.698  57.363  40.308 178.461
116.093   8.146  54.098  37.354 175.296  58.229  39.921 175.735
118.991   7.236  62.415  36.709 175.610  54.042  37.342 175.333
125.055   7.614  53.595  31.744 176.816  62.387  36.697 175.597
120.687   8.483  57.810  32.711 176.740  53.568  31.538 176.841
114.658   9.304  58.340  26.133 176.251  57.810  32.570 176.790
115.027   7.658  60.880  64.890 173.060  58.312  25.928 176.263
117.499   8.726  62.443  70.178 174.806  60.852  64.942 173.110
127.744   9.413  53.763  44.512 176.401  62.415  70.166 174.844
119.232   9.228  55.828  32.002 174.655  53.707  44.371 176.451
124.150   8.292  53.791  44.190 176.401  55.800  31.732 174.693
126.856   9.181  60.601  34.904 175.082  53.763  44.242 176.426
```

FIG. 2.13 – Extrait d'un fichier QUASI1.table issu de QUASI1. Les colonnes correspondent aux valeurs δ_{Ni}, δ_{Hi}, $\delta_{C\alpha i}$, $\delta_{C\beta i}$, δ_{COi}, $\delta_{C\alpha i-1}$, $\delta_{C\beta i-1}$, δ_{COi-1}.

3

QUASI-2 : Placement des fragments sur la séquence primaire

Sommaire

L'objet du programme QUASI-2 est de placer les fragments, obtenus par QUASI-1, sur la séquence primaire. Cette procédure utilise un score de comparaison entre les paramètres expérimentaux liés au fragment et les mêmes paramètres obtenus à partir de l'information de séquence. Ces comparaisons sont présentées sous la forme d'un "tableau de bord" qui permet de visualiser toutes les informations disponibles à un endroit unique. Les données d'entrée nécessaires à ce module sont des fragments constitués de pseudo-résidus enchaînés sur la base des déplacements chimiques observés sur des spectres triple-résonances. Pour

effectuer le placement de ce fragment (autrement dit l'attribution) il faut donc trouver une méthode robuste, indépendante de la composition chimique et de la taille de la protéine, qui permette ce placement. Les fragments sont des enchaînements de déplacements chimiques (δ_{frag}) liés à des atomes, la séquence primaire est un enchaînement de résidus composés d'atomes ; eux mêmes repérables par le biais de leur déplacement chimique respectif (δ_{ref}). On peut ainsi développer un score qui compare les valeurs de (δ_{frag}) aux (δ_{ref}). Alors que les valeurs de δ_{frag} sont issues uniquement des spectres expérimentaux, celles des δ_{ref} peuvent provenir de diverses sources : soit de programmes de prédiction de déplacements chimiques, soit de tables diverses. Dans le prochain chapitre, les indices suivants seront utilisés : i est le numéro du pseudo-résidu traité, j est la position sur la séquence et k est le numéro du fragment étudié.

3.1 L'interface graphique

Lors du développement du programme, une interface graphique a été mise au point en Python (Tkinter). Elle a pour but de faciliter et de rendre conviviale l'utilisation de QUASI-2. Elle se présente sous la forme d'onglets successifs (Figure : 3.1) qui permettent une navigation rapide et aisée entre les pages.

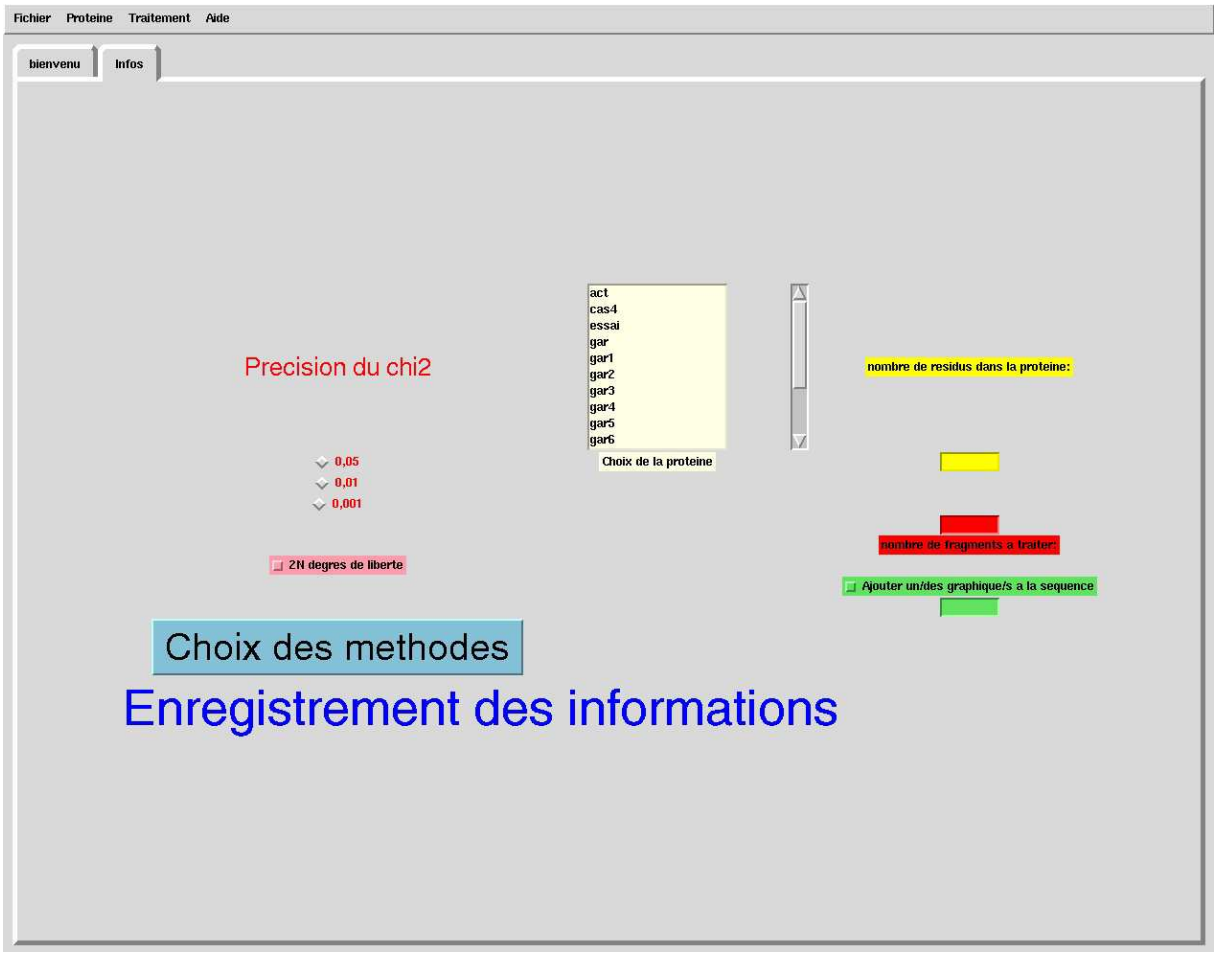

FIG. 3.1 – Première page de QUASI, qui permet la sélection des données d'entrée. Le choix du répertoire, contenant le fichier QUASI 1.table, se fait dans la partie blanche, le nombre de résidus dans la protéine cible est à entrer dans la partie jaune. En fonction du nombre de fragments créés lors de la première étape (QUASI-1), on choisit le nombre d'entre eux que l'on veut traiter (partie rouge). La partie rouge est liée aux caractéristiques choisies pour le test χ^2. La partie verte permet d'intégrer des tracés sous les attributions proposées par QUASI.

Une fois les paramètres liés à la protéine cible enregistrés, il faut choisir les méthodes (Figure : 3.2) qui seront comparées pendant l'utilisation de QUASI-2.

FIG. 3.2 – Choix des méthodes à comparer.

Pour chacune d'entre elles, QUASI crée deux onglets : le premier, sorte de "tableau de bord" présente les attributions acceptées (Figure 3.3), le second présente les histogrammes de chaque fragment traité. A partir du "tableau de bord", l'utilisateur peut accepter l'attribution.

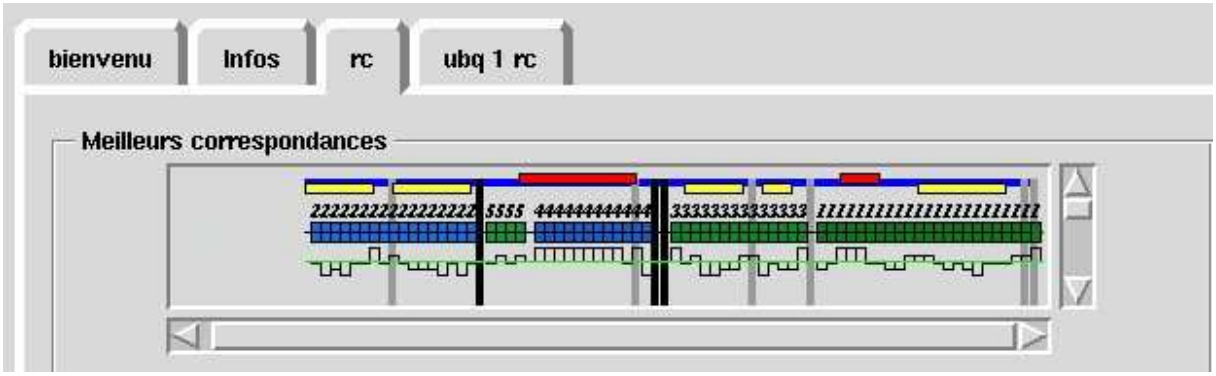

FIG. 3.3 – Tableau de bord dans le cas de l'ubiquitine. La ligne horizontale bleue schématise la séquence primaire de la protéine cible. Les éléments de structure secondaire sont sous la forme de barre jaune pour les feuillets-β, rouge pour les hélices α. Les barres verticales noires indiquent la position des prolines, les grises celles des glycines. La seconde ligne horizontale présente l'attribution faite par QUASI. Chaque pseudo-résidu (avec un NH) est symbolisé par un petit rectangle dont la couleur est spécifique du fragment auquel il appartient. Il est surmonté du numéro du fragment (ordre de taille décroissante). La troisième ligne horizontale (verte) est présente dans les cas où l'utilisateur a sélectionné la méthode CSI. On y trouve la valeur de l'index calculé sous la forme de barres successives.

Si l'attribution ne convient pas à l'utilisateur, il est invité à inspecter le détail du calcul des scores sous forme d'histogrammes (Figure : 3.4) dans l'onglet suivant. Cette analyse permet l'identification d'éventuels problèmes dans les données. Chaque histogramme est un bouton. En cas de doute sur un fragment, l'utilisateur sélectionne l'histogramme

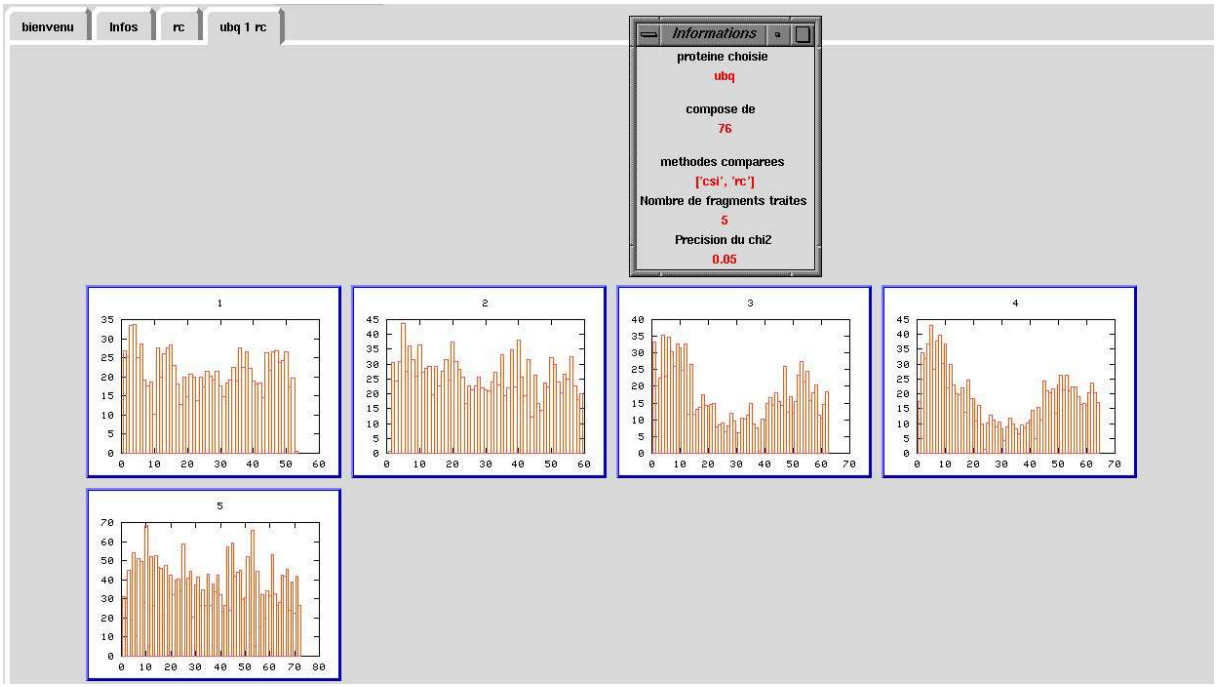

FIG. 3.4 – Histogrammes des fragments traités (ici au nombre de 5). Chaque barre est la valeur de la fonction cible lorsque le fragment est "glissé" sur la séquence. Un petit score (significatif d'un bon accord) est représenté par une petite barre. Sur cette figure, on voit également un élément omniprésent dans QUASI : le récapitulatif des choix fait par l'utilisateur.

correspondant et des histogrammes plus précis apparaissent sur l'onglet suivant. A ce niveau, il y a un histogramme par placement possible sur la séquence primaire. Sur ces histogrammes (Figure : 3.5) est tracée la différence des déplacements chimiques au niveau du pseudo-résidu (une barre par pseudo-résidu à chaque position).

Plus la cohérence entre l'expérience et le modèle est bonne, plus petites seront les barres de l'histogramme.

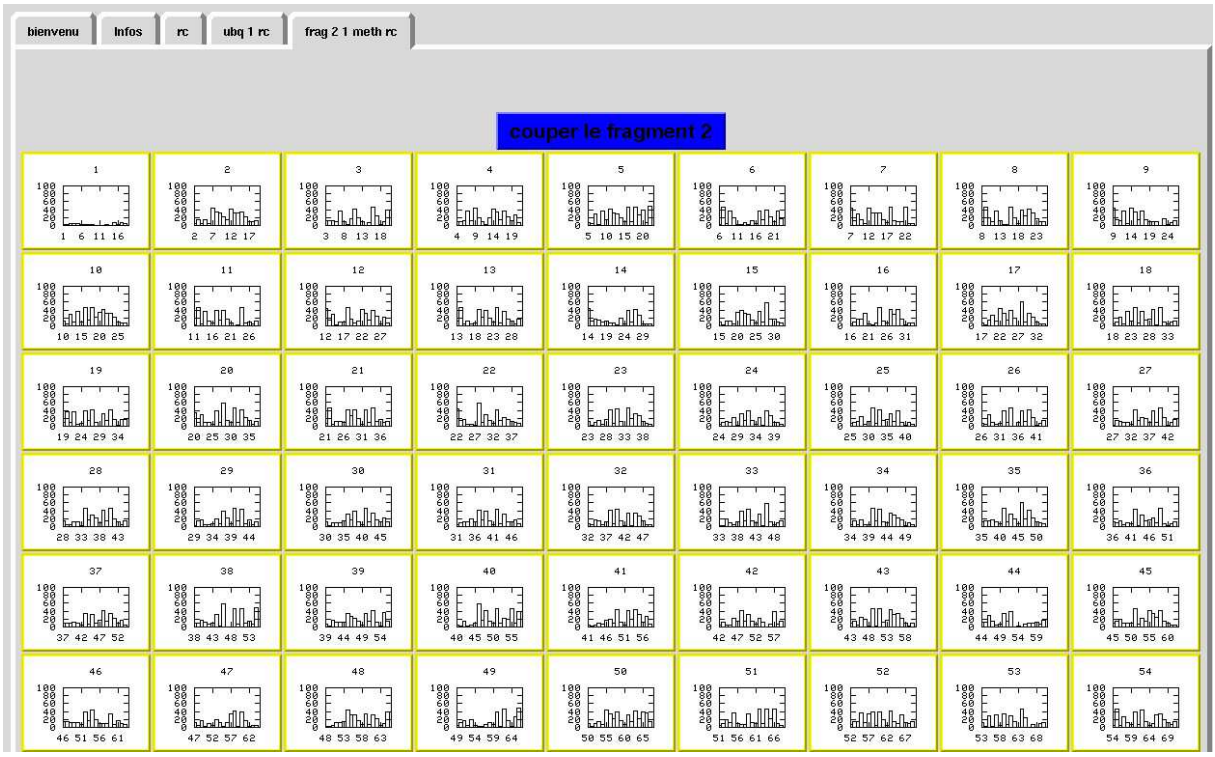

FIG. 3.5 – Histogrammes présentant l'accord local entre les données expérimentales et les références.

A tout moment de cette étude, il est possible de couper un fragment ou de le figer à un endroit de la séquence (Figure : 3.6). Outre l'étude des déplacements chimiques, il est possible de tracer des données (CSI, relaxation...) en fonction de l'attribution faite par QUASI.

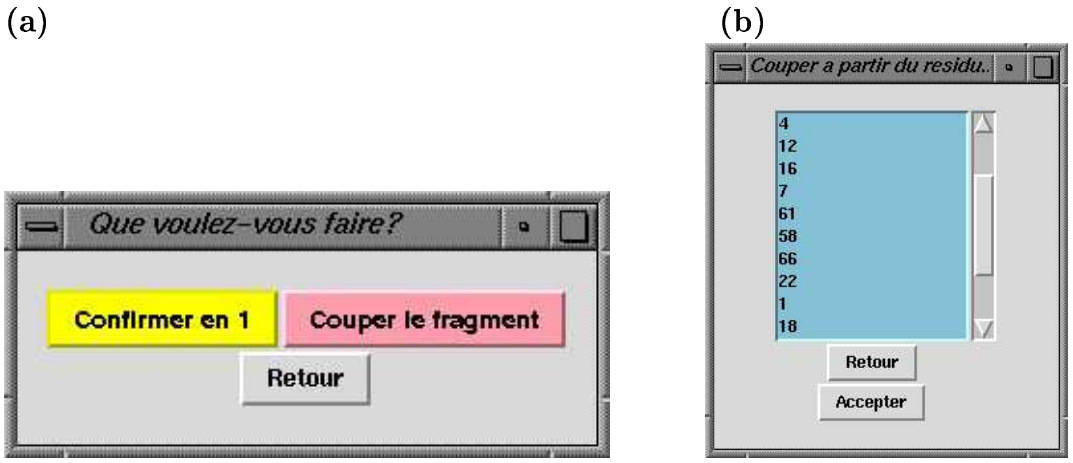

FIG. 3.6 – (a) Fenêtre proposant la coupure du fragment ou l'ancrage de celui-ci à la position choisie. (b) Présentation des pseudo-résidus présents dans le fragment sélectionné. Cette fenêtre permet la coupure du fragment.

3.2 Présentation de données structurales

Outre l'utilisation du déplacement chimique, QUASI-2 permet de regrouper des informations structurales de différents types. Il s'agit de tenir compte de ces informations pour le placement des fragments lorsque la structure tridimensionnelle de la protéine ou d'un homologue est connue. Plusieurs possibilités se présentent : l'utilisation de l'effet de la structure secondaire sur les déplacements chimiques sous la forme d'index de déplacements chimiques (CSI) et/ou l'utilisation des données de relaxation pour délimiter les régions structurées. Lues dans un fichier formé de 2 colonnes, ces données peuvent être tracées en fonction des attributions faites par QUASI-2. Pour le moment ces informations sont présentées sur le "tableau de bord" et permettent à l'utilisateur de juger de l'attribution réalisée par QUASI-2.

3.2.1 Chemical Shift Index

Les déplacements chimiques des noyaux de la chaîne polypeptidique dépendent de la conformation locale de l'acide aminé, et donc de la structure secondaire de la protéine. Les noyaux les plus sensibles sont les protons $^{1}H^{\alpha}$, $^{13}C^{\alpha}$ et ^{13}CO. En 1991, Wishart et al. [Wishart *et al.*, 1991] ont proposé d'utiliser cette information de structure locale sous la forme d'un index défini de la façon suivante (Equation : 3.1) :

$$\begin{aligned} \delta_{obs} < \delta_{table} - marge &\Rightarrow Index = -1 \\ \delta_{table} - marge < \delta_{obs} < \delta_{table} + marge &\Rightarrow Index = 0 \\ \delta_{obs} > \delta_{table} + marge &\Rightarrow Index = 1 \end{aligned} \tag{3.1}$$

Une *marge* est définie pour chaque noyau : 0,5 pour CO, 0,7 pour C^{α} et C^{β} et 0,10 pour le H^{α} excepté pour le cas de l'acide aminé proline dont les marges sont à 4 ppm. Lorsque les déplacements chimiques observés sont dans les marges, l'index est mis à zéro, les cas inférieurs et supérieurs se voient indexés respectivement à "-1" et "1". L'enchaînement continue de plusieurs résidus présentant une valeur d'index de même signe est interprété comme la signature de la présence d'une structure secondaire régulière [Wishart *et al.*, 1991].
QUASI-2 propose l'utilisation de cette méthode pour donner du poids aux attributions enregistrées. C'est la seule référence qui ne fonctionne pas avec un score. Elle se calque sur les autres méthodes et présente les index sur une ligne ; on peut les comparer à la structure secondaire issue du fichier PDB.

3.2.2 Les données de relaxation

Tout comme CSI, les données de relaxation peuvent confirmer ou infirmer une attribution. Ces données permettent de connaître les caractéristiques de mobilité de la protéine cible. En général dans les structures secondaires, la mobilité est restreinte et dans les boucles la mobilité augmente. Le tracé des valeurs S^2 en fonction des propositions de QUASI-2 peuvent aussi donner un premier aperçu sur la cohérence de l'attribution. En effet, les résidus présentant une mobilité interne importante sont souvent regroupés dans des régions distinctes de la protéine, et la variation du paramètre d'ordre en fonction de la séquence est lissée. En tracant ces données dynamiques en fonction des propositions faites par QUASI-2, on peut espérer confirmer ou infirmer les attributions. En effet si on détermine une zone ordonnée, et qu'on y trouve un résidu très mobile, on se devra de retourner aux spectres afin de vérifier les connexions entre les pseudo-résidus. Ainsi dans le cas du fragment 24 kDa de l'ADN-gyrase les données présentées dans l'ordre de la sélection des pics dans le spectre 2D ^{1}H-^{15}N HSQC (Figure : 3.7) offrent un profil irrégulier et désordonné.

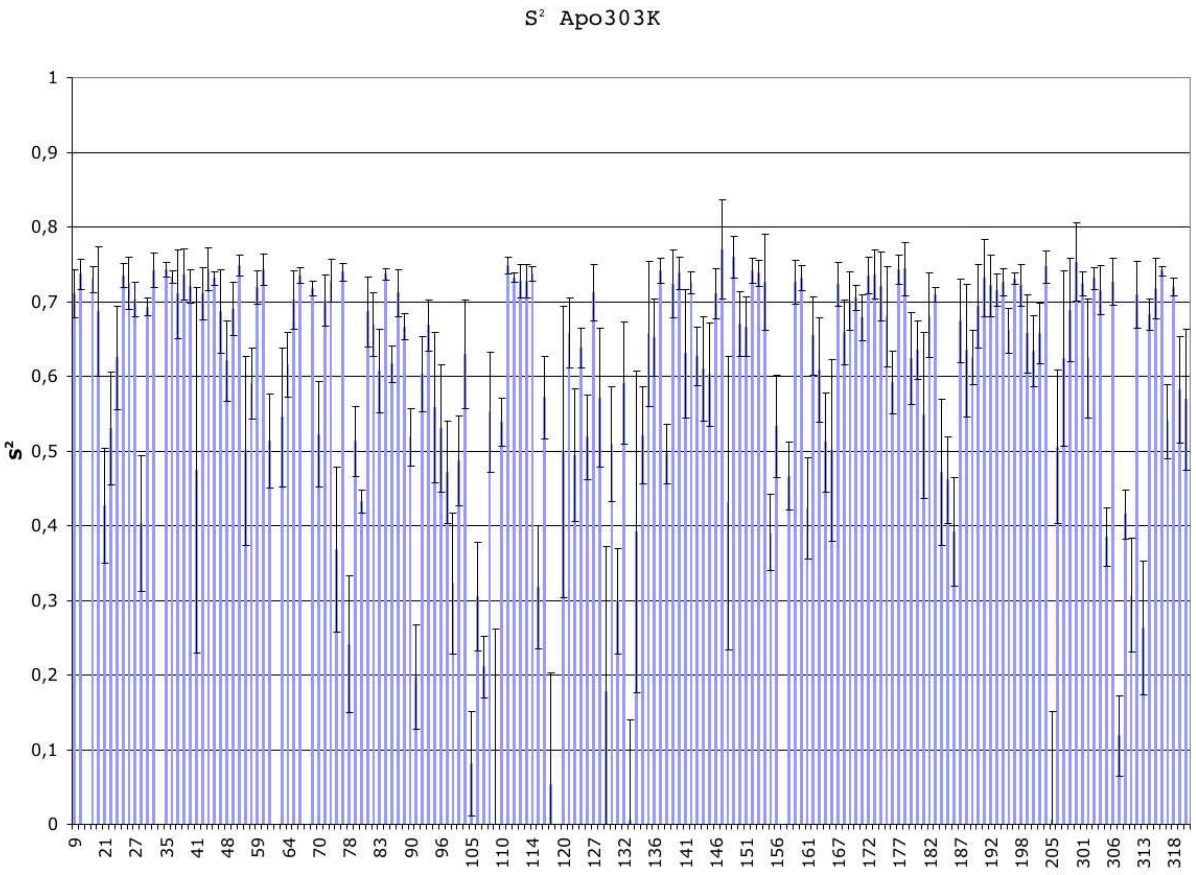

FIG. 3.7 – Données de relaxation tracées en fonction de la numérotation aléatoire des pics du spectre 2D ^{1}H-^{15}N HSQC.

En traçant les mêmes données dans l'ordre suggéré par QUASI-2 (Figure : 3.8), les valeurs élevées du paramètre d'ordre sont regroupées dans les structures secondaires (cas du fragment 24 kDa de l'ADN-gyrase).

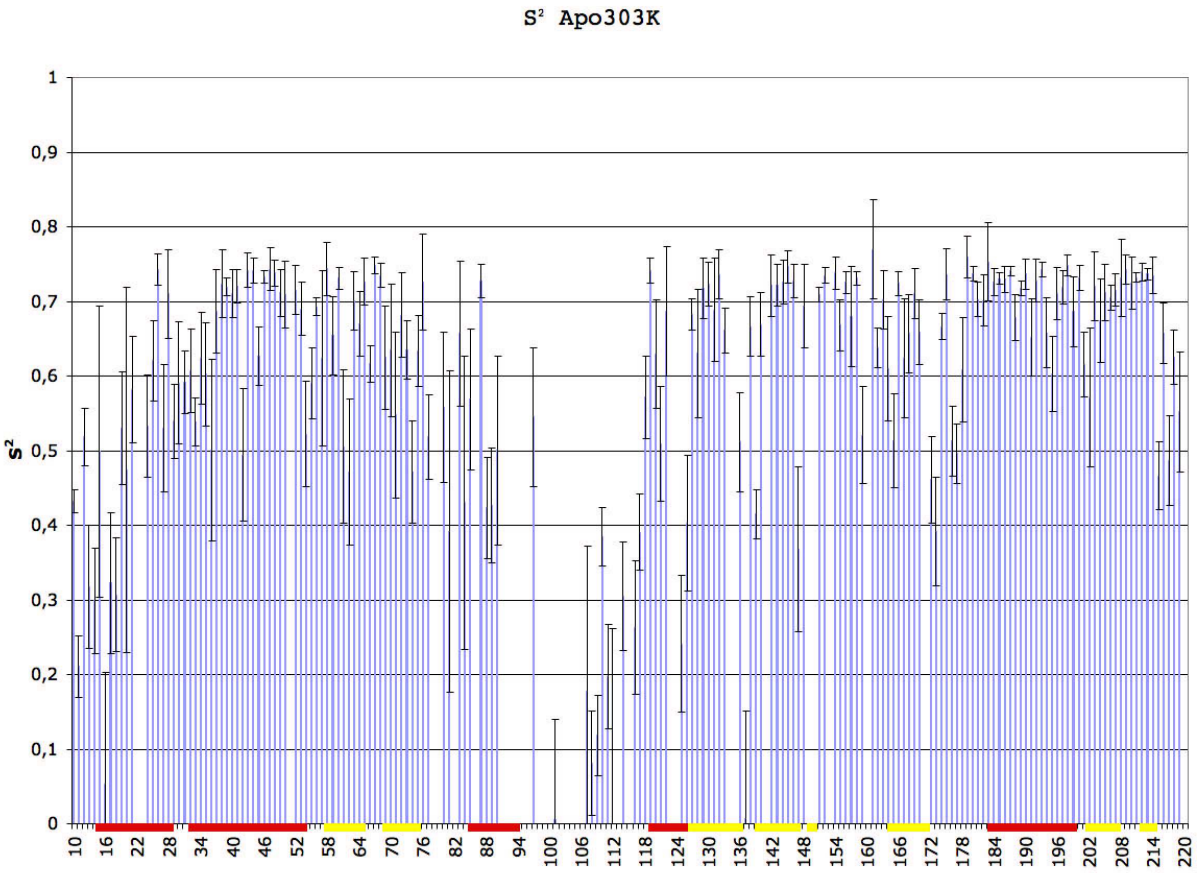

FIG. 3.8 – Données de relaxation tracées en fonction de la numérotation dans la séquence.

Cette interface graphique fait de QUASI un outil agréable à utiliser. De plus, toutes les informations obtenues à partir de spectres RMN sont susceptibles d'être incorporées dans le "tableau de bord". Parmi elles se trouvent les déplacements chimiques. Le placement des fragments se fait suivant l'accord existant entre les données expérimentales et les données de référence.

3.3 Les références

QUASI-2 permet l'utilisation de différentes références en fonction desquelles il va placer les fragments. Les références prises en compte actuellement par QUASI pour calculer la fonction cible peuvent se regrouper en deux groupes :

- les observables qui dépendent de la nature chimique des acides aminés.
- les observables qui vont être sensibles à la structure 3D de la protéine.

3.3.1 Les tables utilisées comme référence

La première observable que l'on peut utiliser pour distinguer les acides aminés est le déplacement chimique. Nous avons utilisé 2 types de valeurs statistiques.

Table Random Coil

La table Random Coil utilisée donne les déplacements chimiques des acides aminés dans les pentapeptides GGXGG enregistrés dans une solution de 8M urée [Schwarzinger *et al.*, 2000]. Ces données ont été préférées à celles du tétrapeptyde H-Gly-Gly-X-L-Ala-OH enregistrées dans une solution aqueuse. Ces valeurs sont dépourvues de tout aspect de structure secondaire, elles reflètent la dépendance des déplacements chimiques à la structure chimique de l'acide aminé. Les valeurs d'écart-types nécessaires à QUASI-2 sont prises aux statistiques issues de la BMRB. Les valeurs nous donneront un premier aperçu des attributions possibles.

Les statistiques sur la BMRB

Les statistiques menées sur la totalité des attributions déposées dans la BMRB donne la moyenne et l'écart-type des déplacements chimiques par atome et par acide aminé. Ces valeurs sont utilisées pour calculer le score. Cette base de données n'est cependant pas exempte d'erreurs et on y trouve des protéines ayant des centres paramagnétiques dont les déplacements chimiques peuvent être très différents de ceux de mêmes atomes dans d'autres protéines.

Une comparaison des deux tables est réalisée (cf Annexes).

3.3.2 Les programmes de prédiction de déplacements chimiques

SHIFTY

Basée sur l'homologie de séquence, l'approche de SHIFTY [Wishart *et al.*, 1997] est inspirée des logiciels utilisés en cristallographie pour assister la détermination expérimentale des structures de protéines [Sali *et al.*, 1990]. Ceci est bien-sûr dépendant de la présence ou non dans la base de données d'une protéine homologue à plus de 35% de la protéine cible (cf Annexes). Les prédictions de déplacements chimiques de SHIFTY sont obtenues sur le site : http ://redpoll.pharmacy.ualberta.ca/shifty/. On sélectionne le nature du noyau dont on souhaite les prédictions de déplacements chimiques ainsi que le format de sortie.

Comme valeurs pour les écarts-types nous utilisons les écart-types issus des statistiques sur la BMRB.

SHIFTS

SHIFTS [Xu Case, 2001] propose un algorithme à partir d'une base de données de déplacements chimiques de peptides calculés au niveau DFT (Density-Functional Theory) qui mène à la prédiction de déplacements chimiques des noyaux ^{15}N et ^{13}C à partir de la structure de la protéine. Accessible à l'adresse : http ://www.scripps.edu/mb/case/qshifts/qshifts.htm le programme SHIFTS est utilisable en ligne. Pour l'utiliser, deux paramètres sont nécessaires : un fichier pdb et les types d'atome dont on veut prédire les déplacements chimiques. De par sa façon de prédire les déplacements chimiques (cf Annexes), le logiciel présente la particularité de ne pas trouver de prédiction pour tous les résidus. Les références manquantes sont remplacées par les valeurs des déplacements chimiques dans la table Random Coil.

3.4 Le calcul d'une fonction cible

Afin de placer les fragments issus de QUASI-1 sur la séquence primaire de la protéine étudiée, il faut définir une fonction cible. Les caractéristiques de la fonction recherchée sont les suivantes : elle doit donner une bonne mesure de l'accord entre les paramètres expérimentaux et les paramètres de référence, elle doit favoriser l'accord local par rapport à l'accord global de façon à faciliter la détection d'erreur, elle doit être la plus discriminante possible. Plusieurs possibilités ont été évaluées.

3.4.1 Les différences de déplacements chimiques

La première possibilité est d'utiliser la valeur absolue des différences entre les déplacements chimiques observés et ceux de la référence choisie (Equation : 3.2). Dans ce cas, la meilleure possibilité correspond à la plus petite différence.

$$\begin{aligned} R_{i,j} &= |\delta C^{\alpha}_{iobs} - \delta C^{\alpha}_{jref}| + |\delta C^{\beta}_{iobs} - \delta C^{\beta}_{jref}| \\ S_{k,j} &= \sum_{i \in k} R_{i,j} \end{aligned} \tag{3.2}$$

où i est le numéro du pseudo-résidu étudié, j est la position sur la séquence et k est le numéro du fragment. Le score $S_{k,j}$ est tracé en fonction de la position j.

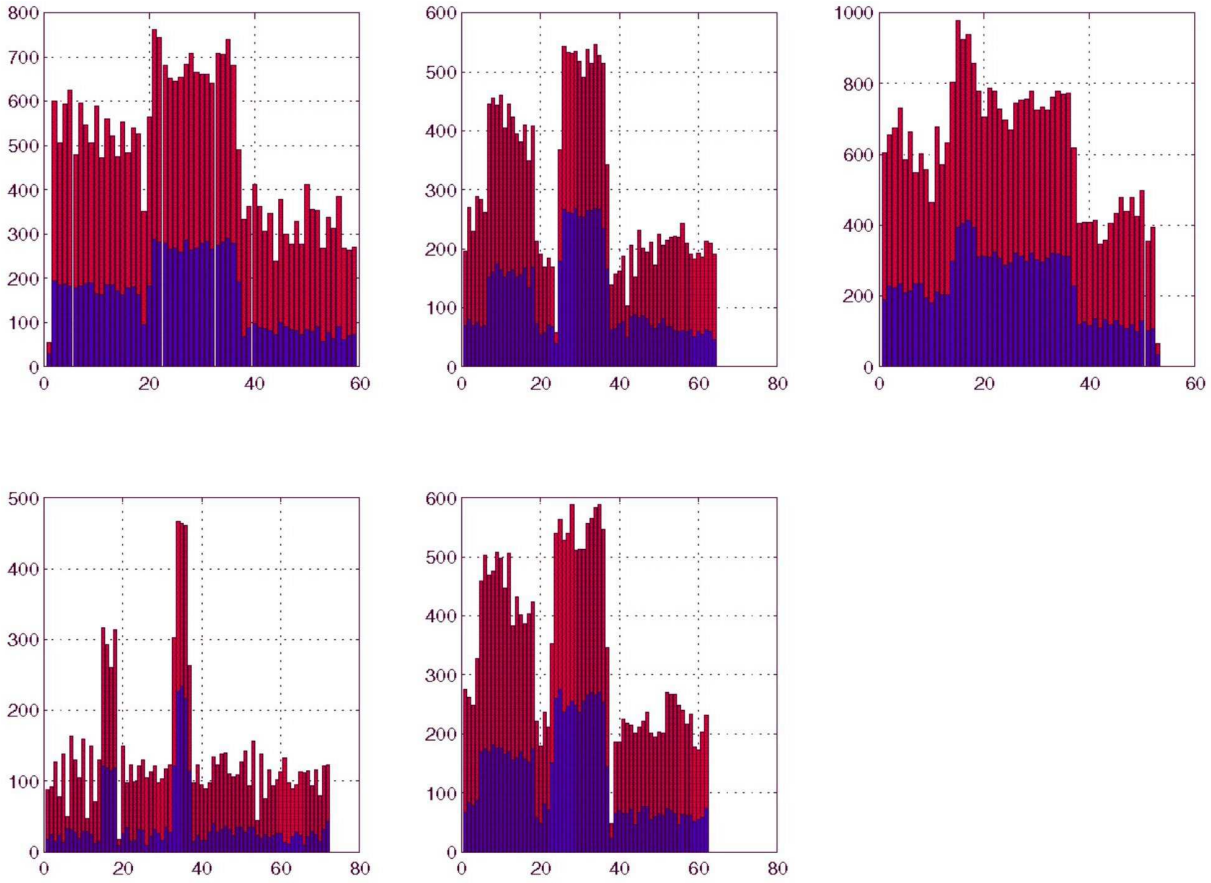

FIG. 3.9 – Représentation des scores $S_{k,j}$ obtenus pour les 5 fragments de l'ubiquitine. La partie bleue des barres verticales indique les sommes de différences observées sur les noyaux C^{α}, la partie rouge correspond aux sommes des différences sur les C^{β}.

Cette fonction cible présente un désavantage majeur : elle ne discrimine pas entre un accord global moyen et des pseudo-résidus qui, successivement, ont un bon accord puis un mauvais (Figure : 3.10).

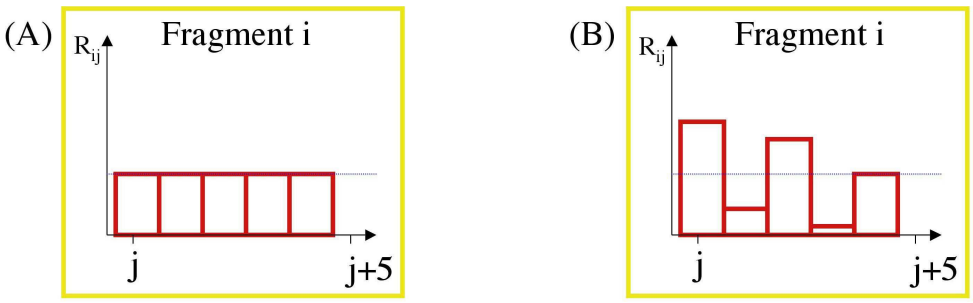

FIG. 3.10 – Présentation de scores $R_{i,j}$. Les cas présentés ici ont le même score $S_{k,j}$. Ils n'ont pourtant pas du tout le même profil pour ce qui est des différences de déplacement chimique au niveau des scores $R_{i,j}$. La ligne pointillée bleue indique la moyenne des scores. Le cas (A) présente des scores moyens pour les cinq pseudo-résidus du fragment i. Le cas (B) présente un mauvais score, suivi d'un bon score (deux fois de suite) puis un score moyen. La fonction présentée ici classe les deux propositions au même rang.

La fonction constituée de simples différences est capable de donner un accord global. Cet accord n'est pas suffisant dans le problème traité ici, car plus que de placer les fragments à partir d'un score global, nous voulons sélectionner les accords observables au niveau des pseudo-résidus.

3.4.2 Privilégier les bons scores consécutifs

Afin de différencier entre les deux cas schématisés dans le chapitre précédent (Figure : 3.10), la fonction cible est modifiée. La différence des déplacements chimiques calculée est exprimée sous la forme d'un écart par rapport à la moyenne des différences observées pour un résidu donné (Equation : 3.3). Sous cette forme, le score sera supérieur à 1 lorsque la différence des déplacements chimiques est plus petite que l'écart-type du noyau concerné dans le résidu référence. Le logarithme népérien de cette valeur, permet d'obtenir un score positif lorsque la cohérence est bonne et négatif lorsque les δ sont éloignés de façon statistiquement significative (Figure : 3.11).

$$
\begin{aligned}
R_{i,j} &= ln\frac{\sigma_\alpha}{|\delta C^\alpha_{iobs} - \delta C^\alpha_{jref}|} + ln\frac{\sigma_\beta}{|\delta C^\beta_{iobs} - \delta C^\beta_{jref}|} \\
S_{k,j} &= \sum_{i \in k} \delta_i R_{i,j} \begin{cases} \delta_i = \delta_i + 1 \text{ si } R_{i,j} \text{ et } R_{i-1,j} > 0 \\ \delta_i = 0 \text{ sinon} \end{cases}
\end{aligned}
\tag{3.3}
$$

Afin de favoriser des fragments qui font apparaitre des scores positifs consécutifs le facteur δ_i est incrémenté lorsque le score pour le résidu i et pour le précédent *i-1* sont positifs. Lorsque le score pour le résidu i est négatif, δ_i redevient nul.

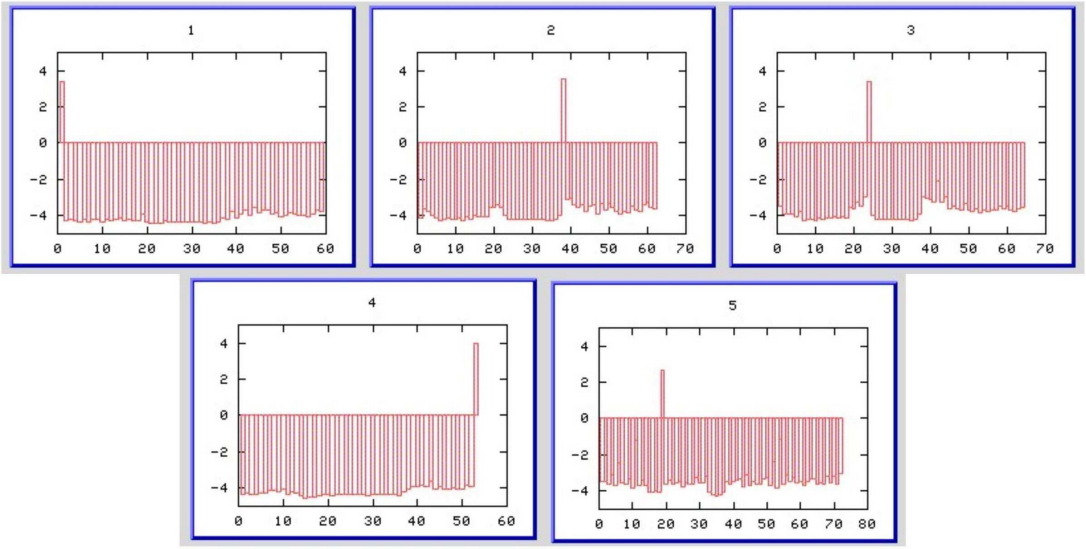

FIG. 3.11 – Présentation des histogrammes obtenus avec le score permettant de privilégier des bons scores consécutifs. Le résultat obtenu est satisfaisant : le score positif est la bonne position pour l'attribution.

Les résultats obtenus avec cette fonction cible ont fait l'objet d'une publication [Coutouly *et al.*, 2004]. Cet article est reproduit en Annexes.

Cette fonction de score donne donc de bons résultats, cependant, il souffre d'un inconvénient majeur : lorsque le score $R_{i,j}$ est composé d'un score élevé (bon accord) pour un noyau (C^{α}) et d'un mauvais accord (score négatif) pour l'autre (C^{β}) La valeur du logarithme népérien très loin de la référence n'est que légèrement négatif (Figure : 3.12).

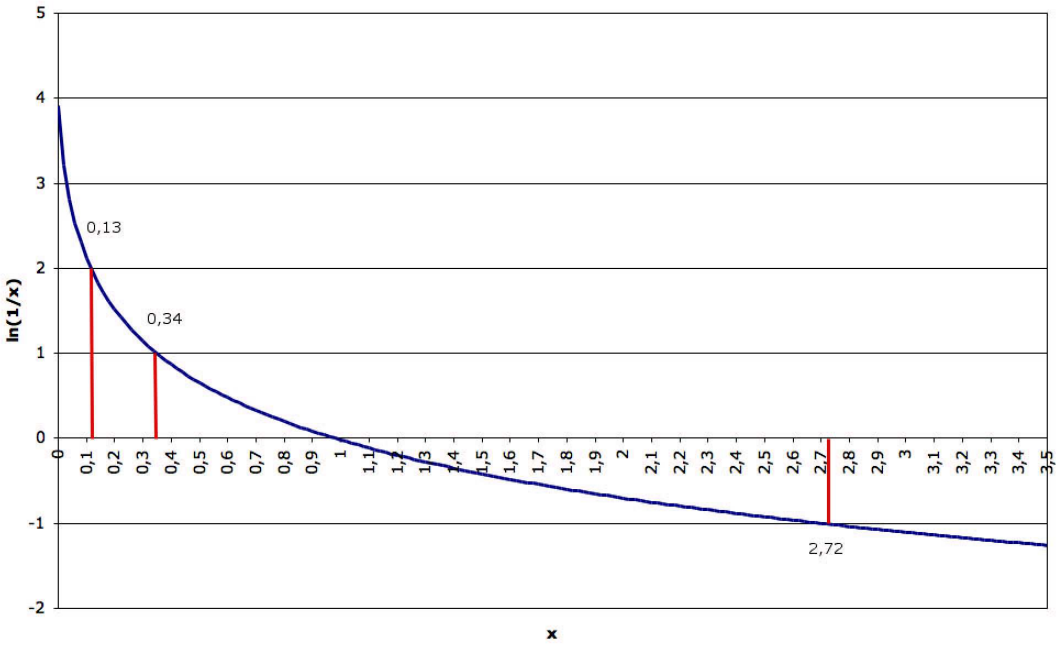

FIG. 3.12 – Représentation de la courbe ln 1/x, pour obtenir un score de -1, il faut une différence de 2,72 ppm. Pour avoir un score de 1, il faut une différence de 0,34 ppm. Pour obtenir un score de 2, il faut une différence de 0,13 ppm, alors que -2 est obtenu pour une différence supérieure à 6 ppm. L'équilibre entre les bons et les mauvais scores n'est pas satisfaisant.

3.4.3 Score basé sur le test du χ^2

L'utilisation d'une fonction cible basée sur le test statistique du χ^2 a également été évaluée. Les fonctions $R_{i,j}$ et $S_{k,j}$ utilisées sont de la forme suivante :

$$R_{i,j} = \frac{(\delta C^{\alpha}_{iobs} - \delta C^{\alpha}_{jref})^2}{\sigma^2 C^{\alpha}_{jref}} + \frac{(\delta C^{\beta}_{iobs} - \delta C^{\beta}_{jref})^2}{\sigma^2 C^{\beta}_{jref}}$$
$$S_{k,l} = \sum_{i \in k, j \in [l; l+N(k)]} R_{i,j} \tag{3.4}$$

où i est le numéro du pseudo-résidu, j est la position sur la séquence primaire et k est le numéro du fragment. La valeur de R peut-être utilisée pour avoir une mesure statistique du choix grâce à un test du χ^2.

Ce test est utilisé, entre autre, lorsque l'on veut comparer une distribution empirique avec une distribution théorique. Le score utilisé est défini comme la distance χ^2. En fonction de cette distance, on trouve la probabilité que l'évènement est compatible avec le modèle. Les variables qui permettent d'utiliser ce test sont l'intervalle de confiance auquel l'utilisateur veut faire son test et le nombre de degrés de liberté. Ce dernier paramètre influe sur la forme de la densité de probabilité (Figure : 3.13). Plus il augmente, plus la densité s'éloigne de l'axe d'ordonnée. L'intervalle de confiance correspond à l'aire sous la courbe que l'on veut considérer, plus il est petit, plus on considère de surface et moins le test sera stringent.

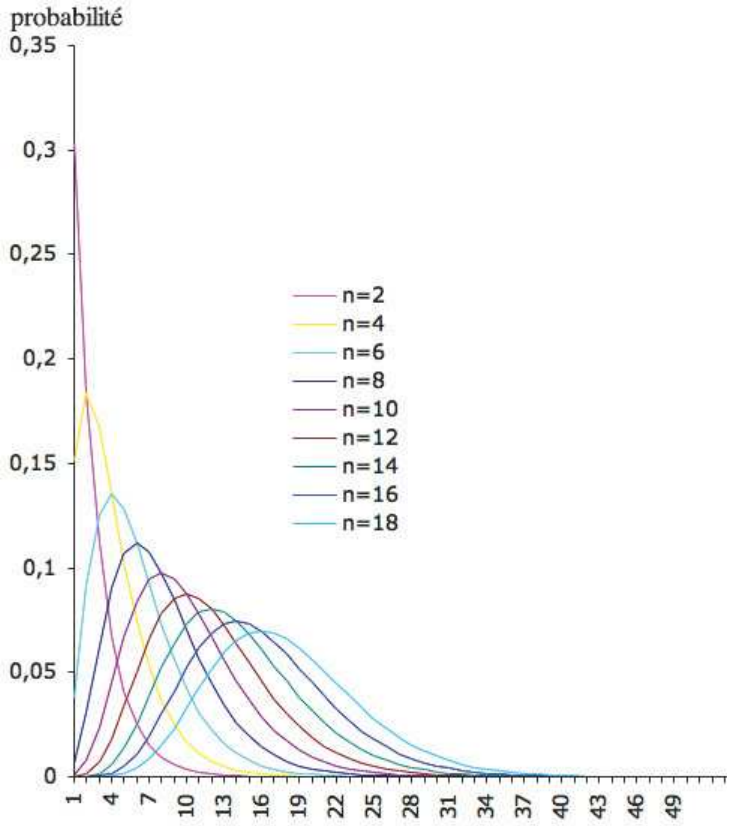

FIG. 3.13 – Densité de probabilité pour différents degrés de liberté (n). Lorsque n augmente, le maximum de la courbe s'éloigne de l'axe d'ordonnées et l'écart-type s'élargit.

Avec ces paramètres, on peut rechercher la valeur critique (V_c). Cette valeur marque le seuil d'acceptation du test. En deçà, l'hypothèse que l'évènement suive le modèle est acceptée ; au-delà, l'hypothèse est rejetée. Les valeurs critiques sont tabulées (cf Annexes). Dans QUASI-2, le test du χ^2 décide si la position est acceptable. Le recours à ce test

permet de ne pas laisser de côté la bonne position qui, pour une raison ou pour une autre, n'a pas le meilleur score.

Le score $S_{k,j}$, qui est la somme des distances $R_{i,j}$, est comparé à la valeur critique du χ^2 à un nombre de degrés de liberté égal au nombre de noyaux C^α et C^β figurant dans le fragment. On retranche le nombre de glycines, car elles n'ont pas de C^β.

Lorsque le score est inférieur à V_c la position est acceptée pour le fragment étudié. Ceci nous conduit, quelques fois à avoir plusieurs positions acceptées pour un même fragment. Suivant la taille des fragments et la référence utilisée il peut donc y avoir, au final, plusieurs possibilités.

3.5 Influence de la longueur des fragments

Les fragments obtenus à l'issue de QUASI-1 sont des enchaînements de pseudo-résidus. Ils ont des tailles variables. Intuitivement, si QUASI cherche à placer un fragment de deux pseudo-résidus sur la séquence, il lui sera aisé de trouver plusieurs paires de résidus dont les déplacements chimiques coïncident avec ceux du fragment. La probabilité de bien placer un fragment si petit est très faible. Afin d'étudier l'influence de la taille du fragment sur le placement de ces derniers, une étude systématique a été menée sur des fragments formés de 2 à 15 pseudo-résidus de l'ubiquitine. Les fragments issus de l'utilisation de QUASI-1 sur l'ubiquitine ont été découpés puis placés sur la séquence par QUASI-2. Tous les scores obtenus ont été enregistrés. Les courbes suivantes (Figure : 3.14) représentent les distributions des bonnes et des mauvaises positions. L'axe des abscisses a comme unité l'écart-type de la distribution des mauvaises positions.

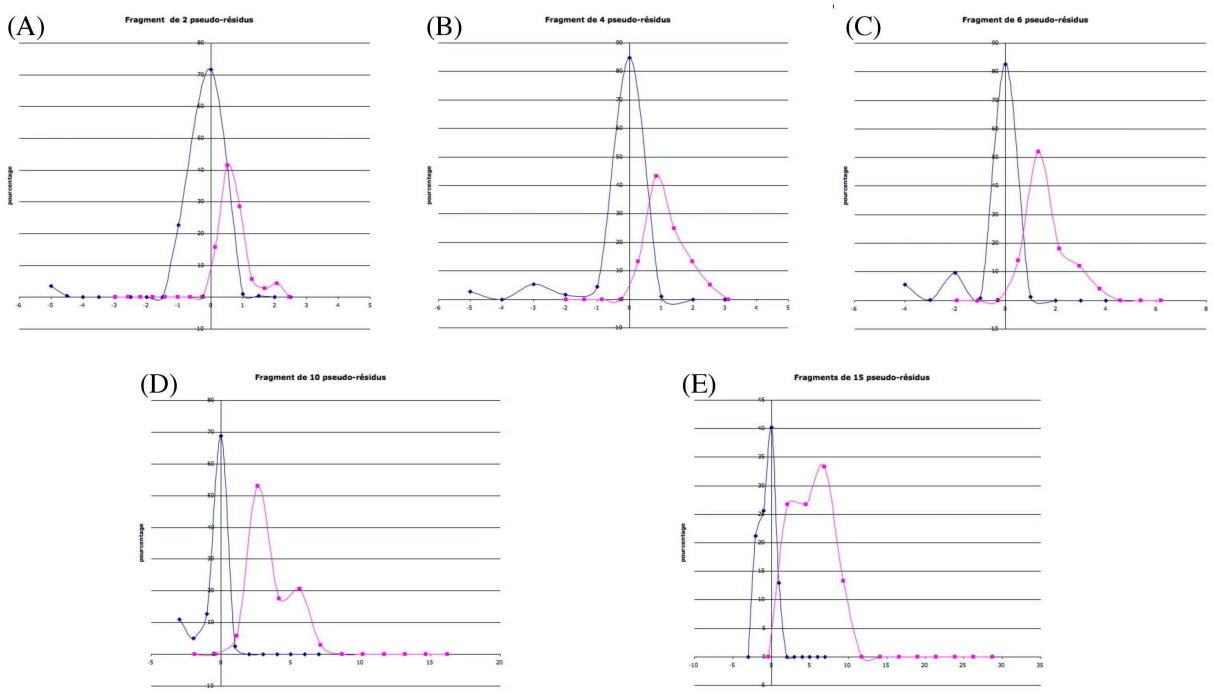

FIG. 3.14 – (A) cas des fragments de 2 pseudo-résidus. La moyenne des bonnes positions se trouve à 0,52 écart-type de la moyenne des mauvaises positions. Les distributions sont confondues.(B) cas des fragments de 4 pseudo-résidus. La moyenne des bonnes positions se trouve à 0,84 écart-type de la moyenne des mauvaises positions.(C) cas des fragments de 6 pseudo-résidus. La moyenne des bonnes positions se trouve à 1,32 écart-type de la moyenne des mauvaises positions.(D) cas des fragments de 10 pseudo-résidus. La moyenne des bonnes positions se trouve à 2,62 écart-type de la moyenne des mauvaises positions.(E) cas des fragments de 15 pseudo-résidus. La moyenne des bonnes positions se trouve à 4,49 écart-type de la moyenne des mauvaises positions.

De ces études taille par taille, on tire l'évolution (Figure : 3.15) de la différence entre les moyennes en fonction de la taille des fragments.

Cette étude nous montre que la taille du fragment à placer est un facteur important dans la confiance que l'on peut apporter aux positions trouvées par QUASI-2. Plus le fragment est important plus la probabilité qu'il soit bien placé est importante. Dès 5 pseudo-résidus enchaînés, il y a plus de 68% de chance que la position acceptée par QUASI-2 soit bonne. Au delà de 11 pseudo-résidus, la probabilité est de plus de 99,7%. C'est pourquoi, QUASI-2 place les fragments par ordre décroissant de taille. Le plus gros fragment est placé en premier, puis les plus petits. Une position occupée par un fragment est inactivée pour les fragments de taille inférieure.

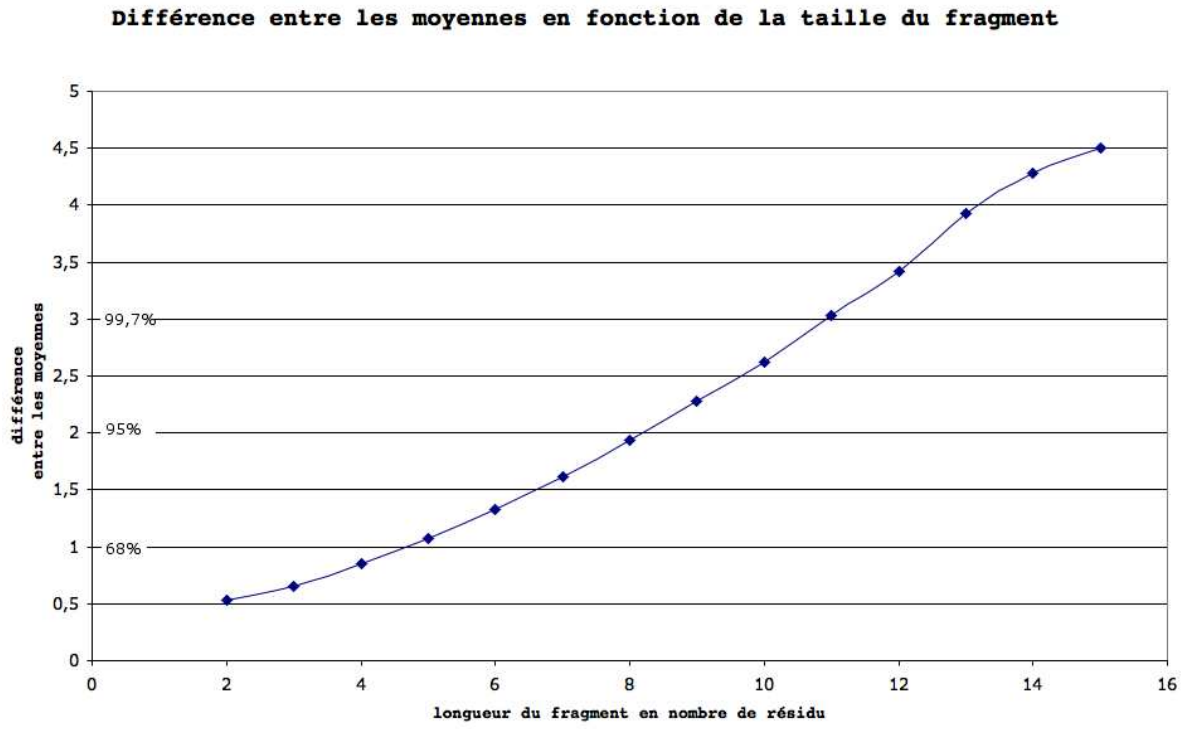

FIG. 3.15 – Evolution de l'écart entre les mauvaises et les bonnes distributions en fonction de la taille du fragment.

3.6 L'application de QUASI-2 à l'ubiquitine

A l'issue de QUASI-1, on a 5 fragments de tailles allant de 24 à 5 résidus (cf Chapitre 2.3.2), les références à notre disposition sont la table des valeurs Random Coil, les statistiques sur la BMRB et les prédictions de SHIFTS et SHIFTY. Vu le petit nombre de fragments et la taille modeste de la protéine (76 résidus), nous pouvons utiliser toutes ces méthodes en même temps afin d'en comparer les résultats. Dans ce cas, chaque fragment a une unique position acceptée par le critère de la valeur critique du test du χ^2, ce qui rend l'attribution plus aisée.

Le fragment numéro 2 se place en position 2, suivi du numéro 5 en position 20, puis le $4^{ème}$ en position 25, le numéro 3 en 39 et enfin le plus grand se place en position 54. L'attribution est faite pour 75 des 76 résidus (présence d'une séquence PP) et est conforme à l'attribution publiée.

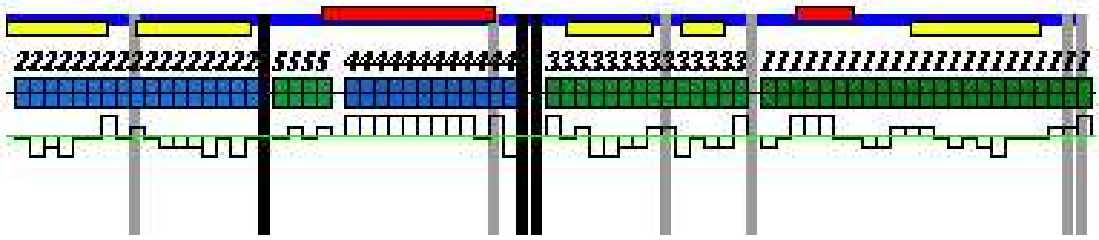

FIG. 3.16 – Présentation du placement des 5 fragments de l'ubiquitine. Les résultats obtenus en utilisant les références RC, SHIFTY, les statistiques issues de la BMRB et SHIFTS se recoupent. La ligne bleue symbolise la séquence de la protéine cible. Dans les cas où la structure secondaire est connue, on représente les hélices-α par des barres rouges et les feuillets β par des barres jaunes. Les barres verticales indiquent les prolines (noires) et les glycines (grises). Chaque pseudo-résidu (avec un NH) placé se trouve sous la forme d'un rectangle surmonté du numéro du fragment dont il fait partie.

4

Application de QUASI sur l'α-actinine EF34

Sommaire

Le cas de l'ubiquitine a permis une première validation de QUASI. Mais l'application n'en est pas toujours aussi aisée. La seconde protéine étudiée est l'α-actinine EF-34. Cet exemple soulève de nouveaux problèmes tels que la gestion d'erreurs éventuelles effectuées par QUASI-1.

4.1 α-actinine

L'α-actinine est présente dans les cellules musculaires et non musculaires. Plusieurs isoformes d'α-actinine sont produites par 3 gènes différents à la suite d'un épissage alternatif. Elle fait partie d'une grande famille de molécules, qui sont les protéines d'ancrage ou de pontage de l'actine. Dans les cellules non musculaires, l'α-actinine rassemble les filaments d'actine en faisceaux ou câbles tels que les fibres de stress. Dans le muscle strié, elle est la composante principale de la strie Z où elle lie, entre autres, les filaments d'actine. Chaque molécule d'α-actinine est composée de trois régions bien distinctes [Blanchard *et al.*, 1989]. Les résidus 1-245 contiennent trois sites de liaison à l'actine du ABS (Actin Binding Site), cette région est conservée dans les divers membres de la famille des ABPs (Actin Binding Proteins). Elle est suivie de quatre unités répétitives, de 120

acides aminés chacune, homologues aux domaines intermédiaires de la spectrine et organisées en structure "coiled-coil". La troisième région correspond à l'extrémité C-terminale qui contient deux domaines de type "EF-Hands". Mais cette protéine a perdu ses propriétés de ligation au calcium car les résidus de cette dernière région ont évolué de façon différente aux autres "EF-Hands" (Figure : 4.1).

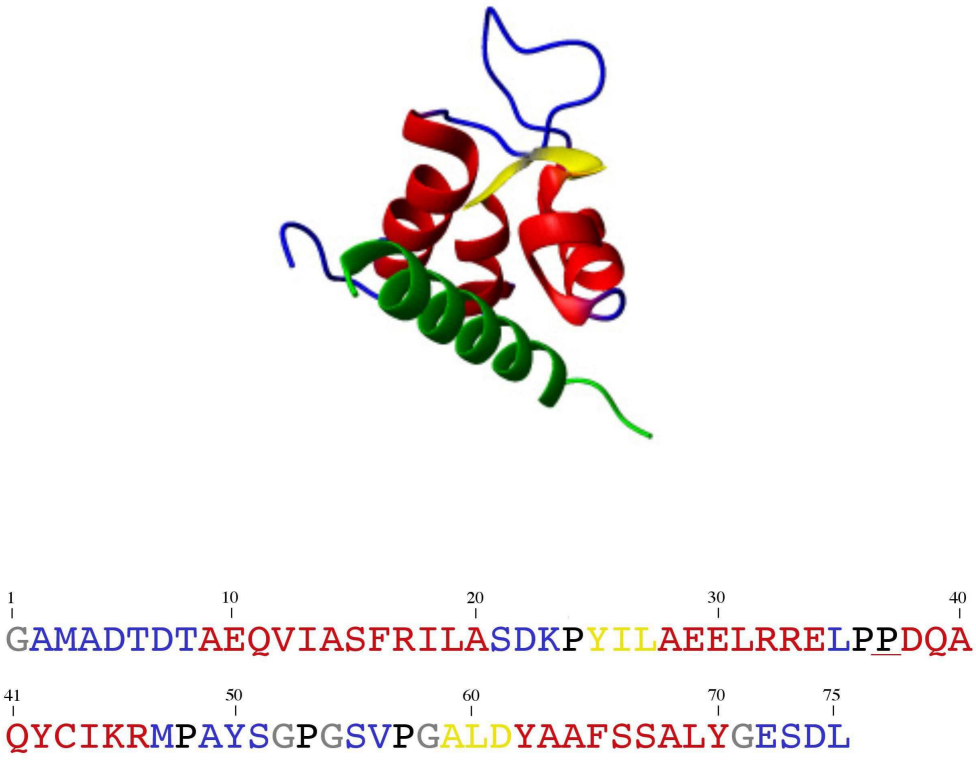

FIG. 4.1 – Représentation en ruban de la structure du complexe α-actinine-EF34 titine Zr7 et séquence primaire de la protéine. Les prolines sont présentées en noir et les glycines en gris. Les hélices-α sont en rouge et les brins-β en jaune. La titine Zr7 est représentée en vert.

Les longueurs attendues pour les fragments sont de 23, 12, 11, 5, 4 et 19 résidus. La structure du complexe de l'α-actinine-2 Act-EF34 avec Zr7 de la titine a été publiée récemment [Atkinson *et al.*, 2001].

4.1.1 Application de QUASI-1

Dans l'étude de cette protéine de 75 résidus, ont été utilisés uniquement les spectres donnant les informations sur les atomes $^{13}C^{\alpha}$ et les $^{13}C^{\beta}$ de la chaîne principale de la protéine.

Spectre	Nombre de pics
HN(CO)CA	63
HNCA	63
CBCA(CO)NH	60
CBCANH	58

TAB. 4.1 – Nombre de pics dans chaque fichier utilisé par QUASI. Le nombre de corrélations sélectionnées dans l'HSQC est 69.

Le nombre de pseudo-résidus est de 63 (Tableau : 4.1). La séquence primaire contient 6 prolines qui forment à priori 6 fragments.

Pic concerné	Au moins une différence au dessus du seuil	Gly	Décision
1	√		Lier au #8
7	√		Coupure
11	√		Coupure
14	√		Coupure
18	√		Coupure
20	√		Accepter la proposition
22		√	Accepter la proposition
33	√		Accepter la proposition
41		√	Coupure
46		√	Coupure
50	√		Lier au #60
61	√		Coupure
63	√		Coupure

TAB. 4.2 – Demandes d'intervention dans la cas de l'α-actinine.

L'utilisateur est invité 13 fois à inspecter les spectres (Tableau : 4.2).

Parmi ces interventions, il y a 2 cas différents. Le premier est qu'il y a au moins une différence de déplacements chimiques au-dessus du seuil d'acceptation (Figure : 4.2).

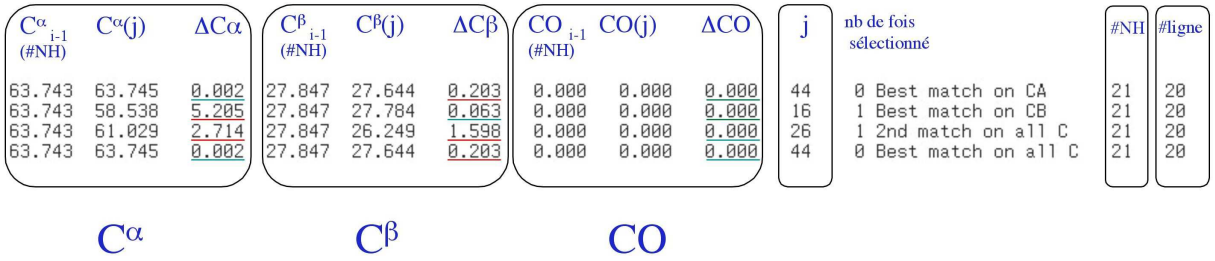

FIG. 4.2 – La différence des déplacements chimiques des C^{β} est juste au-dessus de la marge de 0,2 ppm. Le choix ne se fait qu'après avoir inspecté les spectres.

Le second cas est que les décisions doivent se prendre avec un unique noyau (C^{α}) (Figure : 4.3).

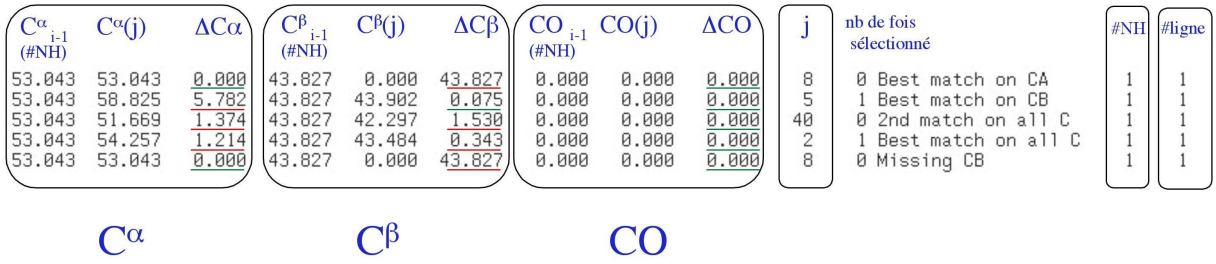

FIG. 4.3 – Le pseudo-résidu 1 trouve un bon candidat comme précédent dans les déplacements chimiques des C^{α} (8) et dans les déplacements chimiques des C^{β} (5), ce dernier n'est pas bon pour les déplacements chimiques des C^{α}. L'inspection des spectres méne à accepter la connexion avec le pseudo-résidu 8 qui n'a pas de déplacement chimique C^{β}. Il est recouvert par le pic C_i^{β}.

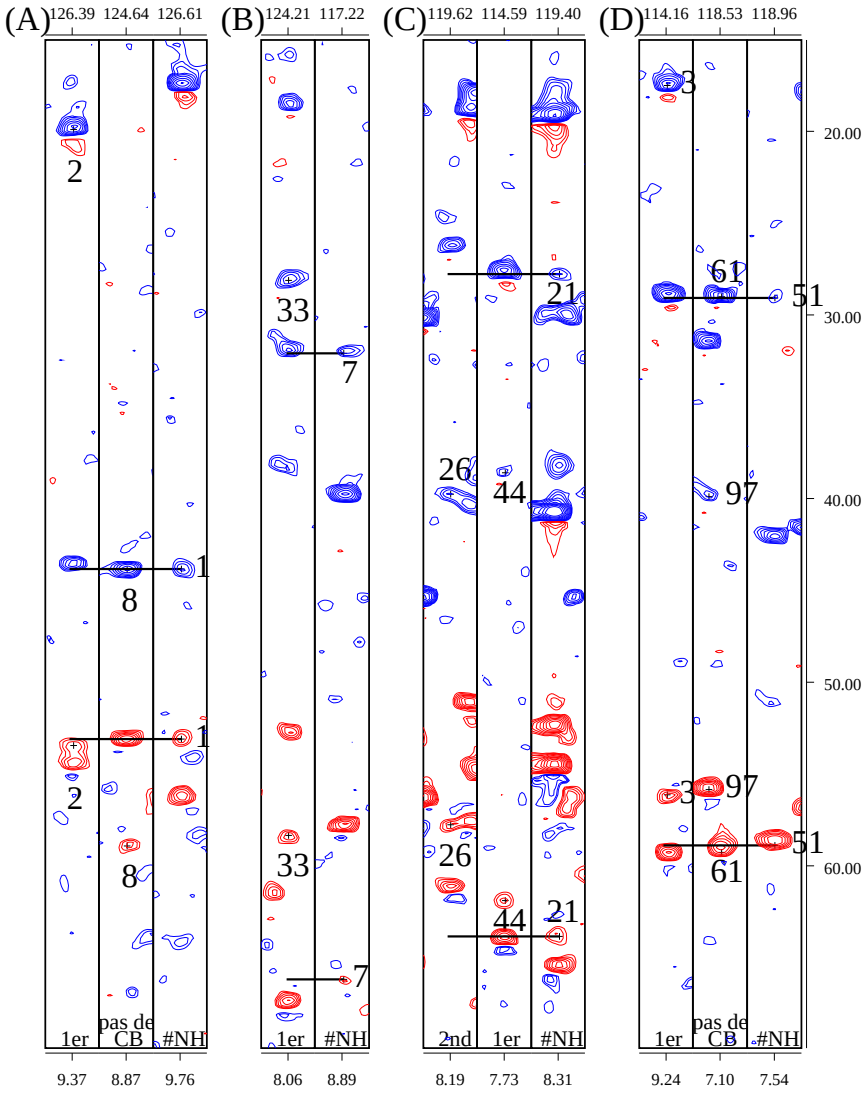

FIG. 4.4 – Extrait du spectre HNCACB de l'α-actinine. (A) Le pseudo-résidu 1 (Figure : 4.3) a comme meilleur précédent le pseudo-résidu 2, mais le pic du C^{α} est très éloigné. Le pseudo-résidu 8 est un très bon candidat au niveau du C^{α} mais il n'a pas de déplacement chimique C_i^{β}. QUASI-2 invite l'utilisateur à inspecter les spectres. La connexion est choisie avec le pseudo-résidu 8. L'utilisateur décide que les pics C_i^{β} et C_{i-1}^{β} se recouvrent. Le cas (B) est une situation où les déplacements chimiques des C_{i-1}^{α} et C_{i-1}^{β} suggèrent que le pseudo-résidu qui précède est une proline. (C) La meilleure possibilité est le pseudo-résidu 44 (Illustration 4.2). La seconde possibilité (26) est éloignée pour les deux noyaux. La connexion est acceptée entre 44 et 21. (D) Le pseudo-résidu 51 a comme meilleure proposition le pseudo-résidu 3 mais l'accord n'est pas bon. Le pseudo-résidu 61 n'a pas de déplacement chimique C_i^{β} identifié. La connexion est choisie avec le pseudo-résidu 61.

On obtient 8 fragments de taille allant de 18 à 2 résidus : 18, 13, 11, 11, 10, 3, 3 et 2 résidus.

4.1.2 Application de QUASI-2

Le placement des 8 fragments de l'α-actinine sur sa séquence pose de nouveaux problèmes qui n'ont pas été rencontrés dans le cas de l'ubiquitine. Les fragments les plus grands se placent de façon unique en bloquant les positions qu'ils occupent, mais dans le cas des petits fragments, il existe plus d'une position acceptée et ceci multiplie le nombre de propositions faites par QUASI-2. De plus, le résultat dépend du type de table de référence utilisé.

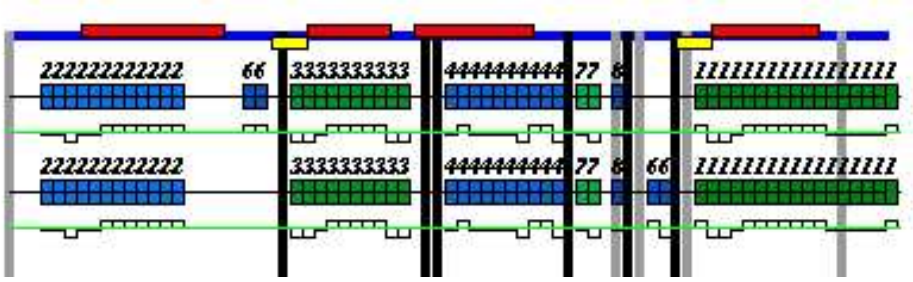

FIG. 4.5 – Présentation des placements des 8 fragments de l'α-actinine avec les prédictions de SHIFTS et les valeurs moyennes de la BMRB. Les deux méthodes donnent deux possibilités différentes pour le placement du fragment 6. De plus, elles ne placent pas du tout le fragment 5.

Un consensus est trouvé avec les prédictions de SHIFTS et l'utilisation des statistiques de la BMRB. Avec aucune de ces deux méthodes, on ne parvient à placer le fragment numéro 5, alors que c'est un fragment de taille conséquente puisqu'il renferme 10 résidus. Par contre, tous les autres fragments trouvent au moins un endroit où leurs déplacements chimiques coïncident avec la référence. Le fragment 6 sélectionne deux positions dans la séquence.

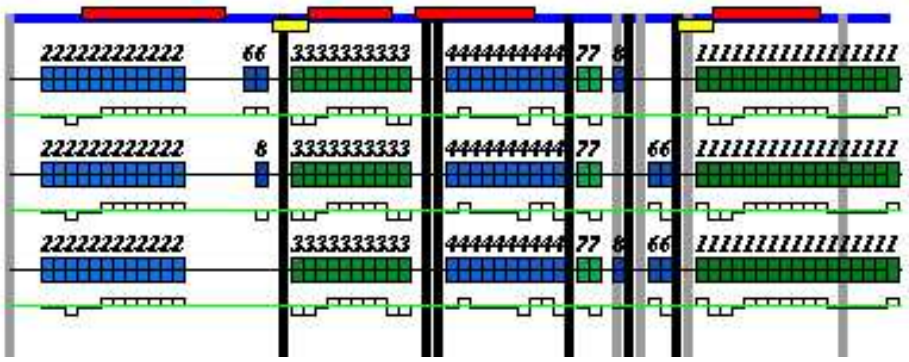

FIG. 4.6 – Présentation des placements des 8 fragments de l'α-actinine avec la table Random Coil. La multiplication des possibilités s'explique car les fragments 6 et 8 présentent un placement multiple.

Avec l'utilisation de la table des valeurs Random Coil (Figure : 4.6), les fragments numéro 6 et 8 se placent chacun à deux endroits de la séquence. Ces placements multiples expliquent qu'il y ait un plus grand nombre de possibilités dans ce cas. En comparant avec les possibilités trouvées par SHIFTS et les statistiques de la BMRB, on peut noter que les positions du fragment 6 coïncident ; une des positions du fragment 8 se retrouve également dans les deux représentations. Avec la dernière méthode (SHIFTY), le fragment

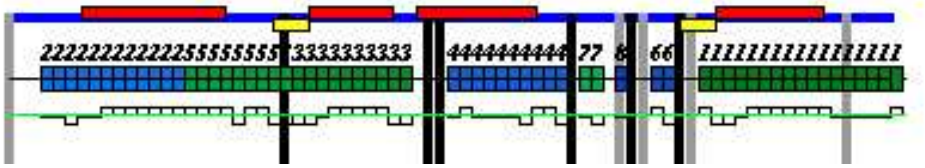

FIG. 4.7 – Présentation des placements des 8 fragments de l'α-actinine avec les prédictions de SHIFTY. Elle place tous les fragments de façon unique.

numéro 5 est placé (contrairement aux trois autres méthodes) et chaque fragment a une unique position (Figure : 4.7). Nous voyons que le dernier pseudo-résidu du fragment 5 se place sur une proline. Le fait de trouver un pseudo-résidu sur un résidu du type proline n'est pas correct. Ce problème peut être résolu en regardant les histogrammes traçant les différences de déplacements chimiques par résidu et par position. Ce type d'histogramme est consultable à chaque instant d'utilisation de QUASI-2. Après avoir recherché le meilleur tracé, c'est à dire le plus petit, parmi les histogrammes du fragment numéro 5, nous étudions la position numéro 15 (Figure : 4.8) qui présente de petites différences pour les 9 premiers pseudo-résidus et une grosse différence pour le dernier pseudo-résidu (qui se place alors sur la proline).

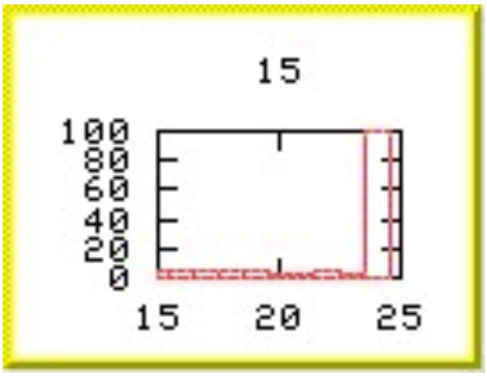

FIG. 4.8 – Histogramme représentant les différences de déplacements chimiques entre les déplacements chimiques observés et ceux qui sont prédits par SHIFTY. Ici le fragment numéro 5 en position 15.

Lorsque nous remontons dans QUASI-1 (Figure : 4.9) puis dans les spectres (Figure : 4.10), nous remarquons qu'effectivement cette connexion est faite de façon automatique par QUASI 1. Le pseudo-résidu accepté à ces deux déplacements chimiques dans la marge

C^{α}_{i-1} (#NH)	C^{α}(j)	$\Delta C\alpha$	C^{β}_{i-1} (#NH)	C^{β}(j)	$\Delta C\beta$	CO_{i-1} (#NH)	CO(j)	ΔCO	j	nb de fois sélectionné	#NH	#ligne
55.750	55.726	0.024	31.322	32.459	1.137	0.000	0.000	0.000	35	0 Best match on CA	40	39
55.750	55.631	0.119	31.322	31.412	0.090	0.000	0.000	0.000	59	1 Best match on CB	40	39
55.750	55.599	0.151	31.322	32.040	0.718	0.000	0.000	0.000	58	1 2nd match on all C	40	39
55.750	55.631	0.119	31.322	31.412	0.090	0.000	0.000	0.000	59	1 Best match on all C	40	39
55.750	55.822	0.072	31.322	0.000	31.322	0.000	0.000	0.000	48	0 Missing CB	40	39

C^{α} C^{β} CO

FIG. 4.9 – QUASI-1 accepte automatiquement la connexion entre les pseudo-résidus 40 et 59. Cette décision est justifiée par les très bons accords observés entre ces deux pseudo-résidus.

d'acceptation du pseudo-résidu 40. Un autre pseudo-résidu est très proche en déplacement chimique C^{α} mais il n'a pas de déplacement chimique C^{β}. QUASI accepte alors la connexion avec le pseudo-résidu le plus complet.

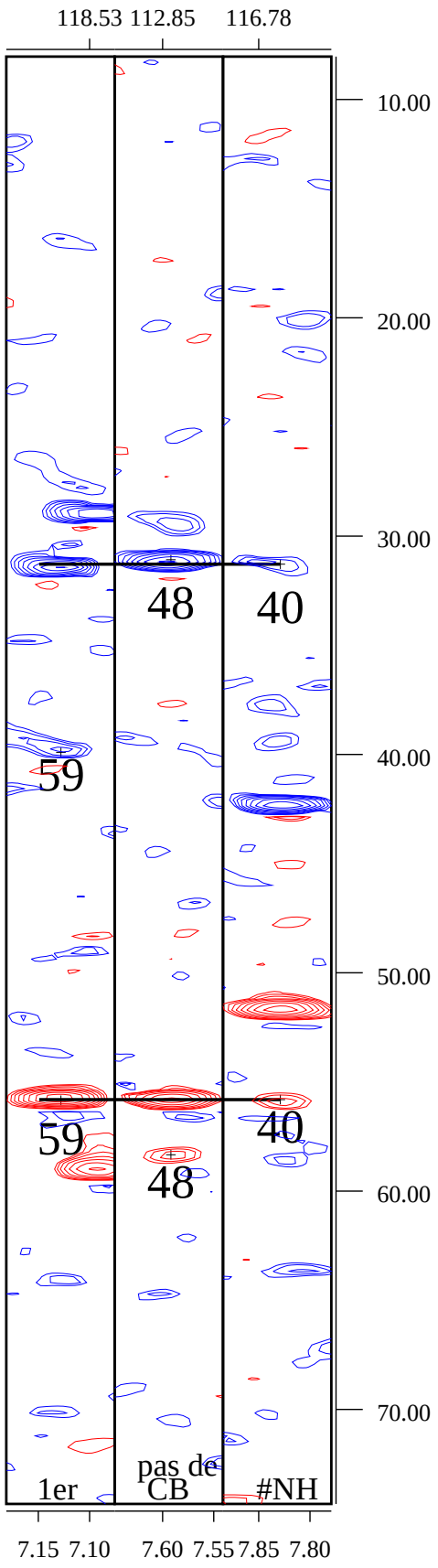

FIG. 4.10 – Lorsque l'on revient aux spectres HNCACB, on voit que les pseudo-résidus 48 et 59 sont tous les deux de bons candidats. Le pseudo-résidu 48 n'a pas de pic C^{β} identifié et QUASI privilégie donc le pseudo-résidu 59.

Il nous est possible de couper ce fragment au niveau du dernier résidu et de tout relancer. Lorsque ceci est fait, un accord parfait est obtenu pour les différentes méthodes (Figure : 4.11). Le fragment 5 se place en position 15 et le résidu qui constitue le fragment 10 est placé à la fin du $3^{ème}$ fragment. Le fait de couper le fragment 5 permet son placement.

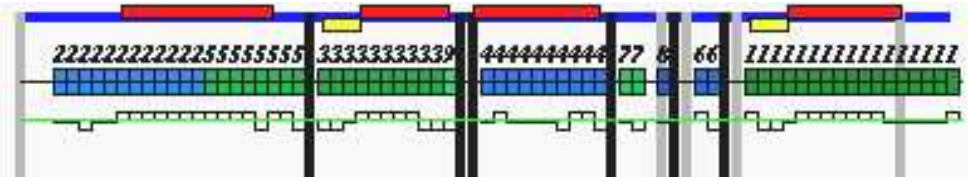

FIG. 4.11 – Placement des 10 fragments de l'α-actinine EF34 après avoir scindé le fragment 5 en deux.

En ancrant ces 9 pseudo-résidus de la position 15 à 24, la mauvaise position qui était privilégiée par le fragment 6 (position 14) n'est plus accessible. Il se place alors à la bonne place selon l'attribution publiée par Atkinson et al. [Atkinson *et al.*, 2000]. Cet exemple nous montre que QUASI-2 peut nous aider à détecter des erreurs commises plus tôt dans QUASI-1. L'utilisateur peut valider tous les choix faits par QUASI : il en a le contrôle complet.

5

Les caractéristiques de QUASI

Sommaire

Une fois les spectres acquis, QUASI regroupe les informations que l'on peut obtenir à partir des spectres RMN. Il présente l'avantage de tout présenter dans un "tableau de bord". Cette approche graphique permet une utilisation rapide et conviviale du programme. Outre cette particularité de présentation, QUASI n'attribue jamais les pseudo-résidus à un type d'acide aminé, tout le processus se déroule en terme de fragments et de position numérique sur la séquence. Enfin, une erreur dans la cascade de décisions prises par QUASI est identifiable et peut-être corrigée sans modifier les données d'entrée, et sans quitter le programme.

5.1 Les innovations

5.1.1 Pas d'identification en terme d'acide aminé

Contrairement à tous les programmes présentés dans la première partie, QUASI n'utilise pas d'étape explicite d'identification entre le pseudo-résidu traité et le type d'acide aminé. Chaque pseudo-résidu est lié à une position numérique dans la séquence. La correspondance avec le type d'acide aminé peut-être aisément réalisée à l'issue du processus.

5.1.2 L'interface graphique

Le tableau de bord utilisé par QUASI permet de regrouper de nombreuses informations, celles ci pouvant éventuellement être visualisées dans différents logiciels. La centralisation de ces données permet une vue globale du problème, l'attribution peut alors se faire en fonction de toutes ces données. Cette sortie graphique permet aussi bien d'évaluer tout de suite la qualité du résultat obtenu que de traiter les données et de les modifier le cas échéant.

5.1.3 La correction des erreurs faites au préalable

QUASI permet de détecter les erreurs réalisées lors du processus. Ceci permet à ce programme de donner une attribution avec 100% de confiance (quitte à ne pas attribuer tous les pseudo-résidus). Cette propriété fait de QUASI un logiciel utile lors de l'étape d'attribution. En fonction de la qualité des données disponibles, il donne une attribution complète (jeu de données complet) ou une attribution partielle (jeu de données morcelé). Le précautions prises lors de l'utilisation de QUASI permettent un taux d'erreur, c'est à dire un nombre de mauvaise attribution, nul. Les embûches au niveau de la mise en place d'un score à la fois satisfaisant sur les plans statistique et pratique ont été explicitées lors de l'introduction de cette partie. Elles ont d'abord été liées à un souci d'obtenir tout de suite la meilleure position, celle que l'oeil humain aurait sélectionnée en favorisant les fragments dont les scores internes sont bons de façon consécutive. En voulant obtenir ce résultat, nous avons rencontré une autre source d'erreurs : l'asymétrie importante entre les scores positifs et les scores négatifs. Après avoir essayé de corriger celle-ci avec des facteurs divers, qui se sont avérés changeant en fonction des propriétés de la protéine traitée, nous nous sommes arrêtés sur un score fondé largement sur la statistique. Ce dernier score nous offre la possibilité de comparer nos résultats à des choses connues en statistiques : en effet le score obtenu, qui est comparé à la valeur critique du χ_2 est significatif de la distance existante entre l'expérience et notre référence.

5.1.4 Les points faibles de QUASI

La principale difficulté rencontrée lors de cette étape reste la sélection des pics. En effet, malgré l'utilisation du spectre triple-résonance HN(CO)CA, la superposition de signaux risque d'être à l'origine d'une perte d'informations. Ceci peut entraîner des erreurs de connexions séquentielles ou même intrarésiduelles : on peut avoir un pic dans l'HN(CO)CA qu'on ne peut pas distinguer et par la suite le lier à la mauvaise paire de $^{1}H/^{15}N$, en outre on risque d'observer un plus grand nombre de fragments de plus petites tailles et ceci complique d'autant l'étape suivante. De plus, la reconnaissance de pics, surtout lorsque les spectres sont très bruités, peut conduire l'utilisateur à sélectionner des artefacts qui risquent de fausser les connexions acceptées par QUASI-1. Lorsque cette étape est bien réalisée, s'il existe bien un et un seul pic sous chaque sélection, cette première partie se fait facilement.
Nous avons décidé de n'accepter aucune connexion automatique lorsque la connexion ne doit se décider que sur un type d'atome. Il nous a paru dangereux d'accepter une telle connexion car elle nous a paru avoir une grande probabilité de fausser de façon importante la suite du processus. Cette décision est contraire à celle prise par les concepteurs de PACES [Coggins Zhou, 2003] qui ont pris le parti de rechercher de façon exhaustive les possibilités d'attribution, ceci passant par le fait de créer tous les pseudo-résidus envisageables et ambigus ainsi que toutes les attributions qui vont avec. Pour QUASI nous avons pris le parti d'utiliser les données les plus propres possibles et nous ne pouvions ainsi courir le risque d'introduire des erreurs de façon automatique par ce biais.

5.2 Incorporation de données structurales dans la fonction cible

Dans l'avenir, QUASI devra offrir la possibilité d'incorporer des informations structurales supplémentaires dans la fonction cible. Comme mentionné dans le chapitre 3.1, QUASI n'offre pour l'instant que la visualisation des informations apportées par le CSI ou par les valeurs de relaxation.

Ce projet nécessite une étude approfondie de la pondération relative des différentes informations, et ceci pour chaque combinaison possible.

5.2.1 CSI

Cet index, de par sa structure reste relativement général, c'est à dire que pour deux valeurs très différentes, les index peuvent être les mêmes que pour des valeurs très proches. Il ne doit donc pas avoir un rôle prépondérant dans la fonction cible. Dans le cas où la position n'est pas la bonne, l'index peut facilement indiquer une structure secondaire sans raison. Si cette information a une pondération importante, elle risque de trop atténuer les autres informations.

5.2.2 Les données de relaxation

L'intégration des données de relaxation dans la fonction cible est encore plus problématique. Il faut trouver la bonne manière de quantifier les différences entre les mesures effectuées. La présence d'une structure secondaire ne se traduit pas par les mêmes valeurs du paramètre d'ordre d'une protéine à l'autre. Il serait vraisemblablement plus efficace d'utiliser des écarts avec la valeur moyenne, la valeur minimale ou la valeur maximale. Une fois cette quantification faite, se pose la problème de la pondération. Doit-on donner plus de poids à ces données qu'à l'index de CSI ?

5.2.3 Les contraintes dipolaires résiduelles

En rendant légèrement anisotrope le mouvement de réorientation de la protéine dans le champ magnétique, certaines informations supplémentaires, tels que les couplages dipolaires résiduels D_{ij} (Equation : 5.1) (usuellement moyennés à zéro en milieu isotrope par le mouvement Brownien) apparaissent sur le spectre.

$$D_{ij} = -S\frac{\gamma_i\gamma_j\mu_0 h}{16\pi^3 r_{ij}^3}\left[A_a(3\cos^2\theta - 1) + \frac{3}{2}A_r \sin^2\theta\cos 2\phi\right] \tag{5.1}$$

Expression du couplage dipolaire résiduel entre les atomes i et j : D_{ij}. A_a et A_r sont les composantes axiale et rhombique du tenseur d'alignement ; $\gamma_i\gamma_j$ sont les rapports gyromagnétiques nucléaires ; $\{\theta, \phi\}$ est le vecteur d'orientation relative par rapport à ce tenseur, r_{ij} est la distance internucléaire et S le paramètre d'ordre local. Leurs amplitudes

dépendent directement de l'orientation polaire de chaque paire de noyaux dans ce repère. En utilisant des milieux très dilués, on conserve un taux d'orientation très faible, donc des couplages dipolaires résiduels très faibles (quelques Hz par rapport à quelques kHz en phase solide). Ainsi la résolution caractéristique de la RMN liquide est conservée.
L'alignement de la protéine peut être obtenu avec différents milieux orienteurs : soit des bicelles de lipides ou des membranes pourpres, soit des suspensions de bactériophages filamenteux qui ont la particularité d'induire l'alignement des protéines. Chaque milieu a des propriétés physiques particulières (charge électrique, phase de transition, etc...) qui le rendront plus ou moins compatible avec la protéine cible.
Dans le futur nous aimerions ajouter les informations qui peuvent être déduites des RDC, il existe de nombreux programmes qui utilisent cette approche.

L'objectif de REDCAT (REsidual Dipolar Coupling Analysing Tool) [Valafar Prestegard, 2004] est de proposer une nouvelle interface graphique capable de traiter les RDC de façon rigoureuse. Pour ce faire les deux objectifs sont isolés l'un de l'autre à tel point que différents langages ont été utilisés : Tcl/Tk pour l'interface graphique et C/C++ ppour la partie calculs. Le coeur de l'algorithme utilise l'analyse par valeurs singulières (SVD) qui extrait les meilleurs solutions d'un système imparfait d'équations linéaires. L'échantillonage Monte Carlo génère des erreurs compatibles avec les données observées.

Le programme MODULE [Dosset *et al.*, 2001] a pour but de déterminer les paramètres du tenseur d'alignement des domaines dans les macromolécules à partir des couplages dipolaires résiduels. Il propose une interface graphique dans laquelle l'utilisateur manipule les domaines des macromolécules comme des corps rigides. Il nécessite les valeurs expérimentales des couplages, leur incertitude respective, une estimation du paramètre d'ordre et les coordonnées issues de la PDB. On peut définir des zones à traiter comme entité séparée, celle-ci ne correspond pas forcément à une partie contiguë de la structure primaire. Le programme facilite la réorientation de chaque domaine et de leur tenseur associé. La dégénérescence des orientations relatives par rapport aux rotations de 180° autour des axes(A_{XX},A_{YY},A_{ZZ}) du tenseur d'alignement. Ces orientations équivalentes peuvent être visualisées grâce à l'interface graphique.

Ces programmes traitent le cas des RDC, QUASI devra donner un poids relativement important à ces informations. Ils pourraient être pris en compte sous la forme de différences entre une valeur de référence et la valeur expérimentale.

Si ces trois types d'informations sont disponibles et coïncident, elles devront prendre un poids important dans la fonction. Ce poids diminuera dès que l'accord ne se fera plus. QUASI sera ainsi une programme capable de prendre en compte des informations provenant de nombreuses sources différentes.

Bibliographie

[Atkinson *et al.*, 2001] Atkinson, R. A., Joseph, C., Kelly, G., Muskett, F. W., Frenkiel, T. A., Nietlispach, D., Pastore, A. (2001). *Nat Struct Biol* **8** (10), 853–857. *Ca2+-independent binding of an EF-hand domain to a novel motif in the alpha-actinin-titin complex.*

[Atkinson *et al.*, 2000] Atkinson, R. A., Joseph, C., Kelly, G., Muskett, F. W., Frenkiel, T. A., Pastore, A. (2000). *J Biomol NMR* **16** (3), 277–278. *Assignment of the 1H, 13C and 15N resonances of the C-terminal EF-hands of alpha-actinin in a 14 kDa complex with Z-repeat 7 of titin.*

[Blanchard *et al.*, 1989] Blanchard, A., Ohanian, V., Critchley, D. (1989). *J Muscle Res Cell Motil* **10** (4), 280–289. *The structure and function of alpha-actinin.*

[Cavanagh *et al.*, 1996] Cavanagh, J., Fairbrother, W. J., Palmer, A. G. r., Skelton, N. J. (1996). *NMR spectroscopy principles and practice.* San Diego : Academic Press.

[Coggins Zhou, 2003] Coggins, B. E. Zhou, P. (2003). *J Biomol NMR* **26** (2), 93–111. *PACES : Protein sequential assignment by computer-assisted exhaustive search.*

[Coutouly *et al.*, 2004] Coutouly, M.-A., Kieffer, B., Atkinson, R. A. (2004). *C R Chimie* **7**, 335–341. *QUASI : Quick Access to Spectral Interpretation.*

[Dosset *et al.*, 2001] Dosset, P., Hus, J. C., Marion, D., Blackledge, M. (2001). *J Biomol NMR* **20** (3), 223–231. *A novel interactive tool for rigid-body modeling of multi-domain macromolecules using residual dipolar couplings.*

[Sali *et al.*, 1990] Sali, A., Overington, J. P., Johnson, M. S., Blundell, T. L. (1990). *Trends Biochem Sci* **15** (6), 235–240. *From comparisons of protein sequences and structures to protein modelling and design.*

[Schwarzinger *et al.*, 2000] Schwarzinger, S., Kroon, G. J., Foss, T. R., Wright, P. E., Dyson, H. J. (2000). *J Biomol NMR* **18** (1), 43–48. *Random coil chemical shifts in acidic 8 M urea : implementation of random coil shift data in NMRView.*

[Valafar Prestegard, 2004] Valafar, H. Prestegard, J. H. (2004). *J Magn Reson* **167** (2), 228–241. *REDCAT : a residual dipolar coupling analysis tool.*

[Wishart *et al.*, 1991] Wishart, D. S., Sykes, B. D., Richards, F. M. (1991). *J Mol Biol* **222** (2), 311–333. *Relationship between nuclear magnetic resonance chemical shift and protein secondary structure.*

[Wishart *et al.*, 1997] Wishart, D. S., Watson, M. S., Boyko, R. F., Sykes, B. D. (1997). *J Biomol NMR* **10** (4), 329–336. *Automated 1H and 13C chemical shift prediction using the BioMagResBank.*

[Xu Case, 2001] Xu, X. P. Case, D. A. (2001). *J Biomol NMR* **21** (4), 321–333. *Automated prediction of 15N, 13Calpha, 13Cbeta and 13C' chemical shifts in proteins using a density functional database.*

Partie III

Etude du fragment 24 kDa de la sous-unité B de l'ADN-gyrase

1

Introduction

Dès la découverte de la structure en double hélice de l'ADN en 1953 par Watson et Crick [Watson Crick, 1953], des questions sur les mécanismes mis en jeux lors des différentes étapes de réplication ont été posées. L'ADN, support de l'information génétique, est chez l'homme une macromolécule double brin de 3.10^9 paires de bases, indispensable au cycle cellulaire. La nature topologique de l'ADN influe sur ses fonctions : au repos, la molécule est sous une forme superenroulée, empaquetée dans la structure chromatinienne, et lors de la phase d'activité (réplication, transcription et recombinaison) elle se déroule et s'ouvre pour permettre aux enzymes d'agir. Si l'ADN parent doit se dupliquer en deux molécules d'ADN filles en séparant ses deux brins et en copiant chacun d'eux, alors les brins doivent se dérouler rapidement pendant la réplication.

Avant les années 60, le déroulement de l'ADN ne semble pas présenter d'intérêt particulier pour la communauté scientifique, mais deux découvertes faites à cette période, conduisent les biologistes à s'y intéresser.

La première est l'accumulation de données sur la longueur des molécules d'ADN qui s'avère plus grande que ce que l'on pensait alors. En effet, une molécule d'ADN correspond à un chromosome, c'est à dire à des millions de paires de bases. Pour une molécule aussi longue, l'idée de démêler les brins enroulés en tournant la molécule autour de son axe est impensable.

La seconde est la découverte de différentes formes circulaires d'ADN. La plus importante des découvertes est que la molécule d'ADN du *polyoma virus* (environ 5000 paires de bases) [Dubelcco Vogt, 1963] est un ADN circulaire double brin. Cette information pose le problème de la séparation des brins parents lors de la réplication.

"Since the two chains in our model are intertwined, it is essential for them to untwist if they are to separate. ..... Although it is difficult at the moment to see how these processes occur without everything getting tangled, we do not feel that this objection would be insuperable."
-J. D. Watson and F. H. C. Crick, 1953

On sait maintenant que ces aspects mécaniques sont gérés par des enzymes particulières, les topoisomérases [Wang Liu, 1979].

Elles contrôlent les contraintes de torsion et d'enchevêtrement de l'ADN. En leur présence, les brins d'ADN et les doubles hélices peuvent se traverser l'un l'autre comme s'il n'y avait aucune barrière physique. Toutes les ADN topoisomérases catalysent la coupure des brins ADN par transestérification (Figure : 1.1).

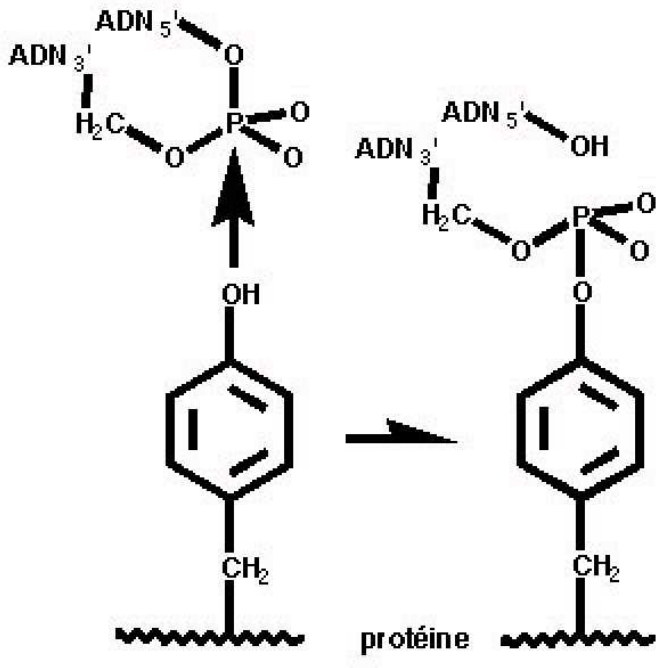

FIG. 1.1 – Transestérification entre l'oxygène phénolique d'un groupe tyrosyle et un groupe phosphorylé du brin d'ADN.

Il existe deux types de topoisomérases :

- les topoisomérases de type I : la première enzyme appartenant à cette famille a été découverte par James WANG [Wang, 1971]. Elle clive l'un des brins d'ADN, ce qui permet sa rotation autour du deuxième brin. Ce mécanisme ne nécessite pas d'apport d'énergie. Le type I peut-être subdivisé en deux familles qui ont des polarités de coupure différentes : la

topoisomérase IA des procaryotes forme une liaison avec l'extrémité 5´ de l'ADN coupé, et la topoisomérase IB des virus et des eucaryotes se fixe sur l'extrémité 3´ de l'ADN. L'oxygène phénolique d'un groupe tyrosyle du site-actif subit une transestérification avec le groupe phosphorylé d'un brin d'ADN.

- les topoisomérases de type II, découvertes par M. GELLERT et al. [Gellert *et al.*, 1976]. Elles agissent en 3 étapes. Tout d'abord, elles clivent les deux brins de l'ADN par la formation d'une liaison phosphodiester entre un phosphate 5´ et le groupement hydroxyle d'une tyrosine de l'enzyme. Ceci permet le passage d'un double brin d'ADN à travers cette coupure qui est refermée par cette même enzyme. Une paire de groupes tyrosyles du site-actif subissent une transestérification avec une paire de groupes phosphorylés de deux brins ADN.

Elles catalysent l'interconversion des différentes topologies de l'ADN en coupant et religant les deux brins d'ADN de façon concomittante pour permettre le passage d'une molécule "duplex" à travers la coupure double-brin. Les topoisomérases II sont intéressantes à plusieurs niveaux : elles détachent les catenanes d'ADN, et elles résolvent les problèmes liés à l'entrelacement des chromosomes durant la mitose ; elles constituent des cibles pour un grand nombre de toxines naturelles, mais aussi pour des médicaments anti-tumeurs.

L'ADN-gyrase est spécifique des procaryotes, et malgré ses similitudes avec les topoisomérases II des eucaryotes, les différences permettent une action spécifique des antibiotiques. Cet enzyme est un hétérotétramère constitué de deux sous-unités A de masse moléculaire de 97 kDa et de deux sous-unités B de masse moléculaire de 90 kDa. Son poids total est de 374 kDa. Chaque sous-unité est constituée de 2 domaines (Figure : 1.2).

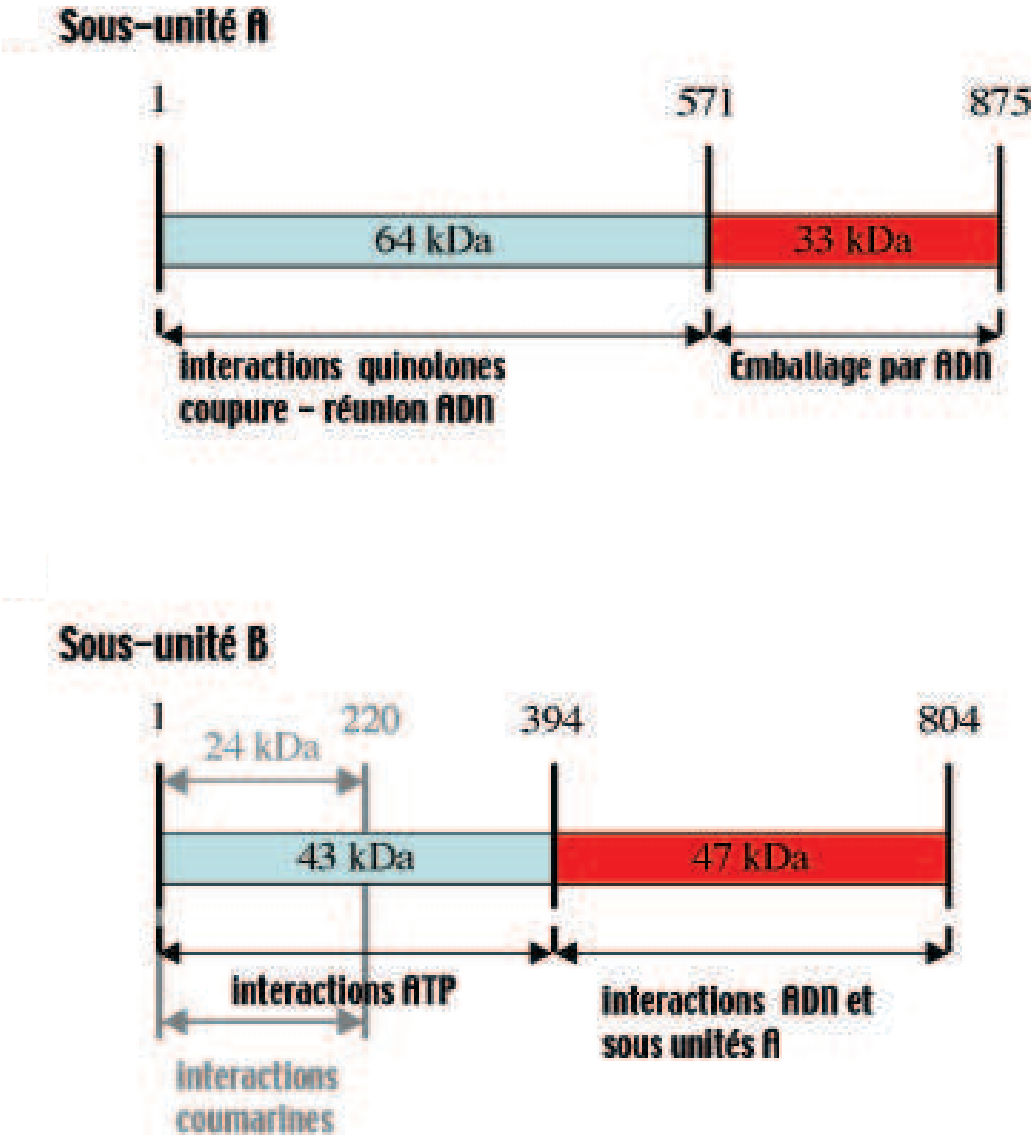

FIG. 1.2 – Représentation schématique des domaines de l'ADN-gyrase. Les acides aminés marquant les bordures des domaines structuraux sont indiqués par leur numéro. Les trait noirs verticaux indiquent les sites de clivages protéolytiques.

Les deux sous-unités A réalisent la réaction de superenroulement de l'ADN. Le fragment N-terminal de la sous-unité A est impliqué dans la réaction de coupure et de réunion de l'ADN. Le domaine C-terminal est impliqué dans les interactions protéine-ADN. Ceci a été confirmé par des études qui ont montré que le domaine C-terminal est impliqué dans l'enroulement par la droite de l'ADN autour de l'enzyme [Reece Maxwell, 1991]. Une mutation dans ce domaine n'impose pas à l'ADN de se tordre de ce côté là, alors que le domaine purifié seul le fait.

Les deux sous-unités B fournit l'énergie nécessaire pour l'introduction du superenroulement négatif en hydrolisant l'ATP. La sous-unité B est constituée d'un domaine N-terminal de 43 kDa présentant une activité ATPasique qui est responsable de la capture du brin d'ADN, et d'un domaine C-terminal de 47 kDa qui interagit avec la sous-unité A et avec l'ADN. Le site catalytique ATPase se situe à l'intérieur du domaine N-terminal et le second sous-domaine connecte les domaines A et B. L'ADN-gyrase à laquelle on enlève la partie N-terminal de la sous-unité B possède une activité de relaxation de l'ADN indépendante de l'ATP. Cette dernière capacité est spécifique du superenroulement négatif de l'ADN.

Ces différentes activités sont étroitement liées à la structure (Figure : 1.3) de la protéine.

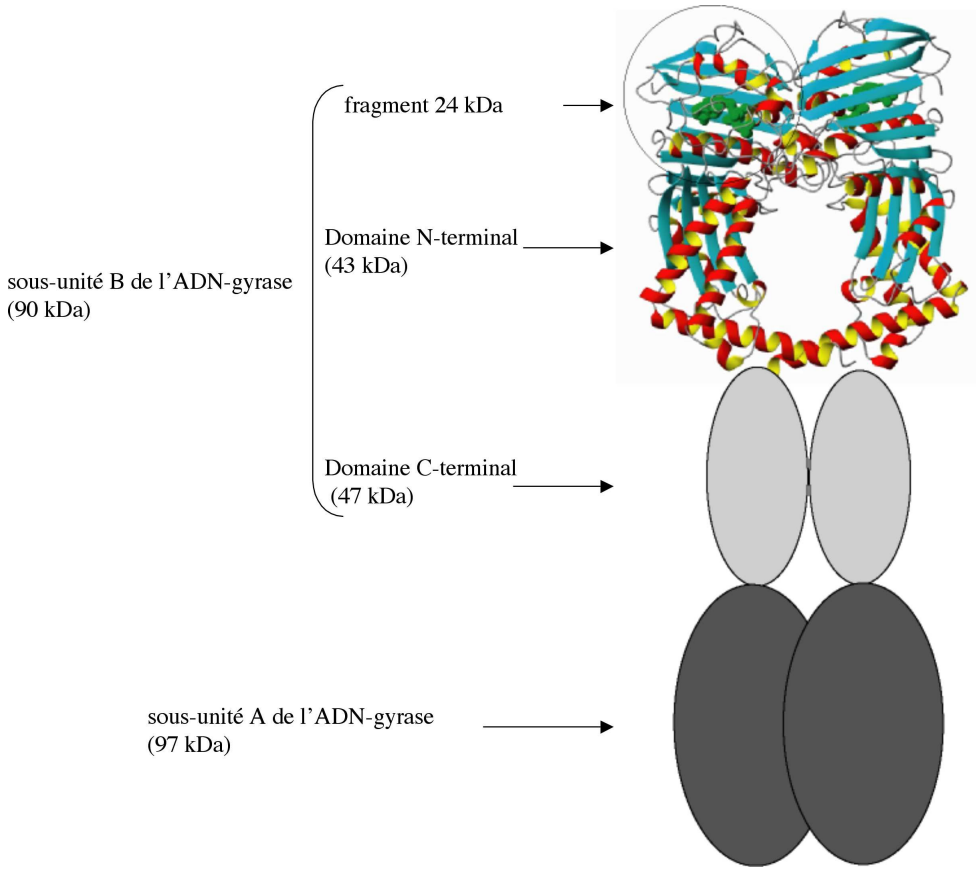

FIG. 1.3 – Schéma de l'ADN-gyrase. Seule la structure du fragment 43 kDa de la sous-unité B est détaillée. Le fragment 24 kDa se situe dans le cercle supérieur.

La structure du domaine de 59 kDa de la sous-unité A a été résolue à 2,8 Å et peut être subdivisée en deux régions : la tête et la queue reliées par 3 hélices [Morais-Cabral *et al.*, 1997].

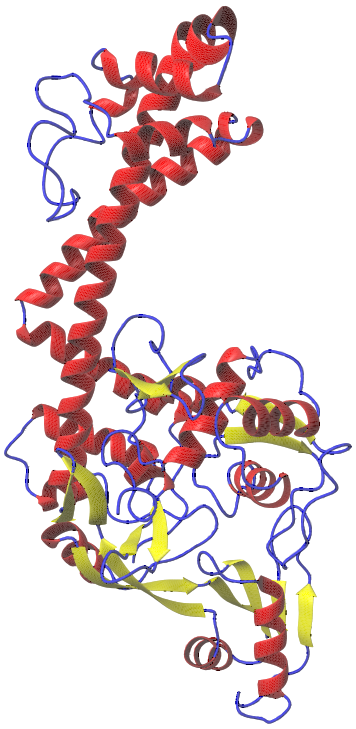

FIG. 1.4 – Représentation en ruban du domaine de 59 kDa de la sous-unité A de l'ADN-gyrase. Les hélices-α sont en rouge et les brins-β en jaune.

La grande flexibilité des trois hélices (Figure : 1.4) pourrait être à l'origine des changements conformationnels importants observés dans la protéine durant le cycle de catalyse. Malgré la connaissance de cette structure, le mécanisme d'interaction avec l'ADN est encore inconnu.

Le fragment N-terminal 43 kDa contient 4 brins-β et 4 hélices-α. L'une de ces hélices saille presque perpendiculairement par rapport au corps de la protéine. Les contacts inter-dimères sont formés principalement par des résidus du premier sous-domaine. Dans le dimère, le second sous-domaine (221-392 résidus) forme un trou de 20 Å de diamètre qui est formé par des résidus polaires, principalement des arginines. Ceci suggère que cette région interagit avec l'ADN. Attendu que la sous-unité A est connue comme étant le principal site de liaison de l'ADN, le trou dans la sous-unité B intervient dans le passage du brin ADN [Sharma Mondragon, 1995] pendant la réalisation du superenroulement. Le fragment 24 kDa, contenant le site de fixation à l'ATP, contient un coeur hydrophobe compact formé par cinq hélices-α et un tonneau formé de 8 brins-β.

L'ADN-gyrase est la cible de deux classes d'inhibiteurs : les coumarines et les quinolones (Figure : 1.5).

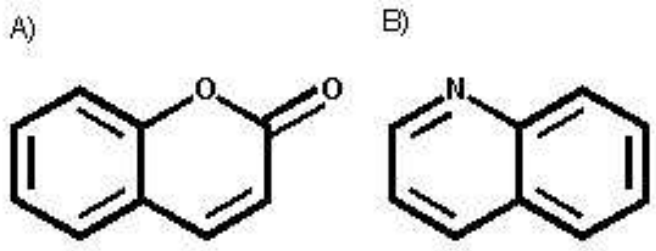

FIG. 1.5 – Formules développées des classes d'inhibiteurs de l'ADN-gyrase : A) les coumarines et B) les quinolones.

Ces composés inhibent le superenroulement de l'ADN en bloquant le complexe sous-unité A-ADN pendant l'étape de coupure-soudure. Le mécanisme exact lié à ce blocage est lui aussi encore inconnu [Kampranis *et al.*, 1999, Lewis *et al.*, 1996b]. Le fragment N-terminal 43 kDa complexé à l'ADPNP a été cristallisé [Wigley *et al.*, 1991, Brino *et al.*, 1999]. Des structures cristallographiques de ce fragment lié à des inhibiteurs dérivés de la coumarine [Tsai *et al.*, 1997] et de la cyclothialidine sont également connues [Lewis *et al.*, 1996a]. Les coumarines se lient à la sous unité B de l'ADN-gyrase et sont des inhibiteurs capables d'entrer en compétition avec l'ATP. Le mécanisme d'inhibition est bien caractérisé grâce à l'existence de structures cristallographiques.

Afin d'étudier l'effet de la fixation de l'ATP ou des inhibiteurs de type coumarine sur la dynamique du fragment 24 kDa, nous avons produit plusieurs échantillons marqués de cette protéine en utilisant un système d'expression basé sur les cyanobactéries. Ce système permet un marquage isotopique à des coûts particulièrement compétitifs. Le fragment 24 kDa de la sous-unité B de l'ADN-gyrase présente un poids à la limite supérieure des tailles pouvant être traitée par la RMN sans échantillon deutéré. Avant d'obtenir un échantillon RMN capable de donner des spectres triples résonances propres, il faut étudier les procédés de production et de marquage des protéines.

2

Expression des échantillons marqués

Sommaire

De nombreux organismes ont été utilisés pour effectuer le marquage isotopique. Des organismes tels que *Dictyostelium discoideum*, *Tolypocladium inflatum*, *Trichoderma viride*, *Streptomyces nashvillensis* ont été adaptés pour se développer dans un milieu enrichi en précurseurs marqués (glycérol, glucose, NH_4Cl, KNO_3..) : ils d'incorporent alors les isotopes ^{13}C et ^{15}N dans les protéines exprimées de façon endogène. Habituellement l'ammonium, le glucose et le D_2O sont utilisés comme précurseurs [Kay Gardner, 1997] dans *Escherichia coli* comme cellule hôte. Le glucose est à la fois une source de carbone et d'énergie. Dans le cas du marquage au deutérium, afin de ne pas incorporer des protons en même temps que la source de carbone, il faut utiliser du glucose marqué non seulement en ^{13}C mais aussi en ^{2}H. Ce substrat s'avère particulièrement onéreux.

L'utilisation de la cyanobactérie *Anabaena* sp.PCC 7120 [Desplancq *et al.*, 2001] présente l'avantage d'utiliser le dioxyde de carbone ambiant comme source de carbone pour exprimer la protéine d'intérêt. Cette source est donc économique et sans protons. Cette propriété rend la production de l'échantillon marqué beaucoup moins onéreuse que celle obtenue par *Escherichia coli*.

2.1 Les cyanobactéries

Les cyanobactéries sont des organismes procaryotes, donc des bactéries, avec des pigments photosynthétiques très proches de ceux que l'on trouve chez les plantes supérieures. Du fait de la présence de ces pigments, elles ont été classées parmi les algues et elles s'appelaient algues bleues ou cyanophycées. Elles font donc l'objet d'une double classification : une classification botanique et une classification bactérienne (Figure : 2.1).

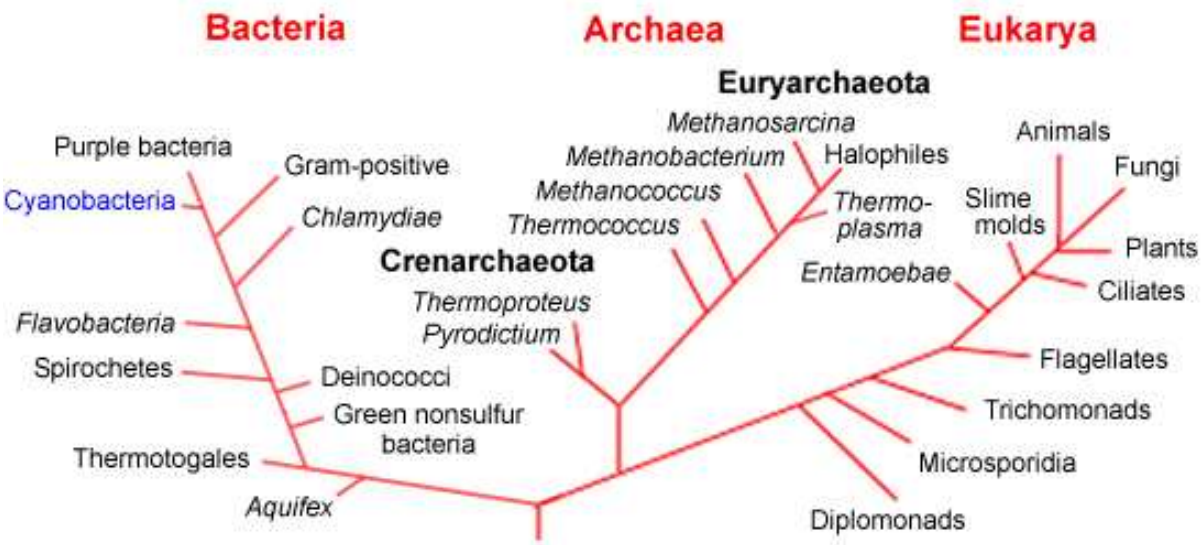

FIG. 2.1 – Arbre phylogénique de la cyanobactérie.

Apparues il y a environ 3,5 milliards d'années, les cyanobactéries représentent un groupe bactérien majeur tant par leur diversité morphologique et physiologique que par le rôle qu'elles ont joué en créant une atmosphère aérobie sur notre planète. Elles jouent, encore de nos jours, un rôle important dans l'équilibre des proportions entre le gaz carbonique et l'oxygène. En effet, ces organismes procaryotes partagent avec les plantes la capacité d'effectuer la photosynthèse en utilisant la lumière et l'eau pour la réduction du gaz carbonique, processus qui s'accompagne d'un dégagement d'oxygène.

Possédant un très grand potentiel d'adaptation à des environnements, même extrêmes, elles colonisent la plupart des écosystèmes aquatiques et terrestres. Ce sont les organismes les plus simples capables de photosynthèse (Figure : 2.2). Elles sont capables de croître en milieu minimal contenant des sels minéraux, du carbonate (CO_3^{2-}) comme seule source de carbone et du nitrite (NO_2^-) du nitrate (NO_3^-) ou de l'ammonium (NH_4^+) comme

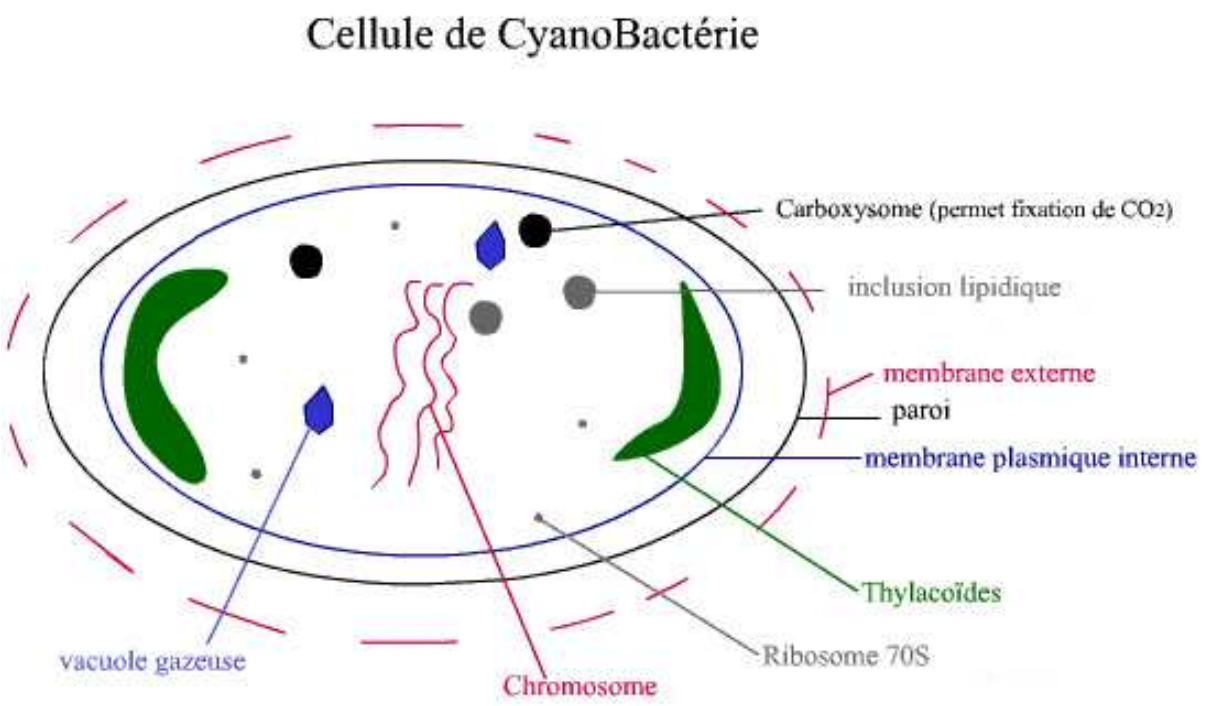

FIG. 2.2 – Schéma de la cellule d'une cyanobactérie.

source d'azote. On peut les trouver soit sous forme de cellule seule soit sous forme de colonies. Ces dernières peuvent former des filaments, des brins ou même des boules creuses. La cyanobactérie *Anabaena* se présente sous forme filamenteuse. *Anabaena*, comme la plupart des cyanobactéries filamenteuses, est capable de fixer l'azote atmosphérique. Cette propriété est liée à la présence de cellules spécifiques appelées *hétérocystes* dont l'intérieur est microaérobique et donc adéquat à la fixation d'azote.

2.2 *Anabaena* sp. PCC 7120

Un premier système d'expression constitutif chez *Anabaena* a été développé [Desplancq *et al.*, 2001]. L'expression de protéines non toxiques, dont le fragment 24 kDa, dans *Anabaena* a été réalisée avec succès, par contre, dans le cas de protéines toxiques, le système échouait. Le manque de système d'expression fort et bien régulé limitait l'utilisation de la cyanobactérie à un petit nombre de protéines non-toxiques pour *Anabaena*. Depuis, un système d'expression inductif a été développé [Desplancq *et al.*, 2004]. Il utilise le promoteur le mieux caractérisé d'*Anabaena* : *nir* [Frias *et al.*, 2003]. Ce dernier contrôle l'assimilation des nitrates et nitrites par les

cellules *Anabaena*. Ce promoteur est actif lorsque la source d'azote dans le milieu est du nitrate ou du nitrite, et réprimé en présence d'ammonium. Il est extrêmement bien régulé [Frias *et al.*, 2000]. De plus, l'utilisation de nitrate marqué comme inducteur est un avantage lorsque l'expérimentateur veut produire une protéine marquée.

2.2.1 Production de fragment marqué 24 kDa de la sous-unité B de l'ADN-gyrase

La culture des cellules d'*Anabaena* sp. PCC 7120 est effectuée dans un bioréacteur équipé de capteurs de pH et O_2. Le fermenteur de 3 litres est éclairé par une source de lumière d'intensité de 3500 à 5000 lux du début à la fin du processus. La culture débute dans un milieu BG11 [Castenholtz, 1988] contenant 1 g $NaH^{13}CO_3$/L et 1 g $Na^{15}NO_3$/L ; la croissance de l'inoculum est faite dans du BG11 NH_4^+. Tout au long de la culture, le dioxygène produit par les cellules est éliminé en faisant circuler de l'argon dans le milieu. La présence d'argon permet également d'éliminer la source de carbone présnet dans l'air sous forme de gaz carbonique. L'agitation est maintenue à 100 rpm. La température et le pH sont maintenus respectivement à 28°C et à 8. La croissance est suivie par mesure de l'absorbance à 700 nm deux fois par jour ; 1 g $NaH^{13}CO_3$ est ajouté à la culture lorsque la concentration de carbonate est proche de 0,1 g/L. L'expression continue du fragment 24 kDa de la sous-unité B de l'ADN-gyrase, dans ces conditions, est vérifiée par analyse d'aliquots d'extraits totaux de cellules. La fermentation est arrêtée quand l'absorbance à 700 nm atteint une valeur de 2. Les cellules sont stockées à -20°C.

Purification du fragment 24 kDa de la sous-unité B de l'ADN-gyrase

Deux culots de cellules congelées sont resuspendus dans 40 mL de tampon A (50 mM Tris-HCl pH 8,0, 10% glycérol, 10 mM β-mercaptoéthanol, 1 mM EDTA 0,5 mM) et soniqués deux fois 2 minutes sur un Branson Sonifier 450 (Branson Ultrasonics, Danbury, CT). Le lysat résultant est centrifugé 10 minutes à 10 000 rpm pour séparer les débris cellulaires. Le surnageant est chargé sur une colonne novobiocine-Sepharose (colonne d'affinité) équilibrée avec du tampon A. Après deux lavages successifs (tampon A contenant 0,5 M KCl puis tampon A contenant 1 M KCl) la protéine liée est éluée par du tampon A contenant de l'urée 5 M.

Les fractions contenant le fragment 24 kDa de la sous-unité B de l'ADN-gyrase sont regroupées. On dialyse l'échantillon contre du tampon A puis du tampon A avec 25 mM

en KCl.

L'échantillon est récupéré et chargé sur une colonne échangeuse d'anion. Cette colonne est équilibrée au préalable avec du tampon A 25 mM KCl. Après le passage de l'échantillon, un lavage est effectué avec du tampon A 25 mM KCl et l'élution est faite avec une solution de tampon A, 333 mM KCl.

Les fractions contenant le fragment 24 kDa de la sous-unité B de l'ADN-gyrase dénaturé sont regroupées. On dialyse l'échantillon contre du tampon P (10 mM NaH_2PO_4, pH 7).

Enfin, la protéine est concentrée avec un Centricon-10 (Sartorius, Palaiseau, France) en centrifugeant à 2500 rpm. L'expression est de 100 mg/L de culture à une absorbance de 2 à 700 nm. La concentration de l'échantillon obtenu est de 0,8 mM. La protéine est conservée dans du tampon P à 4°C.

2.3 Marquage spécifique

2.3.1 Principe

Nous avons souhaité explorer la possibilité de coupler l'interprétation automatique des spectres par QUASI avec l'identification de ceratins types d'acides aminés. L'idée consiste à incorporer de façon séléctive un type d'acide aminé non marqué et de détecter des différences d'intensités spécifiques sur les spctres de corr elation. Cette approche est connue sous le nom de 'reverse labelling'.

Dans le cas présent, le marquage s'apparente à un non marquage en ^{15}N de 4 types d'acide aminé. En pratique, il s'agit de faire exprimer la protéine cible dans la souche *Anabaena* sp. PCC 7120 en ajoutant 100 mg d'acide aminé non marqué en ^{15}N (Ile, Leu, Met, Val). Le reste de la protéine est marquée ^{15}N. La production et la purification de la protéine ont été réalisées en erlenmeyer. Ayant cet acide aminé déjà sous la forme finale dans son milieu, la cyanobactérie ne devrait pas le synthétiser. Tout acide aminé de ce type dans la séquence doit être exempt de ^{15}N et ne devrait pas apparaître sur le spectre 2D ^{1}H-^{15}N HSQC. Ainsi, à moindre coût, on devrait identifier rapidement toutes les Ile, Leu, Met et Val du fragment 24 kDa de la sous-unité B de l'ADN-gyrase.

2.3.2 Résultats

Les spectres 2D ^{1}H-^{15}N HSQC sont enregistrés sur un spectromètre DRX500 avec 8 scans (Figure : 2.3).

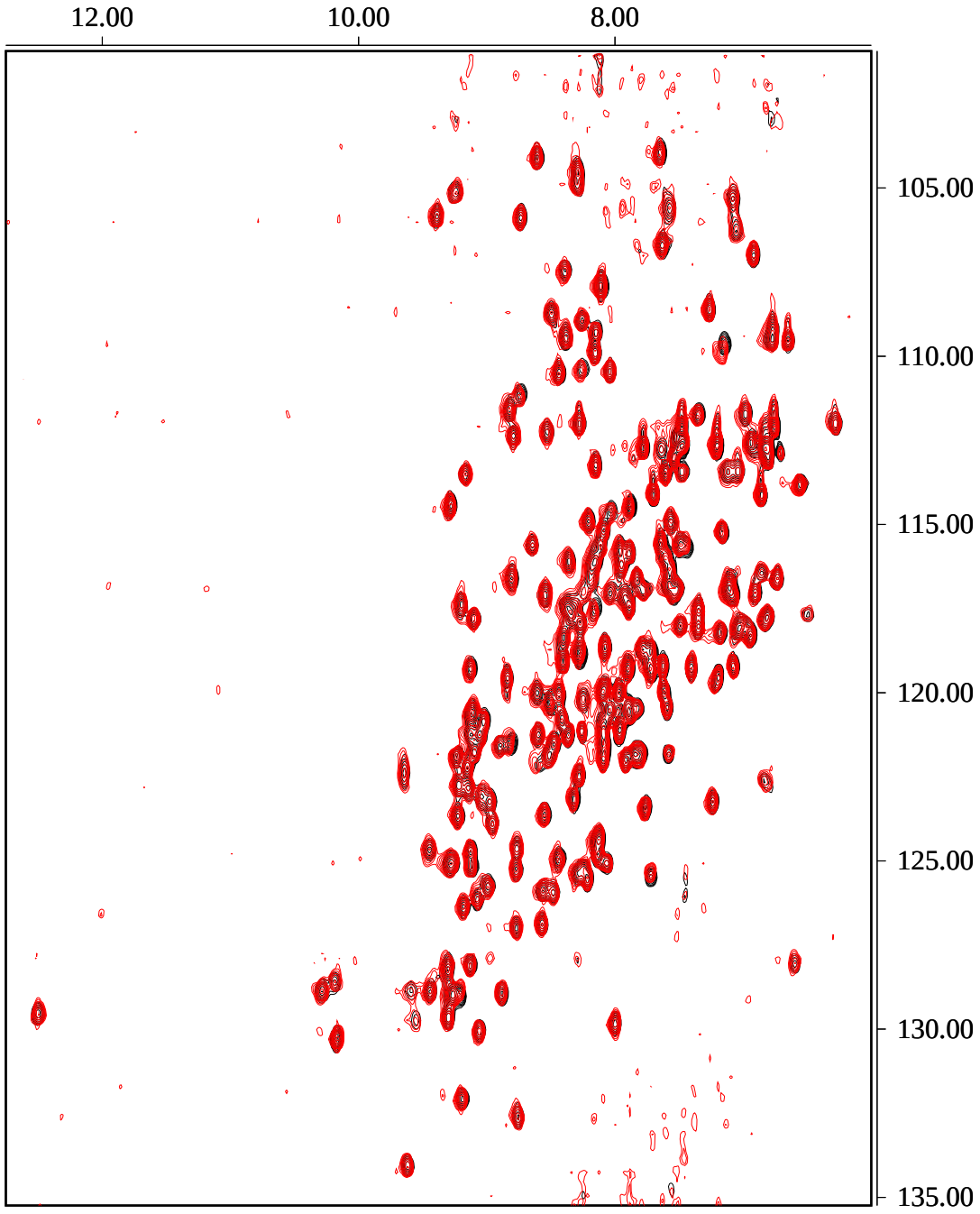

FIG. 2.3 – 2D ^{1}H-^{15}N HSQC superposées. En noir le spectre obtenu par "marquage" des leucines, en rouge celui obtenu par "marquage" des méthionines.

Une fois certains pseudo-résidus attribués, les types de résidus peuvent être regroupés et il est alors possible de tracer les intensités en fonction du type de résidu observé (Figure : 2.4).

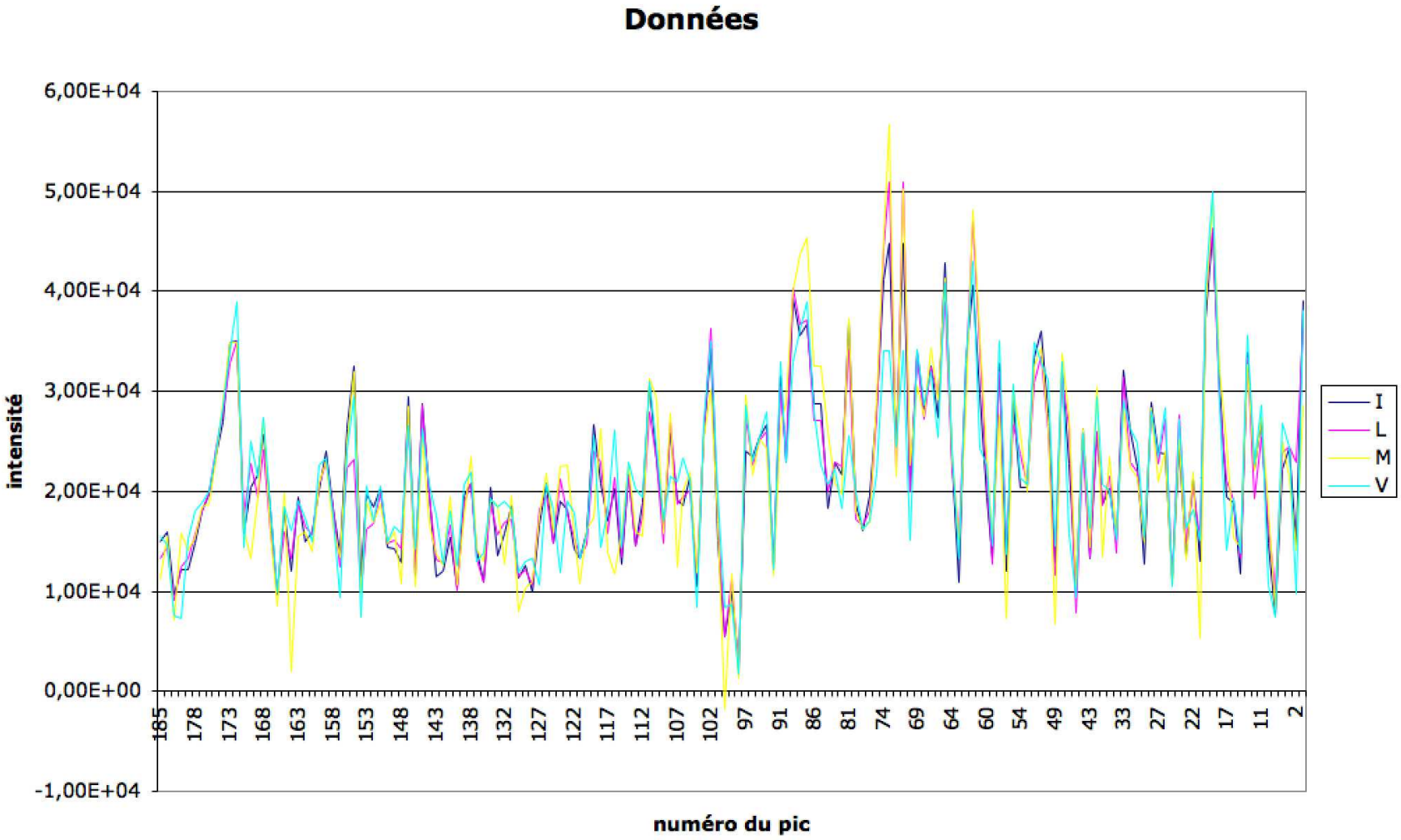

FIG. 2.4 – Intensités obtenues à la suite des marquages des acides aminés : isoleucines, leucines, valines et méthionines. Les types d'acides aminés sont regroupés.

La figure 2.4 montre que ^{14}N n'a pas été incorporé dans la protéine d'une façon spécifique puisqu'aucune variation différentielle de l'intensité des pics n'est observée. Deux hypothèses peuvent expliquer cette observation :

- l'activité des transaminases conduit à une dilution infinie de l'azote 14 sur l'ensemble des résidus.

- l'acide aminé ne traverse pas la membrane de la bactérie.

Afin de tester cette hypothèse, de nouvelles expériences seront tentées en utilisant l'incorporation de ^{12}C comme marqueur d'acide aminé.

3

QUASI sur le fragment 24 kDa de la sous-unité B de l'ADN-gyrase

Sommaire

3.1 Préparation des spectres

Les spectres RMN (HSQC, HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HNCO, HN(CA)CO) ont été enregistrés sur un spectromètre Bruker DRX600 à 303 K. Les séquences d'impulsion utilisées sont celles fournies par Bruker de façon standard. Des légères modifications ont été apportées suivant l'article de Eletsky et al. [Eletsky *et al.*, 2001] pour les protéines partiellement deutérées. Les techniques de deutération et de TROSY ont été couplées afin d'obtenir les meilleurs spectres possibles. Les largeurs spectrales utilisées sont de 18 ppm pour le proton, 33 ppm pour l'azote, 30 ppm pour le $^{13}C^{\alpha}$, 65 ppm pour le $^{13}C^{\beta}$ et 12 ppm pour le ^{13}CO. Les spectres sont traités grâce à NMRPipe [Delaglio *et al.*, 1995], puis l'analyse et les listes de pics sont faites avec le programme

XEASY [Bartels *et al.*, 1995].

Les déplacements chimiques théoriques ont été calculés à l'aide des logiciels SHIFTS et SHIFTY en utilisant la structure contenue dans le fichier 1KZN.pdb. Pour SHIFTS, les prédictions obtenues ne traitent ni les 15 premiers résidus ni les 2 derniers, ni le résidu Cys 56. Pour ce qui est de SHIFTY, ni les 15 premiers résidus, ni le dernier résidu ne sont prédits. On trouve aussi quelques pics isolés pour lesquels le programme ne donne pas de solution : Thr 34 et Asn 107. Dans les deux cas, les données manquantes sont remplacées par les déplacements chimiques issus de la table des valeurs Random Coil.

3.2 Préparation des listes de pics

Le spectre HN(CA)CO n'ayant donné de bons résultats, QUASI a été utilisé avec les listes issues des spectres HN(CO)CA, HNCA, HN(CO)CACB et HNCACB.

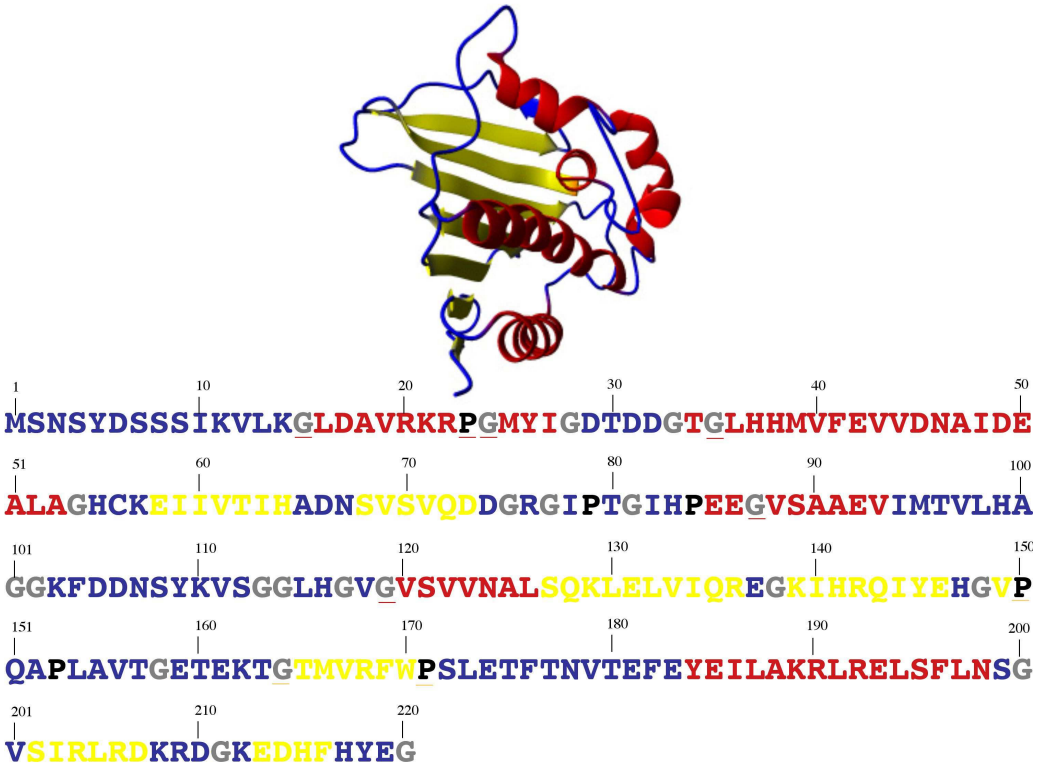

FIG. 3.1 – Représentation en ruban du fragment 24 kDa de la sous-unité B de l'ADN-gyrase. En dessous : séquence primaire de la protéine. Les prolines sont présentées en noir et les glycines en gris. Les hélices-α sont en rouge et les brins-β en jaune.

Sur les 214 pics attendus, car il y a 6 prolines (Figure : 3.1), nous en obtenons 169 (seulement 82% des pics). Le tableau 3.1 récapitule les pics obtenus pour chaque spectre.

Spectre	Nombre de pics
HN(CO)CA	169
HNCA	168
HNCA(CO)CB	152
HNCACB	144

TAB. 3.1 – Nombre de pics issus de chaque spectre utilisé pour le fragment 24 kDa de la gyrase B. Le nombre de pics attendus est de 214.

3.3 QUASI-1

Pendant l'étape de connexion intra-résiduelle, QUASI-1 propose en 71 occasions à l'utilisateur de vérifier les spectres. Dans 11 cas, les deux meilleures solutions se trouvent à l'intérieur des marges fixées ; dans 32 cas, il n'y a aucune proposition dans les marges. Dans 17 cas, le choix doit se faire sur l'unique déplacement chimique de l'atome $^{13}C^{\alpha}$. Enfin, dans les 11 cas restants, la meilleure proposition est déjà acceptée à un ou plusieurs autres endroits.

3.4 Constitution de cycles

Dans ce cas précis, la démarche utilisée est légèrement différente de celle suivie avec les autres protéines testées. Les données utilisées sont en effet de qualité très variable (Figure : 3.2) et cela implique un léger changement de stratégie.

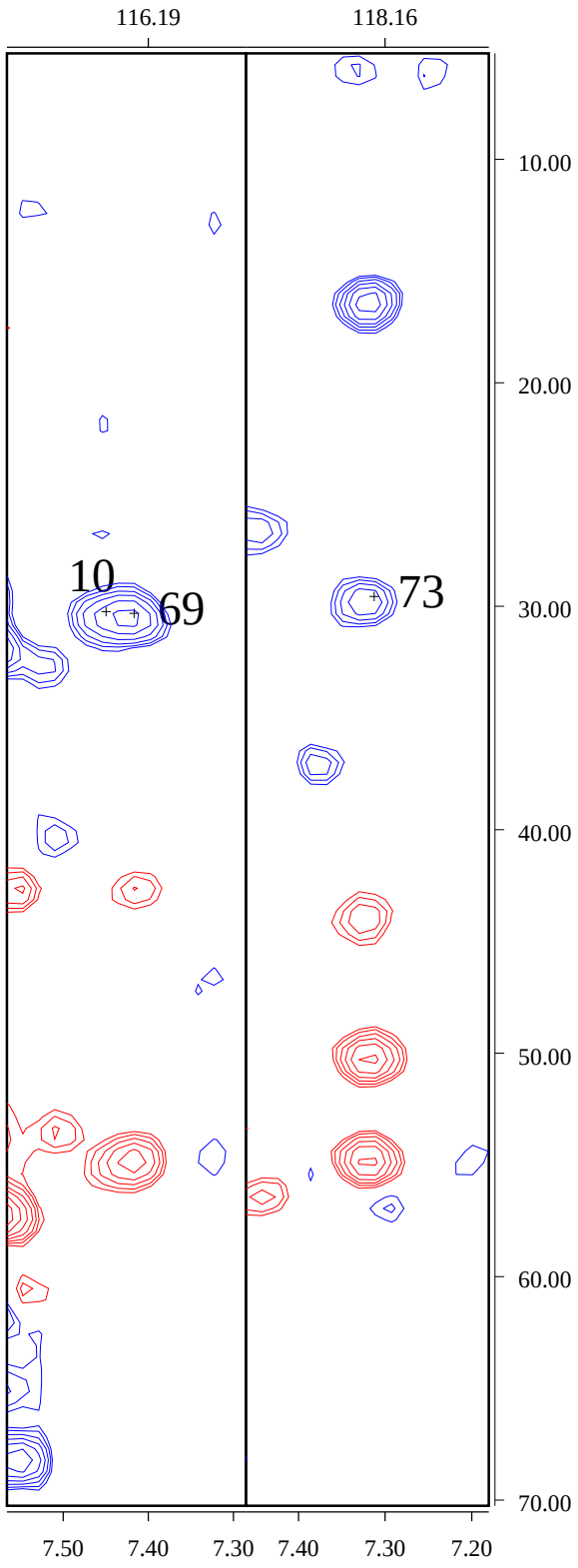

FIG. 3.2 – Exemple de disparité entre les données de départ sur le spectre HNCACB. Alors que le pseudo-résidu 73 est bien défini, le numéro 69 est plus ambiguë.

Au lieu d'utiliser QUASI-2 de façon unique sur tous les fragments obtenus à l'issue de QUASI-1, nous avons utilisé QUASI de façon itérative. Dans un premier temps, QUASI a été appliqué aux données "brutes", c'est à dire que seules les connexions acceptées automatiquement sont faites. Une analyse du résultat obtenu permet d'obtenir et de valider des connexions supplémentaires. Si les connexions sont acceptées, nous relançons QUASI avec le/les nouveaux fragments. Si ce dernier se place de façon unique dans tous les cas, sa position est considérée comme fixée. Ceci permet de réduire le nombre de calculs effectués et de réduire la mémoire utilisée. Les fragments de plus en plus petits peuvent être traités. Cette démarche se répète jusqu'à ce que tous les fragments soient placés. Tout le processus est réalisé avec un test du χ^2 à un taux de confiance de $1/1000^{ème}$.

3.5 Les différents cycles

3.5.1 Les données brutes

Dans un premier temps, nous avons systématiquement coupé les fragments lorsque QUASI-1 nous questionnait et nous avons obtenu 71 fragments de tailles très diverses.

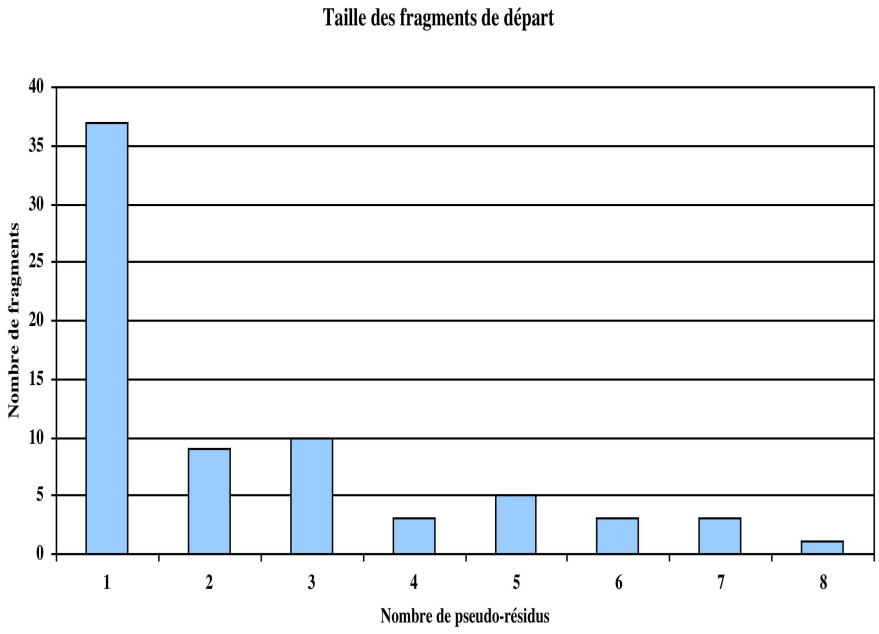

FIG. 3.3 – Histogramme présentant la taille des fragments obtenus au départ du processus d'attribution.

La figure 3.3 montre que la taille moyenne est de 2,38 pseudo-résidus par fragment. Le fragment le plus long est formé de 8 pseudo-résidus. Il existe un nombre important de pseudo-résidus isolés. On traitera d'abord les plus grands fragments qui se placent de façon unique. Dans ce cas, les 15 premiers fragments sont placés (Figure : 3.4, Figure : 3.5).

FIG. 3.4 – Résultats obtenus pour les 15 premiers fragments par rapport aux prédictions faites par le logiciel SHIFTS. Les fragments sont placés de façon unique ou pas du tout. Ainsi les fragments numérotés 14 et 15 ne sont pas placés du tout.

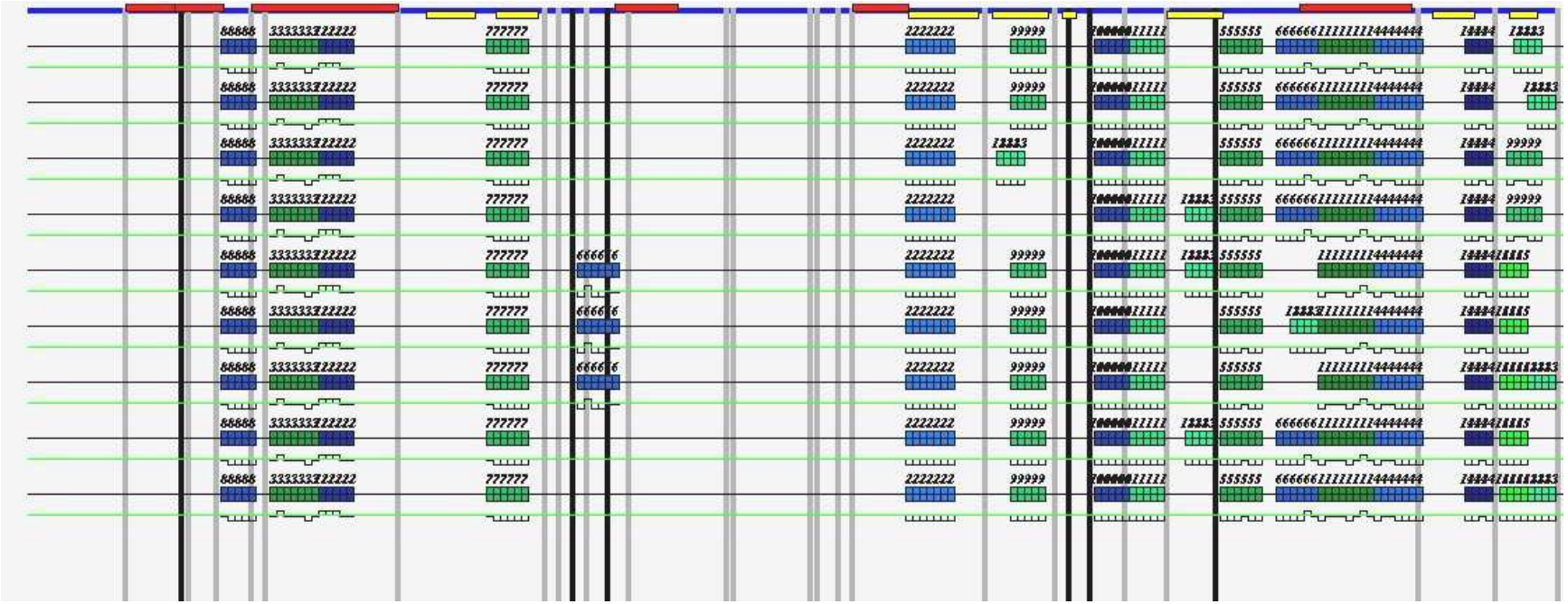

FIG. 3.5 – Résultats obtenus pour les 15 premiers fragments par rapport aux statistiques menées sur la base de données BMRB.

Certaines positions trouvées indiquent des connexions possibles : la connexion entre le $3^{ème}$ et le $12^{ème}$ fragment, celle entre les fragments 10 et 11 et l'enchaînement entre les fragments 6, 1 et 4. Une fois des connexions identifiées, nous les vérifions dans les spectres et nous les validons.

3.5.2 Le second cycle

Avec ces nouveaux fragments, QUASI est relancé. De nouvelles connexions sont proposées et validées : on accepte 2 propositions faites par QUASI-1 et on fait 2 connexions qui s'avèrent être le choix entre les pseudo-résidus 16 et 122 qui sont tous deux acceptés par le pseudo-résidu 197 comme précédents (Figure : 3.6).

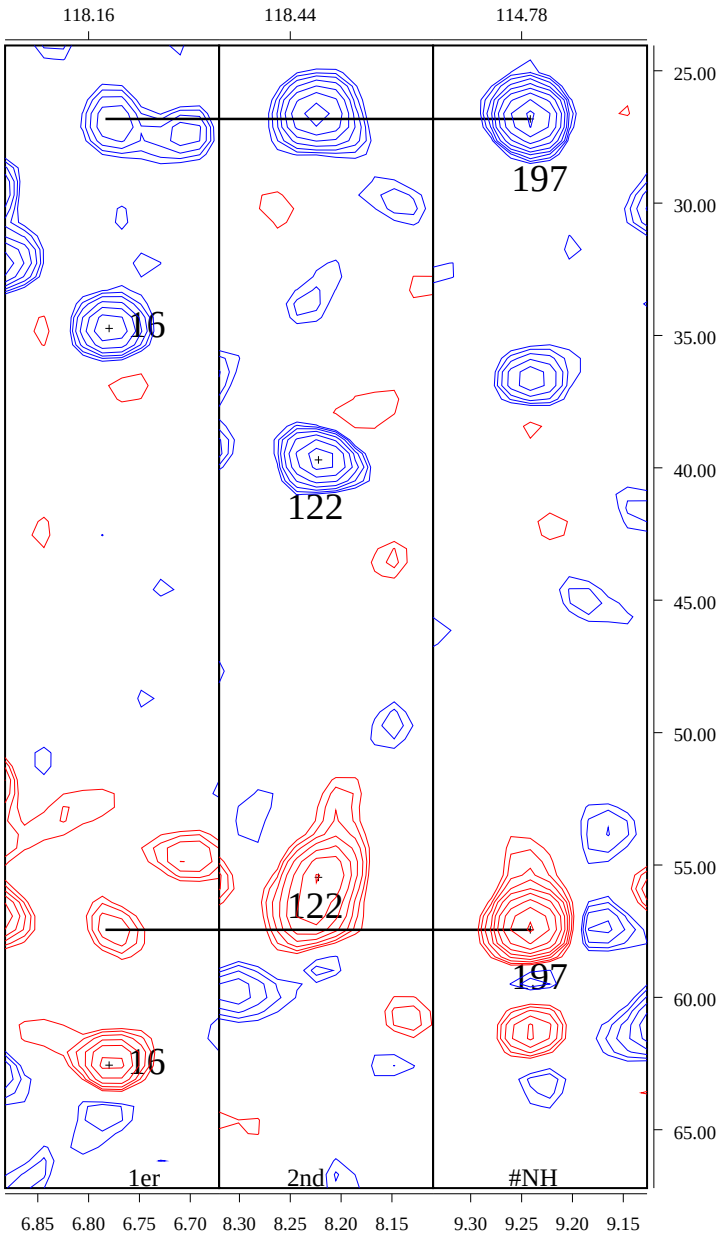

FIG. 3.6 – Présentation sur le spectre HNCACB du cas du pseudo-résidu 197 qui accepte les pseudo-résidus 16 et 122 comme candidats.

Les trois plus grands fragments sont alors formés de 21, 12 et 10 pseudo-résidus. On refait fonctionner QUASI sur les mêmes fragments et on note ceux qui se placent de façon unique dans toutes les méthodes. On ancre alors les fragments 1, 2, 3, 5 et 7 sur la séquence sur les positions 180, 36, 154, 172 et 29. Le nombre total de fragments est maintenant de 66.

Après cette première phase d'attribution, il apparait nettement trois zones où les fragments ne se placent pas : les 18 premiers résidus, les positions allant de 47 à 65 et de 72 à 128.

Le processus d'attribution est répété jusqu'à ce que le programme n'indique plus de nouveau positionnement unique. La figure 3.7 montre la composition différente des frag-

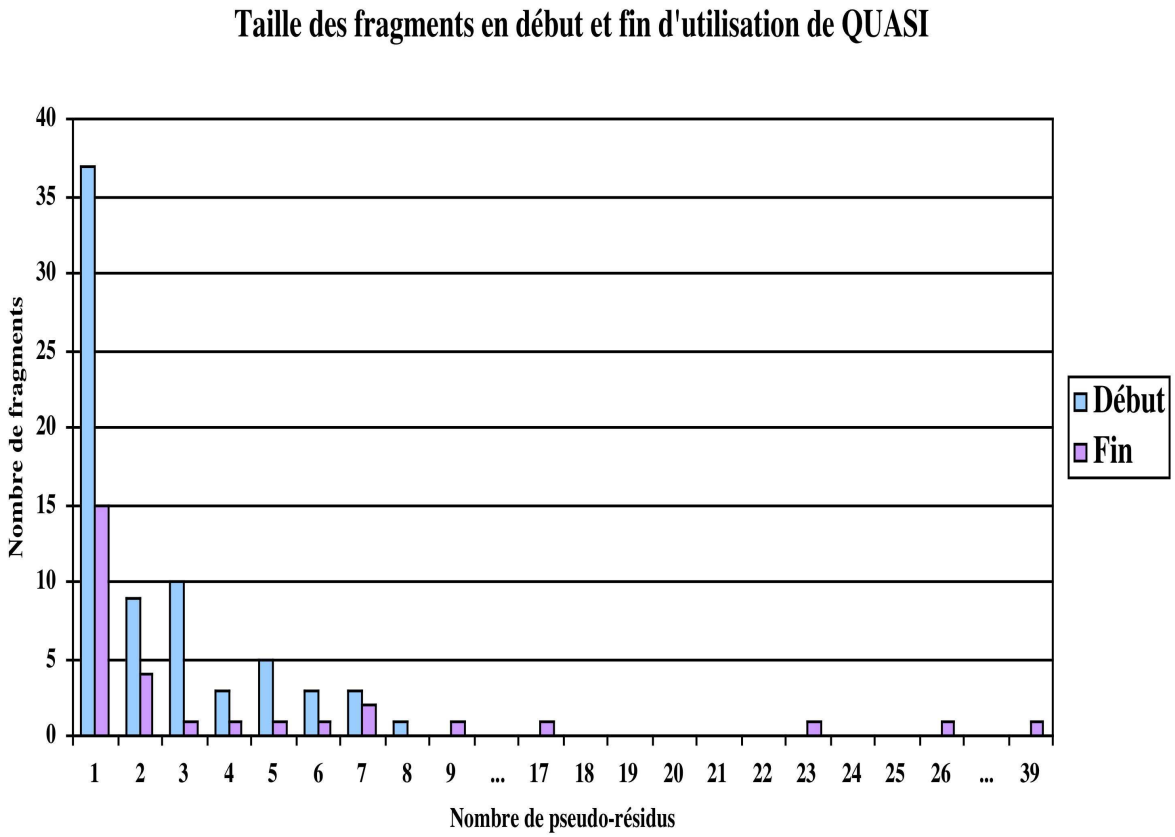

FIG. 3.7 – Histogramme présentant la taille de fragments en début et en fin du processus d'attribution.

ments en début et en fin d'utilisation de QUASI sur le fragment 24 kDa de l'ADN-gyrase. Le nombre de fragments obtenus en fin de processus est bien inférieur (30) au nombre du

départ. La taille moyenne des fragments, à l'issue de l'attribution est de 5,6 pseudo-résidus.

3.6 Attribution finale

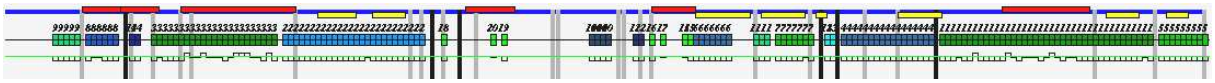

FIG. 3.8 – Résultat obtenu à l'issue de l'utilisation de QUASI sur le fragment 24 kDa de la sous-unité B de l'ADN-gyrase. 159 pseudo-résidus sont placés sur la séquence.

Au final (Figure : 3.8), 159 des 169 pseudo-résidus sont placés sur la séquence. Les zones non attribuées sont les suivantes : les 9 premiers résidus de la protéine ainsi que la zone allant de G78 à S101. Cette dernière zone renferme les deux principales boucles du fragment 24 kDa de l'ADN-gyrase. Ces boucles sont mobiles et rendent l'enregistrement de données RMN difficile.
Une étude, datant de 1998 à été faite au laboratoire afin d'étudier la structure et la dynamique des régions du fragment 24 kDa de la sous-unité B de l'ADN-gyrase, en interaction avec différents inhibiteurs. Parmi ces régions, celle comprise entre G99 et G113, a été particulièrement analysée. La stratégie utilisée pour cette étude est le marquage sélectif des acides aminés présents dans la zone d'intérêt : Gly, Phe, Asp, Lys, Asn et Ser. Les attributions faites à ce moment là, par le Docteur F. Gaudin, sont comparées aux attributions faites dans cette étude par QUASI, ainsi qu'à celles faites par Bellanda et al.[Bellanda *et al.*, 2002].

Dans la figure 3.9 on compare les endroits où les méthodes trouvent une attribution. Les étoiles (*) indiquent les résidus où les attributions faites au laboratoire par le marquage spécifique (référencé comme Gaudin) ne coïncident pas avec l'attribution faite par Bellanda et al.. Par exemple, le pic attribué à K11 par le marquage spécifique est attribué à K16 par Bellanda. Dans la zone non attribuée par QUASI, les attributions de Gaudin et de Bellanda ne concordent pas, par contre, pour les autres endroits, les attributions de QUASI coïncident exactement avec celles faites par Bellanda. C'est pourquoi, afin de mener les calculs de dynamique, nous allons utiliser les attributions faites par Bellanda.
Dans le cas présent QUASI, avec 82% des données au départ, réalise l'attribution de 72% des résidus, soit 94% des données de départ. L'attribution est conforme à l'attribution publiée par Bellanda et al. La principale zone où les données manquent renferme 2 boucles très mobiles. La conformation de ces boucles évolue en fonction du ligand auquel la protéine est complexée. Dans la structure avec l'ADPNP, la boucle formée des résidus 99-119

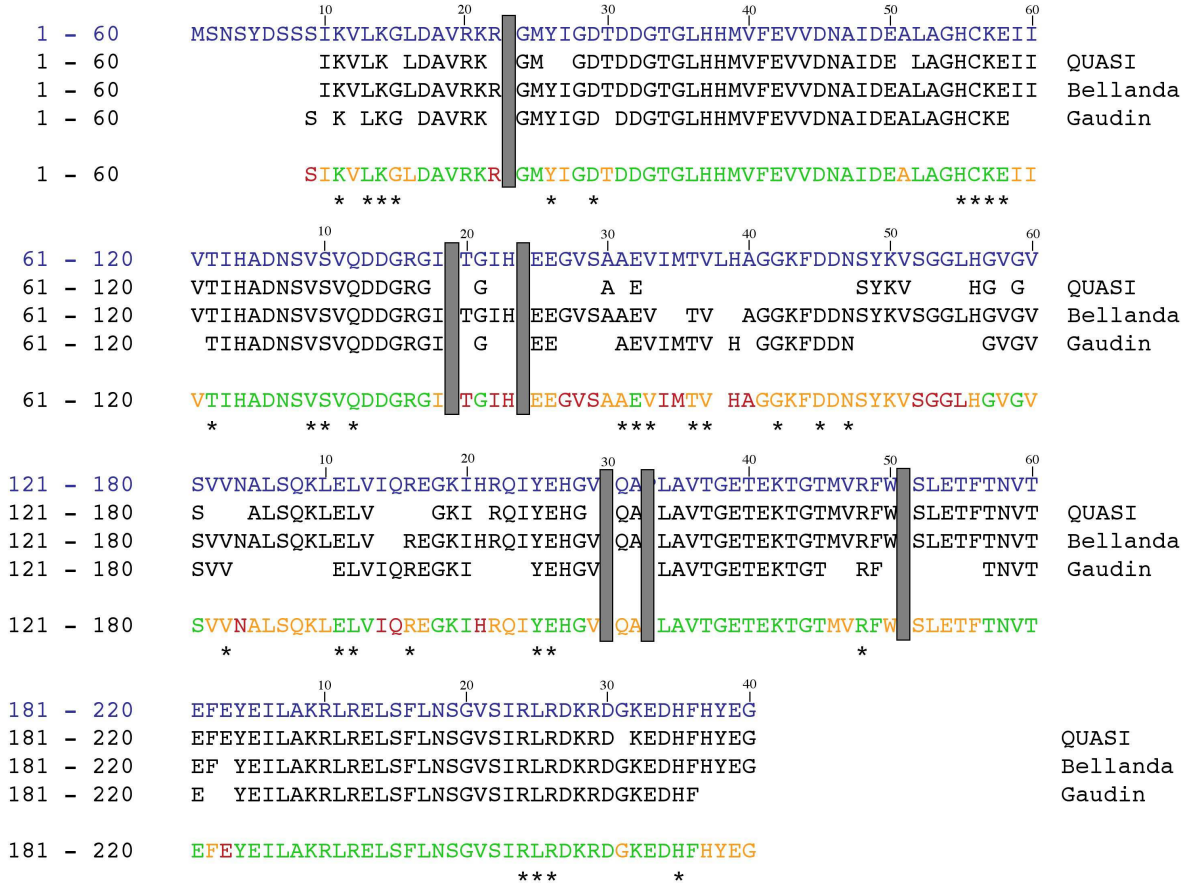

FIG. 3.9 – Attributions faites par les trois méthodes. La première ligne (bleue) est la séquence du fragment 24 kDa de la sous-unité B de l'ADN-gyrase. Les trois lignes noires représentent les résidus attribués par QUASI (première ligne), par Bellanda (deuxième ligne) et par le marquage spécifique (troisième ligne). La cinquième ligne récapitule les attributions. Lorsque la lettre est verte, le résidu correspondant est attribué par les trois méthodes. Lorsque la lettre est orange, seul deux méthodes l'attribuent ; le rouge indique un résidu attribué par une unique méthode. La dernière ligne ($\star$) indique les incohérences relevées entre l'attribution Gaudin et l'attribution de Bellanda et al..

est en position fermée. La structure avec la novobiocine présente une boucle ouverte. Cette dynamique particulière reste encore méconnue.

4

La dynamique du fragment 24 kDa de la sous-unité B de l'ADN-gyrase

Sommaire

4.1 Matériels et méthodes

4.1.1 Préparation des échantillons RMN

L'échantillon avec l'ADPNP est préparé en deux temps, d'abord l'ajout de $MgCl_2$ puis d'ADPNP (rapport 1 : 1 avec la concentration du fragment 24 kDa). L'échantillon avec la novobiocine est préparé à un rapport 1 : 1 avec la concentration du fragment 24 kDa de la sous-unité B de l'ADN-gyrase.

4.1.2 Acquisition des données de relaxation

Les mesures de relaxation ont été effectuées à trois températures : 298 K, 303 K et 310 K sur trois échantillons différents : le fragment 24 kDa de la gyrase Apo, en complexe avec de l'ADPNP (forme non hydrolysable de l'ATP) et en complexe avec de la novobiocine. Les données sont enregistrées à une fréquence proton de 500MHz. Pour les mesures des valeurs R_1 [Farrow *et al.*, 1994], 10 expériences ont été enregistrées avec des délais de relaxation fixés à 0, 202, 404, 605, 807, 1006, 1201, 1394 et 1604 ms. Le point à 404 ms a été enregistré une seconde fois afin d'estimer le niveau de bruit. Pour les valeurs de R_2, 11 expériences ont été enregistrées avec des délais de 0, 16, 31, 47, 63, 79, 94, 110, 126, 142 et 157 ms. La séquence utilisée présente un délai de 1200 μs entre les impulsions 180 °sur le noyau N de 111,3 μs. Les NOE hétéronucléaires ont été mesurées à l'aide de deux expériences, avec et sans saturation du proton [Kay *et al.*, 1989]. Le temps de saturation utilisé est de 4 secondes précédé par un temps de relaxation de 2,5 secondes.

4.1.3 Extraction et traitement des données de relaxation

Les spectres ont été traités avec NMRPipe [Delaglio *et al.*, 1995]. Les valeurs des intensités sont extraites à l'aide du programme XEASY. Le calcul des temps de relaxation est effectué par des scripts MATLAB. Les courbes de décroissance ont été ajustées par une fonction exponentielle à 2 paramètres (la valeur de l'intensité à l'origine et la vitesse de relaxation). Les incertitudes sur les valeurs des paramètres de relaxation R_1, R_2 et NOE ont été estimées à l'aide du logiciel MATLAB.

4.1.4 Interprétation des paramètres

Fonction de densité spectrale

La qualité des données de relaxation est évaluée en premier lieu en calculant (Equation : 4.1) les valeurs de la fonction de densité spectrale aux fréquences 0, ω_N et $\langle\omega_H\rangle$.

$$[J] = A^{-1}\,[R] \tag{4.1}$$

R est un vecteur colonne contenant les vitesses de relaxation longitudinale, transverse et croisée, de l'azote. A est une matrice 3x3 qui contient les coefficients des relations linéaires qui existe entre la valeur de la fonction de densité spectrale [Peng Wagner, 1992] aux 3 fréquences et les vitesses de relaxation.
Les trois temps de relaxation sont utilisés pour cartographier la fonction de densité spectrale aux fréquences 0, ω_N, $\langle\omega_H\rangle$. Puis, une approche plus quantitative, dans le cadre du modèle de Lipari-Szabo a été réalisé à l'aide du programme TENSOR [Dosset *et al.*, 2000].

Les modéles utilisés par TENSOR

La fonction de densité spectrale utilisée dans TENSOR est de la forme suivante :

$$J(\omega) = S_2^2\left(S_1^2\frac{\tau_c}{1+(\omega\tau_c)^2} + (1-S_1^2)\frac{\tau'}{1+(\omega\tau')^2}\right)$$

avec $\tau' = \frac{\tau_c\tau_i}{\tau_c+\tau_i}$(4.2)

Le terme S_1^2 décrit l'amplitude des mobilités internes rapides et τ' est le temps de corrélation effectif pour les mobilités internes rapides. Lorsqu'une mobilité interne est plus lente, le mouvement est caractérisé par $S_1^2 = S_s^2$ et τ_i. Les cinq modèles suivants (Tableau : 4.1) sont testés itérativement sur les données. Le plus simple en premier, puis les modèles deviennent de plus en plus complexes jusqu'à ce que le modèle reproduise les données de relaxation avec un taux de confiance de 95%.

Modèle	Paramètres
1	S^2
2	S^2, τ_i
3	S^2, R_{ex}
4	S^2, τ_i, R_{ex}
5	S_1^2, τ_i=τ_s

TAB. 4.1 – Modéles utilisés par TENSOR pour déterminer la mobilité interne.

-**Modèle 1** Le temps de corrélation τ_i est très court (τ_i<20ps), et la fonction de densité spectrale se réduit à l'équation 4.3.

$$J(\omega) = \frac{2}{5}\frac{S^2\tau_c}{1+\omega^2\tau_c^2} \tag{4.3}$$

-**Modèle 2** Le temps de corrélation τ_i est moins court et il est pris en compte comme variable dans l'ajustement classique de Lipari-Szabo (Equation : 4.4).

$$J(\omega) = \frac{2}{5}\left(S^2\frac{\tau_c}{1+(\omega\tau_c)^2} + (1-S^2)\frac{\tau_e}{1+(\omega\tau_e)^2}\right) \tag{4.4}$$

-**Modèle 3** La fonction de densité spectrale à la même forme que celle utilisée pour le modéle 1 avec un terme R_{ex}. Ce terme s'additionne au R2.

-**Modèle 4** La fonction de densité spectrale à la même forme que celle utilisée pour le modéle 2 avec un terme R_{ex}. Ce terme s'additionne au R2.

-**Modèle 5** Formalisme étendu de Lipari-Szabo [Clore *et al.*, 1990]. Il y a deux mobilités internes (Equation : 4.5), une très rapide et une plus lente.

$$J(\omega) = \frac{2}{5}S_2^2\left(S_1^2\frac{\tau_c}{1+(\omega\tau_c)^2} + (1-S_1^2)\frac{\tau_e}{1+(\omega\tau_e)^2}\right) \tag{4.5}$$

-**Modèle 6** Dans le cas où aucun modèle ne convient, TENSOR note le modèle 6 puis il donne les meilleurs résultats obtenus ainsi que le nom du modèle utilisé.

4.2 Les temps de relaxation dans la protéine libre

Le comportement dynamique du fragment 24 kDa de la sous-unité B de l'ADN-gyrase est étudié en analysant la relaxation des corrélations observées dans le spectre 2D ^{1}H-^{15}N HSQC. Un premier aperçu de la dynamique de la protéine libre est disponible après l'étude des temps de relaxation.

4.2.1 Les données de relaxation à 298 K

Les données de relaxation ne sont pas complètes ; il manque les 9 premiers résidus ainsi que les résidus 20, 22, 23, 51, 79, 91 à 95, 98 à 100, 102 à 106, 115, 123, 124, 134, 135, 141, 149, 150, 153, 160 et 171. En effet, certains protons n'ont pas pu être analysés du fait de problèmes de recouvrement de pics ou de bruit trop important.

FIG. 4.1 – Temps de relaxation R_1 de la protéine libre à 298 K.

La figure 4.1 montre que le profil des valeurs des constantes de relaxation longitudinales est relativement peu contrasté avec une moyenne de $1{,}46 \pm 0,078\ s^{-1}$. Des valeurs plus importantes de R_1 sont observées pour les résidus situés en N-terminal ainsi qu'autour des résidus 100-120 qui est une des boucles située à proximité du site de fixation des antibiotiques. Les résidus 82, 101 et 112 sont des exceptions à ces deux règles d'ensemble.

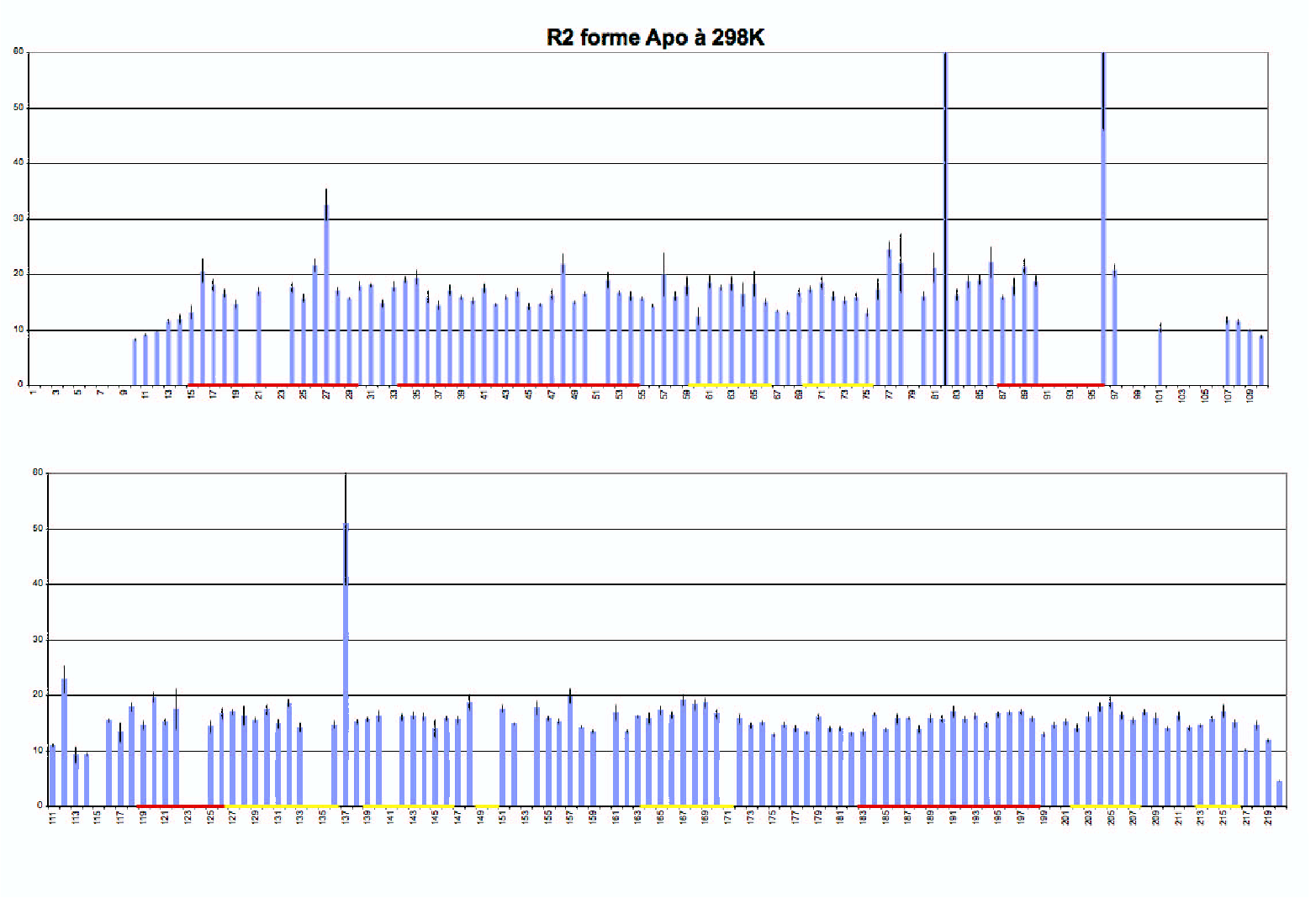

FIG. 4.2 – Temps de relaxation R_2 de la protéine libre à 298 K.

La figure 4.2 montre que le profil des valeurs des constantes de relaxation transverses est relativement peu contrasté avec une moyenne de 16,7 $\pm 1,09 s^{-1}$. La majorité des valeurs fluctuent entre 10 et 20. Les résidus 16, 26, 27, 46, 77, 78, 81, 82, 86, 89, 96, 97, 112 et 137 ont des valeurs de R_2 supérieurs à 20 s^{-1}, ceci est dû à la présence d'échange conformationnel. Les résidus 10, 11, 12, 109, 110, 113, 114 et 220 ont des valeurs de R_2 inférieurs à 10. Les erreurs des résidus 82, 96 et 136 sont très élevées.
Dans le rapport R_1/R_2, seuls quatre résidus sont atypiques : les résidus 27, 82, 96 et 137. La moyenne du rapport R_1/R_2 est de 11,11.

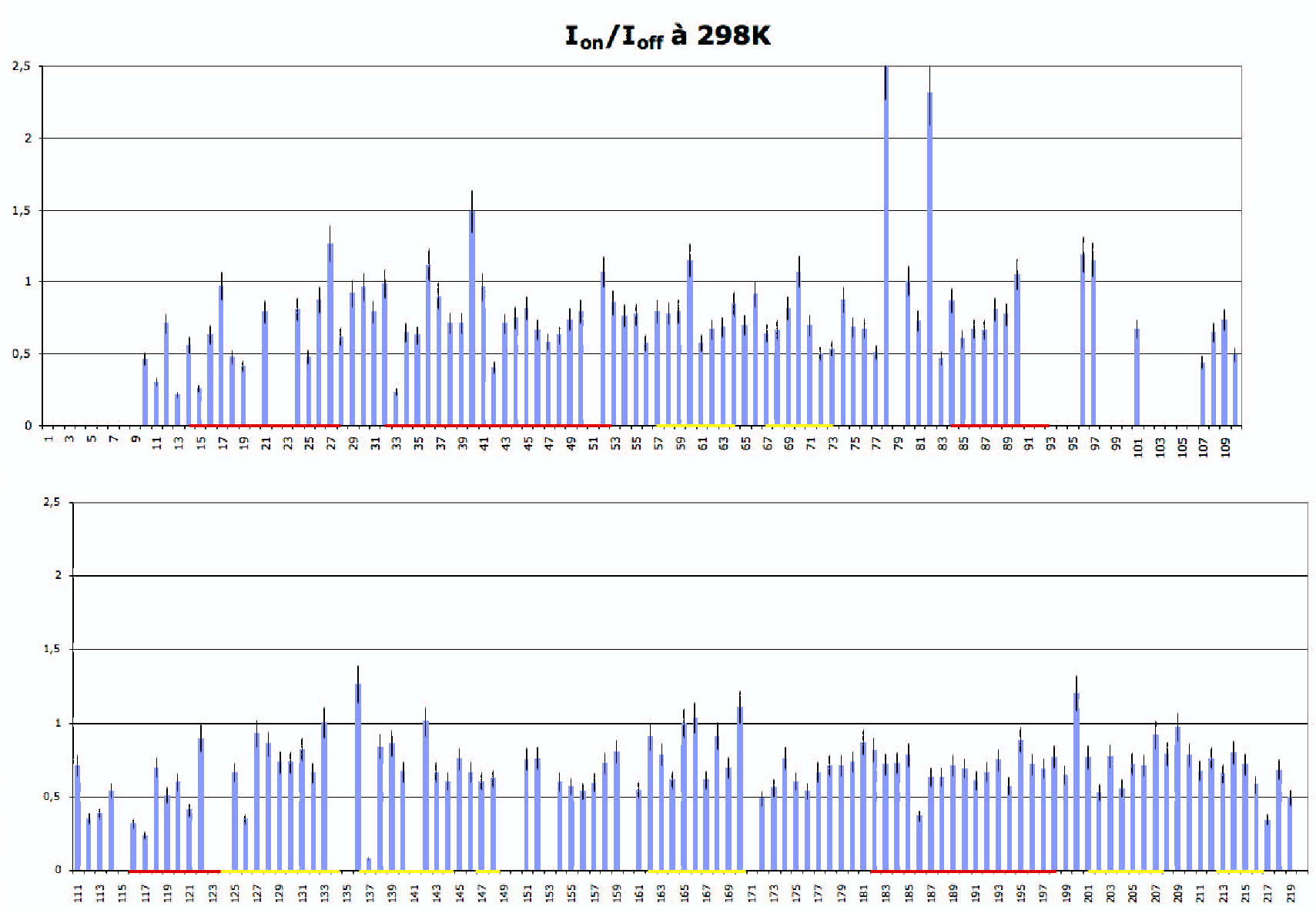

FIG. 4.3 – Rapport des intensités NOE de la protéine libre à 298 K.

La valeur moyenne du rapport entre l'intensité saturée et l'intensité non saturée est de 0,7 et l'erreur moyenne est de 0,14. Les résidus 27, 36, 40, 52, 60, 70, 78, 81, 90, 96, 97, 136, 166, 170 et 200 ont un rapport supérieur à 1, ce qui semble indiquer que l'erreur est supérieure à l'erreur statisitique estimée (la valeur maximale théorique de ce rapport est de 0,82 à 500MHz).

4.2.2 Evolution des temps de relaxation avec la température

Afin d'étudier l'influence de la température sur les paramètres dynamiques, les valeurs de relaxation ont été enregistrées à 3 températures : 298 K, 303 K et 310 K. Ces mesures permettent à la fois d'avoir une idée de l'erreur expérimentale et de voir l'évolution de la protéine lorsque la température varie.

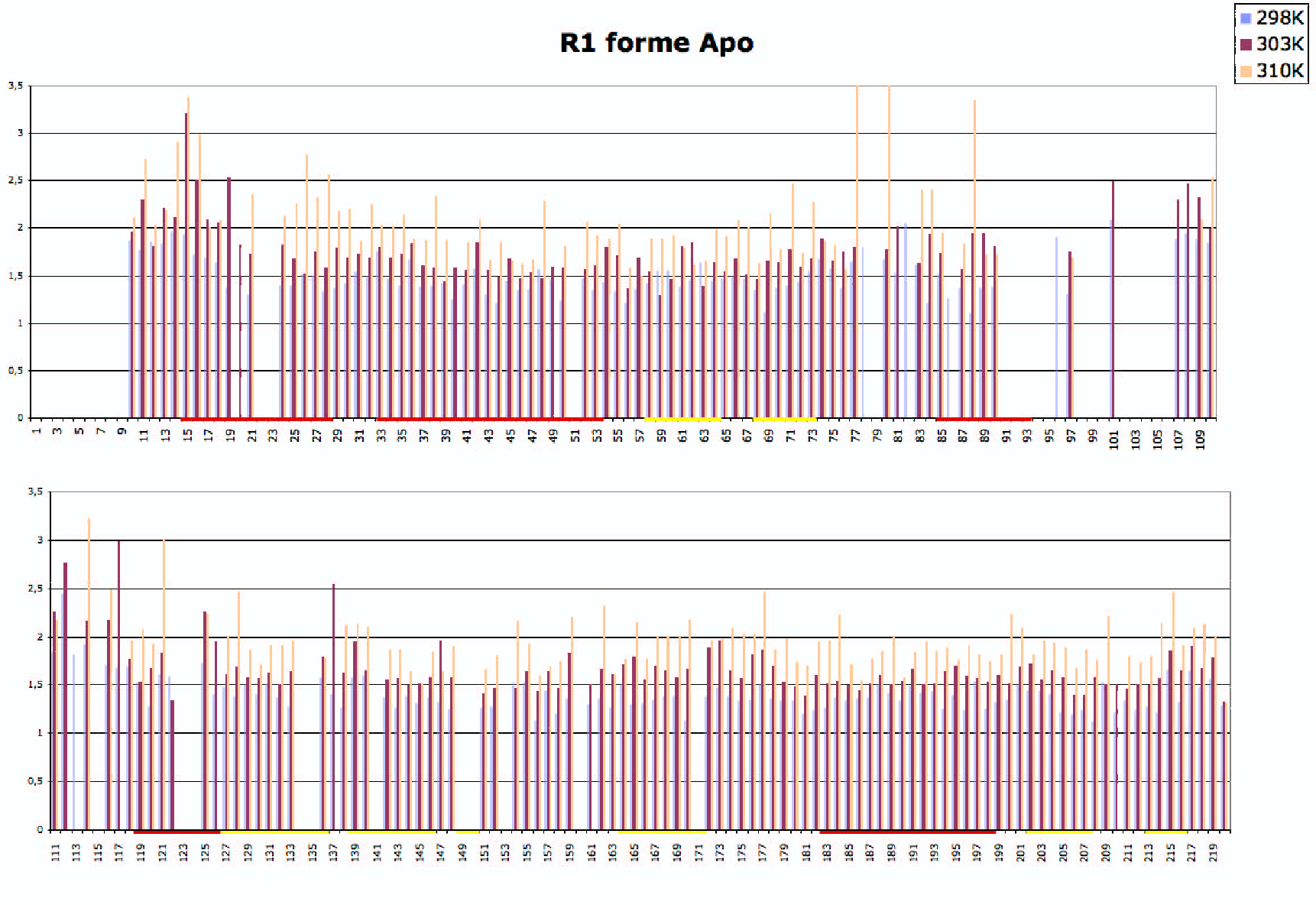

FIG. 4.4 – Temps de relaxation R_1 de la protéine libre à 3 températures : 298 K, 303 K et 310K.

Les valeurs de R_1 augmentent régulièrement avec la température. L'écart observé entre les valeurs à 298 K et 303 K est petit pour les groupes amides situés dans des éléments de structures secondaires. La comparaison des données montre également que le profil des valeurs en fonction de la séquence est globalement conservé à chaque température, ceci atteste d'une bonne reproductibilité des expériences. Cet écart entre 303 K et 310 K augmente de façon significative.

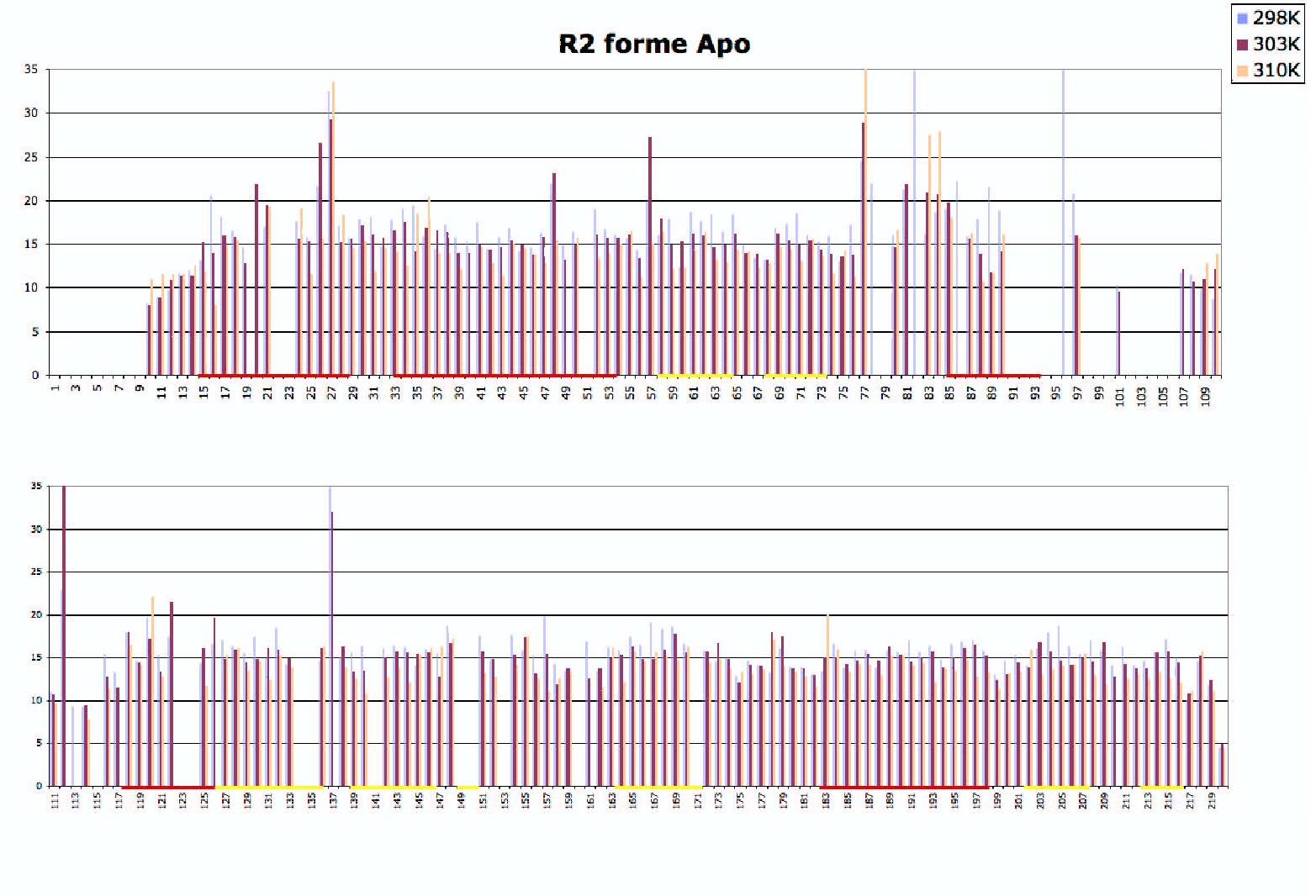

FIG. 4.5 – Temps de relaxation R_2 de la protéine libre à 3 températures : 298 K, 303 K et 310K.

Les valeurs de R_2 diminuent lorsque la température augmente ce qui est conforme au comportement attendu puisque la contribution principale à la relaxation transverse provient du temps de corrélation de rotation. Les résidus 10, 11, 12, 13, 14, 24, 26, 27, 28, 35, 36, 37, 48, 60, 75, 77, 80, 81, 83, 94, 85, 90, 109, 110, 112, 120, 122, 125, 126, 137, 145, 146, 147, 148, 155, 163, 175, 178, 183, 184, 202 et 218 présentent des tendances différentes. Dans ces cas, les valeurs à 310 K sont souvent très éloignées des deux autres jeux de données.

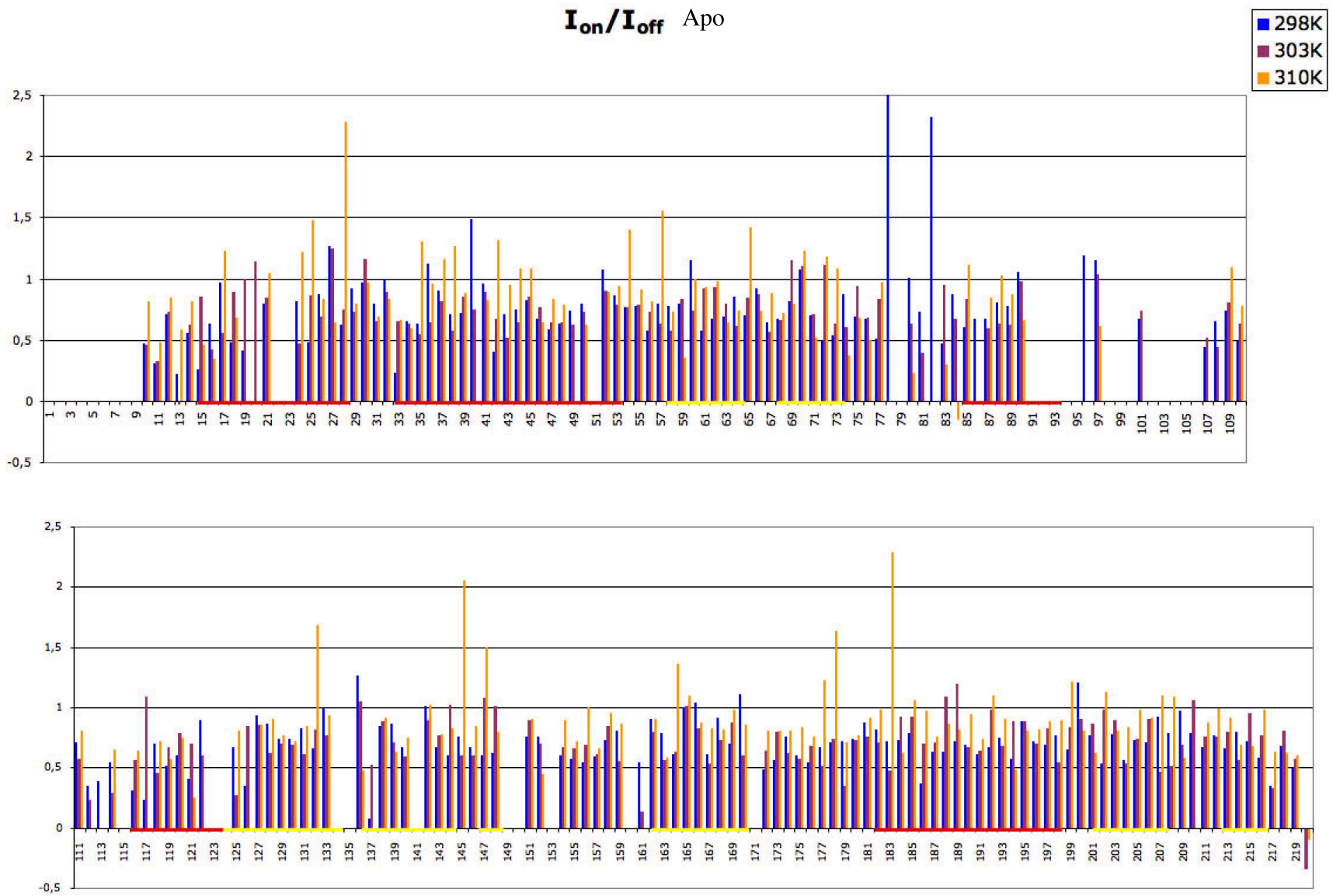

FIG. 4.6 – Evolution du rapport I_{on}/I_{off} de la protéine libre à 3 températures : 298 K, 303 K et 310K.

L'analyse de l'évolution des transferts NOE en fonction de la température montre que les jeux de données obtenus à 298 K et 303 K sont très similaires alors que des différences importantes sont observées à 310 K. Dans ce dernier cas, de nombreux résidus montrent des valeurs supérieures à 0,82 ce qui indique que l'erreur qui entache ce jeu de données est supérieure à celle estimée initialement. D'autre part, la similarité observée entre les jeux enregistrés à 298 K et 303 K semble indiquer que les mouvements rapides $<$100 ps ne sont pas sensibles à la température. Pour les résidus 20, 21, 25, 57, 77, 83, 84, 112, 122, 125, 126, 155, 173, 178, 179, 203 et 209 la valeur de R_2 augmente avec la température. Cela suggère que ces résidus ont des valeurs R_{ex} non nulles.
L'évolution de cette composante à donc été étudiée en fonction de la température.

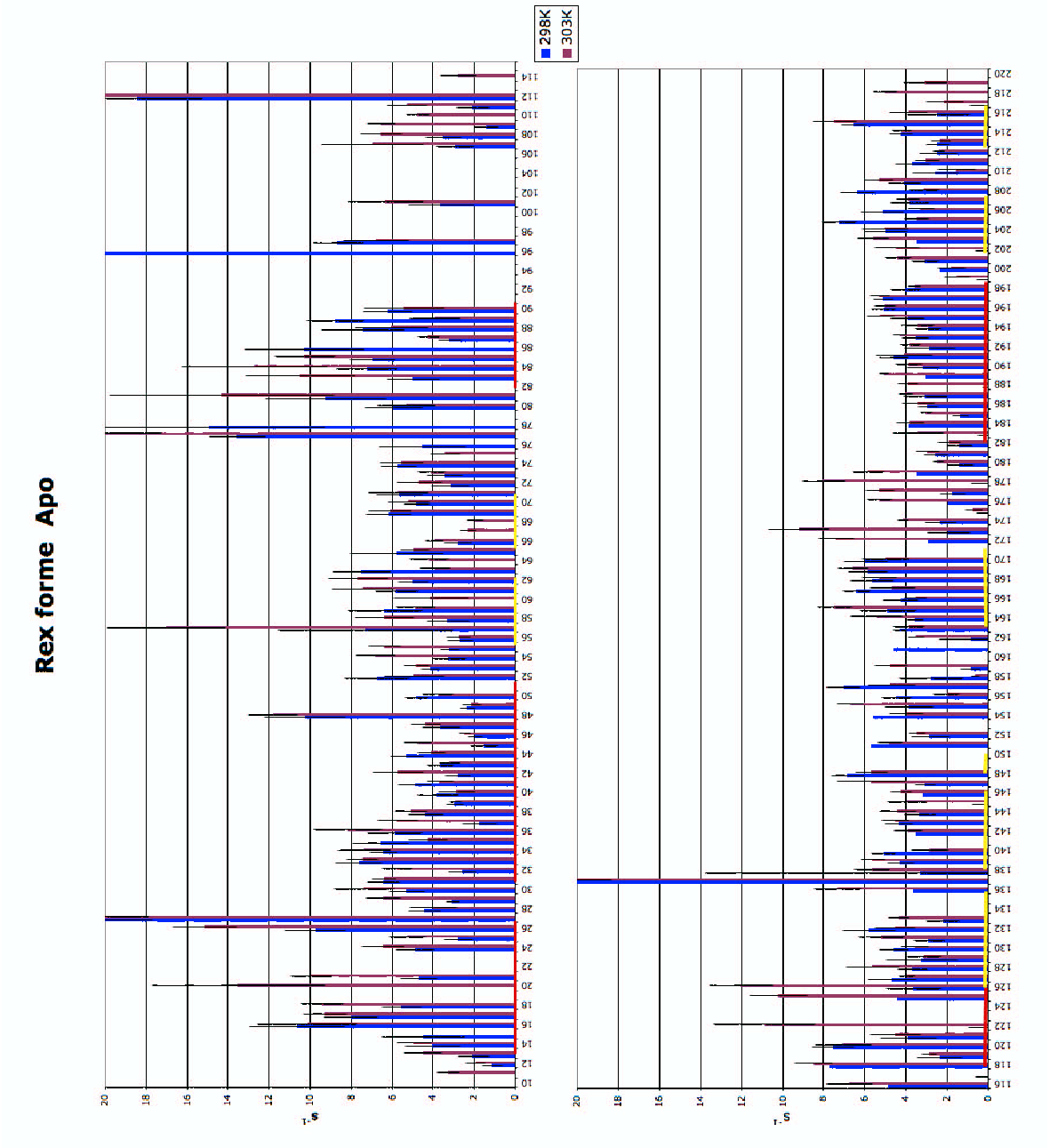

FIG. 4.7 – Evolution de la composante d'échange R_{ex} de la protéine libre à 2 températures : 298 K et 303 K.

Dans la figure 4.7, on voit nettement qu'une majorité des résidus ont des valeurs R_{ex} non nulles. Les profils à 298 K et 303 K sont conservés. Les résidus dans les boucles (54-58 ; 81-85 ; 100-116 ; 172-179) voient leur composante d'échange augmenté de façon significative lorsque la température augmente. Ceci est caractéristique d'un mouvement lent lié à une barrière d'activation élevée. L'observation faite sur ces deux dernières boucles est en accord avec les valeurs de relaxation transverse qui augmentent également.

4.2.3 Etude de la densité spectrale en fonction de la température.

Une première analyse des données de relaxation enregistrées à 298 K, 303 K et 310 K nous a conduit à écarter le jeu à 310 K qui présente manifestement des erreurs importantes. L'analyse du tenseur de diffusion rotationnelle du fragment 24 kDa a donc été réalisée à partir des jeux enregistrés à 298 et 303 K. Une représentation de ces deux jeux est montrée dans la figure 4.9(a) sous la forme de fonctions de densités spectrales. L'analyse des valeurs obtenues pour le fragment 24 kDa à 298 K montre que la majorité des points se regroupent dans une région proche de la courbe théorique pour une valeur de $\tau_c = 0{,}5.10^{-8} \times \frac{5}{2}$ ns. Ces points représentent les éléments rigides de la protéine qui ressentent essentiellement le mouvement de réorientation de la molécule. Deux autres groupes de points indiquent des mouvements particuliers. Le groupe marqué I rassemble des acides aminés pour lesquels une valeur importante du J(0) est obtenue. Ce comportement peut être expliqué par des contributions d'échanges conformationels dans l'échelle de temps μs-ms.
Un deuxième groupe (II) est caractérisé par une diminution de J(0) et des valeurs importantes de $J(\omega_N)$. Ces résidus expérimentent des mouvements à l'échelle de temps d'une centaine de ps.

Ce comportement révèle la présence de mouvements complexes sur des échelles de temps allant de la μs-ms à la centaine de picosecondes. Cette première analyse qualitative des données de relaxation obtenues sur le fragment 24 kDa de la gyrase indique que le squelette de la protéine est animé de mouvements complexes, s'étalant sur une gamme de temps très étendue. Cette analyse préalable nous permet d'identifier un certain nombre de résidus dont les données de relaxation va nécessiter l'utilisation des modèles particuliers.

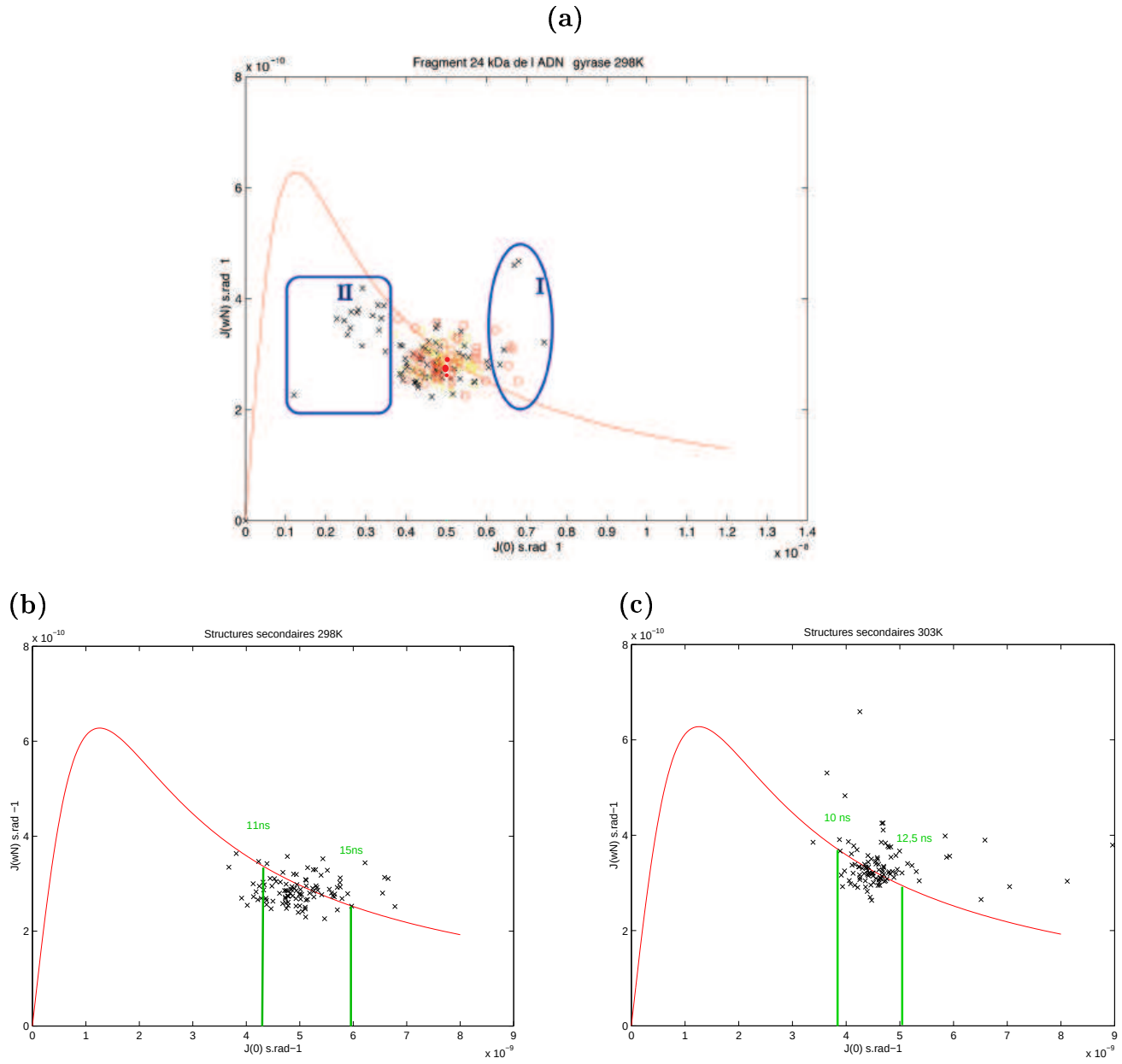

FIG. 4.8 – Tracés de J(ω_N) (en s.rad^{-1}) en fonction de J(0) (en s.rad^{-1}) pour le fragment 24 kDa de la sous-unité B de l'ADN-gyrase. La courbe rouge représente les valeurs de J(0) et J(ω_N) obtenues en considérant une exponentielle unique pour la fonction de corrélation. Cette courbe paramétrique en τ_c permet de visualiser les valeurs limites pour les fonctions de densité spectrale. (a) le fragment est sous sa forme libre à 298 K. Les croix noires indiquent les résidus impliqués dans les boucles, les ronds rouges indiquent ceux dans les hélices-α et les ronds jaunes sont les résidus dans les brins-β. La seconde ligne présente les résidus impliqués dans une des structures secondaires de la protéine. Le cas (b) est enregistré à 298 K. Le cas (c) est enregistré à 303K.

L'analyse en fonction de la température fait apparaitre deux caractéristiques :

- Lorsque la température augmente, la dispersion du nuage de points s'accentue sous l'effet de contributions d'échange. Cette observation explique que les jeux de données enregistrés à 310 K sont entachés d'erreurs importantes. Cette tendance est déjà perceptible en passant de 298 à 303 K.

- La dispersion des points est un indicateur de la forte anisotropie de la diffusion rotationnelle de la molécule.

Nous constatons également que la majorité des points doit pouvoir se prêter à une analyse plus quantitative dans le cadre du modèle de Lipari-Szabo. L'analyse a été réalisée à l'aide du programme TENSOR.

4.3 La diffusion rotationnelle de la molécule

Une première analyse de la plage des valeurs de temps de corrélation possible à partir des fonctions de densité spectrale indique qu'elles se situent entre 11 et 15 ns à 298 K et 10 et 12,5 ns à 303 K (Figure : 4.9(b) et (c)). Ces valeurs sont légèrement inférieures, mais restent compatibles avec les temps de corrélation (isotropes) donnés à ces températures par l'approximation de Daragan [Daragan Mayo, 1997] (16,46 et 14,16 ns) (Equation : 4.6).

$$Testlatemprature, Nestlenombredersidusdanslaprotinecible.\tau_c = \left(\frac{9,8 \times 10^{-}3}{T}\right) \times \exp\left(\frac{2416}{T}\right) \times N^{0,93} \quad (4.6)$$

Les valeurs du rapport R_1/R_2 mesurées pour les résidus pésents dans les structures secondaires ont été utilisées pour calculer les valeurs du tenseur de rotation.
Pour les mesures faites à 298 K, le tenseur de diffusion est de forme ellipsoïdale complètement anisotrope : l'axe principal est de $0,17.10^{-8}$ s^{-1}, D_Y de $0,15.10^{-8}$ s^{-1} et D_X de $0,11.10^{-8}$ s^{-1}. Ces valeurs se comportent conformément aux rapports des valeurs du tenseur d'inertie calculées à partir des coordonnées de la molécule. L'axe principal (Z) du tenseur de diffusion est perpendiculaire (Figure : 4.9(a)) aux brins-β et aux deux hélices principales (des résidus 33 à 53 et 184 à 198). L'axe X coupe les brins-β à la perpendiculaire et l'axe Y se trouve dans le plan formé par ces mêmes brins-β.

L'axe X du tenseur d'inertie se place dans le plan XY du tenseur de diffusion. Un

calcul de type Monte-Carlo de 100 cycles permet de visualiser la dispersion des axes de ce dernier tenseur. Ces calculs sont effectués à partir des erreurs expérimentales des données de relaxation. Les résultats obtenus sont les suivants :

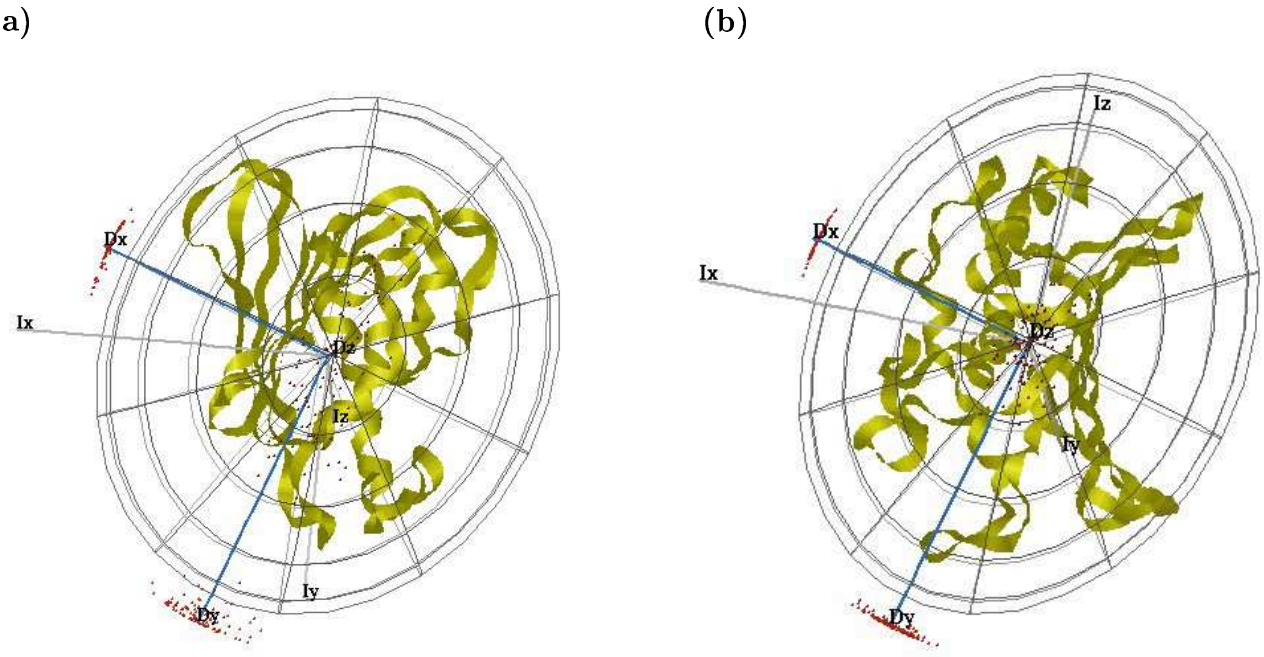

FIG. 4.9 – Figures regroupant le tenseur de diffusion de la protéine à 298 K (a) et à 303 K (b). Les points rouges indiquent les impacts des axes du tenseur d'inertie, obtenus pour chaque cycle de calcul Monte-Carlo.

La dispersion observée pour les axes du tenseur d'inertie est faible. Seul l'axe Z est légèrement dégénéré (les points rouges sont étalés autour de l'axe D_Z). Lorsque les mêmes calculs sont réalisés à la température de 303 K, le tenseur de diffusion est légèrement différent, l'axe principal reste aligné le long du feuillet β mais l'orientation en est légèrement modifiée ($\pi/4$ rad). Il est possible que cette réorientation du tenseur soit liée à l'erreur sur les données de relaxation, ou à la source des contributions d'échange observées. Celles-ci peuvent en effet être dues soit à un équilibre conformationnel suffisament important pour modifier la diffusion globale de la molécule, soit à un équilibre dimère-monomère. Il a en effet été montré [Pfuhl *et al.*, 1999] que la présence d'une faible fraction de dimères peut être suffisante pour modifier significativement les valeurs des vitesses de relaxation. Les dimensions du tenseur évoluent également : D_X de $0{,}12.10^{-8}$ s^{-1}, D_Y de $0{,}16.10^{-8}$ s^{-1}, D_Z de $0{,}18.10^{-8}$ s^{-1} ; il s'élargit légèrement dans les trois dimensions. Le temps de corrélation est maintenant égal à 11,1 ns. La température augmente et la protéine incorpore l'énergie sous la forme de mouvement et le temps de corrélation diminue.

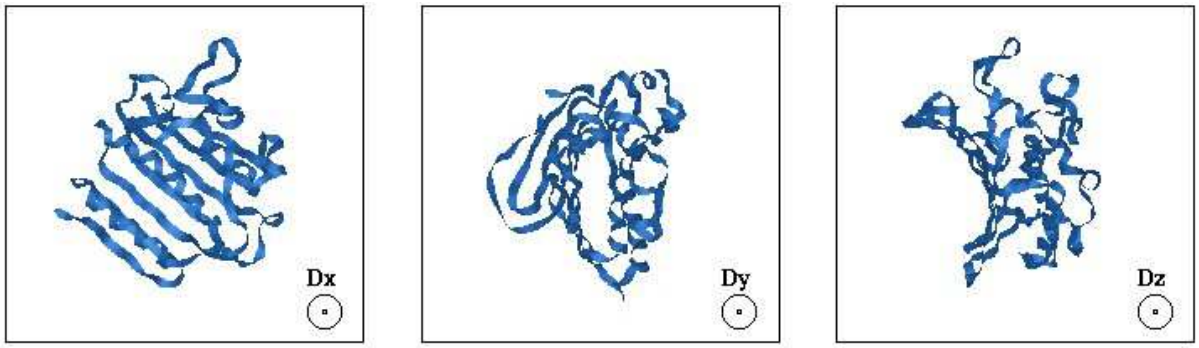

FIG. 4.10 – Vues de la protéine, obtenues le long des trois axes du tenseur de diffusion.

Lorsque l'on compare les dispersions obtenues pour les deux températures (298 K et 303 K), on remarque que pour chaque axe, elle est plus petite à 303 K (Figure : 4.9(b)).

Les tenseurs de diffusion observés aux deux températures ne changent pas beaucoup ; on note une légère augmentation des dimensions entre chaque température. Pour la dispersion des tenseurs d'inertie, les évolutions observées ne sont pas linéaires. Alors qu'à 303 K les axes sont plutôt mieux définis qu'à 298 K (Figure : 4.9(a)).

Lorsqu'on regarde les temps de corrélation obtenus par TENSOR dans ces calculs de mobilité globale, il est notable que plus la température augmente plus la différence avec les τ_c, approximés grâce au rapport R_1/R_2, augmente. L'évolution du temps de corrélation est conforme à nos attentes : il diminue lorsque la température augmente.

4.4 La dynamique interne du fragment 24 kDa de la sous-unité B de l'ADN-gyrase

Une fois que le tenseur de diffusion est connu, la mobilité interne peut-être étudiée avec le formalisme de Lipari-Szabo [Lipari Szabo, 1982a, Lipari Szabo, 1982b]. Dans ce formalisme les valeurs du paramètre d'ordre S^2 sont caractéristiques de l'amplitude de la mobilité interne de la protéine et le temps de corrélation τ_i est le temps caractéristique à la mobilité.

Les résultats obtenus à 298 K montrent un mobilité réduite ($S^2 > 0{,}7$) pour les résidus formant des structures secondaires (Figure : 4.11). La majorité des vecteurs a été traitée avec le modèle 4 qui indique que la prise en compte d'une contribution d'échange est nécessaire pour quasiment l'ensemble des données. Malgré l'absence des nombreuses données dans ces régions, la présence d'une dynamique plus rapide est clairement visible dans la région 95-115 avec une ensemble de paramètres d'ordre inférieurs à 0,5.

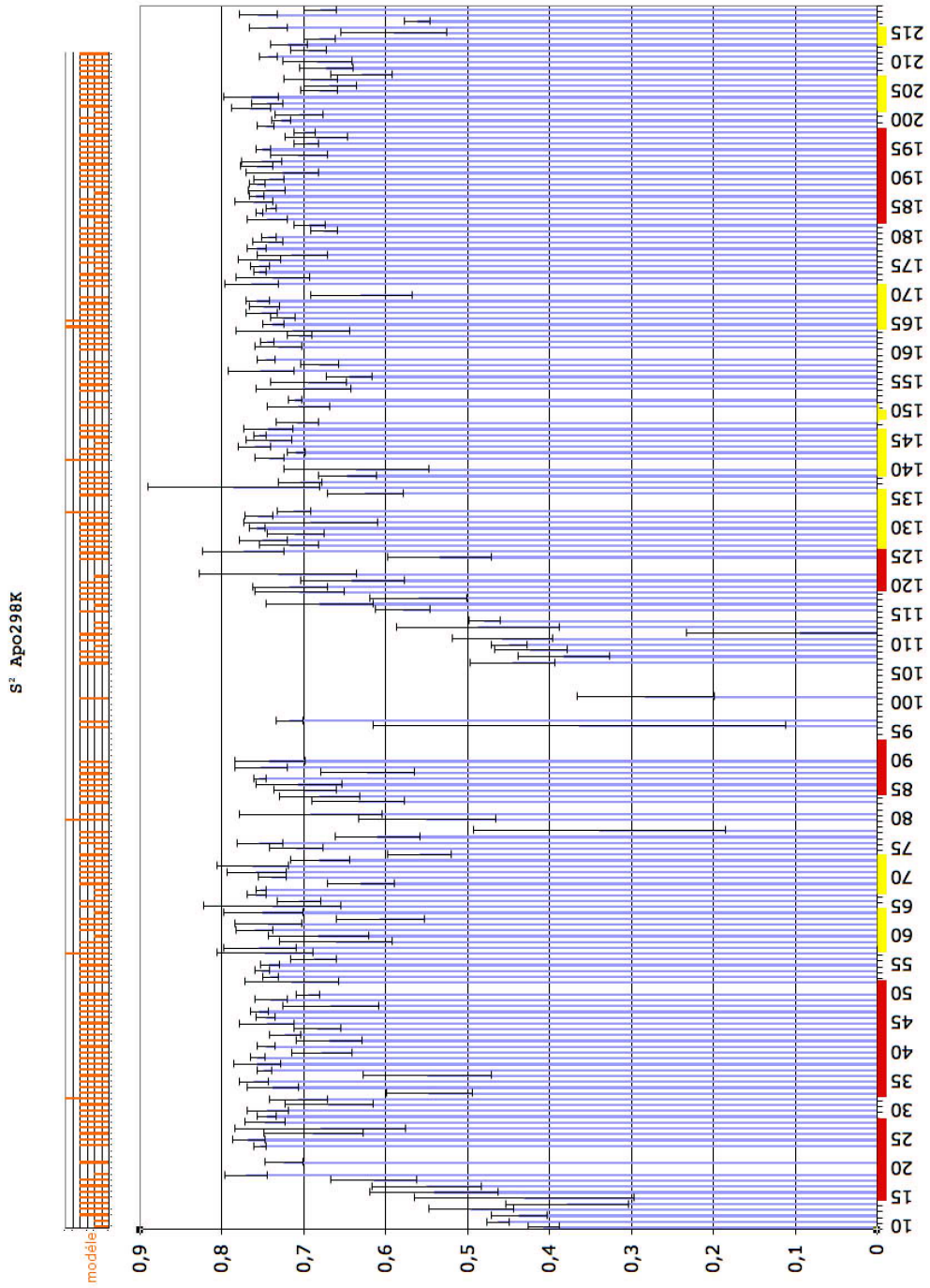

FIG. 4.11 – Tracé des valeurs de S^2 en fonction des résidus de la séquence. Les valeurs et les erreurs (barres noires verticales) sont calculées par TENSOR. Les rectangles rouges sont les hélices-α, les rectangles jaunes sont les brins-β. Au dessus du graphique principal, l'histogramme orange présente le modèle utilisé pour chaque résidu, lors du calcul du paramètre d'ordre.

Les modéles utilisés nécessitent tous l'incorporation d'une constante d'échange (R_{ex}). L'utilisation de ces modèles conforte l'interprétation faite de la fonction de densité spectrale précédemment.

On voit que les 20 premiers résidus, les résidus 33, 36, 73, 79, 96, 136, 215 et 217 ainsi que la zone centrale (103 - 121) sont très flexibles sur l'échelle de temps ps-ns avec des S^2 inférieurs à 0,7. Cette dernière zone est particulièrement mobile et correspond également à l'endroit où les données sont les plus pauvres. La grande mobilité de cette zone, sur plusieurs échelles de temps, explique la difficulté que nous avons eu à observer des signaux RMN. L'autre particularité (Figure : 4.12) de ces résultats se trouve à la fin de la dernière hélice α (positions 195-197) et aux positions 205-215 (les 2 premiers et les 3 derniers résidus de ce groupe se trouvent dans des brins-β ; les autres de trouvent dans une boucle). Ces 11 résidus successifs présentent une baisse des valeurs de S^2 alors que la majorité d'entre eux se trouvent dans un élément de structure secondaire.

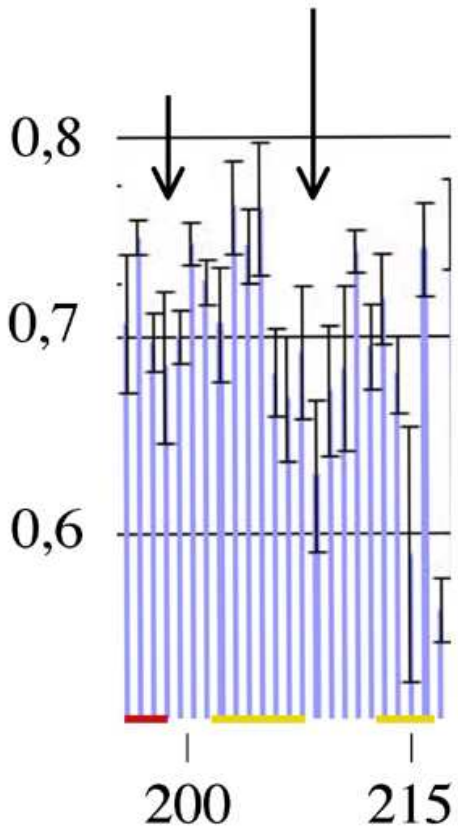

FIG. 4.12 – Agrandissement des zones où les valeurs du S^2 baissent.

D'autres baisses de S^2 sont visibles mais elles correspondent à des boucles, ou lorsqu'elles se trouvent dans une structure secondaire, elles sont isolées (telle qu'en position 35).

Lorsque l'on augmente la température, les valeurs de S^2 diminuent indiquant que la mobilité sur une échelle de temps ps-ns augmente (même dans les structures secondaires). Ceci se vérifie partout sauf pour quelques cas isolés et pour les résidus identifiés à 298 K comme ayant des S^2 bas. Les valeurs calculées aux températures supérieures ne reflètent pas la baisse du paramètre d'ordre pour les résidus 205 à 213.

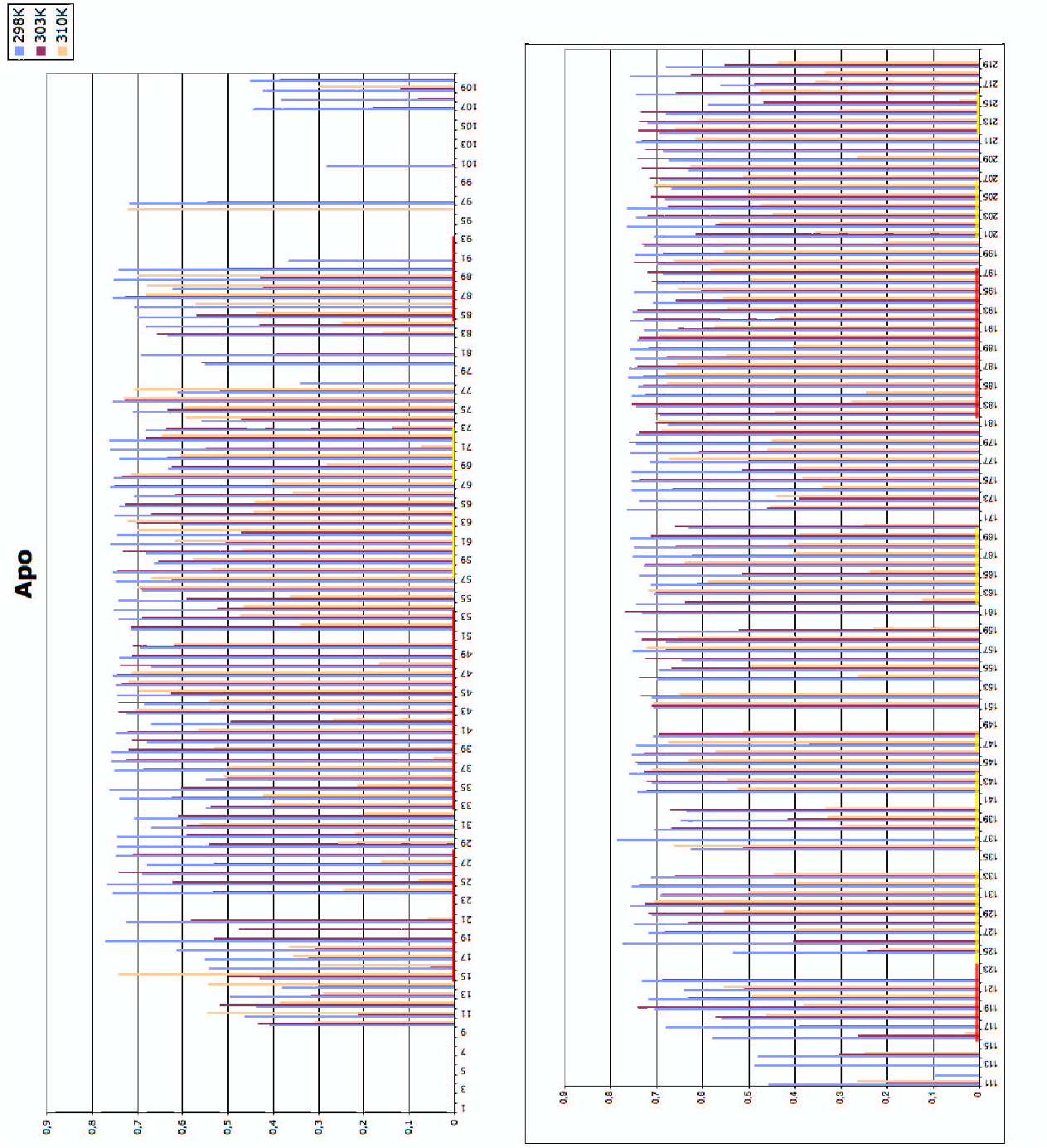

FIG. 4.13 – Comparaison des S^2 observés aux deux températures étudiées : 298 K et 303 K.

Rapportées aux structures tridimensionnelles (Figure : 4.14), ces valeurs indiquent une mobilité interne croissante au niveau de chaque résidu.

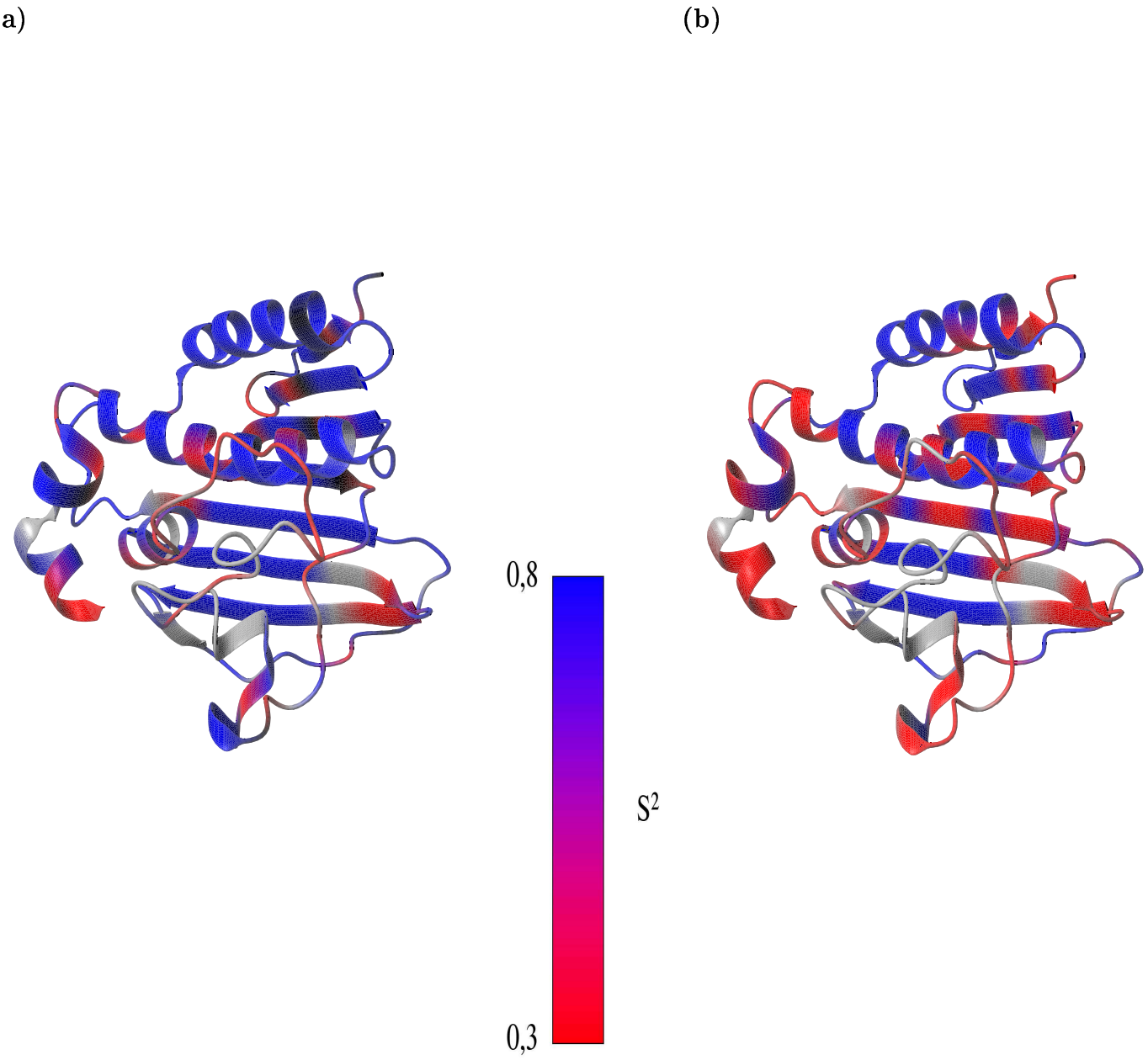

FIG. 4.14 – Représentations en ruban du fragment 24 kDa de la sous-unité B de l'ADN-gyrase. Les mobilités sont reportées sur la structure. Comme l'indique léchelle, les zones rigides sont en bleu et les zones plus mobiles en rouge. Les résidus dont les données ne sont pas disponibles sont en gris. (a) à 298K, (b) à 303K.

4.5 Effet des ligands sur le fragment 24 kDa de l'ADN-gyrase.

Deux autres échantillons ont été étudiés suivant le procédé précédant : un échantillon où la protéine est complexée avec de l'ADPNP et un où elle est complexée avec la novobiocine. Seules les données obtenues à 298 K et 303 K sont utilisables. Les structures cristallographiques [Tsai *et al.*, 1997] placent les deux ligands de manières différentes. Leurs cycles principaux respectifs se recouvrent et forment des liaisons hydrogène avec des résidus communs : Asn 46, Asp 73 et Thr 165 mais le reste des molécules se place (Figure : 4.15) à des endroits opposés. L'ADPNP se lie également aux résidus Tyr 5, Lys 103, Tyr 109,

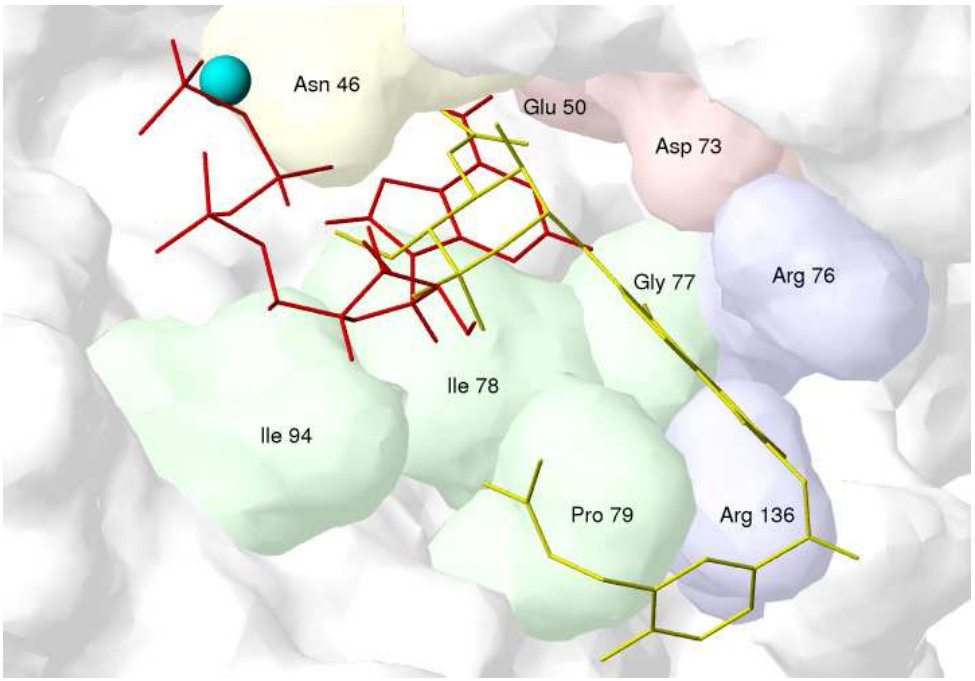

FIG. 4.15 – Site actif du fragment 24 kDa de la sous-unité B de l'ADN gyrase. Le ligand rouge est l'ADPNP ; le jaune est la chlorobiocine (famille des coumarines). La sphère de couleur cyan est le magnésium. [Schechner *et al.*, 2004]

Gly 114, Gly 117 et Gly 119. La novobiocine fait des liaisons avec les résidus Val 43, Gly 77 et Arg 136.

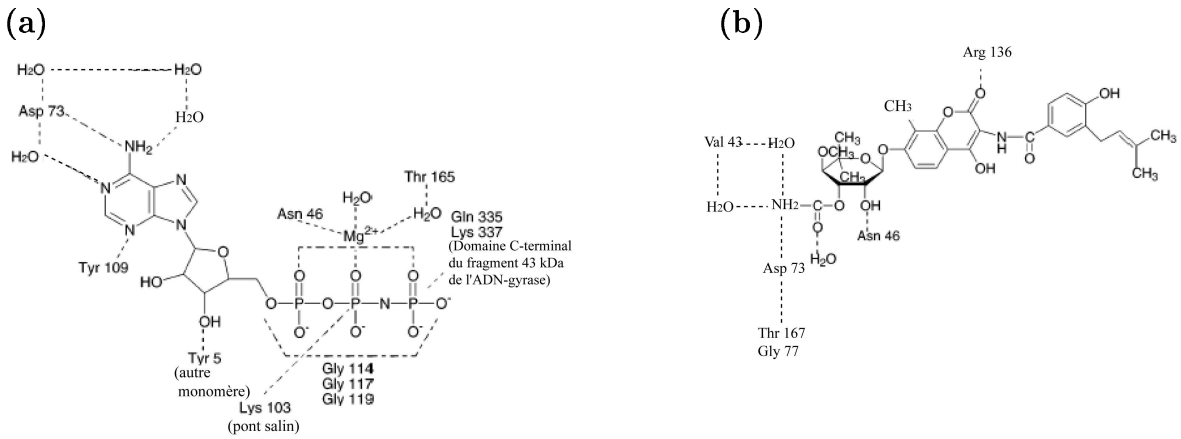

FIG. 4.16 – Liaisons faites par l'ADPNP (a) et la novobiocine (b) avec la sous-unité B de l'ADN-gyrase. Les lignes pointillées représentent des liaisons hydrogène.

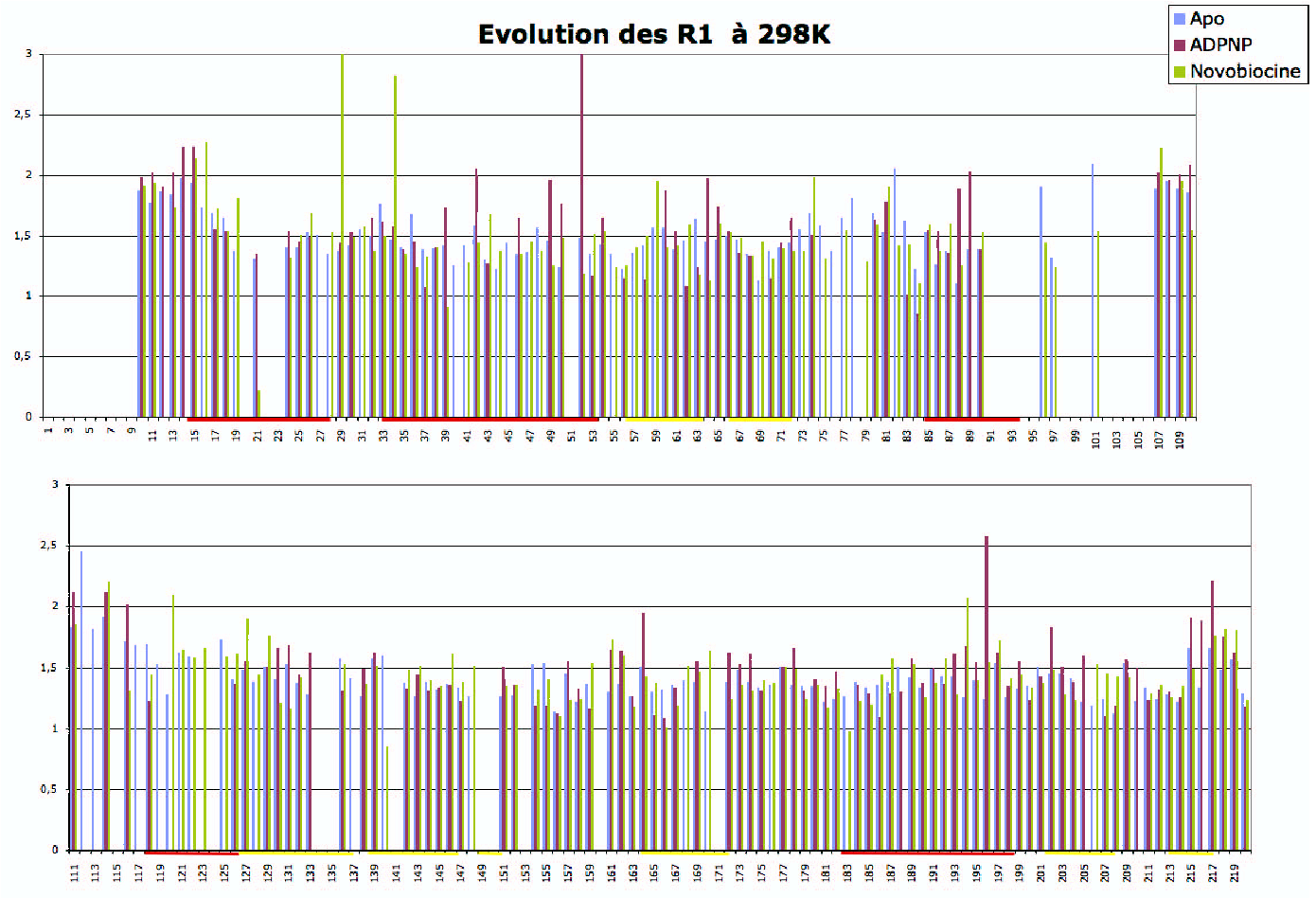

FIG. 4.17 – Comparaison des valeurs du temps de relaxation longitudinal pour le fragment 24 kDa de la sous-unité B de l'ADN-gyrase forme libre et formes complexées avec l'ADPNP et la novobiocine à 298 K

La présence d'inhibiteurs ne modifie pas de façon flagrante le profil des valeurs des temps de relaxation (Figure : 4.17). De manière générale, la forme libre propose des temps de relaxation plus importants que les formes complexées. Parmi les résidus qui réagissent de maniére particuliére, l'Asn 46 réalise des liaisons avec les deux ligands. La valeur de R_1 avec l'ADPNP est alors supérieure aux deux autres valeurs. Le cas du résidu 43 est inversé, la valeur la plus grande est celle liée à la novobiocine, il réalise une liaison avec ce ligand. Pour la relaxation transverse aucun des résidus attendus n'a un comportement particulier.

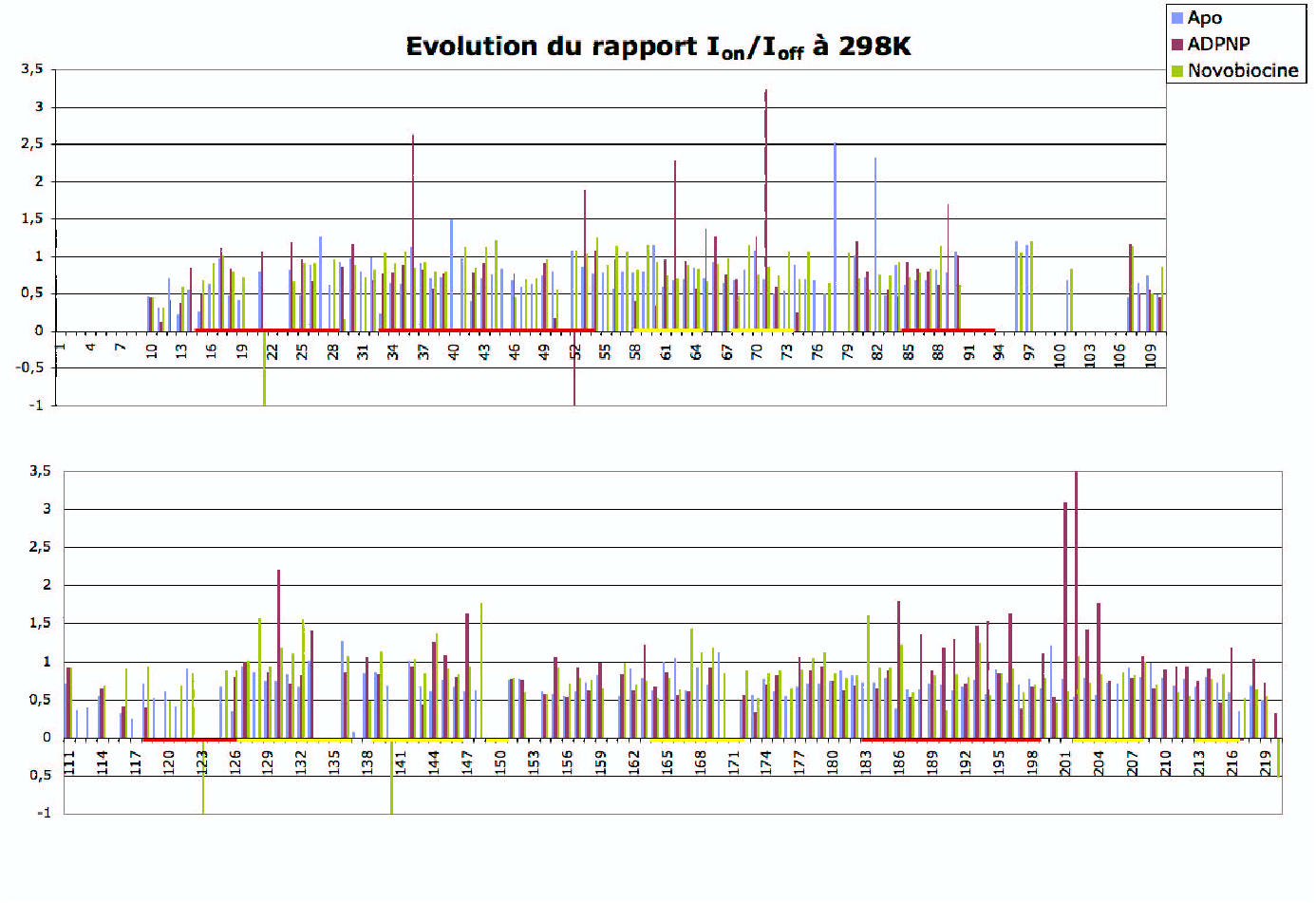

FIG. 4.18 – Comparaison des valeurs du rapport I_{on}/I_{off} pour le fragment 24 kDa de la sous-unité B de l'ADN-gyrase forme libre et formes complexées avec l'ADPNP et la novobiocine à 298 K

Les valeurs du rapport I_{on}/I_{off} entre les différentes formes du fragment 24 kDa de la sous-unité B de l'ADN-gyrase sont beaucoup plus fluctuantes que dans les deux cas précédents. Dans la zone des résidus 53 à 60, les valeurs obtenues avec l'ADPNP sont trés inférieures aux autres. A l'inverse, dans la zone allant de 186 à 219, cette valeur est bien supérieure aux autres. Ceci est également observable pour les résidus 36, 53, 62, 7 et 130.

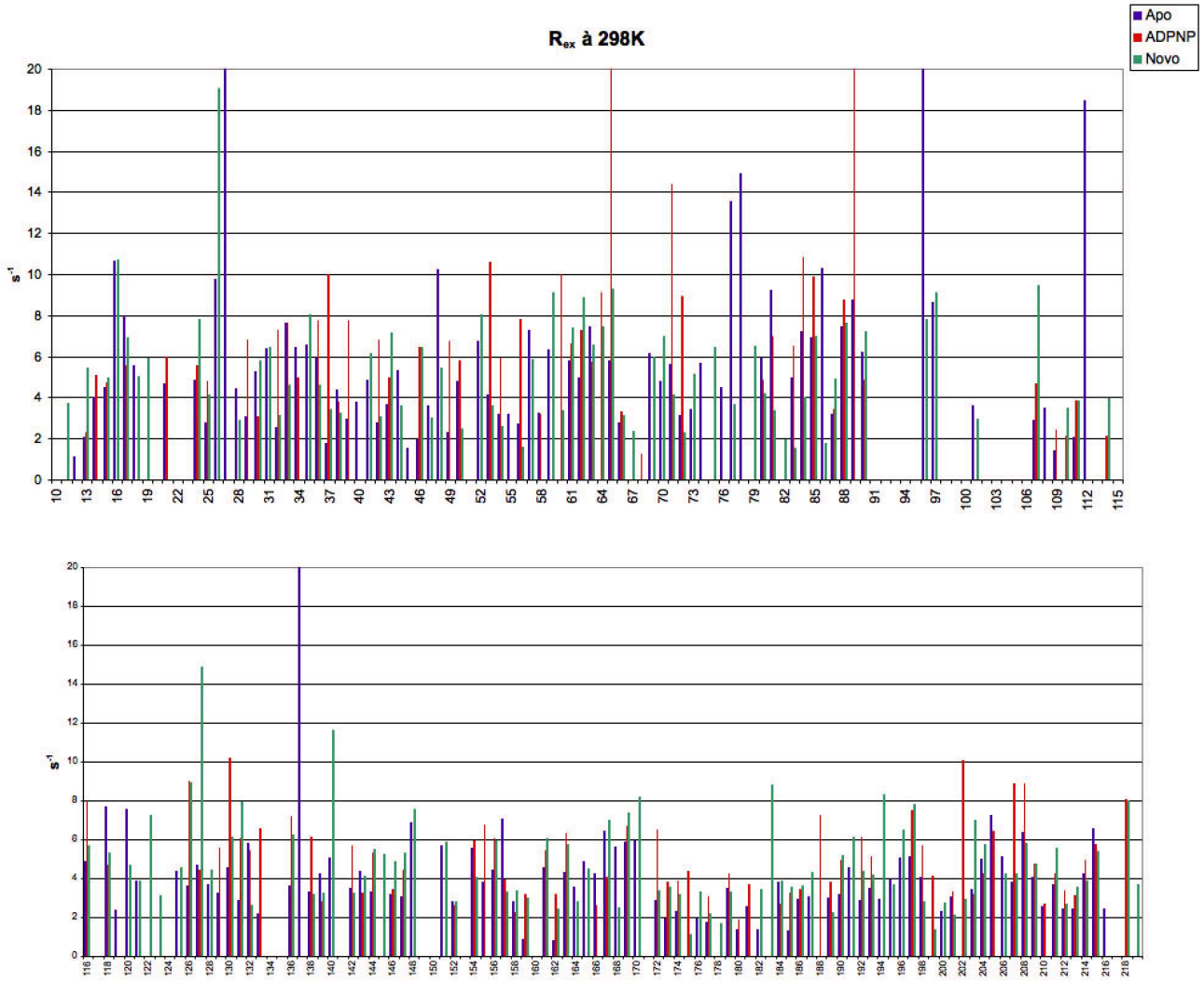

FIG. 4.19 – Comparaison des valeurs de la composante d'échange pour le fragment 24 kDa de la sous-unité B de l'ADN-gyrase forme libre et formes complexées avec l'ADPNP et la novobiocine à 298 K

Parmi les résidus faisant des liaisons avec les ligands l'Asn 46, qui fait des liaisons avec les deux ligands, voit sa composante d'échange tripler. Cette augmentation est observée, dans un moindre mesure, pour les résidus 43 et 136.

4.5.1 Les temps de corrélation

Les valeurs moyennes observées pour le rapport R_1/R_2 sont de 11,11 pour la forme Apo à 298 K, 10,00 pour la forme complexée avec l'ADPNP et 10,30 pour la forme complexée avec la novobiocine. Les temps de corrélation correspondants sont : 12,1 ns ; 11,44 ns et 11,64 ns. Ces valeurs restent très inférieures à celle obtenue par la formule de Daragan (16,46 ns). Les dimensions des tenseurs de diffusion sont référencées dans le tableau 4.2. A 298K, l'addition de magnésium et d'ADPNP dans l'échantillon amène l'axe principal

	Apo 298 K	Apo 303 K	ADPNP 298 K	ADPNP 303 K	Novo 298 K	Novo 303 K
D_X	1,185e-01	1,239e-01	1,280e-01	1,210e-01	1,147e-01	1,359e-01
D_Y	1,417e-01	1,629e-01	1,349e-01	1,748e-01	1,427e-01	1,585e-01
D_Z	1,491e-01	1,773e-01	1,727e-01	1,974e-01	1,709e-01	1,721e-01

TAB. 4.2 – Dimensions des vecteurs qui définissent les tenseurs de diffusion dans les différents échantillons aux différentes températures (10^{-8} s^{-1}).

ainsi que l'axe D_X du tenseur à s'allonger alors que l'axe D_Y se réduit légèrement. La protéine complexée à la novobiocine ne réagit pas de la même manière, l'axe principal s'allonge de manière comparable à celle observée avec l'ADPNP, l'axe D_Y s'allonge très légèrement et l'axe D_X se réduit. A 303 K, les formes complexées du fragment ont un tenseur de diffusion dont l'axe principal est plus petit que dans la forme libre. La forme complexée avec l'ADPNP présente l'axe D_Y le plus long, et pour l'axe D_X, c'est la forme avec la novobiocine qui a le plus long. Le changement le plus significatif est l'allongement de l'axe D_Z. L'ajout de ligands dans la protéine s'accompagne, aux deux températures, de l'allongement de l'axe principal. Celui-ci est plus important en présence d'ADPNP qu'en présence de novobiocine. Dans les structures cristallographiques, la boucle principale (95 à 101) se referme sur l'ADPNP, alors qu'elle reste ouverte en présence de novobiocine. Ceci explique le fait que l'axe principal du tenseur de diffusion soit plus grand en présence d'ADPNP. En comparant l'orientation des différents axes des tenseurs de diffusion, on remarque que les axes principaux dans les cas de l'échantillon avec ADPNP à 303 K et de l'échantillon avec novobiocine à 298 K, quittent le plan perpendiculaire aux principaux éléments de structure secondaire. Malgré ce point commun, ces deux cas ne sont pas similaires. Dans les cas de l'échantillon ADPNP, les axes D_Y et D_Z sont dégénérés. Pour

ce qui est de l'échantillon novobiocine, les 3 axes sont très bien définis, la dispersion est très faible.

A 298 K, la comparaison (Figure 4.20) des S^2 des différents complexes dégage quelques constatations globales : la forme Apo est la plus rigide, puis il y a la forme complexée avec la novobiocine et enfin celle avec l'ADPNP. Les formes Apo et novobiocine sont souvent très proches. Cette tendance générale est également observée sur deux des trois résidus faisant des liaisons avec les ligands : Asp 46 et Thr 165. Une évolution différente est subie par le résidu Asp 73 (rigidification pour les deux complexes, la forme ADPNP est plus rigide que les deux autres formes). Le résidu 136, impliqué dans une liaison à la novobiocine, se rigidifie en présence des deux ligands.

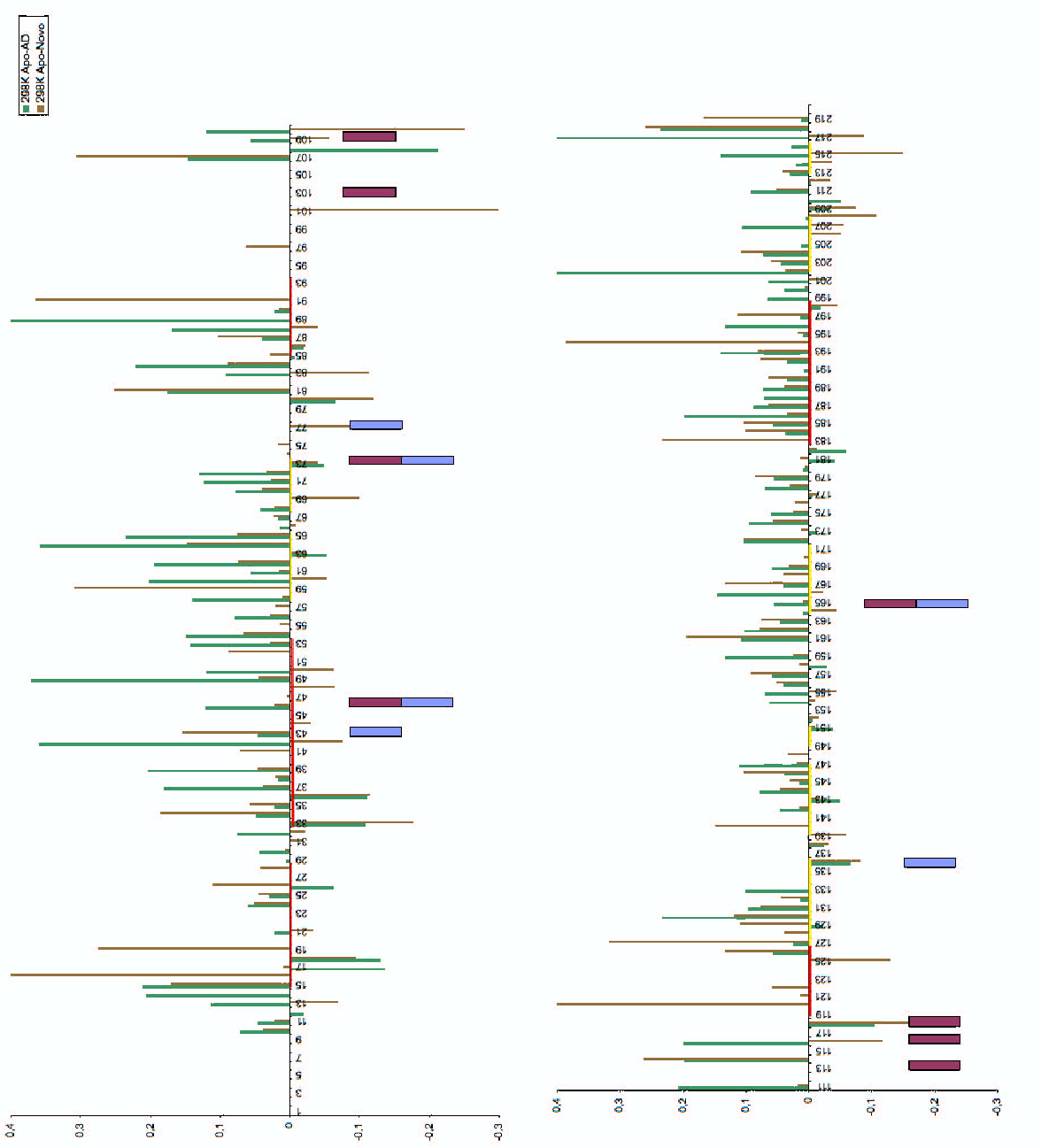

FIG. 4.20 – Comparaison des valeurs du paramètre d'ordre S^2 pour le fragment 24 kDa de la sous-unité B de l'ADN-gyrase forme libre et formes complexées avec l'ADPNP et la novobiocine à 298 K. Les rectangles violets indiquent les résidus impliqués dans la liaisons avec l'ADPNP et les rectangles bleus indiquent ceux liés à la novobiocine.

A 303 K, les évolutions sont beaucoup moins ordonnées. L'évolution générale observée à cette température est la même qu'à 298 K : la forme Apo est la plus rigide puis il y a la forme novobiocine et enfin la forme ADPNP. Pour certaines zones, telles que celle allant du résidu 30 à 40, la valeur du S^2 pour la forme ADPNP est proche de celle obtenue pour les deux autres jeux de données.

4.6 Conclusion

La fragment 24 kDa de la sous-unité B de l'ADN-gyrase est totalement anisotrope. Les mouvements qui l'animent sont très variés et se situent dans une large gamme de temps. La présence de composante d'échange augmentant avec la température suggère pour les boucles, des mouvements entre deux ou plusieurs conformations. Ce mouvement ordonné est en accord avec les dynamiques moléculaires réalisées au laboratoire (ceci fait l'objet d'un article et d'une thèse, tous deux en préparation).
L'effet de la fixation de l'ADPNP sur le fragment 24 kDa de la sous-unité B de l'ADN-gyrase est une déstabilisation générale. Celle-ci est observée essentiellement sur les temps de relaxation. En présence de novobiocine, certains effets spécifiques sur les résidus impliqués dans la fixation du ligand ont été observés.

Bibliographie

[Bartels *et al.*, 1995] Bartels, C., Xia, T.-H., Güntert, P., Wüthrich, K. (1995). *J Biomol NMR* **5**, 1–10. *The program XEASY for computer-supported NMR spectral analysis of biological macromolecules.*

[Bellanda *et al.*, 2002] Bellanda, M., Peggion, E., Otting, G., Weigelt, J., Perdona, E., Domenici, E., Marchioro, C., Mammi, S. (2002). *J Biomol NMR* **22** (4), 369–370. *Backbone 1H, 13C and 15N resonance assignment of the N-terminal 24 kDa fragment of the gyrase B subunit from E. coli.*

[Brino *et al.*, 1999] Brino, L., Bronner, C., Oudet, P., Mousli, M. (1999). *Biochimie* **81** (10), 973–980. *Isoleucine 10 is essential for DNA gyrase B function in Escherichia coli.*

[Castenholtz, 1988] Castenholtz, R. W. (1988). *Methods Enzymol.* **167**, 68–92. *Culturing methods for cyanobacteria.*

[Clore *et al.*, 1990] Clore, G. M., Szabo, A., Bax, A.and Kay, L. E., Driscoll, P. C., Gronenborn, A. M. (1990). *J Am Chem Soc* **112**, 4989–4991. *Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins.*

[Daragan Mayo, 1997] Daragan, V. Mayo, K. (1997). *Prog NMR Spec* **31**, 63–105. *Motional Model Analyses in Proteins and Peptides using 13C and 15N NMR Relaxation.*

[Delaglio *et al.*, 1995] Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., Bax, A. (1995). *J Biomol NMR* **6** (3), 277–293. *NMRPipe : a multidimensional spectral processing system based on UNIX pipes.*

[Desplancq *et al.*, 2004] Desplancq, D., Bernard, C., Sibler, A.-P., Miguet, L., Potier, N., Van Dorsselaer, A., Weiss, E. (2004). . soumis *Combining inductible protein overexpression with triple isotope labelling in the cyanobacterium Anabaena sp. PCC 7120.*

[Desplancq *et al.*, 2001] Desplancq, D., Kieffer, B., Schmidt, K., Posten, C., Forster, A., Oudet, P., Strub, J. M., Van Dorsselaer, A., Weiss, E. (2001). *Protein Expr Purif* **23** (1), 207–217. *Cost-effective and uniform (13)C- and (15)N-labeling of the 24-kDa*

N-terminal domain of the Escherichia coli gyrase B by overexpression in the photoautotrophic cyanobacterium Anabaena sp. PCC 7120.

[Dosset *et al.*, 2000] Dosset, P., Hus, J. C., Blackledge, M., Marion, D. (2000). *J Biomol NMR* **16** (1), 23–28. *Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data.*

[Dubelcco Vogt, 1963] Dubelcco, R. Vogt, M. (1963). *Proc Natl Acad Sci U S A* **50** (2), 236–243. *Evidence for a ring structure of polyoma virus DNA.*

[Eletsky *et al.*, 2001] Eletsky, A., Kienhofer, A., Pervushin, K. (2001). *J Biomol NMR* **20** (2), 177–180. *TROSY NMR with partially deuterated proteins.*

[Farrow *et al.*, 1994] Farrow, N. A., Zhang, O., Forman-Kay, J. D., Kay, L. E. (1994). *J Biomol NMR* **4** (5), 727–734. *A heteronuclear correlation experiment for simultaneous determination of 15N longitudinal decay and chemical exchange rates of systems in slow equilibrium.*

[Frias *et al.*, 2000] Frias, J. E., Flores, E., Herrero, A. (2000). *Mol Microbiol* **38** (3), 613–625. *Activation of the Anabaena nir operon promoter requires both NtcA (CAP family) and NtcB (LysR family) transcription factors.*

[Frias *et al.*, 2003] Frias, J. E., Herrero, A., Flores, E. (2003). *J Bacteriol* **185** (17), 5037–5044. *Open reading frame all0601 from Anabaena sp. strain PCC 7120 represents a novel gene, cnaT, required for expression of the nitrate assimilation nir operon.*

[Gellert *et al.*, 1976] Gellert, M., Mizuuchi, K., O'Dea, M. H., Nash, H. A. (1976). *Proc Natl Acad Sci U S A* **73** (11), 3872–3876. *DNA gyrase : an enzyme that introduces superhelical turns into DNA.*

[Kampranis *et al.*, 1999] Kampranis, S. C., Gormley, N. A., Tranter, R., Orphanides, G., Maxwell, A. (1999). *Biochemistry* **38** (7), 1967–1976. *Probing the binding of coumarins and cyclothialidines to DNA gyrase.*

[Kay Gardner, 1997] Kay, L. E. Gardner, K. H. (1997). *Curr Opin Struct Biol* **7** (5), 722–731. *Solution NMR spectroscopy beyond 25 kDa.*

[Kay *et al.*, 1989] Kay, L. E., Torchia, D. A., Bax, A. (1989). *Biochemistry* **28** (23), 8972–8979. *Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy : application to staphylococcal nuclease.*

[Lewis *et al.*, 1996a] Lewis, R. J., Singh, O. M., Smith, C. V., Skarzynski, T., Maxwell, A., Wonacott, A. J., Wigley, D. B. (1996a). *EMBO J* **15** (6), 1412–1420. *The nature of inhibition of DNA gyrase by the coumarins and the cyclothialidines revealed by X-ray crystallography.*

[Lewis *et al.*, 1996b] Lewis, R. J., Tsai, F. T., Wigley, D. B. (1996b). *Bioessays* **18** (8), 661–671. *Molecular mechanisms of drug inhibition of DNA gyrase.*

[Lipari Szabo, 1982a] Lipari, G. Szabo, A. J. (1982a). *J. Am. Chem. Soc.* **104**, 4546–4558. *Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data.*

[Lipari Szabo, 1982b] Lipari, G. Szabo, A. J. (1982b). *J. Am. Chem. Soc.* **104**, 4559–4570. *Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data.*

[Morais-Cabral *et al.*, 1997] Morais-Cabral, J. H., Jackson, A. P., Smith, C. V., Shikotra, N., Maxwell, A., Liddington, R. C. (1997). *Nature* **388** (6645), 903–906. *Crystal structure of the breakage-reunion domain of DNA gyrase.*

[Peng Wagner, 1992] Peng, J. W. Wagner, G. (1992). *Biochemistry* **31** (36), 8571–8586. *Mapping of the spectral densities of N-H bond motions in eglin c using heteronuclear relaxation experiments.*

[Pfuhl *et al.*, 1999] Pfuhl, M., Chen, H. A., Kristensen, S. M., Driscoll, P. C. (1999). *J Biomol NMR* **14** (4), 307–320. *NMR exchange broadening arising from specific low affinity protein self-association : analysis of nitrogen-15 nuclear relaxation for rat CD2 domain 1.*

[Reece Maxwell, 1991] Reece, R. J. Maxwell, A. (1991). *Nucleic Acids Res* **19** (7), 1399–1405. *The C-terminal domain of the Escherichia coli DNA gyrase A subunit is a DNA-binding protein.*

[Schechner *et al.*, 2004] Schechner, M., Sirockin, F., Stote, R. H., Dejaegere, A. P. (2004). *J Med Chem* **47** (18), 4373–4390. *Functionality maps of the ATP binding site of DNA gyrase B : generation of a consensus model of ligand binding.*

[Sharma Mondragon, 1995] Sharma, A. Mondragon, A. (1995). *Curr Opin Struct Biol* **5** (1), 39–47. *DNA topoisomerases.*

[Tsai *et al.*, 1997] Tsai, F. T., Singh, O. M., Skarzynski, T., Wonacott, A. J., Weston, S., Tucker, A., Pauptit, R. A., Breeze, A. L., Poyser, J. P., O'Brien, R., Ladbury, J. E., Wigley, D. B. (1997). *Proteins* **28** (1), 41–52. *The high-resolution crystal structure of a 24-kDa gyrase B fragment from E. coli complexed with one of the most potent coumarin inhibitors, clorobiocin.*

[Wang, 1971] Wang, J. C. (1971). *J Mol Biol* **55** (3), 523–533. *Interaction between DNA and an Escherichia coli protein omega.*

[Wang Liu, 1979] Wang, J. C. Liu, L. F. (1979). *Molecular Genetics, Part III* , 65–88. *DNA topoisomerases : enzymes that catalyze the concerted breaking and rejoining of DNA backbone bonds.*

[Watson Crick, 1953] Watson, J. D. Crick, F. H. (1953). *Nature* **171** (4356), 737–738. *Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid.*

[Wigley *et al.*, 1991] Wigley, D. B., Davies, G. J., Dodson, E. J., Maxwell, A., Dodson, G. (1991). *Nature* **351** (6328), 624–629. *Crystal structure of an N-terminal fragment of the DNA gyrase B protein.*

Conclusion générale

Lors de mon travail de thèse, un programme d'aide à l'attribution des spectres RMN a vu le jour. Le programme QUASI s'oriente vers une aide à l'attribution qui s'est avérée plus satisfaisante que l'automatisation complète de l'attribution. Il permet de regrouper les informations disponibles à partir des spectres RMN. Avec l'interface graphique développée sous la forme d'un "tableau de bord", toutes les informations sont visualisables en même temps. Le fait de recouper toutes les informations sur les résidus facilite la validation ou le rejet des propositions faites par QUASI. QUASI gère un certain nombre d'informations, prises en compte sous la forme d'une fonction cible ou sous une forme graphique. Le programme n'est pas un produit fini, toutefois il donne des résultats satisfaisants, car il nous a permis l'attribution rapide de plusieurs protéines, dont le fragment 24 kDa de la sous-unité B de l'ADN-gyrase. Il est souhaitable d'y ajouter des informations supplémentaires. L'évolution de la RMN est telle qu'il faut que QUASI puisse évoluer en parallèle et intégrer les développements à venir. Il est déjà prévu d'intégrer certains types d'informations dans ce logiciel, l'utilisation de banques de données, de similarité, d'homologie... Ces informations incorporées sous formes d'informations complémentaires doivent être intégrées à la fonction cible. La difficulté de cette incorporation est le poids à donner aux différentes informations.

En parallèle, le marquage puis l'étude de la dynamique du fragment 24 kDa de la sous-unité B de l'ADN-gyrase ont été réalisés. La mise au point récente d'un nouveau système d'expression nous a permis de marquer cette protéine à moindre coût et de tester également une méthode de 'reverse labelling' qui pourrait faciliter l'attribution des protéines et donc l'utilisation de QUASI. Afin de caractériser la manière dont *Anabaena* incorpore les acides aminés entiers lors de la production de la protéine cible, le marquage en ^{13}C est prévu. Si, dans le cas du marquage ^{15}N, les transaminases sont vraisemblablement à l'origine du fait que les résidus du même type ne sont pas touchés de manière similaire, alors le marquage en ^{13}C peut permettre un résultat satisfaisant.

L'étude dynamique du fragment 24 kDa de la sous-unité B de l'ADN-gyrase réalisée à différentes températures, avec et sans ligands, révèle la présence de nombreux mouvements dans une large gamme d'échelle de temps. Les boucles principales subissent des mouvements lents qui suggèrent des mouvements ordonnées entre plusieurs conformations. Il reste encore à mettre ces résultats de dynamique en rapport avec les calculs de modélisation qui ont déjà été effectués au laboratoire.

Annexes

1

Accessibilité des programmes d'attribution automatique

Méthode	Langages informatiques	Accès
ASSIGN	FORTRAN	http ://www.cmu.edu/nmr-center/links.html
ASSTOOL	NR	Contacter les auteurs
AUTOASSIGN	LISP,Tcl/Tk	http ://www-nmr.cabm.rutgers.edu/NMRsoftware/nmr_software.html
GARANT		http ://www.mol.biol.ethz.ch/wuthrich/software/garant/
IBIS	Langage C	http ://gwagner.med.Harvard.edu/ibis/
MAPPER	FORTRAN	http ://guentert.gsc.riken.go.jp/Software/MapperLicense.html
MARS	C, scripts shell	http ://www.mpibpc.gwdg.de/abteilungen/030/zweckstetter/_links/software_mars.htm
MONTE	Tcl/Tk	http ://www.andrew.cmu.edu/ rule/monte/
PACES	Visual Basic, C, Java	Contacter les auteurs
PASTA	ANSI C, Tcl/Tk	http ://www.org.chemie.tu-muenchen.de/people/jl/PASTAV3.0tk.html
PLATON	C, Tcl/Tk	http ://www.fmp-berlin.de/ labudde/platon.html
PROCESS	Tcl/Tk, PERL, shell scripts	http ://www.pence.ualberta.ca/software/camra/latest/camra.html
QUASI	Fortran, C++, Python	Contacter les auteurs
RESCUE	Langage C, shell scripts	http ://www.infobiosud.cnrs.fr/SERVEUR/RESCUE/welcome.html
Smartnotebook	Tcl/Tk	Contacter les auteurs
st2nmr	NR	http ://arg.cmm.ki.si/ primus/Miscellaneous/st2nmr_tutorial.html
TATAPRO	C	Contacter les auteurs

TAB. 1.1 – Tableau récapitulatif des caractéristiques informatiques des programmes. "NR" indique que le langage utilisé n'est pas renseigné. La seconde colonne indique où se procurer le programme.

2

Les programmes de prédiction de déplacements chimiques

2.1 SHIFTY

SHIFTY, développé par le groupe de David Wishart, est une méthode de prédiction de déplacements chimiques sur la base d'homologie de séquence, c'est un outil qui, par sa conception, est proche des logiciels qui ont fait leur preuve entre les mains des cristallographes. L'idée principale est que deux protéines homologues ont non seulement des structures homologues mais aussi des déplacements chimiques très similaires. En d'autres termes, si une protéine homologue est déjà attribuée, il doit être possible d'utiliser ces mêmes attributions (avec une correction adéquate) pour prédire les déplacements chimiques d'une protéine homologue non attribuée. Ce concept relativement simple est appliqué et décrit par de nombreux groupes. Mais il n'était appliqué qu'à des situations ou le ou les homologues sont identifiés grâce à une recherche faite dans la litérature. Ainsi les créateurs de SHIFTY ont décidé de faire un programme qui sélectionne, aligne et attribue les déplacements chimiques ^{1}H et ^{13}C des protéines non attribuées de façon automatique.

Les bases de données

La BMRB a été scannée, seuls les peptides et les protéines contenant des attributions ^{1}H très complètes collectées en milieu aqueux, dans des pH entre 2 et 7,5 et à des températures allant de 5°C à 60°C ont été sélectionné. Un total de 147 chaînes polypeptidiques sont identifiées. Cette base de données de déplacements chimiques ^{1}H est complétée par 28 protéines issues d'une sélection faite en 1991. Le total est donc de 175 polypeptides.

Les déplacements chimiques sont référencés aussi bien par rapport au TSP que par rapport au DSS.
D'un autre côté, les ambiguïtés dans les références des déplacements chimiques ^{13}C ont rendu beaucoup des attributions ^{13}C de la BMRB inutilisables. La base de données ^{13}C a donc été assemblée à partir des données utilisées pour le développement de la méthode CSI. Cette base de données actualisée contient les déplacements chimiques $^{13}C^{\alpha}$, $^{13}C^{\beta}$ et ^{13}CO de 18 protéines référencés par rapport au DSS.
Les deux bases de données contiennent 193 chaînes polypeptidiques représentant 11 062 résidus et près de 56 000 attributions de déplacements chimiques. Les deux bases de données sont utilisées par le programme pour identifier et aligner la cible sur le/les homologues attribués. Le programme utilise également des données expérimentales obtenues en mesurant les déplacements chimiques d'hexapeptides désordonnés dans des solutions à 1M urée. Cette base de données "Random Coil" est utilisée pour prédire les déplacements chimiques des résidus non identiques.

Les algorithmes

SHIFTY est composé de deux algorithmes : le premier qui gère la comparaison de la séquence non attribuée avec les 193 protéines de la base de données et l'alignement ; le second qui s'occupe de l'attribution séquentielle ou de la prédiction de déplacements chimiques. L'alignement séquentiel et la comparaison sont faits en utilisant la méthode dynamique développée par Needleman et Wunsch. La séquence de la protéine cible est comparée de façon systématique à chacune des 193 protéines. La comparaison se fait en utilisant un matrice de score similaire à la matrice de mutation standard de Dayhoff PAM $_{250}$ (PAM : Point Accepted Mutation). L'alignement et le score de similarité de séquence sont déterminés à partir des scores de similarité entre acides aminés et des informations sur la structure secondaire. Les séquences présentant les scores les plus élevés sont sélectionnées et chaque paire d'alignement est passée au second algorithme.

Une fois la protéine la plus homologue extraite de la base de données, à condition que les deux résidus existent, le second algorithme attribue à chaque acide aminé de la cible un déplacement chimique. Si les deux acides aminés sont identiques, les déplacements chimiques de cet acide aminé sont les déplacements chimiques de la protéine homologue. Par contre si les acides aminés ne sont pas strictement identiques, le déplacement chimique de la protéine homologue est soustrait à sa valeur dans la pelote statistique (pour obtenir le déplacement secondaire) et cette différence est additionnée à la valeur du résidu de la cible dans la pelote statistique. Enfin, si un résidu de la protéine cible se place dans un

trou de la séquence homologue, ou vice-versa, aucune prédiction n'est faite.

Séquence cible	D	T	G	L
δ RC	8,4	8,25	8,41	8,28
Séquence homologue	E		G	L
δ RC	8,87		7,11	7,69
Prédiction	9,03		7,11	7,69

8,87 - 8,4 + 8,56 = 9,03

FIG. 2.1 – Exemples des trois cas rencontrés lors de la prédiction de déplacements chimiques par le second algorithme de SHIFTY. La valeur 8,56 est la valeur RC pour l'acide glutamique.

Le programme est développé en langage C, l'utilisateur peut modifier le fichier de paramètres. Il peut ainsi choisir sa propre base de données, modifier les pénalités, la matrice de score... SHIFTY n'a besoin que de la séquence cible pour fonctionner. Le résultat est obtenu sous forme de fichier texte.

Tests

Ce logiciel a été testé sur 25 protéines ayant au moins une homologue chacune dans la base de données. L'évaluation du programme se fait sur trois critères :

- la cohérence générale entre les déplacements prédits et ceux observés.

- la cohérence entre les déplacements obtenus avec SHIFTY et ceux prédits à partir des structures cristallographiques.

- l'évolution de la cohérence de la prédiction en fonction du pourcentage d'identité de séquence.

Comparaison prédictions/observations

Dans le cas représenté (Figure : 2.2), les déplacements chimiques HN, $^{1}H^{\alpha}$ et $^{13}C^{\alpha}$

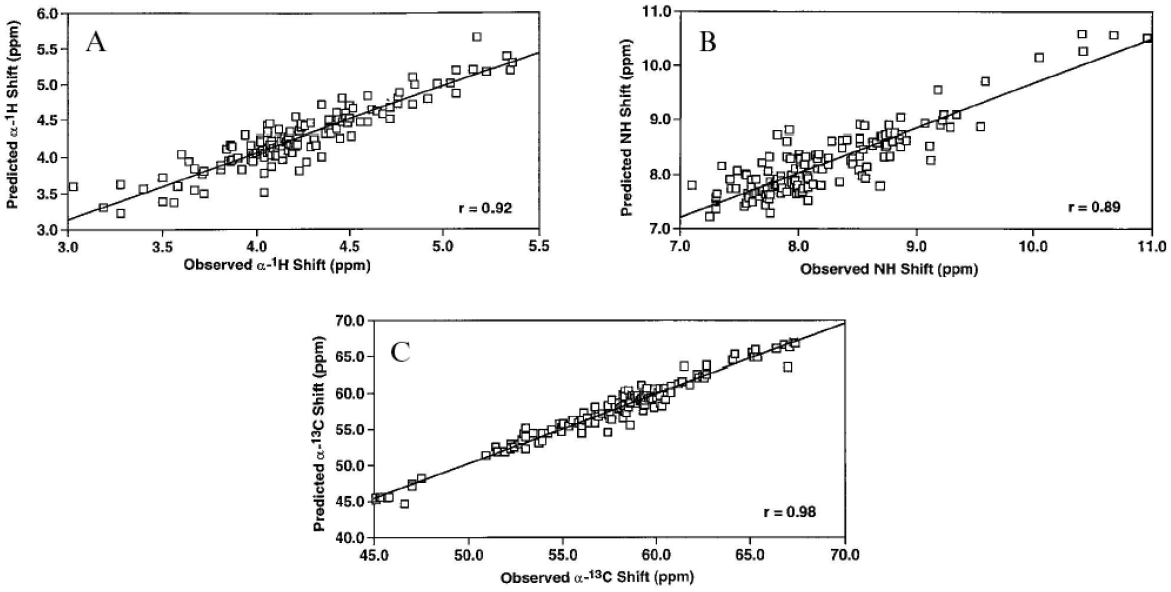

FIG. 2.2 – Comparaison des déplacements chimiques prédits, issus de la calmoduline (Drosophile) et des déplacements chimiques observés de troponine C de dinde. (A)$^{1}H^{\alpha}$, (B)HN, (C)$^{13}C^{\alpha}$. Le coefficent de corrélation(r) est indiqué dans le coin inférieur droit de chaque graphique.

prédits pour la troponine C de dinde sont tracés par rapport aux déplacements chimiques observés expérimentalement. La calmoduline de la Drosophile, qui avait l'identité de séquence la plus élevée avec celle de la troponine C (46,2%) est utilisée comme protéine prédictrice. Comme nous pouvons le voir sur les graphiques, l'accord entre les déplacements chimiques prédits et observés est meilleur pour $^{1}H^{\alpha}$ et $^{13}C^{\alpha}$ (respectivement 0,92 et 0,98) que pour le proton amide (0,89).

Comparaison avec d'autres méthodes

Afin de comparer les résultats à ceux qui auraient pu être observés si on avait utilisé une structure cristallographique de troponine C pour prédire les déplacements chimiques, les développeurs ont utilisé le programme de Ösapay et Case (SHIFTS) et celui de Williamson. Les coordonnées du fichier PDB 5TNC est utilisé à cet effet. Les corrélations entre les déplacements chimiques observés et ceux prédits par la méthode SHIFTS sont de 0,8 pour $^{1}H^{\alpha}$ et 0,17 pour HN. Pour la méthode Williamson, les corrélations sont de 0,81 et 0,36. Malgré le fait que SHIFTY se base sur une protéine différente de la protéine cible, SHIFTY donne de meilleurs résultats que les 2 méthodes basées sur les structures

cristallographiques.
Cette même tendance est observée pour les protéines de Lysozyme de poulet, la Calbindine de bovin, la Bungarotoxine... En moyenne, les coefficients de corrélation des résultats de SHIFTY sont 20 à 40% meilleurs pour $^1H^\alpha$ alors que pour HN, les résultats s'améliorent jusqu'à 300%. Dans ces calculs de corrélations, il est important de noter que seuls les résidus qui sont prédits sont pris en compte.

Pourcentage d'identité de séquence

De par la nature du programme SHIFTY (basé sur les homologies de séquence et de déplacements chimiques), le critère d'évaluation le plus important est l'évolution de la qualité des résultats en fonction du degré d'identité de séquence.

De façon évidente, la présence d'une protéine homologue à 99-100% permet à SHIFTY de prédire les déplacements chimiques de la protéine cible avec une fiabilité de 99% ou 100%. On remarque (Figure : 2.3) que les résultats sont bons tant qu'il y a au moins 35% d'homologie avec la séquence de départ. En plus, en dessous de ce pourcentage d'identité, les méthodes basées sur les structures cristallographiques donnent de meilleurs résultats que SHIFTY.

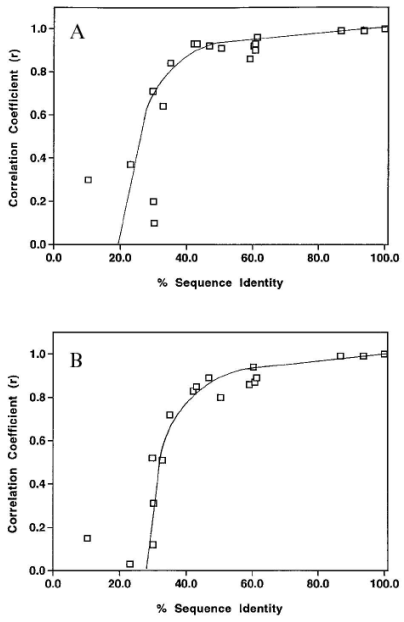

FIG. 2.3 – Graphiques illustrant la variation des coefficients de corrélation entre les déplacements chimiques (A)$^1H^\alpha$, (B)HN prédits et les déplacements chimiques observés en fonction du pourcentage d'identité de séquence.

2.2 SHIFTS

Dès l'année 1991,Ösapay K. et Case D.A. modélisent, avec des équations empiriques, les contributions de groupes peptidiques aux déplacements chimiques des protons H^{α} et H^{β}. Les équations empiriques représentent l'anisotropie magnétique et les interactions électrostatiques. SHIFTS propose un algorithme basé sur une base de données de déplacements chimiques de peptides calculés au niveau DFT qui mène à la prédiction de déplacements chimiques des noyaux ^{15}N et ^{13}C à partir de la structure de la protéine. La base de données est constituée de 1335 peptides dont les angles du squelette sont dans les zones "normales" du diagramme de Ramachadran. Des calculs de DFT on tire 8 contributions au déplacement chimique qui sont supposées additives :

- l'influence du squelette sur l'acide aminé précédent
- l'influence du squelette sur l'acide aminé présent
- l'influence du squelette sur l'acide aminé suivant
- l'influence du type de chaîne latérale sur l'acide aminé précédent
- l'influence du type de chaîne latérale sur l'acide aminé présent
- l'influence des liaisons hydrogène avec le groupe NH sur le déplacement chimique ^{15}N
- l'influence des liaisons hydrogène avec le groupe CO sur le déplacement chimique ^{15}N
- l'influence des liaisons hydrogène avec le groupe NH et avec le groupe CO sur le déplacement chimique ^{15}N

Cette méthode rencontre 2 problèmes majeurs :

- les déplacements chimiques ne sont pas tous calculés avec la même précision par la DFT.
- la base des fonctions utilisée joue sur la qualité du calcul.

De plus SHIFTS est limité car il ne considère que les zones normalement acceptées pour les structures secondaires, ce qui peut représenter une perte d'informations.
La DFT : Théorie de la Fonctionnelle de la Densité est une méthode ab-initio qui permet d'approcher la résolution de l'équation de Schrödinger. Cette théorie est utilisée pour créer une base de données de déplacements chimiques de peptides. La correction liée à la structure secondaire est déterminée grâce à cette base de données.

3

Les tables de référence utilisées par QUASI

3.1 Table Random Coil

acide aminé	atome	δ	acide aminé	atome	δ
ALA	CA	52,82	LEU	CA	55,47
ALA	CB	19,26	LEU	CB	42,46
ARG	CA	56,48	LYS	CA	56,71
ARG	CB	30,93	LYS	CB	33,21
ASP	CA	52,99	MET	CA	55,77
ASP	CB	38,33	MET	CB	32,94
ASN	CA	53,33	PHE	CA	58,09
ASN	CB	39,09	PHE	CB	39,75
CYS	CA	58,63	PRO	CA	63,7
CYS	CB	28,34	PRO	CB	32,22
GLN	CA	56,22	SER	CA	58,67
GLN	CB	29,53	SER	CB	64,06
GLU	CA	56,09	THR	CA	62,01
GLU	CB	28,88	THR	CB	70,01
GLY	CA	45,39	TRP	CA	57,6
HIS	CA	55,39	TRP	CB	29,75
HIS	CB	29,12	TYR	CA	58,28
ILE	CA	61,62	TYR	CB	38,94
ILE	CB	38,91	VAL	CA	62,61
			VAL	CB	32,82

FIG. 3.1 – Déplacements chimiques des acides aminés dans les pentapeptides GGXGG enregistrés dans une solution de 8M urée.

3.2 Table statistique issue de la BMRB

acide aminé	atome	δ	σ	acide aminé	atome	δ	σ
ALA	CA	53,21	2,48	LEU	CA	55,63	2,21
ALA	CB	19,04	2,80	LEU	CB	42,45	2,66
ARG	CA	56,92	2,45	LYS	CA	56,87	2,34
ARG	CB	30,87	3,16	LYS	CB	32,97	2,92
ASP	CA	54,62	2,19	MET	CA	56,22	2,65
ASP	CB	40,94	3,49	MET	CB	33,06	3,58
ASN	CA	53,44	2,19	PHE	CA	58,21	2,75
ASN	CB	38,85	3,18	PHE	CB	40,16	2,95
CYS	CA	57,57	3,49	PRO	CA	63,28	2,08
CYS	CB	33,89	6,82	PRO	CB	32,03	3,21
GLU	CA	57,52	2,36	SER	CA	58,61	2,22
GLU	CB	30,19	2,85	SER	CB	63,85	3,01
GLN	CA	56,58	2,56	THR	CA	62,21	2,83
GLN	CB	29,27	2,79	THR	CB	69,55	2,87
GLY	CA	45,42	2,01	TRP	CA	57,62	2,62
HIS	CA	56,35	2,89	TRP	CB	30,54	4,33
HIS	CB	30,03	2,77	TYR	CA	58,00	2,73
ILE	CA	61,63	2,77	TYR	CB	39,41	3,59
ILE	CB	38,84	3,50	VAL	CA	62,55	3,01
				VAL	CB	32,78	2,69

FIG. 3.2 – Déplacement chimique moyen de chaque atome de chaque acide aminé, calculé à partir des dépôts de structure RMN, ce qui représente 1098240 déplacements chimiques.

3.3 Comparaison entre les tables utilisées comme référence par QUASI

QUASI utilise deux tables de valeurs de déplacements chimiques comme référence pour le placement des fragments. Ces tables sont d'un côté les tables random coil et d'un autre les moyennes et écart-types des valeurs déposées à la BMRB. QUASI n'utilise dans sa fonction de score que deux types de noyaux ($^{13}C^{\alpha}$ et $^{13}C^{\beta}$). En traçant les moyennes,

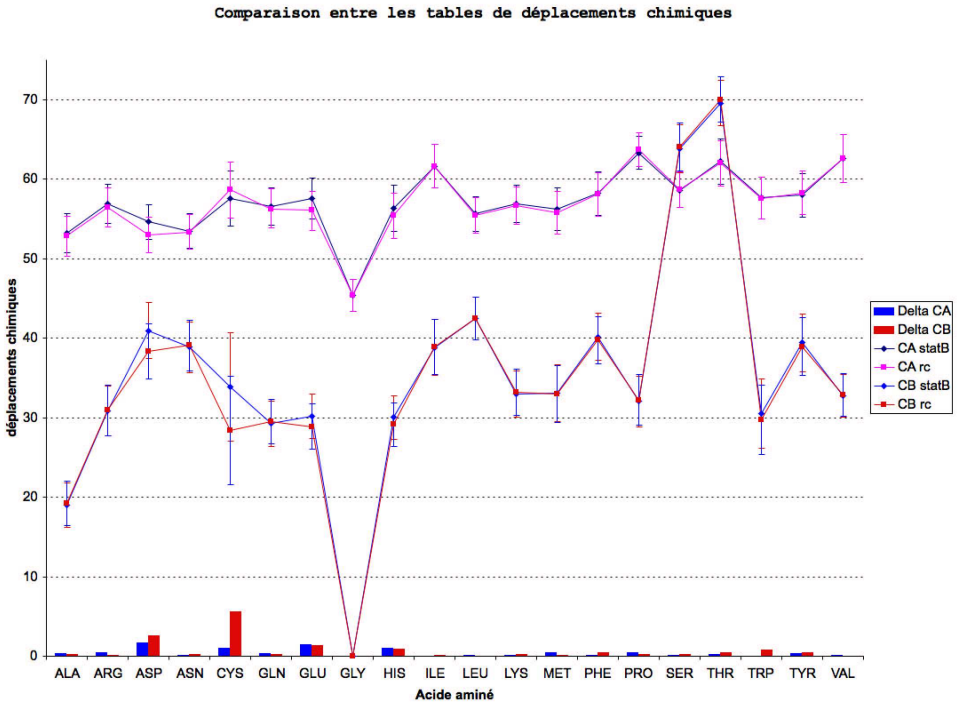

FIG. 3.3 – Comparaison des valeurs moyennes des noyaux $^{13}C^{\alpha}$ et $^{13}C^{\beta}$ des déplacements chimiques issues des tables de Random Coil (courbes bleues) et des statistiques de la BMRB (courbes rouges). Les barres d'erreur sont les écart-types correspondants. Les histogrammes représentent la valeur absolue des différences ($^{13}C^{\alpha}$ en rouge, $^{13}C^{\beta}$ en bleu).

encadrées de leurs écart-types respectifs, en fonction du type d'acide aminé (Figure 3.3), la comparaison des références est réalisable. Pour chaque type d'acide aminé, les références se recoupent. Les plus grandes différences se trouvent au niveau de la cystéine et de l'asparagine, toutefois elles restent inférieures aux écart-types correspondants.
Ces différences, si petites soient elles, entraînent des différences dans les attributions faites par QUASI (cas de l'α-actinine EF34). L'utilisateur a le choix entre ces deux tables.

4

Le test du χ^2

4.1 Table du chi-deux

	seuil				seuil				seuil		
dl	0,05	0,01	0,001	dl	0,05	0,01	0,001	dl	0,05	0,01	0,001
1	3,84	6,64	10,83	34	48,60	56,06	65,25	67	87,11	96,83	108,54
2	5,99	9,21	13,82	35	49,80	57,34	66,62	68	88,25	98,03	109,79
3	7,82	11,35	16,27	36	51,00	58,62	67,99	69	89,39	99,23	111,06
4	9,49	13,28	18,47	37	52,19	59,89	69,35	70	90,53	100,42	112,31
5	11,07	15,09	20,52	38	53,38	61,16	70,71	71	91,67	101,62	113,56
6	12,59	16,81	22,46	39	54,57	62,43	72,06	72	92,81	102,82	114,84
7	14,07	18,48	24,32	40	55,76	63,69	73,41	73	93,95	104,01	116,08
8	15,51	20,09	26,13	41	56,94	64,95	74,75	74	95,08	105,20	117,35
9	16,92	21,67	27,88	42	58,12	66,21	76,09	75	96,22	106,39	118,60
10	18,31	23,21	29,59	43	59,30	67,46	77,42	76	97,35	107,58	119,85
11	19,68	24,73	31,26	44	60,48	68,71	78,75	77	98,49	108,77	121,11
12	21,03	26,22	32,91	45	61,66	69,96	80,08	78	99,62	109,96	122,36
13	22,36	27,69	34,53	46	62,83	71,20	81,40	79	100,75	111,155	123,60
14	23,69	29,14	36,12	47	64,00	72,44	82,72	80	101,88	112,338	124,84
15	25,00	30,58	37,70	48	65,17	73,68	84,03	81	103,01	113,511	126,09
16	26,30	32,00	39,25	49	66,34	74,92	85,35	82	104,14	114,704	127,33
17	27,59	33,41	40,79	50	67,51	76,15	86,66	83	105,27	115,887	128,57
18	28,87	34,81	42,31	51	68,67	77,39	87,97	84	106,40	117,060	129,80
19	30,14	36,19	43,82	52	69,83	78,62	89,27	85	107,52	118,242	131,04
20	31,41	37,57	45,32	53	70,99	79,84	90,57	86	108,65	119,415	132,28
21	32,67	38,93	46,80	54	72,15	81,07	91,88	87	109,77	120,597	133,51
22	33,92	40,29	48,27	55	73,31	82,29	93,17	88	110,90	121,770	134,74
23	35,17	41,64	49,73	56	74,47	83,52	94,47	89	112,02	122,942	135,96
24	36,42	42,98	51,18	57	75,62	84,73	95,75	90	113,15	124,125	137,19
25	37,65	44,31	52,62	58	76,78	85,95	97,03	91	114,27	125,297	138,45
26	38,89	45,64	54,05	59	77,93	87,17	98,34	92	115,39	126,469	139,66
27	40,11	46,96	55,48	60	79,08	88,38	99,62	93	116,51	127,631	140,90
28	41,34	48,28	56,89	61	80,23	89,59	100,88	94	117,63	128,803	142,12
29	42,56	49,59	58,30	62	81,38	90,80	102,15	95	118,75	129,975	143,32
30	43,77	50,89	59,70	63	82,53	92,01	103,46	96	119,87	131,147	144,55
31	44,99	52,19	61,10	64	83,68	93,22	104,72	97	120,99	132,319	145,78
32	46,19	53,49	62,49	65	84,82	94,42	105,97	98	122,11	133,471	146,99
33	47,40	54,78	63,87	66	85,97	95,63	107,26	99	123,23	134,643	148,21
								100	124,34	135,814	149,48

FIG. 4.1 – Valeurs critiques du χ^2 pour les 100 premiers degrés de liberté (dl) à 3 valeurs de seuil : 0,05 ; 0,01 et 0,001.

5

Publication

Résumé

L'identification des signaux RMN en terme d'acide aminé est connu sous le terme d'attribution. Cette étape est incontournable non seulement lorsque l'on veut déterminer la structure d'une protéine par RMN mais également lorsque l'on a une structure cristallographique et que l'on veut étudier plus avant une protéine. Le programme QUASI pour QUick Access to Spectral Interpretation a été développé afin d'assister l'attribution. Ce nouvel outil permet, grâce à son interface graphique, la présentation de résultats obtenus à partir de références issues d'horizons divers. Les résultats obtenus sur des protéines de tailles différentes telles que l'α-actinine EF-34 (75 acide-aminés) et le fragment 24 kDa de la sous-unité B de l'ADN-gyrase (220 acide-aminés) sont précis et fiables.
La seconde partie de ce manuscrit est dédiée à l'étude de la dynamique du fragment 24 kDa de la sous-unité B de l'ADN-gyrase. Cette étude est menée à 3 températures 298 K, 303 K et 310K en présnce de 2 ligands ADPNP ou novobiocine. Le fragment s'avère subir des mouvements à différentes échelles de temps. Les boucles, qui se ferment et s'ouvrent sur la poche active, présentent des mouvements particulièrement difficiles à définir.

Mots-clés: RMN, attribution, programme, automatisation, ADN-gyrase, dynamique, QUASI, protéines, spectres, hétéronucléaires, ADPNP, novobiocine, densité spectrale

Abstract

Establishing the correspondence between observed NMR signals and the nuclei in the covalent structure of a protein is termed assignment. This procedure is obligatory for the interpretation of NMR spectra in atom-specific terms, not only when one wishes to determine the structure of a protein but also when a structure is available and one wishes to pursue further studies of the biological system. QUASI (QUick Access to Spectral Interpretation) is software that has been developed to assist assignment. This new tool is able, via its graphical interface, to present results obtained using different user-determined references. Results obtained on proteins of varying sizes, such as α-actinin EF34 (75 amino acids) and the 24 kDa fragment of DNA-gyrase B subunit (220 amino acids), are accurate and reliable.
Following assignment, the dynamics of the 24 kDa fragment of DNA-gyrase B subunit were investigated in the presence and absence of ligand inhibitors and as a function of temperature. The residues of the fragment undergo motions on a range of different time-scales from pico- to nano-second, up to milli- to micro-second. The loops that open and close over the active site pocket present particularly complex motions that proved difficult to analyse.

Keywords: NMR, assignment, software, automatisation, DNA-gyrase, dynamics, QUASI, proteins,spectra, heteronuclear, ADPNP, novobiocin, spectral density