Déterminer la structure d'une protéine par RMN

Un problème d'optimisation complexe

Yves Nominé et Bruno Kieffer

Résumé Le calcul de structures de protéines en solution par résonance magnétique nucléaire a bénéficié de nombreux développements au cours des dernières années. Ces innovations ont permis d'améliorer l'efficacité de l'ensemble du processus de détermination de structure, de la production d'échantillons de protéines enrichies avec des isotopes stables à la modélisation sous contraintes. Le répertoire des observables RMN utilisées dans le calcul de structures s'est considérablement enrichi et de nouvelles expériences offrent des perspectives inédites et originales sur la structure et la dynamique des assemblages macromoléculaires. L'intégration de données hétérogènes pour construire un modèle rendant compte des multiples réalités des protéines en solution reste une entreprise passionnante.

Mots-clés **Protéines, structure, RMN du liquide, biologie structurale, méthodes.**

Abstract **The determination of protein structures by NMR: a complicated optimization problem**
Nuclear magnetic resonance spectroscopy, already a powerful tool for the determination of protein structures, has recently seen significant improvement as a result of a number of key developments. These advancements have touched all steps of the structure determination process ranging from the production of isotopically labeled protein samples to the development of innovative modelling strategies. The panel of NMR experiments used to obtain exploitable information on proteins has enlarged considerably and now offers an unprecedented view on the structure and dynamics of macromolecular assemblies. Integrating heterogeneous data to build a model that accounts for the multiple aspects of proteins in solution remains a challenging and fascinating task.

Keywords **Protein, structure, liquid state NMR, structural biology, methods.**

Il y a près de trente ans, l'équipe dirigée par Kurt Wüthrich à l'ETH de Zürich démontrait la possibilité de déterminer la structure d'une petite protéine en solution à partir des informations issues des spectres de RMN [1]. En offrant une alternative au cadre rigide et peu physiologique imposé par l'empilement cristallin, cette première a fait naître, dans la jeune communauté de la biologie structurale, l'espoir d'approcher d'encore plus près la nature intime du repliement des protéines. Pourtant, la cristallisation des protéines en vue de résoudre leur structure par diffraction des rayons X reste aujourd'hui le moyen privilégié d'accéder à l'information structurale et la contribution nette en termes de nombre de structures résolues par RMN reste modeste (elle représente près de 12 % des structures de protéines déposées dans la banque PDB[(1)]).

En revanche, la possibilité d'observer individuellement les atomes des protéines en solution par RMN a conduit à de profonds bouleversements de notre compréhension de la relation qui existe entre la structure et la fonction d'une protéine [2]. De nombreuses protéines étudiées depuis les débuts de cette méthode se sont montrées plus flexibles et dynamiques que ce qui était initialement attendu. Dans ce cas, les mesures RMN ne résultent plus d'une organisation géométrique unique des atomes dans l'espace, mais d'un ensemble très large de conformations différentes. Cette situation a conduit les chercheurs à s'interroger sur la façon d'extraire des informations pertinentes des spectres RMN de tels systèmes, ainsi que sur les possibilités de traduire cette complexité dans un modèle moléculaire. Ces travaux ont conduit à développer de nouvelles méthodes permettant d'obtenir des informations sur la structure et la dynamique des protéines, élargissant ainsi considérablement la palette qui s'offre au biologiste.

Initiés au début des années 2000, les programmes de génomique structurale avaient pour ambition la description complète de l'ensemble des repliements de protéines. Si cet objectif n'est pas encore complètement atteint, ces initiatives ont conduit à un développement sans précédent des outils et méthodes permettant d'accélérer les études structurales [3]. Dans les programmes les plus avancés de biologie structurale, l'analyse RMN des échantillons de protéines a été placée au cœur du dispositif associant l'ensemble des techniques de biophysiques structurales. Pourtant, après trente années de développements, la détermination d'une structure de protéine par RMN reste une aventure multidisciplinaire dont la richesse ne s'exprime pas forcément dans le modèle final enregistré au sein de la banque PDB. Ce modèle résulte d'un processus complexe qui intègre les techniques d'analyse de séquences, de biochimie, de spectroscopie RMN et de modélisation les plus avancées (*figure 1*). L'optimisation de l'une ou l'autre de ces étapes est très souvent nécessaire pour mener à bien l'étude d'une protéine particulière. L'objet de cet article est d'illustrer le caractère multidisciplinaire d'une étude structurale de protéine par RMN, où chaque étape peut faire l'objet d'un passionnant travail de recherche et d'optimisation.

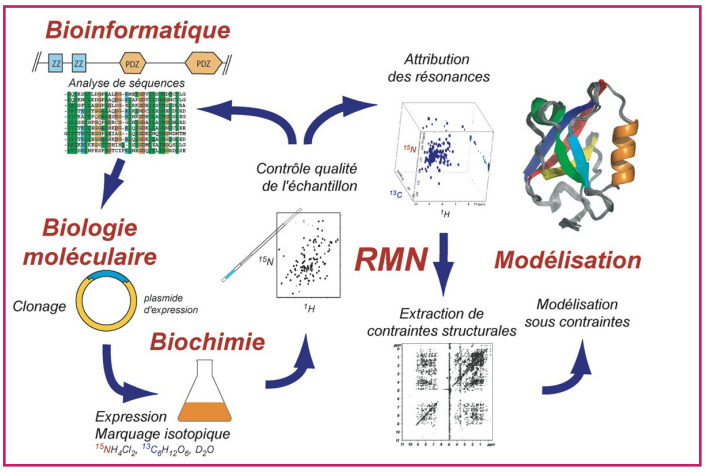


Figure 1 - De la séquence d'une protéine à sa structure.
Résoudre une structure de molécule par RMN ne se limite pas aux seules expériences de RMN. En amont, un travail important d'analyse de séquences, de biologie moléculaire et de biochimie est indispensable pour obtenir des échantillons de qualité. La production de protéines par expression dans un système bactérien permet l'incorporation d'isotopes stables dans la molécule. L'enregistrement de spectres RMN intervient dès les premières étapes de la production afin de guider et d'affiner les choix technologiques en amont (système d'expression, définition précise du système étudié, etc.). Le processus de détermination de structures par RMN se poursuit par l'enregistrement et le traitement de spectres permettant d'attribuer une valeur de déplacement chimique à chacun des atomes de la molécule, puis par l'enregistrement de spectres permettant l'extraction d'informations structurales (distances, angles dièdres, RDC, PRE…). La modélisation de structures sous contraintes et la validation des modèles constituent les étapes finales du processus.

« Exprime-toi et je te dirai qui tu es » : l'étape critique de la préparation de l'échantillon

L'obtention d'un échantillon de protéine pour une étude RMN constitue une étape indissociable du processus qui va conduire à l'obtention d'un modèle moléculaire. Le succès final va être largement conditionné par les choix qui sont réalisés en amont de l'étude proprement dite. Ces choix concernent essentiellement la définition précise du ou des polypeptides à étudier, la stratégie d'attribution des signaux RMN, ainsi que la méthode de production de l'échantillon.

L'analyse couplée bioinformatique/RMN : un préalable indispensable à toute étude structurale et fonctionnelle

Les protéines de grande taille sont souvent composées de plusieurs domaines structuraux, capables de se replier de façon autonome [4]. Ces domaines peuvent parfois être associés à une fonction spécifique et, dans le cas idéal, conservent cette fonction lorsqu'ils sont étudiés en dehors du contexte de la protéine entière. Cette fonction peut être liée, par exemple, à une activité catalytique, ou bien à l'interaction avec une autre protéine ou d'autres composants chimiques de la cellule (ADN, glucides, lipides...). Cette partition en domaines fonctionnels est fondamentale, car elle permet de réduire l'étude globale d'une protéine à celle des domaines individuels qui la composent. Ces domaines sont généralement de taille modeste (quelques dizaines à quelques centaines d'acides aminés), ce qui facilite leur étude structurale par RMN. Le repliement tridimensionnel d'un domaine ainsi que sa fonction imposent des contraintes sur sa composition en acides aminés. Un petit nombre d'acides aminés sont conservés à certaines positions de la séquence primaire, indépendamment de l'organisme vivant considéré. La définition d'un domaine particulier repose donc sur l'identification de ces acides aminés conservés au cours de l'évolution de la protéine. Cette analyse consiste à aligner les séquences des différentes protéines similaires (assurant une fonction identique au sein d'organismes différents), de façon à identifier les acides aminés identiques ou présentant des propriétés physico-chimiques voisines. Dans la plupart des cas, elle permet d'identifier un ou plusieurs domaines et suggère un « découpage » possible de la protéine (*figure 1*).

Cependant, l'expérience montre que de petites variations de séquence, en particulier au sein des régions limitrophes du domaine, peuvent avoir des conséquences importantes sur le comportement du domaine en solution et la qualité des spectres obtenus. C'est pourquoi il est souvent nécessaire d'affiner la définition d'un domaine particulier, en combinant l'analyse de séquences avec la caractérisation par RMN de différents polypeptides à l'aide de spectres de corrélation ^{1}H-^{15}N (*figure 2*).

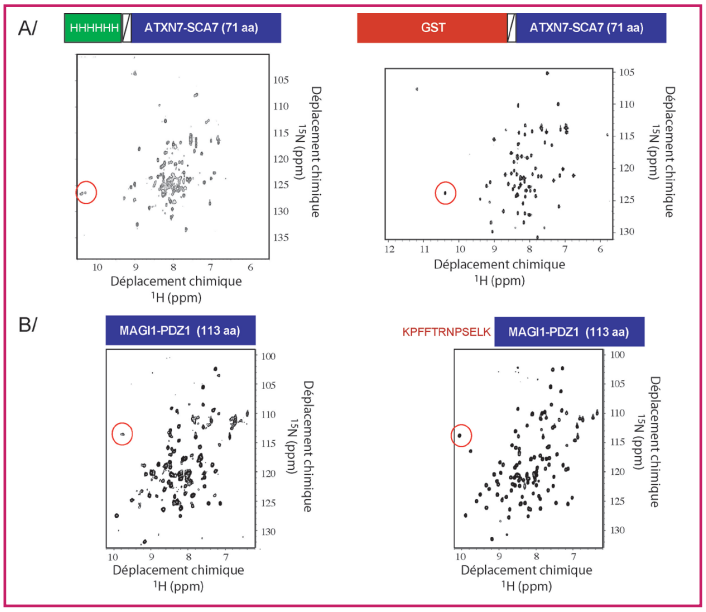


Figure 2 - Les spectres de corrélation ^{1}H-^{15}N, juges arbitres de la qualité d'un échantillon.
Le spectre de corrélation ^{1}H-^{15}N permet d'évaluer rapidement la qualité d'un échantillon de protéine. La très grande sensibilité de cette spectroscopie autorise l'enregistrement de spectres sur des petites quantités d'échantillons, voire sur des extraits cellulaires bruts. Chaque tâche de corrélation correspondant à un acide aminé, le décompte des corrélations donne immédiatement une indication sur l'état de la protéine. Un nombre de corrélations inférieur à celui attendu traduit souvent la présence d'hétérogénéité dynamique de la protéine et interdit la détermination de structure. Il est alors nécessaire de modifier les stratégies utilisées en amont, comme celui de l'étiquette, un polypeptide additionnel fusionné à la protéine et permettant une purification plus aisée. A/ Dans le cas du domaine SCA7 de l'ataxine 7, le doigt de zinc présent dans ce domaine n'était pas compatible avec l'utilisation d'une étiquette poly-histidine. Le passage sur une colonne d'affinité contenant du cobalt conduisait à des échanges entre les ions zinc de la protéine et les ions cobalt de la colonne, entraînant des hétérogénéités dans les spectres. Le problème a été résolu par l'utilisation d'une étiquette composée de la protéine glutathione-S-transférase (GST), comme le montre la plus grande homogénéité de la forme des pics (par exemple ceux entourés en rouge). B/ Dans le cas du domaine PDZ-1 de la protéine MAGI-1[2], il a été nécessaire d'ajouter douze acides aminés en amont de la protéine pour obtenir des spectres homogènes.

Expression bactérienne et marquage isotopique : un couple indissociable

La détermination de la structure d'une protéine (ou de l'un de ses domaines) par RMN nécessite des quantités relativement importantes de matériel (de l'ordre de quelques milligrammes de protéine pure). De plus, l'attribution des

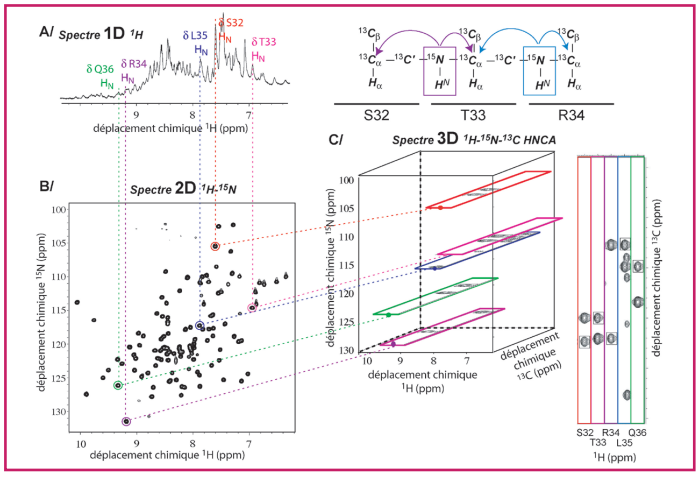


Figure 3 - Attribution des déplacements chimiques.
L'attribution consiste à établir la correspondance entre chaque noyau de la protéine pour lequel on peut observer un signal et la fréquence de ce signal (le déplacement chimique). A/ Dans le cas des protéines, la complexité des spectres 1D ne permet pas une attribution directe. B/ Une première simplification des spectres est obtenue en enregistrant des spectres bidimensionnels de corrélation ^{1}H-^{15}N. Ce type de spectre, qui nécessite un marquage à l'azote-15, permet d'observer individuellement les protons amides du squelette peptidique. C/ Une simplification supplémentaire des spectres peut être obtenue en enregistrant des spectres triple résonance à trois dimensions nécessitant un enrichissement supplémentaire en carbone-13. L'expérience HNCA par exemple permet de visualiser les corrélations entre le couple de fréquences (proton, azote) d'un acide aminé et les fréquences des carbones alpha du même acide aminé ainsi que celui du voisin sur la chaîne. L'enchaînement séquentiel de ces corrélations sur des bandes extraites du spectre 3D permet de propager l'attribution des atomes du squelette peptidique de proche en proche.

fréquences de résonance requiert, dans la plupart des cas, l'incorporation d'un ou plusieurs isotopes stables (^{15}N, ^{13}C, ^{2}H) dans la protéine. Le type de spectroscopie qui va être mise en œuvre pour l'attribution va être déterminé par le marquage isotopique de l'échantillon, de sorte que le choix du système d'expression fait partie intégrante du processus de détermination de structure (*figure 1*). Le processus complet de caractérisation structurale par RMN nécessite souvent la production de plusieurs échantillons différents, ce qui impose l'utilisation d'un système d'expression efficace et peu onéreux. Dans la grande majorité des cas, la protéine est obtenue en exprimant le gène correspondant dans la bactérie *Escherichia coli*, qui permet à la fois de multiples combinaisons de marquage isotopique à partir de précurseurs métaboliques simples (sels d'ammonium pour l'azote-15, glucose pour le carbone-13, eau lourde pour le deutérium), ainsi que l'obtention de rendements élevés. Dans cet organisme, la protéine est souvent exprimée sous la forme d'un polypeptide plus long, comportant, outre la protéine d'intérêt, un second polypeptide (comme une poly-histidine ou la protéine glutathione-S-transferase), appelé « étiquette », qui permet la purification du produit par chromatographie d'affinité. Le choix de cette étiquette dépend de la protéine étudiée et doit également faire l'objet d'optimisation. Une fois la protéine obtenue en quantité suffisante, l'utilisation de techniques de spectroscopie 3D triple résonance permet une simplification considérable des spectres RMN, qui autorise une automatisation partielle du processus d'attribution (*figure 3*).

La RMN n'est pas limitée à l'étude de protéines de petite taille

Il y a quelques années, les difficultés posées par la détermination de structures de protéines ou de complexes protéiques de grande taille semblaient insurmontables. D'une part, l'augmentation du nombre de noyaux entraîne inévitablement un encombrement spectral qui complique l'attribution des signaux. D'autre part, le ralentissement des mouvements de diffusion rotationnelle de la protéine conduit à un élargissement des raies de résonance sous l'effet d'une relaxation dipolaire plus efficace. L'efficacité de la relaxation peut être réduite en diminuant la densité de protons à l'aide d'échantillons de protéines uniformément enrichies au deutérium et marquées à l'azote-15 et au carbone-13. Combiné à l'utilisation d'expériences de corrélation de type TROSY (« transverse relaxation optimized spectroscopy », qui exploitent des propriétés de relaxation de certains noyaux pour réduire la largeur des signaux de résonance, ce type d'approche permet d'obtenir des spectres à haute résolution sur des objets de grande taille (plus de 200 acides aminés) [5]. L'encombrement spectral reste néanmoins souvent un problème pour l'attribution qui peut être résolue en mettant en œuvre des techniques particulières, comme le marquage spécifique de certains types d'acides aminés ou le marquage par fragments [6]. Lorsque la taille de la protéine ou du complexe protéique devient très importante (jusqu'à 1 MDa), l'incorporation sélective et stéréospécifique d'un carbone-13 sur les groupes méthyles des leucines et des valines d'une protéine uniformément perdeutérée devient la seule façon d'obtenir des informations sur sa structure et sa dynamique. Cette incorporation repose sur l'utilisation de précurseurs des voies de biosynthèse de ces acides aminés comme l'acide α-cétoisovalérique marqué sur des positions spécifiques [7]. Un exemple d'utilisation fructueuse de ce type d'approche est donné par l'étude en solution du protéasome, un complexe multi-protéique de 670 kDa [8].

Le bref aperçu des méthodes de marquage isotopique des protéines montre à quel point la production de l'échantillon est indissociable de l'étude structurale d'une protéine par RMN. Le développement des méthodes de marquage isotopique associé à celui des outils spectroscopiques a permis de repousser les limites imposées par la taille des protéines, et il est fort probable que cette tendance perdure dans les années futures avec l'automatisation et la production en parallèle d'échantillons et avec le développement de nouvelles techniques d'incorporation d'acides aminés artificiels, pour ne citer que deux exemples.

Une palette d'expériences très riche

Le chercheur qui aborde l'étude d'une protéine par RMN dispose d'une palette très large d'expériences pour obtenir des informations de structure. L'approche classique repose sur la mesure de la relaxation croisée entre les protons de la molécule (phénomène à l'origine de l'effet Overhauser nucléaire ou NOE) afin d'extraire des distances, ainsi que sur l'analyse des constantes de couplage pour obtenir des informations sur les angles dièdres. Cette combinaison fournit la très grande majorité des contraintes structurales qui sont utilisées pour le calcul des structures déposées dans la banque PDB. Plusieurs développements récents sont venus enrichir la gamme des contraintes à disposition pour la construction de modèles de structures (*figure 4* p. 52). L'analyse des interactions dipolaires résiduelles en milieu anisotrope, l'utilisation de sondes paramagnétiques ou encore le calcul de déplacements chimiques peuvent maintenant être utilisés soit pour fournir des contraintes géométriques supplémentaires, soit pour valider les structures finales.

L'épineux problème de la validation des structures

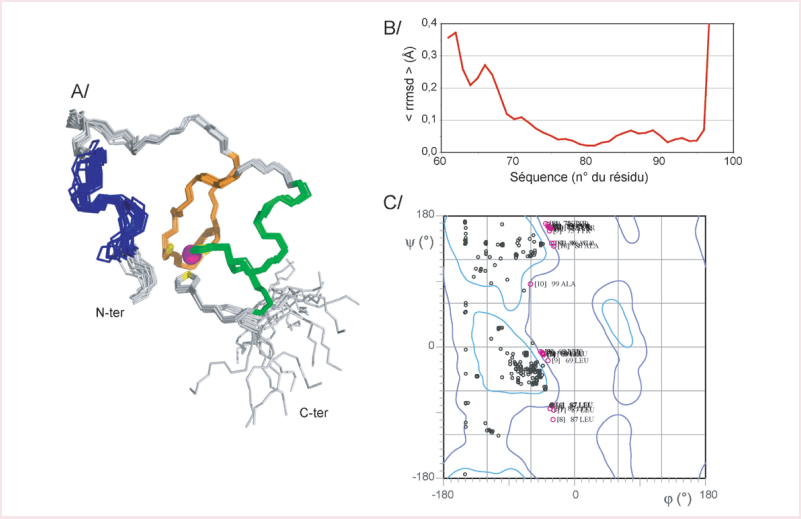
Près d'un quart de siècle après la publication de la première structure de protéine en solution, l'évaluation de la qualité d'un modèle obtenu à partir des données RMN reste l'objet de recherches actives et d'intenses débats. Dans certains domaines, comme la recherche de sites de liaison pour des ligands potentiels à la surface des protéines, les structures RMN sont systématiquement écartées, du fait de l'absence d'indicateurs fiables de la qualité du modèle. Cette situation est intimement liée à la façon dont une structure de protéine est obtenue, ainsi qu'à l'étroite imbrication existant entre les informations sur la structure et la dynamique du système étudié. Les avancées récentes dans les méthodes de modélisation ainsi que l'élargissement de la palette des contraintes expérimentales laissent cependant entrevoir la possibilité d'obtenir un modèle qui rend compte à la fois de la structure et de ses mouvements.

Justesse et précision

L'évaluation de la qualité d'un modèle structural repose sur la réponse à deux questions :
1) Le jeu de contraintes expérimentales permet-il de définir un modèle unique ?
2) Le modèle obtenu rend-il compte de l'ensemble des observations ?
La réponse à ces questions est loin d'être triviale, du fait de l'absence d'une transformation directe permettant de passer des observables RMN à des coordonnées moléculaires. La réponse à la première question est en général obtenue en multipliant le nombre de structures calculées à partir de structures initiales différentes. Les différents jeux de coordonnées sont ensuite comparés en calculant la largeur d'une distribution de distances qui séparent les atomes équivalents, pris dans les différents modèles alignés. Cette valeur (appelée RMSD pour « root mean square deviation ») fournit une mesure de la précision du modèle, qui affiche très souvent des valeurs inférieures à 0,5 Å pour les atomes de la chaîne principale. Le caractère partiel des informations géométriques fournies par les différentes observables RMN ainsi que la multiplicité de ces dernières rendent la réponse à la seconde question très délicate. Le calcul d'un facteur d'accord, tel qu'il est couramment pratiqué pour les structures cristallographiques, n'a pour le moment pas été adopté pour les structures en solution en dépit de plusieurs propositions dans ce sens. En l'absence de critères satisfaisants pour juger de la qualité d'une structure RMN, son appréciation repose souvent sur une analyse qualitative du jeu de données. Ainsi, le nombre moyen de contraintes de distances par acide aminé a été adopté comme critère pour évaluer la qualité d'un modèle. On considère qu'un modèle est fiable si ce nombre est supérieur à 10. Dans certains cas, la mesure d'un jeu de couplages dipolaires résiduels (RDC) est utilisée comme un moyen indépendant de vérifier la justesse des structures obtenues à partir des contraintes de distances et d'angles dièdres. L'utilisation de ce type de données pour obtenir un ensemble représentatif de conformères a également fait l'objet de plusieurs travaux récents. En effet, les valeurs moyennes de RDC proviennent d'ensembles conformationnels plus larges, car le temps caractéristique de cette mesure est plus long que celui du NOE. Une interprétation rigoureuse des données RDC a ainsi permis de proposer récemment un ensemble de structures de l'ubiquitine qui rend compte à la fois de la valeur moyenne des coordonnées et de la dynamique de la chaîne peptidique, ouvrant ainsi la voie à une description complète du système [a].

La tyrannie du Ramachandran

L'un des outils les plus répandus pour juger de la conformité d'un modèle de protéine est le diagramme de Ramachandran, qui présente les angles dièdres φ et ψ du squelette peptidique sous la forme de points dans un graphique bidimensionnel. La présence de points dans des régions dites « défavorables » de ce diagramme permet de repérer rapidement les acides aminés qui ont une conformation locale potentiellement altérée (voir *figure*). Si les premières structures de protéines calculées à partir de contraintes de distances RMN s'écartaient notoirement des critères de qualité communément admis en biologie structurale, la situation a, depuis, beaucoup évolué grâce aux développements des protocoles d'affinement. Cette dernière étape du calcul de structure permet d'incorporer des contraintes visant à favoriser les interactions à l'intérieur de la protéine ou bien entre la protéine et les molécules d'eau du solvant. La prise en compte de ce genre de contraintes qui ne proviennent pas directement de l'expérience, mais de ce qui est généralement connu et admis pour les protéines, a permis d'améliorer considérablement la « qualité » des modèles de structures RMN, et des initiatives proposant de recalculer des structures publiées dans la banque PDB ont été proposées [b].



Évaluation de la qualité d'un jeu de structures RMN.
La résolution de la structure en solution d'un domaine dit C2H2 de la protéine Sgf73 [c] a permis d'obtenir un jeu de structures représenté par la superposition des dix modèles de plus basse énergie (c'est-à-dire présentant la valeur la plus faible possible de la fonction cible). A/ Le repliement de ce domaine est stabilisé par la coordination d'un atome de zinc (en magenta) par les chaînes latérales de deux cystéines et de deux histidines (d'où le nom C2H2). Lors des étapes finales du calcul de structure, des contraintes supplémentaires correspondant à une géométrie de coordination tétraédrique sont ajoutées au jeu de contraintes expérimentales. Ces contraintes ne correspondent pas à une observation expérimentale directe, mais contribuent à la précision des coordonnées. B/ La courbe de la valeur de l'écart quadratique moyen (rmsd) en fonction de la séquence montre que la position de l'hélice située dans la partie N-terminale du polypeptide (en bleu sur la figure) est moins bien définie que celle du motif de coordination du zinc (en vert). C/ Le diagramme de Ramachandran indique que la géométrie locale de quelques acides aminés reste à affiner. Ces acides aminés possèdent des couples de valeurs des angles dièdres (φ, ψ) qui s'écartent des régions dites favorables (contours violets et bleus).

[a] Vendruscolo M., Determination of conformationally heterogeneous states of proteins, *Curr. Opin. Struct. Biol.*, **2007**, *17*, p. 15.
[b] Nederveen A.J., Doreleijers J.F., Vranken W., Miller Z., Spronk C.A., Nabuurs S.B., Guntert P., Livny M., Markley J.L., Nilges M., Ulrich E.L., Kaptein R., Bonvin A.M., RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank, *Proteins*, **2005**, *59*, p. 662.
[c] Bonnet J., Wang Y.H., Spedale G., Atkinson R.A., Romier C., Hamiche A., Pijnappel W.W., Timmers H.T., Tora L., Devys D., Kieffer B., The structural plasticity of SCA7 domains defines their differential nucleosome-binding properties, *EMBO Rep.*, **2010**, *11*, p. 612.

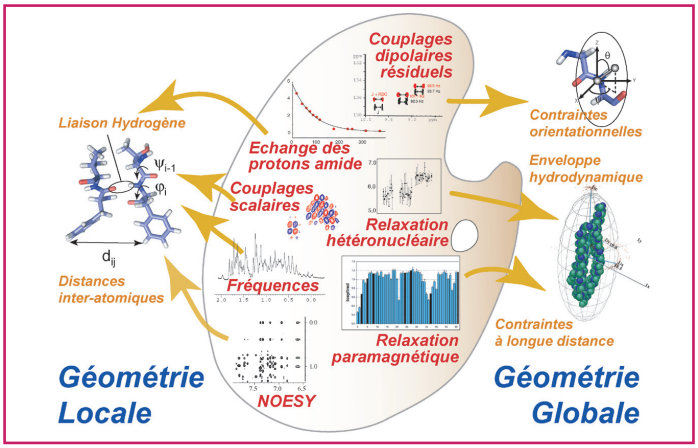


Figure 4 - Mesures RMN et information structurale.
Traditionnellement, la construction d'un modèle moléculaire de protéine en solution repose sur l'interprétation des NOE et des constantes de couplage à trois liaisons pour extraire des informations de distances et d'angles dièdres. Ces informations d'ordre local peuvent être complétées par l'observation de liaisons hydrogène révélées en mesurant la vitesse d'échange des protons amides avec ceux de l'eau. D'autres mesures nous renseignent sur la géométrie globale de la molécule. Ainsi, la forme de l'enveloppe moléculaire peut être obtenue par une analyse détaillée des vitesses de relaxation des noyaux d'azote-15 situés sur le squelette peptidique. L'étude des paramètres de relaxation renseigne également sur l'orientation relative de deux domaines d'une même protéine. Ce type d'information peut également être obtenu en observant des couplages dipolaires résiduels sur des échantillons placés dans des milieux anisotropes. Les ambiguïtés résultant de la faible portée des distances interatomiques extraites de l'analyse des NOE (6 Å) sont levées par des mesures de relaxation paramagnétique électronique, qui donnent une information sur des distances plus longues (20-30 Å).

Une règle d'arpentage incontournable : l'effet Overhauser nucléaire

L'effet Overhauser nucléaire est lié à la relaxation croisée entre les protons d'une même molécule. Il conduit au transfert d'une quantité d'aimantation d'un proton à son voisin au sein d'une même molécule et la vitesse de ce transfert varie en fonction de l'inverse de la puissance sixième de la distance ainsi que de la dynamique du système [9]. Ce phénomène est révélé à l'aide de spectres multidimensionnels comprenant des périodes d'échange d'aimantation de type NOESY. Dans la pratique, ces échanges interviennent entre deux protons si la distance qui les sépare est inférieure à 6 Å. L'observation d'un tel échange entre deux protons ou groupes de protons éloignés les uns des autres au sein de la chaîne peptidique fournit une contrainte topologique très importante. Par exemple, l'observation de quelques NOE entre des protons situés sur deux brins adjacents d'une structure en feuillet β permet d'identifier les acides aminés qui établissent des liaisons hydrogène entre les deux brins. La conversion de l'intensité de l'effet Overhauser en une distance est généralement réalisée en effectuant une calibration à partir des intensités de NOE mesurées pour des couples de protons séparés par une distance connue. La fiabilité d'un modèle tridimensionnel de protéine obtenu à partir de l'analyse des NOE repose sur la redondance importante de ce type d'information. Ces dernières années, plusieurs algorithmes ont été développés afin d'exploiter cette redondance et d'augmenter ainsi la fiabilité des modèles calculés à partir de l'analyse des NOE.

Constantes de couplage et mesure d'angles dièdres

La constante de couplage entre deux noyaux séparés par trois liaisons covalentes dépend de la géométrie du système. Une équation phénoménologique a été proposée par M. Karplus pour décrire ce phénomène. Ainsi, la valeur de la constante de couplage entre les deux uniques protons de la chaîne principale d'un peptide, $^3J_{HN\text{-}H\alpha}$, dépend de la valeur de l'angle dièdre φ, un des trois angles déterminant la géométrie du squelette peptidique. On constate cependant que si l'équation de Karplus donne une valeur unique de constante de couplage pour un angle dièdre donné, la réciproque n'est pas nécessairement vraie, et plusieurs valeurs de l'angle dièdre peuvent correspondre à une même valeur $^3J_{HN\text{-}H\alpha}$ mesurée (*figure 5*). L'obtention de paramètres géométriques à partir de la mesure de constantes de couplage résume très simplement le problème d'indétermination qui se pose lors du calcul de structures de protéines par RMN.

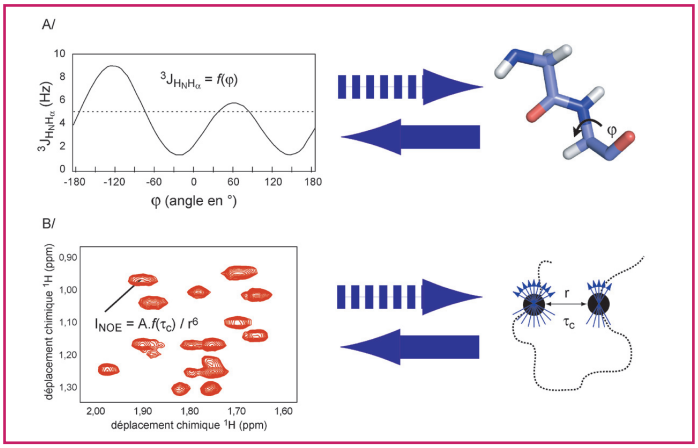


Figure 5 - Indétermination et problème inverse.
L'indétermination des informations structurales obtenues par RMN résulte de plusieurs effets différents. A/ Dans le cas de la constante de couplage, il n'existe pas de fonction réciproque permettant d'obtenir un angle dièdre à partir de la mesure d'une constante de couplage ^3J. Par exemple, la mesure d'un couplage de 5 Hz entre un proton amide et un proton Hα peut correspondre à quatre valeurs différentes de l'angle dièdre φ. B/ Dans le cas d'une distance mesurée à partir de l'intensité d'un pic de corrélation sur une carte NOESY, l'indétermination provient de la présence possible de mouvements internes (qui se traduisent par des variations du temps de corrélation τ_c en fonction de la distance considérée), ou bien de la superposition des tâches de corrélation sur la carte NOESY.

Une palette de contraintes qui s'enrichit

Si les contraintes de distances issues des NOE constituent encore et toujours le noyau central de l'information structurale, le répertoire des observables RMN utilisées dans le calcul de structures s'est considérablement enrichi au cours des dernières années. Parmi les plus utiles et les plus populaires, se trouvent des mesures qui permettent de palier les limitations intrinsèques des contraintes extraites des NOE ou des constantes de couplage. Ainsi, la mesure de constantes de couplage dipolaire résiduel, ou RDC, a été particulièrement développée car elle offre une information qui complète naturellement celle issue des NOE [10]. L'expérience consiste à mesurer l'interaction dipolaire entre deux noyaux séparés par une distance fixe sur des échantillons de protéine placés dans un milieu anisotrope, c'est-à-dire présentant une orientation privilégiée dans l'espace. Les interactions dipolaires, dont l'effet sur les fréquences de résonance est en principe annulé par les mouvements de diffusion rotationnelle de la molécule, se manifestent alors dans le spectre de la protéine sous la forme d'un dédoublement des raies de résonance et de variations des valeurs de constantes de couplage. Cette contribution, qui est de l'ordre de quelques hertz, dépend de l'orientation par rapport au champ magnétique statique du vecteur reliant

les deux noyaux en interaction. L'analyse de l'ensemble de ces contributions mesurées pour une protéine donnée permet de définir l'orientation préférentielle de la protéine par rapport au champ magnétique et fournit des contraintes sur l'orientation relative des vecteurs internucléaires. Différents milieux anisotropes ont été proposés pour l'étude des protéines comme des gels de polyacrylamide contraints, des agents surfactants ou des bicelles phospholipidiques, pour ne citer que les plus courants. Dans la pratique, les contraintes RDC sont surtout utilisées dans l'affinement de structure, ou bien pour définir l'orientation relative de domaines structuraux dans le cas de systèmes plus complexes.

L'utilisation de sondes paramagnétiques permet, dans certains cas, de combler la faible portée des distances obtenues par NOE (5 Å) en offrant des informations à longue distance. La méthode consiste à greffer un centre paramagnétique sur la surface de la protéine et à mesurer la relaxation induite par la présence de l'espèce paramagnétique sur les noyaux. La relaxation paramagnétique s'exerce jusqu'à des distances d'une trentaine d'angströms du centre paramagnétique, ce qui offre un gain considérable par rapport aux contraintes NOE. Les composés comportant un groupe nitroxyde sont en général privilégiés, car ils sont assez compacts et ils permettent un contrôle efficace du phénomène de relaxation paramagnétique, mais il est également possible d'utiliser des métaux paramagnétiques associés à des groupes chélateurs. L'utilisation de ces sondes a évolué depuis son introduction par G. Wagner en 2000 [11], et une des retombées significatives est la mise en évidence de complexes transitoires dans les phénomènes de reconnaissance protéine-protéine [12].

Retour aux sources : extraire l'information structurale contenue dans les fréquences de résonance

L'attribution d'une fréquence de résonance (ou déplacement chimique) à un noyau ou à un groupe de noyaux de la protéine étudiée constitue un préalable obligé du calcul de structure, car elle permet d'associer les paramètres structuraux issus des mesures RMN à des régions spécifiques de la protéine (*figure 3* p. 50). Ces fréquences dépendent fortement de la structure locale et globale de la protéine, et contiennent donc une information structurale essentielle. Les fréquences mesurées pour les noyaux situés sur le squelette peptidique, comme le carbone Cα par exemple, dépendent fortement de la structure secondaire dans laquelle se trouve l'acide aminé. La corrélation étroite entre structure secondaire et déplacements chimiques est exploitée de façon systématique par le programme TALOS, qui propose des valeurs pour les angles dièdres φ et ψ, une fois l'attribution connue [13]. Ce programme recherche, dans une banque validée associant structures et fréquences, l'ensemble des tripeptides présentant des similarités à la fois de séquences et de fréquences avec celles de la protéine étudiée. L'utilisation des fréquences de résonance pour extraire des informations sur le repliement global d'une protéine a été longtemps limitée par la complexité liée au calcul de déplacement chimique dans le cas de macromolécules. Cependant, l'augmentation du nombre de protéines étudiées par RMN a permis la mise au point de modèles semi-empiriques fiables pour le calcul de déplacements chimiques. Ces modèles ont été exploités dans plusieurs approches qui permettent d'obtenir une structure 3D complète à partir des fréquences des noyaux d'une protéine. Le logiciel CS-ROSETTA, par exemple, utilise une approche qui combine la recherche de fragments similaires dans la banque PDB et le calcul de déplacements chimiques à l'aide du logiciel SPARTA [14].

La modélisation : un outil essentiel pour exploiter les données RMN

Le calcul d'une structure à partir de données RMN appartient à la classe des problèmes dits « inverses », mal conditionnés comme cela est illustré par l'absence de fonction réciproque pour l'équation de Karplus (*figure 5*). Dans ce type de problèmes, les paramètres expérimentaux peuvent être calculés à partir de l'ensemble des solutions de façon univoque, mais pas l'inverse. Le principe du calcul de structure consiste alors à explorer l'espace des solutions le plus efficacement possible en ne retenant *in fine* que les modèles dont les paramètres recalculés sont en accord avec l'expérience. Beaucoup de progrès ont été réalisés ces dernières années pour optimiser ce processus. Ils ont permis d'améliorer l'efficacité de l'exploration conformationnelle, la convergence des calculs, la tolérance aux erreurs ou la prise en compte de la dynamique des systèmes.

Une fonction cible hybride

L'accord entre un modèle et les données expérimentales est quantifié par une fonction de plusieurs variables (qui sont les coordonnées du modèle) appelée fonction cible. Cette fonction rend également compte des contraintes liées à la chimie de la protéine comme le respect de la géométrie idéale des liaisons covalentes ou l'optimisation des interactions non liées. Selon le cas, des termes supplémentaires pourront être ajoutés pour favoriser des interactions interatomiques favorables, comme l'établissement de liaisons hydrogène intramoléculaires ou l'exposition de groupes polaires au solvant. La recherche des minima de la fonction cible permet alors d'obtenir des modèles moléculaires qui satisfont à la fois à la géométrie de la chaîne polypeptidique et à son repliement, sous la contrainte des données RMN. La nature différente des informations contenues dans chaque type de données RMN contribue au caractère complexe de la fonction cible. Par exemple, certaines distances issues des NOE apportent des contraintes topologiques très fortes sur la chaîne, alors qu'un angle dièdre issu de la mesure d'une constante de couplage aura un impact plus localisé sur le modèle. Dans le cas d'un équilibre rapide entre plusieurs conformations, le problème devient encore plus complexe, car les valeurs moyennes calculées pour différents types de contraintes deviennent alors incohérentes (*figure 6*). Ce phénomène est lié à un biais systématique pour les petites distances introduit par la moyenne de l'interaction dipolaire. Les travaux les plus récents ont permis de bien comprendre les biais introduits par le calcul de moyenne et de les exploiter pour rendre le calcul de structure plus tolérant vis-à-vis des erreurs d'attribution ou de calibration. Plusieurs distances peuvent ainsi être regroupées et seule la moyenne est prise en compte dans le calcul de la fonction cible. Les distances non respectées (qui sont en général trop longues dans le modèle) ne contribuent alors que faiblement à la moyenne et n'affectent pas la convergence du calcul de structure. Cette méthode est utilisée, par exemple, pour l'interprétation des NOE ambigus qui ne peuvent pas être attribués à un couple unique de protons (*figure 7*) [15]. La forme mathématique de la fonction cible a également fait l'objet de progrès très récents. Ainsi,

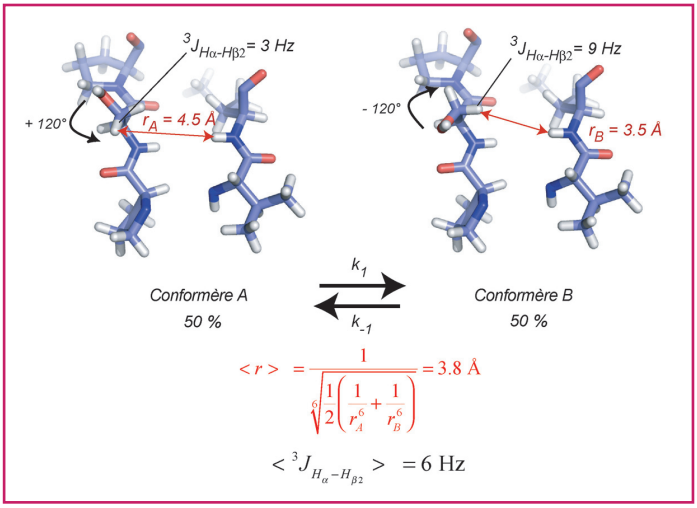


Figure 6 - Des moyennes trompeuses.
Les paramètres RMN donnent des valeurs moyennes qui résultent de la contribution d'un très grand nombre de molécules. Dans le cas d'un système dynamique, en échange entre plusieurs conformations, la valeur moyenne dépend à la fois de la nature de l'interaction à l'origine de la mesure ainsi que de la vitesse de transition entre les différentes conformations de la molécule (k_1). Lorsque cette vitesse est plus grande que la fréquence caractéristique de l'interaction (quelques Hz dans le cas d'un couplage scalaire par exemple), on mesure une moyenne arithmétique à partir des valeurs caractéristiques de chaque état. Dans l'exemple simple de la rotation d'une chaîne latérale d'une thréonine, la fluctuation entre deux rotamères conduit à une valeur moyenne de 6 Hz, dont l'interprétation en termes d'une valeur unique de l'angle dièdre conduira à une conclusion erronée. La moyenne d'une interaction dipolaire, quant à elle, fait intervenir l'inverse de la distance, élevée à la puissance sixième, ce qui favorise la contribution des petites distances. Dans le cas de l'échange ci-dessus, la valeur moyenne favorisera la conformation B, pour laquelle la distance entre un proton amide distant et le proton H_β de la chaîne latérale est la plus courte. L'analyse simultanée de la constante de couplage et des valeurs de NOE fait apparaître une incohérence (la distance extraite du NOE ne correspond pas à l'angle dièdre déduit de la constante de couplage) qui indique la présence de plusieurs conformations.

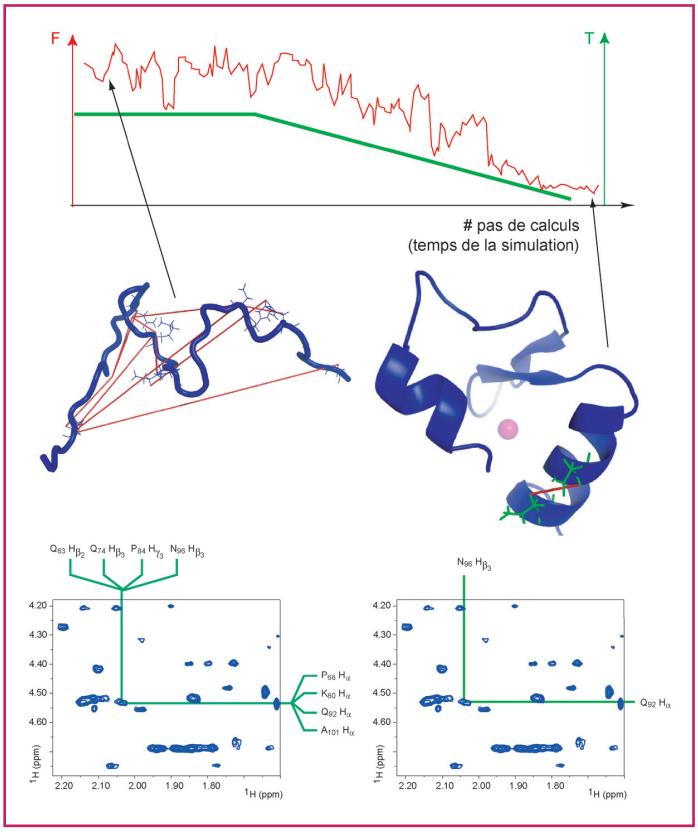


Figure 7 - Recuit simulé et attribution des NOE ambigus.
Le calcul d'un modèle de structure consiste à rechercher le minimum d'une fonction F, qui dépend des coordonnées de la molécule, et qui donne une mesure de l'écart entre les paramètres RMN calculés à partir du modèle et les paramètres expérimentaux. Cette fonction, appelée fonction cible, possède de nombreux minima secondaires, ce qui rend difficile la recherche du minimum principal (qui correspond au jeu de coordonnées reproduisant au mieux les contraintes expérimentales). La recherche de ce minimum est réalisée au cours d'un algorithme de recuit simulé, qui comporte une phase initiale dite à haute température permettant d'explorer un espace très large de solutions. Au cours de cette phase, des structures associées à des valeurs importantes de la fonction cible peuvent être temporairement acceptées, facilitant la sortie des minima locaux. Pendant la phase de refroidissement, seules les solutions conduisant à une réduction de la fonction cible sont considérées. Une des innovations majeures dans le domaine du calcul de structures par RMN a consisté à coupler l'algorithme de recuit simulé avec l'interprétation des NOE. Dans la phase exploratoire à haute température, l'ensemble des combinaisons de couples de protons pouvant expliquer l'observation d'une tâche de corrélation sur une carte NOESY est pris en compte. Dans le cas ci-dessus, la superposition des fréquences de quatre protons différents dans chaque dimension conduit à seize combinaisons possibles. Les propriétés de la moyenne de l'interaction dipolaire permettent d'inclure l'ensemble des hypothèses dans les étapes initiales du calcul, sans affecter les propriétés de convergence. Dans la structure finale, seule une combinaison est retenue.

l'introduction de fonctions de type log-normales dans le calcul de la fonction cible a permis d'améliorer la convergence du calcul de structure [16].

Le protocole de recuit simulé : un algorithme efficace pour l'exploration conformationnelle

La recherche de l'ensemble des conformations en accord avec les données expérimentales repose sur une exploration aussi large que possible de l'espace conformationnel accessible à la protéine. Ce problème est similaire à celui du repliement des protéines qui est à l'origine du paradoxe de Levinthal : en ne considérant qu'un nombre limité de conformations pour chaque acide aminé d'une chaîne peptidique, le nombre de combinaisons devient très vite énorme pour l'ensemble de la chaîne, ce qui interdit son exploration systématique. Pourtant, dans une cellule, les mécanismes de repliement permettent d'obtenir la conformation globale la plus stable en quelques millisecondes. De la même façon, des algorithmes informatiques d'optimisation efficaces permettent de proposer des solutions à des problèmes très complexes dans un temps réduit. Les méthodes employées pour le calcul de structures de protéines par RMN utilisent les algorithmes de simulation de la dynamique moléculaire : ils permettent de reproduire les mouvements aléatoires des atomes pendant un temps donné. L'amplitude de ces mouvements est liée à la température. Le protocole classiquement employé est appelé « recuit simulé », par analogie avec les processus utilisés en métallurgie. C'est une méthode heuristique empirique dans laquelle on procède à un réchauffage (recuit) rapide de la molécule à haute température (qui permet une exploration de l'espace conformationnel), suivi d'un refroidissement contrôlé ayant pour effet de minimiser la valeur de la fonction cible (*figure 7*). Cette fonction contient l'information expérimentale sous la forme de pseudo-potentiels, dont la contribution peut varier au cours du calcul.

Description statistique de l'ensemble des solutions

L'estimation de l'incertitude sur les modèles moléculaires obtenus à l'issue de la procédure décrite ci-dessus est acquise en répétant le calcul d'optimisation un grand nombre de fois, à partir de structures initiales différentes. Des statistiques sont alors effectuées sur l'ensemble des modèles caractérisés par une valeur de la fonction cible la plus basse

possible. Une distribution unimodale caractérise par exemple l'unicité de la solution. Cependant, ces distributions sont fortement dépendantes du poids respectif de chacune des contraintes expérimentales et topologiques dans la fonction cible, ce qui affecte l'estimation de l'incertitude sur le modèle final. Afin de rendre le calcul plus objectif, une approche basée sur les probabilités a été proposée dans le groupe de M. Nilges à l'Institut Pasteur [17]. Cette méthode repose sur une évaluation *a priori* de l'ensemble des solutions possibles (sous la forme de modèles moléculaires). Chaque solution est ensuite associée à une probabilité en fonction de l'accord entre les paramètres calculés à partir du modèle et les données expérimentales. Les modèles qui possèdent les probabilités les plus importantes sont alors sélectionnés et leur analyse permet une estimation non biaisée de l'ensemble des solutions définies par le jeu de données expérimentales. L'utilisation de cette approche est pour le moment limitée par le temps de calcul qui est très important, mais le traitement statistique des données expérimentales associé à l'analyse bayésienne(3) a fait progresser la prise en compte de l'incertitude sur les données RMN dans le calcul de structure par les méthodes traditionnelles.

Conclusions

La détermination de structures de protéines en solution par RMN a bénéficié d'importants efforts de développement au cours des dix dernières années. Les chercheurs disposent à présent d'outils instrumentaux, méthodologiques et logiciels très efficaces et, dans les cas les plus favorables, la résolution d'une structure de protéine de taille modeste (inférieure à 20 kDa) peut être réalisée en quelques semaines. Libérés des contraintes qui mobilisaient, il y a encore peu de temps, des ressources importantes pour la détermination d'une seule structure, les laboratoires se tournent aujourd'hui vers des problèmes plus complexes, comme l'étude de grands complexes moléculaires (protéasome, ribosome, exosome, nucléosome...), la caractérisation structurale d'états excités minoritaires ou encore la caractérisation d'interactions à l'intérieur des cellules. Dans chaque cas, les défis technologiques sont considérables et trouvent généralement leur solution dans l'association étroite entre des méthodes spectroscopiques de pointe et les dernières innovations de la biologie moléculaire. Le regard porté par la RMN révèle alors souvent la nature dynamique et subtile des protéines, et plus généralement des molécules biologiques. En associant structures, mouvements et énergie, la RMN est appelée à jouer un rôle important dans les nouvelles approches intégratives de la biologie structurale.

Notes et références

(1) PDB (« protein data bank ») : site unique de dépôt des structures 3D de protéines.
(2) MAGI-1 : protéine membre de la famille des « membrane-associated guanylate kinase ».
(3) Analyse bayésienne : méthode statistique intégrant explicitement une distribution de probabilités *a priori* fondée sur une opinion subjective ou des données objectives comme les résultats d'une recherche antérieure (HTAGlossary.net).

[1] Wüthrich K., NMR - this other method for protein and nucleic acid structure determination, *Acta Crystallogr. D Biol. Crystallogr.*, **1995**, *51*, p. 249.
[2] Birlirakis N., Bontems F., Guittet E., Leroy J.-L., Lescop E., Louis-Joseph A., Morellet N., Sizun C., Van Heijenoort C., Nuclear magnetic resonance: a tool for structural biology, *L'Act. Chim.*, **2011**, *353-354*, p. 100.
[3] Güntert P., Automated structure determination from NMR spectra, *Eur. Biophys. J.*, **2009**, *38*, p. 129.
[4] Baron M., Norman D.G., Campbell I.D., Protein modules, *Trends Biochem. Sci.*, **1991**, *16*, p. 13.
[5] Fernandez C., Wider G., TROSY in NMR studies of the structure and function of large biological macromolecules, *Curr. Opin. Struct. Biol.*, **2003**, *13*, p. 570.
[6] Liu D., Xu R. Cowburn D., Segmental isotopic labeling of proteins for nuclear magnetic resonance, *Methods Enzymol.*, **2009**, *462*, p. 151.
[7] Gans P., Hamelin O., Sounier R., Ayala I., Dura M.A., Amero C.D., Noirclerc-Savoye M., Franzetti B., Plevin M.J., Boisbouvier J., Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins, *Angew. Chem. Int. Ed. Engl.*, **2010**, *49*, p. 1958.
[8] Sprangers R., Kay L.E., Quantitative dynamics and binding studies of the 20S proteasome by NMR, *Nature*, **2007**, *445*, p. 618.
[9] Canet D., Boubel J., Soulas E.C., *La RMN : Concepts, Méthodes et Applications*, UniverSciences, Dunod, **2002**.
[10] Bax A., Grishaev A., Weak alignment NMR: a hawk-eyed view of biomolecular structure, *Curr. Opin. Struct. Biol.*, **2005**, *15*, p. 563.
[11] Battiste J.L., Wagner G., Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear Overhauser effect data, *Biochemistry*, **2000**, *39*, p. 5355.
[12] Iwahara J., Clore G.M., Detecting transient intermediates in macromolecular binding by paramagnetic NMR, *Nature*, **2006**, *440*, p. 1227.
[13] Cornilescu G., Delaglio F. Bax A., Protein backbone angle restraints from searching a database for chemical shift and sequence homology, *J. Biomol. NMR*, **1999**, *13*, p. 289.
[14] Shen Y., Lange O., Delaglio F., Rossi P., Aramini J.M., Liu G., Eletsky A., Wu Y., Singarapu K.K., Lemak A., Ignatchenko A., Arrowsmith C.H., Szyperski T., Montelione G.T., Baker D., Bax A., Consistent blind protein structure generation from NMR chemical shift data, *Proc. Natl Acad. Sci. USA*, **2008**, *105*, p. 4685.
[15] Nilges M., Macias M.J., O'Donoghue S.I., Oschkinat H., Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin, *J. Mol. Biol.*, **1997**, *269*, p. 408.
[16] Nilges M., Bernard A., Bardiaux B., Malliavin T., Habeck M., Rieping W., Accurate NMR structures through minimization of an extended hybrid energy, *Structure*, **2008**, *16*, p. 1305.
[17] Rieping W., Habeck M., Nilges M., Inferential structure determination, *Science*, **2005**, *309*, p. 303.

Y. Nominé

B. Kieffer

Yves Nominé est maître de conférences, rattaché à l'Université de Strasbourg, dans l'équipe Oncoprotéines de l'Institut de Recherche de l'École supérieure de biotechnologie de Strasbourg*.

Bruno Kieffer (*auteur correspondant*) est professeur à l'École supérieure de biotechnologie de Strasbourg, responsable du groupe de RMN biomoléculaire à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire de l'Université de Strasbourg**.

* Équipe Oncoprotéines, UMR 7242 CNRS/Université de Strasbourg, Institut de Recherche de l'École supérieure de biotechnologie de Strasbourg, Boulevard Sébastien Brandt, BP 10413, F-67412 Illkirch Cedex.

** Biomolecular NMR Group, UMR 7104 CNRS/Université de Strasbourg, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Rue Laurent Fries, F-67400 Illkirch.
Courriel : bruno.kieffer@igbmc.fr