Signature of protein adaptation to warm deep sea environments: the case of Initiation Factor 6 studied by molecular simulation and neutron scattering.

Paolo Calligari

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS 6

Spécialité:

INTERFACES DE LA BIOLOGIE AVEC LA PHYSIQUE, LA CHIMIE ET L'INFORMATIQUE

Pour obtenir le grade de

Docteur de l'Université Pierre et Marie Curie

Presenté par:

PAOLO CALLIGARI

SIGNATURE DE L'ADAPTATION DES PROTÉINES À L'ENVIRONNEMENT DES FONDS MARINS CHAUDS:

le cas du Facteur d'Initiation 6 étudié par simulation moléculaire et diffusion de neutrons

Paris, 18 Dcembre 2008

Devant le jury composé par:

Prof. Cristian Micheletti (Rapporteur)

Prof. Roger Fourme (Rapporteur)

Prof. Françoise Gaill (examinatrice)

Dr. Joel Pothier (examinateur)

Dr. Giuseppe Zaccai (examinateur)

Prof. Mark Johnson (examinateur)

Prof. Gerald R. Kneller (directeur de thse)

THÈSE PRÉPARÉE AU SEIN DE:

INSTITUT LAUE LANGEVIN (Grenoble)

et

LABORATOIRE LEON BRILLOUIN (CEA Saclay, Gif-sur-Yvette)

Dedicated to Francesca Romana.

ABSTRACT

Signature of protein adaptation to *warm deep sea* environments: the case of Initiation Factor 6 studied by molecular simulation and neutron scattering.

The protein Initiation Factor 6 (IF6) takes part in the protein synthesis regulation of several organisms. It was also found in archeaebacteria such as *Methanoccoccus Jannascii* which lives in deep-seas near hydrothermal vents where temperature reaches 80°C and pressure is between 250bar and 500bar. The aim of this work was to study for the first time dynamical and structural properties of IF6 produced by *M.Jannaschii* and comparing them with those of the IF6 homologue present in *Saccharomyces cerevisiae* which lives at "normal" environmental conditions (27°C and 1bar).
Molecular simulation gave here new insights into the adaptation of these two proteins to their respective physiological conditions and showed that the latter induced similar dynamical and structural properties: in their respective "natural" conditions, IF6s show very similar structural fluctuations and the characteristic relaxation times which define their dynamical properties shows similar changes when comparing unfavorable conditions to physiological ones. The creation of these *corresponding states* between the two homologues has been interpreted by the fractional Brownian dynamics model and by a novel method for the characterization of protein secondary structures. The latter is presented here in detail together with some examples of other applications. Experimental data obtained from quasi-elastic neutron scattering seemed to support the results obtained by molecular simulations.

Keywords : Initiation Factor 6, molecular dynamics, neutron scattering, pressure, fractional Brownian dynamics model, warm deep sea, protein secondary structure

RÉSUMÉ

Signature de l'adaptation des protéines à l'environnement des fonds marins chauds: le cas du Facteur d'Initiation 6 étudié par simulation moléculaire et diffusion de neutrons.

Le Facteur d'Initiation 6 (IF6) est une protéine qui participe, dans plusieurs organismes, à la régulation de la synthèse des autres protéines. Elle a été trouvée aussi dans l'archaebactérie *Methanoccoccus Jannascii* qui vit au fond de la mer, près des cheminées hydrothermales, où la température atteint 80°C et la pression hydrostatique est entre 250 et 500bar. L'objectif de ce travail a été celui d'étudier pour la première fois les propriétés dynamiques et structurales de la IF6 issue du *M.Jannaschii* en comparaison avec celles de son homologue présent dans le *Saccharomyces cerevisiae* qui vit dans des conditions environnementales "normales" (27°C et 1bar).
La simulation moléculaire nous a permis de montrer que l'adaptation de ces deux protéines aux conditions physiologiques induit des propriétés dynamiques et structurales similaires: dans leur conditions "naturelles" respectives, les deux protéines montrent des fluctuations structurales très similaires et les temps caractéristiques qui identifient leur propriétés dynamiques subissent les mêmes changements dans la transition d'une condition défavorable vers la condition physiologique.
Cette création d' "états correspondants" entre les deux protéines a été étudiée par le modèle de dynamique Brownienne fractionnaire et par une nouvelle méthode pour la caractérisation des structures secondaires des protéines. Cette dernière est présentée en détail avec des brefs exemples d'autres applications. Les données préliminaires obtenues par diffusion de neutrons semblent confirmer les résultats issues des simulations moléculaires.

Mots clés : Facteur d'Initiation 6, dynamique moléculaire, diffusion de neutrons, pression, modèle Brownien fractionnaire, fond marins chauds, structure secondaire des protéines.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

1) Calligari P.A. et al., Inhibition of viral group-1 and group-2 neuraminidases by oseltamivir: a comparative structural analysis by the ScrewFit algorithm. *Biophysical Chemistry*, accepted for publication (2008).

2) Calligari P.A. and Kneller G.R., ScrewFit : a novel approach for continuum protein secondary structure assessments. Submitted (2008).

3) Calandrini V., Hamon V., Hinsen K., Calligari P., Bellisent-Funel M.-C. and Kneller G.R., Relaxation dynamics of lysozime in solution under pressure: combining molecular dynamics and quasielastic neutron scattering. *Chemical Physics*, **345**, 289-297 (2008).

4) Hamon V., Calligari P., Hinsen K., Kneller G.R. Simulation studies of structural changes and relaxation processes in lysozyme under pressure, *J. of Non-Crystalline Solids*, **352**, 4417-4423 (2006).

5) Kneller G.R. and Calligari P., Efficient characterisation of protein secondary structure in terms of screw motions. *Acta Crystallographica* **D62**, 302-311, (2006).

Il est certain que beaucoup de choses vont bien dans la nature,
c'est-à-dire vont en sorte de pouvoir durer et se conserver, ce
qu'elles ne pourraient faire autrement. Mais une infinité d'autres
vont mal (et peut-être un plus grand nombre), sont mal organisées,
moralement et physiquement, avec un immense inconvénient pour
les créatures; ces choses auraient pu à peu de différence près être bien organisées.

— Giacomo Leopardi, *Zibaldone di pensieri* [N. 4248]

REMERCIEMENTS

Ce travail de thèse a été effectué à l'Institut Laue Langevin de Grenoble et au Laboratoire Léon Brillouin au Commissariat à l'Energie Atomique de Saclay, sous la direction du professeur Gerald KNELLER à qui s'adressent mes premiers remerciements pour m'avoir accueilli dans son équipe et pour avoir su équilibrer, dès le primer jour, mon enthousiasme envers la biologie avec des bonnes doses de physique théorique. Je lui serai toujours reconnaissant de ne pas avoir laisser que j'oublie mes origines de physicien.

Je suis pareillement reconnaissant envers Mark JOHNSON, qui à co-diriger ma thèse. A lui vont mes remerciements pour avoir suivi et coordonner mon travail en me laissant toute la liberté d'expérimenter, d'essayer et de me planter devant les problèmes comme le font les "vrais" chercheurs. Je lui doit tout ma gratitude pour la confiance qu'il m'a réservé ainsi que pour son support dans les moments les plus difficiles.
Envers mes deux directeurs un remerciement supplémentaire pour avoir su être amis en outre que simples "chefs"...

Je voudrais aussi exprimer ma gratitude envers les membres du jury: la présidente Françoise GAILL, Cristian MICHELETTI et Roger FOURME qui ont accepté d'être rapporteurs de mon mémoire, Giuseppe ZACCAI et Joel POTHIER.

Je tiens ensuite à remercier les collaborateurs du professeur Kneller, Vania CALANDRINI et Konrad HINSEN pour tout ce qu'ils m'ont appris et pour leur aide précieuse dans mon travail partagé entre les neutrons et les simulations numériques.

Je suis particulièrement reconnaissant envers Jacques OLLIVIER et Jean-Baptiste ARTERO qui ont partagé avec moi "la quotidienneté" de mon travail (et notamment des expériences de laboratoire, pas toujours encourageantes...!). Je remercie Jacques pour ce qui m'a appris, pour sa patience envers les cotés plus étranges (surtout au monde des neutrons) de mon projet de thèse et pour m'avoir accompagné dans nos expériences de diffusion de neutrons au Paul Scherrer Institut (Zurich, Switzerland) et au NIST (Maryland, USA).
Une partie fondamentale de ce projet a été possible seulement grâce aux enseignements de Jean-Baptiste qui m'a guidé dans le monde de la biologie moléculaire et dans les expériences lourds et souvent peu généreux d'expression et purification des protéines. Je sais qu'il a du prouver un goût très particulier en voyant un pauvre physicien faire des bêtises sur la paillasse mais moi j'ai trouvé ma "vengeance" avec la coupe du monde de football....."Campioni del mondo!!!"

Je ne peux pas oublier les autres collègues des laboratoires qui ont vu ma présence, souvent fugace, durant mon séjour à Grenoble: j'ai parcouru ma mémoire des deux années passées et je vous ai tous retrouvés pour votre amitié, vos encouragements, votre patience (je pense surtout aux collègues du DEUTERATION LAB !!), votre aide direct et constant, pour avoir supporté mon travail et pour beaucoup d'autres choses.... Merci!

A mes amis et ma famille:
A Paris: Andrea et Adrien, merci pour les discussions hors cerveau et/ou hors logique, elles ont souvent sauvé ma stabilité mentale. A Grenoble: Mauro, merci pour nos nombreuses pauses café à mi-chemin entre nos rêves et nos ambitions. Merci à Giada, Bea, Fabrizio, i Fratini pour les beaux moments. Un merci générale à tous mes amis pour avoir supporté ma passion pour l'évolutionnisme et pour avoir écouté avec patience mes contes sur l'histoire du pauvre Kiwi.
Un merci à ma famille et à mes amis à Rome qui ont vécu et participé à distance mes petits succès et faillites...

En fin, au dehors de toute cette liste, une seule personne n'a pas besoin d'être remerciée ici parce que elle a vécu tout ce que j'ai vécu et elle sais bien que tout cela n'aurait jamais pu être sans sa présence, son support, sa compréhension et sa confiance pleins d'amour. Cette personne est devenue, au fil de cette thèse, mon épouse et c'est à elle que tous ces années de travail sont dédiés: "grazie amore".

Part I

RÉSUMÉ SUBSTANTIEL

RÉSUMÉ SUBSTANTIEL

0.1 INTRODUCTION

Dès le début du dernier siècle, les acquis sur la matière inanimée ont amenés les physiciens à se demander comment ces connaissances pouvaient être aussi appliquées à la matière biologique. La matière vivante présentait aux chercheurs une hétérogénéité et complexité de cas que ils n'avaient jamais rencontré en physique. Elle a ainsi donné naissance à un nouveau filon de recherche interdisciplinaire qui, alliant les domaines de la physique, la chimie et la biologie, a fourni des résultats très importants et utiles pour des applications successives en médecine et en biotechnologie.
Ce travail de thèse puise ses sources dans le considérable progrès fait dans le domaine de la *physique biologique* notamment en ce qui concerne le développement de nouvelles méthodes de calcul, ainsi que à l'application des techniques de spectroscopie sur des échantillons biologiques.

0.1.1 *Organismes adaptés aux environnements extrêmes*

Organisms are integrated entities, not collections of discrete objects [1]

Les effets de la pression et de la température sur les propriétés structurales des protéines sont visibles à plusieurs échelles, des interactions non-covalentes des très courte portée, telles-que les liaisons hydrogène, jusqu'aux interactions à plus longue portée comme les interactions électrostatiques. Tout changement dû à une variation des variables thermodynamiques peut être interprété par le principe de Le Chatelier, selon lequel toute perturbation sur un système à l'équilibre provoque une réaction contraire de façon à diminuer l'effet de cette perturbation. Dans le cas considéré, l'application de la pression à une protéine favorise des processus qui s'accompagnent d'une diminution du volume. Ces processus agissent directement sur la structure de la protéine, entraînant des changements des interactions non-covalentes

1 *"Les organismes sont des entités intégrées, ils ne sont pas des collections d'objets discrets"*. The spandrels of San Marco and the Panglossian paradigm: a critic to the adaptationist programme, S.J. Gould and R.C. Lewontin, Proc. Royal Soc. London B, **205**, 581-598, (1979).

qui se forment ou se défont à l'intérieur de la structure. Ces changements produisent des effets différents sur les trois niveaux de structuration des protéines (secondaire, tertiaire et quaternaire)[144].

Dans cette thèse nous avons abordé les effets de la pression et de la température dans un contexte particulier, celui des environnements des fonds marins près des cheminées hydrothermales où l'on trouve normalement des pressions et des température très hautes. En outre, comme suggéré par la phrase citée au début de cette section, nous voulons souligner le fait que l'étude de l'adaptation des protéines aux environnements extrêmes pourrait être largement améliorée par la connaissance du cadre général de l'adaptation cellulaire. Dans le texte de cette thèse nous avons donné un bref aperçu du sujet tout en renvoyant à d'autres publications pour des approfondissements de l'argument [85, 93, 182, 158, 107, 151, 191, 58].

La biosphère, c'est à dire la surface de le terre connue comme lieu accueillant la vie, varie entre les régions abyssales des fonds marins et les hauteurs de l'Himalaya et offre une grande variété d'exemples d'organismes adaptés aux conditions extrêmes. D'un point de vue qualitative les limites physiques et chimiques pour accueillir des formes de vie sont les suivantes[92]:

- $-40°C < T < 115°C$
- $P < 1.2kbar$
- $\sim 1 < pH < 11$

D'un point de vue évolutionniste, les organismes qui vivent dans des conditions proches de ces limites ont dû trouver des façons pour compenser tous les effets chimiques et physiques que ces limites mêmes peuvent générer sur l'ensemble de structures constituant les cellules. Autrement dit, les organismes qui sont devenues *extrêmophiles* ont trouvé des stratégies pour protéger leur système de vie des dégâts produits par l'environnement. D'après le peu qu'on connaît sur le fonctionnement de ces mécanismes, il y a deux démarches principalement suivis par les organismes. La première est la compensation des processus de dégradation à travers d'un réglage fin des taux de synthèse des biomolécules de façon qu'il soient compatibles avec le temps moyen de *survie* des molécules. La deuxième est l'incorporation de mutations ponctuelles dans les séquences des protéines à fin d'augmenter leur stabilité structurale, leur permettant de fonctionner plus longtemps. Apparemment, cette stratégie ne produit pas de différences remarquables dans la distribution globale des acides aminés dans

des protéines extrêmophiles par rapport à celles *mesophiles*, c'est à dire issues d'organismes vivant dans des conditions dites normales (température à 300K et pression à1bar) [48].

Par conséquence, il semble assez évident que l'adaptation des biomolecules doit être liée à la présence de différents motifs locaux dans les séquences des acides aminés entraînant une réorganisation des interactions faibles non-covalentes qui règlent la stabilité et la flexibilité des protéines. Le *new deal* pour la stabilité des protéines extrêmophiles se fait donc par une réorganisation des mécanismes sensibles aux changements environnementaux. D'un point de vue physique, une mesure quantitative de la stabilité des protéines peut être obtenue par les différences d'énergie libre de Gibbs, $\Delta G(T, P, N)$, entre l'état le plus stable et les minima locaux les plus proches. Jaenicke et Böhm [96] ont montré que les protéines mesophiles et extrêmophiles partagent un ΔG du même ordre de grandeur ($\sim 50 kJmol^{-1}$) même si les extrêmophiles montrent une plus large variabilité dans l'intervalle $10 - 100 kJmol^{-1}$. Les différences trouvées entre les protéines mesophiles et extrêmophiles correspondent à l'énergie nécessaire à la formation ou à la rupture de quelques interactions non-covalentes, ce qui confirme le rôle fondamental que ces dernières jouent dans le contexte de l'adaptation moléculaire [157, 147, 151, 158].
Le point central dans l'adaptation des biomolécules reste, de toute façon, la conservation des fonctionnalités biologiques représentant un compromis bien équilibré entre stabilité et flexibilité [197, 92]. Le processus d'adaptation se manifeste, donc, sous forme d'une *transformation* des propriétés des protéines mesophiles vers les conditions extrêmes, c'est à dire, vers des conditions physiologiques qui rendent les propriétés moléculaires des variétés extrêmophiles très similaires à celles des mesophiles. Cette translation est obtenue par la réorganisation des liaisons non-covalentes. Plusieurs expériences ont mis en évidence ce processus [217] et elles ont montré que les propriétés dynamiques de quelques enzymes mesophiles à une température de 25°C sont très similaires à celles des leurs homologues extremophiles à 70°C. Des résultats analogues ont été obtenus par une étude de simulation de dynamique moléculaire sur les différents homologues de la rubredoxine [69].
Dans ce contexte Jaenicke [93] a supposé que l'adaptation moléculaire à des environnements extrêmes puisse fonctionner grâce à un maintien d'*états correspondants* entre les environnements *natifs*, en tenant compte de la topologie générale, de la flexibilité et de l'hydratation des protéines. Même si cette idée a trouvé plusieurs confirmations dans le passé, des études récentes ne semblent pas la confirmer entièrement[158].

0.1.2 *L'environnement typique des fonds marins chauds*

Contrairement à l'opinion générale, les environnements terrestres, où la pression est d'environ 1bar, occupent seulement 1% du volume total de la biosphère. Une partie considérable de la surface terrestre (70%) est couverte par les océans qui ont une profondeur moyenne de 3800 mètres et donc une pression moyenne de 380bar. Plus du 60% de la biosphère marine est à 1000 mètres sous le niveau de la mer. La vie au dessous de cette profondeur, normalement considérée comme le limite pour les environnements dits *fonds marins*, doit faire face à plusieurs conditions défavorables liées aux changements de la pression (de 1 à 1.1 kbar) et de la température (de 1 à 110°C) qui ont des impacts très important sur les mécanismes vitaux. Une nouvelle ligne de recherche sur les organismes qui vivent dans les sédiments marins a été entreprise entre le XIXème et le XXème siècle par Certes[30, 29] suivi par ZoBell et Johnson [220] dont les recherches ont mieux éclairées les stratégies de survie employées par les organismes pour faire face aux conditions extrêmes.

Grâce à ce courant de recherche nous pouvons aujourd'hui distinguer les organismes en fonction de leur capacité d'adaptation aux hautes pressions: les *barophiles* (ou *piezophiles*) qui ont une croissance optimale à des pressions supérieures à celle atmosphérique; les *barotolerants* (ou *piezotolerants*) qui sont des organismes capables de vivre à hautes pressions tout en ayant leur taux de croissance optimale à pression atmosphérique; enfin, les organismes *piezosensibles* qui se distinguent des autres parce que leur croissance est très sensible aux pressions élevées.

A la fin des années 70, les premières colonies d'organismes dans les fonds marins ont été trouvées à proximité des cheminées hydrothermales, des courants qui se forment où la lave extrudée se refroidit en se contractant et permet à l'eau de mer d'entrer dans les fissures des rochers basaltiques à peine formées. L'eau de mer est en suite expulsée très enrichie de métaux lourds formant des courants à très haute température [98].
La vie autour des cheminée hydrothermales est devenue très rapidement l'un des sujets les plus intrigants dans le domaine de la biologie des fonds marins et la plupart des organismes barophiles connus à ce jour sont *de facto* aussi thermophiles, c'est à dire ils ont montré un plus haut taux de croissance à des températures plus élevées (typiquement entre 50 et 90°C) que celles caractéristiques des fonds marins (environ 2°C) [1].

Cet environnement des fonds marins chauds a introduit sous le plan biologique et évolutionniste plusieurs questions de grand intérêt qui ont demandé aussi des nouvelles explications par la biochimie et la physique.

METHANOCOCCUS JANNASCHII Cette thèse a comme objet l'étude d'une protéine produite par une archaebacterie, le *Methanococcus Jannaschii*, qui vit près des cheminées hydrothermales. Cet organisme, découvert en 1963 [102], vit normalement à des températures entre 48 et 94 °C, avec une température optimale de 85°C, et à des pressions supérieures aux 200 bar (profondeur *in situ* 2600 mètres).

Le génome du *M Jannaschii* a été le premier parmi les génomes des archaebacteries dont on a obtenu le sequençage complet [18] ce qui révéla des relations évolutives complexes entre les archaea, les eucaryotes et les bactéries: seulement moins de la moitié des gènes trouvés dans son génome pouvait, en fait, être reliée a ceux des autres organismes. Dans le cadre de cette thèse, d'un point de vue de la biologie moléculaire, la propriété la plus intéressante de cet organisme est due au fait que tout en partageant les mécanismes de biosynthèse des eucaryotes, le M.Jannaschii possède des gènes pour la fonction d'initiation du processus de traduction homologues soit à ceux des eucaryotes soit à ceux des bactéries. Cette fonction est une partie fondamentale de la traduction de l'ARN messager et sera présentée en détail plus avant dans le texte.

Des études ont montré que le métabolisme du *M Jannaschii* et sa croissance à haute température sont favorisés par la pression jusqu'à 750 bar et que les limites pour la croissance sont atteints à 90°C soit à basse ou à haute pression [102]. Ces résultats suggèrent que le *M.Jannaschii* est effectivement un organisme barophile et pas seulement barotolerant. Cette conclusion semble être aussi confirmé par des études de réaction enzymatique qui ont montré que l'application d'une pression de 500bar peut augmenter jusqu'à cinq fois le temps moyen de vie de l'hydrogènase du M.Jannaschii à 90°C [74] et jusqu'à 3 fois celui de la protease à 125 °C [138]. Néanmoins, des études récentes ont montré que la pression inhibait la fonction du 20S proteasome du *M Jannaschii* [55], suggérant ainsi que dans certains cas la moindre activité d'une protéine à la pression *native* pourrait être un mécanisme de régulation conférant des avantages à la cellule entière. Par conséquent, un comportement bien plus complexe de la cellule par rapport aux hautes pressions et aux hautes températures devra être envisagé.

0.1.3 *Le facteur d'anti-association*

Nous présentons ici une brève introduction sur la protéine qui fait l'objet de ce travail de thèse, appelé *Facteur d'anti-association* [170, 199].
Le facteur d'anti-association fait partie de la famille des facteurs d'initiation et il prend aussi le nom de *Facteur d'initiation 6* (IF6).

Le IF6, comme les autres facteurs d'initiation, prend part aux premières étapes du processus de traduction ARN-messager des ribosomes. Le rôle du IF6 dans l'initiation de la traduction est bien plus complexe que celui des autres facteurs d'initiation parce qu'il engendre diverses fonctions qu'on connaît très mal. D'abord, IF6 n'agit pas comme un *vrai* facteur d'initiation parce qu'il ne forme aucun complexe d'initiation, c'est à dire qu'il ne forme pas de complexe macromoléculaire avec les mARN, tARN et les sous-unités du ribosome dans les premières étapes de la traduction. Ceci est prouvé par le fait que la déplétion du IF6 ne bloque pas la traduction du mARN in vitro [178].
Les fonctions principales de cette protéine peuvent être résumées comme suit:

- IF6 est requis dans la biogenèse de la sous-unité ribosomale 60S soit dans les archaebactéries que dans les eukaryotes [173, 178, 177, 212]

- elle agit *de facto* comme un facteur qui règle l'association des sous-unités ribosomales 60S et 40S dans le cytoplasme [178, 28]

Ces fonctions, conservées dans plusieurs organismes, ont été prouvées par un grand nombre d'études et sont vérifiées par la présence de cette protéine dans le cytoplasme et dans les noyau des cellules. Dans le cadre de ce travail de thèse il est important aussi de noter que le IF6 a été montré comme très sensible aux hautes températures : dans certains homologues eukaryotes[178, 10], l'activité du IF6 a affiché un optimum autour de 37°C alors qu'elle était totalement absente entre 50 et 60 °C. A l'état actuel, il parait très difficile d'établir si cet arrêt des fonctions du IF6 correspond aussi à une dénaturation.

D'un point de vue moléculaire, les IF6 sont des protéines d'environ 26kDa qui partagent une séquence conservée de 224 résidus avec une similarité de 30%. Toutes les formes eukaryotes de la IF6, contiennent une partie carboxy-terminale de 21 acides aminés. Même si cette *queue* ne semble pas être directement impliquée dans la fonction d'anti-association du IF6[66], il a été prouvé qu'elle était reliée à cette dernière[28].

En 2000, les premières structures moléculaire du IF6 ont été résolues par des études de cristallographie par rayon-X [66]. Il s'agit des structures de deux homologues du IF6:

celui du *M.Jannaschii* et celui du *Saccharomyces ceraevisie*. Les deux structure ont révélé une pseudo-symétrie interne crée par la disposition, autour d'un centre commun, de cinq copies d'un domaine α/β d'environ 45 résidus. Chaque domaine contient: une longue hélice α, une autre plus courte ou une hélice 3_{10} et trois brin β. Contrairement à la structure des premiers 224 résidus qui est connue et a été prouvée par *homology modeling* être conservée evolutivement[66], très peu a été révélé sur la structure de l'extension C-terminale de 21 acides aminées.

D'un point de vue structural, il faut aussi remarquer que les cinq domaines sont disposés d'une façon qui ne crée pas un centre hydrophobique. Au contraire, dans leur association, ils produisent un "tore" hydrophobe qui, à son tour, forme un centre creux dans la structure de la protéine. Cette cavité est assez large pour faire passer plusieurs molécules d'eau. Seize de ces molécules ont été trouvé aussi dans les structures cristallographiques, dans une configuration très ordonnée à couches pentagonales et dont les hydrogènes interagissent avec les oxygènes carbonyles des résidus dans les brins β voisins. Dans l'homologue provenant de la levure (*Saccharomyces ceraevisie*), le IF6 montre cette cavité fermée par une arginine (résidu 61) dont le groupement guanidinium forme plusieurs liaisons hydrogènes avec les glycines des courtes hélices α de chaque domaine.

0.2 MATÉRIELS ET MÉTHODES

Cette section propose un bref aperçu des techniques utilisées pour la préparation des échantillons qui ont été étudiés par diffusion des neutrons et de ceux qui ont été simulés par dynamique moléculaire. Pour une présentation complète du cadre théorique dans lequel ces techniques ont été appliquées nous renvoyons au deuxième chapitre de la thèse et aux références [3, 56, 13, 129].
Dans le texte qui suit, les deux IF6 homologues seront indiqués par les acronymes suivants:

- **aIF6**: IF6 extremophile produit par *Methanococcus Jannaschii*
- **eIF6**: IF6 mésophile produit par *Saccharomyces cerevisiae*

0.2.1 *Production des échantillons expérimentaux*

Les échantillons expérimentaux étudiés ici par diffusion de neutrons, ont étés produits à travers des étapes préliminaires qui demandaient notamment des connaissances en biologie

moléculaire. Cette thèse rapporte un protocole développé pour le IF6 qui vise à obtenir un rendement suffisant et compatible avec les quantités d'échantillons requis par les expériences de diffusion de neutron. Ce protocole suit le schéma usuel pour la production de protéines et qui peut être résumé comme suit:

- **Clonage** de la partie de l'ADN génomique de *Methanococcus Jannaschii* et *Saccharomyces cerevisiae* qui codifie le IF6.
- **Expression** du gène identifié dans un organisme hôte, dans le cas considéré ici le *E.coli*, à fin de produire une grande quantité de protéines.
- **Purification** de la protéine produite par l'organisme hôte à fin de la séparer des autres protéines et d'obtenir une solution assez pure ne contenant que la IF6.

Afin d'améliorer le rendement, du protocole de purification, nous avons utilisé une queue N-polyhistidine N-terminale (H-Tag) qui a permis de maintenir une concentration suffisante aussi pour des expériences de diffusion de neutrons à hautes pressions, ce qui, comme on le verra après, demande l'utilisation de grands volumes de solution. Néanmoins, des raisons techniques ont fortement limité le clivage de cette queue "marqueur" du IF6 en grands volumes de solution. Pour cette raison, ce clivage a été fait seulement à pression ambiante où les expériences pouvaient être faites avec des volumes adaptés. Il a été montré que la présence de la H-Tag n'a pas d'effets significatifs sur la structure des protéines [23] mais, vu son exposition au solvant qui entoure les protéines, elle pourrait avoir un effet dynamique important. Pour cette raison, afin de vérifier les effets produits dans le cas du IF6, des simulations de dynamique moléculaire ont été réalisés aussi sur un modèle du aIF6 jointe avec la H-Tag.

CLIVAGE DU FRAGMENT C-TERMINAL DU EIF6 Pendant les tests préliminaires pour l'optimisation du protocole (cfr. ci-dessus) un clivage protéolytique du eIF6 a été observé (voir Figure 13, dans le texte de cette thèse, pour vérification SDS-PAGE) soit dans la partie soluble soit dans celle insoluble des fractions de purification du lysate cellulaire. A cause de son poids moléculaire modéré, le fragment clivé a été supposé être le même trouvé par Groft *et al.* [66]. Ces derniers ont rapporté le fait que les tentatives d'expression et purification du eIF6 étaient rendues très difficiles par le clivage protéolytique du fragment C-terminale.
La partie clivée du eIF6 est formée de 21 acides aminés avec une séquence, affichée dans le Tableau 1, présente uniquement dans les homologues eukaryotes du IF6 dont on sait

très peu. En outre, une caractérisation par alignement multiple avec des bases de données d'autres séquences connues a montré que ce fragment n'est présent que dans les IF6, ce qui prouve son importance pour la fonctionnalité du IF6 même ou bien pour son histoire évolutive.

Table 1: La séquence du fragment C-terminal du eIF6 (CTAIL). La numérotation suit celle utilisée dans le PDB 1G62.

Glu225	Asp226	Ala227	Gln228	Pro229	Glu230	Ser231	Ile232
Ser233	Gly234	Asn235	Leu236	Arg237	Asp238	Thr239	Leu240
Ile241	Glu242	Thr243	Tyr244	Ser245			

Afin de mieux comprendre la structure *native* de cette queue C-terminale, nous avons examiné sa structure secondaire à l'aide d'outils de prédiction qui ont vérifié la probable présence d'une hélice-α dans la partie finale du fragment, la région DTLIE. Ce résultat a été confirmé par l'inspection du profil hydrophobe du fragment à l'aide de l'échelle de Kyte/Doolittle [122] qui montre une augmentation de l'hydrophobicité dans la même région. Finalement, cette étude préliminaire nous a convaincus que le fragment C-terminal du eIF6 joue un rôle significatif soit dans la dynamique soit dans la structure du eIF6 et donc aussi dans sa fonction. Cette idée est renforcée par d'autres résultats qui indiquent que les domaines C-terminaux contribuent à la localisation du eIF6 dans le noyau cellulaire [6]. Ces conclusions nous suggèrent l'importance d'exprimer et de purifier le eIF6 avec la queue C-terminale de 21 acides aminés et à ce propos plusieurs tests ont été mis en place pour réduire le clivage pendant les différentes phases de production. Le protocole présenté dans ce texte a donné les meilleurs résultats avec une clivage réduit d'environ 30% (voir la Figure 14 dans le texte pour une vérification par spectrométrie de masse MALDI).

STABILITÉ DU EIF6 Afin d'aborder une étude visant les effets des conditions extrêmes sur les propriétés d'une protéine, il est nécessaire, tout d'abord, d'établir quelles sont les limites qui définissent les conditions normales pour la même protéine. Dans le cas des IF6, la connaissance de ces limites est très limitée à cause d'un manque d'études approfondies sur les propriétés chimiques de cette protéine. Ainsi, si pour le aIF6 on peut facilement supposer

au moins une réponse réversible aux hautes pressions et températures, les conclusions sur les eIF6 sont très limitées.
Dans notre démarche, nous avons d'abord procédé à un examen préliminaire à fin de vérifier la stabilité de eIF6 aux hautes températures. Pour cela, nous avons fait référence au travail de Valenzuela *et al.* qui a montré la cessation de l'activité du eIF6 à des températures au-dela de 60 °C. Une étude par diffusion de lumière, réalisée pour vérifier les conclusions de Valenzuela *et al.*, a montré l'augmentation irréversible du rayon hydrodynamique des molécules au delà de la température de 50°C. Ce résultat semble indiquer pour le eIF6 des conditions non-denaturantes à des températures inférieures à 50°C. Cette limitation a été tenue en compte dans les mesures expérimentales, toutefois, des simulations de dynamique moléculaire ont été quand même réalisées afin de vérifier la présence effective d'un processus de dénaturation ou simplement son amorce.

ECHANTILLONS FINAUX Afin de pouvoir réaliser les expériences de diffusion des neutrons sur aIF6 et eIF6, les deux protéines ont été préparées dans des solutions deuterées avec une concentration d'environ 40 mg/ml et une pD de 7.0. La concentration finale de ces solutions a été vérifiée avec des mesures d'absorption UV-VIS à une longueur d'onde de 280nm. Cette mesure a donné cependant des résultats très imprécis à cause du nombre réduit de chromophores (qui normalement absorbent les UV à 280nm) dans les acides aminés de la séquence des IF6. Le manque de valeur précise a engendré des fortes limitations dans les mesures de diffusion des neutrons.

0.2.2 *Expériences de diffusion des neutrons*

Dans ce travail, les mesures de diffusion des neutrons ont visé les effets de la pression et de la température sur la dynamique des protéines. Pour les expériences de diffusion quasi-elastique de neutrons (QENS), ces mesures sont normalement réalisées sur des échantillons en solution qui sont comprimés dans des porte-échantillons de forme cylindrique. Dans le cas de cette étude, la grande variation de température explorée a nécessité l'utilisation de matériels qui ne subissent pas de changements structuraux importants à hautes températures qui pourraient modifier, par exemple, la résistance mécanique des porte-échantillons. En outre, les expérience de QENS nécessitent des solutions à très haute concentration afin de pouvoir bien distinguer le signal provenant des protéines de celui donné par le solvant.

Ce dernier fait pose un limitation importante dans le volume total disponible pour les expériences sur des molécules biologiques car ces molécules ne sont normalement disponibles qu'en petites quantités. Dans le cas des a/eIF6, cette limitation a constitué un point crucial pour la mise en place de l'instrumentation nécessaire aux mesures sous haute pression. Une description complète de la cellule pour les hautes pressions est donnée dans la thèse.

0.2.3 *Simulation de dynamique moléculaire*

Toutes les étapes de simulation de dynamique moléculaire décrites dans le texte qui suit ont été réalisées avec le programme AMBER9[27]. Le champ de forces utilisé est le AMBER99SB [86], une mise à jour du plus connu champ de forces AMBER94, qui contient des paramètres permettant une meilleure différentiation des éléments de structure secondaire des protéines.
Le schéma général pour la réalisation de ces simulations de dynamique moléculaire peut etre résumé comme suit:

1. La configuration initiale est créée à partir des donnés cristallographiques (codes PDB: 1G61 pour aIF6 et 1G62 pour eIF6). Au système initiale nous avons ajouté des molécules d'eau représentées par le modèle TIP3P et 14 contre-ions de sodium. Ces derniers ont le rôle de rendre nulle la charge totale du système.

2. Le système entier est porté vers un état stable en appliquant des algorithmes de minimisation de l'énergie d'abord sur les seules molécules d'eau et puis sur la totalité du système.

3. Le système est ensuite porté vers les conditions thermodynamiques souhaitées par des courtes simulations de dynamique moléculaire d'équilibration. D'abord le système est équilibré dans un ensemble NVT (volume et température constantes) avec un pas d'intégration de 1fs et pour une durée de 150ps; puis il est équilibré à pression et température constantes (1bar et 300K) pendant 700ps.

4. Une fois obtenue l'équilibration du système, plusieurs simulations dans différentes conditions de pression et de température ont été réalisées. Les cordonnées atomiques sont stockées toutes les 40fs et la longueur des trajectoires générées est de 2ns.

Dans cette démarche la régulation de la température a été réalisée avec un thermostat de Langevin [195] (constante de friction $3.5ps^{-1}$) et la pression moyenne a été maintenue constante par un barostat de Berendsen [14] (temps de relaxation de 1.5ps).

CONFIGURATION INITIALE DU EIF6 La configuration initiale du eIF6 issue des données cristallographiques eu *Saccharomyces cerevisiae* (code PDB 1G62) ne contient pas les positions atomiques des 21 acides aminés qui forment le fragment C-terminale, cette absence étant du aux clivages protéolytiques déjà mentionnés dans les paragraphes précédents. La *construction* de la structure complète du eIF6 exige des étapes supplémentaires par rapport au aIF6.
D'abord la structure du fragment C-terminale (CTAIL, ainsi par la suite) a été modélisée et partiellement repliée. Ensuite la structure du CTAIL a été attachée au reste de la structure du eIF6 et l'ensemble a été équilibré vers une configuration stable. Toutes ces étapes ont été réalisées par des simulations de dynamique moléculaire avec solvant implicite permettant de réduire considérablement les temps de calcul. Cette méthode consiste dans la substitution des molécules d'eau par des termes additionnels, dans le champs de force atomique, qui devraient reproduire les effets du solvant sur la protéine.
Le protocole utilisé pour la modélisation du CTAIL et pour son repliement est basé sur le schéma suivant:

1. Une configuration "linéaire" du CTAIL a été d'abord créée avec la séquence des acides aminés montrée dans le Tableau 1

2. Un repliement initial du CTAIL a été obtenu par une suite de simulations courtes dans un ensemble NVT (volume et température constants) avec un pas d'intégration entre 0.1 et 0.5 fs. Le protocole complet de cette étape est affiché dans le tableau **??** de cette thèse.

3. Le processus de repliement est ensuite obtenu par une simulation de 4ons de longueur.

La structure du CTAIL partiellement repliée est montrée par la Figure 16 (voir le texte de cette thèse). La configuration initiale de la structure complète du eIF6 a été crée par la jonction entre CTAIL et le eIF6. Cette dernière opération a été effectuée par la création d'une liaison covalente entre l'azote N-terminal du CTAIL et le carbone C-terminal du eIF6. La structure ainsi obtenue a été ensuite soumise à une procédure de minimisation et d'équilibration.
Enfin, pour obtenir une meilleure optimisation locale de la structure, nous avons réalisé une simulation moléculaire avec la méthode de *recuit simulé*. Grâce à une succession de simulations à très hautes températures suivies par des autres à 300K, cette méthode permet d'explorer l'espace des configurations des grandes molécules mieux qu'une simple procédure

de minimisation de l'énergie. Dans le cas du eIF6, la structure finale issue de cette méthode (voir Figure 17) a une énergie potentielle légèrement inférieure à celle initiale.
La configuration ainsi trouvée a été utilisée ensuite comme état initial pour la procédure de simulation moléculaire décrite au début de cette section.

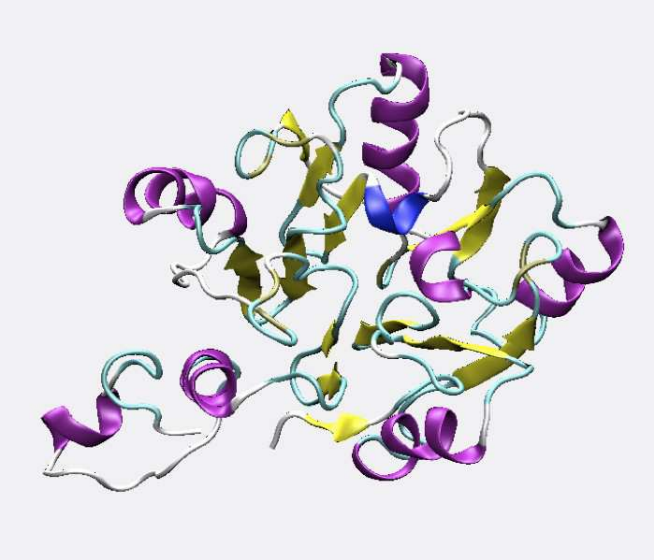

Figure 1: Structure du eIF6 issue de la procédure de *simulated annealing*.

ECHANTILLONS SUPPLÉMENTAIRES Afin de mieux comparer les résultats obtenus à partir des simulations moléculaires avec ceux issues des mesures expérimentales, d'autres échantillons ont été modélisés et simulés. Les résultats issues de ces simulations, ont fourni notamment une meilleure compréhension des effets structuraux et dynamiques du CTAIL et de la queue de poly-histidines (H-Tag) respectivement sur eIF6 et aIF6. Les deux échantillons sont les suivants:

. **eIF6-NoCTAIL**: le eIF6 simulé sans le CTAIL attaché. La structure cristallographique a été utilisé comme configuration initiale et a été soumise à la procédure de simulation moléculaire énoncée au début de cette section.

. **aIF6-HTag**: le aIF6 est simulé avec une queue poly-histidine à l'extrémité N-terminale. Comme déjà expliqué dans ce texte, la présence de ce fragment additionnel a permis une amélioration considérable du rendement des protocoles de production des IF6. Les

expériences de diffusion des neutrons sous pression ont été réalisées sur des échantillons contenant la H-Tag, ce qui a rendu indispensable la réalisation des simulations du même échantillon. La création de la structure, composée par le aIF6 et la H-Tag, a été obtenue avec une procédé similaire à celui utilisé pour le CTAIL.

0.3 Méthode de caractérisation de la structure secondaire des protéines

Dans cette section, nous présentons une nouvelle méthode pour la caractérisation de la structure secondaire des protéines. Le développement de cette méthode, appelée ScrewFit, a été inspiré par la nécessité d'une analyse fine des effets de l'environnement sur les structures des protéines.
Nous avons ensuite trouvé que ScrewFit est capable aussi de détecter les motifs qui caractérisent la structure secondaire et de donner une évaluation des effets locaux et globaux résultants des interactions avec un ligand. La méthode et ses applications sont présentées dans cette thèse par le biais de deux articles, déjà parus ou en cours de publication dans des revues internationales avec comité d'évaluation :

- Kneller, G.R. and Calligari, P. Efficient characterization of protein secondary structure in terms of screw motions. *Acta Crystallographica D*, **62**(3), 302-311(2006).

- Calligari, P. and Kneller G.R., ScrewFit: a novel approach for continuum protein secondary structure assessments. *soumis.*

Une autre application de cette méthode, en dehors du sujet de cette thèse, est exposée ici :

- Calligari, P. et al., Inhibition of viral group-1 and group-2 neuraminidases by oseltamivir: a comparative structural analysis by the ScrewFit algorithm. *Biophysical Chemistry*, accepté pour publication (2008).

Les résumés des deux articles inclus dans la thèse sont ici traduits et une brève introduction de la méthode est donnée avec un exemple d'application.

0.3.1 *Efficient characterisation of protein secondary structure in terms of screw motions*

Nous présentons une méthode simple et efficace pour décrire la structure secondaire en termes de distances d'orientation entre plans peptidiques consécutifs et paramètres

hélicoïdaux locaux. La méthode utilise des fits de superposition de plans peptidiques en combinaison avec le théorème de Chasles qui affirme que tout mouvement de corps rigide peut être décrit par un mouvement du type "mouvement de vis". Pour la superposition de plans peptidiques consécutifs nous avons dérivé les paramètres hélicoïdaux et utilisé le fit le plus mauvais pour définir une distance d'orientation. Nous avons aussi montré les applications de la méthode aux modèles théoriques des structures secondaires, aux protéines appartenant à différentes classes structurales et à la description des changements structuraux induits dans le lysozyme par une haute pression hydrostatique. Dans cette dernière application, nous avons utilisé des donnés déjà publiées, issues de la cristallographie par rayons-X et des mesures par RMN.

0.3.2 *ScrewFit: a novel approach for continuum protein secondary structure assessments*

Nous présentons ici une nouvelle méthode pour la détection des éléments de la structure secondaire des protéines combinant une description de la chaîne principale d'une protéine en terme de "mouvement de vis" (Acta Cryst. 62, 302-311 (2006)) et une approche statistique. L'application de cette méthode produit des intervalles de confiance qui définissant les variations naturelles des paramètres hélicoïdaux qui décrivent la chaîne principale. Afin d'établir ces intervalles pour chaque motif (pattern) de structure secondaire, nous avons analysé plusieurs bases de données contenant des structure de protéines caractérisées par un profil structural bien défini. Cette méthode permet une évaluation "continue" de la structure secondaire d'une protéine et a été démontrée stable par rapport aux variations structurales, trouvées dans les donnés RMN, et à la résolution expérimentale des donnés cristallographiques. La comparaison avec d'autres méthodes confirme sa précision et fiabilité. L'exemple de l'analyse de la structure de l'inhibiteur trypsine pancréatique bovine dans ses trois différentes formes cristallines montre la capacité de notre méthode de détecter et d'analyser des petites variations structurales dans des données expérimentales très bruitées.

0.3.3 *L'algorithme* ScrewFit

A fin de décrire d'une manière simple la structure secondaire des protéines, nous avons utilisé l'algorithme *ScrewFit*, qui est basé sur la superposition optimale de structures moléculaires

[50, 115, 116] et sur le théorème de Chasles [33, 32].

Nous pouvons considérer deux plans peptidiques consécutifs A et B, comme deux corps rigides définis par les positions des atomes $\{O, C, N\}$ de chaque plan. Ces corps rigides peuvent être superposés en minimisant la fonction suivante:

$$m(\mathbf{q}) = \sum_{\alpha=1}^{3} (\mathbf{D} \cdot \mathbf{x}_\alpha - \mathbf{x}'_\alpha)^2, \tag{0.1}$$

où $\{\mathbf{x}_\alpha\}$ et $\{\mathbf{x}'_\alpha\}$ sont respectivement les positions atomiques de la structure de référence (les atomes $\{O, C, N\}$ dans le plan A) et celles de la structure cible (plan B). Le symbole $\mathbf{D}$ dénote une matrice orthogonale qui décrit une rotation. Les deux ensembles des coordonnées sont définis par rapport à un point de référence qui est la position de l'atome de carbone de chaque plan $\{O, C, N\}$. En utilisant le fait qu'une matrice de rotation peut être exprimée par les composantes d'un quaternion normalisé $\mathbf{q} \equiv \{q_0, q_1, q_2, q_3\}$, où $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ [4],

$$\mathbf{D}(q) = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(-q_0 q_3 + q_1 q_2) & 2(q_0 q_2 + q_1 q_3) \\ 2(q_0 q_3 + q_1 q_2) & q_0^2 + q_2^2 - q_1^2 - q_3^2 & 2(-q_0 q_1 + q_2 q_3) \\ 2(-q_0 q_2 + q_1 q_3) & 2(q_0 q_1 + q_2 q_3) & q_0^2 + q_3^2 - q_1^2 - q_2^2 \end{pmatrix}, \tag{0.2}$$

la fonction (6.1) peut être minimisée par rapport à ces quatre composantes. Comme il a été montré dans d'autres études [50, 115, 116], la minimisation avec contraintes peut être transformée dans un problème de valeurs propres d'une matrice $\mathbf{M}$ semi-definie positive $\mathbf{M} \equiv \mathbf{M}(\{\mathbf{x}_\alpha, \mathbf{x}'_\alpha\})$,

$$\mathbf{M} \cdot \mathbf{q} = \lambda \mathbf{q}, \tag{0.3}$$

où les valeurs propres $\lambda_j = m(\mathbf{q}_j)$ sont les possibles erreurs dans la superposition des deux plans peptidiques défini par (6.1). Le quaternion correspondant à la valeur propre la plus petite est la solution pour une superposition optimale et ses composantes décrivent l'orientation relative de $\{\mathbf{x}'_\alpha\}$ par rapport à $\{\mathbf{x}_\alpha\}$. La relation:

$$\mathbf{q} \equiv \begin{pmatrix} q_0 \\ \mathbf{q}_v \end{pmatrix} = \begin{pmatrix} \cos(\phi/2) \\ \sin(\phi/2)\mathbf{n} \end{pmatrix}. \tag{0.4}$$

nous montre que le quaternion correspondant à la valeur propre la plus petite définit aussi un angle de rotation ϕ et un axe de rotation $\mathbf{n}$, ce dernier étant aussi la direction du "mouvement de vis" décrit dans le théorème de Chasles. La preuve de ce dernier théorème peut être trouvée dans la référence [114]. La valeur propre la plus grande λ_{max} décrit la "pire" superposition

possible entre les deux plans peptidiques et donne leur distance euclidienne maximale. Nous avons utilisé ce dernier fait pour définir une *distance d'orientation* unique par

$$\Delta = \sqrt{\frac{\sum_{\alpha=1}^{3}(\mathbf{x}_\alpha - \mathbf{x}'_\alpha)^2}{\lambda_{max}}}. \tag{0.5}$$

Par définition $0 \leqslant \Delta \leqslant 1$.

La caractérisation de la structure secondaire d'une protéine est réalisée dans cette méthode avec les paramètres suivants:

1. La distance d'orientation entre deux plans consécutifs, défini par Eq. (6.5).

2. Le rayon de la surface cylindrique sur laquelle bouge l'atome de référence (atome C) en réalisant le "mouvement de vis" entre deux plans peptidique comme décrit par le théorème de Chasles,

   $$\rho = \frac{|\mathbf{t}_\perp|}{2}\sqrt{1 + \cot^2(\phi/2)}. \tag{0.6}$$

   Ici $\mathbf{t}_\perp$ est la composante perpendiculaire à l'axe de rotation $\mathbf{n}$ du vecteur $\mathbf{t}$ qui relie les atomes C.

3. Le paramètre de "rectitude" σ. Pour chaque résidu i ce dernier est défini comme suit:

   $$\sigma_i = \boldsymbol{\mu}_i^T \cdot \boldsymbol{\mu}_{i+1}, \tag{0.7}$$

   où

   $$\boldsymbol{\mu}_i = \frac{\mathbf{R}_{i+1}^\perp - \mathbf{R}_i^\perp}{|\mathbf{R}_{i+1}^\perp - \mathbf{R}_i^\perp|} \tag{0.8}$$

   et $\mathbf{R}_i^\perp$ est le point de l'axe hélicoïdal le plus proche de l'atome C du plan peptidique i. La "rectitude" donne des renseignements sur la courbure des éléments de la structure secondaire.

0.3.4 *Application de ScrewFit à des structures modèles*

Nous présentons ici un exemple d'application simple de la méthode ScrewFit. Il s'agit de structures-modèles différemment configurées et composées de 10 alanines et qui ont été obtenues par le Image Library of Biological Macromolecules in Jena [2] (les valeurs obtenues sont affichés dans le tableau 2).

Les résultats remarquables de cette application sont:

2 Institute-of-Molecular-Biotechnology-Jena:http://www.imb-jena.de/IMAGE.html

Table 2: Paramètres hélicoïdaux de différentes structures. Ici ρ_{C_α} est la valeur de ρ obtenue en utilisant la position des atomes C_α comme référence. Les paramètres τ, *pitch* et h sont illustrés de manière approfondie dans la thèse.

Motive	ρ [nm]	ρ_{C_α} [nm]	τ	pitch	h	σ	Δ
α-helix (R)	0.171	0.227	3.62	0.556	+	1	0.582
α-helix (L)	0.171	0.227	3.62	0.556	−	1	0.582
3-10 helix	0.146	0.203	3.28	0.589	+	1	0.670
π-helix	0.178	0.258	4.16	0.558	+	1	0.471
β-strand	0.055	0.093	2.03	0.671	−	1	0.875
extended	0.037	0.055	2.00	0.725	−		0.754

1. les hélices-α lévogyre et dextrogyre ont exactement les mêmes valeurs, ce qui montre l'efficacité mathématique de la méthode.

2. Chaque type d'hélice peut être distingué des autres même si, en général, les valeurs des paramètres pour les hélices sont très similaires.

3. La méthode permet une distinction très nette entre hélices et brins-β.

0.4 RÉSULTATS

Nous présentons ici les principaux résultats obtenus dans ce travail de thèse soit par le moyen de simulations de dynamique moléculaire (MD) soit par les expériences de diffusion quasi-élastique des neutrons (QENS). Ces résultats seront utilisés pour déterminer les effets locaux et globaux des conditions environnementales sur les deux homologues du IF6. Le lien entre ces deux types d'effets est assuré par la complémentarité des informations obtenues par dynamique moléculaire et par diffusion des neutrons.
Les conditions d'environnement appliqués dans les deux cas sont affichées dans les Tableaux 4 et 3.

Le nombre et la variété des mesures expérimentales ont été largement limités par des problèmes techniques qui seront discutés plus avant dans le texte. La comparaison entre données QENS et MD a été quand même possible dans un nombre restreint de cas. Pour la raison énoncée, les données affichées dans cette section sont issues des simulations MD sauf

Table 3: Ensemble des configurations environnementales appliquées dans les simulation MD

aIF6	eIF6	eIF6-NoCTAIL
300K - 1bar	300K - 1bar	300K - 1bar
300K - 250bar	320K - 1bar	300K 500bar
300K - 500bar	350K - 1bar	350K - 1bar
350K - 1bar	350K - 500bar	350K - 500bar
350K - 250bar		
350K - 500bar		

Table 4: Ensemble des configurations environnementales appliquées dans les mesures QENS

aIF6	eIF6	aIF6-HTag
300K - 1bar	300K - 1bar	300K - 250bar
350K - 1bar	350K - 1bar	300K - 500bar
		350K - 250bar
		350K - 500bar

exceptions signalés.

0.4.1 *Effets de la pression et de la température sur la structure des IF6*

Les principaux effets de la température et de la pression sur les deux homologues du IF6 ont été caractérisés autant d'un point de vue global que local.

Dans le premier cas, nous avons observé les changements du volume moléculaire des deux protéines ainsi que les changements de leurs rayons de gyration et de leurs surfaces accessibles au solvant. Dans le deuxième cas, une caractérisation fine des changements de la structure secondaire a été obtenue par l'étude des fluctuations des atomes de carbone α et l'analyse effectuée avec la méthode ScrewFit.
Les principaux effets globaux trouvés sont ici résumés:

- Les changements du volume moléculaire et du rayon de gyration montrent que le aIF6 est moins sensible au changements environnementaux que son homologue mesophile. Figures 2 et 3

- Le rayon de gyration et la surface accessible au solvant du aIF6 affichent, aux conditions *naturelles* pour Methanococcus Jannaschii (350K -500bar), des valeurs très similaires à celles du eIF6 dans des conditions *normales*. Ce résultat semble suggérer la présence d' "états correspondants" entre les conditions naturelles des deux homologues.

Les fluctuations des chaînes principales et les variations des paramètres ScrewFit montrent les résultats suivants:

aIF6

- A une haute température les régions autour du résidu 60 et comprises entre les résidus 120 et 130 subissent des distorsions indiquant des changements d'orientation des plans peptidiques de la chaîne principale qui cependant n'entraînent pas de courbures de cette dernière. Nous avons observé que ces changements ne se produisent pas aux hautes pressions.

· La haute pression induit une courbure significative sur la structure principale du aIF6 dans la région comprise entre les résidus 90 et 95. Cet effet engendre aussi une augmentation des fluctuations atomiques dans la même région.

· La combinaison de haute température et haute pression modifie les configurations des plans peptidiques autour du résidu 50 vers des orientations relatives plutôt hélicoïdales.

eIF6

· Des importantes variations dans la région 220-245 sont engendrées par les fluctuations du CTAIL. Comme on pouvait l'attendre, ces variations augmentent en fonction de la température.

· Différentes variations dans la région 170-180 se produisent tant à haute température qu'à haute pression mais elles disparaissent quand la pression et la température sont appliquées de façon simultanée.

· La combinaison d'une haute pression (500bar) et d'une haute température (350K) produit une courbure de la partie C-terminale de la longue hélice-α dans la région 35-45.

Certains des ces effets locaux sur la structure ont été ensuite reliés aux possibles effets sur la fonction du IF6. En particulier, nous avons remarqué que la pression a des effets très importants sur la configuration des sérines 174 et 175 dans le eIF6. Ces deux derniers acides aminés ont un rôle essentiel dans la localisation du IF6 dans le noyau cellulaire parce que leur phosphorylation permet le passage du IF6 entre noyau et cytoplasme.
La présence des états correspondants entre les deux protéines dans leurs conditions naturelles est ultérieurement confirmée par l'étude du facteur de structure élastique incohérent (EISF, voir texte de la thèse) qui donne accès aux fluctuations atomiques en fonction de l'échelle de longueur explorée par les mouvements internes des protéines. La figure 4 montre les fluctuations carrées moyennes en fonction du transfert de quantité de mouvement, q.

Pour mieux comprendre les effets de la présence du CTAIL dans le eIF6 et celle possible de la H-Tag dans le aIF6, nous avons réalisé les mêmes études structurales sur eIF6-NoCTAIL et sur aIF6-HTag. Les résultats les plus remarquables de ces études ont été que le eIF6-NoCTAIL affiche des changements très similaires à ceux qui ont caractérisé le aIF6 et que l'aIF6-HTag montre des propriétés similaire a celles du eIF6. Ces observations suggèrent que la présence d'une "queue" additionnelle dans les deux structures des IF6 joue un rôle crucial dans la

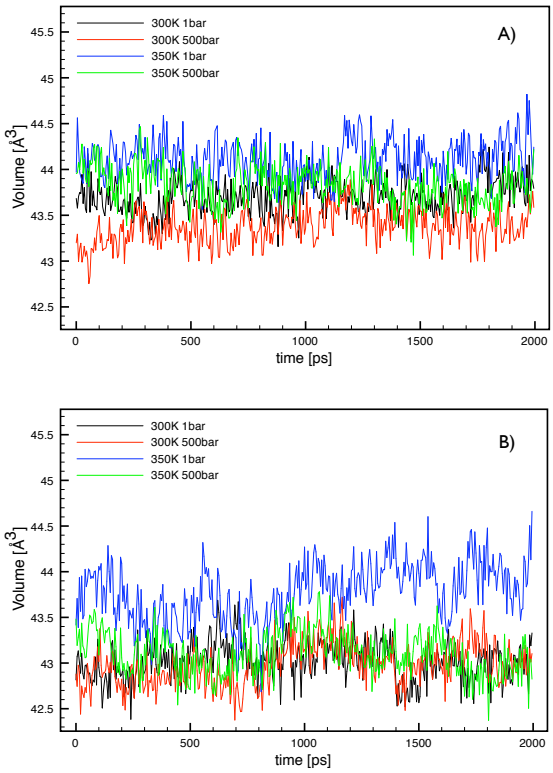

Figure 2: Variation du volume moléculaire en fonction de la pression et de la température. Panel A: aIF6. Panel B: eIF6.

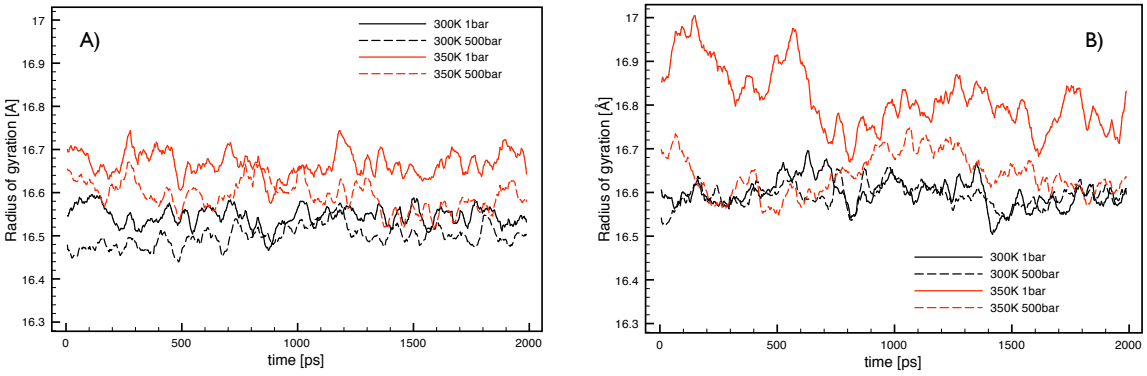

Figure 3: Variation du rayon de gyration en fonction de la pression et de la température. Panel A: aIF6. Panel B: eIF6 (Seuls les résidus entre 1 et 224 ont été pris en compte).

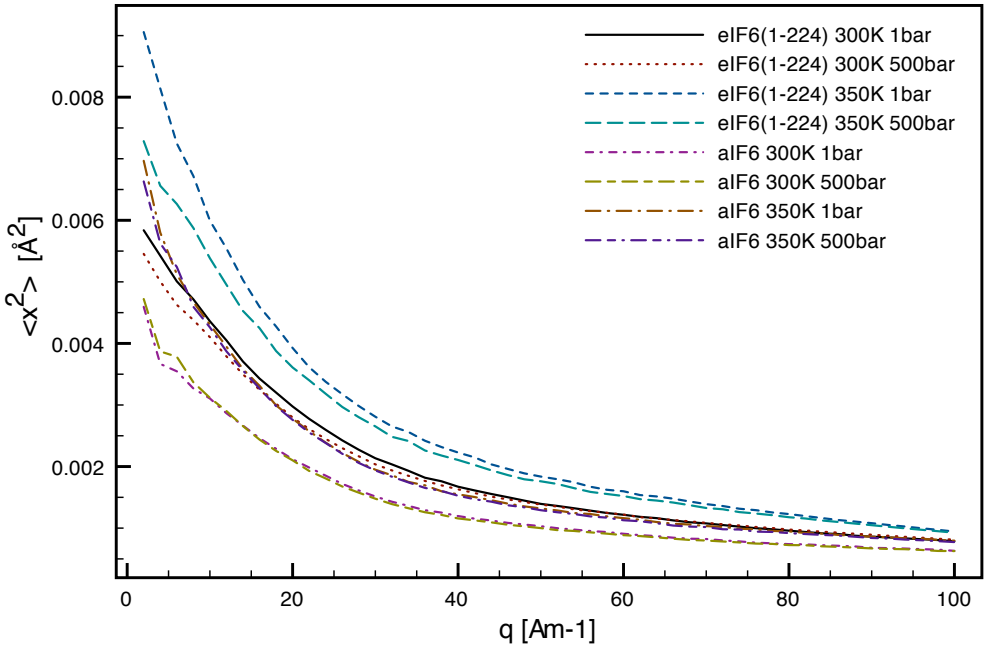

Figure 4: Comparaison des fluctuations atomiques du eIF6 et du aIF6.

détermination des propriétés structurales de chaque homologue.

0.4.2 *Effets dynamiques de la pression et de la température*

Afin de caractériser les effets dynamiques d'un environnement extrême sur les IF6, les donnés issues des simulations MD et celles obtenues par mesures QENS ont été comparées. Les quantités "naturelles" pour l'analyse des simulations MD sont celles reliées à des fonctions dépendant du temps, tels que le déplacement carré moyen, ou bien aux fonctions des corrélations comme la fonction intermédiaire de diffusion. Les mesures expérimentales sont au contraire étudiées par l'analyse du facteur de structure dynamique qui est une quantité reliée au transfert de la quantité de mouvement et aux variations d'énergie. La fonction intermédiaire de diffusion et le facteur de structure dynamique d'une protéine peuvent être bien représentés par le modèle de dynamique Brownienne fractionnaire (fOU, dans la suite) [119, 20]. Ce modèle donne des temps de relaxation qui définissent la dynamique interne des protéines en fonction d'une échelle de longueur.

Les résultats sur les effets dynamiques de la pression et de la température peuvent être résumés comme suit:

Les temps de relaxation définis par le fOU affichent des valeurs très différentes dans les deux protéines. En particulier, dans le cas du aIF6, ils sont systématiquement plus rapides que ceux du eIF6 à toutes les échelles de longueur.

Les mêmes temps caractéristiques des deux protéines changent d'une façon différente en fonction de la pression et de la température. En particulier ils augmentent en fonction de la pression dans le cas du aIF6 et ils baissent dans le cas du eIF6.

Ces dernières observations identifient plutôt les caractères dynamiques de chaque homologue du IF6 et de sa réponse aux changements des conditions environnementales. Cette dernière réponse devient par contre évidente si l'on compare les temps de relaxation de chaque homologue dans ses conditions "naturelles" avec ceux affichés dans des condition défavorables. Les Figures 5 et 6 montrent que dans les deux homologues les conditions les plus défavorables produisent des changements non homogènes dans le temps de relaxation et que ces changements deviennent de plus en plus homogènes en se rapprochant des conditions "naturelles".

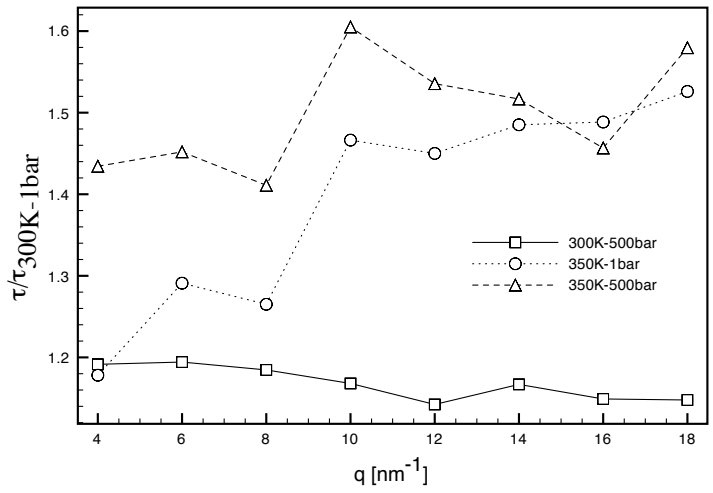

Figure 5: Valeurs du paramètre τ issu du fit fOU de eIF6.

Ces résultats principalement issus des simulations MD ont été aussi confirmés par les expériences de diffusion de neutrons. Ces dernières ont été toutefois limitées par les difficultés techniques suivantes:

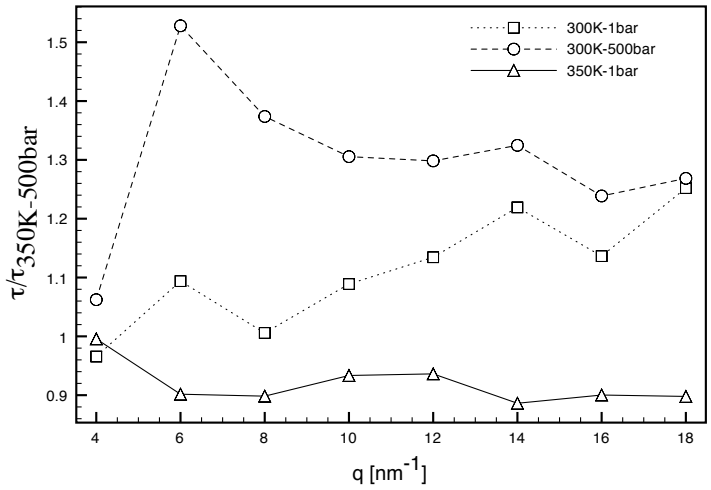

Figure 6: Valeurs du paramètre τ issu du fit fOU de aIF6.

· Le eIF6 a montré une stabilité précaire dans les configurations expérimentales utilisée dans ce travail

· L'évaluation de la concentration des échantillons en solution a été très difficile à cause de l'absorption très limitée dans la gamme des UV-VIS tant du aIF6 que du eIF6.

Ces deux problèmes n'ont pas rendu possible une analyse plus fine des données expérimentales. Celles-ci néanmoins ont pu être utilisées pour supporter les résultats obtenus par dynamique moléculaire.

0.5 CONCLUSIONS

Dans ce travail nous avons caractérisé la réponse des deux homologues du IF6 à différentes conditions extrêmes. Ces réponses ont été caractérisées d'un point de vue autant structural que dynamique.

L'étude structurale a montré que l'aIF6, bien que moins sensible aux changements de température et de pression, assume dans ces conditions "naturelles" (350K -500bar) des caractéristiques structurales très similaires à celle du eIF6 à température et pression ambiante.

Nous avons aussi distingués les effets globaux des effets locaux, qui sont très différents dans les deux IF6. En outre, contrairement aux attentes, les effets des deux variables thermodynamiques ne s'opposent pas les uns aux autres.

D'un point de vue dynamique, le modèle de dynamique Brownienne fractionnaire nous a permis d'associer des temps caractéristiques à chaque protéine. Ces temps, qui

dépendent de l'échelle de longueur considérée, affichent des valeurs différentes dans les deux protéines dans des conditions environnementales diverses alors qu'ils changent d'une façon similaire si les protéines sont dans des conditions défavorables. Ces résultats nous ont permis d'identifier d'abord la présence des "états correspondants" des fluctuations structurales de chaque homologue du IF6 dans ses conditions naturelles. Deuxièmement ils ont rendu possible la distinction entre les propriétés dynamiques reliées à la structure de chaque protéine et les propriétés caractérisant l'état naturel de chaque protéine.

Part II

THESIS

CONTENTS

1

GENERAL INTRODUCTION

The last century owes the majority of its advances in science and technology to the investigation of the structure and dynamics of matter at the atomic and subatomic scale which lead, among others, to a significantly better understanding of the physical and chemical properties of solids and liquids.
Since the beginning of the century, the knowledge acquired about the inanimate matter lead physicists to question how this insight could also be applied also to biological matter. To this end, most of the principles of thermodynamics and classical mechanics were applied to biological systems to gain new insights into their properties, but only the advent of quantum theory and statistical mechanics finally allowed them to be investigated at the atomic scale.
These new theories, along with important developments in the techniques of atomic and molecular spectroscopy lead researchers to revise and interpret from a molecular point of view several results obtained in biology in the last century. [1]

Living matter presented to researchers a degree of heterogeneity and complexity that they had never encountered before in physics and stimulated the birth of a new interdisciplinary field of research which brought together physics, chemistry and biology and which revealed itself as a source for a large number of key results with further applications in medicine and biotechnology.
The work presented in this thesis is based on the enormous progress made in the field of *biological physics*, in particular with respect to the development of computer simulation methods, made possible by the rapid improvements in computers, and the application of spectroscopic techniques to biological samples.
As will be explained more deeply, later in the text, these two techniques can be combined in a *unified* approach in which computer simulations helps to better characterize the experimental results at the atomic scale.

1 As an example, it is worth to cite the famous prediction by E.Schroedinger who inferred from the basic quantum mechanical principles, that the structure of DNA had to be that of an *aperiodic lattice*. His prediction was then confirmed years later by the crystallographic resolution of the DNA structure made by Watson and Crick.

In the following, a general introduction to this thesis work will be given assuming a basic knowledge of protein structure elements the explanation of which would be beyond of the scope of this thesis. This subject is however treated in detail in a number of excellent textbooks [42, 187]

1.1 EFFECTS OF PRESSURE AND TEMPERATURE ON PROTEINS

In the very first efforts to investigate living matter, physicists focused on the application of basic principles of mechanics and thermodynamics to interpret the complexity of biological macromolecules [30].

As this thesis deals essentially with the effects of temperature and pressure on proteins, the following sections will concentrate on this subject, giving a short presentation of the known effects of these two thermodynamic variables on the structure and dynamics of proteins.

Pioneering work on the impact of pressure on proteins was performed by Regnard in 1891 [160] and by Bridgman in 1914 [17]. The former showed that pressure below 1kbar affects reversibly the catalytic functions in bacteria, whereas the second proved that pressures above 8kbar lead to a coagulation of the protein ovalbumin. Moreover, Bridgman also noted that high temperature reduced the effects of high pressure on albumin, i.e. reduced the coagulation.

These studies yielded the important result that two types of pressure effects have to be distinguished: irreversible and reversible effects depending on the magnitude of the applied pressure [75]. The same observations have been made concerning temperature induced changes [42, 189].

When irreversible effects produced by high pressure (as well as by high temperature) are present, a large change in the spatial conformation affects proteins which loose their native structure. In these new *denaturated* states, proteins cannot perform their functions and tend to have unusual interactions with other proteins leading to large aggregates. Partially denaturated states (also known as *molten globules*) can be included among the reversible effects because when pressure or temperature are brought back to their original values proteins fold back in their native configuration [42].

Several more recent works [143, 219, 169, 99], showed also that : *i)* proteins do not all have the same characteristic threshold for *denaturing pressures*: it depends of their size and shape; *ii)* denaturing caused by high temperature is often very different from denaturing caused by

freezing or by high pressure.

In summary, these results, which will be discussed in the following, show that high temperature and high pressure have similar effects on protein stability but they appear to be competitive when applied together. Moreover, the structural changes of proteins due to external pressure are much less homogenous than those induced by temperature changes.

LE CHATELIER'S PRINCIPLE A useful theoretical framework for the comprehension of how biomolecules respond to changes in thermodynamic conditions is given by the criteria for the stability of thermodynamic systems and by Le Chatelier's principle[126]. The latter states that at equilibrium, a system tends to minimize the effect of any external factor by which it is perturbed. As briefly shown in the following, Le Chatelier's principle gives the physical interpretation of the stability criteria for the thermodynamical potentials which describe the macroscopic states of the system. Indeed these criteria reflect those demanded for the intrinsic stability of the thermodynamic system itself. For an equilibrium process connecting two states A and B, the general expression for the variation of Gibbs free energy, G(T,P,N) reads as follows:

$$\Delta G = V\Delta P - S\Delta T + \sum_i \mu_i \, \Delta N_i \tag{1.1}$$

where ΔN_i, ΔP and ΔT are, respectively, the variation in number of particles of type i and the change in pressure and temperature. In the case of the Gibbs free energy, the stability criteria lead to the condition that the function G(T,P,N) be *concave* with respect to variables T and P and *convex* with respect to N_i [21]. From a mathematical point of view, this means that the second order derivatives of the function G(T,P,N) with respect to T and P must be negative. These conditions can be rewritten with respect to the natural thermodynamic variables by means of the Maxwell relations:

$$\left(\frac{\partial^2 G}{\partial P^2}\right)_T = \left(\frac{\partial V}{\partial P}\right)_T \leqslant 0 \tag{1.2}$$

$$\left(\frac{\partial^2 G}{\partial T^2}\right)_P = -\left(\frac{\partial S}{\partial T}\right)_P \leqslant 0 \tag{1.3}$$

As a consequence, an increase in pressure on a system favors a reduction of its conjugated extensive variable, the volume. Hence, in general, pressure favors all processes that are

accompained by negative volume changes.
In the case of proteins, which usually have compact and tightly packed structures, three different types of volumes can be distinguished: the volume of atoms, the void volume due to imperfect packing of atoms which leave small cavities at the interior of molecular structures and the volume variations which result from the degree of solvation of peptide bonds and amino acids residues. All these components influence both the atomic fluctuations and the *weak* intramolecular interactions (i.e. those not involving covalent bonds) which govern the transition between its various conformations.
Besides giving a general framework for the understanding of the effects of pressure on proteins, the Le Chatelier principle also introduces a first explanation of the antagonistic relation between temperature and pressure. The principle states in fact that perturbations on a system directly induce processes that attenuate the perturbations themselves but its response also induces indirectly other processes that reduce the initial perturbation. A phenomenological demonstration of this extension of principle, which is usually called *LeChatelier-Braun principle*, can be found in references [21, 161].

From a molecular point of view, the antagonism between the two variables can be interpreted by the microscopic ordering principle, which states that an increase in pressure at constant temperature leads to a decrease in the entropy of the system, which corresponds to an increase in the ordering of molecules in the system itself.
This is a general property of molecular systems but it has also a strict sense for protein stability as proven by Brandts *et al.* [16] and Hawley [72] who showed that the pressure-temperature phase diagram which determines the transition between functional and denaturated states of proteins has elliptic boundaries, meaning that there exist several combinations of the P,T variables which are compatible with the stability criteria. This peculiar behavior has been ascribed to the exposure of hydrophobic groups to the solvent when the protein folds. It is worth to note that this behavior is found neither in nucleic acids [73] nor in lipids [47]. The elliptic phase boundaries, which are also found in liquid crystals, seems to be produced by a fine balance between aromatic and aliphatic parts of molecules.

EFFECTS OF PRESSURE ON INTERMOLECULAR INTERACTIONS The stability of biomolecules results from a cooperation of interactions involving its components and the external environment [93]. In the case of proteins, once the primary structure is formed

by the peptide bonds, the interactions that play a significant role in this cooperation are essentially weak [94].

The electronic structure of atoms and the covalent bonds are directly modified by pressure only for values greater than 30kbar [51]and 15kbar [67, 144], respectively, but their compression is negligible[62]. Their breaking is responsible for big changes in free volume and leads to a totally irreversible protein denaturation. The formation of covalent bonds has a $\Delta V = -10\mathrm{ml.mol}^{-1}$, whereas ΔV values for the changes in bonds or bond angle changes are nearly zero [200]. Besides the peptide bonds that form the primary structure of proteins, the only type of covalent bond which plays a significant role in protein stability is the interchain disulphide bonds formed by non-contiguous cysteine residues [43].

In contrast to covalent bonds, weak non-bonding interactions give a large contribution to the observed ΔV due to pressure changes.

Ionic interactions Ion pairs in proteins cause attractive, short-range interactions that occur between negatively and positively-charged amino-acid side chains over a distance of 4 Å[42]. The role of ion pairs in proteins is to stabilize the tertiary [76] and quaternary structure. For example, interprotein salt bridges provide a mechanism for solvent exclusion from the interfacial domain of the cytochrome b5-cytochrome c complex, in addition to maintaining the stability and specificity of the complex formed [165].
When an ion is formed in solution, the nearby water dipoles are compressed by the Coulombic field of the ion; this phenomenon, which is accompained by a volume decrease, is usually called *electrostriction*. Because of this volume contraction, pressure is expected to favor the disruption of ion-pair but Michels et al. [139] have showed that this should be eliminated by high temperature due to the disruption of the highly ordered structure of electrostricted water.
The solvation of ions in water produces a $\Delta V = -10\mathrm{ml} \cdot \mathrm{mol}^{-1}$ whereas the dissociation of a neutral molecule into ions induces a volume contraction of about $-20\mathrm{ml} \cdot \mathrm{mol}^{-1}$. These facts confirm that electrostriction is favored by pressure but does not give insights into the effect on protein stability. Indeed, this phenomenon can have opposite results depending on the structural effects it produces locally: it can induce denaturation like in the case of chymotrypsin where pressure produces the dissociation of a salt-bridge in the active site region bringing the protein reversibly to an inactive state[77]; on the other had, it can also engender an increase in protein stability if there are weak intersubunit ion-pair interactions

not exposed to solvent, which may be strengthened by pressure.

Hydrogen bonds Studies on simple model molecules showed that hydrogen bonds are stabilized by high pressures up to 4kbar [105, 200]. This results from the smaller inter-atomic distances in the hydrogen-bonded atoms. The stabilizing effect of pressure on hydrogen bonding in proteins was first detected by the pressure dependence of the infrared spectra of the α-helix in myoglobin[127] and from a comparison of the effect on the intermolecular interactions in hydrogen-bonded versus non-hydrogen-bonded amides[65]. In particular hydrogen bonds play a fundamental role in the formation of the secondary structure due to their role in creating the α-helix and β-strand motifs.
It is worth to note also that pressure may promote intermolecular hydrogen bonds at the expense of intramolecular hydrogen bonds, causing increased conformational fluctuations [38]. However, a very small if not negligible ΔV value is observed for processes in which there is an exchange between the existing hydrogen bonds [200, 24].

Hydrophobic interactions In general hydrophobic interactions describe the tendencies of hydrocarbons to form aggregates in the presence of aqueous media. In proteins, they direct non-polar side chains to the interior of proteins and are a major driving force for proper folding [110]. In 1959, Kauzmann [109] suggested that the volume changes for the association of hydrophobic molecules in water should be positive. His hypothesis was based on the observations made on the large entropic effects of the transfer of hydrocarbons from a nonpolar solvent to water. Weber et al [207], on the other hand, showed that the association of aromatic molecules is enhanced by pressure and later Van Eldik et al.[200] confirmed that interactions that pack aromatic rings parallel to each other engender negative volume changes. These results give account of two different hydrophobic effects, one related to the solubility of apolar molecules into water and another one which concerns the solvent-induced interaction between apolar species.
The former class of hydrophobic effects was initially considered as a good model (called oil-drop model)[7] for protein folding as it was confirmed by works on thermal denaturation but could not be confirmed by the pressure effects on protein unfolding because the volume change upon unfolding is positive at low pressures but negative at pressures of about 1-2 kbar. The transfer of hydrocarbons into water shows exactly the opposite behavior, with ΔV being negative at low pressures and positive at high pressures[110].

The major role in pressure effects in protein denaturation is commonly attributed to the hydration of hydrophobic cores. Indeed, the presence of solvent around the proteins gives a significant importance to the cavities created by the imperfect packing of protein structure. Solvents, and in particular water, can often enter these void volume areas and come into contact with inner regions, destabilizing the hydrophilic/hydrophobic equilibrium which maintained the protein's stability. Due to its higher compressibility compared to proteins, water under pressure can also assume local ordering that can easily cause an important increase in the density of hydrogen bond acceptors/donors, which in turn enhance the hydrophilic properties of those amino acids that are able to produce hydrogen bonds [213]. Although several works confirmed this essential role of hydrophobic interactions, discrepancies in experimental [74, 139, 138] and theoretical [87] results on how their role in protein folding/unfolding is accomplished, have not yet allowed the acceptance of the 'oil-drop' model as a common framework for the interpretation of the different physico-chemical contributions to the related volume variations [168].

It is worth to note here the work by Hummer et al. [87] which proposed a different role of hydrophobic interactions in temperature and pressure-induced denaturations. They suggested that pressure denaturation corresponds to the infiltration of water into the protein, whereas heat denaturation corresponds to the transfer of nonpolar side-chain groups into water. As a consequence, pressure denaturation processes lead to (partially or totally) unfolded protein structures with reduced compactness which are, however, considerably more ordered than those in heat-denatured proteins, as probed by NMR experiments of hydrogen exchange [218].

In the context of hydrophobic interactions, it is well established that a significant contribution is made by the van der Waals (VDW) forces [153, 187, 42]. The effects of pressure on VDW forces present the same contradictory results as hydrophobic interactions in general, which does not permit a simple framework to be defined for their physical interpretation. From a general point of view, VDW forces are likely to be enhanced by pressure since they tend to maximize the packing density of proteins and thus reduce the volume. Indeed, packing density is maximized by VDW interactions because the dipole moment attractions they produce improve the packing of the hydrophobic core [67, 95]. On the other hand, VDW forces can also contribute to protein destabilization under pressure. In particular for oligomeric proteins, pressure below 3 kbar promotes the replacement of some weak non-covalent interactions between amino acid residues with amino acid-water interactions causing changes in peptide

chain conformations and resulting in dissociation [180]. These protein-water interactions are promoted by pressure because they produce stronger and shorter bonds that decrease the overall volume [144, 180].

Compressibility The atomic density in proteins is similar to that in solids and can even exceed the latter locally [87] which makes proteins very insensitive to compression. Most of the data available on protein compressibility are obtained by sound velocimetry measurements and the volume change usually observed is around 1% of the protein's total volume[60]. Such a small volumetric variation is induced only by the weak interactions we presented earlier in this section and does not receive any contribution from covalent bonds in the protein backbone [206]. The major effect is due to the hydrophobic interactions: the higher the hydrophobic content is in a protein, the more the latter is compressible [180].
A significant role in compressibility is played by the effects of pressure on the hydration. In particular, as stated in the previous paragraph, the infiltration of water molecules into void cavities produces a significant change in the local atomic fluctuations which results in the destabilization of the hydrophilic/hydrophobic equilibrium of the protein. In general the cavities filled by the solvent can act as *two-ways-valves* or *one-way-valves*, depending on the local molecular environment in the protein [133], i.e. they can form volumes that solvent molecules can continuously enter or leave or they can instead incorporate water molecules, separating them from the bulk solvent. These two type of behavior engender differences in compressibility of cavities and in the whole protein structure [39].
In contrast to temperature, pressure does not seem to have a uniform effect all over the protein [61]. A molecular simulation study by Paci and Marchi [148] proved that short distances are rather less compressed than longer one. This was also confirmed by x-ray experiments performed by Kundrot et al. [121] which showed that side-chains in lysozyme have a larger response to pressure than backbone.
The physical origin of this heterogeneity in structural response to pressure has been explained by a phenomenological model of non-covalent interactions where each pairwise interaction in the protein's native structure was replaced by an effective harmonic force with a distance dependent force constant. Interestingly the short-ranged interactions were characterized by effective force constants larger than long-ranged ones [82, 80], meaning that the former are stronger and less compressible than the latter.

PRESSURE IMPACT ON PROTEIN STRUCTURE LEVELS To summarize, we can conclude from the previous section that a protein's reaction to pressure increase is often due to a complex balancing of opposite effects which tend to combine differently at different pressures. The effects on the interactions that determine the internal protein structure, have consequences on all the levels of structural organization (i.e. on secondary, tertiary and quaternary structures). In the following, we report schematically these repercussions on each level with reference to the experimental techniques which were used to detect them.

Quaternary structure is mainly maintained by hydrophobic interactions and for this reason it is the most sensitive to pressure. As mentioned in the previous section, moderate pressure below 1.5kbar favours dissociation of oligomeric proteins [180] which results in very large and negative volume changes. Sometimes pressure-induced dissociation leads to formation of individual non-denaturated subunits, as in the case of tryptophane synthethase at 1.5kbar[149], but more frequently it is followed by subsequent conformational changes in individual subunits[150]. When quaternary structure is maintained by either pressure-insensitive interactions like hydrogen-bonds or by pressure-enhanced interactions like aromatic clustering (due to hydrophobic interaction), dissociation can appear also at pressures higher than 3kbar [132].
Pressure effects on quaternary structure are usually investigated by electrophoresis, ultracentrifugation [152], fluorescence spectroscopy[180] and NMR spectroscopy[100].

Tertiary structure can be irreversibly changed by pressures above 2kbar. However, sometimes reversible denaturation can be found also at pressure between 4 and 8kbar[171]. This shows that the volume changes are not completely dominated by hydrophobic interactions but that other type of weak interactions can also play an important role. An important feature of pressure denaturation is the formation of a *molten globule* in which proteins, even though retaining their intact secondary structure, change their tertiary structure adopting conformations with hydrodynamic radii 10-20% higher than that of the native state[154]. For the studies of the tertiary structure changes induced by pressure, the more frequently used methods are NMR spectroscopy, X-ray, UV-visible and fluorescence spectroscopy[121, 180, 100].

Secondary structure undergoes significant changes at very high pressure, above 3-7kbar. These changes, in contrast to those that affect tertiary structure result always in non-reversible protein denaturation. Indeed, such pressures induce a high compression of protein structure resulting in the rupture of hydrogen bonds, which play a major role in the

formation of secondary elements like α-helices and β-sheets [211]. As regards the secondary structure, one of the most suitable methods to detect changes is Fourier transform infrared (FTIR)spectroscopy which gives insights into the protein vibrational spectra where the band corresponding to the vibration of C=O bond of the amide group is easily detectable. The latter is very sensitive to the conformation of the polypeptide. Another technique frequently used to detect changes in secondary structure is circular dichroism (CD) spectroscopy which can rapidly give the percentage of α-helices, β-strands and random-coil in a protein solution [196].

The fact that every pressure range affects roughly only one structure level makes this physical variable a rather better tool to investigate protein structure stability than any other perturbation like temperature or pH changes.

1.2 ORGANISMS ADAPTED TO EXTREME ENVIRONMENTS

Organisms are integrated entities, not collections of discrete objects [2]

While in the previous section the general effects of pressure and temperature on proteins were presented, here I want to focus on the particular aspect of the chemical and physical properties of proteins from organisms which live under extreme conditions. More precisely, a special case of extreme environment, the *warm deep-sea*, in which both high pressure and high temperature can be found, will be presented. Furthermore, inspired by the phrase that opens this section, I want to point out the fact that the study of protein molecular adaptation to extreme environments could be highly improved by the knowledge of the framework of the whole-cell adaptation. Although this section is intended to be an introduction to the subject of the biology of extreme environments, it is inevitably incomplete and mainly reflects the personal view and interests of the author. Many more aspects are described in a number of excellent reviews [85, 93, 182, 158, 107, 151, 191, 58].

The *biosphere*, i.e. the surface of the earth known to host life, ranges from the abyssal region of the deep sea to the heights of the Himalaya and shows a wide range of example of adaptation to 'extreme' environments. From a quantitative point of view, physical limits to life today known are the following[3][92]:

2 From *The spandrels of San Marco and the Panglossian paradigm: a critic to the adaptationist programme*, S.J. Gould and R.C. Lewontin, Proc. Royal Soc. London B, **205**, 581-598, (1979).

3 As a reference: the critical point of water is at 374°C and 2.2kbar

- $-40°C < T < 115°C$
- $P < 1.2kbar$
- $\sim 1 < pH < 11$

In the previous section we have discussed the effects of pressure on proteins structure, here we want to recall some critical effects of temperature on biological chemistry: *i)* at 100 °C thermal degradation outruns biosynthesis[92, 95]; *ii)* at 110-140 °C hydrophobic hydration effects in protein vanish [92]; *iii)* and at 150 °C amino acids start to fall apart [141, 140].
These are obviously the extreme boundaries that do not permit life developing but their knowledge is useful to understand the increasing effects that organisms have to face when approaching these limits. From an evolutionary point of view, this means that organisms should have to find a way to *compensate* these effects.Hence, organisms which *became* extremophiles [167] found some mechanisms to protect themselves against the chemical and physical damages induced by the environment.

Little is known on how this is accomplished[90, 49] but presumably two parallel strategies are followed: the first one is the compensation of degrading processes made through a *tuning* of the synthesis rate of biomolecules that would be compatible with the average lifetime of the molecules [92]; the second one is the incorporation of point mutations in protein sequences in order to increase their structural stability, resulting in a longer lifetime and thus a longer time of functioning[93].
Nevertheless, this *strategy* adopted in molecular adaptation does not seem to engender significant differences between the global distribution of amino acids in *mesophile* proteins, i.e. obtained from 'normal' organisms, as compared to the *extremophile* ones. Deckert et al. showed, for example, that the distribution of amino acids contents in mesophilic and thermophilic proteomes, even though not identical, do not present any significant differences that could lead to a simple method to distinguish them [48].

As a consequence, adaptation of biomolecules seems to be related exclusively to different local patterns in amino acids sequences resulting in a reorganization of the weak non-convalent interactions that govern protein stability and flexibility[147, 94]. The *new deal* for the stability of extremophile proteins is thus played by the re-organization of mechanisms that are, as we have seen in the previous paragraphs, sensitive to environmental changes; they are for example: increased number of hydrogen bonds, salt bridges, improved core

packing, shorter and/or tighter surface loops, enhanced secondary structure propensities, or oligomerization[134, 135, 136].

From a physical point of view, a first quantitative experimental measure of protein stability can be made by means of differences in the Gibbs free energy, $\Delta G(T, P, N)$, between the more stable global minimum and its nearest minima. Jaenicke and Böhm [96] showed that mesophile and extremophile proteins share ΔG of the same order of magnitudes ($\sim 50 kJ mol^{-1}$) even though extremophiles may show a variability in the range $10 - 100 kJ mol^{-1}$. The difference between the typical values for mesophilic proteins and the variation found in extremophilic ones correspond to the energy of a few non-covalent interactions and confirm the significant role the latter play in molecular adaptation[157, 147, 151, 158].

The central issue in the adaptation of biomolecules remains, however, the conservation of biological functionality which means a well-balanced compromise between stability and flexibility[197, 92]. Hence, the process of adaptation has to *translate* the properties of mesophilic proteins toward the relative extreme conditions which means towards physiological conditions that make molecular properties of extremophilic counterparts very similar [93]. This translation is perfomed by a re-organization of non-covalent interactions. Several experiments have provided evidence for this process [217] and showed that dynamical properties of some mesophilic enzymes at temperature of 25°C are very similar to those of their extremophilic homologues at 70°C. Analogous results were obtained on rubredoxin by a molecular dynamics study [69].

In this context Jaenicke [93] proposed that molecular adaptation to extreme environments should tend to maintain *corresponding states* between respective *native* environments, which take into account overall topology, flexibility and hydration of proteins. Even though this idea found confirmation by experiments in the past, recent works seem to not entirely confirm it [158].

1.3 THE WARM DEEP-SEA ENVIRONMENT

In contrast to common opinion, terrestrial environments, where the ambient pressure is 1bar, occupy only the 1% of the total volume of the biosphere. A large fraction of the earth's surface (70%) is covered by oceans which have an average depth of 3800 meters and thus an average ambient pressure of 380 bar[4]. More than 60% of the marine component of the biosphere is

4 Hydrostatic pressure increases at a rate of 10.5kPa per meter depth, compared with 22.6kPa per meter for lithostatic pressure.

below 1000m [5]. Life below this depth, which is usually defined as the upper boundary of the "deep sea", has to face several unfavourable conditions such as high pressure (from 1 to 1.1kbar) or temperature (from 1 to 110°C) [6] which have strong impact to life mechanisms.

After the pioneering studies of Certes in 1884 [30, 29] which helped to establish the existence of microorganisms in deep sea sediments, the first systematic studies of ZoBell and Johnson[220], gave the birth in 1949 to a new research line which helped, during the last century, to better understand the survival strategies employed by organisms to face extreme conditions.

It allowed for example to elucidate the different adaptation to pressure in several groups of microorganisms which can be divided into two distinct groups: *barophiles* (or *piezophiles*) which are micro-organisms that possess optimal growth rates at pressures above atmospheric pressure, whereas *barotolerant* (or *piezotolerant*) microorganisms are capable of growth at high pressure, as well as at atmospheric pressure, but can be distinguished from piezophiles because they do not have optimal growth rates at pressures above one atmosphere.

Both groups of microorganisms can also be distinguished from *piezosensitive* ones, whose growth is sensitive to elevated pressure, because they can grow at 50 MPa at a rate that is above 30% of their growth rate at atmospheric pressure [107, 108]. Moreover, from a molecular point of view, barotolerant organisms differentiate themselves from normal piezosentive ones by the production of some "pressure-shock" proteins induced as reponse to high pressure environment [97]. These proteins have some equivalents in the "heat-shock" proteins found in organisms which undergo thermal stress [67, 174] but their mechanism is far from being completely understood.

In late 70s, the first colonies of "deep-sea" organisms were found also near the hydrothermal vents[8]. The latter are warm currents which form where freshly extruded lava contracts upon cooling and allows seawater to penetrate in the newly formed crust of basaltic rocks. Seawater is then expulsed, highly enriched in heavy metals, creating hot vents which significantly change the temperature in the nearby area [98].

Obviously the life of organisms near hydrothermal vents became very rapidly the most intriguing subject in the deep-sea biology and *de facto* most of the barophilic organisms known today are also thermophilic which means that they have been proven to also have greater growth rates at high temperatures (typically from 50 to 90°C) than at usual deep-sea

5 The greatest depth in oceans, in the Mariana Trench is near 11000 meters whereas the deepest floor found in freshwaters is at 1632 meters in the Lake Baikal

6 From a biochemical point of view it is worth to note for exemple that both light and photosynthesis are available only until 300 meters of depth

temperatures (around 2°C) [1].

This environment, usually called the *warm* deep-sea, introduced several interesting questions from a biological and evolutionary point of view which demanded also new explanation from biochemistry and physics.

PATHWAYS OF ADAPTATION TO WARM DEEP-SEA ENVIRONMENTS Before focusing on some aspects of the molecular adaptation of proteins to this environment, here some well known facts about the whole-cell adaptation are reported schematically, for they play a fundamental role in defining the general framework in which all molecular results should be interpreted. The main features of biological adaptation to high hydrostatic pressure are[182]:

Adaptation to pressure allows deep-living species to thrive under high pressure but it can also reduce their ability to live at moderate depths. This means that deep-living species are usually barophilic and not barotolerants[179].

Common patterns of adaptation appear in widely different types of deep-sea organisms[179].

Most of the physiological and biochemical processes which are in general found to be pressure-sensitive in normal organisms, are rather stable under high pressure in deep-living ones and exhibit, at *in situ* temperatures and pressures, values similar to those found in terrestrail or shallow-living organisms at 1bar pressure[41, 44].

Species living in deep sea hydrothermal vents show that adaptation to both temperature and pressure are essential for cellular growth and physiological function. This means that adaptation to high pressure does not *pre-adapt* organisms from cold deep sea to the environment typical of the hydrothermal vents[182, 45].

While the first three points seem to confirm the idea of the existence of *corresponding states* between mesophilic and extremophilic organisms, the last point reveals the fundamental interplay of pressure and temperature in warm deep-sea environments. This means, however, that evolution has to cope with complex superpositions of parameters which are difficult to separate and renders impossible the definition of strategies of molecular adaptation.

In an extremely simplified view, the complex mechanism which governs the life of a cell can be schematically divided in two parts: protein synthesis, which furnishes all the "tools" (the proteins) necessary for the cell machinery to work; and cellular metabolism, which includes all the process that constitute the cellular life itself.
Following this scheme, we present the principal properties of protein adaptation to a warm deep-sea environment, focusing on proteins that participate in the cellular metabolism and those which have a role in other protein synthesis.

Enzymes, by their role as catalysts and regulators of metabolism, are the mediators by which living systems function and reproduce. Somero et al [183] proposed that the successful adaptation of an organism entering the deep sea would require the maintenance of the same balance of enzyme synthesis, catalytic activity, and regulation as its surface-living counterparts. Indeed, from several studies of enzymatic functions of deep-living organisms, a few convergent common properties were found [130, 98]: *i)* the preservation of an enzyme's capacity to bind substrates under deep-sea conditions is essential in adapting metabolism to that environment; *ii)* structural changes induced at pressures as low as 50-100 bar may modify enzyme function sufficiently to induce evolutional selection for pressure insensitivity; and finally *iii)* the capacity of an enzyme to function independently of pressure is acquired at the cost of a reduction in catalytic efficiency.
Furthermore, in the context of the combined adaptation to temperature and pressure, it is worth noting that for several hydrogenase enzymes from widely different warm deep-sea organisms, pressure was found to significantly increase their thermal stability [45, 188]. This type of enzyme undergoes large conformational change and hydration during its activity and this makes them particularly well suited for investigating the effects of pressure.
In general for deep-sea organisms from the hydrothermal vents, the stabilization of catalytic proteins was often found as the result of some point mutations in primary sequences whose outcome could either be the increase of the size of the hydrophobic core and its compactness (essentially substitutions with hydrophobic amino acids with larger side-chains), the augmentation of hydrogen bonds (substituiting lysine with arginines) or the reduction/deletion of amino acids with side-chains sensitive to high temperatures (asparagines, glutamines, cysteines and tryptophane) [142].

Protein synthesis is ideally divided into two steps: *transcription*, in which the genetic code is copied from DNA to a RNA messenger (mRNA); and *translation* where peptides are bound

together with the help of ribosomes and where genetic code in mRNA is decoded into amino acids sequences. Both process are found to be highly sensitive to pressure [124].

Early studies investigating the effects of pressure on *E. coli* revealed that an essentially complete cessation of protein synthesis occurred at growth-inhibiting pressures of 680 atm [67]. Subsequent comparison of protein synthesis of *E. coli* and known barotolerant organisms permitted the barotolerance properties to be associated to the activity of ribosome [124] and in particular to its 30S subunit [125].

Indeed, ribosomes are made of two subunits (a smaller and a larger one) which are capable of associating and dissociating. As we will discuss later in the text, this phenomenon is important because the initiation of protein synthesis requires free ribosomal subunits in order to allow the formation of a complex containing the mRNA anchored to the small subunit before the joining of the large subunit. The ribosome's subunit association is known to be a rate-limiting process in the protein synthesis [162] and was found to be inhibited at elevated hydrostatic pressures in mesophilic organisms [185].

This effect was previously discussed as a possible reason for the inhibition of protein synthesis and, hence, of cell growth by hydrostatic pressure [91] but later more sophisticated *in vitro* studies showed that ribosomal complexes retain their "associated" state over the whole range of biologically relevant pressures [68]. Other studies on ribosomes from mesophilic organisms showed that association can be regulated also by temperature which usually enhances the association equilibrium towards the associated species[83, 184]. Interestingly, these results show that under estimated physiological conditions in several eukaryotes, ribosomes are present almost exclusively in their associated form. The cell must therefore possess some mechanisms for maintaining a pool of free ribosomal subunits. These mechanisms have to account for protein synthesis inhibition under extreme conditions. Unfortunately, very little is known about the differences between the protein synthesis machinery of extremophiles and mesophiles.

METHANOCOCCUS JANNASCHII This thesis will focus on a protein produced by an archaeon[7] microbe living near the hydrothermal vents, *Methanococcus Jannaschii*. This organism, discovered in 1963 as a "methane-producer" [102], usually lives at temperatures ranging from 48 to 94 °C with an optimum temperature at 85°C and at pressures of more than 200 bar (*in situ* depth 2600 meters).

7 The archaea are known as a primitive form of life significantly different from both prokaryots and eukaryotes. For this reason is often referred to as the *third* domain of life.

M Jannaschii's genome was the first archaeon genome to be completely sequenced [18] and revealed the complex evolutionary relationships between archaea and eukaryotes and bacteria: only less than an half of the genes found in its genome could be matched to those of other organisms. In the context of this thesis work, from a molecular biology point of view, the most interesting feature of this organism is that while sharing the same protein biosynthesis machinery (both transcription and translation) as eukaryotes, it does also have genes homologous to both eukaryotes and bacteria for the initiation part of the translation process. The latter is a fundamental and rate-limiting part of the translation and will be discussed more deeply in the next section.

It has been shown that *M Jannaschii*'s metabolism and growth at high temperatures are enhanced by pressure up to 750 bar whereas the upper limit for growth is 90°C either under low or high pressure [102]. These results suggested that *M.Jannaschii* is effectively a barophilic organism and not just barotolerant. The finding seemed to be confirmed by studies on enzymatic reaction: it was shown that application of 500 bar increased the thermal half-life of hydrogenase from M. jannaschii 4.8-fold at 90 °C °C [74] and increased the half-life of a protease from the same organism 2.7-fold at 125 °C [138]. Nevertheless, recent works showed that pressure inhibits the 20S proteasome function from *M Jannaschii* [55], suggesting that in some cases lower protein activity at the *in situ* pressure could be a regulatory mechanism that confers some advantage to the whole cell. Hence, a more complex behaviour of the cell response to high pressure and high temperature should be envisaged.

1.4 THE ANTI-ASSOCIATION FACTOR

In the following we want to give a short introduction to the protein studied in this thesis, which is called the *Anti-Association Factor*[170, 199]. A complete description of the molecular biology background which is necessary to introduce the function of this protein, would be largely out of the scope of this thesis and we limit our discussion to the essential key points.

The anti-association factor belongs to the familiy of initiation factors and is usually also called *initiation factor 6*. In the following, in order to maintain a coherence with the cited most recent bibliographic references, I will call it by the short name IF6 sometimes preceded by a prefix indicating the domain of life from which it comes: **eIF6**, for eukaryotic initiation factor

6, **pIF6** for its prokaryotic counterpart and **aIF6** for the archaeon one.

IF6, like the other initiation factors, takes part in the first steps of the translation process. Translation initiation can be subdivided into three steps: first, binding of the specific initiator Met-tRNA to the small ribosomal subunit; second, binding of the resulting complex to a mRNA and locating the initiation codon; and third, joining of the large ribosomal subunit to generate a translation competent ribosome. All these steps are facilitated by soluble proteins termed translation initiation factors, and the structures of many of them have been characterized at the atomic level. It is worth noting that while the essential part of this process is shared by both prokaryotes and eukaryotes, the process itself is decidedly more complex in eukaryotes than in prokaryotes and in part reflects the fact that much of the regulation in prokaryotes is via the coupling of transcription and translation. Such coupling is not possible in eukaryotes as transcription occurs in the nucleus while translation occurs in the cytoplasm.

The eukaryotic initiation factors can be divided into two groups: those that bind and operate on ribosomal particles promoting subunit dissociation and association, initiator Met-tRNA binding, and mRNA binding; and those that are aimed at mRNA and engaged in preparing its upstream region for initiation. The first group contains the factors analogous to prokaryotic pIF1, pIF2 and pIF3, namely eIF1 (and eIF1A), eIF2 and eIF3 respectively, as well as several additional factors, such as eIF2B, eIF5 and eIF6. The second group seems to have no analogs in prokaryotes and includes special mRNA-binding and mRNA-unwinding proteins facilitating initiation of translation; these are the factors of eIF4 group, namely eIF4A, eIF4B, eIF4F and eIF4E [186].

The role of IF6 in the translation initiation is more complex than those of the other initiation factors because it engenders different functions about which very little is known. First of all, IF6 does not act as a true initiation factor because it does not form any initiation complex[8] either with RNAs and ribosome subunits as proved by the fact that yeast cells depleted of eIF6 remained active in translation of mRNAs in vitro [178]. Its principal functions can be resumed as follows:

- it is required for biogenesis of the 60S ribosomal subunit both in archaebacteria and eukaryotes [173, 178, 177, 212]

8 Initiation complexes are macromolecular complexes formed by mRNA, tRNA and ribosome subunits which come together to form the machinery that perform the first two steps of the translation. IF6 does not take part directly into this mechanism.

- it acts as *de facto* anti-association factor of the 60S and 40S subunit in the cytoplasm [178, 28]

Moreover, a very recent work also uncovered another evolutionary conserved function of IF6 in the microRNA-mediated post-transcriptional silencing[34] which could extend significantly the peculiarities of IF6.

Si and Maitra showed that depletion of eIF6 in yeast cells resulted first in a decrease in the rate of protein synthesis and then in cessation of cell growth [178] and Basu *et al.* [10] showed also that the protein depletion caused in cells a selective reduction of 60S ribosomal subunits. These two results and the previous one confirmed the evolutionarily conserved main role of IF6 in biogenesis and assembly of ribosomal units and also the absence of its direct participation into the translation process.
The suggestion of these two functions of IF6 is also strengthened by the localization of the protein in both nucleus and cytoplasm[6] where ribosome subunit biogenesis and ribosome assembly take place respectively. Furthermore, Basu and coworkers [10] showed that phosphorylation of eIF6 regulates its distribution in nucleus and cytoplasm [11].

As a whole, the multiple functions of IF6 make of this protein an essential factor for cellular growth and protein synthesis. Very little is known about its participation in the 60S biogenesis whereas several works in the past decades better elucidated its role of anti-association factor at a molecular level [28]. This function has been identified in eukaryotes as part of process which involves a protein promoter of ribosome translation, RACK1 and a protein kinase PKC.

In the context of this work, it is worth noting that eIF6 has been proven to be very sensitive to heat: in homologues from wheat germ [170] and calf liver[199], eIF6 activity had an optimum at around 37°C but was completely abstent between 50 and 60°C. Whether the cessation of activity is due to a partial or complete protein denaturation is not known.

From a molecular point of view, IF6s are 26kDa proteins which share a phylogenetically conserved sequence of 224 residues with 30% similarity. All known eukaryotic IF6 contain an additional carboxyl terminal extension of 21 amino acids. Even though this tail does not seem to be directly involved to the anti-association function of IF6[66], as proved in *in vitro* reactions, it has been proven to be somehow related to this process through the

phosphorylation of a serine residue (Ser235) by the protein kinase PKC [28].

In 2000, the first molecular structure of an IF6 homologues was resolved by x-ray crystallography by Groft and coworkers [66] who obtained the structures of two homologous IF6s issued from *M.Jannaschii* and from *Saccharomyces ceraevisie*. Both structures revealed a very unusual internal pseudo-symmetry made by the disposition around a common axis of five copies of a repeating α/β subdomain of about 45 residues. Each subdomain contains: one long α helix, a shorter α-helix or a 3_{10}-helix and three β-strands. While the global structure of the first 224 residues is well known and has been proven to be evolutionarily conserved by homology modelling [66], nothing is known about the structure of the C-terminal extension of 21 amino acids.

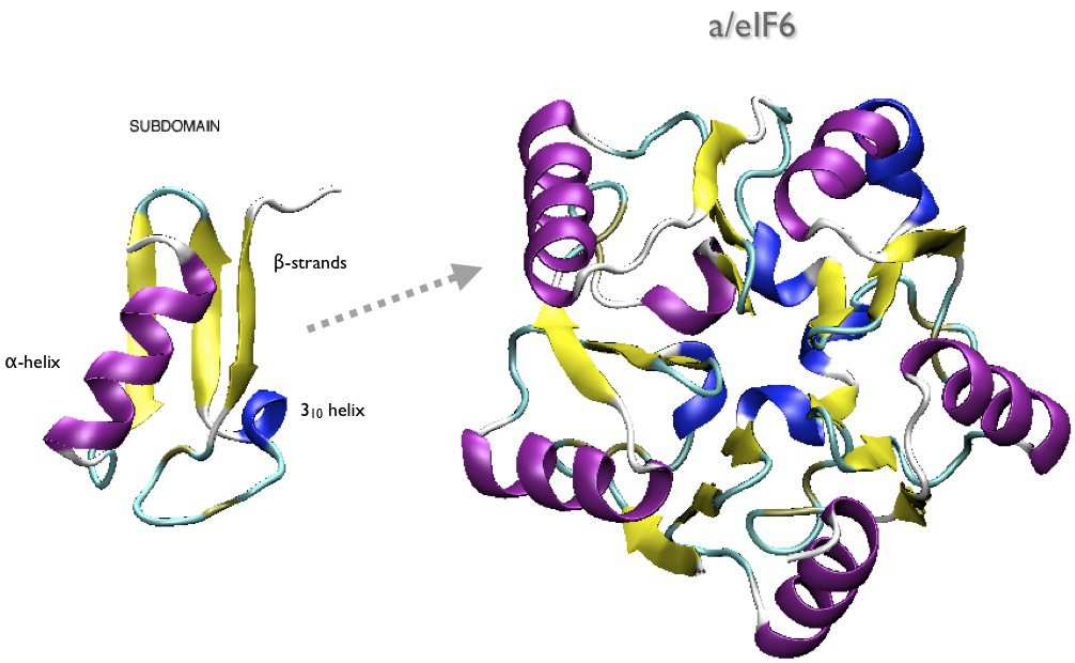

Figure 7: *Cartoon* representation of IF6 backbone. The color scheme highlights different type of protein secondary structure. One of the pseudo-subdomain is shown on the left-hand side as reference.

From a structural point of view is also worth-noting the fact that the five subdomains do not make a unified hydrophobic center of the protein. Instead, they produce an hydrophobic "torus" by their association which in turn creates a cave-like hollow in the center of the structure. The cavity is large enough to let water molecules enter and in the crystallographic data sixteen of them were found in a well-ordered conformation of pentagonal layers and are hydrogen bonded to the carbonyl oxygen atoms of two residues in the neighboring β-strands. In the yeast (*Saccharomyces ceraevisie*) eIF6 this cavity is closed by an arginine (residue 61) whose guanidinium group makes several hydrogen bonds with glycines in the short α-helix of subdomains.

1.5 THE THESIS PROJECT "AT A GLANCE".

As stated earlier in this introduction, the key point for molecular evolution is the maintain of biological function. This is accomplished in proteins either by point mutations or by simple structure re-arrangements. Nonetheless, evolution does not have a unique pathway to follow this procedure.

This thesis focused on the way proteins evolutionarily "react" to large environmental changes, i.e. how they adapt to function even in organisms which live in extreme conditions. Following this idea, this work concentrated on the study of the adaptation of the IF6 from *Methanococcus Jannaschii* to extreme environments. In this context the comparison with proteins in *normal* (mesophilic) conditions would have been essential to finely characterize the ability to adapt to extremes conditions.

IF6s from *Saccaromyces cerevisiae* was chosen as mesophilic counterpart of *Methanococcus Jannaschii*. They share only 33% of identity in sequence but their structures are significantly similar to each other and to most of the other IF6s homologues [66]. As a whole, this means that IF6's function was highly conserved during evolution but it does not explain how function and, apparently, structure have been maintained in the adaptation to a wide variety of environmental conditions.

In this context, the work presented here will try to give a first answer to the following questions:

Where does the extremophilic signature come from ?

If structure cannot be the origin for this, can it be the dynamics??

Of course all answers given here will be limited to the particular case of adaptation to the *warm deep-seas* and will not be neither exhaustive nor general but could introduce a new approach in study of protein evolution.

The main tools used in this work are molecular dynamics simulation and neutron scattering experiments. After a brief introduction to the main concepts of the theoretical frameworks that guided this work, the experimental setups will be presented.

In the following chapter a wide and detailed presentation of a novel method for the characterization of the protein secondary structure is given. The method, called *ScrewFit*, was developped in the context of this thesis but it has rapidly found some different applications of the analysis of protein structures. Some examples already published as scientific articles in international peer-reviwed journals will be given.

Finally the results of this work on IF6 homologues will be presented and discussed. Some general outcomes will be then used in the tentative of giving answers to the questions about protein evolution proposed above.

2

MATERIAL AND METHODS

2.1 MOLECULAR DYNAMICS

2.1.1 *The basic principle*

Molecular simulations used in combination with experimental methods, are a very useful tool to give new insights into the dynamics and structure of complex molecular systems. Simulation methods are used within different approximations, depending on the experimental results with which they are confronted.

In this sense, neutron scattering techniques are well matched to *molecular dynamics* simulation (MD) method which combines the classical equations of motions with empirical force fields obtained from *a priori* quantum-dynamical calculations of atomic interactions in the building blocks of the systems of interest. The basic approximation made in MD is the Born-Oppenheimer approximation which states that due to the large difference in masses between electrons and atomic nuclei, the electron dynamics is orders of magnitude faster than that of the nuclei and, therefore, it can be assumed that electronic shells adapt instantaneously to the positions of the nuclei. Moreover the dynamics of the nuclei is treated by classical mechanics.

De facto, the MD gives information about the dynamics of the same objects that are directly seen by neutrons: the atomic nuclei. As a consequence, MD and neutron scattering give access to the same length and time scales (ranging from 1 Åto 100 Åand from 0.1ps to 10ns, respectively).

In an MD simulation each atom is represented by a point mass whose dynamics is described by the classical Newton's equations of motion:

$$m_i \ddot{\mathbf{r}} = \mathbf{F}_i, \qquad i = 1, \ldots, N \tag{2.1}$$

where m_i is the mass of atom i and $\mathbf{F}_i$ is the total force acting it. The force $\mathbf{F}_i$ is derived from the potential energy $U(\mathbf{r}_1 \dots \mathbf{r}_N)$ through:

$$\mathbf{F}_i = -\frac{\partial U(\mathbf{r}_1 \dots \mathbf{r}_N)}{\partial \mathbf{r}_i} \tag{2.2}$$

As already mentioned, the energy $U(\mathbf{r}_1 \dots \mathbf{r}_N)$ is an effective energy which describes all types of atomic interactions. The generic form of the potential energy is :

$$\begin{aligned} U_{\text{total}} = & \sum_{\text{bonds}} K_r(r - r_{eq})^2 \\ & + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 \\ & + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ & + \sum_{i<j} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \\ & + \sum_{i<j} \frac{q_i q_j}{\epsilon R_{ij}} \end{aligned} \tag{2.3}$$

The terms of this potential can be divided into two groups: *bonded interactions*, which are related to the covalent bond structure taking into account the stretching of bond lengths, the bending of bond angles and dihedral angles described by three consecutive bonds; *non-bonded interactions* describe the repulsion between neighboring atoms, the dispersive forces due to mutually induced atomic dipoles (both contained in a Lennard-Jones 12-6 function) and electrostatic interactions.
Throughout this thesis all molecular dynamics simulations have been performed using the AMBER9 simulation program [26]. The whole set of simulations used the AMBER99SB force field which is a recent upgraded version of the AMBER94 force field, usually used for protein molecular dynamics, and which contains a reparametrization of the backbone torsion terms and achieves a better balance of the different secondary structure elements [86].

MD simulations of biological molecules are mostly performed in solution, typically water, as in real biological environments. In order to obtain a realistic solvation of the sample molecule, the amount of solvent molecules is usually very large and counts for a

substantial part of the computational effort. Several simple models for water molecules have been developed in the past which allow simulations to be performed with a moderate computational cost. A simple and widely used model is the TIP3P model [103, 104]. Here the molecule has a rigid geometry with fixed bond lengths and angles; the intermolecular interactions are modeled by Coulombic and Lennard-Jones terms. Moreover, the model contains three sites for the electrostatic interactions: the partial positive charge of hydrogens is balanced by a negative charge placed at the oxygen atom position; the van der Waals interactions are computed with only one point of interaction per molecule, centered on the oxygen atoms.

The set of ordinary differential equations (2.1) is solved numerically using a discretization method. The general procedure can be summarized as follows. Given the initial dynamic state (position and velocities) at time t, the positions and velocities at time $t + \Delta t$ are calculated. The finite time-step Δt depends on the integration method as well as on the system itself. The basic criteria for a good integrator are: it should be fast, require little computer memory and be time-reversible; it should permit the use of a relatively long time step and guarantee the conservation of the total energy within a reasonable error which is related to the definition of the time step Δt. For later analysis one assumes also the ergodicity of the simulated trajectory, i.e. if A is a function of time then

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T A(\tau) d\tau = \langle A \rangle$$

where T is the simulation length and $\langle \cdot \rangle$ a phase space ensemble average. This condition allows the calculation of the ensemble averages by means of time averages.

PERIODIC BOUNDARY CONDITIONS AND SUMMATION OF NON-BONDED TERMS A real experiment relates to macroscopic samples consisting of $\approx 10^{23}$ atoms whereas the simulated systems usually contains between 10^3 to 10^6 atoms. This discrepancy leads to significant differences in physical quantities of the bulk, due to a larger influence of boundary effects in the simulated systems. For this reason, periodic boundary conditions are applied: the initial unit cell is replicated infinitely in all directions, meaning that each atom i having position $\mathbf{r}_i = (x_i, y_i, z_i)$ in the central box has an infinite number of image atoms with

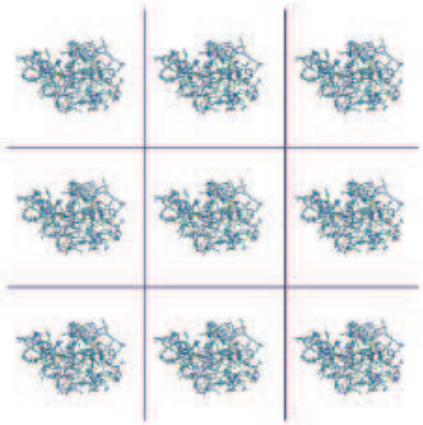

Figure 8: Bidimensional representation of the periodic boundary condition

positions $\mathbf{r}_i + \mathbf{n}_i L = (x_i + n_1 L, y_i + n_2 L, z_i + n_3 L)$ where L is the linear dimension of the simulation box (assuming a cubic shape) and the integer triples $\mathbf{n} = (n_1, n_2, n_3)$define the direction in space. From a practical point of view, this means that during the simulation, whenever an atom leaves the principal simulation box, it enters again from the opposite side thus keeping the average density of the system constant.

When calculating the force acting on an atom of the system, one should consider both the interactions with the other $N-1$ atoms and those with all images lying in the surrounding box. This would lead to a sum of infinite terms which is clearly not possible in practice. For this reason, it is assumed that each atom interacts only with the nearest images of all other atoms. In this approach, called the *minimum image* convention, the distance between the pair of atoms i, j is defined by $d_{ij} = min_{\mathbf{n}}\{||\mathbf{r}_j + \mathbf{n}L - \mathbf{r}_i|\}$, i.e. the distance between atom i and the closest image of atom j. The outcome of this convention is that the calculation of potential energy and forces would involve only $1/2N(N-1)$ terms.

For short-ranged interactions, such as Lennard-Jones ones, the list of atoms interacting with atom i can be restricted to its neighboring pairs taking a cut-off distance which is usually set lower than $L/2$, where L is the linear dimension of the box, in order to ensure that each atom is counted only once. Nevetheless, this truncation of potential energy calculation can be safely applied only to the short-ranged forces. The long-ranged forces, i.e. the electrostatic ones, cannot be calculated within this approach because they contribute significantly to the overall potential also at distances larger than the box dimensions.

The periodic boundary conditions can be used also to calculate the electrostatic interactions between all atoms of the system by means of the Ewald summation method. The term due to these interactions, when written for periodic systems, must take into account the interactions between atoms belonging to different *periodic images* of the original box:

$$U_{elect} = \sum_{n}{}' \sum_{i \leqslant j} \frac{q_i q_j}{\epsilon \|\mathbf{r}_{ij} + \mathbf{n}L\|} \tag{2.4}$$

where the vectors $\mathbf{n} = (n_1, n_2, n_3)$ with $n_1, n_2, n_3 = 1, ..\infty$ are the lattice vectors which define the periodic images of the original box. This sum is conditionally convergent, i.e. the results depend on the order in which the terms are summed up. A suitable choice is to sum boxes in order of their distance from the original one, thus approaching the sum of terms over spherical layers for which the value of U_{elect} in equation (2.4) is known. A rapidly convergent form to this value, can be obtained by the substitution of (2.4) by the sum of two rapidly convergent series, as originally proposed by Ewald [53, 131].

In the general approach, each point charge in the system is surrounded by a charge distribution of equal magnitude and opposite sign. The form of this distribution is usually a Gaussian:

$$\rho(r) = \frac{q_i \beta^3}{\pi^{3/2}} \exp(-\beta^2 r^2) \tag{2.5}$$

This distribution screens the interactions between neighboring charges and summed, over the periodic system, together with the original point charges, leads to a rapidly convergent potential which has the form [46, 78]:

$$\phi(\mathbf{r}; \beta) = \sum_{n} \frac{\mathrm{erfc}(\beta|\mathbf{r} + \mathbf{n}|)}{|\mathbf{r} + \mathbf{n}|} \tag{2.6}$$

where erfc is the error function and β is a parameter which affects the rate of convergence of the series in the direct space of the periodic lattice. Another charge distribution of the same sign of the original charges and with the same shape of $\rho(r)$ is added to reduce the overall potential to the one due to the original set of charges. This *cancelling* distribution is summed in the reciprocal space of the periodic lattice by means of Fourier transforms. To correct the fact that the latter series contains also the interaction of each distribution with itself, a *self-term* is also added. Finally, another additional term, $J(\mathbf{D})$ is usually added to take into account the effects due to the dipole moment $\mathbf{D}$ of the original box as well as the external dielectric constant.

The final rapidly convergent potential reads:

$$
\begin{aligned}
U_{elect} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\Bigg(\sum_{|\mathbf{n}|=0}^{\infty} q_i q_j \frac{erfc(\beta|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|} \\
+ \frac{1}{\pi L^3}\sum_{\mathbf{k}\neq 0} q_i q_j \frac{4\pi^2}{k^2}\exp(-k^2/4\beta^2)\cos(\mathbf{k}\cdot\mathbf{r}_{ij})\Bigg) \\
- \frac{\beta}{\sqrt{\pi}}\sum_{i=1}^{N} q_i^2 + J(\mathbf{D})
\end{aligned}
\tag{2.7}
$$

where $\mathbf{k}$ defines the reciprocal lattice vectors. The rate of convergence of the series in (2.7) is determined by the value of β. Adjusting it can significantly improve the computational effort and transform an initial $\mathcal{O}(N^2)$ algorithm to an $\mathcal{O}(N^{3/2})$ one (if the system is homogenous).

Further improvements in the efficiency of the algorithm can be achieved by choosing β sufficiently large that atom pairs for which r_{ij} exceeds a specific cut-off distance r_{off} are negligible in the first series in (2.7), which reduces this term to an $\mathcal{O}(N)$ calculation. The sum in the reciprocal space in the second series of (2.7) can then be approximated by means of the Fast Fourier Transfom algorithm. The latter algorithm requires a regular discrete grid of points and for this reason the point charges are mapped onto the neighboring grid points by means of interpolation. The spacing of the grid significantly affects the accuracy of this approximation. This improvement of the original Ewald summation is referred to as the *particle-mesh* method and has global complexity of order $\mathcal{O}(N \ln N)$, N being the number of particles.

Throughout this thesis, periodic boundary conditions were applied for all production simulations and the electrostatic interactions were computed using the particle-mesh Ewald method for which the cut-off r_{off} distance in the direct sum was set to 8 Åand the grid for the term in the reciprocal space had a spacing of 0.75 Åwith an B-spline interpolation upto the 4^{th} order.

EXTENDED ENSEMBLE MOLECULAR DYNAMICS The MD simulations presented in this work have been performed to reproduce the environment of experimental setups. For this purpose, they emulated the condition of constant average temperature and pressure which

corresponds to the isothermic-isobaric ensemble, NPT in statistical mechanics. In the case of MD simulations, this means that the number of particles N in the system is constant and the value of the temperature T and of the pressure P are instantaneously readjusted toward the average requested value by the control of some supplementary variables. The latter modify both the velocities (to adjust the temperature) and the interatomic distances (for pressure adjustment), introducing additional degrees of freedom in the original equation of motion (2.1). The following paragraphs give a brief explanation on how these re-adjustements are performed.

Langevin thermostat The *macroscopic* average temperature is maintained constant by means of a coupling of the system with a thermal bath at constant temperature which should work like a real *thermostat*. In the present work this coupling is achieved by inserting in Newton equation of motion a *friction* and a *stochastic* term. This leads to the equation known as the Langevin equation and for each atom i reads:

$$m_i \ddot{\mathbf{v}}(t) = \mathbf{F_i} - \gamma_i \mathbf{v}_i(t) + \mathbf{R}_i(t). \tag{2.8}$$

where $\mathbf{F_i}$ is total force applied on atom i, $\mathbf{R}_i(t)$ is a stochastic force and γ_i represents the intensity of coupling with the thermal bath.
The stochastic term is defined as a *white-noise*, i.e. its ensemble average must be zero and it is completely uncorrelated at all times:

$$\begin{aligned} \langle \mathbf{R}_i(t) \rangle &= 0 \\ \langle \mathbf{R}_i(t) \cdot \mathbf{R}_i(t') \rangle &= 6 k_B T \gamma_i \delta_i(t - t') \end{aligned} \tag{2.9}$$

where $\delta_i(t - t')$ is the Dirac distribution centered in $(t - t')$.
The friction term $-\gamma_i \mathbf{v}_i(t)$ opposes the motion of atom i affecting its velocity: it is proportional to the atom velocity and with opposite sign. This term is strictly related to the stochastic force by the fluctuation-dissipation theorem which leads to:

$$\int dt \langle \mathbf{R}_i(t) \cdot \mathbf{R}_i(0) \rangle = 6 k_B T \gamma_i. \tag{2.10}$$

These assumptions have to be made to avoid any memory effect in the values of $\mathbf{R}_i(t)$ which could bring to unreal drift phenomena in velocity values. This method allows to the

control of the temperature through the mutual balance between the friction and stochastic terms in adjusting the velocity of each atom. It is worth noting here that Langevin thermostat allows a correct sampling of the canonical ensemble but *a priori* the addition of a friction term in the equation of motion could result in some artifacts in the dynamical properties of the simulated system. Nevertheless, this problem can be biased using friction coefficients much smaller than the actual friction coefficient of the solvent. In the case of water which has $\gamma = 50ps^{-1}$ is was shown that a value of $2ps^{-1}$ does not significantly change protein dynamics [195]. Details on the parameter values used in this work are given in the next chapter.

Berendsen Barostat The adjustment of instant pressure values can be approached exactly in the same way as for temperature, by coupling the system to a *pressure bath*.
The value of the system pressure is related microscopically to the internal kinetic energy and the virial:

$$P = \frac{2}{3V}\left[E_{kinetic} - \Xi\right], \tag{2.11}$$

where

$$\Xi = -\frac{1}{2}\langle\sum_{i>j}^{N} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij}\rangle. \tag{2.12}$$

From these equations it is evident that a change in pressure can be obtained by changing the virial through the rescaling of the interatomic distances r_{ij}. The latter leads to a rescaling of the total volume of the simulated system and of its atomic coordinates. The method originally proposed by Berendsen and coworkers [14] uses the rescaling of distances to regulate the macroscopic pressure of the system. For this purpose the equation that relates the time derivative of atomic coordinates with velocities, $\dot{r} = v$ should be modified with the introduction of an extra term : $\dot{x} = v - \alpha x$. The volume of the system should also change accordingly: $\dot{V} = \alpha^3 V$.

The pressure variation in the *pressure bath* is defined by:

$$\left(\frac{dP}{dt}\right)_{bath} = \frac{P_0 - P}{\tau_p} \tag{2.13}$$

where P_0 is the target pressure value. Additionally, the pressure change is generally also related to the isothermal compressibility β by:

$$\left(\frac{dP}{dt}\right)_T = -\frac{1}{\beta V}\frac{dV}{dt} = -\frac{3\alpha}{\beta}, \tag{2.14}$$

where the last equality is obtained by applying the requested volume re-scaling.

Equations (2.13) and (2.14) leads to:

$$\alpha = -\beta\frac{P_0 - P}{3\tau_p} \tag{2.15}$$

Hence, the equation of motion now reads:

$$\ddot{x} = v - \beta\frac{P_0 - P}{3\tau_p}x \tag{2.16}$$

The solution of this equation, as obtained by the finite difference methods used in MD simulations gives a rescaling factor for the atom coordinates and the linear dimension of the simulation box which is equal to (at the first order in the time step Δt)[1]:

$$\mu = 1 - \frac{\beta\Delta t}{3\tau_p}(P_0 - P). \tag{2.17}$$

LEAPFROG INTEGRATOR In the program suite AMBER9, the integration of the equations of motion for the extended system corresponding to the NPT ensemble used in this work, is implemented through the use of the *Leapfrog* integrator [56, 3]. The latter is a symplectic and time-reversible integrator whose general form is derived by the Taylor series expansion of Newton equations for the atom position $\mathbf{r}$, at timestep t and for its velocity at an intermediate time step $t + \Delta t/2$ [56]:

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i(t - \Delta t/2) + \frac{1}{m_i}\mathbf{F}_i\Delta t \tag{2.18}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \Delta t/2)\Delta t \tag{2.19}$$

The velocity $\mathbf{v}_i(t + \Delta t)$ can be computed *a posteriori* with the relation:

$$\mathbf{v}_i(t) = \frac{1}{2}\left[\mathbf{v}_i(t + \Delta t/2) + \mathbf{v}_i(t - \Delta t/2)\right] \tag{2.20}$$

1 The isothermal compressibility β does not need to be known exactly because its value influence only the accuracy of τ_p without any consequence to the dynamics

When applied to an extended system issued by the use of the Langevin thermal bath and Berendsen barostat, (2.19) reads:

$$\mathbf{v}_i(t+\Delta t/2) = \mathbf{v}_i(t-\Delta t/2) + \frac{1}{m_i}\Delta t\left[\mathbf{F_i} - \gamma_i \mathbf{v}_i(t) + \mathbf{R}_i(t)\right] \tag{2.21}$$

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \left[\mathbf{v}_i(t+\Delta t/2) - \beta\frac{P_0 - P}{3\tau_p}\mathbf{r}_i(t)\right]\Delta t \tag{2.22}$$

2.2 NEUTRON SCATTERING

Neutrons are one of the most useful probes to study the structural and dynamical properties in condensed matter. Neutrons for scattering experiments are usually moderated to be at thermal equilibrium at room temperature with typical energies of ~ 25 meV and wavelength around 1.78Å, which corresponds to the same time and length scale of the thermally excited atoms motions. These facts make neutrons very sensitive to both amplitude and frequencies of atomic motions. For macromolecules, cold neutrons which have typical energies around 2.3meV and wavelength ~ 5Å, are however better adapted to explore longer length scale and slower dynamics.

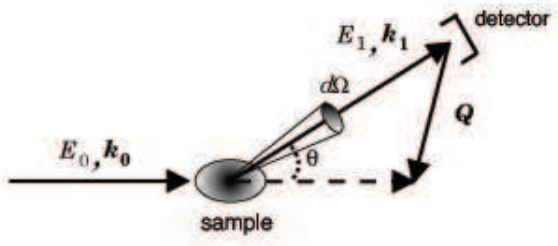

Figure 9: Scheme of neutron scattering

Neutrons interact directly with atomic nuclei via a very short range potential which can be considered as a direct collision between the neutron and a nucleus (see Figure 2.2). An incident monochromatic neutron beam can interact with matter through both absorption and scattering. When scattered, neutrons can be characterized by their wave vector $\mathbf{k_{out}}$ and energy $\mathbf{E_{out}}$ [129, 13].

The number of incident neutrons with wave vector $\mathbf{k}_{out}$ which are scattered with a wave vector $\mathbf{k}_{out}$ in the elementary solid angle $d\Omega$ around the direction of $\mathbf{k}_{out}$ and with an energy exchange $d\omega$ is called *differential cross section* and it is defined as:

$$\frac{d^2\sigma}{d\Omega d\omega} = N \cdot \frac{|\mathbf{k_{out}}|}{|\mathbf{k_{in}}|} S(\mathbf{q}, \omega), \tag{2.23}$$

where N is the number of scatterers in the sample.

The function $S(q, \omega)$ is called the *dynamic structure factor* and it gives access to sample dynamics as a function of $\mathbf{q}$ and ω which represent the momentum and the energy transfer respectively:

$$\Delta\mathbf{p} = \hbar(\mathbf{k_{in}} - \mathbf{k_{out}}) = \hbar\mathbf{q} \tag{2.24}$$

$$\Delta E = E_{in} - E_{out} = \hbar\omega = \frac{\hbar^2}{2m_n}(\mathbf{k_{in}^2} - \mathbf{k_{out}^2}) \tag{2.25}$$

The value of $|\mathbf{q}|$ is related to the energy transfer $\hbar\omega$ by the relation:

$$|\mathbf{q}| = k_{in}\sqrt{2 - \frac{\hbar\omega}{E_{in}} - 2\sqrt{1 - \frac{\hbar\omega}{E_{in}}}\cos\theta} \tag{2.26}$$

This formulation of the problem takes into account the fact that in scattering phenomena neutrons more easily loose energy in interacting with matter. Equation in (2.26) is of crucial importance for the determination of the accessible range in the experimental settings of quasielastic neutron scattering measurements.

Linear response theory allows the dynamic structure factor to be written in terms of equilibrium fluctuations of the sample and thus, using the fluctuation-dissipation theorem [201], $S(q, \omega)$ can be written as the Fourier transform of the intermediate scattering function $F(\mathbf{q}, t)$:

$$S(\mathbf{q}, \omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dt F(\mathbf{q}, t) e^{-i\omega t}, \tag{2.27}$$

$$F(\mathbf{q}, t) = \frac{1}{N}\sum_{\alpha,\beta}\Gamma_{\alpha\beta}\int_{-\infty}^{+\infty} G_{\alpha\beta}(\mathbf{r}, t) e^{-i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r} = \frac{1}{N}\sum_{\alpha,\beta}\Gamma_{\alpha\beta}\left\langle e^{i\mathbf{q}^T\cdot\mathbf{r}_\alpha(t)}\ e^{-i\mathbf{q}^T\cdot\mathbf{r}_\beta(0)}\right\rangle \tag{2.28}$$

$$\Gamma_{\alpha\beta} = \overline{b_\alpha b_\beta} + \delta_{\alpha\beta}(\overline{b_\alpha^2} - \overline{b_\alpha}^2) \tag{2.29}$$

where "$\langle \cdots \rangle$" denotes an ensemble average and if $A(t)$ and $B(t)$ are two time-dependent functions, "$\langle A(t_1)B(t_2)\rangle$" is usually referred to as a *correlation function*. The parameter $b_{\alpha/\beta}$, called the *scattering length*, is an effective linear dimension of the nucleus α/β with respect to its interaction with neutrons. The function $G_{\alpha\beta}(\mathbf{r}, t)$ in (2.28) is referred to as the Van Hove pair correlation function and represents the probability that, given a particle at the origin at time $t = 0$, any particle (including the same one) is at $\mathbf{r}$ at time t.

The intensities of the neutron-nucleus interactions are defined by the scattering lengths b of each nucleus in the sample which depend on the isotope and the relative orientation between the spin of the two interacting particles. If the spins of the nuclei and the neutron are not maintained in a special orientation one can assume a random relative orientation and that spin and position of the nuclei are uncorrelated.
For this, $F(\mathbf{q}, t)$ in (2.28) takes into account the random distribution of nuclear spins in the sample through the parameter $\Gamma_{\alpha\beta}$ defined in (2.29) where the average over isotopes and relative spin orientations of neutron and nucleus is expressed by $\overline{\cdots}$.
The parameter $\Gamma_{\alpha\beta}$ can be divided into two terms:

$$\Gamma_{\alpha\beta} = b_{\alpha,coh} b_{\beta,coh} + \delta_{\alpha\beta}(b_{\alpha,inc} b_{\alpha,inc}) \tag{2.30}$$

where,

$$b_{\alpha,coh} = \overline{b_\alpha} \tag{2.31}$$

$$b_{\alpha,inc} = \sqrt{\overline{b_\alpha^2} - \overline{b_\alpha}^2} \tag{2.32}$$

are linked to the total scattering cross section by the relation:

$$\sigma_\alpha = 4\pi(b_{\alpha,coh}^2 + b_{\alpha,inc}^2).$$

The outcome of equation (2.30) is also that $F(\mathbf{q}, t)$ can now be recast as a sum of two parts: the *coherent* part $F_{coh}(\mathbf{q}, t)$ which results from the correlations in time between the positions of the atom α (autocorrelation term) and those of different atoms β (cross-correlation term); the *incoherent* part $F_{inc}(\mathbf{q}, t)$ which results only from the time auto-correlation of the positions of the same atom α:

$$F_{coh}(\mathbf{q}, t) = \frac{1}{N} \sum_{\alpha,\beta} b_{\alpha,coh} b_{\beta,coh} \langle e^{i\mathbf{q}^T \cdot \mathbf{r}_k(t)} \; e^{-i\mathbf{q}^T \cdot \mathbf{r}_j(0)} \rangle \tag{2.33}$$

$$F_{inc}(\mathbf{q}, t) = \frac{1}{N} \sum_{\alpha} b^2_{\alpha,inc} \langle e^{i\mathbf{q}^T \cdot \mathbf{r}_k(t)} \; e^{-i\mathbf{q}^T \cdot \mathbf{r}_j(0)} \rangle \tag{2.34}$$

2.2.1 *Incoherent Scattering*

Neutron scattering from biological samples is mainly dominated by the incoherent part of the scattering of the hydrogen nucleus. This fact is due to the large difference between the incoherent term from hydrogens and the incoherent and coherent terms from the other isotopes that are usually present in biological samples. This fact is clearly shown in Table 2.2.1

Isotope	b_{coh}(fm)	b_{inc}(fm)	σ_{coh}(barn)	σ_{inc}(barn)
^{1}H	-3.74	25.27	1.76	**80.27**
^{2}H	6.67	4.04	5.59	2.05
^{12}C	6.65	0	5.56	0
^{14}N	9.37	1.98	11.03	0.49
^{16}O	5.80	0	4.23	0
^{32}S	2.80	0	0.99	0

Table 5: Incoherent and coherent scattering lengths of the most common isotopes found in biological samples. Data reported from [175].

The characteristic cross section of hydrogens is mainly due to a large incoherent contribution of the most favorable state of the interacting system composed by the incident neutron and the hydrogen nucleus. From a practical point of view, this fact constitutes a significant advantage in studying biological samples because they present large amounts of hydrogens uniformly spread over their chemical components. Therefore, the scattering from samples will represent the dynamics of the whole molecule even though they are mainly due to only one type of atom. Moreover, the great difference between the cross section of hydrogen (^{1}H) and the one of deuterium (^{2}H) allows some contrast to be created between different parts of the same molecule or between the molecule and the aqueous content of the surrounding solvent yielding a better characterization of the structure and dynamics of the

sample itself.

Elastic and Quasielastic Incoherent Scattering

A typical spectrum from neutron scattering measurements is shown in Figure (2.2.1) where one can easily distinguish three different regions: the *elastic peak* which results from scattering without exchange of energy between the interacting particles and gives insights into the structural configuration of the sample; the *quasi-elastic* part which results from small energy changes and describes all the stochastic dynamics of atoms in the sample, such as rotational and translation diffusion; the *inelastic* part which is present only if the energy change is sufficient to modify the equilibrium state of the sample and gives access to vibrational motions.

As mentioned in the previous paragraph, the large quantity of hydrogens in proteins or DNA lets us perform the following approximation:

$$F(\mathbf{q}, t) \approx \ N \ b^2_{H,inc} \ F_H(\mathbf{q}, t) \tag{2.35}$$

where

$$F_H(\mathbf{q}, t) = \frac{1}{N} \sum_{\alpha \in \{H\}} \langle e^{i\mathbf{q}^T \cdot \mathbf{r}_\alpha(t)} \ e^{-i\mathbf{q}^T \cdot \mathbf{r}_\alpha(0)} \rangle \tag{2.36}$$

The relative dynamic structure factor can be written using the relation (2.27)

$$S_H(\mathbf{q}, \omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt F_H(\mathbf{q}, t) e^{-i\omega t}, \tag{2.37}$$

and it is the main part of the dynamic structure factor that can be directly measured by neutron scattering experiments.

In the case of samples with confined internal motions, such as biological macromolecules, the intermediate scattering function defined in (2.36) can be decomposed into a time-dependent part and a time-independent one:

$$F_H(\mathbf{q}, t) = \ F'_H(\mathbf{q}, t) + F_H(\mathbf{q}, \infty) \tag{2.38}$$

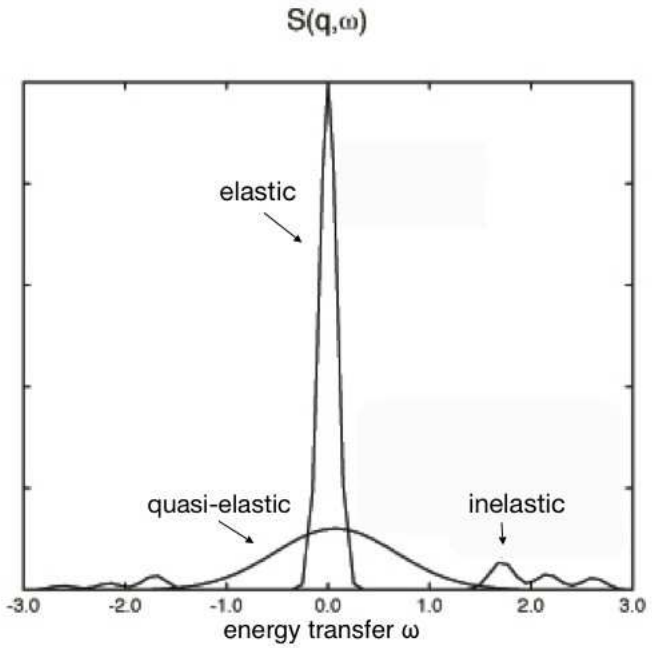

Figure 10: A typical spectrum from neutron scattering measurements.

This is due to the fact that the atom positions at different instants become independent of time as the latter goes to infinity.
In simple liquids, where any confined motion is present, the time-independent term is strictly zero because of the atomic brownian motion which contributes in totally de-correlating atom positions, whereas in proteins, where the atomic internal motions are confined, the time-independent part is a non-zero constant.
The time-independent part of (2.38) is usually called *Elastic Incoherent Structure Factor* (EISF) and can also be used to redefine the dynamic structure factor in the following manner:

$$S_H(\mathbf{q},\omega) = EISF(\mathbf{q})\delta(\omega) \; + \; S'_H(\mathbf{q},\omega) \tag{2.39}$$

In this form, $S_H(\mathbf{q},\omega)$ presents an elastic and $\delta(\omega)$-shaped component related to the EISF along with another part, $S'_H(\mathbf{q},\omega)$ issued from the quasi-elastic and inelastic spectra of the sample. Moreover the two parts on the right-hand side of (2.39) are restrained by the relation $\int_{-\infty}^{+\infty} S_H(\mathbf{q},\omega)d\omega = F(\mathbf{q},o) = 1$ (with appropriate normalization) which leads to:

$$\int_{-\infty}^{+\infty} d\omega S'_H(\mathbf{q},\omega) \; + \; EISF(\mathbf{q})\delta(\omega) = 1 \tag{2.40}$$

By definition, the EISF relates to the ensemble of positions that the scattering nuclei (hydrogens) can attain during an infinite time and it gives insights into the configurational space volume the latter can explore.

In practice, the definition of EISF given above suffers from two main biases: from the experimental point of view $\delta(\omega)$ cannot be obtained exactly because of the finite resolution of the instrument used for the measurements and it typically assumes a larger-width shape (usually fitted as a triangular, gaussian or a lorentzian distribution); from a numerical point of view, the calculation of the EISF as the limit to infinity of the intermediate scattering function is very difficult due to the very low statistics attainable for time scales close to the total time-window used to observe the sample dynamics.
Nevertheless, one can easily overcome these problems by redefining the EISF in both cases: the measured EISF can be rewritten as the ratio of the elastically scattered intensity integrated over the frequencies ω to the total integrated intensity (the sum of the elastic and the quasi-elastic part):

$$EISF_{mes}(\mathbf{q}) = \frac{\int_{-\infty}^{+\infty} d\omega \;\; S_{mes}^{elastic}(\mathbf{q},\omega)}{\int_{-\infty}^{+\infty} d\omega \; S_{mes}(\mathbf{q},\omega)} \tag{2.41}$$

where $S_{mes}(\mathbf{q},\omega) = \int_{-\infty}^{+\infty} S_{mes}(\mathbf{q},\omega')R(\omega-\omega')d\omega'$ takes into account the effects of the finite experimental resolution represented by the function $R(\omega)$ of half-width-at-half-maximum (HWHM) equal to Γ ; the numerical EISF can be written instead as:

$$EISF_{num}(\mathbf{q}) = \frac{1}{N}\sum_{\alpha} b_{\alpha,inc}^2 \langle|exp(i\mathbf{q}\cdot\mathbf{r}_\alpha)|^2\rangle \tag{2.42}$$

2.2.2 *Spectrometers*

The choice of the type of spectrometer to be used to perform the measurements and its resolution is highly related to the time and length scales of interest. In case one wants to

investigate the effects of pressure or temperature on the internal dynamics of biological samples such as proteins, which requires the use of samples in liquid solutions, the spectrometer resolution should be attentively set to reduce the contribution of the global diffusion of the sample to the measured signal.

Time-of-flight spectrometers are typically used for measurements of quasielastic scattering spectra in solids, liquids and molecular crystals.

Time-of-flight spectrometers

In time-of-flight spectrometers, such as the one illustrated in 2.2.2, neutrons from the reactor strike a sequence of choppers: passing through them the beam is first pulsed and then selected with respect to its energy E_0 and wave vector k_0. Therefore the neutrons leave the last chopper, placed at a known distance d_{CS} from the sample, as a pulsed monochromatic beam. An array of ^{3}He detectors is arranged at a known fixed distance d_{SD} from the sample, and scattered neutrons arrive at the detectors at times determined by their scattering energies. The *time-of-flight* of a neutron from the last chopper to one of the detectors, the incident

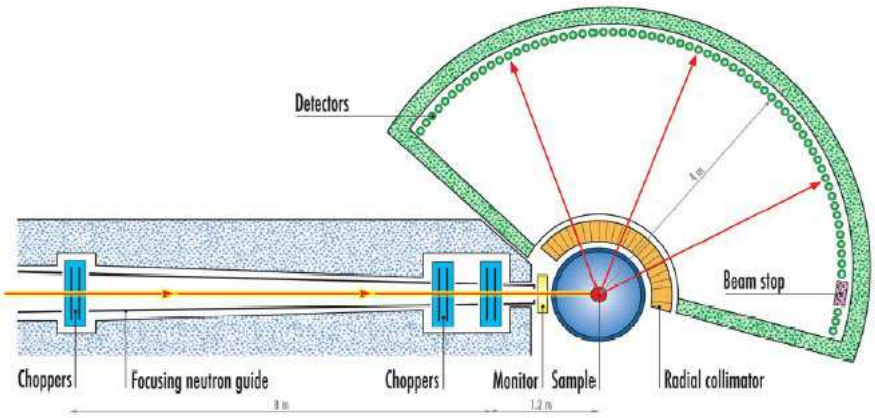

Figure 11: Example of time-of-flight spectrometer.

neutron energy and the distances d_{CS} and d_{SD} are directly obtainable from measurements and allow the calculation of the final energy of scattered neutrons. Moreover, the knowledge of the angles of scattering allows to solve spectra as a function of momentum and energy transfer $\mathbf{q}$ and $\hbar\omega$.

The *time-of-flight* spectrometers typically measure energy transfers in the range $10 - 10^{-2} meV$, hence they are used to investigate dynamics that occur in the time range $10^{-10} - 10^{-13}$s.

All *time-of-flight* measurements for this work have been performed with the following spectrometers:

- FOCUS (Paul Scherrer Institut, Zurich, Switzerland) which is a time-of-flight spectrometer with a variable incident wavelength from 2 to 15 Å. Its energy resolution varies from $50 \mu eV$ to $300 \mu eV$ in the setup with higher incident neutron flux and a maximum momentum transfert of $8 Å^{-1}$.

- IN6 (Institut Laue Langevin, Grenoble, France) which is a time-of-flight spectrometer with incident wavelength of 4 to 5.9 Åwith corresponding energy resolution from 50 μeV to 170 μeV. The maximum momentum transfert available is 2.6 $Å^{-1}$.

3

EXPERIMENTAL AND SIMULATED SYSTEMS SETUP

3.1 SAMPLE PRODUCTION

3.1.1 *Protein expression and purification*

In the following we will discuss the methods used to produce samples for both neutron and *in silico* experiments.
Both approaches demand a wide knowledge of techniques some of which have been already presented in the previous chapters. The experimental samples also needs some preliminary steps for their own production which require knowledge in molecular biology. The discussion of these aspects in detail would be beyond the scope of this thesis, nevertheless, this part of the work has been fundamental for the following experimental steps as it allowed us to setup a stable protocol whose yield was adapted to neutron experiments.
The protocol reported here followed the usual scheme for protein production which can be summarized in a very simplified way, as follow:

- **Cloning** of the part of the genomic DNA from both *Methanococcus Jannaschii* and *Saccaromyces cerevisiae* which encode the a/eIF6 proteins.
- Massive **expression** of the identified gene into an host organism, here *E.coli*, to produce large amount of proteins.
- **Purification** of the protein solution obtained from the host organism to obtain an *as much as possible* pure solution containing only the a/eIF6 proteins.

CLONING Two cloning experiments have been performed to produce the genes encoding eIF6 and aIF6 proteins. Following the nomenclature found in literature we will call TIF6 the gene that encodes for eIF6. For *M. jannaschii*, the genomic DNA fragment encompassing the aIF6 gene was obtained from TIGR/ATCC Bult et al. [18]. The resulting cDNA was then amplified by PCR amplification using the nucleotides showed in Table (6).

For *S. cerevisiae*, TIF6 construct was produced by PCR amplification using genomic DNA from *S. cerevisiae* as a template. Cloning experiments were performed using the nucleotides

Table 6: default

aIF6-forward	AAA CAT ATG ACC ATG ATT ATA AGA AAA TAC TTC TC
aIF6-reverse	TTT TGC GGC CGC TCA TTA AAT CAG GCC TAA AGC ATC TT.
TIF6-forward	CGG GAT CCC ATA TGGCTA CCA GGA CTC AA
TIF6-reverse	GGG AAT TCC TAT GAG TAG GTT TCA ATC AA

showed in Table (6).
The forward primers introduced a **NdeI** restriction site and the reverse primers introduced a **NotI** restriction site. The aIF6 and TIF6 PCR products were cloned into a pET28a expression vector (Novagen). This plasmid contains a kanamycin resistance gene for selection of transformed cells, and a pBR322 replication origin. The target gene is placed under the control of the T7 RNA polymerase promoter, and is expressed as a fusion with a N-terminal hexahistidine tag followed by a thrombin cleavage site. Escherichia coli strain BL21(DE3) (Stratagen) chemically competent cells were transformed with this plasmid in order to express the protein. The DE3 gene encodes for the T7 RNA polymerase (under the control of the galactose operon).

Transformation was done by heat shock: 1ng of plasmid DNA was mixed with 100µl of competent cells, and then incubated on ice for 20min. Following adsorption of the plasmid onto the cell membrane cells are placed at 42°C for 45 sec, and then on ice for 2 min. 500µl of SOC medium is added and the cells were incubated 1 h at 37°C. To select cells which have been properly transformed, the culture was then plated on LB-agarose plates containing kanamycin (10 mg/l).
The plates were incubated at 37°C overnight. The next day 20 ml of LB containing 10 mg/l kanamycin were inoculated with 1 colony. The cells were grown until the OD600 reached 0.6 AU. 1 ml of the cell suspension was complemented with 100µl of sterilized glycerol, and the mix was frozen by immersion into liquid nitrogen. The aliquot was stored at -80°CC.

PROTEIN EXPRESSION The expression was carried out in Erlenmeyer flasks. A 50ml preculture was prepared on the previous day by inoculating 50 ml LB (with 10 mg/l kanamycin) with cells scraped from the top of the frozen stock. For cultures in Erlenmeyer flasks, 2.5 l flasks were filled with 1 l of LB media complemented with 10 mg/l kanamycin.

Each flask was inoculated with 10 ml preculture, then incubated at 37°C with 200 rpm agitation until the OD600 reached 0.6-0.8 AU. The temperature was then decreased to 23°C for 30 min, and 0.5 ml of isopropyl-thiogalactosidase (IPTG) 1 M (0.5 mM final) was added to induce expression of the cloned protein (via induction of the T7 RNA polymerase). Incubation was continued in the same conditions for 5 h. The cells were pooled into 1 l bottles and harvested by centrifugation for 15 min at 5000 g. Cell pellets were frozen by flash cooling in liquid nitrogen and stored at -80°C.

PROTEIN PURIFICATION For each protein preparation 10 g of frozen cellular paste were resuspended in 40 ml of lysis buffer + 40mg of lysozyme (Appendix A). 3 pills of antiprotease complete EDTA free cocktail and 4 µl of Benzonase enzyme were added. The bacterial suspension was incubated 20 min on ice with agitation. Then, the cells were lysed twice by cell disruption (Constant System) at 1.4 Kbar. 1mM β-mercaptoethanol was added to the solution. Here two different approaches were used for aIF6 and eIF6: *i)* the crude extract expressing aIF6 was incubated for 1 hour at 75°C taking advantage of its heat resistance properties; *ii)* the eIF6 was was rapidly brought to the next step to reduce the probability of proteolytic clevage (see next paragraph). The soluble and insoluble fractions of aIF6 and eIF6 were separated by ultra-centrifugation (1.5 h, 250000 g) for both preparations. The soluble part of the crude extract was then incubated with 3 ml Talon superflow Cobalt affinity resin (Clontec) previously equilibrated with lysis buffer; incubation was done at 4°C for 1.5 h with gentle agitation. The resin was transfered in a column and washed successively with 10 column volumes of wash solution (Appendix A, 1M NaCl) by gravity. The histidine tagged IF6 protein was then eluted from the resin using 8 volumes of elution buffer (Appendix A, 250 mM imidazole), and collected in 1.5 ml fractions. Fractions were loaded on a denaturing gel to check for presence of the overexpressed IF6 protein.

Fractions containing the protein were pooled and dialysed twice for 2 h against imidazole free, dialysis buffer (Appendix A). The protein concentration was determined by measuring the OD280 of the solution and thrombin was added (1 u per 150 µg of protein) to selectively cleave the N-terminal histidine tag of the protein. The cleavage step was done by incubation overnight (approx. 16 h) at room temperature, and stopped by addition of 0.2 mM (final concentration) phenylmethylsulfonyl-fluoride (PMSF). The completeness of the proteolysis was checked by SDS-PAGE.The protein was concentrated to 10 mg/ml using a UltraFree (Millipore) ultrafiltration device (5 kDa membrane cutoff) prior to size exclusion

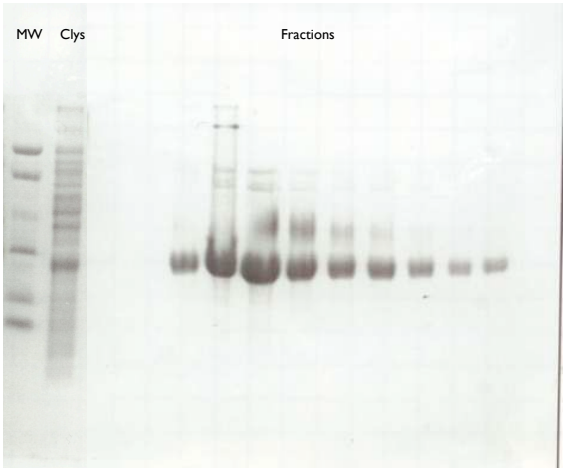

Figure 12: Denaturing gel verification of overexpressed IF6 protein. Cell lysate shown as reference.

chromatography. Gel filtration was performed using a Superdex-75 HR 10/30 (Pharmacia) column previously equilibrated in the protein storage buffer (appendix 1). Both OD280 (optical density at 280nm) and OD260 were monitored during the chromatography to control the absence of nucleic acids and 500 µl fractions were collected. Peak fractions were pooled and concentrated a last time using a Centricon (Millipore) ultra-filtration device with a membrane cutoff of 10 kDa.

POLY-HISTIDINE TAG The use of polyhistidine tags helped to obtain a considerable yield from the above protocol which allowed a reliable protein concentration to be maintained even for neutron scattering measurements in high pressure systems where a large solution volume is necessary. Nevertheless, for technical reasons the selective cleavage of this tag was not suitable for large volume solutions. For this reason, the cleavage procedure explained above was performed only on ambient pressure measurements where more convenient volumes could be used. As a consequence, this protocol must be considered as a first effort toward a more satisfactory procedure adapted to neutron scattering experiments.
The presence of the histidine tag is not expected to have strong effects on protein structures, as proved by Carson and coworkers [23] on a wide set of crystallographic structure, but it could significantly affect the global protein dynamics. For this reason, as shown in Results, some supplemental tests have been performed to quantify the dynamical contribution of the tag in the case of a/eIF6.

CLEVAGE OF CARBOXYL-TERMINAL TAIL IN EIF6 During some preliminary tests for protocol optimization, a proteolytic clevage of the eIF6 protein was observed (see Figure 13 for the corresponding SDS-PAGE verification), either in the soluble and in insoluble fractions of the cell lysates.

Figure 13: SDS-PAGE verification of partial clevage in eIF6 samples. Cell lysate shown as reference.

Due to its apparent small molecular weight, the cleaved fragment detected in this work was supposed to be the same found by Groft and coworkers [66]. The latter reported that attempts to express and purify *S.cerevisiae* eIF6 were complicated by proteolytic cleavage of the divergent C-terminus. They also remarked that truncation beyond residue 224 eliminated this problem, and eIF6(1-224) could be purified using the same aIF6 purification scheme.
Interestingly, the cleaved part of eIF6 contained a tail made of 21 aminoacids, with sequence shown in Table 7, which can be found only in eukaryotes homologous of IF6 and very little is known about its structure.

A preliminary characterization of the fragment was performed by the means of the Basic Local Alignement Search Tool (BLAST) which compares the aminoacid sequence of the fragment against sequence databases in order to find some similarities with other known proteins. For eIF6, the only significant scores were obtained from other IF6 homologous meaning that this sequence fragment must be strictly peculiar for the function of this class of

Table 7: Sequence of the C-terminal fragment of eIF6 (CTAIL). Numbers of residues start at 225, according to the sequence in eIF6.

Glu225	Asp226	Ala227	Gln228	Pro229	Glu230	Ser231	Ile232
Ser233	Gly234	Asn235	Leu236	Arg237	Asp238	Thr239	Leu240
Ile241	Glu242	Thr243	Tyr244	Ser245			

initiation factor or of its evolutionary history.
In order to obtain more insights into the *native* conformation of this C-terminal tail, a test of secondary structure prediction was performed with two distinct methods:

- PSIPRED server [101], which performs structure predictions based on position-specific scoring matrices
- APSSP2 server [156], which predicts secondary structure conformations using nearest neighbor and neural network approaches.

Both methods found a reasonably probable formation of a α-helix in the ending part of the fragment, in the region Asp238 Thr239 Leu240 Glu241. These results are also corroborated by the inspection of the hydrophobic profile of the fragment with the help of the Kyte/Doolittle hydrophilicity scale [122] which shows an increasing hydrophobicity in the regions Pro229-Ser231 and Asp238-Glu242.
These results lead to the convinction that the fragment could have a significant role either in dynamics or structure stability of eIF6 and thus also in protein function. This idea is comforted by recent results which indicate that C-terminal subdomains contribute to the localization in the cellular nucleus of eIF6 [6].
This outcome suggested the importance of maintaining the 21 C-terminal fragment as part of the investigated sample. For this purpose, several tests on the production of eIF6 were performed in order to limit the proteolytic cleavage during the production itself without effect on the total final yield.

The protocol presented here gave the best results with a cleavage reduced to around 30% as resulted from the preliminary MALDI mass spectrometry analysis reported in Figure 14 where two large peaks are evident at molecular weights equal to both the entire protein and the truncated structure (1-224). The height of these peaks gives an estimation of the relative

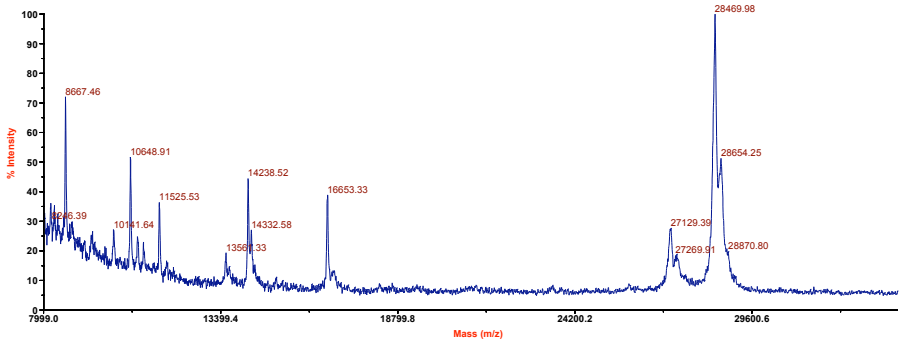

Figure 14: MALDI-MS essay on eIF6 solution.

amount in solution.

STABILITY OF EIF6 When performing studies on the effects of extreme conditions on protein structure and dynamics it is important to know beforehand the boundaries which define the normal conditions for the specific protein of interest.

Nevertheless, in the case of IF6 homologues, there is a scarcity of works on the general chemical properties with respect to temperature and pressure changes. Although aIF6 is presumably able to reversibly respond to high temperature and high pressure values, very little can be said about the eIF6. In an early paper on biological properties of eIF6, Valenzuela and coworkers [199] reported that anti-association factor 6 from calf liver was found to undergo a cessation of its activity at temperature of $\sim 60°C$. As the inactivation of a protein function may not correspond to the total and irreversible denaturation of the protein itself, the latter was tested here by means of dynamic light scattering measurements in the range of interest of the present work. The measurements of the hydrodynamic radius of eIF6 were performed with a DynaPro-Titan© fixed-angle light scattering system. The results, listed in Table 8, shows that, at 50°C, eIF6 forms aggregates or partially unfolds. Moreover, when brought back to 30°C it does not find its initial state, suggesting that the transition is not reversible.

These preliminary reults together with the outcome of the work made by Valenzuela and

coworkers, suggested to limit all neutron scattering experiments on eIF6 into the non-denaturating temperature range. However, as explained in the next sections, some *in silico* exmperiments have been performed on eIF6 also at high temperature and high pressure to verify the presence or more simply the beginning of a denaturating process. It is also worth noting that the reported values of hydrodynamic radius for eIF6 at 20°C correspond only to an effective value due to a spherical approximation of the whole protein structure: the conserved pseudo-globular part formed by residues 1-225 and the C-terminal tail of 21 aminoacids.

Table 8: Variation of the hydrodynamic radius of eukaryotic eIF6

	20°C	30°C	50 °C	back to 30 °C
R_H [nm]	3.32	3.61	29.76	24. 23

FINAL SAMPLES For both aIF6 and eIF6, deuterated protein solutions were prepared at a concentration of $\sim 40\,mg/ml$ and pD $\sim$ 7.0. All labile hydrogen atoms in the samples were exchanged overnight by dialysis and then filtered with an Amicon Filter Ultra (membrane cutoff 5kDa) against a thirty-fold excess of pure deuterated solvent. The final solution was centrifuged to eliminate possible aggregates. The final concentration was measured by UV-VIS absorption at 280nm. The latter measurement gave results with a possible systematic error greater than 10% due to the very low amount of chromophore amino acids (such as tryptophan, tyrosine, phenylalanine and histidine) in the sequence of both proteins. Although the limits for solution concentration of both samples were not known from literature, several preliminary DLS assays seemed to comfort the observation of monodisperse solutions of eIF6 around 40mg/ml.

3.2 NEUTRON SCATTERING MEASUREMENTS SETUP

In this work the reported measurements of neutron scattering spectra were performed in order to investigate the effects of the pressure and temperature changes into the dynamics of

proteins. For quasielastic neutron scattering experiments (QENS), the easiest way to apply pressure to a biological sample is to put it in a liquid solution and then compress the volume in which it is contained. This is usually done in sample containers with a cylindrical geometry. In this work, two different sample containers of this shape have been used for the measurements at ambient and higher pressure, respectively. The relatively wide range of temperature values, explored in measurements, imposed for both sample containers, the choice of materials which do not undergo significant structural changes at high temperature that could modify, for example, their mechanical resistance.
Moreover, when performing QENS experiments on samples in solution, one often needs to use high concentrations to maintain the signal due to the sample itself significantly distinguishable from the one coming from the buffer. For biological macromolecules this fixes a constraint to the total solution volume used in the experiment as the biological samples are often available in small quantity mainly due to expression protocols yields. In the case of a/eIF6, although the protocol presented in the previous section was conceived to obtain a maximum amount of samples, this was still a limiting factor for the total volume available for measurements. For this reason the sample containers used in this work were also chosen for their effective volume.

Ambient pressure An approximate scheme of the sample container used for these measurements is shown in Figure 15 (panel A). The container is made of two concentric hollow aluminum cylinders with a diameter of 19mm and 20mm, respectively. The irradiated region was $\sim 50\text{mm}$ in height. The space between the walls of the two cylinders were filled by the samples and closed on the top by a disc made of Teflon© in order to avoid the sample to exit the space due to capillarity effects. The Teflon© cap was supposed to be out of the irradiated region. The total volume available for samples were limited to 1.2ml.

High pressure The pressure cell was developed at the Institut Laue Langevin in Grenoble (France) and was conceived to carry out experiments on liquid solutions at moderate high pressure. In particular the dimensions of the cylindrical geometry were determined to withstand pressure up to 2kbar: internal diameter of 10mm and wall thickness of 1.5mm. The global geometry of the cell, shown in Figure in 15 (panel B), was inspired by another one previously used for high pressure studies on lysozyme and reported in [20].

In contrast to the latter, however, here the pressure is applied without transmitting media and directly on the sample which is compressed by a pump connected to the pressure cell by a very thin capillary (0.1mm diameter). Moreover, the irradiated part of the cell was made of

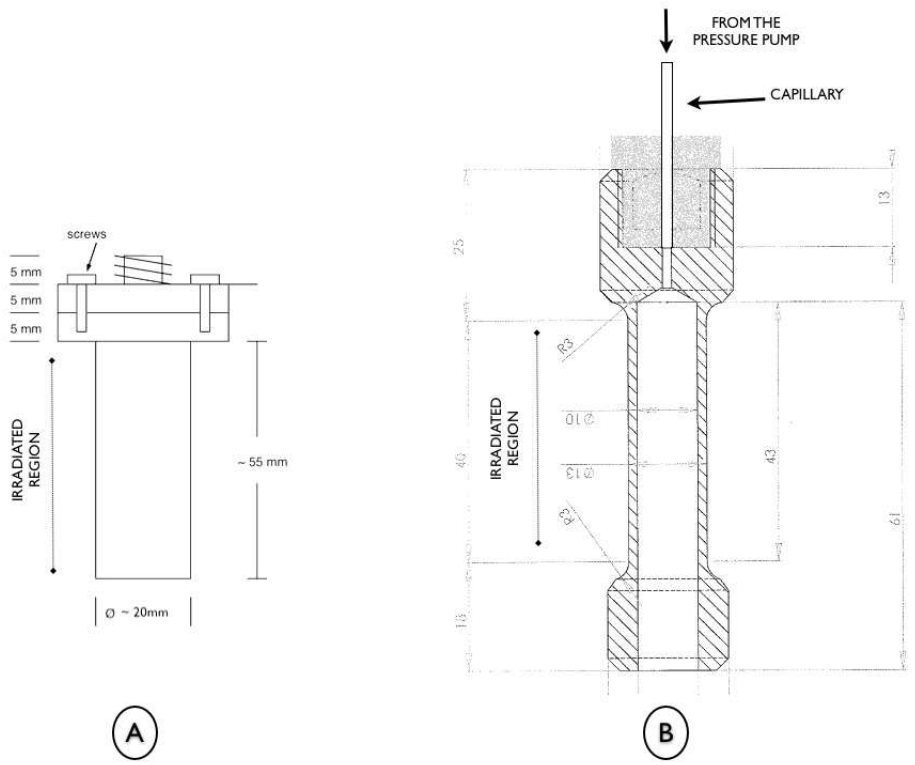

Figure 15: Sample can for high pressure measurements

an alloy of copper-berylium which permitted the wall thickness to be significantly reduced without affect their mechanical resistance. In order to reduce the multiple scattering due to the protein solution an insert with a diameter of 9 mm was used. As a whole the pressure system required a total volume of ~ 3ml of protein solution which is considerably smaller than the volume commonly used in this type of QENS experiments and suitable for the typical amount of a/eIF6 production yields.

INSTRUMENTAL RESOLUTION The QENS measurements reported in this thesis have been performed mainly to investigate the internal dynamics of proteins. For this purpose it was important to filter out of the quesielastic spectra all other contributions coming for example from the global translation and rotation of proteins in solution. The standard method to accomplish this task is the choice of an adequate experimental energy resolution which should be larger than the width HWHM (half-height at half-maximum) of quasielastic signal due to global motions. The energy width of the latter can be estimated from knowledge of the translational and rotational diffusion constant of the protein. Unfortunately in the case

of a/eIF6 no reference can be found in litterature and these values could only be estimated from the available crystallographic structures. Nevertheless, as reported in Results, some *a posteriori* finer estimations have been made by means of molecular dynamics simulations. The values obtained for the conserved structure (residues 1-224) in aIF6 and eIF6 were around $8.4 \cdot 10^{-3}$ $\text{Å}^2/ps$ which correspond, in terms of HWHM, to an energy width of 20 μeV for a value of $q_{el} = 1\text{Å}^{-1}$. This result should be however corrected by a factor 0.20 due to the higher viscosity of D_2O with respect to that of H_2O. The experimental resolution chosen for measurements at the spectrometers FOCUS (Paul-Sherrer Institut) and IN6 (Institut Laue Langevin) reported with other experimental settings in Table 9, was not sufficient to totally eliminate the effects of global motions to the protein spectrum but was enough to significantly reduce it with respect to the quasielastic signal corresponding to internal diffusive motions. In this discussion only the translational diffusion constant was taken into account. To estimate the influence of rotational diffusion on the measured QENS spectra the following formula for the diffusion constant for rotational diffusion [36] was used:

$$\gamma_r = \frac{k_B T}{4\pi\eta R_H^3}. \tag{3.1}$$

The above relation is used assuming that the approximation of eIF6 as a spherical-shaped protein is reliable[1]. Here R_H is the radius of the protein and η is the shear viscosity of the solvent (water). For a/eIF6, which has a radius of $R_H = 2.81\,\text{nm}$, one obtains $\gamma_r = 0.14 \cdot 10^8\ s^{-1}$ at $T = 293K$. This corresponds to a width of $6 \cdot 10^{-5}$ meV, which is far below the instrumental resolution. Therefore, spectral contribution from the global rotational diffusion are expected to be largely within the chosen energy resolutions.

Table 9: Instrumental settings used in this work.

	Incident waveleght [Å]	Energy resolution HWHM [meV]	Sample environment	Time per run [hours]
FOCUS	5.92	~ 0.020	ambient pressure	15
IN6	5.98	~ 0.020	high pressure	~ 10

1 In the Results, I will show that a better approximation of IF6's shape can be given by the Perrin correction to the Stoke's law for the diffusion of a sphere.This refinement will not change however the considerations made here.

3.2.1 *Data analysis*

The usual procedure for the analysis of data from neutron scattering measurements, is preceded by some preliminary as follows:

- Normalization to the total number of incident neutrons during a single measurement. This normalization is needed to account for the differences in time length and in incident neutron flux for each measurement

- Correction for the detector efficiencies. It is performed in two steps. Firstly by the normalization to the spectrum of vanadium performed with the same experimental environment which corrects for relative efficiencies of detectors with respect to each other. Secondly a correction is made with respect to the energy dependent efficiency of detectors. A vanadium spectrum is used for the first calibration because it is a completely elastic incoherent scatterer in the range of q of interest in time-of-flight experiments.

- Grouping of spectra over angles is made to increase the statistics related to data. The latter procedure was crucial because the signal-to-noise ratio in measurements of biological samples in solution is usually very low.

In QENS measurements on protein solutions, the detected signal does not come exclusively from the protein but shows also significant contribution from the bulk solvent and from the sample environment, like for example, the sample containers and background noise. The spurious contributions are usually subtracted from the total signal to obtain scattering functions related only to the protein dynamics. This procedure can be summarized by the following relation:

$$\tilde{S}_{protein}(q,\omega) = (S_S(q,\omega) - \tau * S_E(q,\omega)) - (1-\alpha) * (S_B(q,\omega) - \tau * S_E(q,\omega)) \quad (3.2)$$

where τ takes into account the transmission of the protein solution[2] and α is the volume fraction of the protein and its first hydration shell in the solution sample.The latter must be known to apply a correct subtraction of solvent contribution in protein solution spectra.
The parameter α was estimated from the following relation between the mass density of proteins and their molecular weight ($M(KDa)$) [54]:

$$\rho(M)[g/cm^3] = \left[1.41 + 0.145 \cdot \exp\left(-\frac{M(KDa)}{13}\right)\right] \quad (3.3)$$

2 Through out this thesis the transmission of the protein solution was approximated to the one of the buffer alone

For a/eIF6 one obtains $\rho = 1.43 g/cm^3$, which gives a specific volume $\nu_{prot} = 0.70 cm^3/g$. Multiplying by the estimated concentration of $40 mg/ml$ one has the volume fraction $\alpha = 0.028$. The hydration shell needs to be included as part of the protein because the dynamics of water molecules in this shell is different from the one of bulk water due to the weak interactions occurring between the protein and solvent molecules.
Gerstein and Chothia [63] showed that hydration water volume is equal to 24 $Å^3$ which means it is 20% smaller than the volume of bulk water. Hence, the increase of protein specific volume due to its first hydration shell can be estimated as follows:

$$\nu = \nu_{prot} + \nu_{shell} N$$

where N is the number of protein molecules in the unit volume and

$$\nu_{shell} = r_{shell} \cdot A_{water}$$

In the above equation r_{shell} represents the radius of the spheres with the same volume of water molecules in the first hydration shell and A_{water} is the surface area of the protein accessible to solvent molecules with radius r_{shell}. For lack of more precise estimations, in the case of IF6, the value of ν_{shell} was approximated with the Surface Accessible Surface Area (SASA) calculated on the crystallographic structure of aIF6. The obtained value corresponds to a hydration layer composed by approximately 850 water molecules per protein molecule. Hence, the final volume fraction α was found equal to 0.04.
As stated at the end of the previous section, the measurement of the protein concentration in the final samples was made very difficult by the poor UV-VIS absorption of a/eIF6. As a consequence, all values obtained from the considerations made above, which are strictly dependent to the knowledge of the real protein concentration, have to be considered here as strong approximations and they will be used only as references in the Results.

3.3 MOLECULAR DYNAMICS SETUPS

We performed all the stages of the molecular dynamics simulations using the AMBER9 simulation code[27]. The whole set of final simulations for both the homologues used the AMBER99SB force field [86]. The latter is an upgraded version of the AMBER94 force field, usually used for protein molecular dynamics, and it contains a reparametrization of the back-

bone torsion terms and achieves a better balance of the different secondary structure elements.

3.3.1 *System Setup*

aIF6

The initial configuration of the IF6 protein issued from the *M.Jannaschii* was taken from a crystal structure with a refined resolution of 1.30 Å(available from the Protein DataBank with the code: 1G61)[66] together with all the water molecules found within a distance of 2nm from the protein center of mass in the crystallographic data. The coordinates of the missing hydrogen atoms were added using the algorithms implemented in the LEaP program from the AMBER9 package.

The protein was placed in a orthorhombic periodic box filled with water molecules parametrized as TIP3P. The crystallographic water represented by oxygen atoms, within a distance of 2nm from the protein center of mass, were replaced by the same type of water molecule models. The final total amount of water molecules was 8136. The whole system includes also 14 sodium (Na+) counterions in order to obtain a neutral global charge for the Ewald calculations.

The whole system was initially minimized in two steps with combined use of steepest-descent and conjugated-gradient algorithms: *i)* 200 cycles of conjugated-gradient after 4 steps of steepest-descent with position restraint on protein atoms and counter ions; *ii)* 200 cycles of conjugated-gradient after 4 steps of steepest-descent for the whole system without restraint.

As a preliminary step of the *real* molecular dynamics simulation, we first performed a *system equilibration* step in which the simulated system attained a stable equilibrium conformation compatible with the environmental constraints, i.e. constant pressure and temperature. For this purpose, a short simulation of 150ps in a NVT ensemble, i.e. with fixed total volume and temperature kept equal to 300K was performed, followed by a 700ps long simulation in the final NPT ensemble, with $T = 300K$ and $P = 1bar$. In all simulations performed in this work the time-step used for the integration of motion equation was equal to $\Delta t = 0.001ps = 1fs$ (fs is femto-seconds). The equilibration in the NVT ensemble was made with the calculation of the contribution of the forces slowly-varying in space only every

two steps, resulting in a shorter CPU-time without significant changes in the total energy of the system.

From the equilibration at 300K and 1bar all other simulations branched to other NPT conditions of interest. In every case the equilibration steps were followed by *production* simulations of length equal to 2ns which was used for the calculation of dynamical properties of the system itself.

In each simulation the control of temperature was performed with a Langevin thermostat[3] with a collision rate of $3.5ps^{-1}$, whereas pressure was constrained by a Berendsen barostat relaxation time $\tau_p = 1.5ps$.

3.3.2 *eIF6*

The initial configuration of the IF6 protein issued from the *S.cerevisiae* was taken from a crystal structure with a refined resolution of 2.5 Å(available from the Protein DataBank with the code: 1G62)[66]. Nevertheless, the crystallographic data contained in the original PDB file did not give the atom positions of the 21 amino acids long C-terminal tail which was not crystallized with the rest of the proteolytic cleavage protein because of the structural instability that was encountered in this work during the sample production, as mentioned in Section 3.1.1. The presence of this tail, even though it does not seem to influence the function of eIF6 [66], certainly has some effects on protein mobility [6].
The *construction* of the complete structure of eIF6 required some supplementary steps with respect to the case of aIF6. Firstly the structure of the C-terminal tail (CTAIL, in the following) was modeled and partially folded. Secondly the complete structure was assembled and equilibrated to an appropriate equilibrium state.
Both steps of this preliminary procedure were performed by means of molecular dynamics simulations with implicit solvent in order to significantly reduce the CPU-time [198]. This method consist of the substitution of the explicit calculation of the dynamics of solvent molecules -usually very time-consuming - with an additional *mean forces* term in the protein force-field which should take account of all solvent effects on the protein. This approach

3 The chosen collision rate for Langevin thermostat did not produce artifacts on IF6 dynamics as verified by comparison of effective friction constant calculated from NPT ensemble with the one obtained from a NVE (constant energy) simulation of the same system. Friction constant was estimated by memory function calculation[118].

clearly contains some strong approximations and it does not always give the right description of dynamics if compared with explicit solvent calculations [146] but it is however a valuable tool to rapidly explore the configuration space of large systems in order to find an appropriate equilibrium state. The implicit solvent model used throughout this thesis was the pairwise Generalized Born solvation model (GB), developed by Hawkins and coworkers [70, 71], where mean forces are obtained from the estimation of the total solvation free energy of the molecule into water. A complete description of the method can be found in many textbooks on methods in computational physics and biology [12].

CTAIL FOLDING Here a brief summary of the protocol used for the modeling and the initial folding of CTAIL is reported. The whole procedure was performed using the LEaP and Sander programs from the AMBER9 package.

. A "linear" configuration was firstly created for the polypeptide with sequence shown in Table 7.

. Initial folding of the linear structure was then performed in a NVT ensemble with the protocol for the temperature re-scaling shown in Table 10. The time step for the integration of equation of motion was varied between 0.1 and 0.5 fs, in order to reduce the extent of force variation and thus the probability of unnatural atom contacts which would prevent the CTAIL from folding correctly.

Table 10: default

Total time [ps]	Partial time [ps]	Initial T Kelvin	Final T Kelvin	time step [fs]
10	10	0	50	0.1
260	250	50	100	0.5
510	250	100	150	0.5
560	50	150	200	0.5
610	50	200	250	0.5
660	50	250	300	0.5

. The final step of this procedure was a very long equilibration simulation in which the protein fragment could fold. The total time length of this simulation was $\sim 40ns$. A shorter simulation with explicit solvent was also performed to verify the absence of solvent-specific effects in CTAIL folding.

Interestingly, the final folded structure of CTAIL, shown in Figure 16 (Panel A), contains a small helix in the region 14-18 whose shape is similar to an α-type. The inclusion of this helix into the protein sequence corresponds to the one found in Section 3.1.1 by means of structure prediction and hydrophobic arguments.

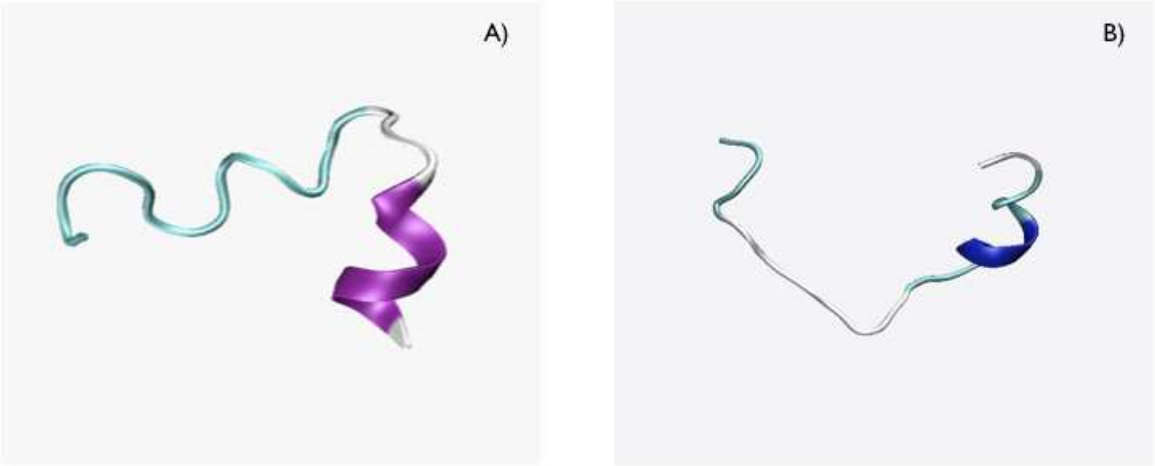

Figure 16: Modeling of CTAIL in implicit solvent. Panel A shows structure folded without contraints. Panel B: structure folded with a position restraint in amino-terminus. Color scheme: purple for α-helix, blue for 3_{10}-helix, cyan for turns and white for random coil.

It is worth noting that the folded configuration obtained here might not correspond to the one that CTAIL would reach when folded together with the rest of the protein sequence. For this reason, the same procedure was repeated with the amino-terminal position fixed in space by an harmonic force, $F = -Kx$, with $K = 10\ kcal\ mol^{-1}\ Å^2$. The resulting final structure was compared with the previous one in order to better understand the effects of the spatial constraints in the folding pathways. The resulting structure after a total simulation time of $40ns$ is shown in Figure (Panel B). Inspection by eye reveals that secondary and tertiary structures are slightly different. This comparison indicate that CTAIL folding is very sensitive

to the number of degrees of freedom available for its three-dimensional configurations.

MODELING OF EIF6'S COMPLETE STRUCTURE The above paragraph showed the importance of an ultimate verification of the actual conformations of CTAIL when the latter is joined to the rest of the protein. A complete study on this subject would be out of the scope of this thesis and would need a more exhaustive simulation of folding of the whole protein. Here I would like to focus on more qualitative considerations about the effects induced on the rest of the eIF6 structure by the presence of CTAIL. For this purpose, in the following it will be supposed that the conserved region of aIF6 and eIF6 (residues 1-224) does not attain a different fold in the presence of CTAIL. This assumption seems to be confirmed as reliable by homology modeling study of different homologues of IF6 from a wide range of different organisms [66].

On these bases, another molecular dynamics simulation with implicit solvent was performed on the system composed by CTAIL directly joined to the rest of "already-folded" eIF6 structure. The initial configuration used for this simulation was composed by CTAIL folded through the procedure described above and the crystallographic structure of eIF6. The binding of the two subunits of the system was performed by the creation of the peptide bond between the nitrogen atom in the C-terminus of the residue 224 and the nitrogen atom in the N-terminus of CTAIL. The latter structure was then minimized with restraints in atom positions in residues from 1 to 200 and with fixed hydrogen covalent bond-length (the SHAKE algorithm[35] for rigid molecule dynamics was used here). The procedure consisted of 900 steps of minimization using the steepest-descent gradient method followed by 1100 steps performed with the conjugated-gradient algorithm.
The energy minimization was followed by a short equilibration run of molecular dynamics (30ps) in a NVT ensemble, i.e. with constant volume and temperature. The Langevin thermostat was used to keep temperature constant. Here, in order to significantly improve the thermal coupling between the system and the thermostat a very high friction constant value ($\gamma = 500\ ps^{-1}$). A short time step ($\Delta t = 0.5\ fs$) for the Leapfrog integrator was used in order to avoid improper fluctuations of potential energy which would result in total energy divergence.

Local optimization of the protein conformation found through the previous steps, was performed using the *simulated annealing* method which consists, as for the real annealing of

matter, in a sequence of heating and cooling phases. The latter, which is commonly used for finding of the global minimum of potential surfaces in the case of small molecules, was shown to not be reliable for systems with a broad distribution of energy scales. Nevertheless, several works demonstrated that it remains very useful method for local optimization of larger molecules conformations, such as proteins, because unlike minimization methods, it is able to locate local minima even far away from the initial conformation[208]. The sequence of annealing steps used in this work is shown in Table 11 and was performed with the same parameters settings as the previous equilibration step.

Table 11: Simulated Annealing of eIF6's complete structure

Total time [ps]	Partial time [ps]	Initial T Kelvin	Final T Kelvin	time step [fs]
50	50	300	400	0.5
100	50	400	400	0.5
150	50	400	500	0.5
200	50	500	500	0.5
250	50	500	300	0.5
300	50	300	300	0.5

Figure 18 shows the time evolution of RMSD for the region 1-224 (hydrogens not taken into account) with respect to the initial equilibrated conformation, expressed as a function of time by the relation:

$$\mathrm{RMSD}(t) = \sqrt{\frac{\sum_i^N (\mathbf{R}_i(t) - \mathbf{R}_i(0))^2}{N}} \tag{3.4}$$

where $\mathbf{R}_i(t)$ represents the position vector of the atom i at time t. In the same figure the evolution of total potential energy during the simulated annealing is shown. Both results give evidence of a new conformation far from the initial one and with a slightly lower energy.The final structure is shown in Figure 17.

It is worth noting that, after minimization and simulated annealing, the root-mean-square-deviation of the eIF6 backbone in region 1-224 with respect to the original crystallographic

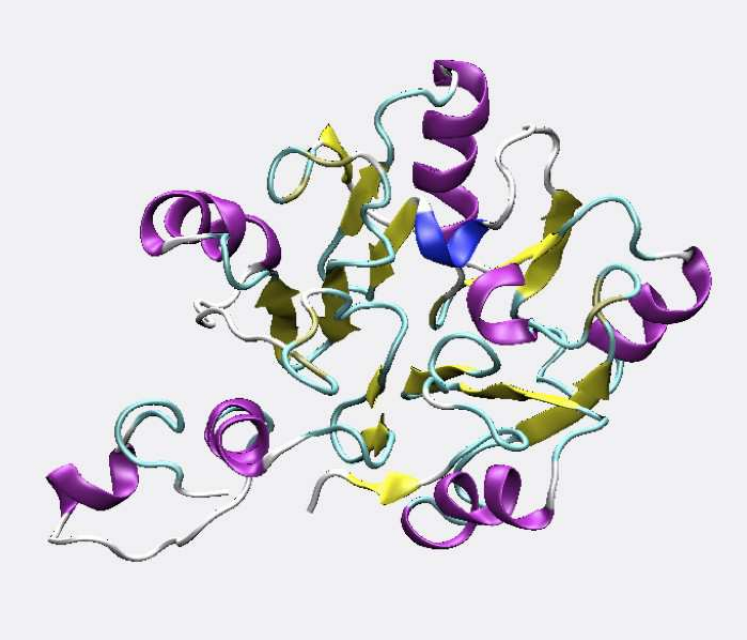

Figure 17: Final structure issued from the minimization and simulated annealing procedures.

structure is however small (1.5Å), meaning that, although the *core* region of eIF6 moved away from its initial minimum, its global arrangement is not significantly changed by the presence of CTAIL, at least on the time scale explored here.

The structure resulting from the sequence of procedures discussed above, was used as initial structure for molecular dynamics simulations performed with explicit solvent and with the same protocol used in the case of aIF6.

SUPPLEMENTAL SAMPLES In order to compare results from molecular simulations with the experimental measurements and to obtain more insights into the dynamical and structural effects of CTAIL, two other samples were modeled and analyzed (they are reported here with the *short names* used through out this thesis to refer to them):

. **eIF6-NoCTAIL**: the eIF6 simulated without the attached CTAIL. The crystallographic structure was used as initial configuration after a preliminary step of minimization to let it attain an equilibrium conformation compatible with the buffer environment. Due to the low resolution of x-ray data, any *crystallographic water* was found in the initial PDB file. All buffer molecules were modeled as made for aIF6.

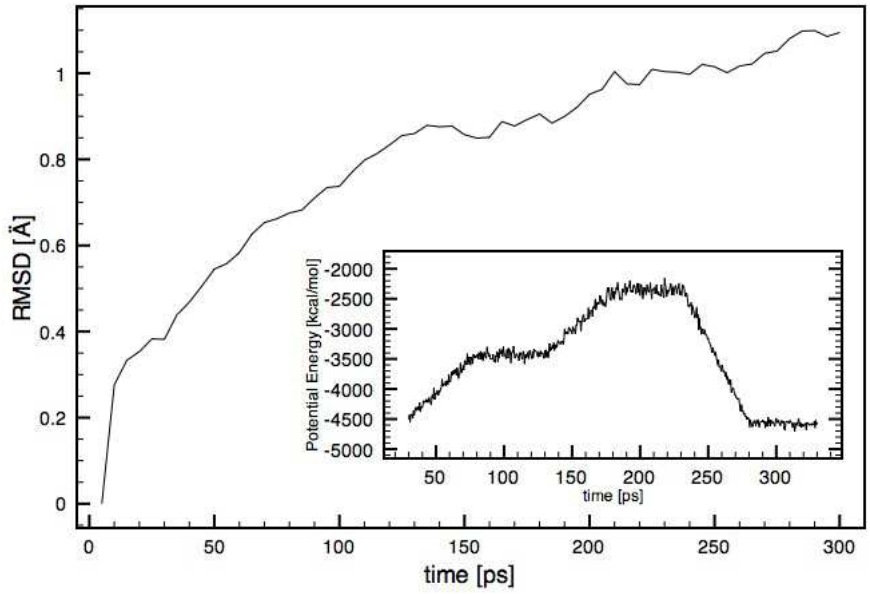

Figure 18: Time evolution of RMSD for the region 1-224. **Insert:** evolution of potential energy of the whole protein during the annealing procedure. Both quantities use the previously equilibrated conformation as reference (see text for more details).

. **aIF6-htagged**: As explained in Section 3.1.1, for technical reasons due to the specific experimental setup, high pressure measurements were performed on samples produced with a specific tag of six consecutive histidines (His-tag) which significanlty improved the yield of the production protocol. In order to compare these measurements with those obtained from molecular dynamics, the structure of aIF6 complexed with the His-tag was modeled following the same procedure used for the global structure of eIF6.

Both samples were simulated following the protocol used for aIF6 and eIF6.

4

CHARACTERIZATION OF PROTEIN STRUCTURE

In the next chapters a new method for the characterization of protein secondary structure will be presented. The development of this method, called *ScrewFit*, was inspired by the task of finely characterize the environmental effects on protein structures.
ScrewFit was then found to be able to make precise assessments on protein secondary structure motifs and also to find local and global structural effects induced by ligand binding. The text proposed here has been already published or is under review as scientific articles in international peer-reviewed journals. The original text is reprinted here together with supplemental data and notes at the end of each article.

Next chapters formerly appeared as the following scientific articles:

Kneller, G.R. and Calligari, P. Efficient characterization of protein secondary structure in terms of screw motions. *Acta Crystallographica D*, **62**, 302-311 (2006).

Calligari, P. and Kneller G.R., ScrewFit: a novel approach for continuum protein secondary assessments. *Submitted* (2008).

Another application of the method ScrewFit can be found in:

Calligari, P. et al., Inhibition of viral group-1 and group-2 neuraminidases by oseltamivir: a comparative structural analysis by the ScrewFit algorithm. *Biophysical Chemistry*, accepted for publication (2008).

5

EFFICIENT CHARACTERISATION OF PROTEIN SECONDARY STRUCTURE IN TERMS OF SCREW MOTIONS

We present a simple and efficient method to describe the secondary structure of proteins in terms of orientational distances between consecutive peptide planes and local helix parameters. The method uses quaternion-based superposition fits of the protein peptide planes in conjunction with Chasles' theorem, which states that any rigid body displacement can be described by a screw motion. From the best superposition of consecutive peptide planes we derive the helix parameters, and the "worst" fit is used to define the orientational distance. Applications are shown for standard secondary structure motifs of peptide chains, for some proteins belonging to different fold classes, and for a description of structural changes in lysozyme under hydrostatic pressure. In the latter case we use published reference data which have been obtained by X-ray crystallography and by structural NMR measurements.

5.1 INTRODUCTION

The determination and characterisation of protein secondary structure is a fundamental task in molecular biology, crystallography and in simulation studies. In many situations arises the necessity to quantify in particular structural changes of a protein, which are due to a change of its environment. The influence of temperature or pressure on the fold of a protein is a typical example. Standard motifs in protein secondary structure are traditionally described in terms of two torsional angles, ϕ and ψ, per residue, which define for each C_α-atom the rotation of the left and right peptide plane about the $N - C_\alpha$ and $C_\alpha - C$ bond, respectively [187]. In the past different methods have been developed to determine secondary structure elements [106, 163, 57, 190] and to describe their geometry in more detail [9, 181, 194]. A rigourous mathematical description of protein secondary structure can be obtained by applying the theory of screw motions, where the winding of the protein backbone is described in terms of local helix parameters. The theory of screw motions goes back to the mathematician M. Chasles [32, 33], and a useful recent introduction can be found in the book by Selig [176]. In a recent paper, Quine [155] uses screw motion theory and constructs local helix parameters for a protein from the torsion angles ϕ and ψ. An important step is the introduction of quaternions which can be related to the (ϕ, ψ)-angles on one hand, and to the rotation/helix axis on the other hand.

In this article we present an efficient method for the characterisation of protein secondary structure, which is based on quaternion superposition fits of consecutive peptide planes. From the resulting quaternion parameters we construct the local helix geometry of the protein backbone, and we show that the superposition method may also be used to define a scalar measure for the orientational distance between consecutive peptide planes. The latter allows to distinguish between all common secondary structure motifs, such as different helix types and β-strands, except for handedness.

In the following section the method is briefly explained and applications are presented in Section 5.3. The first one concerns an illustration for simple model structures, such as right- and left-handed α-helices and β-strands. We show then how our method works for proteins which fall into different fold classes, and discuss finally in more detail how it can be used to quantify changes in the secondary structure of lysozyme which are caused by external pressure. For this purpose we use published reference structures which have been obtained from X-ray crystallography and from structural NMR measurements. The essential results are summarised and discussed in Section 5.4. In the Appendix we recall the essential properties

of quaternions and give a short constructive proof of Chasles' theorem, which demonstrates the usefulness of quaternion calculus.

5.2 METHOD

As stated in the introduction, our method for the description of protein secondary structure relies on quaternion-based superposition fits of molecular structures. The method is well established, and we refer to articles by Kearsley [111] and by Kneller [115] for details. Here we use that the quaternion method does not only yield the "best" fit, from which local helix parameters describing the winding of the protein backbone can be constructed, but also the "worst" fit, from which an orientational distance measure can be derived.

5.2.1 *Quaternion superposition fits*

Suppose that $\{\vec{r}_\alpha\}$ and $\{\vec{r}'_\alpha\}$ are two sets of vectors describing the positions of atoms representing equivalent molecular structures A and B, respectively. Both structures contain the same number of atoms and are somehow placed in space. A rigid-body displacement $A \to B$ can be defined as an optimisation problem, where structure A is fitted onto structure B in a least squares sense. In case that both structures are identical, the resulting fit error will be zero. One starts by constructing the translation vector $\vec{t} = \vec{R}_{c'} - \vec{R}_c$ connecting the two centres of rotation, C and C′, which are to be chosen in the same way for A and B, and computes the coordinate sets $\{\mathbf{x}_\alpha\}$ and $\{\mathbf{x}'_\alpha\}$ containing the relative atomic positions to the respective rotation centres. Here and in the following the prime refers to the target structure B. The optimal rotation is obtained by minimising the target function

$$m(q) = \sum_{\alpha=1}^{N} w_\alpha (\mathbf{D} \cdot \mathbf{x}_\alpha - \mathbf{x}'_\alpha)^2 \tag{5.1}$$

with respect to a set of angular variables which parametrise the orthogonal rotation matrix $\mathbf{D}$. Each atom is assigned a positive weight w_α, with $\sum_\alpha w_\alpha = 1$. A convenient set of angular variables are normalised (real) quaternion parameters, $q \equiv \{q_0, q_1, q_2, q_3\}$, with $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$. In this case $\mathbf{D}$ takes the form [4]

$$\mathbf{D}(q) = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(-q_0 q_3 + q_1 q_2) & 2(q_0 q_2 + q_1 q_3) \\ 2(q_0 q_3 + q_1 q_2) & q_0^2 + q_2^2 - q_1^2 - q_3^2 & 2(-q_0 q_1 + q_2 q_3) \\ 2(-q_0 q_2 + q_1 q_3) & 2(q_0 q_1 + q_2 q_3) & q_0^2 + q_3^2 - q_1^2 - q_2^2 \end{pmatrix} \tag{5.2}$$

and describes a proper rotation with $\det(\mathbf{D}) = +1$. Using the orthogonality of $\mathbf{D}$, the target function $m(q)$ can be written as a quadratic form in the quaternion parameters,

$$m(q) = \mathbf{q}^T \cdot \mathbf{M} \cdot \mathbf{q}, \tag{5.3}$$

where $\mathbf{q} = (q_0, q_1, q_2, q_3)^T$ is a column vector and $\mathbf{M}$ is a positive semi-definite matrix. The superscript "T" denotes a transposition. The matrix $\mathbf{M}$ has the form [111, 115]

$$\mathbf{M} = \sum_{\alpha=1}^{N} w_\alpha \begin{pmatrix} (\mathbf{x}_\alpha - \mathbf{x}'_\alpha)^2 & \mathbf{u}_\alpha^T \\ \mathbf{u}_\alpha & \mathbf{P}_\alpha \end{pmatrix}, \tag{5.4}$$

where $\mathbf{u}_\alpha$ and $\mathbf{P}_\alpha$ are given by

$$\mathbf{u}_\alpha = 2\mathbf{x}_\alpha \wedge \mathbf{x}'_\alpha, \tag{5.5}$$

$$\mathbf{P}_\alpha = \mathbf{x}_\alpha \cdot \mathbf{x'}_\alpha^T + \mathbf{x}'_\alpha \cdot \mathbf{x}_\alpha^T. \tag{5.6}$$

The minimization of $m(q)$ with respect to the quaternion parameters must be performed with the side constraint $\mathbf{q}^T \cdot \mathbf{q} = 1$. Using the method of Lagrange multipliers one is lead to the eigenvector problem

$$\mathbf{M} \cdot \mathbf{q} = \lambda \mathbf{q}. \tag{5.7}$$

Since $m(q) \geqslant 0$ the matrix $\mathbf{M}$ is positive semi-definite, and one obtains a set of four real eigenvalues, $\{\lambda_j\}$, with $\lambda_j \geqslant 0$ ($j = 1, \ldots, 4$), and a set of corresponding orthonormal eigenvectors, $\{\mathbf{q}_j\}$, with $\mathbf{q}_j^T \cdot \mathbf{q}_k = \delta_{jk}$. Here δ_{jk} is the Kronecker symbol. It follows from (5.3) and (5.7) that

$$m(q_j) = \lambda_j. \tag{5.8}$$

The eigenvalues are thus the residuals of the fit and can be ordered such that

$$\lambda_1 \leqslant \lambda_2 < \lambda_3 \leqslant \lambda_4. \tag{5.9}$$

The quaternion corresponding to the smallest eigenvalue, λ_1, is thus the solution for the optimal fit, and the quaternion parameters q_1 describe the relative orientation of $\{\mathbf{x}'_\alpha\}$ with respect to $\{\mathbf{x}_\alpha\}$.

We note that one obtains two twofold degenerate eigenvalues if the structures to be superposed are linear. In this case one has [115]

$$\lambda_{a,b} = \sum_\alpha w_\alpha \left(|\mathbf{x}_\alpha|^2 + |\mathbf{x}'_\alpha|^2 \mp 2|\mathbf{x}_\alpha||\mathbf{x}'_\alpha| \right), \tag{5.10}$$

where $a = 1, 2$, $b = 3, 4$, and both the rotation leading to the minimum and maximum distance are not uniquely determined. Any normalized linear combination of the two eigenvectors associated with λ_a and λ_b, respectively, describes an equivalent rotation.

The use of quaternion parameters in not only very convenient for finding a rigid-body transformation between two sets of coordinates, but the result can also be directly related to conventional representations of rotations. Here the following relation is of importance:

$$\mathbf{q} \equiv \begin{pmatrix} q_0 \\ \mathbf{q}_v \end{pmatrix} = \begin{pmatrix} \cos(\phi/2) \\ \sin(\phi/2)\mathbf{n} \end{pmatrix}. \tag{5.11}$$

From the scalar part of a quaternion, q_0, one obtains thus directly the rotation angle and the rotation axis can be extracted from the vectorial part, $\mathbf{q}_v$. It should be noted that the transformation $\phi \to \phi + 2\pi$, which leaves the rotation matrix $\mathbf{D}(\mathbf{n}, \phi)$ invariant, leads to a global change in sign of the quaternion parameteres. One verifies easily that $\mathbf{q}(\mathbf{n}, \phi + 2\pi) = -\mathbf{q}$. The pair of quaternions $\{Q, -Q\}$ is thus mapped onto the same rotation matrix $\mathbf{D}(q)$.

5.2.2 *Orientational distance*

The eigenvalue describing the "worst" superposition – λ_4 according to the ordering scheme (5.9) – can be used to define an orientational distance between two molecular structures via

$$\Delta_\Omega = \sqrt{\frac{M_{11}}{\lambda_4}}. \tag{5.12}$$

Eq. (5.4) shows that the matrix element M_{11} contains the squared Euclidean distance between the vectors sets $\{\vec{x}_\alpha\}$ and $\{\vec{x}'_\alpha\}$, and therefore Δ_Ω is the Euclidean distance normalised to its maximum possible value. Consequently,

$$0 \leqslant \Delta_\Omega \leqslant 1. \tag{5.13}$$

It is important to note that definition (6.5) yields a unique orientational distance of two linear molecular structures, whose relative orientation has no unique description in terms of angular variables. Supposing that $|\mathbf{x}'_\alpha| = |\mathbf{x}_\alpha|$ for $\alpha = 1, \ldots, N$, we see from eq. (5.10) that for linear rigid bodies $\Delta_\Omega = 0$ in the parallel configuration and $\Delta_\Omega = 1$ in the anti-parallel configuration. We note here that λ_b, as given by eq. (5.10), is a strict upper limit for the Euclidean distance of two molecular structures in general [116].

5.2.3 *Chasles' theorem*

Let $\mathbf{r} = (x, y, z)^T$ be a column vector containing the coordinates of a radius vector $\vec{r} = \overrightarrow{OP}$ of a point P in a rigid body, where O is the origin of the coordinate system. An arbitrary rigid-body displacement is described by a rotation about a point C, which is not necessarily located inside the rigid body, and a subsequent translation. Let $\mathbf{R}_c$ be the coordinates of the radius vector $\vec{R}_c = \overrightarrow{OC}$ and let $\mathbf{t}$ be the coordinates of the translation vector $\vec{t} = \overrightarrow{CC'}$, where C′ is the centre of rotation after the translation. The coordinates of P after a rigid body displacement are then given by

$$\mathbf{r}' = \mathbf{R}_c + \mathbf{D} \cdot (\mathbf{r} - \mathbf{R}_c) + \mathbf{t}, \tag{5.14}$$

where $\mathbf{D}$ is an orthogonal 3×3 matrix. In the following only proper rotations with $\det(\mathbf{D}) = +1$ will be considered. If $\mathbf{n} = (n_x, n_y, n_z)^T$ contains the components of the unit vector $\vec{n}$, pointing in the direction of the rotation axis, and ϕ is the angle of rotation, the corresponding rotation matrix can be written as

$$\mathbf{D}(\mathbf{n}, \phi) = \mathbf{P}_{\parallel} + \cos(\phi)\mathbf{P}_{\perp} + \sin(\phi)\mathbf{N}, \tag{5.15}$$

where $\mathbf{P}_{\parallel} = \mathbf{n} \cdot \mathbf{n}^T$ and $\mathbf{P}_{\perp} = \mathbf{1} - \mathbf{P}_{\parallel}$ are, respectively, the projectors onto $\vec{n}$ and its complement, and $\mathbf{N}$ is the antisymmetric matrix

$$\mathbf{N} = \begin{pmatrix} 0 & -n_z & n_y \\ n_z & 0 & -n_x \\ -n_y & n_x & 0 \end{pmatrix}. \tag{5.16}$$

The theorem of Chasles states that one can find a reference point X, whose radius vector $\vec{R}_x = \overrightarrow{OX}$ has the coordinates $\mathbf{R}_x$, such that

$$\mathbf{r}' = \mathbf{R}_x + \mathbf{D}(\mathbf{n}, \phi) \cdot (\mathbf{r} - \mathbf{R}_x) + \alpha\mathbf{n}. \tag{5.17}$$

This coordinate transformation describes a screw motion, with translation α parallel to the axis of rotation. For the following considerations we introduce the difference vector $\vec{u} = \vec{R}_x - \vec{R}_c$. Equating relations (5.14) and (5.17) and using that $\mathbf{n}$ is an eigenvector of $\mathbf{D}$, one finds that the coordinates of $\vec{u}$ satisfy the following set of linear equations

$$(\mathbf{1} - \mathbf{D}) \cdot \mathbf{u} = \mathbf{t}_{\perp}. \tag{5.18}$$

Here $\mathbf{t}_{\perp} = \mathbf{P}_{\perp} \cdot \mathbf{t}$. As shown in Section 5.5, the above equation has a linear manifold of solutions,

$$\mathbf{u}(\lambda) = \mathbf{u}_{\perp} + \lambda\mathbf{n}, \quad \lambda \in \mathbb{R}, \tag{5.19}$$

where $\mathbf{u}_\perp$ is perpendicular to $\mathbf{n}$ and has the explicit form

$$\mathbf{u}_\perp = \frac{1}{2}\Big(\mathbf{t}_\perp + \cot(\phi/2)\mathbf{n} \wedge \mathbf{t}\Big). \tag{5.20}$$

In absolute coordinates the axis of the screw motion is given by

$$\mathbf{R}_x = \mathbf{R}_c + \mathbf{u}_\perp + \lambda\mathbf{n}, \tag{5.21}$$

and

$$\mathbf{R}_x^\perp = \mathbf{R}_c + \mathbf{u}_\perp \tag{5.22}$$

contains the coordinates of the radius vector $\vec{R}_x^\perp$ relating the origin with the point $X^\perp$ on the helix axis which is closest to the reference point C. In the following $X^\perp$ will be referred to as centre of screw motion. The radius ρ of the corresponding screw motion is given by the Euclidean length of $\vec{u}_\perp$, since the latter is the vector pointing from $\vec{R}_c$ to $X^\perp$. Using (5.20) one finds

$$\rho = \frac{|\mathbf{t}_\perp|}{2}\sqrt{1 + \cot^2(\phi/2)}. \tag{5.23}$$

It should be noted that ρ diverges if ϕ is a multiple of 2π, corresponding to pure translations, and if $|\mathbf{t}_\perp| \neq 0$.

5.3 APPLICATIONS

5.3.1 *Screw motion description of protein main chains*

The method described above, which will be referred to as *ScrewFit* in the following, is now applied to define the local helical structure of polypeptides and proteins. The rigid bodies are here the triangles formed by the atoms $\{O, C, N\}$ in the backbone of polypeptides – see Fig. 19 – which define the so-called peptide planes. Here the C-atoms are the centres of rotation, and the translation vectors are thus the position differences between the C-atoms in consecutive amino acids, $\mathbf{t}_i = \mathbf{R}_{C(i+1)} - \mathbf{R}_{C(i)}$. The quaternion parameters $\mathbf{q}_i$ are obtained from the fit of the $\{O, C, N\}$-triangle of peptide bond i onto the one of peptide bond $i+1$. From each set of quaternion parameters the direction $\vec{n}$ of the rotation axis and the rotation angle ϕ can be computed from relation (6.4).

The following parameters are used to define the local helix structure of a polypeptide:

- The helix radius ρ defined in eq. (6.6).

- The number of amino acids per turn,

$$\tau = \frac{2\pi}{\phi}. \tag{5.24}$$

- The pitch, which is defined as

$$p = |\mathbf{R}^{\perp}_{x,i+1} - \mathbf{R}^{\perp}_{x,i}|\tau. \tag{5.25}$$

  Here $\vec{R}^{\perp}_{x,i}$ is the radius vector pointing from the origin to the centre $X^{\perp}_{i}$ of the screw motion relating peptide plane i and peptide plane $i+1$.

- The handedness, which is defined as the sign of the projection of the translation vector $\vec{t}_i$ onto the direction $\vec{n}_i$ of the local helix axis,

$$h = \text{sign}(\mathbf{n}_i^{\top} \cdot \mathbf{t}_i). \tag{5.26}$$

- The straightness parameter σ of the local helix axis. For residue i the latter is defined as

$$\sigma_i = \boldsymbol{\mu}_i^{\top} \cdot \boldsymbol{\mu}_{i+1}, \tag{5.27}$$

  where

$$\boldsymbol{\mu}_i = \frac{\mathbf{R}^{\perp}_{x,i+1} - \mathbf{R}^{\perp}_{x,i}}{|\mathbf{R}^{\perp}_{x,i+1} - \mathbf{R}^{\perp}_{x,i}|}. \tag{5.28}$$

- The orientational distance between the peptide planes $\{O, C, N\}$ in residues i and $i+1$, which is defined through relation (6.5).

5.3.2 *Model structures*

We apply *ScrewFit* first to well-known model structures for polypeptides which have been taken from the Image Library of Biological Macromolecules in Jena [1]. Table 13 shows the corresponding local helix parameters which have been defined in the previous section. All model peptides are polyalanine molecules containing 10 residues, except for the extended conformation, which is represented by the alanine-tripeptide shown in Fig. 19. In all cases the N-terminus is the starting point of the respective polypeptide chain. In the context of our study all motifs are considered as helices, a β-strand being simply a thin left-handed helix with 2 amino acids per turn[2].

1 Institute-of-Molecular-Biotechnology-Jena:http://www.imb-jena.de/IMAGE.html

2 See also note 5.6 at the end ot this article.

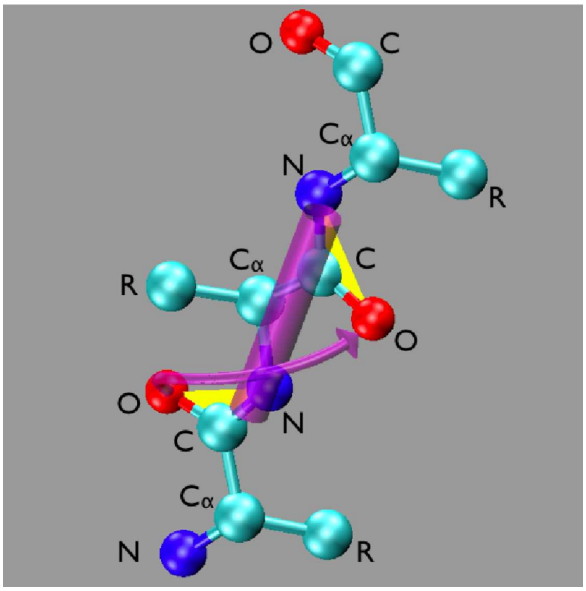

Figure 19: A tri-peptide with two peptide bonds in the extended conformation, where the symbol "R" stands for non-specified side-chains. The screw motion relating the yellow triangles formed by the $\{O, C, N\}$ atoms of the peptide planes defines the local helix which is schematically represented by the cylinder in purple and the corresponding screw arrow. The radius of the cylinder corresponds to the radius of the screw motion.

Table 12: Helix parameters for different model structures. Here ρ is the helix radius with the C-atom of the peptide plane on the helix surface, ρ_{C_α} is the corresponding radius if the C-atom is replaced by the C_α-atom, τ is the number of residues per turn, h is the handedness, and σ the straightness parameter. The latter equals 1 for all model structures, since none is curved. The straightness parameter for the extended conformation cannot be defined, since the model structure for the latter consists of only three residues (see Fig. 19). More explanations are given in the text.

Motive	ρ [nm]	ρ_{C_α} [nm]	τ	pitch	h	σ	Δ_Ω
α-helix (R)	0.171	0.227	3.62	0.556	+	1	0.582
α-helix (L)	0.171	0.227	3.62	0.556	−	1	0.582
3-10 helix	0.146	0.203	3.28	0.589	+	1	0.670
π-helix	0.178	0.258	4.16	0.558	+	1	0.471
β-strand	0.055	0.093	2.03	0.671	−	1	0.875
extended	0.037	0.055	2.00	0.725	−		0.754

The parameters concerning the different secondary structure motifs shown in Table 13 may be compared to those published in the study of Barlow and Thornton [9]. Here one must pay attention to the fact that the helix radius depends on the reference point which is chosen to lie on the helix surface. In our case this is the carbon atom in the $\{O - C - N\}$ peptide plane. If the carbon C_α-atom is chosen instead, we find the values given in the column with the header "ρ_{C_α}".

The parameters we find for the right-handed α-helix are very close to the ones given by Barlow and Thornton, who compare different standard definitions with average values computed from a set of 291 helices in "real" proteins. The parameters listed in the above reference are in the intervals $0.23 \leqslant \rho \leqslant 0.24$, $3.54 \leqslant \tau \leqslant 3.67$, and $0.52 \leqslant p \leqslant 0.55$, respectively, using our notation and units (ρ and p in nm). In case of the 3-10 helix the spread of the parameters given by Barlow and Thornton is $0.18 \leqslant \rho \leqslant 0.20$, $3.0 \leqslant \tau \leqslant 3.2$, and $0.58 \leqslant p \leqslant 0.60$. Parameters for π-helices are not listed. It should be noted that the orientational distance takes well distinguishable values for the different secondary structure motifs, but left and right-handed motifs cannot be distinguished by this parameter.

5.3.3 *Proteins in different fold classes*

In the following we will show the results of *ScrewFit* for proteins which fall into the four main fold classes according to the SCOP scheme [145]:

1. Carbonmonoxy-myoglobin (PDB code 1A6G), which belongs to the "all alpha" class.
2. Protease inhibitor ecotin (PDB code 1ECY), which belongs to the "all beta" class.
3. Triose phosphate isomerase from chicken muscle (PDB code 1TIM), which belongs to the "alpha/beta" class. Proteins falling into this class consist mainly of parallel β-sheets, which are separated by α-helices.
4. Hen egg white lysozyme (PDB code 193L), which falls into the "alpha + beta" class. Proteins of this type contain mainly anti-parallel β-sheets and separated regions containing α-helices.

The latter application is postponed to the next section, where we consider structural changes of lysozyme under pressure. In this context the *ScrewFit* parameters will also be discussed in more detail. Here we give only an impression of the results, as compared to *DSSP*.

Figs. 20 to 22 shows the comparison first three proteins in the list given above. In each figure we give the local orientational distance, Δ_Ω, the local helix radius, ρ, and the straightness

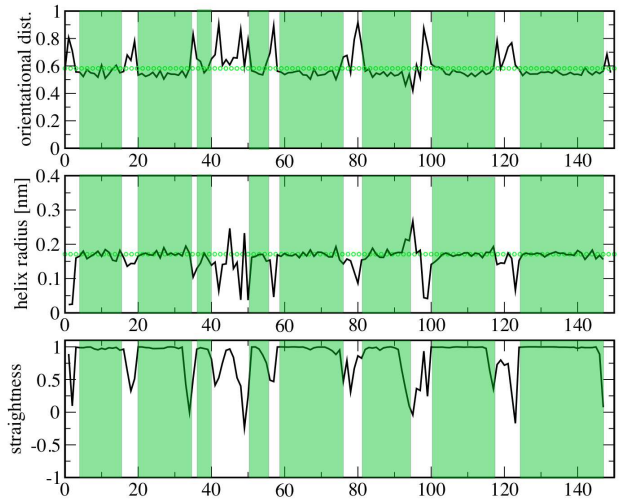

Figure 20: *ScrewFit* description of the main chain of Carbonmonoxy-myoglobin (PDB code 1A6G, "all alpha" in the SCOP scheme). The vertical green stripes indicate α-helices found by the *DSSP* method and the horizontal lines indicate the reference values given Table 13.

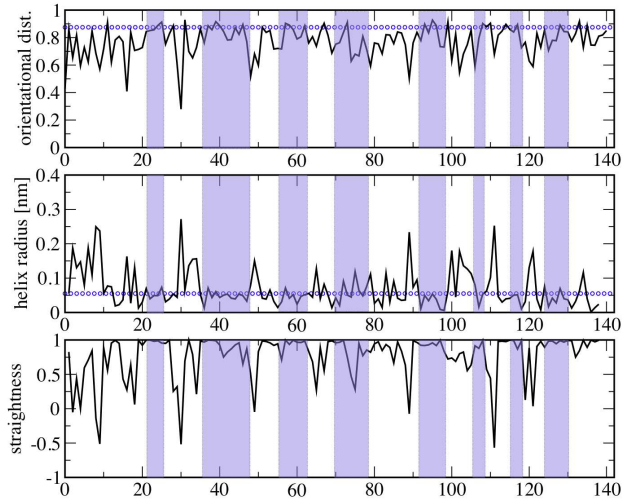

Figure 21: *ScrewFit* description of the main chain of Protease inhibitor ecotin (PDB code 1ECY, "all beta" in the SCOP scheme). The vertical blue stripes indicate β-strands found by the *DSSP* method and the horizontal lines indicate the reference values given in Table 13.

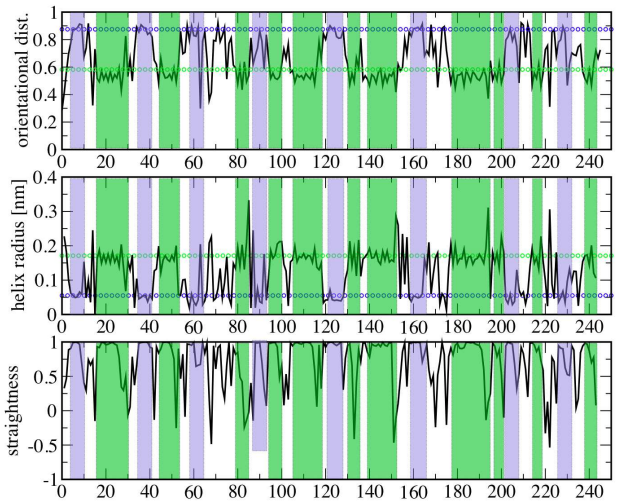

Figure 22: *ScrewFit* description of the main chain of Triose phosphate isomerase (PDB code 1TIM, "alpha/beta" in the SCOP scheme). The vertical green and blue stripes indicate, respectively, α-helices and β-strands found by the *DSSP* method. The green and blue horizontal lines indicate the respective reference values from Table 13.

parameter, σ. All calculations have been performed on the basis of the respective entries in the Brookhaven Protein Data Bank (PDB). The vertical stripes correspond to the secondary structure motifs found by the *DSSP* method by Kabsch and Sander, which is based on hydrogen bonding criteria and which is widely used for the determination of secondary structure elements in proteins [106]. The colouring scheme indicates α-helices in light green and β-strands in light blue.

One recognises that the *ScrewFit* method often leaves some ambiguity concerning the boundaries of secondary structure elements. This is simply due to the fact that it is sensitive to deviations from ideal geometries. This effect is in particular visible in the behaviour of straightness parameter. Similar observations have been made by comparing the method by Barlow and Thornton with *DSSP* [9].

5.3.4 *Lysozyme under hydrostatic pressure*

In the following we apply our method to visualise structural changes in lysozyme due to the application of an external pressure. For this purpose we consider protein structures which have been obtained from X-ray crystallography and from NMR measurements. The X-ray structures are taken from the entries 193L and 3LYM of the Brookhaven Protein Data Bank (PDB), which contain the atomic coordinates of hen egg-white lysozyme at pressures of,

respectively, 1 bar and 1 kbar [202, 121]. The NMR structures are taken from PDB entries 1GXV and 1GXX, corresponding to 1 bar and 2 kbars, respectively [159]. Fig. 23 shows the backbone of lysozyme at 1 bar (blue tube) obtained from the crystal structure together with the line joining the centres of screw motion mapping each peptide plane onto the consecutive one (red line). The centres of the screw motions have been constructed according to expression (5.22). Inspection by eye shows that that the red line passes right through the geometrical centres of the helices.

More details can be obtained from Figs. 24 and 25 which show the same parameters as in Figs. 20 to 22. In both cases the curves corresponding to the structures under pressure are given in red. The green and blue horizontal lines correspond again, respectively, to the reference values for an α-helix and a β-strand given in Table 13. Here the vertical stripes indicate the secondary structures according to the *PDBsum* data base. The latter uses the *PROMOTIF* program for secondary structure determination [89], which is itself based on the *DSSP* method. In addition to α-helices we indicate also 3-10 helices in dark green.

For the crystal structure entries 193L and 3LYM the *PDBsum* database displays three long helices in the residue intervals $\{5-14\}$, $\{25-36\}$, and $\{89-99\}$, and four short ones in the residue intervals $\{80-84\}$, $\{104-107\}$, $\{109-114\}$, and $\{120-123\}$. In addition three short β-strands of 2–3 residues are displayed in the regions $\{43-45\}$, $\{51-53\}$, and $\{58-59\}$, respectively. We note here that only the long helices are described in the work by Barlow and Thornton [9]. Concerning the NMR structures, the *PDBsum* database lists again the three long helices ($\{5-14\}$, $\{25-36\}$, $\{89-98\}$), but only two short ones ($\{80-84\}$, $\{109-114\}$). In contrast, the short β-strands are displayed at almost the same positions as for the crystal structures ($\{44-46\}$, $\{50-53\}$, $\{58-59\}$). As for the crystal structures, the structural motifs are found for both pressures at identical positions.

Looking first at Fig. 24 displaying the parameters corresponding to the crystal structures shows that the orientational distance is a good measure to localise rapidly secondary structure elements in the amino acid sequence of a protein. For the moment we discuss only the structure at ambient pressure. The analysis of the helix radius and the straightness gives more detailed information. The three long helices and also the three short β-strands are easily localised. We find that the first helix ($\{5-14\}$) is straight only in the region $\{5-11\}$. Towards the C-terminus the straightness drops considerably and the orientational distance rises. The helix radius stays approximately constant up to about residue number 15. We find that the second helix ($\{25-36\}$) is deformed as well towards its C-terminus, but here the orientational distances stays more or less constant, whereas the helix radius and the

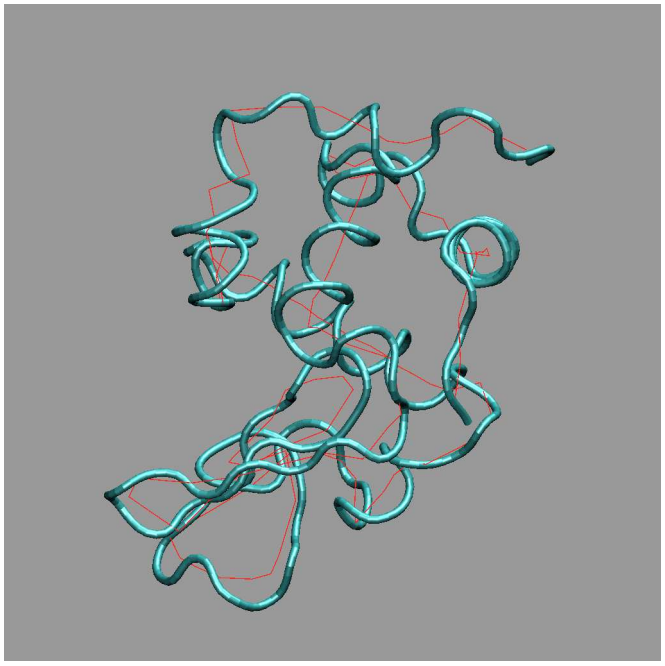

Figure 23: Minimal model for Lysozyme at normal pressure. The red line joins the centres of screw motions, $X^{\perp}$, mapping each peptide plane onto the following one.

straightness change considerably. We consider this helix to be straight in the range $\{25-32\}$. Similar observations can be made for the third long helix, which we find to be straight in the range $\{89-96\}$. We note here that Barlow and Thornton consider the first of the above helices as "irregular" and the others as "curved", using, however, different criteria. Concerning the shorter helices, which are not considered helices by Barlow and Thornton, we confirm less well defined helices in the ranges $\{80-84\}$, $\{104-107\}$, $\{119-123\}$, and $\{109-114\}$. According to the orientational distance, the first three of them are 3-10 helices. As for the β-strands, our analysis would confirm the short one in the range $\{58-59\}$, but yield longer strands in the regions $\{42-46\}$ and $\{50-53\}$. It is worthwhile mentioning that the straightness parameter indicates hairpin turns between the β-strands, leading to antiparallel β-sheets.

Applying *ScrewFit* to the NMR structure of lysozyme at ambient pressure yields the following results: the three long helices indicated by *PDBsum* are retrieved, and, using the orientational distance as criterion, we confirm less well defined helices in the ranges $\{80-84\}$ (3-10 helix) and $\{109-114\}$ (α-helix). According to our analysis the first β-strand is longer than the one displayed in the *PDBsum* database (approximately in $\{41-45\}$).

Let us now look at the changes in secondary structure of lysozyme due to the exertion of an external pressure. We start with the analysis of the crystallographic data by Kundrot and Richards [121]. The black line in Fig. 26 shows that the structural change obtained from the crystal structures is localised at residue no. 72. All parameters show a change in the

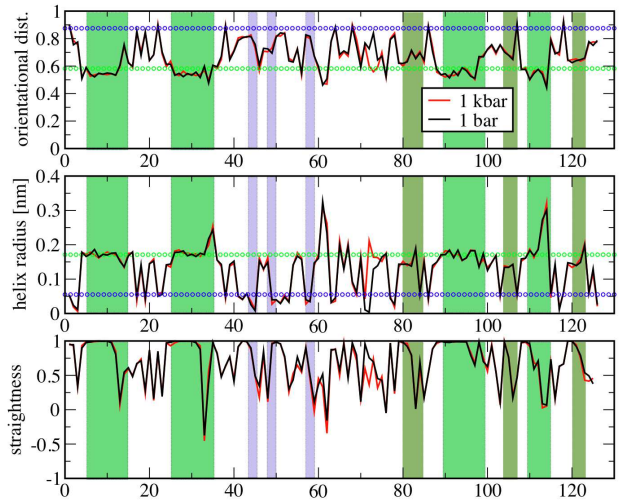

Figure 24: *ScrewFit* description of Lysozyme for crystallographic structures at pressures of 1 bar (black line) and 1 kbar (red line) (PDB entries 193L and 3LYM) [202, 121]. According to the SCOP scheme lysozyme falls into the "alpha+beta" class. The horizontal lines show the reference values given in Table 13 and the vertical stripes indicate here the secondary structures according to the *PDBsum* database. More explanations are given in the text

.

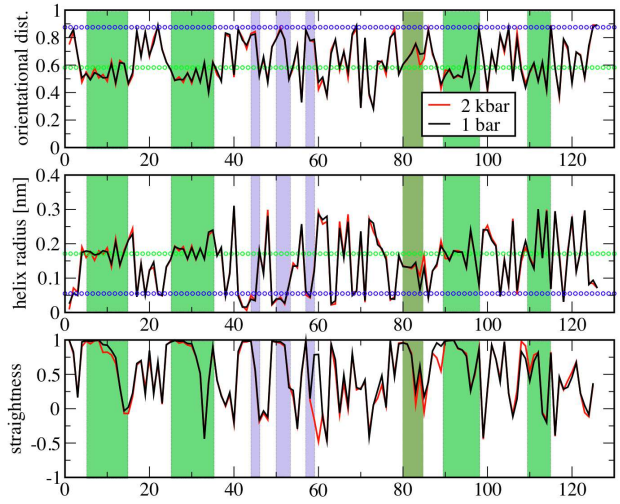

Figure 25: The same as Fig. 24, but for NMR structures at 1 bar (black line) and 2 kbar (red line) [159].

same place. Fig. 27 shows the change of straightness of the crystal structure of lysozyme in a tube representation, using a colouring scheme where red corresponds to a negative change, green to no change, and blue to a positive change. In their study Kundrot and Richards perform a difference distance matrix analysis of the structural changes and report that the least changes are seen in helix 2 ($\{25-36\}$) and in the loop and β-sheet region $\{42-60\}$, whereas a stronger structural change is seen in region $\{61-87\}$, which appears to expand. We note here that Kundrot and Richards call this region "loop region", not counting the short 3-10 helix $\{80-84\}$. This observation is coherent with ours, which shows in particular a strong rise of the local helix radius at residue 72, corresponding to a swelling of the corresponding loop region.

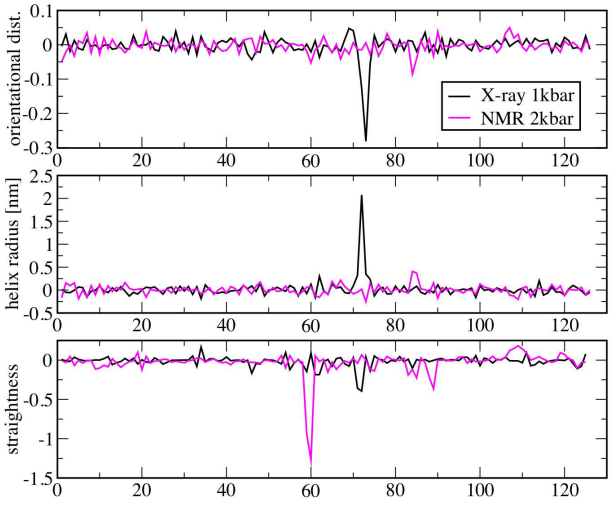

Figure 26: Differences for the parameters shown in Figs. **??** and 25 (X-ray crystallography = black line, NMR = magenta line).

The corresponding analysis for the NMR structures is less clear (see Fig. 25, magenta lines). Here the orientational distance and the helix radius do not exhibit significant changes, whereas the straightness shows a strong decrease at residue no. 60. which is located at the very beginning of the long loop in residue range $\{60-80\}$. Refaee *et al.* report the most extensive deformations in the loop and what they call "β-sheet domain" ($\{40-88\}$), which is certainly in agreement with a very localised change in secondary structure at residue no. 60. We do, however, not see considerable changes in the hairpin turns $\{47-49\}$ and $\{54-57\}$ seen by Refaee *et al.*. Fig. 28 shows the change in straightness for the NMR structure of lysozyme in a tube plot in which the same colouring scheme is used as in Fig. 27.

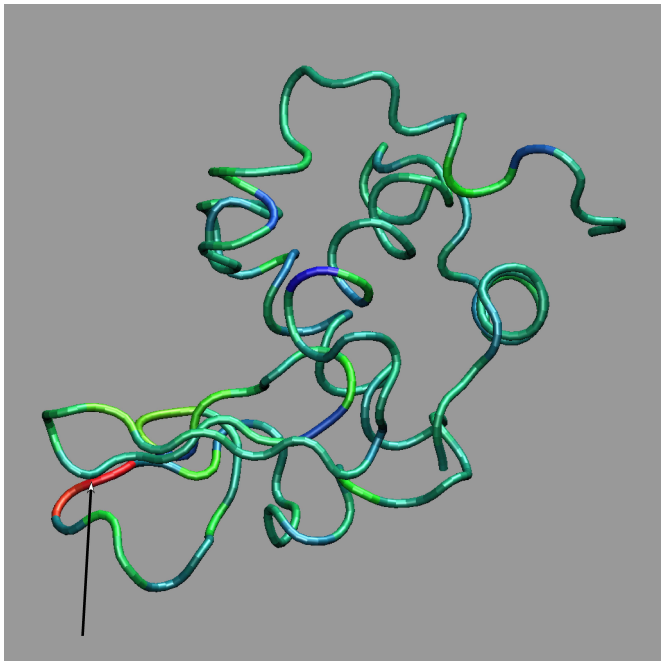

Figure 27: Change in straightness between the crystallographic structure of lysozyme at 3 kbar and 1 bar. The colouring scheme is chosen such that blue, green, red correspond to, respectively +2 (maximum positive change), 0, and -2 (maximum negative change).

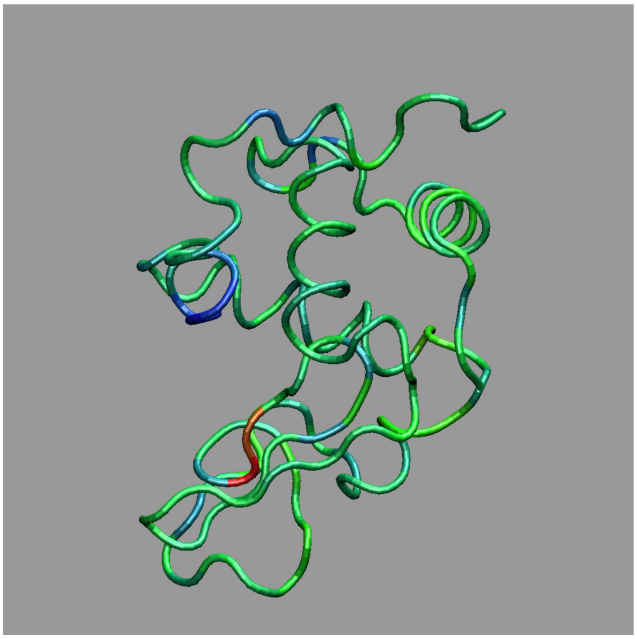

Figure 28: Change in straightness between the NMR structure of lysozyme at 2 kbar and 1 bar. The colouring scheme is the same as in Fig. 27.

5.4 CONCLUSION

We have presented a simple method – *ScrewFit* – for the characterisation of protein secondary structure which uses quaternion-based superposition fits of consecutive peptide planes in

the backbone. The combined use of the quaternion fit method and Chasles' theorem allows to express protein secondary structure in terms of local helix parameters. The superposition method yields also an orientational distance measure for consecutive peptide planes. The latter is obtained from the "worst" possible quaternion fit and yields a simple measure for the rapid localisation of secondary structure elements along the protein backbone. The analysis of standard motifs of protein secondary structure and of proteins belonging to different fold classes showed that all common motifs are well discriminated by the orientational distance measure, and that the straightness parameter and the helix diameter are useful to characterise non-ideal secondary structure elements, keeping a minimal set of parameters.

Using *ScrewFit* to study conformational changes in Lysozyme due to application of an external pressure revealed different localised changes in the loop regions. The structural changes extracted by difference distance matrix analysis from the crystallographic data could be confirmed, giving, however, a more precise description of these changes. Concerning the NMR structures, we find the essential conformational changes in a different position than the authors of the reference article, although both results agree in so far, as the changes are found in the same region. Prior to these analyses we tested that the localisation of the essential secondary structural elements found by crystallography and NMR is confirmed.

ScrewFit allows to pinpoint secondary structure changes precisely, which is more difficult to achieve by the standard analysis of positional differences. The reason is that the latter might indicate important structural differences in a large region, although the corresponding position differences are induced by one single localised change in the winding of the protein backbone.

A point which should also be mentioned is the numerical efficiency of the quaternion-based superposition algorithm we use as a basis of our method. The superposition of two molecular structures can be performed in a few millisceconds [115], and this fact has been exploited in many studies of rigid body motions in molecular systems, using the molecular dynamics analysis package *nMoldyn* [120, 166]. Using the method presented in this article, the characterisation of the secondary structure of a protein can be done in about a second on a normal PC, and this efficiency could for example be used in database-oriented applications and for analyses of molecular dynamics trajectories of proteins. In this context it is important to note that the protein backbone can be completely reconstructed from the helix parameters defined in this article. This is an interesting aspect for homology modelling. Another useful application could be the characterisation of structural variability in different structural models which are used to construct protein 3D structures from NMR distance data.

5.5 MATHEMATICAL BACKGROUND

5.5.1 *Quaternions*

Quaternions are hypercomplex numbers which are composed by linear superposition of one real unit element 1 and three imaginary imaginary unit elements $\mathcal{I}, \mathcal{J}, \mathcal{K}$. The latter satisfy the non-commutative algebra $\mathcal{I}^2 = \mathcal{J}^2 = \mathcal{K}^2 = -1$ and $\mathcal{I}\mathcal{J} = -\mathcal{J}\mathcal{I} = \mathcal{K}$ (cycl.). An arbitrary quaternion $\mathcal{Q}$ is written as $\mathcal{Q} = q_0 1 + q_1\mathcal{I} + q_2\mathcal{J} + q_3\mathcal{K}$, where $q_j \in \mathbb{R}$ $(j = 0, \ldots, 3)$. The component q_0 is called the scalar component, and $\{q_1, q_2, q_3\}$ are the vectorial components. It is useful to introduce the column vector $\mathbf{q}_v = (q_1, q_2, q_3)^T$ comprising the three vectorial components of a quaternion. Analogously to complex numbers, the length of a quaternion is defined as $\|\mathbf{Q}\| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2}$ and its conjugate is given by $\mathbf{Q}^* = q_0 1 - q_1\mathcal{I} - q_2\mathcal{J} - q_3\mathcal{K}$.

Let $\mathcal{A}$ and $\mathcal{B}$ be quaternions with components $\{a_0, a_1, a_2, a_3\}$ and $\{b_0, b_1, b_2, b_3\}$, respectively. The components of $\mathcal{C} = \mathcal{A} \pm \mathcal{B}$ are obtained by $c_j = a_j \pm b_j$ $(j = 0, \ldots, 3)$ and from the algebra of the imaginary elements one finds that the components of the product $\mathcal{C} = \mathcal{A}\mathcal{B}$ are given by

$$\begin{pmatrix} c_0 \\ \mathbf{c}_v \end{pmatrix} = \begin{pmatrix} a_0 b_0 - \mathbf{a}_v^T \cdot \mathbf{b}_v \\ a_0 \mathbf{b}_v + b_0 \mathbf{a}_v + \mathbf{a}_v \wedge \mathbf{b}_v \end{pmatrix},$$

where "$\wedge$" denotes a vector product. In general $\mathcal{A}\mathcal{B} \neq \mathcal{B}\mathcal{A}$. The inverse of a quaternion $\mathcal{A}$ is defined as

$$\mathcal{A}^{-1} = \frac{\mathcal{A}^*}{\|\mathcal{A}\|^2}.$$

Due to the non-commutative algebra of quaternions one has in general $\mathcal{A}^{-1}\mathcal{B} \neq \mathcal{B}\mathcal{A}^{-1}$.

Similarly to complex numbers of unit length, which represent rotations in the plane, normalised quaternions represent rotations in space. Let $\mathbf{r} = (x, y, z)^T$ a column vector comprising the components of a radius vector $\vec{r}$, let $\mathcal{R} = x\mathcal{I} + +y\mathcal{J} + z\mathcal{K}$ be the corresponding spatial quaternion and let $\mathcal{Q}$ be a normalised quaternion with $\|\mathcal{Q}\| = 1$. One finds that the scalar component of $\mathcal{R}' = \mathcal{Q}\mathcal{R}\mathcal{Q}^*$ vanishes too, and that the vectorial components of $\mathcal{R}'$ are given by

$$\mathbf{r}' = \mathbf{D} \cdot \mathbf{r},$$

where $\mathbf{D}$ represents the rotation matrix (6.2). The bilinear transformation $\mathcal{R}' = \mathcal{Q}\mathcal{R}\mathcal{Q}^*$ represents thus a rotation in space.

5.5.2 *Helix parameters in Chasles' theorem*

Chasles' theorem can be easily proven using quaternion algebra. For this purpose we start from Eq. (5.18) and introduce the spatial quaternions $\mathcal{U}$ and $\mathcal{T}_\perp$, representing, respectively, the column vectors $\mathbf{u}$ and $\mathbf{t}_\perp$. Expressed in quaternions, Eq. (5.18) becomes

$$\mathcal{U} - \mathcal{Q}\mathcal{U}\mathcal{Q}^* = \mathcal{T}_\perp.$$

Multiplication with $\mathcal{Q}$ from the right and using that $\mathcal{Q}^*\mathcal{Q} = 1$ yields

$$\mathcal{U}\mathcal{Q} - \mathcal{Q}\mathcal{U} = \mathcal{T}_\perp\mathcal{Q}.$$

Using the multiplication rule for quaternions, the above equation can be expressed in the form

$$\begin{pmatrix} -\mathbf{u}^\mathsf{T}\cdot\mathbf{q}_v \\ q_0\mathbf{u} + \mathbf{u}\wedge\mathbf{q}_v \end{pmatrix} - \begin{pmatrix} -\mathbf{q}_v^\mathsf{T}\cdot\mathbf{u} \\ q_0\mathbf{u} + \mathbf{q}_v\wedge\mathbf{u} \end{pmatrix} = \begin{pmatrix} -\mathbf{t}_\perp\cdot\mathbf{q}_v \\ q_0\mathbf{t}_\perp + \mathbf{t}_\perp\wedge\mathbf{q}_v \end{pmatrix}.$$

Here we can make use of relation (6.4)

$$\begin{pmatrix} q_0 \\ \mathbf{q}_v \end{pmatrix} = \begin{pmatrix} \cos(\phi/2) \\ \sin(\phi/2)\mathbf{n} \end{pmatrix},$$

from which we conclude that $\mathbf{t}_\perp\cdot\mathbf{q}_v = 0$, since $\mathbf{t}_\perp \perp \mathbf{n}$. We are thus left with the vector equation

$$\begin{pmatrix} 0 \\ 2\mathbf{u}\wedge\mathbf{q}_v \end{pmatrix} = \begin{pmatrix} 0 \\ q_0\mathbf{t}_\perp + \mathbf{t}_\perp\wedge\mathbf{q}_v \end{pmatrix},$$

which can be reduced to

$$\mathbf{n}\wedge\mathbf{u} = \frac{1}{2}\Big(-\cot(\phi/2)\mathbf{t}_\perp + \mathbf{n}\wedge\mathbf{t}_\perp\Big) \tag{5.29}$$

if $\phi \neq 2k\pi$ ($k \in Z$). Now one can apply on both sides a vectorial multiplication with $\mathbf{n}$, using that $\mathbf{n}\wedge(\mathbf{n}\wedge\mathbf{a}) = -\mathbf{a}_\perp$ for an arbitrary column vector $\mathbf{a}$. This yields

$$\mathbf{u}_\perp = \frac{1}{2}\Big(\mathbf{t}_\perp + \cot(\phi/2)\mathbf{n}\wedge\mathbf{t}\Big),$$

if one uses that $\mathbf{n}\wedge(\mathbf{n}\wedge\mathbf{t}_\perp) = -\mathbf{t}_\perp$ and that $\mathbf{n}\wedge\mathbf{t}_\perp = \mathbf{n}\wedge\mathbf{t}$. Relation (5.20) is thus proven. The general solution of Eq. (5.29) has obviously the form

$$\mathbf{u}(\lambda) = \mathbf{u}_\perp + \lambda\mathbf{n}, \quad \lambda \in \mathbb{R},$$

which shows that $\mathbf{u}_\perp(\lambda)$ is the solution of minimum length.

Acknowledgements: All figures containing molecular graphics have been generated using the VMD code for molecular dynamics simulation and visualisation of biomolecules [88]. The screw motion calculations have been performed with modules from the MMTK package [79].

5.6 NOTES

[1] It is worth to note here that, in the context of ScrewFit algorithm, all motifs are considered as helices, a β-strand being simply a thin left-handed helix with 2 amino acids per turn. This fact can be easily seen in Fig. 29 where the blue cylinder represent the extent of the helix obtained according to the definition of the C-atom of the peptide plane lying on the helix surface, for the right-handed α-helix and the β-strand.

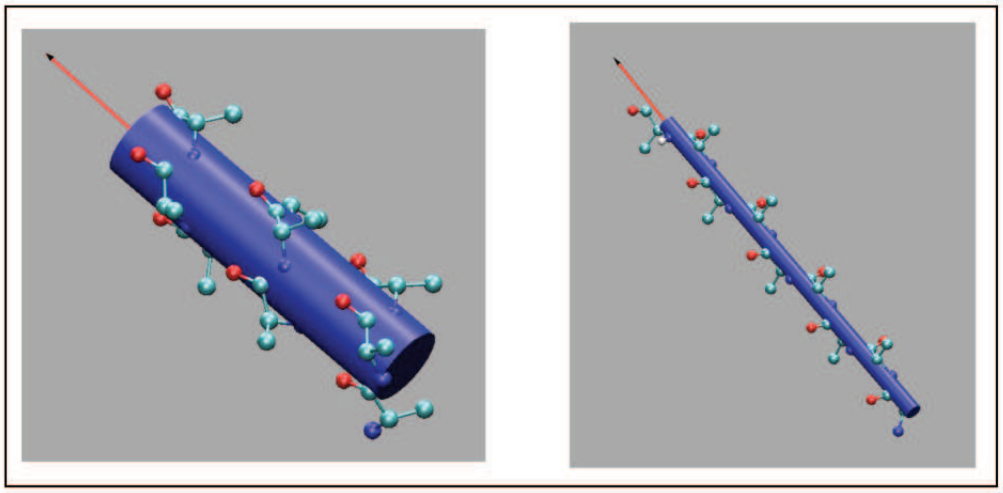

Figure 29: Visualisation of a straight right-handed α-helix (left-side) and a straight β-strand (right-side) in terms of the local helix parameters given in Table 1. The arrow indicates the direction of the helix axis. Here the hydrogen atoms.

6

SCREWFIT: A NOVEL APPROACH FOR CONTINUUM PROTEIN SECONDARY STRUCTURE ASSESSMENTS

We present a novel approach for the detection of protein secondary structure elements, which combines a description of the protein backbone in terms of screw motions (Acta Cryst. 62, p. 302-11 (2006)) with a statistical approach, yielding confidence ranges for the corresponding helix parameters on the basis of natural variations. To establish these ranges for each type of secondary structure element, we analyzed several databases of protein structures, exhibiting each well defined structural profiles. The method allows for a continous assessment of protein secondary structure elements and is proved to be stable with respect to both structural variations found in NMR data and resolution problems in crystallographic data. The comparison with other methods supports its reliability and accuracy. A structural analysis of bovine pancreatic trypsin inhibitor in three different crystal forms illustrates the capability of the method to detect secondary structure elements in noisy data and to describe at the same time small but systematic structural variations in the latter.

6.1 INTRODUCTION

In the last decades, a variety of new methods has been developed for the analysis of experimental and simulation data in structural biology. One of the standard tasks is the determination of secondary structure elements in proteins, in particular for the characterization of changes in protein structure. Such conformational changes may be induced by binding of ligands or by external stress, such as temperature, hydrostatic pressure, or chemical agents.

Traditionally secondary structure elements are described in terms of two torsional angles per residue, ϕ and ψ, which define for each C_α-atom in the protein backbone the rotation of the left and right peptide plane about the $N - C_\alpha$ and $C_\alpha - C$ bond, respectively [187]. This approach is, however, not well suited to describe non-ideal secondary structure elements, such as kinked or curved helices, which occur in any protein fold, nor does it easily allow to detect and to describe structural *changes* in protein structure. In the past, different methods have been developed to determine secondary structure elements in polypeptide chains [106, 163, 57, 190] and to describe their geometry in more detail [9, 181, 194]. One of the most frequently used methods is the dictionary of protein secondary structure (DSSP) which detects secondary structure elements through typical hydrogen bond patterns [106, 89]. The method allows in principle to distinguish eight of the most common motifs, both in secondary and supersecondary structures, but the results are often biased by the fact that DSSP makes assessments on protein structure using average hydrogen-bond distances as references which implicitly include the effects of static and dynamical disorder in crystallographic data. Like other, so-called "discrete methods", DSSP assesses secondary structure elements on the basis of a "true-false" decision, which follows geometrical criteria. All patterns which meet these criteria within certain thresholds are safely detected by all discrete methods [40], but they run into difficulties if the natural variability in the secondary structure of a given protein is too important. Methods like DSSP are, for example, not able to reproduce the variations between different NMR models which correlate with thermal disorder and local mobility of structure motifs. This fact has already been observed by Andersen and coworkers [5] [25] who presented a new method, derived from DSSP, which is called DSSPcont. The suffix "cont" indicates that the method uses a description of protein secondary structure in terms of a set of parameters which vary continuously in a predefined range of values. It has been pointed out in the literature that this *continuous approach* can more easily distinguish between *natural* conformational variations and effective changes in secondary structure profiles [5]. Indeed, Andersen and coworkers proved that a major improvement of the DSSP method

could be obtained simply substituting the single threshold-based definition per secondary structure element by a quasi-continuum spectrum of assessments obtained by running the discrete DSSP method with different hydrogen bond distance thresholds. The resulting series of assessments leads to the final assignment probability of each residue to belong to one of the eight DSSP classes: the three helix types (α, 3_{10} and π); β strands; helix-turns; β bridges; bends and not-structured loops.
With these assumptions, DSSPcont was refined in order to maintain a high consistency with respect to the influences of small structural variations caused by the experimental setups or by the natural thermal fluctuation of structures.

Motivated by the task to pinpoint *changes* in the secondary structure of proteins, we have recently developed the ScrewFit-algorithm [114]. The algorithm uses the Cartesian atomic coordinates of a protein as input and expresses its secondary structure in terms of screw motions relating consecutive peptide planes in the protein backbone. In this work we present further developments of this method which demonstrate that ScrewFit does not only allow to quantify changes in protein configurations but can also be used to *detect* all common secondary structure elements.

By construction, ScrewFit belongs to the so-called continuous methods for secondary structure assignment, since the local helix parameters quantifying the screw motions are continuous functions of the residue number.
Nevertheless, the ScrewFit method proposes a new approach to the continuous assignments as it does not relate to any threshold-based definition but only to simple geometrical criteria that define the set of parameters. The latter run over continuous ranges of values whose interpretation is not biased by an *a priori* division into classes and which are simply obtained from an empirical observation of the natural variations of the parameters in some appropriate protein structure databases. Moreover, this approach does not limit the values of the parameters which can vary beyond the ranges found for the most common secondary structure elements and still be useful to finely characterize the secondary level of uncommon structural configuration.

In the following sections we will first present the ScrewFit-algorithm and then we will show that structure definitions obtained by this algorithm are as precise as those obtained by the DSSP discrete method and are less biased by finite resolution of crystallographic data.

In a second application we will compare ScrewFit with the DSSPcont method mentioned above in order to verify the stability of our set of parameters in respect to the variations observed in NMR models due to experimental technique and thermal fluctuations.

Finally, as an example of a typical application, we will show the secondary structure profile of bovine pancreatic trypsin inhibitor (BPTI) in its three different crystal forms and we will compare the results with those detected with DSSPcont. This example will be useful to illustrate the advantages of this novel approach which allows to make assessments on protein secondary structure and also quantitatively characterize all changes it can undergo as an effect of different external conditions.

6.2 METHODS

6.2.1 *The ScrewFit algorithms*

To synthetically describe protein secondary structure, we use the algorithm *ScrewFit*, which is based on quaternion superposition fits for molecular structures [50, 115, 116] and Chasles' theorem on rigid-body displacements [33, 32]. In this section we sketch the main features of the method and refer to [114] for more details.

In the following two consecutive peptide planes A and B, containing the atoms $\{O, C, N\}$, are considered as rigid bodies. The latter are superimposed by minimizing the target function

$$m(\mathbf{q}) = \sum_{\alpha=1}^{3} (\mathbf{D} \cdot \mathbf{x}_\alpha - \mathbf{x}'_\alpha)^2, \tag{6.1}$$

where $\{\mathbf{x}_\alpha\}$ and $\{\mathbf{x}'_\alpha\}$ are the atomic positions of the reference (the atoms $\{O, C, N\}$ in plane A) and the target structure (plane B), respectively.

The symbol $\mathbf{D}$ denotes an orthogonal matrix describing a proper rotation . Both coordinate sets are defined with respect to a reference point, which is chosen to be the position of atom C in the respective peptide plane $\{O, C, N\}$. Using the fact that a rotation matrix can

be expressed in the components of a normalized quaternion $\mathbf{q} \equiv \{q_0, q_1, q_2, q_3\}$, where $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ [4],

$$\mathbf{D}(q) = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(-q_0q_3 + q_1q_2) & 2(q_0q_2 + q_1q_3) \\ 2(q_0q_3 + q_1q_2) & q_0^2 + q_2^2 - q_1^2 - q_3^2 & 2(-q_0q_1 + q_2q_3) \\ 2(-q_0q_2 + q_1q_3) & 2(q_0q_1 + q_2q_3) & q_0^2 + q_3^2 - q_1^2 - q_2^2 \end{pmatrix}, \quad (6.2)$$

the function (6.1) is to be minimized with respect to these four parameters. As demonstrated in previous work [50, 115, 116], the constrained minimization problem can be mapped onto an eigenvector problem for a positive semi-definite matrix $\mathbf{M} \equiv \mathbf{M}(\{\mathbf{x}_\alpha, \mathbf{x}'_\alpha\})$,

$$\mathbf{M} \cdot \mathbf{q} = \lambda \mathbf{q}, \quad (6.3)$$

whose eigenvalues $\lambda_j = m(\mathbf{q}_j)$ are four possible errors of the superposition fit between the two peptide planes, defined by (6.1). The smallest eigenvalue is the solution for the optimal fit, and the components of the corresponding eigen-quaternion describe the relative orientation of $\{\mathbf{x}'_\alpha\}$ with respect to $\{\mathbf{x}_\alpha\}$. Writing

$$\mathbf{q} \equiv \begin{pmatrix} q_0 \\ \mathbf{q}_v \end{pmatrix} = \begin{pmatrix} \cos(\phi/2) \\ \sin(\phi/2)\mathbf{n} \end{pmatrix}. \quad (6.4)$$

one sees that the resulting quaternion defines the rotation angle ϕ and the corresponding rotation axis $\mathbf{n}$, which is at the same time the direction of the screw motion, according to the theorem of Chasles. The proof of the latter can be elegantly given by using the quaternion calculus [114]. The largest eigenvalue λ_{max} describes the "worst" superposition and gives the maximal Euclidean distance between the two peptide planes. We use the latter to define a unique *orientational distance* via

$$\Delta = \sqrt{\frac{\sum_{\alpha=1}^{3} (\mathbf{x}_\alpha - \mathbf{x}'_\alpha)^2}{\lambda_{max}}}. \quad (6.5)$$

By definition $0 \leqslant \Delta \leqslant 1$.

To characterize protein secondary structure we use the following parameters:

1. The orientational distance of consecutive peptide planes, which is defined in Eq. (6.5).

2. The radius of the cylindrical surface on which the reference atom (atom C) moves performing the screw motion between two consecutive peptide planes,

$$\rho = \frac{|\mathbf{t}_\perp|}{2} \sqrt{1 + \cot^2(\phi/2)}. \quad (6.6)$$

Here $\mathbf{t}_\perp$ is the component of the vector $\mathbf{t}$ relating the C-atoms, which is perpendicular to the rotation axis defined by $\mathbf{n}$.

3. The straightness parameter σ. For residue i the latter is defined as

$$\sigma_i = \boldsymbol{\mu}_i^T \cdot \boldsymbol{\mu}_{i+1}, \tag{6.7}$$

where

$$\boldsymbol{\mu}_i = \frac{\mathbf{R}_{i+1}^{\perp} - \mathbf{R}_i^{\perp}}{|\mathbf{R}_{i+1}^{\perp} - \mathbf{R}_i^{\perp}|} \tag{6.8}$$

and $\mathbf{R}_i^{\perp}$ is the point on the helix axis, which is closest to the C-atom of peptide plane i. The straightness gives information about the curvature of a secondary structure element.

Figure 30(Panel A) gives an illustration of the helix (screw motion) parameters defined above. By definition the C-atoms of the peptide planes are on the surface of the cylinder defining the envelope of the screw motion. Figure 30 (Panel B) gives a sketch of a typical ScrewFit profile of a protein (in this figure, bovine pancreatic trypsin inhibitor).

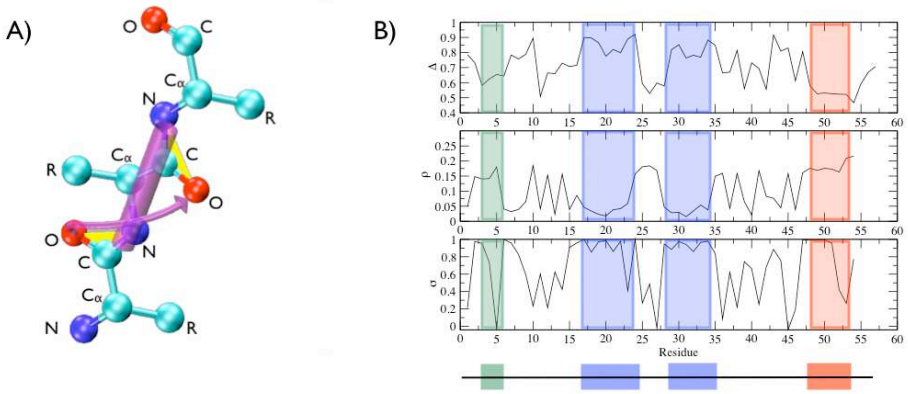

Figure 30: **Panel A)**A tri-peptide with two peptide bonds in the extended conformation. The yellow triangles formed by the atoms $\{O, C, N\}$ of the peptide planes define the local helix structure of the polypeptide. The green spheres, labeled with "R", indicate dummy atoms replacing the side-chains. The radius of the cylinder shown in figure defines the radius ρ of the screw motion relating the two consecutive peptide planes. **Panel B)** A typical ScrewFit profile for a protein structure. In this figure the three parameters refers to the BPTI (PDB code: 4PTI). The colored stripes on the bottom of the figure indicate the secondary structure motifs determined by DSSP method (β-strands: blue; α-helices: red; $3-10$-helices: green). Vertical stripes superimposed on ScrewFit plots represent motifs detected by the latter method (same coloring scheme as for the DSSP).

6.2.2 *Availability*

ScrewFit is implemented in a Python open-source code which uses several modules from the Molecular Modelling Toolkit library [81][1]. A Web-based implementation is under construction.

6.2.3 *Databases*

We created two subsets of the SCOP structural classification with "*all-*α" and "*all-*β" proteins [145], which contain protein domains whose sequence have less than 40% of identity and whose structure elements are essentially α-helices and β-strands, respectively. For this purpose we used a subset of the ASTRAL [31] database which contains PDB-style files for each SCOP classified domain. From the original database, all structures with non-canonical atom notations were corrected or erased.

After these reductions we obtained two databases:

- $\mathbb{A}$, containing 1027 *all-*α domains
- $\mathbb{B}$, containing 1336 *all-*β domains

Every item into these databases is made only by the structurally significant domains and not by the whole proteins structure. The distribution for the Δ and ρ parameters obtained with these databases are represented in Figure 31.

To validate the variation of the ScrewFit parameters and their accuracy obtained by the approach described in the next section we will use three other sets of proteins with a limited overlap with our original databases and with a greater structural heterogeneity. For this purpose we chose:

- $\mathbb{C}$, a subset from the database PDBSelect25[84], containing 2144 chains with sequence homology lower than 25% and non-redundant folds
- $\mathbb{D}$, a subset obtained directly from the PDB by the culling server PISCES [203][204], containing 2477 chains with sequence homology lower than 25% and experimental resolution in the range $0.5 - 2.0$Å
- $\mathbb{E}$, a subset also obtained via PISCES, containing 1829 chains with sequence homology lower than 25% and experimental resolution in the range $2.1 - 3.5$Å

1 Both codes are freely available at http://dirac.cnrs-orleans.fr/software.html

Each entry in these databases $\mathbb{C}, \mathbb{D}, \mathbb{E}$ is chosen according the same rules which were applied to constitute databases $\mathbb{A}$ and $\mathbb{B}$. The database $\mathbb{C}$ is intended for reproduction of most of the structural heterogeneity present in the whole PDB database.

Finally, a finer comparison between ScrewFit and DDSPcont will be performed to verify the consistency of assessments made by ScrewFit. For this purpose, we will refer to the same database of NMR entries used in the original work by Andersen et al. [5]. We will call it :

- $\mathbb{F}$, containing 211 chains from NMR structures each containing at least ten models.

6.3 RESULTS AND DISCUSSION

6.3.1 *Evaluation of* natural *parameters*

The algorithm described above has already been tested in [114] for model structures of polypeptides taken from the Jena Library of Biological Macromolecules (available at [2]) and the resulting parameters will be considered as *ideal* values in the following. Comparison with values obtained from *real* protein structures shows that the parameters fluctuate substantially around the ideal values, even in well definined structural motifs [114]. Secondary structure should, therefore, be associated with a *range* of possible values of the three parameters. In order to find these ranges we have applied ScrewFit to the large number of protein domains of known and well classified structures, which belong to databases $\mathbb{A}$ and $\mathbb{B}$ (see Figure 31).

Each peak of the distributions corresponds to the dominant structural pattern in the respective database and its width gives an estimation of the natural variation of the geometrical parameter which is considered. To quantify the latter we fitted the dominant peaks in each distribution by a Gaussian function,

$$y = a \cdot \exp\left[-\frac{(x-m)^2}{s^2}\right] \tag{6.9}$$

Here m is the position of the peak and s its width, which is to be compared with the parameters for model structures listed in Table 13.

The reader should be aware that 3_{10} and π helices are sometimes present in all-α domains, but they are clearly under-represented in our database $\mathbb{A}$. As shown in [114], these secondary structure elements can however be unambiguously distinguished by ScrewFit applied on their ideal structures. The same is true for the extended conformation with respect to the

2 Available at http://www.imb-jena.de/IMAGE.html

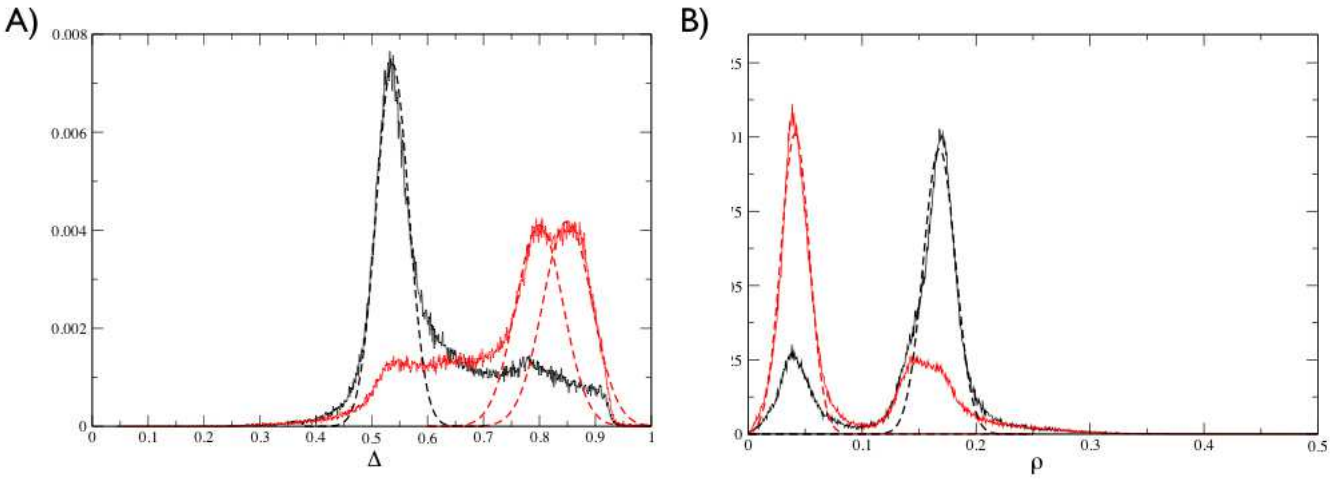

Figure 31: **Panel A) :** Normalized distributions for *Orientational Distance* Δ in 𝔸 (in **black**) and in 𝔹 (in **red**). Fitted Gaussian functions are shown in dashed lines. Correlation coefficients between distributions and fitted curves are between 0.90 and 0.96. **Panel B) :** Nomalized distribution for the *helix radii* ρ in 𝔸 (in **black**) and in 𝔹 (in **red**). Fitted Gaussian functions are shown in dashed lines. Correlation coefficients are between 0.94 and 0.96.

β-strand. From Figure 31 (Panel b), it is also evident that ρ parameter distribution for each database presents an additional peak at values far from the dominant one. This fact is due to the presence of some residues in the extended conformation and reverse turns (α, β and γ type) in both databases.

Table 13: ScrewFit parameters for different structure motifs from model polypeptide and from database evaluation.Values for 3_{10} and π are only derived by the model polypeptide as given in [114]. More details in the text.

Motive	ρ_{ideal}	ρ	Δ_{ideal}	Δ
α-helix	0.171	0.168 ± 0.020	0.582	0.537 ± 0.041
3-10 helix	0.146	*0.146*	0.670	*0.670*
π-helix	0.178	*0.178*	0.471	*0.471*
β-strand	0.055	0.041 ± 0.018	0.875	0.850 ± 0.062
extended	0.037	*(0.041 ± 0.018)*	0.754	0.800 ± 0.057

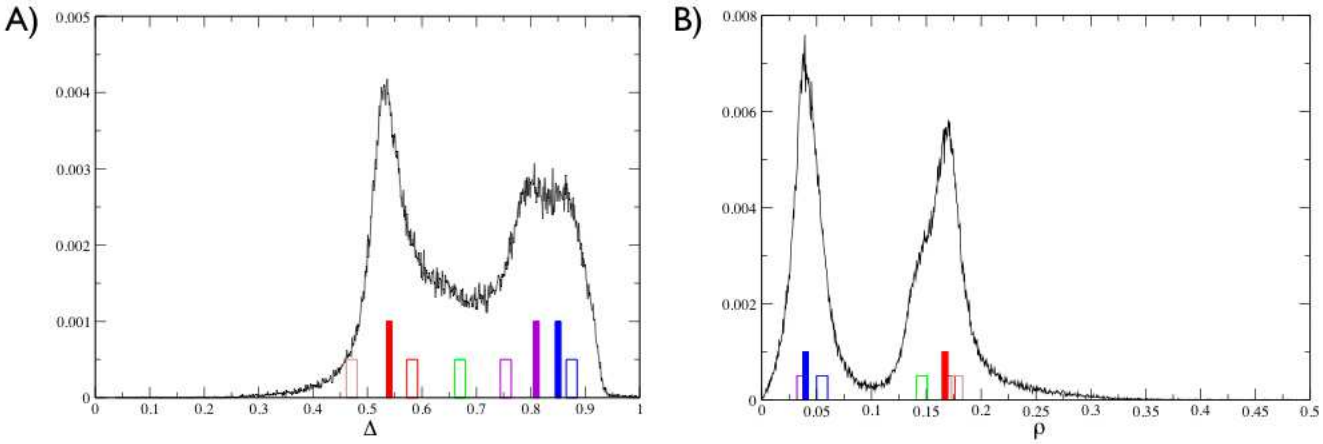

Figure 32: Normalized distributions of Δ (panel **A**) and ρ (panel **B**) parameters calculated on a heterogeneous selection of protein structures obtained from the PDBselect25 database ($\mathbb{C}$). The full-colored bins represent the *natural* values obtained from calculation on $\mathbb{A}$ and $\mathbb{B}$. The hollow bins represent the *ideal* values reprinted from **??**. The color-scheme used is: red for α-helices, blue for β-strands, violet for extended conformations, green for 3_{10}-helices and brown for π-helices.

Figure 32 shows the normalized distributions of the parameters Δ and ρ obtained from the application of ScrewFit to the database $\mathbb{C}$ introduced in the previous section. Despite their origin from a wider spectrum of structural domains, these distributions are very similar to a superposition of those obtained from $\mathbb{A}$ and $\mathbb{B}$. Interestingly, the peaks from $\mathbb{C}$ are also very close to those from $\mathbb{A}$ and $\mathbb{B}$ which are reported in Figure 32 as full-colored bins.

The ideal values reported in Table 13 and obtained from previous works [114] are also reported in Figure 32 as hollows bins. Here it is worth noting that 3_{10} and π helices, which were under-represented in our original databases $\mathbb{A}$ and $\mathbb{B}$, have ideal values slightly different from the *natural* value obtained for the α-helices6.5. This observation supports the assumption that even using these ideal values one can easily characterize an helicoidal pattern and distinguish if it belongs to one of the three most frequent classes.

As a whole, these facts confirm that the empirical definitions made for structural elements and summarized in Table 13 allows to distinguish the structural elements in several types of protein folds.

6.3.2 *Reliability and consistency of ScrewFit assignments*

In the following we want to confirm the reliability of the *natural* parameters in Table 13, comparing ScrewFit to other methods and with respect to the refinement of experimental resolution.

It is worth noting here that a wide variety of methods have been used in the past to make discrete secondary structure assessments based on different approach: intra-backbone hydrogen bonds (DSSP), expert assignements and backbone dihedral angles (STRIDE[57]), C_α coordinates (P-SEA[123] and DEFINE[163]) and protein curvature (P-curve [181]). Observations made by Colloc'h et al. [40] showed that DSSP, DEFINE and P-curve share the same assessment in 63% of cases whereas DSSP and STRIDE agree for 96%. As for several reasons, DSSP is commonly considered the standard reference in the field, we will limit our comparison to it, in its discrete (simply DSSP in the following) and continuous (DSSPcont, in the following) forms, and two other methods with different approach and assume that all other comparisons can be inferred using the results showed in [40].

In their original work, Kabsh and Sander [106] applied DSSP on three different crystallographic structures of decreasing resolution from 1.5 A to 3.0 A in order to study its accurancy against experimental resolutions. In this work we use the same three structures to compare our method with DSSP and with the method by Levitt and Greer (LG) [128] also referred in [106]. The latter is another method based on distance dependent definitions of the secondary structure motifs and all results obtained from its application are simply reprinted from reference [106] . The reader should be aware that two of these structures have now been superseded by new and more accurated ones. Here we use the older structures to refer directly to the original work by Kabsch and Sander as this will not affect our conclusions. Nevertheless we also compare the ScrewFit analysis on one of these structures (PDB code: 2ADK) with its homologue obtained by the more refined structure (PDB code: 3ADK) to make further assessments of the effects of the experimental resolution on our method.

Table 14 shows the results of ScrewFit on pancreatic trypsin inhibitor at 1.5 Å resolution (PDB code: 3PTI) and cytochrome c550 at 2.5 Å resolution (PDB code: 155C).

The comparison of assignments ScrewFit with those taken from [106] for DSSP and LG, shows that the three methods are globally equivalent in detecting secondary structure motifs on the structure at higher resolution but some relevant discrepancies among them begin to appear at the 2.5 Å resolution structure. In particular, this fact is evident in the region 26-31 of cytochrome c550, where LG and DSSP do not find the same β-strand and ScrewFit detects

Table 14: Comparison of ScrewFit assignments with two other methods (Levitt&Greer and DSSP/DDSPcont) for two structures of decreasing resolution. Assessments for both DSSP and DSSPcont are showed only when they differ. *Comments* column refers to additional information obtained by ScrewFit analysis. LG stands for Levitt and Greer method.

Structure	LG	DSSP/DSSPcont	ScrewFit	comments
3PTI Res. 1.5 Å				
3_{10}-helix	2-7	3-6	2-5	
β-strand	14-25	18-24	16-23	
β-strand	28-37	29-35	29-34	
β-strand	43-46	45	45	β-turn
α-helix	47-55	48-55	48-57	
155C Res. 2.5 Å				
α-helix	4-16	6-12	6-10	
3_{10}-helix	-	11-13/ -	11-13	
β-strand	17-23	19-20	19-20	
β-strand	26-31	-	26-30	extended conf.
β-strand	33-39	35-37	-	
α-helix	40-44	-	-	
α-helix	55-65	56-64	57-63	curved C-term.
α-helix	71-80	73-80	72-82	3_{10}-like C-term
α-helix	81-90	-	-	
α-helix	106-118	107-117	106-116	curved C-term.

a region of some β-strand conformations alternated with some extended ones. Equivalent results are obtained also for α-helices in regions 57-63, 72-82 and 106-116 where ScrewFit can give quantitative assessments on the curvature and other terminal deformations of each motif.

Table 15: ScrewFit assignements for the same protein structure with two different experimental resolution. Assessments for both DSSP and DSSPcont are showed only when they differ and assessments by DSSP are printed in italics.

	2ADK Res. 3.0 Å				**3ADK** Res. 2.1 Å			
Struct.	DSSPcont	STRIDE	P-SEA	ScrewFit	DSSPcont	STRIDE	P-SEA	ScrewFit
α-helix	2-7	2-7	2-6	2-6	2-6	1-7	2-7	2-6
β-strand	10-14	10-14	9-13	10-14	10-15	10-15	9-16	7-16
α-helix	*23-30*/23-31	21-31	23-31	23-29	21-31	21-32	21-31	17-30
β-strand	35-38	35-38	34-38	*35-38*[1]	35-38	35-38	34-38	33-38
α-helix	39-48	39-49	41-49	39-47	39-49	39-49	39-49	38-49
α-helix	52-62	52-62	52-62	52-61	52-61	52-63	52-63	52-61
β-strand	-	-	-	-	-	-	65-68	65-68
α-helix	69-83	69-84	69-83	70-83	69-81	69-82	69-83	72-81
β-strand	90-93	90-93	89-92	86-93	90-93	90-93	89-93	88-93
α-helix	101-108	101-107	101-108	101-108	99-108	99-108	99-108	99-108
β-strand	-	-	-	-	-	-	109-112	109-112
β-strand	114-118	114-118	114-121	113-118	114-119	114-119	114-117	113-120
α-helix	122-132	122-133	122-133	122-130	122-136	122-136	122-136	122-136
α-helix	143-157	143-166	144-155	143-155	146-156	145-157	144-155	145-155
α-helix	160-167	-	-	156-165	158-164	159-165	158-167	159-164
β-strand	170-173	170-173	168-175	170-173	170-174	170-174	170-173	170-174
α-helix	179-193	179-193	182-193	179-193	179-191	179-192	179-192	178-191

In both cases assignments made by DSSP and DSSPcont did not show any significant differences. A more detailed comparison with respect to experimental resolution is made in Table 15. The accuracy of ScrewFit assignments on adelynate kinase structure at different resolution (PDB codes: 2ADK and 3ADK) was compared with those made by the means of three other methods: DSSP, STRIDE and P-SEA.
Globally the effects of resolution on these methods are the same and only some discrepancies in N- and C- edges were encountered. This fact is already well known in literature as an artifact due to the different definitions [40].
Nevertheless, a more accurate analysis of this comparison shows different response of each method to the improved resolution. Firstly, at low resolution the assignments made by ScrewFit overlap very well with those made by DSSP except for a slightly deformed β-strand in 35-38 which makes ScrewFit assessment quite uncertain. Secondly, DSSP and ScrewFit detect also two α-helices instead of one unique helix in the region 143-167. This outcome is confirmed for the four methods at higher resolution.
Additionally, from Table 15 is also evident that P-SEA and ScrewFit improve their accuracy with resolution detecting two new β-strands in 65-68 and 109-112.
These results confirm those previously shown in reference [114] and suggest that ScrewFit performance is equivalent to those obtained by the more refined distance-dependent methods if it is applied on a structure at low resolution. If the resolution is improved on the same structure, ScrewFit shows instead a significant conservative refinement with respect to the lower resolution structure and it also adds some new motifs with partial convergence with the common distance-dependent methods.

To confirm the consistency of ScrewFit assignements with respect to the experimental resolution, we plotted the normalized distribution of values obtained for the parameters Δ and ρ calculated over all structures in databases $\mathbb{D}$ (resolution less than 2.0Å) and $\mathbb{E}$ (resolution between 2.1Åand 3.5Å). As shown in Figure 33 the effect of a better resolution is the slight narrowing of some peaks in the distributions, in particular for those corresponding to the α-helices values. Interestingly, assignments on β strands seem to be not affected by the experimental resolution. It is worth noting instead that although the shape of the peak distributions slightly changed, the position of their maxima stands unchanged. As a whole, the outcome of the comparison showed in Table 15 and the calculation over the databases $\mathbb{D}$ and $\mathbb{E}$ proved that the efficiency of ScrewFit assessments is globally quite insensitive to the experimental resolution with which the structural data were obtained.

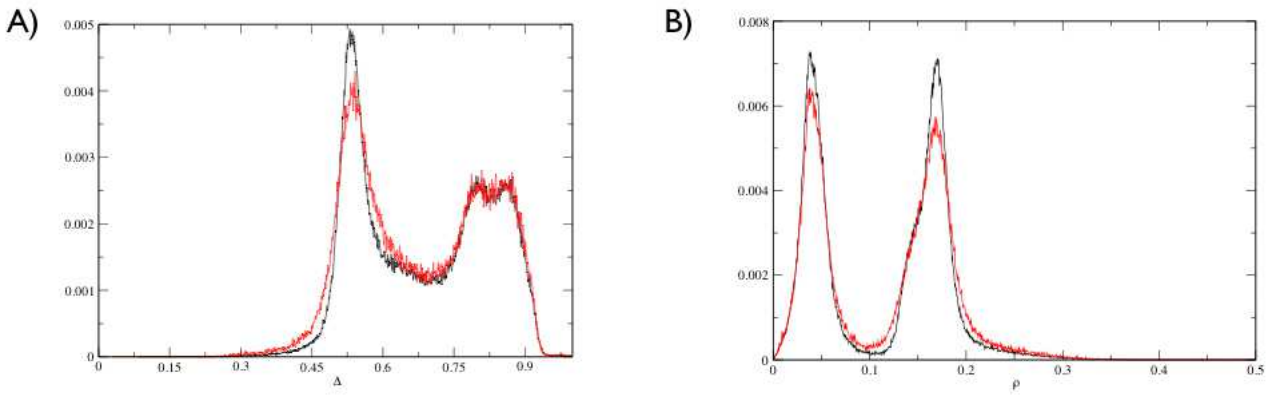

Figure 33: Normalized distributions of Δ (panel **A**) and ρ (panel **B**) parameters calculated over databases $\mathbb{C}$ (experimental resolution lower than 2.0Å; black lines) and $\mathbb{D}$ (experimental resolution greater than 2.1Å; red lines).

6.3.3 *Comparison with DSSPcont*

As we mentioned in the Introduction, the main problem with the most common methods for the secondary structure assignments is the use of threshold-based definitions which recast the structural conformation heterogeneity of protein backbones into eight or less classes without accounting for the effect of the natural variations that could occur into the structures due, for example, to thermal fluctuations.

Andersen and coworkers [5] presented an improved version of the DSSP method in which the discrete assignments were substituted by continuous ones obtained from the former ones variating of the hydrogen-bond distance thresholds defined for each of the eight DSSP structural classes. The efficency of DSSPcont method in capturing the effects of thermal fluctuations in NMR structure models has been shown to be higher than that of the discrete DSSP [5].

Here we discuss the consistency of the ScrewFit assessments with the same criteria used by Andersen and coworkers and in particular we verify the stability of our set of parameters over the database $\mathbb{F}$.

For each NMR structure in $\mathbb{F}$, we first calculated (for both parameters Δ and ρ) the standard deviation of values over all the models. We then grouped these results when the corresponding mean values of the parameters fell within one of the ranges defined in Table 13. The averages over these groups gave an estimation of the stability of parameters Δ and ρ

as a function of their values and could be compared to those obtained for DSSPcont in [5]. Results for α-helices, β-strands and extended conformation are presented in Table 6.3.3.

Table 16: Comparison of assessment consistency between ScrewFit and DSSPcont by the means of the root mean square deviation from average assignments. Value for DSSPcont are reprinted from [5]. Lines labeled with *def* refer to amplitudes of the ranges presented in Table 13 with respect to the corresponding parameter's value. Extended conformation is given with the same values of β-strands when a differentiation between the two cannot be made. All values are in percentage.

	α-helix	β-strand	extended
DSSPcont	13.1	11.3	*11.3*
ScrewFit -ρ - NMR	13.7	29.3	*29.3*
ScrewFit -ρ - *def*	11.9	44.0	*44.0*
ScrewFit -Δ - NMR	9.1	4.4	8.2
ScrewFit -Δ - *def*	7.6	7.2	7.1

The first outcome of this analysis is that consistency of assignments in ScrewFit seems to be higher than in DSSPcont only for parameter Δ and not for ρ. In fact, for the latter, the large deviation from the average values is clearly due to the lack of two distinct ranges for β-strands and for extended conformations. This is a direct effect of using continuous ranges of *natural* values for the definition of secondary structure elements and these differences between parameter consistency prove that in ScrewFit the combination of the two parameters, Δ and ρ, (combined with σ) is essential to make precise assessments.
Moreover, Table 6.3.3 shows that, independently of the type of structural element, the root mean square differences for NMR models is coherent with the relative amplitudes of the ranges defined in Table 13. Nevertheless, it is worth to note that the use of ranges did not permit the same calculation for less represented structural elements like 3_{10} and π helices and the comparison with DSSPcont with respect to the consistency of assignments remains incomplete. Corresponding work is in progress in order to obtain comparison also on the other DSSP structural classes.

6.3.4 *Application*

In the following we will show a simple application of our method. This example underlines the combined approach which allows assessments on the protein secondary structure and a quantitative characterization of its changes within the same tool and without any further data treatment.

BPTI Crystal forms

As reported in our previous work [114], ScrewFit can be used to pinpoint small conformational changes due to several causes ranging from different experimental techniques to external environment changes like, for example, ligand binding or variation in pH. As an example we will consider here the case of the three different crystal forms of bovine pancreatic trypsin inhibitor (BPTI) which are known to be originated by different pH condition: pH values lower than 9.35 favor the form I and II crystals whereas higher pH values favor the form III.
These three different crystal structures induce some small changes in protein conformation as reported in [209, 210]. In the original works, the authors compared the three structures of BPTI by means of inter-atomic distances in the backbone and/or in the side-chains. The outcome of this type of analysis was the finding of several local changes induced by the different crystal forms. Nevertheless, a detailed structural characterization of those changes were not possible.
Here, we will first analyze the secondary structure assignments made by ScrewFit on the structure obtained from the crystal form I and we will then discuss the differences with the other two structures.

For the determination of secondary structure motifs, we confront again the ScrewFit parameters with the secondary structure elements found by the DSSPcont method. The results are shown in Figure 30 (Panel B). The motifs found by DSSPcont are shown on the bottom of the figure. We used the same color scheme for both methods: red for α-helices, blue for β-strands and green for 3-10 helices. Visual inspection shows that most of the motifs found by DSSPcont correspond to the regions where the ScrewFit parameters are enclosed by the *natural* variations. Interestingly, our method is able to identify very well the α-turn between the two beta strands 19-25 and 30-36. These regions are highlighted in Figure 30 (Panel B) with colored vertical bands. Minor discrepancies can, however, be observed in some regions, expecially in the extreme residues of each element. In particular, ScrewFit finds that the assignment of the 3_{10} between residues 3 and 6, is quite uncertain due to the highly

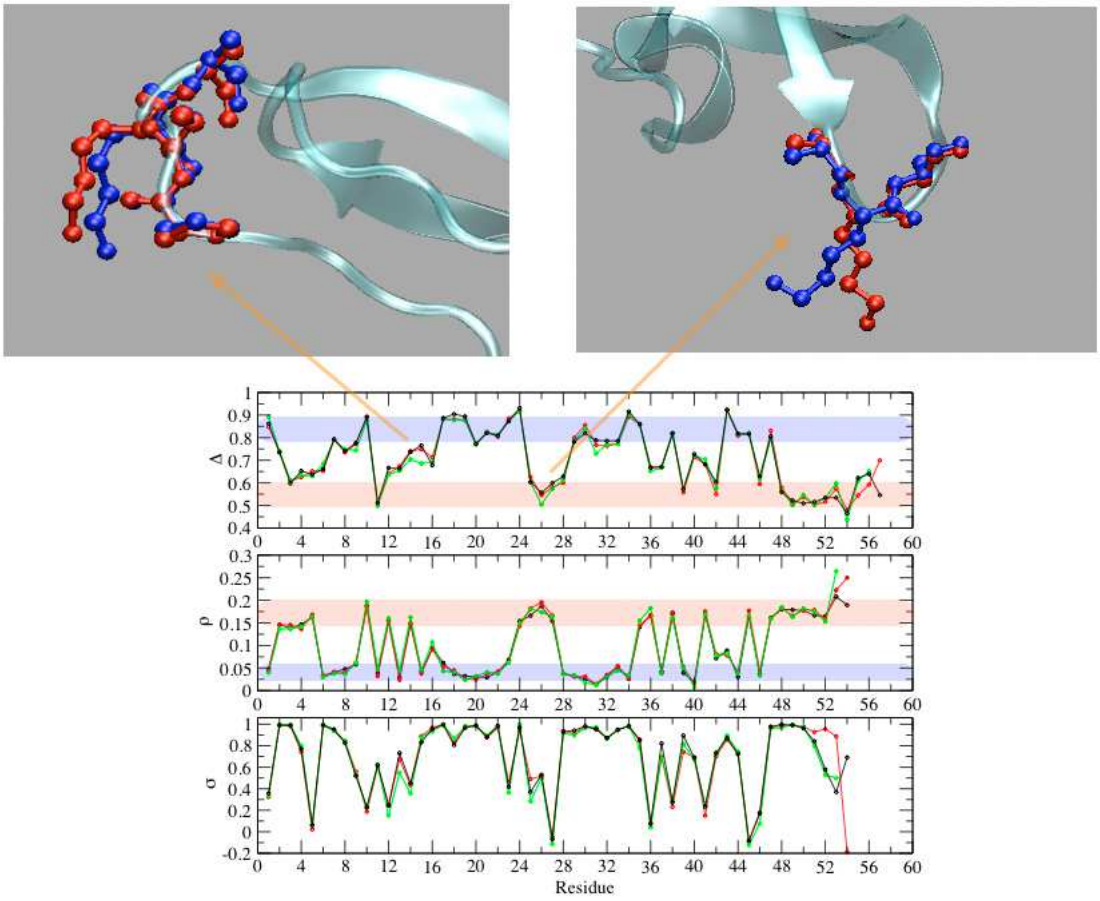

Figure 34: ScrewFit profile for bovine pancreatic trypsin inhibitor (BPTI) in its three different crystal forms. This figure shows how ScrewFit is able to pinpoint small differences between similar structural configuration of the same protein. Here Crystal form I (black solid line, PDB: 4PTI), form II (red solid line, PDB: 5PTI) and forms III (green solid line, PDB: 6PTI) are represented. Horizontal stripes define the *natural* variations for β-strands (colored in blue) and α-helices (colored in red).

deformed C-terminus which is characterized by abnormally high values for this structural element in parameters Δ and ρ and by the value of σ very close to zero in residue 6 which evidently stands for a kink in the backbone profile.

Let us now look at the effects of the three crystal forms on the structure of the BPTI. Figure 34 shows the parameters calculated in the three different conformations of the protein. In each plot, the horizontal stripes indicate the range of variation for the ScrewFit parameters as found in databases $\mathbb{A}$ and $\mathbb{B}$ (the usual color-code is used here: α-helix highlighted in red; β-strands highlighted in blue; 3_{10}-helix highlighted in green). In the original works the authors reported some major differences between form I and form II in the regions 15-19, 26-29, 39-41 and 47-50 and between form II and form III mainly localized in residues 15 and 26. It is worth noting that deviations of backbone close to residue 15 are particularly interesting, since this residue is part of the active site of the BPTI. The application of ScrewFit to the three structures confirms some the differences mentioned above and allows them to be better characterized. In particular, changes around both residue 15 and 26 show major discrepancies between the conformation in crystal form III and the other two. In residue 15, parameter Δ changes, from form I and II to form III, toward values closer to an *helix-turn* conformation but the absence of equivalent changes in the other two parameters proves that this change is only due to a different relative orientation of the peptide planes related to a reorientation of the side-chains bewteen residues 15 and 16. The analysis of parameters' values for residue 26 leads to similar conclusions. A confirmation of these findings can be done with a direct visualization of the structure as showed in Figure 34. Additionally, ScrewFit reveals also a difference between the crystal form II and the other two at the C-terminus of the α-helix in region 47-53. In this case the parameter σ shows that in forms II the helix axis is rather straight whereas in the other two forms it clearly bends.

The application of DSSPcont on the same structures confirmed that the major differences between crystal form I and crystal form II were localized in the regions around residues 15 and 26 where it is reported an higher probability, in form II rather than in form I, of having respectively an not-structured and a helix-turn conformation. Nevertheless, DSSPcont does not find any relevant change between crystal form II and form III which was detected by ScrewFit.

6.4 CONCLUSION

We presented the ScrewFit method for the analysis of the secondary structure level of proteins, which describes the latter in terms of local helix parameters obtained by the screw motions relating consecutive peptide planes. By construction, this method represents, *ipso facto*, a new approach to the so-called continuous methods for the secondary structure assessments because it does not relate to beforehand threshold-based definitions but only to geometrical criteria which are verified *a posteriori* on natural secondary structure elements.

To establish confidence levels for the definition of all types of elements, we applied our method to different well defined selections of protein structures. The natural variations in parameters of ScrewFit have been shown to be coherent with the ideal values obtained in a precedent work.

We have then confirmed the reliability of our definitions by applying our method on a heterogeneous set of proteins with different type of folds. In order to show the efficiency and accuracy of ScrewFit we confronted it to other methods, in particular to DSSP and DSSPcont. The outcome of this comparison showed that ScrewFit can be considered as accurate as DSSP/DSSPcont in function of crystallographic resolution. As an example, we analyzed the three different crystal forms of the bovine pancreatic trypsin inhibitor by means of ScrewFit. This application has shown that ScrewFit finds essentially the same structural elements as DSSP, but gives also a more detailed description of them, leading in some cases to a different assessment of secondary structure elements. It is worth noting that the method is also shown to be efficient in detecting some kind of reverse turns and random coils. Further improvements to permit more detailed assignments also of these motifs are the subject of work in progress.

As a whole, with the same application we showed that ScrewFit is able to pinpoint small structural changes and to give a global view of the structural rearrangements of the protein as a response to external changes.

In this context, another important feature of the method is the fact of combining in the same tool a method for detecting secondary structure and a way to find and characterize any induced structural change.

It should be mentioned that while ScrewFit was developed for the analysis of protein secondary structures, with some minor improvements, it may also be used to characterize the protein supersecondary motifs and the folds of DNA and RNA molecules. Corresponding work is also in progress.

6.5 NOTES

[1] It can be easily verified that the ranges found for the parameter for ρ in α-helices and in 3_{10}-helices are fully compatible with those obtained by Barlow and Thornton [9] [see section 5.3.2] on a rather smaller set of $\sim$ 100 structures.

7

RESULTS

As stated in the introduction, one of the principal aims of this work was the characterization of both dynamical and structural aspects in protein adaptation to environmental conditions. In the present chapter, results obtained in this thesis for the particular case of the IF6s will be presented and discussed.
After a summary on the samples used in this work and the environmental conditions investigated, the discussion will focus on the analysis of structural and dynamical effects of pressure and temperature on each sample. The results will be used to determine both local and global effects of the environmental conditions to IF6 homologues. The *link* between the assessment of both types of effects will be assured by the complementarity of the insights obtained by molecular dynamics (MD) and quasielastic neutron scattering (QENS).

The samples that have been studied either by MD simulations or by QENS experiments are:

- **aIF6**: the extremophile IF6 homologue from *Methanococcus Jannaschii*.
- **aIF6-HTag**: the extremophile homologue with the supplemental N-terminal poly-histidine tag (HTag). This sample was used in the high pressure QENS experiments and was needed during the production phase in order to enhance the expression protocol yield.
- **eIF6**: the modeled mesophile IF6 homologue from *Saccharomyces cerevisiae*.
- **eIF6-NoCTAIL**: the mesophile homologue with C-terminal cleaved.

The different environmental conditions used in MD are listed in Table 17, whereas those used in QENS measurements are shown in Table 18. It is worth noting here that the number and the variety of experimental measurements was largely constrained by several technical facts as will be exposed later in this chapter. Nevertheless, the comparison to results obtained from MD were however possible through a limited set of measurements. The environmental conditions listed in Tables 17 and 18 will be selectively chosen in the next section to present different aspects. Unless explicitly mentioned, the results shown here were obtained from

molecular dyanamics simulations.

Table 17: Set of MD simulations of each sample performed with different environmental configurations

aIF6	eIF6	eIF6-NoCTAIL
300K - 1bar	300K - 1bar	300K - 1bar
300K - 250bar	320K - 1bar	300K 500bar
300K - 500bar	350K - 1bar	350K - 1bar
350K - 1bar	350K - 500bar	350K - 500bar
350K - 250bar		
350K - 500bar		

Table 18: Set of QENS measurements of each sample performed with different environmental configurations

aIF6	eIF6	aIF6-HTag
300K - 1bar	300K - 1bar	300K - 250bar
350K - 1bar	350K - 1bar	300K - 500bar
		350K - 250bar
		350K - 500bar

7.1 EFFECTS OF PRESSURE AND TEMPERATURE CHANGE ON IF6S STRUCTURE

The following section presents a study of the structural response of IF6s to changes in pressure and temperature.

When high pressure and high temperature are applied to protein solution, one of the main effects that one would expect to see is the change in molecular volume and molecular surface exposed to the solvent.

In Figure 36 is shown the volume change induced in aIF6 and eIF6 by a pressure change equivalent to 500bar at 300K and 350K. Here the molecular volume was estimated by the volume related to the surface accessible to solvent molecules. The extremophile IF6 structure seems to be less sensitive to environmental changes than its mesophilic counterpart. This observation is also corroborated by Figure 37 which shows the variation of the radius of gyration (R_{gyr})in eIF6 and aIF6, respectively. This quantity, defined by

$$R_{gyr} = \frac{1}{N}\sum_{i}^{N}(\mathbf{R}_i - \mathbf{R}_{CM})^2, \tag{7.1}$$

is the root mean square distance of atoms from the protein center of mass ($\mathbf{R}_{CM}$). Although R_{gyr} is commonly used to give insight into the global shape of proteins, here it cannot be used to give an exact representation of IF6s, as the latter has a *torus-like* shape (see Figure 35) and the atomic mean distance does not distinguish between changes in either the internal or the external radius of the tours. Nonetheless, R_{gyr} can be used to give a qualitative representation of shape changes in IF6s. Indeed R_{gyr} of aIF6 varies in a narrow range from 16.45 to 16.75 Å, indicating that the protein structure is rather insensitive to environmental changes. In eIF6, R_{gyr} does not significantly change at 300K when pressure is applied whereas it shows a very large variation when pressure is applied at high temperature (350K).

Interestingly radial values of aIF6 at *extremophilic natural conditions* (350K and ~ 500bar) are very similar to those of eIF6 in *mesophilic natural conditions* (300K and 1bar).
Similar observations can be made with respect to the surface accessible area (SASA) shown in Figure 38. The differences in variation of SASA between the two homologues, at 350K reveals that the changes found in R_{gyr} of eIF6 must be mainly related to the increase of the total surface exposed to the solvent.

7.1.1 *Local effects*

The local effects produced by pressure and temperature on IF6s structures can be probed using the root mean square fluctuations $RMSF(i) = \langle \mathbf{R}_i(t) - \langle \mathbf{R}_i \rangle \rangle^2$, where i can refer to either an atom or a group of atoms (e.g., backbones or sidechains); in the latter case an

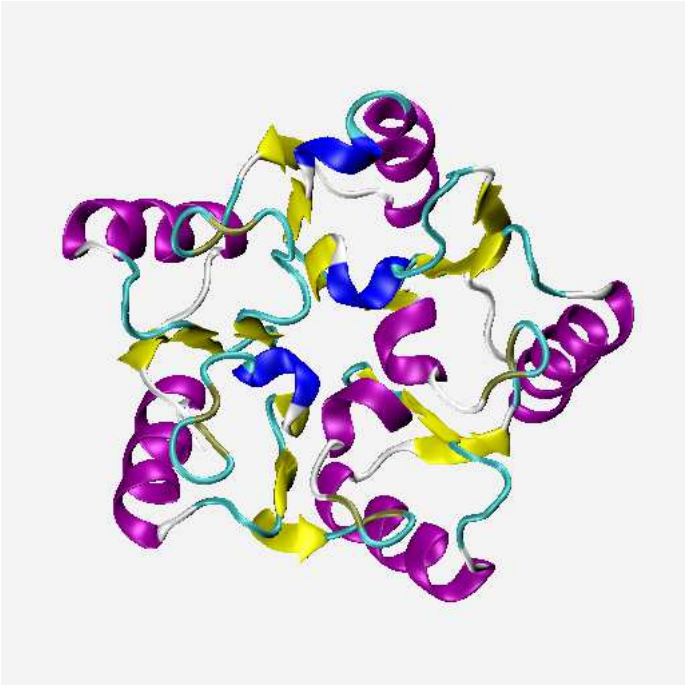

Figure 35: *Cartoons* representation of IF6 structure.

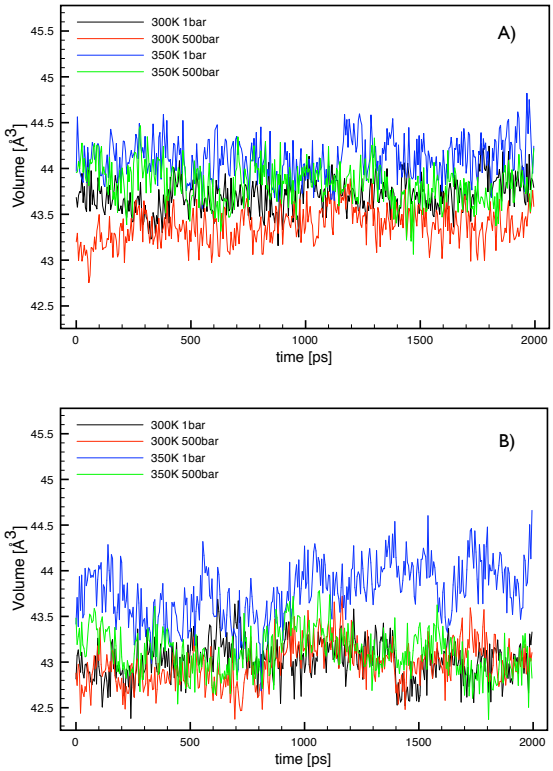

Figure 36: Molecular volume of IF6s with respect of pressure and temperature. Panel A: aIF6. Panel B: eIF6.

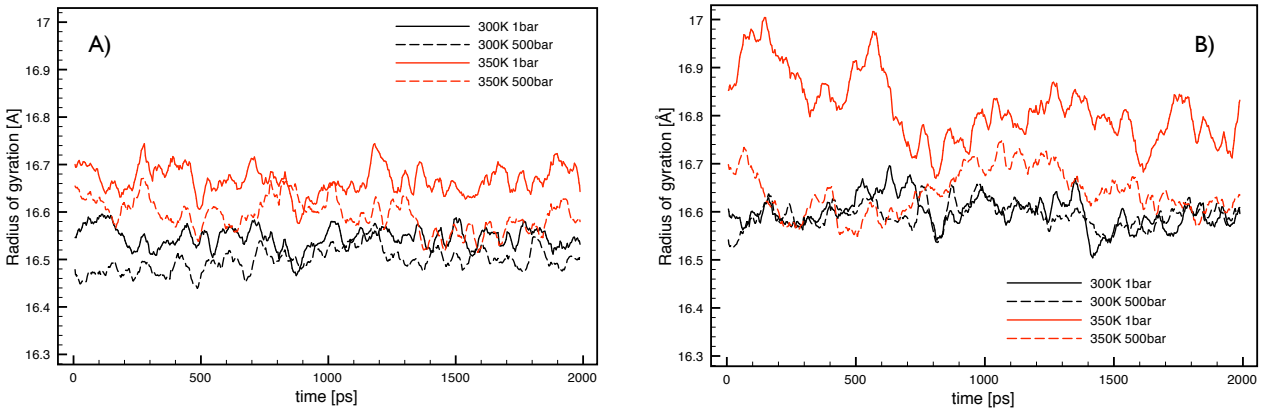

Figure 37: Radius of gyration of IF6s as a function of time. Panel A: aIF6. Panel B: eIF6. Only residues from 1 to 225 are taken into account.

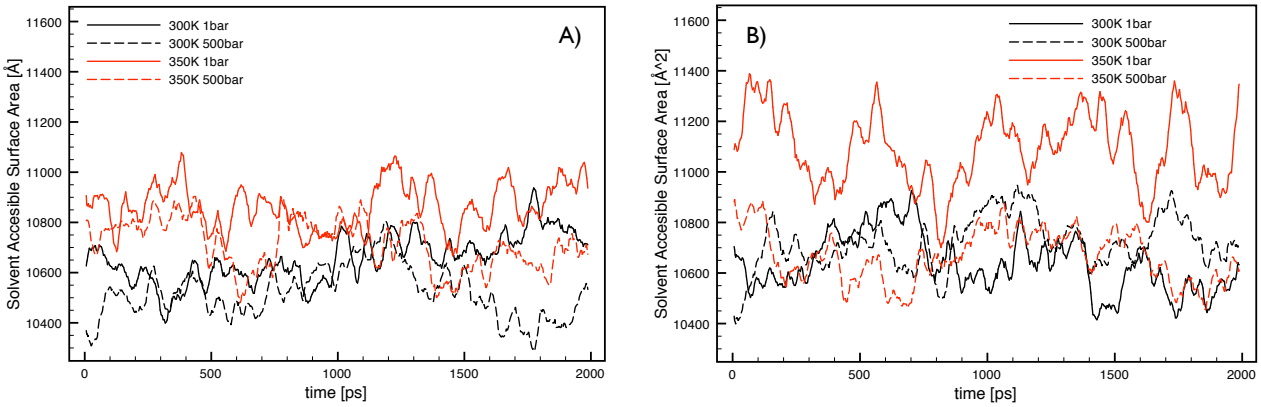

Figure 38: Solvent accessible surface area of IF6s as a function of time. Panel A: aIF6. Panel B: eIF6.

average over the atoms of the group is performed.

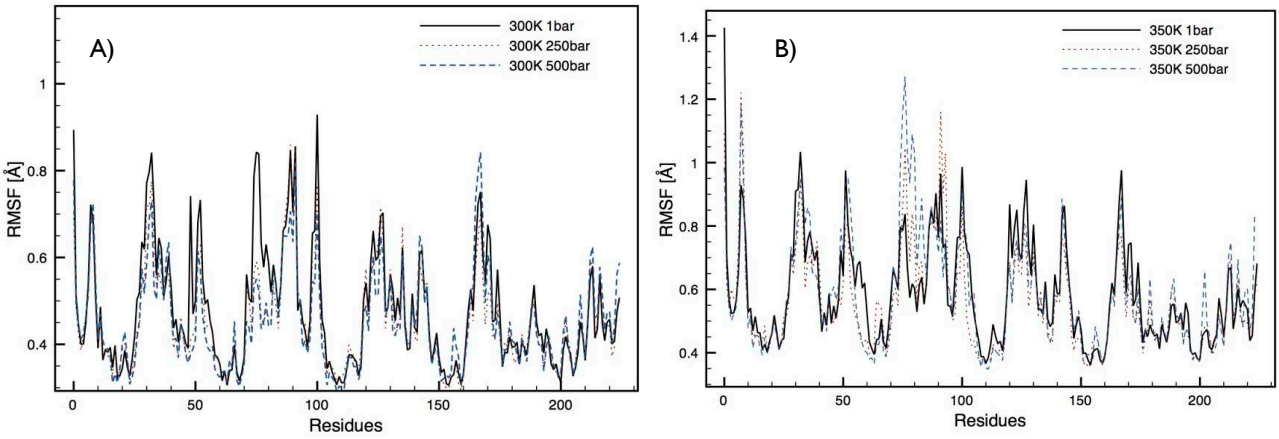

Figure 39: Root-mean-square-fluctuation (RMSF) of carbon C_α in aIF6 backbone.

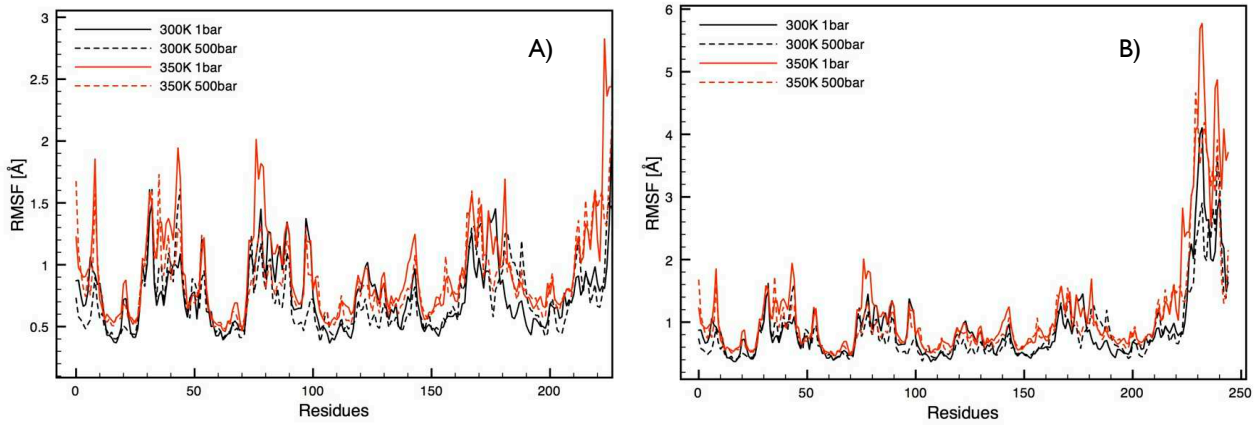

Figure 40: Root-mean-square-fluctuation (RMSF) of carbon C_α in eIF6 backbone. Panel A: Region 1-224. Panel B: Region 1-245.

The RMSF of carbon C_α in aIF6's backbone (Figure 39) shows that at 300K atomic fluctuations of residues are slightly reduced by pressure on the whole structure but a more significant variation is found in the region 80-90 . The latter region corresponds to the α-helix of one of the five pseudo-symmetric subdomains which composed the IF6 structure. No

significant differences are found between RMSF at 250bar and 500bar whereas at 350K the same region 80-90 shows RMSF increased by pressure. Moreover, as one would expect the general effect of temperature is to increase RMSF.
The changes due to pressure found in aIF6 are not present in eIF6 where instead RMSF in region 80-90 is increased by temperature but reduced by pressure.
In both cases, these results lead to the conviction that pressure and temperature can induce large changes in the exposure to the solvent of the α-helix in region 80-90 which results in a higher or lower RMSF of C_α atoms in the region. This observation is supported by visual inspection of both aIF6 and eIF6 sequences which reveals that the α-helix is highly charged and thus it has an high propensity for interaction with solvent.
Figure 40 shows that eIF6, at both 300K and 350K, has an higher *baseline* for C_αs fluctuations, probably due to the presence of the C-terminal tail of 21 amino acids (CTAIL) which could transmit supplemental fluctuations to the rest of the protein structure. In the same context, it is worth noting that at high temperature this *baseline* is particularly increased for residues from 100 to 220, as shown in Figure 40.

7.1.2 *Secondary structure changes*

Even though the calculation of RMSF allowed several qualitative assessments on the effects of pressure and temperature on IF6s' structures, a finer method is necessary to quantify them. For this purpose, the ScrewFit method will be used here to characterize the response of IF6s to the environmental changes. In particular the method will be used to analyze the molecular structures obtained by time-averaging of the atomic trajectories issued from the MD simulations in order to make assessments on the local flexibility of the secondary structure of IF6.
The first application of ScrewFit on the time-averaged structure of both aIF6 and eIF6 at 300K and 1bar allows to distinguish the very similar secondary structure of the five subdomains which compose the IF6 structure (Figure 41 shows the aIF6's profile). The inspection by eye allows the five motifs to be found through their regular profiles. Here, the discussion will be limited to the motif in the region 52-100 but the same conclusions can be obtained also from any of the other structural subdomains.
Groft and coworkers [66], who reported the presence of subdomains in a five-axis pseudo-symmetry in IF6 structures, stated also that the almost invariant secondary structure profile of these subdomains was made of a long α-helix, a shorter one (alternatively a 3_{10}-helix was also

found) and three β-strands. Here the analysis of the orientational distance and of the radius of the helix of screw motion, seems to confirm the presence of a tight α-helix or more likely a 3_{10}-helix in residues 57-60 followed by two β-strands in the region 60-75 (See previous chapters for reference values). The latter have, however, a very variable straightness, meaning that they have a very curved form, as one can verify by visual inspection on molecular structure.

Residues 78-90 clearly form a long and straight α-helix as evident from the comparison of values of the three parameters. Nonetheless, the C-terminus of this long α-helix shows a curved region as shown again by the straightness parameter which attains values close to zero. This result confirms and completes the observation made on the RMSF in the region around residue 90 in both aIF6 and eIF6. Indeed, variations found in the RMSF must correspond in changes in the curvature of the C-terminal region of the α-helix. The subdomain ends with a rather straight short β-strand from residue 90 to 100.

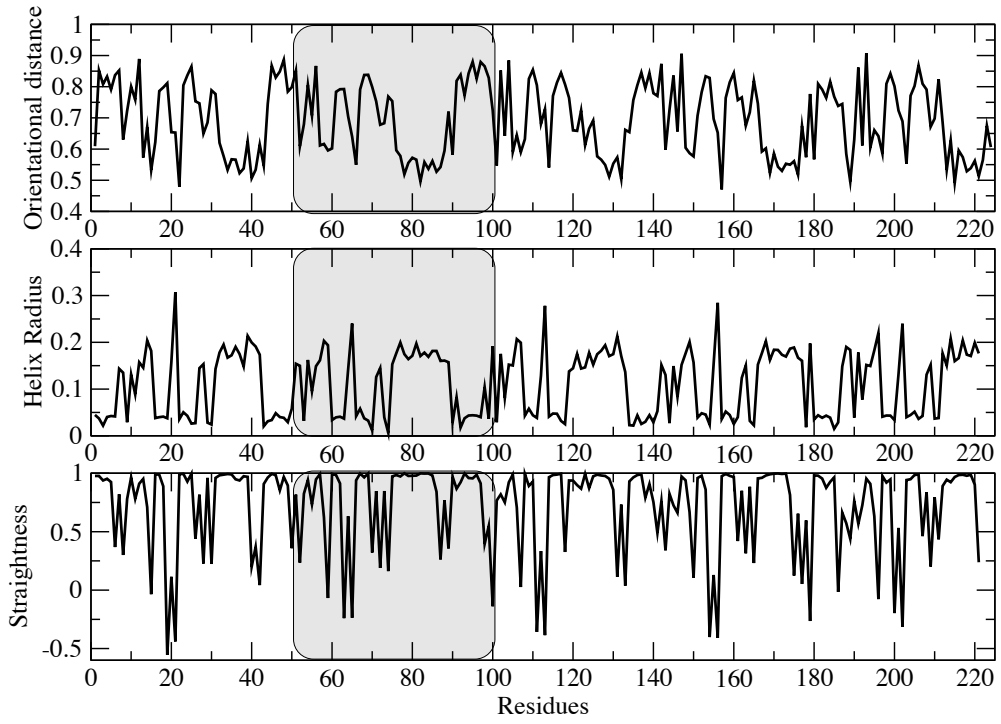

Figure 41: Secondary structure profile of aIF6 as detected by ScrewFit algorithm. The subdomain 52-100 is highlighted by grey stripes.

Once the profile of IF6 secondary structure is characterized by ScrewFit parameters, one can observe the changes of the same parameters on the effect of high pressure and temperature. Nevertheless, as reported in [22], a direct comparison of the structures of eIF6 and aIF6 will not be possible due to the presence of several deletions and insertions in sequences of both homologues which do not permit a direct structural alignment of the two IF6s without gaps.
Figures 42 and 43 show the effects of temperature and pressure on both structures by means of differences of ScrewFit profiles. Profiles at high temperature or pressure were subtracted from the profile at 300K and 1bar to obtain the differences shown in figures. The outcome of the analysis of these figures is listed in the following. In both cases distinct effects due to temperature or pressure have been observed.

aIF6

- At 350K-1bar aIF6 shows a higher value of orientational distance in region around residue 60 and between residue 120 and 130. In both cases this reflects a significant distortion of the helical conformation without change in the local curvature of backbone as evidenced by a rather small variation in straightness values.

- High pressure (500bar), induces a significant change in residues 90-95. The inspection of all parameters seems to confirm that pressure induces a more curved structure and locally residues assume a conformation more similar to the neighbor α-helix 78-90 rather than to β-strand 90-100.

- The combination of high temperature and high pressure has another significant effect around residue 50 where a transition towards a more *curled* conformation is evident from values of orientational distance and helix radius.

eIF6

- Large variations of all parameters in the region 220-245 are clearly due to the large fluctuations of CTAIL. These variations increase, as expected, with temperature.

- Different variations in the region 170-180 are shown as effects of both temperature and pressure when applied separately but they are not present when a simultaneous application of high pressure and high temperature is performed(see also Figure 44).

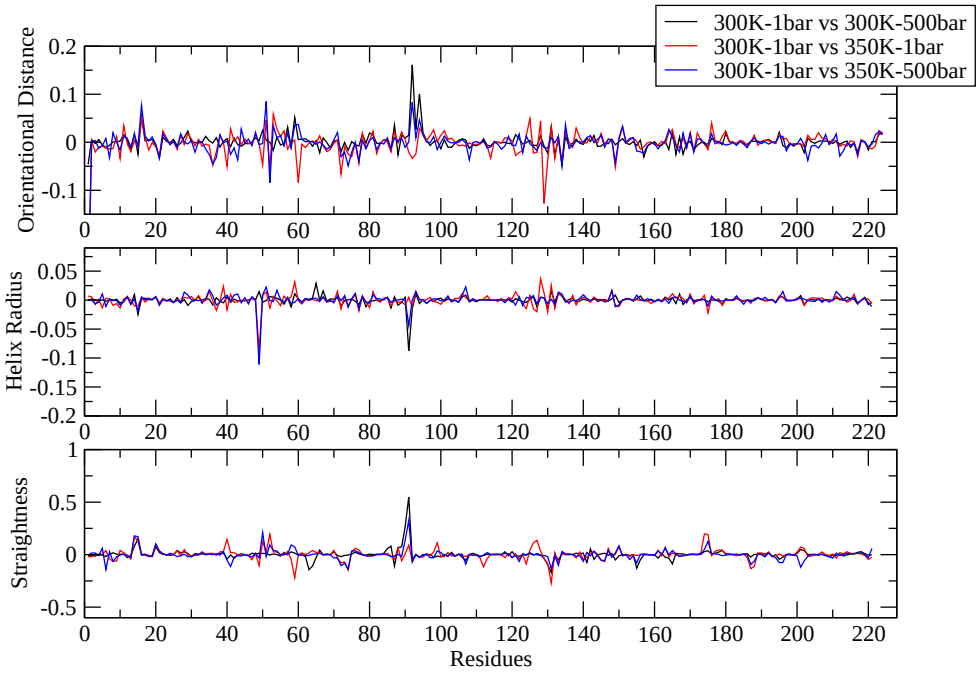

Figure 42: Effects of pressure and temperature on ScrewFit parameters of aIF6 secondary structure.

- Combination of high temperature and high pressure produces an evident change in all parameters in the region 35-45 which correspond to the long α-helix of the first subdomain of IF6 structure. As for region 78-90 in aIF6, also here variations are related to a major curvature of the C-terminus of the α-helix. Several different configurations of peptide planes in region 35-45 are shown by variations in orientational distance and helix radius.

Interestingly, changes in ScrewFit profiles of eIF6 and aIF6 appear in different regions. The former is more affected in residues between 160 and 245 and between 1 and 50, whereas the latter is more significantly affected in the central region 50-140.

7.1.3 *Relation between local structural effects and IF6 function*

Very little is known about how IF6 performs its functions. Only few insights were reported by Basu *et al* [11] and by Groft *et al.* [66]:

- Serines 174 and 175 are found to play an essential role in nuclear localization of IF6s which must be related to the phosphorylation of these two residues.

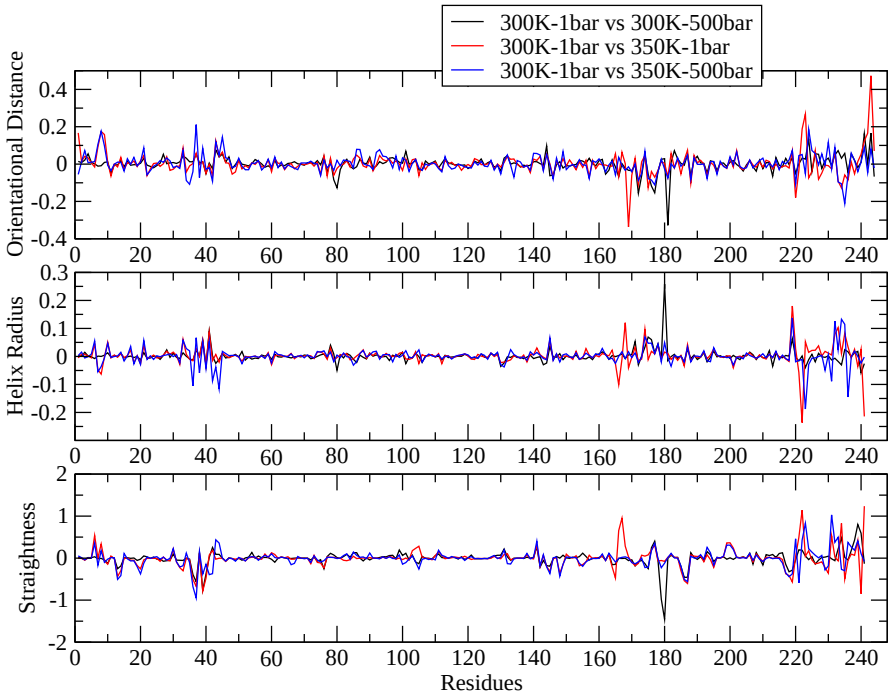

Figure 43: Effects of pressure and temperature on ScrewFit parameters of eIF6 secondary structure.

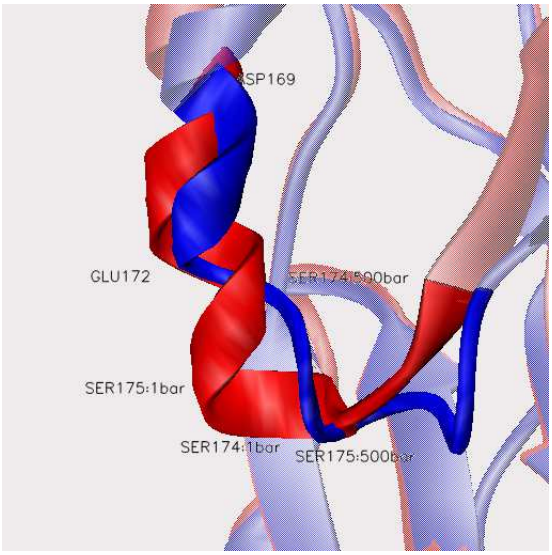

Figure 44: Time-averaged structures of eIF6 at 300K-1bar and 300K-500bar. The structure of the C-terminal region of the α-helix is significantly changed by pressure. This variation affects the position of Serine 174 and Serine 175.Color scheme: red for eIF6 structure at 300K-1bar and blue for structure at 300K-500bar. *Cartoons* representation of secondary structure elements are built following the DDSP assignements.

- Arginine 61 in eIF6 is found to close the central hollow of IF6 structure and prevent water molecules passing throught it. This fact could have effects on the global stability of IF6s structures.

- The C-terminal region (to be distinguished from CTAIL which is formed by residues 224-245) encompassing the evolutionarily conserved arginines and tyrosines Arg 67, Tyr 113, Arg 223 and Tyr 202 represents an exellent candidate for the participation in a biologically relevant protein-protein interaction. Indeed it is suggested as the binding site for the 60S ribosomal subunit.

Some of the structural changes found in this work and induced by pressure and/or temperature involved the residues mentioned above, meaning that the corresponding structural re-arrangements could be related to the functional adaptation of IF6.

In particular it was reported that the region including the two serines 174 and 175 undergoes structural changes at high pressure (Figure 44) and high temperature separately but it assumes its *normal* configuration when both pressure and temperature are increased. This result seems to be crucial for the assessment of the function of eIF6 under extreme conditions similar to those of *warm deep seas*. The C-terminal is obviously also largely influenced by the presence of CTAIL which induces fluctuations in eIF6 that are not present in aIF6. Also this fact will be used in the following to discuss the ability of eIF6 to adapt its function to extreme environments. As a whole, aIF6 seems to show a higher stability with respect to the change of the thermodynamic variables. Nonetheless, it shows some significant changes in backbone conformation in regions apparently not related to protein function.
Finally, the comparison of RMSF and ScrewFit parameters for Arg 61 did not give any significant result in structural changes on the time scales used in this work.

7.1.4 *Comparison between ScrewFit profiles of eIF6 and eIF6-NoCTAIL*

In order to better understand the role of CTAIL in (de)stabilizing eIF6 structure, a supplemental MD simulation was performed on eIF6's structure without its C-terminal tail. The protocol used for this simulation was mentioned in the chapter on system setups.
Comparison of ScrewFit profile of eIF6-NoCTAIL (Figure 45) with that of *complete* eIF6, reveals the absence of variations that characterized eIF6 in the previous paragraphs. Moreover, the variations in ScrewFit parameters in eIF6-NoCTAIL are more similar and closer in sequence localization to those shown for aIF6.

The observations made above lead to the conclusion that CTAIL could play a important role in structural adaptation of IF6 to extreme environment. In particular it seems that the presence of this C-terminal tail induces thermal fluctuations to the backbone of the rest of IF6 structure leading to a lower protein stability. As will be shown later in this chapter, the effects of pressure and temperature induced through CTAIL fluctuations do not cause only structural changes but also a different scheme of IF6 dynamical response to environmental changes.

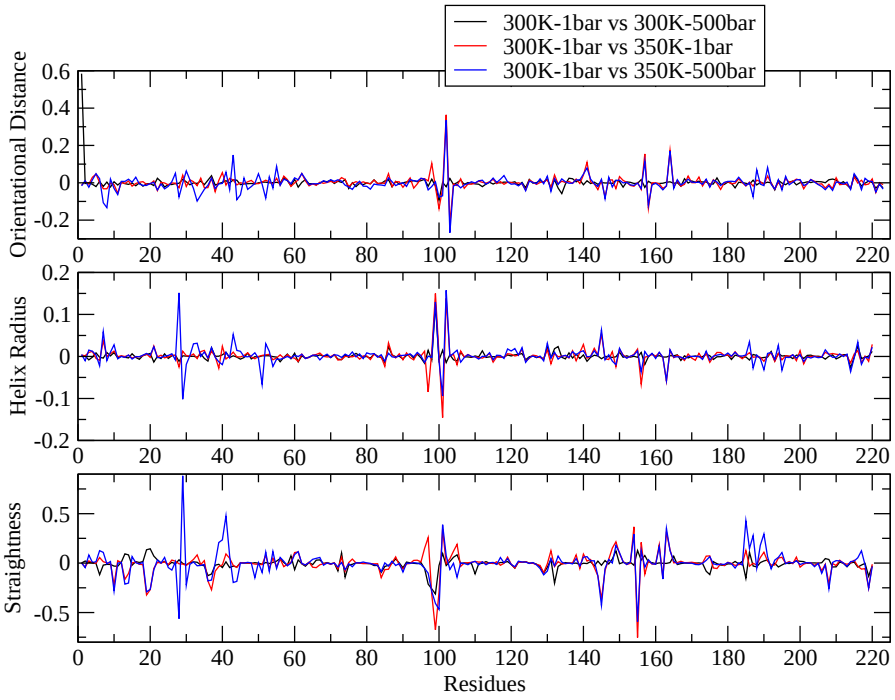

Figure 45: Effects of pressure and temperature on ScrewFit parameters of eIF6-NoCTAIL secondary structure.

7.1.5 *Comparison between ScrewFit profiles of aIF6 and aIF6-HTag*

As recalled at the start of this chapter, high pressure QENS experiments were performed on aIF6 with an attached supplemental N-terminal tail of 21 amino acids (HTag). This tail was essential to improve the yield of protein production. The same sample (aIF6-Htag), was also studied by MD simulations to make comparison with experimental data possible. aIF6-HTag

time-averaged structure was initially analyzed by ScrewFit in order to identify differences with other samples and to verify the structural effect of the HTag on the protein itself.
ScrewFit differences between structures of aIF6-HTag at different pressures and temperatures are shown in Figure 46. Here the beginning of the residue sequence was shifted to -20 in order to maintain the usual numbers in the conserved part of the structure.
First of all, the effect of the HTag is clearly present in all parameters as large variations in the region from -21 to 0 but in contrast to the effects produced by CTAIL on eIF6 structure, here these changes are localized only in the region of the HTag and do not propagate over the rest of the structure.
Moreover, the region between residues 40 and 50 seems to undergo the same changes found in aIF6, meaning that this region is intrinsically sensitive to environmental changes probably due to its exposure to solvent.

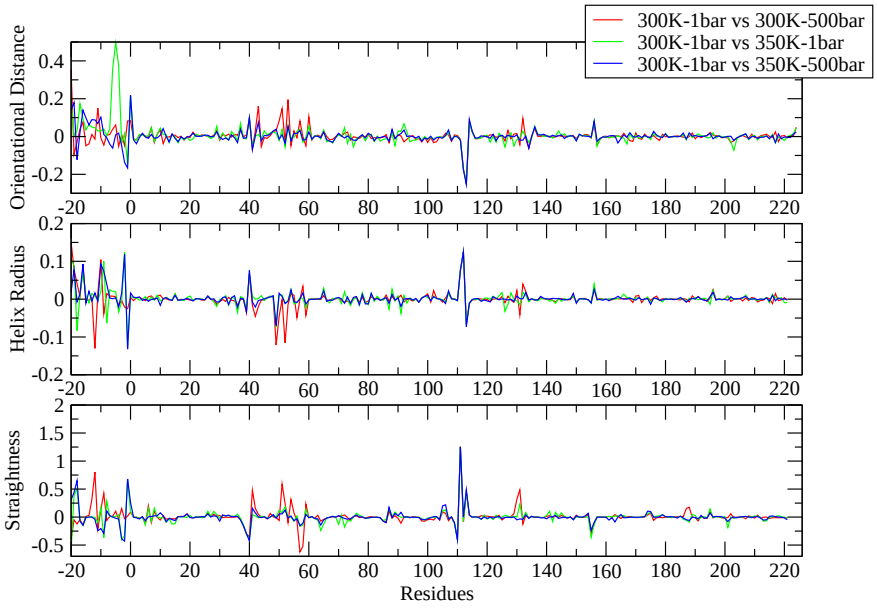

Figure 46: Effects of pressure and temperature on ScrewFit parameters of aIF6-HTag secondary structure.

Finally, it is worth noting that the major change in the whole structure happens in region 110-115 and is present almost in the same way in all conditions. This fact suggests that the

latter is a structural arrangement compatible only with the structure at 300K-1bar.

The outcome of this comparison of ScrewFit profiles as well as the others was that both aIF6 and eIF6 are largely affected by the possible presence of a additional sequence of amino acids either in their N- or C- termini. Nevertheless, the effects produced strictly depend on the rest of the protein structure and can be very different.

7.1.6 *Elastic Incoherent Structure Factor*

As stated in the chapter on Material and Methods, the Elastic Incoherent Strcture Factor gives insights into the configurational space volume explored by atomic motions in macromolecules. As for other scattering functions, in the case of biomolecules, also the EISF is largely dominated by hydrogen contributions and it can be approximated as due to the motions of only one single representative scattering atom.
The EISF has already been used widely to characterize the structural "flexibility" of proteins because it can be directly related to atomic mean square displacement (MSD). Indeed, for small values of the momentum transfer, it can be proved that the equation 2.42 can be approximated by the following expression [15]:

$$EISF(q) = \lim_{t\to\infty} I(q,t) = \exp(-q^2\langle x^2\rangle), \tag{7.2}$$

which has Gaussian form. In reality the Gaussian approximation holds strictly only for $q \to 0$ [15]. The above equation reveals that $\langle x^2\rangle$ can be directly obtained by EISF via:

$$\langle x^2\rangle = -\ln(EISF[q])/q^2 \tag{7.3}$$

Using equation 2.42, the EISF of aIF6 and eIF6 have been calculated. Here and in the following, the quantities of interest are calculated only on the evolutionary "conserved" part of the IF6, i.e. on the region 1-224, filtering out the contribution of the C-terminal tail. A first comparison of EISF from aIF6 (Figure 47 - Panel A) and eIF6 (Figure 47 - Panel B) underlines two main differences between the two samples:

- In both cases pressure makes the slope of EISF only slightly change whereas temperature has a much more evident effect

- Even though aIF6 shows qualitatively the same behavior as eIF6, its variations due to pressure are very limited. Moreover, at high temperature (350K), EISFs at 1bar and 500bar are almost identical.

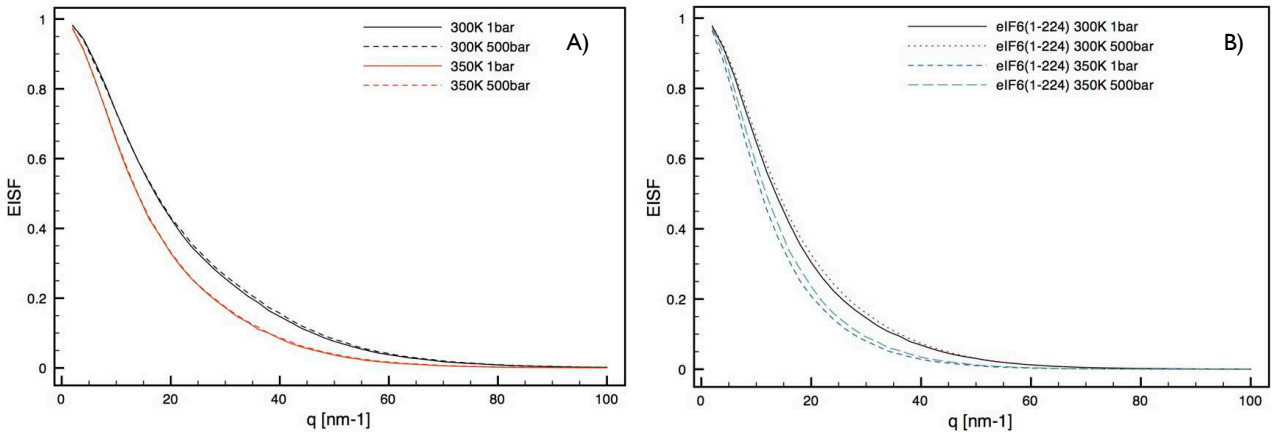

Figure 47: Elastic Incoherent Structure Factor for aIF6 and eIF6.

Looking at these results, one would conclude that the difference between the two samples is mainly related to their stiffness/softness character [216], but the direct comparison of EISFs in Figure 48 shows also another interesting outcome: EISFs of aIF6 at high temperature are very similar to those of eIF6 at 300K. This means that configurational spaces explored have the same extent in both samples when they are both in their *natural* conditions.

Furthermore, EISF from eIF6 at high temperature shows a much more rapid decrease as a function of the momentum transfer with respect to the other EISFs. Another interpretation for this result is given by the atomic position fluctuation, $\langle x^2 \rangle$, obtained from equation 7.3 and shown in Figure 48 (Panel B). From this figure it seems evident that atomic fluctuations can be analyzed as function of the momentum transfer. This does, however, not necessarily mean that the Gaussian approximation given above does not hold, but only that it is not compatible with the assumption of one single representative scattering atom. It has, in fact, been shown that the Gaussian model can represent the EISF of a protein up to moderate q values, since such a model can account for motional heterogeneity.
The $\langle x^2 \rangle$ attains larger values for eIF6 at high temperature (at both pressures), meaning that fluctuations increase and involve the motion of rather large group of atoms or structural domains ($q < 20 nm^{-1}$). It is also worth noting here that the EISF (and the corresponding $\langle x^2 \rangle$) from eIF6 at 300K-500bar superpose slightly better than the one at eIF6-300K-1bar to aIF6 1bar/500bar.

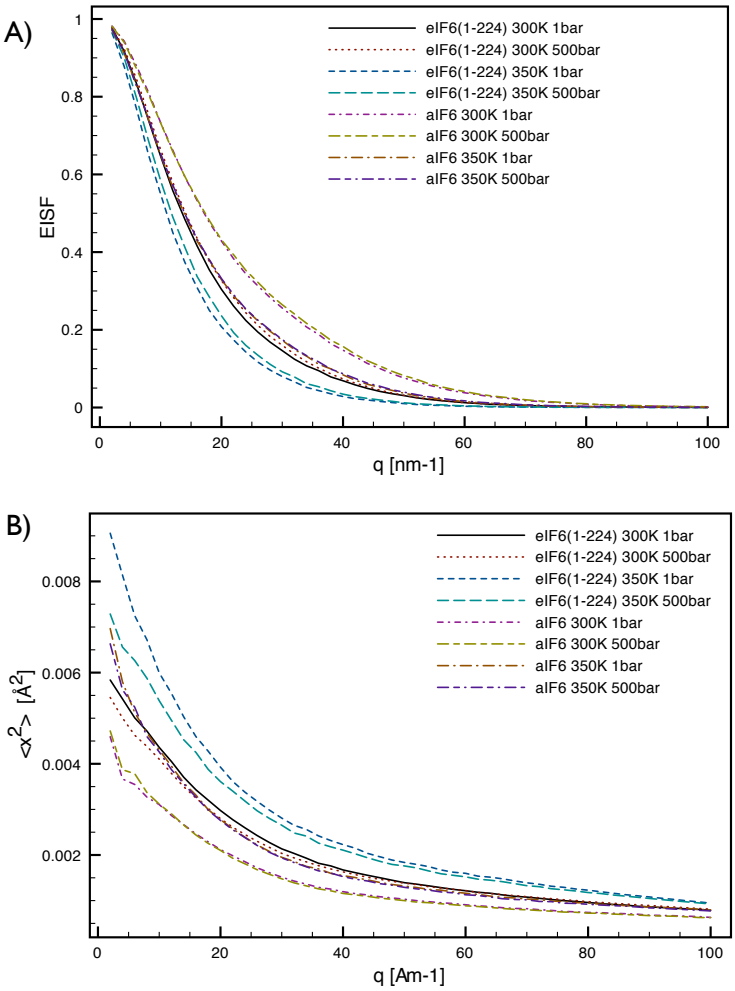

Figure 48: **Panel A):** Comparison of EISFs from aIF6 and eIF6.**Panel B):** Atomic mean fluctuations as obtained from equation 7.3

CHARACTERIZATION OF EIF6-NOCTAIL AND AIF6-HTAG STRUCTURE AND DYNAMICS

In this chapter, an hypothesis on the characterization of structural and dynamical properties of eIF6 and aIF6 is proposed. In particular the comparison of different EISFs will be used to assess the existence of corresponding states between aIF6 and eIF6.
Here, similar characterization will be sketched also for eIF6-NoCTAIL and aIF6-HTag in order to use them latter in a more detailed discussion of the results.

EISFs from aIF6-HTag are compared to those of eIF6 and aIF6 at different temperature and pressure and plotted in Figure 49. Interestingly, this result, together with the analysis made by ScrewFit in the previous sections, shows that aIF6-HTag has the same structural properties as eIF6. An opposite result was obtained from comparison of EISFs from eIF6-NoCTAIL with both eIF6 and aIF6 (Figure 50). Here, it seems that, even though EISFs from eIF6-NoCTAIL are very similar to those from aIF6, they show that the former has a less compact structure as shown in 51.

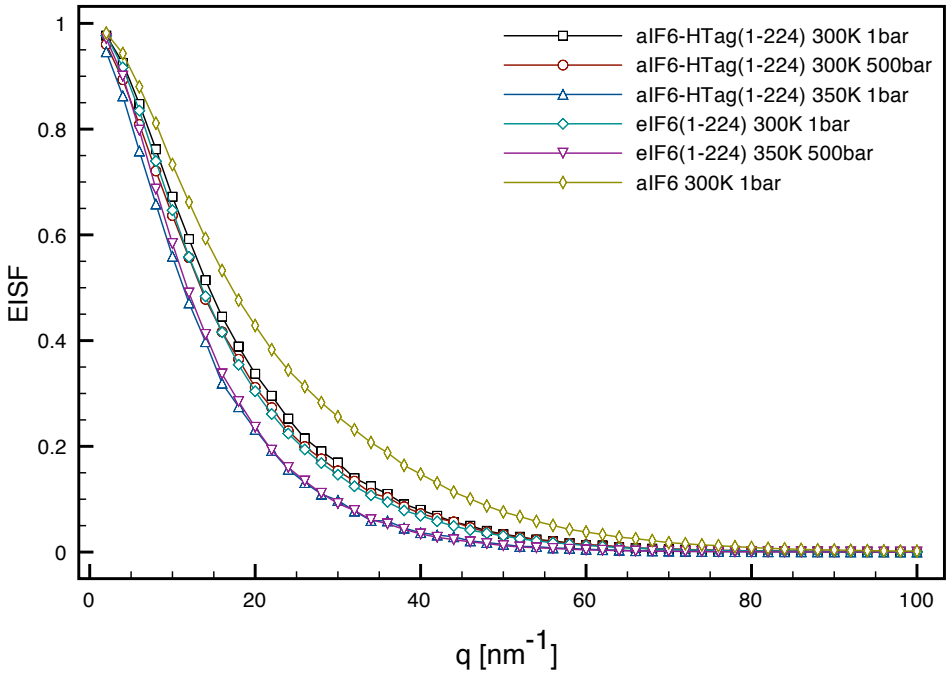

Figure 49: Elastic Incoherent Structure Factor for aIF6-HTag compared to EISFs from eIF6 and aIF6.

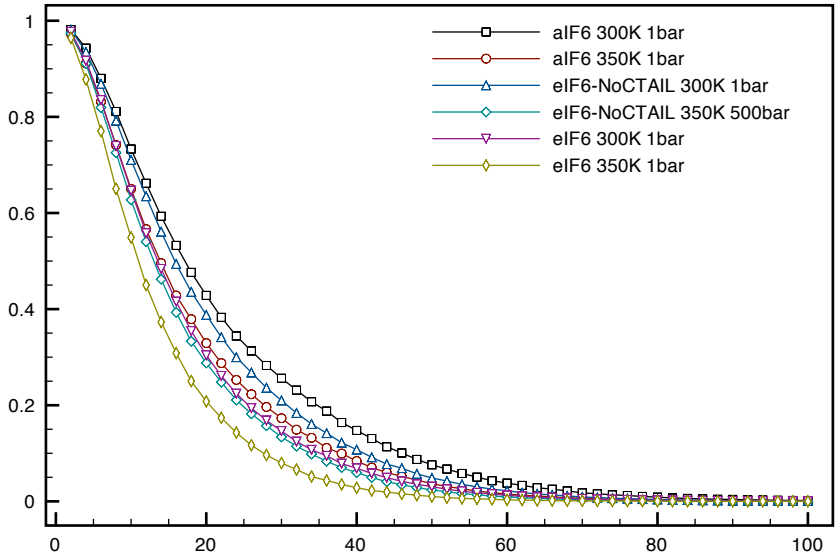

Figure 50: Elastic Incoherent Structure Factor for eIF6-NoCTAIL compared to EISFs from eIF6 and aIF6.

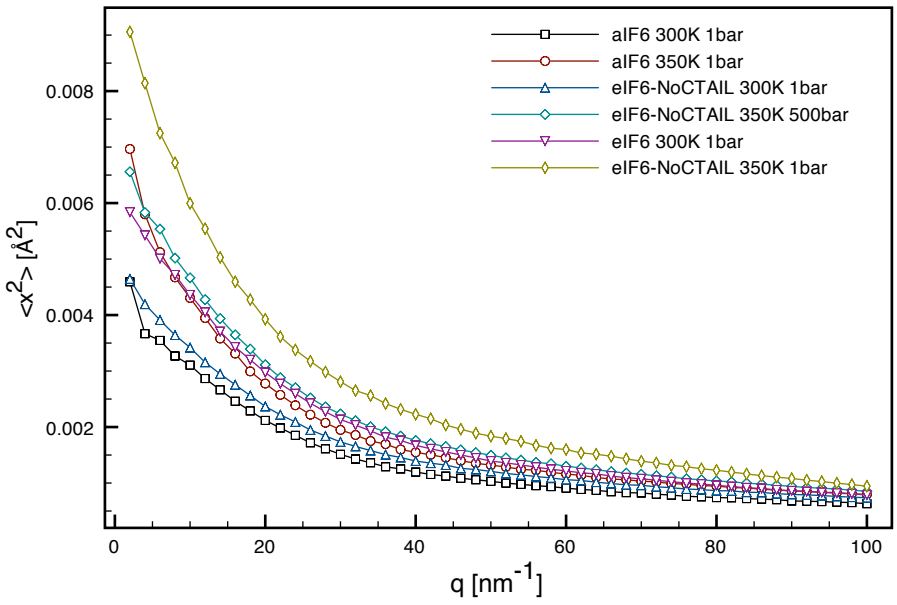

Figure 51: Atomic mean square displacement (MSD) of eIF6-NoCTAIL as function of momentum transfer. Comparison with MSD from eIF6 and aIF6 is shown.

7.2 EFFFECTS OF PRESSURE AND TEMPERATURE ON IF6S DYNAMICS

7.2.1 *Dynamical models*

The data issued from neutron scattering measurements need to be interpreted in a frame of physical models describing the different types of motions which compose the global status of the sample. The easiest assumption which is usually made is to consider every type of motion de-correlated with respect to the others and every atom dynamically equivalent to the others. For this, the position of the atom i reads:

$$\mathbf{R}_i = \mathbf{R}_{CM} + \mathbf{r}_i + \mathbf{u}_i \tag{7.4}$$

where $\mathbf{R}_{CM}$ is the position of the center of mass (CM) of the sample, $\mathbf{r}_i$ is the atom position with respect to CM and $\mathbf{u}_i$ is the deviation from the equlibrium position $\mathbf{R}_{CM} + \mathbf{r}_i$. These assumptions are reliable for small molecules with some degrees of symmetry but becomes an approximation when applied to much more complex molecules like proteins. Assuming that the latter is acceptable, i.e. that atom i represents the average dynamical properties of all atoms in the sample, the intermediate scattering function can be recast as follows:

$$F_H(\mathbf{q}, t) = F_{CM}(\mathbf{q}, t) \cdot F_{rot}(\mathbf{q}, t) \cdot F_{int}(\mathbf{q}, t) \tag{7.5}$$

here the "rot" and "int" indexes refer to global rotational and internal motions, respectively. Fourier transformation of 7.5 gives the relation for the dynamic structure factor which reads now:

$$S_H(\mathbf{q}, t) = S_{CM}(\mathbf{q}, t) \otimes S_{rot}(\mathbf{q}, t) \otimes S_{int}(\mathbf{q}, t) \tag{7.6}$$

The term due to translation of the center of mass can be described following the Frick's law of diffusion of a free particle [13]. For the latter, particles positions are governed by a Gaussian probability and thus, the intermediate scattering function $F(q, t)$ will have a Gaussian form as function of the momentum transfer [164]. The corresponding dynamic structure factor $S(q, \omega)$ will have a Lorentzian form:

$$S(q, \omega) = \frac{1}{\pi} \frac{Dq^2}{(Dq^2)^2 + \omega^2} \tag{7.7}$$

where D is the diffusion constant and the half-width-half-maximum (HWHM) is defined by Dq^2. The contribution to equation 7.6 due to global rotational diffusion is more complicated to obtain because it requires a description of molecules as rigid-bodies [37]. As stated in Chapter 2, this term is not taken into account in our QENS measurements because it is largely beyond the experimental resolution used here. Finally, several models have been developed to analyze the internal motions of atoms in large molecules like proteins. This means that, in the simplest case, all types of atomic diffusion motions which can be found into large molecules, are *averaged* to find a unique *dynamical property* for all atoms that characterize the molecule. This usually results in time-correlation functions characterized by a unique relaxation time associated with all atoms in molecules.
However, a discussion all these models would be out of the scope of this thesis and accounts can be found in many excellent reviews. Here a brief summary will be given of the model used in this work to characterize atomic motions in proteins.

7.2.2 *Fractional Brownian Dynamics*

To interpret both the simulated and experimental data, the fractional Ornstein-Uhlenbeck (OU) process [137] is used as an analytical model for the atomic motions in a protein. The model describes anomalous diffusion in a harmonic potential, where the latter accounts for the fact that atomic motions in a protein are confined in space. The anomalous diffusion describes slow, non-exponential structural relaxation in the functional dynamics of proteins, which has been observed in the past on the microsecond to second time scale by fluorescence correlation spectroscopy[214] and by kinetic studies[64]. The existence of fractional Brownian dynamics in proteins on the nanosecond time scale has been recently demonstrated by analyses of molecular dynamics simulations [119] and the fractional OU process has been introduced in [113] for the interpretation of QENS spectra from proteins. It can be considered as an extension of a simple harmonic protein model, which has been used in the past to describe elastic neutron scattering profiles, in particular to extract the "resilience" of proteins in terms of an average force constant [216]. The fractional OU process (fOU) adds to this a description of the relaxation dynamics, which is measured in QENS experiments.

Full details of fOU model and its application to protein dynamics can be found in previous works [119, 113, 19, 20] and here I will recall only the key points needed for the following

analysis of MD and QENS data.

TIME-DEPENDENT MEAN-SQUARE DISPLACEMENT The most elementary quantity to be considered in the context of diffusion processes is the time-dependent mean-square displacement (tMSD),

$$W(t) := \langle [x(t) - x(0)]^2 \rangle, \tag{7.8}$$

where x is the position of the diffusing particle and the brackets indicate a thermal average. In case that the dynamics of the particle is confined in space, the tMSD will tend to a plateau value, which is given by $2\langle x^2 \rangle$. For the fractional OU process one has

$$W(t) = 2\langle x^2 \rangle \Big(1 - E_\alpha(-[t/\tau]^\alpha) \Big). \tag{7.9}$$

Here $E_\alpha(z)$ is the Mittag-Leffler function [52]

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(1+\alpha k)}, \tag{7.10}$$

where $\Gamma(.)$ denotes the generalized factorial [2]. One recognizes that for $\alpha = 1$, where $\Gamma(1 + \alpha k) = \Gamma(1 + k) = k!$, the exponential function is retrieved from expression (7.10), i.e. $E_1(z) = \exp(z)$. In this case the fractional OU process becomes the well-known standard Markovian OU process, which is characterized by exponential relaxation functions [205, 59, 164]. As indicated in [117], the fractional counterpart is characterized by non-Markovian memory effects, which lead to non-exponential correlation functions.

Expressions (7.10) and (7.9) show that the proposed model contains three parameters:

1. the position fluctuation $\langle x^2 \rangle$,
2. the parameter α indicating the deviation from exponential behavior,
3. the time scale parameter τ.

These parameters will be considered as *q-dependent* and will give account of the multiple relaxation dynamics the characterize the internal motions on proteins.

RELAXATION RATE SPECTRUM The function $E_\alpha(-[t/\tau]^\alpha)$ can be considered as a "stretched" *generalized* exponential function. The non-exponential character of this function can be most easily visualized by writing it as a superposition of normal exponential functions. Using for simplicity a dimensionless time variable we have

$$E_\alpha(-t^\alpha) = \int_0^\infty d\lambda \, p_\alpha(\lambda) \exp(-\lambda t), \tag{7.11}$$

where $p_\alpha(\lambda)$ is a normalized and positive distribution function, which is of the the form [64, 117]

$$p_\alpha(\lambda) = \frac{1}{\pi}\frac{\lambda^{\alpha-1}\sin(\pi\alpha)}{\lambda^{2\alpha}+2\lambda^\alpha\cos(\pi\alpha)+1}, \quad 0<\alpha<1. \tag{7.12}$$

In the limit $\alpha \to 1$ we have [117]

$$\lim_{\alpha\to 1} p_\alpha(\lambda) = \delta(\lambda-1), \tag{7.13}$$

in agreement with $\lim_{\alpha\to 1} E_\alpha(-t^\alpha) = \exp(-t)$.

MODELING INCOHERENT NEUTRON SCATTERING In the following the dynamic structure factor for incoherent neutron scattering will be considered,

$$S(q,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dt\, \exp(-i\omega t) I(q,t), \tag{7.14}$$

where $I(q,t)$ is the incoherent intermediate scattering function, which depends on the position of the scattering atom

$$I(q,t) = \langle \exp(iq[x-x_0])\rangle. \tag{7.15}$$

Here $q = |\mathbf{q}|$ is the modulus of the momentum transfer which the neutron transfers to the scattering atom in the scattering process. Within the model it is assumed that the system under consideration is isotropic and that, as made for EISFs, the protein dynamics, as seen in incoherent neutron scattering, can be described by one "representative" atom. In this case it suffices to consider one coordinate of the scattering atom, which is chosen to be the x-coordinate. In view of the predominance of incoherent scattering by hydrogen atoms, the representative atom in the model is a representative hydrogen atom.

Within the model the intermediate scattering function has the form

$$I(q,t) = \exp(-q^2\langle x^2\rangle)\sum_{n=0}^{\infty}\frac{q^{2n}\langle x^2\rangle^n}{n!}E_\alpha\left(-[t/\tau_n]^\alpha\right), \qquad 0<\alpha\leqslant 1, \tag{7.16}$$

where τ_n is given by

$$\tau_n = \tau n^{-1/\alpha}. \tag{7.17}$$

The dynamic structure factor associated with the intermediate scattering function (7.16) reads

$$S(q,\omega) = \exp(-q^2\langle x^2\rangle)\left\{\delta(\omega)+\sum_{n=1}^{\infty}\frac{q^{2n}\langle x^2\rangle^n}{n!\,2\pi}L_{\alpha,\tau_n}(\omega)\right\}, \tag{7.18}$$

where $L_{\alpha,\tau}(.)$ is the generalized Lorentzian

$$L_{\alpha,\tau}(\omega) = \frac{2\tau\sin(\alpha\pi/2)}{\omega\tau\Big((\omega\tau)^\alpha+2\cos(\alpha\pi/2)+(\omega\tau)^{-\alpha}\Big)}, \quad 0<\alpha\leqslant 1. \tag{7.19}$$

FITTING QENS SPECTRA The model introduced in previous paragraphs describes internal protein dynamics and to be useful for the interpretation of QENS spectra of protein solutions the effects of global diffusion and of finite instrumental resolution must be incorporated. Neglecting multiple scattering effects and absorption, and assuming that global diffusion of the IF6 molecules and internal motions are decoupled, one can write the measured dynamic structure factor as convolution product (defining $(f * g)(\omega) = \int_{-\infty}^{+\infty} d\omega' f(\omega - \omega') g(\omega')$):

$$S_m(q,t) = (S * l * r)(\omega), \tag{7.20}$$

Here S stands for the dynamic structure factor of the model, l is a Lorentzian describing translational diffusion (D is the diffusion constant),

$$l(\omega) = \frac{1}{\pi} \frac{Dq^2}{(Dq^2)^2 + \omega^2} \tag{7.21}$$

and r is the resolution function, which is well described by a Gaussian,

$$r(\omega) = \frac{\exp\left(-\frac{\omega^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}, \tag{7.22}$$

with $\sigma > 0$ and a half-width at half maximum (HWHM) of $\Delta E \approx 1.17\sigma$. Both $r(.)$ and $l(.)$ are normalized such that $\int_{-\infty}^{+\infty} d\omega\, r(\omega) = 1$ and $\int_{-\infty}^{+\infty} d\omega\, l(\omega) = 1$.

The convolution product (7.20) for the measured dynamics structure factor can be written in the following form, using for S the model (7.18),

$$S_m(q,\omega) = \exp(-q^2\langle x^2\rangle) \left\{ (l * r) + \sum_{n=1}^{\infty} \frac{q^{2n}\langle x^2\rangle^n}{n!\, 2\pi} (L^D_{\alpha,\tau_n} * r)(\omega) \right\}. \tag{7.23}$$

Here $L^D_{\alpha,\tau_n}(\omega) = (L_{\alpha,\tau_n} * l)(\omega)$ is the convolution of a generalized Lorentzian with a normal Lorentzian, for which an analytical form can be given. Defining

$$\tilde{\omega} = \sqrt{\omega^2 + (Dq^2)^2}, \quad \phi = \arg(Dq^2 + i\omega), \tag{7.24}$$

one obtains [112]

$$L^D_{\alpha,\tau}(\omega) = \frac{2\{(\tilde{\omega}\tau)^\alpha \cos\phi + \cos([\alpha - 1]\phi)\}}{\tilde{\omega}\{(\tilde{\omega}\tau)^\alpha + 2\cos\alpha\phi + (\tilde{\omega}\tau)^{-\alpha}\}}. \tag{7.25}$$

7.2.3 *Analysis of scattering functions obtained from MD simulations*

The "natural" quantities for the analysis of MD simulations are time-dependent MSDs and time correlation functions, such as the intermediate scattering functions (ISF), which can be directly computed from the trajectories. In this thesis, the MD analysis package nMoldyn

was used for this purpose [166]. All global motions (translation and rotation) were filtered out beforehand from all trajectories in order to avoid the presence of unwanted spurious contributions in the calculation of correlation functions. The procedure used to accomplish this task is well established and has already been explained in [115]. ISFs from aIF6 and eIF6 for four different values of momentum transfer are shown in Figures 53 and 54 for different environmental conditions.

The related time-dependent tMSD is shown in Figure 52. Data were fitted with the fOU model as explained in the previous section. Values for $\langle x^2 \rangle$ in 7.9 were fixed from direct calculation on MD simulations. The results of fits together with calculation of atomic $\langle x^2 \rangle$ are reported in Table 19.

Table 19: Parameters of the fOU model from fits on eIF6 and aIF6.

	aIF6	$\langle x^2 \rangle$ [nm^2]	α [nm^2]	τ [ps]	eIF6	$\langle x^2 \rangle$ [nm^2]	α [nm^2]	τ [ps]
300K 1bar		$2.84 \cdot 10^{-3}$	0.42	115		$4.13 \cdot 10^{-3}$	0.46	225
300K 500bar		$2.96 \cdot 10^{-3}$	0.36	198		$3.88 \cdot 10^{-3}$	0.47	168
350K 1bar		$4.32 \cdot 10^{-3}$	0.40	114		$6.56 \cdot 10^{-3}$	0.56	143
350K 500bar		$4.46 \cdot 10^{-3}$	0.42	142		$5.36 \cdot 10^{-3}$	0.49	168

Inspection by eye of tMSD plots reveals that eIF6 at 300K as well as aIF6 at 350K are quite insensitive to pressure whereas they undergo a much more evident change at the *non-natural* temperatures. Moreover, at high temperature eIF6 shows large variations due to pressure. The fitted ISF from both samples are shown in Figures 53 and 54. Here, ISFs of both eIF6 and aIF6 do not relatively change with respect to pressure when temperature is kept respectively to 300K and 350K whereas significative variations occurred at different temperatures.

Again also the observations made on tMSD and ISF seem to confirm the presence of *corresponding states* between the dynamical properties of the two IF6 homologues in their natural conditions and defined by apparent similar responses to environmental changes. Nevertheless, it is worth noting that here the presence of this correspondence is not found through similar values of these quantities but by means of homologous variations in function of thermodynamical variables. A further and more complete analysis of dynamical properties

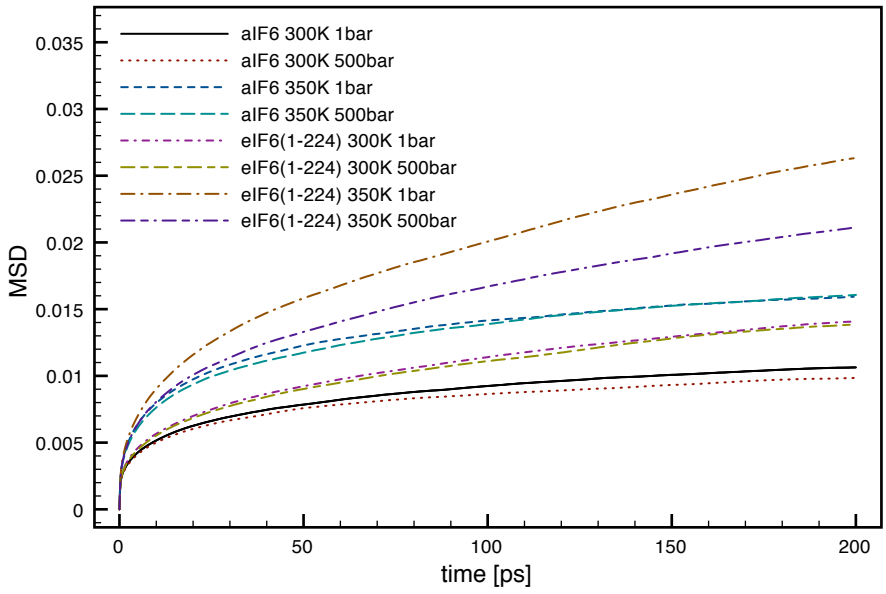

Figure 52: Comparison of time-dependent mean square displacement of aIF6 and eIF6.

can be obtained by means of fractional Brownian dynamics model. As explained in the previous section, parameters α and τ given in the fOU model, define the heterogeneity of relaxation dynamics and are valuable tools to characterize the internal motions in proteins.

Values of these two parameters obtained from fit with the fOU model of ISF are plotted as a function of momentum transfer in Figures 55 and 56. Surprisingly, the fitted parameters show several differences between aIF6 and eIF6:

- τ in eIF6 is generally larger than in aIF6 by a factor of 3, meaning that its internal dynamics is characterized by slower motions. Moreover, its variation as function of pressure is of opposite sign with respect to the one in the extremophile homologue.

- α in aIF6, although similar to the one obtained in eIF6, is always systematically larger than the latter. Following the definition of α and expression 7.12 this could mean that the two holomogues are characterized by different distributions of relaxation rates.

The partial discrepancies between the values of $\{\alpha, \tau\}$ from tMSDs and those from ISFs could be explained from the fact that tMSDs suffer much more than ISFs from the limited time-length of MD trajectories which do not permit a good sampling to be achieved for time-correlation functions.

A key point here is the understanding of which of differences found in $\{\alpha, \tau\}$ are related intrinsically to the differences between the two proteins and which can be taken as a possible

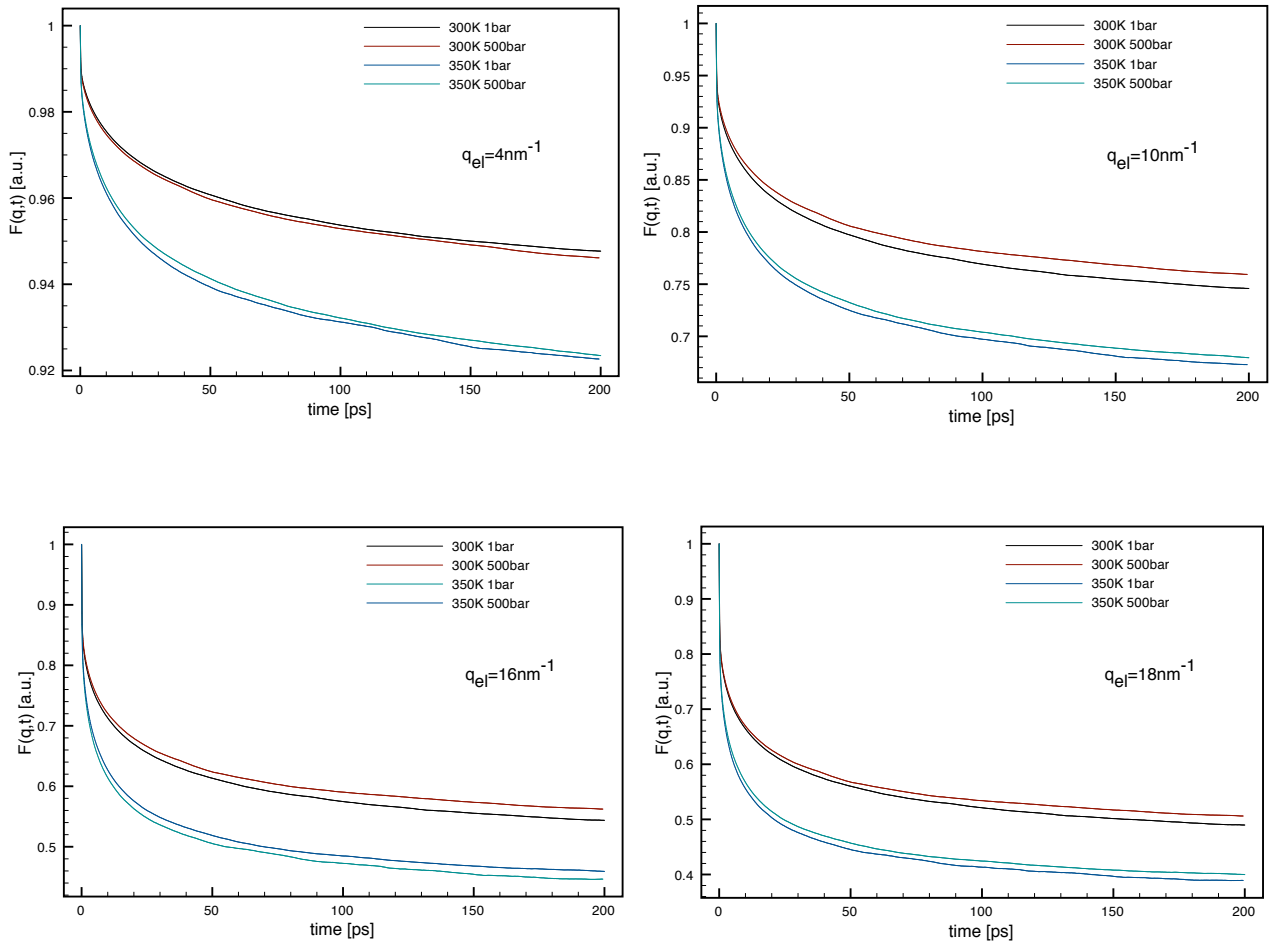

Figure 53: Intermediate Scattering Function from MD simulation on aIF6 at different q-values: $4, 10, 16, 18 nm^{-1}$.

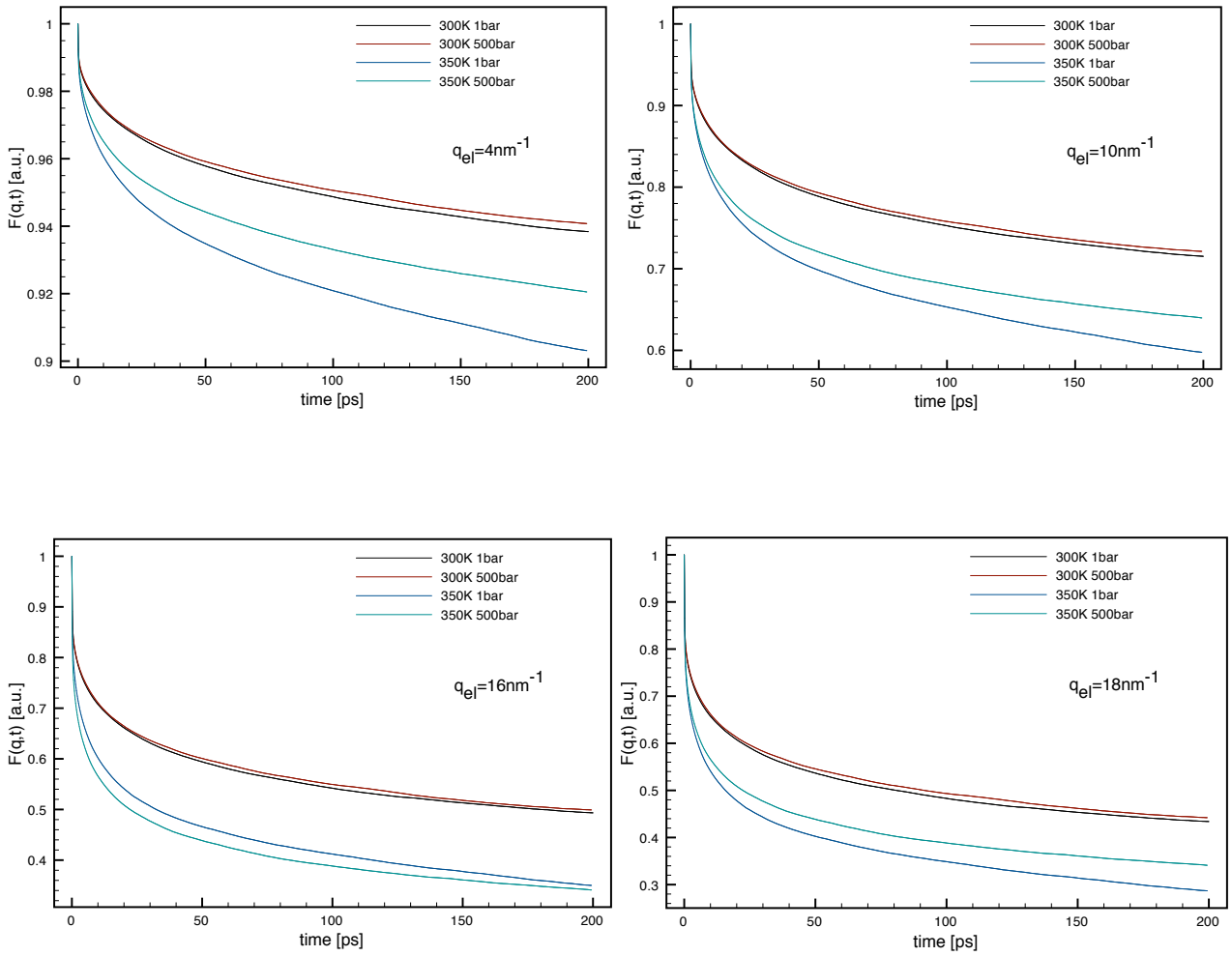

Figure 54: Intermediate Scattering Function from MD simulation on aIF6 at different q-values: $4, 10, 16, 18 nm^{-1}$.

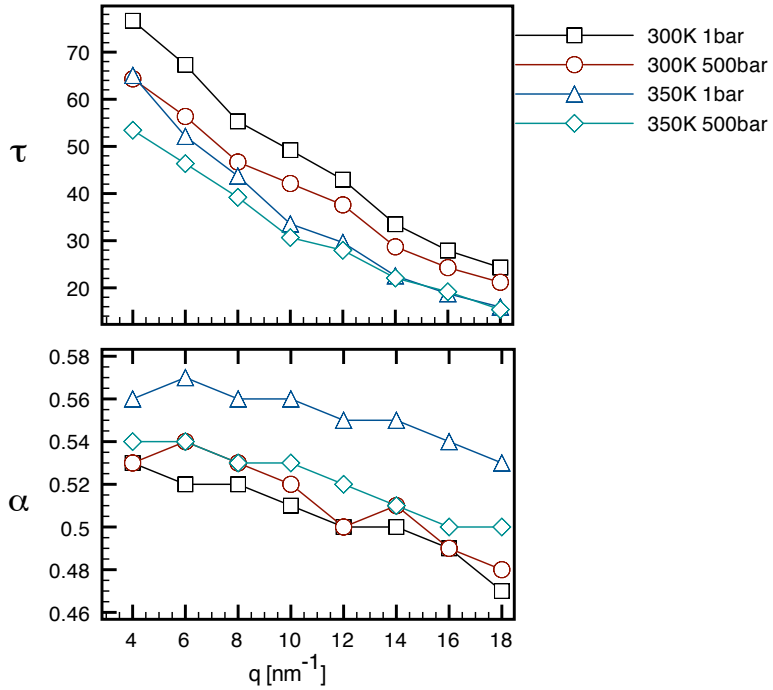

Figure 55: Parameters for the fOU model obtained from fits on intermediate structure factors of eIF6.

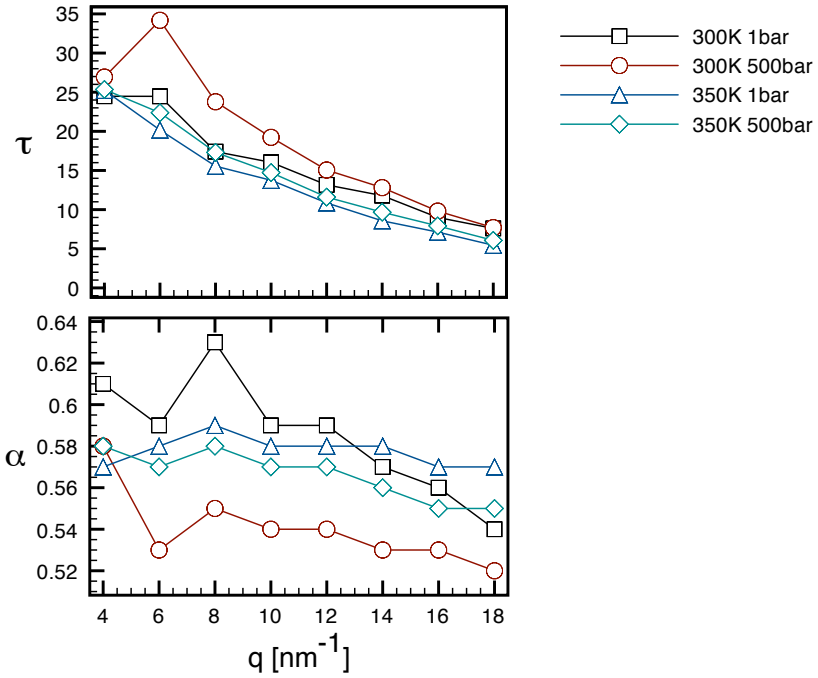

Figure 56: Parameters for the fOU model obtained from fits on intermediate structure factors of aIF6.

"signature" of adaptation. As will explained later in this chapter the sign of variations with respect of temperature and pressure should be related to peculiar properties of each homologue and more generally of each protein, whereas other qualities of these variations, such as their length scale dependancy, could be related to evolutionary molecular adaptation.

The analysis of EISF was very useful for finding the corresponding states but it does not show how dynamical properties of proteins are modified moving from unfavorable conditions to physiological ones. The absolute value of parameter τ which characterizes relaxation time as function of momentum-transfer, thus as function of length-scale, was shown to be a useful tool to identify differences between eIF6 and aIF6. Interestingly, comparing the value of τ of each protein in different environmental conditions to those in *natural* conditions, one finds that in both samples, τ variations are q-dependent and that their extension is related to the *quality of extremeness* of the particular environmental condition. This comparison can be made normalizing values of τ to values at *natural* conditions as shown in Figures 57 and 58. From these figures one can easily note that variations in τ become non-uniform as external conditions move too far from the *natural* (optimal) ones and they mainly involve the region corresponding to small momentum transfer.

7.3 COMPARISON WITH QENS MEASUREMENTS

In the QENS experiments performed for this thesis work a number of technical difficulties were encountered which severely limited the amount and the quality of data. Hence, only a purely qualitative comparison with MD was possible. The main problems related to experimental measurements were due to the lack of previous knowledge of the chemical properties of IF6 samples:

eIF6 turned out to have a very low stability in the experimental setups developed for this work. Indeed, in high pressure experiments after some hours of measurements the protein started to aggregate and very few spectra were reliable for analysis. Probably this was due to variations in protein concentration caused by pressure application and the coexistence of CTAIL and the poly-histidine tag which could interfere with protein stability.

The evaluation of protein concentration in solution was made very difficult by the very low quantity of chromophore aminoacids which resulted in low UV absoprtion. As

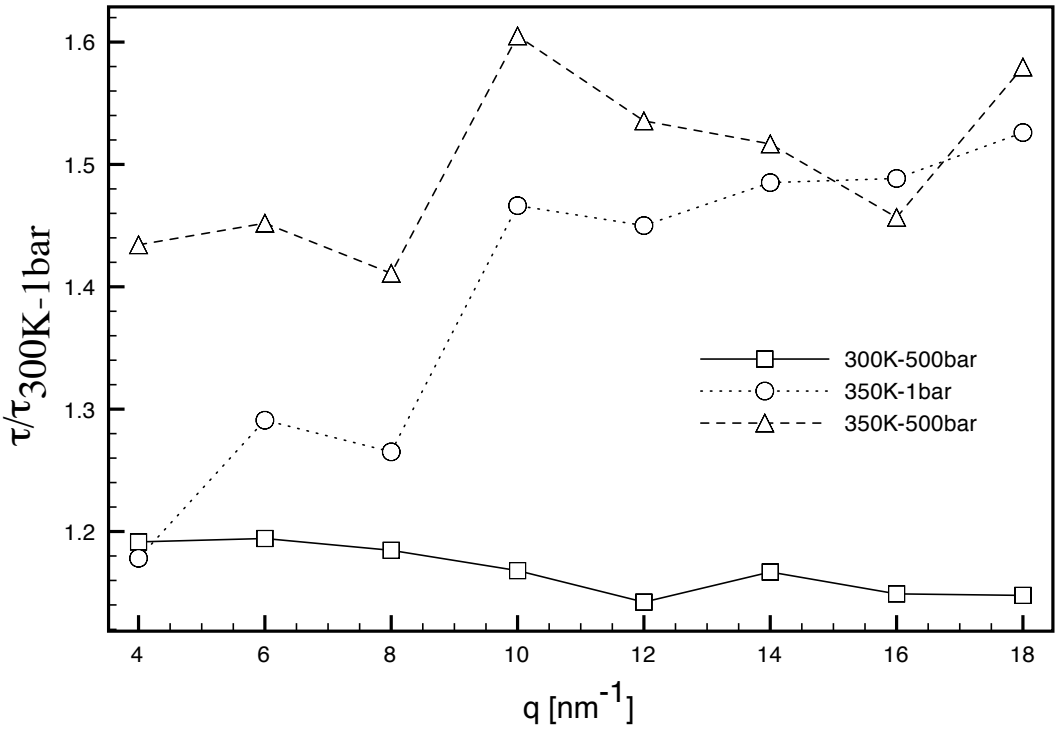

Figure 57: Values of τ from fit of **eIF6** normalized to value at 300K and 1bar.

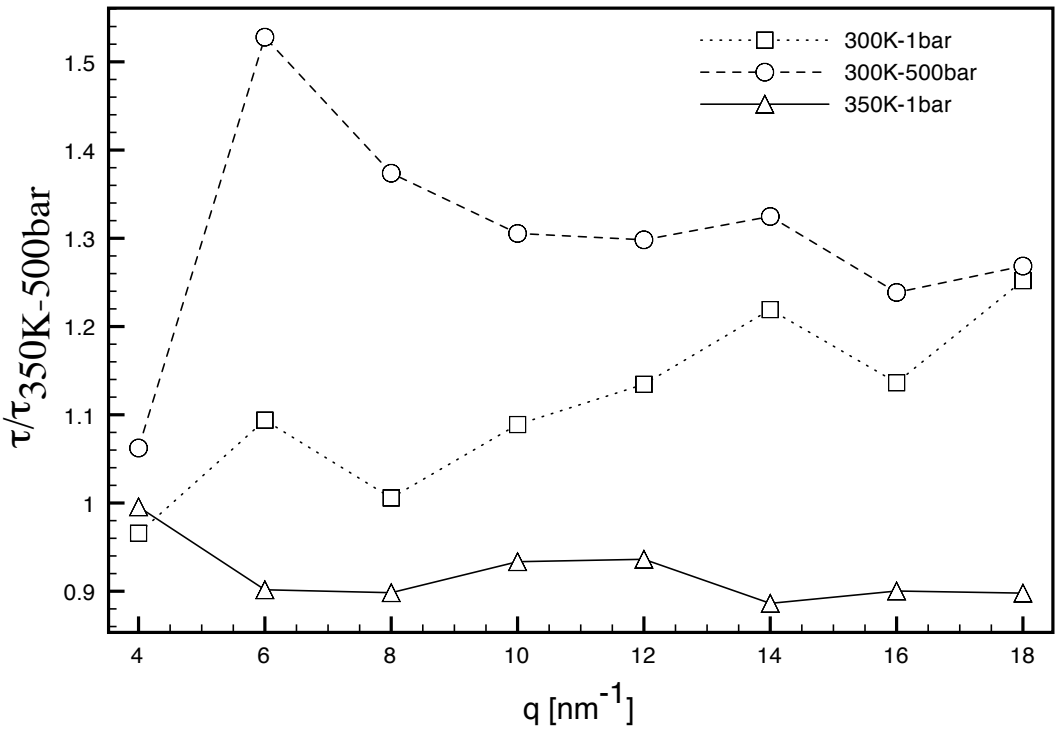

Figure 58: Values of τ from fit of **aIF6** normalized to value at 350K and 500bar.

a consequence, the absorption at 280nm, usually measured to estimate the protein concentration gave systematic errors greater than 10%.

In addition to these limitations, one should recall also the other results found on eIF6 stability and reported in Chapter 3. For this reason, QENS spectra of eIF6 will not be shown here. Some preliminary analysis of spectra from aIF6 was, however, still possible and will be used here to support some of the observations made by means of MD simulations.

7.3.1 *Ambient pressure measurements*

Measurements at ambient pressure were performed on the spectrometer FOCUS (Paul Scherrer Institut, Switzerland) as explained in the Chapter 3. Spectra for the empty sample container, solvent and protein solutions were acquired. The experimental spectra were treated as mentioned in Chapter 3. The following procedure was used to estimate the sample container transmission in the presence of samples: *i)* firstly, the empty cell spectra were fitted with a normalized Gaussian function:

$$G(\omega;\omega_0,\sigma) = \frac{\exp\left(-\frac{(\omega-\omega_0)^2}{2\,\sigma^2}\right)}{\sqrt{2\,\pi}\,\sigma} \tag{7.26}$$

ii) secondly the Gaussian function with parameters $\{\omega_0, \sigma\}$ fixed from the previous fit on empty cell was used to estimate the contribution of the sample container in solvent spectra which were fitted by the following expression:

$$S_{solvent}(q,\omega) = A\ (\tau\ G(\omega;\omega_0,\sigma) + (1-\tau) S_B(q,\omega)) \tag{7.27}$$

where A is a normalization factor and

$$S_B(q,\omega) = \sum_i^2 \frac{1}{\pi} \frac{\Gamma_i^2}{(\Gamma_i)^2 + \omega^2}, \tag{7.28}$$

The expression for $S_B(q,\omega)$ is intended to describe global translation and rotation of solvent molecules. In order to limit the systematic errors due to the technical constraints listed above, the resulting spectra containing the contribution of the sample container, the solvent and the protein itself, were fitted with the fractional Brownian Dynamics model following the expression:

$$S_S^{mes}(q,\omega) = C(\alpha\ ((1-\tau')\ \tilde{S}_P(q,\omega) + \tau'\ \tilde{S}_E(q,\omega)) + (1-\alpha)\ \tilde{S}_B(q,\omega;\Gamma_1,\Gamma_2)) \tag{7.29}$$

where $\tilde{S}_P(q,\omega)$ was defined by expression 7.23, $\tilde{S}_B(q,\omega;\Gamma_1,\Gamma_2)$ was obtained from previous fit and $\tilde{S}_E(q,\omega))$ is the spectrum from the empty sample container. Here, $\tilde{S}$ represents

experimental spectra convoluted with instrumental energy resolution as estimated from vanadium spectra performed with the same setup of the other measurements.

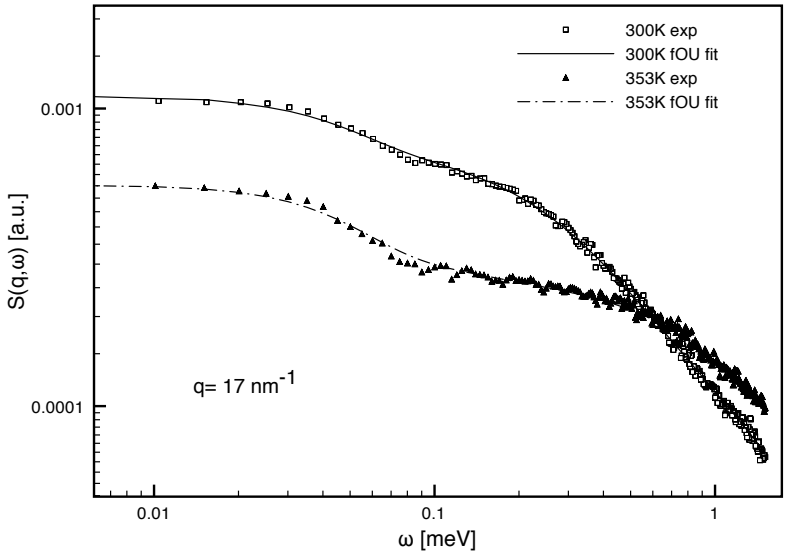

Figure 59: Log–log plot of aIF6 experimental QENS spectra for q = 17 nm^{-1} at 300K (hollow squares) and 353K (filled triangles) as a function of neutron energy gain ω. Lines represent the fits of the fOU model defined in the text.

Data were fitted in the range $[-1.5 : 0.1]meV$. The EISF obtained from this fit is shown in Figure 60 together with the EISFs obtained from MD simulations. Experimental data seem to agree very well with those obtained from MD simulation. The diffusion constant estimated from 7.23 is given in Figure 61 as a function of momentum transfer. The values can be compared with those given by direct calculation from the IF6 molecular dimensions. Here it is worth noting that the spherical approximation for the usual Einstein-Stokes expression of the translation diffusion constant would not fit very well in the case of IF6 as the protein shape can be better approximated by a disk or by a *torus*. These corrections to the spherical case can be performed by means of the Perrin factor and the explicit expression for *torus* diffusion constant (Appendix B). The values obtained from aIF6 were calculated using the volume associated to the solvent accessible surface and are given in Table 20. In the Perrin factor approximation the dimensions of the disk-shape which describe the IF6 structure was estimated by a orthorhombic box which included the whole molecular structure. Experimental values have to be corrected with respect of the viscosity of D_2O which is 20% higher than that

of H_2O. Hence taking into account this correction, values shown in Figure 61 are included between the Perrin and the *torus* approximations.

Table 20: Perrin correction to spherical diffusion constant obtained from bounding box and solvent accessible surface volume.

Temp&Press	axis a [Å]	axis b [Å]	Solv. Vol [10^3Å^3]	Stokes-Einstein D [10^{-3}Å^2/ps]	Perrin D [10^{-3}Å^2/ps]	torus [10^{-3}Å^2/ps]
300K-1bar	33.6	48.0	41.36	11.50	12.39	34.66
350K-1bar	34.2	46.3	40.99	34.23	36.64	95.34

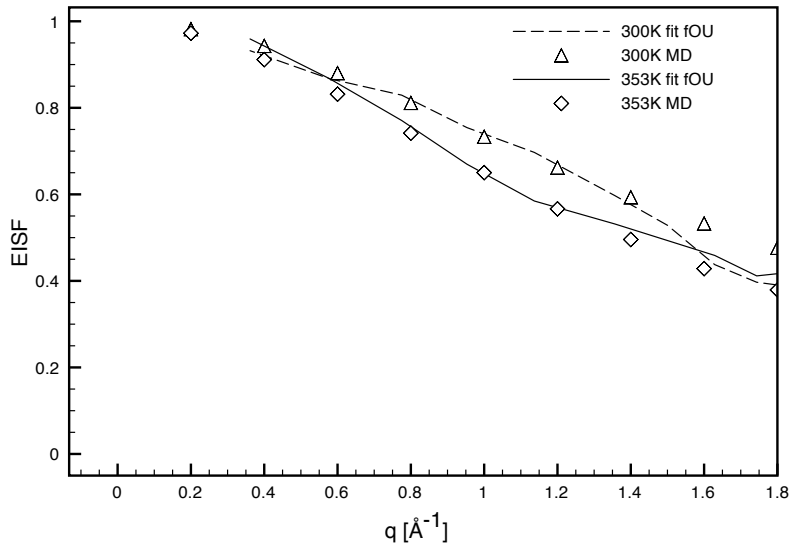

Figure 60: EISF from QENS measurements at ambient pressure and MD simulation of aIF6.

The values of α and τ parameters of the fOU model estimated from the fit of experimental QENS spectra at 300K and 350K are plotted in Figure 62. Interestingly, the values for parameter α are still coherent with those found from MD simulations whereas the values for τ are very different. The discrepancy in the evaluation of τ could be related to the limited energy range

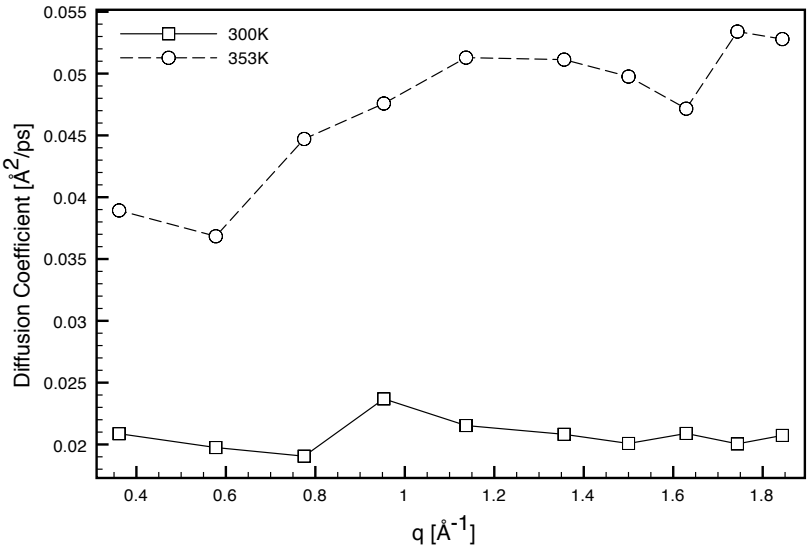

Figure 61: Diffusion coefficient as function of momentum transfer obtained from fractional Ornstein-Uhlenbeck from QENS measurements at ambient pressure.

used to perform this fit which was smaller than the one accessed by the time-correlation functions calculated from MD trajectories.

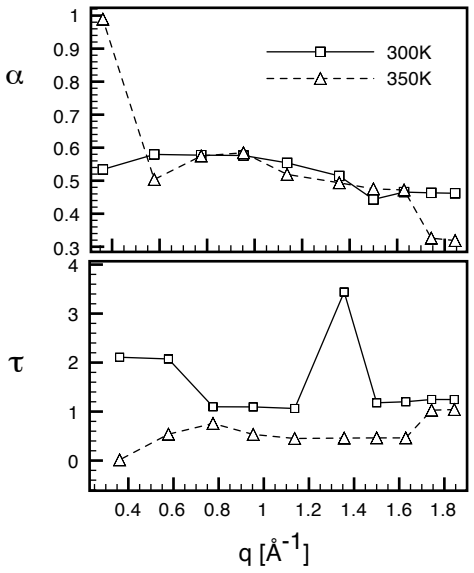

Figure 62: Parameters for the fractional Ornstein-Uhlenbeck process obtained from fits to QENS measurements at ambient pressure.

7.3.2 *High pressure measurements*

As mentioned in the chapter on experimental and simulated systems setups, all the QENS measurements performed in the pressure cell system required large volumes of protein solution. This was a limiting factor for the cleavage of the poly-histidine tag (HTag) used to enhance the expression protocol yield. For this reason, high pressure measurements were performed with aIF6 with HTag (aIF6-HTag). Also for these measurements the knowledge of precise protein concentration in sample solution would be essential for the correct analysis of spectra. As a consequence, a correct treatment for the analysis of quasi-elastic spectra was not possible. Nonetheless, acquired spectra were used to estimate EISF from aIF6-Htag with a qualitative approach which was however useful to partially support the observations made by MD simulations.

Spectra from empty cell, solvent and protein solution were taken and the same procedure for the estimation of sample container transmission used for ambient pressure measurements was used here.

The empty cell contribution was subtracted from solvent and protein solution spectra with a transmission $\tau = 0.90$. EISF from aIF6-Htag was estimated directly from the protein solution spectra from the expression:

$$S_{solution}(q,\omega) = B\ (\beta\ G(\omega;\omega_0,\sigma) + (1-\beta)L(q,\omega)) \tag{7.30}$$

where $L(q,\omega)$ which accounts for IF6 global and internal motions as well as for solvent quasi-elastic contribution and $G(\omega;\omega_0,\sigma)$ is a Gaussian function which should estimate the EISF from aIF6-HTag. The normalized fitted EISFs at different temperatures and pressures are shown in Figure 63. The MD simulation of aIF6-HTag was performed in order to compare experimental data. After filtering out the global motions as the other MD simulations, the scattering functions were calculated using the nMoldyn package. The EISFs from MD simulation are plotted together with experimental ones in Figure 63. Experimental results seem to underestimate EISF with respect to MD but the variation due to pressure and temperature are qualitatively similar. Here, it is worth noting that in general MD simulations tend to overestimate EISF due to the limited time length of trajectories which is far from being sufficiently large to correctly evaluate this quantity. From the experimental point of view, however, the EISF could be susceptible to systematic errors due to an incorrect subtraction of the sample container. Indeed, the latter could give a different contribution from the one

estimated from solvent spectra due to the absorption of protein sample and it could cause a underestimation of EISF values.

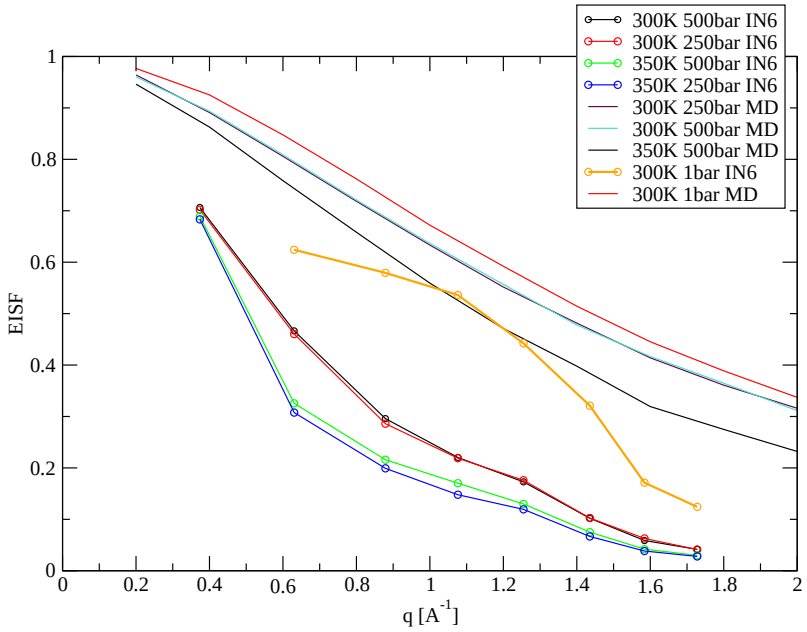

Figure 63: Comparison EISF from IN6 measurements and MD simulation of aIF6-HTag

7.4 DISCUSSION AND CONCLUSION

The effects of pressure and temperature on IF6s have been characterized here both from structural and from a dynamical points of view. Each approach gave different insights into the response of IF6s to environmental changes: *i)* global structural responses were investigated by means of radius of gyration and SASA whereas local effects were studied by RMSF analysis combined with ScrewFit profiles. The latter, besides giving significant insights into the conformational changes induced by pressure and temperature, allowed to assess secondary structure fluctuations induced by the same thermodynamic variables; *ii)* dynamical responses were instead investigated by means of scattering functions which allowed direct comparison between MD simulations and QENS experiments.

For the discussion of results shown in the previous sections, it is important to underline the role of the definition of *extreme conditions*: in contrast to what it is commonly thought, the conditions that request a significant re-adaptation of protein structures are not only those which have unfavorable chemical and thermodynamical conditions in the general sense. In fact, as already stated in the Introduction, some proteins seem to behave better at their own *natural* conditions than in other environmental situations, even though the former are, in principle, chemically and physically less favorable than the latter. This observation is comforted by works on enzymes activity which seemed to maintain *corresponding states* between different environmental conditions that requested evolutionary adaptation. These states are usually characterized by similar conformational flexibility [92].

STRUCTURE As a whole the effects of pressure and temperature on IF6 structures showed that the anti-association factor 6 from *Methanococcus Jannaschii* (aIF6) is much less sensitive than its mesophilic counterpart from *Saccharomyces cerevisiae* (eIF6) with respect to *extreme* conditions, as shown by variation in solvent accessible surface and in atomic root mean square fluctuations. In particular it was also found that its structural properties at high temperature and high pressure were very similar to those of eIF6 at 300K-1bar. This response is characterized by fluctuations of α-carbons in the protein backbone and by variations in secondary structures.

Here, the presence of *corresponding states* was made evident from observations on structural fluctuations and by the comparison of different EISFs which proved by its relation to the mean square displacement that this process is linked to the change on spatial confinement of protein internal motions. The same type of results were obtained by Tehei and coworkers [191, 192] in the comparison of molecular dynamics in thermophilic and mesophilic proteins. As a whole, these results suggest that *corresponding states* are created by local structural re-arrangements that influences atomic motions in proteins. For this purpose, evolution plays on sequence point mutations to give the right adjustment of intramolecular interactions and thus obtain the *desired* effect on protein structure stability.

Comparison between the EISFs of eIF6 and aIF6 gave the evidence that MSDs are very similar in the respective *natural* conditions. The comparison of these results with those previously obtained on the EISF from lysozyme [20] at ambient pressure (Figure 64), shows a

surprising correspondence suggesting the presence of a condition for *optimal* "resilience" of protein structures.

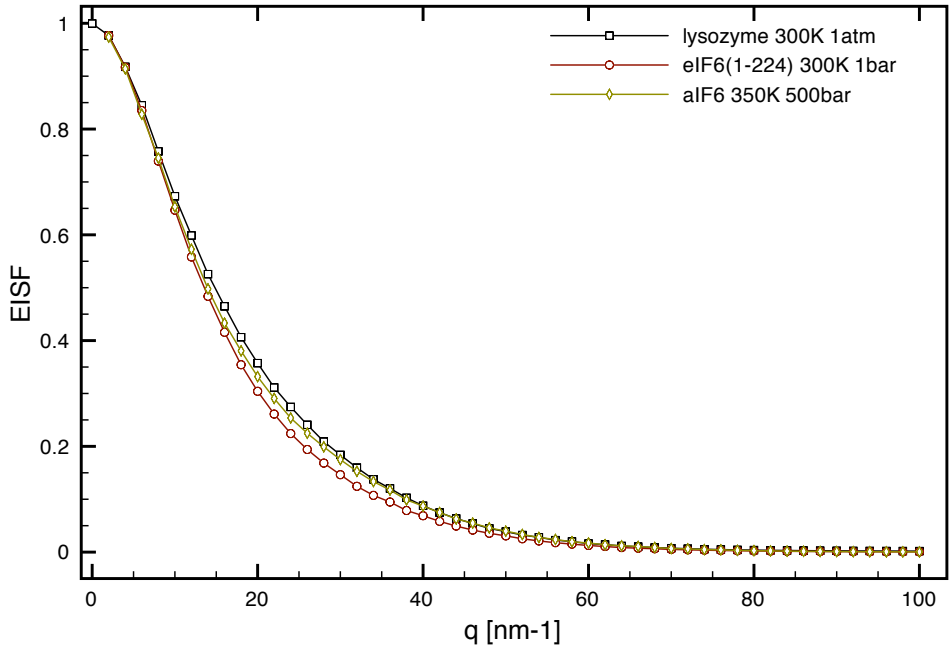

Figure 64: Elastic Incoherent Structure Factor for aIF6, eIF6 and lysozyme.

In IF6's protein family this condition seems to be achieved thanks to a tail of 21 amino acids (CTAIL) attached to the C-terminus of the more evolutionarily conserved part of IF6, because its presence was the source of the lower stability of eIF6 in extreme conditions, whereas its absence made eIF6-NoCTAIL behave like aIF6. Moreover, the role of CTAIL is made even more important by the fact that it induces large thermal fluctuations to structural regions which are supposed to play a direct role in IF6 functions.
Interestingly, similar results have been obtained on EISFs from aIF6-HTag indicating that even the presence of a fluctuating non-structured tail attached on the N-terminus of aIF6 could make the latter not suited for *warm deep sea* environments.

The application of the ScrewFit method to time-averaged structures allowed also some local effects of pressure and temperature to be detected and characterized by comparison with three-dimensional structures. The main outcome of this analysis revealed that pressure

and temperature have locally different effects on IF6 structures. Some of these effects were enhanced by separate application of high pressure and high temperature, whereas they disappeared when both conditions where applied simultaneously.

DYNAMICS The investigation of the dynamical effects refined the structural results. Scattering functions from the MD simulations of both aIF6 and eIF6 were analyzed by means of fractional Brownian dynamics (fOU) model. The latter was proved in other works to be able to finely characterize the multiple relaxation time dynamics of proteins.

The fOU model showed that aIF6 and eIF6 can be distinguished from each other by their dynamical properties. In particular it was shown that they have different characteristic q-dependent relaxation times and give different responses to pressure application. Indeed, when pressure is applied one would expect that slower diffusive motions on large length scales are slowed down since they require large scale spatial rearrangements which are increasingly hindered under pressure. Parameter τ from fit of aIF6 seems to confirm this behaviour whereas it does not for eIF6 in which it is lowered by pressure increase. This fact could be explained by the possible denaturation of eIF6, which has already been observed in real experiments and that could be *visible* already in the nanosecond time-scale.

This outcome suggests that fOU parameters are able to *capture* dynamical properties of each protein, being able to pinpoint differences between two homologues of the same protein family. Moreover, it seems evident that if corresponding states can be found from the atomic MSD, the fOU model can chracterize how unfavorable environmental conditions differentiate from the physiological one by means of the dynamical parameters.

Indeed, if variation of dynamical properties due to changes in pressure and temperature, must be related to each protein characteristic, the general framework in which this variation happens can somehow be a common property of protein families. In this context comparison of relaxation times in different environmental conditions showed that the approaching of a favorable environment corresponds to a non-uniform change in relaxation times. In particular motions involving large domains are less influenced than more localized motions.

The limited amount and quality of experimental data did not permit to completely verify this hypothesis made by means of MD simulations. QENS experiments have shown however consistency with MD results. In particular EISF and the translational diffusion constant found by experiment at ambient pressure was found to be comparable with those estimated by MD simulations. High pressure experiments were largely biased by technical problems and could not give any quantitative result. Nevertheless, a qualitative comparison with MD simulation was possible by estimating the EISF contribution from the protein solution spectra.

8

GENERAL CONCLUSIONS AND PERSPECTIVES

In the introduction was stated that this work was proposed to give an answer to the following questions:

Where does the extremophilic signature come from ?

If structure cannot be the origin for this, can it be the dynamics??

The test case for this extremophilic signature searching was to be found in structural and dynamical properties of the Inititation Factor 6 from *Methanococcus Jannaschii*, an archaebacteria which lives in the warm deep sea, near the hydrothermal chemineys. This protein was chosen for its important role into the synthesis of other proteins, through its participation into the biogenesis of major ribosome subunits and the limitation of association of the two subunits.

IF6 homologue from Saccharomyces cerevisiae was chosen as "normal" conterpart for a detailed comparison. The approach used in this work, which combined MD simulations and QENS experiments, presented several new aspects to both techniques that requested the development of novel theoretical and experimental methods. Firstly, a new computational method was needed to finely pinpoint even small changes in protein structures induced by environmental changes. For this purpose, a method based on quaternion superposition fits in conjunction with Chasles' theorem on rigid body motions was developed. This method, called *ScrewFit*, proved to be efficient in finding variations in secondary structure of proteins and in characterizing them with respect to the deformations in protein structural motifs. The method found also further developments in protein secondary structure assignments and in finding the structural effects due to ligand binding in enzymes.

Secondly, the two IF6 homologues were entirely new samples for QENS experiments and demanded new protocols for expression and purification in order to obtain adequate sample concentration in final solutions. The developed protocol reported here had a yield sufficient for neutron scattering measurements of protein solution. Nevertheless, it also involved other technical complications due to the presence of a N-terminal poly-histidine tag which was needed to improve the total yield of protein purification. The

clavage of the supplemental histidine-tag in large volumes needed for high pressure measurements is prohibitive due to its efficiency, the latter being inversely proportional to volume.

The major results of this work are related to the finding of "corresponding states" in structure and dynamics of each homologue in its respective *natural* condition. These "states" are characterized by very similar atomic mean square displacements, meaning that the atomic internal motions in respective natural conditions are confined to the same length scales. This type of correspondence was already found in other works on other proteins both experimentally [192] and by MD simulations[69]. These results were formerly interpreted as an indication of different "resiliences" of extremophile and mesophile proteins. In this work I showed that even if corresponding states are found, they do not involve the same dynamical behavior in extremophilic and mesophilic homologues. Nevertheless, the variation of characteristic time as function of length-scale seems to be able to capture the optimality of conditions.

Throughout this thesis the dynamical characterization of IF6 samples modeled by MD simulations was pursued using a fractional Brownian dynamics model which proved to efficiently characterize the multi-time-scale heterogeneity in protein internal motions.

Besides the general discussion about "corresponding states", this work showed also that in the case of IF6, pressure and temperature induce different local effects in protein structures. Some of these effects seems to affect regions that probably host functionality of IF6s. The structural analysis by means of Screwfit parameter profiles showed that pressure and temperature effects are neither equivalent nor of opposite sign. They are qualitatively different. The present knowledge of these effects does not permit the association of the latter to the global effects inscribed into the formation of "corresponding states".

The experimental results which were very limited by the absence of adequate protocols and previously unknown sample instability, could however support indirectly results obtained from MD simulations.

PERSPECTIVES This work proposed a new approach to the investigation of evolutionary adaptation of proteins to extreme environments from both a structural and a dynamical point of view. Nevertheless, its novelty had to face some experimental limitations. This did not prevent experimental results being obtained which supported the MD simulations.

Nevertheless, the results shown here should be verified by MD simulations also for longer time-scales in order to better explore dynamical properties of IF6s. In this context, some neutron backscattering experiments, which access longer time-scales, have already been performed during this thesis on the IN16 (Institut Laue Langevin, Grenoble, France) and HFBS (NIST, Gaithersbourg, USA) spectrometers.

As a whole, this work must be considered preliminary and requires further development both in expression/purification protocols of samples and in technical instrumentation for QENS experiments. In fact, the latter may become a limiting factor when performing measurements on protein solutions which require stable samples even at very high concentrations.

This thesis suggested, however, that IF6 constitutes a very interesting sample for studies on molecular evolution and it is worth being further investigated both by QENS and other molecular spectroscopy techniques. In particular small angle scattering (X-ray and neutrons) should be envisaged to obtain more insights into the structural effects of different environmental conditions and the role of CTAIL in eIF6 stability.

In addition to its role in molecular evolution studies, IF6 dynamical properties could be explored by neutron scattering also in relation to its recently uncovered role in regulating the human tumor cell development[172].

Finally this work constitutes a new approach for investigating molecular evolution. Similar approaches have been recently used to investigate the relation between other physical properties of proteins and their evolutionary history[215].

Part III

APPENDIX

A

BUFFERS USED FOR PROTEIN EXPRESSION AND PURIFICATION

Table 21: Buffers used for cell lysis and protein purification

Lysis buffer	50mM Tris-HCl pH 7.4 200mM NaCl 5% Glycerol 1mM PMSF
Wash buffer	50mM Tris-HCl pH 7.4 1M NaCl 5% Glycerol 1mM β-mercaptoethanol
Elution buffer	50mM Tris-HCl pH7.4 200mM NaCl 5% Glycerol 200mM Imidazole 1mM β-mercaptoethanol

Table 22: Buffers for dyalisis and storage

Dialyse buffer	20mM Tris pH 7.9 200mM NaCl 5mM DTT
Storage buffer	50mM Tris-HCl pH 7.4 200mM NaCl 5% glycerol 5mM DTT

B

CORRECTIONS TO THE STOKES' LAW FOR SPHERE DIFFUSION.

PERRIN FRICTION FACTORS Perrin factor is a correction to the translational friction in the case of a rigid spheroid characterized by an **axial ratio** $\sigma = a/b$ where a and b are the axial and the equatorial semiaxis, respectively. For $p > 1$ one has *stick-shaped* bodies and for $p < 1$ *disc-shaped* bodies.

The Perrin factor is defined through a multiplicative correction term to the friction coefficient of a sphere:

$$f_{cor} = f'_{sphere} \cdot f_P \tag{B.1}$$

f'_{sphere} is the friction constant obtained for a sphere of equivalent volume of the spheroid body.

The factor f_P is defined by

$$f_P = \frac{2\sigma^{2/3}}{S} \tag{B.2}$$

where,

$$\sigma = \frac{a}{b}$$

$$S = 2 \cdot \frac{\arctan\left(\frac{\sqrt{|\sigma^2-1|}}{\sigma}\right)}{\frac{\sqrt{|\sigma^2-1|}}{\sigma}}$$

Hence, the diffusion constant for a spheroid reads as follow:

$$\boxed{D_{sp} = \frac{K_B T}{f'_{sphere}} \cdot \frac{S}{2\sigma^(2/3)}}$$

DIFFUSION CONSTANT FOR A TORUS The derivation of the translational diffusion constant for a torus can be found in [193]. Assuming rotational and traslational diffusion as decoupled, one obtains for the translational part:

$$\boxed{D_{sp} = \frac{K_B T}{8\pi\mu a} \cdot \left(\log\left(\frac{8a}{b}\right) + \frac{1}{2}\right)}$$

where a and b are defined in Figure 65. μ is the viscosity.

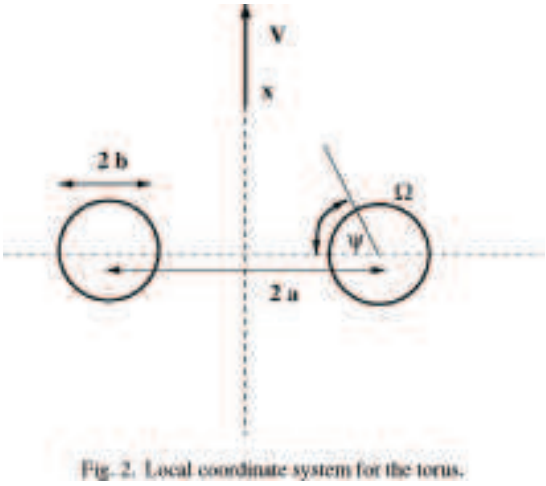

Fig. 2. Local coordinate system for the torus.

Figure 65: Local coordinates of a Torus reprinted from reference [193]

Torus correction To perform the correction due to the toroidal shape of the protein I estimated the values of a and b defined as in Figure 65. The *inner* and *outer* radius are calculated averaging over the distance of α-carbons respectively in the residues flanking the central hollow and in the five opposite sides of the pentameric pseudo-symmetry.

The values shown in the text were obtained using the following viscosities (disregarding the differences between heavy and water):

Table 23: Viscosity of water and heavy water as a function of temperature.

Temp Kelvin	μ_{H_2O} $[10^{-3} Pa \cdot s]$	μ_{D_2O} $[10^{-3} Pa \cdot s]$	μ_{D_2O}/μ_{H_2O}
300	0.89	1.11	1.25
350	0.35	0.45	1.29

BIBLIOGRAPHY

[1] F. Abe and K. Horikoshi. The biotechnological potential of piezophiles. *Trends Biotechnol.*, 19(3):102–108, 2001. (Cité aux pages xviii et 20.)

[2] M Abramowitz and IA Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1972. (Cité à page 136.)

[3] MP Allen and DJ Tildesley. *Computer Simulation of Liquids*. 1987. (Cité aux pages xxi et 37.)

[4] S.L. Altmann. *Rotations, Quaternions, and Double Groups*. Clarendon Press, Oxford, 1986. (Cité aux pages xxx, 73 et 97.)

[5] CAF Andersen, AG Palmer, S Brunak, and B Rost. Continuum secondary structure captures protein flexibility. *Structure*, 10:175–184, 2002. (Cité aux pages 94, 100, 107 et 108.)

[6] A Balbo and S. Bozzaro. Cloning of dictyostelium eif6 (p27bbp) and mapping its nucle(ol)ar localization subdomains. *Eur J Cell Biol*, 85(9-10):1069–78, Sep 2006. doi: 10.1016/j.ejcb.2006.05.010. (Cité aux pages xxiii, 25, 52 et 61.)

[7] R.L. Baldwin. Temperature dependence of the hydrophobic interaction in protein folding. *Proc Natl Acad Sci USA*, 83(21):8069–72, 1986. (Cité à page 12.)

[8] R.D. Ballard. Notes on a major oceanographic find. *Oceanus*, 20:35–40, 1977. (Cité à page 19.)

[9] D.J. Barlow and J.M. Thornton. Helix geometry in proteins. *J. Mol. Biol.*, 201:601–619, 1988. (Cité aux pages 72, 80, 82, 83, 94 et 113.)

[10] U Basu, K Si, J R Warner, and U Maitra. The *saccharomyces cerevisiae* TIF6 gene encoding translation initiation factor 6 is required for 60S ribosomal subunit biogenesis. *Molecular and Cellular Biology*, 75(5):1453–1462, 2001. (Cité aux pages xx et 25.)

[11] U Basu, K Si, H Deng, and U Maitra. Phosphorylation of mammalian eukaryotic translation initiation factor 6 and its saccharomyces cerevisiae homologue tif6p: Evidence that

phosphorylation of tif6p regulates its nucleocytoplasmic distribution and is required for yeast cell growth. *Molecular and Cellular Biology*, 23(17):6187–6199, 2003. (Cité aux pages 25 et 124.)

[12] O. M. Becker, A. D. Jr. MacKerell, B. Roux, and M. Watanabe, editors. *Computational Biochemistry and Biophysics*. CRC Press, 2001. (Cité à page 62.)

[13] M Bée. *Quasielastic Neutron Scattering: Principles and Applications in Solid State Chemistry, Biology and Materials Science*. 1988. (Cité aux pages xxi, 38 et 134.)

[14] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and J.R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984. (Cité aux pages xxvi et 36.)

[15] J.P. Boon and S. Yip. *Molecular Hydrodynamics*. McGraw Hill, 1980. See eqs. (3.5.23), (3.5.33), (3.5.37) and the corresponding references. (Cité à page 129.)

[16] J.F. Brandts, R.J. Oliveira, and C. Westort. Thermodynamics of protein denaturation. effect of pressu on the denaturation of ribonuclease a. *Biochemistry*, 9:1038–47, 1970. (Cité à page 10.)

[17] P.W. Bridgman. The coagulation of albumem by pressure. *J. Biol. Chem.*, 19:511–512, 1914. (Cité à page 8.)

[18] CJ Bult, O White, GJ Olsen, LX Zhou, RD Fleischmann, GG Sutton, JA Blake, LM FitzGerald, RA Clayton, JD Gocayne, AR Kerlavage, BA Dougherty, JF Tomb, MD Adams, CI Reich, R Overbeek, EF Kirkness, KG Weinstock, JM Merrick, A Glodek, JL Scott, NSM Geoghagen, JF Weidman, JL Fuhrmann, D Nguyen, TR Utterback, JM Kelley, JD Peterson, PW Sadow, MC Hanna, MD Cotton, KM Roberts, MA Hurst, BP Kaine, M Borodovsky, HP Klenk, CM Fraser, HO Smith, CR Woese, and JC Venter. Complete genome sequence of the methanogenic archaeon, Methanococcus Jannaschii. *Science*, 273(5278):1058–1073, 1996. ISSN 0036-8075. (Cité aux pages xix, 23 et 47.)

[19] V Calandrini and G Kneller. Influence of pressure on the slow and fast fractional relaxation dynamics in lysozyme: A simulation .... *The Journal of Chemical Physics*, Jan 2008. (Cité à page 135.)

[20] V Calandrini, V Hamon, K Hinsen, P Calligari, M.-C Bellissent-Funel, and G.R.Kneller. Relaxation dynamics of lysozyme in solution under pressure: Combining molecular

dynamics simulations .... *Chemical Physics*, 345:289–297, Jan 2008. (Cité aux pages xxxvii, 55, 135 et 152.)

[21] H.B. Callen. *Thermodynamics and an introduction to Thermostatistics*. John Wiley, 1985. (Cité aux pages 9 et 10.)

[22] PA Calligari, GR Kneller, A Giansanti, P Ascenzi, A Porrello, and A Bocedi. Inhibition of viral group-1 and group-2 neuraminidases by oseltamivir: a comparative structural analysis by the screwfit algorithm. *Submitted*, 2007. (Cité à page 123.)

[23] M Carson, D Johnson, H McDonald, and C Brouillette. His-tag impact on structure. *Acta Crystallographica Section D*, 63(3):295–301, 2007. (Cité aux pages xxii et 50.)

[24] J.V. Carter, D.G. Knox, and A. Rosenberg. Pressure effects on folded proteins in solution. hydrogen exchange at elevated pressures. *J. Biol. Chem.*, 253(6):1947–53, 1978. (Cité à page 12.)

[25] P Carter, CAF Andersen, and B Rost. Dsspcont: continuous secondary structure assignements for proteins. *Nucleic Acids Research*, 31(13):3293–3295, 2003. (Cité à page 94.)

[26] D.A. Case, T.E. CheathamIII, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. Woods. The amber biomolecular simulation programs. *J. Comp. Chem.*, 26:1668–1688, 2005. (Cité à page 30.)

[27] D.A. Case, T.A. Darden, T.E. CheathamIII, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, D.A. Pearlman, M. Crowley, R.C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K.F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D.H. Mathews, C. Schafmeister, W.S. Ross, and P.A. Kollman. *AMBER 9*. University of California, San Francisco., 2006. (Cité aux pages xxv et 59.)

[28] M Ceci, C Gaviraghi, C Gorrini, L A Sala, N Offenhauser, PC Marchisio, and S Biffo. Release of eIF6 (p27BBP) from the 60S subunit allows 80S ribosome assembly. *Nature*, 426:579–584, 2003. (Cité aux pages xx, 25 et 26.)

[29] A. Certes. Sur la culture, à l'abri des germes atmosphériques, des eaux et des sédiments rapportés par les expéditions du travailleur et du talisman. *Comptes Rendus, Accadémie des Sciences*, 98:690–693, 1884. (Cité aux pages xviii et 19.)

[30] A. Certes. De l'action des hautes pressions sur les phénomènes de la putréfaction et sur la vitalité des micro-organismes d'eau douce et d'eau de mer. *Comptes Rendus, Accadémie des Sciences*, 99:385–388, 1884. (Cité aux pages xviii, 8 et 19.)

[31] JM Chandonia, G Hon, NS Walker, L Lo Conte, P Koehl, M Levitt, and S E Brenner. The astral compendium in 2004. *Nucleic Acids Research*, 32:D189–D192, 2004. (Cité à page 99.)

[32] M Chasles. Note sur les propriétés générales du système de deux corps sembables entr'eux, et plaçés d'une manière queconque dans l'espace. *Bulletin des Sciences Mathématiques, Astronomiques, Physiques et Chimiques*, 14:321–326, 1830. (Cité aux pages xxx, 72 et 96.)

[33] M. Chasles. *Rapport sur les Progrès de la Géométrie en France*. Imprimérie Nationale, Paris, 1870. See p. 77. (Cité aux pages xxx, 72 et 96.)

[34] TP Chendrimada, KJ Finn, X Ji, D Baillat, RI Gregory, SA Liebhaber, AE Pasquinelli, and R. Shiekhattar. Microrna silencing through risc recruitment of eif6. *Nature*, 447: 823–8, 2007. (Cité à page 25.)

[35] G. Ciccotti and J.P. Ryckaert. Molecular dynamics simulation of rigid molecules. *Comp. Phys. Rep.*, 4:345–392, 1986. (Cité à page 64.)

[36] B. Cichocki, B.U. Felderhof, and R. Schmitz. Hydrodynamic interactions between two spherical particles. *PhysicoChemical Hydrodynamics*, 10:383, 1988. (Cité à page 57.)

[37] B. Cichocki, B.U. Felderhof, K. Hinsen, E. Wajnryb, and J. Blawdziewicz. Friction and mobility of many spheres in Stokes flow. *J. Phys. Chem.*, 100(5):3780–3790, 1993. (Cité à page 135.)

[38] P. Cioni and G.B. Strambini. Pressure effects on protein flexibility in monomeric proteins. *J. Mol. Biol.*, 242(3):291–301, 1994. (Cité à page 12.)

[39] M. D. Collins, G. Hummer, M.L. Quillin B.W. Matthews, and S. M. Gruner. Cooperative water filling of a nonpolar protein cavity observed by high-pressure crystallography and simulation. *Proc Natl Acad Sci USA*, 102(46):16668–16671, 2005. (Cité à page 14.)

[40] N Colloc'h, C Etchebest, E Thoreau, B Henrissat, and JP Mornon. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.*, 6(4):377–82, 1993. (Cité aux pages 94, 103 et 106.)

[41] A.R. Cossins and A.G. Macdonald. The adaptation of biological membranes to temperature and pressure; fish from the deep and cold. *J. Bioenerg. Biomembr.*, 21:115–35, 1989. (Cité à page 20.)

[42] T.E. Creighton. *Proteins, Structure and Molecular Properties*. Freeman and Company, New York, 5th edition, 1997. (Cité aux pages 8, 11 et 13.)

[43] T.E. Creighton. Disulphide bonds and protein stability. *Bioessays*, 8(2-3):57–63, 2005. (Cité à page 11.)

[44] E.P. Dahlhoff and G.N. Somero. Pressure and temperature adaptation of cytosolic malate dehydrogenases of shallow- and deep-living marine invertebrates: evidence for high body temperature in hydrothermal vents aniumals. *J. Exp. Biol.*, 159:473–87, 1991. (Cité à page 20.)

[45] E.P. Dahlhoff, S. Schneidemann, and G.N. Somero. Pressure-temperature interactions on M_4-lactate dehydrogenases from hydrothermal vents fishes. *Biol. Bull. Woods Hole*, 179:134–139, 1990. (Cité aux pages 20 et 21.)

[46] S.W. de Leeuw, J.W. Perram, and E.R. Smith. Simulation of electrostatic systems in periodic boundary conditions. I. lattice sum and dielectric constant. *Proc. R. Soc. Lond. A*, 373:27–56, 1980. (Cité à page 33.)

[47] H. De Smedt, H. Olbrechts, and K. Heremans. Phase transi- tions in artificial membranes: high pressure light scattering studies. *Europhys. Conf. Abstr. A, Leuwen*, 1:98–99, 1975. (Cité à page 10.)

[48] G. Decket and P.V. Warren. The complete genome sequence of the hyperthermophileic bacterium aquifex aeolicus. *Nature*, 392:353–358, 1996. (Cité aux pages xvii et 17.)

[49] Y. Dehouck, B. Folch, and M. Rooman. Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. *Protein Eng. Des. Sel.*, 21(4):275–278, 2008. (Cité à page 17.)

[50] R Diamond. A note on the rotational superposition problem. *Acta Cryst. A*, 44:211–216, 1988. (Cité aux pages xxx, 96 et 97.)

[51] H.G. Drickamer and Franck C.W. *Electronic transition and high pressure chemistry and physics of solids*. London: Chapman and Hall, 1973. (Cité à page 11.)

[52] A. Erdélyi, W. Magnus, F. Oberhettinger, and F.G. Tricomi. *Higher Transcendental Functions*. McGraw Hill, 1955. (Cité à page 136.)

[53] P. Ewald. Die Berechnung optischer und electrostatischer Gitterpotentiale. *Ann. Phys.*, 64:253–87, 1921. (Cité à page 33.)

[54] H Fischer. Average protein density is a molecular-weight-dependent function. *Protein Science*, 13(10):2825–2828, Oct 2004. doi: 10.1110/ps.04688204. (Cité à page 58.)

[55] R.J. Frankenberg, M. Andersson, and D.S. Clark. Effect of temperature and pressure on the proteolytic specificity of the recombinant 20S proteasome from *methanococcus jannschii*. *Extremophiles*, 7:353–360, 2003. (Cité aux pages xix et 23.)

[56] D. Frenkel and B. Smit. *Understanding Molecular Simulation*. Academic Press, London, San Diego, 1996. (Cité aux pages xxi et 37.)

[57] D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins*, 23:566–579, 1995. (Cité aux pages 72, 94 et 103.)

[58] F. Gaill. Aspects of life development at deep sea hydrothermal vents. *FASEB J.*, 7(6): 558–565, 1993. (Cité aux pages xvi et 16.)

[59] C.W. Gardiner. *Handbook of Stochastic Methods*. Springer Series in Synergetics. Springer, Berlin, Heidelberg, New York, 2nd edition, 1985. (Cité à page 136.)

[60] K. Gekko and Y. Hasegawa. Compressibility/structure relationship of globular proteins. *Biochemistry*, 25:6563–6571, 1986. (Cité à page 14.)

[61] K. Gekko and Y. Hasegawa. Effect of temperature on the compressibility of native globular proteins. *J. Phys. Chem.*, 93:426–429, 1989. (Cité à page 14.)

[62] K. Gekko and H. Noguchi. Compressibility of globular proteins in water at 25 c. *J. Phys. Chem.*, 83:2706–2714, 1979. (Cité à page 11.)

[63] M. Gerstein and C. Chothia. Packing at the protein-water interface. *Proc Natl Acad Sci USA*, 93:101167–101172, 1996. (Cité à page 59.)

[64] W.G. Glöckle and T.F. Nonnenmacher. A fractional calculus approach to self-similar protein dynamics. *Biophys. J.*, 68:46–53, 1995. (Cité aux pages 135 et 137.)

[65] K. Goosens, L. Smeller, and K. Heremans. Pressure tuning spectroscopy of the low-frequency raman spectrum of liquid amides. *J. Chem. Phys.*, 99:5736–5741, 1993. (Cité à page 12.)

[66] C M Groft, R Beckmann, A Sali, and S K Burley. Crystal structures of ribosome anti-association factor IF6. *Nature Structural Biology*, 7(12):1156–64, Dec 2000. doi: 10.1038/82017. (Cité aux pages xx, xxi, xxii, 25, 26, 27, 51, 60, 61, 64, 121 et 124.)

[67] M. Gross and R. Jaenicke. Proteins under pressure. *Eur. J. Biochem.*, 221:617–630, 1994. (Cité aux pages 11, 13, 19 et 22.)

[68] M. Gross, K. Lehle, R. Jaenicke, and K.H. Nierhaus. Pressure-induced dissociation of ribosomes and elongation cycle intermediates. stabilizing conditions and identification of the most sensitive functional state. *Eur. J. Biochem.*, 218:463–468, 1993. (Cité à page 22.)

[69] A Grottesi, MA Ceruso, A Colosimo, and A Di Nola. Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. *Proteins Structure Function and Genetics*, 46(3):287–294, 2002. (Cité aux pages xvii, 18 et 158.)

[70] G.D. Hawkins, C.J. Cramer, and D.G. Truhlar. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.*, 246:122–129, 1995. (Cité à page 62.)

[71] G.D. Hawkins, C.J. Cramer, and D.G. Truhlar. Parametrized models of aqueous free ener- gies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.*, 100:19824–19839, 1996. (Cité à page 62.)

[72] S.A. Hawley. Reversible pressure-temperature denaturation of chymotrypsinogen. *Biochemistry*, 10:2436–42, 1971. (Cité à page 10.)

[73] S.A. Hawley and R.M. Macleod. The effect of base composition on the pressure stability of dna in neutral salt solution. *Biopolymers*, 16:1833–35, 1977. (Cité à page 10.)

[74] D.J. Hei and D.S. Clark. Pressure stabilization of proteins from extreme thermophiles. *Appl. Environ. Microbiol.*, 60:932–939, 1994. (Cité aux pages xix, 13 et 23.)

[75] K. Heremans and L. Smeller. Protein structrure and dynamics at high pressure. *Biochem. et Biophys. Acta*, 1386:353–370, 1998. (Cité à page 8.)

[76] L. Heremans and K. Heremans. Pressure effects on the raman spectrum of proteins: stability of the salt bridge in trypsin and elastase. *J. Mol. Struct.*, 214:305–314, 1989. (Cité à page 11.)

[77] L. Heremans and K. Heremans. Raman spectroscopic study of the changes in secondary structure of chymotrypsin: Effect of ph and pressure on the salt bridge. *Biochem. et Biophys. Acta*, 999:192–197, 1989. (Cité à page 11.)

[78] D.M. Heyes. Electrostatic potentials and fields in infinite point charge lattice. *J. Chem. Phys.*, 74:1924–29, 1981. (Cité à page 33.)

[79] K Hinsen. The molecular modeling toolkit: A new approach to molecular simulations. *J. Comp. Chem.*, 21:79–85, 2000. (Cité à page 91.)

[80] K. Hinsen and G.R Kneller. A simplified force field for describing vibrational protein dynamics over the whole frequency range. *J. Chem. Phys.*, 111(24):10766–10769, 1999. (Cité à page 14.)

[81] K Hinsen and GR Kneller. Projection methods for the analysis of complex motions in macromolecules. *Mol. Sim.*, 23:275–292, 2000. (Cité à page 99.)

[82] K. Hinsen, A.-J. Petrescu, S. Dellerue, M.C. Bellissent-Funel, and G.R. Kneller. Harmonicity in slow protein dynamics. *Chem. Phys.*, 261(1+2):25–38, 2000. Special Issue "Condensed Phase Structure and Dynamics: A combined neutron scattering and molecular modelling approach". (Cité à page 14.)

[83] G. Hui Bon Hoa, M. Graffe, and M. Grunberg-Manago. Thermodynamic studies of the reversible association of escherichia coli ribosomal subunits. *Biochemistry*, 16(12):2800–5, 1977. (Cité à page 22.)

[84] U Hobohm and C Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522, 1994. (Cité à page 99.)

[85] K. Horikoshi. Barophiles: deep-sea microorganisms adapted to an extreme environment barophiles: deep-sea microorganisms adapted to an extreme environment. *Curr. Opin. Microbiol.*, 1:291–295, 1998. (Cité aux pages xvi et 16.)

[86] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, Nov 2006. doi: 10.1002/(ISSN)1097-0134. (Cité aux pages xxv, 30 et 59.)

[87] G. Hummer, S. Garde, A.E. Garcia, M.E. Paulaitis, and L.R. Pratt. The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *PNAS*, 95:1552–1555, 1998. (Cité aux pages 13 et 14.)

[88] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd – visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996. (Cité à page 91.)

[89] G Hutchinson. Promotif - a program to identify structural motifs in proteins. *Protein Science*, 5:212–220, 1996. (Cité aux pages 83 et 94.)

[90] J.K. Ichikawa and S. Clarke. A highly active protein repair enzyme from the extreme thermophile t.marittima. *Arch. Biophys. Biochem.*, 358:222–231, 1998. (Cité à page 17.)

[91] R. Jaenicke. *Current perspectives in high pressure biology*, pages 257–272. Academic Press, 1987. (Cité à page 22.)

[92] R. Jaenicke. Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem.*, 202:715–728, 1991. (Cité aux pages xvi, xvii, 16, 17, 18 et 152.)

[93] R Jaenicke. Stability and stabilization of globular proteins in solution. *J Biotechnol*, 79 (3):193–203, May 2000. (Cité aux pages xvi, xvii, 10, 16, 17 et 18.)

[94] R Jaenicke. Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity? *Proceedings of the National Academy of Sciences*, 97(7):2962, 2000. (Cité aux pages 11 et 17.)

[95] R. Jaenicke and G. Bohm. Stabilization of proteins: what extremophiles teach us about protein stability. *Curr. Opin. Struct. Biol.*, 8:738–748, 1998. (Cité aux pages 13 et 17.)

[96] R. Jaenicke and G. Bohm. Thermostability of proteins from thermotoga maritima. *Methods Enzymol*, 334:438–469, 2001. (Cité aux pages xvii et 18.)

[97] R. Jaenicke, G. Bernhardt, H.D. Ludemann, and K.O. Stetter. Pressure induced alterations in the protein pattern of the thermophilic arcaebacterium methanococcus thermolithotrophicus. *Appl. Environ. Microbiol.*, 1988:2375–80, 54. (Cité à page 19.)

[98] H.W. Jannasch and C.D. Taylor. Deep-sea microbiology. *Ann. Rev. Microbiol.*, 38:487–514, 1984. (Cité aux pages xviii, 19 et 21.)

[99] J. Jonas. High-resolution nuclear magnetic resonance studies of proteins. *Biochimica Biophysica Acta*, 1595(1-2):145–159, 2002. (Cité à page 8.)

[100] J. Jonas, A. Jonas, and Jon. High-pressure nmr spectroscopy of proteins and membranes. *Ann. Rev. Biophys. Biomol. Struct.*, 23:287–318, 1994. (Cité à page 15.)

[101] DT Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999. (Cité à page 52.)

[102] W. Jones, J. A. Leigh, F. Mayer, C. R. Woese, and R. S. Wolfe. *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Arch. Microbiol.*, 136:254, 1963. (Cité aux pages xix, 22 et 23.)

[103] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926, Jan 1983. doi: 10.1063/1.445869. (Cité à page 31.)

[104] WL Jorgensen, J Chandrasekhar, JD Madura, RW Impey, and ML Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926, 1983. (Cité à page 31.)

[105] C. Josefiak and Schneider G.M. Determination of reaction volumes of hydrogen-bonding equilibria by high-pressure near-infrared spectroscopy. 2. Self-association of phenol in CCI, up to 1 kbar. *J. Phys. Chem.*, 84:3004–7, 1980. (Cité à page 12.)

[106] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983. (Cité aux pages 72, 82, 94 et 103.)

[107] C. Kato and D.H. Bartlett. The molecular biology of barophilic bacteria. *Extremophiles*, 1:111–116, 1997. (Cité aux pages xvi, 16 et 19.)

[108] C. Kato, A. Inoue, and K. Horikoshi. Isolating and characterizing deep-sea marine organisms. *Trends Biotechnol.*, 14:6–12, 1996. (Cité à page 19.)

[109] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem*, 14:1–67, 1959. (Cité à page 12.)

[110] W. Kauzmann. Thermodynamics of unfolding. *Nature*, 325:723–724, 1987. (Cité à page 12.)

[111] SE Kearsley. On the orthogonal transformation used for structural comparisons. *Acta Cryst. A*, 45:208–210, 1989. (Cité aux pages 73 et 74.)

[112] G Kneller and V Calandrini. Estimating the influence of finite instrumental resolution on elastic neutron scattering intensities .... *The Journal of Chemical Physics*, Jan 2007. (Cité à page 138.)

[113] Gerald R Kneller. Quasielastic neutron scattering and relaxation processes in proteins: analytical and simulation-based models. *Physical chemistry chemical physics : PCCP*, 7 (13):2641–55, Jul 2005. doi: 10.1039/b502040a. (Cité à page 135.)

[114] Gerald R Kneller and Paolo Calligari. Efficient characterization of protein secondary structure in terms of screw motions. *Acta Crystallographica Section D*, 62(3):302–311, 2006. (Cité aux pages xxx, 95, 96, 97, 100, 101, 102, 106 et 109.)

[115] G.R. Kneller. Superposition of molecular structures using quaternions. *Mol. Sim.*, 7: 113–119, 1991. (Cité aux pages xxx, 73, 74, 88, 96, 97 et 139.)

[116] G.R. Kneller. Comment on "Using quaternions to calculate RMSD" [J. Comp. Chem. 25, 1849 (2004)]. *J. Comp. Chem.*, 26(15):1660–1662, 2005. (Cité aux pages xxx, 75, 96 et 97.)

[117] G.R. Kneller. Quasielastic neutron scattering and relaxation processes in proteins: Analytical and simulation-based models. *Physical Chemistry Chemical Physics*, 7:2641 – 2655, 2005. (Cité aux pages 136 et 137.)

[118] GR Kneller and K Hinsen. Computing memory functions from molecular dynamics simulations. *J. Chem. Phys.*, 115(24):11097–11105, 2001. (Cité à page 61.)

[119] GR Kneller and K Hinsen. Fractional brownian dynamics in proteins. *J .Chem. Phys.*, 121(20):10278–10283, 2004. (Cité aux pages xxxvii et 135.)

[120] GR Kneller, V Keiner, M Kneller, and M Schiller. Nmoldyn, a program package for the calculation and analysis of neutron scattering spectra from md simulations. *Comp. Phys. Comm.*, 91:191–214, 1995. (Cité à page 88.)

[121] CE Kundrot and FM Richards. Crystal structure of hen egg-white lysozyme at a hydrostatic pressure of 1000 atmospheres. *J. Mol. Biol.*, 193:157–170, 1987. (Cité aux pages 14, 15, 83, 84 et 85.)

[122] J Kyte and RF Doolittle. A simple method for displaying the hydropathic character of a protien. *J. Mol. Biol.*, 157:105, 1982. (Cité aux pages xxiii et 52.)

[123] G Labesse, N Colloc'h, J Pothier, and J P Mornon. P-sea: a new efficient assignment of secondary structure from c alpha trace of proteins. *Comput. Appl. Biosci.*, 13:291–295, 1997. (Cité à page 103.)

[124] J.V. Landau and D.H. Pope. Recent advances in the area of barotolerant protein synthesis in bacteria and implications concerning barotolerant and barophilic growth. *Adv. Aquat. Micro.*, 2:49–76, 1980. (Cité à page 22.)

[125] J.V. Landau, W.P. Smith, and D.H. Pope. Role of the 30s ribosomal subunit, initiation factors, and specific ion concentration in barotolerant protein synthesis in pseudomonas bathycetes. *J. Bacteriol.*, 130:154–59, 1977. (Cité à page 22.)

[126] H. Le Chatelier. Sur un énoncé général des lois des équilibres chimiques. *Comptes Rendus, Accadémie des Sciences*, 99:786–89, 1884. (Cité à page 9.)

[127] V. Le Tilly, O. Sire, B. Alpert, and P.T.T. Wong. An infrared study of 2H-bond variation in myoglobin revealed by high pressure. *Eur. J. Biochem.*, 205:1061–1065, 1992. (Cité à page 12.)

[128] M Levitt and J Greer. Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.*, 114:181–239, 1977. (Cité à page 103.)

[129] SW Lovesey. *Theory of Neutron Scattering from Condensed Matter*, volume I. 1984. (Cité aux pages xxi et 38.)

[130] P.S. Low, J.L. Bada, and G.N. Somero. Temperature adaptation of enzymes: Roles of the free energy, the enthalpy, and the entropy of activation. *Proc Natl Acad Sci USA*, 70: 430–43, 1973. (Cité à page 21.)

[131] E. Madelung. Das elektrische Feld in Systemen von regelmassig angeordneten Punktladungen. *Phys. Z.*, 22:861–8, 1918. (Cité à page 33.)

[132] P. Masson and C. Clery. Effect of pressure on structure and activity of cholinesterase. In A.S. Balasubramanian, B.P. Doctor, P. Taylor, and D.M. Quinn, editors, *Enzymes of the Cholinesterase Family*. Plenum, NewYork, 1995. (Cité à page 15.)

[133] P. Masson, C. Clery, P. Guerra, A. Redslob, C. Albaret, and P.L. Fortier. Enzymes - hydration change during the aging of phosphorylated human butyrylcholinesterase: Importance of residues aspartate-70 and glutamate-197 in the water network as probed by hydrostatic and osmotic pressures. *Biochemical Journal*, 343:361–370, 1999. (Cité à page 14.)

[134] B.W. Matthews. Structural and genetic analysis of protein stability. *Ann. Rev. Biochem.*, 62:139–160, 1993. (Cité à page 18.)

[135] B.W. Matthews. Studies on protein stability with t 4 lysozyme. *Adv. Protein Chem*, 46: 249–278, 1995. (Cité à page 18.)

[136] B.W. Matthews. Structural and genetic analysis of the folding and function of t4 lysozyme. *FASEB J.*, 10:35–41, 1996. (Cité à page 18.)

[137] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Phys. Rep.*, 339:1–77, 2000. (Cité à page 135.)

[138] P.C. Michels and D.S. Clark. Pressure enhanced activity and stability of a hyperthermophilic protease from a deep-sea methanogen. *Appl. Environ. Microbiol.*, 63:3985–3991, 1997. (Cité aux pages xix, 13 et 23.)

[139] P.C. Michels, D. Hei, and D.S. Clark. Pressure effects on enzyme activity and stability at high temperature. *Adv. Protein Chem*, 48:341–376, 1996. (Cité aux pages 11 et 13.)

[140] S. Miller and J.L. Bada. Submarine hot springs and the origin of life. *Nature*, 342:255, 1988. (Cité à page 17.)

[141] S. Miller and A. Lazcano. The origin of life,Äîdid it occur at high temperatures? *J. Mol. Evol.*, 41:689, 1995. (Cité à page 17.)

[142] Zoran Minic, Valérie Serre, and Guy Hervé. Adaptation of organisms to extreme conditions of deep-sea hydrothermal vents. *C R Biol*, 329(7):527–40, Jul 2006. doi: 10.1016/j.crvi.2006.02.001. (Cité à page 21.)

[143] E. Mombelli, M. Afshar, P. Fusi, M. Mariani, P. Tortora, J.P. Connelly, and R. Lange. The role of phenylalanine 31 in maintaining the conformational stability of ribonuclease p2 from solfolobus solfataricus under extreme conditions of temperature and pressure. *Biochemistry*, 36(29):8733–42, 1997. (Cité à page 8.)

[144] V.V Mozhaev, K Heremans, J Frank, P Masson, and Claude Balny. High pressure effects on protein structure and function. *Proteins: Structure, Function, and Genetics*, 24:81–91, Nov 1996. (Cité aux pages xvi, 11 et 14.)

[145] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247: 536–540, 1995. (Cité aux pages 80 et 99.)

[146] A. Onufriev, D. Bashford, and D.A. Case. Modification of the generalized born model suitable for macromolecules. *J. Phys. Chem. B*, 104:3712–3720, 2000. (Cité à page 62.)

[147] C.N. Pace, B.A. Shirley, M. McNutt, and K. Gajiwala. Forces contribuiting to the conformational stability of proteins. *FASEB J.*, 10:75–83, 1996. (Cité aux pages xvii, 17 et 18.)

[148] E Paci and M Marchi. Intrinsic compressibility and volume compression in solvated proteins by molecular dynamics simulation at high pressure. *Proc Natl Acad Sci US A*, 93(21):11609–11614, 1996. (Cité à page 14.)

[149] A.A. Paladini, J.L. Silva, and G. Weber. Slab gel electrophoresis of oligomeric proteins under high hydrostatic pressure. *Anal. Biochem.*, 161:358–364, 1987. (Cité à page 15.)

[150] X. Peng, J. Jonas, and J.L. Silva. High-pressure nmr study of the dissociation of arc repressor. *Biochemistry*, 33:8323–8329, 1994. (Cité à page 15.)

[151] G.A. Petsko. Structural basisof thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat". *Methods Enzymol*, 334:469–478, 2001. (Cité aux pages xvi, xvii, 16 et 18.)

[152] M. Potchka and T.M. Schuster. Determination of reaction volumes and polymer distribution characteristic of tobacco mosaic virus coat protein. *Anal. Biochem.*, 161: 70–79, 1987. (Cité à page 15.)

[153] P.L. Privalov. In T.E. Creighton, editor, *Protein folding*, pages 83–126. W.H. Freeman and Company, New York, 1993. (Cité à page 13.)

[154] O.B. Ptitsyn. The molten globule state. In T.E. Creighton, editor, *Protein Folding*, pages 243–300. Freeman and Company, 1992. (Cité à page 15.)

[155] J.R. Quine. Helix parameters and protein structure using quaternions. *J. Mol. Struct. (Theochem)*, 460:53–66, 1999. (Cité à page 72.)

[156] GPS Raghava. APSSP2 : A combination method for protein secondary structure prediction based on neural network and example based learning. *CASP5*, A:132, 2002. (Cité à page 52.)

[157] R. Ragone and G. Colonna. Do globular proteins require some structural peculiarity to best function at high temperatures? *J. Am. Chem. Soc.*, 117:16–20, 1995. (Cité aux pages xvii et 18.)

[158] A. Razvi and J.M. Scholtz. Lessons in stability from thermophilic proteins. *Proteins Science*, 15(7):1569, 2006. (Cité aux pages xvi, xvii, 16 et 18.)

[159] M. Refaee, T. Tezuka, K. Akasaka, and M.P. Williamson. Pressure-dependent changes in the solution structure of hen egg-white lysozyme. *J. Mol. Biol.*, 327:857–865, 2003. (Cité aux pages 83 et 85.)

[160] P. Regnard. *La vie dans les eaux*. Paris, 1891. (Cité à page 8.)

[161] L.E. Reichl. *A modern course in Statistical Mechanics*. Arnold Publishers, 9180. (Cité à page 10.)

[162] R.E. Rhoads. Signal transduction pathways that regulate eukaryotic protein synthesis. *J. Biol. Chem.*, 274:30337–30340, 1999. (Cité à page 22.)

[163] FM Richards and CE Kundrot. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. *Proteins*, 3:71–84, 1988. protein secondary structure recognition. (Cité aux pages 72, 94 et 103.)

[164] H. Risken. *The Fokker-Planck Equation*. Springer Series in Synergetics. Springer, Berlin, Heidelberg, New York, 2nd reprinted edition, 1996. (Cité aux pages 134 et 136.)

[165] K.K. Rodgers and S.G. Sligar. Mapping electrostatic interactions in macromolecular associations. *J. Mol. Biol.*, 221:1453–1460, 1991. (Cité à page 11.)

[166] T. Rog, K. Murzyn, K. Hinsen, and G.R. Kneller. nMoldyn : A Program Package for a Neutron Scattering Oriented Analysis of Molecular Dynamics Simulations. *J. Comp. Chem.*, 24(5):657–667, 2003. (Cité aux pages 88 et 139.)

[167] L.J. Rothschild and R. Mancinelli. Life in extreme environments. *Nature*, 409:1092–1102, 2001. (Cité à page 17.)

[168] C. Royer. Revisiting volume changes in pressure-induced protein unfolding. *BBA Biochem. et Biophys. Acta*, 1595:201–9, 2002. (Cité à page 13.)

[169] K. Ruan, C. Xu, Y. Yu, J. Li, R. Lange, N. Bec, and C. Balny. Pressure-exploration of the 33-kDa protein from the spinach photosystem II particule. *Eur. J. Biochem.*, 268:2742–50, 2001. (Cité à page 8.)

[170] D.W. Russell and L.L. Spremulli. Purification and characterization of a ribosome dissociation factor (eukaryotic initiation factor 6) from wheat germ. *J. Biol. Chem.*, 254 (18):8796–8800, 1979. (Cité aux pages xx, 23 et 25.)

[171] S. Samarasinghe, D.M. Campbell, A. Jonas, and J. Jonas. High-resolution nmr study of the pressure-induced unfolding of lysozyme. *Biochemistry*, 31:7773–7778, 1992. (Cité à page 15.)

[172] F. Sanvito. Expression of a highly conserved protein, p27BBP, during the progression of human colorectal cancer. *Cancer Res.*, 60:510–516, 2000. (Cité à page 159.)

[173] F. Sanvito, S. Piatti, A. Villa, M. Bossi, G. Lucchini, P.C. Marchisio, and S. Biffo. The β4 integrin interactor p27(bbp/eif6) is an essential nuclear matrix protein involved in 60s ribosomal subunit assembly. *J. Cell Biol.*, 144:823–837, 1999. (Cité aux pages xx et 24.)

[174] M.J. Schlesinger. Heat-shock proteins. *J. Biol. Chem.*, 265(21):12111, 1990. (Cité à page 19.)

[175] V. Sears. Neutron scattering lengths and cross sections. *Neutron News*, 3(3):26–37, 1992. (Cité à page 41.)

[176] J.M. Selig. *Geometrical methods in robotics*. Springer, 1996. (Cité à page 72.)

[177] B. Senger, D. L.J. Lafontaine, J-S Graindorge, O Gadal, A Camasses, A Sanni, J-M Garnier, M Breitenbach, E Hurt, and F Fasiolo. The nucle(ol)ar tif6p and efl1p are required for a late cytoplasmic step of ribosome synthesis. *Molecular Cell*, 8:1–20, 2001. (Cité aux pages xx et 24.)

[178] K. Si and U Maitra. The *saccharomyces cerevisiae* homologue of mammalian translation initiation factor 6 does not functions as a translation initiation factor. *Molecular and Cellular Biology*, 19(2):1416–1426, 1999. (Cité aux pages xx, 24 et 25.)

[179] J.F. Siebenaller and G.N. Somero. Biochemical adaptation to the deep sea. *CRC Crit. Rev. Aquat. Sci.*, 1:1–25, 1989. (Cité à page 20.)

[180] J.L. Silva and G. Weber. Pressure stability of proteins. *Ann. Rev. Phys. Chem.*, 44:89–113, 1993. (Cité aux pages 14 et 15.)

[181] H Sklenar, C Etchebest, and R Lavery. Describing protein structure: A general algorithm yielding complete helicoïdal parameters and a unique overall axis. *Proteins: Str. Func. Genet.*, 6:46–60, 1989. (Cité aux pages 72, 94 et 103.)

[182] G.N. Somero. Adaptations to high hydrostatic pressure. *Annu. Rev. Physiol.*, 54:557–577, 1992. (Cité aux pages xvi, 16 et 20.)

[183] G.N. Somero, J.F. Siebenaller, and P.W. Hochachka. In G.T. Rowe, editor, *The Sea*. Wiley, 1983. (Cité à page 21.)

[184] J. M. Sperrazza, D. W. Russell, and L. L. Spremulli. Reversible dissociation of wheat germ ribosomal subunits: cation-dependent equilibriums and thermodynamic parameters. *Biochemistry*, 19(6):1053–1058, 1980. (Cité à page 22.)

[185] A.S. Spirin. On the equilibrium of the association-dissociation reaction of ribosomal subparticles and on the existance of the so-called "6os intermediate" ("swollen 7os") during centrifugation of the equilibrium mixture. *FEBS Letters*, 14:349–353, 1971. (Cité à page 22.)

[186] A.S. Spirin. *Ribosomes*. Cellular Organelles Series. Kluwer Academic/ Plenum Publisher, 1998. (Cité à page 24.)

[187] L. Stryer. *Biochemistry*. W.H. Freeman and Company, 1988. (Cité aux pages 8, 13, 72 et 94.)

[188] M.M.C. Sun, N. Tolliday, C. Vetriani, F. T. Robb, and D. S. Clark. Pressure-induced thermostabilization of glutamate dehydrogenase from the hyper thermophile *pyrococcus furiosus*. *Protein Science*, 8:1056–1063, 1999. (Cité à page 21.)

[189] C. Tanford. Protein denaturation. *Adv. Protein Chem*, 23:121–282, 1968. (Cité à page 8.)

[190] WR Taylor. Defining linear segments in protein structure. *J. Mol. Biol.*, 310:1135–1150, 2001. protein secondary structure recognition. (Cité aux pages 72 et 94.)

[191] M Tehei and G Zaccai. Adaptation to high temperatures through macromolecular dynamics by neutron scattering. *FEBS Journal*, Jan 2007. (Cité aux pages xvi, 16 et 152.)

[192] Moeava Tehei, Bruno Franzetti, Dominique Madern, Margaret Ginzburg, Ben Z Ginzburg, Marie-Thérèse Giudici-Orticoni, Mireille Bruschi, and Giuseppe Zaccai. Adaptation to extreme environments: macromolecular dynamics in bacteria compared in vivo by neutron scattering. *EMBO Rep*, 5(1):66–70, Jan 2004. doi: 10.1038/sj.embor.7400049. (Cité aux pages 152 et 158.)

[193] R.M. Thaorak. Brownian motion of a torus. *Colloids and Surfaces A : Physicochem. Eng. Aspects*, 317:650–657, 2008. (Cité aux pages 165 et 166.)

[194] Thomas. The graduation of secondary structure elements. *J. Mol. Graphics*, 12:146–152, 1994. (Cité aux pages 72 et 94.)

[195] D.J. Tobias, G.J. Martyna, and M.L. Klein. Molecular Dynamics simulations of a protein in the canonical ensemble. *J. Phys. Chem.*, 97:12959–12966, 1993. (Cité aux pages xxvi et 36.)

[196] J Torrent, MT Alvarez-Martinez, F Heitz, JP Liautard, C Balny, and Reinhard Lange. Insights into alternative prion protein topologies induced under high hydrostatic pressure. *J. Phys.: Condens. Matter*, 16:S1059–S1065, 2004. (Cité à page 16.)

[197] C.L. Tsou. Folding of the nascent peptide chain into a biologically active protein. *Biochemistry*, 27:1809–1812, 1988. (Cité aux pages xvii et 18.)

[198] V. Tsui and D.A. Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers (Nucl. Acid. Sci.)*, 56:275–291, 2001. (Cité à page 61.)

[199] D M Valenzuela, A Chaudhuri, and U Maitra. Eukaryotic ribosomal subunit anti-association activity of calf liver is contained in a single polypeptide chain protein of mr = 25,500 (eukaryotic initiation factor 6). *J Biol Chem*, 257(13):7712–9, Jul 1982. (Cité aux pages xx, 23, 25 et 53.)

[200] R. Van Eldik, T. Asano, and W.J. Le Noble. Activation and reaction volume in solution 2. *Chem. Rev.*, 89:549–688, 1989. (Cité aux pages 11 et 12.)

[201] L. van Hove. Correlations in space and time and Born approximation in systems of interacting particles. *Phys. Rev.*, 93(1):249–262, 1954. (Cité à page 39.)

[202] M.C. Vaney, S. Maignan, M. Ries-Kautt, and A. Ducruix. High-resolution structure (1.33 a) of a hew lysozyme tetragonal crystal grown in the apcf apparatus. data and structural comparison with a crystal grown under microgravity from spacehab-01 mission. *Acta Crystallogr. D*, 52:505–517, 1996. (Cité aux pages 83 et 85.)

[203] Guoli Wang and Roland L Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–91, Aug 2003. (Cité à page 99.)

[204] Guoli Wang and Roland L Dunbrack. Pisces: recent improvements to a pdb sequence culling server. *Nucleic Acids Res*, 33(Web Server issue):W94–8, Jul 2005. doi: 10.1093/nar/gki402. (Cité à page 99.)

[205] M.C. Wang and G.E. Uhlenbeck. On the theory of Brownian motion II. *Phys. Rev.*, 93(1):249–262, 1945. (Cité à page 136.)

[206] G. Weber and H.G. Drickamer. The effect of high pressure upon proteins and other biomolecules. *Quart. Rev. of Biophys.*, 16(1):89–112, 1983. (Cité à page 14.)

[207] G. Weber, F. Tanaka, B.Y. Okamoto, and H.G. Drickamer. The effect of pressure on the molecular complex of isoalloxazine and adenine. *Proc Natl Acad Sci USA*, 71:1264–66, 1974. (Cité à page 12.)

[208] C. Wilson and S. Doniach. A computer model to dynamically simulate protein folding: studies with crambin. *Proteins Structure Function and Genetics*, 6:193, 1989. (Cité à page 65.)

[209] A Wlodawer, J Walter, R Huber, and L Sjölin. Structure of bovine pancreatic trypsin inhibitor. results of joint neutron and x-ray refinement of crystal form ii. *J. Mol. Biol.*, 180:301–329, 1984. (Cité à page 109.)

[210] A Wlodawer, J Nachman, GL Gilliland, W Gallagher, and C Woodward. Structure of form iii crystals of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.*, 198:469–480, 1987. (Cité à page 109.)

[211] P.T.T. Wong and K. Heremans. Pressure effects on protein secondary structure and hydrogen deuterium exchange in chymotrypsinogen: a fourier transform infrared spectroscopic study. *BBA Biochimica Biophysica Acta*, 956:1–9, 1988. (Cité à page 16.)

[212] L.C. Wood, M N Ashby, C Grunfeld, and K R Feingold. Cloning of murine translation initiation factor 6 and functional analysis of the homologous sequence YPR016c in *saccharomyces cerevisiae*. *J. Biol. Chem.*, 274:11653–11659, 1999. (Cité aux pages xx et 24.)

[213] B. Wroblowski, J.F. Diaz, K. Heremans, and Y. Engelborghs. Molecular mechanisms of pressure induced conformational changes in bpti. *Proteins Structure Function and Genetics*, 25:446–455, 1996. (Cité à page 13.)

[214] H. Yang, G. Luo, P. Karnchanaphanurach, T.M. Louie, I. Rech, S. Cova, L. Xun, and X.S. Xie. Protein conformational dynamics probed by single-molecule electron transfer. *Science*, 302(5643):262–266, 2003. (Cité à page 135.)

[215] S. Yokoyama, T. Tada, H. Zhang, and L. Britt. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci USA*, 105(36):13480–13485, 2008. (Cité à page 159.)

[216] G Zaccai. How soft is a protein? a protein dynamics force constant measured by neutron scattering. *Science*, 288:1604–1607, 2000. (Cité aux pages 130 et 135.)

[217] P. Závodszky, J. Kardos, A. Svingor, and G. A. Petsko. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci USA*, 95:7406–7411, 1998. (Cité aux pages xvii et 18.)

[218] J. Zhang, X. Peng, A. Jonas, and J. Jonas. Nmr study of the cold, heat, and pressure unfolding of ribonuclease a. *Biochemistry*, 34:8631–8641, 1995. (Cité à page 13.)

[219] J.M. Zhou, L. Zhu, C. Balny, and S. Perret. Pressure denaturation of the yeast prion protein ure2. *Biochem. Biophys. Res. Comm.*, 287:147–152, 2001. (Cité à page 8.)

[220] C.E. ZoBell and F.H. Johnson. The influence of hydrostatic pressure on the growth and viability of terrestrial and marine bacteria. *J. Bacteriol.*, 57:179, 1949. (Cité aux pages xviii et 19.)