Validation of Crystallographic B Factors and Analysis of Ribosomal Crystal Structures

Jacopo Negroni

2012

Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

MSc Bioinformatics Jacopo Negroni

born in: Bergamo, Italy

Oral-examination: .....

Validation of Crystallographic B Factors
and
Analysis of Ribosomal Crystal Structures

Referees: **Dr. Dmitri Svergun**

Prof. Dr. Irmgard Sinning

To my parents

Contents

List of Figures

List of Tables

Nomenclature

$\langle B \rangle$	Average B Factor
$\langle u^2 \rangle$	Mean Square Displacement
$\langle x \rangle$	Sample Mean
μ	Population Mean
σ^2	Population Variance
s_x^2	Sample Variance
B_{wil}	Wilson B factor
na	*not available*
ADP	Atomic Displacement Parameter (B Factor)
IGD	Inverse-Gamma Distribution
IGD*	Shifted Inverse-Gamma distribution

Summary

In X-ray crystallography, validation tools assess the quality and the reliability of the structural models that crystallographers build and refine. These tools check both the consistency of physical, chemical and statistical properties of the model with the prior knowledge available in structural databases, and the agreement of the model with the diffraction data.

B factors give important information about the spatial disorder of each atom around its rest position in a crystal, allowing one to infer the precision of atomic coordinates and dynamical properties of the macromolecule.

The first part of the thesis work is focused on the development of a new validation tool for the distribution of isotropic B factors in crystallographic models. By means of a Bayesian approach the shifted Inverse-Gamma distribution (IGD*) is proposed as a reference distribution and a validation protocol is designed and developed to test this hypothesis. Starting from an empirical B factor distribution, the protocol returns the parameters estimates of the IGD* that best fits the B factor distribution and a p-value that is used to label the distribution as *acceptable* or *suspicious*. The protocol is then tested on a large data set of high-resolution protein structures from the PDB. From the distribution of the IGD* parameters it is possible to identify different groups of outliers, each characterized by peculiar features. Moreover, from the analysis of the distribution of p-values, the majority of the structures analysed have an *acceptable* B factor distribution and the agreement to the IGD* follows a hierarchical organization (whole asymmetric unit content, single chains and single domains). B factor distributions that do not satisfy the IGD* assumption usually correspond to models with problems with the deposited coordinates or diffraction data. In light of these results the developed protocol is proposed as an effective tool for the validation of B factor distributions in macromolecular crystallography. Furthermore, provided that the diffraction data are deposited in the PDB, a standard re-refinement protocol is confirmed to be a valid approach to rescue a B factor distribution from *suspicious* to *acceptable*, and to improve the quality of the results of the ensemble analysis performed with the ESCET framework if the starting data set contains models with *suspicious* B factor distributions.

The validation protocol for B factor distributions finds a direct application in the second part of the thesis work, which is focused on the ensemble analysis with the ESCET

framework of a selected data set of twenty-nine 30S ribosomal subunits from *Thermus thermophilus*. Thirteen refinement protocols are tested to improve, normalise and de-bias the selected structures, and to rescue models with *suspicious* B factor distributions. A comparative ensemble analysis is performed between the ribosomal models as deposited into the PDB and those obtained from the best refinement protocol in terms of refinement statistics and distribution of B factors. The cluster analysis is confirmed to be an effective method to automatically rationalise the structural information content of the data set. The observation that after re-refinement some structures moved to a different cluster confirms the existence of structural bias in the originally deposited structures and leads to the discovery of electron density that was not modelled in the deposited structure. Improvements of refinement statistics after re-refinement result in lower coordinate uncertainty estimates with positive effects on the results of the rigid body analysis. The main rigid bodies found on the 16S rRNA correspond to the domains known in the literature to move during the decoding process.

Final remarks are given about the possible application of the presented validation tool for B factor distributions and about the importance of the availability of experimental data.

Zusammenfassung

In der Röntgenkristallographie werden Validierungswerkzeuge benutzt, um die Qualität und Zuverlässigkeit von Strukturmodellen, die Kristallographen erstellen und verfeinern, zu bemessen. Diese Werkzeuge überprüfen sowohl die Konsistenz physikalischer, chemischer und statistischer Parameter des Modells mit denen bereits bekannter Strukturen, als auch die Übereinstimmung mit den Beugungsdaten.

Die B-Faktoren enthalten wichtige Informationen über die Genauigkeit der räumlichen Positionierung rund um die im Kristall bestimmte Position. Dies ermöglicht einerseits die Analyse der Genauigkeit der Atomkoordinaten und andererseits eine Einschätzung der dynamischen Eigenschaften des Makromoleküls.

Der erste Teil dieser Dissertation beschätigt sich mit der Entwicklung eines neuen Werkzeugs zur Validierung der Verteilung isotroper B-Faktoren in kristallographischen Modellen. Ausgehend von einem bayesschen Ansatz wird die “shifted Inverse-Gamma distribution” (IGD*) als Referenzverteilung vorgeschlagen und ein Validierungsprotokoll zum Testen dieser Hypothese entwickelt. Ausgehend von der empirisch ermittelten Verteilung der B-Faktoren, schätzt das Protokoll die Paramter der IGD*, welche die B-Faktor Verteilung am besten beschreiben, und berechnet einen p-Wert, um die Verteilung als aktzeptabel oder verdächtig zu klassifizieren. Das Protokoll wird gegen einen grossen Datensatz hochauflösender Strukturen aus der PDB getestet. Aus der Verteilung der IGD*-Parameter lassen sich verschiedene Gruppen von Ausreissern erkennen, die sich durch charakteristische Eigenschaften von anderen abgrenzen. Darüber hinaus zeigt die Analyse der p-Werte, dass die meisten der untersuchten Strukturen eine akzeptable Verteilung der B-Faktoren aufzeigen. Die Übereinstimmung mit der IGD* folgt hierbei einer hierarchischen. Organisation (gesamte asymmetrische Einheit, einzelne Molekülketten und einzelne Domänen). Modelle, deren B-Faktorverteilung nicht der IGD*-Annahme entsprichen, zeigen meist auch Unregelmässigkeiten bei den in der Datenbank hinterlegten Koordinatendaten oder aber auch bei den Streuungsdaten. Basierend auf den erhaltenen Ergebnissen wird das entwickelte Protokoll als effektives Werkzeug für die Validierung der B-Faktor Verteilung in der makromolekularen Kristallographie vorgeschlagen. Des Weiteren wird bestätigt, dass ein Standard Protokoll zum Re-Refinement (erneute Verfeinerung) ein geeignetes Instrument ist, die B-Faktor Verteilung von verdächtig nach akzeptabel zu korrigieren. Dies führt ausserdem zu

einer Qualitätsverbesserung der Ergbnisse für die Ensemble Analyse mit dem ESCET Paket für Fälle, bei denen der Startdatensatz Modelle mit verdächtigen B-Faktor Verteilungen enthält.

Im zweiten Teil dieser Arbeit wird das Protokoll zur Validierung der B-Faktor Verteilung angewandt, um eine Ensemble Analyse mit dem ESCET Paket für 29 Datensätze der 30S Untereinheit des Ribosoms aus *Thermus thermophilus* durchzuführen. Dreizehn Protokolle werden getestet, um die jeweiligen Strukturen zu verbessern, zu normalisieren, Beeinflussungen durch benutzte Modelle zu korrigieren und Modelle mit verdächtiger B-Faktor Verteilung anzupassen. Es wird eine vergleichende Ensemble Analyse zwischen den in der PDB hinterlegeten Strukturmodellen und den optimal verfeinerten (Refinement Statistik / B-Faktor Verteilung) Modellen vorgestellt. Es wird bestätigt, dass die Cluster Analyse die geeignete Methode ist, um automatisch die strukturspezifischen Informationen des Datensatzes systematisch zu erfassen. Durch das Re-Refinement werden einzelne Strukturen anderen Clustern zugeordnet, was zum einen die Existenz von strukturellen Fehlern in den hinterlegten Daten zeigt, und zum anderen zur Identifikation von Elektronendichte führt, die nicht zum Erstellen des veröffentlichten Strukturmodells benutzt wurde. Die Verbesserung der Verfeinerungsstatistiken führen zu höherer Genauigkeit der Koordinaten, was wiederum einen positiven Effekt auf die Ergebnisse der Rigid-Body Analyse hat. Die wichtigsten Rigid-Bodies der 16S rRNA, die gefunden wurden, korrespondieren mit den Domänen, für die in der Literatur eine koformationelle Bewegung während des Dekodierungs Prozesses beschrieben ist.

Die Arbeit schliesst mit einem Ausblick auf mögliche Anwendungsfelder für das vorgestellte Protokoll zur Valdierung der B-Faktor Verteilung ab. Hierbei wird insbesondere die Bedeutung der Verfügbarkeit experimenteller Daten hervorgehoben.

Chapter 1

Introduction

1.1 X-ray Crystallography in Structural Biology

The determination of the three-dimensional structure of biological macromolecules plays an important role in molecular biology since their functions are strongly related to their structures [63,96,116]. Thus, structural information can be used to infer the mechanism of function of a given macromolecule (whether it is purely structural or enzymatic), to validate experimental data obtained from different experimental techniques, and to formulate new hypotheses for further experiments (e.g. with applications in drug discovery [6,88]).

X-ray crystallography is a well established experimental technique [43,106] that allows the determination of the three-dimensional structure of macromolecules by taking advantage of the interaction between X-rays and the scattering matter in a crystal (a description of the crystallographic experiment is given in section 3.1.1). At the moment it is still the most widely used experimental method for the determination of atomic models of biological macromolecules as shown by the total number of models deposited in the Protein Data Bank (PDB) [14] (as of October 2011, 67414 out of 77101 deposited models where determined using X-ray crystallography). X-ray crystallography has a wide applicability ranging from small organic molecules [5] to large macromolecular complexes of several mega daltons like the ribosome [73,102,114] (from 2.5 to 3.3 MDa) or the fatty acid synthase [64] ($\sim$ 2.6 MDa). The analysis of the results obtained from structural genomics projects allowed recently to quantify the difficulties of each step of the structure determination process using X-ray crystallography [124]. The main drawback of this technique still resides in the difficulty to obtain diffracting crystals for highly flexible macromolecules or large complexes due to the high structural order required in the crystal for generating high-resolution diffraction data. For these reasons X-ray crystallography can be complemented by other techniques used in structural biology like Nuclear Magnetic Resonance (NMR) [52], cryo-Electron Microscopy (cryo-EM) [53] and Small Angle X-ray Scattering

(SAXS) [81] that have the advantage that no crystals are required. Amongst these techniques NMR is the only one that allows the determination of macromolecular structures at atomic level, but it has limitations in the maximum size of the macromolecule under analysis and until now the number of structures determined using NMR and deposited in the PDB are much lower than those obtained using X-ray crystallography (as of October 2011, 9149 out of 77101 deposited models where determined using NMR).

Atomic resolution models obtained using X-ray crystallography or NMR are important for the interpretation of the results obtained from low resolution experimental techniques like cryo-EM and SAXS for the production of the so called "hybrid models" [118]. Furthermore, also predictive methods like homology modeling, threading and ab-initio modeling benefit from the availability of atomic models from X-ray crystallography or NMR [7]. The increase in number of the structures deposited in the PDB allows the extraction of more accurate features that can be consequently used to improve the quality of the model predictions.

It follows that it is of extreme importance to have tools that check for the reliability and quality of the deposited models.

1.2 Validation of Crystallographic Structures

Given the scientific importance of crystallographic structures, it is necessary to have validation tools for new crystallographic models prior to deposition and for those already deposited in the PDB. The reliability of the deposited structures is in fact indispensable for the extraction of physical, chemical and biological information from the structures [42]. Structure validation has also become a significant and routine step in automatic model building [1,35].

The atomic model deposited in the PDB is the result of the interpretation by the crystallographer of the average electron density in a crystal obtained during the crystallographic experiment, and as such it is the result of a subjective interpretation of the experimental data. Validation tools or quality indicators are then necessary to assess the validity of the models and to limit the subjectivity introduced during model building and refinement in favor of a more objective interpretation of the experimental data [22,42,70,107]. Such goal is achieved by judging how well a given model is in agreement with the prior knowledge available from small and large molecules structural databases [39,77] and by evaluating the agreement between the model and the collected diffraction experimental data [72,128].

The quality indicators used in macromolecular crystallography can be divided in two main classes: global and local statistics [70]. The first class includes all those statistics that provide information about the overall quality of the model like: R_{work} and R_{free} statis-

tics, average, minimum, maximum and root mean square of local statistics, percentage of outliers for each local statistics [70]. This statistics are used to monitor the progress of model building and refinement. The second class includes all those statistics that provide information at amino-acid or atomic level like: real-space fit, Ramachandram plot, pep-flip value, rotamer side chain fit value, hydrogen-bonding analysis, unusual B factors, unusual occupancies, unusual bond lengths or angles, unusual torsion angles, atomic clashes. This statistics are especially useful during model building to identify local problems in the model [70].

Depending on their use in model building and refinement these quality indicators can be also divided in 'weak' and 'strong' validation criteria [70, 71]. The first category assess how well a model reproduces the information that was used during model building (e.g. deviations from ideal geometry, R_{work}). Since these statistics have no predictive power they are classified as 'weak'. The second category instead tests how well a model predicts information that was not used in the model building process (e.g. R_{free} , patterns of non-bonded interactions, conformational torsion-angle distributions). Since these statistics have predictive power they are classified as 'strong' and make the model more reliable [71].

The availability of the experimental diffraction data is of extreme importance in the validation process, since they are the only evidence that can be used to evaluate the quality of a given model and to explain any outlier identified by local statistics (i.e. an outlier for a local statistic is not necessary an error if it is supported by experimental data). Moreover, it has been shown that an automatic re-refinement of the models deposited in the PDB, for which experimental data are available, can deliver better models thanks to the improvement of the crystallographic software [65, 66, 104]. The lack of deposited structure factors thus not only represents a huge loss in the validation process, making the deposited model less reliable, but it also limits the possibility to derive better models in the future. Fortunately this problem will be not present in the future since the deposition of diffraction data is now mandatory for X-ray crystallography [104].

Since the number of structures deposited in the PDB increased of nearly one order-of-magnitude from the time the current validation tools were introduced, new and more advanced validation criteria have been recently proposed [104].

1.3 B Factors and their use in Structural Analysis

The result of a crystallographic experiment is an average over time and space of the electron density content in a crystal [43, 106]. This means that any disorder in the crystal will affect the final average electron density: well ordered regions will result in in well defined and strong electron density, while disordered regions will result in noisy and weak electron den-

sity. However, because the obtained electron density is the result of an averaging process, any information about the nature of the disorder is lost, this is either due to thermal vibrations (dynamic disorder) or to alternative conformations (static disorder) of the atoms of the macromolecules inside the crystal. The goal of a crystallographer is then to interpret the electron density by building an atomic model in agreement with it (see also section 3.1.1). For each atom of the model, together with the three-dimensional coordinates, a fourth parameter called B factor is used to model the disorder of the electron density and thus to model the uncertainty of to atomic position (see also section 3.1.3). The higher the B factor of an atom, the lower the probability the atom is located at the coordinates given in the model, indicating a lower degree of order. In addition, depending on the data to parameter ratio, different models have been proposed for the B factors [2, 133]. From low- to high-resolution crystal structures the most common B factor models are: global isotropic or anisotropic overall B factor, TLS groups, grouped B factor, individual isotropic or anisotropic B factor. In the past, several studies tried to interpret the disorder of the electron density by generating ensemble of structures in agreement with the experimental data instead of only one model. However the interpretation of this ensembles is still debatable. It is in fact not clear if the generated ensemble is representative of all the different conformations of the macromolecule in the crystal or if, alternatively, each model in the ensemble is an agreement with the crystallographic data but not necessarily it represents a conformation present in the crystal [8, 40, 41, 55, 74, 123].

B factors give then important information about the flexibility of a macromolecule inside a crystal and can be used to infer structural and mechanical properties. Since they are related to the positional uncertainty of the atomic coordinates, several statistical methods were proposed in the past for the use of B factors in the structural comparison of models of the same macromolecule obtained from different crystals [19, 33, 59, 98]. However, one of the main drawbacks of the crystallographic experiment is that no direct measure about the precision of the atomic coordinates is available. Even if B factors are related by the theory of thermal vibrations to the atomic displacements in a crystal, they are nevertheless the result of the fitting of a model to the experimental diffraction data and they are not measured directly, becoming dependent on the model used to describe them and on the parameterisations used by the refinement program. Thus, it follows that B factors from different structural models should not be directly compared and used as a direct measure of the coordinate uncertanties of the model. In small molecule crystallography a routine approach to estimate the precision of coordinates in a model consists at the inversion of the Hessian matrix at convergence of refinement [15,36]. However this approach is not feasible in macromolecule crystallography since the inversion of the Hessian matrix becomes computationally a gigantic task. To partially overcome this problem, the diffraction-component precision index (DPI) [36] was introduced to estimate the coordinate error for an atom whose B factor is equal to the average B factor of the model (see also section 3.1.4), al-

lowing the comparison on a common scale of the errors in different models. Assuming a linear relationship between the coordinate uncertainty of an atom and its B factor, the DPI was used in the statistical framework ESCET to estimate the standard uncertainties of the atomic coordinates in crystallographic model, allowing the analysis and comparison of ensemble of structures [84, 110–112] (see also section 3.1.5).

Despite the fact B factors give important information about the flexibility and the uncertainties of the coordinates of a macromolecule in a crystal, at the moment a validation tool for their distribution still does not exist. The only controls performed on B factors are focused on extreme values (too low or too high) or on their average value in a crystallographic model when compared to structures solved at a similar resolution [127]. However the analysis of their distribution in a structure could give important information about the quality of the model. Firstly it could be useful to identify artefacts introduced by the refinement procedures. Secondly the validation of the distribution of B factors would be very important in methods that use them to compare different models like in the ESCET framework. In fact, if the distribution of B factors contains strong artefacts it will affect the subsequent analysis. In the best scenario this will just introduce noise in the data, making the interpretation of the results more difficult. In the worse scenario it could also introduce artefacts in the output of the ensemble analysis leading to wrong conclusions from the interpretation of the results. In the past different models have been proposed for the distribution of B factors in macromolecules obtained via X-ray crystallography [38, 97, 130] but they were never used for the validation of the distribution itself.

1.4 The Ribosome Complex

The ribosome complex, which is responsible for the synthesis of proteins in all living organisms, is one of the most exciting structures of the last two decades obtained by X-ray crystallography, as illustrated by the 2009 Nobel prize in chemistry. The integration of the crystallographic models [75, 79, 109, 117] with the results from cryo-EM [51, 82, 129] and single-molecule fluorescence resonance energy transfer (smFRET) [99] allowed the rationalization of the dynamics of the protein synthesis process from a structural point of view [46].

Ribosomes from eukaryotes and archaea consist of a large (50S) and a small (30S) subunit, which together compose the 2.5 MDa 70S ribosome. The 50S subunit consists of 23S rRNA, 5S rRNA and about 30 proteins. The 30S subunit consists of 16S rRNA and about 20 proteins [109, 117]. Similarly, the eukaryotic ribosomes consist of a large (60S) and a small (40S) subunit, which together compose the 3.3 MDa 80S ribosome. The 60S consists of 25S-28S rRNA, 5.8S rRNA, 5S rRNA and about 46 proteins. The 40S consists

18S rRNA and about 32 proteins [12, 73, 102, 103].

The rRNA moieties are the largest components of the ribosome complex and are responsible for the the main functional steps of protein synthesis, while the ribosomal proteins have mostly a regulatory and a structural function by assisting the correct folding of the ribosomal subunits and at the same time by stabilizing with positively charged amino-acids the negative charges of phosphate groups of the RNA backbone [24, 60, 69, 121].

The current knowledge about ribosomal structures and their dynamics is mainly based on bacterial stuctures since only recently the structures of the 80S ribosome from *Tetrahymena thermophila* [73, 102] and *Saccharomyces cerevisiae* [12] have been obtained via X-ray crystallography. Despite the fact that the core structure of the ribosome is conserved amongst all three domains of life, the increased complexity of the eukaryotic ribosome reflects functional differences between between prokaryotes and eukaryotes and a more involved regulation of the protein synthesis process [73, 102]. Further structures at atomic resolution of the eukaryotic ribosome in different functional states are nevertheless necessary to reach the same level of knowledge available for the bacterial counterpart.

In bacteria the small 30S subunit binds the mRNA and contains the decoding center, which is responsible for the recognition and discrimination of the codon-anticodon interaction [91, 92, 94]. The large 50S subunit contains instead the peptidyl-transferase center (PTC), which is responsible for the synthesis of the peptide bond between two amino-acids. The fact that no proteins are present in the PTC brought to the conclusion that the ribosome is a ribozyme since only RNA moieties are involved in the peptide-bond synthesis reaction [9]. In addition, the PTC has been proposed as the most ancient component of the ribosome [3, 11, 17].

The interface between the two subunits consists mainly of RNA and the molecule of mRNA binds in a cleft between the head and body domains of the 30S subunit, where its codons interact with the anticodons of tRNA molecules. Both subunits contain three binding sites for tRNA molecules that are in three different functional states: The A site binds the aminoacyl-tRNA that is going to be added to the growing polypeptide chain, the P site holds the peptidyl-tRNA attached to the nascent polypeptide chain and the E site is occupied the deacylated tRNA before it is ejected from the ribosome [109, 117].

Initiation, elongation and termination are the three main stages in which translation can be divided in bacteria [109].

In the initiation phase the selection of the start site on the mRNA is facilitated by base pairing between the Shine-Dalgarno (SD) sequence upstream the start codon on the mRNA and the anti-Shine-Dalgarno (aSD) sequence at the 3' end of 16S rRNA of the small subunit [75, 115]. Initiation requires the ribosome to position the initiator fMet-tRNA^{fMet} over the start codon of mRNA in the P site [109]. Three initiation factors (IF1,IF2 and IF3) are

required in this process [79]. The elongation cycle consists of the steps involved in sequentially adding amino acids to the polypeptide chain, starting from a 70S ribosome containing a peptidyl-tRNA with a nacent polypeptide chain in the P site and an empty A site [109]. Decoding is the first step of the elongation and ensures that the correct aminoacyl-tRNA is selected in the A site. The amino acid is delivered in a ternary complex of elongation factor Tu (EF-Tu), GTP and aminoacyl-tRNA [109]. Decoding is followed by peptide-bond formation, where the amino acid carried by the aminoacyl-tRNA is added to the nascent polypeptide chain. During translocation, the third and last step of elongation, the elongation factor G (EF-G) promotes the movement of the tRNAs and mRNA with respect to the 30S subunit. The deacylated tRNA moves from the P site to the E site and the peptydil-tRNA moves from the A site to the P site. At the end of this process a new codon is presented in the empy A site. The elongation cycle continues until an mRNA stop codon moves into the A site, signalling the end of the coding sequence [109]. Termination is the last phase of the translation process. When the stop codon is positioned in the A site, a class I release factor recognizes it and cleaves the nascent polypeptide chain from the P-site tRNA, resulting in the release of the newly synthesized protein from the ribosome [109].

Protein synthesis is intrinsecally a dynamic process since it requires both small-scale and large-scale movements of tRNA and mRNA moieties [75]. In addition, also the two 30S and 50s subunits are charachterized by movements and structural changes at different scale levels [75]. On the large-scale, the 30S subunit has been observed to rotate in respect of the 50S subunit in all three phases of protein translation. The most studied of these movements is the ratchet-like rotation of the 30S subunit relative to the 50S subunit in the direction of the mRNA movement induced by the binding of EF-G to the ribosome during the translocation phase. After hydrolisis of the GTP the 30S subunit rotates back to the classic conformation [54, 75, 109, 129]. On the small-scale, the most characterized and significant movements are observed in the small subunit during the decoding mechanism. The decoding center is located in the A site of the 30S ribosomal subunit and is surrounded by four different domains: the head, shoulder, platform and helix 44 on the ribosome body. Crystal structures have revealed that 16S RNA bases of the decoding centre specifically contact the cognate codon-anticodon pair by an induced fit [91–94]. Nucleotides G530 from the shoulder domain, and A1492 and A1493 in helix H44 come together to span the minor groove of the codon-anticodon duplex at the first two codon positions. This results in a closed conformation of the 30S subunit, in which the shoulder, the head and the platform domains are rotated towards the subunit centre, compared to a more open structure when the A site is unoccupied [92, 134]. The sensing of the Watson-Crick base-pairing by nucleotides G530, A1492 and A1493 is responsible for the high accuracy of the ribosome in the discrimination between cognate and near-cognate codon-anticodon pairing and it has been proposed that the stabilization of a closed conformation is required for the tRNA selection. The cognate codon-anticodon pairing is characterized by Watson-Crick base pairing in

the first two codon positions, while a certain degree of variability (non-Watson-Crick base pairing) is allowed at the third codon position, called also wobble position. The allowed variability for the wobble position is due to the fact that the contacts of the ribosome at this position do not depend upon the precise shape of the minor groove [92]. Near-cognate codon-anticodon pairing is characterized instead by non-Watson-Crick base-pairing in the first or in the second codon position, affecting the geometry of the minor groove of the codon-anticodon helix. This mismatch has been shown to be unfavorable for a transition to a closed conformation of the small subunit and therefore the near-cognate tRNA is not selected. The antibiotic Paromomycin affects the rates of aminoacyl-tRNA selection by inducing a conformational change for A1492 and A1493 similar to the one observed when cognate tRNA is present in the decoding site. This results in a lower accuracy in the selection of the aminoacyl-tRNA and in the closure of the small subunit also in presence of near-cognate tRNA. [92]

Many structures are now available for procaryotic ribosomes, representing a vast amount of structural information. The novelty and the complexity of the ribosomal structure makes then the ribosome an interesting case for a structural comparative analysis performed with the ESCET framework.

Chapter 2

Aim of the work

The ESCET framework [110–112] has been developed for the comparative analysis of ensembles of structural models (protein or RNA macromolecules) obtained via X-ray crystallography.

Amongst the crystallographic parameters taken into account by ESCET, the B factors play and important role since they are used for the estimation of the coordinate uncertainties at atomic level, which are then propagated through the subsequent steps of the analysis. In order to avoid aftefacts being introduced into the results of the analysis the correctness of the B factors must be assessed. Since a validation tool for B factor distributions is still missing, the first part of the thesis is focused on the ***development of a new method for the validation of the distribution of isotropic B factors in crystallographic structural models***.

In addition, provided that the diffraction data are deposited into the PDB together with the atomic model, a re-refinement procedure is tested as a valid approach to improve those models with *suspicious* B factor distributions.

The methods developed for the validation of B factors find a direct application in the second part of the thesis, which is focused on the ***ensemble analysis of a selected data set of 30S ribosomal subunits from T. thermophilus***. Here, *suspicious* structures are detected and remedied via re-refinement. Different re-refinement protocols are tested and a comparative analysis is performed between the results of the ensemble analysis of the ribosomal models before and after re-refinement.

An outline of the actions performed in the two studies follows.

2.1 New Validation Method for B Factor Distributions

- A reference distribution for the distribution of isotropic B factors in crystallographic models is proposed.
- A validation protocol for the distribution of isotropic B factors is proposed, designed and implemented in R language.
- The validation protocol is tested on a data set of 15998 protein models at high resolution (equal to or higher than 2 Å).
- Guidelines for discriminating between *acceptable* and *suspicious* isotropic B factor distributions are given.
- A standard re-refinement protocol is tested as a valid procedure to rescue B factor distributions from *suspicious* to *acceptable*.
- A set of 12 L-alanine dehydrogenase (L-AlaDH) protein models from *Mycobacterium tubercolosis*, whose B factor distributions are rescued or improved after re-refinement, is selected for a comparative analysis of the outcome of the ensemble analysis performed on the models before and after re-refinement, characterized by *suspicious* and *acceptable* B factor distributions, respectively.

This study is the first to develop a validation procedure for B factor distributions in macromolecular X-ray crystallography.

2.2 Ensemble Analysis of Ribosomal Structures

- A working data set of 30S ribosomal structures from *Thermus thermophilus* is selected from the PDB.
- To improve, normalise and de-bias the selected structural models, 13 different refinement protocols are applied.
- The refinement protocol that provides the best models in terms of refinement statistics and distribution of isotropic B factors (judged with the validation protocol described above) is selected.
- A comparative analysis is performed on the outcomes of The ESCET framework applied to the deposited models and to those selected at the previous point. Particular emphasis is given to the de-biasing effect of the selected re-refinement procedure on the 30S structural models.

- The results obtained from the clustering and the rigid body analyses are related to the structural and functional information available in the literature.
- Considerations are made on the efficacy of the re-refinement as a tool to improve, normalise and de-bias low-resolution structural models and on how it affects the results obtained from the ESCET framework.

The novelty of this study is twofold:

1. It is the first assessment of the performance of the ESCET framework on a large data set of RNA structural models at low-resolution.
2. It is the first study on the effect of a re-refinement protocol on the results obtained from the ensemble analysis performed with the ESCET framework.

Chapter 3

Theoretical Background

In this chapter the crystallographic and statistical bases of the methods used in the thesis work are presented.

3.1 Crystal Structures and their Analysis

3.1.1 The Crystallographic Experiment

A macromolecular crystal can be defined as a periodic assembly in three-dimensional space of identical molecules [43, 106]. The unit cell is the building block from which the whole crystal is generated by applying translation operations, while the asymmetric unit is the smallest unit of a crystal structure from which by applying the space group symmetry operations the unit cell of the crystal is generated [106].

The goal of a crystallographic experiment is to compute from the diffraction pattern in reciprocal space, which is given by the interaction of X-rays with the scattering matter inside a crystal, an interpretable electron density distribution in real space that can be used to build a reasonable atomic model of the content of the asymmetric unit of the crystal (see Figure 3.1).

After the integration [78] and the scaling [50] of the collected diffraction images [20, 37], an inverse Fourier transformation can be used to compute from the observed diffraction amplitudes $F_{obs}(h,k,l)$ the electron density $\rho(x,y,z)$ via the electron density function [43, 106], defined as:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} F_{obs}(h,k,l)\exp[-2\pi i(hx+ky+lz-\alpha_{hkl})], \tag{3.1}$$

where x, y, z are coordinates in real space, V is the volume of the unit cell and α_{hkl} is

the phase angle of each reflection. The observed structure factor amplitude $F_{obs}(h,k,l)$ is proportional to the square root of the measured intensity of the diffraction reflection, defined by h,k,l coordinates in reciprocal space. Since in a crystallographic experiment the phase α_{hkl} cannot be directly measured together [122] with the diffraction intensities (also known as "crystallographic phase problem") other methods were developed to overcome this drawback. The three main classes of methods are: direct methods, experimental phasing and molecular replacement.

Once an interpretable electron density is calculated, several cycles of model building and refinement are usually necessary to obtain a model in a good agreement with the experimental data. The agreement between the observed data and the model is measured in reciprocal space by comparing the observed structure factors with those computed from the atomic model with a Fourier transformation via the structure factor equation

$$F_{calc}(h,k,l) = \sum_{i=1}^{n} q_i f_i \exp[2\pi i(hx_i + ky_i + lz_i)] \exp[-B_i \sin^2\theta/\lambda^2], \tag{3.2}$$

where n is the total number of atoms in the unit cell, q_i, f_i and B_i are the occupancy value, the atomic scattering factor and the isotropic B factor of the i^{th} atom (see section 3.1.3), respectively, θ is the scattering angle and λ is the wave length of the X-ray beam used during data collection. The R factor is the main measure used to evaluate the agreement between $F_{obs}(h,k,l)$ and $F_{calc}(h,k,l)$ (see section 3.1.2).

After each cycle of refinement, new phases are computed from the new coordinates and used in equation (3.1) to compute an updated electron density. The refinement is basically an optimization procedure that tries to maximise the agreement between $F_{obs}(h,k,l)$ and $F_{calc}(h,k,l)$ in the reciprocal space by varying the parameter of the model. When no substantial changes are observed in the parameters between two consecutive refinement iterations, the refinement is said to have reached convergence [72, 125].

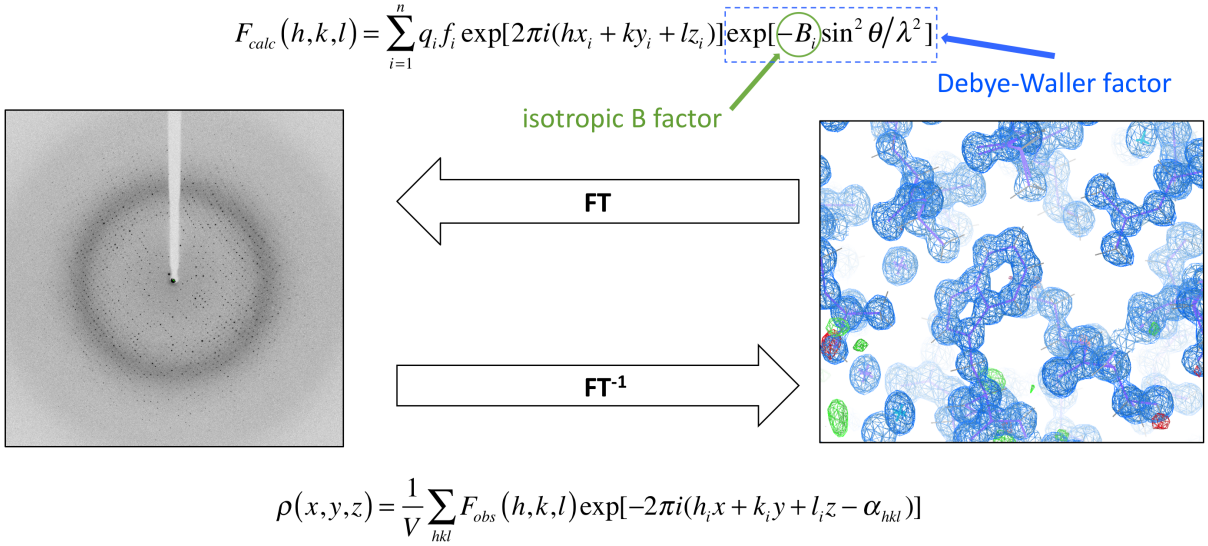


Figure 3.1: Summary of the crystallographic process. On the left a diffraction image is shown. On the right the computed electron density and the model fitted inside the electron density are shown. The blue density is the σ weighted map 2mFo-DFc contoured at +1.0 σ, the green density is the σ-weighted mFo-DFc contoured at +3.0 σ, the red density is the σ-weighted mFo-DFc contoured at -3.0 σ. The electron density figure was made with the Coot program [49]. The diffraction and the electron density images were kindly provided by Florian Sauer and they refer to the pdb-entry 3PUC, a structure of the titin domain M7 at 0.96 Å resolution.

3.1.2 The R Factors

The main statistic used to measure the global agreement between a crystallographic model and the experimental data is the R-value, defined as [26, 28, 106]:

$$R = \frac{\sum_{h,k,l} |F_{obs}(h,k,l) - F_{calc}(h,k,l)|}{\sum_{h,k,l} F_{obs}(h,k,l)}, \tag{3.3}$$

where h,k,l are the reciprocal lattice points of crystal, $F_{obs}(h,k,l)$ and $F_{calc}(h,k,l)$ are the observed and calculated structure factor amplitudes, respectively [26, 28].

Two different versions of the R factor are nowadays used: the R_{work} and the R_{free} . The first is computed with the structure factors that are used in model refinement. The second is computed with a smaller subset of structure factors (usually 5% or not less then 1000) that are omitted in the building and refinement process [26, 28]. The R_{free} is used as a cross-validation tool to monitor the degree of potential over-refinement. Analogously with the cross-validation method of testing statistical models the R_{free} can be considered as a parameter that express the prediction power of the model [26, 28, 106]. Moreover it has been shown that it is related to the mean phase error [27, 106] and thus it can be used as a measure of the quality of the model.

3.1.3 The Isotropic B Factor

The atomic isotropic B factor (B_{iso}) is directly related to the the mean square isotropic displacement $\langle u_{iso}^2 \rangle$ of an atom from its equilibrium position due to static or dynamic disorder inside the crystal lattice [106, 126]:

$$B_{iso} = 8\pi^2 \langle u_{iso}^2 \rangle. \tag{3.4}$$

In the approximation mostly used in macromolecular crystallography the atomic displacement is considered to be isotropic and a univariate Gaussian centred in the rest position of the atom is used to model it. The choice of the Gaussian distribution is supported by the harmonic approximation used to describe lattice dynamics [126, 133].

The effect of the B factor is to attenuate the magnitude of the atomic scattering factor in the structure-factor equation via the Debye Waller factor (see Figure 3.1), defined as [126]:

$$T_{iso} = \exp[-8\pi^2 \langle u_{iso}^2 \rangle (\sin^2 \theta)/\lambda^2]. \tag{3.5}$$

The effect of increasing values of B factors on the atomic scattering factor for selected atomic types is shown in Figure 3.2. The higher the B factor, the higher the attenuation of the atomic scattering factors, with significant effects especially at high resolution.

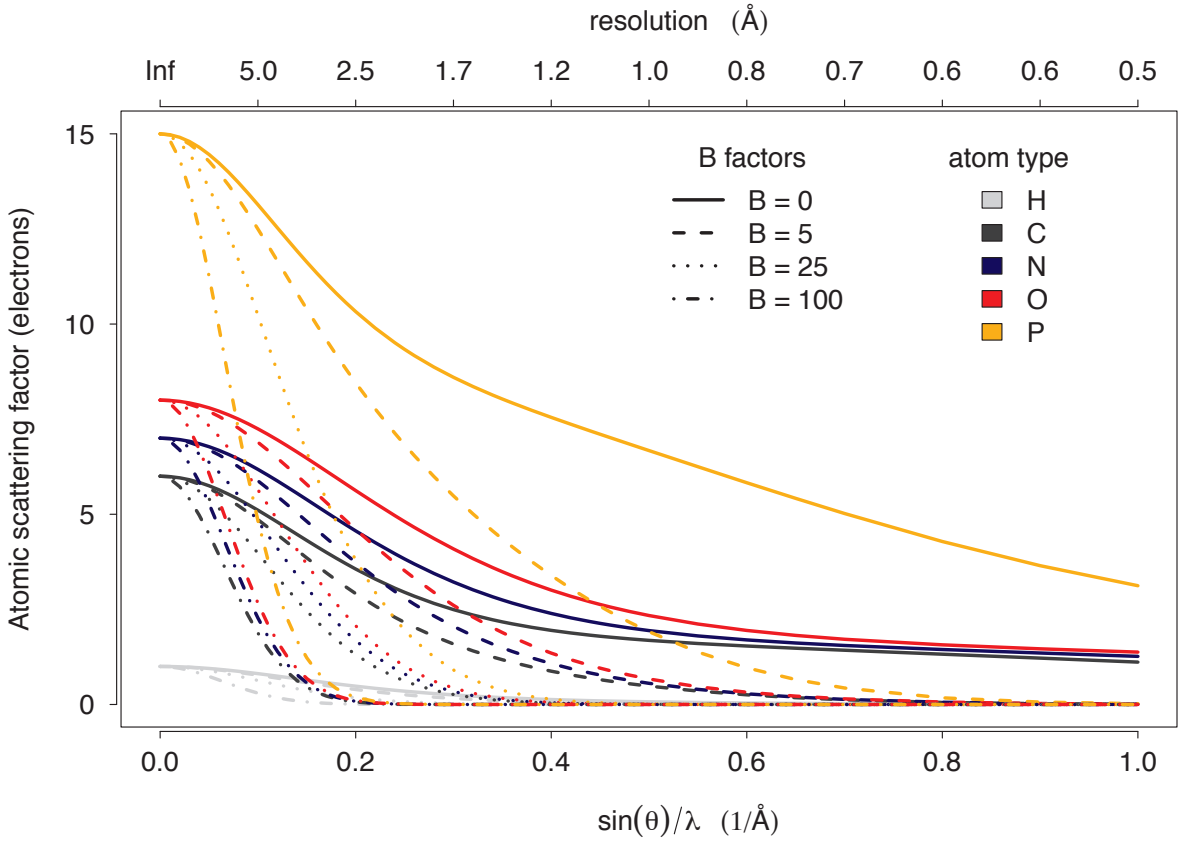


Figure 3.2: Effect of B factors on atomic scattering factors of the most common atoms in biological macromolecules (reproduced from Figure 6.13 in [106]). The atomic scattering factors for selected atom types were taken from Table 6.1.1.1 in [25]. The attenuated atomic scattering factor $f^{B_{iso}}$ is computed as $f^{B_{iso}} = f \cdot T_{iso} = f \cdot e^{-B_{iso}(\sin\theta/\lambda)^2}$, where f is the atomic scattering factor in absence of thermal vibration, T_{iso} is the isotropic Debye-Waller factor of equation (3.5). The abscissa at the bottom is plotted in units of $\sin\theta/\lambda = 1/2d\,(\text{Å}^{-1})$, where θ is the scattering angle and d is the interplanar spacing of the diffracting lattice planes [106]. The abscissa at the top reports the corresponding nominal resolution d_{min} in Å.

3.1.4 The Diffraction Precision Index (DPI)

The diffraction-component precision index (DPI) is an empirical formula that allows the calculation of the coordinate error $\sigma(x, \langle B \rangle)$ for an atom whose B factor is equal to the average B ($\langle B \rangle$) in a macromolecular crystal structure [36]. The DPI formula based on the R_{free} value is [36]:

$$\sigma(x, \langle B \rangle) = (N_i/n_{obs})^{1/2} C^{-1/3} R_{free} d_{min}, \tag{3.6}$$

where N_i is the number of sites in the model, n_{obs} is the number of reflections measured, C and d_{min} are the completeness and the maximum resolution of the diffraction data used in refinement and R_{free} is the free R value for the final model. The two main advantages of the DPI are that it can be computed without having the diffraction data available and that it allows to put onto a common absolute scale the errors in different models [36, 110].

3.1.5 The ESCET Framework

ESCET is a statistical framework that automatically identifies invariant and flexible regions in an ensemble of macromolecular crystal structures by taking into account experimental coordinate uncertainties [110–112]. The main workflow of the framework is as follows.

Starting from an ensemble of structural models (of the same or homologous macromolecules), for each pair of models a and b, an error-scaled difference distance matrix (EDD matrix) is used as an objective measure of the structural differences without superposition between the two structures [110]:

$$E_{ij}^{ab} = \frac{\Delta_{ij}^{ab}}{\sigma(\Delta_{ij}^{ab})}, \tag{3.7}$$

where $\Delta_{ij}^{ab} = d_{ij}^{a} - d_{ij}^{b}$ is the difference distance matrix (DD matrix) that corresponds to the difference in distance between atoms i and j in the two conformations a and b, and $\sigma(\Delta_{ij}^{ab})$ is the related experimental uncertainty. The experimental uncertainty $\sigma(\Delta_{ij}^{ab})$ can be computed by error propagation from the coordinate uncertainties $\sigma_i^a, \sigma_j^a, \sigma_i^b$ and σ_j^b of the individual atoms i and j in models a and b as follows:

$$\sigma(\Delta_{ij}^{ab}) = [(\sigma_i^a)^2 + (\sigma_j^a)^2 + (\sigma_i^b)^2 + (\sigma_j^b)^2]^{1/2}. \tag{3.8}$$

A modified version of the DPI [36] (equation (3.6)) is then used to estimate the coordinate uncertainty of each atom of the models in the ensemble. By assuming a linear relationship between the coordinate uncertainty of an atom i and its B value, B_i, an error estimate $\hat{\sigma}_{x,i}$ for the coordinate error $\sigma_{x,i}$ can be obtained as [110, 112]:

$$\hat{\sigma}_{x,i} = \frac{\sigma(x, \langle B \rangle)}{\langle B \rangle} B_i = \frac{B_i}{\langle B \rangle} (N_i / n_{obs})^{1/2} C^{-1/3} R_{free} d_{min} \tag{3.9}$$

and used in equation (3.8). In ESCET, the error model shown in equation (3.9) takes the name of DPIU.

Recently a new conformational similarity index (CSI) was introduced in the framework (data not published) to quantitatively measure the similarity of two structural models. The CSI is defined as:

$$\mathrm{CSI} = \frac{n_{id}}{n_{tot}}, \tag{3.10}$$

where n_{tot} is the total number of interatomic distances analysed in the EDD matrix (see equation 3.7) and n_{id} is the number of interatomic distances identical within the error ϵ_{low} (a threshold ϵ_{low} of $\pm 2\sigma$ is usually chosen for protein models). If two models are identical within the error, the CSI is equal to 1.0. If two models show instead significant conformational differences the CSI is smaller than 1.0.

The CSI is then used to compute the similarity between each pair of models in the ensemble and the resulting similarity matrix is analysed with an agglomerative hierarchical clustering algorithm [56]. An adaptive cutting ruled based in the KGS penalty function [68] is then used to cut the dendrogram obtained from the cluster analysis and to partition the ensemble of models in clusters of similar conformers.

From each cluster a representative structure is selected (usually the one with the lowest mean estimated standard uncertainty). The obtained subset of representative structures are then analysed by a genetic algorithm [111] for the identification of invariant regions, or rigid bodies, and flexible regions. In ESCET an invariant region or rigid body is defined as a group of atoms whose interatomic distances are identical within error. It follows that depending on the value of the threshold ϵ_{low} the number and the dimension of the identified rigid bodies change. In general, larger the threshold ϵ_{low}, fewer and larger rigid bodies are found.

3.2 Bayesian Statistics and Hypothesis Testing

3.2.1 Frequentist *versus* Bayesian Approach

Two main interpretations of probability are used by statisticians: the *frequentist*, or *classical approach*, and the *Bayesian approach* [18, 105, 113].

In the frequentist approach, the probability of an event is taken to be equal to the limit of the relative frequency of the chosen event with respect to all possible events as the number of trials goes to infinity [113]. In addition, the parameters, or numerical characteristics, of the population under analysis are considered fixed but unknown and thus they are not random quantities. Since probability statements are only allowed for random quantities, in the frequentist approach it is not possible to make probability statements about the value itself of the parameters. Instead, a sample is drawn from the population and the sample statistic is computed [18].

In the Bayesian approach, degrees of belief or knowledge are included in the statistical propositions [113]. This means that, since there is no certainty about the true value of the parameters, they are considered as random variables. In contrast to the frequentist approach, the laws of probability are directly used to make inferences about the parameters and the outcomes are interpreted as degrees of belief (i.e. no certainty is guaranteed). Moreover, every time new data are available the beliefs about the parameter are revised by using the Bayes' theorem, reflecting the dynamical nature of the Bayesian approach. While in the Bayesian approach the inference is based only on the occurring data, in the classical approach it is based on all possible data sets that could have occurred for the fixed

parameters [18].

Apart from a change of emphasis, the majority of the statistical procedures of the two approaches are identical since the axioms used to define the mathematical properties of probability remain unchanged [113].

3.2.2 Bayes' Theorem

Given two events A and B the Bayes' theorem is defined as [80, 106]:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}, \tag{3.11}$$

where the probability $P(A|B)$, which is called *posterior probability*, is the probability of the event A conditioned to the occurrence of the event B. The probability $P(A)$, which is called *prior probability*, is the probability of the event A without any knowledge about the occurrence of the event B. The probability $P(B|A)$, which is called *likelihood probability*, is the probability of the event B conditioned to the occurrence of the event A. The probability $P(B)$ is instead a proportionality constant, which ensures that the posterior probability is a valid probability with values between zero and one. The likelihood probability is usually known or can be easily computed, while the posterior probability is inferred from the combination of the prior probability with the likelihood probability. The terms *prior* and *posterior* refer to the change of knowledge about the event A after the occurrence, or the observation, of the event B [106].

In the context of a crystallographic experiment the Bayes' theorem can be rewritten as:

$$P(model|data) = \frac{P(model) \times P(data|model)}{P(data)}, \tag{3.12}$$

where the posterior probability $P(model|data)$ is the probability of the model after the observation of the data, the prior probability $P(model)$ is the probability of the model before the observation of the data, and the likelihood probability $P(data|model)$ is the probability of the observed data given the model under consideration [80, 106]. Since in experimental situation $P(data)$ is usually constant [80] and the goal of a crystallographic experiment is to maximise the probability of the model given the collected data (posterior probability), a version of the Bayes' theorem without $P(data)$ is normally used:

$$P(model|data) \propto P(model) \times P(data|model). \tag{3.13}$$

Bayes' theorem is also called the rule of inverse probability since it shows how to turn $P(data|model)$, that can by easily computed, into $P(model|data)$ [80]. The Bayes' the-

orem of equation 3.13 is applied at different stages of a crystallographic experiment from phasing to refinement. The power of Bayesian statistics is to incorporate prior information in the computation of the posterior probability. This is quite important at refinement stage where it is quite common to not have enough observations to refine all the parameters included in the model. In this case the prior distribution is given by the stereochemistry knowledge available from structures solved at higher resolution and coded under the form of restraints during refinement [80].

3.2.3 Conjugate Prior Distributions

One of most controversial concept in Bayesian statistics is the one corresponding to prior distributions since they express the degree of belief of the statistician regarding the distribution of the unknown parameter before actually seeing the data (i.e. if no information is available a uniform distribution can be used, otherwise any other distribution that express the belief of the statistician about the functional form of the prior distribution is acceptable).

The conjugate distributions in Bayesian statistics are a mathematical convenient choice for the functional form of the prior distribution. In general, arbitrary priors make for intractable mathematics. Since knowledge of the functional form of the prior is itself often vague, this has led to the development of a class of conjugate prior distributions, for which the prior and the posterior are of the same functional form [119]. It follows that the main advantage of using a conjugate prior distribution is that the posterior distribution can be computed analytically since it is of the same functional form of the conjugate prior distribution [95].

3.2.4 The Normal (or Gaussian) Distribution

The normal, or Gaussian distribution, $N(\mu, \sigma^2)$ is a continuous distribution with probability density function (pdf) defined as [83]:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{3.14}$$

where μ, defined in $-\infty < \mu < +\infty$, and σ^2, defined in $\sigma^2 > 0$, are the mean and variance parameters, respectively [83].

The Gaussian distribution is a very common and widely used distribution function in physical sciences, since random errors are expected to be normally distributed [15].

3.2.5 Inverse-Gamma Distribution (IGD)

The Inverse-Gamma distribution (IGD) is known from statistical theory to be a valid conjugate prior for the variance parameter σ^2 in a normal model $N(\mu, \sigma^2)$ with known mean μ and unknown variance σ^2 [95]. The IGD belongs to the family of continuous exponential distributions and its probability density function (pdf) is defined as [95]:

$$f(x|\alpha,\beta) = IG(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \quad \forall \alpha,\beta,x > 0 \tag{3.15}$$

where α and β are called the shape and scale parameters of the distribution, and $\Gamma(\alpha)$ stands for the gamma function with argument α (some representative IGDs are shown in figure 3.3).

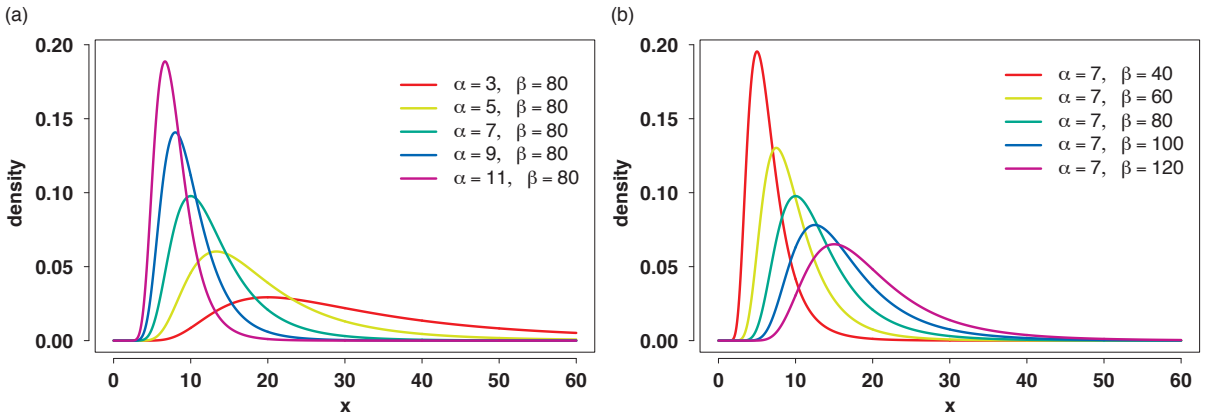


Figure 3.3: Examples of IGDs with variable α and constant β (panel (a)) and with constant α and variable β (panel (b)).

The mean value μ and the variance σ^2 for a given $IG(\alpha,\beta)$ can be computed based on the α- and β-parameters [95] as:

$$\mu = \frac{\beta}{\alpha - 1} \quad \forall \alpha \in \mathbb{R} > 1 \tag{3.16}$$

and

$$\sigma^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \quad \forall \alpha \in \mathbb{R} > 2. \tag{3.17}$$

It should be noted that if $0 < \alpha \leq 1$ the distribution is proper but the mean does not exist, while the variance exists only for $\alpha > 2$ [95].

If the statistics μ and σ^2 of an IGD are defined (note the existence conditions on α in equations (3.16) and (3.17)) and known, equations (3.16) and (3.17) can be rearranged

such that it is possible to calculate the shape (α) parameter

$$\alpha = \frac{\mu^2}{\sigma^2} + 2 \quad \forall \sigma \in \mathbb{R} > 0 \tag{3.18}$$

and the scale (β) parameter

$$\beta = \mu(\alpha - 1) \quad \forall \alpha \in \mathbb{R} > 1 \tag{3.19}$$

from the known quantities.

3.2.6 Shifted Inverse-Gamma Distribution (IGD*)

A shifted version of the IGD, called shifted Inverse-Gamma distribution and denoted hereafter as IGD* or IG*(α,β,γ), can be derived from equation (3.15). Its pdf is defined as:

$$f^*(x|\alpha,\beta,\gamma) = IG^*(\alpha,\beta,\gamma) = \frac{\beta^\alpha}{\Gamma(\alpha)}(x-\gamma)^{-\alpha-1}\exp\left(-\frac{\beta}{(x-\gamma)}\right) \tag{3.20}$$

where α and β are the shape and scale parameters, respectively, $\Gamma(\alpha)$ is the gamma function with argument α, and γ is the new shifting parameter. The new conditions of existence become: $\forall \gamma \in \mathbb{R} \geq 0$ and $\forall x > \gamma$, $\forall \alpha, \beta \in \mathbb{R} > 0$. The effect of the third parameter γ is simply to translate a classic IG(α, β) towards higher values of x. It should be noted that, while the classic IGD is defined only for positive values of its argument x, the interval on which the IGD* is defined becomes now from γ to $+\infty$.

From the combination of equations 3.18 and 3.19 together with the properties of mean and variance (see appendix B) it is possible to obtain the definitions of the shape (α) and the scale (β) parameters for the IGD*:

$$\alpha = \frac{(\mu-\gamma)^2}{\sigma^2} + 2 \quad \forall \sigma \in \mathbb{R} > 0, \forall \gamma \in 0 \leq \mathbb{R} < \mu \tag{3.21}$$

and

$$\beta = (\mu-\gamma)(\alpha - 1) \quad \forall \gamma \in 0 \leq \mathbb{R} < \mu, \forall \alpha \in \mathbb{R} > 1 \tag{3.22}$$

A similar formulation of the IGD* has been recently used as appropriate prior distribution for the unknown scaling of signal coefficient variances relative to noise variance for noise-floor estimation in archived audio recordings [57].

3.2.7 Maximum Likelihood Estimation (MLE)

Given n independent observations from the same distribution, the joint probability of the observations as a function of a single unknown parameter θ is called the likelihood function (LF) of the of the sample [120]:

$$L(x|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n}(x_i|\theta). \tag{3.23}$$

The principle of Maximum Likelihood states that among different choices for the unknown parameter θ the one ($\hat{\theta}$) that maximises the likelihood function in equation (3.23) should be chosen [119]:

$$L(x|\hat{\theta}) \geq L(x|\theta). \tag{3.24}$$

Computationally it is often more convenient to work with the logarithm of the likelihood function because in this way the product in equation (3.23) becomes a sum and then it is easier to handle within the machine precision. This approach takes advantage of the property of logarithms and of the fact that, being the logarithm a monotone function, the likelihood function and its logarithm have maxima at the same values of their arguments [120]. From equation (3.23) the so-called log-likelihood function is defined as:

$$\log L(x|\theta) = \sum_{i=1}^{n} \log L(x_i|\theta) \tag{3.25}$$

A normalized version of equation (3.25), called average-log-likelihood, is often used and defined as:

$$\ell = \frac{1}{n}\log L(x|\theta) = \frac{1}{n}\sum_{i=1}^{n} \log L(x_i|\theta) \tag{3.26}$$

where the log likelihood of equation (3.25) is normalized by the size of the sample. The advantage of the average-log-likelihood is that it allows one to compare log-likelihood values obtained from samples of different sizes.

By definition the goal of the MLE is to find the estimator $\hat{\theta}$ that maximises equation (3.25). However, by changing the sign of the log-likelihood function, the problem is reversed to find the estimator $\hat{\theta}$ that minimizes that function, allowing the use of the available optimization algorithms. Here, the so-called negative average-log-likelihood function is defined as:

$$-\ell = -\frac{1}{n}\log L(x|\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log L(x_i|\theta). \tag{3.27}$$

3.2.8 Tests of Hypotheses

A statistical hypothesis is an assertion about the distribution of one or more random variables [83] and it is called hypothesis since it is not known in advance whether it is true or not [105].

If the parameters of the underlying distribution are fully specified the test of hypotheses is called parametric and non-parametric otherwise. Further, if not even the underlying form of the distribution is specified, then the test is called distribution-free.

Usually in hypotheses-testing problems two hypotheses are discussed [83]: the *null hypothesis (H0)* and the *alternative hypothesis (H1)*. H0 corresponds to the hypothesis being tested, while H1 corresponds in general to its negation. In fact, the main idea is that if H0 is false than H1 is true and vice versa [83].

The outcome of a test of hypothesis is the acceptance or the rejection of the null hypothesis H0. In this decision making process two errors can occur. The first, called *type I error*, is the rejection of H0 when it is true, while the second, called *type II error*, is the acceptance of H0 when it is false [83]. The usual approach is to reject H0 only when very unlikely. The objective of a statistical test is then not to explicitly determine whether or not the hypothesis under analysis is true but rather to determine if its validity is consistent with the data [105]. To accomplish that a value α, called the *level of significance of the test*, is specified before performing the statistical test and it is required that the test has the property that whenever H0 is true its probability of being rejected is never greater than α. This corresponds to say that the probability of a type I error can never be greater than α [105].

3.2.9 Kolmogorov-Smirnov Test (KS-test)

The two-sided two-sample Kolmogorov-Smirnov test (KS-test) belongs to the category of nonparametric statistical methods [61]. It is a distribution-free test of the general null hypothesis that two samples are identical in all respects or, in other words, that they are drawn from the same distribution [61]. In terms of hypothesis testing this can be reformulated as:

- H0: The two samples *belong* to the same distribution
- H1: The two samples *do not belong* to the same distribution

The basic assumption of the test is that if two samples $X = x_1, x_2, \ldots, x_m$ and $Y = y_1, y_2, \ldots, y_n$ are drawn from the same distribution their cumulative distribution functions

$$F_m(t) = \frac{\text{number of sample } X\text{'s} \leq t}{m} \tag{3.28}$$

and

$$G_n(t) = \frac{\text{number of sample } Y\text{'s} \leq t}{n} \tag{3.29}$$

should be identical or at least very similar [61]. The largest distance between the two cumulative distribution functions is then computed as:

$$D = \max_{-\infty \leq t \leq +\infty} |F_m(t) - G_n(t)| \tag{3.30}$$

and the KS-test statistics J^* computed by weighting the distance D with the size of the two samples as follows:

$$J^* = \sqrt{\frac{m \times n}{m+n}} \times D = \sqrt{\frac{m \times n}{m+n}} \times \max_{-\infty \leq t \leq +\infty} |F_m(t) - G_n(t)| \tag{3.31}$$

It's worth noting that this is the definition of KS-test statistics that has been used in the thesis work and it is supported by the large-sample approximation. In case of small samples the equation (3.31) takes a slightly different form (see pages 178-179 in Hollander M. and Wolfe D.A. (1999) [61]). For each comparison between two samples X and Y the test returns a *p-value* that is an estimate of the probability to obtain by chance a *KS-test* statistic J^* at least as extreme as the one computed, assuming that H_0 is true. For large samples an asymptotic distribution is used for the estimation of the p-value. Thus, the lower the p-value the less likely the null hypothesis is. If the p-value is lower then the the *level of significance* α, then the null hypothesis is rejected and considered very unlikely that the two samples belong to the same population.

Chapter 4

Materials and Methods

4.1 Selection of Protein Structures Data Set

For the selection of a large data set of protein crystal structures at high resolution from the Protein Data Bank (PDB) [14] the following criteria were applied :

- The resolution $d_{\min}$ of the crystal structure is equal to 2.0 Å, or higher.
- Experimental structure factors are available.
- The model contains more than 1000 non-hydrogen (non-H) protein atoms with occupancy equal to one and B factors higher than zero. Only atoms that belong to standard amino-acids are taken into consideration.
- Models refined with one constant B factor for all atoms are excluded.
- Models for which only backbone atoms were refined are excluded.
- Models for which TLS-refinement had been applied are excluded.

The analyses based on full models of crystal structures which can contain multiple chains are denoted as *all-chains* and analyses based on individual chains as *single-chains*.

For the *single-chains* data set a lower bound limit 500 atoms was fixed as a compromise to guaranty robustness of the statistical tools applied and at the same time to take into account as many structures as possible for the analysis.

The advanced search query tool of the RCSB PDB (www.pdb.org) [14] was used to retrieve the PDB codes of the protein structures that fulfill the resolution and availability of the structure factors criteria.

The coordinate files with the respective structure factors were downloaded from the worldwide PDB (www.wwpdb.org) [13].

Perl (www.perl.org) and Python (www.python.org) scripts were used to download the structure coordinates and structure factors, and to extract refinement information from the downloaded coordinate files. The selection for the number of atoms was done during the analysis phase with in R [101]. The R package *bio3d* [58] was used to upload and extract information from the pdb-files.

4.2 Re-refinement of Protein Structures

The CCP4 software suite [135] was used for all the steps required for the re-refinement of the protein structures. A modified version of the *pdb_redo.csh* script (version 2.5) from the PDB_REDO project [65] was used to coordinate and perform the conversion of structure factors files and the refinement of PDB models.

The main changes introduced in the script are in the setup of the refinement protocol. Before refinement the B factors of the PDB model were set to the Wilson B value computed with the *sfcheck* program [128] to avoid possible bias from the B factors present in the deposited model. The refinement of each protein model was then performed with the REFMAC5.6 version 75 [89,90]. The following parameterisations were used: The allowed B factor range was defined from 0.1 $Å^2$ to 500 $Å^2$; The number of cycles of refinement was set to 50 to guarantee convergence of the refinement procedure; the automatic detwinning procedure implemented in REFMAC5.6 was activated to properly handle twinned structures during the refinement.

A fresh re-refinement of the deposited structures was preferred to the re-refined structures available from the PDB_REDO [65] project for several reasons. Firstly the structures available for download were not usually refined with the same protocol. Secondly the majority of these structures were obtained by applying TLS refinement. This contradicts the last criterion used for the selection of the data set of protein structures (see section 4.1).

4.3 Validation of B Factor Distributions

Starting from the assumption that the IGD* is the reference distribution to be used for validating B factor distributions (see section 5.1), a validation protocol was designed and implemented in R language [101] (see Fig. 4.1 and the script is available in Appendix A). A detailed description of each step of the protocol follows.

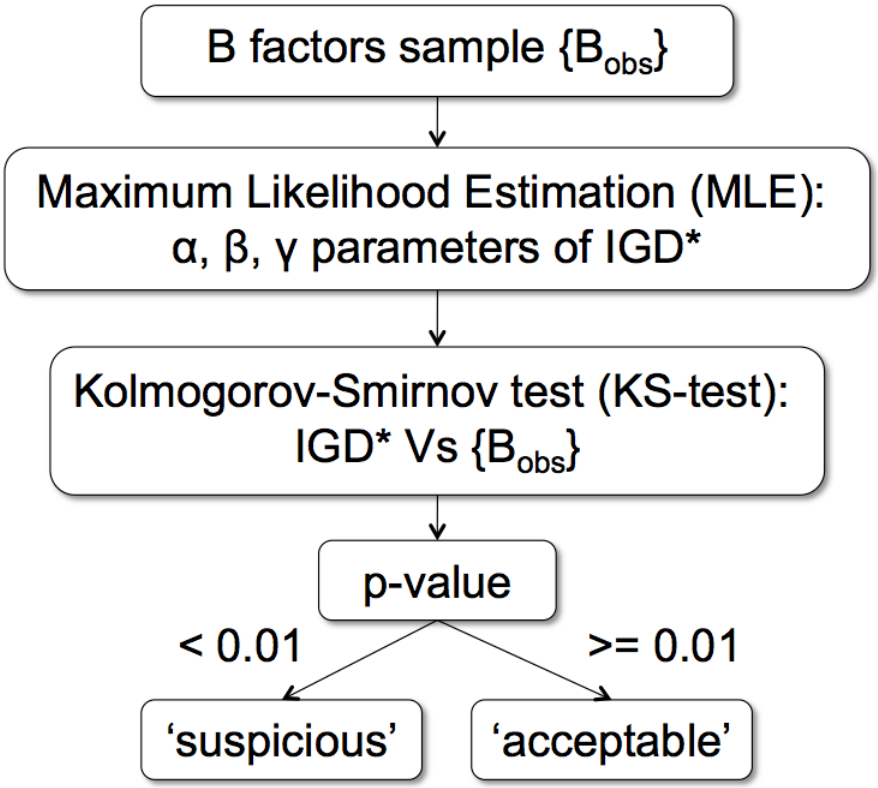


Figure 4.1: Validation protocol flowchart. The main steps of the validation protocol are here summarized.

4.3.1 Determination of IGD* Parameters

Given a sample of isotropic B factors $\boldsymbol{B}_{obs}$ from a protein model, a Maximum Likelihood Estimation (MLE) procedure (see section 3.2.7 and script in Appendix A) was used to obtain the estimators $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ for α, β and γ parameters of the IGD* that best fit the empirical distribution of $\boldsymbol{B}_{obs}$.

From equation (3.20) the negative log likelihood function and the average negative log-likelihood function (as defined in section 3.2.7) are respectively:

$$\begin{aligned} &-\log L(\boldsymbol{x}\,|\,\hat{\alpha},\hat{\beta},\hat{\gamma}) \\ &= -n\,\hat{\alpha}\log(\hat{\beta}) + n\log\left(\Gamma(\hat{\alpha})\right) + (\hat{\alpha}+1)\sum_{i}^{n}\log(x_i-\hat{\gamma}) + \hat{\beta}\sum_{i}^{n}\frac{1}{x_i-\hat{\gamma}} \end{aligned} \tag{4.1}$$

and

$$-\ell = -\frac{1}{n}\log L(\boldsymbol{x}|\hat{\alpha},\hat{\beta},\hat{\gamma})$$
$$= -\hat{\alpha}\log(\hat{\beta}) + \log(\Gamma(\hat{\alpha})) + \frac{(\hat{\alpha}+1)\sum_{i}^{n}\log(x_i-\hat{\gamma}) + \hat{\beta}\sum_{i}^{n}\frac{1}{x_i-\hat{\gamma}}}{n}, \tag{4.2}$$

where n is the size of the B factors sample and $\boldsymbol{x}$ is defined as $\boldsymbol{x} = \boldsymbol{B}_{obs} - \hat{\gamma}$. Equation (4.2) is the target function used in the minimization procedure performed by using an L-BFGS-B quasi-Newton method [136]. This algorithm needs initial estimates of α, β and γ parameters at the start of the optimization. However, as shown in equations (3.21) and (3.22), both α and β parameters of the IGD* are functions of γ parameter. In order to make the estimation of the optimal IGD* parameters more robust and avoid that the minimization procedure becomes trapped in a local minimum the minimization procedure was combined with a grid search along a set of reasonable γ parameter values as follows:

1. Define a set of initial values $\boldsymbol{\gamma_0}$ for the γ parameter in the interval $[0, min(\boldsymbol{B}_{obs}) - 0.01]$, by using a step size of 2 Å^2.

2. For each of the initial $\boldsymbol{\gamma_0}$ parameters computed in step 1:

   (a) Compute the initial values α_0 and β_0 for estimates of α and β parameters from equations (3.21) and (3.22). The sample mean and the variance of the experimental B factors were respectively used as estimates for the population mean (μ) and variance (σ^2) in equations (3.21) and (3.22).

   (b) Minimize function (4.2) starting from γ_0 and from α_0 and β_0 parameter estimates computed at step 2a.

   (c) Store final IGD* parameters estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, and the relative negative average-log-likelihood.

3. Choose the set of final IGD* parameter estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, that give the lowest value of equation (4.2).

4.3.2 Evaluation of the Goodness of Fit

A two-sample two-sided Kolmogorov-Smirnov test (KS-test) [61] (see section 3.2.9) was used to quantitatively evaluate the agreement between the observed B factor distribution and the fitted IGD* obtained via MLE.

In the context of the validation of B factor distributions the *null hypothesis* (H0) and the *alternative hypothesis* (H1) are defined as follows:

- H0: The B factor distribution under analysis belongs to an IGD* whose parameters α, β and γ were estimated via MLE.
- H1: The B factor distribution under analysis does not belong to an IGD* whose parameters α, β and γ were estimated via MLE.

Since the parameters of the IGD* are estimated from the target sample, the null hypothesis is not anymore simple but becomes composite. This complicates the evaluation of the p-value from the KS-test [83, 119]. To obtain a robust estimation of the p-value under these circumstances, the following parametric bootstrap [47, 48] approach was applied:

1. Obtain via MLE the estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ for the α, β and γ parameters of the IGD* that best fit the empirical B factor distribution.
2. For 1000 times:
   (a) Assuming that n is the size of the B factors' sample $\boldsymbol{B}_{obs}$, draw n points from an IG*($\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$) where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ parameters were computed at step 1.
   (b) Perform a two-sample two-sided KS-test by comparing the B factors' sample and the population of n points generated at step 2a.
   (c) Store the p-value given by the KS-test.
3. Compute a bootstrapped estimate of the p-value by averaging the 1000 p-values produced at step 2.

Since in R $2.220446e-16$ is the lowest positive floating-point number x for which $1+x \neq 1$ [34, 101], if the bootstrapped estimate of the p-value is lower than this number, '$< 2.2e-16$' is returned by the protocol instead of a precise estimate. The iteration of the bootstrap procedure was fixed to 1000 as a compromise between a reasonably low standard deviation of the estimated average p-value and acceptable CPU time consumption.

For the interpretation of the p-value a significance level of 0.01 is used. If the mean p-value resulting from the bootstrap procedure is less than 0.01, the null-hypothesis is rejected and the distribution of B factors flagged as *suspicious*; conversely, if the mean p-value is equal to or higher than 0.01, the null-hypothesis is not rejected, and the the distribution of B factors flagged as *acceptable*.

In hypothesis testing four different categories of situations are identifiable:

1. True positives are those structures that do not follow an IGD* and have a p-value lower than 0.01. We refer to them as *suspicious* structures.
2. True negatives are those structures that follow an IGD* and have a p-value equal or higher than 0.01. We refer to them as *acceptable* structures.

3. False positives (incorrectly called *suspicious*) are those structures that globally follow an IGD* but that have a p-value lower than 0.01. They populate the type I error.

4. False negatives (incorrectly called *acceptable*) are those structures that do not follow an IGD* but have a p-value equal or higher than 0.01. They populate the type II error.

It should also be noted that an exact p-value from a KS-test can only be calculated in the absence of ties, where ties stand for repeated values [61]. Since we are dealing with experimental distributions such as the observed B factor distributions, here ties are a common phenomenon (see below). When ties are present, the measured D will be larger or equal to the D measured in the absence of ties. Correspondingly, the derived p-values will become a lower limit estimate of the true p-value. In the context of this study an underestimated p-value will give rise to false positives, i.e. acceptable structures will be incorrectly categorized as suspicious. Such cases actually do not create a problem since they can be re-categorized by eye or by applying orthogonal statistics (see below).

4.3.3 Orthogonal Statistics

The estimates of α, β and γ parameters of the fitted IGD* and the p-value from the KS-test are the main measures for the detection of structures with a suspicious B factor distribution. Two orthogonal statistics were introduced to further characterise the *suspicious* B factor distributions detected by a p-value lower than 0.01.

- *B factor equal to or lower than zero*. The IGD* is defined only for positive values (see equation (3.20)) and then, if present, B factors equal to or lower than zero are not included in the MLE analysis and later in the KS-test step. B factors equal to zero are meaningless since they would correspond to atoms steady in space that are not vibrating. Such an event is impossible. Their presence in a protein model can be due to errors introduced during the refinement or model building (i.e. wrong atom type).

- *Strong ties*. Depending on the magnitude and the nature of the process that produced them, their presence will negatively bias the outcome of the KS-test (see above).

In the protocol the statistic B_{null} is used to take into account the presence of B factors equal to or lower than zero. It is defined as:

$$B_{null} = \sum_{i=1}^{n} f(B_i) \tag{4.3}$$

where

$$f(B_i) = \begin{cases} 1 & \text{if } B_i \leq 0, \\ 0 & \text{if } B_i > 0. \end{cases} \tag{4.4}$$

and n is the total number of non-H atoms, while B_i is the B Factor value of the i^{th} atom. Every time a structure has $B_{null} > 0$ a warning flag should be raised.

To detect strong ties the following procedure was introduced. Let $\boldsymbol{B}_{obs}$ be the set of experimental B factors from a crystallographic structure with n non-H atoms and $\boldsymbol{B}_{unique}$ be the subset of $\boldsymbol{B}_{obs}$ where all the duplicated values have been removed. If n is the size of $\boldsymbol{B}_{obs}$ and j the size of $\boldsymbol{B}_{unique}$, it follows by definition that $j \leq n$. Let us introduce now a third set called $\boldsymbol{B}_{ties}$ of size j, the same as $\boldsymbol{B}_{unique}$, that counts the number of times each value in $\boldsymbol{B}_{unique}$ appears in $\boldsymbol{B}_{obs}$. Any i^{th} element in $\boldsymbol{B}_{ties}$ is defined as follows:

$$B_{ties,\ i} = \sum_{k=1}^{n} f(B_{obs,k}) \tag{4.5}$$

where

$$f(B_{obs,k}) = \begin{cases} 1 & \text{if } B_{obs,k} = B_{unique,\ i}, \\ 0 & \text{if } B_{obs,k} \neq B_{unique,\ i}. \end{cases} \tag{4.6}$$

Considering that B Factors are rounded until the second decimal digit in a PDB model, that they contain noise and errors, and that they are usually restrained during refinement (i.e. two atoms close to each other in three-dimensional space or connected by a bond have similar B factors) there is a non-zero probability to find duplicated values. This is particularly true when the number of refined atoms is large (i.e. large complexes, multiple chains). In this study only those structures that are strongly biased by the presence of ties were identified. To achieve this an empirical rule based on a *z-score* approach was applied by considering as outliers those structures for which the highest value in $\boldsymbol{B}_{ties}$ is higher than the arithmetic mean of $\boldsymbol{B}_{ties}$ plus 9 times its standard deviations. This gives the equation:

$$\text{if } \max(\boldsymbol{B}_{ties}) > \left(\langle \boldsymbol{B}_{ties} \rangle + 9\sqrt{\text{Var}(\boldsymbol{B}_{ties})} \right) \Rightarrow \boldsymbol{B}_{ties} \text{ is suspicious} \tag{4.7}$$

The multiplicative coefficient of 9 was chosen empirically in order to maximize the detection of large ties at the lower and upper ends of the B factors range.

4.4 Selection of Ribosomal Structures

For the selection, downloading and analysis of all ribosomal structures available at a resolution equal to or higher than 4 Å, the SOAP and RESTful web services at RCSB PDB database were used [14].

The program *blastn* was used to query the PDB. The sequences of chains A from PDB codes 1J5E and 1J01 were used to retrieve all 30S and 50S subunits from *Thermus thermophilus*, respectively, the sequences of chains A from PDB codes 2QAL and 2QAM were used to retrieve all 30S and 50S subunits from *Escherichia coli*, respectively, the sequences of chains 0 from PDB codes 1NKW and 1VQ8 were used to retrieve all 50S subunits from *Deinococcus radiodurans* and *Haloarcula marismortui*, respectively.

Python scripts were used to automate and coordinate the analysis process.

4.5 Re-refinement of Ribosomal Structures

The coordinates and the structure factors of the 29 selected ribosomal structures reported in table 7.2 were downloaded from the PDB. The *phenix.cif_as_mtz* program [1] was used to convert the structure factors from the mmCif format to the mtz format.

The *phenix.refine* program [1] was then used to refine the ribosomal structure. The following 13 different protocols were applied:

- **Protocol 1**:
  - Processed all input PDB files with *phenix.ready_set* [1] program before refinement
  - Run 3 macro-cycles of *phenix.refine* program
  - Assigned B_{wil} to all atoms before refinement
  - Refined individual ADP and occupancies
- **Protocol 2**:
  - Taken as input files the output files from protocol 1
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined individual ADP, TLS groups (1 group per chain), individual sites, occupancies
  - Applied simulated annealing refinement for individual sites (2^{nd} and 4^{rd} macro-cycles)

- **Protocol 3**:
  - Taken as input files the output files from protocol 1
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined individual ADP, individual sites, occupancies
  - Applied simulated annealing refinement for individual sites (2^{nd} and 4^{rd} macro-cycles)
- **Protocol 4**:
  - Taken as input files the output files from protocol 1
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined individual ADP, occupancies
- **Protocol 5**:
  - Taken as input files the PDB files as deposited in the PDB
  - Run 3 macro-cycles of *phenix.refine* program
  - Computed only bulk solvent contribution
- **Protocol 6**:
  - Processed all input PDB files with *phenix.ready_set* program before refinement
  - Run 3 macro-cycles of *phenix.refine* program
  - Assigned B_{wil} to all atoms before refinement
  - Refined grouped ADP (2 groups per residue), occupancies
- **Protocol 7**:
  - Taken as input files the output files from protocol 6
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined grouped ADP (2 groups per residue), occupancies
- **Protocol 8**:
  - Taken as input files the output files from protocol 6
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined grouped ADP (2 groups per residue), individual sites, occupancies
  - Applied simulated annealing refinement for individual sites (2^{nd} and 4^{rd} macro-cycles)

- **Protocol 9**:
  - Processed all input PDB files with *phenix.ready_set* program before refinement
  - Run 3 macro-cycles of *phenix.refine* program
  - Assigned B_{wil} to all atoms before refinement
  - Refined grouped ADP (1 group per residue), occupancies
- **Protocol 10**:
  - Taken as input files the output files from protocol 9
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined grouped ADP (1 group per residue), occupancies
- **Protocol 11**:
  - Taken as input files the output files from protocol 9
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined grouped ADP (1 group per residue), individual sites, occupancies
  - Applied simulated annealing refinement for individual sites (2^{nd} and 4^{rd} macro-cycles)
- **Protocol 12**:
  - Processed all input PDB files with *phenix.ready_set* program before refinement
  - Run 3 macro-cycles of *phenix.refine* program
  - Computed only bulk solvent contribution
- **Procotol 13**:
  - Taken as input files the output files from protocol 12
  - Run 5 macro-cycles of *phenix.refine* program
  - Refined individual sites, occupancies
  - Applied simulated annealing refinement for individual sites (2^{nd} and 4^{rd} macro-cycles)

Python scripts were used to automate and coordinate the re-refinement processes.

4.6 ESCET Protocols

4.6.1 Protein Structures

The framework ESCET was used for the analysis of the ensemble of L-Alanine dehydrogenase (L-AlaDH) structures from *M. tubercolosis* discussed in chapter 6.

Since the R_{free} values are available in the deposited and re-refined structures, the DPIU error model (see section 3.1.5) was used to assign coordinate uncertainties to all atoms in the structures.

The CA-atoms were used as representative atoms for the structural comparison between the structures in the ensemble and for the identification of rigid bodies and flexible regions.

The minimum size for a rigid body was set to 10 amino-acids and a ϵ_{low} of 2 was used for the rigid body analysis.

The ensemble type was set to *similar* to identify the consistent CA-atoms in the ensemble of protein structures with the structural alignment algorithm implemented in RAPIDO [84, 85]. This guaranteed that the comparisons were made by using only structural information and no bias in sequence order or numbering would affect the final results.

4.6.2 Ribosomal Structures

The ESCET framework was used for the analysis of the ensemble of 29 30S ribosomal subunit structures from *T. thermophilus* discussed in chapter 7.

Since the R_{free} values are available in the deposited and re-refined structures, the DPIU error model (see section 3.1.5) was used to assign coordinate uncertainties to all atoms in the structures.

The P-atoms from the 16S rRNA moieties were used as representative atoms for the comparison between the structures in the ensemble and for the identification of rigid bodies and flexible regions. The phosphate atoms were chosen as the representative atoms for the analysis because they are the strongest scatterers present in the ribosomal structures as shown in Figure 4.2. Since the outcome of a crystallographic experiment is an electron density map and the phosphate atoms give the strongest signal, they are considered the most reliable atoms for structural comparisons.

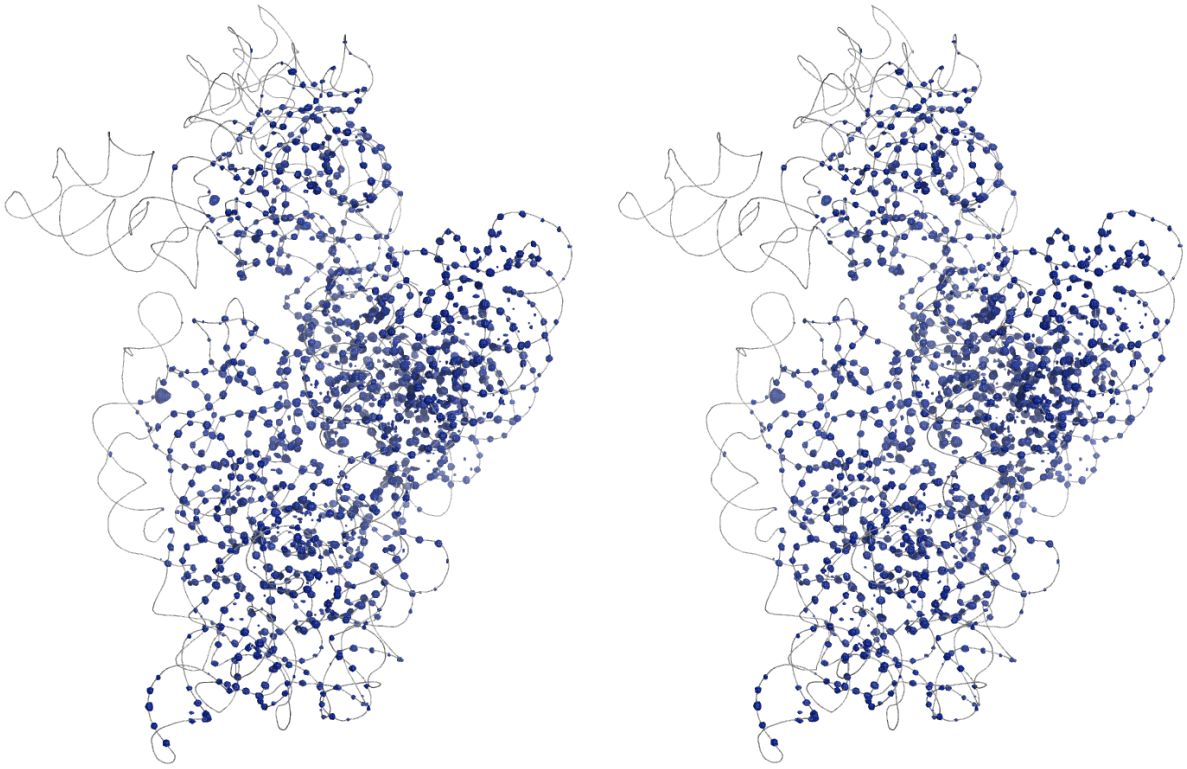


Figure 4.2: Wall-eye stereo picture of the σ-weighted 2mFO-DFc electron density map for the pdb-entry 1J5E after re-refinement with protocol 2 (section 4.5) contoured at 6.5σ and carved at 2 Å from each non-ion atom in the asymmetric unit. The 16S rRNA is shown in ribbon representation and gray coloured. It should be noted that clear density is visible only for phosphate atoms along the ribbon representation of the rRNA moiety. No density is visible for any protein present in the 30S complex.

The minimum size for a rigid body was set to 20 nucleotides and a ϵ_{low} of 1.5 was used for the rigid body analysis.

The ensemble type was set to *similar* to identify the consistent P-atoms in the ensemble of 16S rRNA ribosomal structures with the structural alignment algorithm implemented in RAPIDO [84,85]. This guaranteed that the comparisons were made by using only structural information and no bias in sequence order or numbering would affect the final results.

4.7 Graphics

All three-dimensional figures of protein and ribosomal structures shown in this thesis were created with Pymol (www.pymol.org), or with Coot if differently specified [49]. The graphical output of the validation method for B factors, bar-plots and scatter-plot were produced with R. For correlograms the R package *corrgram* was used. Microsoft PowerPoint®, Adobe® Illustrator® and Adobe® Photoshop® suites were used for the final preparation of the images present in this thesis.

4.8 Computational Resources

All the analyses reported in this thesis were performed by using the following computational infrastructures: iMac with Intel® Core™ 2 Duo CPU at 2.4 GHz, Linux workstation with 8 CPUs Intel® Xeon® at 2.80GHz, Linux cluster with 68 Dual-Core AMD Opteron™ Processors at 2.01 GHz.

Chapter 5

A New Validation Method for B Factor Distributions

This chapter describes the statistical assumption on which the validation method for B factor distributions is built. To test the validity of the new approach proposed here, a large data set of protein structures at high resolution was selected and analysed with the validation protocol. The results obtained from the analysis are then discussed.

5.1 The Statistical Rationale

In order to develop a validation protocol for B factor distributions, it is first necessary to identify a suitable distribution to be used as a reference in the validation process. In the attempt to find such a distribution, the model used to describe isotropic displacements in a crystal is related to some statistical concepts from Bayesian statistics.

5.1.1 The Frequentist View

From the definition of the isotropic B factor B_i for the i^{th} atom in a crystal structure with n atoms [126]

$$B_i = 8\pi^2\langle u_i^2\rangle, \tag{5.1}$$

it follows that

$$B_i = 8\pi^2\langle u_i^2\rangle = 8\pi^2 s_i^2 \approx 8\pi^2\sigma_i^2 \propto \sigma_i^2, \tag{5.2}$$

where $\langle u_i^2\rangle$ is the mean square displacement of the i^{th} atom around its rest position, s_i^2 and σ_i^2 are the corresponding sample variance and population variance, respectively. If the size of the sample tends to infinity then the sample variance converges to the population variance, as the case for many unit cells over the duration of the crystallographic experiment.

It should be noted that from now on the expression 'B factors' will imply isotropic B factors, unless differently specified. As shown in equation (5.2), the atomic B factor B_i of the i^{th} atom is, by definition, proportional to the population variance σ_i^2 of its atomic displacements. If $\boldsymbol{\sigma}^2$ is defined as the set of variances of atomic displacements in a protein model and $\boldsymbol{B}$ as the set of the corresponding B factors, it follows that any consideration about the nature of the distribution of $\boldsymbol{\sigma}^2$ is also valid for the distribution of $\boldsymbol{B}$.

Let us assume that it would be possible in a crystallographic experiment to record the individual atomic positions for all the atoms while they are vibrating and/or displaced around their rest position, and let $\boldsymbol{c}_{i,1}, \ldots, \boldsymbol{c}_{i,k}$ be the k hypothetical observations for the i^{th} atom in the protein model. For sake of convenience, each $\boldsymbol{c}_i$ stands for a three-dimensional vector with the $x_{i,j}, y_{i,j}$ and $z_{i,j}$ coordinates of the single observation with $j = 1, \ldots, k$.

In the Gaussian approximation, which is used for the isotropic refinement of B factors, the atomic displacements of the i^{th} atom are modelled with a uni-variate Gaussian distribution $N(\mu_i, \sigma_i^2)$. The mean parameter μ_i is usually set to zero since the Gaussian distribution is supposed to be centred at the rest position of the i^{th} atom, while the variance σ_i^2 is a parameter that needs to be estimated.

5.1.2 The Bayesian View

Following a Bayesian approach, the atomic displacements for the i^{th} atom in a protein model can be described by a Bayesian normal model with known mean μ_i and unknown variance σ_i^2. By means of Bayes' theorem, the distribution of the variance parameter σ_i^2 of the i^{th} atom given the $\boldsymbol{c}_i$ observation can be written as

$$\begin{aligned} &\text{posterior probability } \propto \text{prior probability } \times \text{ likelihood probability} \\ &\Rightarrow \quad f(\sigma_i^2|\boldsymbol{c}_{i,1}, \ldots, \boldsymbol{c}_{i,k}) \propto f(\sigma_i^2) \times f(\boldsymbol{c}_{i,1}, \ldots, \boldsymbol{c}_{i,k}|\sigma_i^2). \end{aligned} \tag{5.3}$$

Given that the distribution of the atomic displacements in an isotropic model is considered to be Gaussian, the Inverse-gamma distribution (IGD) can be chosen as a possible conjugate prior distribution for the σ_i^2 parameter (see section 3.2.5). It then follows that $f(\sigma_i^2) \sim$ IGD. Moreover, thanks to the definition of conjugacy (see section 3.2.3), the posterior distribution $f(\sigma_i^2|\boldsymbol{c}_{i,1}, \ldots, \boldsymbol{c}_{i,k})$ is expected to follow an IGD. According to the Gaussian model used to describe the atomic displacements in a protein model, it follows that for each atom the variance that describes the atomic disorder is expected to belong to a certain IGD.

In a first approximation the IGD that describes the distribution of the variances of atomic displacements $\boldsymbol{\sigma}^2$ in a protein model is assumed to be the same for all atoms. It should be noted that this does not mean that all the variances $\boldsymbol{\sigma}^2$ are identical, but simply that they

are considered to be independent and identically distributed according to a common IGD:

$$f(\boldsymbol{\sigma}^2) \sim \text{IGD}. \tag{5.4}$$

The fact that the unknown parameter σ_i^2 in the posterior probability $f(\sigma_i^2|\boldsymbol{c}_{i,1},\ldots,\boldsymbol{c}_{i,k})$ of equation (5.3) is not considered to be a constant value but instead an unknown, random variable is an intrinsic feature of Bayesian statistics.

Since any consideration about the nature of the distribution of $\boldsymbol{\sigma}^2$ is also valid for the distribution of $\boldsymbol{B}$, it follows that

$$f(\boldsymbol{B}) \sim \text{IGD}. \tag{5.5}$$

The idea that the distribution of isotropic B factors in a protein model follows an IGD has been previously proposed [38], but never used to validate the B factor distribution itself.

5.1.3 The Reference Distribution

In the previous paragraph the central assumption of the proposed validation method was that isotropic B factors in protein X-ray structures should follow an IGD (see equation (3.15)):

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$$

However, experimental B factor distributions are often systematically shifted towards higher B factor values. This can be due to the fact that the average B factor of a structural model is related to the resolution d_{min} to which the respective crystal diffracted, or it can be a consequence of not fully accounting for systematic error in data collection, such as those caused by X-ray absorption [38, 62]. To model such systematic shifts, a third parameter γ is introduced into the classic equation of the IGD and the result is a shifted inverse gamma distribution (see equation (3.20)):

$$f^*(x|\alpha,\beta,\gamma) = \frac{\beta^\alpha}{\Gamma(\alpha)} (x-\gamma)^{-\alpha-1} \exp\left(-\frac{\beta}{(x-\gamma)}\right)$$

hereafter named IGD* or IG*(α,β,γ), to distinguish it from the classic IGD.

Assuming that the IGD* is the reference distribution to be used for the validation of B factor distribution in protein structures, a protocol was designed and implemented in R language as described in section 4.3.

5.2 Comparison of Observed B Factor Distributions to IGD*

The search protocol described in section 4.1 was applied to retrieve a large data set of protein structures that fulfilled all the required criteria as of June 2010. The query resulted in 15998 crystal structures containing a total of 30441 protein chains with more than 500 non-H atoms. For sake of clarity, it should be acknowledged that the analyses based on full models of crystal structures, which can contain multiple chains, are denoted as *all-chains* and analyses based on individual chains as *single-chains*.

The validation protocol described in section 4.3 was then applied to both *all-chains* and *single-chains* data sets. For each of the 15998 protein structures and 30441 protein chains the protocol delivered the α, β and γ parameters of the IGD* that best fit the empirical B factor distribution and a p-value that expresses the goodness of fit. The outcome of the validation protocol is summarised by a graphical plot for inspection and analysis as shown in Figure 5.1.

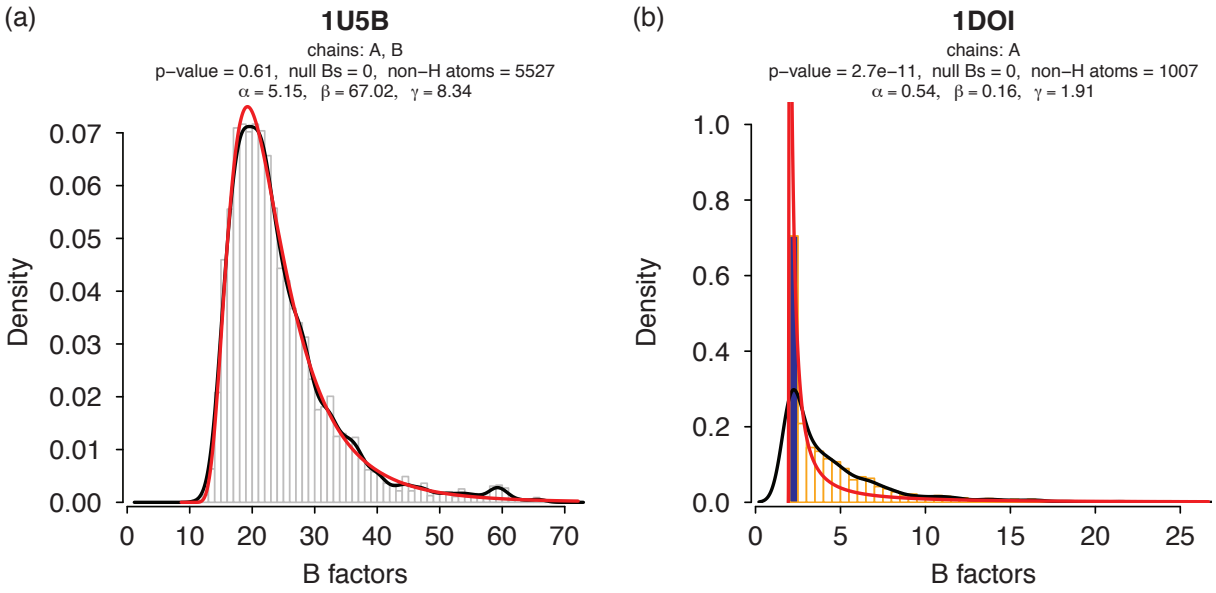


Figure 5.1: Example of graphical output from the validation protocol. Examples of *acceptable* (a) and *suspicious* (b) B factors distributions from pdb-entries 1U5B and 1DOI, respectively. Each plot produced by the validation protocol described in section 4.3 contains the following information. For comparison, the histogram of the observed B factors normalized to the total number of atoms, the curve representing its kernel density estimation (black line), and the density curve of the IGD* with the best fit to the observed B factors (red line) are represented together. The PDB code, the identifiers for the polypeptide chains under analysis, the p-value as derived from the KS-test, the number of B-values lower than or equal to zero, the total number of non-H atoms, and the α, β and γ parameters used for the calculation of the IGD* are noted at the top of the plot. In case the p-value is higher than or equal to 0.01, the B factor distribution is considered *acceptable* and the histogram is coloured grey (e.g. panel (a)); if instead the p-value is smaller than 0.01, the B factor distribution is considered *suspicious* and the histogram is coloured orange (e.g. panel (b)). Strong ties in the B-factor distribution are indicated by filled blue bars (e.g. panel (b)).

5.3 Analysis of α, β and γ IGD* Parameters

Estimates for α, β and γ parameters were determined for 15998 B factor distributions for entire crystal structures (*all-chains* data set) and for 30441 B factor distributions for individual peptide chains (*single-chains* data set) via MLE as described in section 4.3.1. The correlations between the IGD* parameters and some selected statistics were computed and a correlogram plot for the *all-chains* data set is shown in Figure 5.2. In this section the observed trends for the IGD* parameters are discussed.

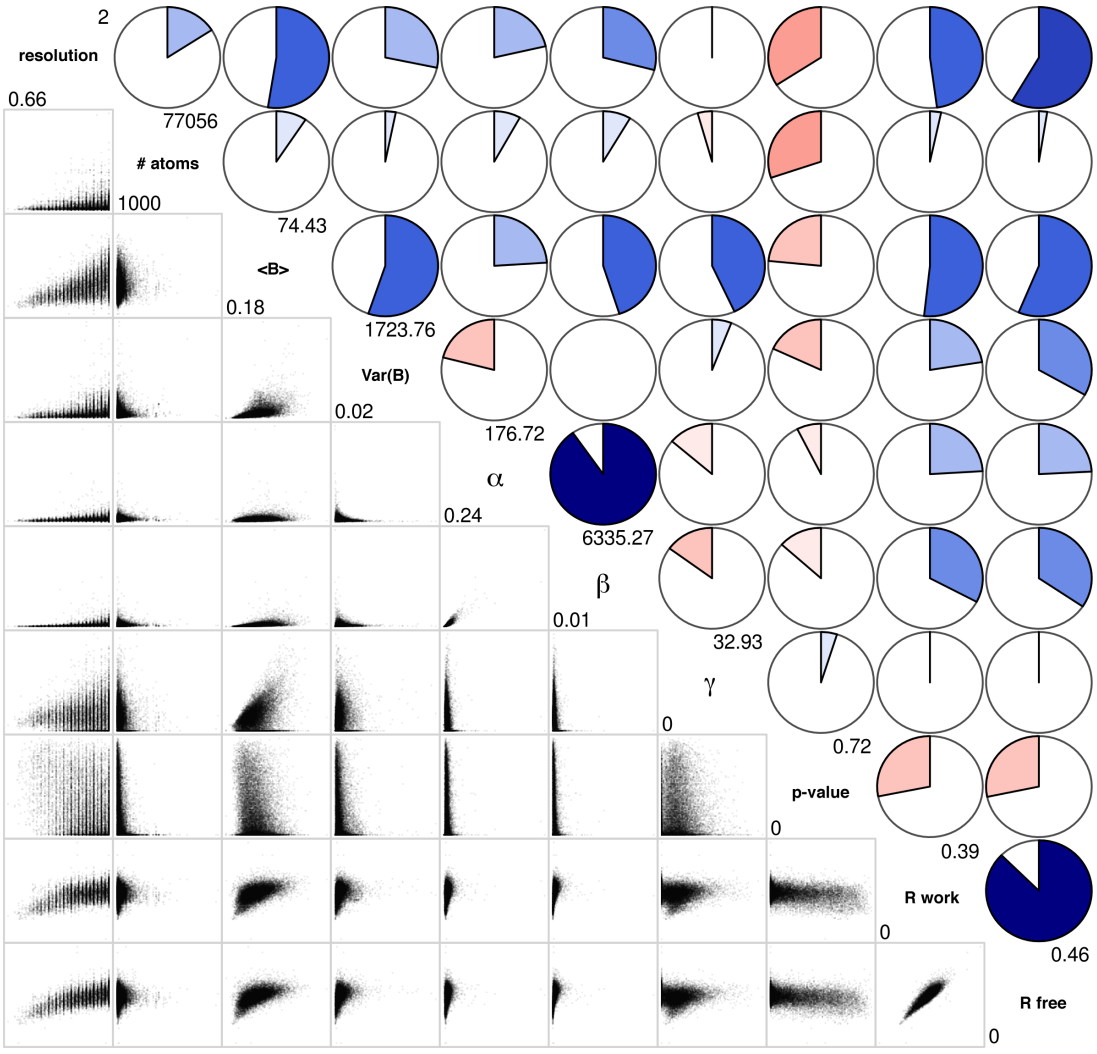


Figure 5.2: Correlogram for selected statistics from the *all-chains* data set. Along the diagonal the name of the statistic under analysis is reported. For each statistic the minimum and maximum values are shown on the lower-left and upper-right corners, respectively. On the lower matrix a scatter-plot representation is shown for each pair of statistics, while in the upper matrix a pie-chart representation is used to express the correlation between each pair of statistics. Blue and red colours indicate positive and negative correlation, respectively. A white and empty pie chart corresponds to a correlation of zero, while a full pie chart corresponds to a correlation of 1 (if blue) or -1 (if red). The intensity of the colours is proportional to magnitude of the correlation: greater the absolute value of the correlation, greater the intensity of the colour. Similar trends are observed for the *single-chains* data set (not shown).

The γ parameter is constrained by definition (see section 3.2.6) to assume values between 0.0 Å^2 and the minimum B factor found in the protein model under analysis. This gives rise to fitted γ parameters, varying between 0.0 Å^2 and 32.9 Å^2 for the *all-chains* data set and between 0.0 Å^2 and 45.6 Å^2 for the *single-chains* data set. From the analysis of the correlogram plot computed for some selected statistics from the protein data set (see Figure 5.2) it is possible to observe that the γ parameter is slightly positively correlated to the $\langle B \rangle$ of the structure (Pearson correlation coefficient of 0.43 and 0.39 for the *all-chains* and *single-chains*, respectively). Such positive correlation is in agreement with the fact that experimental B factor distributions are often systematically shifted towards higher B factor values as discussed in section 5.1.3.

Moving to the analysis of α and β IGD* parameters, they both vary over several orders of magnitude. The α parameter varies from 0.24 to 176.72 with mean 5.89 ± 3.95 for the *all-chains* data set and from 0.23 to 499.57 with mean 6.38 ± 5.39 for the *single-chains* data set, while the β parameter varies from 0.01 to 6335.27 with mean 96.15 ± 123.14 for the *all-chains* data set and from 0.01 to 10668.74 with mean 111.66 ± 170.19 for the *single-chains* data set. In addition, they are strongly correlated to each other, resulting in a Pearson correlation coefficients of 0.90 and 0.88 for *all-chains* and *single-chains* data sets, respectively. They both show also a slightly positive correlation with the maximum resolution of the deposited structures and with the $\langle B \rangle$ (see Figure 5.2). This is in agreement with the fact that structures at higher resolution are usually characterized by lower $\langle B \rangle$ as shown by the positive correlation between the resolution and the $\langle B \rangle$ in Figure 5.2.

The strong correlation between α and β parameters becomes clear if the two quantities are plotted against each other in a logarithmic scale as shown in Figure 5.3.

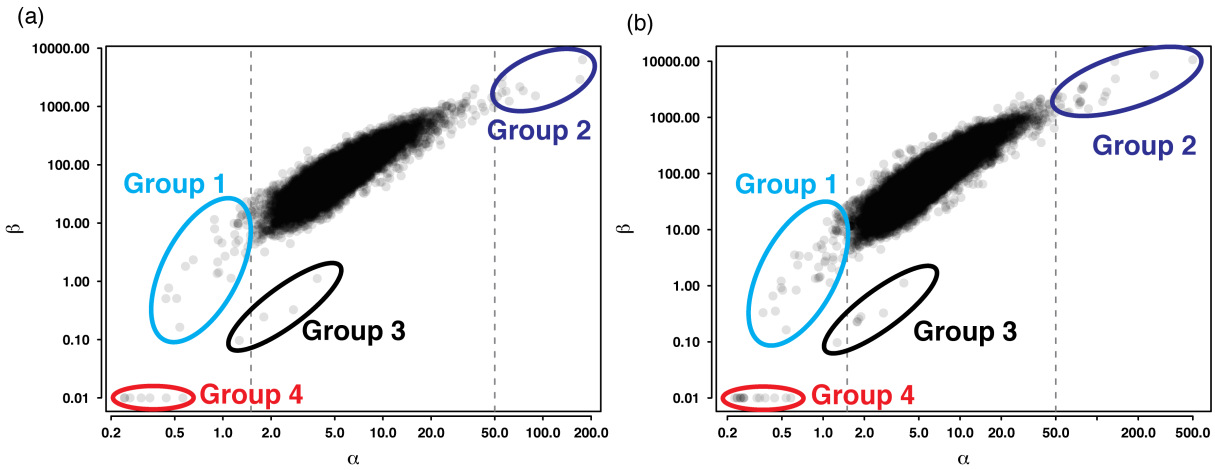


Figure 5.3: Distribution of the β IGD* parameter *versus* the α IGD* parameter for *all-chains* (panel (a)) and *single-chains* (panel (b)) data sets. A logarithmic scale is used for both abscissa and ordinate. Two vertical gray dashed lines denote the two empirical boundaries at 1.5 and 50 used for the detection of outliers in function of the α parameter and four different groups of outliers are highlighted with a coloured ellipse. For further details please refer to text in section 5.3.

From the conditions of existence of equation (3.20), in principle any combination of positive α and β parameters is allowed. However, in protein structures such strong correlation is somehow expected for the following reasons.

In equation (3.21), the mean value μ and the variance σ^2 of the underlying IGD* can be approximated by the observed mean B value $\langle B \rangle$ and the observed variance s_B^2, respectively:

$$\alpha = \frac{(\mu-\gamma)^2}{\sigma^2} + 2 \approx \frac{(\langle B \rangle - \gamma)^2}{s_B^2} + 2 \propto \frac{\langle B \rangle - \gamma}{s_B}. \tag{5.6}$$

Given that the range for the mean B values is limited to positive and physically reasonable values and that the variances are consequently limited to a defined range as well, the variation in the possible values for α is also limited. Besides, as shown in Figure 5.2 the mean and variance of B factors in protein structures at high resolution show a positive correlation that is propagated in equation (5.6). In addition, from equation (3.22) the β parameter is related to the α parameter by a constant factor $(\mu-\gamma)$ that in turn can assume only a limited range of values:

$$\beta = (\mu-\gamma)(\alpha-1) \approx (\langle B \rangle - \gamma)(\alpha-1) \propto \alpha. \tag{5.7}$$

The advantage of such strong linear correlation between α and β is that it is usually sufficient to use only one of them (i.e. α) for the detection of *suspicious* B factor distributions in terms of α and β IGD* parameters. Here, for practical purposes in the distribution of α and β parameters of Figure. 5.3, two empirical boundaries for low and high α parameters are defined respectively at 1.5 and 50. All the structures whose B factor distributions are fitted by a IGD* with an α parameter lower than 1.5 or higher than 50 are considered outliers or *suspicious*. This results in 48 and 11 structures for the *all-chains* data set and in 78 and 24 structures for the *single-chains*, respectively. From the analysis of the localization of these outliers in the distribution of α and β parameters it is possible to identify three different groups (groups 1, 2 and 4 in Figure 5.3). A fourth group is identified in a set of four pdb-entries whose α parameter is in the allowed range (except one), but the β parameter is much lower than expected (group 3 in Figure 5.3). A discussion on the composition of these groups for the *all-chains* data set follows.

5.3.1 Outliers, Group 1

This group, highlighted by a cyan ellipse in Figure 5.3, contains a total of forty protein structures with B factor distributions usually characterized by a strong tie at the lowest B factor value.

The pdb-entry 2YYH is an illustrative example of outlier from this group (see Fig-

ure 5.4).

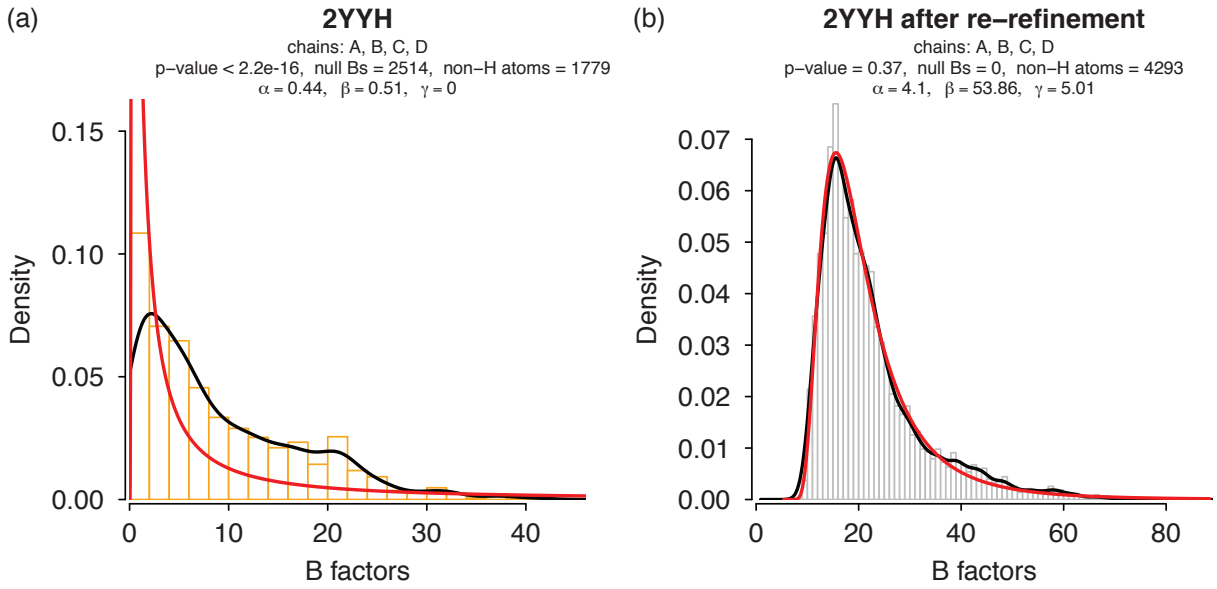


Figure 5.4: Example of an outlier structure from group 1: analysis of the pdb-entry 2YYH as taken from the PDB (a) and after re-refinement (b). For a description of the plot generated by the validation protocol please refer to Figure 5.1.

The deposited structure (panel (a) in Figure 5.4) is an extreme example of a protein model with a large number of B factors equal to zero (2514 out of 4293). It should be noted that even if B factors equal to zero are not taken directly into consideration during the MLE procedure (since they are physically meaningless and contradict the existence conditions of the IGD*), their presence in the the model greatly affects the overall distribution of the remaining B factors in the model, resulting in a p-value lower than 0.01 and thus in a *suspicious* distribution. To see if a standard refinement protocol is sufficient to recover the distribution of B factors, the refinement protocol described in section 4.2 was applied to the deposited 2YYH model and the distribution of B factors from the re-refined model is shown in panel (b) of Figure 5.4. After re-refinement all B factors in the model were observed to be higher than zero and the overall distribution of B factors was labeled as *acceptable*, with a p-value from the KS-test of 0.37. It is also worth noting that the α and β parameters increased from 0.44 and 0.51 to 4.1 and 53.86 respectively, in agreement with the trend of the majority of the α and β parameters computed from the protein data set and shown in Figure 5.3. No significant difference was observed in R factors from the default and the re-refined structure. The R_{work} and R_{free} went from 0.21 and 0.25 to 0.19 and 0.24, respectively.

Another example of a *suspicious* structure from this group is the pdb-entry 1DOI shown previously in panel (b) of Figure 5.1. Also in this case a strong tie is present at the lowest B factor value (214 out of 1007 atoms have a B factor of 2 Å^2).

For both pdb-entries 2YYH and 1DOI the p-value from the KS-test was enough to flag their B factor distributions as *suspicious*. This is due to the fact that usually a strong tie at the lowest B factor value significantly affects the estimation of the IGD* that best fits the empirical distribution of B factors by lowering the values of α and β parameters. Furthermore, the observation that a default re-refinement protocol was sufficient to fix the distribution of B factors for the pdb-entries 2YYH (see Figure 5.4) and 1DOI (not shown) can be explained by the fact that strong ties localized at the lowest B factor value in the deposited structures are usually artefacts introduced in the refinement stage before the deposition of the structures.

5.3.2 Outliers, Group 2

This group, highlighted by a blue ellipse in Figure 5.3, contains a total of eleven structures samples with narrow and symmetric B factor distributions characterized by very high α and β parameters. The reason for such narrow distributions is at the moment not clear, but it is suspected that during the refinement of these structures the B factors were too tightly restrained.

The pdb-entry 1XT8 is an extreme example of an outlier from this group (see Figure 5.5). In the *all-chains* data set it is the structure with the highest α and β IGD* parameters, respectively equal to 176.72 and 6335.27.

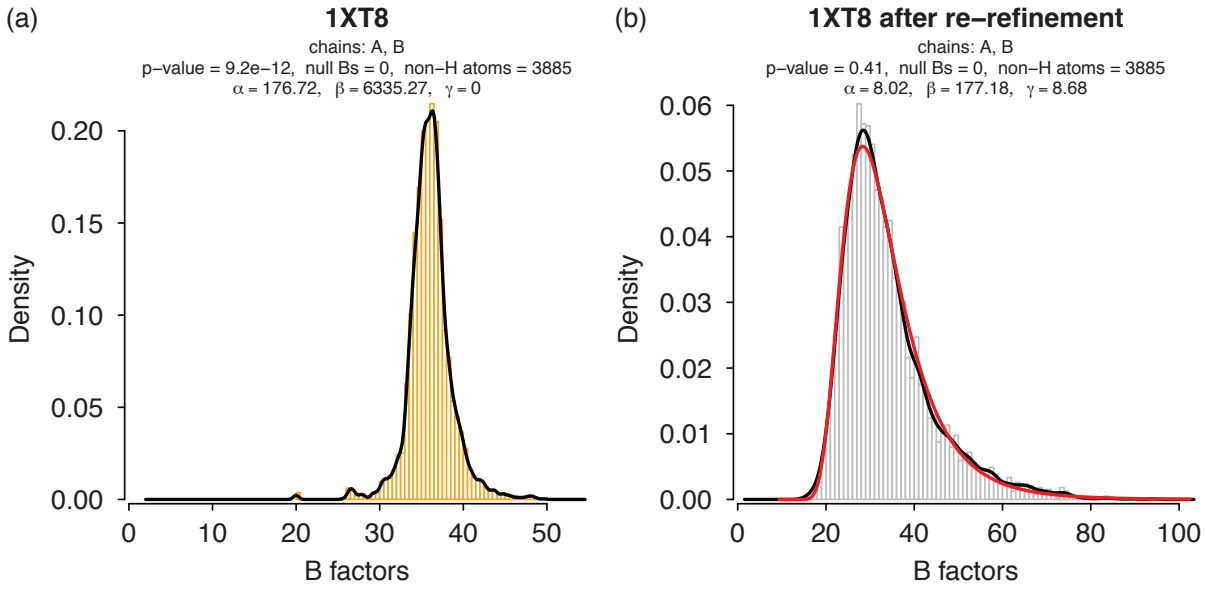


Figure 5.5: Example of outlier structure from group 2: analysis of PDB code 1XT8 as taken from the PDB (a) and after re-refinement (b). For a description of the plot generated by the validation protocol please refer to Figure 5.1.

The deposited model (panel (a) in Figure 5.5) is the structure with the highest α and

β IGD* parameters in the *all-chains* data set with values equal to 176.72 and 6335.27, respectively. Similar to section 5.3.1, the deposited model was re-refined with the refinement protocol described in section 4.2 and the distribution of B factors from the re-refined model is shown in panel (b) of Figure 5.5. Also in this case the distribution of B factors became *acceptable* after re-refinement and the α and β parameters lowered from 176.72 and 6335.27 to 8.02 and 177.18, respectively. It should be noted that the gap between R_{free} and R_{work} slightly increased after re-refinement. In fact, the R_{work} and R_{free} factors went from 0.18 and 0.23 to 0.18 and 0.25, respectively.

5.3.3 Outliers, Group 3

This group, highlighted by a black ellipse in Figure 5.3, contains four structures (PDB codes: 1H7R, 4GCR, 1IDS and 1PPO) for which the α parameters lie in the acceptable range between 1.5 and 50 (except for the pdb-entry 1PPO, with $\alpha = 1.27$), but the β parameters are systematically shifted to lower values if compared to the expected trend in Figure 5.3. The common feature of the four structures is that their B factors are very low, ranging between 0 and 2.5 $Å^2$.

A careful analysis of the four structures reveals that they were all refined using the refinement program RESTRAIN [44], which has an option to write mean square displacements $\langle u^2 \rangle$ instead of B factors to the pdb-file containing the refined model. Apparently this option had been used to produce the model deposited in the PDB. In fact, multiplying the numbers given in the B factor column of the respective pdb-files by a constant of $8\pi^2$ and repeating the analysis, the obtained α and β parameters fall in the expected range, except for the pdb-entry 1PPO (data not shown). This is explained by the fact that a multiplicative constant applied to the B factors affects only the β parameter of the fitted IGD* (a derivation of how a scaling of B factors by a multiplicative constant is reflected in the estimated α and β parameters can be found in appendix C).

A representative example for this group is the pdb-entry 4GCR (see Figure 5.6).

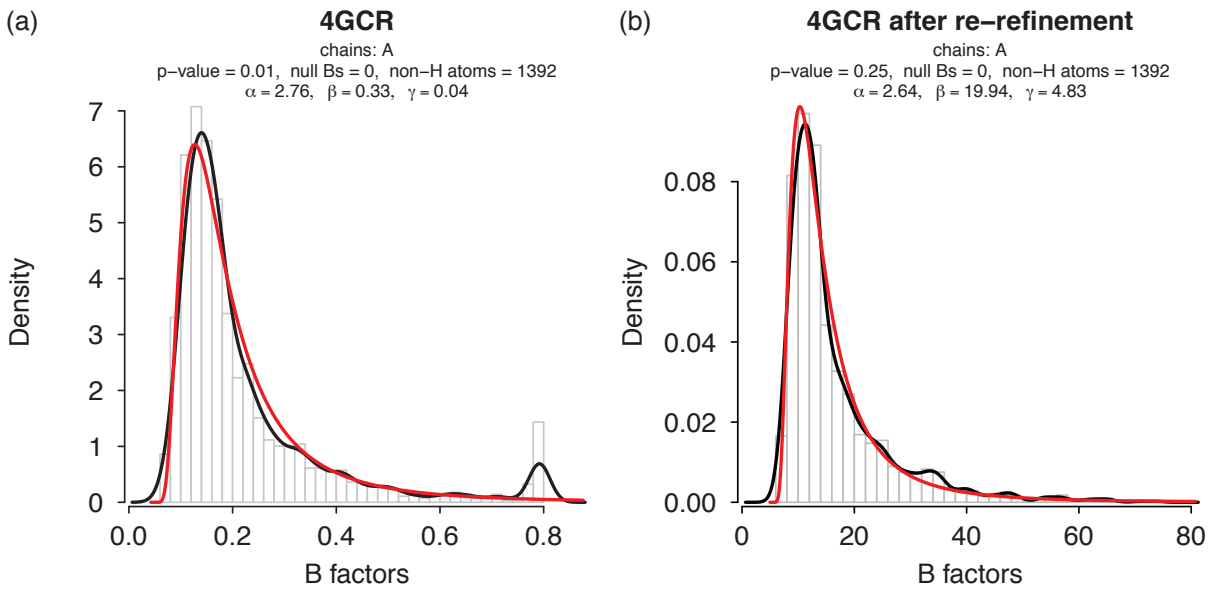


Figure 5.6: Example of an outlier structure from group 3: analysis of the pdb-entry 4GCR as taken from the PDB (a) and after re-refinement (b). For this structure $\langle u^2 \rangle$ values were erroneously stored instead of B values. If the $\langle u^2 \rangle$ in (a) are converted to B factors by multiplying them by a constant of $8\pi^2$, the p-value from the KS-test applied to the obtained B factors is 0.011 (data not shown). The fact that the p-value did not significantly change ($0.01 \simeq 0.011$) is expected, as shown in appendix C. For a description of the plot generated by the validation protocol please refer to Figure 5.1.

Panel (a) in Figure 5.6 shows the distribution of $\langle u^2 \rangle$ values for the deposited 4GCR model. The mean square displacement values lie in the range between 0.06 and 0.80 Å^2 resulting in a p-value for the agreement between the fitted IGD* and the observed $\langle u^2 \rangle$ values of 0.01 and thus an *acceptable* distribution. When the 49 atoms with $\langle u^2 \rangle$ higher than 0.75 Å^2, giving rise to the second maximum in the B factor distribution are removed from the model, the β parameter still remains rather small (0.45) while the p-value increases to 0.023 (data not shown) indicating that, in principle, the agreement to the fitted IGD* is good, despite the fact that the $\langle u^2 \rangle$ values are much too small if compared to the expected B factors. Also in this case the deposited model was re-refined with the refinement protocol described in section 4.2 and the distribution of B factors from the re-refined model is shown in panel (b) of Figure 5.6. After re-refinement the B factors lie in a reasonable range between 6.09 Å^2 and 73.87 Å^2, resulting in an *acceptable* distribution with a p-value from the KS-test of 0.25. Since the deposited structure factors did not contain the *free* set it is difficult to say if the R factor statistics improved or worsened. Nevertheless, after re-refinement the R_{work} and R_{free} went from 0.18 and *na* to 0.16 and 0.21, respectively.

The standard re-refinement protocol was applied also to the other three protein models in this group. For two of them (PDB codes: 1H7R and 1IDS) the re-refinement protocol significantly altered the B factor distributions resulting in p-values > 0.25 throughout. For 1PPO the refinement protocol failed, indicating a serious problem with the model and/or

the diffraction data.

5.3.4 Outliers, Group 4

This group, which is highlighted by a red ellipse in Figure 5.3, includes seven structures (PDB codes: 3B8D, 3ENV, 3KRG, 3ENW, 1EPT, 1YJB and 3F3R) that share the feature of giving rise to very low β parameters of 0.01 (the lower limit allowed for this parameter by the MLE-procedure).

The corresponding B factor distributions all exhibit strong ties at a value of 2 Å^2 which corresponds to the lower limit imposed by several refinement programs, indicating problems with the parameterisation of the models. The number of atoms with a B factor of 2 Å^2 are: 2600 out of 11004 atoms ($\sim$ 24%) for 3B8D, 841 out of 3204 atoms ($\sim$ 26%) for 3ENV, 666 out of 3055 atoms ($\sim$ 22%) for 3KRG, 1363 out of 3204 atoms ($\sim$ 43%) for 3ENW, 774 out of 1526 atoms ($\sim$ 51%) for 1EPT, 863 out of 1934 atoms ($\sim$ 45%) for 1YJB, 995 out of 1592 atoms ($\sim$ 63%) for 3F3R.

As an example of an outlier structure from this group, the B factor distribution for the pdb-entry 3B8D is shown in Figure 5.7

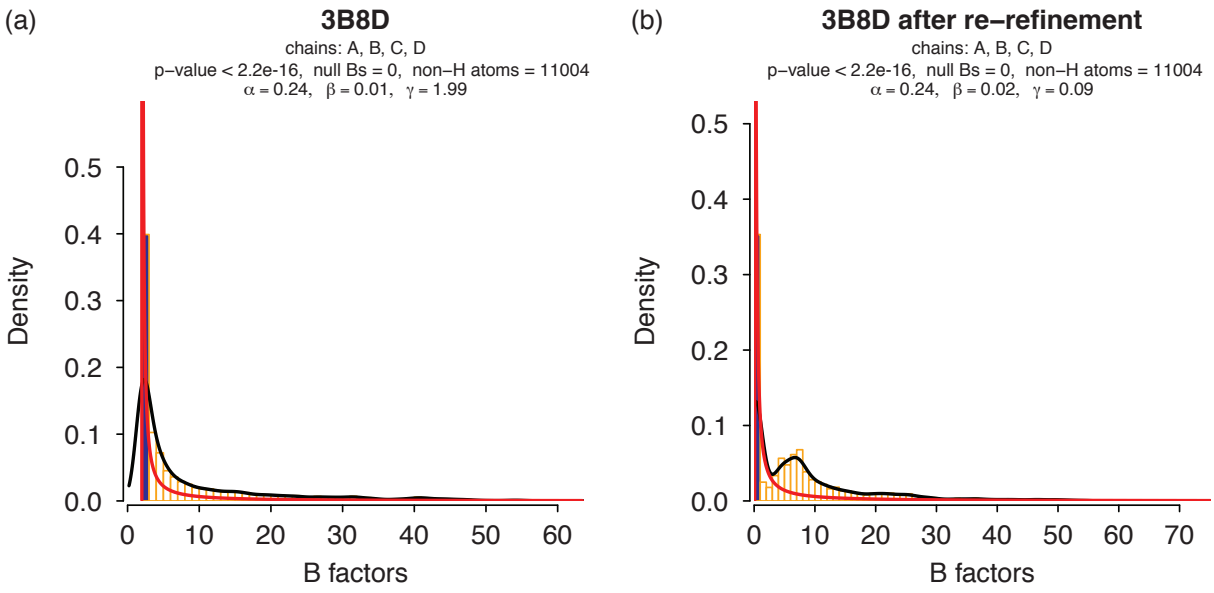


Figure 5.7: Example of an outlier structure from group 4: analysis of pdb-entry 3B8D as taken from the PDB (a) and after re-refinement (b). For a description of the plot generated by the validation protocol please refer to Figure 5.1.

For four out of the seven structures (PDB codes: 3ENV, 3ENW, 3F3R and 1DOI), the standard refinement protocol described in section 4.2 was successful to bring α and β parameters into the expected range. Three models were not rescued by the re-refinement

procedure (as shown in panel (b) of Figure 5.7 for the pdb-entry 3B8D). These models had a very low B_{wil} including one case (PDB code 1YJB) with a negative B_{wil} of -4.0 Å^2, as determined by the *phenix.xtriage* program [1]. This clearly indicates problems with the diffraction data (e.g. errors in detector calibration or incorrect absorption corrections being applied to the data).

5.4 Agreement Between B Factor Distributions and the IGD*

After estimation of the α, β and γ parameters of the IGD* that best fit the B factor distributions for the 15998 structures of the *all-chains* data set and the 30441 chains in the *single-chains* data set, a parametric bootstrapped two-sample two-sided KS-test was applied to evaluate the goodness of fit between the observed B factor distributions and the estimated IGD*, as described in section 4.3.2. This resulted in 15998 p-values from the *all-chains* data set and 30441 p-values from the *single-chains* data set.

5.4.1 Distribution of P-values

As described in section 4.3.2, the null hypothesis H0 adopted in the KS-test states that the experimentally determined B factors follow a IGD* whose parameters are obtained via MLE, while the alternative hypothesis is its negation (i.e. the the experimental B factors do not follow a IGD*). A significance level of 0.01 (1%) was used for the interpretation of the outcome from the KS-test. If the p-value obtained from the KS-test is smaller than 1% the null-hypothesis is rejected, indicating that it is highly improbable that the observed distribution is well represented by the estimated IGD*. This corresponds to what in statistics is called a *positive* case and the B factor distribution under analysis is labelled as *suspicious*. If instead the p-value is greater than or equal to 1% the null-hypothsis is not rejected. This corresponds to what in statistics is called a *negative* case and the B factor under analysis is labelled as *acceptable*.

For 12694 out of 15998 crystal structures from the *all-chains* data set, the p-value is equal to or greater than 0.01, indicating that for the majority of structures inspected (i.e 79%) the null-hypothesis should not be rejected and the corresponding B factor distributions are considered *acceptable*. This observation became even more pronounced when the distributions for individual chains were inspected. In this case, 26962 out of 30441 single chains analysed from the *single-chains* had a p-value higher than or equal to 0.01, leading to 89% of the structures having B-factor distributions that could be satisfactorily described by a IGD* and then considered *acceptable* (see Figure 5.8).

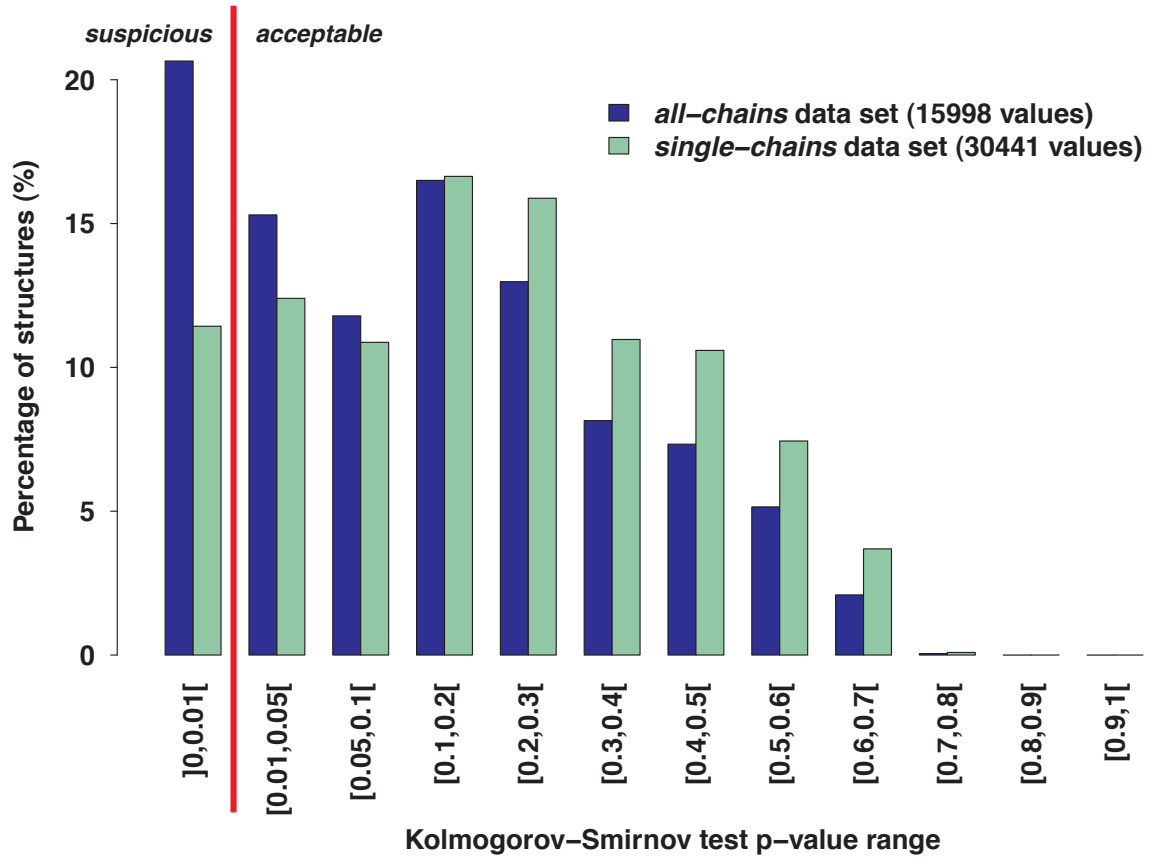


Figure 5.8: Distribution of p-values for *all-chains* and *single-chains* data sets. Blue and green colours are used to distinguish the *all-chains* data set from the *single-chains* data set, respectively. A vertical red line is used to highlight the boundary between the *suspicious* structures (to the left) and the *acceptable* structures (to the right).

5.4.2 Some Observations about Sample Size Bias

It should be noted that, during the analysis of the distribution of p-values from the protein data set, the KS-test showed a different stringency in function of the sample size. Even if a correction for the size of the samples is present in the KS-test statistic with the multiplicative factor $\sqrt{(m \times n)/(m+n)}$ (see equation 3.31), there are cases in which the test seems too *permissive*, especially with samples of small size, and other cases in which the test seems too *strict*, especially with samples of large size (see Figure 5.9 for some indicative examples). This can give rise to false negatives in the first case (see panel (d) in Figure 5.9) or false positives (see panel (a) in Figure 5.9). It must be noted anyway that all these cases lie in the category of B factor distributions with a p-value of 0.01. Such significance level should be considered a sort of "twilight zone" and all the structures with a p-value close to 0.01 should be checked manually. Complementarily, the orthogonal statistics discussed in section 5.5 can be useful for the detection of false negatives.

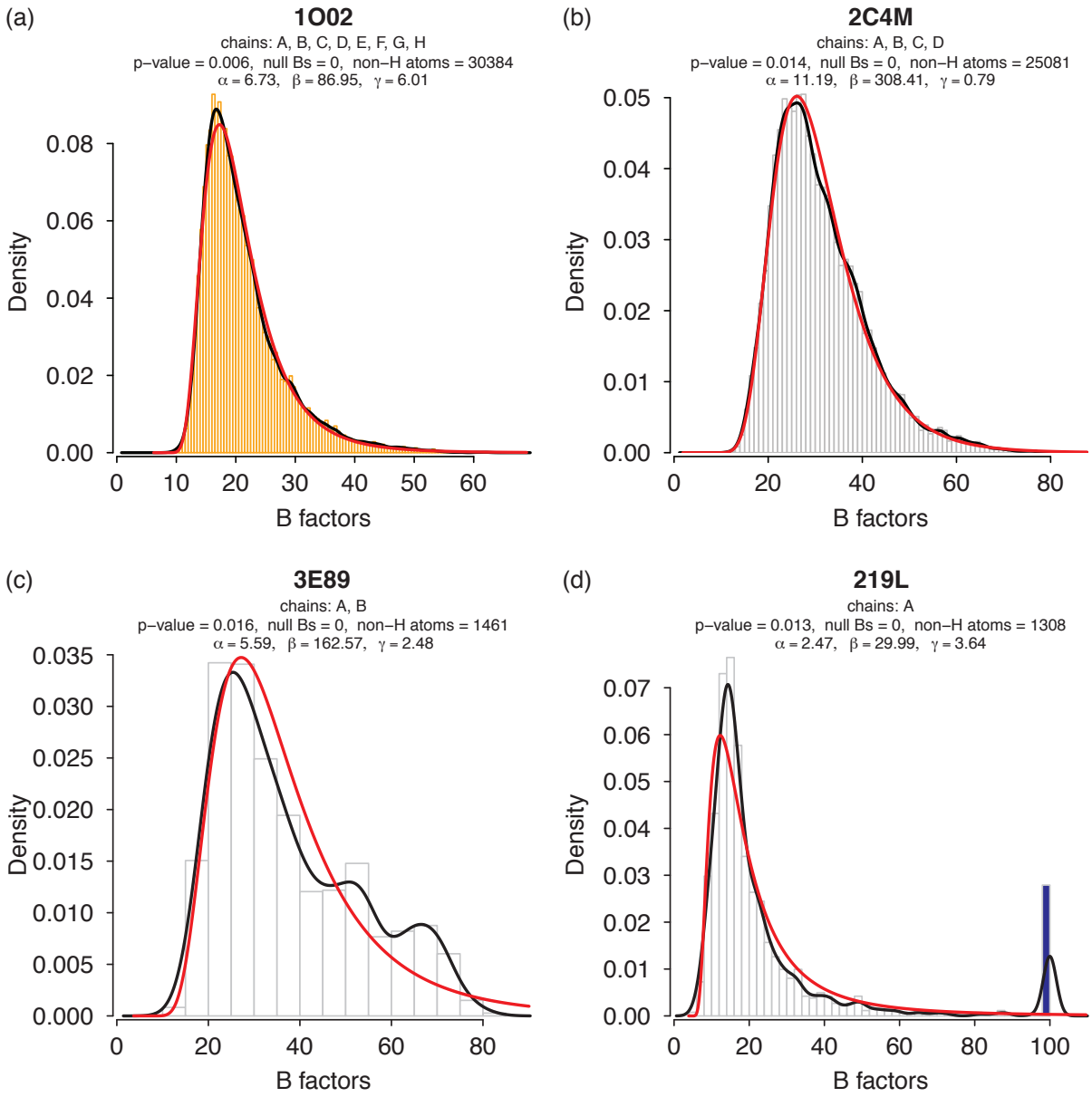


Figure 5.9: Examples of the effect of the population size on the outcome of the KS-test. In panels (a) and (b) two examples of large B factor samples of 30384 and 25081 atoms, respectively; in panels (c) and (d) two examples of small B factor samples of 1461 and 1308 atoms, respectively. It is worth noting that all four distributions have a p-value close to the significance level 0.01: the B factor distribution from the pdb-entry 1O02 is considered *suspicious* since its p-value is lower than 0.01, while the B factor distributions from the pdb-entries 2CM4, 3E89 and 219L are considered *acceptable* since their p-values are all higher than 0.01. However the qualitative agreement between the empirical distributions and the fitted IGD* greatly changes when moving from large (panels (a) and (b)) to small (panels (c) and (d)) B factor samples.

The fact that the stringency changes with the size of the sample is an intrinsic property of statistical tests. This has also a logical basis: by randomly drawing a low number of points from a IG*(α, β, γ) it is quite probable that the empirical distribution will look quite different from the expected one and usually it will be affected by multimodality. Instead, when drawing a larger number of points, the empirical distributions will get closer to the

expected distribution. This is intrinsically related to the random process that is used to draw the points.

For practical purposes one could decide to use different significance levels, depending on the size of the sample of B factors under analysis, in order to finely tune the outcome of the KS-test. It should be noted that this would be a completely pragmatic approach without any support from the statistical theory. It would be an attempt to find a compromise between theory and real life but at the risk of breaking the statistical assumptions behind hypothesis testing.

Since some approximations were already introduced to adapt the KS-test for the validation of B factors (see section 4.3.2), for the analysis of the p-values from the KS-test only one significance level of 0.01 was used.

5.4.3 Refinement Programs Distribution

The distribution of p-value was studied as a function of the refinement program used to refine the structures in the *all-chains* data set to see if any interesting trend was detectable (see Figure 5.10). For this analysis only those refinement programs for which at least more than one hundred structures are available in the data were taken into considerations. This resulted in the selection of the following refinement programs: REFMAC (6607 structures), CNS (5535 structures), SHELXL (1194 structures), X-PLOR (1124 structures), TNT (609 structures), PHENIX (391 structures), PROLSQ (175 structures) and CNX (166 structures). Together they account for 98.8% of the total number of structures in the *all-chains* data set. From the analysis of the p-value distribution in panel (a) of Figure 5.10 it emerges that SHELXL and PHENIX depart mostly from the general trend shown by the other refinement programs. Their p-value median are centered respectively at 0.29 and 0.20, while the p-value median for the other programs is centered around 0.1 (REFMAC at 0.11, CNS at 0.09, X-PLOR at 0.11, PROLSQ at 0.10, CNX at 0.11). The reason for such difference, especially for SHELXL, becomes clear if the distribution of resolution is taken into consideration (panel (b) in Figure 5.10). It fact, protein structures that were refined with SHELXL are on average characterized by a higher maximum resolution (1.4 Å) when compared to structures refined with the other refinement programs ($\sim$ 1.7 Å). This is reflected also by the lowest percentage of *suspicious* structures with a p-value lower than 0.01 (84 out of 609, corresponding to 7%). For PHENIX the average resolution of the structures (1.8) is not as high as for SHELXL. Nevertheless the percentage of *suspicious* structures (47 out of 391, corresponding to 12%) is lower than the average of *suspicious* structures present in the data sets ($\sim$20%) refined with the remaining refinement programs. This observation can be partially biased by the rather small size of the data set since it contains only 391 structures.

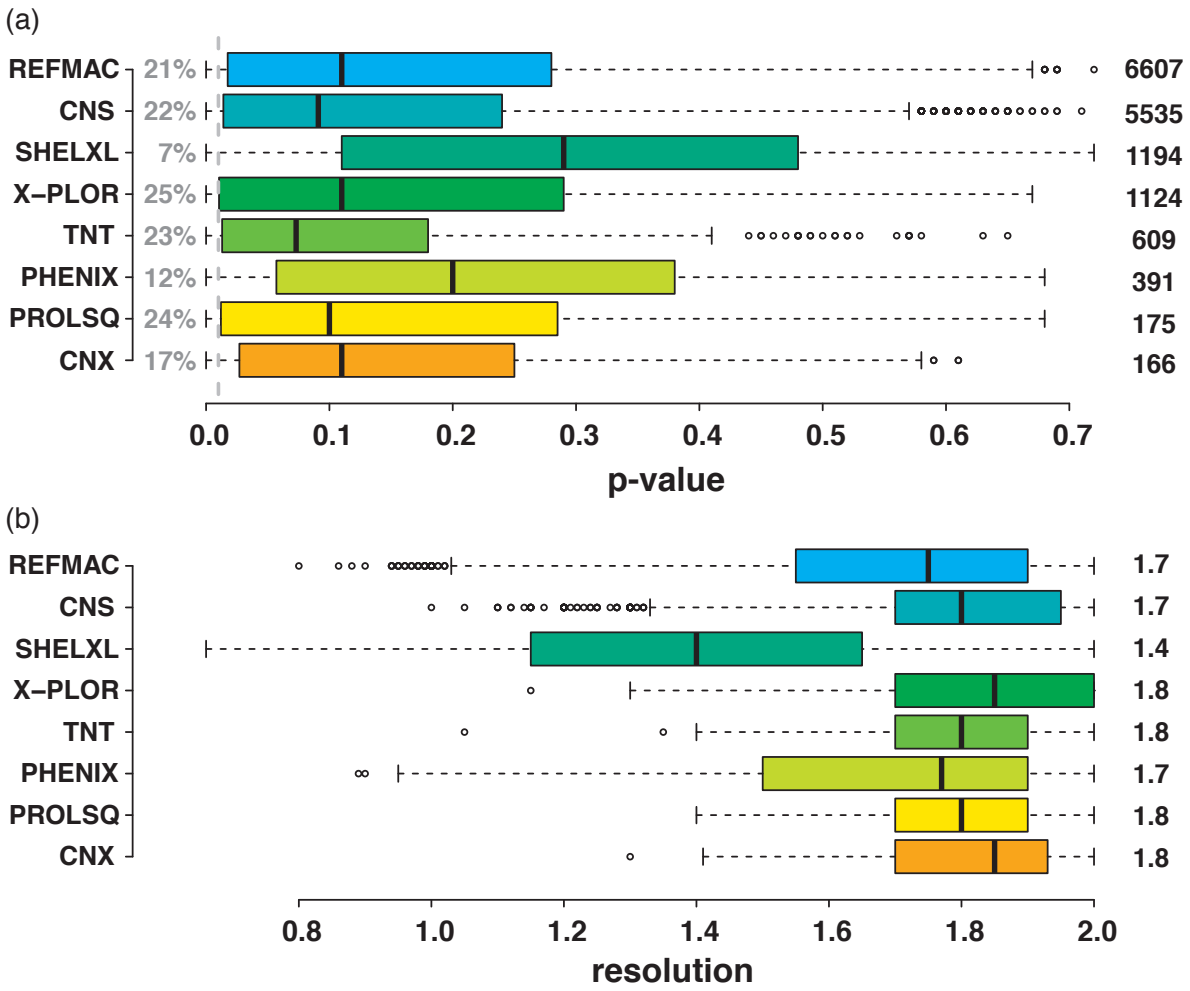


Figure 5.10: Distribution of p-value (panel (a)) and resolution (panel (b)) as a function of the refinement program used to refine the protein structures in the *all-chains* data set. On the right side of the distribution of p-value (panel (a)) the total amount of structures for each program is reported in black digits, while on the left side the relative percentage of structures with a p-value lower than 0.01 is reported in gray digits. On the right side of the distribution of resolution (panel (b)) the average resolution is reported for each refinement program in black digits.

5.4.4 Multimodality 1, Chain Level

After dealing with the structures which are outliers in terms of α and β IGD* parameters (see section 5.3), there were still many structures which gave rise to a bad agreement between observed B factor distributions and derived IGD* as measured by the p-value (see Figure 5.8). Of particular interest is the observation that, when moving from the analysis of the *all-chains* data set to the analysis of the *single-chains* data set, the percentage of *acceptable* structures increases from 79% to 89%. Looking at the distribution of B factors of those structures from the *all-chains* data set with a p-value lower than 0.01 it became clear that the majority of them were affected by a strong multimodality. Nevertheless, for some of them, each single chain in the asymmetric unit satisfied the IGD* assumption if anal-

ysed independently, showing a p-value higher than 0.01. This explains the difference in percentage of *suspicious* structures shown in Figure 5.8 between the *all-chains* data set and the *single-chains* data set. The presence of multimodal B factor distributions is physically reasonable as often different molecules in the same asymmetric unit are in a different packing environment, giving rise to different global and/or local disorder/mobility, resulting in systematically different B factor distributions.

An extreme example from this category of structures is the pdb-entry 2R8Y [16] (see Figure 5.11). The asymmetric unit of this structure contains 16 molecules of YrbI phosphatase from *E. coli* containing 188 amino acid residues each. If the B factor distribution from all chains (21549 atoms) is analysed the resulting p-value is lower than 0.01 and thus the distribution is considered *suspicious* (see panel (b) in Figure 5.11). If the model is divided into 16 chains of ca. 1350 atoms each, all 16 chains exhibit p-values equal or higher than 0.01 and then they are all considered *acceptable* (see panels (d) and (f) in Figure 5.11 and in Table 5.1, default columns). Similar results are obtained if the model is re-refined with the protocol described in section 4.2 (see Table 5.1, re-refined columns).

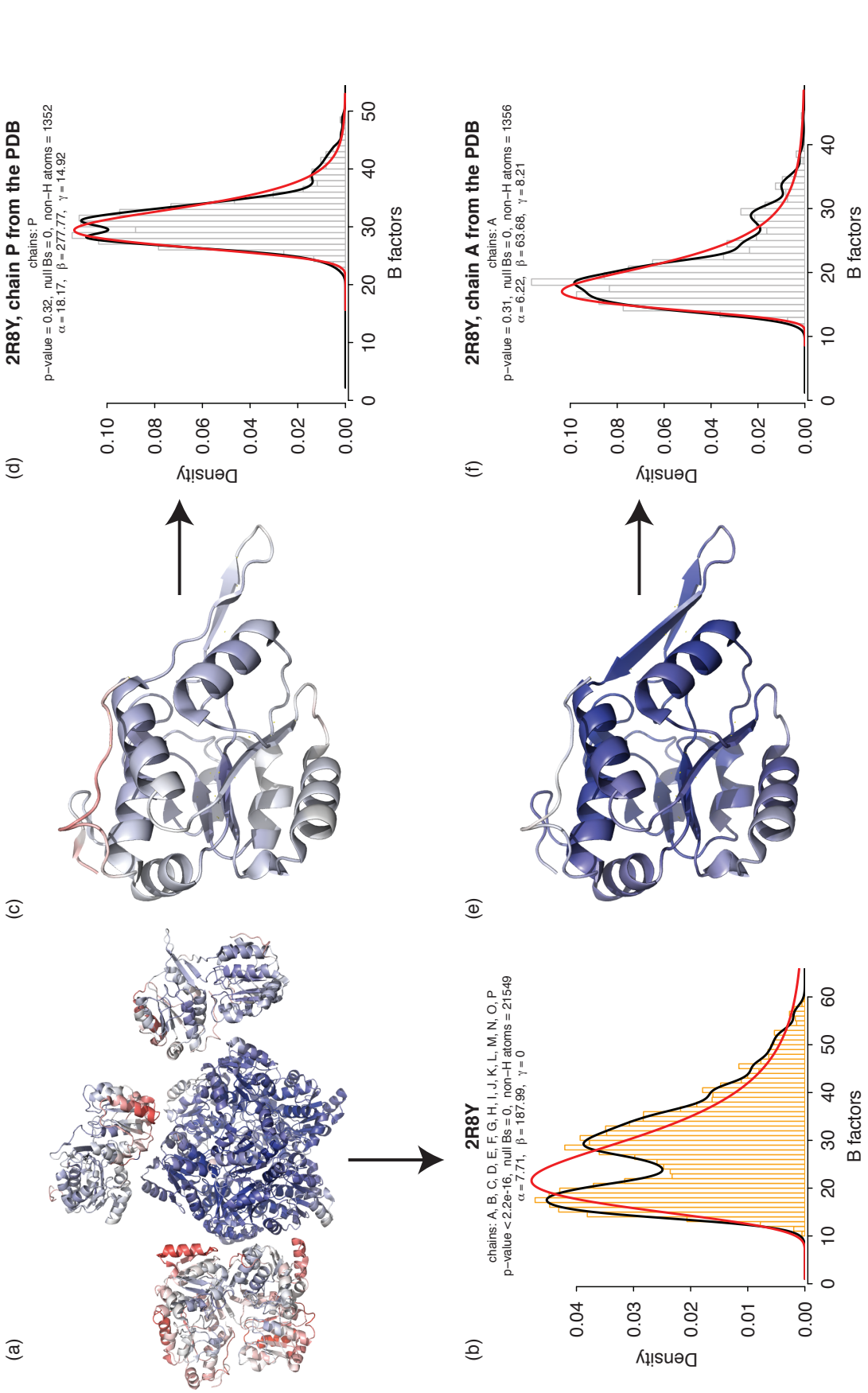


Figure 5.11: Example of multimodal B factor distribution which can be resolved at chain level. Cartoon representation of the PDB model 2R8Y (a) in the asymmetric unit and its B factor distribution (b) if all chains are taken into account. If a colour gradient based on B factor values is applied, it is possible to identify two different sub-populations of B factors: one colder (dark-blue) and one warmer (light-blue). (e),(f) Example of a "cold" chain. (c),(d) Example of a "warm" chain. For a description of the plot generated by the validation protocol please refer to Figure 5.1.

chain	default				re-refined			
	α	β	γ	p-value	α	β	γ	p-value
A	6.22	63.68	8.21	0.31	4.75	51.42	8.10	0.54
B	10.83	120.11	6.26	0.21	6.06	62.53	7.35	0.18
C	9.38	105.61	6.40	0.51	5.46	57.26	7.57	0.23
D	8.70	132.42	5.55	0.15	5.24	72.28	7.28	0.65
E	7.32	85.81	7.74	0.37	5.37	66.83	7.30	0.55
F	7.63	89.40	6.66	0.30	5.11	56.89	7.54	0.57
G	9.56	110.38	5.27	0.31	5.44	56.75	6.73	0.30
H	8.81	120.03	6.13	0.25	5.43	69.02	7.33	0.62
I	10.52	180.84	16.05	0.07	7.11	136.06	13.91	0.48
J	9.26	182.47	14.15	0.03	6.16	125.33	13.18	0.33
K	32.54	1281.22	0.00	0.03	11.56	331.59	9.68	0.50
L	27.94	1099.99	0.69	0.10	11.22	338.42	8.84	0.60
M	14.76	313.45	11.16	0.24	8.06	159.95	12.40	0.57
N	17.04	285.10	11.94	0.42	9.55	166.16	11.30	0.30
O	24.90	820.73	0.61	0.01	8.85	202.27	9.95	0.46
P	18.17	277.77	14.92	0.32	10.38	195.73	11.43	0.07
A-P	7.71	187.99	0.00	< 2.2e-16*	6.94	173.08	0	< 2.2e-16*

Table 5.1: IGD* statistics for single chains from the pdb-entry 2R8Y. For each of the 16 chains present in the asymmetric unit the α, β and γ IGD* parameters are reported, together with the p-value from the KS-test for the default structure and for the re-refined structure. The last row refers to the whole deposited model. P-values lower than 0.01 are highlighted with and asterisk '*' symbol.

5.4.5 Re-refinement of *Suspicious* Structures

Since the standard refinement protocol described in section 4.2 was found to be sufficient to rescue the majority of those structures that were considered outliers for the distribution of α and β parameters (see section 5.3), the same refinement protocol was used to re-refine 2255 structures, which based on their p-value were classified as *suspicious* for the *all-chains* data set and *suspicious* for at least one chain in the *single-chains* data set.

For 1959 structures the refinement protocol converged, while for the remaining 296 structures it was not successful, indicating problems with the diffraction data and/or the deposited model. The R factor statistics before and after the re-refinement procedure follow the same general trend observed in similar comparative analysis available in the literature [65, 66] (data not shown). The distribution of p-values for the resulting models is shown in Figure 5.12. 69% and 89% of the *all-chains* and *single-chains* cases respectively resulted in p-values higher than 0.01 and could therefore be re-classified as *acceptable*. This is a quite impressive result if it is considered that before re-refinement all the structures in the selected subset were considered *suspicious*.

Also in this case, the percentage of *suspicious* structures lowers while moving from the all-chains data set to the *single-chains* data set as previously observed for the complete data sets (see Figure 5.8). As expected, such difference in percentage was due to multimodal

B factor distributions that could be decomposed into unimodal B factor distributions if the single chains were taken into consideration.

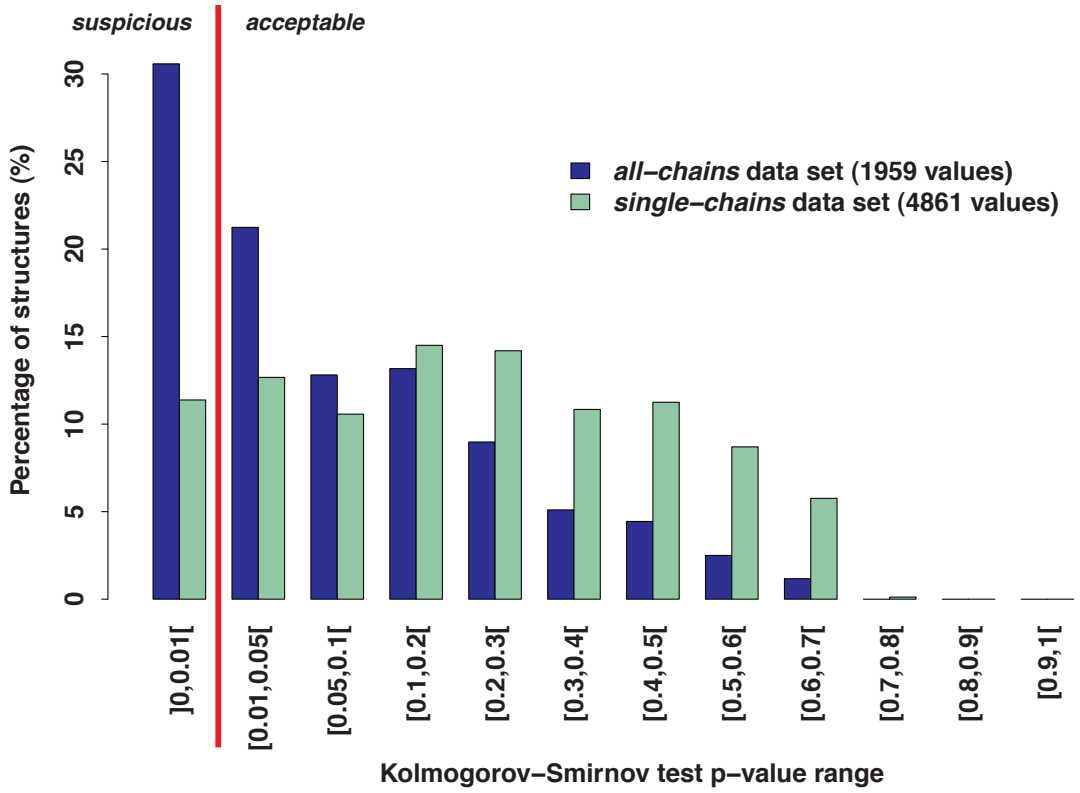


Figure 5.12: Distribution of p-values for *all-chains* and *single-chains* data sets after re-refinement of a selected subset of *suspicious* structures. Blue and green colours are used to distinguish the *all-chains* data set from the *single-chains* data set, respectively. A vertical red line is used to highlight the boundary between the *suspicious* models (to the left) and the *acceptable* models (to the right).

5.4.6 Multimodality 2, Domain Level

However, even if the re-refinement procedure rescued the distribution of B factors for around 70% of the structures in the *all-chains* data set, there were still 11% of *suspicious* B factor distributions in the *single-chains* data set (see *suspicious* structures in Figure 5.12). Looking at these distributions it became clear that also in the *single-chains* data set it was possible to identify some strong multimodal distributions. They usually corresponded to structures where it is possible to fit a IGD* to the B factors from individual domains of the protein.

An extreme example for this class of structures is the GroEl complex (PDB code 1KP8 [131]) where the whole complex and all the single-chains in the deposited model have a p-value close to zero (see Table 5.2, default columns). If the structure of the complex is visualized in a cartoon representation and coloured according to the atomic B-factors, it is

possible to notice a different distribution of B factors among the three different domains composing each protein chain (data shown only for the re-refined model, see below). The B factors seem in fact to follow a gradient. From low values in the equatorial domain to higher values in intermediate and apical domains. In particular, for all chains the equatorial domain the B factors seem to be colder than in the other two domains. The distribution of B factors was then analysed at single domain level and it was observed that the equatorial and the intermediate domains usually satisfy the IGD assumption. However, none of the apical domains comply to the IGD*. Is should be noted that a gradient similar to the one observed for the B factors was detected also for the α and β parameters of the IGD* fitted on the B factor distributions of each domain. The average α and β parameters are equal to 8.18 ± 1.02 and 191.63 ± 40.95 for the equatorial domains, 36.04 ± 16.12 and 3007.89 ± 1622.81 for the intermediate domains, 62.43 ± 39.83 and 7612.66 ± 5749.09 for the apical domains, respectively (see Table 5.2, default columns).

After re-refinement, a strong multimodality was still detected on the B factor distribution at both *all-chains* and *single-chains* levels. However, differently from the deposited model, all the individual domains satisfied the IGD* assumption (see Table 5.2, re-refined columns) and then could be classified as *acceptable*. The gradient of B factors values from the equatorial to the apical domains was still noticeable (see Figure 5.13, panels (e) and (f)). The same holds for the distribution of α and β parameters from the IGD* fitted to each domain. After re-refinement the average α and β parameters are equal to 6.57 ± 0.45 and 154.51 ± 22.08 for the equatorial domains, 15.74 ± 5.83 and 1128.54 ± 614.97 for the intermediate domains, 22.92 ± 7.44 and 2377.65 ± 1060.90 for the apical domains, respectively (see Table 5.2, re-refined columns). From these results it seems that the magnitude of the α and β parameters is related to the optical resolution of the domains that in turn is related to the quality of the electron density map and thus to the magnitude of the B factors.

The fact that the multimodal behaviour at *all-* and *single-chain* level did not disappear after re-refinement means that the observed multimodality is an intrinsic property of the distribution of B factors of the structure and it is related to the disorder due to the mobility of the domains in the crystal lattice. It is in fact known that the GroEl complex undergoes conformational changes during its activity [131]. The same behaviour was observed in other structures where it is possible to identify in the same chain more than one domain with different level of disorder. It would be then possible to improve the statistics by identifying among the *suspicious* structures those where each single domain satisfies the IGD assumption. However, unless an automated procedure is implemented, this step must be conducted manually.

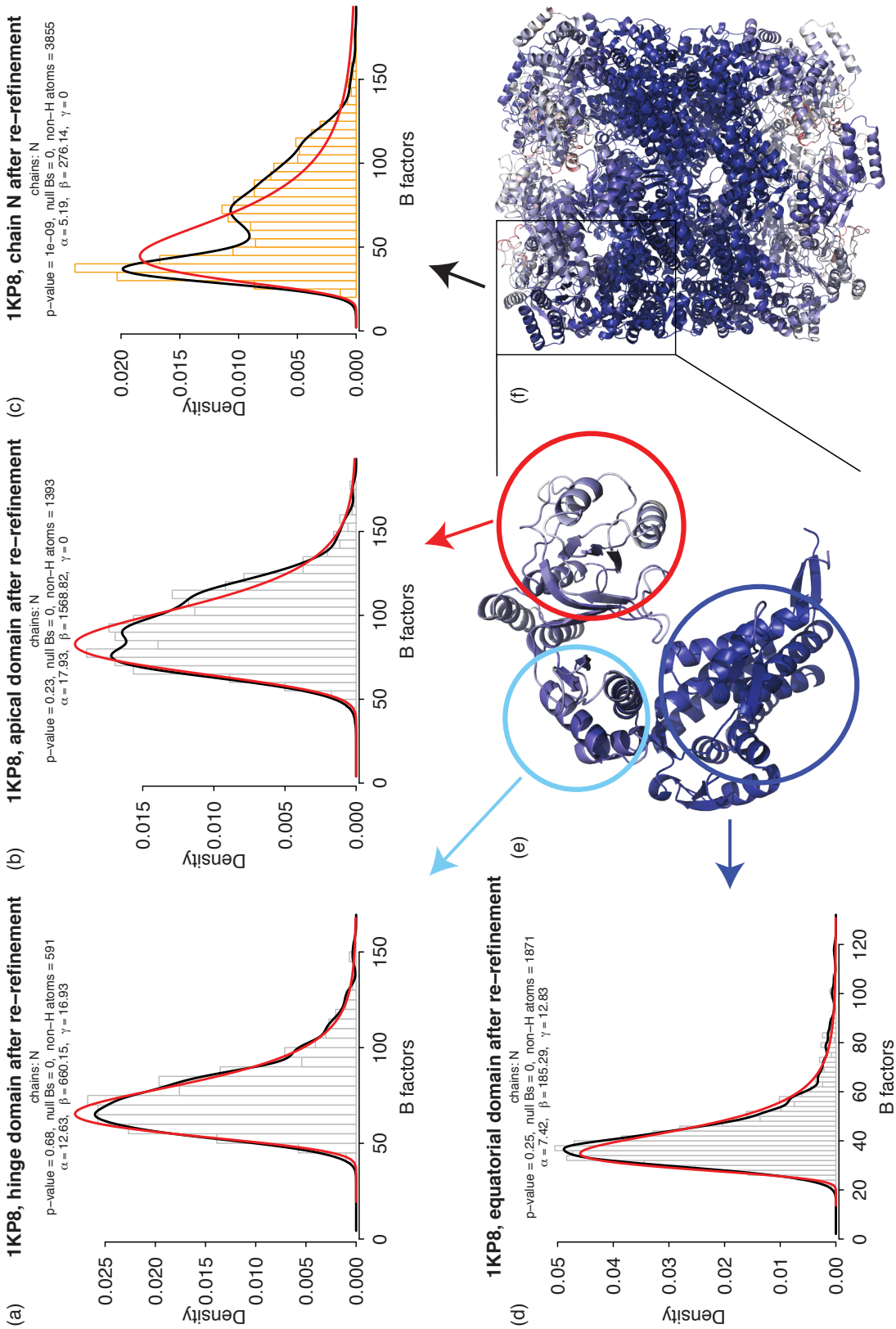


Figure 5.13: Example of a multimodal B factor distribution which can be resolved at domain level. In panels (f) and (e) the cartoon representation for the whole GroEL complex and the chain N is shown, respectively. A sort of gradient of disorder corresponding to increasing values of B factors is visible (the color system used is the same as in Figure 5.11).In panels (a), (b) and (d) the results of the B factor validation protocol for the equatorial, hing and apical domain from chain N, respectively are shown. The definition of the three domains was taken from the accompanying publication [131]: equatorial domain (residues 2–136 and 410–525), intermediate domain (residues 137–188 and 378–409) and apical domain (residues 189–377). In panel (c) the same is shown for the whole chain N. For a description of the plot generated by the validation protocol please refer to Figure 5.1.

chain	domain	default				re-refined			
		α	β	γ	p-value	α	β	γ	p-value
A	equatorial	8.68	207.84	15.23	0.60	6.79	162.92	12.28	0.25
	intermediate	27.88	2240.54	0.00	0.27	16.27	1167.85	0.00	0.57
	apical	45.25	4631.00	0.00	2.3e-04*	20.97	1909.09	0.00	0.41
	all	4.67	267.19	0.00	< 2.2e-16*	4.52	236.79	0.00	1.1e-12*
B	equatorial	6.76	131.59	14.86	0.27	5.97	125.42	11.04	0.10
	intermediate	25.43	2029.86	0.00	0.14	16.56	1191.99	0.00	0.62
	apical	69.37	9026.66	0.00	1.1e-05*	25.72	2973.30	0.00	0.25
	all	1.87	71.48	11.46	< 2.2e-16*	2.49	112.29	3.77	< 2.2e-16*
C	equatorial	8.06	203.29	15.18	0.20	6.51	164.01	11.66	0.57
	intermediate	42.36	3782.38	0.00	0.01	19.50	1560.82	0.00	0.59
	apical	102.62	12976.23	0.00	4.7e-03*	29.90	3326.68	0.00	0.19
	all	3.65	224.64	0.00	< 2.2e-16*	3.54	197.45	0.00	< 2.2e-16*
D	equatorial	10.01	233.49	12.19	0.16	6.53	135.71	11.35	0.04
	intermediate	10.31	439.66	12.98	0.10	6.67	249.04	13.02	0.25
	apical	16.30	1181.78	0.00	1.2e-03*	9.02	517.60	5.94	0.10
	all	4.65	175.56	8.3	2.0e-05*	4.34	152.29	6.71	5.2e-03*
E	equatorial	6.74	130.15	13.52	0.35	5.93	119.74	10.41	0.27
	intermediate	23.65	1691.67	0.00	0.07	10.87	625.07	7.45	0.41
	apical	52.64	6031.11	0.00	2.2e-04*	23.72	2391.02	0.00	0.26
	all	2.37	96.84	7.49	< 2.2e-16*	3.01	136.99	1.2	< 2.2e-16*
F	equatorial	7.93	193.19	13.57	0.21	6.68	166.93	10.26	0.43
	intermediate	29.24	2497.76	0.00	0.02	19.15	1439.93	0.00	0.61
	apical	66.03	8621.81	0.00	2.7e-03*	25.76	2985.99	0.00	0.17
	all	2.77	146.91	4.33	< 2.2e-16*	3.18	171.51	0.00	< 2.2e-16*
G	equatorial	7.43	142.83	16.32	0.25	7.00	154.83	11.05	0.02
	intermediate	9.72	399.74	17.26	0.16	6.10	214.54	16.54	0.29
	apical	14.71	1164.91	0.00	1.3e-03*	10.50	729.44	0.00	0.07
	all	3.51	122.45	11.72	4.0e-07*	4.00	144.91	7.23	2.6e-04*
H	equatorial	7.54	154.76	16.51	0.60	6.33	136.59	13.18	0.12
	intermediate	27.76	1888.23	0.00	0.08	9.27	432.51	15.52	0.58
	apical	36.94	3621.96	0.00	4.1e-04*	21.44	1857.73	0.00	0.20
	all	3.87	180.79	5.67	< 2.2e-16*	4.72	234.55	0.08	7.9e-11*
I	equatorial	6.94	156.35	18.19	0.21	5.94	135.10	14.92	0.54
	intermediate	50.95	4289.64	0.00	0.24	21.24	1624.24	0.00	0.52
	apical	60.22	6824.89	0.00	8.7e-04*	26.99	2699.66	0.00	0.24
	all	4.39	265.76	0.00	< 2.2e-16*	4.33	240.38	0.00	< 2.2e-16*
J	equatorial	8.66	212.69	18.73	0.26	6.27	149.28	16.00	0.19
	intermediate	55.89	5052.82	0.00	0.07	23.67	1931.64	0.00	0.58
	apical	104.70	12916.52	0.00	2.3e-03*	32.05	3454.02	0.00	0.32
	all	4.29	276.42	0.00	< 2.2e-16*	4.28	251.74	0.00	< 2.2e-16*
K	equatorial	9.06	239.41	17.52	0.19	6.77	171.48	15.06	0.52
	intermediate	54.78	5348.94	0.00	0.04	21.84	1890.20	0.00	0.63
	apical	161.62	22356.28	0.00	3.4e-03*	33.38	4093.37	0.00	0.22
	all	3.43	218.20	2.58	< 2.2e-16*	3.64	222.67	0.00	< 2.2e-16*
L	equatorial	8.40	200.36	16.32	0.29	6.71	161.12	13.09	0.40
	intermediate	46.47	3762.17	0.00	0.14	15.02	938.69	10.72	0.66
	apical	30.73	3319.30	0.00	9.1e-04*	16.09	1530.19	0.00	0.16
	all	4.40	225.08	0.74	< 2.2e-16*	4.35	234.76	0.00	2.6e-10*
M	equatorial	8.63	231.30	15.79	0.41	7.13	194.66	12.38	0.51
	intermediate	44.15	4298.20	0.00	0.03	21.60	1872.84	0.00	0.46
	apical	77.02	10285.01	0.00	6.4e-06*	27.45	3250.13	0.00	0.10
	all	3.59	233.71	0.00	< 2.2e-16*	3.60	215.27	0.00	< 2.2e-16*
N	equatorial	9.64	245.52	15.06	0.40	7.42	185.29	12.83	0.25
	intermediate	55.94	4388.81	0.00	0.21	12.63	660.15	16.93	0.68
	apical	36.82	3619.73	0.00	2.4e-07*	17.93	1568.82	0.00	0.23
	all	5.35	308.75	0.00	< 2.2e-16*	5.19	276.14	0.00	1.0e-09*
A-N	all	3.38	172.99	4.56	< 2.2e-16*	3.75	192.26	0.83	< 2.2e-16*

Table 5.2: IGD* statistics for single domains from the pdb-entry 1KP8. For each of the 14 chains the α, β and γ IGD* parameters are reported together with the p-value from the KS-test for each domain and for the whole chain. The last row refers to the whole deposited model. P-values lower than 0.01 are highlighted with and asterisk '*' symbol.

5.4.7 Hierarchical Agreement to the IGD*

From the results obtained so far it is possible to state that the IGD* assumption is satisfied at different hierarchical levels as summarized in Figure 5.14. As shown in Figure 5.8 for the retrieved data set of protein structures around 79% of the *all-chains* data set satisfy the IGD* assumption. The agreement increases to 89% if *single-chains* are taken into consideration. If the remaining 11% of *suspicious* structures for the *single-chains* data set is re-refined (see section 5.4.5), around 89% of the re-refined structures satisfy the IGD* assumption at chain level (see Figure 5.12). This result increases the total percentage of agreement to a IGD* by around 9% (89% of 11%). The remaining 2% (11% of 11%) are instead *suspicious* cases characterized by multimodal distribution of B factors or for which the distribution of B factors is in disagreement with the IGD* assumption, indicating problems in the model or in the deposited data.

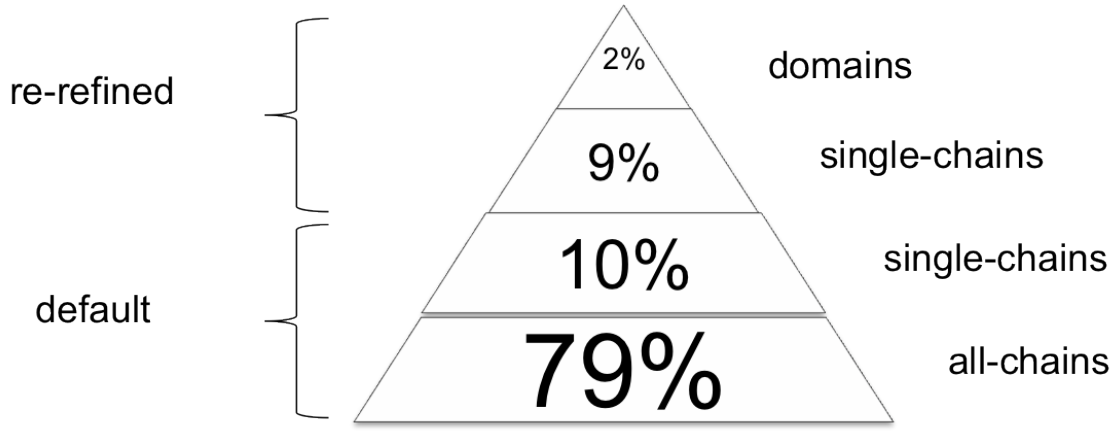


Figure 5.14: Hierarchical organization of the agreement to the IGD* assumption in protein structures at high resolution. The cumulative percentage of structures that satisfy the IGD* is shown for *all-chains* and *single-chains* data sets. Please note that the percentages are used interchangeably between the *all-chains* and *single-chains* data sets. For an explanation of the graph please refer to text in section 5.4.7.

From a combined analysis of the results obtained for the *all-chains* and *single-chains* data sets, four different categories of deposited structures can be identified, depending on the outcome of the KS-test from the *all-* and *single-chains* data sets:

1. Structures that are *acceptable* for both *all-* and *single-chains* data sets.
2. Structures that are *acceptable* for *all-chains* data set and *suspicious* for at least one chain in the *single-chains* data set.
3. Structures that are *suspicious* for *all-chains* data set and *acceptable* for all individual chains in the *single-chains* data set.

4. Structures that are *suspicious* for *all-chains* data set and at least one chain in the *single-chains* data set.

Category 1 is populated by structures made of only one chain or if they contain multiple chains they are homogeneously packed in the crystal lattice. Category 2 is populated by those structures that contain multiple chain and one or more of them are individually *suspicious*, but when analysed together they are masked by those chains that satisfy the IGD* (an example will be shown in chapter 6). An alternative scenario is the one where all the chain are *suspicious*, but they complement each other when analysed together. Category 3 is mainly populated by those structures that show a multimodal B factor distribution at chain level. These usually correspond to crystal structures where the level of order and packing inside the crystal lattice is different for the individual chains, as shown in section 5.4.4 for the pdb-entry 2R8Y. Category 4 is mainly populated by two different subgroups of structures. The first subgroup contains those structures that are affected by serious artefacts in the B factor distribution. The second subgroup contains instead those structures that show a multimodal B factor distribution at domain level. These usually correspond to multidomain crystal structures where the individual domains have a different packing order in the crystal lattice, as shown in section 5.4.6 for the pdb-entry 1KP8.

5.5 Common Features of Suspicious Structures

Detailed inspection of the structures marked as *suspicious* identified some common pathologies of these structures. In Tables 5.3 and 5.4 the distribution of outliers for ties at the minimum or maximum B factor, null B factors and extremely low (< 1.5) or high (> 50) α IGD* parameter are reported in function of p-value ranges (the same used in Figures 5.8 and 5.12).

p-value range	# total	# ties min(B)	# ties max(B)	# null B	# $\alpha < 1.5$	# $\alpha > 50$	# p-value
]0, 0.01[	3304	193	158	30	47	6	2918
[0.01, 0.05[	2448	24	105	4	1	1	2313
[0.05, 0.1[	1886	8	92	3	0	0	1783
[0.1, 0.2[	2640	10	85	5	0	2	2538
[0.2, 0.3[	2077	6	39	2	0	1	2029
[0.3, 0.4[	1304	5	15	2	0	1	1281
[0.4, 0.5[	1172	3	11	2	0	0	1156
[0.5, 0.6[	824	3	6	2	0	0	813
[0.6, 0.7[	335	0	1	4	0	0	330
[0.7, 0.8[	8	0	0	0	0	0	8
[0.8, 0.9[	0	0	0	0	0	0	0
[0.9, 1]	0	0	0	0	0	0	0

Table 5.3: Distribution of outliers for orthogonal statistics and α parameter values for *all-chains* data set. Column content: '# total' stands for total number of structures in the selected p-value range; '# ties min(B)' stands for number of structures with a strong tie at the lowest B factor value; '# ties max(B)' stands for number of structures with a strong tie at the highest B factor value; '# null B' stands for number of structures with one or more B factors equal to or lower than zero; '# $\alpha < 1.5$' stands for number of structures whose fitted IGD* has an α parameter lower than 1.5; '# $\alpha > 50$' stands for number of structures whose fitted IGD* has an α parameter higher than 50; '# p-value' stands for number of structures selected solely by the p-value.

p-value range	# total	# ties min(B)	# ties max(B)	# null B	# $\alpha < 1.5$	# $\alpha > 50$	# p-value
]0, 0.01[	3479	239	251	50	77	13	2922
[0.01, 0.05[	3775	44	139	2	0	6	3584
[0.05, 0.1[	3309	17	110	3	0	0	3179
[0.1, 0.2[	5066	21	111	6	0	3	4925
[0.2, 0.3[	4835	20	57	4	1	1	4752
[0.3, 0.4[	3340	16	23	5	0	1	3295
[0.4, 0.5[	3223	8	18	3	0	0	3194
[0.5, 0.6[	2264	6	16	2	0	0	2240
[0.6, 0.7[	1122	0	5	4	0	0	1113
[0.7, 0.8[	28	0	0	0	0	0	28
[0.8, 0.9[	0	0	0	0	0	0	0
[0.9, 1]	0	0	0	0	0	0	0

Table 5.4: Distribution of outliers for orthogonal statistics and α parameter values for *single-chains* data set. For a description of the content of each column please refer to Table 5.3.

It is worth noting that the significance level of 0.01 is sufficient to detect automatically the majority of the structures whose B factor distributions result to be outliers for the orthogonal statistics (null B factors, and strong ties) and extreme α parameter values. In detail the p-value range between 0 and 0.01 contains around 77% (193 out of 252 structures for the *all-chains* data set) and 64% (239 out of 371 structures for the *single-chains* data set) of the outliers for strong ties at the lowest B factor value, around 31% (158 out of 512 structures for the *all-chains* data set) and 34% (251 out of 730 structures for the *single-chains* data set) of the outliers for strong ties at the highest B factor value, around 56% (30 out of 54 structures for the *all-chains* data set) and 63% (50 out of 79 structures for the *single-chains* data set) of the outliers for null B factors (equal to or lower than zero), around 98% (47 out of 48 structures for the *all-chains* data set) and 99% (77 out of 78 structures for the *single-chains* data set) of the outliers for α IGD* parameter lower than 1.5, around 55% (6 out of 11 structures for the *all-chains* data set) and 54% (13 out of 24 structures for the *single-chains* data set) of the outliers for α IGD* parameter higher than 50. For cases with a p-value lower than 0.01, the orthogonal statistics provide further information on the possible problems that caused the B factor distributions to be considered *suspicious*.

However there are still some outliers for orthogonal statistics that are characterized by a p-value higher than 0.01. An analysis of such cases revealed that they usually correspond to protein structures with a low number of atoms (as shown in section 5.9) or where the artefacts do not greatly affect the overall distribution of B factors (e.g. less than 5 atoms with B factors equal to zero). As discussed in section 5.9 the stringency of the KS-test is different depending on the size of the samples. This implies that for small sample sizes the statistical test is more permissive.

Regarding the presence of ties at the extremes of the B factor distributions, it should be noted that for cases characterized by a p-value higher than 0.01 the number of structures with strong ties at the highest B factor value is always higher than the number of structures with strong ties at the lowest B factor value (see Tables 5.3 and 5.4). This reflects the fact that in general, ties at the highest B factor value are caused by a constraint in the refinement parameterisations and they do not affect the distribution of the remaining B factors in the sample (unless the size of the sample is large), resulting in a higher number of structures with p-value higher than 0.01. These ties result from the fact that high B factors, which would be located along the tail of a fitted IGD*, are refined to a common unique value. Differently, ties at the lowest B factor value, even if still caused by a constraint in the refinement parameterisation, usually reflect an *anomalous* tendency of B factors to move towards unreasonable values (zero or negative). This can explain why their number is lower in models characterized by a p-value from the KS-test higher than 0.01.

Outlier structures for orthogonal statistics with a p-value higher than 0.01 can be considered a sort of *false negatives* since they contain artefacts that are not detected by the

p-value. Nevertheless these artefacts usually do not severely affect the distribution of B factors and can be fixed by performing a re-refinement of the protein model, provided the experimental data are available. In addition these outliers correspond to a small percentage of the total number of structures present in the considered p-value ranges (with a maximum of 5% for the ties at the maximum B factor in the p-value range $[0.05, 0.1[$).

The p-value range $]0, 0.01[$ contains 2918 and 2922 structures from the *all-chains* and the *single-chains* data sets, respectively, flagged as *suspicious* solely by the p-value. An analysis of the B factor distributions from these structures revealed that the majority of them are affected by a strong multimodality. These cases have been already taken into considerations and discussed in sections 5.4.4 and 5.4.6. A smaller part of the distributions showed instead a unimodal distribution similar to the one shown in panel (a) of Figure 5.5 but with smaller α and β parameters. As explained in section 5.3.2 the reason for the presence of such distribution is still not clear at the moment. A possible interpretation for such outliers is that tight restraints were used for the refinement of B factors.

As a closing remark, it should be noted that since the KS-test is a statistical test, and thus it obeys to the law of probability, there is always a non-zero chance to find false negatives in the analysis. Moreover a p-value of 0.5 does not mean automatically that a B factor distribution is "truly" *acceptable*, but instead "very likely" to be *acceptable*. The lack of absolute certainty is intrinsic to the nature of any statistical test and it should be accepted as it is.

Chapter 6

Ensemble Analysis of a Set of Protein Structures

In this chapter it is investigated to which extent a re-refinement protocol improves the results of an ensemble analysis performed with ESCET when the starting structures are affected by suspicious B factor distributions. A test case is selected from the large protein sample analysed in section 5.2 and the validation protocol introduced and discussed in section 5.1 is used to evaluate the quality of the B factor distributions before and after re-refinement. A comparative analysis with ESCET is then performed on both deposited and re-refined models.

6.1 The Selected Protein Structures

A test case was identified in the comparison of two protein complexes of the L-alanine dehydrogenase (L-AlaDH) from *Mycobacterium tubercolosis* (PDB codes 2VHW and 2VHX). This enzyme catalyzes the NADH-dependent reversible conversion of pyruvate and ammonia to L-alanine and it has been observed to undergo a conformational change upon coenzyme binding from an open to a closed conformation [4].

As shown in Figure 6.1 and described in the accompanying paper, each macromolecule of L-AlaDH consists of two distinct domains: The substrate-binding domain (residues 1–128 and 309–371) and the NAD-binding domain (residues 129–308). This two domains are connected by a hinge region consisting of two α-helices, here called H1 (residues 126–133) and H2 (residues 304–320). The authors report that this hinge region was determined by the program DynDom [4, 100].

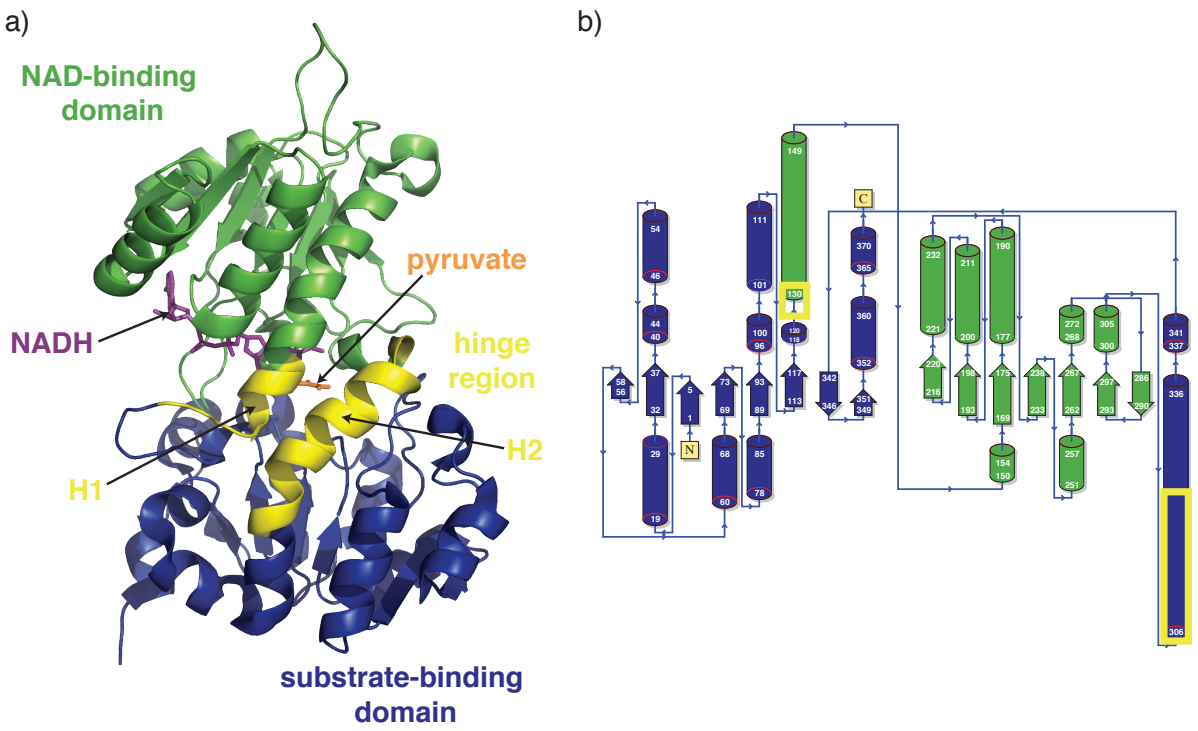


Figure 6.1: Domain composition for the L-AlaDH. In panel (a) the three-dimensional structure for the chain E from the pdb-entry 2VHX is shown in cartoon representation. The NAD-binding domain, the substrate-binding domain and the hinge region are green, blue and yellow coloured, respectively. For consistency, this color code is kept constant thoughout the chapter. In panel (b) the topology diagram obtained from PDBsum is shown. The two helices of the hinge region are depicted as yellow boxes along the topology diagram.

In solution, the L-AlaDH is a hexamer and the binding of the NADH moiety stabilizes the closed conformation of the holo-enzyme where the substrate-binding domains are rotated by about 16° toward the dinucleotide binding domains, compared to the open structure of the apo-enzyme [4].

The pdb-entry 2VHW contains the hexameric holo-form of the L-AlaDH in complex with NADH at 2.0 Å resolution. Chains A, B, C and D are in open conformation, while chains E and F are in closed conformation. The pdb-entry 2VHX contains the hexameric holo-form of L-AlaDH in complex with NAD+ and pyruvate at 2.0 Å resolution. Chains A, B, C and D are in open conformation and they bind only pyruvate, while chains E and F are in closed conformation and they bind both NAD+ and pyruvate.

6.2 Analysis of B Factor Distributions

The validation protocol for B factor distributions described in section 4.3 and extensively tested in chapter 5 was applied to the B factor distributions from pdb-entries 2VHW and 2VHX before and after re-refinement, respectively. The result of the analysis is shown in

Figure 6.2.

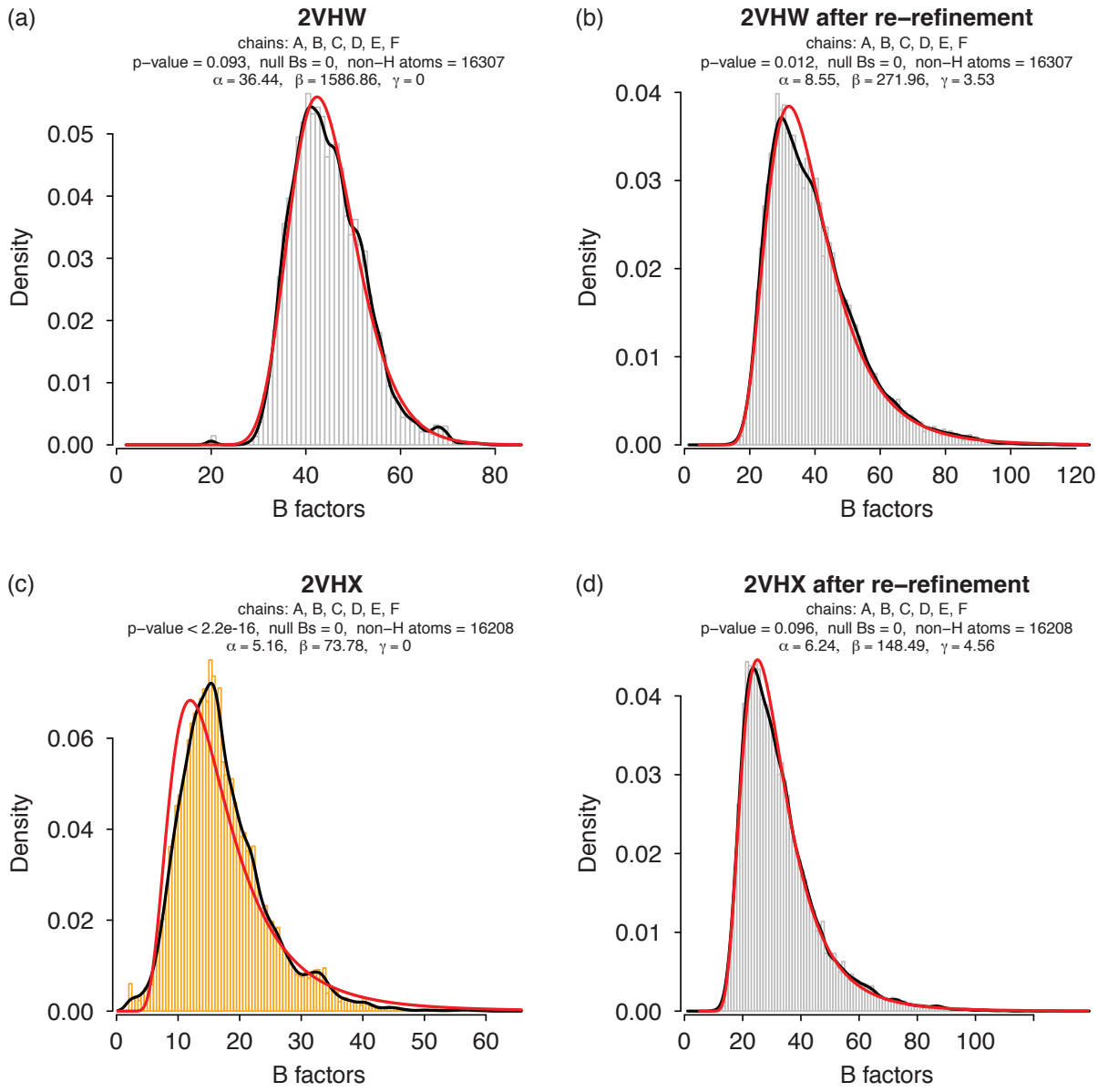


Figure 6.2: B factor distributions for PDB codes 2VHW and 2VHX as deposited into the PDB (panels (a) and (c)) and after re-refinement (panels (b) and (d)) respectively. For a description of the plot generated by the validation protocol please refer to Figure 5.1.

When the whole crystal structures in the asymmetric unit are taken into consideration for the analysis (*all-chains* data set), the B factor distribution of the default 2VHW structure is *acceptable*, with a p-value of 0.093, while the B factor distribution of the default 2VHX structure is *suspicious*, with a p-value lower than 0.01. After the re-refinement process, the R factor statistics R_{work} and R_{free} increased from 0.176 and 0.212 to 0.181 and 0.234 respectively for the pdb-entry 2VHW and from 0.172 and 0.210 to 0.174 and 0.225 respectively for the pdb-entry 2VHX. After re-refinement both structures have a p-value higher than the significance level of 0.01 (0.012 and 0.096 for PDB codes 2VHW and 2VHX, respectively)

and therefore can be considered *acceptable*.

Following the hierarchical model proposed in section 5.4.7 for the agreement of the B factor distributions in a protein model to the IGD* assumption, if the analysis is performed at single chain level (*single-chains* data set) three chains out of six from the pdb-entry 2VHW result to have *suspicious* B factor distributions (see Table 6.1). For the pdb-entry 2VHX all six chains have a *suspicious* B factor distribution (see Table 6.2). The standard refinement protocol is anyway sufficient to move all *suspicious* distributions to *acceptable* and confirm those that were already *acceptable*. It should be noted that especially for the pdb-entry 2VHW all B factor distributions from single chains as deposited into the PDB are characterized by a relatively high individual α parameter. The mean of the α parameter ($\langle\alpha\rangle = 46.52 \pm 6.81$) is in fact close to the upper bound limit of 50 defined in section 5.3, indicating that probably too tight restraints were used for the refinement of B factors. After re-refinement they all show a lower value of α parameter ($\langle\alpha\rangle = 9.07 \pm 1.91$), more in agreement with the general trend observed during the analysis of a large data set of protein structures at high resolution (see Figure 5.3).

chain	**atoms**	**default**				**re-refined**			
		α	β	γ	**p-value**	α	β	γ	**p-value**
A	2719	48.46	2217.24	0.00	0.043	9.63	345.71	2.07	0.240
B	2698	40.18	1504.62	2.15	0.038	8.17	213.00	5.11	0.560
C	2719	51.90	2449.36	0.00	0.021	6.90	212.77	7.74	0.380
D	2714	44.56	1720.59	1.83	0.003*	7.32	180.18	7.38	0.400
E	2731	38.22	1858.12	0.00	1.4e-05*	11.59	471.20	0.00	0.056
F	2726	55.82	2299.16	0.00	0.007*	10.82	316.73	3.92	0.130

Table 6.1: IGD* α, β and γ parameters and p-value from the KS-test for each single chain from the PDB code 2VHW before (default column) and after re-refinement (re-refined column). P-values lower than 0.01 are highlighted with an asterisk '*' symbol.

chain	**atoms**	**default**				**re-refined**			
		α	β	γ	**p-value**	α	β	γ	**p-value**
A	2679	5.75	87.84	0.00	6.4e-10*	7.06	186.90	3.70	0.450
B	2695	3.63	43.15	0.00	1.4e-09*	5.74	125.93	4.53	0.620
C	2682	6.46	99.04	0.00	8.3e-07*	5.16	116.09	7.31	0.380
D	2680	5.22	68.20	0.00	1.2e-05*	4.59	83.16	8.01	0.290
E	2736	7.83	141.19	0.00	4.6e-11*	8.54	259.15	1.22	0.150
F	2736	7.25	98.95	0.00	1.3e-04*	8.00	182.88	3.77	0.400

Table 6.2: IGD* α, β and γ parameters and p-value from the KS-test for each single chain from the PDB code 2VHX before (default column) and after re-refinement (re-refined column). P-values lower than 0.01 are highlighted with an asterisk '*' symbol.

6.3 Rigid Body Analysis with ESCET

The protocol described in section 4.6 was used to perform ensemble analysis with ESCET on the total data set of twelve chains of L-AlaDH (six from the pdb-entry 2VHW and six from the pdb-entry 2VHX). The analysis was performed separately on the structures as found in the PDB (default data set) and after re-refinement (re-refined data set).

In the default data set the average of the mean estimated standard uncertainties is 0.14 ± 0.01 Å. Chain B from PDB code 2VHW is the chain with the lowest mean estimated standard uncertainty (0.12 Å), while chain E from PDB code 2VHX is the chain with the highest mean estimated standard uncertainty (0.16 Å). The mean CSI for all 66 pair-wise comparisons (((12 x 12) - 12) / 2) , which can be used as a global measure of the structural diversity inside the ensemble, for the default data set is equal to 0.741 ± 0.186.

The cluster analysis based on the CSI matrix gave three different clusters of structures: the first two clusters contain structures in two slightly different open conformations, while the third cluster contains structures in closed conformations (data not shown). Three representative structures with the lowest mean estimated standard uncertainty (chains B, C and F from 2VHX) from each of the clusters were then used by the genetic algorithm to identify flexible and structurally invariant regions (i.e. rigid bodies). The result of the rigid body analysis is shown in panel (a) of Figure 6.3.

Using a ϵ_{low} of 2.0, the rigid body analysis identified five different rigid bodies. The first rigid body consists of 170 amino-acids (blue fragment in panel (a) of Figure 6.3) and corresponds to the substrate-binding domain; the second rigid body comprises 101 amino-acids (green fragment in panel (a) of Figure 6.3) and corresponds to the NAD-binding domain; the third rigid body (salmon fragment in panel (a) of Figure 6.3) corresponds to a small loop 11 amino-acids long (residues 149–159) located in the NAD binding domain; the fourth rigid body (yellow fragment in panel (a) of Figure 6.3) corresponds to a loop (residues 123–129) and an α-helix (residues 305–317) that have the function of hinge region between the NAD-binding domain and the substrate-binding domain; the fifth rigid body (cyan fragment in panel (a) of Figure 6.3) corresponds to an α-helix 11 amino-acids long (residues 200–210) located in the NAD-binding domain.

In the re-refined data set the average of the mean estimated standard uncertainties is 0.15 ± 0.01 Å. This indicates that the uncertainties are slightly increased if compared to the ones from the default data set. The increase in the mean of estimated standard uncertainties is due to the fact that after re-refinement the R factor statistics slightly worsened, especially the R_{free} that is used in the DPIU error model. Moreover some inconsistencies for the reflection data were detected between the deposited model and the obtained after re-refinement. In fact for the pdb-entry 2VHW the completeness of deposited reflection data moved from 99.40% for the deposited model to 94.40% for the re-refined model, causing

an increase of the computed DPI (see section 3.1.5 for more details about the error model).

The clusters obtained from the cluster analysis were the same as for the default data set regarding their members and the conformations they represented (data not shown). The only difference detected was the selection of the chain B from the pdb-entry 2VHW instead of the chain B from the pdb-entry 2VHX as representative structure for the first cluster. The genetic algorithm was then applied to find rigid bodies and the result is shown in panel (b) of Figure 6.3.

Using a ϵ_{low} of 2.0, the rigid body analysis identified four different rigid bodies. The first rigid body consists of 172 amino-acids (blue fragment in panel (b) of Figure 6.3) and corresponds to the substrate-binding domain; the second rigid body consists of 130 amino-acids (green fragment in panel (b) of Figure 6.3) and corresponds to the NAD-binding domain; the third rigid body (cyan fragment in panel (b) of Figure 6.3) corresponds to an α-helix 11 amino-acids long (residues 200–210) and is the same rigid body as the fifth rigid body in the default data set; the fourth rigid body comprises 19 amino-acids (yellow fragment in panel (b) of Figure 6.3) and as in the default data set it corresponds to the hinge region between the substrate-binding domain and the NAD-binding domain.

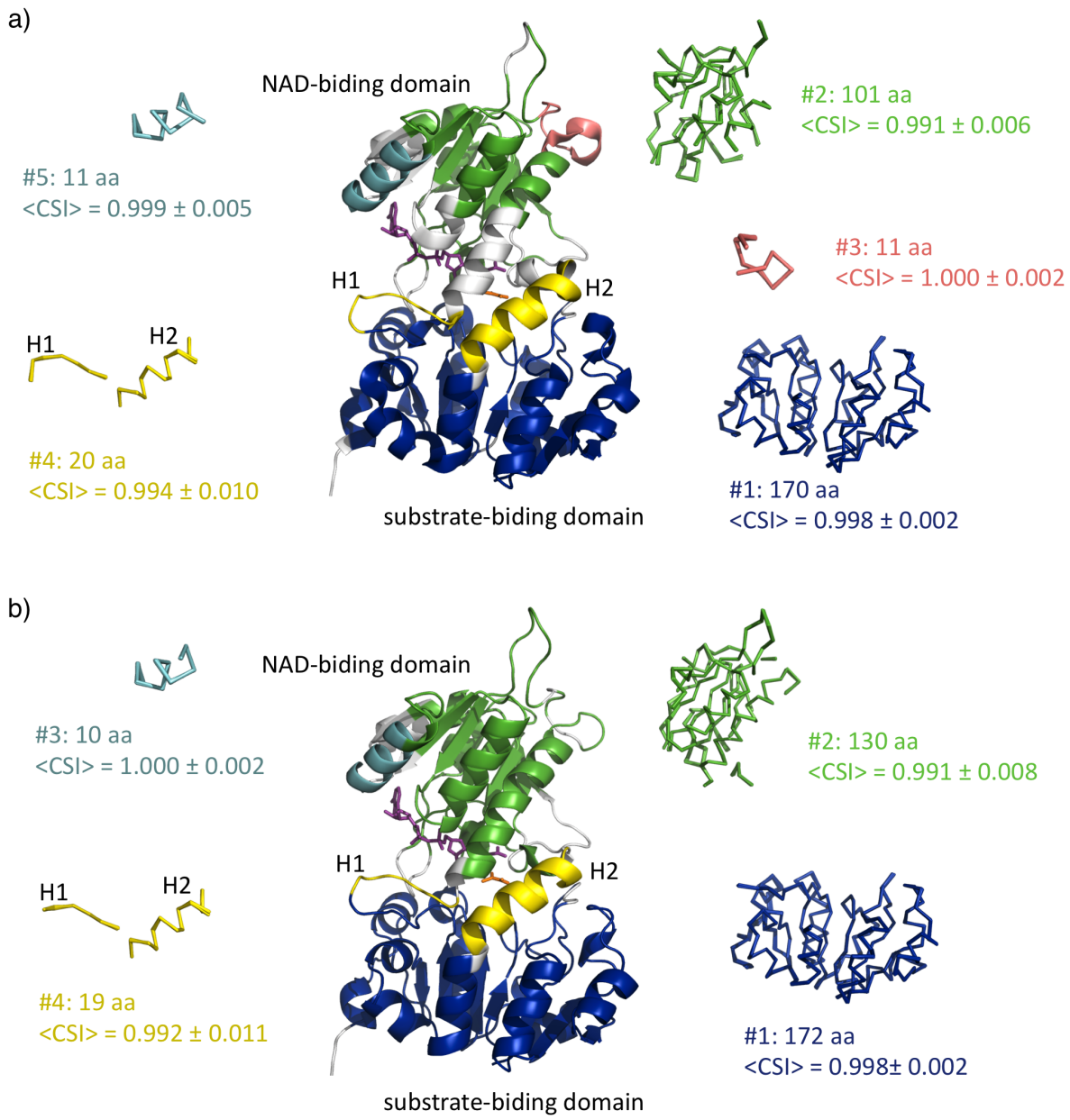


Figure 6.3: Identification of rigid bodies using PDB codes 2VHW and 2VHX before (panel (a)) and after re-refinement with REFMAC5.6 (panel (b)). In the center of each panel a schematic view of the L-AlaDH is shown using a cartoon representation. Each rigid body is coloured with a different colour. NAD+ and pyruvate molecules are shown as stick models in purple and orange, respectively. Flexible regions are light grey coloured. For each rigid body, the superposition of the 12 fragments is shown in a ribbon representation together with the number of amino-acids and the average CSI for all 66 pair-wise comparisons.

The results obtained from the rigid analysis are in agreement with the overall architecture of the L-AlaDH reported in the original publication [4], consisting of two domains (substrate-binding domain and NAD domain) connected by an hinge domain. Nevertheless the hinge region identified with ESCET is slightly different from the one reported by the authors and determined by the program DynDom [100]. In fact the DynDom program identified the hinge region in two α-helices (residues 126–133 and residues 304–320 cor-

responding to fragments H1 and H2, respectively), while the ESCET framework identified it in one loop (residues 124–129, corresponding to fragment H1) and an α-helix (residues 305–317, corresponding to fragment H2). It should be noted that when describing the results obtained from ESCET, the numbering refers to the results from the re-refined data set. The biggest difference between the results from DynDom and ESCET is in the first fragment. In the results from ESCET it ends before the long α-helix defined from Pro132 to Leu149. In the analysis of the re-refined data set this helix is part of the rigid body corresponding to the NAD-binding domain. This is consistent with the literature since it is reported that the pyrophosphate moiety of the NAD molecule makes a hydrogen bond with the side chain of Ser134 and the nicotinamide ring is bound in a pocket that includes residues Met133, Ser134, Ala137 [4] (see Figure 6.4). It is then reasonable to assume that the fragment H1 of the hinge region ends at Leu129 before Pro132, since from Met133 a number of interactions with the NAD molecule stabilize the nascent helix in a compact form with the NAD-binding domain. Moreover the amino acid proline is known to cause bends when located in alpha-helices and to be characterized by a low conformational freedom [21] (Pro132 is the first amino-acid in the second rigid body corresponding to the NAD-binding domain after the H1 fragment of the hinge region, separated by a flexible stretch consisting of Leu130 and Ala131).

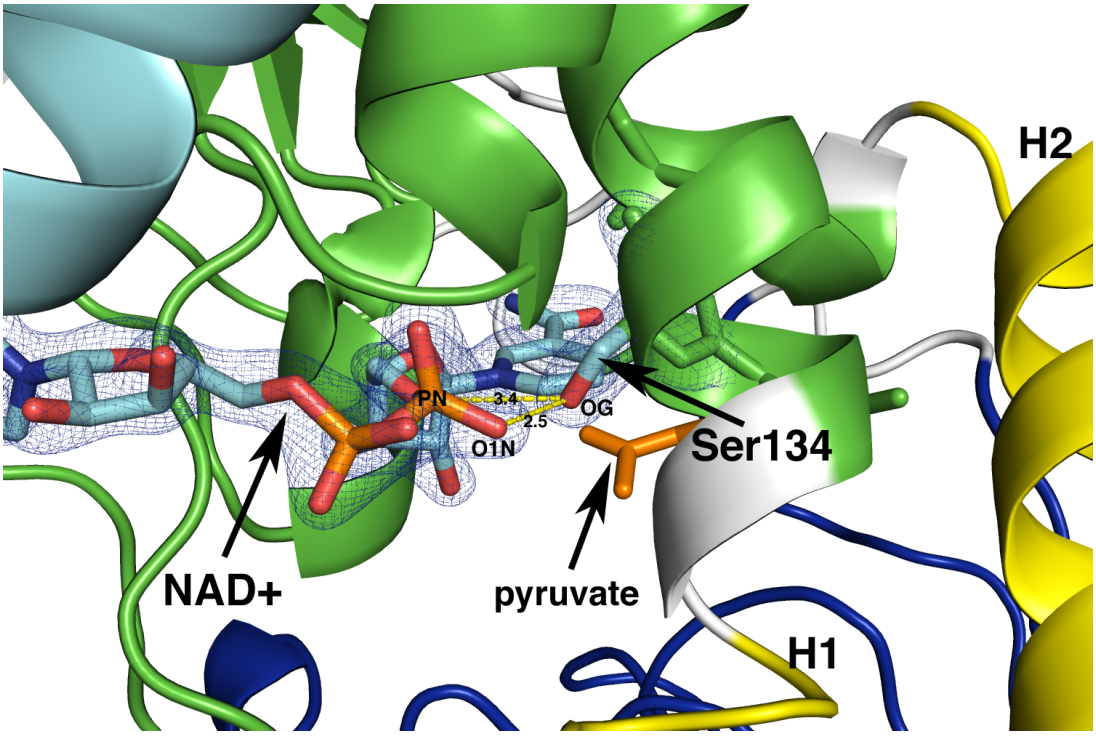


Figure 6.4: Interactions between the NAD+ molecule and Ser134. The overal L-AlaDH model is shown in cartoon representation. The NAD+ and Ser134 moieties are shown in stick representation and their σ-weighted 2mFo-DFc electron density contoured at 2 σ and carved at 1.5 Å is shown in blue. The same color code of panel (b) in Figure 6.3 is used. Dashed yellow lines are used to highlight the two hydrogen bond between Ser134 and the pyrophosphate group of the NAD+ molecule.

It should be noted that after re-refinement, the rigid bodies found are on average larger in size and lower in number if compared to those obtained in the default data set (e.g. the third rigid body found in the default data set was not found in the re-refined data set. Instead it was included in the second rigid body) and in general the flexible regions are less and smaller. This can be interpreted as a reduction of noise in the weighted distance matrices used by ESCET, where the B factors and then their distributions play an important role.

The RMSDs without superposition for CA-atoms before and after refinement for the pdb-entries 2VHW and 2VHX are 0.10 Å and 0.10 Å, respectively. Since the mean estimated standard uncertainties for the default and re-refined data sets are 0.135 and 0.149 respectively, the coordinate differences are within the error. Thus, it is reasonable to consider the differences observed in the cluster analysis as the result of different estimated uncertainties combined with different B factor distributions. However it is difficult to decouple the contribution of each of them since they are related to each other (i.e. the distribution of B factors has effects on the refinement statistics from which the DPI and subsequently the DPIU error model are computed).

Chapter 7

Ensemble Analysis of Ribosomal Structures

In this chapter a data set of 29 crystal structures of the 30S ribosomal subunit from T. thermophilus is defined. Since in chapter 6 it was shown that a re-refinement of the deposited models can improve both B factor distributions and results from the ESCET framework, 13 different refinement protocols are applied to the selected ribosomal structures. To find the protocol that provides the best re-refined models, the results from the re-refinement procedures are analysed together with the results from the validation protocol for B factor distributions, described in chapter 5, applied to each re-refined model. The ESCET framework is then used to perform a comparative ensemble analysis between the default data set and the best data set from the refinement procedure.

7.1 Selection of Ribosomal Structures

In the attempt to draw a summary picture of all the available ribosome structures at medium resolution, the protocol described in section 4.4 was used to query the PDB. As of June 2010 it resulted in 197 ribosomal complexes at a resolution of 4 Å or higher, resulting in a total of 263 individual subunits deposited into the PDB. A complete list of the retrieved structures is available in Table E.1 in Appendix E and a summary of the composition of the data set is shown in Table 7.1.

organism	70S	50S	30S	total
E. coli	28 [3.2Å, 4.0Å]			28 [3.2Å, 4.0Å]
T. thermophilus	38 [2.8Å, 3.8Å]		32 [2.5Å, 3.8Å]	70 [2.5Å, 3.8Å]
H. marismortui		67 [2.2Å, 3.5Å]		67 [2.2Å, 3.5Å]
D. radiodurans		32 [2.9Å, 3.8Å]		32 [2.9Å, 3.8Å]
total	66 [2.8Å, 4.0Å]	99 [2.2Å, 3.8Å]	32 [2.5Å, 3.8Å]	197 [2.2Å, 4.0Å]

Table 7.1: Summary of available ribosomal structures at resolution equal to or higher than 4.0 Å. For each data set the highest and the lowest resolution available are reported in square brackets.

During some preliminary analysis no relevant conformational changes were detected in the 50S data sets from *D. radiodurans* and *H. marismortui* organisms. Moreover it was found that some of these models are incomplete (e.g. in the pdb-entry 2AAR [10] all the protein moieties are missing). This affects the reliability of the deposited structures since the reported refinement statistics should be reproducible by default or at least after re-refinement, which is impossible if the deposited models are incomplete or the structure factors were not deposited.

The choice of the ribosomal data sets to analyse moved then to the 30S subunits from the 70S complexes or the 30S complexes, because it is known from the literature that it is the 30S subunit that is mostly affected by conformational changes during the protein synthesis process.

However during the characterization of the retrieved structures it was found that the majority of the deposited structures were solved by molecular replacement or by difference Fourier methods. Considering the average resolution of the available structures (3.38 ± 0.29 Å) it is likely that the search model used to solve the new structures introduced a bias in the final model. Furthermore, the search model is usually the same for all the structures from an organism and it usually corresponds to the first structure ever solved for that particular organism.

One possible way to efficiently deal with structural bias is to re-refine the structures with an *aggressive* refinement protocol (e.g. by using simulated annealing refinement). However the refinement of ribosome structures becomes a large task while moving from 30S to 70S complexes. In fact the 70S complexes usually contain two complete copies of the entire 70S ribosome per asymmetric unit. This causes an increase in the CPU time (i.e. three macro cycles of refinement with *phenix.refine* program including simulated annealing refinement in the second macro cycle took ca. eight hours for a 30S complex and ca. two days for a 70S complex on a Linux workstation with 8 CPUs Intel® Xeon® 2.80 GHz) and in the memory (the maximum peak of memory usage is around 5.7 GB for the 30S subunit and around 19.7 GB for the 70S complex) required for the refinement. In addition, for the 70S complexes a larger number of macro-cycles is probably required to reach convergence if

compared to the 30S subunit. For sake of convenience and efficiency the analysis was thus focused on the 32 *T. thermophilus* 30S ribosomal subunits from 30S complexes (a list of structures is listed in Table7.2).

The working data set was further reduced from 32 to 29 structures since for three structures (PDB codes: 1I94, 1FKA and 2F4V) the structure factors are not available and it is then impossible to perform any re-refinement (see Table 7.2). The relatively small size of the data set makes it ideal for studying the effect of different refinement protocols. Moreover the 30S complexes from *T. thermophilus* are the structures that have been used to study the decoding process and the conformational changes related to it.

In ascending order of deposition date, the final data set of 29 30S subunit structures from *Thermus thermophilus* is defined as follows. The pdb-entry 1J5E is the first structure of 30S subunit ever solved at medium resolution (3.05 Å) [134] and it corresponds to a 30S subunit in open conformation without any ligand (group I in Table 7.2). It should be noted that the deposition date (04/08/02) is misleading since this structure supersedes the pdb-entry 1FJF (deposited on 08/08/00). The main difference is that the pdb-entry 1J5E contains a more complete model for the 16S rRNA, especially for the 3' folded back along the mRNA path. The pdb-entry 1FJG was solved to study the interaction of three different antibiotics (paromomycin, streptomycin and spectinomycin) [32] (group II). Similarly, the pdb-entries 1HNW, 1HNX and 1HNZ were used to study the structural basis for the mechanism of action of other three antibiotics (tetracycline, pactamycin and hygromycin) [23] (group IV). The pdb-entry 1HR0 was solved to obtain the first structure of a 30S subunit from *T. thermophilus* in complex with the initiation factor 1 (IF1) [31] (group V). The pdb-entries 1IBK, 1IBL and 1IBM were the structures used to study the recognition of cognate tRNA in the decoding center [91] (group VII). The pdb-entries 1N32, 1N33, 1N34 and 1N36 were used to study the transition from an open to a closed conformation of the 30S subunit during the selection of tRNA [93] (group VIII). The pdb-entries 1XMQ and 1XMO were determined to study the role of base modifications in codon discrimination by tRNA [87] (group IX), while the pdb-entries 1XNR and 1XNQ were used to study the structure of a purine-purine wobble base pair in the decoding center of the ribosome [86] (group X). The pdb-entry 2HHH was used to study the mechanism of action of the antibiotic kasugamycin that inhibits translation initiation of canonical but not of leaderless messenger RNAs [108] (group XII). The pdb-entry 2E5L was solved to study the structural mechanism of the interaction between the Shine-Dalgarno (SD) sequence present in the 5' of the mRNA with the anti-Shine-Dalgarno (aSD) sequence present in the 3' or the 16S rRNA [67] (group XIII). The pdb-entries 2UU9, 2UUA, 2UUB and 2UUC were used to study the mechanism for expanding the decoding capacity of tRNA by modification of uridines in the first position of the anticodon which is involved in the formation of the wobble base-pair [132] (group XIV). The pdb-entries 2UXC, 2UXD and 2UXB were used to study structures of tRNA with an ex-

panded anticodon loop in the decoding center of the 30S ribosomal subunit. All anticodons used for these structures recognise codons four nucleotides long [45] (group XV). The two pdb-entries 2VQE and 2VQF were solved to study the wobble base pair when the first position of the anticodon on the tRNA contains the modified base 5-taurinomethyluridine [76] (group XVI). The pdb-entry 2ZM6 contains the crystal structure of an empty 30S ribosomal structures from *Thermus thermophilus* (group XVII). However no reference is available for this structure and the only source of information is the header section of the corresponding pdb-file.

id	PDB	gr	d_{min}	R_{work}	R_{free}	SF	dep	sol	SM	<esu>† [Å]	<esu>‡ [Å]	Δ<esu>
1	1J5E	I	3.05	0.21	0.25	yes	4/8/02	MIRAS	/	0.378	0.332	↘
2	1FJG	II	3.00	0.22	0.26	yes	8/8/00	DF	**1J5E**	0.362	0.318	↘
3	~~1FKA~~	III	3.30	0.30	0.30	no	8/9/00	MIRAS	/	*na*	*na*	*na*
4	1HNW	IV	3.40	0.22	0.26	yes	12/8/00	DF	**1J5E**	0.511	0.365	↘
5	1HNX	IV	3.40	0.23	0.28	yes	12/8/00	DF	**1J5E**	0.579	0.423	↘
6	1HNZ	IV	3.30	0.22	0.26	yes	12/8/00	DF	**1J5E**	0.516	0.410	↘
7	1HR0	V	3.20	0.22	0.26	yes	12/20/00	DF	**1J5E**	0.448	0.368	↘
8	~~1I94~~	VI	3.20	0.20	0.24	no	3/18/01	MIR	/	*na*	*na*	*na*
9	1IBK	VII	3.31	0.23	0.28	yes	3/28/01	DF	**1J5E**	0.522	0.445	↘
10	1IBL	VII	3.11	0.23	0.28	yes	3/28/01	DF	**1J5E**	0.433	0.366	↘
11	1IBM	VII	3.31	0.23	0.29	yes	3/28/01	DF	**1J5E**	0.528	0.430	↘
12	1N32	VIII	3.00	0.23	0.27	yes	10/25/02	DF	**1J5E**	0.379	0.350	↘
13	1N33	VIII	3.35	0.22	0.28	yes	10/25/02	DF	**1J5E**	0.619	0.520	↘
14	1N34	VIII	3.80	0.24	0.31	yes	10/25/02	DF	**1J5E**	0.967	0.758	↘
15	1N36	VIII	3.65	0.26	0.32	yes	10/25/02	DF	**1J5E**	0.905	0.783	↘
16	1XMO	IX	3.25	0.23	0.28	yes	10/4/04	MR	**1J5E***	0.515	0.443	↘
17	1XMQ	IX	3.00	0.22	0.24	yes	10/4/04	MR	**1J5E***	0.317	0.319	↗
18	1XNQ	X	3.05	0.23	0.27	yes	10/5/04	MR	**1J5E***	0.395	0.347	↘
19	1XNR	X	3.10	0.23	0.27	yes	10/5/04	MR	**1J5E***	0.421	0.379	↘
20	~~2F4V~~	XI	3.80	0.26	0.32	no	11/24/05	MR	**1J5E**	*na*	*na*	*na*
21	2HHH	XII	3.35	0.26	0.29	yes	6/28/06	MR	?	0.478	0.484	↗
22	2E5L	XIII	3.30	0.26	0.30	yes	12/21/06	MR	?	0.563	0.475	↘
23	2UU9	XIV	3.10	0.23	0.27	yes	3/1/07	MR	**1J5E**	0.377	0.385	↗
24	2UUA	XIV	2.90	0.22	0.25	yes	3/1/07	MR	**1J5E**	0.312	0.288	↘
25	2UUB	XIV	2.80	0.22	0.24	yes	3/1/07	MR	**1J5E**	0.269	0.256	↘
26	2UUC	XIV	3.10	0.21	0.24	yes	3/1/07	MR	**1J5E**	0.355	0.314	↘
27	2UXB	XV	3.10	0.30	0.33	yes	3/28/07	DF	**1J5E***	0.494	0.403	↘
28	2UXC	XV	2.90	0.22	0.26	yes	3/28/07	DF	**1J5E***	0.320	0.302	↘
29	2UXD	XV	3.20	0.24	0.28	yes	3/28/07	DF	**1J5E***	0.481	0.342	↘
30	2VQE	XVI	2.50	0.26	0.28	yes	3/13/08	MR	**1J5E**	0.226	0.203	↘
31	2VQF	XVI	2.90	0.22	0.26	yes	3/14/08	MR	**1J5E**	0.325	0.279	↘
32	2ZM6	XVII	3.30	0.29	0.32	yes	4/11/08	MR	2E5L	0.606	0.534	↘

Table 7.2: *Thermus thermophilus* 30S working data set. The following abbreviations are used in the column names: PDB stands for PDB identification code, gr stands for group (a cardinal roman number is used to group together structures that appeared in the same publication), d_{min} stands for maximum resolution, R_{work} stands for refinement R factor for the working reflection data set, R_{free} stands for refinement R factor for the validation reflection data set, SF stands for structure factors availability, dep stands for deposition date, sol stands for solution method, SM stands for search model, <esu>† stands for mean estimated standard uncertainty for the structures as found in the PDB, <esu>‡ stands for mean estimated standard uncertainty for the structures after re-refinement with protocol 2 (see section 4.5 for details), Δ<esu> stands for gradient of the difference between <esu>‡ and <esu>†. For the solution method the following abbreviations are used: MR stands for molecular replacement, DF stands for difference fourier, MIRAS stands for multiple isomorphous replacement with anomalous scattering, MIR stands for multiple isomorphous replacement. Regarding the search model used for the solution of the structure, a question mark (?) is used when it was not possible to find any information in the literature . An asterisk (*) is instead used to mark the most probable search model used, but for which no clear and explicit evidence is present in the PDB file nor in the accompanying article. Since for the pdb-entries 1I94, 1FKA and 2F4V the structures factors are not available, it was not possible to perform any re-refinement protocol and thus they were not used for the ensemble analysis. For that reason the <esu>† and the <esu>‡ were set to *na* (not available) and their PDB codes were crossed out. It is worth noting that, except for the PDB codes 1XMQ, 2UU9 and 2HHH, in all cases the mean estimated uncertainty lowered after re-refinement with protocol 2 (described in section 4.5). For the definition of the groups see text in section 7.1.

7.2 Refinement of Selected Ribosomal Structures

In order to find the refinement strategy that gives the best R factor statistics and at the same time produces the highest number of structures with B factor distribution in agreement with the IGD* assumption described in chapter 5, the 13 refinement protocols defined in section 4.5 were applied to the data set of 29 30S ribosomal subunit reported in Table 7.2. The re-refinement procedure led to a final pool of 377 models (29×13). A complete list of the R factor statistics for the 29 structures before and after re-refinement is reported in Table D.1 in appendix D.

In Figure 7.1 the distribution of the R factor statistics for selected refinement protocols is analysed for each PDB code. For sake of clarity, only the default structures and those from refinement protocols 2, 3, 8, 11 and 12 were taken into consideration for this comparative analysis. Protocols 2, 3, 8, 11 and 12 can in fact be considered as the representative protocols for the different refinement parameterisations tested in this study (for a definition of the refinement protocols please refer to section 4.5).

The refinement protocol 2 is the one that gives the lowest average R_{work} and R_{free} statistics in the data set (0.20 ± 0.02 and 0.25 ± 0.02, respectively), indicating a better agreement between the re-refined models and the deposited experimental data. This is due to the fact that TLS refinement was used in combination with individual ADP refinement, where ADP stands for atomic displacement parameter (term used in *phenix.refine* program to refer to B factors). The fact that in general TLS refinement gives better R factor statistics is in agreement to what was observed in the outcomes of the PDB_REDO project [65, 66] and in some comparative analyses made by the PHENIX team for different ADP parameterisations applied to several real cases at medium and low resolutions [2].

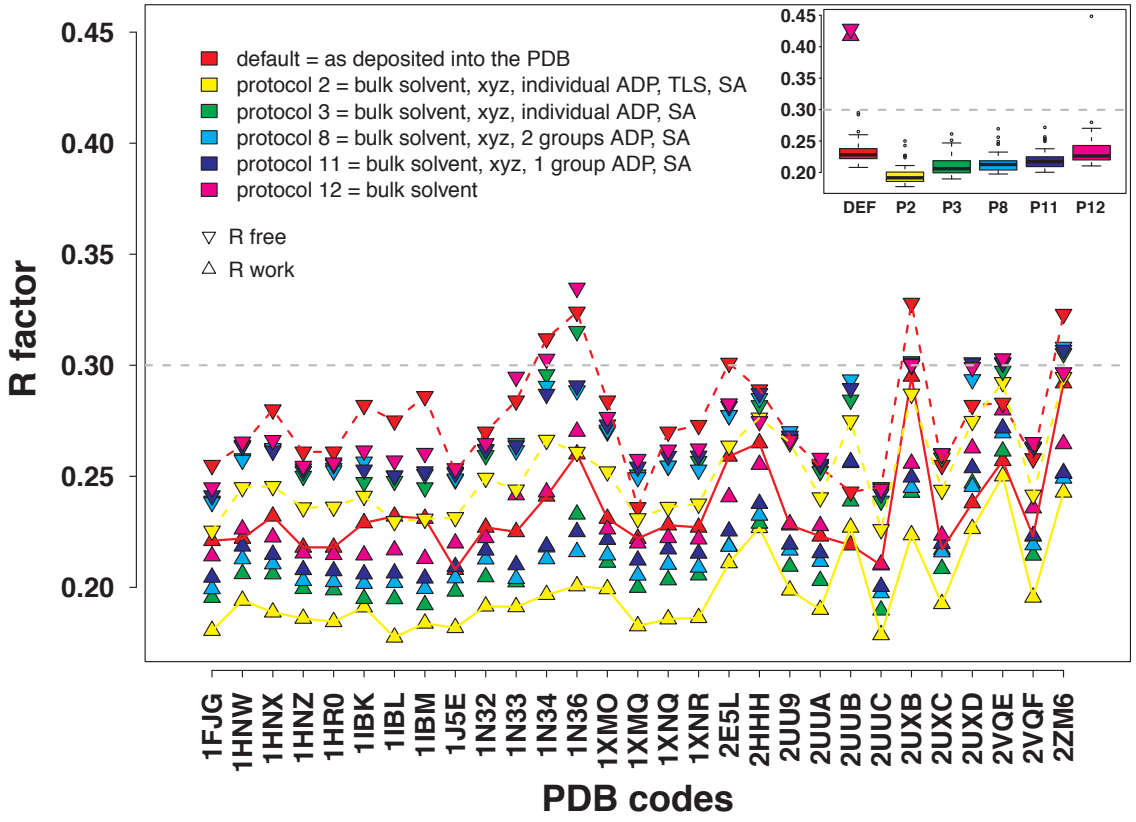


Figure 7.1: R factors R_{work} and R_{free} for default and re-refined ribosomal structures from selected re-refinement protocols. Triangles pointing upwards are used to represent R_{work} values, while triangles pointing downwards are used to represent R_{free} values. Different colours are used for the different protocols (see legend). A dashed gray line indicates the limit of 0.30 which is usually used to discriminate between reliable (< 0.30) and unreliable (≥ 0.30) models. R_{work} and R_{free} for the default structures and for the results of protocol 2 are highlighted with a solid and a dashed line, respectively. The insert shows a box-plot for the R_{work} values. The following abbreviations are used in the annotation of the box-plot for the refinement protocols: DEF stands for default and P2, P3, P8, P11, P12 stand respectively for protocols 2, 3, 8, 11 and 12.

Looking at the comparative distributions of R_{work} and R_{free} statistics it is worth noting that a simple update of the bulk solvent correction as in protocol 12 gives on average better R factor statistics if compared to those from the default structures. The only exception in this respect is the pdb-entry 2UUB, for which a re-refinement of the atomic coordinates xyz is necessary to obtain R factors close to the ones deposited in the PDB (see protocol 13 in Table D.1 in Appendix D). This is verified also in protocols 2, 3, 8 and 11 shown in Figure 7.1. In agreement with the average R factors computed for the data sets, the refinement protocol 2 outperforms all the other refinement protocols and it is able to lower all the R_{free} statistics below the threshold value of 0.30, while maintaining reasonable stereochemistry (data not shown). Improvements of more than 0.04 in both R_{work} and R_{free} are observed for the pdb-entries 1N34, 1N36, 2E5L, 2UXB and 2ZM6. It should be noted that for the pdb-entries 1N34 and 2UXB the ΔR, defined as the difference between the

R_{free} and the R_{work} statistics, starts to be important (~ 0.07) indicating a possible problem of over-refinement. Nevertheless, since it is lower than the usually accepted maximum limit of 0.08 for reliable structures, they are considered acceptable.

7.3 Analysis of B Factor Distributions from Default and Re-refined Structures

7.3.1 Analysis of α and β Parameters from Fitted IGD*

The validation protocol for B factor distributions described in section 4.3 and extensively tested for protein structure in chapter 5 was applied to the B factor distributions for the phosphate atoms from the 16S rRNA moieties of the 30S ribosomal subunits before and after application of the 13 refinement protocols. The models contain between 1472 (PDB code 2UXD) and 1517 (PDB code 2E5L) phosphate atoms in the 16S rRNA. In the analysis, for the models refined with TLS groups (refinement protocol 2) the equivalent isotropic B factors were taken into consideration. The reason why phosphate atoms were taken into account as representative is explained in section 4.6.2.

The distribution of the α IGD* parameter for each model from the default data sets and five selected refinement protocols (the same used in Figure 7.1) is shown in Figure 7.2.

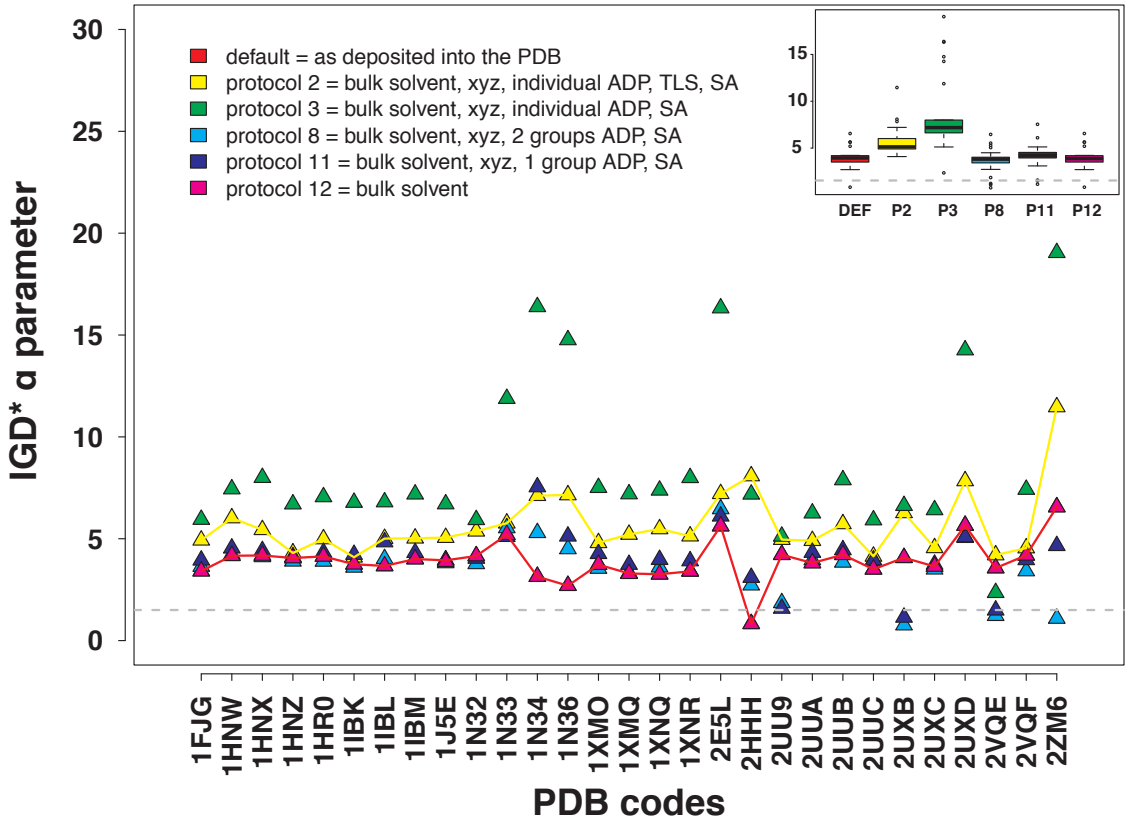


Figure 7.2: Distribution of α parameter values for the models resulting from selected refinement protocols. The same colour code and abbreviations as in Figure 7.1 are used for different refinement protocols. The dashed gray line is used to highlight the lower bound limit of 1.5 for the α IGD* parameter as described in section 5.3. If the IGD* that fits the B factor distribution is characterized by an α parameter lower than 1.5, it is considered *suspicious*. The α parameters for the default structures and for protocol 2 are highlighted by a solid line. The insert shows a box-plot of the distribution of α parameter for the selected refinement protocols.

If α values of less than 1.5 are considered outliers as in the protein case (see section 5.3), in the default data set the pdb-entry 2HHH is an outlier since the IGD* fitted to the B factor distribution of phosphate atoms is characterized by an α parameter of 0.82. The reason for such a low value is a strong tie at the lowest B factor (24 P atoms out of 1502 with a B factor equal to 1 Å^2) as shown in panel (c) of Figure 7.4. This is equivalent to what was observed for protein outliers in group 1 of Figure 5.3 and discussed in section 5.3.1. In fact, it has been observed that the presence of a strong tie at the lowest B factor value considerably affects the α parameter of the fitted IGD* and usually reflects a problem also for the distribution of the remaining B factors in the sample (see section 5.3.1).

The distribution of the α parameter for all the models obtained from refinement protocol 2 are characterized by acceptable α parameters (between 1.5 and 50, as proposed for the protein data set analysis in section 5.3). The mean α from refinement protocol 2 is 5.68 ± 1.55 and the mean β is 347.25 ± 224.28. The mean α is very close to the average α in the

all-chains data set of protein structures (i.e. 5.89) analysed in section 5.3. This positions the models from protocol 2 in the allowed region for α parameters in Figure 5.3 obtained from the analysis of a large data set of proteins structures at high resolution. The average β is instead at least three times higher than the one observed in the *all-chains* data set of protein structures (i.e. 96.15) in section 5.3. The reason for such a large difference for the β parameter is due to the fact that the average B factor for the phosphate atoms in ribosomal structures is much higher than the average B factor observed in the protein data set (88.02 Å^2 against 22.30 Å^2). From equation (5.7) it becomes clear how the average B factor affects the value of the β parameter:

$$\beta = (\mu - \gamma)(\alpha - 1) \approx (\langle B \rangle - \gamma)(\alpha - 1).$$

The fact that the β parameter is proportional to the average B factor in the model explains the observed difference between the protein and ribosome data sets. It should be noted that also the γ parameter affects the final value of the β parameter. The mean γ parameter is equal to 4.76 ± 4.19 and 18.83 ± 12.39 for the *all-chains* protein data set and ribosomal models from protocol 2, respectively. Thus, even if present in the equation for the β parameter, it does not greatly affect the observed difference between the two data sets.

Looking at the distribution of the α parameter in function of the refinement protocol used (box-plot insert in Figure 7.2) it is possible to observe that protocols 2 and 3, where the individual ADP model was used in refinement, are characterized by α values on average higher than protocols where a group B factor model was used (i.e. default, protocol8, protocol9, protocol11, protocol12). This behaviour can be explained by the fact that the type of B factor model and restraint used for the refinement of B factors affect their overall distribution. On average, the stronger the restraint, the lower the corresponding variance of B factors in the model. The effect of the variance of B factor distributions on the α parameter is clear from equation (5.6):

$$\alpha = \frac{(\mu - \gamma)^2}{\sigma^2} + 2 \approx \frac{(\langle B \rangle - \gamma)^2}{s_B^2} + 2 \propto \frac{\langle B \rangle - \gamma}{s_B}.$$

This is in agreement with what is observed from the comparative analysis shown in the insert of Figure 7.2. In fact, group B factors are usually not restrained at all and they result in B factor distributions with the highest variances. Instead, individual B factors, at least in *phenix.refine* program, are spatially restrained as described in reference [2]. This results in B factor distributions with the lowest variances. In *phenix.refine* the cut-off distance used to restrain individual ADP is set to 1.55 Å and to 5 Å depending if TLS groups are used (protocol 2) or not (protocol 3), respectively. This is reflected in the distribution of the α parameter. In fact, when TLS are not used (protocol 3) the effect of the restraints is stronger (due to the longer distance cut-off), resulting in a lower variance of B factors and then in a

higher α parameter.

From a methodological point of view it should be pointed out that in the analysis of B factor distributions from protein structures discussed in chapter 5 the models refined with TLS groups were explicitly discarded. This was done to avoid strong correlations between B factors that could break the assumption of independence made in the statistical rationale described in section 5.1. The fact that for the ribosomal structures from *T. thermophilus* re-refined with protocol 2 the IGD* assumption is valid even in presence of TLS groups can be explained by the fact that in *phenix.refine* program the TLS refinement is always combined in two subsequent steps with individual ADP refinement [2] and by the fact that only phosphate atoms are taken into account, which are not directly bound to each other. Moreover the limited number of phosphate atoms in the 16S rRNA allows the KS-test to be quite permissive in the validation of the distribution of B factors as discussed in section 5.9.

7.3.2 Analysis of P-value Statistics from KS-test

The distribution of p-values for each structure from the default data sets and five selected refinement protocols (the same used in Figure 7.1) is shown in Figure 7.3.

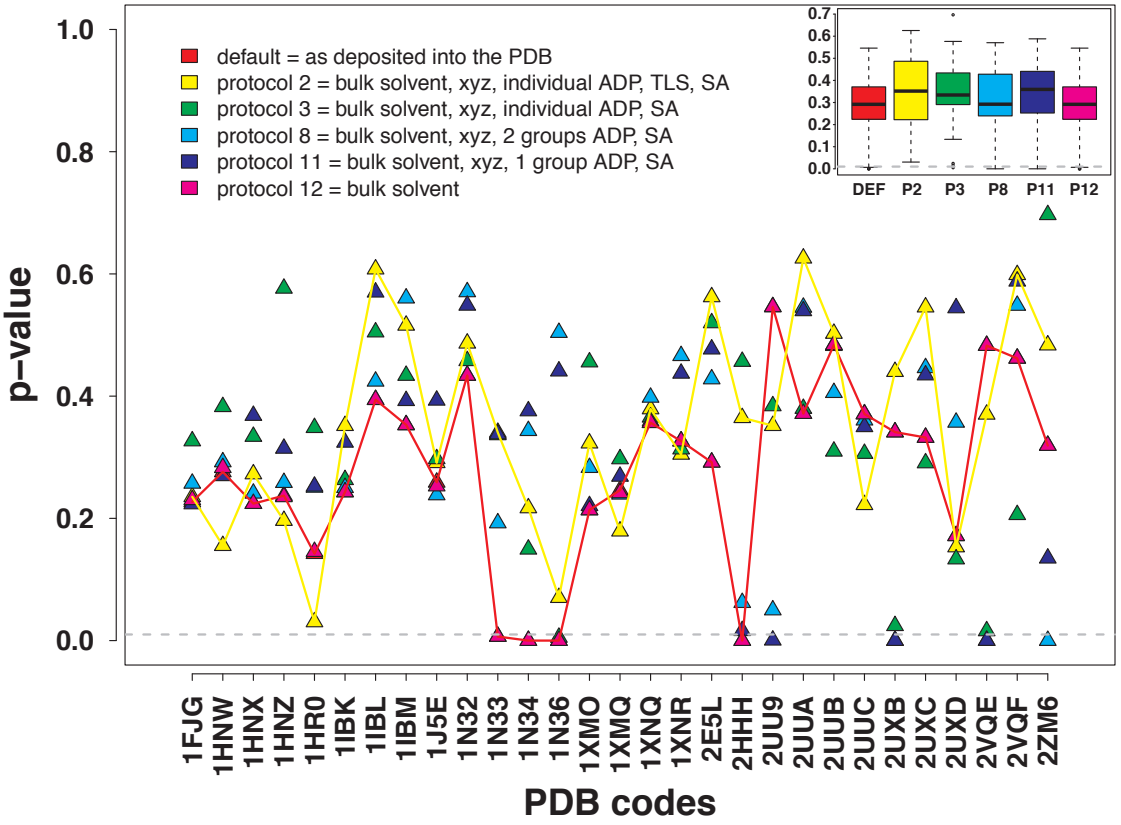


Figure 7.3: P-values from KS-test for P-atoms from 16S rRNA moieties. The same colour code and abbreviations as in Figure 7.1 are used for different refinement protocols. The dashed gray line is used to highlight the confidence level of 0.01 used to discriminate between *acceptable* (p-value $\geq$ 0.01) and *suspicious* (p-value $<$ 0.01) B factor distributions. P-values for the default structures and for the protocol 2 are highlighted by a solid line. The insert with a box-plot of the distribution of the p-values in functions of the selected refinement protocols is shown.

In the default data set 25 out of 29 ribosomal structures exhibit *acceptable* B factor distributions for P-atoms in the 16S rRNA moiety. This is the case although for 18 structures a strong tie is detected at the highest B factor value around 200 Å^2 (pdb-entries 1FJG, 1HNW, 1HNX, 1HNZ, 1HR0, 1IBK, 1IBL, 1IBM, 1N32, 1XMO, 1XMQ, 1XNQ, 1XNR, 2E5L, 2UXB, 2UXC, 2UXD and 2ZM6). This is probably due to the introduction of a constraint for B factor values at refinement stage and it has been observed also in protein structures. Nevertheless, since the number of P-atoms is relatively low (around 1507, as discussed above), the presence of a strong tie at the upper limit of the B factor distribution does not considerably affect the outcome of the KS-test. In addition, as observed for in protein structures, ties at the highest B factor value correspond to less severe problems in the B factor distribution than ties at the lowest B factor value (see section 5.5).

It is worth noting that, as previously observed for the distribution of R factor statistics in section 7.2, also for the distribution of p-values from the KS-test the refinement protocol 2 performs better than the other protocols. In fact it is the only protocol for which all B

factor distributions from the 16S rRNA have a p-value higher than 0.01 after re-refinement and thus they are all considered *acceptable* for the IGD* assumption. All four structures (PDB codes 1N33, 1N34, 1N36 and 2HHH) that in the default data had a *suspicious* B factor distribution, were rescued after re-refinement with protocol 2. Two examples (PDB codes 1N33 and 2HHH) are shown in Figure 7.4 for which the highest $\Delta p\text{-}value$ (~ 0.3) was observed in the comparative analysis between the default data set and the data set obtained from re-refinement with protocol 2.

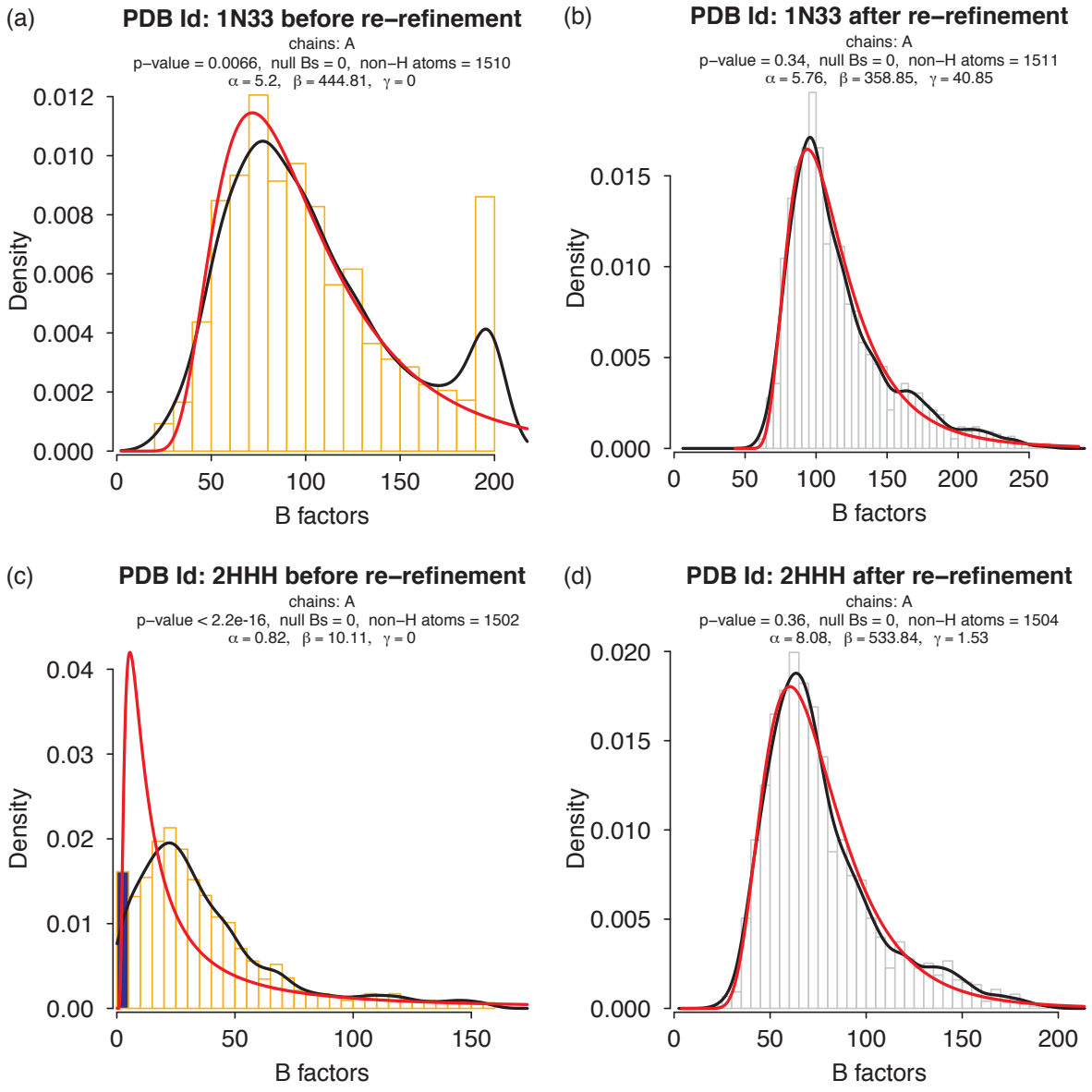


Figure 7.4: Example of B factor distributions for P atoms moved from *suspicious* (panels (a) and (c)) to *acceptable* (panels (b) and (d)) after re-refinement with protocol 2 defined in section 4.5. For a description of the plot generated by the validation protocol please refer to Figure 5.1.

The default model for PDB code 1N33 (panel (a) of Figure 7.4) is characterized by a strong bimodal distribution probably due to a constraint for the maximum allowed B factor

in the structure during refinement (around 200 $Å^2$). After re-refinement with protocol 2 (panel (b) of Figure 7.4) the agreement between the empirical B factor distribution and the fitted IGD* improves very considerably from 0.006 to 0.34 indicating that the B factor distribution of the re-refined structure is *acceptable*.

The default model for PDB code 2HHH (panel (c) of Figure 7.4) is instead characterized by a strong tie at the lowest B factor value as discussed in section 7.3.1. Similarly to the pdb-entry 1N33, after re-refinement with protocol 2 the B factor distribution of P-atoms becomes *acceptable* (panel (d) of Figure 7.4).

In contrast to what was observed for the distribution of R factor statistics in Figure 7.1, refinement protocol 2 did not always lead to B factor distributions with lower p-values than distributions from the default data set. This holds for the pdb-entries 1HNW, 1HNZ, 1HR0, 1XMQ, 1XNR, 2UU9, 2UUC and 2VQF, indicating a worsening of the fit between the empirical distributions of B factors and the estimated IGD*. However it is the only refinement protocol that guarantees that all B factor distributions for phosphate atoms are *acceptable*.

7.4 Ensemble Analysis with ESCET

7.4.1 Choice of the Re-refined Set of Models

The 29 ribosomal models obtained from protocol 2 (hereafter referred to as re-refined data set), which have the best R factor statistics (see section 7.2) and for which all B factor distributions for the phosphate atoms of the 16S rRNA are *acceptable* (see section 7.3), and the corresponding 29 deposited ribosomal models (hereafter referred to as default data set) were chosen for a comparative ensemble analysis with ESCET as described in section 4.6.

All the parameters required for the error model implemented in ESCET were extracted automatically from the pdb-file, with the exception of the pdb-enty 1XMO in the default set, where the data completeness parameter was specified manually. For this structure only 5% completeness was reported in the deposited PDB file, obviously impossible. The completeness was thus set to 93.1% as reported in the accompanying paper [87].

The average of the mean estimated uncertainties went from 0.47 ± 0.17 Å to 0.40 ± 0.13 Å for the default and the re-refined data sets, respectively, indicating an improvement in the precision of the atomic coordinates upon re-refinement. As shown in Table 7.2 for 26 out of 29 structures in the *T. thermophilus* data set the mean estimated uncertainty decreased after re-refinement with protocol 2. Only for the pdb-entries 1XMQ, 2UU9 and 2HHH it slightly increased.

7.4.2 Cluster Analysis

The CSI values computed for each pair of models in the data sets are presented in matrix form in Figure 7.5. The average CSI for 406 pairwise comparisons $(((29x29) - 29)/2)$ resulted to be 0.939 ± 0.056 and 0.929 ± 0.064 for the two data sets, respectively, indicating an increase in structural diversity in the re-refined data set with respect to the default set of models. The increase is probably related to a similarly lower uncertainties in the re-refined data set. This shows one of the features of the ESCET framework: when the models compared are more reliable in terms of coordinate errors, smaller absolute differences become more significant, reflecting an increase in the information content.

From the analysis of the ΔCSI values (panels (e) and (f)) different patterns of variability inside the data set after re-refinement emerge. In particular the pdb-entries 2UUB and 2UXD are the ones that increase most their structural diversity with the other structures in the data set.

The similarity matrices based on the CSI pairwise analysis shown in panels (a) and (c) of Figure 7.5 were then analysed by the hierarchical clustering algorithm implemented in ESCET and the resulting dendrograms for the default and re-refined data sets are shown in Figure 7.6.

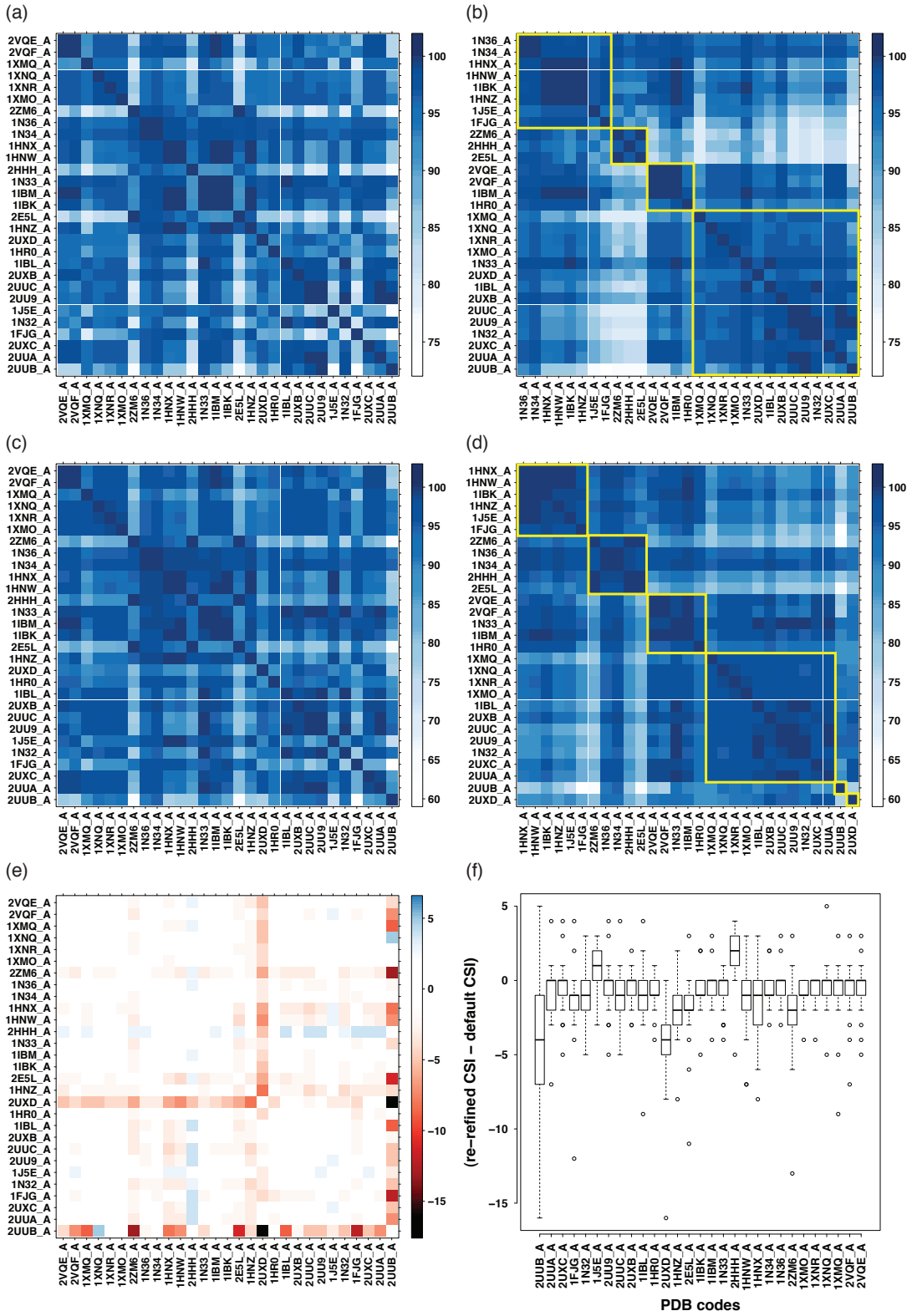


Figure 7.5: CSI matrix before (panels (a) and (c)) and after (panels (b) and (d)) clustering for the default data set (panels (a) and (b)) and the re-refined data set (panels (c) and (d)). For convenience, CSI-values were multiplied by 100. In panels (b) and (d) the clusters defined from the clustering procedure and shown in Figure 7.6 are highlighted by yellow boxes. The ΔCSI, defined as the difference between the matrix in panel (c) and the matrix in panel (a), is shown in panel (e) in matrix form and in panel (d) in box-plot form for each PDB code in the data set.

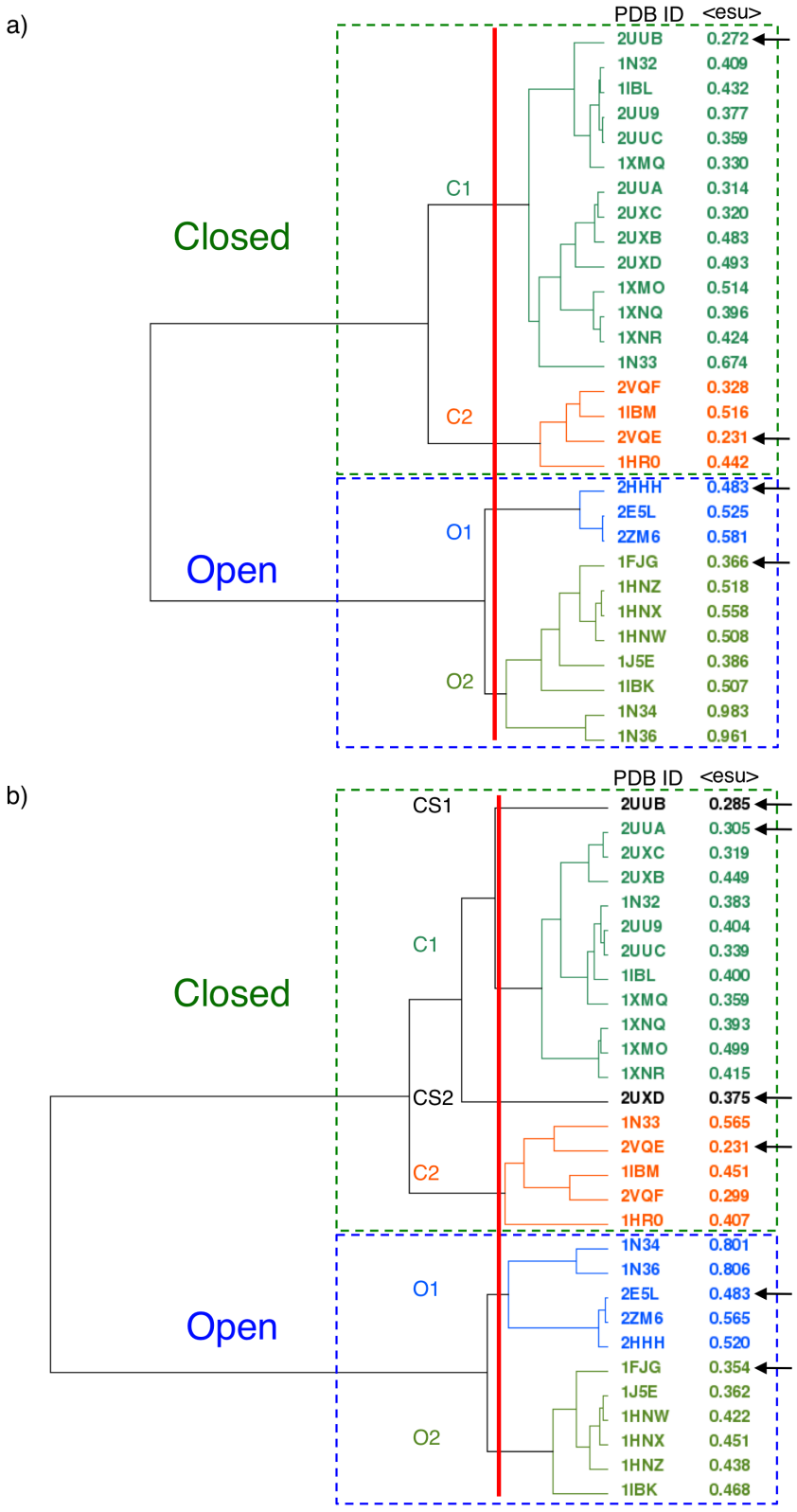


Figure 7.6: Dendrograms obtained from cluster analysis for default (panel (a)) and re-refined (panel (b)) data sets. For each pdb-entry the PDB code and the mean estimated uncertainty for phosphate atoms in Å are reported. Structures from the same cluster have the same colour. The clusters that contain closed and open conformations are highlighted with green and blue dashed boxes, respectively. A vertical red line is used to highlight the position at which the dendrograms were cut by the adaptive cutting rule. The representative structures from each group that are used in the rigid body analysis (see Figure 7.9) are highlighted with a black arrow. For the analysis of the results please refer to text in section 7.4.2.

For both default and re-refined data sets the two main branches of the dendrograms correspond to closed and open conformations of the 30S ribosomal subunits from *T. thermophilus*, respectively.

Further, the adaptive cutting rule [68] implemented in ESCET partitioned the ensemble of ribosomal structures in four clusters of similar conformers: two clusters of structures in closed conformation (clusters C1 and C2 in Figure 7.6) and two clusters of structures in open conformation (clusters O1 and O2 in Figure 7.6). However some differences are observed in the composition of the clusters when comparing the default data set to the re-refined data set. The pdb-entries 2UUB and 2UXD, which were included in the cluster C1 in the default data set, become singletons in the re-refined data set (clusters CS1 and CS2 in panel (b) of Figure 7.6). In addition, the pdb-entry 1N33 moves from cluster C1 to cluster C2 and the pdb-entries 1N34 and 1N36 move from cluster O2 to the cluster O1.

In the following sections the relationship between members of the different clusters is discussed with respect to the molecular components contained in the respective structures (see Table 7.3).

CL	PDB	gr	mRNA	tRNA	CG	N-CG	IF-1	SD	PAR	SCM	SRY	TAC	PCY	HYG	KSG
CS1	2UUB	XIV	✓	✓	✓†				✓						
C1	2UUA	XIV	✓	✓	✓†				✓						
	2UXC	XV	✓	✓	✓†			*	✓						
	2UXB	XV	✓	✓	✓†				✓						
	1N32	VIII	✓	✓		✓		*	✓						
	2UU9	XIV	✓	✓	✓†			*	✓						
	2UUC	XIV	✓	✓	✓†				✓						
	1IBL	VII	✓	✓	✓				✓						
	1XMQ	IX	✓	✓	✓†				✓						
	1XNQ	X	✓	✓	✓†				✓						
	1XMO	IX	✓	✓	✓†				✓						
	1XNR	X	✓	✓	✓†				✓						
CS2	2UXD	XV	✓	✓	✓†				✓						
C2	1N33	VIII	✓	✓		✓		*	✓						
	2VQE	XVI	✓	✓	✓†				✓						
	1IBM	VII	✓	✓	✓										
	2VQF	XVI	✓	✓	✓†				✓						
	1HR0	V					✓								
O1	1N34	VIII						*							
	1N36	VIII						*							
	2E5L	XIII						✓							
	2ZM6	XVII													
	2HHH	XII						*							✓
O2	1FJG	II							✓	✓	✓				
	1J5E	I													
	1HNW	IV										✓			
	1HNX	IV											✓		
	1HNZ	IV												✓	
	1IBK	VII							✓						

Table 7.3: Ligand content in 30S *T. thermophilus* ribosomal structures. CL stands for cluster name (the same used in panel (b) of Figure 7.6), gr stands for group (the same used in Table 7.2), mRNA stands for codon mRNA in the A site, tRNA stands for tRNA anticodon stem loop (tRNA ASL), CG stands for cognate tRNA ASL, N-CG stands for near-cognate tRNA ASL, IF-1 stands for initiation factor 1, SCM stands for SPECTINOMYCIN; SRY stands for STREPTOMYCIN; TAC stands for TETRACYCLINE; PCY stands for PACTAMYCIN; HYG stands for HYGROMYCIN B; KSG stands for KASUGAMYCIN, PAR stands for PAROMOMYCIN. The symbol '†' is used to higlight those structures where the ASL is not canonical or contain modified bases. The symbol '*' is used to highlight those structures where there is strong evidence in the electron density for a Shine-Dalgarno anti-Shine-Dalgarno (SD-aSD) interaction but no model has been built inside the density. Only for the pdb-entry 2HHH the presence of the SD-aSD interaction was mentioned in the accompanying publication. The order in which the pdb-entries are reported is the same of the dendrogram from cluster analysis shown in panel (b) of Figure 7.6.

Cluster C1 – Closed Conformations

Starting from the branch of the dendrogram with structures in closed conformation (green dashed boxes in Figure 7.6), the cluster C1 contains structures that show a most pronounced closed conformation. All these structures contain a codon mRNA moiety in the A site, a cognate or near-cognate tRNA ASL, and more importantly a molecule of paromomycin that it is known to facilitate 30S domain closure, resulting in better diffracting

crystal and a better defined ASL density [45, 91]. As anticipated above, small differences are detectable in the composition of the cluster C1 between the default and the re-refined data sets. The major differences are in the moving of the structures 2UUB and 2UXD to two distinct singletons (clusters CS1 and CS2 in Figure 7.6) and the moving of the pdb-entry 1N33 from the cluster C1 to the cluster C2 (discussed below).

Singleton CS1 – Closed Conformation

Regarding the PDB code 2UUB, already during the analysis of refinement statistics (see Figure 7.1) it became clear that the deposited structure did not show a good agreement with the experimental data since the refinement protocol 12, in which only bulk solvent correction is performed, gave R_{work} and R_{free} statistics equal to 0.448 and 0.451, respectively (as shown in table D.1 in Appendix D). Only protocols refining the atomic coordinates were able to reduce the R factor statistics below 0.30 (see also Figure 7.1).

The movements of coordinates are reflected by an RMSD without superposition of 1.36 Å for the phosphate atoms in the 16S rRNA moieties before and after re-refinement with protocol 2. This is very significant in comparison to the experimental uncertainties for a typical structure at this resolution ($\sim$ 0.40 Å). A comparative analysis of the default and re-refined structures for 2UUB revealed that a global shift and a rotation comparable to a rigid body movement of the whole 16S rRNA may have been occurred during the re-refinement. This observation is surprising since in the re-refinement protocol 2 no rigid body refinement was performed.

To cross validate the model obtained from protocol 2, the re-refined model was compared with the one available in the PDB_REDO database for which a rigid body re-refinement is always performed in the first stage of the refinement protocol. The RMSD without superposition between the phosphate atoms in the two 16S rRNA moieties from the refinement protocol 2 e from the PDB_REDO project is 0.32 Å, less than the coordinate error. If these two 16S rRNA moieties are analysed with ESCET they result to be identical within the error and one single rigid body is found, as shown in panel (a) of Figure 7.7. This is an interesting observation given that two different routes for refinement were taken. In fact, in the PDB_REDO project no simulated annealing refinement is performed and a different refinement program (REFMAC) is used.

Furthermore, if ESCET is instead used to compare the re-refined model from protocol 2 and the deposited model for 2UUB, besides the body and the head domain that form a unique rigid body, two rigid bodies are identified in the shoulder and the platform, respectively, and a concerted motion of the shoulder and platform domains is detected as shown in panel (b) of Figure 7.7.

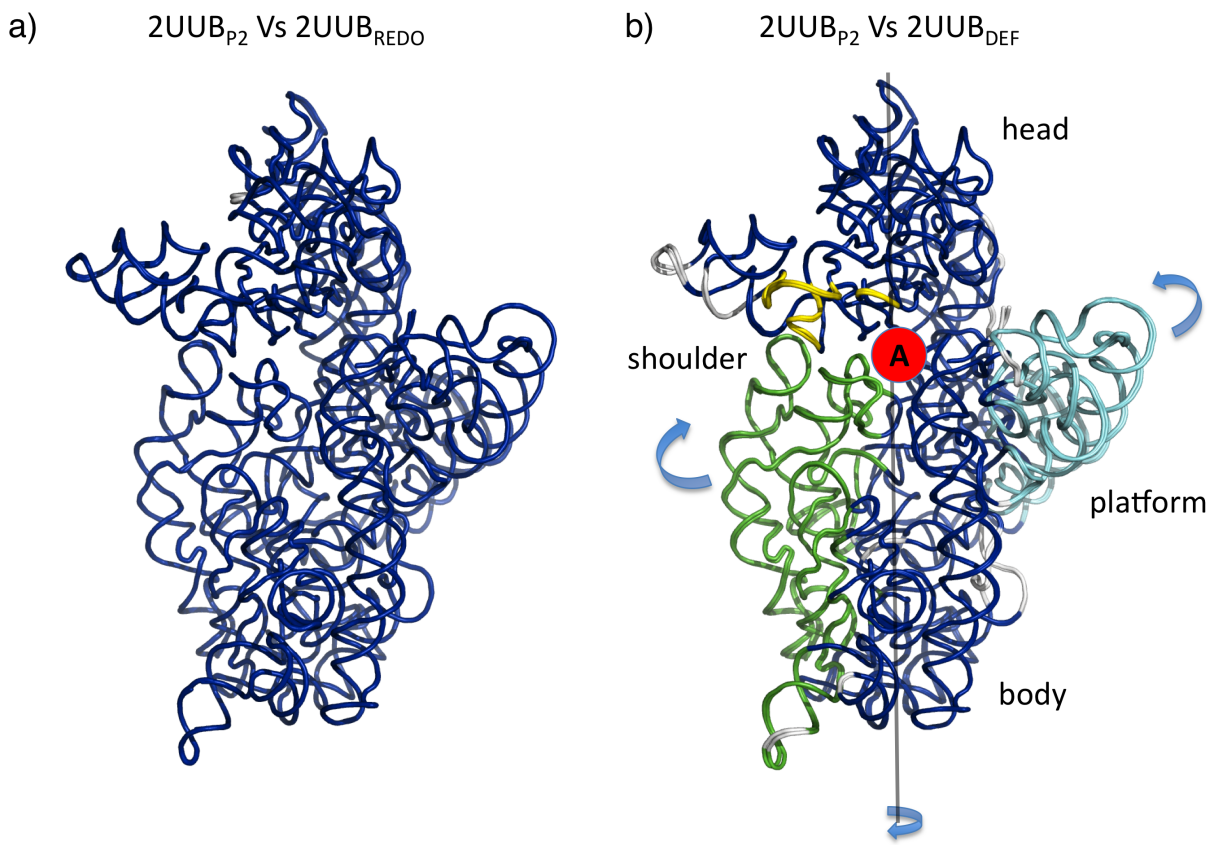


Figure 7.7: Rigid body analysis with ESCET of different models of PDB code 2UUB. Panel (a) shows the result of the analysis for the 16S rRNA moieties from protocol 2 and PDB_REDO. Panel (b) shows the result of the analysis for the 16S rRNA from protocol 2 and the deposited model. The decoding A site is highlighted with a red circle.

In particular, if the deposited structure is taken as reference, the shoulder domain moves away from the A site, while the platform moves towards from the A site. This can explain why after re-refinement the structure increases its structural diversity compared to the other structures in closed conformation.

Singleton CS2 – Closed Conformation

Similarly to 2UUB, also the pdb-entry 2UXD after re-refinement shows a larger conformational diversity when compared to the other structures belonging to cluster C1. A possible explanation can be found in the decrease in the mean estimated standard uncertainty value after re-refinement with protocol 2 (as shown in Table 7.2), in fact one of the largest decreases observed in the data set. It moves from 0.48 Å in the deposited model to 0.34 Å in the re-refined model. This significantly lowers the coordinate uncertainties in the model and makes smaller structural differences more significant in the ensemble comparison. Even if not as obvious as for the pdb-entry 2UUB (see above), probably during re-refinement the accumulation of many small differences possibly related to bias due to the incomplete-

ness of the model (it contains 1472 phosphate atoms against an average of 1507 in the data set) increased its structural diversity with the other models in closed conformation of cluster C1. The combination of the variations in atomic coordinates and the decrease in the coordinate uncertainties explains why after re-refinement the pdb-entry 2UXD becomes a singleton.

Cluster C2 – Closed Conformations

The cluster C2 contains structures that show a less closed conformation. In contrast to what is observed for the cluster C1, the composition of this group in terms of ligand content is more heterogeneous. In fact, in both data sets it contains structures with (pdb-entries 2VQE and 2VQF) and without paromomycin (pdb-entries 1IBM and 1HR0). As shown in Table 7.3 both pdb-entries 2VQE and 2VQF contain a codon mRNA in the A site, a modified tRNA ASL, and a molecule of paromomycin, while the pdb-entry 1IBM contains a codon mRNA in the A site and a tRNA ASL, and the pdb-entry 1HR0 contains only one molecule of IF1. In the clustering results from the re-refined data set this cluster contains also the pdb-entry 1N33, which in the default data set is located in the cluster C1. Similarly to pdb-entries 2VQE and 2VQF, the pdb-entry 1N33 contains a codon mRNA in the A site, a near-cognate tRNA ASL, and a molecule of paromomycin.

Even if the composition of cluster C2 is slightly heterogeneous, the results are in agreement with the data from the literature. It is in fact known that the recognition of the correct codon-anticodon base pair requires a change from an open to a closed conformation of the 30S ribosomal subunit and that the magnitude of the closure is smaller if paromomycin is not present in the crystal structure (pdb-entry 1IBM) [91]. A similar closure of the 30S subunit was observed if the IF1 was present in the crystal structure. Pdb-entries 2VQE and 2VQF should be in cluster C1 for their ligand content. Nevertheless, it has been reported that the 5' of the tRNA ASL in these structures is not visible in the electron density map, indicating a certain degree of disorder. This could be related to a reduced closure of the 30S subunit. The same holds for the pdb-entry 1N33 for which in the presence of paromomycin the affinity between the codon and near-cognate tRNA is lower than for the pdb-entry 1N32 (included in cluster C1) [93]. This would be in agreement with the observation that after re-refinement, the model 1N33 moved from cluster C1 to cluster C2.

Cluster O1 – Open Conformations

Taking now into consideration the branch of the dendrogram that includes 30S structures in open conformation (blue dashed boxes in Figure 7.6), the cluster O1 contains 30S ribosomal subunit crystallized with a Shine-Dalgarno anti-Shine-Dalgarno (SD-aSD) interaction. In the default data set it includes pdb-entries 2E5L, 2HHH and 2ZM6, while in the re-refined

data set also the pdb-entries 1N34 and 1N36 are included.

2E5L was solved to study the SD-aSD interaction in the ribosome (see panel (a) in Figure 7.8). In 2HHH a SD-aSD was not modelled in the electron density but reported in the accompanying article (see panel (b) in Figure 7.8). 2ZM6 instead is a dubious case since no accompanying publication is present and it has been solved by molecular replacement from 2E5L. Nevertheless no SD-aSD interaction is visible in the electron density map (see panel (c) in Figure 7.8). Instead electron density is clearly visible for the 3' rRNA folded back along the mRNA path even if not modelled (see panel (d) in Figure 7.8). It is worth noting that this structure should be in the O2 cluster (see below), but in both default and re-refined data set it is in the O1 cluster. It is not clear if this is an example of structural bias that the re-refinement protocol was not able to fix (since it does not contain the SD-aSD interaction) or instead an alternative form of open conformation.

As anticipated before, in the dendrogram obtained from the re-refined data set the pdb-entries 1N34 and 1N36 are also included in the cluster O1. An analysis of the electron density maps revealed that they both show clear density for a helix in the cavity where usually the SD-aSD is observed although it is not modelled here (see panels (e) and (f) in Figure 7.8).

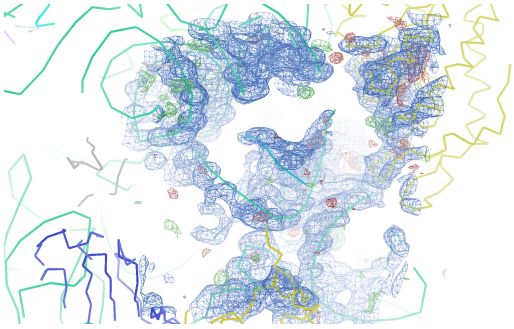
It is then reasonable to assume that this a successful example of structural de-bias thanks to the re-refinement procedure. In fact in the default data set these two structures are in the O2 cluster (discussed below) that contains also the pdb-entry 1J5E that was used to solve both 1N34 and 1N36 structures by difference Fourier method (see Table 7.1). In the accompanying paper the authors state that no density is visible for the near-cognate tRNA ASL but the head moves to a similar extent as with cognate ASL and no movement is observed in the shoulder. This movement would be consistent with conformationally disordered binding in the A site [93]. In light of the results obtained with ESCET it is reasonable to assume that the observed movement of the head is also affected by the presence of a SD-aSD-like interaction on the cavity between the head and the shoulder, where the canonical SD-aSD interaction is expected. This is supported by the fact that after re-refinement the 1N34 and 1N36 structures are structurally more similar to structures that contain the SD-aSD interaction than to structures in a closed (clusters C1 or C2) or completely open conformation (cluster O2). Surprisingly, the authors do not mention the presence of clear electron density for a SD-aSD interaction in the accompanying paper [93].

It should be noted that clear density for the SD-aSD interaction is visible also in the electron density maps for pdb-entries 1N32 (cluster C1 in the re-refined data set) and 1N33 (cluster C2 in the refined data set) that contain the same near-cognate of 1N34 and 1N36, respectively, but in presence of paromomycin (see panels (g) and (h) in Figure 7.8). Since no specific SD sequence was added during the crystallization procedure, a possible explanation for the observed unmodelled electron density in the SD-aSD pocket is that

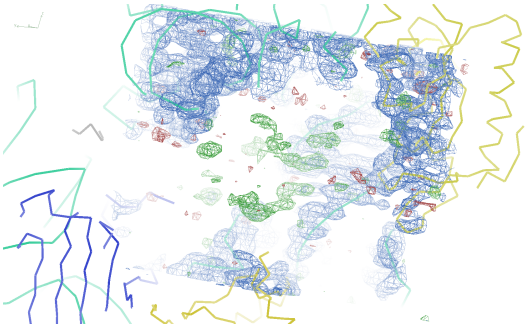
the near cognate tRNA ASL, with which the crystals were soaked, hybridizes with the aSD sequence on the 3' of the 16S rRNA.

To test this hypothesis, the hybridization energy between the aSD sequence CCUUUCU on the 3' of the 16S rRNA and the two near-cognate tRNA ASLs CUACCUUGAGGUGGUAG (present in the pdb-entries 1N32 and 1N34) and CACGCCUGGAAAGUGUG (present in the pdb-entries 1N33 and 1N36) was computed using the IntaRNA web server [29]. The hybridization energy obtained for the first near-cognate is -3 kcal/mol while for the second is -4.3 kcal/mol, indicating a favorable energy of hybridization between the two oligonucleotides. The computed favorable energy of hybridization makes then plausible the hypothesis of an interaction between the near-cognate tRNA moieties and the aSD sequence on the 3' of the 16S rRNA.

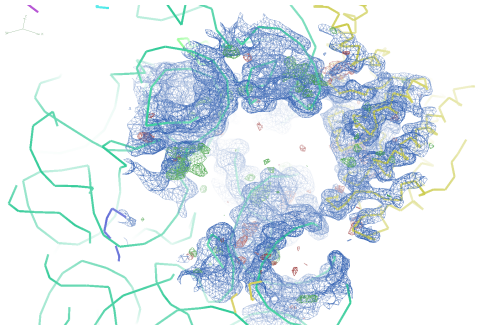
Since the presence of the SD-aSD like interaction in the pdb-codes 1N32 and 1N33 did not affect the closure of the 30S subunit in presence of paromomycin it is reasonable to assume that when the antibiotic is present in the crystal it masks the effect of any SD-aSD interaction, similarly to how paromomycin masks any difference in the closure of the 30S subunit between cognate and near-cognate tRNA (as shown in cluster C1).



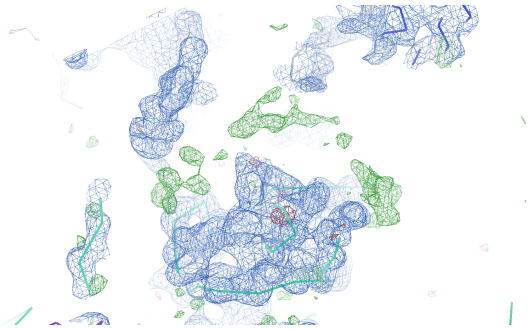
(a) PDB 2E5L, SD-aSD



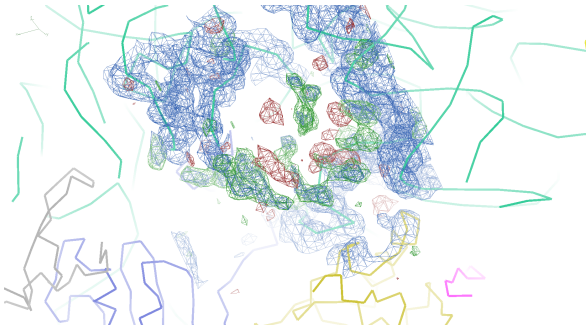
(b) PDB 2HHH, SD-aSD



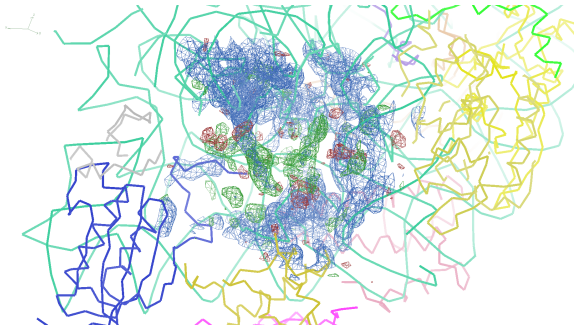
(c) PDB 2ZM6, SD-aSD



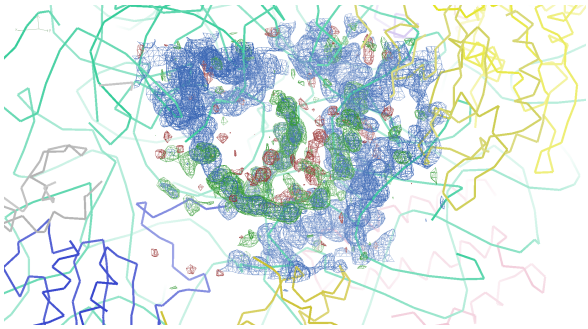
(d) PDB 2ZM6, P site



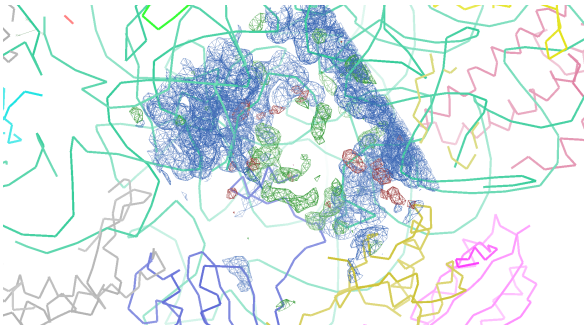
(e) PDB 1N34, SD-aSD



(f) PDB 1N36, SD-aSD



(g) PDB 1N32, SD-aSD



(h) PDB 1N33, SD-aSD

Figure 7.8: Electron density maps of selected cases. The blue density is the σ-weighted electron density map 2mFo-DFc contoured at +1.6 σ, the green and the red densities are the σ-weighted electron density maps 2mFo-DFc contoured at +3.0 σ and -3.0 σ, respectively. The rRNA and protein moieties are shown in ribbon representation and a different colour is used for each chain. These figures were made with Coot program [49].

Cluster O2 – Open Conformations

The cluster O2 contains structures in the classical open conformation. These models do not contain any mRNA or tRNA moieties, but they contain different types of antibiotics (pdb-entries 1FJG, 1HNW, 1HNX, 1HNZ and 1IBK) or are empty (pdb-entry 1J5E). In contrast to the models in cluster O1, in all the models in the cluster O2 the 3' of the 16S rRNA is folded back along the mRNA path in the P site. The main difference in the composition of this cluster between the default and the re-refined data sets is the moving of the pdb-entries 1N34 and 1N36 to the cluster O1 as described above. Also in this case the results are in agreement with the literature since it has been observed that the antibiotics alone are not sufficient to induce a closure of the small subunit [23, 32, 91, 134].

7.4.3 Rigid Body Analysis

The representative models from each group determined by the cluster analysis (pdb-entries highlighted with a black arrow in Figure 7.6) were then used by the genetic algorithm to identify rigid bodies and flexible regions on the 16S rRNA. The results for the default and the re-refined data sets obtained by using a ϵ_{low} of 1.5 are shown in Figure 7.9 (panel (a) and (b), respectively).

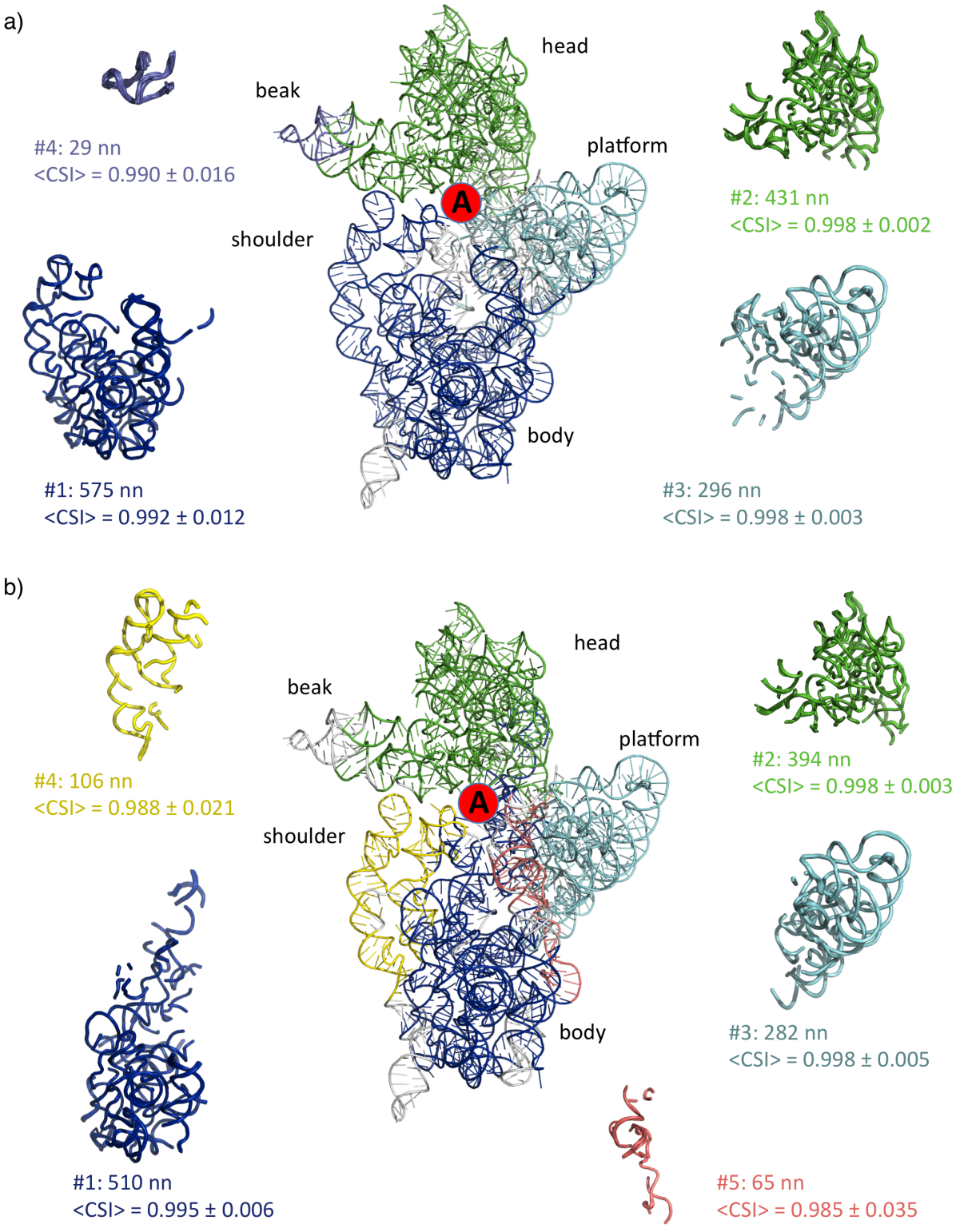


Figure 7.9: Identification of rigid bodies on the 16S rRNAs from the 30S data set before (panel (a)) and after re-refinement (panel (b)). In the center of each panel a schematic view of the 16S rRNA is shown with a cartoon representation. Each rigid body is coloured with a different colour, while flexible regions are light grey coloured. The decoding A site is highlighted with a red circle. For each rigid body, the superposition of the 29 fragments from each structure in the data set is shown in a ribbon representation together with the number of nucleotides and the average CSI for all 406 pair-wise comparisons.

In the default data set ESCET identified four rigid bodies from the comparison of the representative pdb-entries 2UUB, 2VQE, 2E5L and 1FJG. The first rigid body (blue fragment in panel (a) of Figure 7.9) consists of 575 nucleotides and includes the body and the shoulder domains. The second rigid body (green fragment in panel (a) of Figure 7.9) comprises 431 nucleotides and includes the head domain. The third rigid body (cyan fragment in panel (a) of Figure 7.9) consists of 296 nucleotides and corresponds to the platform domain. The fourth rigid body (slate blue fragment in panel (a) of Figure 7.9) consists of 29 nucleotides and includes the terminal part of the beak of the head domain. The presence of the beak as a rigid body is an artefact due to the fact that this region of the 16S rRNA is usually highly disordered and consequently is characterized by the highest B factor values in the structure. This affects the weighted distance matrix and it results in the identification of the beak as a rigid body. This is anyway consistent with the definition of rigid body in ESCET as conformationally invariant region (i.e. a region for which all interatomic distances are identical within the error) [110–112].

In the re-refined data set ESCET identified five rigid bodies from the comparison of the representative pdb-entries 2UUB, 2UUA, 2UXD, 2VQE, 2E5L and 1FJG. The first rigid body (blue fragment in panel (b) of Figure 7.9) consists of 510 nucleotides and includes the body domain and fragments of helices from the shoulder domain (H17 and H18), the platform domain (H21 and H26a) and the head (H36 and H40), as shown in Figure 7.10. The second rigid body (green fragment in panel (b) of Figure 7.9) consists of 394 nucleotides and includes the head domain. The third rigid body (cyan fragment in panel (b) of Figure 7.9) comprises 282 nucleotides and corresponds to the platform domain. The fourth rigid body (yellow fragment in panel (b) of Figure 7.9) consists of 106 nucleotides and corresponds to the shoulder domain. The fifth rigid body (salmon fragment in panel (b) of Figure 7.9) comprises 65 nucleotides and includes part of helix 44 (nucleotides 1401-1413, 1416-1417, 1482-1489) that is important for the interaction with the 50S subunit in the 70S ribosome complex. Other fragments of this rigid body are located in the body domain. Given its relatively small size and its fragmentation it is probably due to some bias introduced by one of the representative structures (i.e. by removing the pdb-entry 2E5L the fifth rigid body appears only on the H44 without any other fragment on the body domain).

Moving from the default data set to the re-refined data set a new rigid body appears on the shoulder of the 16S rRNA, the rigid body on the beak (fourth rigid body on panel (a) of Figure 7.9) disappears and a new rigid body appears on the H44 (fifth rigid body on panel (b) of Figure 7.9). While these two small rigid bodies (on the beak and on the helix 44) can be considered as artifact due to noisy regions (i.e. high B factors or local structural differences in one of the representative structures) the new rigid body on the shoulder is in agreement to what was previously observed in the literature (see below). This corresponds to a gain of structural information after the re-refinement. In fact, as previously discussed

in section 7.4.2, when the models compared are more reliable, smaller absolute differences become more significant, reflecting an increase in the information content.

Moreover if the distribution of rigid bodies is mapped on the secondary structure of the 16S rRNA as shown in Figure 7.10, the location of the rigid bodies from the re-refined data set shows a good agreement with the modern nomenclature for domains on the 16S rRNA. The first (blue) and the fourth (yellow) rigid bodies, which in the classical nomenclature correspond to the body and the shoulder domain respectively, include the majority of the 5' domain. The second rigid body (green), which in the classical nomenclature corresponds to the head domain, includes the majority of the 3' major domain. The third rigid body (cyan), which in the classical nomenclature corresponds to the platform domain, includes the majority of the central domain. From the analysis of the mapping of the rigid bodies on the secondary structure in panel (b) of Figure 7.10 it emerges also that there is a strong correspondence between the definition of a rigid body and the patterns of base-base and base-backbone interactions in the 16S rRNA, although these interactions are not directly used in the rigid body analysis.

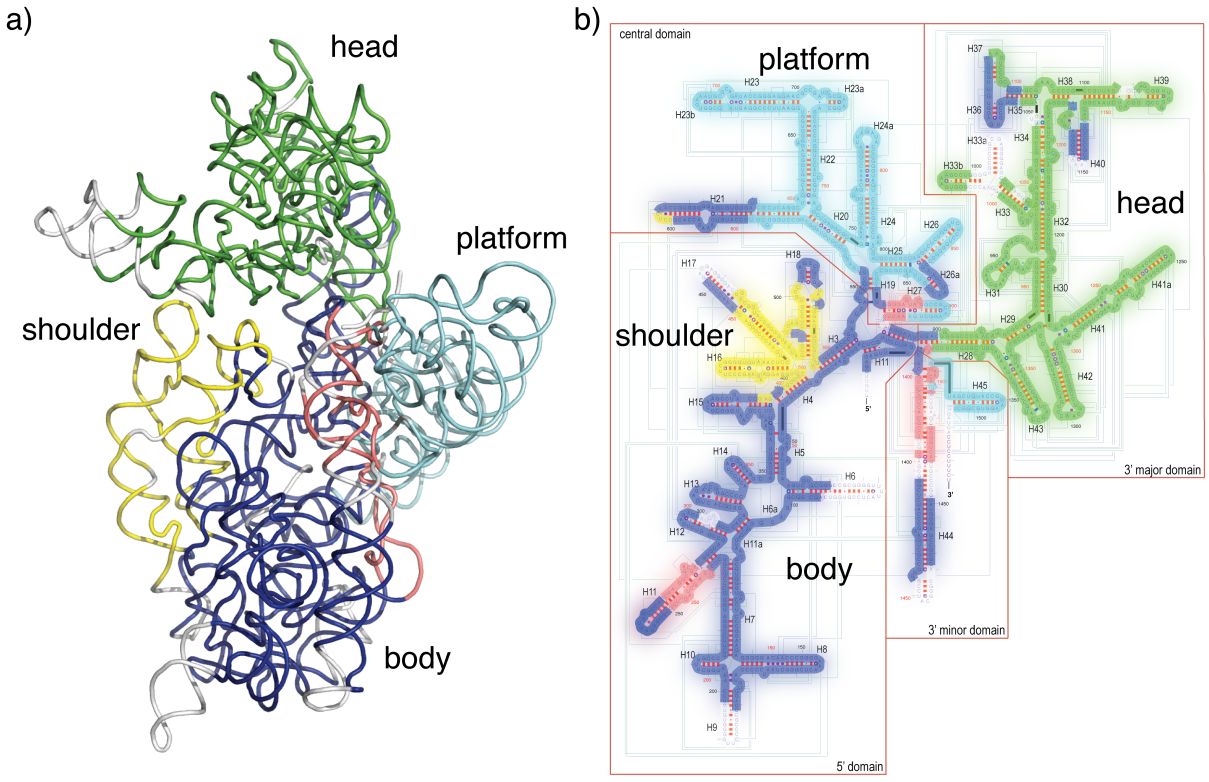


Figure 7.10: Mapping of the rigid body obtained from ESCET (panel (a)) on the secondary structure obtained from the comparative RNA web (CRW) site [30] (panel (b)). In panel B both classical names (body, shoulder, platform and head) and modern names (5' domain, central domain, 3' minor domain and 3' major domain) for the different 16S rRNA domains are given. In panel (b) the light blue and red lines represent the base-base and base-backbone interactions, respectively.

The four main rigid bodies mapped to the body, the shoulder, the platform and the head

of the small subunit correspond to domains that have been observed to undergo conformational changes during the decoding process.

Looking at the results obtained from ESCET on the models from protocol 2 and using the pdb-enty 1J5E (empty structure in open conformation) as the reference, all the models in the cluster C1 and the singletons CS1 and CS1 show the strongest closure of the 16S rRNA. This conformational change involves that, with respect to their position in the empty form (PDB code 1J5E), the head, the shoulder and the platform domains move towards the A site, where the decoding process takes place. A rotation angle of 2.0°, 1.6°, 1.0° is measured if the representative conformer 1IBL is compared to the fully open conformer 1J5E for the head, platform and shoulder domains, respectively, in relation to the body domain. As previously reported all these structures contain a molecule of the antibiotic paromomycin. Its property of enhancing the closure of the small subunit was observed in 2001 with the pdb-entry 1IBL [91] and thereafter it has been used extensively in structural studies where the codon-anticodon interaction was investigated [45, 86, 87, 93, 132]. It should be noted that in the accompanying paper of the pdb-entries 1N32 and 1N33, only the movement of the head and the shoulder is reported, while from the results obtained with ESCET also a discrete movement of the platform is observed, in agreement with the other models in the cluster C1.

Similarly to cluster C1, taking as reference the empty 30S structure (PDB code 1J5E) also the conformational changes observed from the cluster analysis for models in the C2 cluster are in agreement with those reported in the accompanying publications. For all the models in this group the closure movement of the shoulder, the head and the platform towards the A site is still present but it is not as pronounced as for models in cluster C1. A rotation angle of 1.0°, 1.1°, 0.7° is measured if the representative conformer 2VQE is compared to the fully open conformer 1J5E for the head, platform and shoulder domains, respectively, in relation to the body domain. Even if for the pdb-entries 1N33, 2VQE and VQF a molecule of paromomycin is present in the crystal complex, the codon-anticodon interaction is not strong enough to induce the strongest closure conformation [76, 93]. These conformational changes are of the same magnitude of those observed in pdb-entries 1IBM and 1HR0 where no paromomycin is present, but only mRNA and cognate tRNA (for the PDB code 1IBM) or initiation factor I (for the PDB code 1HR0).

As previously discussed, the cluster O1 contains 30S ribosomal models that contain the SD-aSD interaction. However in the accompanying publications no global conformational changes are described [67, 108]. From the results obtained with ESCET, it is possible to observe that in the re-refined models for the pdb-entries 2HHH and 2E5L, which were solved with a specific SD sequence, the shoulder domain moves towards the A site while the head and the platform move backwards towards the pocket on the back side of the 16S rRNA, between the head and the platform. The head domain is also affected by a lateral

translation towards the platform domain. Similar conformational changes are observed for the pdb-entries 1N34 and 1N36, in agreement with the result of the clustering in section 7.4.2. In these models the head domain shows a more pronounced movement toward the A site, probably due to the presence of a disordered near-cognate tRNA ASL in the A site as suggested by the authors [93]. A rotation angle of 2.5°, 2.3°, 0.4° is measured if the representative conformer 2E5L is compared to the fully open conformer 1J5E for the head, platform and shoulder domains, respectively, in relation to the body domain. It should be noted that only after re-refinement the conformational changes observed for the pdb-entries 1N34 and 1N36 are of the same nature as those observed for the pdb-entries 2HHH and 2E5L where the SD-aSD interaction is documented. The reliability of these findings is further supported by the fact that after re-refinement the R factors for 1N34 and 1N36 greatly improved (as shown in Figure 7.1), indicating a better agreement between the models and their experimental data.

The cluster O2 contains models in the classical open conformation for which the pdb-entry 1J5E is the representative model [134]. For some of these models only small conformational changes are detectable. Pdb-entries 1HNW (which binds tetracycline) and 1HNX (which binds pactamycin) do not show significant conformational changes and are very similar to 1J5E. In the pdb-entry 1HNZ (which binds hygromycin) instead the head, the shoulder and the platform move slightly towards the A site (not reported in the accompanying publication). This can be due to the fact that hygromycin binds close to the top of helix 44, in a region that contains the A, P and E sites for tRNA [23]. It is probable that its presence is sensed by the three domains (head, shoulder and platform). Very similar conformational changes are observed for the pdb-entry 1IBK (not observed in the accompanying paper), which binds only paromomycin that is known to bind the 16S rRNA in the major groove of helix 44 and to flip out bases A1492 and A1493 involved in the sensing of the first two positions of the codon-anticodon interaction [32, 91]. Even if the presence of paromomycing is not sufficient to induce a closed conformation of the small subunit, it looks like it prepares the 16S in a conformation favorable for the closure of the subunit. In pdb-entry 1FJG (which binds streptomycin, spectinomycin and paromomycin) the shoulder and platform tilt forward the A site while the head tilts back away from it, in agreement with what is reported in the accompanying publication [32].

Chapter 8

Conclusions and Perspectives

8.1 Validation of B Factor Distributions

In chapter 5 by means of a Bayesian approach it was hypothesised that the distribution of isotropic B factors in a crystallographic model should follow an IGD*. Since the majority of the B factor distributions in the large data set of 15998 protein models fulfilled the IGD* assumption at different hierarchical levels (whole asymmetric unit cell content, single chains and single domains), the IGD* can be considered a valid reference distribution for B factor distributions and it can be used for their validation as implemented in the protocol described in section 4.3. The time required by the validation protocol to validate a distribution of B factors varies in function of the number of atoms, but it has been found to be in the order of seconds, allowing the use of the protocol in routine validation procedures. Furthermore the fact that all the protein models analysed are at a resolution equal to or higher than 2 Å ensures that, given the high *data-to-parameter* ratio, the observed B factors reflect more the experimental diffraction data than the restraints applied during refinement. This observation guarantees that the IGD* is a property itself of the B factors and not an artefact produced by the restraints used in refinement.

The strong correlation observed between the estimates of the α and β parameters can be used to identify, group and categorise outliers in function of extreme values for the α and β IGD* parameters. In section 5.3 it was found that models in a given group of outliers are characterized by peculiar common artefacts. These artefacts include strong ties at the lowest B factor value (see sections 5.3.1 and 5.3.4), too low variances of the B factors probably due to too tight restraints (see section 5.3.2), or mean square displacements $\langle u^2 \rangle$ saved in the coordinate files instead of B factor values (see section 5.3.3). These observations can be used in the future to automatically flag a newly determined structure if the α and β parameters of the IGD* fitted to its B factor distribution lie in one of these groups.

The p-value obtained from the KS-test is the main measure used to flag a given empirical

B factor distribution as *suspicious* (if p-value < 0.01) or *acceptable* (if p-value ≥ 0.01). The analysis should be performed at different hierarchical level: whole asymmetric unit content, single chains, single domains. It was in fact observed that several B factor distributions in a crystal model are affect by a strong multimodality due to a different packing environment. This strong multimodality usually can be solved by taking into account single chains or single domains. If the p-value remains lower than 0.01 then the B factor distribution is likely to be affected by some strong artefact. The orthogonal statistics introduced in section 5.5 can then be used to find the reasons that caused the distribution under analysis to be *suspicious*.

A standard re-refinement procedure resulted to be a valid approach to rescue those structures with a *suspicious* B factor distribution (see section 5.4.5). Those models whose B factor distribution was *suspicious* and could not be explained by the hierarchical model proposed, even after re-refinement, corresponded usually to cases with serious problems with the deposited atomic coordinates or structure factors. If the B factor distribution remains *suspicious* even after re-refinement of the structural model, then a more detailed analysis is required to understand the deviance from the expected IGD*.

As shown in chapter 6, the validation protocol for B factor distributions can be useful in an ensemble analysis performed with ESCET. In fact, it allows to check the quality of the distribution of B factors from the models in the ensemble under analysis and it allows the identification of strong artefacts in the B factor distribution that can affect the results of the rigid body analysis. Moreover a standard re-refinement protocol is observed to improve the quality of the results of the ensemble analysis performed with the ESCET framework if the starting data set contains models with *suspicious* B factor distributions. These results confirm that the reliability of the outcome of a comparative structural analysis strongly depends on the reliability of the models under analysis in the starting ensemble. Since each macromolecular model deposited into the PDB is one of several possible interpretations of the diffraction data, before any comparative analysis it is important to use validation methods to judge its reliability and to maximise its agreement with the experimental data and with the prior information available in structural databases.

8.2 Ensemble Analysis of Ribosomal Structures

The large number of ribosomal structures available in the PDB allowed the definition of a working set of 29 structures of the small subunit from *T. thermophilus* for testing the validation protocols for B factor distributions and the ESCET framework on RNA structures.

Since the presence of structural bias due to the phasing method used was suspected, and since some models were found to have *suspicious* B factors, 13 re-refinement protocols

were applied to the selected data set of ribosomal structures. Amongst the re-refinement protocols applied, the protocol 2, which combines individual ADP refinement with TLS refinement, is the one that outperforms all the others, confirming, as already observed in the PDB_REDO project, that TLS refinement on average gives better refinement statistics. In addition, protocol 2 is the only refinement protocol that produced *acceptable* B factor distributions for all the models in the data set.

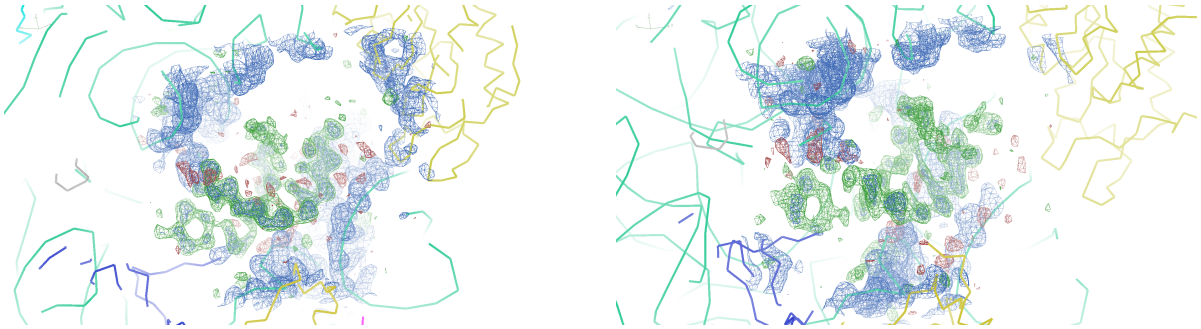
The large differences in R factor statistics and p-values from the KS-test between models obtained from different re-refinement protocols highlight how at low resolution different refinement parameterisations can greatly affect the final refinement statistics and the distribution of B factors. Therefore at low resolution the application of the same re-refinement protocol to all the structures in an ensemble before any comparative analysis can be used as a normalisation procedure, with effects at both coordinates and B factors levels. In addition, at low resolution there could be issues of structural bias if all the structures in the data set were solved by molecular replacement using the same searching model. These suspects were confirmed by the results obtained from the cluster analysis (see below) applied to the ensemble of ribosomal structures. In these cases an aggressive refinement protocol (e.g. simulated annealing refinement) can be used in the attempt of de-biasing as much as possible the models in the data set.

The comparative analysis of the IGD* parameters computed by the validation method for B factor distributions applied to the models obtained from different refinement protocols allowed to detect a correlation between the magnitude of the α parameter and the model used to refine the B factors. On average it has been observed that the stronger the restraint used the higher is the α IGD* parameter estimated. In the particular case of the data set of models obtained with TLS refinement, the α parameter shows a distribution similar to the one observed in the analysis of the protein data set (shown in Figure 5.3), while the β parameters are at least three times higher than the corresponding values. This difference has been attributed to a higher average B factor in ribosomal structures.

Moving to the results obtained with the ESCET framework, the cluster analysis was confirmed as an effective way to automatically rationalise the structural information content of the data set of ribosomal structures. In fact structures grouped together in the same cluster resulted to have similar conformations as a result of the ligand content, for a total of two clusters containing small ribosomal subunits in closed conformation (clusters C1 and C2 in Figure 7.6) and two clusters containing small ribosomal subunits in open conformation (clusters O1 and O2 in Figure 7.6).

Besides being effective in rationalising the information content in an ensemble of structures, from the comparative analysis of the results of the cluster analysis on the data set of models as deposited into the PDB and after re-refinement with protocol 2 (which produced the best models in terms of refinement statistics and B factor distribution), the ESCET

framework resulted also a useful tool in discovering unexpected features of the data and extracting new information that can be used for the formulation of new hypothesis (see section 7.4.2). The observation that after re-refinement the pdb-entries 1N34 and 1N36 moved from cluster O2 (containing structures in fully open conformation) to cluster O1 (containing structures in open conformation and with a modelled Shine-Dalgarno / anti-Shine-Dalgarno interaction) led to the discovery of new electron density in these structures in correspondence of the Shine-Dalgarno / anti-Shine-Dalgarno (SD-aSD) interaction site. This density was not modelled in the deposited structures and not mentioned in the accompanying paper. Since these pdb-entries were obtained by molecular replacement from the pdb-entry 1J5E (contained in the cluster O2), the change of cluster after re-refinement confirmed the existence of structural bias in the deposited ribosomal structures. This finding brought also to the discovery of other models with unexpected electron density for the SD-aSD interaction as shown in Figure 8.1.



(a) PDB 2UU9, SD-aSD **(b)** PDB 2UXC, SD-aSD

Figure 8.1: Electron density maps of selected cases. The blue density is the σ-weighted electron density map 2mFo-DFc contoured at +1.6 σ, the green and the red densities are the σ-weighted electron density maps 2mFo-DFc contoured at +3.0 σ and -3.0 σ, respectively. The rRNA and protein moieties are shown in ribbon representation and a different colour is used for each chain. These figures were made with Coot program [49].

The interaction between the anti-Shine-Dalgarno sequence on the 3' of the 16S rRNA and the near-cognate tRNA has been proposed as the possible cause for the presence of not modelled electron density in correspondence of the SD-aSD interaction site. A SD-aSD interaction was found also in pdb-entries 2UXC, 1N32, 1N33 and 2UU9 but no movement to different clusters was observed for these models after re-refinement with protocol 2. It has been hypothesised that the absence of change of cluster is due to the presence of the antibiotic paromomycin that is known to enhance the closure of the small subunit and to mask the structural differences between the presence of cognate or near-cognate tRNA ASL.

Also the results from the rigid body analysis improved after re-refinement of the ribosomal models with protocol 2, resulting in the description by a different rigid body of

each of the four classical domains of the 30S subunit (i.e. body, platform, head and shoulder domains) as shown in figures 7.9 and 7.10. All these domains are known to undergo structural movements upon codon-anticodon interaction as discussed in section 7.4.3. The observation that after re-refinement with protocol 2 a new rigid body appeared on the shoulder domain of the ribosome is explained by the fact that when the models compared are more reliable (all models with better refinement statistics and with *acceptable* B factor distributions), smaller absolute differences become more significant. This is also reflected by a decrease in the average CSI computed across the models in the ensemble, indicating an increase in detectable structural diversity.

8.3 Perspectives

The validation protocol for B factor distribution described here can be used in the future for the validation of models already deposited into the PDB or of newly solved macromolecular structures prior to the deposition in the PDB. Moreover, it can be integrated in frameworks for structural analysis, such as the ESCET framework, and used for the evaluation of the quality of the distribution of B factors in an ensemble of models before performing any analysis.

Following the classification of quality indicators reviewed in [70], the proposed method can be considered a global validation statistic, since it gives information about the overall distribution of B factors in a protein model. At the same time, since the IGD* assumption is not used in any step of refinement, it would be considered a 'strong' validation criterion.

Because the IGD* assumption was satisfied by the majority of protein structures at high resolution analysed in chapter 6 it can be used in refinement under the form of prior information as source of restraints for B factors. The refinement of structures at low resolution will particularly benefit of this new type of restraint for the reason that at low resolution there are no sufficient observations to allow the refinement of B factors at atomic level. Moreover, due to the fact that the α and β IGD* parameters are highly correlated, the knowledge of one of these parameters is sufficient to estimate the second, with computational advantages. In case the IGD* assumption is used in refinement as source of restraints for B factors, the proposed validation method will be considered a 'weak' validation criterion [70] since it will just check the consistency between the observed distribution of B factors and the prior information used to restrain them.

The statistical model behind the IGD* validation method is based in strong assumptions like the independence between B factors in a model and their belonging to the same IGD*. This limits its applicability in case TLS refinement (except particular cases like the ensemble analysis of ribosomal structures, where only phosphate atoms where taken into account).

In the future more complex statistical models should be introduced in order to consider the correlations between B factors, extending the applicability of the validation tool to structures refined with TLS groups.

The results of the analyses showed in the presented work confirm the importance of the availability of the experimental data. In fact for both protein (see chapters 5 and 6) and ribosomal (see chapter 7) data sets an automated re-refinement procedure has been shown to be sufficient to improve the quality of the models in terms of refinement statistics and B factor distributions, normalizing the models and allowing the extraction of more and new structural information from the ensemble of models.

When experimental data are available, it is possible to rebuild and re-refine models with the current best technology/knowledge. This is not only useful for individual models but also for the analysis of ensemble of models, producing a broad picture of the structural properties of the macromolecule under investigation and providing new information that can bring new knowledge on the relation between the structure of the macromolecule and its function.

Appendices

Appendix A

Bvalid.R script

```
#  Bvalid.R, an R script for the validation of B factor distributions. The main function if Bvalid().
#  To use this script open an R session and type: source("/<path_to_script>/Bvalid.R")
#
#  Created by Jacopo Negroni.
#  Copyright (c) 2011 EMBL. All rights reserved.

# Importing required packages
library(pscl)       # Available in Cran (http://mirrors.softliste.de/cran/)
library(flexmix)    # Available in Cran (http://mirrors.softliste.de/cran/)
library(bio3d)      # Not available in Cran. Please visit: http://mccammon.ucsd.edu/~bgrant/bio3d/

# Function to compute a selection vector for residue ID.
#  -Input: 1) a bio3d atom structure
#          2) a vector with the residue codes to be selected
#  -Output: 1) a boolean vector of the same length of the bio3d atom structure given in input
selectResid <- function (atom.struct, code) {
  selection <- rep(FALSE,length(atom.struct[,"resid"]))
  for (i in code) {
    selection <- selection | as.character(atom.struct[,"resid"]) == i
  }
  selection
}

# Function to compute a selection vector for atom name.
#  -Input: 1) a bio3d atom structure
#          2) a vector with the atom types to be selected
#  -Output: 1) a boolean vector of the same length of the bio3d atom structure given in input
selectElety <- function (atom.struct, elety) {
  selection <- rep(FALSE,length(atom.struct[,"elety"]))
  for (i in elety) {
    selection <- selection | as.character(atom.struct[,"elety"]) == i
  }
  selection
}

# Function to select hydrogen atoms in a Bio3D atom structure.
#  -Input: 1) a bio3d atom structure
#  -Output: 1) a boolean vector of the same length of the bio3d atom structure given in input
selectHydrogens <- function(atom.struct) {
  selection <- rep(FALSE,length(atom.struct[,"elety"]))
  for (i in 1:length(selection)) {
    elety <- substr(atom.struct[i,"elety"],1,1)
    if (elety == "H") {
      selection[i] <- TRUE
    }
  }
  selection
}

# Vectors of IUPAC codes for amino- and nucleic-acids. Only unambiguous codes are taken into
#   account (e.g: each code identifies only a single chemical compound).

# Vector of IUPAC three letter aminoacid nomenclature. Source:
#   http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/iupac_aa_abbreviations.html
aa.codes <- c("ALA", "ARG", "ASN", "ASP", "CYS", "GLU", "GLN", "GLY", "HIS", "ILE",
    "LEU", "LYS", "MET", "PHE", "PRO", "SER", "THR", "TRP", "TYR", "VAL"
)

# Vector of IUPAC three letter nucleic acid nomenclature. Source:
#   http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/iupac_nt_abbreviations.html
nn.codes <- c("U", "A", "T", "C", "G")

# Vector of atom types for backbone atoms.
backbone.elety <- c("CA", "C", "O","N")

# Function for the computation of the alpha and beta parameters of an Inverse-Gamma Distribution
# (IGD).
#  -Input: 1) a vector x of positive and non-zero numeric values
#  -Output: 1) a vector with the alpha and beta IGD parameters computed from x
igdParams <- function(x) {
  alpha <- (((mean(x))^2)/var(x)) + 2
  beta <- mean(x)*(alpha - 1)
  parameters <- c(alpha,beta)
```

```
  parameters
}

# Function to compute the vector of partial first derivatives of the negative log likelihood
#  function for a shifted Inverse-Gamma Distribution (IGD*).
#  -Input: 1) alpha IGD* parameter
#          2) beta IGD* parameter
#          3) gamma IGD* parameter
#          4) a vector x of positive and non-zero numeric values.
#  -Output: 1) vector of partial first derivatives
sigdLogLikFD <- function(alpha, beta, gamma, x) {
  n  <- length(x)
  dp <- vector(len=3)
  dp[1] <- -n*log(beta) +sum(log(x+gamma)) + n*digamma(alpha)
  dp[2] <-  -n*alpha/beta + sum(1/(x+gamma))
  dp[3] <- (1+alpha)*sum(1/(x+gamma)) - beta*sum(1/(x+gamma)^2)
  return(dp)
}


# Function to compute the matrix of second derivatives of the negative log likelihood function
#  -Input: 1) alpha IGD* parameter
#          2) beta IGD* parameter
#          3) gamma IGD* parameter
#          4) a vector x of positive and non-zero numeric values.
#  -Output: 1) vector of partial second derivatives
sigdLogLikSD <- function(alpha, beta, gamma, x) {
  n   <- length(x)
  dp2 <- matrix(vector(len=9),nrow=3)
  dp2[1,1] <- n*trigamma(alpha)
  dp2[1,2] <- -n/beta
  dp2[1,3] <- sum(1/(x+gamma))
  dp2[2,3] <- -sum(1/(x+gamma)^2)
  dp2[2,1] <- dp2[1,2]
  dp2[3,1] <- dp2[1,3]
  dp2[3,2] <- dp2[2,3]
  dp2[2,2] <- n*alpha/beta^2
  dp2[3,3] <- -(1+alpha)*sum(1/(x+gamma)^2) + 2*beta*sum(1/(x+gamma)^3)
  return(dp2)
}

# Function to compute the average negative log likelihood for a IGD*.
#  -Input: 1) a vector with the three alpha, beta and gamma IGD* parameters
#          2) a vector with positive values for which compute the average negative log likelihood.
#  -Output: 1) the computed average negative log likelihood for a IGD*
avrgSigdLogLik <- function(sigd.params, x) {
  n <- length(x)
  alpha <- sigd.params[1]
  beta <- sigd.params[2]
  gamma <- sigd.params[3]
  x <- x + gamma
  avrg.sigd.log.lik <- -alpha*log(beta) + lgamma(alpha) + ((alpha+1)*sum(log(x)) + beta*sum(1/x))/n
  return(avrg.sigd.log.lik)
}

# Function to compute a robust MLE to fit a IGD* to a given vector of positive values.
#  -Input: 1) a vector x of positive and non-zero numeric values.
#          2) a boolean value to switch the select as best fit the one with the lowest
#             mean variance of the estimated IGD* parameters (default to FALSE).
#  -Output: 1) a list with the following elements:
#             1.a) alpha = alpha parameter of the estimated IGD* that best fits the data
#             1.b) beta = beta parameter of the estimated IGD* that best fits the data
#             1.c) gamma = gamma parameter of the estimated IGD* that best fits the data
#             1.d) dp2 = matrix of the partial second derivatives
#             1.e) stdev.params = vector of the estimated standard deviations of the estimated IGD*
#                                 parameters
#             1.f) igd.loglik = average negative log-likelihood value
robustMLE <- function (y, filter.variances=FALSE) {

  max.gamma <- 0.01
  if (min(y) > max.gamma) {
    max.gamma <- (min(y)-0.01)
  }
```

```
gamma.starts <- seq(0.01,max.gamma,2)
if (gamma.starts[length(gamma.starts)] < max.gamma) {
  gamma.starts <- c(gamma.starts, max.gamma)
}

# Variables used to store different statistics computed during the MLE for different values of gamma
sigd.log.liks <- vector(length=(length(gamma.starts)))
sigd.mean.stdev.strict <- vector(length=(length(gamma.starts)))
final.alphas <- vector(length=(length(gamma.starts)))
final.betas <- vector(length=(length(gamma.starts)))
final.gammas <- vector(length=(length(gamma.starts)))
final.analytical.hessian.strict <- list()
final.stdev.strict <- list()
list.mles <- list()



for (gamma.start in (1:length(gamma.starts))) {

  y.start <- y - gamma.starts[gamma.start]

  alpha.start <- igdParams(y.start)[1]
  beta.start <- igdParams(y.start)[2]

  control <- list(maxit=500, ndeps=c(1e-3, 1e-3, 1e-3))

  sigd.mle <- optim(
      c(alpha.start, beta.start, -gamma.starts[gamma.start]),
      avrgSigdLogLik,
      x=y,
      method="L-BFGS-B",
      lower=c(0.01, 0.01, -(min(y) -0.0001)),
      upper=c(Inf, Inf, 0),
      control=control,
      hessian = TRUE)

  list.mles[[gamma.start]] <- sigd.mle
  final.alphas[gamma.start] <- sigd.mle$par[1]
  final.betas[gamma.start] <- sigd.mle$par[2]
  final.gammas[gamma.start] <- sigd.mle$par[3]
  sigd.log.liks[gamma.start] <- sigd.mle$value
  final.analytical.hessian.strict[[gamma.start]] <- sigdLogLikSD(sigd.mle$par[1],
      sigd.mle$par[2],
      sigd.mle$par[3],
      y)

  final.stdev.strict[[gamma.start]] <- estimStandardError(final.analytical.hessian.strict[[gamma.start]])

  sigd.mean.stdev.strict[gamma.start] <- Inf
  if (sum(is.na(final.stdev.strict[[gamma.start]])) == 0) {
    sigd.mean.stdev.strict[gamma.start] <- mean(final.stdev.strict[[gamma.start]])
  }
}

index.min.log.lik <- which.min(sigd.log.liks)
index.min.stdev <- which.min(sigd.mean.stdev.strict)

alpha.estim <- NA
beta.estim <- NA
gamma.estim <- NA
dp2.strict <- NA
stdev.params.strict <- NA
igd.loglik <- NA

if (filter.variances) {
  alpha.estim <- final.alphas[index.min.stdev]
  beta.estim <- final.betas[index.min.stdev]
  gamma.estim <- final.gammas[index.min.stdev]
  dp2.strict <- final.analytical.hessian.strict[[index.min.stdev]]
  stdev.params.strict <- final.stdev.strict[[index.min.stdev]]
  igd.loglik <- sigd.log.liks[index.min.stdev]
}
```

```
  else {
    alpha.estim <- final.alphas[index.min.log.lik]
    beta.estim <- final.betas[index.min.log.lik]
    gamma.estim <- final.gammas[index.min.log.lik]
    dp2.strict <- final.analytical.hessian.strict[[index.min.log.lik]]
    stdev.params.strict <- final.stdev.strict[[index.min.log.lik]]
    igd.loglik <- sigd.log.liks[index.min.log.lik]
  }

  return(list(alpha=alpha.estim,
          beta=beta.estim,
          gamma=gamma.estim,
          dp2=dp2.strict,
          stdev.params=stdev.params.strict,
          igd.loglik=igd.loglik
      )
  )
}

# Function to compute the standard errors of the alpha, beta, gamma estimates from the inversion of
#  the Hessian matrix
# -Input: 1) an Hessian matrix
# -Output: 1) a vector with the standard deviations for the diagonal elements of the Hessian matrix
estimStandardError <- function(Hessian) {
  if (det(Hessian) != 0) {
    return(sqrt(diag(solve(Hessian))))
  }
  else {
    return(rep(NA, length(diag(Hessian))))
  }
}

# Function to compute the average p-value from a parametric bootstrapped KS-test
#  -Input: 1) the number of iterations for the bootstrap procedure
#          2) a vector x of positive and non-zero numeric values
#          3) alpha IGD* parameter
#          4) beta IGD* parameter
#          5) gamma IGD* parameter
#          6) boolean value for the computation of an exact p-value (please see help page for
#             ks.test() function for further information)
#          7) seed number for the number random generator
#  -Output: 1) a list with the following elements:
#             1.a) bootstrap estimator for the p-value from the KS-test
#             1.b) variance of the bootstrap estimator
kspvalue <- function(iterations, x, alpha, beta, gamma, exact=NULL, seed=3){
  set.seed(seed)
  ks <- vector(len=iterations)
  for (i in 1:iterations) {
    ks[i] <- ks.test(x,rigamma(length(x),alpha,beta)-gamma,exact=exact)$p.value
  }
  return(list(mean=mean(ks), var=var(ks)))
}

# Function to validate the distribution of isotropic B factors in crystallographic models deposited
#   into the PDB.
Bvalid <- function(
    pdb,                                  # A valid PDB code or filename.
    download=FALSE,                       # Whether to download the PDB model.
    water=FALSE,                          # Whether water should be taken into account.
    hydrogen=FALSE,                       # Whether hydrogen atoms should be taken into account.
    chain="",                             # The chain Id to analyse. "" corresponds to all chains.
    compound=c("p","n","PN"),             # Compound to analyse. "p" stands for protein, "n" stands
                                          #   for nucleic acids. "PN" stands for both (default="p").
    selection=c("all","backbone","P","CA"),  # Selection for the atom type (default="all").
    plot=TRUE,                            # Whether to plot the validation histogram.
    to.screen=TRUE,                       # Whether to plot to screen.
    output.pdf.filename=NULL,             # Filename of the pdf file to use to save the histogram.
    plot.header=NULL,                     # String to use as header in the plot (default = pdb)
    seed=3,                               # Seed for the random numbers generator.
    ties.mult=9,                          # Multiplicative factor for the z-score used to detect
                                          #   strong ties in the distribution of B factors.
    mult.B.const=NULL,                    # Multiplicative constant for B factors (used fo <u^2>).
    ks.iterations=1000) {                 # Number of iterations for the bootstrapped KS-test
```

```
pdb.repository <- "ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/"
fullPath <- c()

if (download) {
  downloadString <- paste("wget ", pdb.repository, "pdb",tolower(pdb),".ent.gz",sep="")
  gunzipString <- paste("gunzip pdb", tolower(pdb), ".ent.gz",sep="")
  system(downloadString)
  system(gunzipString)
  fullPath <- paste("pdb", tolower(pdb), ".ent",sep="")
}
else {
  fullPath <- pdb
  pdb <- sub(".*/","",pdb,perl=TRUE)   # Selecting only the file name. Not the path to the file.
}

compound <- match.arg(compound)
selection <- match.arg(selection)

# Performing some consistency checks for selections
if (((compound == "p") & (selection == "P")) | ((compound == "n") & (selection == "CA"))) {
  stop(paste("Consistency error! Impossible to select CA in RNA/DNA moieties or\n",
          "P in protein moieties. Please check the parameters.", sep=""))
}
if (compound == "n" & selection == "backbone") {
  stop(paste("Error! At the moment it is possible to select backbone only for protein moieties.\n",
          "Sorry for the inconvinience. Hopefully it will be added soon.", sep=""))
}

# Uploading all atoms, including water (het2atom=TRUE). The HETATM records are
#   converted to ATOM records to simplify the data structure to analyse.
pdbContent <- c()
if (download) {
  pdbContent <- read.pdb(fullPath, het2atom=TRUE, maxlines=1000000,rm.alt=FALSE)
}
else {
  pdbContent <- read.pdb(fullPath, het2atom=TRUE, maxlines=1000000,rm.alt=FALSE)
}
upAtoms <- pdbContent$atom
lengthRawUploadedAtoms <- length(upAtoms[,"b"])
cat(paste("\nTotal uploaded atoms: ",lengthRawUploadedAtoms , "\n\n", sep=""))

# Extracting chains from uploaded atoms
chains <- unique(upAtoms[,"chain"])
chains[is.na(chains)] <- " "

if ((sum(chains == chain) == 0) & (chain != "")) {
  stop(paste("Error! The chain you selected is not present in the structure.\n",
          "Please remember that the chain selection is case sensitive.",,sep=""))
}

# Creating a boolean vector for the selection of desired subsets
selectionVector <- rep(TRUE,lengthRawUploadedAtoms)

# Dealing with hydrogen atoms
selectionHydrogen <- rep(TRUE,lengthRawUploadedAtoms)
for (i in 1:length(selectionHydrogen)) {
  elety <- substr(upAtoms[i,"elety"],1,1)
  if (elety == "H") {
    selectionHydrogen[i] <- FALSE
  }
}
cat(paste("\nTotal hydrogen atoms: ", sum(!selectionHydrogen) , "\n\n", sep=""))
if ((!hydrogen) & (sum(!selectionHydrogen) > 0)) {
  selectionVector <- selectionVector & selectionHydrogen
  cat(paste(" -- The hydrogen atoms have been discarded from the analysis\n\n", sep=""))
}
else {
  cat(paste(" -- The hydrogen atoms have been included in the analysis\n\n", sep=""))
}

# Selecting only atoms with occupancy == 1
occupancySel <- as.double(upAtoms[,"o"]) == 1
```

```
outOccupancy <- !occupancySel & selectionVector
cat(paste("Number of atoms with occupancy not equal to 1: ", sum(outOccupancy), "\n\n", sep=""))
if (sum(outOccupancy) >= 1) {
  cat("**** !!!Warning!!! The following atoms have an occupancy value different from 1 ****\n\n")
  print(upAtoms[outOccupancy,c("eleno","elety","resid","chain","resno","o","b")])
  cat("\n\n")
}
selectionVector <- selectionVector & occupancySel # Updating selection vector

# Applying chain selection
chainSelection <- c()
if (chain != "") {
  chainSelection <- upAtoms[,"chain"] == chain
}
else {
  chainSelection <- rep(TRUE,lengthRawUploadedAtoms)
}
selectionVector <- selectionVector & chainSelection # Updating selection vector
if (chain == "") {
  cat(paste(" -- Number of atoms from all chains: ",
          sum(selectionVector,na.rm=TRUE), "\n\n", sep=""))
}
else {
  cat(paste(" -- Number of atoms from chain ", chain, ": ",
          sum(selectionVector,na.rm=TRUE), "\n\n", sep=""))
}

# Applying compound selection
compoundSelection <- c()
if (compound == "p") {
  compoundSelection <- selectResid(upAtoms, aa.codes)
}
else if (compound == "n") {
  compoundSelection <- selectResid(upAtoms, nn.codes)
}
else if (compound == "PN") {
  compoundSelection <- selectResid(upAtoms, c(aa.codes,nn.codes))
}
selectionVector <- selectionVector & compoundSelection # Updating selection vector
cat(paste("  -- Number of atoms from selection ", compound,
        ": ", sum(selectionVector, na.rm=TRUE), "\n\n", sep=""))

# Applying selection criteria
selectionOpt <- c()
if (selection == "backbone") {
  selectionOpt <- selectElety(upAtoms, backbone.elety)
}
else if (selection == "CA") {
  selectionOpt <- selectElety(upAtoms, c("CA"))
}
else if (selection == "P") {
  selectionOpt <- selectElety(upAtoms, c("P"))
}
else if (selection == "all") {
  selectionOpt <- rep(TRUE, lengthRawUploadedAtoms)
}
selectionVector <- selectionVector & selectionOpt # Updating selection vector
cat(paste("   -- Number of atoms from selection ", selection, ": ",
        sum(selectionVector, na.rm=TRUE), "\n\n", sep=""))

# Selecting water molecules... If requested.
waterSelection <- c()
waterSelection <- selectResid(upAtoms, c("HOH"))
if (water) {
  cat(paste("   -- Number of water molecules included into the analysis: ",
          sum(waterSelection), "\n\n", sep=""))
  selectionVector <- selectionVector | waterSelection
}
else {
  cat(paste("   -- Number of water molecules excluded from the analysis: ",
          sum(waterSelection), "\n\n", sep=""))
  selectionVector <- selectionVector &  !(waterSelection)
}
```

```
# Selecting only atoms with B-factor higher than zero
bfacSelection <- as.double(upAtoms[,"b"]) > 0
nullBfactors = sum(!bfacSelection & selectionVector)
cat(paste("Number of atoms with B-factor values equal to zero (or less...): ",
        nullBfactors, "\n\n", sep=""))
if (nullBfactors >= 1) {
  cat("**** !!!Warning!!!: found atoms with B-factors values equal to zero (or less...) ****\n\n")
  print(upAtoms[!bfacSelection & selectionVector,c("eleno","elety","resid","chain","resno","o","b")])
  cat("\n\n")
}
selectionVector <- selectionVector & bfacSelection # Updating selection vector

cat(paste("Number of atoms with occupancy equal to one and Bfactor values higher than zero: ",
        sum(selectionVector), "\n\n", sep=""))

bFactors <- as.double(upAtoms[selectionVector,"b"])

# If mult.B.const is not null, all the B factors in the sample are multiplied by a constant
if (!(is.null(mult.B.const))) {
  bFactors = bFactors * mult.B.const
}

uniqueBfactors <- unique(bFactors)
rawBfacStat <- getBfactorStatistics(bFactors)

# Computation of IGD parameters by MLE
cat("Computing Inverse-Gamma parameters by Maximum Likelihood Estimation\n\n")
mlestimation <- robustMLE(bFactors, filter.variances=FALSE)
alphaML <- mlestimation$alpha
betaML <- mlestimation$beta
gammaML <- mlestimation$gamma

# Computation of a parametric bootrstrapped two-sided two-sample Kolmogorov-Smirnov test
cat("Computing Kolmogorov-Smirnov test p-value\n")
ksTwoSidedMLPvalue <- kspvalue(ks.iterations, bFactors, alphaML, betaML, gammaML, seed=seed)
cat(paste("Kolmogorov-Smirnov test p-value: ",
        signif(ksTwoSidedMLPvalue[["mean"]],4),"\n\n",sep=""))

# computation of graphical parameters
print("Computing graphical parameters")
breakswidth <- signif(((rawBfacStat$max*1.1 - rawBfacStat$min*0.1)/500),3)
xML <- seq((rawBfacStat$min+gammaML)*0.1, (rawBfacStat$max+gammaML)*1.1, breakswidth)
inverseGammaML <- densigamma(xML, alphaML, betaML)
smoothedBML <- density(bFactors,kernel="gaussian",from=rawBfacStat$min*0.1,
    to=rawBfacStat$max*1.1,n=length(xML))

# Looking for ties in the data
sigmaMultiplier <- ties.mult
ta <- getTiesStatistics(bFactors, sigmaMultiplier)
if (ta$mean > 2 & sqrt(ta$var) > 0) {
  cat(paste("Number of strong ties detetected with a treshold of ", sigmaMultiplier, " sigma: ",
          paste(ta$outliers.counts,collapse=" "),"\n",sep=""))
  if (ta$outliers.counts > 0) {
    cat(paste(" -- B-factor values at which the ties have been detected: ",
            paste(round(ta$outliers.bfactors,2),collapse=" "),"\n",sep=""))
    cat(paste(" -- Corresponding pecentiles: ",
            paste(round(ta$outliers.percentiles,2),collapse=" "),"\n",sep=""))
  }
}

# Plotting section
if (plot) {

  histInfo <- hist(bFactors, breaks="FD", plot=FALSE, freq=FALSE)
  histMax <- max((histInfo$density),(inverseGammaML),(smoothedBML$y), na.rm=TRUE)*1.01

  if (max(inverseGammaML, na.rm=TRUE) > 2*max((histInfo$density),(smoothedBML$y))) {
    histMax <- max((histInfo$density),(smoothedBML$y)) * 1.5
  }

  # Checking for outliers in ties distribution
  selectionTiesOutliers <- ta$mean > 2 & sqrt(ta$var) > 0
```

```
selectionHistCol <- rep("white",length(histInfo$mids))
if (selectionTiesOutliers) {
  for (i in ta$outliers.bfactors) {
    index <- min(which(histInfo$breaks >= i))
    if (index == 1) {
      selectionHistCol[min(which(histInfo$breaks >= i))] <-"blue"
    }
    else {
      selectionHistCol[(min(which(histInfo$breaks >= i)) - 1)] <-"blue"
    }
  }
}

# opening devices
if (to.screen & plot) {
  X11()
}
if ((!is.null(output.pdf.filename)) & (!to.screen)) {
  pdf(paste(output.pdf.filename, "_", pdb,"_Bfactors_distribution.pdf",sep=""),
      width = 8.2, height = 11.6, colormodel="cmyk", paper="a4")
  par(mfrow=c(3,2))
}

histCol <- "gray"
if (ksTwoSidedMLPvalue[["mean"]] < 0.01) {
  histCol <- "orange"
}

chains = ""

if (chain != "") {
  chains = chain
}
else {
  chains = paste(unique(upAtoms[selectionVector,"chain"]),collapse=", ",sep="")
}

par(
    mar=c(4.5,6,5,2),
    las = 1,
    yaxs="i",
    xaxs="i",
    tcl = -0.3,
    font.axis=1,
    font.main=1,
    font.lab=1,
    cex.axis=1.5,
    mgp=c(4, 1, 0.2)
)

xrange <- c(0, rawBfacStat$max*1.1)
yrange <- c(0, histMax)

header = pdb

if (!is.null(plot.header)) {
  header = plot.header
}

hist(
    bFactors,
    breaks="FD",
    freq=FALSE,
    main=NULL,
    xlab=NULL,
    ylab=NULL,
    col=selectionHistCol,
    border=histCol,
    ylim=yrange,
    xlim=xrange
)
```

```
    # Drawing y axis label
    mtext("Density", side=2, font=1, line=4.5, las=3, cex=1 )

    # Drawing x axis labels
    mtext("B factors", side=1, font=1, line=3, cex=1)

    # Drawing main title
    title_igd=header
    mtext(title_igd, side=3, font=2, line=3.2, cex=1)

    # Drawing IGD estimates
    IGD_parameters = as.expression(substitute(list(alpha == a, ~~beta == b, ~~gamma == c),
            list(a = round(mlestimation$alpha,2), b = round(mlestimation$beta,2),
                c = round(- mlestimation$gamma,2))))
    mtext(IGD_parameters, side=3, font=1, line=0.1, cex=0.7)

    # Drawing orthogonal statistics
    orthogonal_stats=paste("p-value = ", signif(ksTwoSidedMLPvalue[["mean"]],2), ",  null Bs = ",
        nullBfactors, ",  non-H atoms = ", length(bFactors), sep="")
    mtext(orthogonal_stats, side=3, font=1, line=1, cex=0.7)

    # Drawing chain selection:
    chain_selection = paste( "chains: ", chains, sep="")
    mtext(chain_selection, side=3, font=1, line=2, cex=0.7)

    # Plotting the density function of the derived Inverse Gamma distribution
    lines(smoothedBML$x,smoothedBML$y,col="black", lwd=2)
    lines(xML-gammaML,inverseGammaML,col="red",lwd=2)

    if ((!is.null(output.pdf.filename)) & (!to.screen)) {
      dev.off()
    }

  }

  if (download) {
    rmString <- paste("rm pdb", tolower(pdb), ".ent",sep="")
    system(rmString)
  }

  return(
      list(
          B.factors.stats=rawBfacStat,
          ta=ta,
          p.value=ksTwoSidedMLPvalue,
          sigd.mle=mlestimation,
          tot.null.B=nullBfactors,
          length=length(bFactors)
      )
  )
}

# Function to compute the number of times a given number is present in a vector
#  -Input: 1) a vector x of numerical values
#  -Output: 1) a list with two elements:
#             1.a) vector of unique values
#             1.b) vector of how many times each unique value has been counted in the original x vector
checkTies <- function (x) {
  # Converting Bfactors values to double entities
  x <- as.double(x)

  # Obtaining the vector of unique values
  uniqueValues <- unique(x)
  counts <- vector(len=length(uniqueValues))

  # counting how many times each doubled value appears in the original data vector x
  for (i in 1:length(uniqueValues)){
    counts[i] <- length(which(x == uniqueValues[i]))
  }
  list(unique=uniqueValues, counts=counts)
}

# Function to compute the percentile values for a vector of values in a data set.
```

```
#  -Input: 1) a vector x of numerical values (the population o values)
#          2) a vector y of selected values from x for which compute the correspoding percentiles
#  -Output: 1) a vector of percentiles
getPercentile <- function(x,y) {
  percentiles <- vector(length=length(y))
  for (i in 1:length(y)) {
    percentiles[i] <- round((sum(x <= y[i], na.rm=TRUE)/length(x))*100,2)
  }
  percentiles
}


# Function to compute several statistics on ties, if present in the data.
#  -Input: 1) a vector x of numerical values
#          2) a costant k to be used for the detection of strong ties in the data set x. All the
#             ties counts that are higher than the mean values plus k times the standard deviation
#             are considered outliers
#  -Output: 1) a list with several stastistics computed from the population x
getTiesStatistics <- function (x,k=3) {

  # Extracting ties data
  tiesData <- checkTies(x)

  # Extracting number of unique Bfact
  uniqueBfact <- tiesData$unique
  numberUniqueBfact <- length(uniqueBfact)

  # Extracting number of counts for each unique B-factor value
  #  and computing theis mean and variance
  counts <- tiesData$counts
  counts.mean <- mean(counts)
  counts.var <- var(counts)
  counts.max <- max(counts)
  counts.firstIqr <- quantile(counts,1/4)
  counts.median <- median(counts)
  counts.thirdIqr <- quantile(counts,3/4)

  # Extracting minimum value from ties count
  minTies <- 1
  if (sum(counts > 1) != 0) {
    minTies <- min(counts[counts > 1])
  }

  # Extracting maximum value from ties count
  maxTies <- 1
  if (sum(counts > 1) != 0) {
    maxTies <- max(counts[counts > 1])
  }

  # Extracting average number of duplicated values (for counts > 1)
  avrgTies <- 1
  if (sum(counts > 1) != 0) {
    avrgTies <- mean(counts[counts > 1])
  }

  # Extracting variance value duplicated values (for counts > 1)
  varTies <- 0
  if (sum(counts > 1) > 1) {
    varTies <- var(counts[counts > 1])
  }

  # Extracting first interquartile for ties data
  firstIQties <- 1
  if (sum(counts > 1) != 0) {
    firstIQties <- quantile(counts[counts > 1],1/4)
  }

  # Extracting median for ties data
  medianTies <- 1
  if (sum(counts > 1) != 0) {
    medianTies <- median(counts[counts > 1])
  }
```

```
  # Extracting third interquartile for ties data
  thirdIQties <- 1
  if (sum(counts > 1) != 0) {
    thirdIQties <- quantile(counts[counts > 1],3/4)
  }

  counts.selectionOut <- (counts > (counts.mean + k*sqrt(counts.var)))

  counts.outNumber <- sum(counts.selectionOut)
  counts.tiedBfact <- NA
  counts.tiedCounts <- 0
  counts.tiedPercentiles <- NA

  if (counts.outNumber > 0 & length(counts) > 0) {
    counts.tiedBfact <- uniqueBfact[counts.selectionOut]
    counts.tiedCounts <- counts[counts.selectionOut]
    counts.tiedPercentiles <- getPercentile(x,counts.tiedBfact)
  }

  selectionOut <- (counts > (avrgTies + k*sqrt(varTies)))

  outNumber <- sum(selectionOut)
  tiedBfact <- NA
  tiedCounts <- 0
  tiedPercentiles <- NA

  if (outNumber > 0 & length(counts) > 0) {
    tiedBfact <- uniqueBfact[selectionOut]
    tiedCounts <- counts[selectionOut]
    tiedPercentiles <- getPercentile(x,tiedBfact)
  }

  list(firstQ=firstIQties[[1]],
      median=medianTies,
      min=minTies,
      max=maxTies,
      mean=round(avrgTies,2),
      var=round(varTies,4),
      counts=counts,
      counts.mean=counts.mean,
      counts.var=counts.var,
      counts.max=counts.max,
      counts.median=counts.median,
      counts.firstQ=counts.firstIqr[[1]],
      counts.thirdQ=counts.thirdIqr[[1]],
      counts.outNumber=counts.outNumber,
      counts.tiedBfact=counts.tiedBfact,
      counts.tiedCounts=counts.tiedCounts,
      counts.tiedPercentiles=round(counts.tiedPercentiles,2),
      thirdQ=thirdIQties[[1]],
      outliers.number=outNumber,
      outliers.bFactors=tiedBfact,
      outliers.counts=tiedCounts,
      outliers.k=k,
      outliers.percentiles=round(tiedPercentiles,2),
      number.uniqueBfact=numberUniqueBfact
  )
}

# Function to compute several parametric and non-parametric statistics from a vector of numerical values
#  -Input: 1) a vector x of numerical values
#  -Output: 1) a list with parametric and non-parametric statistics
getxtatistics <- function (x) {
  list(min=round(min(x),2),
      firstQ=round(quantile(x,1/4),2),
      median=round(median(x),2),
      mean=round(mean(x),2),
      var=round(var(x),4),
      thirdQ=round(quantile(x,3/4),2),
      max=round(max(x),2))
}
```

Appendix B

Properties of mean and variance in presence of additive and multiplicative constants

Let $\boldsymbol{x}$ be a sample of size n drawn from a population having distribution D. The sample mean is then defined as:

$$\langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{B.1}$$

and the sample variance as:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 \tag{B.2}$$

By definition the sample mean $\langle x \rangle$ and sample variance s_x^2 are respectively related to population mean μ and population variance σ^2 as follows

$$\mu = \lim_{n \to \text{inf}} \langle x \rangle = \lim_{n \to \text{inf}} \frac{1}{n} \sum_{i=1}^{n} x_i \tag{B.3}$$

$$\sigma^2 = \lim_{n \to \text{inf}} s_x^2 = \lim_{n \to \text{inf}} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 = \lim_{n \to \text{inf}} \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{B.4}$$

Let now define a multiplicative constant a and an additive constant b. From (B.1) we obtain:

$$\langle ax + b \rangle = \frac{1}{n} \sum_{i=1}^{n} (ax_i + b) = \frac{a}{n} \sum_{i=1}^{n} x_i + \frac{nb}{n} = a\langle x \rangle + b \tag{B.5}$$

and from (B.2) we obtain:

$$\begin{aligned} s_{ax+b}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \big((ax_i + b) - (a\langle x \rangle + b)\big)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \big(a(x_i - \langle x \rangle)\big)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} a^2 (x_i - \langle x \rangle)^2 \\ &= \frac{a^2}{n-1} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 = a^2 s_{ax+b}^2 \end{aligned} \tag{B.6}$$

Similarly, from equations (B.5) and (B.3)

$$\lim_{n \to \text{inf}} \langle ax + b \rangle = a\mu + b \tag{B.7}$$

and from equations (B.6) and (B.4)

$$\lim_{n \to \text{inf}} s_{ax+b}^2 = a^2 \sigma^2 \tag{B.8}$$

Appendix C

Effect of a multiplicative constant on α and β IGD* parameters

Let a be a multiplicative constant and α' the α parameter of the IGD* whose sample values were multiplied by a. From the combination of equation (3.21) with equations (B.7) and (B.8) we obtain:

$$\alpha' = \frac{\big(a(\mu+\gamma)\big)^2}{a^2\sigma^2} + 2 = \frac{a^2(\mu-\gamma)^2}{a^2\sigma^2} + 2 = \frac{(\mu-\gamma)^2}{\sigma^2} + 2 = \alpha. \tag{C.1}$$

Similarly, let be β' the β parameter of the IGD* whose sample values were multiplied by a. From the combination of equation (3.22) with equation (B.7) we obtain

$$\beta' = a(\mu-\gamma)(\alpha-1) = a\beta. \tag{C.2}$$

It follows that if we multiply an IGD* by a constant a, only the β parameter is affected (equation (C.2)), while the α parameter remains constant (equation (C.1)).

Appendix D

R factors statistics for default and re-refined 30S subunits from *Thermus thermophilus*

PDB	default		protocol 1		protocol 2		protocol 3		protocol 4		protocol 5		protocol 6		protocol 7		protocol 8		protocol 9		protocol 10		protocol 11		protocol 12		protocol 13	
code	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}	R_{work}	R_{free}
1FJG	0.221	0.255	0.210	0.245	0.181	0.225	0.195	0.239	0.205	0.240	0.210	0.238	0.212	0.242	0.212	0.242	0.199	0.238	0.220	0.246	0.220	0.246	0.205	0.241	0.214	0.245	0.203	0.241
1HNW	0.222	0.264	0.221	0.263	0.194	0.245	0.206	0.258	0.212	0.255	0.222	0.260	0.224	0.262	0.224	0.262	0.213	0.257	0.232	0.266	0.232	0.266	0.219	0.264	0.226	0.266	0.217	0.263
1HNX	0.232	0.280	0.221	0.267	0.189	0.246	0.206	0.262	0.214	0.260	0.219	0.260	0.222	0.264	0.222	0.265	0.211	0.263	0.229	0.265	0.228	0.264	0.215	0.261	0.223	0.266	0.214	0.264
1HNZ	0.218	0.261	0.215	0.260	0.186	0.236	0.199	0.250	0.211	0.255	0.210	0.249	0.214	0.253	0.214	0.253	0.203	0.252	0.221	0.255	0.221	0.255	0.208	0.253	0.215	0.255	0.207	0.253
1HR0	0.218	0.261	0.213	0.257	0.185	0.236	0.199	0.254	0.205	0.250	0.210	0.251	0.214	0.255	0.214	0.255	0.203	0.253	0.221	0.257	0.221	0.257	0.207	0.255	0.215	0.256	0.207	0.256
1IBK	0.229	0.282	0.211	0.259	0.191	0.241	0.195	0.247	0.205	0.254	0.209	0.257	0.215	0.260	0.212	0.258	0.202	0.257	0.224	0.262	0.219	0.258	0.206	0.253	0.215	0.262	0.206	0.257
1IBL	0.232	0.275	0.213	0.257	0.178	0.230	0.195	0.248	0.208	0.253	0.212	0.251	0.214	0.255	0.214	0.255	0.202	0.250	0.222	0.257	0.221	0.256	0.206	0.250	0.217	0.257	0.206	0.251
1IBM	0.231	0.286	0.211	0.261	0.184	0.231	0.192	0.245	0.201	0.249	0.206	0.251	0.213	0.260	0.213	0.260	0.199	0.251	0.221	0.263	0.217	0.258	0.204	0.252	0.213	0.260	0.203	0.253
1J5E	0.208	0.252	0.216	0.257	0.182	0.232	0.198	0.249	0.209	0.249	0.215	0.248	0.217	0.251	0.217	0.252	0.204	0.249	0.225	0.255	0.224	0.255	0.209	0.251	0.220	0.254	0.208	0.252
1N32	0.227	0.270	0.219	0.266	0.192	0.249	0.205	0.259	0.214	0.261	0.218	0.261	0.223	0.263	0.223	0.264	0.213	0.262	0.229	0.264	0.229	0.264	0.217	0.262	0.222	0.265	0.215	0.263
1N33	0.225	0.284	0.224	0.280	0.191	0.244	0.202	0.265	0.216	0.272	0.235	0.285	0.223	0.270	0.222	0.269	0.204	0.262	0.231	0.272	0.230	0.272	0.210	0.264	0.242	0.295	0.229	0.294
1N34	0.241	0.312	0.240	0.305	0.197	0.266	0.218	0.296	0.236	0.301	0.238	0.307	0.230	0.288	0.232	0.296	0.213	0.291	0.245	0.296	0.244	0.297	0.219	0.287	0.243	0.303	0.227	0.300
1N36	0.260	0.324	0.260	0.326	0.201	0.261	0.233	0.315	0.254	0.320	0.264	0.328	0.245	0.304	0.245	0.305	0.216	0.289	0.255	0.308	0.255	0.309	0.225	0.291	0.270	0.335	0.251	0.333
1XMO	0.231	0.284	0.223	0.276	0.199	0.252	0.211	0.270	0.220	0.273	0.222	0.273	0.224	0.275	0.225	0.275	0.215	0.271	0.234	0.278	0.234	0.275	0.222	0.273	0.226	0.277	0.219	0.274
1XMQ	0.222	0.236	0.217	0.259	0.183	0.231	0.200	0.251	0.211	0.253	0.216	0.253	0.218	0.255	0.218	0.256	0.205	0.250	0.226	0.259	0.226	0.259	0.212	0.254	0.220	0.258	0.209	0.252
1XNQ	0.228	0.270	0.222	0.268	0.186	0.236	0.203	0.255	0.211	0.257	0.219	0.257	0.222	0.260	0.219	0.259	0.210	0.255	0.230	0.264	0.228	0.262	0.217	0.259	0.222	0.262	0.215	0.258
1XNR	0.227	0.273	0.222	0.268	0.186	0.238	0.206	0.257	0.212	0.257	0.216	0.255	0.218	0.259	0.219	0.259	0.209	0.253	0.228	0.263	0.227	0.263	0.215	0.259	0.222	0.262	0.214	0.258
2E5L	0.259	0.301	0.244	0.285	0.211	0.264	0.219	0.278	0.238	0.283	0.238	0.277	0.236	0.278	0.237	0.283	0.218	0.277	0.245	0.281	0.248	0.287	0.225	0.282	0.241	0.283	0.226	0.279
2HHH	0.265	0.289	0.248	0.279	0.227	0.276	0.229	0.282	0.239	0.270	0.255	0.274	0.251	0.276	0.251	0.276	0.232	0.285	0.256	0.278	0.257	0.278	0.238	0.287	0.255	0.275	0.240	0.283
2UU9	0.228	0.268	0.221	0.269	0.199	0.265	0.209	0.267	0.218	0.266	0.225	0.263	0.227	0.267	0.227	0.267	0.217	0.270	0.231	0.268	0.231	0.268	0.219	0.268	0.229	0.266	0.220	0.266
2UUA	0.223	0.254	0.219	0.258	0.190	0.240	0.203	0.252	0.215	0.254	0.223	0.253	0.225	0.256	0.225	0.256	0.212	0.255	0.229	0.258	0.229	0.258	0.216	0.256	0.228	0.258	0.215	0.256
2UUB	0.219	0.243	0.424	0.442	0.227	0.275	0.239	0.284	0.418	0.441	0.447	0.450	0.442	0.459	0.435	0.456	0.256	0.294	0.450	0.459	0.448	0.458	0.257	0.290	0.448	0.451	0.248	0.284
2UUC	0.210	0.245	0.205	0.247	0.179	0.226	0.190	0.239	0.199	0.242	0.206	0.239	0.209	0.244	0.209	0.244	0.198	0.243	0.213	0.245	0.212	0.244	0.200	0.242	0.210	0.244	0.200	0.242
2UXB	0.295	0.328	0.256	0.309	0.224	0.287	0.243	0.302	0.249	0.301	0.253	0.296	0.253	0.298	0.254	0.299	0.245	0.300	0.261	0.301	0.261	0.301	0.250	0.301	0.256	0.300	0.248	0.302
2UXC	0.217	0.255	0.220	0.264	0.193	0.244	0.208	0.258	0.214	0.256	0.217	0.254	0.223	0.259	0.223	0.260	0.216	0.260	0.228	0.261	0.228	0.261	0.220	0.260	0.223	0.260	0.217	0.260
2UXD	0.238	0.282	0.266	0.310	0.226	0.275	0.247	0.301	0.257	0.301	0.260	0.294	0.259	0.298	0.258	0.296	0.245	0.294	0.266	0.302	0.266	0.303	0.254	0.301	0.263	0.299	0.253	0.297
2VQE	0.257	0.283	0.270	0.300	0.250	0.292	0.261	0.297	0.268	0.298	0.278	0.300	0.278	0.302	0.277	0.302	0.270	0.301	0.280	0.302	0.279	0.302	0.272	0.301	0.280	0.303	0.273	0.302
2VQF	0.223	0.258	0.229	0.268	0.196	0.242	0.214	0.263	0.221	0.259	0.230	0.261	0.231	0.264	0.230	0.262	0.219	0.263	0.235	0.264	0.234	0.262	0.223	0.263	0.236	0.265	0.225	0.266
2ZM6	0.292	0.323	0.265	0.299	0.243	0.295	0.251	0.305	0.265	0.299	0.264	0.294	0.260	0.296	0.261	0.296	0.249	0.308	0.270	0.300	0.271	0.300	0.252	0.307	0.265	0.297	0.250	0.301

Table D.1: Summary of R factor statistics for all 13 re-refinement protocols applied to the 30S data set from *Thermus thermophilus* as described in section 4.5.

Appendix E

Ribosomal structures retrieved from the PDB

Table E.1: List of ribosomal structures retrieved in the PDB at a resolution higher than or equal to 4 Å.

	PDB code	organism	subunit	SF	resolution	R_{work}	R_{free}	date	split codes
1	2ZJR	*D.radiodurans*	50S	yes	2.91	0.28	0.31	3/8/08	
2	1NKW	*D.radiodurans*	50S	no	3.10	0.24	0.27	1/5/03	
3	1JZX	*D.radiodurans*	50S	no	3.10	0.27	0.30	9/17/01	
4	3CF5	*D.radiodurans*	50S	yes	3.30	0.28	0.32	3/2/08	
5	1NWY	*D.radiodurans*	50S	no	3.30	0.28	0.30	2/7/03	
6	2ZJQ	*D.radiodurans*	50S	yes	3.30	0.30	0.34	3/8/08	
7	2O44	*D.radiodurans*	50S	no	3.30		0.33	12/3/06	
8	1Y69	*D.radiodurans*	50S	no	3.33	0.28	0.34	12/4/04	
9	2D3O	*D.radiodurans*	50S	no	3.35	0.30	0.32	9/30/05	
10	1OND	*D.radiodurans*	50S	yes	3.40	0.26	0.31	2/27/03	
11	1P9X	*D.radiodurans*	50S	no	3.40	0.27	0.34	5/13/03	
12	1SM1	*D.radiodurans*	50S	no	3.42	0.28	0.35	3/8/04	
13	1NJP	*D.radiodurans*	50S	no	3.50	0.24	0.30	1/2/03	
14	2AAR	*D.radiodurans*	50S	yes	3.50	0.25	0.32	7/14/05	
15	3DLL	*D.radiodurans*	50S	yes	3.50	0.26	0.28	6/27/08	
16	1JZY	*D.radiodurans*	50S	no	3.50	0.27	0.30	9/17/01	
17	1J5A	*D.radiodurans*	50S	no	3.50	0.27	0.32	3/6/02	
18	1K01	*D.radiodurans*	50S	no	3.50	0.28	0.32	9/17/01	
19	2OGM	*D.radiodurans*	50S	yes	3.50	0.28	0.33	1/7/07	
20	1NWX	*D.radiodurans*	50S	no	3.50	0.28	0.31	2/7/03	
21	1XBP	*D.radiodurans*	50S	no	3.50	0.29	0.36	8/31/04	
22	2OGN	*D.radiodurans*	50S	yes	3.56	0.28	0.34	1/7/07	
23	1NJM	*D.radiodurans*	50S	no	3.60	0.28	0.31	1/2/03	
24	2O43	*D.radiodurans*	50S	no	3.60		0.34	12/3/06	
25	2O45	*D.radiodurans*	50S	no	3.60		0.36	12/3/06	
26	2OGO	*D.radiodurans*	50S	yes	3.66	0.26	0.33	1/7/07	
27	1NJO	*D.radiodurans*	50S	no	3.70	0.28	0.30	1/2/03	
28	1NJN	*D.radiodurans*	50S	no	3.70	0.28	0.31	1/2/03	
29	2ZJP	*D.radiodurans*	50S	yes	3.70	0.30	0.34	3/7/08	
30	3FWO	*D.radiodurans*	50S	yes	3.71	0.28	0.34	1/19/09	
31	1JZZ	*D.radiodurans*	50S	no	3.80	0.21	0.27	9/17/01	
32	1Z58	*D.radiodurans*	50S	no	3.80	0.27	0.37	3/17/05	
33	3I1M	*E.coli*	30S	yes	3.19	0.20	0.25	6/27/09	3I1M,3I1N,3I1O,3I1P
34	3I1N	*E.coli*	50S	yes	3.19	0.20	0.25	6/27/09	3I1M,3I1N,3I1O,3I1P
35	3I1O	*E.coli*	30S	yes	3.19	0.20	0.25	6/27/09	3I1M,3I1N,3I1O,3I1P
36	3I1P	*E.coli*	50S	yes	3.19	0.20	0.25	6/27/09	3I1M,3I1N,3I1O,3I1P
37	2QAL	*E.coli*	30S	yes	3.21	0.27	0.31	6/15/07	2QAL,2QAM,2QAN,2QAO
38	2QAM	*E.coli*	50S	yes	3.21	0.27	0.31	6/15/07	2QAL,2QAM,2QAN,2QAO
39	2QAN	*E.coli*	30S	yes	3.21	0.27	0.31	6/15/07	2QAL,2QAM,2QAN,2QAO
40	2QAO	*E.coli*	50S	yes	3.21	0.27	0.31	6/15/07	2QAL,2QAM,2QAN,2QAO
41	2I2P	*E.coli*	30S	yes	3.22	0.29	0.32	8/16/06	2I2P,2I2T,2I2U,2I2V
42	2I2T	*E.coli*	50S	yes	3.22	0.29	0.32	8/16/06	2I2P,2I2T,2I2U,2I2V
43	2I2U	*E.coli*	30S	yes	3.22	0.29	0.32	8/16/06	2I2P,2I2T,2I2U,2I2V
44	2I2V	*E.coli*	50S	yes	3.22	0.29	0.32	8/16/06	2I2P,2I2T,2I2U,2I2V
45	2QBD	*E.coli*	30S	yes	3.30	0.28	0.30	6/16/07	2QBD,2QBE,2QBF,2QBG
46	2QBE	*E.coli*	50S	yes	3.30	0.28	0.30	6/16/07	2QBD,2QBE,2QBF,2QBG
47	2QBF	*E.coli*	30S	yes	3.30	0.28	0.30	6/16/07	2QBD,2QBE,2QBF,2QBG
48	2QBG	*E.coli*	50S	yes	3.30	0.28	0.30	6/16/07	2QBD,2QBE,2QBF,2QBG
49	1VS5	*E.coli*	30S	yes	3.46	0.28	0.33	8/4/06	1VS5,1VS6,1VS7,1VS8
50	1VS6	*E.coli*	50S	yes	3.46	0.28	0.33	8/4/06	1VS5,1VS6,1VS7,1VS8
51	1VS7	*E.coli*	30S	yes	3.46	0.28	0.33	8/4/06	1VS5,1VS6,1VS7,1VS8
52	1VS8	*E.coli*	50S	yes	3.46	0.28	0.33	8/4/06	1VS5,1VS6,1VS7,1VS8
53	2AVY	*E.coli*	30S	yes	3.46	0.28	0.33	8/30/05	2AVY,2AW4,2AW7,2AWB
54	2AW4	*E.coli*	50S	yes	3.46	0.28	0.33	8/31/05	2AVY,2AW4,2AW7,2AWB
55	2AW7	*E.coli*	30S	yes	3.46	0.28	0.33	8/31/05	2AVY,2AW4,2AW7,2AWB
56	2AWB	*E.coli*	50S	yes	3.46	0.28	0.33	8/31/05	2AVY,2AW4,2AW7,2AWB
57	2QOY	*E.coli*	30S	yes	3.50	0.26	0.31	7/21/07	2QOY,2QOZ,2QP0,2QP1
58	2QOZ	*E.coli*	50S	yes	3.50	0.26	0.31	7/21/07	2QOY,2QOZ,2QP0,2QP1
59	2QP0	*E.coli*	30S	yes	3.50	0.26	0.31	7/21/07	2QOY,2QOZ,2QP0,2QP1
60	2QP1	*E.coli*	50S	yes	3.50	0.26	0.31	7/21/07	2QOY,2QOZ,2QP0,2QP1
61	3DF1	*E.coli*	30S	yes	3.50	0.27	0.32	6/11/08	3DF1,3DF2,3DF3,3DF4
62	3DF2	*E.coli*	50S	yes	3.50	0.27	0.32	6/11/08	3DF1,3DF2,3DF3,3DF4
63	3DF3	*E.coli*	30S	yes	3.50	0.27	0.32	6/11/08	3DF1,3DF2,3DF3,3DF4
64	3DF4	*E.coli*	50S	yes	3.50	0.27	0.32	6/11/08	3DF1,3DF2,3DF3,3DF4
65	2QB9	*E.coli*	30S	yes	3.54	0.28	0.32	6/16/07	2QB9,2QBA,2QBB,2QBC
66	2QBA	*E.coli*	50S	yes	3.54	0.28	0.32	6/16/07	2QB9,2QBA,2QBB,2QBC
67	2QBB	*E.coli*	30S	yes	3.54	0.28	0.32	6/16/07	2QB9,2QBA,2QBB,2QBC
68	2QBC	*E.coli*	50S	yes	3.54	0.28	0.32	6/16/07	2QB9,2QBA,2QBB,2QBC
69	3I1Z	*E.coli*	30S	yes	3.71	0.23	0.27	6/28/09	3I1Z,3I20,3I21,3I22

Table E.1: List of ribosomal structures retrieved in the PDB at a resolution higher than or equal to 4 Å.

	PDB code	organism	subunit	SF	resolution	R_{work}	R_{free}	date	split codes
70	3I20	*E.coli*	50S	yes	3.71	0.23	0.27	6/28/09	3I1Z,3I20,3I21,3I22
71	3I21	*E.coli*	30S	yes	3.71	0.23	0.27	6/28/09	3I1Z,3I20,3I21,3I22
72	3I22	*E.coli*	50S	yes	3.71	0.23	0.27	6/28/09	3I1Z,3I20,3I21,3I22
73	2VHM	*E.coli*	50S	yes	3.74	0.26	0.32	11/22/07	2VHM,2VHN,2VHO,2VHP
74	2VHN	*E.coli*	50S	yes	3.74	0.26	0.32	11/22/07	2VHM,2VHN,2VHO,2VHP
75	2VHO	*E.coli*	30S	yes	3.74	0.26	0.32	11/22/07	2VHM,2VHN,2VHO,2VHP
76	2VHP	*E.coli*	30S	yes	3.74	0.26	0.32	11/22/07	2VHM,2VHN,2VHO,2VHP
77	3I1Q	*E.coli*	30S	yes	3.81	0.21	0.25	6/27/09	3I1Q,3I1R,3I1S,3I1T
78	3I1R	*E.coli*	50S	yes	3.81	0.21	0.25	6/27/09	3I1Q,3I1R,3I1S,3I1T
79	3I1S	*E.coli*	30S	yes	3.81	0.21	0.25	6/27/09	3I1Q,3I1R,3I1S,3I1T
80	3I1T	*E.coli*	50S	yes	3.81	0.21	0.25	6/27/09	3I1Q,3I1R,3I1S,3I1T
81	2QOU	*E.coli*	30S	yes	3.93	0.26	0.31	7/21/07	2QOU,2QOV,2QOW,2QOX
82	2QOV	*E.coli*	50S	yes	3.93	0.26	0.31	7/21/07	2QOU,2QOV,2QOW,2QOX
83	2QOW	*E.coli*	30S	yes	3.93	0.26	0.31	7/21/07	2QOU,2QOV,2QOW,2QOX
84	2QOX	*E.coli*	50S	yes	3.93	0.26	0.31	7/21/07	2QOU,2QOV,2QOW,2QOX
85	2QBH	*E.coli*	30S	yes	4.00	0.26	0.30	6/17/07	2QBH,2QBI,2QBJ,2QBK
86	2QBI	*E.coli*	50S	yes	4.00	0.26	0.30	6/17/07	2QBH,2QBI,2QBJ,2QBK
87	2QBJ	*E.coli*	30S	yes	4.00	0.26	0.30	6/17/07	2QBH,2QBI,2QBJ,2QBK
88	2QBK	*E.coli*	50S	yes	4.00	0.26	0.30	6/17/07	2QBH,2QBI,2QBJ,2QBK
89	1VQO	*H.marismortui*	50S	yes	2.20	0.22	0.25	12/16/04	
90	1VQ8	*H.marismortui*	50S	yes	2.20	0.22	0.25	12/16/04	
91	1VQP	*H.marismortui*	50S	yes	2.25	0.22	0.25	12/16/04	
92	1VQM	*H.marismortui*	50S	yes	2.30	0.21	0.25	12/16/04	
93	1VQK	*H.marismortui*	50S	yes	2.30	0.22	0.25	12/16/04	
94	1VQL	*H.marismortui*	50S	yes	2.30	0.22	0.25	12/16/04	
95	1S72	*H.marismortui*	50S	yes	2.40	0.19	0.22	1/28/04	
96	1JJ2	*H.marismortui*	50S	yes	2.40	0.19	0.22	7/3/01	
97	1YHQ	*H.marismortui*	50S	yes	2.40	0.19	0.23	1/10/05	
98	3CC2	*H.marismortui*	50S	yes	2.40	0.20	0.23	2/23/08	
99	1VQN	*H.marismortui*	50S	yes	2.40	0.21	0.25	12/16/04	
100	1VQ9	*H.marismortui*	50S	yes	2.40	0.22	0.26	12/16/04	
101	1FFK	*H.marismortui*	50S	yes	2.40	0.25	0.26	7/25/00	
102	1VQ7	*H.marismortui*	50S	yes	2.50	0.21	0.24	12/16/04	
103	3CCM	*H.marismortui*	50S	yes	2.55	0.20	0.24	2/26/08	
104	1YIJ	*H.marismortui*	50S	yes	2.60	0.18	0.22	1/12/05	
105	1VQ5	*H.marismortui*	50S	yes	2.60	0.20	0.24	12/16/04	
106	1YI2	*H.marismortui*	50S	yes	2.65	0.18	0.21	1/11/05	
107	3CC7	*H.marismortui*	50S	yes	2.70	0.18	0.23	2/25/08	
108	3CCJ	*H.marismortui*	50S	yes	2.70	0.18	0.23	2/26/08	
109	3G6E	*H.marismortui*	50S	yes	2.70	0.19	0.23	2/6/09	
110	1VQ4	*H.marismortui*	50S	yes	2.70	0.19	0.23	12/16/04	
111	3CPW	*H.marismortui*	50S	yes	2.70	0.19	0.23	4/1/08	
112	1VQ6	*H.marismortui*	50S	yes	2.70	0.19	0.23	12/16/04	
113	3CC4	*H.marismortui*	50S	yes	2.70	0.20	0.24	2/24/08	
114	2OTL	*H.marismortui*	50S	yes	2.70	0.20	0.25	2/8/07	
115	3CCE	*H.marismortui*	50S	yes	2.75	0.18	0.23	2/25/08	
116	3CD6	*H.marismortui*	50S	yes	2.75	0.19	0.24	2/26/08	
117	1YIT	*H.marismortui*	50S	yes	2.80	0.18	0.22	1/13/05	
118	3CCU	*H.marismortui*	50S	yes	2.80	0.18	0.22	2/26/08	
119	1M90	*H.marismortui*	50S	yes	2.80	0.18	0.22	7/26/02	
120	1YJ9	*H.marismortui*	50S	yes	2.80	0.18	0.24	1/13/05	
121	3CMA	*H.marismortui*	50S	yes	2.80	0.19	0.24	3/21/08	
122	3G71	*H.marismortui*	50S	yes	2.85	0.19	0.23	2/9/09	
123	3CCL	*H.marismortui*	50S	yes	2.90	0.17	0.22	2/26/08	
124	1YJW	*H.marismortui*	50S	yes	2.90	0.17	0.22	1/15/05	
125	3CCV	*H.marismortui*	50S	yes	2.90	0.18	0.22	2/26/08	
126	3CCQ	*H.marismortui*	50S	yes	2.90	0.19	0.23	2/26/08	
127	3I56	*H.marismortui*	50S	yes	2.90	0.19	0.24	7/3/09	
128	2OTJ	*H.marismortui*	50S	yes	2.90	0.19	0.24	2/8/07	
129	1QVG	*H.marismortui*	50S	yes	2.90	0.20	0.26	8/27/03	
130	2QEX	*H.marismortui*	50S	yes	2.90	0.20	0.24	6/26/07	
131	3CCS	*H.marismortui*	50S	yes	2.95	0.18	0.24	2/26/08	
132	3CME	*H.marismortui*	50S	yes	2.95	0.20	0.26	3/21/08	
133	1Q81	*H.marismortui*	50S	yes	2.95	0.21	0.26	8/20/03	
134	1Q82	*H.marismortui*	50S	yes	2.98	0.21	0.25	8/20/03	
135	1YJN	*H.marismortui*	50S	yes	3.00	0.17	0.23	1/14/05	
136	1NJI	*H.marismortui*	50S	yes	3.00	0.18	0.21	12/31/02	
137	3CCR	*H.marismortui*	50S	yes	3.00	0.18	0.25	2/26/08	
138	1N8R	*H.marismortui*	50S	no	3.00	0.20	0.24	11/21/02	

Table E.1: List of ribosomal structures retrieved in the PDB at a resolution higher than or equal to 4 Å.

	PDB code	organism	subunit	SF	resolution	R_{work}	R_{free}	date	split codes
139	1K9M	*H.marismortui*	50S	yes	3.00	0.22	0.26	10/29/01	
140	1KD1	*H.marismortui*	50S	yes	3.00	0.22	0.27	11/12/01	
141	1K8A	*H.marismortui*	50S	yes	3.00	0.23	0.26	10/23/01	
142	1Q86	*H.marismortui*	50S	yes	3.00	0.23	0.26	8/20/03	
143	2QA4	*H.marismortui*	50S	yes	3.00	0.24	0.29	6/14/07	
144	1FG0	*H.marismortui*	50S	yes	3.00			7/26/00	
145	3CXC	*H.marismortui*	50S	yes	3.00			4/24/08	
146	1KC8	*H.marismortui*	50S	yes	3.01	0.20	0.24	11/7/01	
147	1K73	*H.marismortui*	50S	yes	3.01	0.21	0.25	10/18/01	
148	1KQS	*H.marismortui*	50S	yes	3.10	0.17	0.22	1/7/02	
149	1QVF	*H.marismortui*	50S	yes	3.10	0.20	0.24	8/27/03	
150	3I55	*H.marismortui*	50S	yes	3.11	0.21	0.26	7/3/09	
151	1M1K	*H.marismortui*	50S	yes	3.20	0.21	0.25	6/19/02	
152	3G4S	*H.marismortui*	50S	yes	3.20	0.22	0.29	2/4/09	
153	1Q7Y	*H.marismortui*	50S	yes	3.20	0.22	0.28	8/20/03	
154	1FFZ	*H.marismortui*	50S	yes	3.20			7/26/00	
155	1W2B	*H.marismortui*	50S	no	3.50	0.19	0.27	7/1/04	
156	2VQE	*T.thermophilus*	30S	yes	2.50	0.26	0.28	3/13/08	
157	2UUB	*T.thermophilus*	30S	yes	2.80	0.22	0.24	3/1/07	
158	2J00	*T.thermophilus*	30S	yes	2.80	0.27	0.31	7/31/06	2J00,2J01,2J02,2J03
159	2J01	*T.thermophilus*	50S	yes	2.80	0.27	0.31	7/31/06	2J00,2J01,2J02,2J03
160	2J02	*T.thermophilus*	30S	yes	2.80	0.27	0.31	7/31/06	2J00,2J01,2J02,2J03
161	2J03	*T.thermophilus*	50S	yes	2.80	0.27	0.31	7/31/06	2J00,2J01,2J02,2J03
162	2UXC	*T.thermophilus*	30S	yes	2.90	0.22	0.26	3/28/07	
163	2UUA	*T.thermophilus*	30S	yes	2.90	0.22	0.25	3/1/07	
164	2VQF	*T.thermophilus*	30S	yes	2.90	0.22	0.26	3/14/08	
165	1FJG	*T.thermophilus*	30S	yes	3.00	0.22	0.26	8/8/00	
166	1XMQ	*T.thermophilus*	30S	yes	3.00	0.22	0.24	10/4/04	
167	1N32	*T.thermophilus*	30S	yes	3.00	0.23	0.27	10/25/02	
168	3F1E	*T.thermophilus*	30S	yes	3.00	0.28	0.32	10/27/08	3F1E,3F1F,3F1G,3F1H
169	3F1F	*T.thermophilus*	50S	yes	3.00	0.28	0.32	10/27/08	3F1E,3F1F,3F1G,3F1H
170	3F1G	*T.thermophilus*	30S	yes	3.00	0.28	0.32	10/27/08	3F1E,3F1F,3F1G,3F1H
171	3F1H	*T.thermophilus*	50S	yes	3.00	0.28	0.32	10/27/08	3F1E,3F1F,3F1G,3F1H
172	3KNH	*T.thermophilus*	30S	yes	3.00	0.25	0.27	11/12/09	3KNH,3KNI,3KNJ,3KNK
173	3KNI	*T.thermophilus*	50S	yes	3.00	0.25	0.27	11/12/09	3KNH,3KNI,3KNJ,3KNK
174	3KNK	*T.thermophilus*	50S	yes	3.00	0.25	0.27	11/12/09	3KNH,3KNI,3KNJ,3KNK
175	1J5E	*T.thermophilus*	30S	yes	3.05	0.21	0.25	4/8/02	
176	1XNQ	*T.thermophilus*	30S	yes	3.05	0.23	0.27	10/5/04	
177	2UUC	*T.thermophilus*	30S	yes	3.10	0.21	0.24	3/1/07	
178	1XNR	*T.thermophilus*	30S	yes	3.10	0.23	0.27	10/5/04	
179	2UU9	*T.thermophilus*	30S	yes	3.10	0.23	0.27	3/1/07	
180	2UXB	*T.thermophilus*	30S	yes	3.10	0.30	0.33	3/28/07	
181	2X9R	*T.thermophilus*	30S	yes	3.10	0.22	0.26	3/24/10	2X9R,2X9S,2X9T,2X9U
182	2X9S	*T.thermophilus*	50S	yes	3.10	0.22	0.26	3/24/10	2X9R,2X9S,2X9T,2X9U
183	2X9T	*T.thermophilus*	30S	yes	3.10	0.22	0.26	3/24/10	2X9R,2X9S,2X9T,2X9U
184	2X9U	*T.thermophilus*	50S	yes	3.10	0.22	0.26	3/24/10	2X9R,2X9S,2X9T,2X9U
185	3HUW	*T.thermophilus*	30S	yes	3.10	0.25	0.30	6/15/09	3HUW,3HUX,3HUY,3HUZ
186	3HUX	*T.thermophilus*	50S	yes	3.10	0.25	0.30	6/15/09	3HUW,3HUX,3HUY,3HUZ
187	3HUY	*T.thermophilus*	30S	yes	3.10	0.25	0.30	6/15/09	3HUW,3HUX,3HUY,3HUZ
188	3HUZ	*T.thermophilus*	50S	yes	3.10	0.25	0.30	6/15/09	3HUW,3HUX,3HUY,3HUZ
189	3I8F	*T.thermophilus*	50S	yes	3.10	0.22	0.26	7/9/09	3I8F,3I8G,3I8H,3I8I
190	3I8G	*T.thermophilus*	30S	yes	3.10	0.22	0.26	7/9/09	3I8F,3I8G,3I8H,3I8I
191	3I8H	*T.thermophilus*	30S	yes	3.10	0.22	0.26	7/9/09	3I8F,3I8G,3I8H,3I8I
192	3I8I	*T.thermophilus*	50S	yes	3.10	0.22	0.26	7/9/09	3I8F,3I8G,3I8H,3I8I
193	1IBL	*T.thermophilus*	30S	yes	3.11	0.23	0.28	3/28/01	
194	3KNJ	*T.thermophilus*	30S	yes	3.15	0.25	0.27	11/12/09	3KNH,3KNI,3KNJ,3KNK
195	1I94	*T.thermophilus*	30S	no	3.20	0.20	0.24	3/18/01	
196	1HR0	*T.thermophilus*	30S	yes	3.20	0.22	0.26	12/20/00	
197	2UXD	*T.thermophilus*	30S	yes	3.20	0.24	0.28	3/28/07	
198	3D5A	*T.thermophilus*	30S	yes	3.21	0.29	0.32	5/16/08	3D5A,3D5B,3D5C,3D5D
199	3D5B	*T.thermophilus*	50S	yes	3.21	0.29	0.32	5/16/08	3D5A,3D5B,3D5C,3D5D
200	3D5C	*T.thermophilus*	30S	yes	3.21	0.29	0.32	5/16/08	3D5A,3D5B,3D5C,3D5D
201	3D5D	*T.thermophilus*	50S	yes	3.21	0.29	0.32	5/16/08	3D5A,3D5B,3D5C,3D5D
202	1XMO	*T.thermophilus*	30S	yes	3.25	0.23	0.28	10/4/04	
203	1HNZ	*T.thermophilus*	30S	yes	3.30	0.22	0.26	12/8/00	
204	2E5L	*T.thermophilus*	30S	yes	3.30	0.26	0.30	12/21/06	
205	2ZM6	*T.thermophilus*	30S	yes	3.30	0.29	0.32	4/11/08	
206	1FKA	*T.thermophilus*	30S	no	3.30	0.30	0.30	8/9/00	
207	2WDI	*T.thermophilus*	50S	yes	3.30	0.22	0.27	3/24/09	2WDG,2WDH,2WDI,2WDJ

Table E.1: List of ribosomal structures retrieved in the PDB at a resolution higher than or equal to 4 Å.

	PDB code	organism	subunit	SF	resolution	R_{work}	R_{free}	date	split codes
208	2WDJ	*T.thermophilus*	50S	yes	3.30	0.22	0.27	3/24/09	2WDG,2WDH,2WDI,2WDJ
209	2WDG	*T.thermophilus*	30S	yes	3.30	0.22	0.27	3/24/09	2WDG,2WDH,2WDI,2WDJ
210	2WDH	*T.thermophilus*	30S	yes	3.30	0.22	0.27	3/24/09	2WDG,2WDH,2WDI,2WDJ
211	3KIQ	*T.thermophilus*	30S	yes	3.30	0.22	0.25	11/2/09	3KIQ,3KIR,3KIS,3KIT
212	3KIR	*T.thermophilus*	50S	yes	3.30	0.22	0.25	11/2/09	3KIQ,3KIR,3KIS,3KIT
213	3KIS	*T.thermophilus*	30S	yes	3.30	0.22	0.25	11/2/09	3KIQ,3KIR,3KIS,3KIT
214	3KIT	*T.thermophilus*	50S	yes	3.30	0.22	0.25	11/2/09	3KIQ,3KIR,3KIS,3KIT
215	1IBK	*T.thermophilus*	30S	yes	3.31	0.23	0.28	3/28/01	
216	1IBM	*T.thermophilus*	30S	yes	3.31	0.23	0.29	3/28/01	
217	1N33	*T.thermophilus*	30S	yes	3.35	0.22	0.28	10/25/02	
218	2HHH	*T.thermophilus*	30S	yes	3.35	0.26	0.29	6/28/06	
219	1HNW	*T.thermophilus*	30S	yes	3.40	0.22	0.26	12/8/00	
220	1HNX	*T.thermophilus*	30S	yes	3.40	0.23	0.28	12/8/00	
221	2WH1	*T.thermophilus*	30S	yes	3.45	0.21	0.26	4/30/09	2WH1,2WH2,2WH3,2WH4
222	2WH2	*T.thermophilus*	50S	yes	3.45	0.21	0.26	4/30/09	2WH1,2WH2,2WH3,2WH4
223	2WH3	*T.thermophilus*	30S	yes	3.45	0.21	0.26	4/30/09	2WH1,2WH2,2WH3,2WH4
224	2WH4	*T.thermophilus*	50S	yes	3.45	0.21	0.26	4/30/09	2WH1,2WH2,2WH3,2WH4
225	3KNL	*T.thermophilus*	30S	yes	3.45	0.22	0.27	11/12/09	3KNL,3KNM,3KNN,3KNO
226	3KNM	*T.thermophilus*	50S	yes	3.45	0.22	0.27	11/12/09	3KNL,3KNM,3KNN,3KNO
227	3KNN	*T.thermophilus*	30S	yes	3.45	0.22	0.27	11/12/09	3KNL,3KNM,3KNN,3KNO
228	3KNO	*T.thermophilus*	50S	yes	3.45	0.22	0.27	11/12/09	3KNL,3KNM,3KNN,3KNO
229	2V46	*T.thermophilus*	30S	yes	3.50	0.26	0.33	6/28/07	2V46,2V47,2V48,2V49
230	2V47	*T.thermophilus*	50S	yes	3.50	0.26	0.33	6/28/07	2V46,2V47,2V48,2V49
231	2V48	*T.thermophilus*	30S	yes	3.50	0.26	0.33	6/28/07	2V46,2V47,2V48,2V49
232	2V49	*T.thermophilus*	50S	yes	3.50	0.26	0.33	6/28/07	2V46,2V47,2V48,2V49
233	2WDK	*T.thermophilus*	30S	yes	3.50	0.21	0.26	3/24/09	2WDK,2WDL,2WDM,2WDN
234	2WDL	*T.thermophilus*	50S	yes	3.50	0.21	0.26	3/24/09	2WDK,2WDL,2WDM,2WDN
235	2WDM	*T.thermophilus*	30S	yes	3.50	0.21	0.26	3/24/09	2WDK,2WDL,2WDM,2WDN
236	2WDN	*T.thermophilus*	50S	yes	3.50	0.21	0.26	3/24/09	2WDK,2WDL,2WDM,2WDN
237	3I9B	*T.thermophilus*	30S	yes	3.50	0.21	0.25	7/10/09	3I9B,3I9C,3I9D,3I9E
238	3I9C	*T.thermophilus*	50S	yes	3.50	0.21	0.25	7/10/09	3I9B,3I9C,3I9D,3I9E
239	3I9D	*T.thermophilus*	30S	yes	3.50	0.21	0.25	7/10/09	3I9B,3I9C,3I9D,3I9E
240	3I9E	*T.thermophilus*	50S	yes	3.50	0.21	0.25	7/10/09	3I9B,3I9C,3I9D,3I9E
241	2WRI	*T.thermophilus*	30S	yes	3.60	0.23	0.26	9/1/09	2WRI,2WRJ,2WRK,2WRL
242	2WRJ	*T.thermophilus*	50S	yes	3.60	0.23	0.26	9/1/09	2WRI,2WRJ,2WRK,2WRL
243	2WRK	*T.thermophilus*	30S	yes	3.60	0.23	0.26	9/1/09	2WRI,2WRJ,2WRK,2WRL
244	2WRL	*T.thermophilus*	50S	yes	3.60	0.23	0.26	9/1/09	2WRI,2WRJ,2WRK,2WRL
245	2WRN	*T.thermophilus*	30S	yes	3.60	0.28	0.32	9/1/09	2WRN,2WRO,2WRQ,2WRR
246	2WRO	*T.thermophilus*	50S	yes	3.60	0.28	0.32	9/1/09	2WRN,2WRO,2WRQ,2WRR
247	2WRQ	*T.thermophilus*	30S	yes	3.60	0.28	0.32	9/1/09	2WRN,2WRO,2WRQ,2WRR
248	2WRR	*T.thermophilus*	50S	yes	3.60	0.28	0.32	9/1/09	2WRN,2WRO,2WRQ,2WRR
249	3KIU	*T.thermophilus*	30S	yes	3.60	0.22	0.24	11/2/09	3KIU,3KIW,3KIX,3KIY
250	3KIW	*T.thermophilus*	50S	yes	3.60	0.22	0.24	11/2/09	3KIU,3KIW,3KIX,3KIY
251	3KIX	*T.thermophilus*	30S	yes	3.60	0.22	0.24	11/2/09	3KIU,3KIW,3KIX,3KIY
252	3KIY	*T.thermophilus*	50S	yes	3.60	0.22	0.24	11/2/09	3KIU,3KIW,3KIX,3KIY
253	3MR8	*T.thermophilus*	30S	yes	3.62	0.26	0.29	4/29/10	3MR8,3MRZ,3MS0,3MS1
254	3MRZ	*T.thermophilus*	50S	yes	3.62	0.26	0.29	4/29/10	3MR8,3MRZ,3MS0,3MS1
255	3MS0	*T.thermophilus*	30S	yes	3.62	0.26	0.29	4/29/10	3MR8,3MRZ,3MS0,3MS1
256	3MS1	*T.thermophilus*	50S	yes	3.62	0.26	0.29	4/29/10	3MR8,3MRZ,3MS0,3MS1
257	1N36	*T.thermophilus*	30S	yes	3.65	0.26	0.32	10/25/02	
258	1VSA	*T.thermophilus*	50S	yes	3.71	0.35	0.35	2/15/07	1VSA,2OW8
259	2OW8	*T.thermophilus*	30S	yes	3.71	0.35	0.35	2/15/07	1VSA,2OW8
260	1N34	*T.thermophilus*	30S	yes	3.80	0.24	0.31	10/25/02	
261	2F4V	*T.thermophilus*	30S	no	3.80	0.26	0.32	11/24/05	
262	1VSP	*T.thermophilus*	50S	yes	3.83	0.33	0.35	7/18/07	1VSP,2QNH
263	2QNH	*T.thermophilus*	30S	yes	3.83	0.33	0.35	7/18/07	1VSP,2QNH

Publications and Conferences

Publications

"Validation of B Factor Distributions in Protein Crystal Structures by Comparison to Shifted Inverse-Gamma Distributions". In preparation.

Conferences

XXII Congress and General Assembly, International Union of Crystallography (IUCr2011), Madrid, Spain (August 22 – August 29, 2011)

Selected oral presentation: *"Validation of B factor Distributions in Protein Crystal Structures"*.

26th European Crystallographic Meeting (ECM26), Darmstadt, Germany (August 29 – September 2, 2010)

Poster presentation: *"Validation of B factor Distributions in Protein Crystal Structures"*.

EMBO Conference: 8th International Conference on Ribosome Synthesis, Regensburg, Germany (August 26 – August 30, 2009)

25th European Crystallographic Meeting (ECM25), Istanbul, Turkey (August 16 – August 21, 2009)

Poster presentation: *"Coordinate Variations in Structural Ensembles from Different Refinement Methods"*.

Europass Curriculum Vitae

Personal information

Surname(s) / First name(s)	**Negroni Jacopo**
Address(es)	Notkestrasse 85, EMBL c/o DESY Building 25A, 22607, Hamburg, Germany
Telephone(s)	0049-(0)40-89902-182 Mobile 0049-(0)176-96135105
E-mail	jacopo.negroni@embl-hamburg.de
Nationality	Italian
Date of birth	06/06/1982
Gender	Male

Education and training

Dates	01/10/2007 – Ongoing
Position	PhD student at EMBL-Hamburg in Thomas R. Schneider's Group
Principal subjects/occupational skills covered	PhD Project: "Validation of Crystallographic B Factors and Analysis of Ribosomal Crystal Structures"
Name and type of organisation providing education and training	European Molecular Biology Laboratory Hamburg, Notkestrasse 85, EMBL c/o DESY Building 25A, 22607, Hamburg, Germany. Thomas R. Schneider's Group.

Dates	2004-2007
Title of qualification awarded	Two-year specialisation Degree (MSc) in Bioinformatics.
Principal subjects/occupational skills covered	Thesis title: "Development and Automation of a Bioinformatics Protocol for the Analysis of Ligand-Protein Interactions using High Performance Computing.", a molecular docking study. Mark: 110 cum laude. Supervisor: Dr. Luciano Milanesi.
Name and type of organisation providing education and training	Università degli Studi di Milano-Bicocca, edificio U3, Piazza della scienza 2, Milano, Italy. Institute for Biomedical Technologies – National Research Council (ITB-CNR), Via Fratelli Cervi 93, 20090 Segrate (MI).

Dates	2001-2004
Title of qualification awarded	Three year First Degree (BSc) in Molecular Biotechnologies
Principal subjects/occupational skills covered	Thesis title: "Computational Study of the 3D Structural Model of Ethe1p, a Protein Involved in Ethylmalonic Encephalopathy.", an homology modelling study. Mark: 110 cum laude. Supervisor: Prof. Luca de Gioia.
Name and type of organisation providing education and training	Università degli Studi di Milano-Bicocca, edificio U3, Piazza della scienza 2, Milano, Italy.

Dates	1996-2001
Title of qualification awarded	High school Degree
Name and type of organisation providing education and training	Liceo Scientifico Statale Filippo Lussana, via Angelo Maj 1, Bergamo, Italy

Acknowledgements

The work presented in this PhD thesis was carried out in Thomas R. Schneider's group at EMBL-Hamburg.

I would like to thank my supervisor Thomas R. Schneider for giving me the opportunity to work in his group and for helping and supporting me during my PhD. I am thankful to him for initiating me to the field of macromolecular crystallography, especially for allowing me to investigate the crystallographic topics in which I got interested at the beginning of my doctorate (i.e. macromolecular refinement, atomic coordinate precision and B factors). Moreover I would like to thank him for very useful and constructive criticisms during the writing of this thesis.

I am deeply grateful to Garib N. Murshudov for accepting to collaborate with us in the project focused on the validation of B factor distributions in crystallographic macromolecules. Without his help and assistance, especially for the statistical and computational part, the success of the project would had been way more uncertain. I would also like to thank him for guesting me for one week in the campus of the University of York at the beginning of our collaboration.

A special thank goes to the members of my thesis advisory committee Thomas R. Schneider, Dmitri Svergun, Janet Thornton from EMBL and Irmgard Sinning from Heidelberg University for the useful feedback and suggestions they gave me at the end of each annual report. I would also like to thank Dmitri Svergun and Irmgard Sinning for accepting to be the referees of my PhD thesis, and Robert Russell and Sabine Strahl from Heidelberg university for accepting to be part of my PhD thesis defence committee.

I am thankful to Roberto Mosca, a former member of Thomas R Schneider's lab, for helping me with the ESCET framework and its code, and for fruitful suggestions and support at the beginning of my doctorate. I would like to thank Fabio dall'Antonia for useful discussion about the analysis of ensemble of structures with the ESCET framework and for helping me during the analysis of the ensemble of ribosomal structures from *T. thermophilus*. I am grateful to Philipp Heuser and Florian Sauer for translating the summary of this thesis from English to German.

I would also like to thank Fabio dall'Antonia, Rosemary Wilson, Jonathan Rapley, Chris

Williams, and Philipp Heuser for proofreading parts of my thesis and for giving me useful feedback.

A huge thank to my family, my friends from Italy, and to all the wonderful people I had the chance to interact with during these four years at EMBL-Hamburg for their friendship and for their support in moments of despair, which are not so uncommon during a PhD.

Last but not least, I would like to thank the European Molecular Biology Laboratory for funding my PhD and for giving me the opportunity to work in such a stimulating and multicultural environment.

Bibliography

[1] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart. PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography*, 66(2):213–221, Jan. 2010.

[2] P. V. Afonine, A. Urzhumtsev, R. W. Grosse-Kunstleve, and P. D. Adams. Atomic displacement parameters (ADPs), their parameterization and refinement in phenix. *Crystallography Computational Newsletter*, 1(1):24–31, July 2010.

[3] I. Agmon, A. Bashan, R. Zarivach, and A. Yonath. Symmetry at the active site of the ribosome: structural and functional implications. *Biological Chemistry*, 386(9):833–844, Sept. 2005.

[4] D. Ågren, M. Stehr, C. L. Berthold, S. Kapoor, W. Oehlmann, M. Singh, and G. Schneider. Three-Dimensional structures of apo- and Holo-l-Alanine dehydrogenase from mycobacterium tuberculosis reveal conformational changes upon coenzyme binding. *Journal of Molecular Biology*, 377(4):1161–1173, Apr. 2008.

[5] F. H. Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B Structural Science*, 58(3):380–388, May 2002.

[6] A. Amy C. The process of Structure-Based drug design. *Chemistry & Biology*, 10(9):787–797, Sept. 2003.

[7] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93 –96, Oct. 2001.

[8] P. I. d. Bakker, N. Furnham, T. L. Blundell, and M. A. DePristo. Conformer generation under restraints. *Current Opinion in Structural Biology*, 16(2):160–165, Apr. 2006.

[9] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481):905–920, Aug. 2000.

[10] D. Baram, E. Pyetan, A. Sittner, T. Auerbach-Nevo, A. Bashan, and A. Yonath. Structure of trigger factor binding domain in biologically homologous complex with eubacterial ribosome reveals its chaperone action. *Proceedings of the National Academy of Sciences of the United States of America*, 102(34):12017 –12022, 2005.

[11] M. Belousoff, C. Davidovich, E. Zimmerman, Y. Caspi, I. Wekselman, L. Rozenszajn, T. Shapira, O. Sade-Falk, L. Taha, A. Bashan, M. Weiss, and A. Yonath. Ancient machinery embedded in the contemporary ribosome. *Biochemical Society Transactions*, 38:422, Apr. 2010.

[12] A. Ben-Shem, L. Jenner, G. Yusupova, and M. Yusupov. Crystal structure of the eukaryotic ribosome. *Science*, 330(6008):1203 –1209, Nov. 2010.

[13] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nat Struct Mol Biol*, 10(12):980, Dec. 2003.

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, Jan. 2000.

[15] P. Bevington and D. K. Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill Science/Engineering/Math, 3rd edition, July 2002.

[16] T. Biswas, L. Yi, P. Aggarwal, J. Wu, J. R. Rubin, J. A. Stuckey, R. W. Woodard, and O. V. Tsodikov. The tail of KdsC: conformational changes control the activity of a haloacid dehalogenase superfamily phosphatase. *Journal of Biological Chemistry*, 284(44):30594 –30603, Oct. 2009.

[17] K. Bokov and S. V. Steinberg. A hierarchical model for evolution of 23S ribosomal RNA. *Nature*, 457(7232):977–980, Feb. 2009.

[18] W. M. Bolstad. *Introduction to Bayesian Statistics*. Wiley-Interscience, 2 edition, Aug. 2007.

[19] R. Bott and J. Frane. Incorporation of crystallographic temperature factors in the statistical analysis of protein tertiary structures. *Protein Engineering*, 3(8):649–657, 1990.

[20] G. P. Bourenkov and A. N. Popov. A quantitative approach to data-collection strategies. *Acta Crystallographica Section D Biological Crystallography*, 62:58–64, Dec. 2005.

[21] C. Brändén, , and J. Tooze. *Introduction to Protein Structure*. Garland Science, 2 edition, Jan. 1999.

[22] C. Brändén and T. A. Jones. Between objectivity and subjectivity. *Nature*, 343(6260):687–689, Feb. 1990.

[23] D. E. Brodersen, W. M. Clemons Jr., A. P. Carter, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin b on the 30S ribosomal subunit. *Cell*, 103(7):1143–1154, Dec. 2000.

[24] D. E. Brodersen, W. M. Clemons Jr, A. P. Carter, B. T. Wimberly, and V. Ramakrishnan. Crystal structure of the 30 s ribosomal subunit from thermus thermophilus: structure of the proteins and their interactions with 16 s RNA. *Journal of Molecular Biology*, 316(3):725–768, Feb. 2002.

[25] P. J. Brown, A. G. Fox, E. N. Maslen, M. A. O'Keefe, and B. T. M. Willis. Intensity of diffracted intensities. In E. Prince, editor, *International Tables for Crystallography*, volume C, pages 554–595. International Union of Crystallography, Chester, England, 1 edition, Oct. 2006.

[26] A. T. Brunger. Free r value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, Jan. 1992.

[27] A. T. Brunger. Assessment of phase accuracy by cross validation: the free r value. methods and applications. *Acta Crystallographica Section D Biological Crystallography*, 49:24–36, Jan. 1993.

[28] A. T. Brunger. Free r value: Cross-validation in crystallography. In *Methods in Enzymology*, volume Volume 277, pages 366–396. Academic Press, 1997.

[29] A. Busch, A. S. Richter, and R. Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849 –2856, Dec. 2008.

[30] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3:2, Jan. 2002. PMID: 11869452 PMCID: 65690.

[31] A. P. Carter, W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, T. Hartsch, B. T. Wimberly, and V. Ramakrishnan. Crystal structure of an initiation factor bound to the 30S ribosomal subunit. *Science*, 291(5503):498 –501, Jan. 2001.

[32] A. P. Carter, W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, 407(6802):340–348, 2000.

[33] J. L. Chambers and R. M. Stroud. The accuracy of refined protein structures: comparison of two independently refined models of bovine trypsin. *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry*, 35:1861–1874, Aug. 1979.

[34] W. J. Cody. Algorithm 665: Machar: a subroutine to dynamically determined machine parameters. *ACM Trans. Math. Softw.*, 14(4):303–311, Dec. 1988.

[35] S. X. Cohen, R. J. Morris, F. J. Fernandez, M. Ben Jelloul, M. Kakaris, V. Parthasarathy, V. S. Lamzin, G. J. Kleywegt, and A. Perrakis. Towards complete validated models in the next generation of ARP/wARP. *Acta Crystallographica Section D Biological Crystallography*, 60(12):2222–2229, Nov. 2004.

[36] D. W. J. Cruickshank. Remarks about protein structure precision. *Acta Crystallographica Section D Biological Crystallography*, 55(3):583–601, Mar. 1999.

[37] Z. Dauter. Data-collection strategies. *Acta Crystallographica Section D Biological Crystallography*, 55:1703–1717, Oct. 1999.

[38] Z. Dauter, G. N. Murshudov, and K. S. Wilson. Refinement at atomic resolution. In M. G. Rossmann and E. Arnold, editors, *International Tables for Crystallography*, volume F, pages 393–402. International Union of Crystallography, Chester, England, 1 edition, Oct. 2006.

[39] I. W. Davis, L. W. Murray, J. S. Richardson, and D. C. Richardson. MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Research*, 32(Web Server):W615–W619, July 2004.

[40] M. A. DePristo, P. I. de Bakker, and T. L. Blundell. Heterogeneity and inaccuracy in protein structures solved by X-Ray crystallography. *Structure*, 12(5):831–838, May 2004.

[41] M. A. DePristo, P. I. de Bakker, R. J. Johnson, and T. L. Blundell. Crystallographic refinement by Knowledge-Based exploration of complex energy landscapes. *Structure*, 13(9):1311–1319, Sept. 2005.

[42] E. Dodson, G. J. Kleywegt, and K. Wilson. Report of a workshop on the use of statistical validators in protein x-ray crystallography. *Acta Crystallographica Section D Biological Crystallography*, 52(1):228–234, Jan. 1996.

[43] J. Drenth. Introduction to basic crystallography. In M. G. Rossmann and E. Arnold, editors, *International Tables for Crystallography*, volume F, pages 45–63. International Union of Crystallography, Chester, England, 1 edition, Oct. 2006.

[44] H. Driessen, M. I. J. Haneef, G. W. Harris, B. Howlin, G. Khan, and D. S. Moss. RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures. *Journal of Applied Crystallography*, 22(5):510–516, 1989.

[45] C. M. Dunham, M. Selmer, S. S. Phelps, A. C. Kelley, T. Suzuki, S. Joseph, and V. Ramakrishnan. Structures of tRNAs with an expanded anticodon loop in the decoding center of the 30S ribosomal subunit. *RNA*, 13(6):817 –823, June 2007.

[46] J. A. Dunkle and J. H. Cate. Ribosome structure and dynamics during translocation and termination. *Annual Review of Biophysics*, 39(1):227–244, Apr. 2010.

[47] B. Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, Dec. 1981.

[48] B. Efron. Second thoughts on the bootstrap. *Statistical Science*, 18(2):135–140, May 2003.

[49] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):486–501, Apr. 2010.

[50] P. Evans. Scaling and assessment of data quality. *Acta Crystallographica Section D Biological Crystallography*, 62(1):72–82, Dec. 2005.

[51] N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, and H. Stark. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature*, 466(7304):329–333, July 2010.

[52] R. R. Forseth and F. C. Schroeder. NMR-spectroscopic analysis of mixtures: from structure to function. *Current Opinion in Chemical Biology*, 15(1):38–47, Feb. 2011.

[53] J. Frank. Single-particle reconstruction of biological macromolecules in electron microscopy – 30 years. *Quarterly reviews of biophysics*, 42(3):139–158, Aug. 2009. PMID: 20025794 PMCID: 2844734.

[54] J. Frank and R. K. Agrawal. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, 406(6793):318–322, July 2000.

[55] N. Furnham, T. L. Blundell, M. A. DePristo, and T. C. Terwilliger. Is one solution good enough? *Nat Struct Mol Biol*, 13(3):184–185, Mar. 2006.

[56] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. SIAM, Society for Industrial and Applied Mathematics, May 2007.

[57] S. Godsill. The shifted inverse-gamma model for noise-floor estimation in archived audio recordings. *Signal Processing*, 90(4):991–999, Apr. 2010.

[58] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves. Bio3d: an r package for the comparative analysis of protein structures. *Bioinformatics*, 22(21):2695 –2696, Nov. 2006.

[59] J. Guss, P. R. Harrowell, M. Murata, V. A. Norris, and H. C. Freeman. Crystal structure analyses of reduced (CuI) poplar plastocyanin at six pH values. *Journal of Molecular Biology*, 192(2):361–387, Nov. 1986.

[60] S. R. Holbrook. Structural principles from large RNAs. *Annual Review of Biophysics*, 37(1):445–464, June 2008.

[61] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods, 2nd Edition*. Wiley-Interscience, 2 edition, Jan. 1999.

[62] J. M. Holton. A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, 16(2):133–142, 2009.

[63] T. R. Hvidsten, A. Lægreid, A. Kryshtafovych, G. Andersson, K. Fidelis, and J. Komorowski. A comprehensive analysis of the Structure-Function relationship in proteins based on local structure similarity. *PLoS ONE*, 4(7):e6266, July 2009.

[64] S. Jenni, M. Leibundgut, D. Boehringer, C. Frick, B. Mikolsek, and N. Ban. Structure of fungal fatty acid synthase and implications for iterative substrate shuttling. *Science*, 316(5822):254 –261, Apr. 2007.

[65] R. P. Joosten, J. Salzemann, V. Bloch, H. Stockinger, A. Berglund, C. Blanchet, E. Bongcam-Rudloff, C. Combet, A. L. D. Costa, G. Deleage, M. Diarena, R. Fabbretti, G. Fettahi, V. Flegel, A. Gisel, V. Kasam, T. Kervinen, E. Korpelainen, K. Mattila, M. Pagni, M. Reichstadt, V. Breton, I. J. Tickle, and G. Vriend. PDB_REDO: automated re-refinement of x-ray structure models in the PDB. *Journal of Applied Crystallography*, 42(3):376–384, 2009.

[66] R. P. Joosten, T. Womack, G. Vriend, and G. Bricogne. Re-refinement from deposited x-ray data can deliver improved models for most PDB entries. *Acta Crystallographica Section D Biological Crystallography*, 65(2):176–185, 2009.

[67] T. Kaminishi, D. N. Wilson, C. Takemoto, J. M. Harms, M. Kawazoe, F. Schluenzen, K. Hanawa-Suetsugu, M. Shirouzu, P. Fucini, and S. Yokoyama. A snapshot of the 30S

ribosomal subunit capturing mRNA via the Shine-Dalgarno interaction. *Structure*, 15(3):289–297, Mar. 2007.

[68] L. A. Kelley, S. P. Gardner, and M. J. Sutcliffe. An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. *Protein Engineering*, 10(6):737 –741, June 1997.

[69] D. Klein, P. Moore, and T. Steitz. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *Journal of Molecular Biology*, 340(1):141–177, June 2004.

[70] G. J. Kleywegt. Validation of protein crystal structures. *Acta Crystallographica Section D Biological Crystallography*, 56(3):249–265, Mar. 2000.

[71] G. J. Kleywegt. On vital aid: the why, what and how of validation. *Acta Crystallographica Section D Biological Crystallography*, 65(2):134–139, Jan. 2009.

[72] G. J. Kleywegt, T. Alwyn Jones, and R. M. S. Charles W. Carter Jr. Model building and refinement practice. In *Macromolecular Crystallography Part B*, volume Volume 277, pages 208–230. Academic Press, 1997.

[73] S. Klinge, F. Voigts-Hoffmann, M. Leibundgut, S. Arpagaus, and N. Ban. Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science*, 334(6058):941 –948, Nov. 2011.

[74] J. L. Knight, Z. Zhou, E. Gallicchio, D. M. Himmel, R. A. Friesner, E. Arnold, and R. M. Levy. Exploring structural variability in x-ray crystallographic models using protein local optimization by torsion-angle sampling. *Acta Crystallographica Section D Biological Crystallography*, 64(4):383–396, Mar. 2008.

[75] A. Korostelev, D. N. Ermolenko, and H. F. Noller. Structural dynamics of the ribosome. *Current Opinion in Chemical Biology*, 12(6):674–683, Dec. 2008.

[76] S. Kurata, A. Weixlbaumer, T. Ohtsuki, T. Shimazaki, T. Wada, Y. Kirino, K. Takai, K. Watanabe, V. Ramakrishnan, and T. Suzuki. Modified uridines with c5-methylene substituents at the first position of the tRNA anticodon stabilize UG wobble pairing during decoding. *Journal of Biological Chemistry*, 283(27):18801 –18811, July 2008.

[77] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26:283–291, Apr. 1993.

[78] A. G. W. Leslie. The integration of macromolecular diffraction data. *Acta Crystallographica Section D Biological Crystallography*, 62(1):48–57, Dec. 2005.

[79] A. Liljas. *Structural Aspects Of Protein Synthesis*. World Scientific Pub Co Inc, 1 edition, Nov. 2004.

[80] A. J. McCoy. Liking likelihood. *Acta Crystallographica Section D Biological Crystallography*, 60(12):2169–2183, Nov. 2004.

[81] H. D. Mertens and D. I. Svergun. Structural characterization of proteins and complexes using small-angle x-ray solution scattering. *Journal of Structural Biology*, 172(1):128–141, Oct. 2010.

[82] K. Mitra and J. Frank. RIBOSOME DYNAMICS: insights from atomic structure modeling into Cryo-Electron microscopy maps. *Annual Review of Biophysics and Biomolecular Structure*, 35(1):299–317, June 2006.

[83] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill Higher Education, 3 edition, June 1974.

[84] R. Mosca, B. Brannetti, and T. R. Schneider. Alignment of protein structures in the presence of domain motions. *BMC Bioinformatics*, 9(1):352, 2008.

[85] R. Mosca and T. R. Schneider. RAPIDO: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic Acids Research*, 36(Web Server):W42–W46, May 2008.

[86] F. V. Murphy and V. Ramakrishnan. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nat Struct Mol Biol*, 11(12):1251–1252, Dec. 2004.

[87] F. V. Murphy, V. Ramakrishnan, A. Malkiewicz, and P. F. Agris. The role of modifications in codon discrimination by tRNALysUUU. *Nat Struct Mol Biol*, 11(12):1186–1191, Dec. 2004.

[88] C. W. Murray and T. L. Blundell. Structural biology in fragment-based drug design. *Current Opinion in Structural Biology*, 20(4):497–507, Aug. 2010.

[89] G. N. Murshudov, P. Skubk, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D Biological Crystallography*, 67(4):355–367, Mar. 2011.

[90] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Refinement of macromolecular structures by the Maximum-Likelihood method. *Acta Crystallographica Section D Biological Crystallography*, 53(3):240–255, May 1997.

[91] J. M. Ogle, D. E. Brodersen, W. M. Clemons, M. J. Tarry, A. P. Carter, and V. Ramakrishnan. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, 292(5518):897–902, May 2001.

[92] J. M. Ogle, A. P. Carter, and V. Ramakrishnan. Insights into the decoding mechanism from recent ribosome structures. *Trends in Biochemical Sciences*, 28(5):259–266, May 2003.

[93] J. M. Ogle, F. V. Murphy IV, M. J. Tarry, and V. Ramakrishnan. Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell*, 111(5):721–732, Nov. 2002.

[94] J. M. Ogle and V. Ramakrishnan. STRUCTURAL INSIGHTS INTO TRANSLATIONAL FIDELITY. *Annual Review of Biochemistry*, 74(1):129–177, July 2005.

[95] A. O'Hagan and F. Jonathan. *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Wiley-Blackwell, 2 edition, July 2004.

[96] C. A. Orengo, A. E. Todd, and J. M. Thornton. From protein structure to function. *Current Opinion in Structural Biology*, 9(3):374–382, June 1999.

[97] S. Parthasarathy and M. Murthy. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Science*, 6(12):2561–2567, Dec. 1997.

[98] K. M. Perry, E. B. Fauman, J. S. Finer-Moore, W. R. Montfort, G. F. Maley, F. Maley, and R. M. Stroud. Plastic adaptation toward mutations in proteins: Structural comparison of thymidylate synthases. *Proteins: Structure, Function, and Bioinformatics*, 8(4):315–333, Jan. 1990.

[99] A. Petrov, G. Kornberg, S. O'Leary, A. Tsai, S. Uemura, and J. D. Puglisi. Dynamics of the translational machinery. *Current Opinion in Structural Biology*, 21(1):137–145, Feb. 2011.

[100] G. Qi, R. Lee, and S. Hayward. A comprehensive and non-redundant database of protein domain movements. *Bioinformatics*, 21(12):2832 –2838, June 2005.

[101] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[102] J. Rabl, M. Leibundgut, S. F. Ataide, A. Haag, and N. Ban. Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. *Science*, 331(6018):730 –736, Feb. 2011.

[103] V. Ramakrishnan. The eukaryotic ribosome. *Science*, 331(6018):681 –682, Feb. 2011.

[104] R. J. Read, P. D. Adams, W. B. Arendall III, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart. A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, Oct. 2011.

[105] S. M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 4th ed. edition, Feb. 2009.

[106] B. Rupp. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Taylor & Francis Ltd., 1 edition, Jan. 2007.

[107] B. Rupp. Scientific inquiry, inference and critical reasoning in the macromolecular crystallography curriculum. *Journal of Applied Crystallography*, 43(5):1242–1249, Aug. 2010.

[108] F. Schluenzen, C. Takemoto, D. N. Wilson, T. Kaminishi, J. M. Harms, K. Hanawa-Suetsugu, W. Szaflarski, M. Kawazoe, M. Shirouzo, K. H. Nierhaus, S. Yokoyama, and P. Fucini. The antibiotic kasugamycin mimics mRNA nucleotides to destabilize tRNA binding and inhibit canonical translation initiation. *Nat Struct Mol Biol*, 13(10):871–878, Oct. 2006.

[109] T. M. Schmeing and V. Ramakrishnan. What recent ribosome structures have revealed about the mechanism of translation. *Nature*, 461(7268):1234–1242, Oct. 2009.

[110] T. R. Schneider. Objective comparison of protein structures: error-scaled difference distance matrices. *Acta Crystallographica Section D Biological Crystallography*, 56(6):714–721, June 2000.

[111] T. R. Schneider. A genetic algorithm for the identification of conformationally invariant regions in protein molecules. *Acta Crystallographica Section D Biological Crystallography*, 58(2):195–208, Jan. 2002.

[112] T. R. Schneider. Domain identification by iterative analysis of error-scaled difference distance matrices. *Acta Crystallographica Section D Biological Crystallography*, 60(12):2269–2275, Nov. 2004.

[113] D. Schwarzenbach, S. C. Abrahams, H. D. Flack, W. Gonschorek, T. Hahn, K. Huml, R. E. Marsh, E. Prince, B. E. Robertson, J. S. Rollett, and A. J. C. Wilson. Statistical descriptors in crystallography: Report of the IUCr subcommittee on statistical descriptors. *Acta Crystallographica Section A Foundations of Crystallography*, 45(1):63–75, Jan. 1989.

[114] M. Selmer, C. M. Dunham, F. V. Murphy, A. Weixlbaumer, S. Petry, A. C. Kelley, J. R. Weir, and V. Ramakrishnan. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, 313(5795):1935 –1942, 2006.

[115] J. Shine and L. Dalgarno. Terminal-Sequence analysis of bacterial ribosomal RNA. *European Journal of Biochemistry*, 57(1):221–230, Sept. 1975.

[116] J. Skolnick and J. S. Fetrow. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in Biotechnology*, 18(1):34–39, Jan. 2000.

[117] T. A. Steitz. A structural understanding of the dynamic ribosome machine. *Nat Rev Mol Cell Biol*, 9(3):242–253, Mar. 2008.

[118] A. C. Steven and W. Baumeister. The future is hybrid. *Journal of Structural Biology*, 163(3):186–195, Sept. 2008.

[119] A. Stuart and K. Ord. *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Wiley-Blackwell, 6 edition, June 1994.

[120] A. Stuart, K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*. Wiley-Blackwell, 6 edition, Dec. 1999.

[121] M. T. Sykes and J. R. Williamson. A complex assembly landscape for the 30S ribosomal subunit. *Annual Review of Biophysics*, 38:197–215, June 2009.

[122] G. L. Taylor. Introduction to phasing. *Acta Crystallographica Section D Biological Crystallography*, 66:325–338, Mar. 2010.

[123] T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V. Afonine, P. D. Adams, N. W. Moriarty, P. Zwart, R. J. Read, D. Turk, and L. Hung. Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. *Acta Crystallographica Section D Biological Crystallography*, 63(5):597–610, Apr. 2007.

[124] T. C. Terwilliger, D. Stuart, and S. Yokoyama. Lessons from structural genomics. *Annual review of biophysics*, 38:371–383, Jan. 2009.

[125] D. E. Tronrud. Introduction to macromolecular refinement. *Acta Crystallographica Section D Biological Crystallography*, 60(12):2156–2168, Nov. 2004.

[126] K. N. Trueblood, H. B. Bürgi, H. Burzlaff, J. D. Dunitz, C. M. Gramaccioli, H. H. Schulz, U. Shmueli, and S. C. Abrahams. Atomic dispacement parameter nomenclature. report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallographica Section A Foundations of Crystallography*, 52(5):770–781, Sept. 1996.

[127] L. Urzhumtseva, P. V. Afonine, P. D. Adams, and A. Urzhumtsev. Crystallographic model quality at a glance. *Acta Crystallographica Section D Biological Crystallography*, 65(3):297–300, Feb. 2009.

[128] A. A. Vaguine, J. Richelle, and S. J. Wodak. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallographica Section D Biological Crystallography*, 55(1):191–205, Jan. 1999.

[129] M. Valle, A. Zavialov, J. Sengupta, U. Rawat, M. Ehrenberg, and J. Frank. Locking and unlocking of ribosomal motions. *Cell*, 114(1):123–134, July 2003.

[130] J. E. Wampler. Distribution analysis of the variation of B-Factors of x-ray crystal structures: Temperature and structural variations in lysozyme. *Journal of Chemical Information and Computer Sciences*, 37(6):1171–1180, Nov. 1997.

[131] J. Wang and D. Boisvert. Structural basis for GroEL-assisted protein folding from the crystal structure of $(\text{GroEL-KMgATP})_{14}$ at 2.0 Å resolution. *Journal of Molecular Biology*, 327(4):843–855, Apr. 2003.

[132] A. Weixlbaumer, F. V. Murphy, A. Dziergowska, A. Malkiewicz, F. A. P. Vendeix, P. F. Agris, and V. Ramakrishnan. Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. *Nat Struct Mol Biol*, 14(6):498–502, June 2007.

[133] B. T. M. Willis and A. W. Pryor. *Thermal Vibrations in Crystallography*. Cambridge University Press, 1 edition, Mar. 1975.

[134] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan. Structure of the 30S ribosomal subunit. *Nature*, 407(6802):327–339, 2000.

[135] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D Biological Crystallography*, 67:235–242, Mar. 2011.

[136] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, 1997.