BETTER MODELS IN MACROMOLECULAR CRYSTAL STRUCTURE DETERMINATION

DISSERTATION

zur Erlangung des akademischen Grades eines Doktors der
Naturwissenschaften (Dr. rer. nat.) im

Fachbereich Chemie

der Universität Hamburg

vorgelegt von

TIM WIEGELS
geboren am 29.10.1982 in
Stralsund, Deutschland

Hamburg, den 13. Juli 2012

Die vorliegende Arbeit wurde in der Zeit von September 2008 bis Juni 2012 unter der externen Leitung von

Herrn Dr. Victor S. Lamzin

im Europäischen Laboratorium für Molekularbiologie (im folgenden EMBL) Hamburg Outstation angefertigt. Die universitäre Betreuung erfolgte durch

Herrn Prof. Dr. Andrew E. Torda

in der Abteilung für Biomolekulare Modellierung des Zentrums für Bioinformatik der Fakultät für Mathematik, Informatik und Naturwissenschaften an der Universität Hamburg. Mitglieder des, durch das EMBL vorgeschriebene, Thesis Advisory Committees (TAC) waren neben Herrn Dr. Victor Lamzin und Herrn Prof. Dr. Andrew Torda: Frau Dr. Anne-Claude Gavin (Structural Biology, EMBL Heidelberg), Herr Dr. Richard Morris (Computational and Systems Biology, John Innes Centre, Norwich), Herr Dr. Thomas Schneider (Structural Biology, EMBL Hamburg) und Herr Dr. Manfred Weiss (Makromolekulare Kristallographie, Helmholtz Zentrum Berlin).

1. Gutachter: Prof. Dr. Andrew E. Torda[1]

2. Gutachter: Prof. Dr. Dr. Christian Betzel[2]

Externer Betreuer: Dr. Victor S. Lamzin[3]

Tag der Disputation: 7. September 2012

---

[1] Zentrum für Bioinformatik, Biomolekulare Modellierung, Bundesstrasse 43 - 20146 Hamburg
[2] Abteilung für Biochemie und Molekularbiologie, Martin-Luther-King Platz 6 - 20146 Hamburg
[3] EMBL, Hamburg Outstation, Notkestrasse 85 - 22603 Hamburg

Preface

Parts of this thesis (text and figures) have been published in peer reviewed journal and have further been presented as posters and talks at conferences and workshops:

Peer-reviewed Publications

- T Wiegels and VS Lamzin. Use of non-crystallographic symmetry for automated model building at medium to low resolution. *Acta Crystallogr. D Biol. Crystallogr*. 68:446-453, **2012** [1]

Oral Presentations at Conferences

- “Exploiting synergy between computational biology and X-ray crystallography for solving challenging macromolecular structures”; 1st European Student Council Symposium (ESCS1), 9th European Conference on Computational Biology (ECCB 2010), Ghent Belgium, Sep **2010**
- “Release 7.2 of ARP/wARP Software Suite”; MS058 - New Computational Approaches to Structure Solution and Refinement, XXII Congress and General Assembly of the International Union of Crystallography (IUCr2011), Madrid, Spain, Aug **2011** [2]

Oral Presentations at Workshops

- “Structure Solution: CCP4 Seminar and Workshop”, Osaka University, Japan, Nov **2010**
- “CCP4 APS Summer School 2011”, Argonne National Laboratory, USA, Jun **2011** & Jun **2012**
- “Software Fayre” at XXII Congress and General Assembly of the International Union of Crystallography (IUCr2011), Madrid, Spain, Aug **2011**
- “CCP4 School on Advanced X-ray crystal structure analysis”, Australian Synchrotron, Melbourne, Australia, Feb **2012**
- “EMBO Practical Course ’Computational aspects of protein structure determination and analysis: from data to structure to function”; EMBL-EBI, Hinxton, United Kingdom, Nov **2012** & Apr **2012**

Poster Presentations at Conferences

- “Automatic Completion of auto-traced protein fragments”; EMBL-EBI Bioinformatics Workshop 2009, EMBL-EBI, Hinxton, United Kingdom, Nov **2009**
- “Exploiting synergy between computational biology and X-Ray crystallography for solving challenging macromolecular structures”; 9th European Conference on Computational Biology (ECCB 2010), Ghent, Belgium, Sep **2010**
- “Towards more complete protein models in macromolecular crystal structure determination”; 3DSig, 19th Annual International Conference on Intelligent Systems for Molecular Biology & 10th European Conference on Computational Biology (ISMB/ECCB 2011), Vienna, Austria, Jul **2011**
- “Towards more complete protein models in macromolecular crystal structure determination”; MS058 - New Computational Approaches to Structure Solution and Refinement, XXII Congress and General Assembly of the International Union of Crystallography (IUCr2011), Madrid, Spain, Aug **2011** [3]

Molecular graphics images were produced using the ARPnavigator and the UCSF Chimera package [4].

Contents

Abbreviations and Notation 9

1 Introduction 11

1.1 Overview 11

1.2 Protein structure 12

1.3 Macromolecular crystallography 13

1.3.1 Automated model building 16

1.3.2 Structure refinement 19

1.3.3 Model validation 20

1.4 Challenges in macromolecular structure determination 21

1.5 *ARP/wARP* 23

1.6 Non-crystallographic symmetry - NCS 28

1.7 Theoretical modelling 30

1.7.1 Homology modelling 31

1.7.2 *de novo* protein structure prediction 31

1.8 Where do we stand? 34

1.8.1 Automatic model building at medium-to-low resolution 34

1.8.2 Theoretical modelling in macromolecular crystallography 36
1.9 Scope of this thesis 38

2 Methodological Background **41**
2.1 Rotations in $\mathbb{R}^3$ 41
2.1.1 Rotation matrices 41
2.1.2 Euler angles 43
2.1.3 Quaternions 44
2.2 Translations in $\mathbb{R}^3$ 45
2.3 Rigid body superposition 46
2.3.1 Root mean square deviation 47
2.4 Clustering 48
2.5 Protein secondary structure - assignment and prediction 49
2.5.1 Assignment of secondary structure to protein chains 50
2.5.2 Protein secondary structure prediction 50
2.6 String matching algorithms 52

3 Materials and Methods Developed **53**
3.1 PNSextender - Automatic NCS identification for extension and restraints . 53
3.1.1 Clustering of transformations between chain fragments and identification of NCS-related copies 54
3.1.2 Improving structural information by transformation of NCS copies 57
3.1.3 Weighting the detected *extensions* 58
3.1.4 Use of identified NCS-based chain fragment *extensions* for model building 59
3.1.5 Derivation of NCS-based stereochemical restraints: 59
3.2 FittOFF - Fragment extension by motif comparison 60
3.2.1 Identification of gaps in intermediate models 63
3.2.2 FRAGRA 70

3.2.3 Application of fitted fragments to model building 72
3.3 Data 73
3.4 Tests environments 73
3.5 Implementation and complexity 74

4 Results **77**
4.1 The importance of scoring *extensions* derived by NCS-identification 77
4.2 The necessity of using secondary structure content for relating gap length to distance 78
4.3 Completeness in the absence of the coordinate error 79
4.4 Application to *ARP/wARP* protein model building 86

5 Discussion **95**
5.1 Considerations for further research 99
5.2 Conclusion 102
5.2.1 Availability to the community 103

6 Zusammenfassung **105**

7 Summary **107**

Bibliography **109**

A Testcases **123**

B Detailed results **125**

C Acknowledgements **131**

D Gefahrstoffe und KMR-Substanzen **133**

E Erklärungen **135**

Abbreviations and Notation

General abbreviations

ADP	Atomic displacement parameter
CCP4	Collaborative Computational Project, Number 4
EM	Electron microscopy
EMBL	European Molecular Biology Laboratory
FittOFF	Fitting Of Fragments
MAD	Multi-wavelength anomalous dispersion
MX	Macromolecular crystallography
NCS	Non-crystallographic symmetry
NMR	Nuclear magnetic resonance
PDB	RCSB protein database
PDB ID	Identifier of a structure in the PDB
PNSextender	Protein NCS-based Structure extender
SAD	Single wavelength anomalous dispersion
SAXS	Small angle X-ray scattering
SIRAS	Single isomorphous replacement with anomalous scattering

Biochemistry

$C\alpha$	Carbon alpha atom
N	Nitrogen atom
O	Oxygen atom

Mathematical terms

$P(x \mid y)$	Probability of x given y
$P(x)$	Probability of x
$rmsd$	Root mean square deviation

Resolution

high resolution	Better (higher) than 2.0 Å
low resolution	Worse (lower) than 3.0 Å
medium resolution	Between 2.0 Å and 3.0 Å

Chapter

1

Introduction

1.1 Overview

Macromolecular structures, involving proteins, DNA RNA or complexes thereof, are the main focus of attention in structural biology. This can be attributed to their high biomedical significance and their role as major players in the key processes of life. In order to obtain a full understanding of their function and to gain new insights, it is crucial to have a complete knowledge of the spatial arrangement of their constituent atomic blocks. Important applications of 3D macromolecular structures can be found in diverse areas of pharmaceutical and biotechnological industry and research.

There are a number of methods that can be used to obtain structural knowledge of a macromolecule. Macromolecular crystallography (MX) is the most important technique for the determination of biomolecular structures at an atomic of detail. MX has provided over 85% of all entries in the Protein Data Bank [5, 6] and over 90% of proteins that are larger than 80 amino acids. The continuous growth in the number of PDB entries demonstrates the increasing demand for crystallographic 3D models of biological macromolecules. Other experimental methods providing structural information include nuclear magnetic resonance (NMR) [7], electron microscopy (EM) [8], electron tomography [9], electron diffraction [10], neutron diffraction [11] and small angle scattering (SAXS) [12]. Structure prediction and molecular modelling is also gaining popularity [13].

Within the scope of this thesis several methods have been developed to increase the completeness and accuracy of models obtained from automatic model building of proteins in MX. Significant improvements have been obtained, particularly for model building using low resolution crystallographic electron density maps, from 2.4 to 3.8 Å.

1.2 Protein structure

Proteins are the most abundant and versatile macromolecules in all living systems. They provide stability to cells and tissues, immune protection, transport and storage of other molecules (such as oxygen), they control and regulate pathways and metabolic networks and they catalyse almost all chemical reactions occurring in living organisms [14]. The 20 naturally occurring amino acids are the basic building blocks of proteins. In an amino acid the central carbon atom, which is called the $C\alpha$ atom, is linked to an amino group, carboxylic group, a hydrogen atom and a side chain (square brackets in Figure 1.1, where the side chain is referred to as R). The link of the carboxyl group of one residue to the amino group of another residue is called peptide bond. The resulting molecule is a dipeptide. This reaction is catalysed by the ribosome that compiles polypeptide chains with a specific amino acid sequence that is determined by a messenger RNA. Proteins are polypeptides that usually consist of a few hundred amino acids. The amino acid sequence of a protein is also called the primary structure. Polypeptides contain a repeating part, the protein backbone, and a variable part, the side chains. The partial double-bond character of the peptide bond prevents a rotation around this bond. Thus, the only degrees of freedom in the protein backbone are the torsion angle ϕ between the $C\alpha$ atom and the amino group and the torsion angle ψ between the $C\alpha$ atom and the carboxyl group (Figure 1.1).

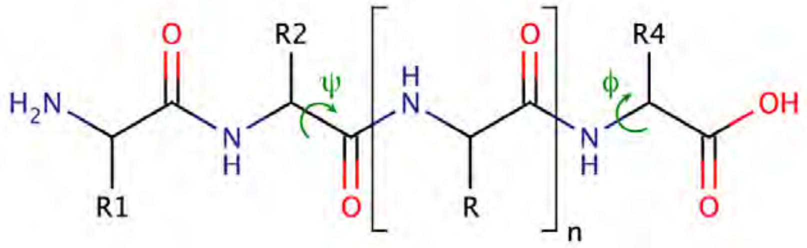


Figure 1.1: Structure of a generic peptide. Square brackets denote one residue, side chains are referred to as R_n. The torsion angles ψ and ϕ are coloured in green.

Hydrogen bonding between backbone atoms define the secondary structure of a protein. The most common secondary structure elements are the α-helix and the β-strand (described in more detail in section 2.5). The three-dimensional conformation of a protein is defined by spatial arrangement of these secondary structure elements and chain sections that link them and is referred to as the tertiary structure. Proteins usually consist of more than one polypeptide chain. Interactions between these chains form and stabilize structures containing several protein subunits. The arrangement the subunits assemble is called quaternary structure [15].

1.3 Macromolecular crystallography

The procedure for obtaining a protein structure in an MX experiment can be seen as consisting of four major steps (Figure 1.2). First, the protein has to be expressed in sufficiently large quantities and purified so that a protein crystal can be grown. This step can take up to several years. During the crystallographic experiment, known as the data collection step, such a protein crystal is mounted in front of a detector and rotated stepwise while being exposed to an incoming X-ray beam. X-rays interact in a specific way with crystalline matter. The result is a set of reflections, collected by the detector, that make up a specific diffraction pattern. An electron density map, which is needed to identify the atomic positions of a macromolecular structure, can be represented as the three-dimensional Fourier transform of an infinite set of complex structure factors [16]. The measured intensities of the collected reflections are proportional to the squared amplitudes of the structure factors. In a unit cell volume, V, the electron density, ρ_{xyz}, at location (x, y, z) can be represented by the following Fourier equation:

$$\rho_{xyz} = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |F_{hkl}| \cos 2\pi(hx + ky + lz - \alpha_{hkl}) \tag{1.1}$$

The equation includes the summation of the amplitudes of the structure factors, F_{hkl} and the phase angle α_{hkl} at location (h, k, l) in reciprocal space [17]. However, during data collection, the phase angles are not directly obtainable. This is known as the phase problem [18] and constitutes a significant challenge in structure determination, especially in the initial stages. At the same time, the phases cannot be computed in the absence of the structure factor amplitudes. The importance of phases is shown in Figure 1.3.

There are indeed many computational and experimental techniques to recover the otherwise lost phase information. For crystals at very high resolution, the phases can be grad-

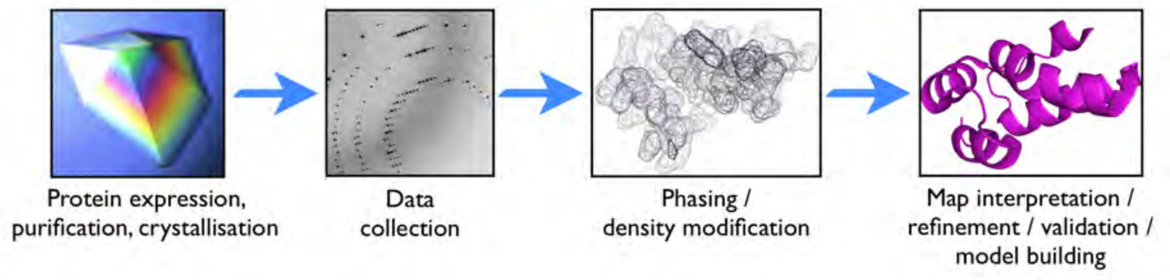

Protein expression, purification, crystallisation
Data collection
Phasing / density modification
Map interpretation / refinement / validation / model building

Figure 1.2: The main steps in protein crystal structure determination.

ually obtained *ab initio* from the measured amplitudes using so-called direct methods. Here the positions of some atoms can be derived directly from the collected magnitudes given the structure under consideration is smaller than 100 residues. The phases generated from these atomic positions are subsequently used to derive the phases for the remaining parts of the structure [19–22]. Isomorphous replacement, MAD and SAD all use the positions of a few atoms to derive the (initial) phase information of the entire macromolecule. Isomorphous replacement exploits additional data collected from the same structure but with one or a few electron-rich atoms added to the structure [23–25], whereas in MAD or SAD the signal of the anomalous scattering of atoms, such as sulphur and phosphorus, can be used to determine their positions [26–28]. The combination of isomorphous replacement and anomalous scattering, SIRAS, makes simultaneous use of the position of atoms derived from anomalous scattering and heavy atom derivatives [29, 30]. This approach is rarely used when the data are collected at only one wavelength, since it requires high quality diffraction data [31]. In the most frequently used approach, molecular replacement, the phase information is obtained by transforming a homologous molecule into the expected location and orientation [32–38]. After computing the initial phase information from that positioned model an electron density map can be computed. Density modification techniques can be applied to improve this map [39–42].

The final step, which comprises the transformation from an electron density map to a chemically sensible model of a protein structure is called map interpretation or model building. Here, the electron density has to be interpreted in terms of atoms and bonds based on the prior knowledge of the chemical nature of the molecule and, indeed, molecules in general and properties of the map. Map interpretation is thus, fundamentally, a pattern recognition problem, which becomes more difficult when less information can be deduced from the electron density map. Before 1970, crystallographers had to equip themselves with rulers, screwdrivers and rods and switch on their overhead projectors

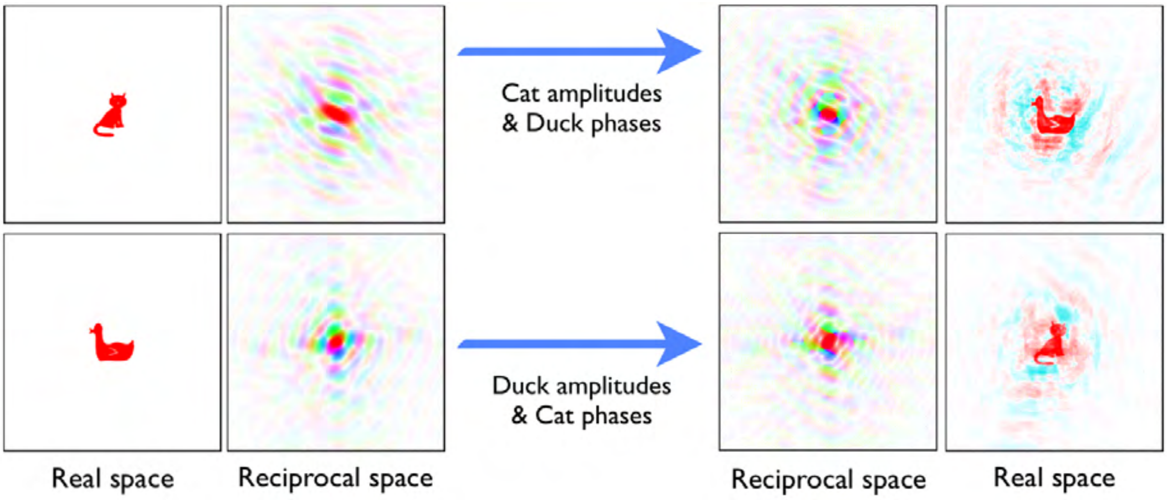


Figure 1.3: Importance of phases. Pictures on the left in real space were transformed to reciprocal space, the phases of each picture exchanged with each other and back transformed, yielding the real space pictures in the right column. From Cowtan [43].

to investigate the right slide or slice of density or use a Richards box, which projected a electron density map upon the model with semi-transparent mirrors [44], to build a so-called Kendrew model of their structure [45]. At a scale between 5 and 1cm/Å they sometimes even had to climb ladders. Advances in computer graphics allowed ladders to be replaced with deckchairs and the use of molecular graphics programs such as FRODO [46], O [47], Xtalview [48] and Coot [49, 50]. Although molecular graphics made manual interpretation of electron-density maps less tedious, it is still a very labour intensive and subjective process. The need to speed up macromolecular structure building and to provide at least some elements of objectivity into the model building process gave rise to automated model building procedures. Macromolecular structure refinement [51] is repeatedly applied during automated model building. Here, the intermediate model is adjusted with regard to the experimental data and the current set of phases is changed according to the intermediate model, respectively. Thus, in case of a correct intermediate model, the electron density is improved and will give rise to a better model in the next iteration. To further improve and assess the accuracy of the structure model, prior information regarding molecular structure in general - encoded in the form of restraints or constraints on atomic bond lengths, angles and general stereochemistry - is used to give a measure of its validity. Software suites like CCP4 [52] and PHENIX [53] provide collections of programs for many of the mentioned approaches, from indexing of diffraction patterns and phasing, through map interpretation and refinement, to model validation.

Today, there is a manifold of iterative methods that combine tracing of the protein chain with density refinement. The most important ones, from the author's point of view, will be described in detail in the next section. Since the methods presented in this thesis (the PNSextender [1] and FittOFF [3]) have been developed for their application within the *ARP/wARP* project [54, 55], special attention will be given to this software suite in section 1.5. Refinement and validation are described in more detail in sections 1.3.2 and 1.3.3.

1.3.1 Automated model building

A first step towards automated model building was taken in 1974 with the publication of the skeletonisation method by Jonathan Greer [56]. A skeleton representation helps to obtain a more interpretable "image" of a an electron density map, by its automatic reduction to a set of connected thin line segments that follow the density profile. This is achieved by placing points at density peaks and then deleting those with lower electron density values, unless this breaks the connectivity or affects the end of a connected region. The obtained skeletal representation of the map can be used to derive potential $C\alpha$-positions using the known $C\alpha - C\alpha$-distance.
More sophisticated pattern recognition approaches for identifying the positions of $C\alpha$ atoms of a protein backbone in a skeleton representation have been implemented in the programs QUANTA and CAPRA. CAPRA, the $C\alpha$ Pattern Recognition Algorithm [57, 58], uses a range of electron density-feature scores combined in a neural network and rotation invariant numerical features to predict the positions of $C\alpha$ atoms and connects them into chains by an heuristic search method. Coupling this method with modelling of side chains, sequence alignment and real space refinement gave rise to the TEXTAL method for automated building of proteins [59, 60] QUANTA [61] uses a principal component analysis on the skeleton representation in order to identify regions that correspond to regular secondary structure features, i.e. α-helices and β-sheets. Afterwards, identified segments are be used to define the positions of $C\alpha$ atoms in order to build a polypeptide chain [62].

Crystallographic template matching methods aim to recognise small search models with low structural variation within electron density maps. They can thus be called 'mini'-molecular replacement [63]. In the first of these methods, ESSENS [64], Kleywegt and Jones use penta-alanine templates in 'ideal' α-helix or β-sheet conformations to detect secondary structure elements in electron density maps. These search fragments were tried in all possible positions locations and orientations. The best fit was chosen by

evaluating the densities calculated at the atomic centres of the fragments. Detected secondary structure elements can then be used to improve the phases or to judge whether the map is interpretable at all. A shortcoming of ESSENS is its exhaustive search over six dimensions in real space (three translational and three rotational parameters) and hence, large demands on computation time. This problem is to a large extent alleviated in FFFEAR [65]. Instead of using the density at atomic centres, a target function compares the electron density map with density shapes computed from nine-residue long search fragments, carrying out the translation searches in reciprocal Fourier space and thus reducing the computation time. In BUCCANEER [66], this search function is repeatedly applied to locate possible $C\alpha$ atoms. In the subsequent applications putative $C\alpha$ atoms are subsequently refined before being extended into chains using an exhaustive search over torsion angles allowed in the Ramachandran plot [67]. Finally they are assigned probabilities for each amino acid type at each $C\alpha$-position. Recently, BUCCANEER was updated with a library of protein fragments to build chains from identified $C\alpha$ atoms [68] more efficiently, especially in terminal regions and loops.
RESOLVE [42, 69, 70], now called phenix.resolve and part of the PHENIX project [53, 71, 72], employs a search function similar to FFFEAR in order to locate map regions containing secondary-structural features. Identified helices and strands are extended with additional residues using a tripeptide-fragment library. In the last step probabilities of side chains are derived using 20 electron density templates. The most likely side chain is assigned in accordance with an alignment of the protein model to its sequence.
In the *ARP/wARP* protein model building, ‘free atoms’ (similar to the ones described in [73]) are used in an iterative approach together with real- and reciprocal-space refinement to build up the protein chain from (di-)peptides identified in the electron density. This is described in more detail in section 1.5.

Other novel approaches have been undertaken in ARCIMBOLDO and ACMI. In ACMI (Automatic Crystallographic Map Interpreter), residues are not constrained to a single location during the process of model building, but are instead represented as a probability distribution, the Markov field, over the whole electron density map. Physically possible, incomplete models from this distribution are extended step by step to construct an all-atom protein model using a statistical sampling method called particle filtering [74, 75]. ARCIMBOLDO (named after the artist who assembled portraits from fruit and vegetables) employs direct methods to generate phase information for structures of medium size (< 2000 atoms) at resolution higher than 2.0 Å [76]. As mentioned earlier, direct methods can usually only be applied to structures of sizes below 100 residues with data extending to atomic resolution. ARCIMBOLDO circumvents the missing atomicity at resolution between 1.0 Å and 2.0 Å by a multi-solution framework that combines the location of small model fragments ('ideal' polyalanine α-helices and β-strands of 10 to 14 residues) with density modification and autotracing of the resulting maps in SHELXE [77]. This results in several thousand structures based on numerous positioning of model fragments in space. To extend the applicability of the method to larger structures and lower resolution more sophisticated fragments with modeled side-chains or extracted from low-homology models can be added as model fragments [78]. Given the massive computation demand of the method, so far it only runs on a dedicated 100 CPU grid.

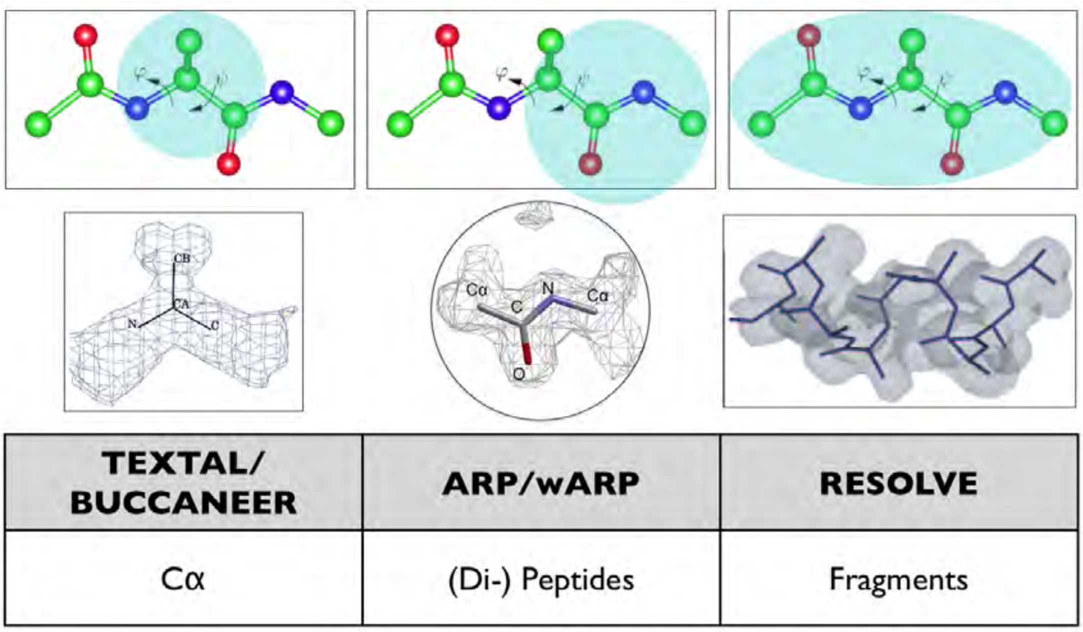


Figure 1.4: Different density templates in automated model building approaches. Shown are the density templates employed in BUCCANEER, TEXTAL (spheres around $C\alpha$ atoms), *ARP/wARP* (shapes of peptides and dipeptides) and RESOLVE (densities of standard protein structure fragments of different lengths).

In conclusion, it becomes apparent that most current automatic model building methods use similar techniques, such as density search functions and the combination of model building with structure refinement, mimicking the steps a crystallographer would take when building a model manually. What distinguishes them are the density search shapes or patterns used for identifying the positions of the main-chain atoms during chain tracing. As shown in Figure 1.4, these shapes range from 4 Å spheres to identify $C\alpha$ atoms in BUCCANEER and TEXTAL, peptide (or dipeptide) units in *ARP/wARP* or longer fragments in RESOLVE/phenix.resolve. Often, these templates lead to to models built to a different extend of completeness for the same electron density map; their strengths at different resolution will be discussed in section 1.8.1.

1.3.2 Structure refinement

Refinement of a macromolecular model aims at optimising the agreement between the structure factors calculated from the model parameters (F_{calc}) and the structure factors observed in the experimental data (F_{obs}). Model parameters include atomic coordinates, atomic displacement parameters (ADP), scale factors and, if appropriate, twin fractions [79]. A common problem, especially at low resolution, is that the number of parameters of the model exceed the number of experimental observations ($\frac{\text{observations}}{\text{parameter}} < 1$). In such cases, additional information is required; otherwise the refinement becomes underdetermined and the model overfitted. Such additional information can comprise *a priori* structural knowledge about bond lengths and angles [80], chirality and planarity of atomic groups, similar orientation or non-crystallographic symmetry between molecular fragments or substructures [81] or any experimental phase information. Refinement thus is the process of adjusting the model parameters so as to minimise the difference between calculated properties and experimental data. This makes it a complex optimisation problem. The agreement between the experimental data and the structural model is commonly measured by the R-factor value [82], Equation 1.2.

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \tag{1.2}$$

Historically, least-squares procedures in real and reciprocal space [83, 84] were the first methods applied to minimise the residual between the observed and calculated data. However, the need to account for the uncertainty in model parameters lead to the increasing popularity of maximum likelihood methods in refinement. For example, REFMAC [51] maximises the probability of observing the current model given the set of measurements and additional knowledge. REFMAC excels in refinement at a broad range of resolution, due to the use of different likelihood functions depending on the input diffraction data. The target function that is minimised, f_{total} has two components, f_{geom} utilising geometry or prior knowledge, and f_{X-ray} including the likelihood of observing the current model given the observed experimental X-ray data [79], equation 1.3 and 1.4. The optimum weight between the contributions, w, can be selected automatically on-the-fly.

$$\begin{aligned} f_{total} &= -log[P_{posterior}(model; obs)] \\ f_{geom} &= -log[P_{prior}(model)] \\ f_{xray} &= -log[P_{likelihood}(obs; model)] \end{aligned} \tag{1.3}$$

$$f_{total} = f_{geom} + w f_{xray} \tag{1.4}$$

To ensure reliable models at resolution as low as 4 Å, REFMAC employs a wide range of specific refinement tools, such as secondary structure restraints, restraints to known homologous structures, automatic global and local NCS restraints [85]. A very important feature in REFMAC, which is used in one of the methods presented in this thesis, is the possibility of adding known non-crystallographic symmetry relations as restraints to f_{geom}. Originally, refinement procedures have been designed for the final stages of MX analysis. Nowadays they are frequently used to improve partial models and to obtain better electron density maps for further rounds of model building. Examples are refinement with REFMAC in *ARP/wARP* [54] and phenix.refine in the PHENIX suite [81].

1.3.3 Model validation

After formally successful refinement the model might exhibit correct bond lengths and angles but still contain errors. These errors might hail from incorrect tracing of a chain, flexible loops, presence of peptide flips or incorrect side chain conformation and could have been reinforced by the refinement. To account for such errors, it is very important to evaluate the model with regard to *a priori* biochemical knowledge that has not been used in the refinement. One of the most widely-used validation methods is the Ramachandran plot [67]. It describes the occurrence of combinations of protein torsion angles ϕ and ψ, which define protein main-chain conformation. Residues with ψ- and ϕ-angles lying outside of highly populated areas in the Ramachandran plot are often incorrectly built or contain peptide-flip errors [86].
Unfortunately, the R-factor value itself cannot always be consulted to assess the validity of the model, since it is very similar to the function minimised during the refinement and thus is biased towards errors present in the model. The R_{free} factor was introduced to give a more reliable and unbiased global quality index [87]. R_{free} is computed from a small subset of structure factors, usually 5% of the data, that is not used during refinement and model building. Thus, only changes to the model that lead to a better

explanation of the experimental data will improve R_{free}. One should note that since R_{free} is computed from a relatively small number of reflections, its value is subject to higher statistical variation, compared to the plain R-factor. A number of papers have been devoted to the discussion on this topic, e.g. [88–90].

Another approach for validation is taken in the PDB_REDO project [91–93], which aims at improving structures in the PDB by applying re-refinement and some model rebuilding. Structures have been deposited in the PDB over the years and they have been determined using the methods available at the time. Many crystallographic methods have improved since then and can make a better use of the same X-ray data. PDB_REDO has been tested on more than 12,000 PDB entries and could improve the majority of these structures with regard to R_{free} and geometric validation criteria [94].

The protein data bank itself is also taking actions to improve validation measures during the structure deposition into the PDB. To achieve this, several validation task forces have been convened to advise on methods and standards, with the recommendations of the X-ray task force currently being implemented [95]. These recommendations include, among others, assessment of the X-ray data Wilson plot, amplitude mislabeling and missed symmetry to be used as validation criteria for diffraction data as well as analyses of the Ramachandran plot and rotamers, assessment of the covalent geometry for the validation of models. Additionally, the agreement between the model and the data will be evaluated globally by R and R_{free} and per-residue with the real-space R value (RSR, [47]).

1.4 Challenges in macromolecular structure determination

The most limiting factor in crystal structure determination is the resolution to which the crystal of a protein structure diffracts in the diffraction experiment. The current state of the art is such that many challenging structure determination projects cannot be brought to a satisfactory result (i.e. the determination of a structure). In particular, crystals of large proteins and their complexes may not diffract to a resolution where an atomic model can be straightforwardly constructed. This issue is confirmed by the average size of structures in the PDB solved at certain resolution. As the resolution decreases, the average structure size increases significantly (Figure 1.5). Indeed, even after semi-high-throughput sample screening, the crystals of a typical protein of interest diffract on average to about 4 Å resolution on synchrotron beamlines [96], and only a small fraction of the measured X-ray data results in a structure being deposited in the PDB. More precisely, the ratio of collected data sets and published structures is about 50 to 1 [97].

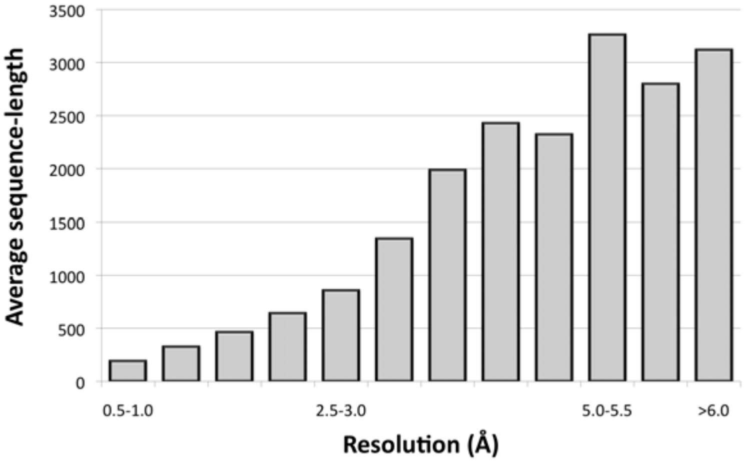


Figure 1.5: Average size of structures solved at different resolution (data derived from the PDB, January 2012).

An apparent problem with low-resolution X-ray diffraction is that the amount of the observed data that can be used for structure refinement and calculation of an electron density map is limited. For example, for a protein crystal with 50% solvent content that diffracts to a resolution of 2 Å, there are 9 reflections per atom. If four atomic parameters, e.g. $xyzB$, are to be refined, the observation-to-parameter ratio is two, and the task is numerically overdetermined. However, for the same structure at a resolution of 4 Å there is only one observation per atom, which is insufficient to refine several atomic parameters [98] (Table 1.1). This lack of observations at reduced resolution of the data requires the use of additional parameters in the form of constraints or restraints, and causes smoothing of density maps and a loss of detectable atomic features as shown in Figure 1.6. Dependent on the resolution of the measured data, the following problems may arise: at about 3 Å to 4 Å peptide groups cannot be seen anymore (Figure 1.6a, b). Between 5.0 Å and about 10 Å α-helices appear as tubes of density (Figure 1.6c, d, e) and at lower than 6 Å individual β-strands may not be visible at all (Figure 1.6d). The development of automated structure determination methods in MX has been predominantly focused on high-resolution data, where bonded or at least angle-bonded atoms are resolved. Thus, the determination of low-resolution structures is usually beyond the normal operational range of crystallographic software and necessitates a large, if not excessive, amount of manual intervention (an example being Rapper [99], where the user has to conduct an initial placement of $C\alpha$ atoms himself).

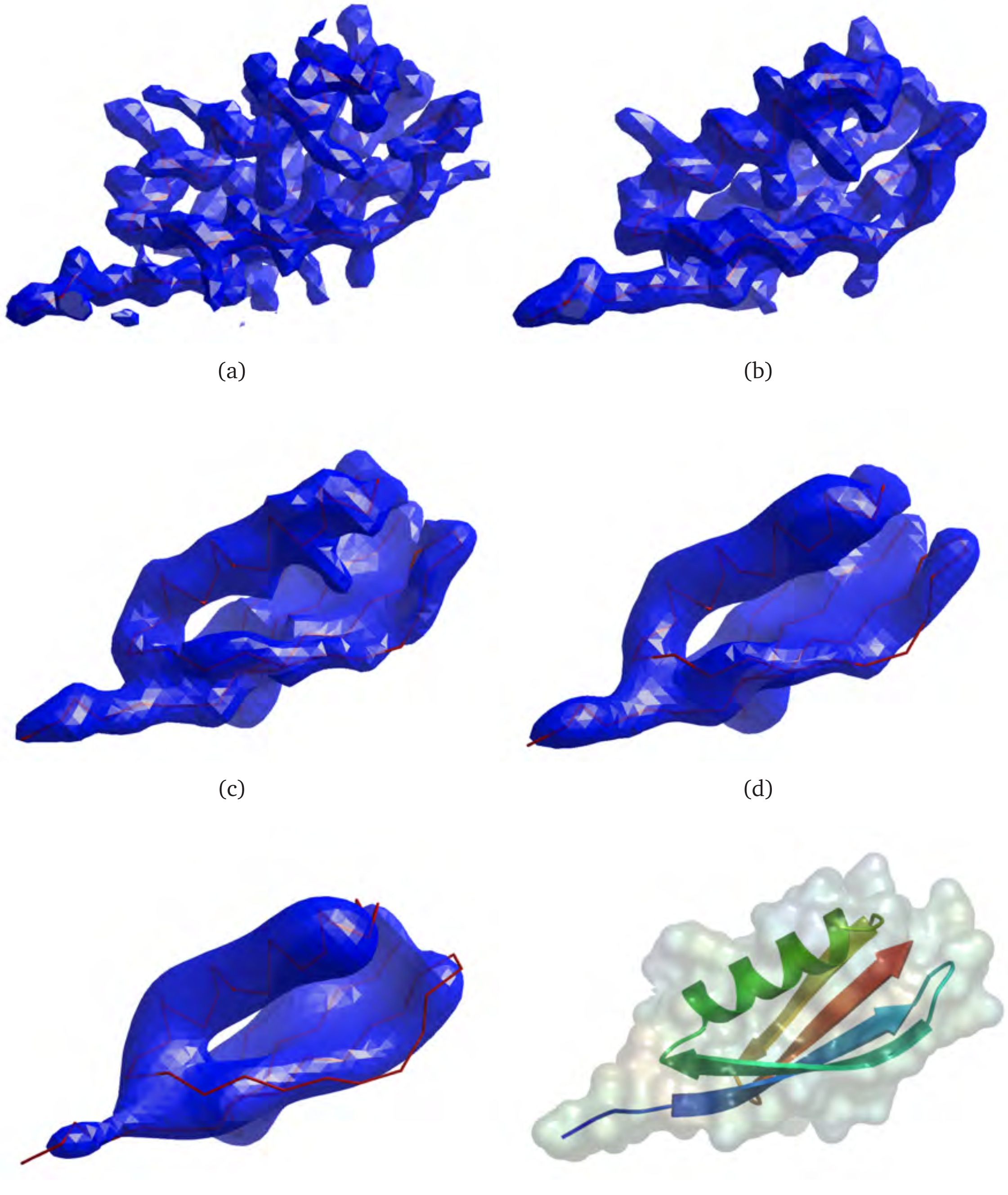
Resolution	Reflections per atom	Reflections per residue
2.0 Å	9	70
2.3 Å	6	46
2.6 Å	4	32
3.0 Å	3	21
3.5 Å	2	13
4.0 Å	1	9

Table 1.1: Overview of measured reflections to be expected per atom ($N_{refl/atom} \approx \frac{70}{d^3}$) and per residue ($N_{refl/residue} \approx \frac{500}{d^3}$) for resolution between 2.0 Å and 4.0 Å

However, despite the enormous effort that has to be undertaken to solve a structure at low-resolution, the yearly percentage of structures deposited being in the PDB with a resolution of lower than 3.5 Å is steadily increasing, see Figure 1.7. While the percentage of depositions in this low resolution range - 1 to 2.5 % - may appear to be low, the increase in terms of the raw number of depositions is much more evident. Thus, in 1992, 1% of structures solved below 3.5 Å corresponded to just two structures, whereas 2.3% in 2009 were equivalent to 170 structures. This shows that there is an increasing interest and need for structural information even at reduced levels of data resolution.

1.5 *ARP/wARP*

The *ARP/wARP* project [54, 55] is one of the leading software projects in macromolecular structure determination. The goal of the project is to facilitate automated building of the three-dimensional structure of proteins [101–108], nucleotides [109], ligands [110–112], as well as their complexes into electron density maps obtained from MX experiments using pattern recognition approaches. The foundation of *ARP/wARP* is the idea of coupling the interpretation of an electron density map with the iterative refinement of the atomic parameters [55, 113]. The initial model used for describing the electron density, calculated from the measured amplitudes and initial set of phase estimates, consists of a set of unconnected atoms of uniform atomic type. Reminiscent of the approach of Agarwal and Isaacs they are referred to as ‘free atoms’ [73]. In each iteration this set of ‘free atoms’ is chosen to reproduce the electron density as closely as possible while retaining an overall protein-like conformation. *ARP/wARP* then proceeds



(a) (b)

(c) (d)

(e) (f)

Figure 1.6: Electron density maps at different resolution: a) shows a map at 3 Å, b) at 4 Å, c) at 5 Å, d) at 6 Å and e) at 8 Å. The cases shown in a) to c) mark the resolution regime where density maps become difficult to automatically interpret. All maps have been computed from the structure of protein G (f, 2igd). Structure factors were calculated from the refined model and then truncated; B-factors were adjusted to their expected value at each resolution. Maps were generated using the ARPnavigator. From Langer [100].

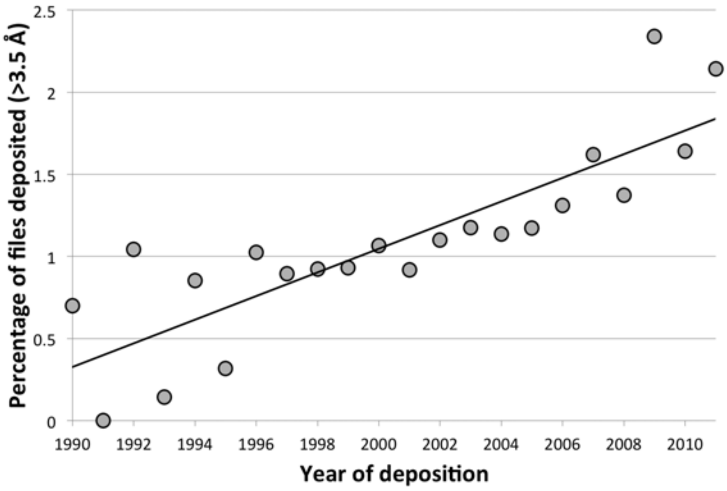


Figure 1.7: Percentage of structures being deposited in the PDB determined at a resolution of less than 3.5 Å. Values are given for the years 1990 to 2011 (derived from PDB data, January 2012).

to extend this model by evaluating the density at atomic centres. If the density falls under a given threshold, the atom under consideration is deleted. Likewise, if there are areas of high density within binding distance of a valid 'free atom', a new atom is added. The model is then used in refinement to improve the positions of the 'free atoms'. This leads to a phase improvement and thus to a better map. This model update procedure largely improves refinement: conventional refinement might use wrong atoms, whereas *ARP/wARP* will simply delete them and add them somewhere else later on.

The 'free atoms' model is further used to build the protein model using pattern recognition techniques. Ideal peptide search density shapes are mapped on 'free atoms' that are located within the expected $C\alpha - C\alpha$ distance from each other (3.8 ± 1.0 Å). At this point, the correct direction of the peptide is still unknown - the density shape is thus tried for in both directions. Subsequently, all peptides that share a common $C\alpha$ atom are mapped to dipeptide templates. The resulting dipeptides are then used to build up the longest possible polypeptide chain. This chain is saved and all other peptides that give rise to sterical clashes are removed. Iteratively, every next-longest chain is saved until no more chains longer than four peptides (five $C\alpha$-residues) can be found anymore [55, 104]. Partial side chains of four types - glycine, alanine, serine and valine are built if there is sufficient density support.

This gives rise to another fundamental concept of *ARP/wARP* - the 'hybrid model' [102]. In each building cycle some 'free atoms' gain chemical identity and are recognised as part of a protein chain fragment. Others remain free (Figure 1.8). The evolving hybrid model combines two sources of information: It incorporates chemical knowledge from the partially built model and the ‘free atoms’ continue to interpret the electron density in areas where no model is yet available. A restrained refinement of the chemically-assigned parts will improve the electron density map, which then allows the building of another hybrid model that should yield more chemically assigned parts and in turn leads to an even better electron density map. Thus, *ARP/wARP* combines model building and refinement, as depicted in Figure 1.9, in which the scheme of restraints and ‘free atoms’ are iteratively updated and the hybrid model converges to the final model. The recent addition to ARP/wARP was the incorporation of automatically-detected non-crystallographic symmetry into both the model building and refinement stages of the procedure [1]. These developments are an essential part of this thesis.

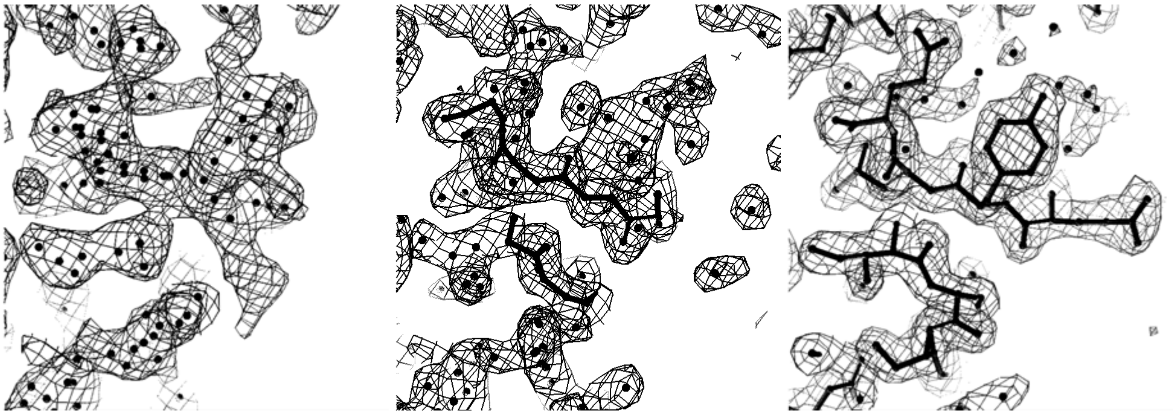

(a) (b) (c)

Figure 1.8: Evolution of the *ARP/wARP* hybrid model. At first, the density is filled with ‘free atoms’, placed to retain a protein-like interatomic distance distribution (a). During the model building process, some ‘free atoms’ are recognised as parts of a protein chain, others remain free (b). At the end of *ARP/wARP* model building a large part of the model is built. Some ‘free atoms’ remain, which can be attributed to solvent (c). Additionally, (a)-(c) show the improvement of the density as the hybrid model advances.

As the final step, the peptide backbone is decorated with side chains [102, 105–107]. To achieve this, the partially built side chains as well as other ‘free atoms’ present around every $C\alpha$ atom are described as a connectivity vector. For each polypeptide fragment, a matrix of such connectivity vectors is generated. This ’observed connectivity’ is then slid over a ’precomputed connectivity’ matrix describing the input amino acid sequence (similar to the approach described in [114]). Subsequently the polypeptide fragment is docked to the position in the sequence with the best agreement.

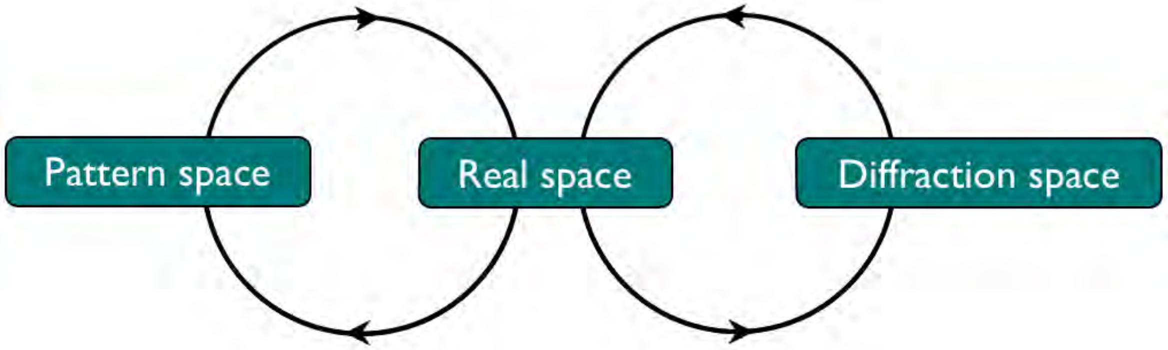


Figure 1.9: *ARP/wARP* circulates between pattern space, real space and diffraction space; forming a unified process of model building and refinement.

In addition to automatically building protein structures, *ARP/wARP* employs pattern recognition techniques for a number of other tasks. With the helix building method it is possible to efficiently identify secondary structure elements in electron density maps. The method delivers accurate results for α-helices for data extending to 4.5 Å resolution and for sheets down to 4.0 Å. The *Loopy* module allows the building of up to 14 residues-long loops between protein fragments that have been docked into sequence and this increases the completeness of the resulting models [115]. *Loopy* fits $C\alpha$ atoms of several template pentapeptides to the model termini and thereby extends the peptide segment iteratively. Subsequently backbone conformations are constructed. At the end the electron density correlation is used as a criterion to select the best loop. To build DNA or RNA into an electron density map with *ARP/wARP* a slightly different pattern recognition approach is used. Phosphates and base planes are identified in a map as balls and planar disks of defined volumes. Both shapes have been chosen so they can only be detected in nucleotide structures. This information is used to build up the nucleotide backbone [109].

Another application in *ARP/wARP* is the building of ligands [110, 111]. After model building of the protein is completed, a difference density map is built using the complete protein model. In the case of a known ligand, leftover density is analysed to identify the most likely ligand-binding site. Then, the selected density cluster is modelled as a 'sparse grid', a thinned set of gridpoints at approximately inter-atomic distances that correspond to the actual ligand to be built. The ligand is matched to this grid using graph-matching protocols that also take into account the automatically-generated ligand topology. In parallel, a Metropolis (Monte-Carlo) simulation generates ligand models in the same density. The ligand molecule is firstly aligned to the density using PCA, and then 'grown' into the density by rotations around the appropriate bonds and dihedral angles of the ligand. The best model is then selected from the ensemble generated by both methods and subjected to real space refinement to satisfy geometric restraints and the fit to the density. This provides the final model that is output to the user [112]. If the density cluster cannot account for the whole input ligand, a cocktail of substructures can be generated to fit the most likely partial ligand. In addition to this, a database containing the 50 most abundant ligands in the PDB can be used to identify the most likely one for a known binding site.

1.6 Non-crystallographic symmetry - NCS

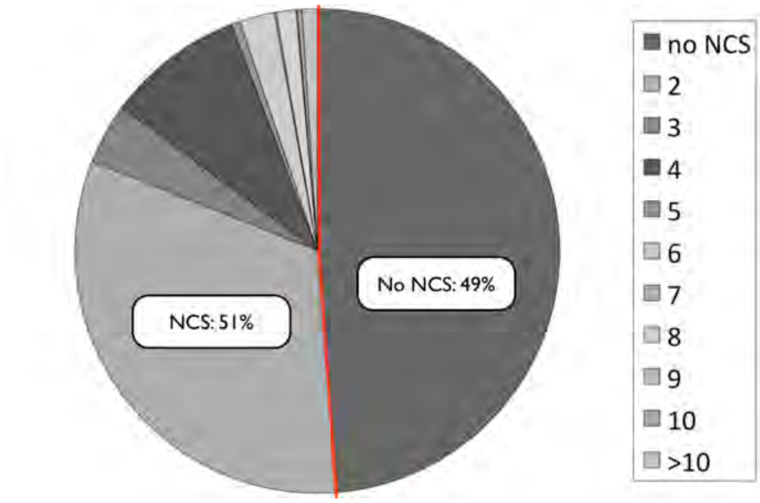


Figure 1.10: Proportion of databank structures with and without non-crystallographic symmetry (derived from the PDB data, January 2012).

A statistical survey carried out in 1993 showed that a vast amount of intrinsic information - the so-called non-crystallographic symmetry (NCS) existed in about one-third of all protein crystals [116]. Kleywegt later noted that about half of the proteins diffracting worse than 2.5 Å had NCS relations [117]. In the current release of the PDB more than 50% of structure contain NCS (Figure 1.10). NCS occurs if there are multiple subunits in the asymmetric unit of a crystal, and some of these adopt - at least in part - almost the same tertiary structure. The subunits related by NCS may have slight differences due to different crystal environments as identity is not enforced by crystallographic restraints. The NCS order may be as high as 60 (e.g. lumazine synthase from *Salmonella*

typhimurium LT2 [118], PDB ID 3mk3). NCS is more likely to appear in very large structures. This means that, while only 51% of all structures have NCS-relations, more than 70% of all residues in the PDB are involved in an NCS relation.
One distinguishes between two types of NCS [37]. An element, which is independent in the sense of rotation, is defined as 'proper'. An example would be a molecule exhibiting an N-fold axis, with each element rotated by $(360/N)^\circ$ to the next one. 'Improper' NCS is referred to in cases of arbitrary rotation or translation between two molecules in the same asymmetric unit. Any NCS operation that includes a translation must therefore be 'improper'.

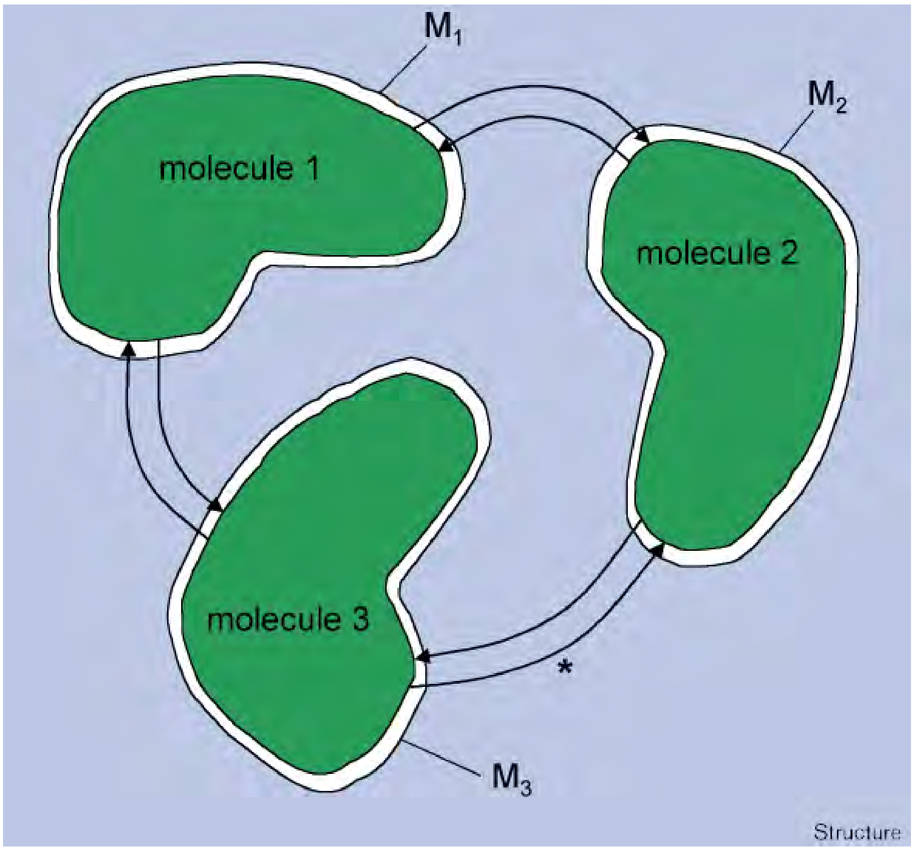


Figure 1.11: NCS averaging in case of improper NCS. Within the masks M_{1-3}, the electron density of the molecules (shown in green) is replaced by the average density calculated from all NCS-related subunits. From Kleywegt[119].

The use of NCS is an extremely valuable asset in crystallographic structure determination [35, 41, 63, 119]. Perhaps its most frequent application is in density modification, where NCS-averaging helps improve and extend phases to higher resolution as well as reduce bias in cases where initial maps have been derived from an incomplete model [119].

Here, the NCS relations can be specified by the user or derived from the determined heavy atom positions. The electron density map is segmented into areas related by NCS-operators and for each operator a mask or envelope function is generated. Within these masks, the initial electron density for each operator is replaced by, e.g. an average density over all NCS-related copies. A schematic overview of NCS averaging is shown in Figure 1.11. NCS-averaging is one of the most powerful constraints available for phase improvement. However, it is also the least automated one, since the symmetry operator has to be known and the masks must be defined manually by the user.

A more recent use of NCS-relations in MX includes their automatic detection and addition as restraints during structure refinement [53, 79]. Here, further restraints are applied to the chains or parts thereof that have been defined as NCS-related. This information is added to the geometrical prior probability function in order to treat related chains in the same way during refinement. For automatic NCS-detection in REFMAC, the sequences of all chains are aligned with all other chains. If the alignment is longer than 15 residues and has a sequence identity >80%, the chains are superposed. If the $rmsd$ of the superposition is less than a defined threshold (default 2.5 Å) the chains are deemed to be NCS-related. The approach does not work for chains that have not been sequence-assigned.

1.7 Theoretical modelling

As described in section 1.4, models of macromolecular structures cannot always be obtained by an experimental technique such as MX, NMR or EM. In these cases, theoretical models may provide further insights [120]. All approaches in theoretical modelling are based on the premise that protein structures with high sequence similarity have very similar structures [121, 122]. To obtain the tertiary or 3D-structure of a macromolecule, one may differentiate between homology modelling and *de novo* protein structure prediction. Both methods are described in more detail in the following sections. If no approach gives rise to a model of the three-dimensional structure of the macromolecule, one may take advantage of secondary structure prediction tools, available for protein and RNA, that suggest which segments of a primary sequence are likely to form helix, sheet or loop structures in 3D. An overview of protein secondary structure prediction is given in section 2.5.2.

1.7.1 Homology modelling

Homology modelling, also known as comparative modelling, is a widely-used method for the prediction of protein structure. It takes advantage of the ever-growing abundance of structural information in databases such as the PDB, and sequence information such as obtainable from UniProt [123, 124]. Known protein structures are used as templates to predict structures of target sequences, which are evolutionarily and/or functionally related [125]. The method itself is straightforward (see Figure 1.12). Initially, possible template sequences related to the desired target sequence have to be identified using large-scale sequence-alignment tools, like the ones supplied by the structural databases or an implementation of BLAST or PSI-BLAST [126]. In the next step, the target sequence has to be aligned to all template sequences to build a structural model based on identical or highly similar areas in both target and template sequences. Finally, this model is assessed by different criteria [127–131]. The accuracy of homology modelling entirely depends on the identification of the correct templates for the considered target sequence, as wrong templates will generate a wrong model. Since the structure of the target sequence will always be similar to the structure of the template sequence and cannot fully compensate for fold mutations, the introduction of bias might also be a problem. Nowadays, all tasks involved in homology modelling can conveniently be executed by a web-server such as SWISS-MODEL [132–135].

1.7.2 *de novo* protein structure prediction

If there are no sufficiently-related homologues for the target structure under consideration, the problem of building a model can be addressed by *de novo* protein structure prediction [136]. In contrast to homology modelling, the protein structure here is completely built from scratch using energy functions or statistical potentials based on the analysis of recurrent patterns in known structures and sequences. *De novo* structure prediction can be considered as more objective, since it uses only physiochemical properties, and thus reduces the risk of model bias as it may occur in homology modelling. One distinguishes between two approaches:

In *ab initio* modelling suitable models of the protein are solely derived from the sequence [137]. Furthermore, geometrical information, similar to that used in macromolecular refinement and validation (bond lengths and angles, agreement with the Ramachandran plot, etc.), is used to derive the target structures. Afterwards the models are analysed by an energy or score function to determine whether the obtained fold is a native-like conformation or, in other words, corresponds to an energy minimum. If not, the model

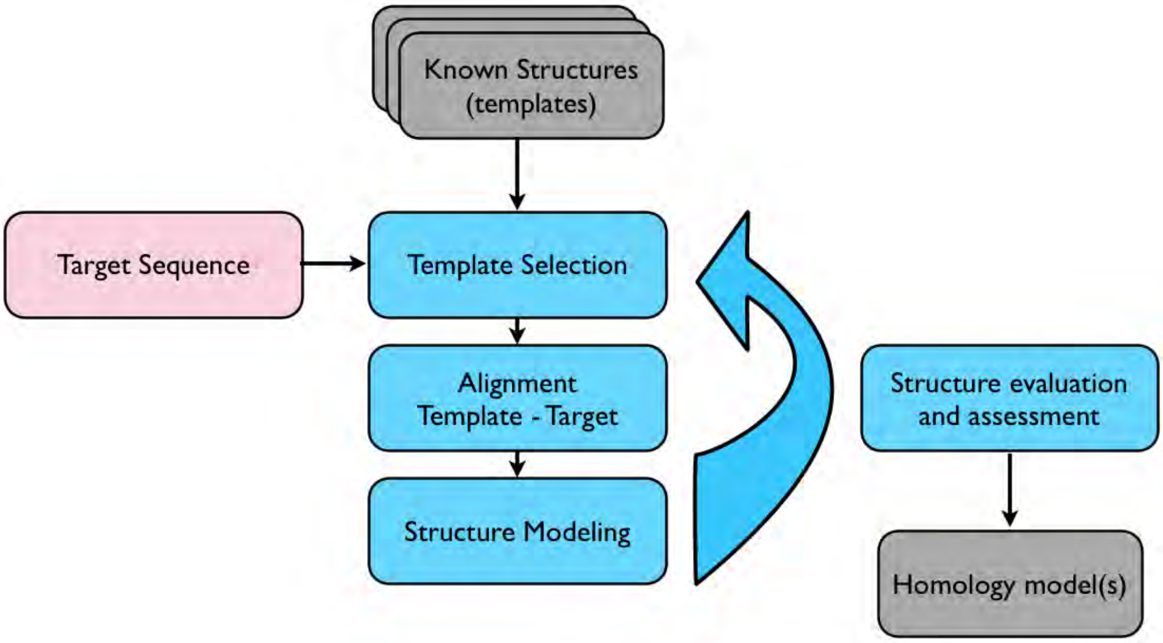


Figure 1.12: Homology modelling flow chart.

is modified to minimise the energy function. This usually involves the generation of thousands of models. Often, the protein is modelled in a reduced representation and successively extended. An example are lattice models that represent the protein as a sequence of hydrophobic and hydrophilic states and exploit the hydrophobic effect [138]. Once such models satisfy the energy function, they are extended to peptides and finally full proteins, while being iteratively modified to satisfy the energy function.
The second approach also uses energy functions to find the most native conformation of the protein. However, the way the models are generated is somewhat different. These methods are called "knowledge-based" and use properties derived from the ever-increasing amount of structures in structural databases [139]. Models are built using small fragments of a few residues in length. These represent the 'ideal' conformations for the considered sequence, derived from available structure by secondary structure prediction and multiple sequence alignments.

The pool of protein structure prediction methods is regularly evaluated by the Critical Assessment of Protein Structure Prediction (CASP, www.predictioncenter.org/). In each iteration of CASP, the sequences of a wide variety of solved but unpublished structures [140] are distributed to all participating groups, who are then asked to build the best possible model from this sequence using their method. This provides a means of objectively testing the methods via blind prediction [141, 142]. From CASP 7 onwards, all targets are divided into domains and then classified into two categories for assessment. The first one, template-based modelling (TBM), comprises all cases for which sequence-related structures exist, whereas the second one, template free modelling (FM), contains all cases without identifiable templates [143]. The free modelling category replaced the *ab initio* or "New Fold"-category. This change was due to the strong hybridisation of all prediction methods towards a combination of *ab initio* and knowledge-based methods, as well as most new folds being covered by templates. Results from the latest iteration of CASP (number 9, [144]) show that many methods are able to generate models for targets from the TBM category that are significantly better than a model built from the closest sequence-related structure. Targets of up to 200 residues from the TBM category can often be modeled with an $rmsd$ of less than 2.0 Å for backbone superposition to the reference structure. Unfortunately, no method has been developed that permits the accurate postulation of any model's validity to date [145]. However, there are several methods in development addressing this problem, such as QMEAN [146]. For the FM category, results are considerably worse, with a backbone accuracy of less than 2.0 Å only being achieved for stretches of structure of less than 50 residues [147].

One of the most widely-used method for protein structure prediction, which is regularly scoring within the highest ranks in CASP experiments, is ROSETTA [148, 149]. This knowledge-based approach treats the considered sequence as two sets of sequence segments, with a length of three and nine residues. For each of these segments, structural fragments from a library are selected based on sequence similarity and secondary structure prediction. The best protein conformation is derived by randomly inserting fragments into the protein chain applying a Monte Carlo simulated annealing search strategy and evaluating the resulting models with a database-derived scoring function that rewards nonlocal properties of protein structures (such as hydrophobic burial, compactness and pairing of β-sheets) [150]. Other methods scoring high in CASP experiments include HHPRED [151] or the I-TASSER pipeline [152].

The methods used in *de novo* protein structure prediction demand a much higher computation time compared to homology modelling. However, they can lead to results that are impossible to achieve with other methods. ROSETTA was tried in several experiments to improve structure determination in MX, and this is described in more detail in section 1.8.2.

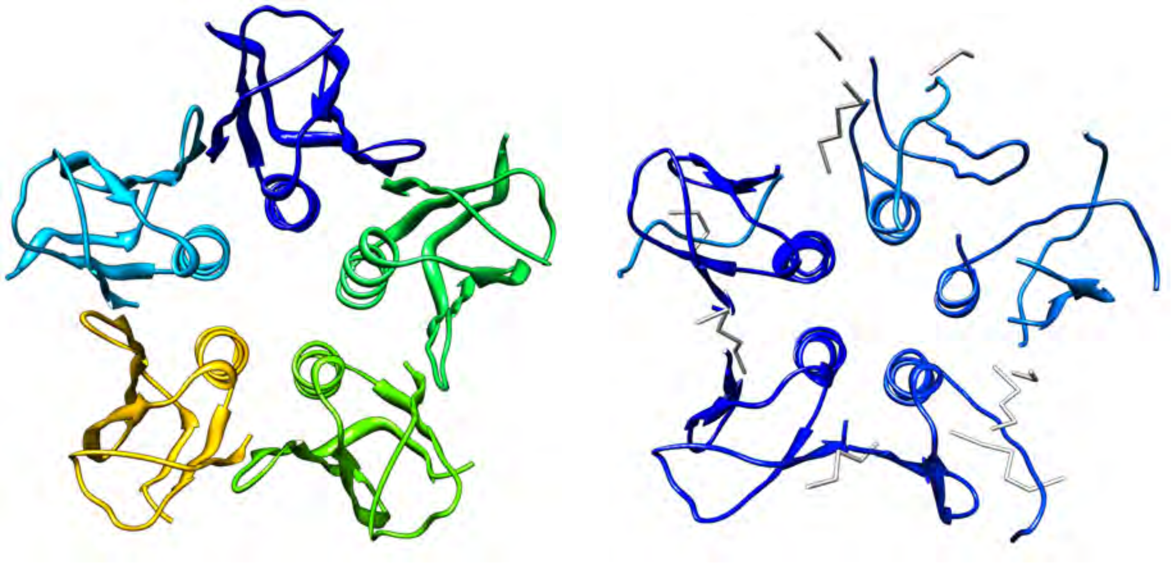
1.8 Where do we stand?

1.8.1 Automatic model building at medium-to-low resolution

As mentioned in section 1.4, one issue in structure determination at medium-to-low resolution is that the development of automatic model building procedures has been focussed primarily on solving structures with data extending to high resolution. However, recent developments in the MX field do address automation in this resolution regime. Often, impressive results are reported for low-resolution structure determination, although a complete structure can rarely be built without user intervention. As an example, usage of the PHENIX AutoBuild wizard [72] showed that structures with data extending to resolution around 2.8 Å could be built automatically to a completeness of 80% and more. At a resolution of 3.3 Å, the model completeness drops to 60%. A comparable performance is obtained with *ARP/wARP* [54], version 7.1 (Table 1.2). Estimates from the *ARP/wARP* remote model-building web service suggest that structures at a resolution around 2.6 Å are typically built to a completeness of 80%. At a resolution of 3.0 Å, the model completeness decreases to 75%. For cases with a resolution of 3.5 Å and lower one might still obtain a structure with 65% model completeness. The Buccaneer software can build up to 80% of the model at a resolution down to 3.2 Å provided an initial map correlation is higher than 0.6 [66].

The ability to automatically build 75% of the model of a structure at resolution of 3.0 Å and lower might lead to the conclusion that current methods work sufficiently well at low resolution. However, this number is rather deceptive as structures built in this resolution regime are often highly fragmented. An example is given in Figure 1.13, which shows a shiga-like toxin (PDB ID 1c48) built in 10 model building cycles with *ARP/wARP* at different resolution. In both Figure 1.13a and 1.13b, the reflections which have originally been deposited extending to 1.6 Å, were cut to 2.0 Å and 3.0 Å respectively, without introducing any phase error. As shown in Figure 1.13a, the structure at 2.0 Å data has been built completely with an average number of residues per chain of 70. The structure at 3.0 Å data has also been built with 80% model completeness. However, as can be seen in Figure 1.13b, it is highly fragmented with an average fragment length of only 14 residues while the amount of chains fragments quadrupled compared to the data at 2.0 Å. A more detailed overview of the levels of fragmentation that might be expected for typical automated model building runs at various resolution is given in Table 1.2.

This shows that automated interpretation of MX data in general and model building in particular, in a resolution range from 2.5 to 3.5 Å and lower, requires more research if



(a) (b)

Figure 1.13: Comparison of standard automated protein model building of test case shiga-like toxin (PDB ID 1c48) with *ARP/wARP* and X-ray data truncated at 2.0 Å (a) and 3.0 Å (b).

Resolution (in Å)	Estimated fraction of automatically built protein structure	Average length of built fragments
< 2.0 Å	over 90%	70
2.3 Å	84%	47
2.6 Å	80%	23
3.0 Å	74%	13
3.5 Å	65%	6

Table 1.2: Results from the *ARP/wARP* 7.1 web-service (tracing performance, obtained in May 2011)

it is to be generally successful. All approaches are limited by the quality of the initial phases. Reduction of model completeness at medium-to-low resolution implies an increase in the number of shorter, unconnected fragments built, which shows a need for novel approaches that will increase the completeness and the quality of derived macromolecular structural information in this resolution regime.

1.8.2 Theoretical modelling in macromolecular crystallography

The most widely-utilised approach from theoretical modelling to aid structure determination in MX is the application of homology modelling in molecular replacement. For models obtained from automatic protein model building various forms of loop predictions are employed to rebuild flexible regions. Also, there have been a few high impact experiments over the recent years that used *de novo* protein structure prediction to obtain structural models.

Search models for phasing by molecular replacement can often be detected and improved by homology modelling. An example of a procedure automating this is MODELLER [153], which is included in the CCP4 software project [52]. MODELLER aims at deriving the best molecular replacement solution for an input sequence and potential template structures. Another approach has been implemented in the CaspR web service [154]. Here, a combination of programs (including MODELLER) is used to generate high-quality homology models, again obtained from sequence and one potential template, that are each screened, giving rise to a number of MR solutions from which the best is chosen for subsequent steps. Additionally, websites like the Protein Structure Initiative's Structural Biology Knowledgebase [155] give comprehensive information about input sequences or PDB IDs such as related proteins, annotations and homology models to ease the search for applicable templates for molecular replacement.

Another application of theoretical modelling in MX has been the prediction of loops. Usually, secondary structure elements, such as helices and sheets can be built quite reliably at a broad range of resolution. However there are difficulties with building less ordered sections between secondary structure elements, which are commonly referred to as loops. The flexibility of such regions of the protein backbone leads to either very low or smeared density and thus, prevents its automatic interpretation. Many methods have been developed to address this problem, examples being *Loopy* in *ARP/wARP* [115], XPLEO [156], LAFIRE [157] or phenix.fit_loops in the PHENIX suite [53].

A major problem in using theoretical modelling in MX is the introduction of bias towards known structures. It could be attributed to the small amount of non-redundant structures in the structural databases that have been used for homology modelling or to the features of the prediction methods. However, there has been a rapid growth in structural databases and their non-redundant subsets, reducing the problem of bias. Recently, the total number of structures in the PDB has surpassed 80,000. One could estimate the amount of non redundancy by clustering proteins such that the members of a group are at least 50% sequence homologous with another. If one does this, the number of clusters has grown more than 12-fold since 2000 (Table 1.3). These circumstances gave rise to several high impact experiments which are described in the following.

Year	Clusters
2000	1813
2005	5873
2010	15743
2012	21758

Table 1.3: Number of sequence clusters in the PDB in which members of a cluster have at least 50% sequence identity with each other, data taken from the PDB in May 2012

In 2007, the feasibility of using *de novo*-calculated structures from ROSETTA for molecular replacement was demonstrated [158]. Blind predictions were generated for a target sequence that had no sequence homologues. A consensus core model of the five best blind predictions was then used for molecular replacement with Phaser [159]. Using this solution, all 112 residues could be automatically traced with *ARP/wARP* with a $C\alpha$ $rmsd$ of 0.13 Å to the reference crystal structure in the PDB.

The approach was subsequently applied to 15 further examples with sizes below 100 residues [160]. For each target structure up to 5×10^7 models were generated with ROSETTA in 10^5 CPU hours and for the best 200 of the models, as well as 200 randomly chosen ones, molecular replacement was executed with Phaser. Again, *ARP/wARP* was able to build most of the residues and assign the sequence for all cases. These results are technically impressive, but perhaps impractical due to the vast amount of computing time used.

A related approach showed that even on a desktop computer, ROSETTA can produce models which allow solutions to be found with molecular replacement [161]. In this work, only 3000 models were produced for each target. These models were pure polyalanine backbone models without any side chains. This required only 20 CPU hours per target. Of 16 test cases, 10 provided acceptable MR solutions (no more than 2.8 Å $rmsd$

to the crystal structure). For two cases, *ARP/wARP* could build and sequence-assign 95% of the structure. For three further cases, tracing could be conducted but a complete structure could not be built.

These results showed that *de novo*-built models can be used to phase diffraction data for many structures of up to 100 residues. This led to a method combining ROSETTA and PHENIX : phenix.mr_rosetta [162]. The method appears successful [163], but remains computationally very expensive (using the approach from [158, 160]).

1.9 Scope of this thesis

The current state of the art in macromolecular structure determination is such that algorithmic methods and databases from structural bioinformatics are rarely exploited in an integrative manner, despite their increasing scope. In turn, theoretical modelling does not make extensive use of all of the information available from MX experiments. There are also large differences in the computation time required to obtain a protein model in different approaches: On the ROSETTA server it takes about 400 CPU hours to build a structure *de novo* for a 150-residue query, whereas model building for structures at a wide range of sizes and resolution takes only a few hours or even minutes with *ARP/wARP* if phases are available. It has been shown that employing a coordinated use of structural bioinformatics and modern X-ray data interpretation software can lead to impressive results, although conceived methods are not yet fully applicable to everyday use (section 1.8.2). To fully take advantage of the technical possibilities of both experimental and theoretical methods, novel, sophisticated software solutions are required. The complementary use of knowledge-based approaches from protein structure prediction could aid structure completion in MX, while the already built fragments and available electron density maps can be used as starting points for, or as restraints in, database searches. This would considerably reduce the amount of required computations and allow more difficult cases to be successfully tackled.

The aim of this thesis is to develop computational methods to improve completeness and connectivity of models obtained from automated crystallographic protein model building using *ARP/wARP*. The focus is put on cases with data extending only to medium-to-low resolution data and thus meeting the challenges described in section 1.8.1. Another important goal of this work is to accomplish novel methodology without introducing a massive overhead on the computation time. Approaches were followed that exploit intrinsic information from intermediate models obtained from *ARP/wARP* and complementary information obtained from structural databases such as the PDB. Available techniques from both MX and computational biology have also been exploited.

Two distinct methods have been developed: The Protein NCS-based Structure (PNS) extender and the Fitting OF Fragments (FittOFF) method. The PNSextender uses intra-structural relationships to identify non-crystallographic symmetry between chain fragments in intermediate models resulting from *ARP/wARP* protein model building. Following an all-versus-all least squares superposition between all chain fragments, potential NCS matches are clustered according to their rotational relationships. Identified and validated NCS-relations are used to generate additional $C\alpha$-seeds used in the subsequent model building steps. Identified NCS-relations are also used within the refinement engine REFMAC as restraints (sections 1.3.2 and 1.6).

The FittOFF method, utilising the experience accumulated in the Lamzin and Schwede groups, identifies chain breaks between partially built fragments from *ARP/wARP* intermediate models and uses structural information obtained from the PDB to fill these structural gaps. As opposed to loop-building approaches commonly used in MX [115, 156, 157], the identification of structural gaps does not require the fragments to be sequence-assigned. Gap identification is achieved by docking the partially built protein chains to a secondary structure predicted from the input amino acid sequence. Structural gaps identified in this process are filtered using a knowledge-based approach that provides probability values for the number of residues enclosed in a gap given the distances between the anchoring residues. Further, an evaluation of uninterpreted density between the fragments to be connected was applied. For all gaps of a certain confidence, backbone conformations are sampled from a large fragment database and scored by spatially correlating them to the residual density. Similar to the PNSextender, the best fitting fragments are fed back to the *ARP/wARP* model building process as new seed points for further main-chain tracing. Both methods have been implemented in the *ARP/wARP* software suite (with the PNSextender being publicly available since version 7.2 [1]). The increase in time taken for *ARP/wARP* model building since the incorporation of both methods is negligible. At the same time, there is a substantial improvement in completeness and fragmentation for various testcases at resolution ranging from 1.9 to 3.8 Å.

Chapter

2

Methodological Background

In this chapter, the methodologies used for the development of PNSextender and FittOFF are described. We begin with the basis of the methods used for describing rotations and translations, which are essential for the calculation of object superposition in three-dimensional space, and proceed through data clustering to protein secondary structure prediction. The chapter ends with a short introduction to string-matching algorithms.

2.1 Rotations in $\mathbb{R}^3$

Generally, a rotation is "the turning of an object or coordinate system by an angle around a fixed point." [164]. In 3D-space a rotation is about an axis that runs through that fixed point. "Euler's rotation theorem states that an arbitrary rotation [in 3D-space] can be parameterised using three parameters. These parameters are commonly taken as the Euler angles. Rotations can be implemented using rotation matrices"[164].

2.1.1 Rotation matrices

In linear algebra, a rotation matrix is used to describe a rotation in Euclidean space. For simplification, let us consider an example in 2D-space ($\mathbb{R}^2$). The matrix shown in Equation 2.1 rotates a point (x) in the xy-Cartesian plane clockwise through an angle

θ about the origin of the coordinate system (resulting in the point moving to a location described by x', Eq. 2.2).

$$\mathbf{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{2.1}$$

$$\mathbf{x}' = \mathbf{R}_\theta \mathbf{x} \tag{2.2}$$

Vectors are rotated by the means of matrix multiplication:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix} \tag{2.3}$$

A rotation matrix is a special orthogonal matrix, meaning it has the following properties:

$$\mathbf{R}^T = \mathbf{R}^{-1} \tag{2.4}$$

$$\mathbf{R}^T \cdot \mathbf{R} = I \tag{2.5}$$

$$det(\mathbf{R}) = 1 \tag{2.6}$$

where $\mathbf{R}^T$ is the transpose of $\mathbf{R}$, $\mathbf{R}^{-1}$ is its inverse, I is the identity matrix and $det(\mathbf{R})$ is its determinant.

In $\mathbb{R}^3$, rotations around the x-, y-, and z-axes (for an angle α, β and γ) give the matrices:

$$\mathbf{R}_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \tag{2.7}$$

$$\mathbf{R}_y(\beta) = \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \tag{2.8}$$

$$\mathbf{R}_z(\gamma) = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.9}$$

Eigen properties of rotation matrices

Let us consider a rotation matrix **R**. A vector **x** is the eigenvector of **R** with the corresponding eigenvalue described as a scalar λ if

$$\mathbf{R}\mathbf{x} = \lambda\mathbf{x} \tag{2.10}$$

Let **R** be a 3 x 3 matrix, then the eigenvector **x** and eigenvalue λ satisfy

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{2.11}$$

The rotation axis and its corresponding angle can be derived from any rotation matrix. A vector **u** that parallel to the rotation axis of a rotation matrix **R** must satisfy

$$\mathbf{R}\mathbf{u} = \mathbf{u} \tag{2.12}$$

in other words, **u** is an eigenvector of **R** corresponding to the eigenvalue $\lambda = 1$. The rotation angle can then be calculated from $arccos(\mathbf{v} \cdot \mathbf{R}\mathbf{v})$, were **v** is a unit vector perpendicular to **u**.

At the same time one can calculate the matrix **R** of a rotation of a rotation angle θ around a rotation axis defined as a unit vector $\mathbf{u} = (u_x, u_y, u_z)$ as described in the following:

$$\mathbf{R} = \begin{bmatrix} \cos\theta + u_x^2(1-\cos\theta) & u_x u_y(1-\cos\theta) u_z \sin\theta & u_x u_z(1-\cos\theta) + u_y \sin\theta \\ u_y u_x(1-\cos\theta) + u_z \sin\theta & \cos\theta + u_y^2(1-\cos\theta) & u_y u_z(1-\cos\theta) + u_x \sin\theta \\ u_z u_x(1-\cos\theta) + u_y \sin\theta & u_z u_y(1-\cos\theta) + u_x \sin\theta & \cos\theta + u_z^2(1-\cos\theta) \end{bmatrix} \tag{2.13}$$

2.1.2 Euler angles

According to Euler's Rotation Theorem, any rotation can be achieved by composing three elemental rotations around a single axis. If the rotations are written in terms of elemental rotation matrices $\mathbf{R}_B$, $\mathbf{R}_C$ and $\mathbf{R}_D$, then a general rotation $\mathbf{R}_A$ can be written as:

$$\mathbf{R}_A = \mathbf{R}_B \mathbf{R}_C \mathbf{R}_D \tag{2.14}$$

The three angles giving the three rotation matrices are called Euler angles. There are several conventions for Euler angles, depending on the axes about which the rotations are carried out. In the “xyz” (pitch-roll-yaw) convention, the first rotation is performed around the x-axis, followed by rotations around the y-axis and subsequently around the z-axis. For

$$\mathbf{R}_D \equiv \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.15}$$

$$\mathbf{R}_C \equiv \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \tag{2.16}$$

$$\mathbf{R}_B \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{bmatrix} \tag{2.17}$$

where θ is pitch, ψ is roll and ϕ is yaw, R_A is given by:

$$\mathbf{R}_A \equiv \begin{bmatrix} \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ \sin\psi\sin\theta\cos\phi - \cos\psi\sin\phi & \sin\psi\sin\theta\sin\phi + \cos\psi\cos\phi & \cos\theta\sin\psi \\ \cos\psi\sin\theta\cos\phi + \sin\psi\sin\phi & \cos\psi\sin\theta\sin\phi - \sin\psi\cos\phi & \cos\theta\cos\psi \end{bmatrix} \tag{2.18}$$

2.1.3 Quaternions

Quaternions, also called Euler parameters, can be described as vectors in four dimensions q, extending from the origin onto the surface of a 3D sphere within a 4D space with unit radius (Eq. 2.19). Thus a quaternion has unit length and only three parameters; the fourth one can be computed from the other three.

$$q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1 \tag{2.19}$$

Three parameters of q form a vector in imaginary ijk space, whereas the fourth one is a real scalar.

$$q = q_0 + iq_1 + jq_2 + kq_3 \tag{2.20}$$

Quaternions can be used to represent axis angles (Eq. 2.21, for an angle θ, around the x, z and z-axes).

$$q = \cos\frac{\theta}{2} + i(x\sin\frac{\theta}{2}) + j(y\sin\frac{\theta}{2}) + k(z\sin\frac{\theta}{2}) \tag{2.21}$$

In addition, unit quaternions can be used as rotation operators as in q operating on the shape $\mathbf{X}$,

$$\mathbf{X}^{\mathrm{R}} = q\mathbf{X}q^{-1} \tag{2.22}$$

The most relevant feature of quaternions is their relation to Euler angles and therefore the possibility to substitute them in three-dimensional rotation matrices (Eq. 2.23 shows a rotation matrix from Eq. 2.18 using quaternions). The use of quaternions instead of Euler angles makes the handling of rotations much more convenient, since common vector algebra can be applied.

$$\mathbf{R} = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_0q_1) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \tag{2.23}$$

These and other properties [165] make quaternions a convenient tool for a description of three-dimensional rotations as well as animation in computer graphics, computer vision, robotics, etc. The cosine of the angular difference between two rotations can be determined by calculating the dot product of their two respective quaternions. This quality led to their application for the identification of NCS-relations (section 3.1.1), since the rotation difference, by implication, is the clustering criterion for matched fragments.

2.2 Translations in $\mathbb{R}^3$

In euclidean space, a translation is a geometric transformation that moves an object over a certain distance in a certain direction. Translations are denoted by a translation vector.

A translation takes a point

$$\begin{bmatrix} x \\ y \end{bmatrix} \tag{2.24}$$

to the point

$$\begin{bmatrix} x + a \\ y + b \end{bmatrix} \tag{2.25}$$

for fixed values a and b.

The application of rotations and translations is important for superposition of structures, as described in the next section.

2.3 Rigid body superposition

To identify relations between macromolecular structures (or fragments thereof), one needs to superimpose them. In the methods developed as part of this thesis, this is especially important in the process of finding NCS-relations, as described in 3.1.1. For objects between which a point-by-point correspondence is not known, the problem of finding the optimal superposition is NP-Hard [166]. To compare one molecule (template) to another (target), the template has to be rotated and translated until the best solution is found (Eq. 2.26, Figure 2.1). If, however, the correspondence is known, the superposition problem reduces to an eigen decomposition. This is, in essence, is a least-squares procedure to find the optimal rotation matrix R_0 and the optimal translation vector t_0 by minimizing the sum of the squared distances over a set of selected atoms ($X = \{x_i\}, Y = \{y_i\}, i = 1..n$), as shown in Equation 2.27.

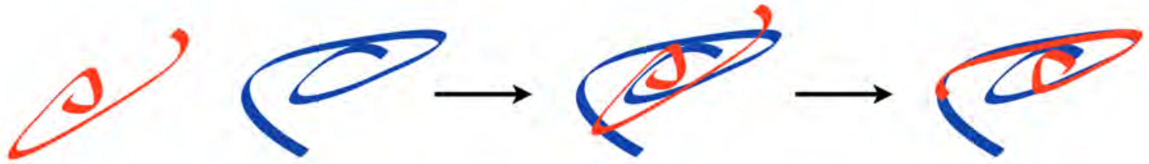


Figure 2.1: For the best superposition, one has to rotate and translate the template (red) onto the target (blue) until the optimal solution is found.

$$\mathbf{x}' = \mathbf{R}\mathbf{x} + \mathbf{t} \tag{2.26}$$

$$min_{\mathbf{R},\mathbf{x}} \sum_i |\mathbf{R}x_i + \mathbf{t} - y_i|^2 = \sum_i |\mathbf{R}_0 x_i + \mathbf{t}_0 - y_i|^2 \tag{2.27}$$

For computing superpositions in the methods presented in this thesis, i.e. the detection of NCS-relations (section 3.1.1), the approach of Kearsley is followed closely [167]. This method solves the superposition problem directly by an eigenvalue decomposition of a 4 x 4 matrix **M** constructed using quaternion algebra. The optimal rotation can then be found by calculation of the normalised eigenvector belonging to the smallest non-negative eigenvalue of **M**. This eigenvalue denotes the $rmsd$ ($rmsd = \sqrt{(\lambda_1/n)}$) and the eigenvector represents a quaternion describing the corresponding rotation.

2.3.1 Root mean square deviation

The $rmsd$ is the root mean square deviation between corresponding atoms in superimposed structures and is calculated as in the following:

$$rmsd(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{\sum_{i=1}^{N}(|x_i - y_i|)^2}{N}} \tag{2.28}$$

where x_i are the residues in structure **X**, y_i are those in structure **Y** and N is the number of aligned residues. Thus, an $rmsd$ value of 0 Å means that the structures are identical. Generally, $rmsd$ values of 2 Å between $C\alpha$ atoms in proteins have been adopted as the limit for a 'good' superposition (also a standard criterion in CASP, as described section 1.7.2). It becomes more obvious what a reasonable $rmsd$ value is when we consider some of the distances common in protein structure. Distances of about 1.5 Å denote the average bond length between two sp3 C atoms, whereas the distance between two adjacent $C\alpha$ atoms is 3.8 Å. Hence, a superposition with an $rmsd$ of less than 0.75 Å can be considered as very good, since every sp3 C atom in the template is matched to the related one in the target. For $rmsd$ values higher than 2 Å, the $C\alpha$ atoms in template and target are not aligned anymore, thus such a superposition can be considered as poor.

The $rmsd$ is a classic structural similarity measure since it directly reflects the quality of the structural superposition. $rmsd$ scales with the cube root of the number of aligned residues (N_{align}), therefore one can use $rmsd_{adj} = \frac{rmsd}{(N_{align})^{1/3}}$ to better compare alignments with varying N_{align} [168].

2.4 Clustering

"Data clustering (or clustering), also called cluster analysis, segmentation analysis, taxonomy analysis, or unsupervised classification, is a method of creating groups of objects, or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct" [169]. Objects in a cluster generally share the same or closely related properties or show only small mutual dissimilarities. Clustering techniques are subdivided into hierarchical and partitional algorithms. Partitional algorithms cluster the objects into a single partition of clusters, while hierarchical algorithms divide the objects into a sequence of nested clusters. The approach for clustering matches between potentially NCS-related fragments (algorithm 3.2) combines features from both techniques, such as the bottom-up approach from single linkage clustering [170], a hierarchical clustering method, which starts with each object in its own cluster and iteratively merges them to bigger ones. Also, iteratively updated cluster means are employed, reminiscent of k-means clustering [171], which is a partitional clustering method. The standard algorithms for single linkage and k-means clustering are given in algorithm 2.1 and 2.2, respectively.

Algorithm 2.1: Single-linkage clustering algorithm

```
Input  : Distance matrix $D$ containing all distances between objects ($d_{ij}$)
Output : cluster sequence $C$
Every object denotes one cluster
while Number of clusters > 1 do
    Find most similar pair of clusters $C_x$ and $C_y$, with $d_{x,y} = min(d_{i,j})$
    Merge clusters $C_x$ and $C_y$ into $C_z$
    Update $D$ by deleting rows and columns corresponding to $C_x$ and $C_y$
    add rows and columns for $C_z$, with distance to all other $C_a$ being $min(d_{a,x}, d_{a,y})$
end
```

Algorithm 2.2: Conventional k-means algorithm

```
Input  : Data set $D$, Number of Clusters $k$
Output : List of $k$ clusters $C$
initial $k$ clusters = $(C_1, C_2, ..., C_k)$
while cluster means change do
    $d_{i,j}$ = distance between $D_i$ and mean of $C_j$
    $n_i = \arg\min_{1 \leq j \leq k} d_{ij}$
    add $D_i$ to cluster $n_i$
    Update cluster means for all changed clusters
end
```

2.5 Protein secondary structure: assignment and prediction

The secondary structure of a protein describes the interactions between backbone atoms, which are stabilised by hydrogen bonds either within the peptide or between neighbouring chains. It results in repetitive local structures such as α-helices [172] and β-strands [173]. In an α-helical conformation, the peptide chain is wound like a screw, with each turn of the screw covering approximately 3.6 residues. In β-sheets, built up of β-strands, the peptide planes are arranged like a regularly folded sheet of paper [15]. Kabsch and Sander have defined seven different types of secondary structure: the 3-turn helix (commonly denoted by G), the standard 4-turn helix (H), the 5-turn helix (I), a hydrogen bonded turn (T), an extended strand in β-sheet conformation (E), residues in isolated β-bridges and a bend (S) [174]. Other definitions have been used over the years but are not further discussed here. In the FittOFF method, secondary structure information is used in two ways. Firstly, secondary structure states are assigned to each residue of all partially built protein chain fragments. These assignments are used to "dock" the fragments into a secondary structure prediction obtained from the amino acid sequence of the considered structure (section 3.2.1) or to validate the results in PNSextender (section 3.1.2). The methods employed for these approaches are described in the following sections. For simplification, the seven states (H, I, G, E, B, S and T) are reduced to three (H, E and C) according the scheme outlined in [175]; i.e. H and G are denoted as H (helices), B and E are denoted as E (sheets) and all others as C (coils).

2.5.1 Assignment of secondary structure to protein chains

Once the tertiary structure of a protein has been resolved the secondary structure state of each residue can be straightforwardly derived from the atomic positions of its backbone. Historically this was done by the crystallographer, which often led to highly subjective or incomplete assignments. The need to improve the objectivity and obtain a physically meaningful definition of secondary structure assignments gave rise to several computational approaches. The standard method, DSSP (Define Secondary Structure of Proteins), calculates the H-bond energy between all donor/acceptor pairs. A set of predefined geometrical features is used to determine the most likely secondary structure state of each residue based on the two best H-bonds for each atom [174].
Another approach has been taken by Zhang & Skolnick [176], where for a given residue a secondary structure state is assigned based on the $C\alpha$ coordinates of five neighbouring residues. The ith residue is assigned as α-helix or β-sheet when

$$|d_{j,j+k} - \lambda_k^{\alpha(\beta)}| < \delta^{\alpha(\beta)}, (j = i-2, i-1, i; k = 2, 3, 4) \tag{2.29}$$

is satisfied for all $d_{j,j+k}$ ($C\alpha$ distance between residues j and $j+k$), otherwise it is assigned as coil. The three λ_k^{α} and λ_k^{β}, as well as δ^{α} and δ^{β} are optimised parameters that denote the secondary structure. If not all $d_{j,j+k}$ satisfy Eq. 2.29, probability values can be derived for the given residue to be in each secondary structure state. Finally, the assignment is smoothed by merging and removing singlet secondary structure states, such as a single α-helix residue between five β-sheet residues on each side. The method was tested by Zhang and Skolnick on a set of 1489 non-homologous structures and obtained 85% agreement with the assignments from DSSP. This is perfect agreement with the average differences between secondary structure assignment methods [177].

For the secondary structure assignment needed in the developed methods, we use method of Zhang & Skolnick (referred to as the Zhang algorithm in the following). It only requires the coordinates of $C\alpha$ atoms, which is more convenient. The nature of the method imposes a clear restriction - it cannot be applied to fragments shorter than seven residues.

2.5.2 Protein secondary structure prediction

These methodologies attempt to predict the secondary structure of proteins (or nucleic acids) solely from the knowledge of their sequence, which is also called primary structure. In the protein case, regions in the amino acid sequence that are likely to form alpha-helices, beta-strands or loops (coils) are subject of the prediction [178–180]. Early methods, such as the Chow-Fasmann algorithm [181], used statistical probabilities (derived from less than 30 X-ray structures) of each amino acid to take part in an alpha-helix, beta-strand or loop. This resulted in three probability values for each amino acid, which were combined to identify helix and strand regions and then extended in both directions. Subsequent methods used scoring matrices, which took the influences of up to eight neighbors on each side of the to be investigated residue into account [182], which resulted in three 17×20 scoring matrices. Using these matrices, sequences could be evaluated for their propensity towards secondary structure elements. For the development and testing the performance of the FittOFF method, the two most reliable and popular methods were used and compared (section 3.2.1 and 4.4). Both methods have repeatedly been ranked highly in the EVA project [183, 184], which continuously evaluates the discriminatory power of secondary structure prediction methods. The first method is the PSIPRED web server [185–187] and the other the SSpro web server [188, 189]. For the generation of secondary structure predictions applied to FittOFF, described in section 3.3, we used PSIPRED version 3.0 and SSpro version 4.1.

PSIPRED

PSIPRED begins by generating a PSI-BLAST sequence profile [126] for the protein of interest. Each position is no longer a single amino acid, but has the probability of each amino acid type, considering a set of aligned sequences. This profile is then used as the input to a neural network [186]. The accuracy of the prediction is improved by generating a consensus from four independently trained sets of neural networks. Using the web-server version, the average accuracy with respect to the DSSP assignment [174] is 80%. A trimmed version of PSIPRED (*single_sequence*) can also be downloaded. However, this version uses only the input sequence instead of a sequence profile and hence can, on average, only achieve 70% accuracy compared to the DSSP assignment.

SSpro

SSpro uses ensembles of bidirectional recurrent neural network architectures and sequence profiles [126] to predict the secondary structure of an input amino acid sequence [190]. Although it is quite similar in its approach to PSIPRED, SSpro employs a different architecture and training set for its neural networks. Secondary structure predictions from SSPro have been compared to the DSSP assignment, with a sustained performance of 78% correct prediction [189].

2.6 String matching algorithms

To dock a secondary structure assigned chain fragment to a secondary structure prediction, its most-likely position in the prediction sequence has to be identified (section 3.2.1). In computer science, string matching (string searching) algorithms, are used to find one string (a pattern) that is embedded in another (a template or search string) [191]. In computational biology these algorithms are primarily used for sequence alignments [192]. In a naïve approach (shown in Algorithm 2.3) a pattern P of length m is slid over a template T of length n (in a sliding window approach); all elements of the pattern string are compared one by one to the corresponding elements of the template. If these elements are non-similar (a so-called mismatch is found), the pattern is slid one position further. If the pattern matches, the current starting position (offset) in the template is saved. The algorithm proceeds until the whole template has been evaluated for the occurrence of the pattern. Assuming that the pattern is only embedded once in the template, the algorithm takes $n+m$ steps in the average case. This is due to mismatches usually being identified at the first or second position of the pattern (for example, P = "baab" and T = "aaaabaab"). In the worst case, the algorithm takes $n \cdot m$ steps (for example, P = "ab" and T = "aaaaaaaaaaab"). There are many other algorithms tackling this problem with a better running time, such as the Knuth-Morris-Pratt algorithm [193], etc. However, for identifying the position of a chain fragment in a secondary structure prediction, only the number of agreements between the pattern and all positions in the template are required. Thus, a slightly amended version of the naïve approach was found to be most applicable.

Algorithm 2.3: Naïve string search algorithm

```
Input  : template $T = T_1..T_n$, pattern $P = P_1..P_m$
Output : List of offsets $offset_{list}$, at which $P$ appears in $T$
for $i = 0$ to $n - m$ do
    if $P_1 = T_{q+1}$ and $P_2 = T_{q+2}$ and ... and $P_m = t_{q+m}$ then
        Add $i$ to $offset_{list}$
    end
end
return $offset_{list}$
```

Chapter

3

Materials and Methods Developed

This chapter gives an overview of the methods developed as part of the work described in this thesis - the PNSextender method for NCS-based structure extension as well as automatic implementation of NCS-based restraints for refinement (PNSextend and PNS-restrain) and the FittOFF method for identifying structural gaps and fitting structural fragments into them. The chapter finishes with a description of the data and designed environments for testing both methods as well as an overview of their implementation and complexity.

3.1 PNSextender - Automatic NCS identification for extension and restraints

The PNSextender is a novel method for the automatic detection of NCS during automated model building with X-ray data at medium-to-low resolution. To circumvent the computationally intensive examination of electron density, it is solely based on the comparison of partially built protein chains (referred to as chain fragments in the following) from an intermediate model constructed during *ARP/wARP* automated model building. The derived NCS-relations are used to improve the model so that the built chain fragments can be extended and become more accurate. In essence, the approach builds up on the observation that during automated model building NCS-related parts of the structure are seldom built in exactly the same manner. This is highlighted in Figure 1.13b and such a situation is especially the case in the early stages of model building. Causes for this can be manifold and include differences in local solvent accessibility or the quality of phases and the electron-density throughout the unit cell. During the process of model building each NCS-related copy of the structure thus holds information that may be lacking in another copy. Combining the information from several copies helps to increase the overall structural completeness. Additionally, the identified NCS-relations are used as restraints on the model parameters during structure refinement with REFMAC5. An overview of the method that uses NCS-relations for model extension (PNSextend) can be found in algorithm 3.1. The steps described in the pseudocode are explained in the following sections.

Algorithm 3.1: Overview of the PNSextender method for NCS extension (PNSextend)

```
Input  : intermediate coordinate file $coord_{gaps}$, $rmsd$ threshold $rmsd_{thresh}$, initital
         search length $frag_{len}$, number of extensions to use $top_{percent}$
Output : extended coordinate file $coord_{extended}$
search for matches of $frag_{len}$ between chain fragments in $coord_{gaps}$ with $rmsd < rmsd_{thresh}$
cluster matches after rotational difference
extend matches
transform additional information from each extended match to related NCS-copy
rank extensions according to accuracy criteria
add $top_{percent}$ extensions into improved $coord_{extended}$
remove atoms clashing sterically with already-built model
```

3.1.1 Clustering of transformations between chain fragments and identification of NCS-related copies

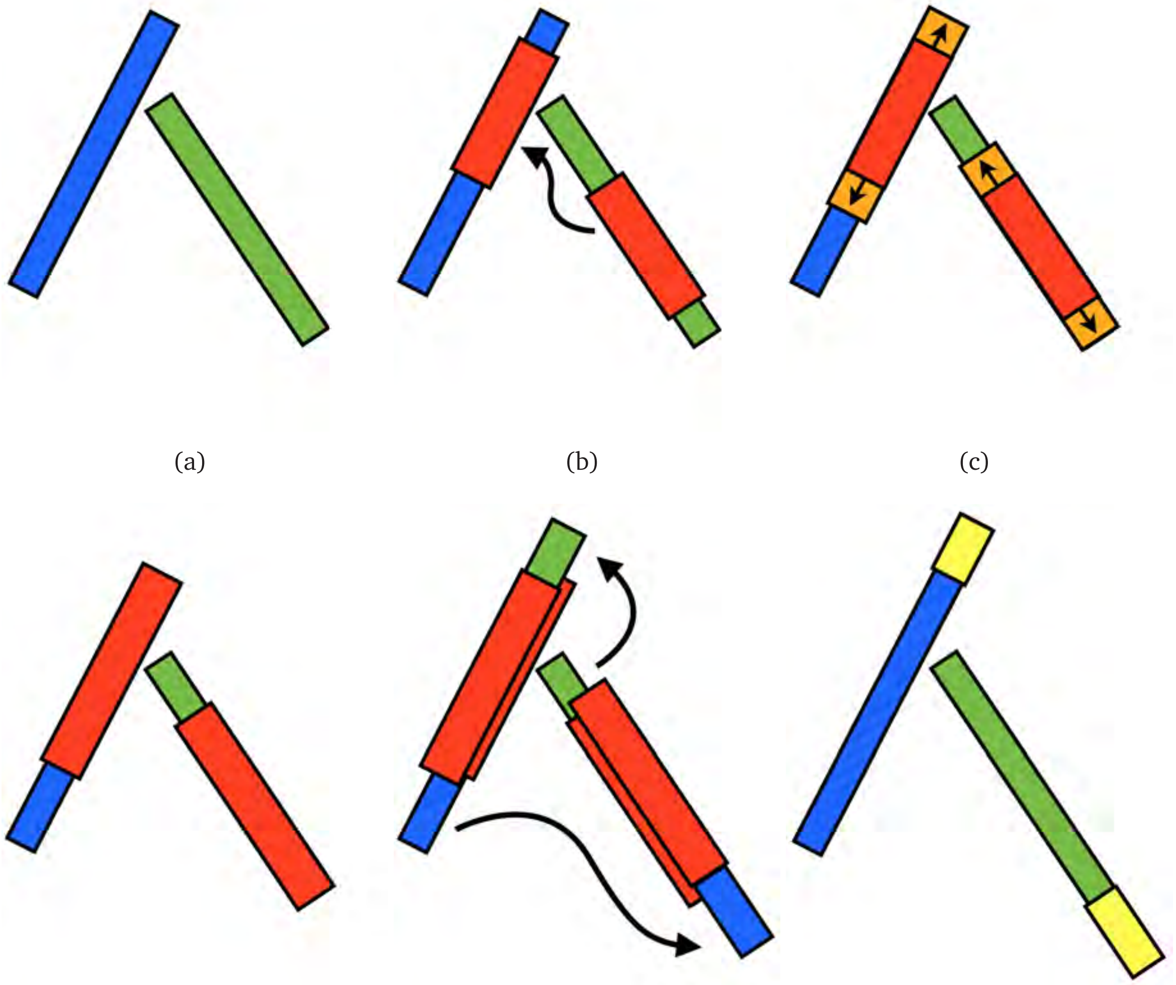
The first step of the PNSextender method, shown in figures 3.1a and 3.1b, involves an analysis of the chain fragments for their possible symmetry-related dependencies. Each stretch of a fixed number of $C\alpha$ atoms of each chain fragment ($frag_{len}$, between 5 and 15 residues) is least-squares superposed to each stretch of the same length of every other chain fragment. The approach is described in pseudocode in algorithm 3.2.

Pairs of stretches, which have been superimposed with an $rmsd$ between $C\alpha$ atoms below a fixed threshold ($rmsd_{thresh}$, 0.4 Å for resolution better than 2.8 Å, otherwise 0.5) are selected for further analysis and sorted in ascending order according to the $rmsd$ of

Algorithm 3.2: Identification of initial matches with NCS-relations

```
Input  : intermediate coordinate file $coord_{gaps}$, $rmsd$ threshold $rmsd_{thresh}$, initital
         search length $frag_{len}$
Output : set of initial matches $matches_{initial}$
identify all chain fragments $fragment$ in $coord_{gaps}$
for $i = 0$ to $length(fragment) - 1$ do
    for $j = i + 1$ to $length(fragment)$ do
        for $k = 0$ to $length(fragment_i) - frag_{len}$ do
            for $l = 0$ to $length(fragment_j) - frag_{len}$ do
                $rmsd_{k,l}$ = superposition of $fragment_i[k...k + frag_{len}]$ with
                $fragment_j[l...l + frag_{len}]$
                if $rmsd_{k,l} < rmsd_{thresh}$ then
                    add $match_{k,l}$ and superposition information to $matches_{initial}$
                end
            end
        end
    end
end
sort $matches_{initial}$ ranked in ascending order by $rmsd$
```

their superposition. If two matches are related by the same NCS-operator their superpositions must have a similar rotation angle, thus all matches are clustered according to their rotational components. As described in section 2.3, quaternions are used for the rapid superposition of chain fragments following the formulation of [167]. The rotation angle between two matched chain fragments can be determined from the diagonal elements of the rotation matrix. The cosine of the difference between two angles corresponding to two pairs of matched fragments can be conveniently calculated as the dot product of their respective quaternions. If such difference is below 5°, the two rotations relating respective pairs of superposed $C\alpha$ atom stretches are deemed to belong to the same NCS operator and are assigned to the same cluster. The applied clustering technique uses features from the methods described in section 2.4 such as a hierarchical selection and iteratively updated cluster centres and is described in algorithm 3.3. A rotation difference of five degrees was chosen to allow some variation in the derived NCS operators. This parameter is dependent on the accuracy of the built chain fragments and may vary as a function of resolution. Highly populated clusters of rotations point to a correspondence between NCS-related copies. Since only pair-wise NCS relations are considered, the method is able to detect both proper and improper symmetries.



(a) (b) (c)

(d) (e) (f)

Figure 3.1: Workflow of the PNSextender. Intermediate partial models are examined for symmetric dependencies between stretches of two chain fragments (a); an initial match is found between green and blue regions (b, red blocks); the initial match is extended in both directions of the chain fragments (c, orange blocks); once the extension is finished and the $rmsd$ of the extended matches (red blocks in (d)) is still below the acceptance threshold, each extension (e, overlayed blocks) is NCS-transformed, as shown by arrows, onto the other chain fragment; finally longer, extended green and blue chain fragments are obtained (f), and the extended parts of the them (f, yellow blocks) are used as $C\alpha$ seeds for the next iteration of protein chain tracing.

Algorithm 3.3: Clustering of initial matches

```
Input  : list of matches $matches$, list of quaternions of matches $quaternion$
Output : list of clusters $clusters$
for $i = 0$ to $length(matches)$ do
    if $matches_i$ is not part of a cluster then
        $matches_i$ starts $cluster_x$
        $matches_i$ is part of a cluster
        $cluster_x \rightarrow quaternion = quaternion_i$
        for $j = i + 1$ to $length(matches)$ do
            $rotation\_difference_{ij} = (quaternion_i \cdot quaternion_j)$
            if $rotation\_difference < 5°$ then
                add $matches_j$ to $cluster$
                $matches_j$ is part of a cluster
                $cluster_x \rightarrow quaternion = cluster_x \rightarrow quaternion + \frac{quaternion_j}{length(cluster_x)}$
            end
        end
    end
end
```

3.1.2 Improving structural information by transformation of NCS copies

To find the longest continuous region of the NCS match between two chain fragments, each initial overlapping stretch (as shown in Figure 3.1b) is adjusted by extending the matching region in both directions along the chain (Figure 3.1c). During this extension the $rmsd$ is re-computed over the increased length of the fragment, L_{ext}. Should the $rmsd$ exceed a predefined threshold of $rmsd_{thresh} 0.2 L_{ext}$ Å, the inspected NCS match is not considered further. This helps to reduce false positives by avoiding arbitrary or unlikely matches. During fragment extension, the fragments are also tested for their secondary structure content using the algorithm of Zhang & Skolnick [176], described in section 2.5.1. This helps to avoid superimposing purely alpha-helical fragments to each other, since their abundance in proteins structures can often lead to false-positive matches. Once the extension is complete, the remaining 'tails' (Figure 3.1e, blue and green 'leftover'-tubes) are considered on both sides of the overlap region. All $C\alpha$ atoms from the tails of each chain fragment are NCS-transformed to the end-part of the corresponding chain fragment, Figure 3.1d,e. All extended parts of the chain fragments (referred to as *extensions* in the following) are assigned a weight as described in the next section. Should there be stereo-chemical clashes between an NCS-transformed $C\alpha$ atom and any other atom from existing protein chain fragments, the former is deleted. A stereo-chemical clash was defined as two atoms being separated by less than 0.7 Å. This value was chosen to allow some variation in the next iteration of chain tracing, i.e. allowing the *ARP/wARP* model update (section 1.5) to improve already built chains.

3.1.3 Weighting the detected *extensions*

The *extensions* obtained from the tails of the superimposed chain fragments (Figure 3.1f) are not error-free and therefore need to be weighted according to their estimated accuracy. The errors may originate from the detection of matches between common structural motifs, such as helices or any other repeating shapes, which may not necessarily be related by NCS. In addition, in case of an NCS order higher than two, more than one copy of the same extension can be obtained (e.g. for a trimer fragment1 transferred to fragment2 and also fragment3 to fragment2). Therefore, a weighting scheme for extended chain fragments (Figure 3.1d) was implemented to emphasise the most accurate *extensions* while reducing the influences of *extensions* that are less-confidently predicted. This weighting accounts for the clustering of initial rotational transformations (section 2.2 and Figure 3.1b), as well as the preliminary chain fragment extension (section 2.3, Figure 3.1c) (eq. 1):

$$W = \frac{(S_{Cluster} + S_{NCS}) + C^{\frac{(N_{Matches}-1)}{2}}}{rmsd_{ext}} \tag{3.1}$$

This equation contains two parameters reflecting the relative size of the cluster of rotations: the cluster size compared to all other clusters ($S_{Cluster}$) and the cluster size compared to the cluster size expected for an NCS-related part of the molecule (S_{NCS}). These parameters can take values between 0 and 3, as follows. $S_{Cluster}$ is 0 if the considered cluster ($cluster_i$) is smaller than the average cluster size, 1 if it is larger, 2 if it is at least twice as large, and 3 if it is three (or more) times larger. Similarly, S_{NCS} is 3 if the considered cluster is bigger than the cluster size expected for an NCS-related part ($cluster_{NCS}$); 2 if $cluster_i$ is greater than or equal to $0.5(cluster_{NCS})$; 1 if cluster $cluster_i$ is greater than or equal to $0.25(cluster_{NCS})$ and otherwise 0. $N_{matches}$ amounts to the total number of initially superimposed residues (Figure 3.1b) that have led to the construction of the extended chain fragment (Figure 3.1c). C is a scaling coefficient, usually set to 1. The denominator $rmsd_{ext}$ can take values between 0 and $rmsd_{thresh}0.2L_{ext}$ Å (the $rmsd$ of the extended chain fragment, as described in section 3.1.2).

Typically the weights vary between 0 and 100. The higher the weight, the more likely the extension obtained from this extended chain fragment is a valid NCS hit. The weights are then used in subsequent steps of the procedure to rank the *extensions*. A limited number of top-ranked *extensions* are fed back into the subsequent model building process.

3.1.4 Use of identified NCS-based chain fragment *extensions* for model building

Within the *ARP/wARP* workflow, the PNSextender method for automatic NCS-detection used for model improvement is applied to the intermediate model as depicted in Figure 3.2. Specifically, the obtained NCS-based *extensions* of partially built protein chain fragments are added to the current hybrid model as $C\alpha$ seed points (a more detailed explanation on why the *extensions* are taken as $C\alpha$ seeds and not as chain fragments can be found in Chapter 5). This hybrid model is used for subsequent tracing of protein chains in the main-chain building block.

3.1.5 Derivation of NCS-based stereochemical restraints:

Information about the identified NCS-related copies is also used to construct stereochemical restraints for the refinement of the hybrid model with REFMAC [194].During the refinement (section 1.3.2), the parameters of the model are adjusted to better fit the experimental data and *a priori* stereo-chemical restraints. If the NCS-relations detected by the PNSextender are valid, restraining those areas of the model in a similar way should improve the agreement to the experimental data. The input of NCS restraints into REFMAC is realised by specifying chains or fragments of chains that are related through NCS-operations. To ensure that NCS-based restraints are formulated only for highly reliable protein chain fragments, only extended overlaps that are more than 15 residues long are taken into account. This length was chosen since same is used as a cutoff for the automatic NCS-detection in REFMAC (section 1.6). Furthermore, since NCS restraints are applied to both main and side chain atoms (medium restraints, allowed positional deviation of 0.5 Å), they are only generated from chain fragments that *ARP/wARP* has docked into the sequence. Obviously, NCS-relations between chain fragments that have not been sequence assigned will thus be ignored. However, the use of REFMAC as a “black box” justifies this trade-off that avoids possible errors but might miss a few improvements of the model. The PNSextender for defining NCS relations used during structure refinement is invoked prior to each refinement step during *ARP/wARP*'s automated model building protocol, as depicted in Figure 3.2.

3.2 FittOFF - Fragment extension by motif comparison

As described in section 1.6, more than 50% of structures in the PDB, and especially large ones, contain NCS-relations. However, the remaining part does not. Thus, the problem of model fragmentation at medium-to-low resolution cannot be solved solely using the PNSextender or indeed, any other method solely relying on model improvement via the use of NCS-based information. Furthermore, some degree of fragmentation might still exist after the successful application of the PNSextender, due to, for example, parts of the structure missing in all NCS-related copies. Here, we focus on using structural motifs from the PDB to close gaps between partially built protein chain fragments in intermediate models obtained from *ARP/wARP* model building. These structural gaps are defined by the stem residues - the residues that anchor the motif from the PDB to the intermediate model - and the number of residues missing (an example is shown in Figure 3.3a). In the software framework OpenStructure [195] a method has been developed to find candidate motifs for such structural gaps by sampling backbone conformations from a large database of motifs or fragments extracted from high-resolution X-ray structures [196]. In this approach, candidates are automatically selected based on their agreement with the geometry of the stem residues and their stereo-chemical validity. Usually several hundred candidates are found, Figure 3.3b. In collaboration with Marco Biasini from the Schwede group, this method was enhanced (called FRAGRA, more thoroughly explained in section 3.2.2) by spatially correlating the candidates to residual density and thereby scoring them, Figure 3.3c. This introduces a rigid ranking and drastically decreases the number of applicable candidate motifs to be fitted into the structural gap (see Figure 3.3d). In the following, the FittOFF method (Fitting OF Fragments) is introduced. It was implemented to join the identification of the position and number of residues contained in structural gaps in the intermediate model (described in the next section) and the derivation of applicable motifs with the FRAGRA method (more thoroughly explained in section 3.2.2). An overview of the method is given in algorithm 3.4.

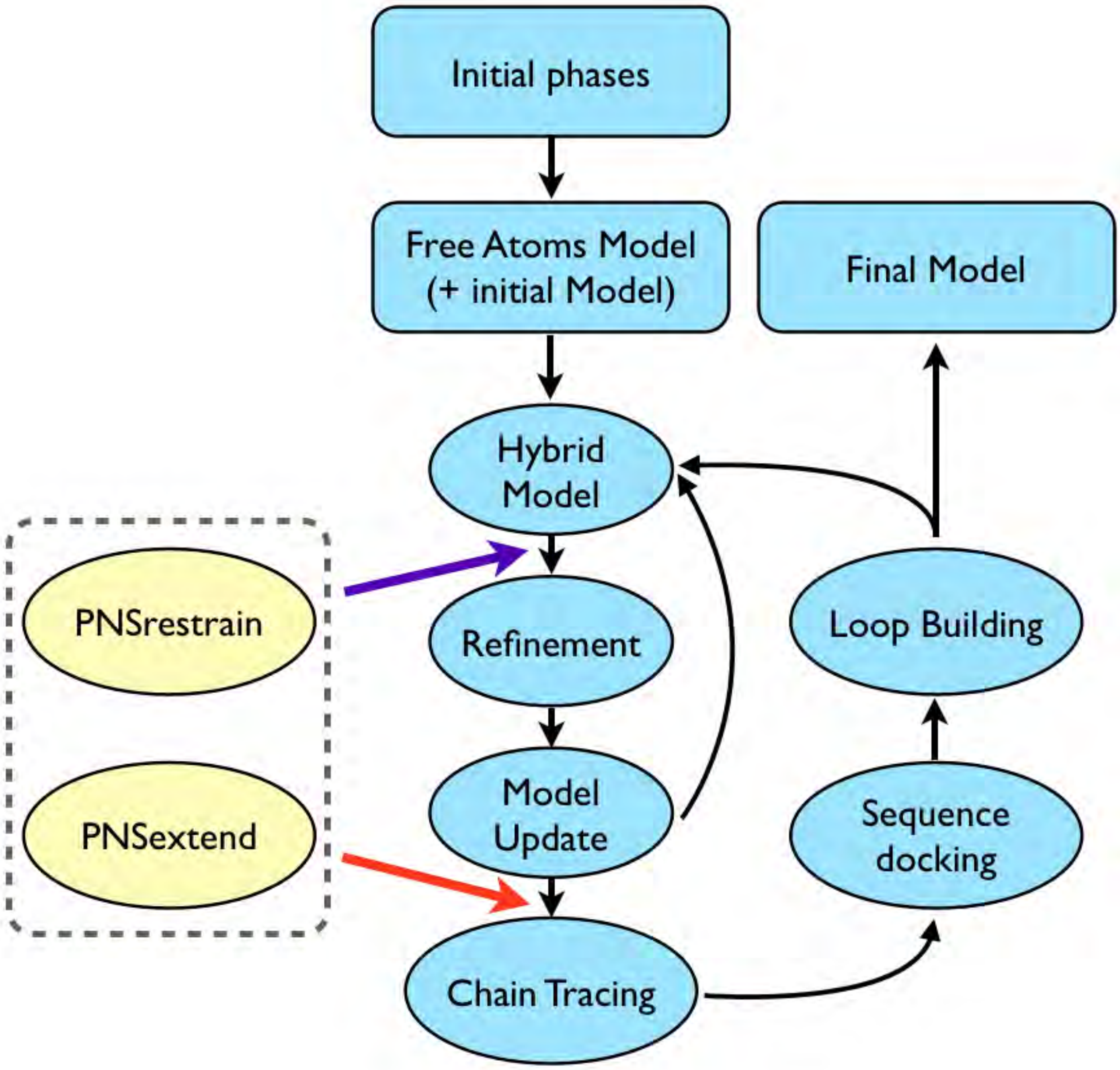
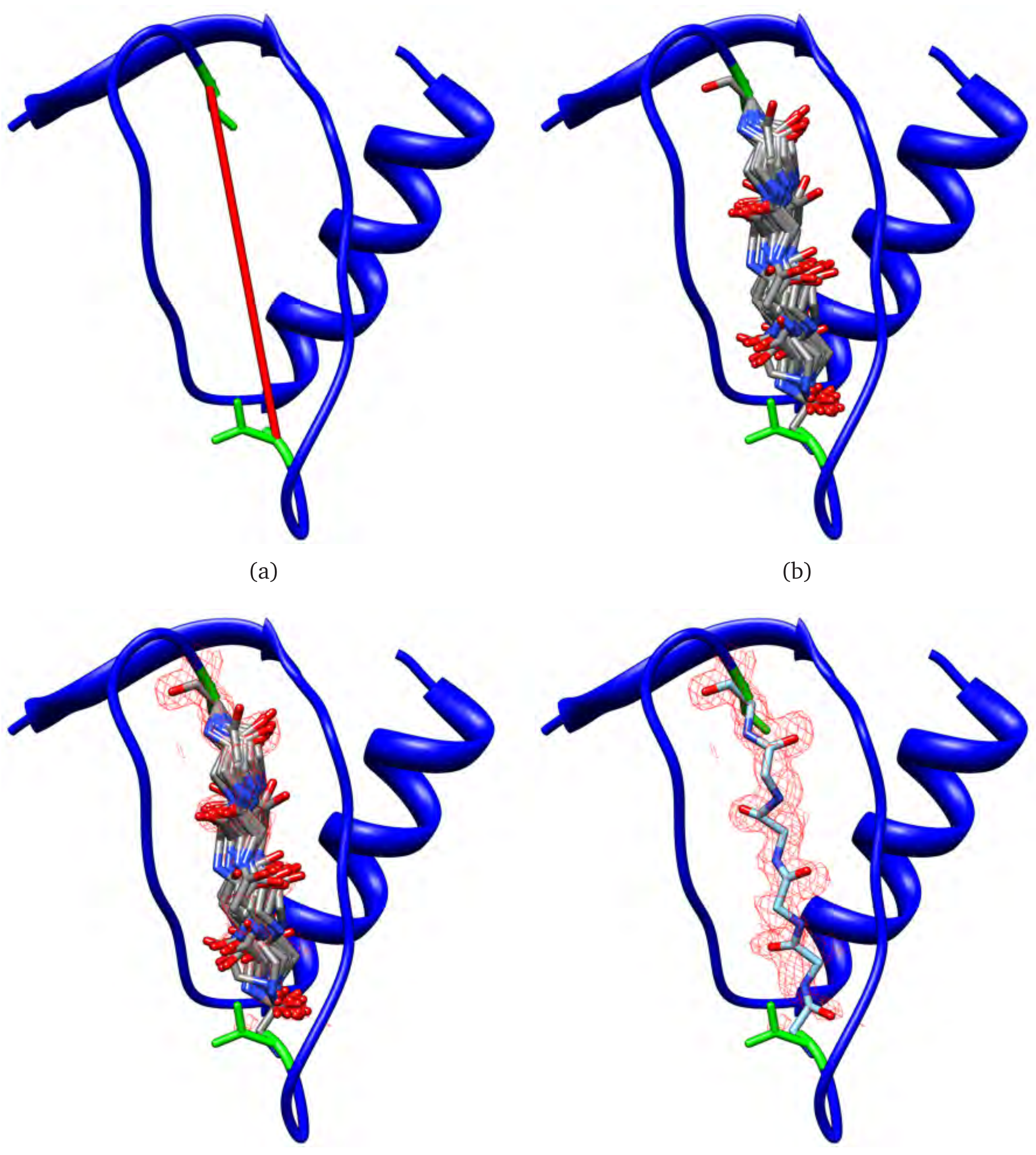


Figure 3.2: Flowchart of the *ARP/wARP* protein model building, including the PNS extender for automatic NCS detection (dotted box) and indicating its application for model extension (red arrow) and refinement restraints (purple arrow) of the intermediate model.



(a) (b)

(c) (d)

Figure 3.3: Overview of FittOFF method. A structural gap (a), stem residues are marked in green, gap length indicated by a red pseudo bond, is found and defined by the means described in section 3.2.1 and used for fragment fitting, the initial database search results in a huge number of candidates (b). By taking the residual density into account (c), it is possible to rank the candidates by map correlation and define one, or a few, top scoring results (d). The map correlation can further be used to identify regions that have been incorrectly identified as gaps.

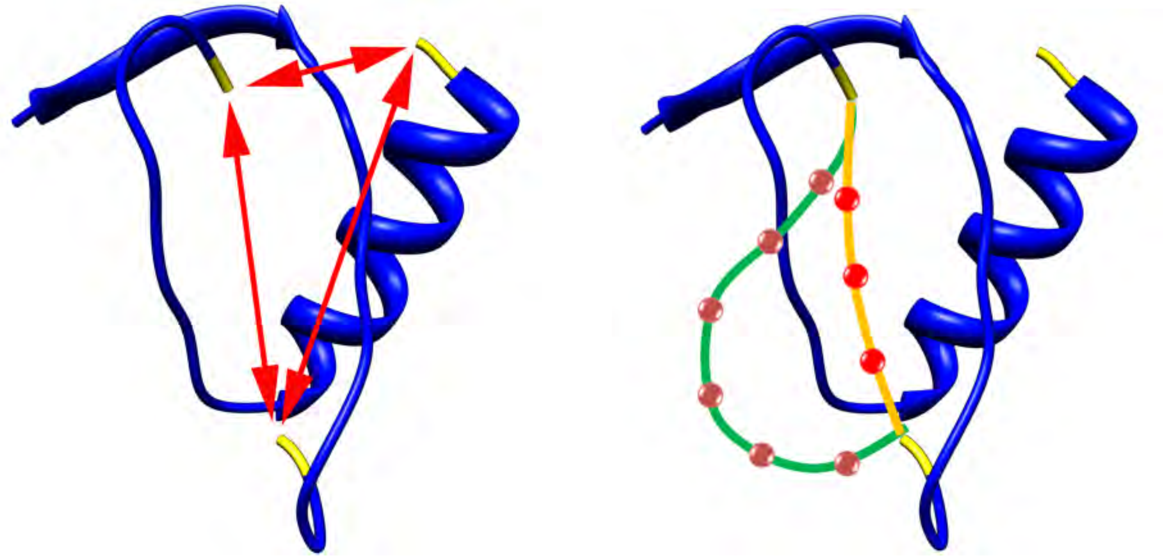
Algorithm 3.4: Overview of FittOFF method in pseudocode

```
Input  : intermediate coordinate file $coord_{gaps}$, electron density map $map$,
         secondary structure prediction $ss\_pred$, $\#high\_correl_{ranks}$
Output : extended coordinate file $coord_{extended}$
identify_gaps($coord_{gaps}$, $map$, $ss\_pred$)
foreach $gap_i$ in $list_{gaps}$ do
    $fit\_candidates$ = FRAGRA($gap$)
    save $fit\_canditates_x$ with highest correlation to $map$ as $fitted\_motif_i$
end
sort $fitted\_motif$s after map correlation
add best $\#high\_correl_{ranks}$ number of $fitted\_motif$s to $coord_{extended}$
remove those atoms clashing sterically with already-built model
```

3.2.1 Identification of gaps in intermediate models

The time-consuming process of generating a fragment database and implementing a method to find the best fitting fragment for given stem residues, length and residual density had already been accomplished by the Schwede group at the outset of the described work [196]. Nevertheless, there was still the rather challenging task of defining chain fragments from an intermediate model which are connectable by such motifs from the PDB. A way had to be found to identify potential stem residues and the number of residues contained in a structural gap, in order to solve the problem depicted in Figure 3.4. To achieve this, several approaches were combined into a sequential approach that is shown in algorithm 3.5.

At high resolution, the mutual location of chain fragments with respect to each other can easily be derived from their sequence assignment. However, at resolution lower than 2.5 Å the sequence docking algorithm in *ARP/wARP* (described in section 1.5) does not work sufficiently accurately and additional tools are being sought. For the identification of structural gaps between chain fragments, docking them to the sequence and thereby identifying their location is a tool too powerful to be ignored. In the FittOFF method, stem residues are identified by an evaluation of the partially built protein chains for their propensity towards secondary structure elements (following the algorithm described in 2.5.1). Thereby their location in a secondary structure predicted from an input sequence can be identified. The best three docking positions are stored for each fragment, this can result in different lengths (number of residues missing) for the same gap. The potential gap lengths are filtered for false-positives using a knowledge-based approach relating the number of residues contained in a gap to the distances between the $C\alpha$ atoms of the



(a)

(b)

Figure 3.4: Difficulties for defining stem residues and gap length. The question which stem residues can be connected by a gap is depicted in a). Additionally, there is the question of how many residues are missing in the potential gap, shown in b).

stem residues. As a validation criterion, the density between each pair of partially built protein chains is examined. Only structural gaps that are supported by a certain amount of density (which can be small, but must be significantly higher than zero) are used for fragment fitting with FRAGRA.

Secondary structure docking

As described in section 2.5.1, to derive the secondary structure of a macromolecule or fragments thereof, only the $C\alpha$ atoms of the protein backbone are needed. These atoms are the basic building blocks of any chain fragment one obtains during model building. We developed a method that detects the best agreement between a segment of secondary structure assigned to a chain fragment and a secondary structure predicted from an amino acid sequence to obtain results that are similar to sequence docking even at low resolution.

The secondary structure can be obtained from an amino acid sequence by either using the method described in section 2.5.2 or by a manual annotation of the sequence with specific information of domains or subunits that has been gathered so far. The

Algorithm 3.5: Overview of the method for the identification of gaps in intermediate models from *ARP/wARP* (identify_gaps)

```
Input  : intermediate coordinate file $coord_{gaps}$, electron density map $map$,
         secondary structure prediction $ss\_pred$
Output : list of gaps $list_{gaps}$
foreach $fragment_i$ in $coord_{gaps}$ do
    if $length(fragment_i) \geq 7$ then
        assign secondary structure to $fragment$
        dock $fragment$ to $ss\_pred$
    end
end
generate $list_{gaps}$ from docking results
foreach $gap$ in $list_{gaps}$ do
    compute probabilities for $gap_{length}$ based on distance statistics
    search for uninterpreted density between neighbouring $fragments$
end
filter $list_{gaps}$ with probabilities and uninterpreted density
rank $list_{gaps}$ after confidence scores
return $list_{gaps}$
```

assignment of the three secondary structure states (H, E, C) to each chain fragment is done automatically following the algorithm of Zhang [176], described in section 2.5.1; due to the required number of neighbouring $C\alpha$ atoms only chain fragments of at least seven residues receive an assignment. All assigned chain fragments are compared to the corresponding secondary structure prediction using an amended naïve string search algorithm. As described in section 2.6, the assigned secondary structure of the chain fragment (the pattern) is slid over the secondary structure prediction (the template) and at each offset, the number of matches or similarities between pattern and template is computed (see Figure 3.5). To obtain a better judgement, the number of matches and mismatches is evaluated at each position. The alignment with the highest amount of matches denotes the best fitting position. A predefined number of best fits (three in the current implementation) are kept for each chain fragment. For each of fit the percentage of matches with the secondary structure prediction is stored in the variable $conf_{dock}$. A description of the method in pseudocode can be found in Algorithm 3.6.

In order to gain insight into the location of possible structural gaps, all docked chain fragments are compared to each other and their relative positions analysed. In short, if one chain fragment ($fragment_i$) has been docked from position 0 to 14 in the sequence

and another one ($fragment_j$) from position 19 to 34, a gap of length four is assumed between the last $C\alpha$ atom of $fragment_i$ and the first $C\alpha$ atom of $fragment_j$.
If some chain fragments have been indeed sequence-assigned by *ARP/wARP*, this information is taken into account to derive further structural gaps or validate the results from the secondary structure docking.

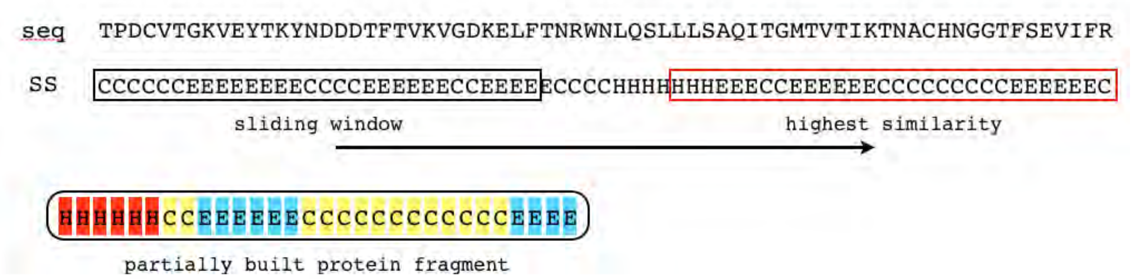

seq
TPDCVTGKVEYTKYNDDDTFTVKVGDKELFTNRWNLQSLLLSAQITGMTVTIKTNACHNGGTFSEVIFR
ss
CCCCCCEEEEEEEEECCCCEEEEEEECCEEEEECCCCHHHHHHHEEECCEEEEEEECCCCCCCCCEEEEEEEC
sliding window
highest similarity
HHHHHHCCEEEEEEECCCCCCCCCCCCEEEE
partially built protein fragment

Figure 3.5: Schematic overview of the sliding window method for sequence and secondary structure docking.

```
Algorithm 3.6: Overview of the method for secondary structure docking
Input  : chain fragment $fragment_i$, secondary structure prediction $ss\_pred$,
         number of top ranks $ss\_rank_{depth}$
Output : list of docking probabilities $dock\_prob_{SS}$
for $i = 0$ to $(length(ss\_pred) - length(fragment_i))$ do
    compare$(fragment, ss\_pred[i..i + length(fragment_i)])$
    save matches $offset_{sim}$ and mismatches $offset_{penalty}$
end
add $ss\_rank_{depth}$ best $offset_{sim}/offset_{penalty}$ to $dock\_prob_{SS}$
```

Relating gap length to the distance between stem residues

Saving the three best positions for each chain fragment following secondary structure docking can, in the worst case, lead to nine different lengths for the same structural gap. To decide which length is the most likely another source of intrinsic information is exploited - the spatial distance between the terminal $C\alpha$ atoms of chain fragments in the intermediate model. This is based on the expectation that due to the intrinsic properties and geometrical features of the peptide backbone, it must be possible to identify a relation of the distance between two potential stem residues to the number of residues between them.

A knowledge-based method has been developed to provide a probability value for the number of residues to be enclosed in a gap (gap_{length}) given a certain distance between the stem residues anchoring it to the chain fragments ($gap_{distance}$), or:

$$P\left(gap_{length} \mid gap_{distance}\right) \tag{3.2}$$

A survey of a large set of structures from the PDB (selected as described below) was undertaken where we investigated the distances between two $C\alpha$ atoms with none to 14 residues enclosed between them. For each of these gaps the number of residues being part of an α-helix, β-sheet or loop was also stored. To avoid bias towards redundant structures, the survey was performed with the PDB50 subset from the PDB. This subset is generated by clustering all protein chains of at least 20 amino acids at 50% sequence identity (i.e. all chains sharing at least half of their sequence information belong to one cluster). All objects in each cluster are ranked according to resolution and deposition date. The highest ranked chain is then included in the PDB50 subset. The set used in this study was selected from the PDB in January 2011 and contained 6,613 chains that were solved by MX at resolution of 2.0 Å or better.

This way we obtained a database relating distances between the stem residues to the number of instances in which 0 to 14 residues were inside the gap (within the limits of 0 Å to 40 Å and a step size of 0.1 Å). This also made it feasible to construct 15-dimensional probability vectors $\vec{P}$ denoting a list of probabilities P_i for a gap of a certain size to occur at each distance in the database. All 15 elements in each vector $\vec{P}$ sum to one:

$$\sum_{i=0}^{14} \vec{P}_i = 1 \tag{3.3}$$

A vector $\vec{P}$, reduced to four dimensions for simplification, for the distance of 10 Å and taking only the occurrences for gaps with two, four, six or eight residues into account, would be the following:

$$\vec{P}(10\,\text{Å}) \begin{bmatrix} P(2) \\ P(4) \\ P(6) \\ P(8) \end{bmatrix} = \begin{bmatrix} 0.48 \\ 0.22 \\ 0.24 \\ 0.06 \end{bmatrix} \tag{3.4}$$

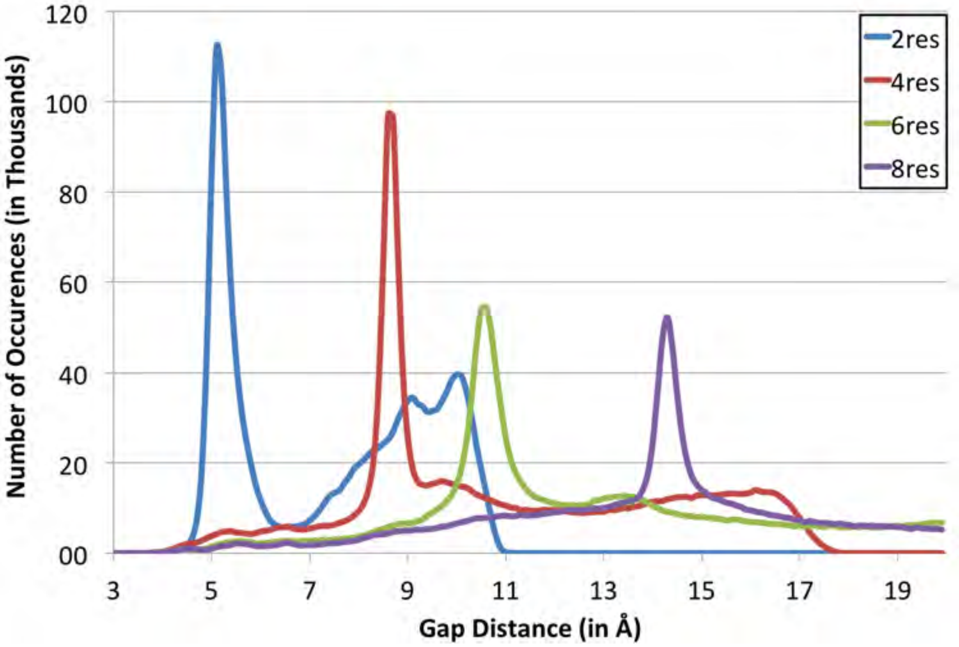
An overview of occurrences for gaps with two, four, six, or eight residues at distances between 0 Å to 20 Å and the related distribution of probabilities is given in Figure 3.6. Distinctive peaks are evident, for example, for gaps of two residues containing helices at ∼5.1 Å a gap of two residues containing a β-sheet at 10.0 Å residues and a helical gap of eight residues at ∼14.3 Å.

Using this database, the identification of incorrectly assigned lengths for structural gaps obtained from secondary structure docking can be accomplished by evaluating the vector of probabilities for the distance between the $C\alpha$ atoms of their stem residues. The probability of observing each assigned gap length is stored in the variable $conf_{pvec}$. Furthermore, all potential structural gaps with a physically impossible number of suggested missing residues ($\frac{gap_{distance}}{gap_{length}} > 4.5$ Å, including some error tolerance) are not taken into account. In addition to the identification of stem residues and gap length, the docking also gives information on the secondary structure content of the proposed gap. To facilitate easy retrieval of this information and obtain databases with higher discriminative power, the initial database was divided into 10 distinct ones. The first three databases denote all gaps with more than 50% / 75% or 100% helical content, databases four to six and seven to nine denote the same for sheet and coil content. The last database comprises all gaps that contain less than 50% of any secondary structure element. Examples for the higher discriminatory power of the new databases are given in section 4.2.

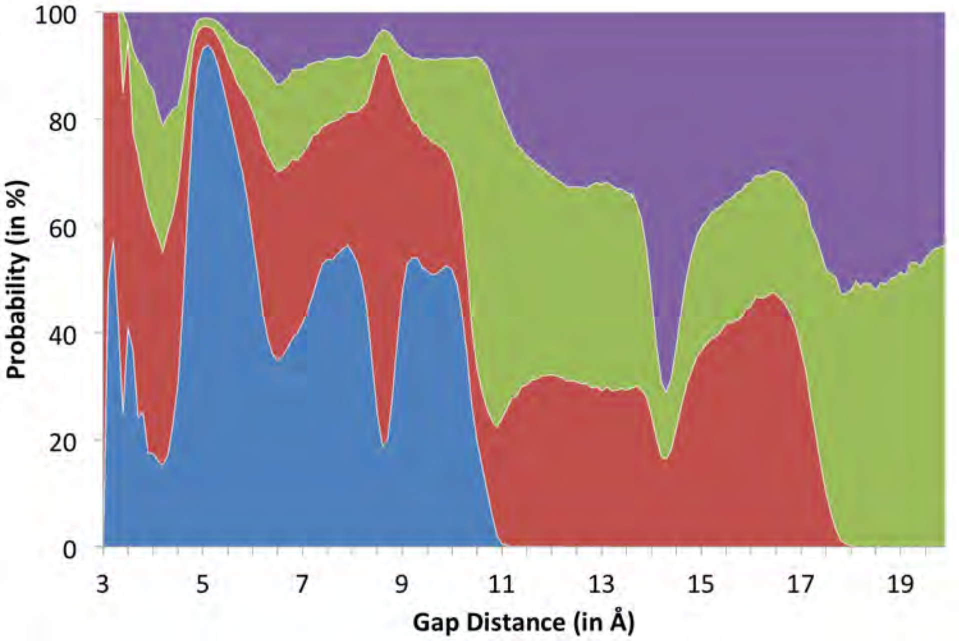
Uninterpreted density

Results from FittOFF are used as $C\alpha$-seeds for subsequent chain tracing. Hence, the gaps selected for fragment fitting with FRAGRA should be supported by residual electron density if *ARP/wARP* is to incorporate them into the next model. A lack of supporting density would lead to ambiguous results in FRAGRA, since density is used to rank the identified candidates (see Figure 3.3c). Furthermore, any effort spent on analysing gaps that lack the support of experimental density is essentially wasted computation time.

Therefore, all proposed structural gaps are analysed for the amount of uninterpreted density between the two partially built protein chains that are to be connected. To evaluate the density, a set of points (called a pointcloud) between the $C\alpha$ atoms of the stem residues is generated. This set places points at 1 Å intervals along a straight line between the $C\alpha$ atoms - in the following referred to as centre points. Eight additional points are placed on a circle with a radius of 1 Å around each centre point (Figure 3.7a,b). These points $i = (1..8)$ are positioned at angles of $i\frac{\pi}{4}$ on the described cycle. The density level is computed for every point of the pointcloud and the average density is calculated ($< dens_{cloud} >$, where the angle brackets denote the average). To decide whether the



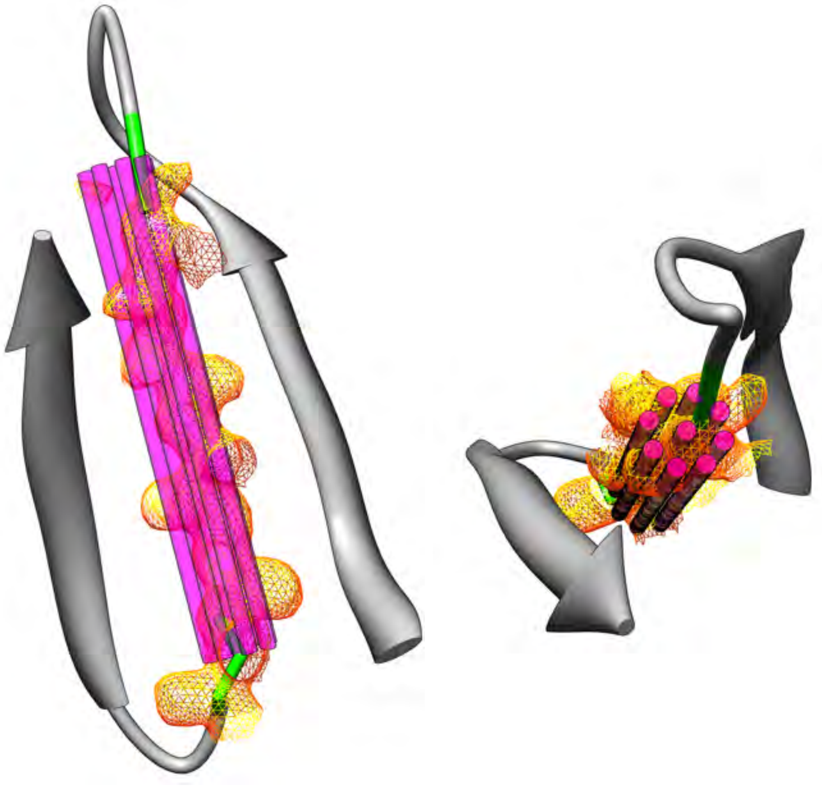
(a)



(b)

Figure 3.6: Relations between gap length and distance between stem residues for gaps of 2, 4, 6, and 8 residues. Occurrences of a certain distance (in Å) are shown in (a). The related probability distributions are shown in (b).

density between the stem residues is strong enough to support the fit of a fragment into its related structural gap, $< dens_{cloud} >$ is compared to the average density over all built $C\alpha$ atoms in the intermediate model ($< dens_{model} >$). Since *ARP/wARP* will only keep $C\alpha$ atoms in high levels of electron density, all points with density levels of less than 0.1 electrons/Å^3 will not be taken into account for the computation of $< dens_{cloud} >$. If the $< dens_{cloud} >$ is higher than $\frac{<dens_{model}>}{4}$, the gap will be used for fragment fitting in FRAGRA. These thresholds were chosen empirically to also allow regions with significant gaps in the electron density. The pseudocode is shown in algorithm 3.7.



(a) (b)

Figure 3.7: Evaluation of uninterpreted density between partially built protein chain fragments that were identified as enclosing a structural gap.

3.2.2 FRAGRA

FRAGRA is a knowledge-based method to remodel structural gaps and has been developed by our collaborator Marco Biasini from the Schwede group at the Biozentrum in

Algorithm 3.7: Evaluate uninterpreted density between chain fragments to be connected by a gap

```
Input  : stem residues $gap_{start}$, $gap_{end}$, electron density map $map$, average
         density over all $C\alpha$ atoms $< dens_{model} >$
Output : Bool $use\_gap$
Build up $pointcloud$ between $gap_{start}$ and $gap_{end}$
foreach point in pointcloud do
    $point_{dens} = \text{lookup\_density\_at}(point)$
end
calculate $< dens_{cloud} >$ over all $point_{dens}$
if $< dens_{cloud} > \geq \frac{<dens_{model}>}{4}$ then
    $use\_gap = 1$
end
```

Basel. The method was originally designed to produce accurate results for short loops, to complete homology models in seconds rather than hours and to close structural gaps up to 14 residues in length. To achieve this, FRAGRA incorporates a large database of backbone conformations (hereafter called fragment database). FRAGRA differs from related methods such as XPLEO [156], as it uses existing fragments from the PDB instead of rebuilding gaps with physical approaches. This results in a drastic decrease of the required computation time (under a minute for FRAGRA compared to up to two hours with XPLEO for a gap of 12 residues). In principal, FRAGRA follows the concept of loop modelling as described in [197]. Within the FittOFF method, these backbone conformations are sampled to find the fragment fitting best into a structural gap identified by the means described in section 3.2.1. In the following the fragment database, the sampling of backbone candidates from the fragment database, as well as the scoring of identified candidates are described.

Fragment database

The fragment database has been constructed from about 60,000 protein chains solved by X-ray crystallography. Only structures with experimental data extending to a resolution of 2.2 Å or better have been included, which provided a good trade-off between the quality of the backbone models and the number of chains included in the database. The database uses a hash generated from the geometry of the two residues lining each fragment, the so-called stem geometry. Here, $C\alpha - C\alpha$ distances as well as the angle between the $C\alpha - C\alpha$ and the planes formed by $N - C\alpha - C$ of the N-terminal residue and the $C\alpha - C - O$ plane of the C-terminal residue are used as descriptors.

Sampling of backbone fragments

During the sampling a list of fragments that are suitable candidates to remodel the backbone of a structural gap, is provided. In this step no scoring is applied. As input, the loop length and six coordinates are required (the N, $C\alpha$, C positions of the N-terminal stem residue and the $C\alpha$, C and O positions of the C-terminal stem residue). This information is used to calculate the stem geometry for the structural gap. All fragments from the database that are in agreement with this stem geometry are selected. To improve the fit at the stems, small fragments with a length of three residues are used to bridge between the stem residues and the backbone.

Scoring the candidates

The list of candidates found during backbone sampling contains 1500 fragments on average. To decide which one fits the gap best, a scoring scheme is applied. At first, fragments that clash with the already built protein structure are filtered out. In the next step validation measures from the QMEAN scoring function [146] are applied. As mentioned in section 3.2, the method was enhanced to incorporate the electron density information in the region of the structural gap. A finer ranking is achieved by spatially correlating the candidates to the residual density. The expected density is computed by placing a Gaussian sphere of density at each atom and the real-space correlation to the experimental density is calculated as described in [198]. The fragments are then output in PDB-format ranked according to the real-space correlation value. In the case, where the gap length cannot be determined exactly by FittOFF, different lengths can be used in FRAGRA and the real gap length can be identified by analysing the map correlations for different trials.

3.2.3 Application of fitted fragments to model building

FittOFF is applied to the *ARP/wARP* workflow in the same way as PNSextend; fitted motifs are added to the considered hybrid model as $C\alpha$ seed points for subsequent tracing of protein chains. For more information refer to Figure 3.2 and section 3.1.4.

3.3 Data

To examine whether the introduced methodologies improved automated model building in *ARP/wARP* and, if so, to evaluate the obtained improvement, high-resolution structures from the PDB were used as well as some structures that had been submitted to the *ARP/wARP* model building web service [199] and made available to us for testing purposes. A good representative example is the 1.6 Å structure of the B subunit of a mutated shiga-like toxin from *Bacteriophage h30*, expressed in *Escherichia coli*, PDB ID 1c48 [200]. The molecule is arranged as a homo-pentamer, with each subunit composed of 69 residues. This structure was predominantly used for the basic development of both the PNSextender and the FittOFF methods. The full test set used for subsequent evaluation of the PNSextender consisted of 13 multimeric structures that were determined by molecular replacement or isomorphous replacement at resolution ranging from 1.9 to 3.2 Å. These structures were chosen since they exhibited clear NCS, were of different sizes and comprised a wide range of secondary structure content. More specifically, the structures had an asymmetric unit content varying between 300 and 2300 residues in 2 to 10 NCS-related subunits and were characterised by various secondary structure content, so that there were predominantly helical, predominantly stranded or mixed alpha-beta models (for a complete overview refer to table A.1).

For the FittOFF method, the test set of ten structures was chosen based on different features. Only structures with low molecular weight (15 - 25 kDa) were selected to ensure fast tests. Also, the structures had to have been solved at a resolution lower than 3.0 Å and contain a variety of secondary structure elements (for a detailed overview refer to table A.2). For all test cases secondary structure predictions were generated with PSIPRED version 3.0 and SSpro version 4.1. Additionally, secondary structure assignments with the Zhang algorithm were also generated for each test case.

3.4 Tests environments

Initially, we tested the performance of the PNSextender module to automatically identify and apply NCS relations to the appropriate parts of the model - the ’exclusion’ test. A single model - the mutated shiga-like toxin B-subunit (PDB 1c48) - was used for this purpose. The structure was artificially fragmented by cutting out parts of the model to mimic real cases where intermediate models may contain a large number of unconnected fragments. To generate cases with various degrees of fragmentation, ten differently fragmented structures were built. Starting from the complete structure, 5% of residues

were successively deleted from each model. Hence, 95% of the structure left in the first test case, 90% in the second case through to 50% in the 10^{th} case. The models were fragmented by cutting out blocks of residues (15 to 30 amino acids, see figure 4.2b for the 7^{th} case, 65% of the model left) from different parts of the structure.
A similar approach was chosen for testing the FittOFF method. Again the mutated shiga-like toxin was taken and parts of it were excluded. This was done to emulate gaps, which could then be identified and filled using the FittOFF method. The model was fragmented to obtain six gaps of various length and secondary structure content (between two and nine residues, with mainly helical, mainly sheet, mainly loop and mixed content). For the use in FittOFF during this test the electron density of 1c48 at 1.6 Å and the correct secondary structure information was used.

Subsequently, the results were evaluated when all of the test structures described in section 3.3 were built using the automated model building protocol of *ARP/wARP*. One batch of tests was executed with the PNSextender for NCS extension and restraints, another one with the FittOFF method. Each protocol was executed with five cycles of model update and refinement after each of the ten model building cycles. For FittOFF, additional tests were conducted for each of the generated secondary structure predictions (PSIPRED, SSpro and the Zhang assignment).

For the published tests [1] of the PNSextender module, *ARP/wARP* [54] version 7.2, REFMAC [194] version 5.5.0109 and CCP4 [52] version 6.1.13 were used. Later tests of an advanced version of the PNSextender, as well as the tests of the FittOFF method, were executed with *ARP/wARP* version 7.3, REFMAC 5.7.0028 and CCP4 6.2.0. Both methods have been tested on an Apple iMac (quad-core, 2.8GHz, 10GB RAM), running MacOSX 10.6.8 (Snow Leopard).

3.5 Implementation and complexity

Computationally demanding core functions of the PNSextender and FittOFF methods are written in the C programming language. To simplify access, both methods are called via a Perl wrapper which also takes care of extensive file handling. Many functions called from the C-routines are part of the f77/f95 fortran library *arplib* and the C-library *mapread* that have both been developed for many years as part of the *ARP/wARP* software project. Specifically, parsing of PDB-files and electron density maps has been accomplished using functionality in the *mapread* library. This also includes functions for the lookup of density values at given coordinates. Functions used for the computation of superpositions of structural fragments and derivation of quaternions are those from *arplib*. Fortran functions from *arplib* are called via a C-interface (arplibc.h), while *mapread* functions are called directly. Parsing of the PDB50 subset for relating gap length to distance databases, as well as binning of occurrences after distance is implemented in Python. The databases are generated with Microsoft Excel and saved in the cvs file format.

The required CPU time for both the PNSextender and FittOFF methods rises with the size and degree of fragmentation of the structure under consideration. For the PNSextender, the most computationally demanding subroutines are the identification of initial matches with NCS-relations (algorithm 3.2) and the clustering of those matches. The former compares all stretches of a fixed length of $C\alpha$ atoms to all other stretches, thus arriving at a complexity of $O(n^2)$, where n is the number of $C\alpha$ atoms. A similar complexity class is achieved for the latter - the first cluster begins with the first match and all succeeding matches that obey the clustering criterion are added to same cluster. In the worst case, where the number of clusters equals the number of matches, we arrive at complexity class $O(n^2)$, but in the average case a complexity class of $O(n \log n)$ can be expected. The usual CPU time for executing PNSextend is in the area of a few milliseconds - which is only a small additional overhead compared to the model building without it. The generation of NCS-restraints in PNSrestrain is even faster, since only the longest overlaps between NCS-related fragments need to be identified.

The computationally most exhaustive routines in the FittOFF method are the secondary structure docking and FRAGRA itself. In the secondary structure docking, every fragment is compared to a secondary structure prediction that has the length of the amino acid sequence of the structure. In the worst case, with very short fragments, the complexity is $O(n^2)$. The complexity of FRAGRA is similar, but due to the large number of fragments that have to be evaluated every time, the computation time needed is considerably longer. Thus the execution of the FittOFF method takes from a few seconds to just under a minute, correlating to the number of gaps that are examined for fragment fitting with FRAGRA.

Chapter

4

Results

This chapter describes the tests of the PNSextender and FittOFF methods and is structured as follows. In the first section, the importance of scoring the NCS-extensions is explained and justified with results. Similarly, the splitting of the distance database used by FittOFF is justified by referencing the results of the tests. Subsequently the effects of both methods in idealised test cases with no coordinate errors are shown: these serve as a proof of principle study for the function of the methods. The last part deals with the incorporation of both methods into the *ARP/wARP* model building protocol and their performance on real world test cases.

4.1 The importance of scoring *extensions* derived by NCS-identification

To support the validity of the weighting scheme (as described in 3.1.3), $rmsd$ values between the NCS-extended parts of a model (NCS-extensions) and a reference structure were calculated for a number of cases. These were compared with the weights assigned to the *extensions* (eq. 3.1). As expected, small deviations from the reference structure, in the order of 0.2 Å or less, corresponded to *extensions* with high weights, Figure 4.1. *Extensions* with low weights display larger deviations (∼0.7 Å and more) from the reference structure. This concludes the validity of the suggested weighting scheme and demonstrates that *extensions* with higher weights are, indeed, more accurate.

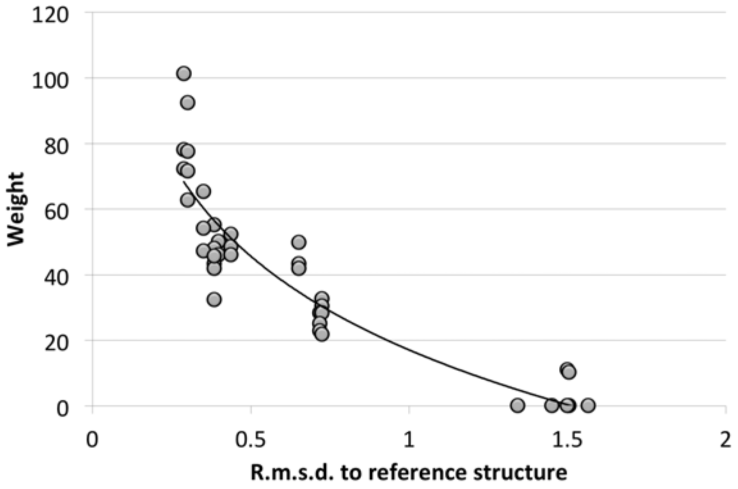


Figure 4.1: Estimated reliability of the derived weights used for the NCS-extension and the accuracy of the obtained extended parts of the model.

4.2 The necessity of using secondary structure content for relating gap length to distance

As described in section 3.2.1, it was decided to divide the distance database using the content of secondary structure elements in the gaps. In the following the necessity of such a partition is shown by comparing the overall database (Figure 4.1d) to the three databases with more than 75% helix (Figure 4.1a), sheet (Figure 4.1b) or coil content (Figure 4.1c). For simplification, the databases have been reduced to include only gaps containing two, four, six or eight residues.

Let us consider three gaps missing the following fragments: an eight-residue α-helix with a distance of 16.0 Å between the stem residues, a four-residue β-strand with a distance of 14.6 Å and a two-residue loop with a distance of 8.6 Å. For the first gap, using the overall database (blue bar in Figure 4.1d) would indicate 45% probability for a gap of four residues and only 30% for the right gap length. However, since this is a helical gap, the 75%-helix database can be used. Here, the highest probability (93%, blue bar in Figure 4.1b) denotes the right length of eight residues. Similarly, for the second and third gap the probabilities obtained from the overall database indicate gaps of length eight (red bar in Figure 4.1d) and four (yellow bar in Figure 4.1d), respectively. Using the 75%-sheet and loop databases gives the highest probability to the expected number of enclosed residues of four (red bar in Figure 4.1b) and two (yellow bar in Figure 4.1c). Hence, the databases for more than 75% of a given secondary structure content can indicate the right gap lengths with higher probability. Similar results have been obtained for the sets with more than 50% and 100% of a given secondary structure content (data not shown).

4.3 Effect on the completeness of the structure in the absence of the coordinate error

Using NCS-extensions and restraints

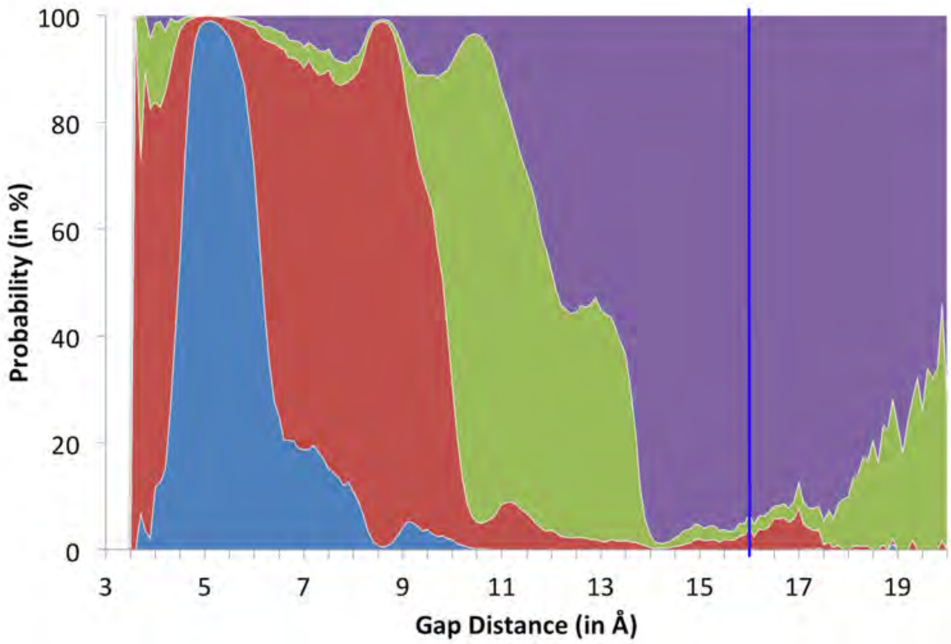
For each of the artificially fragmented test structures from the mutated shiga-like toxin B-subunit 1c48 (Figures 4.2a,b; fragmentation described in section 3.4), the PNSextender was used to retrieve the missing $C\alpha$ atoms. Each structure was checked against the full reference model to examine the accuracy of the retrieval. For the first seven test structures (5% to 35% of the residues excluded) our method rebuilt the complete model with an $rmsd$ from the $C\alpha$ atoms of the reference structure of 0.33 Å or better, Figure 4.2b,c. As would be expected, the accuracy of the retrieved parts of the structure decreased gradually as a larger fraction of the model was excluded (Table 4.1). For the last three cases, it was not possible to retrieve the complete structure because some elements of the structure were missing in all five subunits. Nevertheless, the accuracy of the retrieved parts was still very high.

Using fragment fitting

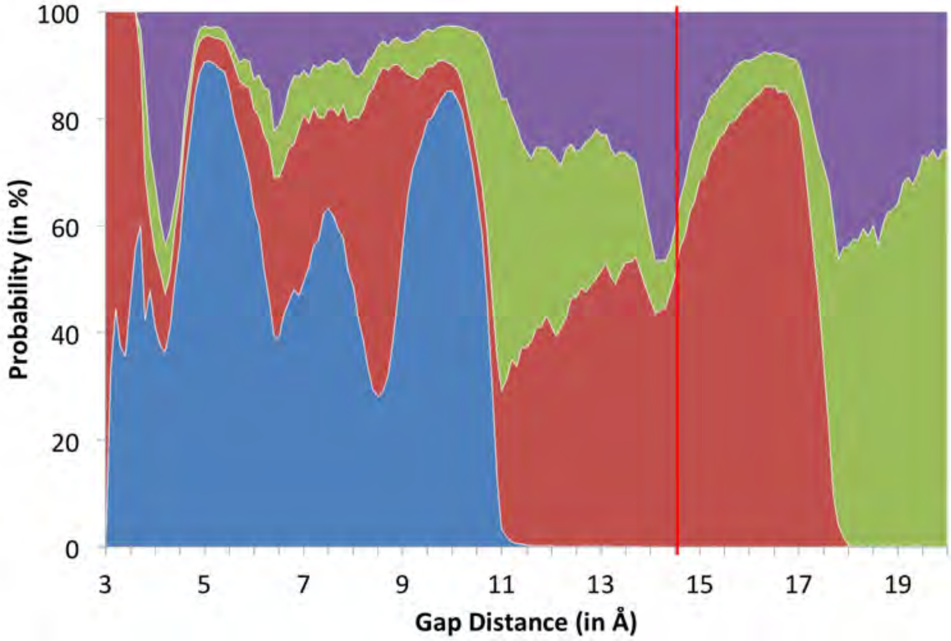
To assess its effectiveness, the FittOFF method was used to improve the completeness of the artificially “gapped” structure of 1c48 (described in section 3.4). The identification of gaps and the fitting of the highest-ranking fragments have been evaluated independently.

Of the six gaps in the test structure, five could be successfully detected using secondary structure docking. The sixth gap had a size of only five residues, thus preventing it from assignment of secondary structure.

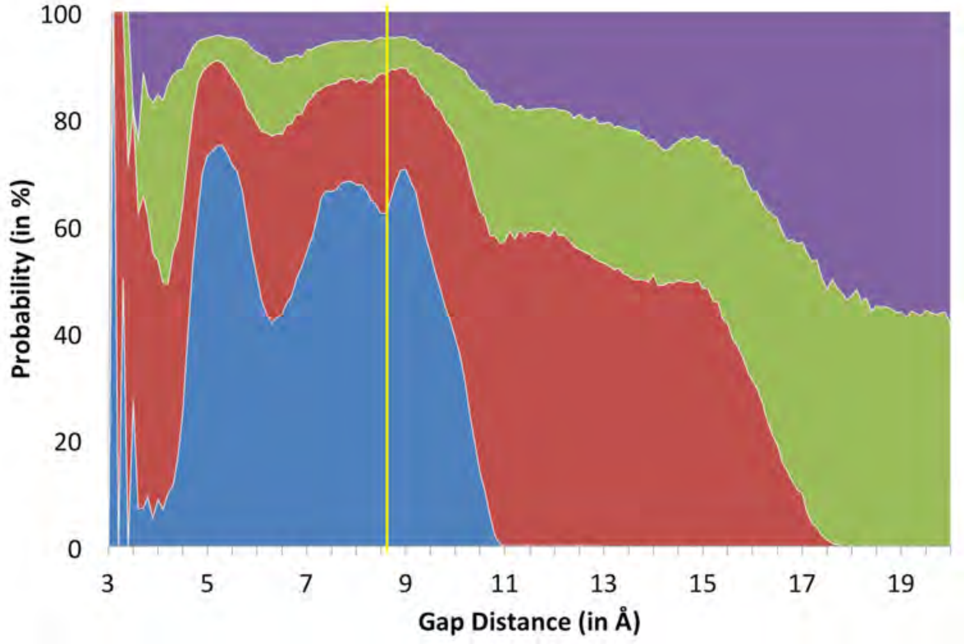
The retrieval of the five gaps was attempted using two different protocols. Initially, a rigid gap filtering was tried, requiring a secondary structure docking of the anchoring fragments with at least 60% confidence (referred to as $conf_{dock}$) and a probability for the suggested number of residues missing in the gap of at least 50% ($conf_{pvec}$). Using this protocol, three of the five gaps were automatically detected without any mistakes.



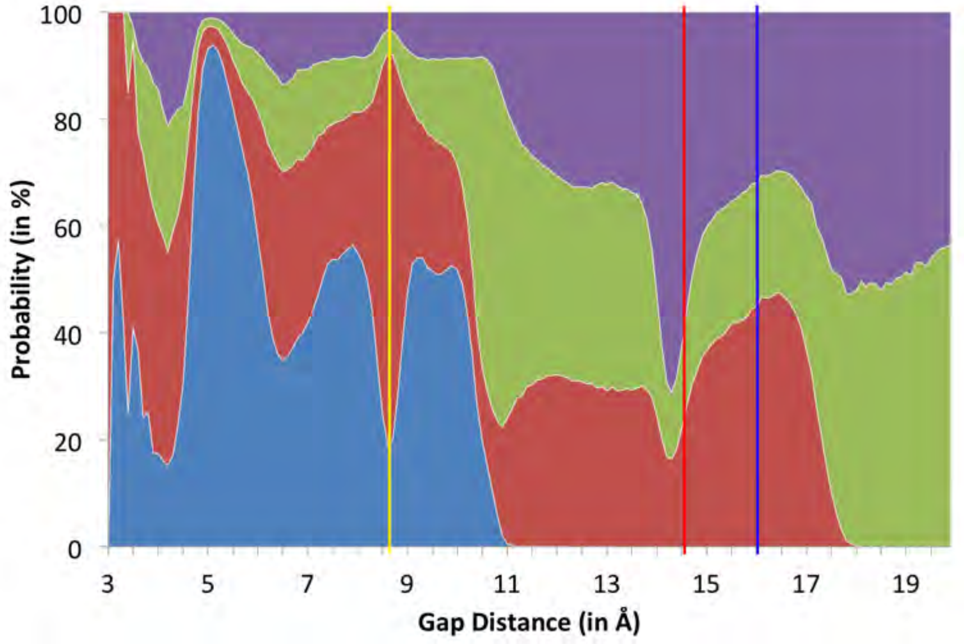
(a) Gaps > 75% Helix



(b) Gaps > 75% Sheet



(c) Gaps > 75% Loop

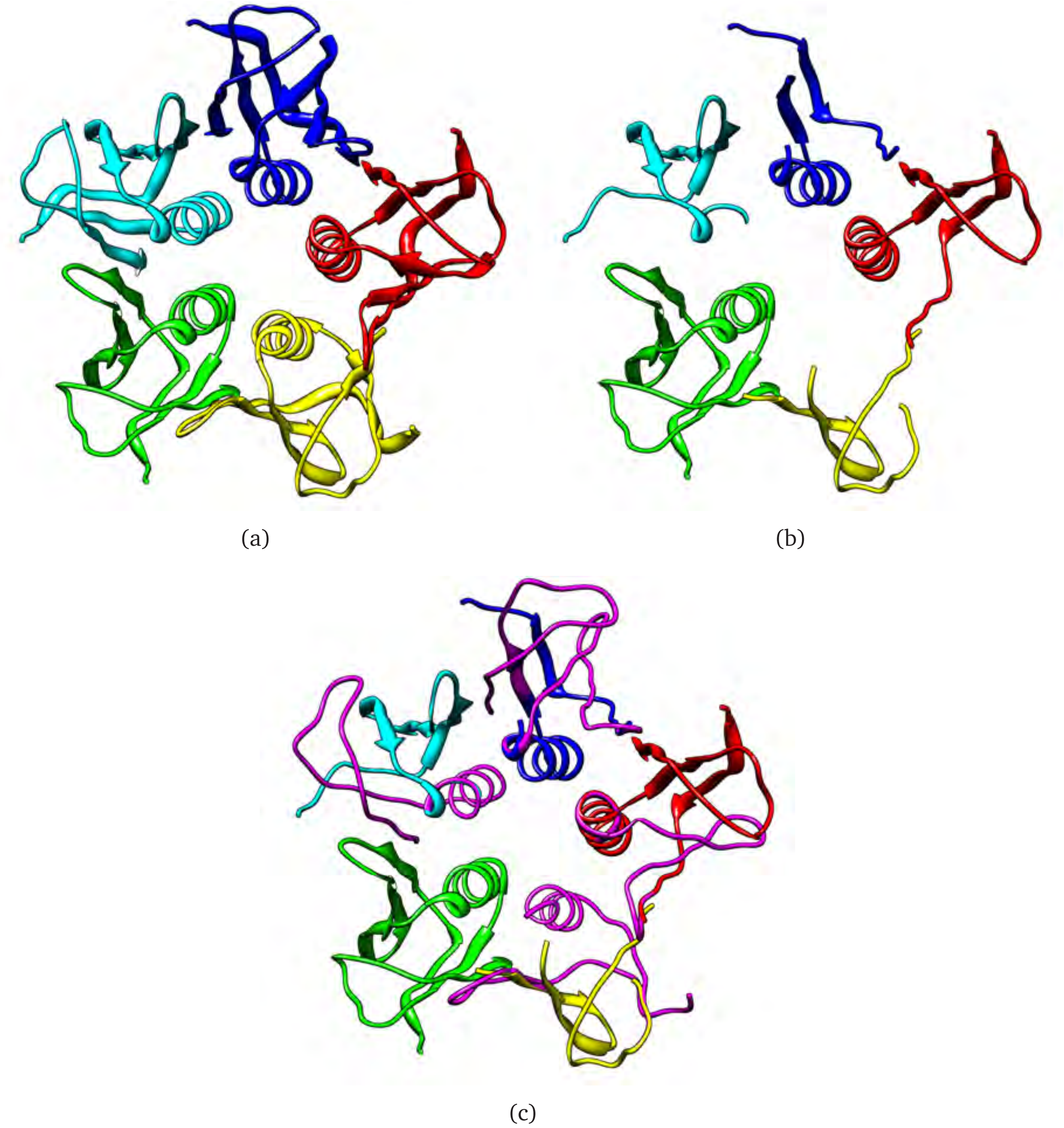


(d) Not divided using content of secondary structure element in the gap

🟦 2res 🟥 4res 🟩 6res 🟪 8res

(e)

Figure 4.1: Comparison of distance databases split according to secondary structure content for gaps of two, four, six and eight residues. Plot a) shows the probability distributions for a gap with at least 75% α-helical content. The same is shown for β-sheets and coil in b) and c), respectively. The last plot shows the distribution for the original database not split according to secondary structure content. The blue bar denotes the helix example, the red bar the sheet example and the yellow bar the coil example. White areas are related to steps without any occurrence.



(a)

(b)

(c)

Figure 4.2: Validation test of the PNSextender - exclusion of residues. a) the original structure (pdb ID 1c48); b) the same structure, with 35% of all residues excluded; c) all 35% of the missing residues are retrieved. Retrieved parts of the structure in c) are coloured in magenta.

Percent of the model excluded	Completeness of the initial model (%)	Residues retrieved	$rmsd$ of the retrieved structure to reference model (Å)	Completeness of the retrieved structure (%)
5	95	17	0.08	100
10	90	34	0.09	100
15	85	52	0.14	100
20	80	69	0.21	100
25	75	86	0.29	100
30	70	103	0.30	100
35	65	121	0.32	100
40	60	107	0.31	91
45	55	122	0.37	90
50	50	126	0.45	87

Table 4.1: Test of the PNSextender - exclusion of residues. The table shows the number of residues retrieved and their $rmsd$ to the reference crystal structure.

FittOFF identified 15 gap candidates and 45 corresponding unique lengths by secondary structure docking. Of these candidate solutions, four gaps and 12 lengths were immediately discarded, because their values were beyond the limits of the distance databases, such as distance between anchors longer than 40 Å or more than 14 missing residues. Furthermore, another gap and five lengths were removed from consideration, since the relation between gap distance and missing residues required an impossible average $C\alpha$-$C\alpha$ distance of more than 4.5 Å. Thus 10 gaps and 28 unique gap lengths were subsequently subjected to the filtering using the relations between gap distance and the number of missing residues. In this step seven gaps and 25 lengths were filtered out, since the probabilities for the suggested number of residues missing were lower than 50%, leaving the three gaps and corresponding lengths mentioned above. No potential gap or length was filtered out by the check for uninterpreted density.

By applying a protocol that required a lower value of $conf_{pvec}$, of only 10%, all five gaps could be identified correctly. As might be expected, loosening the filtering cutoff led to more than one potential gap length being suggested for three of the five gaps. The filtering started from the same 10 gap candidates and 28 corresponding unique gap lengths obtained after discarding those that were either physically impossible or beyond the database limits. In the subsequent filtering two gaps and 15 lengths were removed from consideration using the distance databases and another three gaps and four lengths were eliminated after checks for uninterpreted density. This finally resulted in five gaps being identified, with altogether nine unique estimates for the number of missing residues. An overview of the number of gap candidates and corresponding lengths filtered out in both protocols is given in Table 4.2.

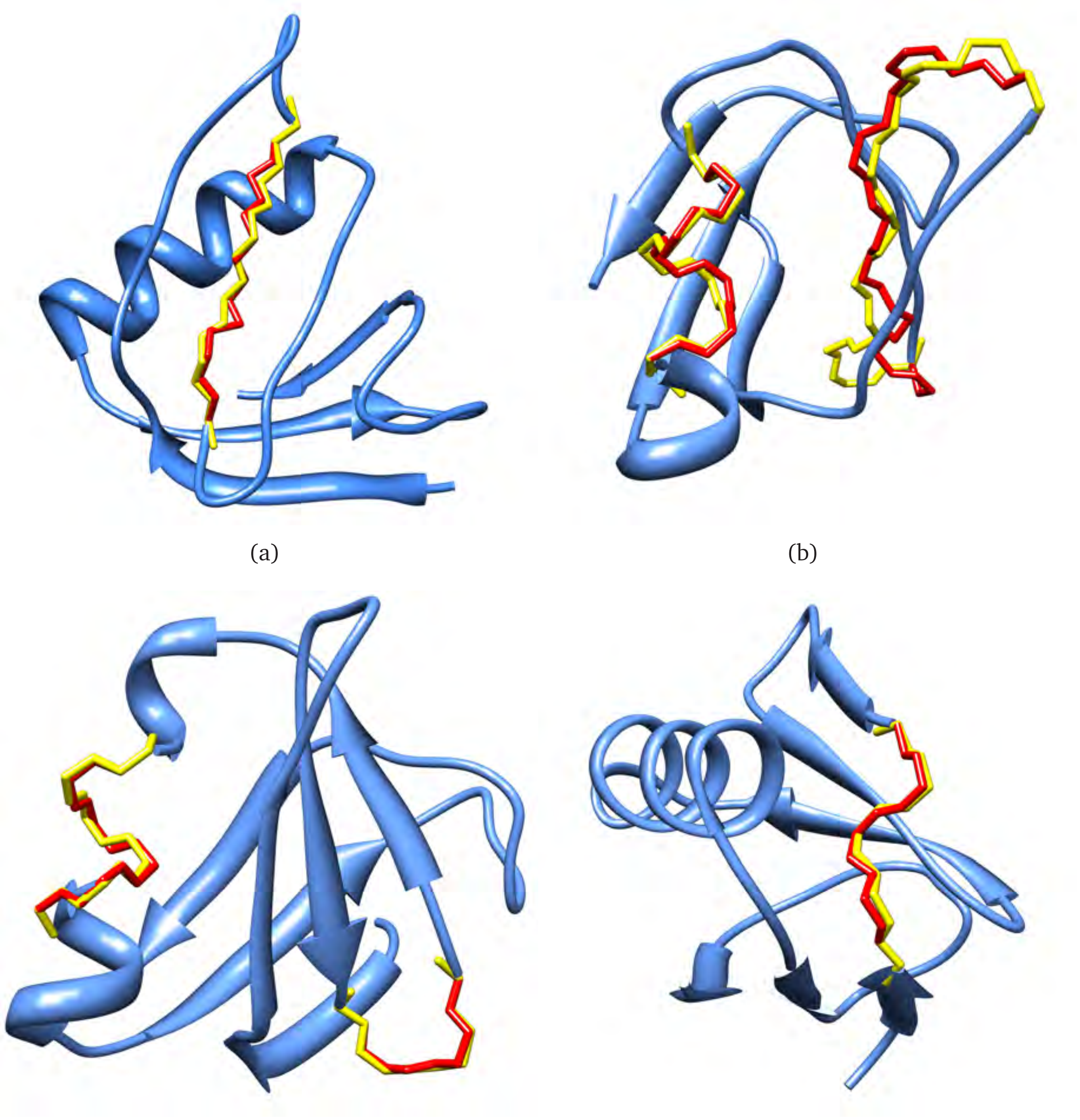
	Rigid protocol	Loose protocol
Detected by SS-docking	15 (45)	15 (45)
Beyond database limits	4 (12)	4 (12)
Physically impossible	1 (5)	1 (5)
Filtered out due to distances	7 (25)	2 (15)
Filtered out by density check	0 (0)	3 (4)
Final results	3 (3)	5 (12)

Table 4.2: Validation test of FittOFF - ‘artificial gapping’. The table shows the number of gap candidates (corresponding gap lengths given in brackets) filtered out by the applied protocols.

For testing the fitting of the fragments, the gaps obtained from the loose protocol were fed into FRAGRA. By applying the map correlation as a ranking criterion, all false-positive solutions could be eliminated, leaving only the best-ranked fragments for the expected gaps and gap lengths. To obtain a good estimate of the validity of these fitted fragments, they were superposed against the corresponding areas in the reference structure. Evidently, poorer results are achieved for long gaps (Table 4.3). However, even for larger deviations the fitted fragments follow a path very similar to the protein backbone in the reference structure (Figure 4.3).

Gap length	Secondary structure	Map correlation	$rmsd$ to reference structure	$rmsd_{adj}$
4	E	0.18	0.5 Å	0.31
9	E / C	0.10	1.8 Å	0.87
5	H	0.16	0.4 Å	0.23
2	C	0.23	0.3 Å	0.24
3	H	0.26	0.4 Å	0.28
3	E	0.21	0.4 Å	0.28

Table 4.3: $rmsd$ values of fitted fragments to reference structure. The last fragment is the best fit for the gap that could not be automatically detected due to short anchoring fragments. $rmsd_{adj}$ is the $rmsd$ scaled to the cube root of the number of aligned residues.



(a) (b)

(c) (d)

Figure 4.3: Validation test of FittOFF method with artificially "gapped" test case 1c48. Part a) to d) show the different structural gaps, with d) showing the gap that could not be automatically detected due to short anchoring fragments. Fitted fragments are shown in stick representation for the minimal backbone. The fitted fragment is colored in red, the reference structure is shown in yellow. The biggest deviation can be seen in b), for a gap including parts of a β-sheet and a loop.

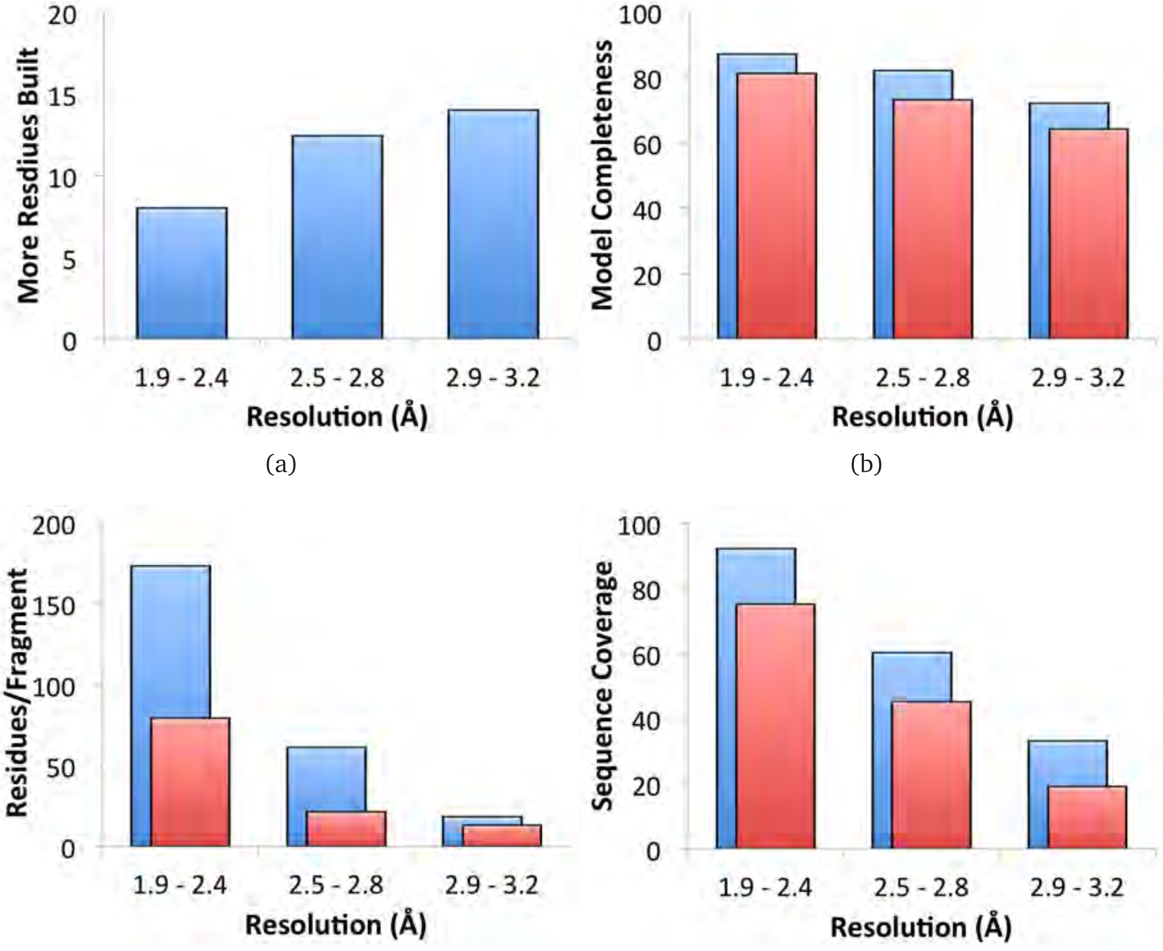
4.4 Application to *ARP/wARP* protein model building

The main application for both methods described in this thesis is improving the completeness of the built model and reducing its fragmentation at medium-to-low resolution, specifically in the *ARP/wARP* protein model building protocol. Both methods were tested with a wide range of parameters, as described below. The best results obtained and their dependence on resolution is shown. Additionally, the PNSextender protocols and results are described both as they were at the time of the *ARP/wARP* version 7.2 release to the community and as intended for the release of version 7.3.

NCS-extensions and restraints in *ARP/wARP*

For its evaluation, the PNSextender was tested on a wide range of parameters in PNSextend and PNSrestrain. The parameters included the $rmsd$ threshold, below which pairs were deemed to match, the initial length for the identification of NCS-related chain fragments and the amount of located *extensions* to be fed back into the model building process (ranked according to the weights described in sections 3.1.3 and 4.1). Moreover, an option was included to remove short matches that have been identified as helix-only using the Zhang algorithm, section 2.5.1.

In most of the tested cases, a higher number of residues built and a higher average length of protein chain fragments was observed when the PNSextender was employed. The relative improvement in model building was almost independent of the resolution of the data within the range of 2.5 - 3.8 Å. As expected, the improvement in model completeness diminished at higher resolution - between 1.9 and 2.4 Å - since the structures are already built well using the standard *ARP/wARP* protein model building protocol. Notably at any resolution, the resulting models became less fragmented which should simplify their completion by manual intervention. It was also noticed that the amount of built residues that have automatically been docked to the sequence (sequence coverage) improved in all cases. The improvements for the best cases are shown in Figure 4.4 and a detailed overview is given in Tables B.1 and B.2. There were also decreases in R-factor of up to 7.5%, increases of up to 15% in model completeness at a resolution around 3.2 Å and tripling of the average length of the resulting protein chain fragments at a resolution of 2.5 Å. On average, the length of the built fragments was more than doubled for the test cases at resolution from 1.9 to 2.8 Å (Figure 4.4c).



(a) (b)

(c) (d)

Figure 4.4: PNSextender applied to *ARP/wARP*. The best results are shown for tests with variable *rmsd* thresholds for acceptance of identified NCS matches (0.4/0.5 Å) and a variable amount of top-ranking fragments to be fed back into the model building process. The red columns denote the values obtained with the standard *ARP/wARP* model building protocol, whereas the blue columns show the best values obtained with the *ARP/wARP* incorporating the PNSextender. a) The percentage of extra residues built compared to the standard *ARP/wARP* protocol; b) Average completeness of the built model; c) Average length of built fragments and (d) Residues that have been assigned to sequence.

Release in *ARP/wARP* version 7.2

Subsequently, parameters were identified that gave the best improvement for all tested structures at their various sizes and data resolution. These parameters were used in the protocols released in *ARP/wARP* version 7.2, which was the most recent software release at the time of writing this thesis. It was observed that during protein chain tracing, smaller fragments are more likely to contain mistakes. This could be due to the connectivity and non-branching nature of the protein chain serving as an extremely powerful constraint in model building with *ARP/wARP* and helps to eliminate incorrect chain diversions. The use of small chain fragments introduces noise into the derivation of the NCS operators and smears out their clusters during identification of NCS-related copies. This in turn disturbs the ranking of the NCS matches and, in the end, it may result in incorrect *extensions* being sent back to the model building process and thus introduce additional complexity in the chain tracing procedure. To avoid such problems, the minimum number of residues of $C\alpha$ stretches used for initial least-squares superposition was set to the current average length of built chain fragments in the structure.

It was also found that a lower $rmsd$ threshold for acceptance of NCS matches provided better results at medium rather than at lower resolution as the accuracy of the matches likely correlates with the coordinate error. Thus, for data higher than 2.8 Å resolution, the threshold was set to 0.4 Å and for lower resolution to 0.5 Å. More elaborate dependencies may be sought in the future. Additionally, only a limited number of top-ranked *extensions* - typically three - are fed back into the model building process.

Overall, the use of the method with the optimised parameters applied at resolution lower than 2.4 Å results in models with 5% higher model completeness, 25% longer chain fragments and 10% higher sequence coverage than those models built without the use of the PNSextender module. A detailed overview of the results is given in Table B.3.

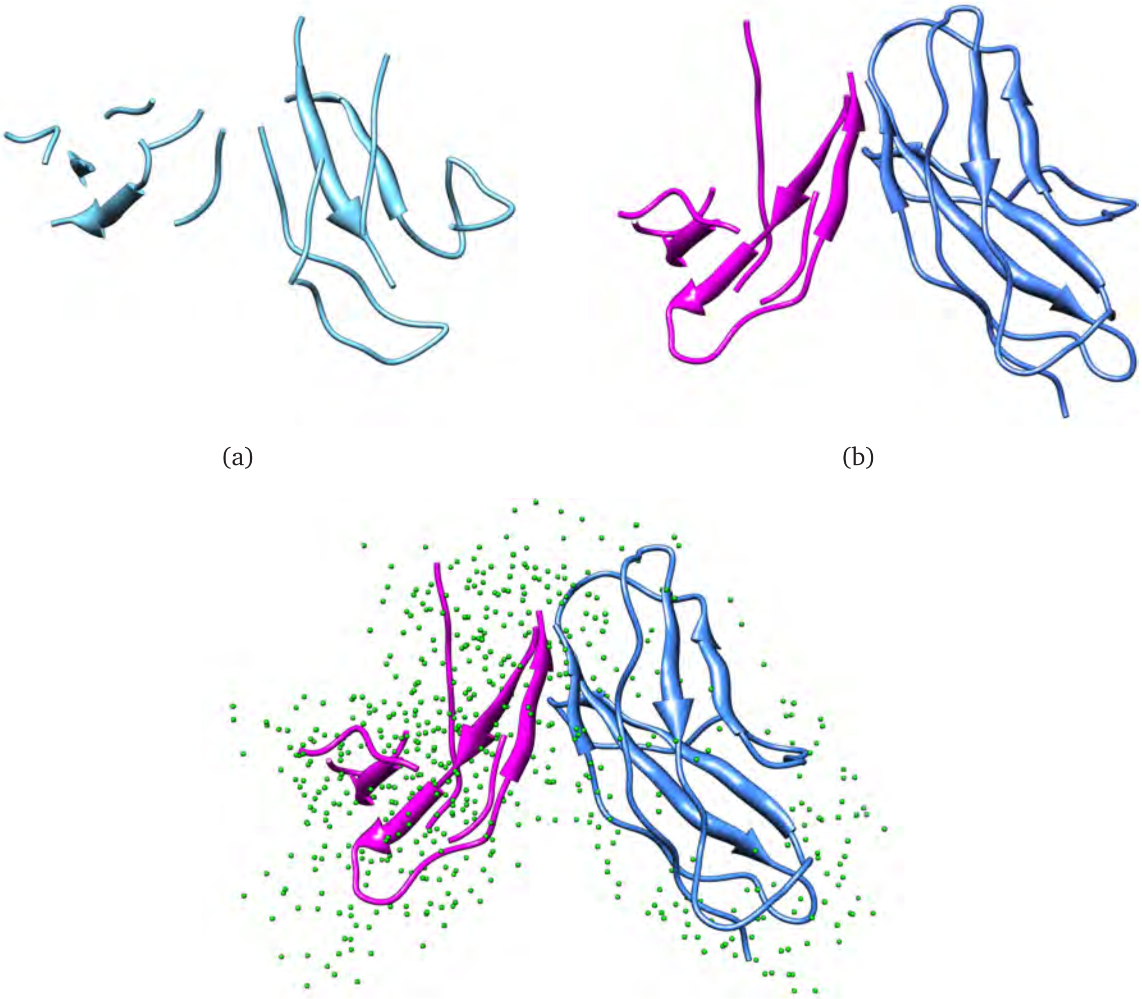
Changes for the upcoming release in *ARP/wARP* version 7.3

For the upcoming release of version 7.3 of *ARP/wARP* we decided to change the initial length for the identification of NCS-related fragments, which, in version 7.2, was set to the average length of all fragments in the intermediate model. At medium and high resolution, models of structures that could be built to a high level of completeness using standard model building protocols often did not show the amount of improvement that could be expected to result from the addition of NCS-based *extensions* and restraints to those protocols. More specifically, structures with one or more well-built NCS-related copies were not extended at all. To understand this problem, let us consider a dimer, of

which one subunit is built to a high percentage (one chain of 100 residues). The second subunit is built to a low percentage (four chains of 10 residues each), which could be accounted to varying map or phase quality. Using the average chain length, one would arrive at a value of $(100+10+10+10+10)/5=28$. Given one well-built subunit one would expect matches to the second subunit. However, there is no single pair of chain fragments with the length of 28 or more in the structure. Hence, we decided to use the 50^{th} percentile, which would set the initial length in the described example to 10 and thus permit the finding of several matches between the first and the second subunit.
This change lead to an improvement in a few cases, with the most significant one being noted for a dimeric structure of the fifth domain of human myomesin-1, mutant F700S with 196 residues and data extending to 1.95 Å [201]. The default protocol of *ARP/wARP* model building was able to deliver 42% of the structure in 11 fragments. No sequence could be docked and the R-factor was 32%. The resulting, highly-fragmented model is shown Figure 4.5a. With the PNSextender using the 50^{th} percentile as the initial length limit, it was possible to improve the model completeness by 24%, the average fragment length by 75% and the sequence coverage to 46%. The R-factor dropped to 27%. By applying a looser $rmsd$ threshold of 0.8 Å, it was possible to improve the model even further. Finally, a model completeness of 75% was achieved - with almost doubled number of residues compared to the standard ARP/ wARP protocol. The sequence coverage increased to 63%, the average number of residues per chain increased by another 62%. Overall, the use of the PNSextender increased the average chain length from 7 to 21 residues; the R-factor dropped to 26%. In the resulting model one subunit has been built completely and large parts of the second subunit have been built as well (Figure 4.5b). Areas of the second subunit that have not been built are likely to correspond to low density, since the $C\alpha$-seeds required to build these chains were generated during model building (Figure 4.5c).

Fragment fitting in *ARP/wARP*

The application of the FittOFF method to standard *ARP/wARP* protein model building was investigated using three protocols each with differing parameters. The first two protocols are the rigid and the loose ones described above in section 4.3. In these protocols, the average map correlation over all top-fitting fragments was calculated and only fragments with a map correlation higher than the average were admitted to *ARP/wARP* model building as $C\alpha$ seeds. In addition, a third protocol (also loose in regard to filtering) that fed back all fitted fragments into *ARP/wARP* was used, denoted the loosest



(a)

(b)

(c)

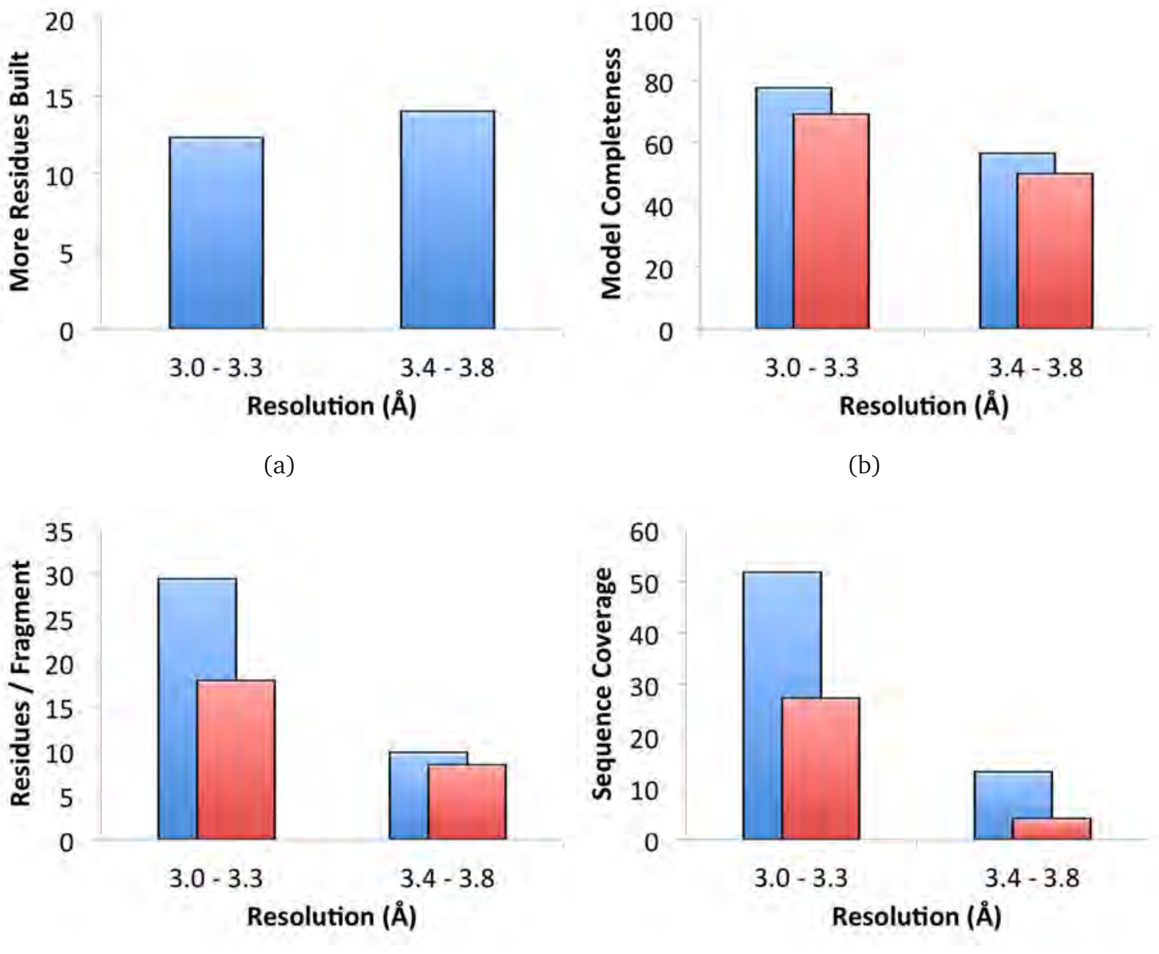
Figure 4.5: Significant improvement in structure building with PNSextender. a) shows the result of the standard *ARP/wARP* model building, b) shows the result using the amended PNSextender. c) shows the result using the amended PNSextender including all remaining ‘free atoms’ that could not be used for model building.

protocol. The test structures used are described in section 3.3. Every protocol was executed with two different secondary structure predictions, generated with PSIPRED and SSPro. Additionally, secondary structure assignments from the Zhang methodology were also tested.
Similarly to the results for the PNSextender, higher model completeness with less fragments was observed in almost all cases. The improvement for the best cases is shown in Figure 4.6. A detailed overview is given in Tables B.4 and B.5. Up to 25% more residues could be built for resolution as low as 3.8 Å (Figure 4.6a and 4.6b). In addition, decreases in R-factor by 4% and doubling of the average length of fragments were observed in several cases (Figure 4.6c); improvement in sequence coverage of up to 50% (resulting in doubled and tripled sequence coverage, Figure 4.6d) was apparent in the best cases. The best results for extra residues and sequence coverage were obtained following the application of the loose protocol, although the best improvement in fragmentation was seen following application of the rigid protocol as shown in Table 4.4. Regarding the source of secondary structure information, it was found that the prediction methods delivered results comparable to those obtained with the Zhang assignment. Nevertheless, as might be expected, the best results were achieved using the Zhang assignment (Table 4.5). No apparent relation was detected between the secondary structure content of a structure and the extent of improvement.

Protocol	Residues built	Model completeness	Fragment length	Sequence coverage	R-factor
Rigid	+18.8%	+11.0%	+98.2%	+27.2%	-4.3
Loose	+25.3%	+11.7%	+78.1%	+51.9%	-3.8
Loosest	+20.5%	+11.0%	+78.1%	+35.2%	-3.8
Overall best	+25.3%	+11.7%	+98.2%	+59.1%	-4.3

Table 4.4: Comparison of the best result obtained by model building with FittOFF for each of three different protocols for gap filtering.

For testing of FittOFF, only cases from the PDB were used. This made it was possible to compare the models built by the standard *ARP/wARP* protocol and that incorporating the FittOFF method to the crystal structure from the PDB. The results for the superpositions of three models are shown in Table 4.6. It is shown that all models built to a higher extent with FittOFF also had a smaller $rmsd$ and $rmsd_{adj}$ ($rmsd$ scaled to the cube root of the number of aligned residues) to the reference structure. In the case of 2aj2, it was even possible to obtain an R-factor lower than the one for the structure from the PDB (24.9% compared to 25.8%).



(a) (b)

(c) (d)

Figure 4.6: FittOFF applied to *ARP/wARP*. The best results are shown for tests with the different protocols and sources of secondary structure information described in section 4.4. The red columns denote the values obtained with the standard *ARP/wARP* model building protocol, whereas the blue columns show the best values obtained with the *ARP/wARP* incorporating FittOFF. a) The percentage of extra residues built compared to the standard *ARP/wARP* protocol; b) Average completeness of the built model; c) Average length of built fragments and (d) Residues that have been assigned to sequence. Better results with the standard *ARP/wARP* protocol as compared to the tests of the PNSextender (Figure 4.4) originate from the smaller structures used as test cases for FittOFF.

Secondary Structure source	Residues built	Model completeness	Fragment length	Sequence coverage	R-factor
PSIPRED	+18.1%	+11.0%	+78.1%	+27.8%	-3.8
SSPro	+18.8%	+10.1%	+60.0%	+52.9%	-3.4
Zhang assignment	+25.3%	+11.7%	+98.2%	+29.8%	-4.3
Overall best	+25.3%	+11.7%	+98.2%	+59.1%	-4.3

Table 4.5: Comparison of the best result obtained by model building with FittOFF for each of three different sources of secondary structure information.

Testcase	Size	Resolution	Aligned residues	$rmsd$	$rmsd_{adj}$
1plr	258	3.0 Å	225 / 240	0.80 Å / 0.75 Å	0.13 / 0.12
2qsr	173	3.1 Å	119 / 138	0.93 Å / 0.84 Å	0.19 / 0.16
2aj2	165	3.2 Å	144 / 165	0.67 Å / 0.67 Å	0.13 / 0.12

Table 4.6: Best results obtained with FittOFF; results of the structural alignments of the model built with the standard *ARP/wARP* protocol and the protocol incorporating FittOFF and the reference structure. The $rmsd$ has been calculated over all $C\alpha$ atoms in each model.

Chapter

5

Discussion

In the test scenarios described in section 4.3, all chain fragments are free of phase-dependent coordinate error. There are also no mistakes in traced chain fragments such as route shortcuts or spurious loops. Moreover, the electron density and secondary structure information used for testing FittOFF were of high quality, meaning the test cases were somewhat idealised (the best case scenario). In case of NCS, there may however be inherent differences between NCS-related parts of the structure as there are between chain E and all other chains in the model of mutated shiga-like toxin B-subunit. Indeed, the NCS operators are rarely exact across all copies of a fragment [202, 203]. This also was one reason to implement somewhat loose $rmsd$ thresholds for the initial identification of NCS-relations. Nevertheless, in the exclusion test, the retrieval of the full pentameric structure to a very high accuracy was possible even when the initial model was highly fragmented and contained only 65% of its Cα atoms, Table 4.1. It is thus estimated that in the best-case scenario - in which all structural information is available, NCS matches are accurate and there are no coordinate errors - it may become possible to retrieve the full structure of a protein at 3.5 Å resolution in a single building cycle, even with the current performance of the *ARP/wARP* protein model building module. It was also shown that with FittOFF, all gaps surrounded by fragments docked into the secondary structure could be identified without introducing any false-positive gaps. Furthermore, wrongly recognised gap lengths could be eliminated using a threshold applied to the map correlation of the fitted fragment to the residual density. Although the deviation to the reference structure rises with longer fitted fragments, they are generally highly similar to the path taken by the protein backbone.

The way the results of both methods are used in the protein model building protocol has certain advantages over other possible approaches. For example, plain averaging of the coordinates of the *extensions* derived from NCS-relations or the fragments fitted into identified structural gaps may not be the best option as it introduces a certain degree of model bias. It may also move some parts of the averaged model out of the density. In our implementation all *extensions* of fitted fragments are only used as potential Cα seeds (suggestions) to *ARP/wARP* for subsequent building of longer chain fragments. Therefore, the method is not expected to build parts of the structure that lack support for coordinate placement in terms of electron density and plausible stereochemistry. If the additional $C\alpha$ atoms admitted to further chain tracing by the PNSextender or FittOFF are in agreement with the density, longer chains will be built. If, on the contrary, the suggestions do not match the density they will not be used for building a chain. This is especially important for the additional information derived from FittOFF, which has been shown to be less accurate for longer gaps. Seeds that deviate too much from the density would simply be ignored by the chain tracing module and only those parts of the fitted fragments that are in agreement with the electron density will be used.

There may still be small decreases in model completeness or higher fragmentation for some cases. This may occur for models which are built only to a modest extent and can be caused by different paths that will be followed during chain tracing. Taking another (incorrect) path could always lead the tracing to areas of low density and thus the building of shorter chains.

The filtering of structural gaps obtained from secondary structure docking with the methods described in section 3.2.1 has many important features. Considering the use of FittOFF on structures containing NCS, it is to be expected that fragments from different, spatially divided subunits, may be docked closely to each other in the sequence, suggesting a structural gap between them. However, evaluating the gap distance between the given stem residues shows that there is no relation to the proposed number of missing residues. For the case that two fragments from NCS-related subunits are close to each other in space, the check for density between their stem residues should filter this gap out if there is not enough density to support it.

The use of secondary structure predictions in FittOFF does not lead to significantly poorer results compared to the use of the Zhang assignment. However, as would be expected, the best improvement has been achieved using the Zhang assignment (Table 4.5). As mentioned in section 3.2.1, the sequence docking algorithm used by *ARP/wARP* is unlikely to produce valid results at a resolution used for the tests of FittOFF. Thus, the improvement of the number of sequence-assigned residues in the final model obtained when using SSpro (Table 4.5) mainly results in incorrectly docked residues. In the considered case only one fifth of all sequence-assigned residues were docked correctly. Actually, in all observed cases at least half of the sequence-assigned residues are docked incorrectly. It might thus be best to abandon the improvement of sequence coverage as a criterion for future testing of the effectiveness of any method for test cases at resolution lower than 2.5 Å.

The most favoured scenario regarding the use of secondary structure information would be one reviewed by the crystallographer and comprising all information gathered so far for domains or subunits of the considered structure under the given conditions. If for some domains, the atomic coordinates are already known, their secondary structure should ideally be derived using the Zhang algorithm. However, there is no agreed definition of a perfect secondary structure assignment. Even when the coordinates are known, the average agreement between all established secondary structure assignment methods is only 85% [177].

Fitting fragments into long gaps with FittOFF might result in only marginally reliable fragments, which will certainly not lead to *ARP/wARP* building a connection between the anchoring fragments. However, the host fragments may be partially extended, thus leading to a shorter gap and result in a more reliably fitted fragment in the next iteration. This is supported by the results obtained for the "loose" protocols in section 4.4. Moreover, we found that in more than 70% of the tests, there were no differences between the loose protocol with and without the application of the map correlation threshold for fragment elimination. This again proves that admitting more, probably unreliable fragments (or $C\alpha$ seeds) into the *ARP/wARP* model building process does not necessarily result in poorer results, since the wrong seeds will plainly be ignored. Notwithstanding, an evaluation and elimination of the prospective gaps and gap lengths using the map correlation threshold is beneficial. As described in section 3.5, the use of FRAGRA introduces the most noticeable computational overhead compared to standard *ARP/wARP* model building. Thus evaluating and admitting less gaps to FRAGRA would result in a reduction of computation time, although with a potential trade-off against some improvement in model completeness.

The accuracy of both methods depends predominantly on the degree of fragmentation of the initial model and its coordinate accuracy. For NCS extension, a more complete initial model will yield better results (i.e. many residues missing but one subunit with NCS-relations built to a significant extent), whereas for the identification of structural gaps between chain fragments with FittOFF, correctly built, long chain fragments with defined secondary structure are of benefit. The use of NCS extension and fragment fitting in model building (at least in the current implementation of the *ARP/wARP* protein chain tracing) is always advantageous, but the degree of improvement depends even more strongly on the completeness, fragmentation and correctness of the model, which all in turn depend on the quality of the initial phases and the data. More specifically, for a model consisting only of chain fragments shorter than seven residues, both methods cannot provide any further improvement: In FittOFF no chain fragments could be docked to the secondary structure prediction and the PNSextender would not find meaningful *extensions* or restraints. Additionally, for such a model, there is also a high probability that most, if not all, chain fragments are modelled incorrectly or with high positional error.

Another problem will arise at resolution lower than 5.0 Å. In this resolution regime, it is impossible to reliably detect all $C\alpha$ atoms required to build the protein backbone. This can be accounted to the pattern recognition approach currently implemented in *ARP/wARP* that uses the expected $C\alpha$-$C\alpha$ distance (3.8 Å) to define which free atoms should be considered as candidate $C\alpha$ atoms. Furthermore, the most commonly used refinement restraints are bonded and angle-bonded distances or planarity restraints that span moderate distances between 2.2 Å and 3.8 Å [204]. Such pattern recognition approaches and refinement restraints are sufficient to aid model building at a resolution of 1 to 4 Å, since here a matching grid of placed information is provided. In other words, restraints or patterns corresponding to a distance smaller than the smallest spacing of the experimental data will not be seen (directly) and thus may not be helpful. Hence, while modelling of a structure at 5 Å resolution there is little sense to introduce restraints and patterns between adjacent $C\alpha$ atoms. However, it would be possible to follow the idea of approaches like RESOLVE [42]. As described in section 1.3.1, a combination of placing ideal α-helix and β-sheet fragments into the electron density and connecting them using a database of short fragments should work to a certain extent even at resolution below 5 Å; obviously, the resulting model would be biased, at least to a certain degree, towards the ideal fragments. Such an approach would also break down as soon as the resolution becomes too low to recognise secondary structure elements, which is about 8 to 10 Å (section 1.4). One could develop this idea even further. Once the secondary structure cannot be recognised anymore, fragments to be placed in the density could comprise ideal super secondary structure elements, for example, β-hairpins [205] or β-barrels [206, 207]. At even lower resolution, 15-20 Å, automated model building could be realised by placing known substructures or domains into the electron density [208]. All this would ultimately lead to an automatic modelling of large molecular machines or even cellular compartments, which is an application envisaged with the newly developed X-ray free electron lasers (FEL) [209]. However, the amount of structural information from databases like the PDB is currently insufficient to ensure the building of models that are free of bias with such approaches.

5.1 Considerations for further research

There is still a wide array of conditions which should be investigated in order to make the most from the identification of NCS-relations, filtering and identification of structural gaps and the way fragments are selected to be fitted in FRAGRA.

As described in section 4.4, the best initial length for the identification of NCS-related fragments is still being investigated. Further approaches could be based on enhancing the clustering of the transformations between chain fragments (section 3.1.1). To obtain a more accurate clustering, the translation vector could be used as a clustering criterion in addition to the rotational component between two transformations. This would give rise to clusters that draw a clearer picture of the relation between two NCS-related subunits, but would also imply the need for a complete rethinking of the weighting scheme applied to NCS-extensions. To filter out clusters that do not denote NCS-relations, a list of allowed angles between NCS-related subunits of the structure under consideration could prove to be very beneficial. This list could be derived by applying the self-rotation function [34].

In the current implementation, only NCS-based refinement restraints of high confidence are admitted to REFMAC. These are limited to NCS-relations between fragments docked into sequence by *ARP/wARP* with a length of at least 15 residues. This is based on the fact that, contrary to the *extensions* used for model improvement, there is no further evaluation of the quality of restraints, meaning every restraint generated will be used in REFMAC. A quality criterion for the acceptance of an NCS-relation to be used as refinement restraint could use a list of allowed angles between NCS-related subunits, as described above. Another more technical problem is related to the way *ARP/wARP* labels undocked fragments and the input format for NCS-relations required by REFMAC. Currently, ARP/ wARP assigns all undocked fragments to one chain (Q) with continuous residue numbers. REFMAC on the other hand requires the input format "A B 10 35" for NCS-relations, with A and B being the chains and 10 to 35 the residues in these chains related by NCS. Thus, an NCS-relation, even with high confidence, between residues 10 to 45 of chain C and residues 56 to 81 of chain Q cannot be used, since the residues numbers have to be the same in both chains. There are two solutions to this problem: either one has to write some code that offers a possibility of a smart change of residue numbers (and chain identifiers) of the intermediate model in a suitable manner to permit the input of identified NCS relationships to REFMAC or one would have to persuade the REFMAC-developers to change the input format. The latter may become a reality in the light of coming acceptance of the mmCIF format in MX.

Although the PNSextender has been developed for proteins, the symmetric nature of complementary strands in DNA calls for an investigation of its applicability to model building of poly-nucleotide structures. Likewise, the PNSextender methods should be applied to structures with high symmetries in their subunits, such as repetitions of secondary or super-secondary structure motifs (beta-hairpins, helix-turn-helix, etc.). An even more ambitious step could be the extension of incomplete helices and sheets using ideal conformations. However, if this path is to be followed, more rigid $rmsd$ thresholds will have to be applied to avoid errors and false positives.

The next step in the development of the FittOFF method must be its integration and release within the next version of the *ARP/wARP* software suite. For this a a number technical issues have to be resolved. Firstly, the current implementation of FittOFF requires OpenStructure to be installed, which may not be straightforward for a non-expert user. It should also be investigated whether the currently used fragment database in FRAGRA, which has a size of around 600 MB, can be further scaled-down or compressed by about an order of magnitude.
Once a convenient incorporation into *ARP/wARP* has been accomplished, it will also be necessary to evaluate protocols combining fragment fitting and NCS-extension on an array of test cases containing NCS-relations to see if their combined effectiveness is more than the effects of either addition singly.
Considering the method itself, there are a few areas with room for improvements. One shortcoming might be the use of only one stem residue from each fragment anchoring the structural gap for the backbone sampling (section 3.2.2), since the terminal ends of chain fragments are often built by *ARP/wARP* with significant positional deviation. The use of only a single stem residue on either side of the gap permits the longest fragments to be built, although perhaps to a smaller degree of accuracy as may be the case if more residues were used. If one could settle for the maximum length of fragments to be limited to ten residues, it would be possible to take three stem residues on each side into account, meaning fragments would be fitted to more reliable parts of the intermediate model from *ARP/wARP*.
An improvement of the secondary structure docking could be achieved by using a confidence score for each docked residue. PSIPRED gives confidence values between one and nine for each predicted secondary structure element. A similar confidence score can be derived for secondary structure state assigned by the Zhang algorithm by evaluating the amount of $d_{i,j+k}$ that satisfy either λ_k^α or λ_k^β in Eq. 2.29. Accounting for these scores in secondary structure docking should improve the accuracy of the method by providing a more valid positioning of the chain fragments.
As it has been shown in section 4.3, it is possible to deduce the correct structural gaps and number of missing residues using the rigid protocol in FittOFF. This requires that

the fragments anchoring these gaps are docked into the sequence at the correct positions. Hence, the secondary structure docking should also be tested for its application to aid the sequence assignment in *ARP/wARP*, especially at medium-to-low resolution. Ambiguous dockings could be further improved by a combination of secondary structure docking and identification of large side chains in the density.

The placing of nine tubes for evaluating the density between two stem residues (section 3.2.1) delivers good results. However a possible improvement in the evaluation should be investigated, since the density of kinked structural elements will be missed in the current 'straight-line' implementation. The use of spheres might be more accurate, although they might include the density from neighbouring fragments. A better strategy may involve the division of the area between the anchors into several slices. Evaluating the density of each slice would allow the generation of histograms of density content. This would enable us to make assumptions about the secondary structure in the gap based on the density, hence gaining an even better understanding of the number of residues missing. However, the FRAGRA method already provides the best fragment for the residual density. Hence, the main application of such an improved density evaluation would be the identification of additional structural gaps that would be missed otherwise.

Another approach, involving the placement of all protein atoms as seeds for model building as opposed to merely $C\alpha$-candidates, warrants investigation although the incorporation of such an approach into *ARP/wARP* would require a complete overhaul of the PNSextender and FittOFF modules. For the PNSextender, a mode could be implemented in which "NCS-copy-paste" is applied to the whole fragment. Here, instead of using only the $C\alpha$ atoms as seeds, all copied backbone atoms could be directly admitted to the *ARP/wARP* model update. If they are not supported by the density or cause steric clashes, they would be deleted anyway. An analogous approach could be implemented with FittOFF, applied after *ARP/wARP*s final model building cycle. Following the final execution of *Loopy*, fitted fragments could be used to build very difficult loops and kept in the model as described above. Although such an approach would not be appropriate with regard to proper statistical validation of the data, as fitted residues would instantly be admitted to REFMAC, without any validation of their appropriateness, this is actually the procedure undertaken by the *Loopy* module. It thus seems that such a violation can be justified if the result is a better model. In this regard one could even go so far as to investigate the substitution of *Loopy* by FittOFF.

In section 4.4 and Table B.3, it was shown that the fixed protocol for the PNSextender gives improvements for all cases with resolution of the data worse than 2.4 Å. However, these improvements are less impressive compared to the best ones obtained for variable

parameters (Table B.1). This can be accounted to the fact that each test case gave the highest improvements for its ‘individual’ set of parameters (of those described in section 4.4).
If large amounts of computer time would be available, for example by using large-scale computer clusters, the strategy should be to model each structure with all possible combinations of parameters. More precisely, a model building job would be executed for each combination of parameters. The best model could then be chosen from the pool of solutions according to the best model completeness, lowest degree of fragmentation, the best R-factor or a combination of these values. Thus, for every structure the best model that could possibly be obtained with the application of NCS extension and restraints would be found. Such a brute-force approach would be similar to ARCIMBOLDO [78], which also generates several thousand models for each structure (section 1.3.1). Similar to the current implementation of ARCIMBOLDO, the computation time would be immense. For just five different $rmsd$-thresholds, five different initial lengths for the identification of NCS-relations and five different numbers of generated *extensions* to be fed back into the model building process one would arrive at 5^3 model building jobs, which, even for a modest-size structure, would translate to a requirement of 125 CPU hours. A similar approach could be designed for FittOFF, where all reasonable values for $conf_{dock}$ and $conf_{pvec}$, described in section 3.2.1, could be combined with map correlation thresholds for admitting the fitted fragments or not.

Clearly, substantially modified or even completely different approaches will be required for model building at resolution of lower than 5 Å, where adjacent $C\alpha$ atoms cannot be resolved anymore, since any information obtained from the experimental data will be placed on a grid at least 5 Å apart from each other. Building models at such resolution with the current experimental methods would always include a trade-off to a certain degree of bias and thus losses in the uniqueness of the resulting structural model.

5.2 Conclusion

The obtained results support the general benefit of the combination of intrinsic information (NCS-based extension and refinement restraints) with, or the application of methods from theoretical modelling (fitted fragments) to, automatic protein model building in macromolecular crystallography. For the PNSextender, a protocol has been developed that provides notable improvements within the resolution range from 1.9 to 3.2 Å. Especially at resolution around 3.1-3.2 Å, the use of the method gave rise to a 20% increase in the length of the built chain fragments; their length was typically higher than 10 residues - the value sometimes quoted as an indicator of a ‘good’ model. Even at higher resolution, 1.9 to 2.4 Å, the method gives significant improvement in terms of fragmentation and sequence coverage. The application of the FittOFF method showed notable results at resolution between 3.0 and 3.8 Å, pushing model building for some case studies towards 80% completeness and a significantly better $rmsd$ to the crystal structure from the PDB (Tables 4.6). Importantly, both methods impose negligible overhead on the computation time required by standard *ARP/wARP* protein model building protocol and are thus applicable for the general use (section 3.5). Further optimisation of the parameters specific to each method will certainly provide additional enhancement (as was already shown in section 4.4). Continuous evaluation of the methods on a wide variety of cases will be performed automatically in the future due to the invocation of the PNSextender in the *ARP/wARP* web-based model building (as of version 7.2). A similar approach will be taken for the FittOFF method once it has been released within the *ARP/wARP* software suite.

The developments presented in this thesis will create the capability not only of solving structures at a higher rate, but also of producing higher-quality structural results, especially for challenging structures with data to limited resolution. A major deliverable of this work is a provision of the developed software to world-wide user community. The software will allow achieving increasing levels of automation and implementing “smart” structure determination protocols capable of delivering expert-quality results to non-expert users. These developments have the potential to find their use in fields as diverse as biochemistry, medicine, bioinformatics and computational drug design to name just a few.

5.2.1 Availability to the community

The PNSextender has been incorporated into the *ARP/wARP* software project (from version 7.2 onwards), the software is available from http://www.arp-warp.org. The method, which has been published [1] in a peer-reviewed journal, has also been presented at several conferences and workshops in terms of oral and poster presentations. The FittOFF method will be incorporated into a future release of *ARP/wARP*. However, preliminary results have already been presented at conferences and workshops, and the publication in a peer reviewed journal is intended after the submission of this thesis.

Chapter

6

Zusammenfassung

Das Ziel der Makromolekularen Röntgenbeugung (MX) ist die Bestimmung der dreidimensionalen Strukturen von Molekülen. Eine besondere Herausforderung stellt die Strukturbestimmung von grossen Makromolekülen und deren Komplexen dar, wel- che bislang oft gar nicht möglich oder mit grossem Aufwand verbunden ist. Das Hauptproblem liegt darin, dass für die Kristalle solcher Moleküle während eines Diffraktionsexperimentes nur selten Daten mit hoher Auflösung gemessen werden können. Das Ergebnis sind oft verrauschte und ungenaue Elektronendichtekarten. Ein weiteres Problem liegt darin, dass die bislang entwickelte Software für automatische Modellierung in MX weitgehend auf hochaufgelöste Daten ausgelegt ist. Es ist zwar möglich diese auf niedrigaufgelöste Daten (unter 3.0 Å) anzuwenden, die resultierenden Strukturmodelle sind jedoch meist unvollständig und stark fragmentiert. Es besteht also der dringende Bedarf für robuste und effiziente Methoden, welche die Vollständigkeit und Genauigkeit von niedrigaufgelösten Strukturmodellen verbessern.

In dieser Dissertation werden zwei Methoden vorgestellt, welche die Qualität von Strukturmodellen basierend auf niedrigaufgelösten Daten deutlich verbessern. Hierfür werden vorhandene Informationen, die entweder intrinsisch, also in den zu analysierenden Daten bereits enthalten, oder komplementär, aus Datenbanken gewonnen, genutzt. Die erste Methode basiert darauf, dass viele Makromoleküle multiple Kopien ihrer Teilstrukturen in der asymmetrischen Einheit aufweisen. Im Jahr 2012 beinhalteten mehr als 50% aller Kristallstrukturen in der Proteindatenbank (PDB) jene sogenannte Nichtkristalline Symmetrie (NCS). Bei der automatischen Modellierung in *ARP/wARP* werden diese NCS-Teilstrukturen selten im gleichen Umfang rekonstruiert, insbesondere in den anfänglichen Zyklen. Die Gründe hierfür können von limitierter Auflösung bis hin zu schlechten initialen Phasen reichen. Die Tatsache, dass NCS-Teilstrukturen zu unterschiedlichen Graden modelliert werden, hat den Vorteil, dass jede dieser Teilstrukturen Informationen beinhalten kann die in einer anderen fehlen. Die Kombination dieser (intrinsischen) Informationen führt zu einer Verbesserung der Vollständigkeit der resultierenden Strukturmodelle, besonders wenn Daten mit niedriger Auflösung zu Grunde liegen.

Die Fragmentierung von Strukturmodellen, basierend auf niedrigaufgelösten Daten, beruht auf der oft nicht ausreichenden Qualität der Elektronendichte um Peptide eindeutig zu erkennen, und somit eine kontinuierliche Proteinkette aufbauen zu können. Insbesondere zu Beginn der automatischen Modellierung betrifft dies nicht nur Loops, sondern auch Helices oder Faltblätter. In der zweiten Methode, die im Zuge dieser Dissertation vorgestellt wird, werden diese strukturellen Lücken mit Strukturfragmenten aus der PDB aufgefüllt. Hierfür ist eine Verbindung der richtigen Fragmente essentiell. Zur Identifikation der zu verbindenden Ankergruppen werden hier zwei Ansätze kombiniert: Zum einen das Docken von Fragmenten in eine Sekundärstrukturvorhersage und zum anderen statistische Relationen zwischen der Distanz der ankernden Fragmenten zueinander und der Anzahl der fehlenden Residuen in einer strukturellen Lücke.

Die beiden im Rahmen dieser Dissertation entwickelten, neuen Methoden wurden in das *ARP/wARP* Proteinmodellierungsprotokoll integriert. Der Protein NCS-basierte Struktur (PNS) Extender, identifiziert NCS-Relationen automatisch und nutzt diese für die Komplettierung von Strukturmodellen und als Restraints für das Strukturrefinement.

FittOFF (Fitten von Fragmenten) identifiziert strukturelle Lücken in unvollständigen Strukturmodellen und füllt diese mit Strukturfragmenten aus der PDB auf.

Durch die Integration beider Methoden in die *ARP/wARP* Proteinmodellierung werden signifikante Verbesserungen erzielt. Der PNSextender ist in der Lage die Vollständigkeit von Strukturmodellen bei Auflösungen um 3.2 Å von 56% auf 72% zu verbessern. Des weiteren sind die resultierenden Strukturmodelle weniger fragmentiert und deutlich mehr Seitenketten werden erkannt. Mit FittOFF wird die Vollständigkeit von Strukturmodellen um bis zu 12% erhöht und die durchschnitte Länge aller Fragmente verdoppelt.

Chapter

7

Summary

Determining the three-dimensional structures of large molecular assemblies is a challenging task in macromolecular X-ray crystallography (MX). Crystals of such molecules rarely diffract to high resolution. Often only noisy and inaccurate electron density maps can be obtained. Computational approaches for model building in MX have historically been focused on high-resolution data. Thus their application to data extending to lower than 3.0 Å resolution is limited and typically results in incomplete and highly fragmented models. Hence, robust and fast methods that improve the completeness and the accuracy of models obtained from automated crystallographic model building routines are urgently needed, particularly to aid solution of low-resolution MX structures.

In this thesis, this challenge has been addressed by the development of two approaches that use intrinsic information, which is already encoded in the model, and complementary information derived from structural databases. The first one exploits the fact that a significant proportion of crystal structures contain multiple copies of subunits or their assemblies in the asymmetric unit; based on the current content of the Protein Databank, more than 50% of structures contain such non-crystallographic symmetry (NCS). It was noticed that during automated model building with *ARP/wARP*, particularly in its initial steps, NCS-related parts of the structure are often built to different extents. The reasons for that are manifold and include limited resolution of the data and poor initial phases. However, this also has a beneficial side effect. Each NCS-related copy can provide information that is not present in another one; combining this (intrinsic) information helps to advance the model building process and significantly increases the overall completeness of built structures, especially with low-resolution data.

Often, the density between two built chain fragments is too poorly defined to be interpreted as part of a protein chain. Especially in the early stages of model building, this is the case for not only loops but also helices or strands. A method is introduced to fill these structural gaps with structural fragments from the PDB. It makes use of secondary structure predictions and statistical descriptions of the relationship between gap size and and the number of missing residues to identify connectable chains fragments.

The two novel methods that were developed in this thesis have been integrated into the *ARP/wARP* protein model building; the Protein NCS-based Structure (PNS) extender for using automatically detected NCS-relations for model extension and restraints in structure refinement and FittOFF (Fitting OF Fragments) for identifying structural gaps and filling them with fragments from the PDB. The application of both methods during model building with *ARP/wARP* provides a significant improvement. In the best case for the PNSextender, model completeness improves from 56% to 72% at 3.2 Å resolution. Additionally, more side chains are docked in sequence, and the length of the built fragments increases. For FittOFF, a noticeable increase in model completeness of up to 12% and doubling of the average fragment length was observed.

Bibliography

[1] Wiegels, T. & Lamzin, V. S. Use of non-crystallographic symmetry for automated model building at medium to low resolution. *Acta Crystallogr D Biol Crystallogr* **68** 446–453 (2012). 3, 16, 26, 39, 74, 103

[2] Hazledine, S. *et al.* Release 7.2 of *ARP/wARP* software suite. *Acta Crystallogr A* **67** C135 (2011). 3

[3] Wiegels, T., Biasini, M. & Lamzin, V. S. Towards more complete models in macromolecular crystal structure determination. *Acta Crystallogr A* **67** C592 (2011). 4, 16

[4] Pettersen, E. *et al.* UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **25** 1605–1612 (2004). 4

[5] Berman, H. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28** 235–242 (2000). 11

[6] Rose, P. *et al.* The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* **39** D392–D401 (2011). 11

[7] Bieri, M. *et al.* Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. *FEBS J* **278** 704–15 (2011). 11

[8] Zhou, Z. H. Atomic resolution cryo electron microscopy of macromolecular complexes. *Adv Protein Chem Struct Biol* **82** 1–35 (2011). 11

[9] Yahav, T., Maimon, T., Grossman, E., Dahan, I. & Medalia, O. Cryo-electron tomography: gaining insight into cellular processes by structural approaches. *Curr Opin Struct Biol* **21** 670–7 (2011). 11

[10] Zewail, A. H. 4D ultrafast electron diffraction, crystallography, and microscopy. *Annu Rev Phys Chem* **57** 65–103 (2006). 11

[11] Kovalevsky, A. *et al.* Macromolecular neutron crystallography at the Protein

Crystallography Station (PCS). *Acta Crystallogr D Biol Crystallogr* **66** 1206–12 (2010). 11

[12] Mertens, H. D. T. & Svergun, D. I. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J Struct Biol* **172** 128–41 (2010). 11

[13] Leach, A. R. Molecular modelling: principles and applications (Prentice Hall, Harlow, England, 2001), 2nd edn. 11

[14] Berg, J. M., Tymoczko, J. L., Stryer, L. & Stryer, L. Biochemistry (W.H. Freeman, New York, 2002), 5th edn. 12

[15] Koolman, J. & Röhm, K.-H. Color atlas of biochemistry. Flexibooks (Thieme, Stuttgart, 2005), 2nd edn. 13, 49

[16] Drenth, J. & Mesters, J. Principles of protein x-ray crystallography (Springer, New York, 2007), 3rd edn. 13

[17] Ooi, L.-l. Principles of x-ray crystallography (Oxford University Press, Oxford, 2010). 13

[18] Taylor, G. The phase problem. *Acta Crystallogr D Biol Crystallogr* **59** 1881–90 (2003). 13

[19] Foadi, J. *et al.* A flexible and efficient procedure for the solution and phase refinement of protein structures. *Acta Crystallogr D Biol Crystallogr* **56** 1137–47 (2000). 14

[20] Xu, H. & Weeks, C. M. Rapid and automated substructure solution by Shake-and-Bake. *Acta Crystallogr D Biol Crystallogr* **64** 172–7 (2008). 14

[21] Karle, J. & Hauptman, H. A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Crystallogr A* **9** 635–651 (1956). 14

[22] Weeks, C. & Miller, R. Optimizing Shake-and-Bake for proteins. *Acta Crystallogr D Biol Crystallogr* **55** 492–500 (1999). 14

[23] Blow, D. M. & Crick, F. H. C. The treatment of errors in the isomorphous replacement method. *Acta Crystallogr A* **12** 794–802 (1959). 14

[24] Blow, D. M. & Rossmann, M. G. The single isomorphous replacement method. *Acta Crystallogr A* **14** 1195–1202 (1961). 14

[25] Harker, D. The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. *Acta Crystallogr A* **9** 1–9 (1956). 14

[26] Dauter, Z., Dauter, M. & Dodson, E. Jolly SAD. *Acta Crystallogr D Biol Crystallogr* **58** 494–506 (2002). 14

[27] Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254** 51–58 (1991). 14

[28] Rossmann, M. G. The position of anomalous scatterers in protein crystals. *Acta Crystallogr A* **14** 383–388 (1961). 14

[29] North, A. C. T. The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions. *Acta Crystallogr A* **18** 212–216 (1965). 14

[30] Mathews, B. W. The determination of the position of anomalously scattering

heavy atom groups in protein crystals. *Acta Crystallogr A* **20** 230–239 (1966). 14

[31] Wang, B. C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol* **115** 90–112 (1985). 14

[32] Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. BALBES: a molecular-replacement pipeline. *Acta Crystallogr D Biol Crystallogr* **64** 125–32 (2008). 14

[33] Navaza, J. Implementation of molecular replacement in AMoRe. *Acta Crystallogr D Biol Crystallogr* **57** 1367–72 (2001). 14

[34] Rossmann, M. G. & Blow, D. M. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr A* **15** 24–31 (1962). 14, 99

[35] Rossmann, M. G. The Molecular Replacement Method (Gordon and Breach, New York, 1972). 14, 29

[36] Rossmann, M. G. The molecular replacement method. *Acta Crystallogr A* **46** 73–82 (1990). 14

[37] Rossmann, M. G. Molecular replacement - historical background. *Acta Crystallogr D Biol Crystallogr* **57** 1360–1366 (2001). 14, 29

[38] Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr D Biol Crystallogr* **66** 22–5 (2010). 14

[39] Cowtan, K. & Main, P. Miscellaneous algorithms for density modification. *Acta Crystallogr D Biol Crystallogr* **54** 487–93 (1998). 14

[40] Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr* **56** 965–72 (2000). 14

[41] Terwilliger, T. Statistical density modification with non-crystallographic symmetry. *Acta Crystallogr D Biol Crystallogr* **58** 2082–2086 (2002). 14, 29

[42] Terwilliger, T. C. SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol* **374** 22–37 (2003). 14, 17, 98

[43] Cowtan, K. Kevin Cowtan's Book of Fourier. http://www.ysbl.york.ac.uk/cowtan/fourier/fourier.html (2007). 15

[44] Richards, F. M. The matching of physical models to three-dimensional electron-density maps: a simple optical device. *J Mol Biol* **37** 225–30 (1968). 15

[45] Jones, T. A. Interactive electron-density map interpretation: from INTER to O. *Acta Crystallogr D Biol Crystallogr* **60** 2115–25 (2004). 15

[46] Jones, T. A graphics model building and refinement system for macromolecules. *J Appl Crystallogr* **11** 268–272 (1978). 15

[47] Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A* **47** 110–9 (1991). 15, 21

[48] McRee, D. E. XtalView/Xfit–A versatile program for manipulating atomic coordinates and electron density. *J Struct Biol* **125** 156–65 (1999). 15

[49] Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60** 2126–32 (2004). 15

[50] Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66** 486–501 (2010). 15

[51] Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53** 240–55 (1997). 15, 19

[52] Collaborative Computational Project, N. . The CCP4 Suite: Programs for Protein Crystallography. *Acta Crystallogr D Biol Crystallogr* **50** 760–763 (1994). 15, 36, 74

[53] Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66** 213–21 (2010). 15, 17, 30, 36

[54] Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using *ARP/wARP* version 7. *Nat Protoc* **3** 1171–1179 (2008). 16, 20, 23, 34, 74

[55] Lamzin, V. & Wilson, K. Automated refinement for protein crystallography. *Methods Enzymol* **277** 269–305 (1997). 16, 23, 25

[56] Greer, J. Three-dimensional pattern recognition: an approach to automated interpretation of electron density maps of proteins. *J Mol Biol* **82** 279–301 (1974). 16

[57] Holton, T., Ioerger, T. R., Christopher, J. A. & Sacchettini, J. C. Determining protein structure from electron-density maps using pattern matching. *Acta Crystallogr D Biol Crystallogr* **56** 722–34 (2000). 16

[58] Ioerger, T. R. & Sacchettini, J. C. Automatic modeling of protein backbones in electron-density maps via prediction of Calpha coordinates. *Acta Crystallogr D Biol Crystallogr* **58** 2043–54 (2002). 16

[59] Romo, T. D., Sacchettini, J. C. & Ioerger, T. R. Improving amino-acid identification, fit and C-a prediction using the Simplex method in automated model building. *Acta Crystallogr D* **62** 1401–1406 (2006). 16

[60] Ioerger, T. R. & Sacchettini, J. C. TEXTAL system: artificial intelligence techniques for automated protein model building. *Methods Enzymol* **374** 244–70 (2003). 16

[61] Oldfield, T. Pattern-recognition methods to identify secondary structure within X-ray crystallographic electron-density maps. *Acta Crystallogr D Biol Crystallogr* **58** 487–93 (2002). 16

[62] Oldfield, T. J. Automated tracing of electron-density maps of proteins. *Acta Crystallogr D Biol Crystallogr* **59** 483–91 (2003). 16

[63] Bricogne, G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallogr A* **30** 395–405 (1974). 16, 29

[64] Kleywegt, G. J. & Jones, T. A. Template convolution to enhance or detect structural features in macromolecular electron-density maps. *Acta Crystallogr D Biol Crystallogr* **53** 179–85 (1997). 16

[65] Cowtan, K. Fast Fourier feature recognition. *Acta Crystallogr D Biol Crystallogr* **57** 1435–44 (2001). 17

[66] Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62** 1002–1011 (2006). 17, 34

[67] Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7** 95–9 (1963). 17, 20

[68] Cowtan, K. Completion of autobuilt protein models using a database of protein fragments. *Acta Crystallogr D* **68** 328–335 (2012). 17

[69] Terwilliger, T. C. Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallogr D Biol Crystallogr* **59** 38–44 (2003). 17

[70] Terwilliger, T. C. Automated side-chain model building and sequence assignment by template matching. *Acta Crystallogr D Biol Crystallogr* **59** 45–9 (2003). 17

[71] Zwart, P. H. *et al.* Automated structure solution with the PHENIX suite. *Methods Mol Biol* **426** 419–435 (2008). 17

[72] Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* **64** 61–69 (2008). 17, 34

[73] Isaacs, N. W. & Agarwal, R. C. Free atom insertion and refinement as a means of extending and refining phases. *Methods Enzymol* **115** 112–7 (1985). 17, 23

[74] DiMaio, F. *et al.* Creating protein models from electron-density maps using particle-filtering methods. *Bioinformatics* **23** 2851–8 (2007). 17

[75] DiMaio, F., Shavlik, J. & Phillips, G. N. A probabilistic approach to protein backbone tracing in electron density maps. *Bioinformatics* **22** e81–e89 (2006). 17

[76] Rodríguez, D. D. *et al.* Crystallographic ab initio protein structure solution below atomic resolution. *Nat Methods* **6** 651–653 (2009). 17

[77] Sheldrick, G. M. A short history of SHELX. *Acta Crystallogr A* **64** 112–22 (2008). 18

[78] Rodriguez, D. *et al.* Practical structure solution with ARCIMBOLDO. *Acta Crystallogr D* **68** 336–343 (2012). 18, 102

[79] Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67** 355–367 (2011). 19, 30

[80] Engh, R. A. & Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A* **47** 392–400 (1991). 19

[81] Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D* **68** 352–367 (2012). 19, 20

[82] Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins* **12** 345–364 (1992). 19

[83] Diamond, R. A real-space refinement procedure for proteins. *Acta Crystallogr A* **27** 436–452 (1971). 19

[84] Sayre, D. On least-squares refinement of the phases of crystallographic structure factors. *Acta Crystallogr A* **28** 210–212 (1972). 19

[85] Nicholls, R. A., Long, F. & Murshudov, G. N. Low-resolution refinement tools in REFMAC5. *Acta Crystallogr D* **68** 404–417 (2012). 20

[86] Kleywegt, G. J. & Jones, T. A. Efficient rebuilding of protein structures. *Acta Crystallogr D Biol Crystallogr* **52** 829–32 (1996). 20

[87] Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355** 472–475 (1992). 20

[88] Tickle, I. J., Laskowski, R. A. & Moss, D. S. Rfree and the Rfree ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. *Acta Crystallogr D Biol Crystallogr* **54** 547–57 (1998). 21

[89] Brünger, A. T. Free R value: cross-validation in crystallography. *Methods Enzymol* **277** 366–96 (1997). 21

[90] Dodson, E., Kleywegt, G. J. & Wilson, K. Report of a workshop on the use of statistical validators in protein X-ray crystallography. *Acta Crystallogr D Biol Crystallogr* **52** 228–34 (1996). 21

[91] Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G. & Perrakis, A. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* **27** 3392–8 (2011). 21

[92] Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* **39** D411–D419 (2011). 21

[93] Joosten, R. P. *et al.* PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J Appl Crystallogr* **42** 376–384 (2009). 21

[94] Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr D* **68** 484–496 (2012). 21

[95] Gore, S., Velankar, S. & Kleywegt, G. J. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr D* **68** 478–483 (2012). 21

[96] Holton, J. Abstract. *Annual Meeting of the American Crystallographic Association* Abstract W0308 (2005). 21

[97] Stroud, R. M. *et al.* 2007 Annual progress report synopsis of the Center for Structures of Membrane Proteins. *J Struct Funct Genomics* **10** 193–208 (2009). 22

[98] Morris, R., Perrakis, A. & Lamzin, V. Macromolecular Crystallograpy – conventional and high-throughput methods (Oxford University Press, 2007). 22

[99] Furnham, N. *et al.* Knowledge-based real-space explorations for low-resolution structure determination. *Structure* **14** 1313–1320 (2006). 23

[100] Langer, G. G. personal communication (2010). 24

[101] Perrakis, A., Sixma, T., Wilson, K. & Lamzin, V. wARP: Improvement and

extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Crystallogr D Biol Crystallogr* **53** 448–455 (1997). 23

[102] Perrakis, A., Morris, R. & Lamzin, V. Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6** 458–463 (1999). 23, 25, 26

[103] Perrakis, A., Harkiolaki, M., Wilson, K. & Lamzin, V. *ARP/wARP* and molecular replacement. *Acta Crystallogr D Biol Crystallogr* **57** 1445–1450 (2001). 23

[104] Morris, R., Perrakis, A. & Lamzin, V. *ARP/wARP*'s model-building algorithms. I. The main chain. *Acta Crystallogr D Biol Crystallogr* **58** 968–975 (2002). 23, 25

[105] Morris, R., Perrakis, A. & Lamzin, V. *ARP/wARP* and automatic interpretation of protein electron density maps. *Methods Enzymol* **374** 229–244 (2003). 23, 26

[106] Morris, R. *et al.* Breaking good resolutions with *ARP/wARP*. *J Synchrotron Radiat* **11** 56–59 (2004). 23, 26

[107] Cohen, S. *et al.* Towards complete validated models in the next generation of *ARP/wARP*. *Acta Crystallogr D Biol Crystallogr* **60** 2222–2229 (2004). 23, 26

[108] Cohen, S. X. *et al.* *ARP/wARP* and molecular replacement: the next generation. *Acta Crystallogr D Biol Crystallogr* **64** 49–60 (2008). 23

[109] Hattne, J. & Lamzin, V. S. Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps. *Acta Crystallogr D Biol Crystallogr* **64** 834–842 (2008). 23, 27

[110] Zwart, P., Langer, G. & Lamzin, V. Modelling bound ligands in protein crystal structures. *Acta Crystallogr D Biol Crystallogr* **60** 2230–2239 (2004). 23, 27

[111] Evrard, G. X., Langer, G. G., Perrakis, A. & Lamzin, V. S. Assessment of automatic ligand building in *ARP/wARP*. *Acta Crystallogr D Biol Crystallogr* **63** 108–117 (2007). 23, 27

[112] Langer, G. G., Evrard, G. X., Carolan, C. G. & Lamzin, V. S. Fragmentation-Tree Density Representation for Crystallographic Modelling of Bound Ligands. *J Mol Biol* **419** 211–222 (2012). 23, 28

[113] Lamzin, V. & Wilson, K. Automated refinement of protein models. *Acta Crystallogr D Biol Crystallogr* **49** 129–147 (1993). 23

[114] Zou, J. Y. & Jones, T. A. Towards the automatic interpretation of macromolecular electron-density maps: qualitative and quantitative matching of protein sequence to map. *Acta Crystallogr D Biol Crystallogr* **52** 833–41 (1996). 26

[115] Joosten, K. *et al.* A knowledge-driven approach for crystallographic protein model completion. *Acta Crystallogr D Biol Crystallogr* **64** 416–424 (2008). 27, 36, 39

[116] Wang, X. & Janin, J. Orientation of Non-crystallographic symmetry axes in protein crystals. *Acta Crystallogr D Biol Crystallogr* **49** 505–512 (1993). 28

[117] Kleywegt, G. Use of non-crystallographic symmetry in protein

structure refinement. *Acta Crystallogr D Biol Crystallogr* **52** 842 – 857 (1996). 28

[118] Kumar, P., Singh, M. & Karthikeyan, S. Crystal structure analysis of icosahedral lumazine synthase from Salmonella typhimurium, an antibacterial drug target. *Acta Crystallogr D Biol Crystallogr* **67** 131–139 (2011). 29

[119] Kleywegt, G. & Read, R. Not your average density. *Structure* **5** 1557 – 1569 (1997). 29

[120] Schwede, T. *et al.* Outcome of a workshop on applications of protein models in biomedical research. *Structure* **17** 151–159 (2009). 30

[121] Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5** 823–6 (1986). 30

[122] Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12** 85–94 (1999). 30

[123] Jain, E. *et al.* Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10** 136 (2009). 31

[124] UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39** D214–D219 (2011). 31

[125] Sánchez, R. & Sali, A. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* **7** 206–14 (1997). 31

[126] Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25** 3389–3402 (1997). 31, 51

[127] Sali, A. Modeling mutations and homologous proteins. *Curr Opin Biotechnol* **6** 437–51 (1995). 31

[128] Greer, J. Comparative model-building of the mammalian serine proteases. *J Mol Biol* **153** 1027–42 (1981). 31

[129] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. & Thornton, J. M. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326** 347–352 (1987). 31

[130] Fiser, A. & Sali, A. Modeller: generation and refinement of homology models. *Methods Enzymol* **374** 461–491 (2003). 31

[131] Fiser, A., Feig, M., Brooks, C. L., 3rd & Sali, A. Evolution and physics in comparative protein structure modeling. *Acc Chem Res* **35** 413–21 (2002). 31

[132] Kiefer, F., Arnold, K., Kuenzli, M., Bordoli, L. & Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* **37** D387–D392 (2009). 31

[133] Kopp, J. & Schwede, T. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res* **32** D230–D234 (2004). 31

[134] Kopp, J. & Schwede, T. The SWISS-MODEL repository: new features and functionalities. *Nucleic Acids Res* **34** D315–D318 (2006). 31

[135] Schwede, T., Kopp, J., Guex, N. & Peitsch, M. SWISS-MODEL: an

automated protein homology-modeling server. *Nucleic Acids Res* **31** 3381–3385 (2003). 31

[136] Jones, D. T. Progress in protein structure prediction. *Curr Opin Struct Biol* **7** 377–87 (1997). 31

[137] Bonneau, R. & Baker, D. Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct* **30** 173–189 (2001). 31

[138] Chan, H. S. & Dill, K. A. The protein folding problem. *Physics today* **46** 24–32 (1993). 32

[139] Ngan, S.-C., Inouye, M. T. & Samudrala, R. A knowledge-based scoring function based on residue triplets for protein structure prediction. *Protein Eng* **19** 187–193 (2006). 32

[140] Kryshtafovych, A. *et al.* Target highlights in CASP9: Experimental target structures for the critical assessment of techniques for protein structure prediction. *Proteins* **79 Suppl 10** 6–20 (2011). 32

[141] Cozzetto, D. *et al.* Evaluation of template-based models in CASP8 with standard measures. *Proteins* **77 Suppl 9** 18–28 (2009). 32

[142] Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z. & Fidelis, K. Protein structure prediction center in CASP8. *Proteins* **77 Suppl 9** 5–9 (2009). 32

[143] Kinch, L. N. *et al.* CASP9 target classification. *Proteins* **79 Suppl 10** 21–36 (2011). 33

[144] Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)–round IX. *Proteins* **79 Suppl 10** 1–5 (2011). 33

[145] Mariani, V., Kiefer, F., Schmidt, T., Haas, J. & Schwede, T. Assessment of template based protein structure predictions in CASP9. *Proteins* **79 Suppl 10** 37–58 (2011). 33

[146] Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27** 343–50 (2011). 33, 72

[147] CASP. Results from CASP, Number 9. http://predictioncenter.org/casp9/results.cgi (2010). 33

[148] Das, R. & Baker, D. Macromolecular modeling with Rosetta. *Annu Rev Biochem* **77** 363 – 382 (2008). 33

[149] Simons, K. T. *et al.* Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34** 82–95 (1999). 33

[150] Rohl, C. A., Strauss, C. E. M., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55** 656–677 (2004). 33

[151] Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21** 951–60 (2005). 33

[152] Xu, D., Zhang, J., Roy, A. & Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **79 Suppl 10** 147–160 (2011). 33

[153] Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234** 779–815 (1993). 36

[154] Claude, J.-B., Suhre, K., Notredame, C., Claverie, J.-M. & Abergel, C. CaspR: a web server for automated molecular replacement using homology modelling. *Nucleic Acids Res* **32** W606–W609 (2004). 36

[155] Gabanyi, M. J. *et al.* The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J Struct Funct Genomics* **12** 45–54 (2011). 36

[156] van den Bedem, H., Lotan, I., Latombe, J. C. & Deacon, A. M. Real-space protein-model completion: an inverse-kinematics approach. *Acta Crystallogr D Biol Crystallogr* **61** 2–13 (2005). 36, 39, 71

[157] Yao, M., Zhou, Y. & Tanaka, I. LAFIRE: software for automating the refinement process of protein-structure analysis. *Acta Crystallogr D Biol Crystallogr* **62** 189–96 (2006). 36, 39

[158] Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450** 259–264 (2007). 37, 38

[159] McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40** 658–674 (2007). 37

[160] Das, R. & Baker, D. Prospects for de novo phasing with de novo protein models. *Acta Crystallogr D Biol Crystallogr* **65** 169–175 (2009). 37, 38

[161] Rigden, D. J., Keegan, R. M. & Winn, M. D. Molecular replacement using ab initio polyalanine models generated with ROSETTA. *Acta Crystallogr D Biol Crystallogr* **64** 1288–91 (2008). 37

[162] DiMaio, F. *et al.* Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **473** 540–543 (2011). 38

[163] Terwilliger, T. C. *et al.* phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta. *J Struct Funct Genomics* **13** preprint (2012). 38

[164] Weisstein, E. W. CRC Concise Encyclopedia of Mathematics (Chapman & Hall/CRC, 1999). 41

[165] Mackay, A. Quaternion transformation of molecular-orientation. *Acta Crystallogr A* **40** 165–166 (1984). 45

[166] Diamond, R. A note on the rotational superposition problem. *Acta Crystallogr A* **A** 211–216 (1988). 46

[167] Kearsley, S. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr A* **45** 208–210 (1989). 47, 55

[168] Maiorov, V. N. & Crippen, G. M. Size-independent comparison of protein three-dimensional structures. *Proteins* **22** 273–283 (1995). 48

[169] Gan, G., Ma, C. & Wu, J. Data clustering: theory, algorithms, and applications, vol. 20 of *ASA-SIAM series on statistics and applied probability* (SIAM, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 2007). 48

[170] Sibson, R. SLINK: an optimally efficient algorithm for the single-link clus-

ter method. *The Computer Journal* **16** 30–34 (1973). 48

[171] Hartigan, J. & Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *J Roy Statist Soc Ser C* **28** 100–108 (1979). 48

[172] Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37** 205–11 (1951). 49

[173] Pauling, L. & Corey, R. B. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proc Natl Acad Sci U S A* **37** 729–40 (1951). 49

[174] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22** 2577–637 (1983). 49, 50, 51

[175] Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232** 584–99 (1993). 49

[176] Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33** 2302–2309 (2005). 50, 57, 65

[177] Zhang, W., Dunker, A. K. & Zhou, Y. Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* **71** 61–7 (2008). 50, 97

[178] King, R. D. & Sternberg, M. J. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* **5** 2298–2310 (1996). 51

[179] Rost, B. & Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19** 55–72 (1994). 51

[180] Rost, B. Review: protein secondary structure prediction continues to rise. *J Struct Biol* **134** 204–18 (2001). 51

[181] Chou, P. Y. & Fasman, G. D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **13** 211–22 (1974). 51

[182] Garnier, J., Osguthorpe, D. J. & Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120** 97–120 (1978). 51

[183] Eyrich, V. *et al.* EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **17** 1242–1243 (2001). 51

[184] Koh, I. Y. Y. *et al.* EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res* **31** 3311–3315 (2003). 51

[185] Bryson, K. *et al.* Protein structure prediction servers at University College London. *Nucleic Acids Res* **33** W36–38 (2005). 51

[186] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292** 195–202 (1999). 51

[187] McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16** 404–5 (2000). 51

[188] Cheng, J., Randall, A. Z., Sweredoski, M. J. & Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res* **33** W72–76 (2005). 51

[189] Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47** 228–235 (2002). 51, 52

[190] Baldi, P., Brunak, S., Frasconi, P., Soda, G. & Pollastri, G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* **15** 937–46 (1999). 51

[191] Cormen, T. H. Introduction to algorithms (MIT Press, Cambridge, Mass., 2009), 3rd edn. 52

[192] Gusfield, D. Algorithms on strings, trees, and sequences: computer science and computational biology (Cambridge University Press, Cambridge, 1997). 52

[193] Knuth, D., Morris, J., James H & Pratt, V. Fast pattern matching in strings. *SIAM journal on computing* **6** 323–350 (1977). 52

[194] Murshudov, G., Vagin, A. & Dodson, E. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53** 240–255 (1997). 59, 74

[195] Biasini, M. *et al.* OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics* **26** 2626–8 (2010). 60

[196] Biasini, M. & Schwede, T. personal communication (2011). 60, 63

[197] Heuser, P., Wohlfahrt, G. & Schomburg, D. Efficient methods for filtering and ranking fragments for the prediction of structurally variable regions in proteins. *Proteins* **54** 583–595 (2004). 71

[198] DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of protein structures into low-resolution density maps using rosetta. *J Mol Biol* **392** 181–90 (2009). 72

[199] Hazledine, S. *et al.* *ARP/wARP* web service. www.arp-warp.com (2012). 73

[200] Ling, H. *et al.* Structure of the Shiga-like toxin I B-pentamer complexed with an analogue of its receptor Gb(3). *Biochemistry* **37** 1777–1788 (1998). 73

[201] Sauer, F. personal communication (2012). 89

[202] Poon, B. K., Grosse-Kunstleve, R. W., Zwart, P. H. & Sauter, N. K. Detection and correction of underassigned rotational symmetry prior to structure deposition. *Acta Crystallogr D Biol Crystallogr* **66** 503–513 (2010). 95

[203] Tête-Favier, F., Rondeau, J. & Podjarny. . . , A. Structure determination of aldose reductase: joys and traps of local symmetry averaging. *Acta Crystallogr D Biol Crystallogr* **49** 246–256 (1993). 95

[204] Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr D Biol Crystallogr* **60** 2184–2195 (2004). 98

[205] Sibanda, B. L. & Thornton, J. M. Beta-hairpin families in globular proteins. *Nature* **316** 170–174 (1985). 98

[206] Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of beta-sheet barrels in proteins. ii. the observed structures. *J Mol Biol* **236** 1382–1400 (1994). 98

[207] Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of beta-sheet barrels in proteins. i. a theoretical analysis. *J Mol Biol* **236** 1369–1381 (1994). 98

[208] Heuser, P., Langer, G. G. & Lamzin, V. S. Interpretation of very low resolution X-ray electron-density maps using core objects. *Acta Crystallogr D Biol Crystallogr* **65** 690–696 (2009). 98

[209] Seibert, M. M. *et al.* Single mimivirus particles intercepted and imaged with an x-ray laser. *Nature* **470** 78–81 (2011). 98

Appendix

A

Testcases

Testcase	Resolution (Å)	Residues	NCS operators	α-content	β-content
Cc_1	1.9	2064	4	24%	10%
Cc_2	2.1	800	4	62%	2%
Cc_3	2.3	261	3	0%	55%
Cc_4	2.4	345	3	19%	39%
Cc_5	2.5	360	3	10%	0%
Cc_6	2.8	440	4	39%	15%
Cc_7	2.9	2300	10	8%	45%
Cc_8	3.0	616	2	65%	0%
Cc_9	3.0	480	2	0%	45%
Cc_{10}	3.0	580	2	35%	14%
1duv$_{cut}$	3.0	999	3	46%	15%
1o14	3.2	662	2	35%	26%
Cc_{11}	3.2	460	2	49%	16%

Table A.1: Testdata for used for the PNSextender method. To warrant anonymity, job number of the cluster cases used have been changed to Cc_1 to Cc_{11}. Shown secondary structure content has been computed with PSIPRED or taken from the DSSP assignment where applicable. Cases from the PDB for which the data has been cut are indicated by a $_{cut}$ subscript.

Testcase	Resolution (Å)	Residues	α-content	β-content
2pjw	3.0	179	82%	0%
2o8x	3.0	210	61%	0%
3i78	3.0	229	9%	38%
1plr	3.0	258	19%	44%
2qsr	3.1	173	44%	16%
2v6t	3.1	212	42%	19%
2aj2	3.2	208	16%	28%
3eb6	3.4	223	24%	18%
2ing	3.6	213	29%	16%
1xn4	3.8	192	35%	19%

Table A.2: Testdata used for the FittOFF method. Secondary structure content shown has been derived from the DSSP assignment.

Appendix

B

Detailed results

ID	Size	Resolution (Å)	Model completeness (%)	Residues built (%)	Residues docked in sequence (%)	Average residues per chain (%)	R-Factor
Cc_1	4 × 516	1.9	+0.8%	+0.8%	+0.8%	+127%	-0.004
Cc_2	4 × 200	2.1	+4.9%	+6.0%	+7.3%	+80%	-0.013
Cc_3	3 × 87	2.3	+9.6%	+15.2%	+16.5%	+79%	-0.075
Cc_4	3 × 115	2.4	+7.5%	+9.1%	+49.3%	+160%	-0.042
Cc_5	3 × 120	2.5	+8.1%	+9.3%	+17.5%	+228%	-0.058
Cc_6	4 × 110	2.8	+9.3%	+15.6%	+13.2%	+37%	-0.032
Cc_7	10 × 230	2.9	+4.2%	+7.0%	+7.1%	+8%	-0.013
Cc_8	2 × 308	3.0	+7.1%	+12.2%	+9.3%	+50%	-0.030
Cc_9	2 × 240	3.0	+12.5%	+19.7%	+6.3%	+25%	-0.044
Cc_{10}	2 × 290	3.0	+5.5%	+11.2%	+3.3%	+18%	-0.013
1duv$_{cut}$	3 × 333	3.0	+5.2%	+5.8%	+10.7%	+99%	-0.050
1o14	2 × 331	3.2	+10.4%	+15.2%	+32.9%	+57%	-0.024
Cc_{11}	2 × 230	3.2	+15.2%	+27.0%	+5.4%	+9%	-0.097
		Ø	+7.7%	+11.9%	+13.8%	+75%	-0.038

Table B.1: Best results obtained during testing of the PNSextender method with variable parameters. In all cases more residues have been built and docked, the average number of residues per chain is higher and the R-Factor decreases. This table shows the improvements.

ID	Size	Resolution (Å)	Residues built	Residues docked in sequence	Average residues per chain	R-Factor
Cc_1	4 × 516	1.9	2004 / 2020	2004 / 2020	223 / 505	0.196 / 0.192
Cc_2	4 × 200	2.1	653 / 692	621 / 680	54 / 98	0.277 / 0.264
Cc_3	3 × 87	2.3	164 / 189	103 / 146	16 / 29	0.397 / 0.322
Cc_4	3 × 115	2.4	286 / 312	116 / 286	24 / 62	0.352 / 0.310
Cc_5	3 × 120	2.5	312 / 341	278 / 341	35 / 114	0.293 / 0.235
Cc_6	4 × 110	2.8	263 / 304	0 / 59	7 / 9	0.329 / 0.297
Cc_7	10 × 230	2.9	1363 / 1459	56 / 219	10 / 11	0.212 / 0.199
Cc_8	2 × 308	3.0	362 / 406	104 / 161	8 / 12	0.291 / 0.261
Cc_9	2 × 240	3.0	304 / 364	0 / 31	6 / 8	0.409 / 0.364
Cc_{10}	2 × 290	3.0	286 / 318	0 / 20	6 / 7	0.320 / 0.307
$1duv_{cut}$	3 × 333	3.0	897 / 949	809 / 916	43 / 85	0.283 / 0.233
1o14	2 × 331	3.2	453 / 522	57 / 275	10 / 16	0.263 / 0.239
Cc_{11}	2 × 230	3.2	259 / 329	0 / 26	6 / 7	0.448 / 0.351

Table B.2: Best results obtained during testing of the PNSextender method with variable parameters. This table shows the results from the ARP/wARP protocol executed without the PNSextender compared to the one incorporating it.

ID	Size	Resolution (Å)	Model completeness (%)	Residues built (%)	Residues docked in sequence (%)	Average residues per chain (%)	R-Factor
Cc_1	4 × 516	+1.9	+0%	+0%	+0%	+0%	+0.000
Cc_2	4 × 200	+2.1	+3.4%	+4.1%	+3.4%	+25%	+0.007
Cc_3	3 × 87	+2.3	+0%	+0%	+0%	+0%	+0.000
Cc_4	3 × 115	+2.4	+3.8%	+4.5%	+43%	+25.5%	+0.028
Cc_5	3 × 120	+2.5	+4.5%	+5.1%	+14%	+58%	-0.009
Cc_6	4 × 110	+2.8	+3.4%	+5.7%	+3.4%	+29%	+0.001
Cc_7	10 × 230	+2.9	+4.0%	+6.7%	+5%	+8%	+0.017
Cc_8	2 × 308	+3.0	+6.5%	+11.0%	+6%	+28%	+0.000
Cc_9	2 × 240	+3.0	+8.8%	+13.8%	+0%	+7%	-0.040
Cc_{10}	2 × 290	+3.0	+0.3%	+0.7%	+0 %	-1%	+0.010
1duv$_{cut}$	3 × 333	+3.0	+3.9%	+4.3%	+11%	+83%	-0.050
1o14	2 × 331	+3.2	+6.8%	+9.9%	+21%	+37%	-0.008
Cc_{11}	2 × 230	+3.2	+6.1%	+10.8%	+0%	-1%	-0.023
		Ø	+4.0% (+4.8%)	+5.9% (+7.3%)	+8% (+10%)	+23% (+27%)	-0.005 (-0.007)

Table B.3: Average results for tests of the PNSextender method for fixed protocols released in ARP/wARP version 7.2. In most cases more residues have been built and docked and the average number of residues per chain is higher. There are a few declines for the R-Factor. This table shows the improvements. Values in brackets are the average for application at resolution lower than 2.4 Å.

ID	Size	Resolution (Å)	Model completeness (%)	Residues built (%)	Residues docked in sequence (%)	Average residues per chain (%)	R-Factor
2pjw	179	3.0	+11.7%	+25.3%	+0%	+29%	-0.038
2o8x	210	3.0	+9.0%	+11.8%	+15%	+86%	-0.036
3i78	229	3.0	+4.4%	+6.3%	+30%	+15%	-0.003
1plr	258	3.0	+5.8%	+6.7%	+27%	+60%	+0.001
2qsr	173	3.1	+11.0%	+10%	+11%	+98%	-0.043
2v6t	212	3.1	+7.5%	+11.4%	+53%	+55%	-0.016
2aj2	208	3.2	+10.1%	+14.6%	+35%	+47%	-0.022
3eb6	223	3.4	+6.7%	+9.9%	+19%	+31%	-0.032
2ing	213	3.6	+6.1%	+13.3%	+0%	+13%	-0.002
1xn4	192	3.8	+6.8%	+18.8%	+8%	+47%	-0.023
		Ø	8.0%	12.8%	20%	48%	-0.021

Table B.4: Best results during testing of the FittOFF method with variable parameters. This table shows the improvements.

ID	Size	Resolution (Å)	Residues built	Residues docked in sequence	Average residues per chain	R-Factor
2pjw	179	3.0	83 / 104	0 / 0	7 / 9	0.348 / 0.310
2o8x	210	3.0	161 / 180	85 / 100	32 / 60	0.259 / 0.223
3i78	229	3.0	159 / 169	0 / 50	12 / 14	0.278 / 0.278
1plr	258	3.0	225 / 240	145 / 220	38 / 50	0.249 / 0.250
2qsr	173	3.1	119 / 138	0 / 13	13 / 26	0.353 / 0.310
2v6t	212	3.1	140 / 156	0 / 74	8 / 13	0.246 / 0.230
2aj2	208	3.2	144 / 165	60 / 116	16 / 24	0.271 / 0.249
3eb6	223	3.4	151 / 166	18 / 50	9 / 12	0.296 / 0.264
2ing	213	3.6	98 / 111	0 / 0	9 / 10	0.430 / 0.428
1xn4	192	3.8	69 / 82	0 / 7	7 / 8	0.377 / 0.348

Table B.5: Best results for each category for tests of the FittOFF method with variable parameters. This table shows the results from the ARP/wARP protocol executed without the FittOFF compared to the one incorporating it.

Appendix

C

Acknowledgements

A tremendous amount of thanks go to Silke, the one and only girl in my life, who helps me in too many ways to tell and, most importantly, always knows how to make me smile (*fistbump*)! More thanks go to my parents, who paved, padded and smoothed the way leading here (also to my sister and our mentally challenged, furry and monochromatic brother).

But to be serious, all this would not have been possible without the continuing support I receive from Victor Lamzin and him giving me the opportunity to work in his group. Over the years, Victor made real effort to teach me about math and crystallography (of which one third still seems a lot like magic to me). I am also really grateful that he encouraged me to give loads of lectures, talks and tutorials on *ARP/wARP* literally all over the world (and also paid for it). Next to Victor I also have to mention Philipp Heuser and Gerrit Langer, who quite often helped me to develop new ideas or simply to find bugs. Another essential part during my PhD was discussions and brainstormings with Ciaran Carolan, Saul Hazledine, Matthew Groves and Santosh Panjikar (and the rest of the Lamzin group).
EMBL regulations state that every PhD student needs to have a “proper” PhD supervisor at a University - this allowed me to keep in touch with my Diplom supervisor Andrew Torda, who was a great help guiding me through the jungle of forms and guidelines at Uni Hamburg or for discussions (sometimes even scientific) at our favourite Greek restaurant (Corfu, Hellkamp 43). I would also like to thank my Thesis Advisory Committee, which included Victor, Andrew, Richard Morris, Thomas Schneider, Manfred Weiss

and Anne-Claude Gavin. They were a great help and always gave me amazing feedback during each of our yearly meetings. These "TAC" meetings were really helpful since I had to formulate what I had done (always good if you give talks or...write a thesis) and always got a good idea how to go on. So, when shall we have the “whattodowithmylife”-meeting?
Another huge Thank You goes down to Basel to Marco Biasini and Torsten Schwede. Our collaboration saved me a big load of effort (which wouldn’t have been remotely as good as what Marco did)!

I wouldn’t even have thought about taking this way, would it not have been for my dear nerd-friends (Jens Kleesiek, Thomas Margraf, Joern Lenz and my LATEX magician Stefan “Bienchen” Bienert), this for you:

Computerfreak, Computerfreak
Lebst in deiner Welt,
hackst mit dicken Fingern in die Tastatur.

Everyone who read and helped correcting this thing called thesis (no sense in writing your names twice) can count on some free drinks (even alcohol-free Weizen).

It’s all about work-life-balance, so thanks for keeping me down to earth: Mario, Charlotte, Nadine, Sushi, Axel, Klaas, Jan, Sina, Krissi, Lea and everyone I like, forgot, who contributed to this in any way or feels upset because I didn’t mention him/her (Yeah, you, sorry for that - but feel acknowledged)!

Last but not least, money makes the world go round, so thank you, dear EMBL, for the nice stipend that made my time as a PhD student easier and sometimes luxurious, allowed me a colored right arm and made me lose 25kg somewhere on the way.

`s/\s*\d*\S*[hH][!\s]al[!a-su-z\s]//g`

Appendix

D

Gefahrstoffe und KMR-Substanzen

Die vorliegende Arbeit ist rein theoretischer Natur. Es wurden daher keinerlei Laborexperimente mit chemischen oder biologischen Materialien durchgeführt. Aus diesem Grund werden keine Gefahrstoffe, krebserzeugende, erbgutverändernde oder fortpflanzungsgefährdende (KMR) Stoffe angegeben.

Appendix

E

Erklärungen

Ich versichere an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht.
Ferner versichere ich, dass dies mein erster Promotionsversuch ist und dass ich diese Dissertation noch an keiner anderen Universität eingereicht habe um ein Promotionsverfahren eröffnen zu lassen.

(Tim Wiegels)

Curriculum vitae

In der elektronischen Version entfällt der Lebenslauf aus Datenschutzgründen.

In the electronic version of this document, the CV has been omitted for privacy reasons.