

Software developments in automated structure solutions and crystallographic studies of the Sso10a2 and human C1 inhibitor protein

Waterreus, W.J.

Citation
Waterreus, W. J. (2013, December 5). *Software developments in automated structure solutions and crystallographic studies of the Sso10a2 and human C1 inhibitor protein*. Retrieved from https://hdl.handle.net/1887/22715

Version:	Not Applicable (or Unknown)
License:	Leiden University Non-exclusive license
Downloaded from:	https://hdl.handle.net/1887/22715

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle http://hdl.handle.net/1887/22715 holds various files of this Leiden University dissertation.

Author: Waterreus, Willem-Jan
Title: Software developments in automated structure solution and crystallographic studies of the Sso10a2 and human C1 inhibitor protein
Issue Date: 2013-12-05

SOFTWARE DEVELOPMENTS

in automated structure solution and

CRYSTALLOGRAPHIC STUDIES

of the Sso10a2 and human C1 inhibitor protein

Willem-Jan Waterreus

November 5, 2013

SOFTWARE DEVELOPMENTS

in automated structure solution and

CRYSTALLOGRAPHIC STUDIES

of the Sso10a2 and human C1 inhibitor protein

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College van Promoties
te verdedigen op donderdag 5 december 2013
klokke 13:45 uur

door

Willem-Jan Waterreus
geboren te Rotterdam, Nederland
7 juli 1982

Promotiecommissie

prof. dr. Jan Pieter Abrahams (promotor)
dr. Navraj Pannu (copromotor)
prof. dr. Jaap Brouwer
dr. Remus Dame
prof. dr. Mathieu Noteborn
dr. Irakli Sikharulidze, *Diamond Light Source, Oxford, UK*
prof. dr. Gerard Canters

Cover illustration reproduced with permission from:
Towards the Infinitesimal by Joan Modderkolk

Printed and bound by Wöhrmann Print Service, Zutphen, The Netherlands

The work reported in this thesis was performed at the Leiden Institute of Chemistry, Gorlaeus Laboratories, Einsteinweg 55, Leiden, The Netherlands. Funding was provided by the Netherlands Organization for Scientific Research (NWO), grant number 700.55.425 and Leiden University.

© 2013 Willem-Jan Waterreus All rights reserved.

Contents

List of Figures vii

List of Tables viii

List of Abbreviations ix

1 Introduction 3
- 1.1 The importance of structural biology . . . 3
  - 1.1.1 The molecular basis of life . . . 4
  - 1.1.2 Protein structure and human disease . . . 5
- 1.2 An introduction to X-ray structure solution . . . 7
  - 1.2.1 The basics of diffraction and the phase problem . . . 8
  - 1.2.2 Solving the phase problem . . . 12
    - Direct methods . . . 12
    - Molecular replacement . . . 13
    - Experimental phasing methods . . . 14
- 1.3 Automation in X-ray structure solution . . . 23
  - 1.3.1 Substructure detection . . . 24
  - 1.3.2 Substructure phasing . . . 26
  - 1.3.3 Density modification . . . 27
  - 1.3.4 Automated model building and refinement . . . 32

2 Recent advances in Crank 37
- 2.1 Introduction . . . 38
- 2.2 Recent developments in Crank . . . 40
  - 2.2.1 Substructure determination . . . 40
  - 2.2.2 Substructure phasing . . . 41
  - 2.2.3 Density modification . . . 42
  - 2.2.4 Automated model building and refinement . . . 42
  - 2.2.5 Integration of programs and steps . . . 43

2.3 Methods . . . 45
2.4 Results and Discussion . . . 46
2.4.1 Analysis of data sets that were not automatically built . . . 47
2.5 Conclusions and future developments . . . 49

3 A new method for phase combination 51
3.1 Introduction . . . 52
3.2 Methods . . . 54
3.3 Results . . . 56
3.4 Discussion . . . 59

4 Structural analysis of DNA binding protein Sso10a2 63
4.1 Introduction . . . 64
4.2 Methods . . . 68
4.2.1 Cloning and overproduction . . . 68
4.2.2 Purification . . . 68
4.2.3 Crystallisation . . . 69
4.2.4 X-ray diffraction analysis . . . 70
4.2.5 Analysis of the crystal structure . . . 71
4.3 Results and Discussion . . . 71
4.3.1 Sequence analysis . . . 71
4.3.2 Structure solution . . . 74
4.3.3 Global structure . . . 77
4.3.4 Comparison of Sso10a and Sso10a2 . . . 77
4.3.5 Concluding remarks . . . 79

5 Crystallisation of human recombinant C1 inhibitor 83
5.1 Introduction . . . 84
5.2 Methods . . . 88
5.2.1 Protein expression and purification . . . 88
5.2.2 Crystallisation . . . 89
5.3 Results and discussion . . . 90
5.3.1 Crystallisation screening . . . 90
5.3.2 Diffraction targets . . . 91
5.3.3 Concluding remarks . . . 92

A Data sets used for testing 95

Bibliography 101

Nederlandse samenvatting 123

List of Figures

1.1 Bragg diffraction. 9
1.2 Diffraction image from the Sso10a2 selenomethionine data set 11
1.3 The Harker construction for a SIR experiment. 17
1.4 The Harker construction for a SAD experiment. 19
1.5 Lack of closure in a SIR experiment. 21
1.6 The Harker construction for a MIR experiment. 22
1.7 The Harker construction for a SIRAS experiment. 23
1.8 Typical steps in experimental phasing and model building. . . 25
1.9 Steps in traditional density modification. 28
1.10 Steps in statistical density modification. 31

2.1 Flow chart of the programs CRANK interfaces with. 39
2.2 Screenshot of the CRANK graphical user interface. 44
2.3 Fraction of the model built for CRANK 1.3 versus 1.4. 46
2.4 The Bijvoet ratios as a function of resolution for the peak and inflection wavelength data of the GerE data set. 48

3.1 Flow chart of the different steps in density modification. . . . 53
3.2 Map correlation for *SAD-DM* versus the σ_A. 57
3.3 Distribution of improvement in map correlation after density modification as a function of resolution and map correlation after experimental phasing for *SAD-DM* versus σ_A. 58
3.4 Fraction of the model built correctly for *SAD-DM* versus σ_A. 60

4.1 Sequence alignment of archaeal Sso10a2 homologs. 73
4.2 Representative electron-density for each chain in asymmetric unit of the Sso10a2 crystal. 80
4.3 Superposition of the Sso10a and Sso10a2 structures. 81

5.1 Typical C1 inhibitor crystals selected for diffraction analysis. 93

List of Tables

3.1 Average fraction of the model correctly built for BUCCANEER and ARP/wARP. 59

4.1 Data collection statistics for Sso10a2. 74
4.2 Refinement statistics for Sso10a2. 76

A.1 The map correlation, fraction of the model built correctly and miscellaneous raw data comparing the performance of σ_A versus *SAD-DM*. 96

List of Abbreviations

Alba	acetylation lowers binding affinity
ASU	asymmetric unit
Bk2R	bradykinin type 2 receptor
C1INH	C1 inhibitor
DM	density modification
EMSA	electrophoretic mobility shift assay
ESRF	European Synchrotron Radiation Facility
FOM	figure of merit
fXIa	activated factor XI
fXII	factor XII
fXIIa	activated factor XII
GAG	glycosaminoglycan
GUI	graphical user interface
HAE	hereditary angioedema
HMWK	high-molecular-weight kinogen
IPTG	isopropyl β-D-1-thiogalactopyranoside
LMW1	low molecular weight rhC1INH species 1
LMW2	low molecular weight rhC1INH species 2
MAD	multiple-wavelength anomalous diffraction

MIB	sodium malonate, imidazole and boric acid
MIR	multiple isomorphous replacement
MLHL	maximum likelihood Hendrickson-Lattman
MMT	DL-malic acid, MES and Tris base
MR	molecular replacement
NCS	non-crystallographic symmetry
NMR	nuclear magnetic resonance spectroscopy
PCB	sodium propionate, sodium cacodylate and Bis-tris propane
PDB	Protein Data Bank
PEG	polyethylene glycol
PEG MME	polyethylene glycol monomethyl ether
PNGase F	peptide-N4- (N-acetyl-beta-glucosaminyl) asparagine amidase
RCL	reactive center loop
rhC1INH	recombinant human C1 inhibitor
RMSD	root-mean-square deviation
SAD	single-wavelength anomalous diffraction
serpin	serine protease inhibitor
SIR	single isomorphous replacement
SIRAS	single isomorphous replacement with anomalous scattering
SPG	succinic acid, sodium dihydrogen phosphate and glycine
TLS	translation libration screw-motion
TPM	tethered particle motion

1

Introduction

1.1 The importance of structural biology

"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." With this seemingly unassuming conclusion Watson and Crick (1953) have charmed many in the paper describing the double-helix structure of DNA. A model that elegantly explained X-ray diffraction experiments on DNA fibres by Franklin and Gosling (1953). The discovery of the DNA double helix laid the foundation for a molecular understanding of heredity. While the relation between structure and function of a macromolecule may not always be so apparent, more often than not, a three dimensional atomic model gives valuable insight into the molecular basis for its biological activity. Only a few years after the discovery of the structure of DNA, Kendrew et al. (1958) published the first X-ray structure of a protein: myoglobin, an iron containing protein related to haemoglobin.

Both discoveries set off a research effort to unravel the molecular structure of life and improve understanding of the molecular defects in human disease at an atomic level, an endeavour that has resulted in more than eighty-thousand crystal structures of protein and nucleic acid molecules deposited to the Protein Data Bank (PDB) only five decades later.

1.1.1 The molecular basis of life

Life is characterized by a number of innate unique qualities that set it apart from ordinary, lifeless matter. The self-organisation of well defined structural components into systems of enormous organisational complexity is the most ubiquitous property of life and required to perform many of the functions that are also considered to be essential traits of living systems, such as the ability to interact with the environment, transform energy, metabolize matter, and reproduce (Palade 1964). At the heart of these processes are proteins whose function is determined by their molecular structure and composition.

The genes needed to produce all proteins are stored on either the strand of what is without question the best known structure of a biological molecule: the DNA double helix. The amino acid sequence of a protein is encoded in a gene using a simple cypher consisting of a triplet of the bases adenine (A), thymine (T), cytosine (C), and guanine (G) (e.g. Lodish et al. 2000). The highly specific pairing of the A/T and G/C base-pairs, not only ensures that each strand can be reliably copied from the other, but also allows a gene to be converted into an RNA intermediate by a process called transcription. In the last stage of protein expression, each triplet in the RNA sequence is translated to the one of the twenty-one amino acids that can occur in a protein until a triplet signalling the end of the protein is encountered (e.g. Lodish et al. 2000).

The human genome is composed of roughly three billion base-pairs distributed over twenty-two chromosomes common to both men and women plus the two sex chromosomes. The DNA content of a single human cell is about two meters stretched end-to-end (Lander et al. 2001; Watson and Crick 1953). To some extent the length of this molecule may be reduced by introducing additional coils into the double helix causing the molecule to kink and fold up on itself (Champoux 2001). This however is not enough to confine the DNA to the small volume of the cell or cellular nucleus. Thus, most organisms express proteins that pack the DNA into a compact structure called chromatin by wrapping the DNA,

introducing bends, or looping the DNA by intra-strand forming bridges (Luijsterburg et al. 2008).

An important functional aspect of DNA packing is in the control of gene activity; usually local unpacking is associated with increased expression of the genes in that region and vice versa (Li et al. 2005; Luijsterburg et al. 2008). In addition to the role in gene activity, chromatin organisation has an important function in promoting the genomic stability, which is critical to survival and proliferation of an organism (Duboule et al. 2007; Oberdoerffer and Sinclair 2007). One example is the involvement of chromatin organisation and mobility in DNA damage repair. DNA lesions, caused by, for instance, UV- or ionizing radiation can have detrimental effects on the cell by blocking the transcription of the DNA into RNA, hindering DNA-replication or introducing mutations in the genetic code. While chromatin mobility was considered of little functional relevance, and chromatin packing often regarded as an obstacle to DNA repair, both are now recognized to have an active role in the repair of DNA lesions (Duboule et al. 2007; Groth et al. 2007; Misteli and Soutoglou 2009; Soria et al. 2012).

Homologous recombination, one of the processes involved in maintaining genomic integrity, sometimes inadvertently causes duplication of whole stretches of DNA. The genes duplicated through recombination or other mechanisms, can be passed on to an organism's offspring and may diverge in function over time (Näsvall et al. 2012; Ohno 1970). In Chapter 4 the crystal structure of the Sso10a2 protein is presented, a member of three highly homologous DNA binding proteins that have co-evolved in *Sulfolobus solfataricus*. Even though the structure of Sso10a2 is very similar to that of Sso10a, there are a number of striking differences. Chapter 4 speculates about the different functional role of Sso10a2 by comparing the structure with the homologous structure of Sso10a.

1.1.2 Protein structure and human disease

As well as unravelling the molecular basis of life's processes, knowledge of protein structure is important to understanding disease processes caused

by molecular defects in proteins. Normal cellular activity is disrupted when a protein that performs an essential function is not active or not expressed, which in some cases can lead to serious disease.

The same processes that cause gene duplication can result in deletion of a gene and consequently reduce or completely abolish the expression of the encoded proteins. Various other types of mutation can affect protein activity either by disrupting the global fold, often the result of a frameshift or nonsense mutation, or by missense mutations that result in the substitution of a different amino acid. Especially when located in the hydrophobic core, an amino acid substitution can have a dramatic effect on the stability and global fold of a protein (Dill 1990; Eriksson et al. 1992; Sharp 1991), whereas localized changes in structure often have the greatest effect near the catalytic centre of an enzyme or in interaction sites (Schaefer and Rost 2012). Several diseases, such as familial Alzheimer's and autosomal recessive Parkinson's disease, are associated with mutations that decrease protein stability (Stefl et al. 2013; Wang and Moult 2001; Yue et al. 2005).

While the substitution of an amino acid does not always lead to appreciable changes in the protein's structure, the altered chemistry can have a marked influence on the activity of the protein. A well known example of such a mutation is the substitution of glutamic acid at position 6 by valine in haemoglobin that causes Sickle-cell disease in people homozygous for the mutation (Ingram 1958; Neel 1949). Under low oxygen conditions, a hydrophobic patch is exposed on the haemoglobin surface that binds the valine at position 6 resulting in the polymerization of haemoglobin into rod-like structures. The sickle shape and reduced flexibility of the red blood cells, induced by the fibrous haemoglobin precipitates, causes blockage of capillaries. Consequently, the symptoms of sickle cell disease are mostly a direct result of restriction of blood flow to organs and tissues.

Sometimes small molecule drugs can mitigate the effect of inactive or malfunctioning proteins by targeting other parts of the disease causing pathway. While difficult to achieve with a small molecule, restoration of

the biological function of an inactive protein can sometimes be accomplished by simply substituting it with an active equivalent. For instance, hereditary angioedema (HAE), a condition which is accompanied by life-threatening swelling in the upper airways, may be treated by substituting endogenous inactive or absent C1 inhibitor (C1INH) though intravenous injection of Ruconest®, a recombinant C1INH (Plosker 2012; Varga and Farkas 2011; Zuraw et al. 2010). Chapter 5 reports the outcome of a high-throughput crystallization screen of three forms of recombinant human C1 inhibitor (rhC1INH) in over one thousand different conditions. Several high quality crystals were obtained and preliminary evidence suggests that these crystals may contain rhC1INH in a conformation that is currently unknown. Knowledge of both active and latent (Beinrohr et al. 2007) forms of C1INH will give insight into the molecular mechanisms underpinning C1INH's activity, and could provide a molecular explanation for HAE and other diseases associated with molecular defects in C1INH.

1.2 An introduction to X-ray structure solution

Nowadays, a structural biologist has access to an entire arsenal of technologies for biophysical characterisation of a protein, including powerful methods like nuclear magnetic resonance spectroscopy (NMR) and high resolution cryo-electron microscopy. Nonetheless, X-ray crystallography remains the method of choice for determining the structure of a protein at atomic resolution. In order to do so, however, one has to accomplish the often difficult task of expressing, purifying and crystallizing the molecule of interest. Even if a crystal is obtained and diffracts, confounding factors such as high mosaicity, twinning, non-isomorphism or a weak (anomalous) signal may make structure solution difficult or impossible. New algorithms that intelligently handle these situations in conjunction with the integration of individual programs into software suites have enabled the crystallographer to rapidly try different struc-

ture solution strategies and obtain a three dimensional model from even the most challenging diffraction data.

1.2.1 The basics of diffraction and the phase problem

The behaviour of X-ray photons can be understood in terms of properties associated with waves, as well as particle-like properties. When an X-ray photon interacts with an electron bound to an atomic nucleus it can scatter elastically. This means the energy of the photon, and thus its wavelength, is conserved and only the direction of the photon is changed. Alternatively, X-ray scattering can be thought of as the interaction of electromagnetic waves with the electrons in a solid material. Irradiation of the solid with an X-ray beam causes the electrons to oscillate with the same period as the incoming beam. The accelerated electrons emit their own electromagnetic field with the same phase and wavelength as the incident X-rays that propagates radially outward from every scatterer. The resultant field composed of contributions of individual scatterers is called the scattered wave.

Due to the periodic structure of a crystalline solid, diffraction of X-rays by a crystal can be described as scattering of electromagnetic waves by a series of equidistant planes, denoted by three integers h, k and ℓ, known as Miller indices (Miller 1839). As illustrated in Section 1.2.1 the incident X-ray beam hitting the surface of an $(hk\ell)$ plane at angle θ is partially scattered at the same angle θ away from the plane. The non-scattered radiation is transmitted deeper into the solid until it interacts with the electrons in the second plane, where the process repeats. The X-rays reflected off a given $(hk\ell)$ plane will travel different path lengths through the solid. Constructive interference of the scattered waves only occurs at certain angles θ for which the path lengths are equal to an integer number n times the wavelength λ. This relationship, shown in Equation 1.1 was derived by William Lawrence Bragg and is known as Bragg's law (Bragg 1913).

$$n\lambda = 2d\sin(\theta) \tag{1.1}$$

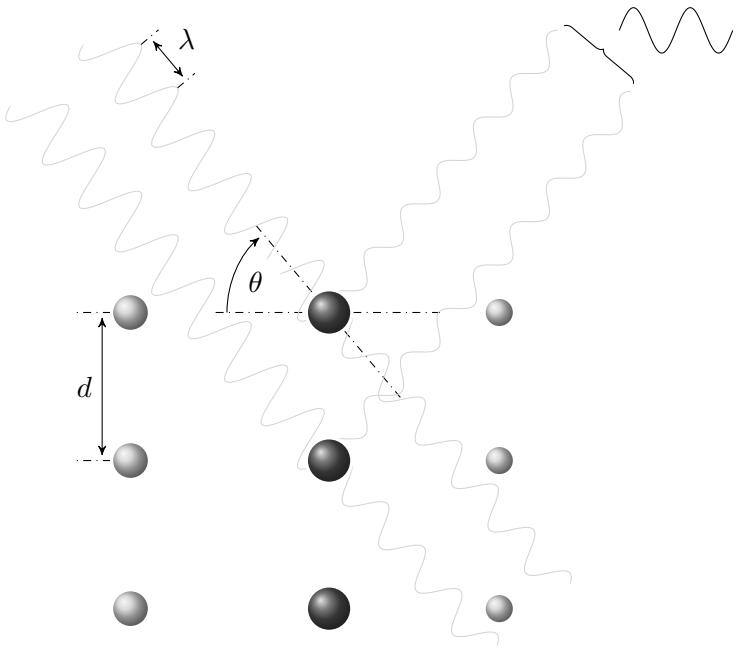


Figure 1.1: Bragg diffraction. At the given angle of incidence θ the X-rays scattered off the parallel $(hk\ell)$ planes separated by a distance d are shifted in phase such that they interfere constructively and give rise to a detectable reflection. For most other values of θ the reflected waves are out of phase and do not interfere constructively.

The distance between successive identical planes in the crystal is denoted by d. It is immediately apparent from the reformulation of Bragg's law shown in Equation 1.2 that the maximum resolution at which a crystal can be sampled, that is the minimum distance d_{min} between $(hk\ell)$ planes, is determined by θ_{max}, the maximum angle to which the crystal still diffracts. Theoretically the maximum resolution would be obtained when $\theta_{max} = \frac{1}{2}\pi$, but due to decreasing atomic scattering factors at increasing Bragg angles the maximum scattering angle is lower ($\theta_{max} \ll \frac{1}{2}\pi$).

$$d_{min} = \frac{1}{2}\frac{n\lambda}{sin(\theta_{max})} \tag{1.2}$$

In a diffraction experiment the angle of the primary X-ray beam with respect to the crystal planes is changed by rotating the crystal. Whilst rotating, the position and intensity of the diffracted waves are recorded on an X-ray detector. The result is a regular arrangement of spots of varying intensity, known as *reflections* (see Section 1.2.1). Each reflection can be assigned a set of indices $(hk\ell)$ associated with the direction of the set of reflecting parallel planes in the crystal. The type and number of scattering atoms in the $(hk\ell)$ planes, or more precisely the associated electrons, determine the magnitude of the corresponding $(hk\ell)$ reflection. The squared *structure factor amplitude* is proportional to the intensity value of a reflection $|F_{hk\ell}|^2 \propto I_{hk\ell}$.

$$F_{hk\ell} = |F_{hk\ell}|e^{i\alpha_{hk\ell}} = \sum_{j=1}^{N} f_j e^{2\pi i(hx_j + ky_j + \ell z_j)} \tag{1.3}$$

The *structure factor* is a complex number, as shown in Equation 1.3, and can be obtained by a simple summation over all N atoms j in the unit cell with positional coordinates x_j, y_j and z_j. The atomic scattering factor f_j depends on the element and the diffraction angle for reflection $(hk\ell)$. Equation 1.3 shows that if the atomic coordinates are known, the structure factor amplitudes $F_{hk\ell}$ and phase $\alpha_{hk\ell}$ can easily be obtained. However, in X-ray crystallography one tries to solve the inverse problem, that is determining the *electron density* $\rho(x, y, z)$ in the unit cell. Note

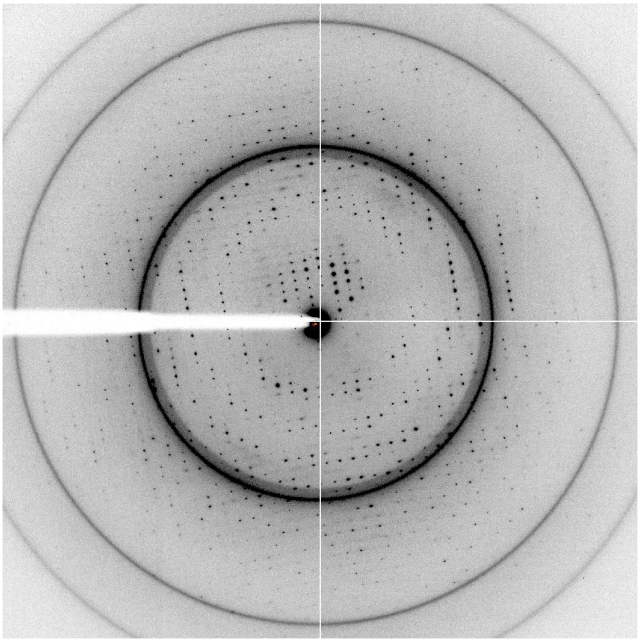


Figure 1.2: Diffraction image from the data set used to solve the Sso10a2 structure presented in Chapter 4. Each dot represents a reflection that is assigned an $(hk\ell)$ index. The structure factor amplitude $|F_{hk\ell}|$ is calculated from the intensity value $I_{hk\ell}$ of each reflection. Most X-rays pass through the crystal without being scattered, which shows up as a large high intensity circle in the centre of the image. The further away from the centre the reflections are, the higher the resolution of the data set. Some of the reflections are obscured by the large ring around 2.2 Å, a problem that results from scattering by the ice that can accumulate on the crystal.

that the structure factors are a *reciprocal space* quantity as the Miller indices are actually derived from the fractional numbers $1/h, 1/k$ and $1/\ell$ denoting the intersections of the associated lattice planes with the unit cell axes. By applying a Fourier transform to Equation 1.3 the reciprocal space may be transformed to *real space* leading to the following relationship.

$$\rho(xyz) = \frac{1}{V} \sum_{hk\ell} |F_{hk\ell}| e^{-2\pi i(hx+ky+\ell z)+i\alpha_{hk\ell}} \tag{1.4}$$

As mentioned earlier the structure factor amplitudes are proportional to the square root of the intensity $|F_{hk\ell}|^2 \propto I_{hk\ell}$. However, the phase $\alpha_{hk\ell}$ in Equation 1.4 *cannot* be measured experimentally and methods are needed to estimate the phase. This is known as the *phase problem* in crystallography.

1.2.2 Solving the phase problem

Below, three methods to solve the phase problem in X-ray crystallography are discussed: *Direct methods* refer to mathematical methods that try to estimate the phases of the Fourier transform of the scattering density from the corresponding structure factor amplitudes alone. Another increasingly popular method in macromolecular X-ray crystallography is *molecular replacement*, which makes use of a similar structure to the structure one is trying to solve to obtain the initial phase estimates. The third family of methods exploit the scattering by one or more heavy atoms in the crystal to obtain the phases. These effects consider additional information from crystals containing a heavy atom that can be measured experimentally and are hence referred to as *experimental phasing methods*.

Direct methods

Direct methods encompass a broad range of mathematical relations between phases based on knowing the (normalized) structure factor amplitudes that stem from the fact that the electron density of the correct model must be greater than zero and that the structure is composed of discrete atoms (e.g. Giacovazzo 2008). An important concept in direct methods is that of structure invariants, which are linear combinations of phases that do not depend on the choice of origin and whose value depends solely on the crystal structure and may in principle be estimated from the structure factor amplitudes (e.g. Giacovazzo 2008). Currently, the vast majority of X-ray structures of small molecules are solved by direct methods. Based on the experience with solving a large number of such structures, Sheldrick (1990) concluded that direct methods rarely

lead to a solution unless at least half of the theoretically measurable reflections in a 1.1 to 1.2 Å range are observed. This empirical rule, which has come to be known as "Sheldrick's rule", was given a structural basis for proteins by Morris and Bricogne (2003) who argue that the rule has its origin in bonding distances typical of such molecules and the occurrence of families of inter-atomic distances that differ by 1.1 to 1.2 Å. Resolutions above 1.2 Å however are seldom attained in macromolecular X-ray crystallography. Moreover, the application of classical direct methods is limited to structures of several hundred non-hydrogen atoms. The advent of dual space iteration methods as implemented in SnB (Rappleye et al. 2002) and SHELXD (Schneider and Sheldrick 2002) allows larger structures up to a few thousand non-hydrogen atoms to be solved (Usón and Sheldrick 1999) and is enhanced by the presence of intrinsic metal ions or other heavy atoms (Weeks and Miller 1999). Nonetheless direct methods are of limited practical use in protein crystallography, with the exception of the determination of the heavy-atom substructure in experimental phasing methods, discussed in more detail in Section 1.3.1. Perhaps the combination of direct methods with other phase improvement procedures may relax the resolution restraints somewhat and allow larger structures to be phased in the future.

Molecular replacement

Molecular replacement (MR) leverages the abundance of structural information present in the PDB to solve a new structure. An initial electron density map is obtained by combination of the structure factor amplitudes of the unknown structure with the phase information of one or more known structures that are believed to have a similar fold. As a rule of thumb MR has a good chance of being successful if the unknown structure has more than 25 % sequence identity with the MR search model and their C^{α} atoms have a root-mean-square deviation smaller than 2.0 Å (Giorgetti et al. 2005; Taylor 2010).

The principal problem in molecular replacement is finding the correct orientation and position of the search model in the unit cell of the

new structure. Instead of exploring six-dimensional space for the correct solution, many programs split this problem into finding the right rotation and translation separately. Classically, these search methods employ the Patterson function shown in Equation 1.5.

$$P(uvw) = \frac{1}{V} \sum_{hk\ell} |F_{hk\ell}|^2 cos(2\pi(hu + kv + \ell w)) \tag{1.5}$$

Generating a Patterson map does not require knowledge of the phases and can simply be calculated from the squared structure factor amplitudes, which are proportional to the intensities. The peaks in a Patterson map correspond to interatomic vectors, that is for a crystal with N atoms in the unit cell the map has $N(N-1)$ maxima. This high-density of peaks and the partial overlap due to thermal vibration of the atoms makes it impossible to generate distance restraints for all atoms in case of larger structures like proteins. However the correlation of a Patterson map of the experimental data with Patterson maps of a MR search model in different orientations can be used to find the correct rotation (Crowther 1972; Rossmann 1990). Similarly, Patterson-based methods can be used to find the appropriate translation of the origin and using the self-rotation function non-crystallographic symmetry (NCS) can be identified. The phases calculated from the correctly oriented MR search model are used to generate an initial electron density map for the new structure. These phases can be subsequently improved by model building iterated with refinement.

The MR approach is so powerful that to date nearly two-thirds of the structures deposited in the PDB have been solved[1] by MR. As the size of the PDB continues to grow so will the probability that a given new structure can be solved by MR.

Experimental phasing methods

The use of heavy-atom derivatives was amongst the first methods employed to solve the phase problem in protein crystallography (Kendrew

[1] Molecular replacement was the structure determination method for 51 151 of the 82 022 entries in the PDB as of August 10th 2013.

et al. 1958; Perutz 1956). The method relies on the intensity differences between the native protein crystal and an isomorphous heavy-atom derivative to obtain the correct phase of the native protein structure.

Heavy atoms, such as lanthanides, (transition) metals, can be introduced into a protein crystal by soaking it in a solution containing a heavy atom salt. Frequently these heavy-atoms bind on well defined places in the protein. For instance Hg^{2+} ions bind the cysteine thiol groups and uranyl salts like UO_2NO_3 preferentially bind between the carboxyl groups of glutamic acid and aspartate. Other examples include Pb and $PtCl_4^{2-}$-ions that bind cysteine and histidine, respectively (e.g. Rould 1997).

The use of heavy atom solutions, however, carries one major practical downside. Many of substances containing heavy atoms are highly toxic, which in combination with the tendency of many of these compounds to accumulate in the human body, poses a health risk to exposed workers. Noble gases, such as Xe, are non-hazardous heavy atoms that can be introduced into the crystal in a high-pressure environment. Note, that there is no guarantee that an heavy atom can even be introduced in the crystal. In fact often crystals are seen to disintegrate in the heavy atom soaking solution. When the effect on the crystal packing is more subtle, poor diffraction compared to the native crystal can be the result.

The intensity differences arising from incorporation of a heavy atom are relatively easy to determine, considering a single Hg atom in a 1000 atom protein can give a fractional change in intensity of as much as 25 % (Crick and Magdoff 1956; Taylor 2010).

Defining the structure factor obtained from both heavy atoms and protein as the derivative structure factor F_{PH} and assuming it is a vector sum of the native F_P and the heavy atom F_H structure factor, an initial estimate of the heavy atom structure factor amplitude may be obtained with the following approximation: $|F_H| \simeq |F_{PH}| - |F_P|$. Using this estimate and the interatomic distance restraints that can readily be obtained from the large heavy atom peaks in the Patterson map, the positions of the heavy atoms can be calculated. Alternatively, the locations

of the heavy atoms in the unit cell may be obtained by direct methods (see Section 1.3.1). Once the heavy atom substructure is known the heavy atom coordinates and other parameters such as thermal B-factors and occupancy can be refined to give a more accurate estimate of the heavy atom structure factor F_H and consequently the phase of F_P.

In an error-free, idealized experiment, obtaining the native protein phase is a simple geometric problem that is solved by applying the cosine rule shown in Equation 1.6, which leads to two possible solutions for α_P that are distributed symmetrically about the heavy atom phase as is illustrated by the Harker construction shown in Figure 1.3.

$$\alpha_P = \alpha_H \pm \arccos \frac{F_{PH}^2 - F_P^2 - F_H^2}{2F_P F_H} \tag{1.6}$$

Isomorphism of the heavy atom derivative is a critical condition to the success of phase determination by isomorphous replacement that may not be met. A seemingly small change in unit cell dimensions, orientation of the protein in the unit cell, or other form of non-isomorphism result in intensity differences that interfere or completely mask the signal from the heavy atom(s) alone (Crick and Magdoff 1956). Non-isomorphism can be caused by the soaking process itself, that, in many cases causes shrinkage of the crystal by dehydration due to the difference in osmotic concentration between the mother liquor and the heavy atom solution. Furthermore. soaking can induce reorientation and changes in conformation of the protein, as well as exchange of salt-ions other than the heavy atoms.

In anomalous scattering wavelength dependent differences in scattering behaviour of certain atoms in the protein are used to obtain initial phase estimates. In a classical description of scattering X-ray photons are assumed to only change direction. In other words, the structure factor amplitude $|F_{hk\ell}|$ is taken to only depend on the positions and *not* on the scattering behaviour of individual atoms. This is the basis for Friedel's law, which states that $|F_{hk\ell}| = |F_{-h-k-\ell}|$. However, the assumption that photons only change direction when scattered is an oversimplified model. While a high-energy X-ray photon may be scat-

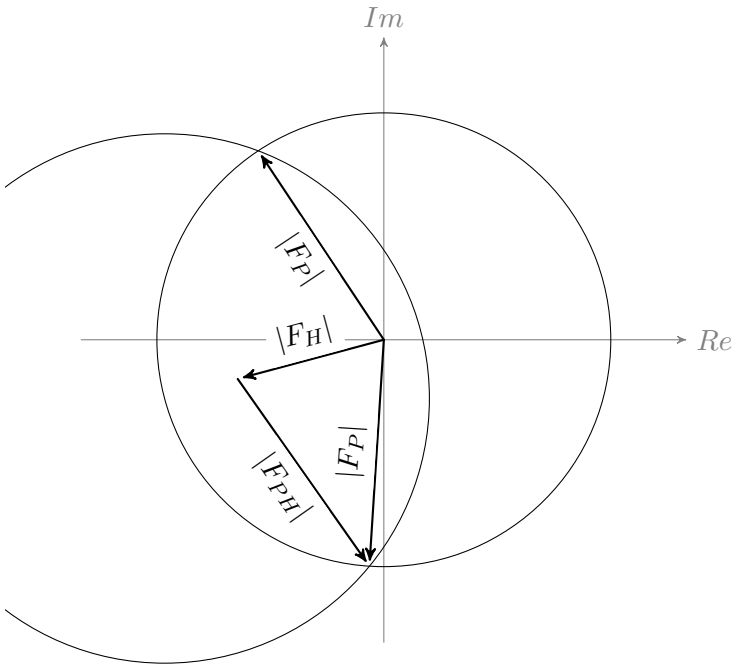


Figure 1.3: The Harker construction for a single isomorphous replacement (SIR) experiment. The circles indicate possible values for the phase. In the idealized case there are two possible choices for the phase, found where the circles intersect. However, in reality one needs to account for errors in the measurements and in the atomic parameters and the possible phase is smeared out.

tered normally when interacting with the electron cloud surrounding the atomic nucleus it may also be absorbed and promote an electron from an inner shell. The absorbed photon can be re-emitted at a lower energy, known as fluorescence; immediately re-emitted at the same energy; or retarded compared to the normally scattered photon. The latter can be modelled by the addition of an imaginary component to the photon's phase represented by the if" term in Equation 1.7.

$$f(\theta, \lambda) = f_0(\theta) + f'(\lambda) + if''(\lambda) \tag{1.7}$$

The scattering behaviour of individual atoms is given by the atomic scattering factor f shown in Equation 1.7. In the simplified description the wavelength dependent contributions f' and f'' can be ignored and the scattering factor only depends on the Bragg angle θ and Friedel's law holds.

The anomalous scattering effect is most noticeable and easy to measure close the absorption edge(s) where the absorption term f'' becomes large and consequently the difference between $|F_{hk\ell}|$ and $|F_{-h-k-\ell}|$ are heightened. Note that the term "anomalous scattering" can be considered a misnomer as the resonance effects on the atomic scattering factor are always present, but are generally to small to be accurately measured. As illustrated in Figure 1.4 the anomalous, or Bijvoet difference is similar to the isomorphous difference and can be used in the same way to locate the anomalous substructure and recover the native protein phase.

The sulphur SAD phasing method only requires the sulphur atoms present in almost any protein to obtain initial phase estimates and solve the structure. The crambin structure, solved in 1981 by Hendrickson and Teeter, was one of the first structures to be determined by the sulphur SAD phasing method. However, to date relatively few structures have been solved by using the anomalous differences from the sulphur atoms alone (Doutch et al. 2012; Ramagopal et al. 2003). There are several reasons for this apparent lack in progress. Typically synchrotron radiation sources are operated at a wavelength well away from the sulphur K-edge at 5.0155 Å. Even though modern quantum area detectors are substantially more sensitive than older detectors such as film and image plates, the Bijvoet difference is usually very small and the anomalous signal may be lost in detector noise, requiring long exposure times. On the other hand the small Bijvoet differences can easily be overwhelmed by the effects of radiation damage or overexposure (Dauter 2010). Data collection strategy programs like BEST (Bourenkov and Popov 2006), that estimate different measures of data quality from the first images of a diffraction experiment, may be helpful in optimizing collection of SAD data set.

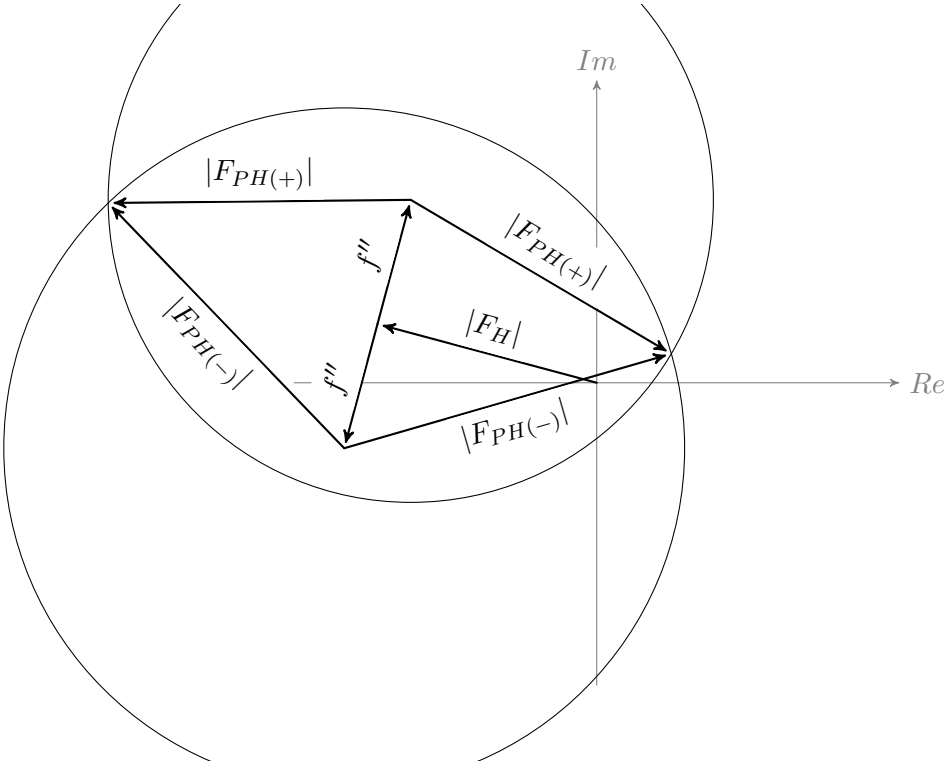


Figure 1.4: The Harker construction single-wavelength anomalous diffraction (SAD) experiment. In the presence of an anomalous scatterer Friedel's law breaks down. The f'' term is advanced 90° in phase, which gives rise to the Bijvoet difference $\Delta F^{\pm} = |F_{PH(+)}| - |F_{PH(-)}|$. If the measurement errors are ignored, the phase choices are given by the intersections of the phase circles.

Nowadays high-efficiency selenomethionine labelling protocols are available for bacterial, yeast, insect and mammalian cell expression systems (Barton et al. 2006). The major advantage of selenium over sulphur in SAD and multiple-wavelength anomalous diffraction (MAD) phasing is that the selenium absorption edge (0.9795 Å) is within 1.5 to 2.5 Å the typical range at a tunable macromolecular beamline (Doutch et al. 2012). Note that approximately 80 % of all proteins have a methionine content of 1 % or more (Strub et al. 2003), the percentage required for a successful MAD experiment (Hendrickson and Ogata 1997).

An idealized error-free SIR or SAD experiment would both result in two possible choices for the phase, only one of which is correct. This phase ambiguity cannot be resolved without introducing additional information. An earlier procedure to address the phase ambiguity is the resolved anomalous phasing approach, as used by Hendrickson and Teeter (1981) to solve the crambin structure, which uses the combined probability resulting from the partial structure and from anomalous scattering, unless the latter gives rise to a strong unimodal phase probability distribution. An alternative approach to choose the right set of phases is based on direct methods and employs the relations between large normalized structure factors to refine the initial phase estimates (Fan et al. 1990; Hauptmann 1982, 1996). Finally, the iterative single-wavelength anomalous scattering method introduced by Wang (1985) starts from the average phase and enhances meaningful features in the macromolecular crystal through iterative real-space smoothing of the electron density in the solvent region, much like ordinary solvent flattening in density modification (DM).

Another difficulty associated with phase determination by the SIR or SAD is that the initial phase estimates are often simply too inaccurate to generate an electron density map with interpretable features, or even one that could be improved by density modification. This inaccuracy has its origin in the experimental errors in the measurement of the structure factor amplitudes, errors due to scaling, non-isomorphism, errors in heavy atom parameters or even that the isomorphous and/or anomalous signal is not strong enough.

To account for these effects and define a phase probability Blow and Crick (1959) introduced the concept of lack of closure ε as graphically illustrated in the with the Argand diagram in Figure 1.5. Simply put ε is the difference between derivative structure factor amplitude $F_{PH(calc)}$, calculated from the assumed values of structure factor amplitudes for the native protein and heavy atom contributions, and the measured $F_{PH(obs)}$.

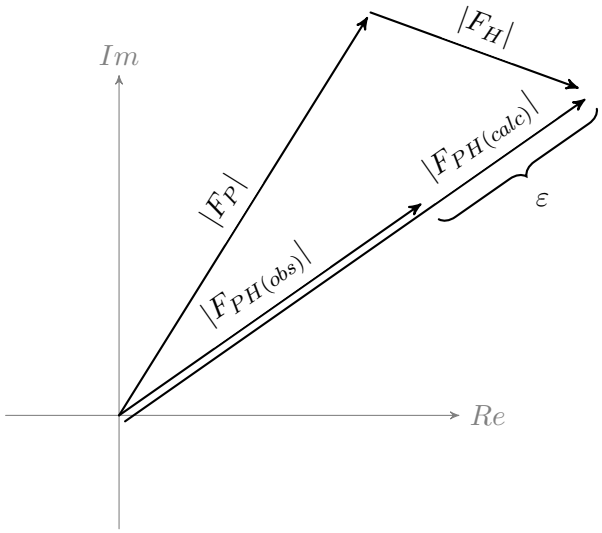


Figure 1.5: The lack of closure ε between the calculated derivative structure factor amplitude $F_{PH(calc)}$ and the observed $F_{PH(obs)}$ illustrated for a SIR experiment.

$$\varepsilon = |F_{PH(obs)}| - |F_{PH(calc)}| = |F_{PH(obs)}| - ||F_P|e^{i\alpha_P} + |F_H|e^{i\alpha_H}|| \quad (1.8)$$

In the case of a SIR experiment the phase probability distribution is obtained as shown in Equation 1.9. This assumes the measurement errors are contained in the derivative structure factor amplitude $F_{PH(calc)}$ and follow a Gaussian distribution (Blow and Crick 1959). The lack of closure variance E^2 is given by Equation 1.10.

$$P(\alpha_P) \propto e^{-\varepsilon^2/2E^2} \quad (1.9)$$

$$E^2 = \langle (F_{PH(obs)} - F_{PH(calc)})^2 \rangle \quad (1.10)$$

Collecting additional data, in the form of MAD data or several heavy-atom data set, is often worthwhile. As illustrated in Figure 1.6 for a MIR experiment, the combination of several heavy-atom data sets breaks the phase ambiguity by restricting the phase angle and sharpening the phase probability. Likewise, the phase information from the

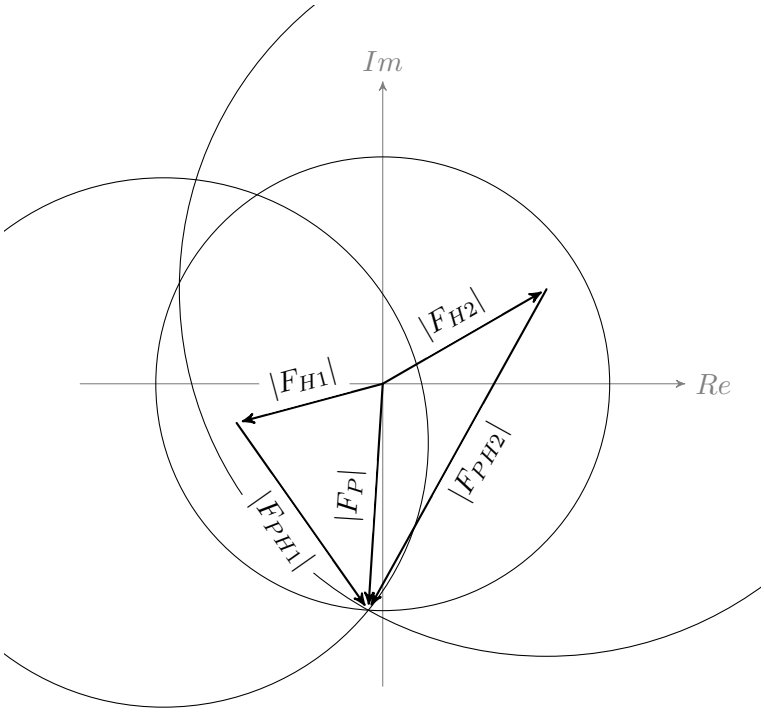


Figure 1.6: The Harker construction for a multiple isomorphous replacement (MIR) experiment. The additional heavy atom derivative restricts the phase from two in the case of SIR to a single choice.

anomalous differences from a heavy-atom derivative are used in a single isomorphous replacement with anomalous scattering (SIRAS) experiment to the same effect (Figure 1.7). However, the relative ease with which selenomethionine may be incorporated in the protein and the availability of macromolecular crystallography beamlines with tunable wavelength at many of the major synchrotrons has made collection of MAD data routine. More than 50 % of the experimentally phased structures deposited to the PDB in 2006 were determined by MAD (Barton et al. 2006) from selenomethionines or natively present metal atoms. MAD phasing makes use of the wavelength dependence of the dispersive f' and absorption f'' effect of the atomic scattering factor (see Equation 1.7), to provide distinct sets of possible phases. Similar to MIR

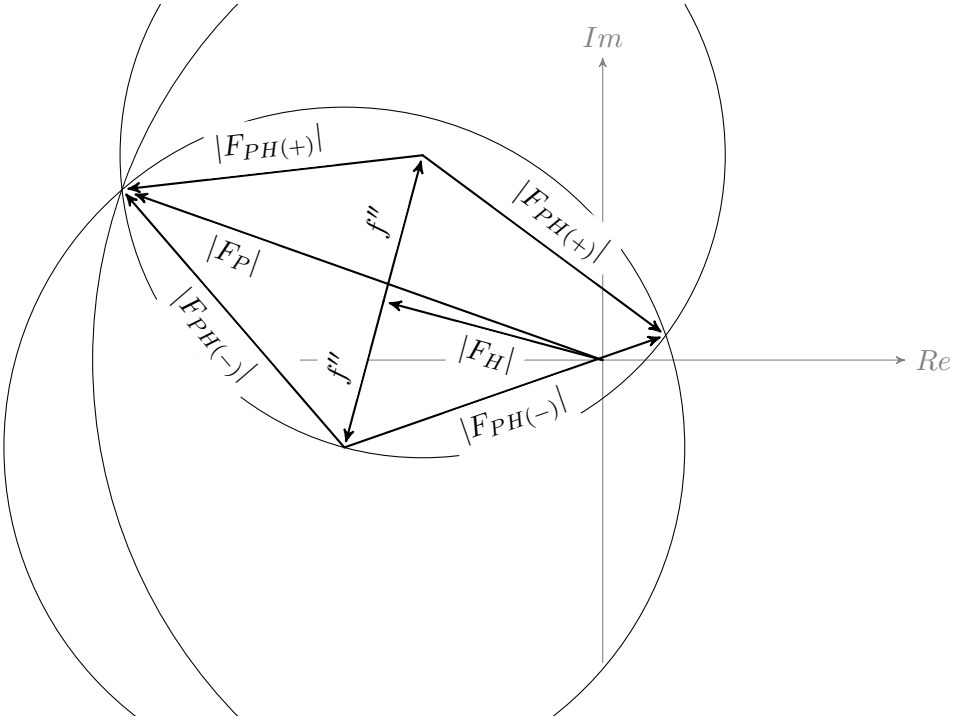


Figure 1.7: The Harker construction for a SIRAS experiment. The phase choices are restricted from two options for SAD to one choice by the combination isomorphous replacement and anomalous data.

the phase circles in a MAD Harker diagram have only one intersection. The added benefit of MAD is that the data can be collected on a single crystal.

1.3 Automation in X-ray structure solution

The growing need for automation in the structure solution process is fed by two trends in X-ray crystallography. Easy access to synchrotron radiation sources and the availability of commercial crystallization reagents, consumables and equipment have made protein X-ray crystallography available to an ever expanding audience of life scientists. Bundling and integration of individual software tools into graphical user interface (GUI) driven user friendly software suites, like CCP4 (Winn et al. 2011)

or Phenix (Adams et al. 2010), that use sensible default settings and a high degree of automation allow non-crystallographers to solve a crystal structure. Secondly, X-ray crystallography technology has matured to a stage where it is being deployed in structural genomics consortia across the world for high-throughput structure determination. The large number of diffraction experiments performed by such consortia call for automation of data processing to enable efficient evaluation of candidate data sets for structure solution.

Many research groups involved in development of software for macromolecular crystallography are attempting to provide fully automated programs for structure solutions by integrating MR strategies, pipelines for structure solution from experimentally obtained phase estimates, and increasingly combinations of both. Structure solution from phase estimates obtained from an anomalous scattering or isomorphous replacement experiment always follows the same basic separation of steps shown in Figure 1.8 even though the programs used may differ.

1.3.1 Substructure detection

Solving a structure by experimental phasing rests on the ability to successfully detect the anomalous substructure. Without it, calculation of substructure phases is impossible. Even if the phases recovered from the substructure cannot be extended to the full model automatically, it is clear that if substructure detection is successful there is some anomalous signal that may be exploited to provide better phase estimates, by for instance combination with phase information from MR or DM.

The Patterson function has an important application in MR as discussed in Section 1.2.2. However, it is also used to locate heavy atoms in anomalous scattering and isomorphous replacement experiments. Considering the height of a peak in a Patterson map is determined by the number electrons in an atom, atoms with a higher atomic number are relatively easy to detect. The substructure may be readily calculated from the distances vectors of the heavy atoms. The program Solve (Terwilliger and Berendzen 1999) implements Patterson based routines

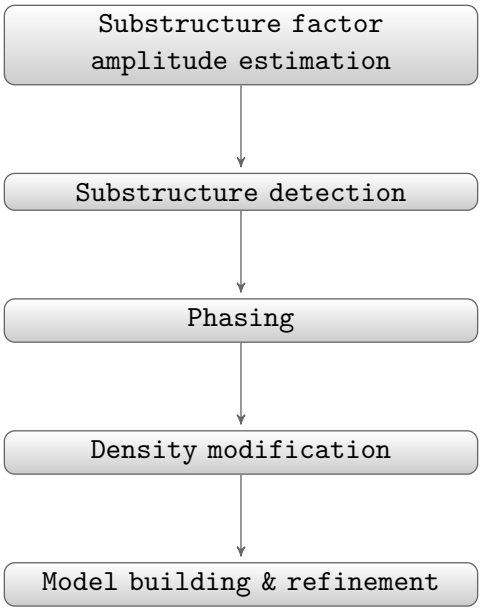


Figure 1.8: Typical steps in the (automated) solution of a macromolecular crystal structure by experimental phase determination. In contrast to phasing by molecular replacement the initial phase estimates are recovered from the anomalous scattering component, or the isomorphous difference between a native- and heavy atom derivative data set.

for determining the substructure for MAD and MIR.

Direct methods, first popularized by Hauptmann and Karle (1953) for phasing of small molecule X-ray crystal structures *ab initio*, have also found application in the determination of the heavy-atom substructure in macromolecular crystallography. The introduction of dual-space iteration, also known as 'Shake-and-Bake', by Miller et al. (1993) was a major advance in this area. First, starting phases are calculated from a set of atoms, that may either be randomly placed or whose position is constrained by for example agreement with the Patterson function. In the 'shaking' half of the method the starting phases are refined in reciprocal space either through minimisation of the minimal function (Debaerdemaeker and Woolfson 1983; Miller et al. 1993) or by using

the tangent formula (Karle and Hauptmann 1956). The map obtained by combination of the improved phases with the normalized structure factors is searched for the highest peaks that could be possible atoms. Several different approaches exist for the real space, or 'baking' part of the 'Shake-and-Bake' routine (Usón and Sheldrick 1999). The program SnB implements the dual-space iteration algorithm in its purest form, though most substructure detection programs, such as SHELXD, HySS (Grosse-Kunstleve and Adams 2003) and SIR2011 (Burla et al. 2012) employ a hybrid approach combining 'Shake-and-Bake' type procedures, other direct methods inspired approaches, as well as Patterson based procedures. A noteworthy exception is the program Crunch2 (de Graaff et al. 2001), which takes a different approach to 'Shake-and-Bake' that is based on Karle–Hauptman determinants (van der Plas et al. 1998).

1.3.2 Substructure phasing

Obtaining accurate initial experimental phases estimates is an important aspect of automated structure solution pipeline. Advanced maximum-likelihood methods offer substantially better models of an anomalous scattering or isomorphous replacement experiment and associated errors.

The univariate Gaussian phase probability defined by Blow and Crick (1959) was an important conceptual advance in thinking about phase estimation and modelling errors. Many researchers expanded on the lack of closure work using more advanced probabilistic models based on the maximum-likelihood formalism. The first step in this direction was taken by Otwinowski et al. (1991) who, in Mlphare (Collaborative Computational Project Number 4 1994; Otwinowski et al. 1991), applied the phase probability as a weight in heavy-atom refinement as well as during the structure factor calculation. In contrast to many others at that time who used only one phase for each reflection to estimate the heavy-atom parameters.

The maximum-likelihood based approach chosen in the program Sharp (Bricogne et al. 2003) numerically integrates over all possible

phases and the "true" structure factor amplitude which, for example, is not measured in a SAD or MAD experiment. While the program refines errors, scale and atomic parameters simultaneously with a Gaussian error in the isomorphic and anomalous terms, it does not consider correlations between data sets and errors. In contrast, the program BP3 (Pannu and Read 2004) directly considers the correlations between data sets and correlated errors for a SAD experiment, which has shown to be better over other approaches.

1.3.3 Density modification

The experimentally phased electron density map may initially be of insufficient quality to allow automatic tracing of a three dimensional atomic model. DM aims to improve phase estimates with an iterative procedure that cycles between real space incorporation of chemical knowledge and combination of the real space modified phases or phase distributions with the experimentally obtained phase probabilities in reciprocal space.

Figure 1.9 illustrates DM in its most traditional form, starting with the generation of the electron density map by Fourier summation of the structure factors calculated with the centroid phases from the experimentally obtained phase probabilities (see Equation 1.4). Three procedures frequently used to modify the resulting electron density map are solvent flattening (Leslie 1987; Wang 1985), NCS averaging (Muirhead et al. 1967) and histogram matching (Zhang et al. 1997).

Solvent flattening (Leslie 1987; Wang 1985) is one of the earlier density modification procedures and relies on the disorder of the solvent in a protein crystal. In theory the disordered solvent molecules do not give rise to constructive interference in the diffraction experiment and hence the electron density in the protein crystal solvent channels should be featureless and flat. However, at the start of DM the experimental phase estimates are not very accurate and the initial electron density map is noisy. This may make it difficult to distinguish the boundary between solvent and protein. Rather than looking at the absolute value of

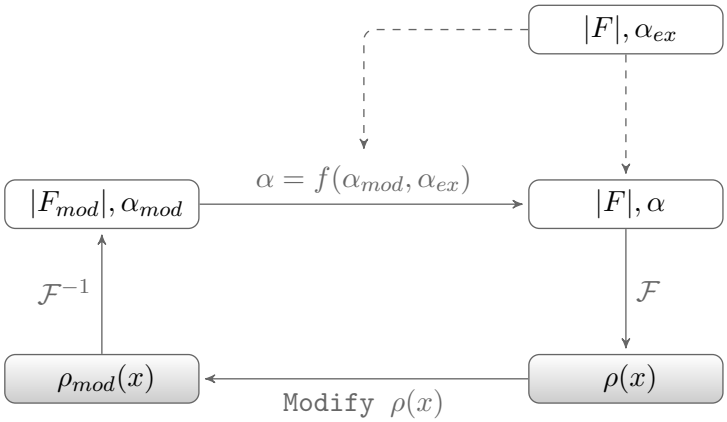


Figure 1.9: Steps in traditional density modification. Structure factor amplitudes, phases and electron density are indicated with $|F|$, α and $\rho(x)$, respectively. The updated phase probabilities are obtained by multiplication of the original phase probabilities with the modified phases weighted according to the agreement between original and modified structure factor magnitudes.

the electron density to distinguish solvent and protein, the solvent mask may be defined by looking at local variance throughout the unit cell. Indeed, the featureless solvent region will have a low variance compared to the regions with protein. In the flattening stage, the electron density in the solvent region is set to the average solvent density, effectively removing the noise not associated with any structural information.

In cases where there are several copies of a molecule or domain in the asymmetric unit, the multiple redundant samples of highly similar electron density may be combined to improve the signal-to-noise ratio. The usefulness of non-crystallographic symmetry (NCS) averaging in generating improved electron density maps has long been known (Muirhead et al. 1967). By comparison, histogram matching (Zhang et al. 1997) has been described only recently. The premiss in histogram matching is that the electron density values for a well phased map follow characteristic distributions as opposed to the electron density values of a randomly phased map that follow a normal distribution. By non-linear rescaling

of the electron density the histogram is made to look more like that of a well phased map. This suppresses negative electron density and tends to sharpen electron density peaks (Cowtan 2010).

Note that the application of scaling and masking functions in real space is equivalent to combining, in reciprocal space, many structure factors through a convolution. This is an intrinsic property of the relation between reciprocal and direct space by Fourier transformation. Consequently, the random error component of the structure factors will average out, whereas the true values of the structure factors will add up systematically.

After modification of the electron density map the modified structure factors, obtained by inverse Fourier transform, need to be combined with the experimental phase probabilities. Because there is no explicit modified phase probability distribution, one is generated by estimating the errors in the density-modified phases from the agreement from the between the observed and modified structure factor magnitudes. In the final stage of a DM cycle the probability distribution of modified phases and original experimental phases are multiplied to give an updated distribution. In the popular σ_A (Lunin and Urzhumtsev 1984; Read 1986) algorithm, the original experimentally determined phases, represented by Hendrickson-Lattman coefficients (Hendrickson and Lattman 1970), are combined with the density-modified structure factors through a heuristic weighting scheme (Read 1997).

There are two major problems associated with the traditional DM approach. All information about the phase probabilities is discarded by using only the centroid phases when calculating the electron density map. Furthermore, when multiplying probabilities care must be taken to ensure that the those probabilities are independent so as to avoid introducing bias. Yet, clearly the modified phase probabilities ultimately depends on the experimental phase probabilities through the DM cycle, resulting in ever narrower distributions without a real improvement of the phase estimates.

Several methods have been developed to reduce the correlation between the original and density-modified structure factors. Some of the earlier approaches include the reflection omit method (Cowtan and Main 1996), solvent flipping (Abrahams and Leslie 1996) and the γ-correction (Abrahams 1997) method. The γ parameter is an estimate of the contribution of the initial experimental structure factor to the density-modified structure factor in solvent flattening. By subtracting this contribution from the density-modified structure factor the correlation with the initial structure factor is suppressed. The γ-perturbation method (Cowtan 1999) is a generalization of this method for any type of density modification. Recently, Skubák and Pannu (2011) introduced the β correction parameter, a novel cross-validation method aiming to reduce bias in maximum-likelihood phase combination functions.

The statistical density modification techniques pioneered by Terwilliger (2004); Terwilliger and Berendzen (1999) and Cowtan (2000) are a different approach to reduce bias in DM. Rather than using a single modified map to express the *a priori* knowledge incorporated in the electron density map, statistical DM expresses the newly introduced information in terms of probability distributions, which are subsequently carried forward into reciprocal space illustrated in Figure 1.10. Since the density-modified phase probabilities are no longer estimated from the agreement between the observed and density-modified structure factor magnitudes, the link between the experimental and density-modified phase information is weakened.

Maximum likelihood allows for a rational incorporation of other sources of information. The use of maximum-likelihood in the phase-combination scheme described by Pannu et al. (1998) allows for incorporation of experimentally determined phase information in the form of Hendrickson-Lattman coefficients. The method has been shown to outperform the σ_A phase combination traditionally used in classical density modification (Cowtan 2010).

It is clear that the problem of a poor electron density map can be partially mitigated by using prior knowledge about the expected distri-

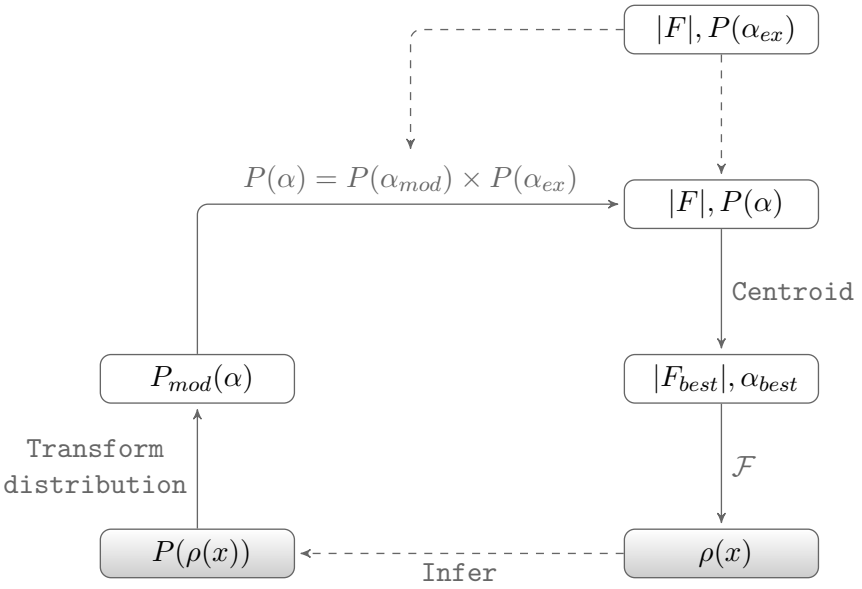


Figure 1.10: Steps in statistical density modification. Structure factor amplitudes and phases are indicated with $|F|$ and α, respectively. The *a priori* information inferred from the electron density, $\rho(x)$ is expressed as probability distributions, $P(\rho(x))$. These are carried forward into reciprocal space, $P(\alpha_{mod})$ and multiplied with the experimental phase probabilities, $P(\alpha_{ex})$ to obtain the updated phase probabilities, $P(\alpha)$.

bution of electron density and typical arrangement of a protein crystal and combining this density-modified phases with the experimental phases. Current methods assume that these sources of information are independent, yet it is clear that they are not, which results in model bias and escalated figures of merit. Chapter 3 describes a DM scheme based on a novel multivariate equation that does not assume independence of initial and density-modified phases. Large scale testing of the new DM method convincingly shows that this procedure generates dramatically improved electron density maps, allowing many more macromolecular structures to be traced without manual intervention from the user.

1.3.4 Automated model building and refinement

Once the diffraction data is collected, manual model building and refinement of a protein structure are the single most time consuming tasks in solving a protein structure. Automatic chain tracing has made the overall process of building a structural model of a protein substantially more efficient, which has benefited both high-throughput structure solution and structure solution in a non-automated setting. The computer programs ARP/wARP (Langer et al. 2008), Resolve (Terwilliger 2000, 2002), Buccaneer (Cowtan 2006) and SHELXE (Sheldrick 2002) are designed to build protein structures into electron density maps without user intervention. Although the implementation between the different programs differs, the basic premise is the same. Often the only input required is the amino acid sequence and a file with the structure factor amplitudes and phases. In the first step towards model building the electron density map is populated with a number of seed atoms. Subsequently, the C^{α} atom seeds are connected into chain fragments and individual amino acids or short peptides are docked into the fragments. The final stage often includes connecting the fragments into chains, building loops and resolving clashes.

To date the ARP/wARP program is the most comprehensive solution available, including high-level decision maker driven iterative protein model building, fast building of the secondary structure of a protein, tracing of flexible loops in alternate conformations, fully automated placement of ligands, and locating ordered water molecules (Langer et al. 2008). The program initially required high resolution data (Perrakis et al. 1999), but later developments (Morris et al. 2002, 2003) have relaxed the resolution requirements and the recent inclusion of a secondary structure recognition routine allows partial model building down to 4.5 Å (Langer et al. 2008).

The Buccaneer program combines a novel oriented electron-density likelihood target function for positioning C^{α} atoms with a several existing routines. By using the same likelihood function in different ways throughout the chain-tracing procedure has resulted in a program that

is relatively simple and very suitable for incorporation into automation pipelines (Cowtan 2006). The method is reasonably fast and performs well on data of moderate to low resolution (Cowtan 2006).

The program RESOLVE carries out model building in somewhat different, highly hierarchical fashion. Firstly, α-helices and β-sheets are located and fitted, followed by extension of these secondary structure elements with tripeptides, computed from a set of refined protein structures. In the final stage individual fragments are connected into chains (Terwilliger 2002). In addition to performing model building RESOLVE includes a likelihood-based density modification procedure, that yields better phase improvement than conventional procedures such as solvent flattening and histogram matching (Terwilliger 2000).

To improve a partial model and obtain the best electron density map, model building is often iterated with reciprocal space refinement. REFMAC (Murshudov et al. 2011) is a popular refinement program that uses different likelihood-based functions to best fit the experimental conditions and the availability of *prior* information, for instance the presence experimental phase information in the form of a SAD or SIRAS signal. The various restraints and model parametrizations allow customization of REFMAC to suit data quality. For instance, the use of secondary structure restraint or local and global NCS restraints can greatly benefit lower resolution data sets (Nicholls et al. 2012), whereas optional refinement of anisotropic B-factors can give the fullest description of a high resolution structure. The latter though is typically only used in the final stage of model refinement and not during the model (re)building stage.

The CRANK (Ness et al. 2004) program is an attempt to streamline and automate structure solution by experimental phasing methods and there are plans to include MR in the near future. For its release in 2004, CRANK had already shown to effectively detect and phase anomalous scatterers with only minimal user input and provide better quality solutions over existing programs. Currently, CRANK integrates programs in substructure detection, refinement, phasing, DM and automated model

building in a convenient and easy-to-use CCP4i graphical user interface that sit on top of a powerful scriptable back-end, though a high-level decision maker, for optimisation of pipeline selection and parameters to the different plugins, as found in programs such as AutoRickshaw (Panjikar et al. 2005) and Phenix is currently not implemented. Chapter 2 discusses the many enhancements incorporated since the first release of Crank that have made the program more robust and have significantly improved the structure solution process.

2

Recent advances in CRANK

Abstract

For its first release in 2004, CRANK was shown to effectively detect and phase anomalous scatterers from SAD data. Since then, CRANK has significantly improved and many more structures can be built automatically with single or multiple wavelength anomalous diffraction or SIRAS data. Here, we discuss the new algorithms we have developed that lead to the substantial improvements and show CRANK's performance on over one hundred real data sets. The latest version of CRANK is freely available for download at http://www.bfsc.leidenuniv.nl/software/crank/ and from CCP4 (http://www.ccp4.ac.uk/).

The work in this chapter was published in N. S. Pannu, W.-J. Waterreus, P. Skubák, I. Sikharulidze, J. P. Abrahams, and R. A. G. de Graaff (Apr. 2011). "Recent advances in the CRANK software suite for experimental phasing." In: *Acta Crystallographica, Section D: Biological Crystallography* 67.Pt 4, pp. 331–7. DOI: 10.1107/S0907444910052224

The author of this thesis implemented utilities for deployment of massive parallel tests on openPBS; implemented a generic framework for parsing log files and automated analysis of extracted statistics; performed minor debugging of the CRANK program support scripts; evaluated, summarized and presented the results of parallel tests.

2.1 Introduction

Currently, many software packages are available to automatically solve structures. The main aim of Crank (Ness et al. 2004) is to provide a user friendly and automated system incorporating the latest computational developments in all stages of structure solution by experimental phasing. Crank is not a monolithic system: users can define pipelines from a choice of many different programs. Figure 2.1 shows the current steps that Crank can perform and the programs that users can select to perform the task. The externally developed programs that Crank can interface with are SHELXC (Sheldrick 2008), SHELXD (Schneider and Sheldrick 2002), SHELXE (Sheldrick 2002), dm (Cowtan 1994), Parrot (Cowtan 2010), Pirate (Cowtan 2000), Buccaneer (Cowtan 2006), ARP/wARP (Langer et al. 2008) both of which iterate with Refmac (Murshudov et al. 2011) and Resolve (Terwilliger 2000, 2002).

We are the main authors of the programs Afro (Pannu et al., in preparation) for F_A calculation, Crunch2 (de Graaff et al. 2001) for substructure detection, Bp3 (Pannu and Read 2004) for substructure phasing, Solomon (Abrahams and Leslie 1996) for density modification, Multicomb (Skubák et al. 2010) for phase combination and co-authors of the program Refmac. These programs use multivariate maximum likelihood methods that allow the observed diffraction data and any current models to be considered simultaneously at any stage in the structure solution process. Thus, the wealth of information contained in the observed diffraction data can be used directly throughout the structure solution process and not approximated or ignored as current approaches do after constructing an initial electron density map.

Below, we provide a brief intuitive description of the novel methods in various steps in experimental phasing that we have developed since our first publication on Crank. We show the power of combining all of these new methods on over one hundred real single- and multiple-wavelength anomalous diffraction (SAD/MAD) and single isomorphous replacement with anomalous scattering (SIRAS) data sets run automatically with minimal user input in Crank.

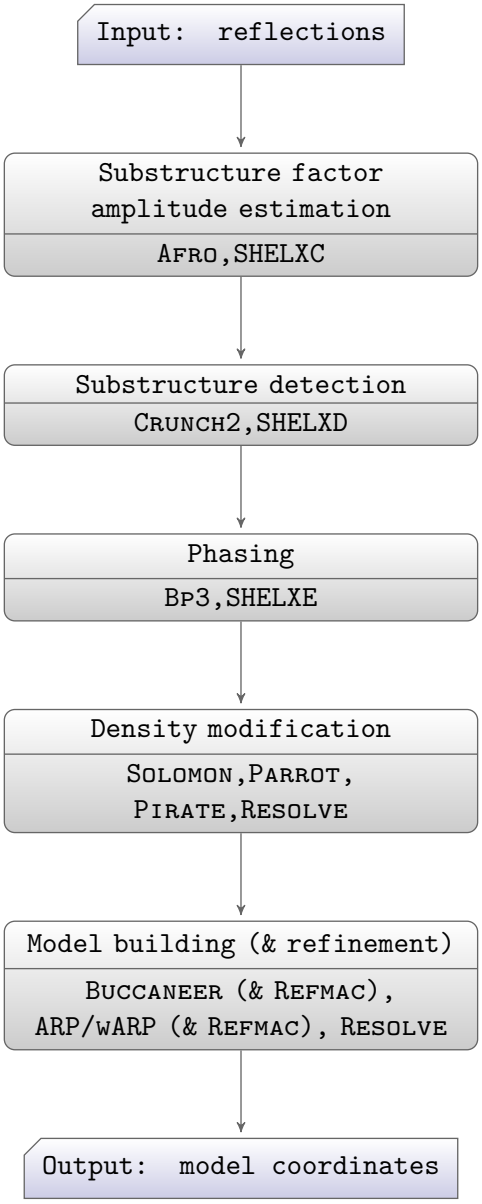


Figure 2.1: The steps in the structure solution process from substructure factor amplitude estimation to model building. The bottom half of each node shows the programs that Crank can interface with in that step. In the final phase the model can be refined by iterating either Buccaneer or ARP/wARP with Refmac.

The programs and methods we develop are not only available in Crank, but also AutoRickshaw (Panjikar et al. 2005) and ARP/wARP. Furthermore, the original methods we have developed have also been re-written in mathematically identical forms in both phenix.refine and phaser (Adams et al. 2010).

2.2 Recent developments in Crank

2.2.1 Substructure determination

After diffraction data has been indexed and merged, $|F_A|$ values are calculated for input to substructure detection programs. $|F_A|$ values are the amplitudes of structure factors corresponding to the heavy atoms to be located. For single-wavelength anomalous diffraction (SAD) data, most programs use the absolute value of Bijvoet differences, $\Delta F = ||F^+| - |F^-||$ as an estimate for $|F_A|$. Burla et al. (2002) proposed employing multivariate joint probability distributions to obtain the expected value for $|F_A|$ in an equation that contains three integrals. In order to obtain an analytical solution to the integrals, Burla et al. (2002) assume the "Bijvoet phases" are equal. We have obtained an expression requiring only one numerical integration without making this assumption. This approach has been implemented in the program Afro and performs satisfactorily. Details of the implementation and test results will be shown elsewhere (Pannu et al., in preparation). The development version of Afro containing the multivariate $|F_A|$ calculation is available in the latest version of Crank and can be used as input for either Crunch2 or SHELXD.

Within Crank, methods exist to validate whether a correct substructure has been determined and to terminate the substructure detection step early. If a threshold value for a statistic used by the substructure detection program has been reached or if a significant deviation exists between the best and worst score in different trials, the substructure detection program will successfully terminate before running all trials. Crank also provides an alternate and independent assessment

of whether a correct substructure solution has been located: an option exists to run the substructure phasing program Bp3 quickly in "check" mode and examine likelihood based statistics to determine whether a correct and complete substructure has been found. The statistic that Crank uses is a Luzzati (1952) parameter: if the average Luzzati parameter is greater than a threshold value (the default is 0.7), it is assumed that the full substructure has been found and substructure detection is terminated. Using likelihood methods to validate substructure detection has been available in Crank for over three years (Pannu et al. 2007) and this approach has been appreciated by phenix developers, who recently adopted it in their own suite (Paul Adams, CCP4 bulletin board, 31 July 2010).

2.2.2 Substructure phasing

To incorporate anomalous phase information, heavy atom refinement programs such as Sharp (Bricogne et al. 2003) or Mlphare (Collaborative Computational Project Number 4 1994; Otwinowski et al. 1991) use a Gaussian function on observed Bijvoet differences ($\Delta F = |F^+| - |F^-|$) centered on the "calculated" Bijvoet difference that is determined from an assumed value of the "true" structure factor and the heavy atom structure factor (Matthews 1966; North 1965). Since, in general, the "true" structure factor is not known for a SAD or multiple-wavelength anomalous diffraction (MAD) experiment, Sharp integrates out the amplitude and phase of the true structure factor. Furthermore,the estimate of measurement error for Bijvoet differences is determined by merging the measurement errors for Friedel pairs ($\sigma_{\Delta F} = \sqrt{\sigma_{F^+}^2 + \sigma_{F^-}^2}$), leading to suboptimal use of experimental information.

To input the observed structure factors directly, it is necessary to consider a joint probability of all observations given a current model. We have previously shown that this method provides better results over other approaches for the case of SAD (Ness et al. 2004; Pannu and Read 2004) as implemented in Bp3. We have recently shown that better results may be obtained by deriving a multivariate function for SIRAS

(Skubák et al. 2009) which will be released in the next version of Crank.

2.2.3 Density modification

In the density modification (DM) procedure, the density modified map is iteratively combined with the initial map obtained from experimental phasing. Current methods assume that these two maps are independent and propagate the initial map's phase information indirectly through Hendrickson-Lattman coefficients (Hendrickson and Lattman 1970). We have applied a multivariate analysis that considers the observed Friedel pairs directly for a SAD experiment, accounts for the correlation between the initial and density modified map and refines the errors that can occur in a SAD experiment. Results on many test cases show a significant improvement over the current state of the art (Skubák et al. 2010): the maps produced by the multivariate phase combination algorithm lead to many more structures being built automatically.

Despite the improvements in the quality of electron density maps, figures of merit remained escalated after DM. To obtain more accurate FOMs, we have recently developed and implemented a new cross-validated scheme for accurate error parameter estimation in likelihood based phase combination. The method leads to more reliable phase probability statistics from DM and results in a further improvement in subsequent model building. In addition, the more accurate FOMs enable a more reliable hand determination or identification of incorrect non-crystallographic symmetry (NCS) operators used in DM (Skubák and Pannu 2011). These developments have been implemented in a new phase combination program called Multicomb and can be used in conjunction with either Solomon or Parrot.

2.2.4 Automated model building and refinement

The incorporation of experimental phase information was previously shown to improve refinement (Pannu et al. 1998). However, the likelihood function developed, typically denoted maximum likelihood Hendrickson-Lattman (MLHL), propagates the external phase informa-

tion via Hendrickson-Lattman coefficients (Hendrickson and Lattman 1970). Thus, the MLHL function is dependent on the accuracy and reliability of the coefficients that are input. Furthermore, in its derivation, the MLHL function assumes that the experimental phase information (represented by Hendrickson-Lattman coefficients) is independent from the calculated structure factor. This assumption is questionable, as the experimental phase information is used to build an initial model. To overcome these issues, we considered and derived a multivariate likelihood function for SAD (Skubák et al. 2005) and SIRAS (Skubák et al. 2009) experiments. The likelihood functions take as input the diffraction data directly, the heavy atom coordinates and the calculated structure factors and accounts for correlation between them. Compared to the other likelihood functions in Refmac, more models are built automatically in ARP/wARP with the multivariate functions. The SAD and SIRAS functions in Refmac are available in Crank both in model building with ARP/wARP and Buccaneer.

2.2.5 Integration of programs and steps

To support the integration of the different programs it interfaces with, Crank has a plug-in architecture and communicates between plugins via an XML file. At the moment, there are two methods available to generate an XML file that Crank uses to run a pipeline: the program gcx (Ness et al. 2004) and a CCP4i graphical user interface. Both interfaces to Crank can be run with only minimal input: an mtz file with the relevant column labels specified, a sequence file and the name, expected number and f' and f" values for the heavy atoms. However, users can customize settings for individual programs, define custom made pipelines using any programs at each step and define the start and end step for a particular pipeline.

The program gcx allows Crank to be run from command line with a simple unix script: more information on it can be obtained from the program's documentation (`http://www.ccp4.ac.uk/html/gcx.html`). The test cases described below were run with gcx. Most users are

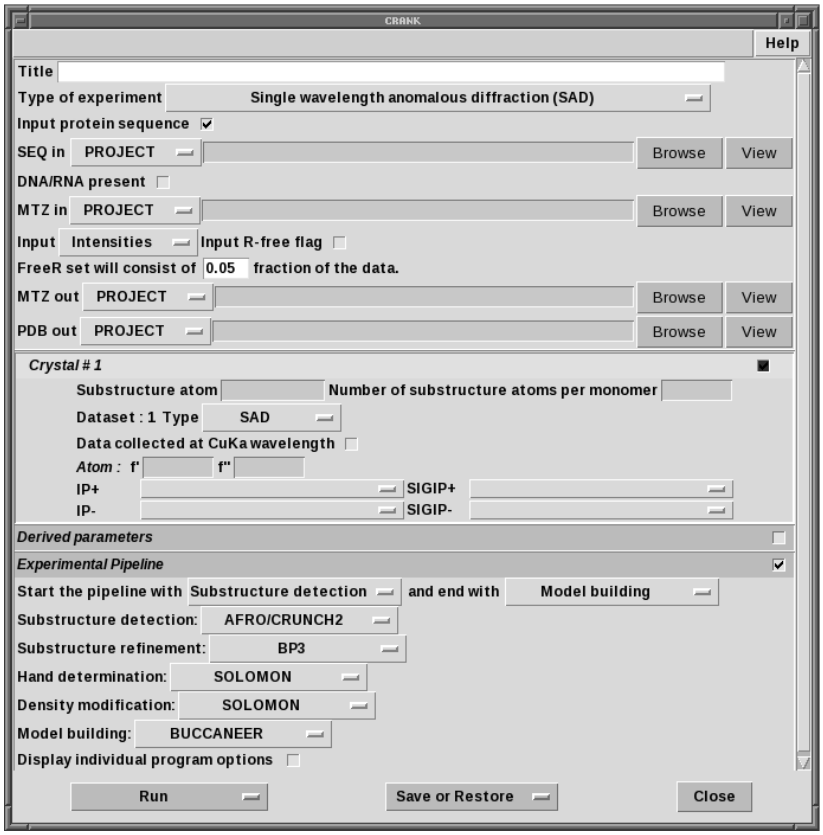


Figure 2.2: Screenshot of the CCP4i GUI for Crank showing the few required fields and the possibility to customize pipelines for automatic stucture solution.

likely to run Crank via the CCP4i GUI. The most convenient way to view a Crank log file is via the baubles system that can be initiated with the "View Annotated Logfile in a Web Browser" option in the CCP4i GUI. Documentation for Crank can be found at the CCP4 wiki (http://www.ccp4wiki.org/) which includes information on how to best interpret the log files.Figure 2.2 shows the CCP4i GUI with its few required fields.

2.3 Methods

Here, we test the new methods described above on a wide range of real SAD, MAD and SIRAS merged diffraction data sets. For our tests, only the intensities or structure factor amplitudes, along with the sequence for a protein monomer, the number of substructure atoms expected per monomer and the f' and f" values for the substructure atoms were input. Crank performed substructure detection using Afro and Crunch2, Bp3 for substructure phasing and Solomon with Multicomb was used for DM. Three cycles of Buccaneer iterated with Refmac were used for automated model building with iterative refinement. All options or parameters were default in all programs. The defaults set by Crank depend upon the particular experiment: for SAD data, Afro uses the multivariate $|F_A|$ value calculation, Multicomb uses the multivariate SAD function for phase combination in DM while Buccaneer uses the SAD function implemented in Refmac. For SIRAS data, Afro calculates $|F_A|$ from either the anomalous signal or using isomorphous differences by determining which signal is greater. Bp3 uses the uncorrelated SIRAS function described previously (Pannu et al. 2003), Solomon uses MLHL phase combination in Multicomb, while Buccaneer uses the multivariate SIRAS function in Refmac. Finally, for MAD data, Afro chooses the wavelength with the greatest anomalous signal and calculates multivariate F_A values from it. Similar to SIRAS data, Solomon uses MLHL phase combination in Multicomb to perform DM and Buccaneer uses the MLHL likelihood function in Refmac for model refinement.

In the test cases below, the previous version of Crank, 1.3, is tested with the current version, 1.4 The main differences between the two versions is the development version of Afro that calculates multivariate $|F_A|$ values given SAD data and the use of Multicomb for phase combination in DM which are both introduced in version 1.4.

In total, we report results from 116 real data sets from several different sources listed in Appendix A. The data sets provide a wide range of resolution (from 0.94 to 3.29 Å) and anomalous scatterers, including se-

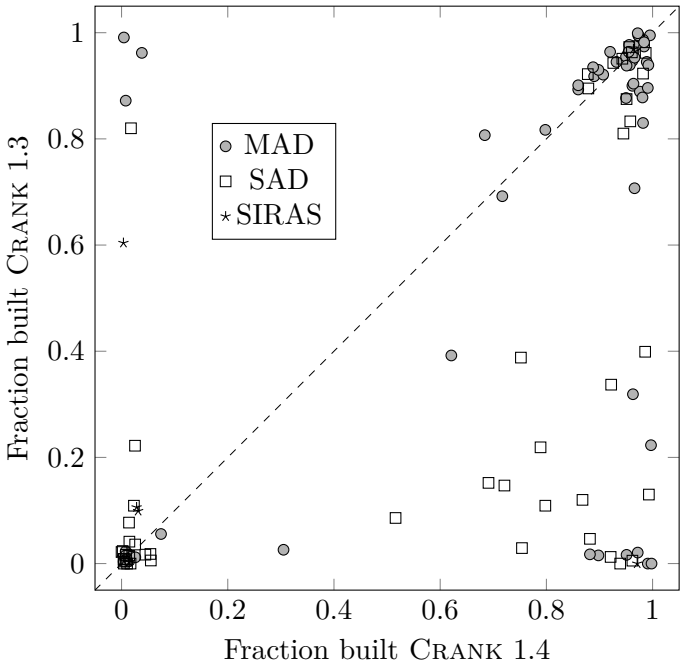


Figure 2.3: Improvement of Crank version 1.4 compared with Crank version 1.3 in terms of the fraction of the model built for SAD, MAD and SIRAS data sets. The fraction of the model built is defined as the fraction of the coordinates that are within 1 Å of the coordinates of the structure deposited in the PDB.

lenium, sulfur, chloride, sulfate, manganese, bromide, calcium and zinc. Of the 116 data sets, 63 are MAD data sets, 46 are SAD data sets, and 7 are SIRAS data sets.

2.4 Results and Discussion

Figure 2.3 shows the fraction of the backbone built within 1 Å of the final deposited structure for each of these data sets for the current version of Crank (1.4) versus the previous version (1.3). In total, 77 out of 116 structures have greater than sixty percent of the structure built correctly and of these 77 structures, 66 are built to over eighty percent completeness.

An example of an automatically built structure with a weak signal is GerE (Ducros et al. 2001). The structure of GerE was originally solved with a four wavelength seleno-methionine MAD data set collected at 2.7 Å and a native data set diffracting to 2.1 Å. Crank version 1.3 could build the structure from just the peak data set to a high degree, but failed to build the structure with just the SAD inflection data set. Crank version 1.4 can build the structure to a high degree using either the peak or inflection data set. We are unaware of any other automated package or collection of algorithms that can build GerE using either the peak or inflection data set automatically. To give an indication of the anomalous signal, Figure 2.4 plots the Bijvoet ratio (i.e. $\frac{|\Delta F|}{|F|}$) as a function of resolution bin for the GerE peak and inflection wavelength data: the overall Bijvoet ratios for the peak and inflection data are 0.167 and 0.139, respectively.

For the 77 structures that were built automatically, substructure determination successfully terminated early in 69 of the cases. For 33 of the 69 cases, the Luzzati parameter statistics in Bp3 allowed the early termination, while the remaining 36 cases the complete substructure was validated by an analysis of the Crunch2 statistics.

2.4.1 Analysis of data sets that were not automatically built

39 of the 116 data sets could not be built automatically by Crank. 19 of the 39 data sets failed at substructure detection and could be built automatically if the resolution cutoff in Crunch2 was changed or SHELXC and SHELXD was used in substructure detection. It should also be noted that the five cases that could not be built in version 1.4 but were successful in version 1.3 are all due to the changes in the substructure detection algorithm. These tests will be used to further debug and improve the development version of the multivariate $|F_A|$ calculation in Afro.

For 5 of the 39 cases, Crank in conjunction with a new SIRAS function for phasing leads to building when the current "uncorrelated"

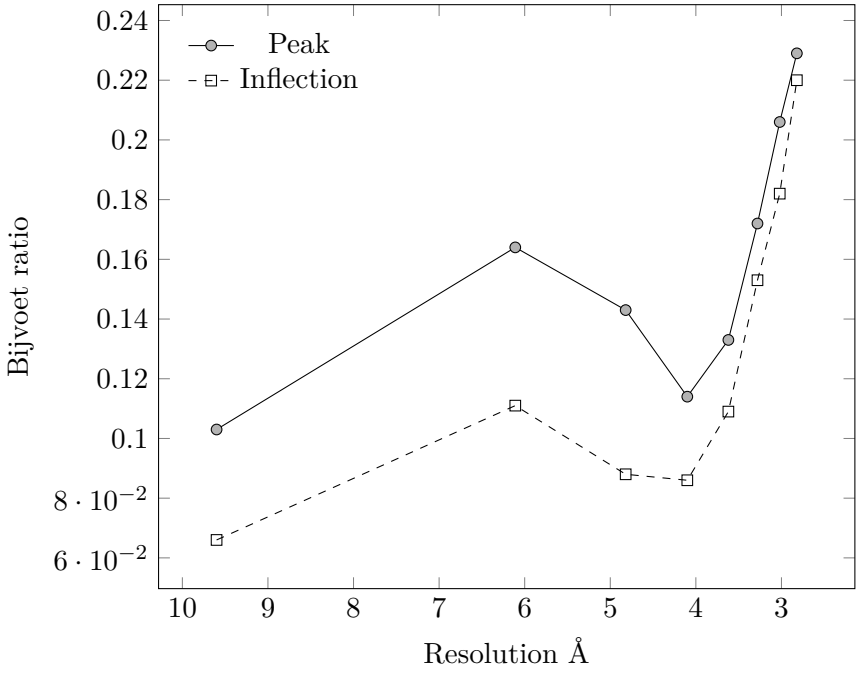


Figure 2.4: Plot of the Bijvoet ratios as a function of resolution for the peak and inflection wavelength data of the GerE test case. The peak and inflection wavelength are shown with a solid and dashed line, respectively.

function in Bp3 had failed to produce an automatically traceable map. The multivariate SIRAS function for phasing will be released in the next version of Crank.

The remaining 15 cases could not be built automatically or manually in Crank. For 7 of theses cases, Mueller-Dieckmann et al. (2007) had also failed to built the structures. Similarly, 4 other cases consisted of SAD experiments using derivative data sets from SIRAS experiments also containing a very weak signal. It is very likely that no currently available methods can build these structures and new methods need to be developed to build structures from such weak data. The remaining 4 cases that could not be built are from the JCSG repository: these structures can be built with currently available methods and the given data. The reasons why Crank fails to build these data sets are yet to

be determined.

2.5 Conclusions and future developments

Because of the new methods we have developed, Crank can build many more structures automatically and can build structures where current methods fail. Crank's robustness is shown by the large number of data sets we use in this test that require very minimal input.

Crank's CCP4i GUI is easy to use, but does have some limitations. Firstly, log files are only updated once a particular step in the pipeline has finished. Secondly, users can not manually stop a current step and proceed to a next step. Instead the pipeline can only be terminated and the Crank run must be restarted from the beginning. Furthermore, although, Crank has an interface to Coot (Emsley et al. 2010), it cannot show real time updates of a model as a Crank run proceeds. All of these shortcomings are being addressed and a new pyQt (http://www.riverbankcomputing.co.uk/software/pyqt/intro) interface for Crank is currently being developed in collaboration with CCP4.

Although having an easy to use and powerful interface is important, the first priority for Crank will always be developing better methods to solve data sets that elude current methods. For the case of MAD data, current approaches in Crank and elsewhere use univariate, uncorrelated likelihood functions for F_A calculation, substructure phasing and the MLHL function for DM and automated model building and refinement. Obviously, a multivariate MAD function could address the shortcomings in current approaches and could lead to structures where current methods fail.

In the case of SAD data, the multivariate functions used in substructure phasing, DM and model refinement only differ in the number of input variables and the parametrization. Although current algorithms separate these steps, the common mathematical framework suggests that all the information could be used simultaneously and combined optimally in a unified process using a single mathematical function, possibly resulting in substantial improvements.

3

A new method for phase combination

Abstract

DM is a standard technique in macromolecular crystallography that can significantly improve an initial electron density map. To obtain optimal results, the initial and density modified map are combined. Current methods assume these two maps are independent and propagate the initial map information and its accuracy indirectly through previously determined coefficients. We have derived a multivariate equation that no longer assumes independence between the initial and density modified map, considers the observed diffraction data directly, and refines the errors that can occur in a single wavelength anomalous diffraction experiment. We have implemented and tested the equation on over one hundred real data sets. Results are dramatic: the method provides significantly improved maps over the current state of the art and leads to many more structures built automatically.

The work in this chapter was published in P. Skubák, W.-J. Waterreus, and N. S. Pannu (June 2010). "Multivariate phase combination improves automated crystallographic model building". In: *Acta Crystallographica, Section D: Biological Crystallography* 66.7, pp. 783–788. DOI: 10.1107/S0907444910014642 * The first two authors contributed equally.

The author of this thesis implemented utilities for deployment of massive parallel tests on openPBS; implemented a generic framework for parsing log files and automated analysis of extracted statistics; evaluated, summarized and presented the results of parallel tests.

3.1 Introduction

The single-wavelength anomalous diffraction (SAD) experiment is a widely used technique to provide an estimate of the unknown structure factor phases and thus experimentally solve a macromolecular structure. To improve an electron density map obtained by SAD or another experimental method providing initial phase estimates, crystallographers employ density modification (DM) techniques. DM relies on prior information. For example, if the experimentally determined map is of sufficient quality to distinguish between regions of ordered macromolecule and disordered solvent, the electron density of the solvent region can be flattened (Leslie 1987; Wang 1985). This density modified map is then combined with the initial density map and the process is iterated as illustrated in Figure 3.1.

"Phase combination" or combining the original and density modified electron density map is important for the success of DM. Many currently used density modification programs (e.g. DM (Cowtan 1994), Solomon (Abrahams and Leslie 1996), SHELXE (Sheldrick 2002), CNS (Brünger et al. 1998)) employ the σ_A (Lunin and Urzhumtsev 1984; Read 1986) algorithm. In the σ_A algorithm, the original experimentally determined phases, represented by Hendrickson-Lattman coefficients (Hendrickson and Lattman 1970), are assumed to be independent from the density modified structure factors and combined with them through a heuristic weighting scheme (Read 1997). However, the density modified map is obtained from the experimental map, invalidating this assumption of independence. Methods have been developed to reduce the correlation between the original and density modified structure factors such as γ-correction (Abrahams 1997) or statistical density modification (Terwilliger 2000). Yet, the correlation is not considered explicitly and still remains.

Most density modification programs obtain prior phase information in the form of Hendrickson-Lattman coefficients. When these coefficients are used, it is assumed that errors in structure factors can be represented by a one-dimensional probability distribution over phases.

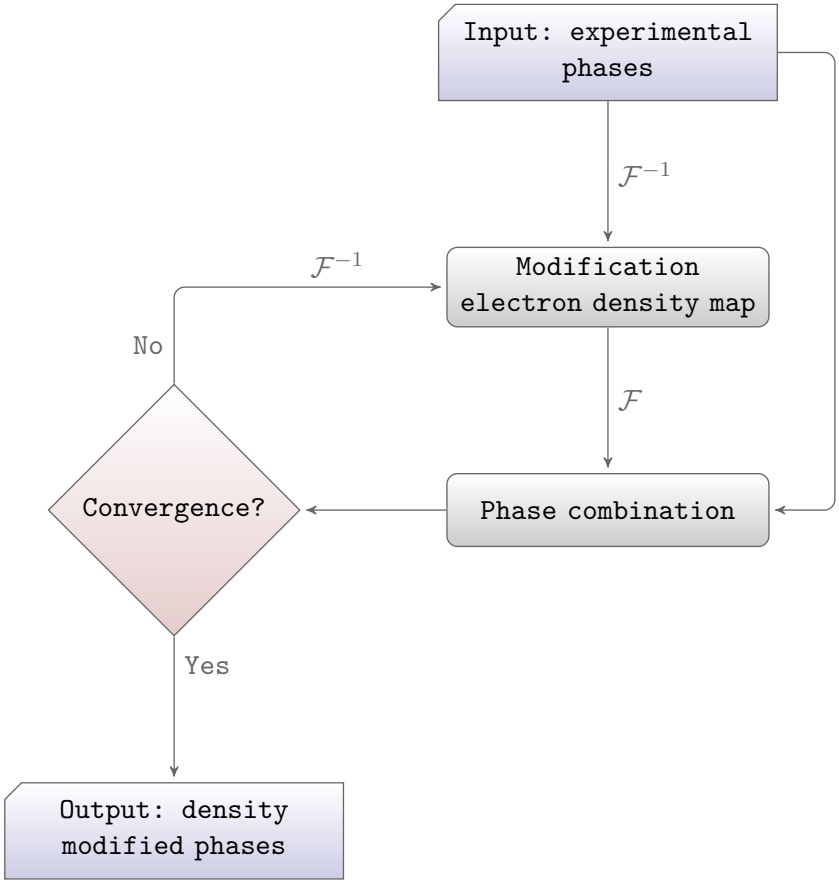


Figure 3.1: Flow chart of the different steps in density modification and phase combination.

Furthermore, any computer program that inputs Hendrickson-Lattman coefficients assumes that the program that generated them can provide a reliable and accurate estimate of the phase probability.

We have developed a multivariate probability distribution that no longer assumes independence between the initial and combined electron density map, but considers the correlation between the observed structure factor amplitudes, the density modified structure factor and a heavy atom substructure for a single wavelength anomalous diffraction experiment. The equation generates phase information directly from the heavy atom substructure and the observed diffraction data and thus does not require input of Hendrickson-Lattman coefficients. The equation also models and refines the errors in the density modified and heavy atom structure factors to provide an advanced multivariate model for a single wavelength anomalous diffraction experiment and phase combination. We have implemented this "*SAD-DM*" function in a new program called Multicomb (Skubák et al. 2010). We were motivated to consider these methods for phase combination since our previous work in applying similar multivariate methods to experimental phasing (Pannu and Read 2004) and in model building with iterative refinement (Skubák et al. 2005) produced better results compared to previous approaches.

3.2 Methods

The multivariate *SAD-DM* function used for correlated DM phase combination implemented in Multicomb is show in Equation 3.1. The observed Bijvoet/Friedel pairs from a SAD experiment are denoted by $|F^+|$ and $|F^-|$. $|F_{DM}|$, α_{DM} and $|F_{HA}|$, α_{HA} are the amplitude and phase for the density modified and heavy atom structure factors, respectively. Σ is the covariance matrix with its elements denoted by a_{ij} and Σ_2 is the bottom right submatrix of Σ with its elements denoted by c_{ij}. Equation 3.1 has previously been derived in Appendix A of (Pannu and Read 2004; Skubák et al. 2004) in the context of SAD heavy atom and protein refinement.

$$
\begin{aligned}
&P_{SAD\text{-}DM} = P(|F^+|, |F^-|; |F_{DM}|, \alpha_{DM}, |F_{HA}|, \alpha_{HA}) \\
&= \frac{2|F^+||F^-|\det(\Sigma_2)}{\pi \det(\Sigma)} \exp(-a_{11}|F^+|^2 - a_{22}|F^-|^2 - (a_{33} - c_{33})|F_{DM}|^2) \times \\
&\exp(-(a_{44} - c_{44})|F_{HA}|^2 - 2|F_{DM}||F_{HA}|(a_{34} - c_{34})\cos(\alpha_{DM} - \alpha_{HA})) \times \\
&\int_0^{2\pi} \exp(-2|F^-||F_{DM}|a_{23}\cos(\alpha^- - \alpha_{DM})) \times \\
&\exp(-2|F^-||F_{HA}|a_{24}\cos(\alpha^- - \alpha_{HA})) \times \\
&\mathrm{I}_0(2|F^+|\sqrt{(a_{12}|F^-|\cos(\alpha^-) + a_{13}|F_{DM}|\cos(\alpha_{DM}) + a_{14}|F_{HA}|\cos(\alpha_{HA}))^2} \\
&\overline{+(a_{12}|F^-|\sin(\alpha^-) + a_{13}|F_{DM}|\sin(\alpha_{DM}) + a_{14}|F_{HA}|\sin(\alpha_{HA}))^2})d\alpha^-
\end{aligned}
\tag{3.1}
$$

To compare MULTICOMB against the current SIGMAA (Read 1986) program, we used SOLOMON from CCP4 (Winn et al. 2011). The modular design of SOLOMON allowed an easy switch between MULTICOMB and SIGMAA for testing purposes.

For our tests, we started with merged diffraction data from a wide range of real SAD data sets. Only the intensities or structure factor amplitudes, along with the sequence for a protein monomer, the number of substructure atoms expected per monomer and the f' and f'' values for the substructure atoms were input into the CRANK (Ness et al. 2004) structure solution suite. CRANK performed substructure detection using either AFRO (Pannu et al., in preparation) and CRUNCH2 (de Graaff et al. 2001) or SHELXC (Sheldrick 2008), SHELXD (Schneider and Sheldrick 2002) and SHELXE. BP3 (Pannu and Read 2004) was used for substructure phasing and twenty cycles of density modification in SOLOMON were performed. Information about non-crystallographic symmetry was not used in density modification. Either three cycles of BUCCANEER (Cowtan 2006) or ten cycles of ARP/wARP (Langer et al. 2008) both iterated with REFMAC (Murshudov et al. 2011) (Murshudov et al. 1997) using the SAD refinement target (Skubák et al. 2004) were used for automated model building. A resolution cut-off or specifying the number of disulphide bonds was needed for successful substructure

determination for a few data sets. All other options or parameters were default in all programs. In the results shown below, for any data set, the changes shown were caused only by running either SigmaA or Multicomb with Solomon.

In total, we used 132 real data sets from several different sources listed in Appendix A. The majority of these structures were originally solved by multiple-wavelength anomalous diffraction (MAD), single isomorphous replacement with anomalous scattering (SIRAS) or molecular replacement (MR). Data sets where we could not determine the substructure or where a program within any pipeline terminated abnormally were excluded from the statistics presented, resulting in 102 data sets. In cases where multiple SAD data sets were available (i.e. collected either at different wavelengths or processed with different redundancies), the SAD data set corresponding to the highest f" value and to the highest data redundancy was used. The data sets provided a wide range of resolution (from 0.94 to 3.29 Å) and anomalous scatterers, including selenium, sulphur, chloride, sulphate, manganese, bromide, calcium and zinc.

3.3 Results

Figure 3.2 compares the quality of the electron density maps obtained by density modification using the σ_A algorithm with the maps obtained from the multivariate *SAD-DM* function for the 102 SAD data sets. The quality of a map is measured by its correlation with the map constructed from the deposited model. In general, the graph can be divided into three distinct regions. The first region is where the map correlation for both algorithms is under 0.2. This region shows test cases where the initial experimental phase information is too weak and results in an uninterpretable electron density map. In the second region, with map correlations over 0.8 after DM using both algorithms, all the points lie very close to the diagonal line, indicating that both algorithms performed similarly, producing high quality phases for these data sets. The remaining region shows the greatest variation between the two methods:

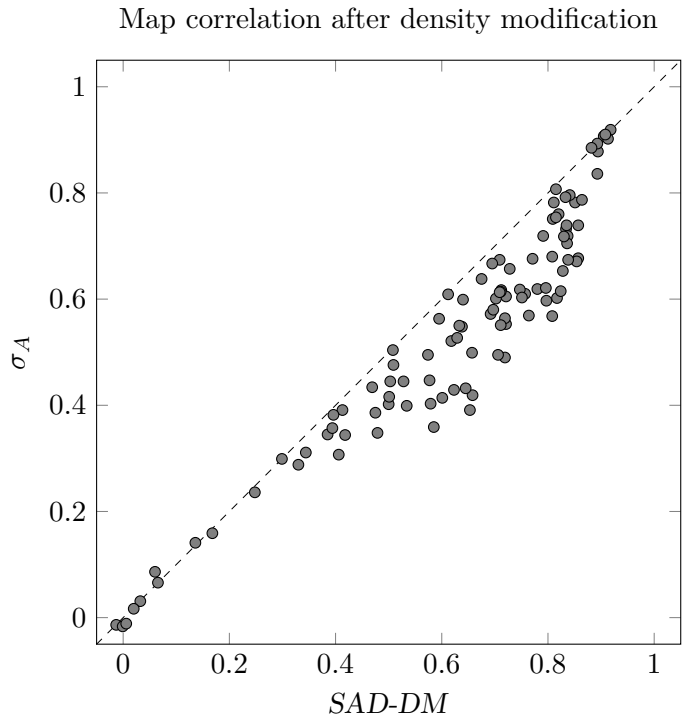


Figure 3.2: Map correlation after density modification compared to the final deposited models using the multivariate *SAD-DM* function (X-axis) and the σ_A algorithm (Y-axis).

in this region, there is a 21.4 % increase in average map correlation for the multivariate *SAD-DM* function. The overall increase in average map correlation calculated from all data sets is 16.9 % (from 0.539 to 0.630).

In Figure 3.3, the data sets are divided into two categories based on the improvement of *SAD-DM* over σ_A maps and shown as a function of data resolution and map correlation after experimental phasing. The figure indicates that when the input map correlation is below 0.3, usually neither method can significantly improve the map. Furthermore, when the input map correlation is above 0.4 and the data set has a 2.0 Å or better resolution, both algorithms usually produce equally high quality maps. A region where *SAD-DM* clearly outperforms σ_A are data sets with a resolution lower than 2.1 Å and an input map correlation higher than 0.35.

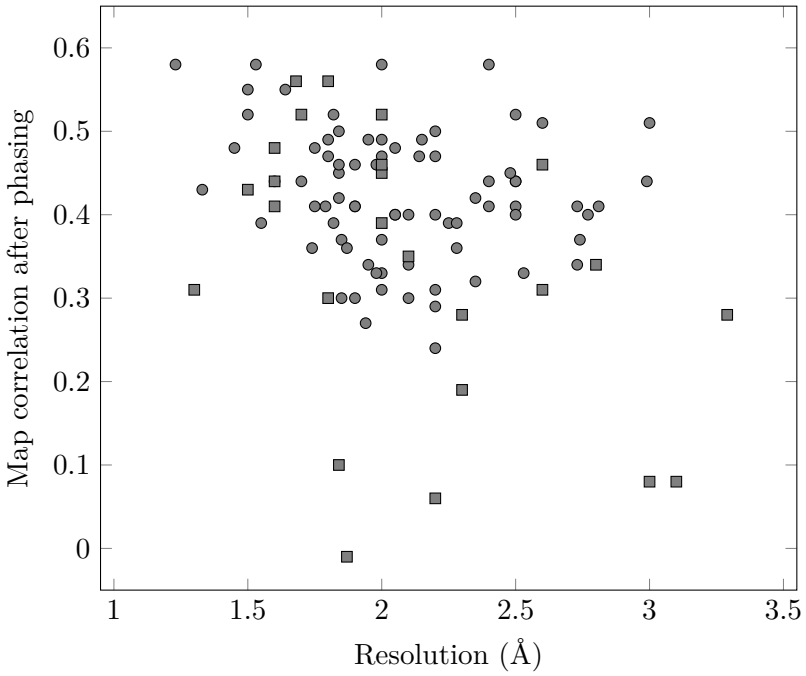


Figure 3.3: Map improvement of the *SAD-DM* algorithm over σ_A as a function of resolution (X-axis) and map correlation after experimental phasing (Y-axis). The data sets were divided into two groups based on the difference between the *SAD-DM* and σ_A map correlations with the deposited structure: the circles represent data sets where the *SAD-DM* map correlation is better than σ_A by 3 % or more, while a correlation difference less than 3 % is shown in squares.

Next, we examine the effect of the different phase combination algorithms on automated model building with iterative model refinement by using ARP/wARP (Perrakis et al. 1999) and Buccaneer (Cowtan 2006). Since ARP/wARP and Buccaneer use different protein tracing algorithms, examining the behaviour of both programs (Figure 3.4) provides a better comparison of the two different phase combination algorithms. The graph again can be divided into three distinct regions: the region where the fraction of the model correctly built is under 10 % indicating model building has failed for both phase combination algorithms, the region where both phase combination algorithms lead to a

Table 3.1: Average fraction of the model correctly built of all data sets for Buccaneer and ARP/wARP. A residue is regarded as "correctly built" if it is within 1 Å of the deposited model.

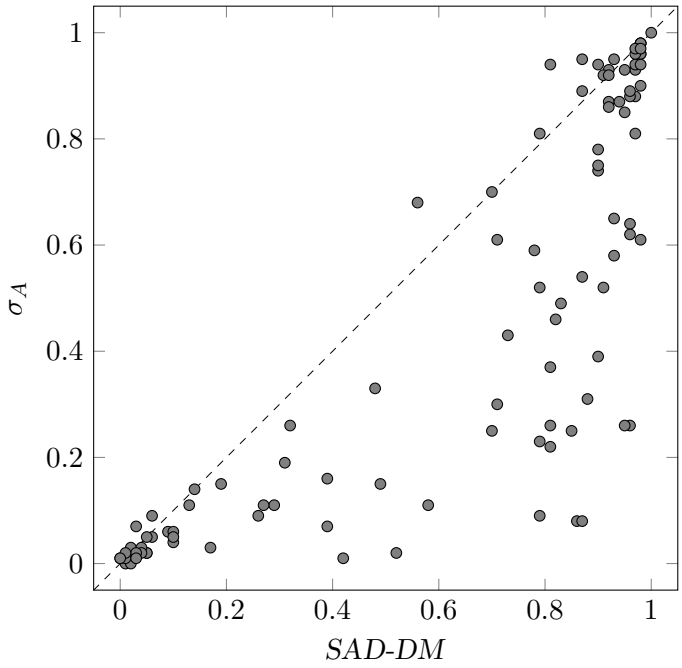
	σ_A	*SAD-DM*
Buccaneer	0.446	0.610
ARP/wARP	0.533	0.620

model that is over 80 % correctly built and the remaining region which shows the greatest disparity between the two phase combination methods. This last region includes 36 data sets where maps produced by the multivariate *SAD-DM* function result in a 59.6 % increase in the average fraction of the model built by ARP/wARP. The same region contains 50 data sets when using Buccaneer resulting in a 93.0 % increase in the fraction of the model built using multivariate methods over σ_A.

By comparing the quality of the electron density maps and the extent to which model building programs could build a model, the performance of the σ_A algorithm and *SAD-DM* function have been compared for 102 SAD data sets. The data summarized in Figures 3.2 to 3.4 and Table 3.1 are given in Table A.1 in Appendix A.

3.4 Discussion

The results clearly show that the multivariate *SAD-DM* function, in general, significantly outperforms the existing phase combination method in SigmaA. The improved map correlation provided by the multivariate *SAD-DM* function led to 16.4 % improvement of the overall average fraction built using ARP/wARP and 36.7 % improvement when using Buccaneer, compared to the state of the art (Table 3.1). The difference in the degree of improvement between the model building programs is mainly caused by the multivariate SAD refinement target used with ARP/wARP which enables successful model building after σ_A for many data sets that would otherwise fail: with the default Rice target, the ARP/wARP improvement is similar to that of Buccaneer (44 %).

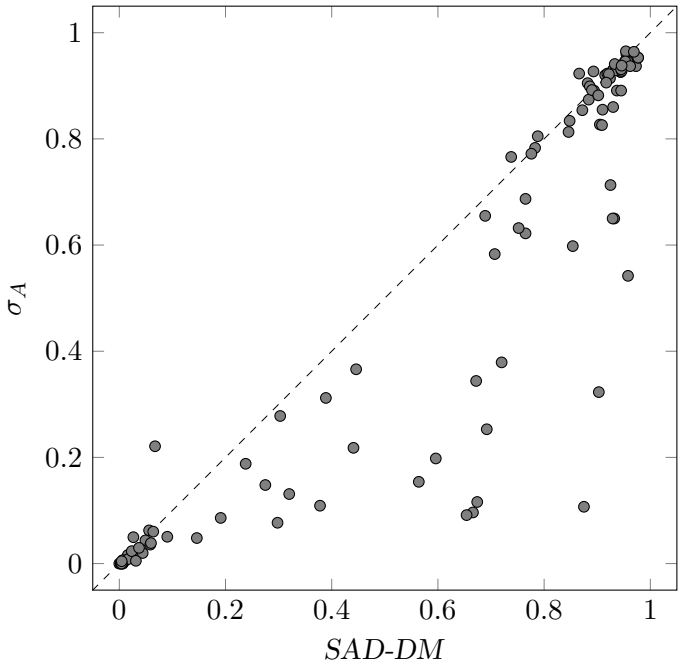


(a) Fraction built by Buccaneer

Figure 3.4: Performance of the multivariate *SAD-DM* function versus the σ_A algorithm in terms of fraction correctly built for the model building programs (a) Buccaneer and (b) ARP/wARP. The X-axis and Y-axis are the fraction of the model built after phase combination with Multicomb and SigmaA, respectively. A residue is regarded as "correctly built" if it is within 1 Å of the deposited model.

Although only SAD data was used in the above tests, the multivariate framework laid out is by no means limited to only a SAD experiment. Indeed, we have already implemented a multivariate SIRAS function for phase combination and initial results are very promising (Skubák and Pannu, in preparation). Furthermore, a multivariate MAD function is currently being implemented.

The introduction of the multivariate function to phase combination has important consequences for automated structure solution and manual structure solution at lower resolutions. The same multivariate function described above is used in substructure phasing, density modification and model refinement. Thus, at any step in the process up



(b) Fraction built by ARP/wARP

Figure 3.3: Continued

to or including substructure phasing, if a correct anomalous or even non-anomalous portion of the structure is added, significantly improved phases can be generated via simultaneously considering phasing, density modification and model refinement together. In essence, the successful implementation of multivariate methods to phase combination should lead to a new paradigm where the crystallographer no longer considers phasing, density modification, and model refinement as separate parts of structure determination, but one process using one mathematical function that can lead to combined and significant improvements over considering them separately.

4

Structural analysis of DNA binding protein Sso10a2

Abstract

Archaeal chromatin proteins play an important role in functionally organising the genomic DNA into a compact nucleoid. *Sulfolobus solfataricus* expresses a small dimeric protein Sso10a, that is thought to play an important role in chromatin organisation and dynamics. The dimer is constituted of a highly stable coiled-coil and two winged helix DNA binding domains. The two other Sso10a family members, Sso10a2 and Sso10a3, expressed by *S. solfataricus* are highly similar yet their biological roles are poorly understood. To gain insight into the molecular differences between Sso10a and Sso10a2 we crystallized the Sso10a2 protein and solved the crystal structure to 2 Å resolution. Superposition of the Sso10a and Sso10a2 crystal structures reveals significant divergence in the winged helix DNA binding domains.

The author of this thesis performed the final purification of the Sso10a2 protein by gel filtration; crystallisation of the Sso10a2 protein and diffraction analysis of its crystals at the ESRF; solution, description and interpretation of the Sso10a2 structure.

4.1 Introduction

The extreme conditions under which the hyperthermophilic Crenarchaeon *Sulfolobus solfataricus* and other extremophiles live pose unique challenges to maintaining DNA duplex stability and genomic organisation. To this end crenarchaeal species express several highly abundant small DNA binding proteins, which *bend* or *wrap* the DNA or *bridge* duplexes to fold the DNA into a compact structure known as the nucleoid (Driessen and Dame 2011). In addition confining the genome to the cell volume, these DNA compacting proteins play important roles in regulating essential cellular processes, such as DNA repair, replication and transcription (Driessen and Dame 2013). In contrast to their euryarchaeal counterparts who express true tetrameric histone homologues that wrap DNA into structures reminiscent of eukaryotic nucleosomes (Pereira et al. 1997), the small DNA organising proteins expressed by Crenarchaea are more similar to those of bacteria (Driessen and Dame 2011). Throughout the domains of life the amino acid sequence similarity between chromatin proteins is minimal. Despite this apparent lack of homology, the architectural properties of chromatin proteins are highly conserved (Luijsterburg et al. 2008). Knowledge of the structure of these proteins may reveal the mechanisms of molecular recognition that explain how *bridgers* engage in long range interactions to promote higher-order genomic organisation and how *wrappers* and *benders* exert their effect on the DNA structure locally. Furthermore, molecular models of bacterial DNA organising proteins, and in particular the differences with their eukaryotic counterparts, may enable rational design of new antibiotics that act by specifically disrupting bacterial DNA compaction.

S. solfataricus DNA binding proteins separate into three groups according to their size (Dijk and Reinhardt 1986). The 7 kDa class that includes the well studied Sso7d (Agback et al. 1998; Baumann et al. 1994; Choli et al. 1988; Edmondson and Shriver 2001; Knapp et al. 1996; Lundbäck et al. 1998; Mai et al. 1998; Robinson et al. 1998; Shriver et al. 2001), the recently discovered Sso7c4 (Hsu and Wang 2011) and the

Cren7 protein (Guo et al. 2008); the 8 kDa class about which little is known other than that is composed of the Sso8a and Sso8b species (Hsu and Wang 2011; Teale et al. 2003); and lastly the 10 kDa group that contains Sso10a, Sso10b and Sso10b2 (Dijk and Reinhardt 1986; Forterre et al. 1999).

Insights into non-specific DNA binding in *S. solfataricus* is primarily based on the protein–DNA complexes of Sso7d (Gao et al. 1998; Su et al. 2000) and Cren7 (Feng et al. 2010). The Sso7d protein strongly binds DNA in a sequence non-specific manner by docking into the minor groove. The residues involved in the interaction with the DNA are located in a three stranded β-sheet. Binding is accompanied by intercalation of hydrophobic side chains, which cause a single sharp bend of approximately 60° (Gao et al. 1998). The formation of the complex is accomplished by several salt-bridges. However, non-electrostatic interactions, primarily carried by hydrogen bonds, contribute significantly to the free energy of binding (Gao et al. 1998). Interestingly, some of these hydrogen bonds occur through bridging waters, which act as a filler to accommodate different base pairs. This mechanism may in part explain the sequence generality of DNA binding (Gao et al. 1998).

The Cren7 protein is a highly conserved crenarchaeal chromatin protein, that bends double-stranded DNA and constrains negative supercoils (Driessen et al. 2013; Feng et al. 2010). Cren7 and Sso7d share many functional characteristics and have a similar fold. Both compact DNA molecules in a comparable manner and to approximately the same extent (Driessen et al. 2013). Like Sso7d, Cren7 has a DNA binding surface composed of a three stranded β-sheet that contains positively charged, as well as hydrophobic residues (Feng et al. 2010). The crystal structure of the Cren7 protein in complex with double-stranded DNA shows that intercalation of hydrophobic side chains induce a bend of approximately 60° (Feng et al. 2010), which is agreement molecular dynamics simulations of the protein–DNA complex (Driessen et al. 2013). A notable difference with Sso7d is the presence of the $\beta 3 - \beta 4$ loop that interacts strongly with the DNA, as evidenced by the observations that the loop

becomes more rigid upon DNA binding in solution (Guo et al. 2008), and that deletion of this loop gives a sixty fold reduction in DNA binding affinity (Feng et al. 2010). Interestingly, several sites in this loop may be methylated (Guo et al. 2008) and both methylated and unmethylated forms of a lysine in this loop have been reported *in vivo*, suggesting that DNA binding may be regulated by methylation (Feng et al. 2010).

Sso10b, also known as Alba (acetylation lowers binding affinity) (Bell et al. 2002), is a well-characterized conserved archaeal protein that tightly binds double-stranded DNA in a sequence non-specific manner. The protein draws its name from the post-translational acetylation on Lys16 by the acetyltransferase Pat that lowers the affinity for DNA (Marsh et al. 2005). Conversely, a homologue of the eukaryotic histone deacetylase Sir2 can deacetylate acetylation lowers binding affinity (Alba) and mediate transcriptional repression *in vitro* (Bell et al. 2002). The structure of Sso10b (Wardleworth et al. 2002) and its smaller isoform Sso10b2 (Chou et al. 2003) have been determined by X-ray crystallography and models for the protein–DNA complex have been put forward that suggest an important role in DNA binding for Lys16 and Lys17 in Sso10b and the corresponding residues Lys12 and Lys14 in Sso10b2. However, in the protein–DNA co-crystal structure of Ape10b2, a Alba2 homologue from *Aeropyrum pernix* K1, the conserved Lys14 – which corresponds to Lys17 and Lys14 in Sso10b and Sso10b2, respectively – has been shown to have only an indirect role in coordination of bound DNA and seems primarily involved in stabilizing the interaction between individual Alba2 dimers (Tanaka et al. 2012). DNA binding by Alba2 is effected by positively charged and polar residues located in the monomer-monomer interface that pack into the opposite successive minor grooves of the DNA (Tanaka et al. 2012). Crystal contacts brought about by hydrophobic interactions between the two α-helices between Alba2 dimers suggest a possible mechanism for DNA compaction by bridging of strands (Tanaka et al. 2012). This is in line with the single molecule manipulation studies of Laurens et al. (2012) who, in addition to DNA bridging, report stiffening of the DNA by cooperative side-by-

side biding of Alba dimers.

Sso10a, the last member of the 10 kDa class, was identified by its homology to the Sac10a protein from *Sulfolobus acidocaldarius*. Electron microscopy studies with reconstituted Sac10a–DNA complexes suggest that Sac10a binds DNA in a sequence non-specific manner and can introduce supercoiling (Lurz et al. 1986). The DNA binding properties of Sso10a itself are poorly characterized, however, the 1.47 Å crystal structure of Sso10a dimer (Chen et al. 2004) contains two copies of the winged helix fold which is observed in many DNA binding proteins. The winged helix fold is a compact α/β structure consisting of a helix-turn-helix DNA binding motif and two antiparallel β-strands that form the wing ($w1$), arranged in order $\alpha1 - \beta1 - \alpha2 - \alpha3 - \beta2 - w1 - \beta3 - w2$. In many winged helix proteins a second wing ($w2$) is formed by the C-terminus of the DNA binding domain, however in the Sso10a dimer the C-terminus of each chain is and extended α-helix. These helices interact to form a very stable antiparallel left-handed coiled-coil. Based on homology with $\gamma\delta$resolvase, Chen et al. (2004) suggest that Sso10a binding of DNA is primarily mediated by insertion of $\alpha3$ of the helix-turn-helix DNA binding motif in the major groove of the DNA with only minor roles in coordination of DNA binding by insertion of $w1$ and N-terminus into the outer and inner minor groove, respectively. Nonetheless, the winged helix fold is an extremely versatile motif with distinct DNA binding modes (Gajiwala and Burley 2000), making predictions about DNA binding highly speculative.

The Sso10a protein has two close homologues which, while showing strong sequence conservation, have shown differential DNA binding behaviour in preliminary experiments. Here we describe the X-ray crystal structure solution and compare the global structure of Sso10a2 with that of the previously solved Sso10a (Chen et al. 2004; Kahsai et al. 2005) to identify residues and structural elements that may be involved in DNA binding.

4.2 Methods

4.2.1 Cloning and overproduction

The sequence coding for the Sso10a2 protein was amplified from *S. solfataricus* P2 strain genomic DNA by PCR with the forward primer 5' ATA CAT ATG CAA CTT GAA CGG CGT AAA AGA GGA ACA ATG G 3' and reverse primer 5' GGT GGA TCC TTA CTC ACT GTC TGT TCT TAA AAG TTC ACT AAC 3' and subsequently cloned into the NdeI and BamHI sites of the pET11a cloning vector via a classical restriction–ligation procedure. After sequence verification selenomethionine substituted Sso10a2 was expressed in a methionine auxotrophic derivative of *Escherichia coli* BL21 as described by Sohi et al. (2000). Briefly, *E. coli* were grown in minimal medium containing $50\,\mathrm{mg\,L^{-1}}$ selenomethionine and $30\,\mathrm{mg\,L^{-1}}$ ampicillin to an OD600 of 0.4 upon which Sso10a2 expression was induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of $120\,\mathrm{mg\,L^{-1}}$. Cells were harvested after 2 h by centrifugation.

4.2.2 Purification

The cell pellet was resuspended in 20 mL 50 mM Tris (pH 8.0), 2 mM $MgCl_2$, 0.1 % v/v Triton X-100, 3 mM benzamidine, 10 mM 2-mercaptoethanol. After sonication to promote cell disruption and DNA shearing, 1000 units of Omnicleave™ endonuclease (Epicentre) were added per gram wet cell weight and the DNA was digested by incubation at 293 K for 30 min. By heating the sample to 343 K for 40 min non thermostable proteins will denature and aggregate. Following cooling down of the sample, EDTA and NaCl were added to a final concentration of 20 mM and 100 mM respectively. The soluble faction obtained by ultracentrifugation of the cell lysate was filtered with a 0.45 µm filter (Filtropur S, SARSTEDT) to remove residual aggregates.

The clarified lysate was loaded onto a 1 ml RESOURCE S (GE Healthcare, Little Chalfont, United Kingdom) cation-exchange column equilibrated with 10 mM K_2HPO_4/KH_2PO_4 (pH 7.0), 100 mM NaCl,

10 % v/v glycerol and 5 mM 2-mercaptoethanol and bound material was eluted with a linear gradient of 0.1 M to 1 M NaCl. The pooled peak fractions were dialysed against 10 mM K_2HPO_4/KH_2PO_4 (pH 7.0), 100 mM NaCl and loaded onto a 1 mL HiTrap Heparin HP column (GE Healthcare), equilibrated with the same buffer and eluted with the same gradient.

The fractions containing Sso10a2, kindly provided by Dr. Nora Goossen[1], were pooled and the residual impurities removed by purification on a Superdex 200 gel-filtration column (GE Healthcare). The column was equilibrated with 20 mM Na_2HPO_4/NaH_2PO_4 pH 7.3, 150 mM NaCl and resolved at a flow rate 0.8 mL min^{-1}.

At every stage in the purification the peak fractions were examined on 17% tricine-SDS-PAGE gels stained with Coomassie Blue.

4.2.3 Crystallisation

Sso10a2 was concentrated to 10 mg mL^{-1} with an 3 kDa molecular weight cutoff centrifugal filter unit (Millipore). Sitting drop vapour diffusion crystallisation trials were set up using the JCSG+ and PACT screens purchased from Qiagen (Hilden, Germany). The reservoir solution (75 µL) was dispensed in MRC Crystallisation 2 drop plates (Molecular Dimensions, Newmarket, United Kingdom)) using a Genesis RS200 liquid handling robot (Tecan, Männedorf, Switzerland)). Dispensing of 0.5 µL drops was performed with an Oryx robot (Douglas Instruments, East Garston, United Kingdom). The protein to reservoir solution ratio of the drops were 1:1 and 3:1, respectively. Over the course of 30 days, crystals appeared in several conditions. Those crystals confirmed to contain protein by UV fluorescence microscopy using a U-MWU2 filter (Olympus, Tokyo, Japan) and were considered for X-ray diffraction analysis.

[1] Faculty of Science, Leiden Institute of Chemistry, Molecular Genetics, Gorlaeus Laboratories, Einsteinweg 55, 2333 CC Leiden

4.2.4 X-ray diffraction analysis

Crystals were picked up in cryoloops and directly flash-cooled in liquid nitrogen for those conditions containing known cryoprotectants. Crystals from conditions without cryoprotectant were briefly soaked in a solution composed of 10 % v/v glycerol in mother liquor prior to flash-cooling. X-ray diffraction experiments were conducted on ID 14-1 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. Crystals grown in PACT condition no. 71, consisting of 1 M succinic acid, 0.1 M HEPES pH 7.0 and 0.1 % w/v PEG MME 2000 (final pH 7.0) in a protein to reservoir solution ratio 3:1 showed strong diffraction. 378 images were collected at a wavelength of 0.933 Å with an oscillation angle of 0.95° and an exposure time of 7 s per frame at 100 K. The diffraction images were processed with XDS (Kabsch 2010), excluding the reflections around the 2.249 Å ice-ring. Scaling and merging was done with Aimless (Evans 2011) and the Crank (Ness et al. 2004) suite was used to obtain an initial model with the following toolchain: Afro (Pannu et al., in preparation), Crunch2 (de Graaff et al. 2001) and Bp3 (Pannu and Read 2004) for substructure detection and refinement. This was followed by density modification with Parrot (Cowtan 2010) using the novel phase combination algorithm described in Chapter 3, and Buccaneer (Cowtan 2006) in combination with Refmac (Murshudov et al. 2011) for model building iterated with refinement. Manual adjustments and completion of the model were performed in Coot (Emsley et al. 2010), also iterated with translation libration screw-motion (TLS) and restrained refinement in Refmac with automatically generated TLS groups and non-crystallographic symmetry (NCS) restraints. Both Parrot and all refinement cycles in Refmac were run with direct phase restraints (Skubák et al. 2004, 2010). All programs except XDS and Coot were run from the CCP4i graphical user interface. Data reduction and refinement statistics are shown in Tables 4.1 and 4.2, respectively. The atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank under PDB identifier 4HW0.

4.2.5 Analysis of the crystal structure

The THESEUS (Theobald and Wuttke 2006) program was used for superpositions and root-mean-square deviation calculations. Prior to superposition with theseus the primary sequences were aligned with CLUSTALW (Larkin et al. 2007), which was also used to align different homologues of Sso10a2. Crystal packing and possible macromolecular assemblies were examined using the PDBe PISA (Krissinel and Henrick 2007) web interface. The PYMOL (Schrödinger LLC 2010) package was used to create views of Sso10a2 and superpositions with homologous structures.

4.3 Results and Discussion

4.3.1 Sequence analysis

The alignment of the Sso10a2 amino acid sequence (Figure 4.1) with those of its paralogs Sso10a (Q97VX8), Sso10a3 (Q97UL8) and several similar sequences from crenarchaeal species shows that the first twenty-six residues are not present in any of the homologous proteins. Furthermore, a BLAST (Altschul et al. 1990) search with these twenty-six residues did not uncover any sequences with significant similarity and it is unclear whether the full-length Sso10a2 protein or transcript exists. A N-terminal α-helix predicted with the PSIPRED (Buchan and Ward 2010) server for secondary structure prediction server, is not essential for DNA binding or protein stability, as we have found that the Sso10a2 protein truncated to Met27 to Glu130 is stable in solution and binds DNA in both electrophoretic mobility shift assay (EMSA) and tethered particle motion (TPM) experiments (results not shown).

α1
β1
α2
α3
10
20
30
40
50
60
70
Q97V10_SULSO MTYGYNIFTVIKFSDNVYKLQLIYYSMQLERRKRGTMEIMFDILRNCEPK...CGITRVIYGAGINYVVAQKYLD 72
Q97VX8_SULSO .....MAKKRSKLEIIQAILEAC.KSG..SPKTRIMYGANLSYALTGRYIK 43
Q97UL8_SULSO .....MR..RTRDDIISDILEAIDSKI..NRISSIMKNVNLSASLAKKYIS 42
Q4J7Y2_SULAC .....MSEMPKHRSKIEIMYTILKVLRESETEVKKTHLMYLVGLNYDVFSKYIS 49
Q4JAD3_SULAC .....MMAEVTTGRKKRTRYEIIHDILAQC.ENG..AKKTWLMYKANLSYELTNNYIN 50
A4YG75_METS5 .....MTTKRRGKLEIYLDILSIC.AMG..GKKTEIMYRANLSYDLLNKYLN 44
A4YH26_METS5 .....MRSRRTSIEIIADILEVVSSGS..GSKSSIMKNANLSEGLTEKYLS 44
A4YG66_METS5 .....MMIRRTRQEIVLSMLEGCSGIG..TNKTNLMYKSGLNYDSFLRYLN 44
Q973Y2_SULTO .....MSKRLKRTSLEIIYSILGACNNN...STKTRIMYNAGVNLLELTKYLE 45
Q973B3_SULTO .....MRSKRDKFDIVADILENIKDGR..NSKSALMKNANLSFSILKKYID 44
F9VNH4_SULTO .....MKVKFNKKKRTDYEIIYDILNALKDGP..IPKTRLIYRAGLTYLTATKYIP 49
Q970P2_SULTO .....MRISFNKRKRTEYQIMFQILDALSRGP..LPKTKLMYKADLTYVVTEKYLP 49
Q974K1_SULTO .....MTEAQVGRKKRTRYEIIHDILSQC.ENG..AKKTWLMYKANLSYELTNNYIN 49
C3MSQ1_SULIM .....MRVERRKRGTMEIMLDILRSCDPE...CGVTRVIYGAGINYSVAQKYLD 46
C3MV70_SULIM .....MHKKRSSIEVIVSILEACNKGD..IKKTRIMFITYLNPRSFYRYIE 44
C3MUB9_SULIM .....MAKKRSKLEIIQAILEAC.KSG..SPKTRIMYGANLSYALTGRYIK 43
C3MSF1_SULIM .....MR..RTRDDIIGDILEAIDSNI..NRISSIMKNVNLSASLAKKYLS 42
C3MVN3_SULIM .....MIEN.RRRTKTEIIYSILRGCSEG...SKKTRLMYLAKLNYSVFMKYIY 45
C3MVP1_SULIM .....MSRKRNRMEIVANILEGC.KGG..IKKTRLMYKTNLSYRLFCKYTR 43
consensus ****!****!* *!!*** ****!* **!** * *!*

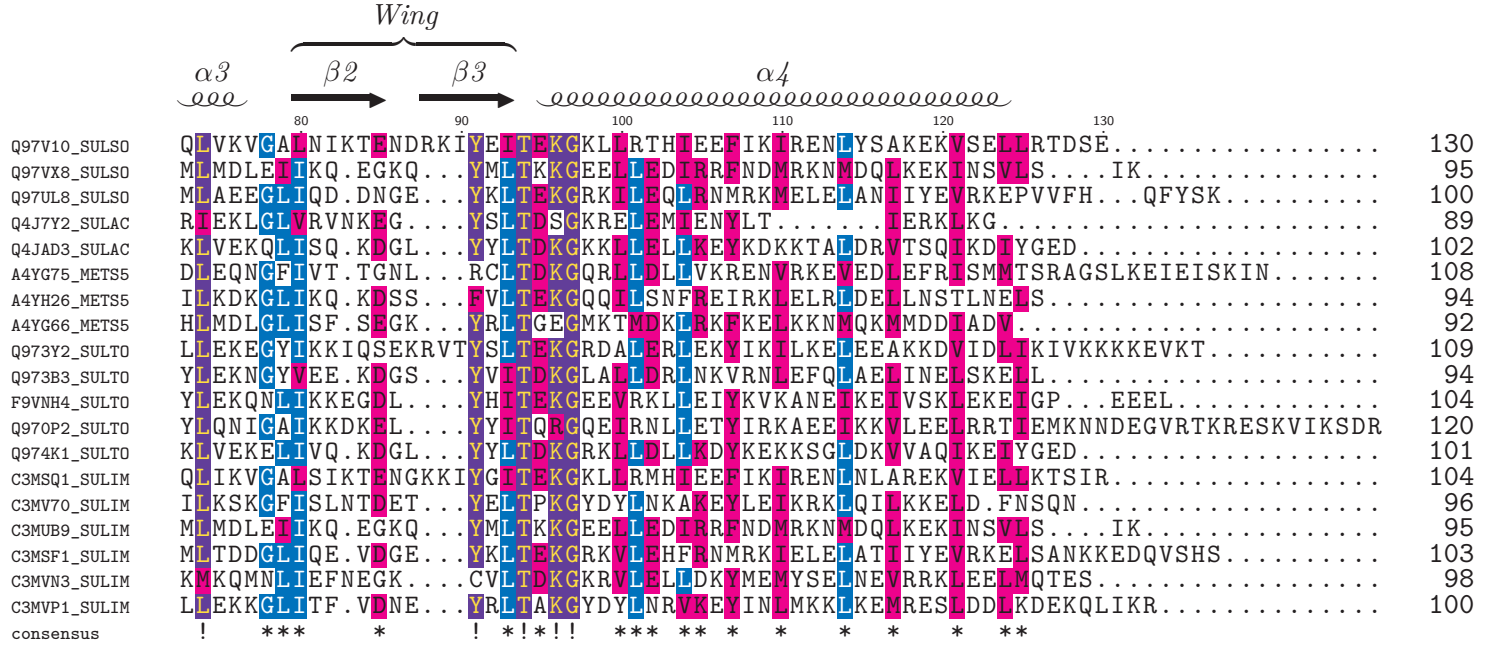


Figure 4.1: ClustalW sequence alignment of Sso10a (Q97VX8) archaeal homologs, including Sso10a2 (Q97V10) and Sso10a3 (Q97UL8) from *Sulfolobus solfataricus*; Sac10a (Q4JAD3) and a paralog (Q4J7Y2) from *Sulfolobus acidocaldarius*; similar sequences from *Metallosphaera sedula* (A4YG75, A4YH26 and A4YG66), *Sulfolobus tokodaii* (Q973Y2, Q973B3, F9VNH4, Q970P2 and Q974K1), and *Sulfolobus islandicus* (C3MSQ1, C3MV70, C3MUB9, C3MSF1, C3MVN3 and C3MVP1). The secondary stucture of Sso10a2 is depicted in the first line. Violet and blue shading denote a consensus of more than 80 % and 50 to 80 %, respectively. Residues that are chemically similar, but not identical are shaded in magenta if the similarity exceeds 50 %.

4.3.2 Structure solution

Initial attempts to solve the structure of Sso10a2 from a native data set using Sso10a as search model molecular replacement (MR) were unsuccessful. Interestingly a MR solution could not readily be obtained on the selenomethionine data either using Phaser (McCoy et al. 2007) or Molrep (Vagin and Teplyakov 1997) with one or both chains of Sso10a as search model. Even with the A chain from the refined Sso10a2 model a solution could not be found using the default setting for finding multiple copies of the search model in the asymmetric unit.

Table 4.1: Data collection statistics for Sso10a2 as reported by the Scala (Evans 2006) program. The R_{merge} factor was multiplicity weighted (Diederichs and Karplus 1997; Weiss and Hilgenfeld 1997)

X-ray source	ESRF ID14-1
X-ray detector	ADSC Quantum 210
Wavelength (Å)	0.9334
Space group	$P2_122$
Unit-cell parameters	$a = 60.22$, $b = 69.54$, $c = 80.73$, $\alpha = \beta = \gamma = 90.00$
Matthews coefficient ($\text{Å}^3\,\text{Da}^{-1}$)	2.31
Solvent content (%)	46.73
Resolution (Å)	40.36 – 2.00 (2.11 – 2.00)
Wilson plot B factor (Å^2)	32.24
R_{merge} (%)	3.9 (48.2)
R_{rim} (%)	5.2 (54.1)
R_{pim} (%)	1.4 (16.8)
Mean $I/\sigma(I)$	39.6 (4.4)
Completeness (%)	93.3 (95.2)
Multiplicity	13.5 (9.6)
Anomalous completeness (%)	92.9 (94.0)
Anomalous multiplicity	7.2 (5.0)
Total No. of observations	296078
No. of unique reflections	21998
No. of reflections in R_{free} set	1118

Given that both the completeness and multiplicity were higher in the selenomethionine data set compared to the native data set, we built

and refined the structure using the anomalous data. The data collection statistics shown in Table 4.1 are consistent with those of a good quality data set (Evans 2006), with high completeness, redundancy and a strong anomalous signal. Unfortunately, even though the crystal was soaked in 10 % v/v glycerol prior to flash freezing, the diffraction data showed strong ice-rings. Weak reflections close to ice-rings are usually treated properly by the data processing programs. Nonetheless during refinement it was observed that around the 2.249 Å ice-ring the R factors increase sharply, with a respective mean R and R_{free} factor of 40.5 % and 46.5 % in the 2.27 to 2.21 Å resolution bin, as opposed to a mean overall R of 25.6 % and R_{free} of 21.4 %. The large difference between the overall R factors and the R factors of reflections between 2.27 and 2.21 Å suggests that the reflections near the ice-rings carry little information. Moreover, the concomitant rise in the mean R factors, as compared to the situation where the reflections near the ice-rings are removed, is such that these values are no longer a reflection of the quality of the refined model. Excluding the reflections around 2.249 Å only slightly affected the (anomalous) completeness without noticeably decreasing the quality of the electron-density map.

The initial model output by the automatic structure solution suite Crank was very complete. The R and R_{free} were 28.9 % and 34.6 %, respectively. The RMSD of the C^{α}-backbone with respect to the final refined model was just 0.44 Å. After manual model building in Coot Chain A was most complete with residue 7 to 104 out of 104 residues sequenced correctly. While some features of the first six N-terminal amino acids could be distinguished the electron-density map was to poor in this region to place the residues unambiguously. The same was true for the first seven residues of the B and C chains and portions of the C-termini of these chains. After placement of waters and several rounds of manual model building, iterated with restrained refinement in Refmac with TLS and loose NCS restraints the R_{free} decreased to 24.5 % (Table 4.2).

With the exception of both termini the quality of the electron-density map for chain A was excellent. In contrast, the electron-density map for

Table 4.2: Refinement statistics for Sso10a2.

Resolution (Å)	40.36 – 2.00
No. of reflections	20862
No. of molecules in the ASU	3
R_{work} (%)	20.5
R_{free} (%)	24.5
No. of atoms	
Protein	2127
Water	318
B factors (Å^2)	
Protein	45.72
Chain A	37.81
Chain B	45.30
Chain C	54.04
Water	43.44
Root-mean-square deviations	
Bond lengths (Å)	0.0167
Bond angles (°)	1.7139
No. of TLS bodies	3
Ramachandran favoured (%)	98.57
Ramachandran outliers (%)	0
Rotamer outliers (%)	2.34

the B and C chains was poor around Lys57 to Lys63. Overall the quality of the map around chain B was almost equal to that of chain A, whereas it was noticeably worse for chain C, of which several residues have poorly resolved side-chains. The decline in quality of the electron-density map is clearly illustrated by Figure 4.2 which shows the electron-density map contoured at 1.5σ around the Lys57 to Lys63 region for each of the chains. This is also consistent with the rise in the mean temperature factors (Table 4.2), which suggest increasing disorder going from chain A to chain C. Because of the absence of any striking differences between the chains and the notable better quality of the chain A model it seems reasonable to discuss the structure of Sso10a2 with reference to chain A only.

4.3.3 Global structure

The structure of Sso10a2 has been determined to 2 Å resolution and contains three molecules in the ASU. The average interchain RMSD after a least-squares superposition in Coot was 0.341 Å (0.302 and 0.391 Å for chain A with respect to chain B and C and 0.331 Å between chain B and chain C), which indicates that the overall structure of each of the chains is very similar.

Analogous to Sso10a the Sso10a2 monomer's winged helix domain is arranged in a $\alpha1-\beta1-\alpha2-\alpha3-\beta2-\beta3$ topology. The protein is composed of mostly α-helices (69.7 %) and approximately 14.1 % β-sheets. The first α-helix ($\alpha1$, 2.5 turns, Thr10 to Asn20) is preceded by a short negatively charged stretch that becomes increasingly disordered toward the distal end of the N-terminus. The first α-helix is directly followed by Cys21, which forms a sulphur bridge with Cys25, the only other cysteine in the monomer. The aforementioned residues are connected by a short loop that ends with Pro23 in the cis configuration, followed by a short β-strand. Together with $\alpha1$, a second short helix ($\alpha2$, 2 turns, Ile27 to Ala34) is packed against a third slightly longer α-helix ($\alpha3$, 3.5 turns, Tyr38 to Val51) to form the hydrophobic core, stabilised by the sulphur bridge formed between the two cysteines. Finally, the wing of the winged helix domain is a β-hairpin composed of two six residue β-strands ($\beta2$, Leu54 to Glu59 and $\beta3$, Arg62 to Ile67). The second domain in Sso10a2 is a 30 residue C-terminal α-helix (8 turns, Glu69 to Leu98) that forms an antiparallel coiled-coil in the dimer. Similar to the N-terminus the last six C-terminal residues appear to be part of a flexible region in the protein.

4.3.4 Comparison of Sso10a and Sso10a2

Like the Sso10a dimer, the Sso10a2 dimer is defined by the coiled-coil interaction which closely follows the typical "knobs-into-holes" packing (Crick 1953). The presence of the non-hydrophobic Asp69 and Lys86$*$ pair in Sso10a disrupts the heptad repeat pattern (Chen et al. 2004). However, the atypical salt bridge between these residues is not formed

between the corresponding residues His77 – Lys94∗ in Sso10a2. At physiological pH histidine carries little charge and no strong interaction will occur with the positively charged lysine. Indeed the distance of 4.2 Å between the primary amine NZ of the lysine and tertiary amine NE2 of the histidine precludes a hydrogen bond. Furthermore the conformational freedom of the Lys94 side chain does not appear to be constrained by strong interactions as is clear from the low contrast of the side chain electron density.

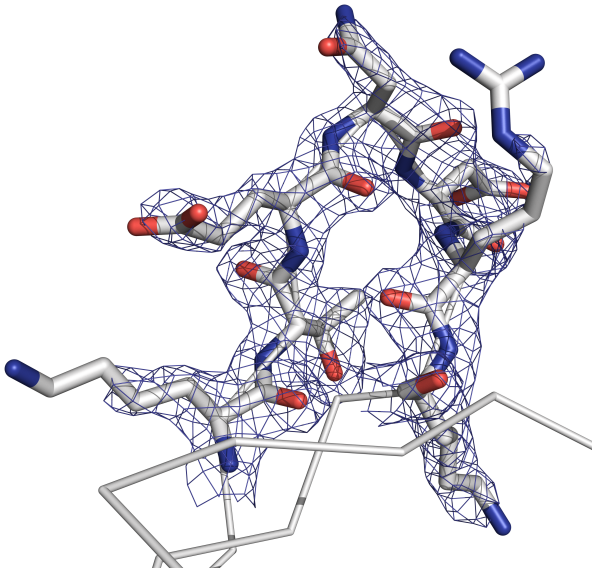
Roughly two-thirds of the coiled-coil is solvent exposed with the majority of the solvent exposed residues being hydrophilic and conversely the inward facing residues hydrophobic. A noteworthy exception is Leu82 that is found in a position where one would expect a hydrophilic residue according to the heptad repeat. Together with Leu74 and Leu78 from H1 and Phe15 and Leu18 in H4 this residue partakes in the hydrophobic interface between both helices to form a contiguous hydrophobic core between the N-terminal end of the coiled-coil. The interface is further stabilized by hydrogen bonds between Glu12 – Arg85 and Glu22 – Arg75, similar to the Gln11 – Asn74 and Glu15 – Arg71 hydrogen bonded pairs in Sso10a. The coiled-coil and winged helix DNA binding domains are more closely associated in Sso10a2 compared to Sso10a, which causes the globular head to be tilted approximately 5° toward the H4 helical axis with respect to the orientation in Sso10a, as shown in Figure 4.3. Consequently, the DNA binding domains in the Sso10a2 dimer are closer together than in Sso10a. For example, within the dimer the distance between the C^{α}atoms of the highly conserved tyrosine at position 34 in Sso10a and 38 in Sso10a2 is 61.1 Å for Sso10a and 56.2 Å for Sso10a2. The conserved tyrosine is found in the N-terminal end of the α3 helix that has previously been speculated to be involved in DNA binding (Chen et al. 2004). Presumably the different orientation of the DNA binding domains and the α3 helices in particular has consequences for DNA binding by Sso10a2.

4.3.5 Concluding remarks

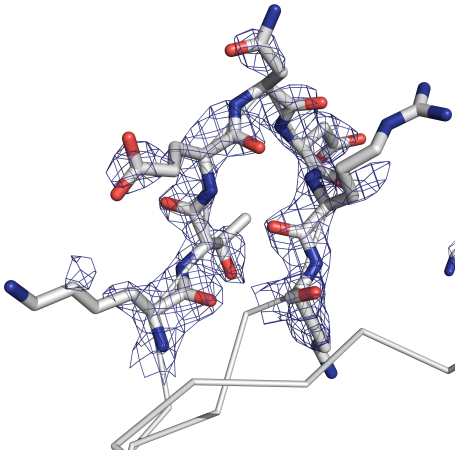
Sso10a2 is a member of three highly similar proteins in *S. solfataricus*, yet to some extent the protein sequences have diverged. In the case of Sso10a and Sso10a2 this leads to significant biochemical and structural differences, which may have consequences for the biological function, most importantly DNA binding. As of yet the exact mechanism of DNA binding remain unclear and predictions regarding the effects of single amino acid substitutions or the observed global shift of secondary structure elements remain highly speculative.

The similarity between the "knob-and-hole" pattern between Sso10a and Sso10a2 may allow formation of heterodimers. This intriguing possibility, potentially offers further means for fine tuning DNA organisation by Sso10a family members. The existence and relative distribution of the different putative Sso10a dimer species is therefore an interesting avenue for investigation. Additionally, *in vivo* studies at protein level should also conclusively demonstrate whether truncation of Sso10a2 based on the sequence similarity with Sso10a and Sso10a3 was indeed justified.

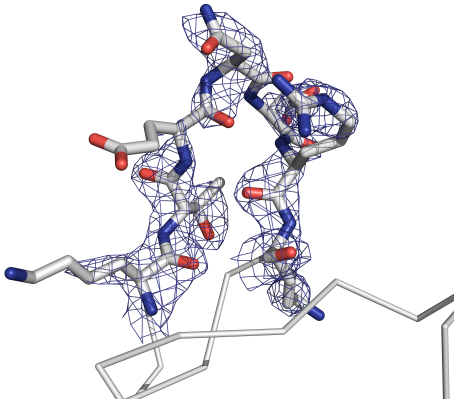
The structures of Sso10a and Sso10a2 have opened up the possibility of further comparative analyses by for instance molecular dynamics simulations of DNA binding and multimerisation, as well as, enabling rational design of point mutants to identify residues that influence the stability of the coiled-coil or higher order assemblies, as well as DNA interacting residues. However, the precise coordination of DNA binding requires co-crystallisation of Sso10a2 with a suitable DNA substrate.



(a) Chain A



(b) Chain B

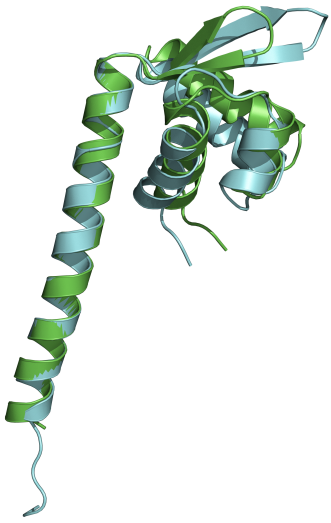


(c) Chain C

Figure 4.2: The electron-density map contoured at 1.5σ around the Lys57 to Lys63 loop representative for each of the NCS related chains in Sso10a2



(a) Front



(b) Back

Figure 4.3: THESEUS maximum likelihood superpositioning of the Sso10a and Sso10a2 monomers (a) Front view (b) Back view

5

Crystallisation of human recombinant C1 inhibitor

Abstract

C1 inhibitor (C1INH), a member of the serpin superfamily of protease inhibitors, is an important down-regulator of inflammatory processes in the blood by preventing spontaneous activation of the complement system. In addition C1INH acts on the coagulation contact activation pathway by inhibition of kallikrein, activated factor XI, and activated factor XII. Genetic C1INH deficiency results in hereditary angioedema (HAE), a serious illness associated with recurring swelling in the upper airways that may be life threatening. Ruconest® is a recombinant human C1 inhibitor (rhC1INH) that was recently approved for replacement therapy in HAE. Here we present extensive crystallisation screening of rhC1INH, as well as two low molecular weight variants that both retain their inhibitory activity *in vitro*. Crystals were formed in several conditions and a number of high quality crystals were selected for analysis at the European Synchrotron Radiation Facility. These crystals are believed to contain rhC1INH in the *active* form, whose conformation is unknown for this serpin.

The author of this thesis designed and evaluated high-throughput crystallisation studies of the full length rhC1INH protein and its two low molecular weight species.

5.1 Introduction

C1INH is an acute phase plasma protein essential for down-regulation of inflammation that acts through inhibiting activation of the complement system, as well as, the coagulation contact (kallikrein-kinin) activation pathway. As part of the innate immune system, the complement system "complements" the ability of phagocytic cells and antibodies to defend the host against pathogens and altered host cells. Its activation occurs through three distinct routes: the classical complement pathway, the alternative complement pathway, and the lectin pathway. C1INH is the only known inhibitor that down-regulates early components in the complement system activation cascade by inhibiting both C1r and C1s serine proteases of the classical complement pathway (Cooper 1985; Schapira et al. 1985; Sim et al. 1979), as well as, the highly similar MASP-1 and MASP-2 proteins of the lectin pathway (Matsushita et al. 2000). In addition to its role in the complement system, C1INH also targets activated factor XI (fXIa), activated factor XII (fXIIa) and kallikrein of the contact activation system (Chan et al. 1977; Pixley et al. 1985; Schapira et al. 1982; van der Graaf et al. 1983; Wuillemin et al. 1995). C1INH has several other anti-inflammatory functions, including suppression of gram negative sepsis and endotoxin shock (Liu et al. 2003, 2004; Nuijens et al. 1989), as well as direct and indirect interactions that influence leukocyte behaviour (Cai and Davis 2003; Cai et al. 2005). Finally C1INH treatment has show to improve the clinical outcome in a number of inflammation related conditions, including sepsis and other bacterial infections, ischaemia-reperfusion injury and acute transplant rejection as reviewed by Davis et al. (2010).

The physiological importance of C1INH is perhaps best illustrated by its absence in HAE, a genetic C1INH deficiency that is responsible for localized increases in vascular permeability. The resulting episodic bouts of tissue swelling may be life-threatening when occurring in the upper airways. This essential role of C1INH in the pathology of HAE was established with the discovery by Donaldson and Evans (1963), who showed that patients with HAE have low blood plasma levels of C1INH.

The disease is known to be autosomal dominantly inherited, in fact the hereditary nature of HAE has been firmly established for well over a century (Osler 1888). To date, a variety of mutations, deletions and translocations have been identified that lead to the disease (Bock et al. 1986; Bowen et al. 2001). On this genetic basis two variants of HAE can be distinguished. The type 1 form of HAE affects roughly 85 % of the patients, in which expression of C1INH from one of the alleles is abolished altogether. In the less common type 2 form patients have a mutation leading to dysfunctional protein (Bos et al. 2003). Both types of the disease cause a decrease in the level of *functional* C1INH, a trait HAE has in common with acquired angioedema, a disease with a similar clinical presentation, but caused by aberrant antibody activity against C1INH that effectively blocks its activity, rather than a genetic C1INH deficiency.

While it is clear that C1INH is a key player in HAE, the exact biochemical mechanism explaining the recurrent acute localised increases in vascular permeability remains somewhat unclear. Davis et al. (2010) reviewed the various biological functions of C1INH also in relation to HAE. One theory suggests involvement of factors derived from complement activation (da Silva 1967; Donaldson et al. 1969, 1977; Klemperer et al. 1968; Strang et al. 1986, 1988). However this theory has been abandoned, at least in part, in favour of an explanation that suggests a role for the contact activation pathway in the pathogenesis of HAE. Hallmarks of contact system activation, such as decrease of pre-kallikrein and high-molecular-weight kinogen (HMWK) plasma levels (Schapira et al. 1983), cleavage of HMWK (Berrettini et al. 1986; Lämmle et al. 1988), and presence of bradykinin in plasma (Nussberger et al. 1998), have been observed during HAE attacks. Additionally several studies show that bradykinin is released during *in vitro* incubation of HAE plasma (Curd et al. 1982; Fields et al. 1983) and that this activity depends on the presence of pre-kallikrein, HMWK, and factor XII (fXII) (Shoemaker et al. 1994). Furthermore HAE does not occur in cases if the mutation in C1INH only affects inhibition of C1r and C1s, but not

fXIIa and plasma kallikrein, ruling out disruption of the complement system as the sole cause for HAE (Wisnieski et al. 1994; Zahedi et al. 1995, 1997). Finally, increased vascular permeability observed in C1INH knockout mice can be reversed by administering C1INH, but also with a kallikrein inhibitor or a bradykinin type 2 receptor (Bk2R) antagonist. Interestingly C1INH/Bk2R double knockout mice did not develop increased vascular permeability. In summary it would seem that disruption of the contact system activation is responsible for the angioedema attacks in HAE and that bradykinin is it principle mediator.

Like all other members of the serine protease inhibitor superfamily, C1INH can exist in two states, a metastable conformation in which the flexible reactive center loop (RCL) is exposed and acts as a bait to trap one of C1INH's target proteases. Cleavage of the peptide bond by the targeted protease triggers the trap and results in a large conformational change in the serpin-domain that distorts the active site of the protease and locks the covalently linked protease in a C1INH-protease complex (Huntington et al. 2000). During this architectural transition the RCL is inserted in the A β-sheet of the serpin domain forming an additional β-strand. Spontaneous insertion of the RCL into β-sheet A transforms the serpin in *latent* conformation and is accompanied by the loss in inhibitory activity as the RCL is no longer exposed to bind its target protease.

C1INH has an unusual architecture, with a C-terminal serpin domain and a heavily glycosylated N-terminal domain that it not found in other serpins (Bock et al. 1986). The structure of the *latent* form of C1INH, solved by Beinrohr et al. (2007), is similar to that of other serpins. The exception is the complete insertion of the RCL into the β-sheet A, which represents a novel conformation comprised of seven parallel β-strands. The other unique feature is the unusual arrangement of the six C-terminal residues. In particular, the C-terminal proline, which is found in a different position from almost all other serpins. Beinrohr et al. (2007) speculate that the C-terminus may act as a barrier to RCL insertion as it needs to move considerably with respect to its presumed

location in active C1INH.

Heparin and other glycosaminoglycans have an distinct effect on C1INH activity (Caldwell et al. 1999; Caughman et al. 1982), though the precise mechanism of C1INH potentiation remains unclear (Bos et al. 2002). Beinrohr et al. (2007) put forward a mechanism for C1INH potentiation whereby polyanions are "sandwiched" between C1INH and the target protease, neutralising the positive charge surrounding the RCL and the active site of the protease.

While the effect of frameshifts, introduced stop codons and large deletions in HAE can be readily explained, knowledge of the *active*, as well as the *latent* form of C1INH is needed to provide a structural explanation for the effect of certain point mutations. In particular, mutations that only partly affect the inhibitory activity of C1INH are of interest, for example the Ala443Val mutation that significantly impairs binding with C1r, but is able to complex with Cls, kallikrein, and fXIIa (Zahedi et al. 1995). Additionally, insight in C1INH structure and its intermediate states may also offer means to prevent spontaneous *active* to *latent* transition and stabilize therapeutic formulations of (recombinant) C1INH. Finally, as the inhibitory activity of C1INH on several of its targets is modulated by glycosaminoglycans such as heparin, it may be possible to design therapeutic glycosaminoglycan mimics that potentiate C1INH towards specific protease targets or speed up formation of transition state complexes thereby effectively boosting C1INH's specific activity. This could improve clinical outcomes in situations where (endogenous) levels of functional C1INH are low, such as HAE. Determination of the structure of *active* C1INH is the first step in this process.

Infusion of C1INH isolated from human plasma has been used in Europe for over four decades to successfully treat HAE attacks (Waytes et al. 1996). Recently the safety and efficacy of rhC1INH (Ruconest®) for the treatment of angioedema attacks in patients with HAE have been demonstrated (Plosker 2012; Varga and Farkas 2011; Zuraw et al. 2010). rhC1INH is commercially produced in high quantities from the milk of transgenic rabbits by the Dutch company Pharming. A

small fraction of the expressed rhC1INH protein lack either the first 84 or the first 96 N-terminal residues. An unwanted impurity from a pharmacological perspective, however, as both retain their inhibitory activity (van Veen et al. 2012), these truncated forms present interesting targets for structure determination. Here, we describe the crystallisation of both truncated *fully glycosylated active* forms of rhC1INH and report the outcome of extensive crystallization trials with full-length rhC1INH.

5.2 Methods

5.2.1 Protein expression and purification

Expression and purification of rhC1INH was as described before by van Veen et al. (2012). In brief, rhC1INH was secreted into the milk of transgenic New Zealand White rabbits. The transgenic rabbit milk was defatted by centrifugation and applied to a SP Sepharose column (GE Healthcare, Uppsala, Sweden), followed by a Q Sepharose column (GE Healthcare, Uppsala, Sweden) to separate rhC1INH from rabbit milk proteins including trace amounts of endogenous C1INH and to remove chemicals added during purification. Residual rabbit protein impurities were removed in a chromatographic step consisting of a zinc-charged chelating Sepharose column (GE Healthcare, Uppsala, Sweden). The purified rhC1INH was filtered over Planova™15N hollow fibers (Asahi Kasei Bioprocess, Tokyo, Japan) and formulated in 20 mM sodium citrate, 6.5 % w/v sucrose of pH 6.8 by ultrafiltration using Biomax-10 membranes (Millipore, Billerica, MA). Analytical size-exclusion chromatography of purified rhC1INH revealed a purity of 99 %. Beside multimeric rhC1INH species, the 1 % impurities contained two N-terminal cleaved C1INH species starting at amino acid 85 (LMW1) and 97 (LMW2), respectively. The LMW species were isolated from purified rhC1INH by a purification procedure that will be described elsewhere. All purified rhC1INH species were buffer exchanged using Vivaspin filter units (10 kDa MWCO; Sartorius, Goettingen, Germany) to a solution composed of 0.15 M NaCl 20 mM sodium citrate (pH 7.0).

5.2.2 Crystallisation

The inhibitory effect of intact rhC1INH on the target proteases C1s, Factor XIIa, kallikrein and Factor XIa has been shown previously, as has inhibition of C1s by both truncated forms of rhC1INH (van Veen et al. 2012). All three purified C1INH species were kindly provided by Pharming Technologies B.V. (Leiden, The Netherlands). Initial sitting drop vapour diffusion crystallisation trials were set up using the JCSG+ and PACT screens purchased from Qiagen (Hilden, Germany) in Corning® 3550 96 well protein crystallisation plates (Sigma-Aldrich, St. Louis, MO). A comprehensive screen was performed varying the protein to precipitant ratio between 10 to 90 % v/v with a 10 % v/v step and a drop size of 0.8 µL using three different concentrations (17, 50 and 100 mg mL^{-1}) of full-length rhC1INH protein. Additionally, crystallisation trials were set up with 50 mg mL^{-1} full-length rhC1INH pretreated with 1 U PNGase F per 100 µg protein for 24 h at 37 °C to partially remove the N-glycans and enhance the propensity of rhC1INH to crystallise.

Considerably less material was available for setting up trials with the truncated rhC1INH species and the initial screens were set up in Corning® 3550 plates using a fraction containing a mixture of LMW1 and LMW2. Three drops of 0.66 µL were dispensed composed of 30, 50 and 70 % v/v protein solution with concentration of 13.4 mg mL^{-1}.

Optimisation of initial hits with the rhC1INH truncated species were set up in MRC Crystallisation 2 drop plates (Molecular Dimensions, Newmarket, United Kingdom). These grid screens were formulated from stock solutions using the Rock Maker software package (Formulatrix, Waltham, MA) by varying the polyethylene glycol (PEG) 3350 and KF concentration from 10 to 30 % w/v and from 100 to 300 mM, respectively. The concentration of LMW1 and LMW2 was approximately 10 mg mL^{-1} and the drop size 1 µL with a protein content in each of the drops of 60 and 70 % v/v, respectively.

Where required the protein was concentrated with an 15 kDa molecular weight cut-off centrifugal filter unit (Millipore, Billerica, MA). The

reservoir solution (75 µL) was dispensed with a Genesis RS200 liquid handling platform (Tecan, Männedorf, Switzerland) and an Oryx robot (Douglas Instruments, East Garston, United Kingdom) was used to set the crystallisation drops. Over the course of 27 days crystals appeared in several conditions, which were verified to contain protein by UV fluorescence microscopy using a U-MWU2 filter (Olympus, Tokyo, Japan).

5.3 Results and discussion

5.3.1 Crystallisation screening

rhC1INH remains soluble to at least 180 mg mL^{-1} in PBS without loss in activity (Van Veen 2012, personal communication). This exceptional solubility was confirmed by the results of the crystallisation experiments with the full-length protein. While some of drops remained clear, even at a concentration of 100 mg mL^{-1} protein and 90 % v/v precipitant, the drops went from completely clear to full precipitate for the vast majority of the 192 tested conditions, indicating adequate sampling of the crystallisation space. No crystals could be detected in any of the drops, though phase separation and skin formation on the drops was observed for a number of conditions. Glycosylation has been shown to both stabilize and promote protein solubility (Solá and Griebenow 2009), which may explain the high solubility of the full-length rhC1INH whos N-terminus is extensively N- and O-glycosylated (Bock et al. 1986). Nevertheless no difference in crystallisation behaviour of the PNGase F treated protein compared to non-deglycosylated rhC1INH was observed. It should be noted though that the deglycosylation of the native rhC1INH was not monitored and it is generally accepted that the efficiency of PNGase F under non-denaturing conditions is limited.

The crystallisation experiments with a mixture of the N-terminally truncated forms of rhC1INH produced several hits. In the JCSG+ screen crystals were found in conditions with a drop composition of 70 % v/v 10 % w/v PEG 1000, 10 % w/v PEG 8000 and 30 % v/v 0.1 M imidazole (pH 8), 10 % w/v PEG 8000 (JCSG+ conditions No. 36 and 60), as well

as JCSG+ conditions No. 81 and 82 containing 30 % w/v polyethylene glycol monomethyl ether (PEG MME) 2000 and 0.1 M KSCN or 0.15 M KBr. In the latter two conditions the crystals were found in drops containing 30 to 50 % v/v and 50 to 70 % v/v protein solution, respectively. In the PACT screen all crystals suitable for diffraction analysis were found in drops composed of 70 % v/v well solution in PACT conditions No. 3 to 6, 18, 29, 30, 39, 40, 42, and 49. These conditions contained 25 % w/v PEG 1500 in all but one case, where the well solution was composed of 0.2 M NaF and 20 % w/v PEG 3350. The nine remaining drops that produced suitable crystals were buffered with 0.1 M SPG (succinic acid, sodium dihydrogen phosphate and glycine) at pH 6, 7, 8, and 9; MIB (succinic acid, sodium dihydrogen phosphate and glycine) at pH 9; PCB (sodium propionate, sodium cacodylate and Bis-tris propane) at pH 8 and 9; and MMT (DL-malic acid, MES and Tris base) at pH 6, 7, and 9.

Varying the PEG 3350 and KF concentration resulted in crystallisation in both drops for almost all conditions with the truncated forms of rhC1INH. Though crystals suitable for diffraction analysis were predominantly found in the grid rectangle defined by well E8 (22.7 % w/v PEG 3350, 214 mM KF) to F11 (28.2 % w/v PEG 3350, 243 mM KF) and very few of such crystals were found in conditions with a PEG concentration below 17.3 % w/v or a KF concentration below 157 mM.

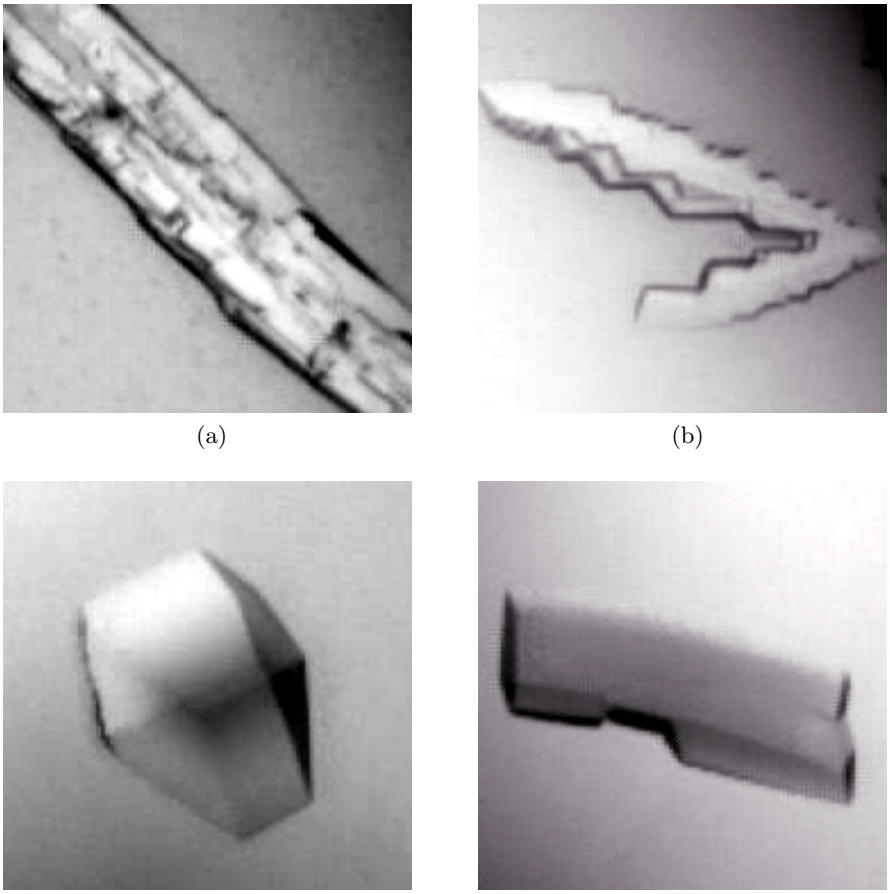
5.3.2 Diffraction targets

Crystals of appropriate size for X-ray diffraction all showed strong fluorescence under UV illumination, making it very likely that they contain protein. The morphology typical of the crystals selected for X-ray diffraction analysis (Figure 5.1) includes highly regular crystals composed of a single crystal domain (Figure 5.1c) or with minor packing defects (Figure 5.1d). In both cases the smallest dimension of the crystals ranged from approximately 50 to 250 µm. It is generally accepted that single crystals offer the best chances for obtaining good quality diffraction. Nonetheless, several crystals with irregular packing such as

those shown in Figures 5.1a and 5.1b were also selected for diffraction analysis. While the diffraction pattern of these macroscopically twinned crystals will likely be uninterpretable, it will definitively show whether the crystal contains protein and whether the protein is ordered in the individual crystal domains. Further optimisation of the crystallization conditions possibly in combination with micro-seeding may improve the quality of these crystals. Additionally, the crystal clusters were dissected into individual crystals yielding fragments of 20 to 50 µm along the shortest edge, that may yet provide good quality diffraction.

5.3.3 Concluding remarks

We have reported conditions to successfully crystallise two *active* forms of rhC1INH lacking the first 84 and 96 N-terminal residues, respectively. As both truncations have previously been shown to inhibit their target proteases (van Veen et al. 2012) *in vitro*, it is possible that the crystals contain rhC1INH in an active form. In contrast to the structure of *latent* C1INH previously reported by Beinrohr et al. (2007) both truncated forms were not deglycosylated prior to crystallisation possibly providing additional structural information. While the planned diffraction experiments may not yet provide a high resolution model, their outcome can certainly be used to guide future crystallisation experiments by optimising only those conditions that produce diffracting crystals.



(a) (b)

(c) (d)

Figure 5.1: Crystals of truncated forms of rhC1INH representative of those selected for X-ray diffraction analysis at the European Synchrotron Radiation Facility (ESRF). (a) a 170×700 µm sized crystal composed of parallelly arranged rod-like crystal domains with an approximate mean width of 30 µm (b) crystal roughly 80×470 µm composed of stacked plates with roughly 50 µm in both directions of the plate plane (c) highly-regular crystal measuring approximately 250×350 µm (d) rod-like crystal of 130×420 µm with minor packing defects

A

Data sets used for testing

For testing of the new algorithms presented in Chapters 2 and 3 a total of 134 data sets were used. A majority of 78 data sets was provided by the Joint Center for Structural Genomics (JCSG, http://www.jcsg.org/): 1vjn, 1vjr, 1vjz, 1vk4, 1vkm, 1vlm, 1vqr, 1z82, 1zy9, 1zyb, 2a2m, 2a2o, 2a3n, 2a6b, 2aml, 2avn, 2b8m, 2etd, 2etj, 2ets, 2etv, 2evr, 2f4p, 2fea, 2ffj, 2fg0, 2fg9, 2fna, 2fqp, 2fur, 2fzt, 2g0t, 2g42, 2gc9, 2nlv, 2nuj, 2nwv, 2o08, 2o1q, 2o2x, 2o2z, 2o3l, 2o62, 2o7t, 2o8q, 2obp, 2oc5, 2od5, 2od6, 2oh3, 2okc, 2okf, 2ooj, 2opk, 2osd, 2otm, 2ozg, 2ozj, 2p10, 2p4o, 2p7h, 2p7i, 2p97, 2pg3, 2pg4, 2pgc, 2pim, 2pn1, 2pnk, 2ppv, 2pr7, 2prr, 2prv, 2prx, 2pv4, 2pw4, 2b78, 2b79; 23 data sets from Mueller-Dieckmann et al. (2007): 2g4h, 2g4i, 2g4j, 2g4k, 2g4p, 2g4q, 2g4l, 2g4n, 2g4o, 2g4r, 2g4s, 2g4t, 2g4u, 2g4v, 2g4w, 2g4x, 2g4y, 2g4z, 2ill, 2g51, 2g52, 2g54, 2g55; and 33 from various other individual data set contributors: 1e42 (Owen, Vallis, et al. 2000), 1e6i (Owen, Ornaghi, et al. 2000), 1hf8 (Ford et al. 2001), 2ahy (Shi et al. 2006), 2hba, 2o0h (Sun et al. 2007), 2rkk (Xiao et al. 2008), 3bpj, 2fdn (Dauter et al. 1997), 1of3 (Boraston et al. 2003), 1i4u (Gordon et al. 2001), 1dw9 (Walsh et al. 2000), 2v0o (Henne et al. 2007), 1fse (Ducros et al. 2001), 1xib (Carrell et al. 1989), 1fj2 (Devedjiev et al. 2000), 1h29 (Matias et al. 2002), 1c8u (Li et al. 2000), 1lvy (Schiltz et al. 1997), 1lz8 (Dauter et al. 1999), 1e3m (Lamers et al. 2000), 1ga1 (Dauter et al. 2001), 1djl (White et al. 2000), 1dtx (Skarzyński 1992), 1dpx (Weiss 2001), 1mso (Smith et al. 2003), 1ocy (Thomassen et al. 2003), 1rju (Calderone 2004), 1rgg & 2sar (Sevcik et al. 1996, 1991), 1m32 (Chen et al. 2002). Finally, two data sets that were used have not been deposited in the PDB: a subtilisin savinase (sav3, Betzel et al. 1988; Dauter et al. 2002) and glucose isomerase (gilu, Dauter et al. 2002) data set.

Table A.1: The raw data summarized in Chapter 3. The PDB code is given by the first column. For data sets SAD data is taken from peak wavelength MAD data (※) or the inflection wavelength (†) or were phased on a partial heavy atom substructure or substructure from another data set (⋆). data sets not deposited in the PDB are shown in italics. The maximum resolution is denoted by R, followed by the fraction of the substructure that was detected and the map correlation after experimental phasing as S and M_{Ph}, respectively. M_{σ_A} and $M_{SAD\text{-}DM}$ denote the map correlation with the map generated from the deposited model phases with the map after density modification and phase combination with the σ_A and the *SAD-DM* algorithm, respectively. The difference between the two is given by ΔM. Similarly the fraction built with Buccaneer and ARP/wARP are denoted by B and A, respectively.

PDB ID	R (Å)	S	M_{Ph}	M_{σ_A}	$M_{SAD\text{-}DM}$	ΔM	$B_{SAD\text{-}DM}$	B_{σ_A}	$A_{SAD\text{-}DM}$	A_{σ_A}
1vjn	3	0.38	0.08	0.03	0.03	0	0.02	0.02	0	0
1vjr	2.4	1	0.41	0.8	0.6	0.2	0.87	0.54	0.89	0.93
1vjz	2.5	1	0.41	0.72	0.55	0.17	0.81	0.37	0.3	0.08
1vk4	2	1	0.58	0.85	0.78	0.07	0.92	0.93	0.85	0.83
1vkm	2.6	0.98	0.51	0.76	0.61	0.15	0.96	0.64	0.77	0.62
1vlm	2.8	1	0.34	0	0	0	0.01	0	0	0
1vqr	2.77	0.86	0.4	0.6	0.41	0.19	0.58	0.11	0.15	0.05
1z82	2.05	0.85	0.4	0.47	0.43	0.04	0.09	0.06	0.06	0.06
1zyb	2.1	1	0.34	0.53	0.4	0.14	0.39	0.16	0.38	0.11
2a3n	1.23	1	0.58	0.84	0.8	0.05	0.98	0.98	0.97	0.96
2a6b	2.1	1	0.3	0.58	0.4	0.18	0.86	0.08	0.85	0.6
2aml	1.5	0.75	0.52	0.71	0.62	0.1	0.96	0.62	0.94	0.93
2avn	2.35	0.9	0.42	0.83	0.73	0.1	0.98	0.96	0.87	0.92
2etd	2.3	1	0.28	0.17	0.16	0.01	0.05	0.02	0.03	0.02

Table A.1 Continued from previous page

PDB ID	R (Å)	S	M_{Ph}	M_{σ_A}	$M_{SAD\text{-}DM}$	ΔM	$B_{SAD\text{-}DM}$	B_{σ_A}	$A_{SAD\text{-}DM}$	A_{σ_A}
2etj	1.74	1	0.36	0.5	0.4	0.1	0.06	0.09	0.44	0.22
2ets	2.25	1	0.39	0.81	0.68	0.13	0.97	0.93	0.93	0.93
2evr	1.6	1	0.48	0.91	0.91	0	0.91	0.92	0.96	0.96
2f4p	1.9	1	0.3	0.39	0.35	0.04	0.02	0.03	0.06	0.06
2fea	2	1	0.49	0.84	0.72	0.12	0.81	0.94	0.92	0.92
2fg0	1.79	1	0.41	0.86	0.74	0.12	0.97	0.94	0.94	0.93
2fg9	2.2	1	0.5	0.83	0.65	0.18	0.9	0.74	0.77	0.69
2fna	2	0.89	0.33	0.34	0.31	0.03	0.05	0.02	0.04	0.02
2fqp	1.8	1	0.47	0.6	0.56	0.03	0.17	0.03	0.45	0.37
2fur	1.9	1	0.41	0.51	0.48	0.03	0.1	0.04	0.24	0.19
2gc9	1.7	1	0.44	0.64	0.6	0.04	0.78	0.59	0.89	0.89
2nlv	1.33	0.53	0.43	0.68	0.64	0.04	0.56	0.68	0.94	0.89
2nuj	2	1	0.45	0.89	0.88	0.02	0.98	0.96	0.88	0.87
2nwv	1.85	1	0.3	0.66	0.42	0.24	0.52	0.02	0.88	0.11
2o08	2.5	1	0.4	0.72	0.49	0.23	0.79	0.23	0.67	0.1
2o1q	2	0.89	0.46	0.51	0.5	0	0.03	0.07	0.39	0.31
2o2x	1.5	1	0.55	0.71	0.67	0.04	0.87	0.89	0.92	0.92
2o2z	2.6	1	0.46	0.81	0.78	0.03	0.97	0.96	0.88	0.91
2o3l	2.05	0.92	0.48	0.84	0.74	0.1	0.87	0.95	0.87	0.85
2o62	1.75	1	0.41	0.76	0.57	0.2	0.88	0.31	0.92	0.91
2o7t	2.1	1	0.35	0.3	0.3	0	0.06	0.05	0.02	0.02
2o8q	2.5	1	0.44	0.81	0.75	0.06	0.9	0.75	0.74	0.77

Table A.1 Continued from previous page

PDB ID	R (Å)	S	M_{Ph}	M_{σ_A}	$M_{SAD\text{-}DM}$	ΔM	$B_{SAD\text{-}DM}$	B_{σ_A}	$A_{SAD\text{-}DM}$	A_{σ_A}
2obp	1.7	1	0.52	0.89	0.89	0	0.98	0.98	0.94	0.93
2oc5	1.68	1	0.56	0.92	0.92	0	1	1	0.97	0.94
2od5	1.98	1	0.33	0.86	0.68	0.18	0.97	0.88	0.85	0.81
2od6	1.9	0.95	0.46	0.72	0.61	0.12	0.9	0.39	0.91	0.83
2oh3	2	0.8	0.47	0.82	0.76	0.06	0.95	0.93	0.89	0.9
2okc	2.2	1	0.47	0.69	0.57	0.12	0.71	0.3	0.28	0.15
2okf	1.6	1	0.44	0.82	0.75	0.06	0.92	0.87	0.78	0.78
2ooj	1.84	0.92	0.42	0.72	0.56	0.16	0.7	0.25	0.93	0.71
2opk	2.1	1	0.4	0.82	0.6	0.22	0.96	0.26	0.93	0.65
2osd	2.3	0.57	0.19	0.14	0.14	0	0.04	0.03	0.01	0
2otm	1.9	0.83	0.41	0.85	0.67	0.18	0.97	0.81	0.95	0.94
2ozg	2.05	1	0.4	0.57	0.5	0.08	0.31	0.19	0.32	0.13
2ozj	1.6	0.67	0.44	0.7	0.67	0.03	0.32	0.26	0.78	0.77
2p4o	1.95	1	0.34	0.39	0.36	0.04	0.05	0.05	0.03	0.05
2p7h	1.85	1	0.37	0.81	0.57	0.24	0.95	0.26	0.79	0.81
2p7i	1.84	1	0.45	0.84	0.67	0.16	0.95	0.85	0.89	0.89
2p97	1.6	1	0.41	0.82	0.81	0.01	0.98	0.94	0.93	0.93
2pg3	2.4	0.75	0.44	0.75	0.62	0.13	0.83	0.49	0.69	0.25
2pg4	2.2	0.33	0.31	0.33	0.29	0.04	0.13	0.11	0.05	0.04
2pim	2.2	0.8	0.4	0.62	0.52	0.1	0.81	0.26	0.72	0.38
2pn1	2	1	0.39	0.41	0.39	0.02	0.02	0.01	0.06	0.04
2pr7	1.55	1	0.39	0.7	0.6	0.1	0.96	0.88	0.95	0.89

Table A.1 Continued from previous page

PDB ID	R (Å)	S	M_{Ph}	M_{σ_A}	$M_{SAD\text{-}DM}$	ΔM	$B_{SAD\text{-}DM}$	B_{σ_A}	$A_{SAD\text{-}DM}$	A_{σ_A}
2prv	1.3	1	0.31	0.4	0.38	0.01	0.04	0.02	0.75	0.63
2pv4	1.95	1	0.49	0.84	0.71	0.13	0.98	0.9	0.91	0.86
2pw4	1.45	1	0.48	0.83	0.79	0.04	0.93	0.95	0.96	0.95
1e42	2.2	0.5	0.06	0.02	0.02	0	0.02	0	0.01	0.01
1e6i	1.87	0.83	0.36	0.62	0.43	0.19	0.87	0.08	0.93	0.86
1hf8	2	0.5	0.31	0.65	0.39	0.26	0.79	0.09	0.65	0.09
2b78	1.94	1	0.27	0.41	0.31	0.1	0	0.01	0.09	0.05
2b79	2.35	1	0.32	0.58	0.45	0.13	0.26	0.09	0.06	0.04
2hba	2.28	0.75	0.36	0.48	0.39	0.09	0.39	0.07	0.07	0.22
2o0h	3.29	0.64	0.28	0	0	0.02	0.03	0.02	0.01	0
2rkk	2.99	1	0.44	0.64	0.55	0.09	0.82	0.46	0.19	0.09
3bpj	2.28	0.78	0.39	0.63	0.55	0.08	0.79	0.81	0.71	0.58
2ill	2.2	1	0.29	0.65	0.43	0.21	0.29	0.11	0.56	0.15
2g4h	2	1	0.52	0.88	0.89	0	0.98	0.98	0.94	0.93
1of3	2	0.75	0.37	0.89	0.84	0.06	0.94	0.87	0.95	0.95
1i4u	2.6	0.5	0.31	0.25	0.24	0.01	0.01	0.01	0.02	0.01
1dw9[※]	2.4	0.97	0.58	0.86	0.79	0.08	0.92	0.92	0.95	0.97
2v0o	3.1	1	0.09	0.07	0.07	0	0.01	0.02	0	0
1fse[†]	2.73	0.83	0.34	0.48	0.35	0.13	0.27	0.11	0.02	0.02
1fse[※]	2.73	1	0.41	0.71	0.5	0.21	0.93	0.58	0.67	0.34
gilu	1.5	1	0.43	0.91	0.9	0.01	0.97	0.97	0.95	0.95
1fj2	1.8	1	0.49	0.73	0.66	0.07	0.98	0.97	0.92	0.92

Table A.1 Continued from previous page

PDB ID	R (Å)	S	M_{Ph}	M_{σ_A}	$M_{SAD\text{-}DM}$	ΔM	$B_{SAD\text{-}DM}$	B_{σ_A}	$A_{SAD\text{-}DM}$	A_{σ_A}
2g4k	1.82	0.95	0.39	0.53	0.45	0.08	0.14	0.14	0.6	0.2
1h29※	2.81	0.91	0.41	0.5	0.42	0.09	0.1	0.06	0.04	0.03
2g4m	1.8	1	0.56	0.91	0.91	0	0.92	0.86	0.9	0.88
1c8u†	2.5	1	0.44	0.78	0.62	0.16	0.91	0.52	0.91	0.83
1c8u※	2.5	1	0.52	0.83	0.72	0.11	0.96	0.89	0.95	0.93
1lvy	1.87	1	0	0.01	0	0.02	0	0.01	0	0
1lz8	1.53	0.94	0.58	0.77	0.68	0.1	0.71	0.61	0.98	0.95
2g4q	1.84	0.8	0.5	0.7	0.58	0.12	0.73	0.43	0.98	0.95
1e3m※	3	0.98	0.51	0.79	0.72	0.07	0.9	0.78	0.69	0.66
2g4s	2.2	0.75	0.24	0.59	0.36	0.23	0.49	0.15	0.67	0.12
2g4u on Ca⋆	1.84	1	0.46	0.66	0.5	0.16	0.42	0.01	0.93	0.65
2g4u on Na⋆	2.15	1	0.49	0.71	0.55	0.16	0.85	0.25	0.96	0.54
2g4v	2.14	0.93	0.47	0.63	0.53	0.1	0.48	0.33	0.9	0.32
1ga1※	1.8	1	0.3	0.61	0.61	0	0.9	0.94	0.92	0.91
1rgg & 2sar⋆	2.53	0.57	0.33	0.42	0.34	0.07	0.1	0.05	0.03	0.01
sav3	1.75	1	0.48	0.71	0.61	0.1	0.7	0.7	0.93	0.94
1djl	2.48	0.94	0.45	0.82	0.62	0.21	0.98	0.61	0.95	0.93
2g4z	1.98	0.54	0.46	0.8	0.62	0.18	0.93	0.65	0.96	0.94
2g51	1.84	0.29	0.1	0.06	0.09	0	0.03	0.01	0	0
2g55	1.82	0.83	0.52	0.75	0.6	0.15	0.79	0.52	0.97	0.96
1dpx	1.64	0.75	0.55	0.71	0.61	0.1	0.81	0.22	0.95	0.94
1ocy	2.74	1	0.37	0.5	0.45	0.06	0.19	0.15	0.3	0.28

Bibliography

Abrahams, J. P. (July 1997). "Bias reduction in phase refinement by modified interference functions: introducing the gamma correction." In: *Acta Crystallographica, Section D: Biological Crystallography* 53.Pt 4, pp. 371–6. DOI: 10.1107/S0907444996015272 (cit. on pp. 30, 52).

Abrahams, J. P. and A. G. W. Leslie (Jan. 1996). "Methods used in the structure determination of bovine mitochondrial F1 ATPase." In: *Acta Crystallographica, Section D: Biological Crystallography* 52.Pt 1, pp. 30–42. DOI: 10.1107/S0907444995008754 (cit. on pp. 30, 38, 52).

Adams, P. D., P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart (Feb. 2010). "PHENIX: a comprehensive Python-based system for macromolecular structure solution." In: *Acta Crystallographica, Section D: Biological Crystallography* 66.Pt 2, pp. 213–21. DOI: 10.1107/S0907444909052925 (cit. on pp. 24, 40).

Agback, P., H. Baumann, S. Knapp, R. Ladenstein, and T. Härd (July 1998). "Architecture of nonspecific protein-DNA interactions in the Sso7d-DNA complex." In: *Nature structural biology* 5.7, pp. 579–84. DOI: 10.1038/836 (cit. on p. 64).

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (Oct. 1990). "Basic local alignment search tool." In: *Journal of molecular biology* 215.3, pp. 403–10. DOI: 10.1016/S0022-2836(05)80360-2 (cit. on p. 71).

Barton, W. A., D. Tzvetkova-Robev, H. Erdjument-Bromage, P. Tempst, and D. B. Nikolov (Aug. 2006). "Highly efficient selenomethionine labeling of recombinant proteins produced in mammalian cells". In: *Protein science : a publication of the Protein Society* 15.8, pp. 2008–2013. DOI: 10.1110/ps.062244206 (cit. on pp. 19, 22).

Baumann, H., S. Knapp, T. Lundbäck, R. Ladenstein, and T. Härd (Nov. 1994). "Solution structure and DNA-binding properties of a thermostable protein from the archaeon Sulfolobus solfataricus". In: *Nature structural biology* 1.11, pp. 808–819. DOI: 10.1038/nsb1194-808 (cit. on p. 64).

Beinrohr, L., V. Harmat, J. Dobó, Z. Lörincz, P. Gál, and P. Závodszky (July 2007). "C1 inhibitor serpin domain structure reveals the likely mechanism of heparin potentiation and conformational disease." In: *The Journal of*

biological chemistry 282.29, pp. 21100–9. DOI: 10.1074/jbc.M700841200 (cit. on pp. 7, 86, 87, 92).

Bell, S. D., C. H. Botting, B. N. Wardleworth, S. P. Jackson, and M. F. White (Apr. 2002). "The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation." In: *Science* 296.5565, pp. 148–51. DOI: 10.1126/science.1070506 (cit. on p. 66).

Berrettini, M., B. Lämmle, T. White, M. J. Heeb, H. P. Schwarz, B. L. Zuraw, J. G. Curd, and J. H. Griffin (Aug. 1986). "Detection of in vitro and in vivo cleavage of high molecular weight kininogen in human plasma by immunoblotting with monoclonal antibodies." In: *Blood* 68.2, pp. 455–62 (cit. on p. 85).

Betzel, C., Z. Dauter, M. Dauter, M. Ingelman, G. Papendorf, K. S. Wilson, and S. Branner (Dec. 1988). "Crystallization and preliminary X-ray diffraction studies of an alkaline protease from Bacillus lentus." In: *Journal of molecular biology* 204.3, pp. 803–4 (cit. on p. 95).

Blow, D. M. and F. H. C. Crick (Oct. 1959). "The treatment of errors in the isomorphous replacement method". In: *Acta crystallographica* 12.10, pp. 794–802. DOI: 10.1107/S0365110X59002274 (cit. on pp. 20, 21, 26).

Bock, S. C., K. Skriver, E. Nielsen, H. C. Thøgersen, B. Wiman, V. H. Donaldson, R. L. Eddy, J. Marrinan, E. Radziejewska, and R. Huber (July 1986). "Human C1 inhibitor: primary structure, cDNA cloning, and chromosomal localization." In: *Biochemistry* 25.15, pp. 4292–301 (cit. on pp. 85, 86, 90).

Boraston, A. B., T. J. Revett, C. M. Boraston, D. Nurizzo, and G. J. Davies (June 2003). "Structural and thermodynamic dissection of specific mannan recognition by a carbohydrate binding module, TmCBM27." In: *Structure (London, England : 1993)* 11.6, pp. 665–75 (cit. on p. 95).

Bos, I. G. A., C. E. Hack, and J. P. Abrahams (Nov. 2002). "Structural and Functional Aspects of C1-Inhibitor". In: *Immunobiology* 205.4-5, pp. 518–533. DOI: 10.1078/0171-2985-00151 (cit. on p. 87).

Bos, I. G. A., Y. T. P. Lubbers, D. Roem, J. P. Abrahams, C. E. Hack, and E. Eldering (Aug. 2003). "The functional integrity of the serpin domain of C1-inhibitor depends on the unique N-terminal domain, as revealed by a pathological mutant." In: *The Journal of biological chemistry* 278.32, pp. 29463–70. DOI: 10.1074/jbc.M302977200 (cit. on p. 85).

Bourenkov, G. P. and A. N. Popov (Jan. 2006). "A quantitative approach to data-collection strategies." en. In: *Acta Crystallographica, Section D: Biological Crystallography* 62.Pt 1, pp. 58–64. DOI: 10.1107/S0907444905033998 (cit. on p. 18).

Bowen, B., J. J. Hawk, S. Sibunka, S. Hovick, and J. M. Weiler (Feb. 2001). "A review of the reported defects in the human C1 esterase inhibitor gene producing hereditary angioedema including four new mutations." In: *Clinical immunology (Orlando, Fla.)* 98.2, pp. 157–63. DOI: 10.1006/clim.2000.4947 (cit. on p. 85).

Bragg, W. L. (1913). "The Diffraction of Short Electromagnetic Waves by a Crystal". In: *Proceedings Of The Cambridge Philosophical Society* 17, pp. 43–57 (cit. on p. 8).

Bricogne, G., C. Vonrhein, C. Flensburg, M. Schiltz, and W. Paciorek (Oct. 2003). "Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0". In: *Acta Crystallographica, Section D: Biological Crystallography* 59.11, pp. 2023–2030. DOI: 10.1107/S0907444903017694 (cit. on pp. 26, 41).

Brünger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren (Sept. 1998). "Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination". In: *Acta Crystallographica, Section D: Biological Crystallography* 54.5, pp. 905–921. DOI: 10.1107/S0907444998003254 (cit. on p. 52).

Buchan, D. W. A. and S. M. Ward (July 2010). "Protein annotation and modelling servers at University College London". In: *Nucleic acids research* 38.Web Server issue, W563–8. DOI: 10.1093/nar/gkq427 (cit. on p. 71).

Burla, M. C., R. Caliandro, M. Camalli, B. Carrozzini, G. L. Cascarano, C. Giacovazzo, M. Mallamo, A. Mazzone, G. Polidori, and R. Spagna (Feb. 2012). "SIR2011 : a new package for crystal structure determination and refinement". In: *Journal of applied crystallography* 45.2, pp. 357–361. DOI: 10.1107/S0021889812001124 (cit. on p. 26).

Burla, M. C., B. Carrozzini, G. L. Cascarano, C. Giacovazzo, G. Polidori, and D. Siliqi (June 2002). "MAD phasing: probabilistic estimate of [F (oa)]." In: *Acta Crystallographica, Section D: Biological Crystallography* 58.Pt 6 Pt 2, pp. 928–35 (cit. on p. 40).

Cai, S. and A. E. Davis (Nov. 2003). "Complement regulatory protein C1 inhibitor binds to selectins and interferes with endothelial-leukocyte adhesion." In: *Journal of immunology (Baltimore, Md. : 1950)* 171.9, pp. 4786–91 (cit. on p. 84).

Cai, S., V. S. Dole, W. Bergmeier, J. Scafidi, H. Feng, D. D. Wagner, and A. E. Davis (May 2005). "A direct role for C1 inhibitor in regulation of leukocyte adhesion." In: *Journal of immunology (Baltimore, Md. : 1950)* 174.10, pp. 6462–6 (cit. on p. 84).

Calderone, V. (Dec. 2004). "Practical aspects of the integration of different software in protein structure solution." In: *Acta Crystallographica, Section D: Biological Crystallography* 60.Pt 12 Pt 1, pp. 2150–5. DOI: 10.1107/S0907444904019055 (cit. on p. 95).

Caldwell, E. E., A. M. Andreasen, M. A. Blietz, J. N. Serrahn, V. VanderNoot, Y. Park, G. Yu, R. J. Linhardt, and J. M. Weiler (Jan. 1999). "Heparin binding and augmentation of C1 inhibitor activity." In: *Archives of biochemistry and biophysics* 361.2, pp. 215–22. DOI: 10.1006/abbi.1998.0996 (cit. on p. 87).

Carrell, H. L., J. P. Glusker, V. Burger, F. Manfre, D. Tritsch, and J. F. Biellmann (June 1989). "X-ray analysis of D-xylose isomerase at 1.9 A: native enzyme in complex with substrate and with a mechanism-designed inactivator." In: *Proceedings of the National Academy of Sciences of the United States of America* 86.12, pp. 4440–4 (cit. on p. 95).

Caughman, G. B., R. J. Boackle, and J. Vesely (Feb. 1982). "A postulated mechanism for heparin's potentiation of C1 inhibitor function." In: *Molecular immunology* 19.2, pp. 287–95 (cit. on p. 87).

Champoux, J. J. (Jan. 2001). "DNA topoisomerases: structure, function, and mechanism." en. In: *Annual review of biochemistry* 70, pp. 369–413. DOI: 10.1146/annurev.biochem.70.1.369 (cit. on p. 4).

Chan, J. Y., C. E. Burrowes, F. M. Habal, and H. Z. Movat (Jan. 1977). "The inhibition of activated factor XII (Hageman factor) by antithrombin III: the effect of other plasma proteinase inhibitors." In: *Biochemical and biophysical research communications* 74.1, pp. 150–8 (cit. on p. 84).

Chen, C. C. H., H. Zhang, A. D. Kim, A. Howard, G. M. Sheldrick, D. Mariano-Dunaway, and O. Herzberg (Nov. 2002). "Degradation pathway of the phosphonate ciliatine: crystal structure of 2-aminoethylphosphonate transaminase." In: *Biochemistry* 41.44, pp. 13162–9 (cit. on p. 95).

Chen, L., r. Chen, X. E. Zhou, Y. Wang, M. A. Kahsai, A. T. Clark, S. P. Edmondson, Z.-j. Liu, J. P. Rose, c. Wang, E. J. Meehan, and J. W. Shriver (July 2004). "The hyperthermophile protein Sso10a is a dimer of winged helix DNA-binding domains linked by an antiparallel coiled coil rod." In: *Journal of molecular biology* 341.1, pp. 73–91. DOI: 10.1016/j.jmb.2004.05.044 (cit. on pp. 67, 77, 78).

Choli, T., P. Henning, B. Wittmann-Liebold, and R. Reinhardt (July 1988). "Isolation, characterization and microsequence analysis of a small basic methylated DNA-binding protein from the Archaebacterium, Sulfolobus solfataricus." In: *Biochimica et biophysica acta* 950.2, pp. 193–203 (cit. on p. 64).

Chou, C.-C., T.-W. Lin, C.-Y. Chen, and A. H.-J. Wang (July 2003). "Crystal Structure of the Hyperthermophilic Archaeal DNA-Binding Protein Sso10b2 at a Resolution of 1.85 Angstroms". In: *Journal of bacteriology* 185.14, pp. 4066–4073. DOI: 10.1128/JB.185.14.4066-4073.2003 (cit. on p. 66).

Collaborative Computational Project Number 4 (Sept. 1994). "The CCP4 suite: programs for protein crystallography." In: *Acta Crystallographica, Section D: Biological Crystallography* 50.Pt 5, pp. 760–3. DOI: 10.1107/S0907444994003112 (cit. on pp. 26, 41).

Cooper, N. R. (Jan. 1985). "The classical complement pathway: activation and regulation of the first complement component." In: *Advances in immunology* 37, pp. 151–216 (cit. on p. 84).

Cowtan, K. D. (1994). "'dm': An Automated Procedure for Phase Improvement by Density Modification". In: *Joint CCP4 AND ESF-EACBM Newsletter on Protein Crystallography Number 31* 32. Ed. by S. (L. Bailey and K. (c. D. Wilson, pp. 34–38 (cit. on pp. 38, 52).

— (Sept. 1999). "Error estimation and bias correction in phase-improvement calculations". In: *Acta Crystallographica, Section D: Biological Crystallography* 55.9, pp. 1555–1567. DOI: 10.1107/S0907444999007416 (cit. on p. 30).

— (Dec. 2000). “General quadratic functions in real and reciprocal space and their application to likelihood phasing.” In: *Acta Crystallographica, Section D: Biological Crystallography* 56.Pt 12, pp. 1612–21 (cit. on pp. 30, 38).
— (Sept. 2006). “The Buccaneer software for automated model building. 1. Tracing protein chains.” In: *Acta Crystallographica, Section D: Biological Crystallography* 62.Pt 9, pp. 1002–11. DOI: 10.1107/S0907444906022116 (cit. on pp. 32, 33, 38, 55, 58, 70).
— (Apr. 2010). “Recent developments in classical density modification.” In: *Acta Crystallographica, Section D: Biological Crystallography* 66.Pt 4, pp. 470–8. DOI: 10.1107/S090744490903947X (cit. on pp. 29, 30, 38, 70).
Cowtan, K. D. and P. Main (Jan. 1996). “Phase combination and cross validation in iterated density-modification calculations.” In: *Acta Crystallographica, Section D: Biological Crystallography* 52.Pt 1, pp. 43–8. DOI: 10.1107/S090744499500761X (cit. on p. 30).
Crick, F. H. C. (Sept. 1953). “The packing of α-helices: simple coiled-coils”. en. In: *Acta crystallographica* 6.8, pp. 689–697. DOI: 10.1107/S0365110X53001964 (cit. on p. 77).
Crick, F. H. C. and B. S. Magdoff (Nov. 1956). “The theory of the method of isomorphous replacement for protein crystals. I”. en. In: *Acta crystallographica* 9.11, pp. 901–908. DOI: 10.1107/S0365110X56002552 (cit. on pp. 15, 16).
Crowther, R. A. (1972). “The Molecular Replacement Method”. In: *New York: Gordon and Breach*. Ed. by M. G. Rossmann, pp. 173–178 (cit. on p. 14).
Curd, J. G., M. Yelvington, N. Burridge, N. P. Stimler, C. Gerard, L. J. Prograis, C. G. Cochrane, and H. J. Müller-Eberhard (1982). “Generation of bradykinin during incubation of hereditary angioedema plasma”. In: *Molecular immunology* 19.11, p. 1365 (cit. on p. 85).
Da Silva, W. D. (May 1967). “Complement as a mediator of inflammation: II. Biological properties of anaphylatoxin prepared with purified components of human complement”. In: *The Journal of experimental medicine* 125.5, pp. 921–946. DOI: 10.1084/jem.125.5.921 (cit. on p. 85).
Dauter, Z. (Apr. 2010). “Carrying out an optimal experiment.” en. In: *Acta Crystallographica, Section D: Biological Crystallography* 66.Pt 4, pp. 389–92. DOI: 10.1107/S0907444909038578 (cit. on p. 18).
Dauter, Z., M. Dauter, E. de La Fortelle, G. Bricogne, and G. M. Sheldrick (May 1999). “Can anomalous signal of sulfur become a tool for solving protein crystal structures?” In: *Journal of molecular biology* 289.1, pp. 83–92. DOI: 10.1006/jmbi.1999.2743 (cit. on p. 95).
Dauter, Z., M. Dauter, and E. J. Dodson (Feb. 2002). “Jolly SAD”. In: *Acta Crystallographica, Section D: Biological Crystallography* 58.3, pp. 494–506. DOI: 10.1107/S090744490200118X (cit. on p. 95).
Dauter, Z., M. Li, and A. Wlodawer (Mar. 2001). “Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics.” In: *Acta Crystallographica, Section D: Biological Crystallography* 57.Pt 2, pp. 239–49 (cit. on p. 95).

Dauter, Z., K. S. Wilson, L. C. Sieker, J. Meyer, and J. M. Moulis (Dec. 1997). "Atomic resolution (0.94 A) structure of Clostridium acidurici ferredoxin. Detailed geometry of [4Fe-4S] clusters in a protein." In: *Biochemistry* 36.51, pp. 16065–73. DOI: 10.1021/bi972155y (cit. on p. 95).

Davis, A. E., F. Lu, and P. Mejia (Nov. 2010). "C1 inhibitor, a multi-functional serine protease inhibitor." In: *Thrombosis and haemostasis* 104.5, pp. 886–93. DOI: 10.1160/TH10-01-0073 (cit. on pp. 84, 85).

Debaerdemaeker, T. and M. M. Woolfson (Mar. 1983). "On the application of phase relationships to complex structures. XXII. Techniques for random phase refinement". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 39.2, pp. 193–196. DOI: 10.1107/S0108767383000434 (cit. on p. 25).

De Graaff, R. A. G., M. Hilge, J. L. van der Plas, and J. P. Abrahams (Nov. 2001). "Matrix methods for solving protein substructures of chlorine and sulfur from anomalous data". In: *Acta Crystallographica, Section D: Biological Crystallography* 57.12, pp. 1857–1862. DOI: 10.1107/S0907444901016535 (cit. on pp. 26, 38, 55, 70).

Devedjiev, Y., Z. Dauter, S. R. Kuznetsov, T. L. Jones, and Z. S. Derewenda (Nov. 2000). "Crystal structure of the human acyl protein thioesterase I from a single X-ray data set to 1.5 A." In: *Structure (London, England : 1993)* 8.11, pp. 1137–46 (cit. on p. 95).

Diederichs, K. and P. A. Karplus (Apr. 1997). "Improved R-factors for diffraction data analysis in macromolecular crystallography". In: *Nature structural biology* 4.4, pp. 269–275. DOI: 10.1038/nsb0497-269 (cit. on p. 74).

Dijk, J. and R. Reinhardt (1986). *Bacterial chromatin*. Ed. by C. O. Gualerzi and C. L. Pon. Proceedings in life sciences. Berlin: Springer-Verlag, pp. 185–218 (cit. on pp. 64, 65).

Dill, K. A. (Aug. 1990). "Dominant forces in protein folding". In: *Biochemistry* 29.31, pp. 7133–7155. DOI: 10.1021/bi00483a001 (cit. on p. 6).

Donaldson, V. H. and R. R. Evans (July 1963). *A biochemical abnormality in hereditary angioneurotic edema: Absence of serum inhibitor of C'1-esterase* (cit. on p. 84).

Donaldson, V. H., O. D. Ratnoff, W. Dias Da Silva, and F. S. Rosen (1969). "Permeability-increasing activity in hereditary angioneurotic edema plasma. II. Mechanism of formation and partial characterization." In: *The Journal of clinical investigation* 48.4, pp. 642–653 (cit. on p. 85).

Donaldson, V. H., F. S. Rosen, and D. H. Bing (Jan. 1977). "Role of the second component of complement (C2) and plasmin in kinin release in hereditary angioneurotic edema (H.A.N.E.) plasma." In: *Transactions of the Association of American Physicians* 90, pp. 174–83 (cit. on p. 85).

Doutch, J., M. A. Hough, S. S. Hasnain, and R. W. Strange (Jan. 2012). "Challenges of sulfur SAD phasing as a routine method in macromolecular crystallography." In: *Journal of synchrotron radiation* 19.Pt 1, pp. 19–29. DOI: 10.1107/S0909049511049004 (cit. on pp. 18, 19).

Driessen, R. P. C. and R. T. Dame (Jan. 2011). "Nucleoid-associated proteins in Crenarchaea." In: *Biochemical Society transactions* 39.1, pp. 116–21. DOI: 10.1042/BST0390116 (cit. on p. 64).

— (Feb. 2013). "Structure and dynamics of the crenarchaeal nucleoid." In: *Biochemical Society transactions* 41.1, pp. 321–5. DOI: 10.1042/BST20120336 (cit. on p. 64).

Driessen, R. P. C., H. Meng, G. Suresh, R. Shahapure, G. Lanzani, U. D. Priyakumar, M. F. White, H. Schiessel, J. van Noort, and R. T. Dame (Jan. 2013). "Crenarchaeal chromatin proteins Cren7 and Sul7 compact DNA by inducing rigid bends." In: *Nucleic acids research* 41.1, pp. 196–205. DOI: 10.1093/nar/gks1053 (cit. on p. 65).

Duboule, D., F. Grosveld, E. Soutoglou, and T. Misteli (2007). "Mobility and immobility of chromatin in transcription and genome stability". In: *Current Opinion in Genetics & Development* 17.5, pp. 435–442 (cit. on p. 5).

Ducros, V. M., R. J. Lewis, C. S. Verma, E. J. Dodson, G. Leonard, J. P. Turkenburg, G. N. Murshudov, A. J. Wilkinson, and J. A. Brannigan (Mar. 2001). "Crystal structure of GerE, the ultimate transcriptional regulator of spore formation in Bacillus subtilis." In: *Journal of molecular biology* 306.4, pp. 759–71. DOI: 10.1006/jmbi.2001.4443 (cit. on pp. 47, 95).

Edmondson, S. P. and J. W. Shriver (2001). "DNA-binding proteins Sac7d and Sso7d from Sulfolobus". In: *Hyperthermophilic Enzymes, Part C*. Ed. by R. M. K. Michael W. W. Adams. Vol. 334. Methods in Enzymology. Academic Press, pp. 129–145. DOI: http://dx.doi.org/10.1016/S0076-6879(01)34463-4 (cit. on p. 64).

Emsley, P., B. Lohkamp, W. G. Scott, and K. D. Cowtan (Apr. 2010). "Features and development of Coot." In: *Acta Crystallographica, Section D: Biological Crystallography* 66.Pt 4, pp. 486–501. DOI: 10.1107/S0907444910007493 (cit. on pp. 49, 70).

Eriksson, a. E., W. a. Baase, X. J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, and B. W. Matthews (Jan. 1992). "Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect." In: *Science* 255.5041, pp. 178–83 (cit. on p. 6).

Evans, P. R. (Jan. 2006). "Scaling and assessment of data quality." In: *Acta Crystallographica, Section D: Biological Crystallography* 62.Pt 1, pp. 72–82. DOI: 10.1107/S0907444905036693 (cit. on pp. 74, 75).

— (Apr. 2011). "An introduction to data reduction: space-group determination, scaling and intensity statistics." In: *Acta Crystallographica, Section D: Biological Crystallography* 67.Pt 4, pp. 282–92. DOI: 10.1107/S090744491003982X (cit. on p. 70).

Fan, H.-F., Q. Hao, Y.-X. Gu, J.-Z. Qian, C.-D. Zheng, and H.-M. Ke (Nov. 1990). "Combining direct methods with isomorphous replacement or anomalous scattering data. VII. Ab initio phasing of one-wavelength anomalous scattering data from a small protein". en. In: *Acta Crystallographica, Section A: Foundations of Crystallography* 46.11, pp. 935–939. DOI: 10.1107/S0108767390008030 (cit. on p. 20).

Feng, Y., H. Yao, and J. Wang (June 2010). “Crystal structure of the crenarchaeal conserved chromatin protein Cren7 and double-stranded DNA complex.” In: *Protein science : a publication of the Protein Society* 19.6, pp. 1253–7. DOI: 10.1002/pro.385 (cit. on pp. 65, 66).

Fields, T., B. Ghebrehiwet, and A. P. Kaplan (July 1983). “Kinin formation in hereditary angioedema plasma: evidence against kinin derivation from C2 and in support of "spontaneous" formation of bradykinin.” In: *The Journal of allergy and clinical immunology* 72.1, pp. 54–60 (cit. on p. 85).

Ford, M. G., B. M. Pearse, M. K. Higgins, Y. Vallis, D. J. Owen, A. Gibson, C. R. Hopkins, P. R. Evans, and H. T. McMahon (Feb. 2001). “Simultaneous binding of PtdIns(4,5)P2 and clathrin by AP180 in the nucleation of clathrin lattices on membranes.” In: *Science* 291.5506, pp. 1051–5. DOI: 10.1126/science.291.5506.1051 (cit. on p. 95).

Forterre, P., F. Confalonieri, and S. Knapp (May 1999). “Identification of the gene encoding archeal-specific DNA-binding proteins of the Sac10b family.” In: *Molecular microbiology* 32.3, pp. 669–70 (cit. on p. 65).

Franklin, R. E. and R. G. Gosling (Apr. 1953). “Molecular Configuration in Sodium Thymonucleate”. In: *Nature* 171.4356, pp. 740–741. DOI: 10.1038/171740a0 (cit. on p. 3).

Gajiwala, K. S. and S. K. Burley (Feb. 2000). “Winged helix proteins.” In: *Current opinion in structural biology* 10.1, pp. 110–6 (cit. on p. 67).

Gao, Y. G., S. Y. Su, H. Robinson, S. Padmanabhan, L. Lim, B. S. McCrary, S. P. Edmondson, J. W. Shriver, and A. H.-J. Wang (Sept. 1998). “The crystal structure of the hyperthermophile chromosomal protein Sso7d bound to DNA.” In: *Nature structural biology* 5.9, pp. 782–6. DOI: 10.1038/1822 (cit. on p. 65).

Giacovazzo, C. (Oct. 2008). “Direct methods”. In: *International Tables for Crystallography Volume B: Reciprocal space*. Ed. by U. Shmueli. 3rd ed. Vol. B. International Tables for Crystallography. Dordrecht, The Netherlands: Springer. Chap. 2.2, pp. 215–235. DOI: 10.1107/97809553602060000102 (cit. on p. 12).

Giorgetti, A., D. Raimondo, A. E. Miele, and A. Tramontano (Sept. 2005). “Evaluating the usefulness of protein structure models for molecular replacement.” In: *Bioinformatics (Oxford, England)* 21 Suppl 2, pp. ii72–6. DOI: 10.1093/bioinformatics/bti1112 (cit. on p. 13).

Gordon, E. J., G. A. Leonard, S. McSweeney, and P. F. Zagalsky (Sept. 2001). “The C1 subunit of alpha-crustacyanin: the de novo phasing of the crystal structure of a 40 kDa homodimeric protein using the anomalous scattering from S atoms combined with direct methods.” In: *Acta Crystallographica, Section D: Biological Crystallography* 57.Pt 9, pp. 1230–7 (cit. on p. 95).

Grosse-Kunstleve, R. W. and P. D. Adams (Oct. 2003). “Substructure search procedures for macromolecular structures”. In: *Acta Crystallographica, Section D: Biological Crystallography* 59.11, pp. 1966–1973. DOI: 10.1107/S0907444903018043 (cit. on p. 26).

Groth, A., W. Rocha, A. Verreault, and G. Almouzni (2007). "Chromatin Challenges during DNA Replication and Repair". In: *Cell* 128.4, pp. 721–733 (cit. on p. 5).

Guo, L., Y. Feng, Z. Zhang, H. Yao, Y. Luo, J. Wang, and L. Huang (Mar. 2008). "Biochemical and structural characterization of Cren7, a novel chromatin protein conserved among Crenarchaea." In: *Nucleic acids research* 36.4, pp. 1129–37. DOI: 10.1093/nar/gkm1128 (cit. on pp. 65, 66).

Hauptmann, H. A. (May 1982). "On integrating the techniques of direct methods and isomorphous replacement. I. The theoretical basis". en. In: *Acta Crystallographica, Section A: Foundations of Crystallography* 38.3, pp. 289–294. DOI: 10.1107/S0567739482000631 (cit. on p. 20).

— (May 1996). "The SAS Maximal Principle: a New Approach to the Phase Problem". en. In: *Acta Crystallographica, Section A: Foundations of Crystallography* 52.3, pp. 490–496. DOI: 10.1107/S0108767396001237 (cit. on p. 20).

Hauptmann, H. A. and J. Karle (1953). *Solution of the Phase Problem: 1. The Centrosymmetric Crystal*. Acta monograph v. 1. Pittsburgh, PA: American Crystallographic Association (cit. on p. 25).

Hendrickson, W. A. and E. E. Lattman (Feb. 1970). "Representation of phase probability distributions for simplified combination of independent phase information". In: *Acta Crystallographica, Section B: Structural Crystallography and Crystal Chemistry* 26.2, pp. 136–143. DOI: 10.1107/S0567740870002078 (cit. on pp. 29, 42, 43, 52).

Hendrickson, W. A. and C. M. Ogata (1997). "Phase determination from multiwavelength anomalous diffraction measurements". In: *Macromolecular Crystallography Part A*. Ed. by J. Charles W. Carter. Vol. 276. Methods in Enzymology. Academic Press. Chap. 28, pp. 494–523. DOI: 10.1016/S0076-6879(97)76074-9 (cit. on p. 19).

Hendrickson, W. A. and M. M. Teeter (Mar. 1981). "Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur". In: *Nature* 290.5802, pp. 107–113. DOI: 10.1038/290107a0 (cit. on pp. 18, 20).

Henne, W. M., H. M. Kent, M. G. J. Ford, B. G. Hegde, O. Daumke, P. J. G. Butler, R. Mittal, R. Langen, P. R. Evans, and H. T. McMahon (July 2007). "Structure and analysis of FCHo2 F-BAR domain: a dimerizing and membrane recruitment module that effects membrane curvature." In: *Structure (London, England : 1993)* 15.7, pp. 839–52. DOI: 10.1016/j.str.2007.05.002 (cit. on p. 95).

Hsu, C.-H. and A. H.-J. Wang (Aug. 2011). "The DNA-recognition fold of Sso7c4 suggests a new member of SpoVT-AbrB superfamily from archaea." In: *Nucleic acids research* 39.15, pp. 6764–74. DOI: 10.1093/nar/gkr283 (cit. on pp. 64, 65).

Huntington, J. A., R. J. Read, and R. W. Carrell (2000). "Structure of a serpin-protease complex shows inhibition by deformation." In: *Nature* 407.6806, pp. 923–6. DOI: 10.1038/35038119 (cit. on p. 86).

Ingram, V. M. (June 1958). "Abnormal human haemoglobins. I. The comparison of normal human and sickle-cell haemoglobins by fingerprinting." In: *Biochimica et biophysica acta* 28.3, pp. 539–45. DOI: 10.1016/0006-3002(58)90516-X (cit. on p. 6).

Kabsch, W. (Feb. 2010). "XDS." In: *Acta Crystallographica, Section D: Biological Crystallography* 66.Pt 2, pp. 125–32. DOI: 10.1107/S0907444909047337 (cit. on p. 70).

Kahsai, M. A., B. Vogler, A. T. Clark, S. P. Edmondson, and J. W. Shriver (Mar. 2005). "Solution structure, stability, and flexibility of Sso10a: a hyperthermophile coiled-coil DNA-binding protein." In: *Biochemistry* 44.8, pp. 2822–32. DOI: 10.1021/bi047669t (cit. on p. 67).

Karle, J. and H. A. Hauptmann (Aug. 1956). "A theory of phase determination for the four types of non-centrosymmetric space groups 1 P 222, 2 P 22, 3 P 1 2, 3 P 2 2". In: *Acta crystallographica* 9.8, pp. 635–651. DOI: 10.1107/S0365110X56001741 (cit. on p. 26).

Kendrew, J. C., G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips (Mar. 1958). "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". In: *Nature* 181.4610, pp. 662–666. DOI: 10.1038/181662a0 (cit. on pp. 3, 14).

Klemperer, M. R., V. H. Donaldson, and F. S. Rosen (1968). "Effect of C'1 esterase on vascular permeability in man: studies in normal and complement-deficient individuals and in patients with hereditary angioneurotic edema". In: *The Journal of clinical investigation* 47.3, pp. 604–611 (cit. on p. 85).

Knapp, S., A. Karshikoff, K. D. Berndt, P. Christova, B. Atanasov, and R. Ladenstein (Dec. 1996). "Thermal unfolding of the DNA-binding protein Sso7d from the hyperthermophile Sulfolobus solfataricus." In: *Journal of molecular biology* 264.5, pp. 1132–44. DOI: 10.1006/jmbi.1996.0701 (cit. on p. 64).

Krissinel, E. and K. Henrick (Sept. 2007). "Inference of macromolecular assemblies from crystalline state." In: *Journal of molecular biology* 372.3, pp. 774–97. DOI: 10.1016/j.jmb.2007.05.022 (cit. on p. 71).

Lamers, M. H., A. Perrakis, J. H. Enzlin, H. H. Winterwerp, N. de Wind, and T. K. Sixma (Oct. 2000). "The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch." In: *Nature* 407.6805, pp. 711–7. DOI: 10.1038/35037523 (cit. on p. 95).

Lämmle, B., B. L. Zuraw, M. J. Heeb, H. P. Schwarz, M. Berrettini, J. G. Curd, and J. H. Griffin (Apr. 1988). "Detection and quantitation of cleaved and uncleaved high molecular weight kininogen in plasma by ligand blotting with radiolabeled plasma prekallikrein or factor XI." In: *Thrombosis and haemostasis* 59.2, pp. 151–61 (cit. on p. 85).

Lander, E. S. et al. (Feb. 2001). "Initial sequencing and analysis of the human genome." In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062 (cit. on p. 4).

Langer, G., S. X. Cohen, V. S. Lamzin, and A. Perrakis (Jan. 2008). "Automated macromolecular model building for X-ray crystallography using

ARP/wARP version 7." In: *Nature protocols* 3.7, pp. 1171–9. DOI: 10.1038/nprot.2008.91 (cit. on pp. 32, 38, 55).

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins (Nov. 2007). "Clustal W and Clustal X version 2.0." In: *Bioinformatics (Oxford, England)* 23.21, pp. 2947–8. DOI: 10.1093/bioinformatics/btm404 (cit. on p. 71).

Laurens, N., R. P. C. Driessen, I. Heller, D. Vorselen, M. C. Noom, F. J. H. Hol, M. F. White, R. T. Dame, and G. J. L. Wuite (Jan. 2012). "Alba shapes the archaeal genome using a delicate balance of bridging and stiffening the DNA." In: *Nature communications* 3, p. 1328. DOI: 10.1038/ncomms2330 (cit. on p. 66).

Leslie, A. G. W. (Jan. 1987). "A reciprocal-space method for calculating a molecular envelope using the algorithm of B.C. Wang". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 43.1, pp. 134–136. DOI: 10.1107/S0108767387099720 (cit. on pp. 27, 52).

Li, G., M. Levitus, C. Bustamante, and J. Widom (Jan. 2005). "Rapid spontaneous accessibility of nucleosomal DNA." In: *Nature structural & molecular biology* 12.1, pp. 46–53. DOI: 10.1038/nsmb869 (cit. on p. 5).

Li, J., U. Derewenda, Z. Dauter, S. Smith, and Z. S. Derewenda (July 2000). "Crystal structure of the Escherichia coli thioesterase II, a homolog of the human Nef binding enzyme." In: *Nature structural biology* 7.7, pp. 555–9. DOI: 10.1038/76776 (cit. on p. 95).

Liu, D., S. Cai, X. Gu, J. Scafidi, X. Wu, and A. E. Davis (Sept. 2003). "C1 inhibitor prevents endotoxin shock via a direct interaction with lipopolysaccharide." In: *Journal of immunology (Baltimore, Md. : 1950)* 171.5, pp. 2594–601 (cit. on p. 84).

Liu, D., X. Gu, J. Scafidi, and A. E. Davis (Apr. 2004). "N-linked glycosylation is required for c1 inhibitor-mediated protection from endotoxin shock in mice." In: *Infection and immunity* 72.4, pp. 1946–55 (cit. on p. 84).

Liu, Z., F. Galli, W.-J. Waterreus, E. M. Meulenbroek, R. I. Koning, G. E. M. Lamers, R. C. L. Olsthoorn, N. S. Pannu, T. H. Oosterkamp, A. J. Koster, R. T. Dame, and J. P. Abrahams (Apr. 2012). "Single-walled carbon nanotubes as scaffolds to concentrate DNA for the study of DNA-protein interactions." In: *Chemphyschem : a European journal of chemical physics and physical chemistry* 13.6, pp. 1569–75. DOI: 10.1002/cphc.201100896.

Lodish, H. F., A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell (2000). *Molecular Cell Biology*. 4th ed. New York: W. H. Freeman and Company (cit. on p. 4).

Luijsterburg, M. S., M. F. White, R. van Driel, and R. T. Dame (2008). "The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes." In: *Critical reviews in biochemistry and molecular biology* 43.6, pp. 393–418. DOI: 10.1080/10409230802528488 (cit. on pp. 5, 64).

Lundbäck, T., H. Hansson, S. Knapp, R. Ladenstein, and T. Härd (Mar. 1998). "Thermodynamic characterization of non-sequence-specific DNA-binding

by the Sso7d protein from Sulfolobus solfataricus." In: *Journal of molecular biology* 276.4, pp. 775–86. DOI: 10.1006/jmbi.1997.1558 (cit. on p. 64).

Lunin, V. Y. and A. G. Urzhumtsev (May 1984). "Improvement of protein phases by coarse model modification". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 40.3, pp. 269–277. DOI: 10.1107/S0108767384000544 (cit. on pp. 29, 52).

Lurz, R., M. Grote, J. Dijk, R. Reinhardt, and B. Dobrinski (Dec. 1986). "Electron microscopic study of DNA complexes with proteins from the Archaebacterium Sulfolobus acidocaldarius." In: *The EMBO journal* 5.13, pp. 3715–21 (cit. on p. 67).

Luzzati, V. (Nov. 1952). "Traitement statistique des erreurs dans la determination des structures cristallines". In: *Acta crystallographica* 5.6, pp. 802–810. DOI: 10.1107/S0365110X52002161 (cit. on p. 41).

Mai, V. Q., X. Chen, R. Hong, and L. Huang (May 1998). "Small abundant DNA binding proteins from the thermoacidophilic archaeon Sulfolobus shibatae constrain negative DNA supercoils." In: *Journal of bacteriology* 180.9, pp. 2560–3 (cit. on p. 64).

Marsh, V. L., S. Y. Peak-Chew, and S. D. Bell (June 2005). "Sir2 and the acetyltransferase, Pat, regulate the archaeal chromatin protein, Alba." In: *The Journal of biological chemistry* 280.22, pp. 21122–8. DOI: 10.1074/jbc.M501280200 (cit. on p. 66).

Matias, P. M., A. V. Coelho, F. M. A. Valente, D. Plácido, J. LeGall, A. V. Xavier, I. A. C. Pereira, and M. A. Carrondo (Dec. 2002). "Sulfate respiration in Desulfovibrio vulgaris Hildenborough. Structure of the 16-heme cytochrome c HmcA AT 2.5-A resolution and a view of its role in transmembrane electron transfer." In: *The Journal of biological chemistry* 277.49, pp. 47907–16. DOI: 10.1074/jbc.M207465200 (cit. on p. 95).

Matsushita, M., S. Thiel, J. C. Jensenius, I. Terai, and T. Fujita (Sept. 2000). "Proteolytic activities of two types of mannose-binding lectin-associated serine protease." In: *Journal of immunology (Baltimore, Md. : 1950)* 165.5, pp. 2637–42 (cit. on p. 84).

Matthews, B. W. (Jan. 1966). "The extension of the isomorphous replacement method to include anomalous scattering measurements". In: *Acta crystallographica* 20.1, pp. 82–86. DOI: 10.1107/S0365110X6600015X (cit. on p. 41).

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read (Aug. 2007). "Phaser crystallographic software." In: *Journal of applied crystallography* 40.Pt 4, pp. 658–674. DOI: 10.1107/S0021889807021206 (cit. on p. 74).

Miller, R., G. T. DeTitta, R. Jones, D. A. Langs, C. M. Weeks, and H. A. Hauptmann (Mar. 1993). "On the application of the minimal principle to solve unknown structures." In: *Science* 259.5100, pp. 1430–3 (cit. on p. 25).

Miller, W. H. (1839). *A Treatise on Crystallography*. Cambridge: J. & J. J. Deighton (cit. on p. 8).

Misteli, T. and E. Soutoglou (Apr. 2009). "The emerging role of nuclear architecture in DNA repair and genome maintenance." In: *Nature reviews.*

Molecular cell biology 10.4, pp. 243–54. DOI: 10.1038/nrm2651 (cit. on p. 5).

Morris, R. J. and G. Bricogne (Feb. 2003). "Sheldrick's 1.2 Årule and beyond". In: *Acta Crystallographica, Section D: Biological Crystallography* 59.3, pp. 615–617. DOI: 10.1107/S090744490300163X (cit. on p. 13).

Morris, R. J., A. Perrakis, and V. S. Lamzin (May 2002). "ARP / wARP 's model-building algorithms. I. The main chain". In: *Acta Crystallographica, Section D: Biological Crystallography* 58.6, pp. 968–975. DOI: 10.1107/S0907444902005462 (cit. on p. 32).

Morris, R. J., P. H. Zwart, S. X. Cohen, F. J. Fernandez, M. Kakaris, O. Kirillova, C. Vonrhein, A. Perrakis, and V. S. Lamzin (Nov. 2003). "Breaking good resolutions with ARP/wARP". In: *Journal of synchrotron radiation* 11.1, pp. 56–59. DOI: 10.1107/S090904950302394X (cit. on p. 32).

Mueller-Dieckmann, C., S. Panjikar, A. Schmidt, S. Mueller, J. Kuper, A. Geerlof, M. Wilmanns, R. K. Singh, P. A. Tucker, and M. S. Weiss (Mar. 2007). "On the routine use of soft X-rays in macromolecular crystallography. Part IV. Efficient determination of anomalous substructures in biomacromolecules using longer X-ray wavelengths." en. In: *Acta Crystallographica, Section D: Biological Crystallography* 63.Pt 3, pp. 366–80. DOI: 10.1107/S0907444906055624 (cit. on pp. 48, 95).

Muirhead, H., J. M. Cox, L. Mazzarella, and M. F. Perutz (1967). "Structure and function of haemoglobin: III. A three-dimensional fourier synthesis of human deoxyhaemoglbin at 5·5 Åresolution". In: *Journal of molecular biology* 28.1, pp. 117–150. DOI: http://dx.doi.org/10.1016/S0022-2836(67)80082-2 (cit. on pp. 27, 28).

Murshudov, G. N., P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin (Apr. 2011). "REFMAC5 for the refinement of macromolecular crystal structures." In: *Acta Crystallographica, Section D: Biological Crystallography* 67.Pt 4, pp. 355–67. DOI: 10.1107/S0907444911001314 (cit. on pp. 33, 38, 55, 70).

Murshudov, G. N., A. A. Vagin, and E. J. Dodson (May 1997). "Refinement of macromolecular structures by the maximum-likelihood method." In: *Acta Crystallographica, Section D: Biological Crystallography* 53.Pt 3, pp. 240–55. DOI: 10.1107/S0907444996012255 (cit. on p. 55).

Näsvall, J., L. Sun, J. R. Roth, and D. I. Andersson (Oct. 2012). "Real-time evolution of new genes by innovation, amplification, and divergence." In: *Science* 338.6105, pp. 384–7. DOI: 10.1126/science.1226521 (cit. on p. 5).

Neel, J. V. (July 1949). "The Inheritance of Sickle Cell Anemia." In: *Science* 110.2846, pp. 64–6. DOI: 10.1126/science.110.2846.64 (cit. on p. 6).

Ness, S. R., R. A. G. de Graaff, J. P. Abrahams, and N. S. Pannu (Oct. 2004). "CRANK: new methods for automated macromolecular crystal structure solution." In: *Structure (London, England : 1993)* 12.10, pp. 1753–61. DOI: 10.1016/j.str.2004.07.018 (cit. on pp. 33, 38, 41, 43, 55, 70).

Nicholls, R. A., F. Long, and G. N. Murshudov (Apr. 2012). "Low-resolution refinement tools in REFMAC5." In: *Acta Crystallographica, Section*

D: Biological Crystallography 68.Pt 4, pp. 404–17. DOI: 10.1107/S090744491105606X (cit. on p. 33).

North, A. C. T. (Feb. 1965). "The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions". In: *Acta crystallographica* 18.2, pp. 212–216. DOI: 10.1107/S0365110X65000488 (cit. on p. 41).

Nuijens, J. H., A. J. Eerenberg-Belmer, C. C. Huijbregts, W. O. Schreuder, R. J. Felt-Bersma, J. J. Abbink, L. G. Thijs, and C. E. Hack (Aug. 1989). "Proteolytic inactivation of plasma C1- inhibitor in sepsis." In: *The Journal of clinical investigation* 84.2, pp. 443–50. DOI: 10.1172/JCI114185 (cit. on p. 84).

Nussberger, J., M. Cugno, C. Amstutz, M. Cicardi, A. Pellacani, and A. Agostoni (June 1998). "Plasma bradykinin in angio-oedema." In: *Lancet* 351.9117, pp. 1693–7. DOI: 10.1016/S0140-6736(97)09137-X (cit. on p. 85).

Oberdoerffer, P. and D. A. Sinclair (Sept. 2007). "The role of nuclear architecture in genomic instability and ageing." In: *Nature reviews. Molecular cell biology* 8.9, pp. 692–702. DOI: 10.1038/nrm2238 (cit. on p. 5).

Ohno, S. (1970). *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag. (cit. on p. 5).

Osler, W. (1888). "Hereditary Angio-Neurotic Oedema". In: *The American journal of the medical sciences* 95.4 (cit. on p. 85).

Otwinowski, Z., W. Wolf, P. R. Evans, and A. G. W. Leslie (1991). "Isomorphous replacement and anomalous scattering". In: *Proceedings of the CCP4 Study Weekend, 25-26 Jan 1991.* Ed. by W. (L. Wolf, P. R. (C. Evans, and A. G. W. (C. Leslie. Vol. None, pp. 80–86 (cit. on pp. 26, 41).

Owen, D. J., P. Ornaghi, J. C. Yang, N. Lowe, P. R. Evans, P. Ballario, D. Neuhaus, P. Filetici, and A. A. Travers (Nov. 2000). "The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p." In: *The EMBO journal* 19.22, pp. 6141–9. DOI: 10.1093/emboj/19.22.6141 (cit. on p. 95).

Owen, D. J., Y. Vallis, B. M. Pearse, H. T. McMahon, and P. R. Evans (Aug. 2000). "The structure and function of the beta 2-adaptin appendage domain." In: *The EMBO journal* 19.16, pp. 4216–27. DOI: 10.1093/emboj/19.16.4216 (cit. on p. 95).

Palade, G. E. (Aug. 1964). "The organization of living matter". In: *Proceedings of the National Academy of Sciences of the United States of America* 52.5, pp. 613–34 (cit. on p. 4).

Panjikar, S., V. Parthasarathy, V. S. Lamzin, M. S. Weiss, and P. A. Tucker (Apr. 2005). "Auto-rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment." In: *Acta Crystallographica, Section D: Biological Crystallography* 61.Pt 4, pp. 449–57. DOI: 10.1107/S0907444905001307 (cit. on pp. 34, 40).

Pannu, N. S., A. J. McCoy, and R. J. Read (Sept. 2003). "Application of the complex multivariate normal distribution to crystallographic methods

with insights into multiple isomorphous replacement phasing". In: *Acta Crystallographica, Section D: Biological Crystallography* 59.10, pp. 1801–1808. DOI: 10.1107/S090744490301936X (cit. on p. 45).

Pannu, N. S., G. N. Murshudov, E. J. Dodson, and R. J. Read (Nov. 1998). "Incorporation of Prior Phase Information Strengthens Maximum-Likelihood Structure Refinement". In: *Acta Crystallographica, Section D: Biological Crystallography* 54.6, pp. 1285–1294. DOI: 10.1107/S0907444998004119 (cit. on pp. 30, 42).

Pannu, N. S. and R. J. Read (Dec. 2004). "The application of multivariate statistical techniques improves single-wavelength anomalous diffraction phasing". In: *Acta Crystallographica, Section D: Biological Crystallography* 60.1, pp. 22–27. DOI: 10.1107/S0907444903020808 (cit. on pp. 27, 38, 41, 54, 55, 70).

Pannu, N. S., P. Skubák, I. Sikharulidze, J. P. Abrahams, and R. A. G. de Graaff (Aug. 2007). "Recent advances in the CRANK automated structure solution suite". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 63.a1, s116. DOI: 10.1107/S0108767307097516 (cit. on p. 41).

Pannu, N. S., W.-J. Waterreus, P. Skubák, I. Sikharulidze, J. P. Abrahams, and R. A. G. de Graaff (Apr. 2011). "Recent advances in the CRANK software suite for experimental phasing." In: *Acta Crystallographica, Section D: Biological Crystallography* 67.Pt 4, pp. 331–7. DOI: 10.1107/S0907444910052224 (cit. on p. 37).

Pereira, S. L., R. A. Grayling, R. Lurz, and J. N. Reeve (Nov. 1997). "Archaeal nucleosomes". In: *Proceedings of the National Academy of Sciences of the United States of America* 94.23, pp. 12633–7 (cit. on p. 64).

Perrakis, A., R. J. Morris, and V. S. Lamzin (May 1999). "Automated protein model building combined with iterative structure refinement." In: *Nature structural biology* 6.5, pp. 458–63. DOI: 10.1038/8263 (cit. on pp. 32, 58).

Perutz, M. F. (Nov. 1956). "Isomorphous replacement and phase determination in non-centrosymmetric space groups". en. In: *Acta crystallographica* 9.11, pp. 867–873. DOI: 10.1107/S0365110X56002485 (cit. on p. 15).

Pixley, R. A., M. Schapira, and R. W. Colman (Feb. 1985). "The regulation of human factor XIIa by plasma proteinase inhibitors." In: *The Journal of biological chemistry* 260.3, pp. 1723–9 (cit. on p. 84).

Plosker, G. L. (Oct. 2012). "Recombinant human c1 inhibitor (conestat alfa): in the treatment of angioedema attacks in hereditary angioedema." In: *BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy* 26.5, pp. 315–23. DOI: 10.2165/11206880-000000000-00000 (cit. on pp. 7, 87).

Ramagopal, U. A., M. Dauter, and Z. Dauter (May 2003). "Phasing on anomalous signal of sulfurs: what is the limit?" In: *Acta Crystallographica, Section D: Biological Crystallography* 59.6, pp. 1020–1027. DOI: 10.1107/S0907444903007467 (cit. on p. 18).

Rappleye, J., M. Innus, C. M. Weeks, and R. Miller (May 2002). "SnB version 2.2: an example of crystallographic multiprocessing". In: *Journal of applied*

crystallography 35.3, pp. 374–376. DOI: 10.1107/S0021889802005782 (cit. on p. 13).

Read, R. J. (May 1986). "Improved Fourier coefficients for maps using phases from partial structures with errors". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 42.3, pp. 140–149. DOI: 10.1107/S0108767386099622 (cit. on pp. 29, 52, 55).

— (1997). "Model phases: Probabilities and bias". In: *Macromolecular Crystallography Part B*. Ed. by R. M. S. Charles W. Carter Jr. Vol. 277. Methods in Enzymology. Academic Press. Chap. 7, pp. 110–128. DOI: 10.1016/S0076-6879(97)77009-5 (cit. on pp. 29, 52).

Robinson, H., Y. G. Gao, B. S. McCrary, S. P. Edmondson, J. W. Shriver, and A. H.-J. Wang (Mar. 1998). "The hyperthermophile chromosomal protein Sac7d sharply kinks DNA." In: *Nature* 392.6672, pp. 202–5. DOI: 10.1038/32455 (cit. on p. 64).

Rossmann, M. G. (Feb. 1990). "The molecular replacement method". en. In: *Acta Crystallographica, Section A: Foundations of Crystallography* 46.2, pp. 73–82. DOI: 10.1107/S0108767389009815 (cit. on p. 14).

Rould, M. A. (1997). "Screening for heavy-atom derivatives and obtaining accurate isomorphous differences". In: *Macromolecular Crystallography Part A*. Ed. by J. Charles W. Carter. Vol. 276. Methods in Enzymology. Academic Press. Chap. 26, pp. 461–472. DOI: 10.1016/S0076-6879(97)76072-5 (cit. on p. 15).

Ruhaak, L. R., C. Huhn, W.-J. Waterreus, A. R. de Boer, C. Neusüss, C. H. Hokke, A. M. Deelder, and M. Wuhrer (Aug. 2008). "Hydrophilic interaction chromatography-based high-throughput sample preparation method for N-glycan analysis from total human plasma glycoproteins." In: *Analytical chemistry* 80.15, pp. 6119–26. DOI: 10.1021/ac800630x.

Schaefer, C. and B. Rost (Jan. 2012). "Predict impact of single amino acid change upon protein structure." In: *BMC genomics* 13 Suppl 4, S4. DOI: 10.1186/1471-2164-13-S4-S4 (cit. on p. 6).

Schapira, M., A. de Agostini, J. A. Schifferli, and R. W. Colman (Jan. 1985). "Biochemistry and pathophysiology of human C1 inhibitor: current issues." In: *Complement (Basel, Switzerland)* 2.2-3, pp. 111–26 (cit. on p. 84).

Schapira, M., C. F. Scott, and R. W. Colman (Feb. 1982). "Contribution of plasma protease inhibitors to the inactivation of kallikrein in plasma." In: *The Journal of clinical investigation* 69.2, pp. 462–8 (cit. on p. 84).

Schapira, M., L. D. Silver, C. F. Scott, A. H. Schmaier, L. J. Prograis, J. G. Curd, and R. W. Colman (1983). "Prekallikrein activation and high-molecular-weight kininogen consumption in hereditary angioedema." In: *The New England journal of medicine* 308.18, pp. 1050–1053 (cit. on p. 85).

Schiltz, M., W. Shepard, R. Fourme, T. Prangé, E. de La Fortelle, and G. Bricogne (Jan. 1997). "High-pressure krypton gas and statistical heavy-atom refinement: a successful combination of tools for macromolecular structure determination." In: *Acta Crystallographica, Section D: Biological Crystallography* 53.Pt 1, pp. 78–92. DOI: 10.1107/S0907444996009705 (cit. on p. 95).

Schneider, T. R. and G. M. Sheldrick (Sept. 2002). "Substructure solution with SHELXD". In: *Acta Crystallographica, Section D: Biological Crystallography* 58.10, pp. 1772–1779. DOI: 10.1107/S0907444902011678 (cit. on pp. 13, 38, 55).

Schrödinger LLC (Aug. 2010). "The PyMOL Molecular Graphics System, Version 1.5.0.3" (cit. on p. 71).

Sevcik, J., Z. Dauter, V. S. Lamzin, and K. S. Wilson (Mar. 1996). "Ribonuclease from Streptomyces aureofaciens at atomic resolution." In: *Acta Crystallographica, Section D: Biological Crystallography* 52.Pt 2, pp. 327–44. DOI: 10.1107/S0907444995007669 (cit. on p. 95).

Sevcik, J., E. J. Dodson, and G. G. Dodson (Apr. 1991). "Determination and restrained least-squares refinement of the structures of ribonuclease Sa and its complex with 3'-guanylic acid at 1.8 Åresolution". In: *Acta Crystallographica, Section B: Structural Science* 47.2, pp. 240–253. DOI: 10.1107/S0108768190009569 (cit. on p. 95).

Sharp, K. A. (Apr. 1991). "The hydrophobic effect". In: *Current opinion in structural biology* 1.2, pp. 171–174. DOI: 10.1016/0959-440X(91)90057-Z (cit. on p. 6).

Sheldrick, G. M. (1990). "Phase annealing in SHELX-90: direct methods for larger structures". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 46.6, pp. 467–473. DOI: 10.1107/S0108767390000277 (cit. on p. 12).

— (Dec. 2002). "Macromolecular phasing with SHELXE". In: *Zeitschrift für Kristallographie* 217.12-2002, pp. 644–650. DOI: 10.1524/zkri.217.12.644.20662 (cit. on pp. 32, 38, 52).

— (Jan. 2008). "A short history of SHELX." In: *Acta Crystallographica, Section A: Foundations of Crystallography* 64.Pt 1, pp. 112–22. DOI: 10.1107/S0108767307043930 (cit. on pp. 38, 55).

Shi, N., S. Ye, A. Alam, L. Chen, and Y. Jiang (Mar. 2006). "Atomic structure of a Na+- and K+-conducting channel." In: *Nature* 440.7083, pp. 570–4. DOI: 10.1038/nature04508 (cit. on p. 95).

Shoemaker, L. R., S. J. Schurman, V. H. Donaldson, and A. E. Davis (1994). "Hereditary angioneurotic oedema: characterization of plasma kinin and vascular permeability-enhancing activities." In: *Clinical and experimental immunology* 95.1, pp. 22–28 (cit. on p. 85).

Shriver, J. W., W. B. Peters, N. Szary, A. T. Clark, and S. P. Edmondson (2001). "Calorimetric analyses of hyperthermophile proteins". In: *Hyperthermophilic Enzymes, Part C*. Ed. by R. M. K. Michael W. W. Adams. Vol. 334. Methods in Enzymology. Academic Press, pp. 389–422. DOI: http://dx.doi.org/10.1016/S0076-6879(01)34483-X (cit. on p. 64).

Sim, R. B., A. Reboul, G. J. Arlaud, C. L. Villiers, and M. G. Colomb (Jan. 1979). "Interaction of 125I-labelled complement subcomponents C-1r and C-1s with protease inhibitors in plasma." In: *FEBS letters* 97.1, pp. 111–5 (cit. on p. 84).

Skarzyński, T. (Apr. 1992). "Crystal structure of alpha-dendrotoxin from the green mamba venom and its comparison with the structure of bovine pan-

creatic trypsin inhibitor." In: *Journal of molecular biology* 224.3, pp. 671–83 (cit. on p. 95).

Skubák, P., G. N. Murshudov, and N. S. Pannu (Dec. 2004). "Direct incorporation of experimental phase information in model refinement." In: *Acta Crystallographica, Section D: Biological Crystallography* 60.Pt 12 Pt 1, pp. 2196–201. DOI: 10.1107/S0907444904019079 (cit. on pp. 54, 55, 70).

— (Oct. 2009). "A multivariate likelihood SIRAS function for phasing and model refinement." In: *Acta Crystallographica, Section D: Biological Crystallography* 65.Pt 10, pp. 1051–61. DOI: 10.1107/S0907444909028078 (cit. on pp. 42, 43).

Skubák, P., S. R. Ness, and N. S. Pannu (Dec. 2005). "Extending the resolution and phase-quality limits in automated model building with iterative refinement." In: *Acta Crystallographica, Section D: Biological Crystallography* 61.Pt 12, pp. 1626–35. DOI: 10.1107/S0907444905032233 (cit. on pp. 43, 54).

Skubák, P. and N. S. Pannu (Mar. 2011). "Reduction of density-modification bias by β correction". In: *Acta Crystallographica, Section D: Biological Crystallography* 67.4, pp. 345–354. DOI: 10.1107/S0907444911002083 (cit. on pp. 30, 42).

Skubák, P., W.-J. Waterreus, and N. S. Pannu (June 2010). "Multivariate phase combination improves automated crystallographic model building". In: *Acta Crystallographica, Section D: Biological Crystallography* 66.7, pp. 783–788. DOI: 10.1107/S0907444910014642 (cit. on pp. 38, 42, 51, 54, 70).

Smith, G. D., W. A. Pangborn, and R. H. Blessing (Mar. 2003). "The structure of T6 human insulin at 1.0 A resolution." In: *Acta Crystallographica, Section D: Biological Crystallography* 59.Pt 3, pp. 474–82 (cit. on p. 95).

Sohi, M., A. Alexandrovich, G. F. Moolenaar, R. Visse, N. Goosen, X. Vernede, J. C. Fontecilla-Camps, J. Champness, and M. R. Sanderson (Jan. 2000). "Crystal structure of Escherichia coli UvrB C-terminal domain, and a model for UvrB-uvrC interaction." In: *FEBS letters* 465.2-3, pp. 161–4 (cit. on p. 68).

Solá, R. J. and K. Griebenow (Apr. 2009). "Effects of glycosylation on the stability of protein pharmaceuticals." In: *Journal of pharmaceutical sciences* 98.4, pp. 1223–45. DOI: 10.1002/jps.21504 (cit. on p. 90).

Soria, G., S. E. Polo, and G. Almouzni (2012). "Prime, Repair, Restore: The Active Role of Chromatin in the DNA Damage Response". In: *Molecular cell* 46.6, pp. 722–734 (cit. on p. 5).

Stefl, S., H. Nishi, M. Petukh, A. R. Panchenko, and E. Alexov (July 2013). "Molecular Mechanisms of Disease-Causing Missense Mutations." In: *Journal of molecular biology*. DOI: 10.1016/j.jmb.2013.07.014 (cit. on p. 6).

Strang, C. J., H. S. Auerbach, and F. S. Rosen (1986). "C1s-induced vascular permeability in C2-deficient guinea pigs." In: *The Journal of Immunology* 137.2, pp. 631–635 (cit. on p. 85).

Strang, C. J., S. Cholin, J. Spragg, A. E. Davis, E. E. Schneeberger, V. H. Donaldson, and F. S. Rosen (Nov. 1988). "Angioedema induced by a peptide

derived from complement component C2." In: *The Journal of experimental medicine* 168.5, pp. 1685–98 (cit. on p. 85).

Strub, M. P., F. Hoh, J. F. Sanchez, J. M. Strub, A. Böck, A. Aumelas, and C. Dumas (Nov. 2003). "Selenomethionine and selenocysteine double labeling strategy for crystallographic phasing." In: *Structure (London, England : 1993)* 11.11, pp. 1359–67 (cit. on p. 19).

Sun, S., K. Kondabagil, P. M. Gentz, M. G. Rossmann, and V. B. Rao (Mar. 2007). "The structure of the ATPase that powers DNA packaging into bacteriophage T4 procapsids." In: *Molecular cell* 25.6, pp. 943–9. DOI: 10.1016/j.molcel.2007.02.013 (cit. on p. 95).

Su, S., Y. G. Gao, H. Robinson, Y. C. Liaw, S. P. Edmondson, J. W. Shriver, and A. H.-J. Wang (Oct. 2000). "Crystal structures of the chromosomal proteins Sso7d/Sac7d bound to DNA containing T-G mismatched base-pairs." In: *Journal of molecular biology* 303.3, pp. 395–403. DOI: 10.1006/jmbi.2000.4112 (cit. on p. 65).

Tanaka, T., S. Padavattan, and T. Kumarevel (Mar. 2012). "Crystal structure of archaeal chromatin protein Alba2-double-stranded DNA complex from Aeropyrum pernix K1." In: *The Journal of biological chemistry* 287.13, pp. 10394–402. DOI: 10.1074/jbc.M112.343210 (cit. on p. 66).

Taylor, G. L. (Apr. 2010). "Introduction to phasing." In: *Acta Crystallographica, Section D: Biological Crystallography* 66.Pt 4, pp. 325–38. DOI: 10.1107/S0907444910006694 (cit. on pp. 13, 15).

Teale, M. J., M. Kahsai, S. K. Singh, S. P. Edmondson, R. Gupta, J. W. Shriver, and E. J. Meehan (June 2003). "Cloning, expression, crystallization and preliminary X-ray analysis of the DNA-binding protein Sso10a from Sulfolobus solfataricus". In: *Acta Crystallographica, Section D: Biological Crystallography* 59.7, pp. 1320–1322. DOI: 10.1107/S090744490301062X (cit. on p. 65).

Terwilliger, T. C. (Aug. 2000). "Maximum-likelihood density modification." In: *Acta Crystallographica, Section D: Biological Crystallography* 56.Pt 8, pp. 965–72. DOI: 10.1107/S0907444900005072 (cit. on pp. 32, 33, 38, 52).

— (Dec. 2002). "Automated main-chain model building by template matching and iterative fragment extension". In: *Acta Crystallographica, Section D: Biological Crystallography* 59.1, pp. 38–44. DOI: 10.1107/S0907444902018036 (cit. on pp. 32, 33, 38).

— (Dec. 2004). "Using prime-and-switch phasing to reduce model bias in molecular replacement." In: *Acta Crystallographica, Section D: Biological Crystallography* 60.Pt 12 Pt 1, pp. 2144–9. DOI: 10.1107/S0907444904019535 (cit. on p. 30).

Terwilliger, T. C. and J. Berendzen (Apr. 1999). "Automated MAD and MIR structure solution". In: *Acta Crystallographica, Section D: Biological Crystallography* 55.4, pp. 849–861. DOI: 10.1107/S0907444999000839 (cit. on pp. 24, 30).

Theobald, D. L. and D. S. Wuttke (Sept. 2006). "THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures."

In: *Bioinformatics (Oxford, England)* 22.17, pp. 2171–2. DOI: 10.1093/bioinformatics/btl332 (cit. on p. 71).

Thomassen, E., G. Gielen, M. Schütz, G. Schoehn, J. P. Abrahams, S. Miller, and M. J. van Raaij (Aug. 2003). "The structure of the receptor-binding domain of the bacteriophage T4 short tail fibre reveals a knitted trimeric metal-binding fold." In: *Journal of molecular biology* 331.2, pp. 361–73 (cit. on p. 95).

Usón, I. and G. M. Sheldrick (Oct. 1999). "Advances in direct methods for protein crystallography." In: *Current opinion in structural biology* 9.5, pp. 643–8 (cit. on pp. 13, 26).

Vagin, A. A. and A. Teplyakov (Dec. 1997). "MOLREP : an Automated Program for Molecular Replacement". In: *Journal of applied crystallography* 30.6, pp. 1022–1025. DOI: 10.1107/S0021889897006766 (cit. on p. 74).

Van der Graaf, F., J. A. Koedam, and B. N. Bouma (Jan. 1983). "Inactivation of kallikrein in human plasma." In: *The Journal of clinical investigation* 71.1, pp. 149–58 (cit. on p. 84).

Van der Plas, J. L., R. A. G. de Graaff, and H. Schenk (May 1998). "Karle–Hauptman Matrices and Eigenvalues: a Practical Approach". In: *Acta Crystallographica, Section A: Foundations of Crystallography* 54.3, pp. 267–272. DOI: 10.1107/S0108767397013603 (cit. on p. 26).

Van Veen, H. (Nov. 2012). Personal communication. E-mail correspondence of November 6th 2012: "a concentration of 178 $\mathrm{mg\,mL^{-1}}$ can be reached without loss in activity." (cit. on p. 90).

Van Veen, H. A., J. Koiter, C. J. M. Vogelezang, N. van Wessel, T. van Dam, I. Velterop, K. van Houdt, L. Kupers, D. Horbach, M. Salaheddine, J. H. Nuijens, and M. L. M. Mannesse (Dec. 2012). "Characterization of recombinant human C1 inhibitor secreted in milk of transgenic rabbits." In: *Journal of biotechnology* 162.2-3, pp. 319–26. DOI: 10.1016/j.jbiotec.2012.09.005 (cit. on pp. 88, 89, 92).

Varga, L. and H. Farkas (Mar. 2011). "rhC1INH: a new drug for the treatment of attacks in hereditary angioedema caused by C1-inhibitor deficiency." In: *Expert review of clinical immunology* 7.2, pp. 143–53. DOI: 10.1586/eci.11.5 (cit. on pp. 7, 87).

Walsh, M. A., Z. Otwinowski, A. Perrakis, P. M. Anderson, and A. Joachimiak (May 2000). "Structure of cyanase reveals that a novel dimeric and decameric arrangement of subunits is required for formation of the enzyme active site." In: *Structure (London, England : 1993)* 8.5, pp. 505–14 (cit. on p. 95).

Wang, C. (1985). "Resolution of phase ambiguity in macromolecular crystallography". In: *Diffraction Methods for Biological Macromolecules Part B*. Ed. by S. N. T. Harold W. Wyckoff C. H. W. Hirs. Vol. 115. Methods in Enzymology. Academic Press, pp. 90–112. DOI: http://dx.doi.org/10.1016/0076-6879(85)15009-3 (cit. on pp. 20, 27, 52).

Wang, Z. and J. Moult (Apr. 2001). "SNPs, protein structure, and disease." In: *Human mutation* 17.4, pp. 263–70. DOI: 10.1002/humu.22 (cit. on p. 6).

Wardleworth, B. N., R. J. M. Russell, S. D. Bell, G. L. Taylor, and M. F. White (Sept. 2002). "Structure of Alba: an archaeal chromatin protein modulated by acetylation." In: *The EMBO journal* 21.17, pp. 4654–62 (cit. on p. 66).

Watson, J. D. and F. H. C. Crick (Apr. 1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." In: *Nature* 171.4356, pp. 737–8 (cit. on pp. 3, 4).

Waytes, A. T., F. S. Rosen, and M. M. Frank (June 1996). "Treatment of hereditary angioedema with a vapor-heated C1 inhibitor concentrate." In: *The New England journal of medicine* 334.25, pp. 1630–4. DOI: 10.1056/NEJM199606203342503 (cit. on p. 87).

Weeks, C. M. and R. Miller (Feb. 1999). "Optimizing Shake-and-Bake for proteins." In: *Acta Crystallographica, Section D: Biological Crystallography* 55.Pt 2, pp. 492–500 (cit. on p. 13).

Weiss, M. S. (Apr. 2001). "Global indicators of X-ray data quality". In: *Journal of applied crystallography* 34.2, pp. 130–135. DOI: 10.1107/S0021889800018227 (cit. on p. 95).

Weiss, M. S. and R. Hilgenfeld (Apr. 1997). "On the use of the merging R factor as a quality indicator for X-ray data". In: *Journal of applied crystallography* 30.2, pp. 203–205. DOI: 10.1107/S0021889897003907 (cit. on p. 74).

White, S. A., S. J. Peake, S. McSweeney, G. Leonard, N. P. Cotton, and J. B. Jackson (Jan. 2000). "The high-resolution structure of the NADP(H)-binding component (dIII) of proton-translocating transhydrogenase from human heart mitochondria." In: *Structure (London, England : 1993)* 8.1, pp. 1–12 (cit. on p. 95).

Winn, M. D., C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson (Apr. 2011). "Overview of the CCP4 suite and current developments." In: *Acta Crystallographica, Section D: Biological Crystallography* 67.Pt 4, pp. 235–42. DOI: 10.1107/S0907444910045749 (cit. on pp. 23, 55).

Wisnieski, J. J., T. C. Knauss, I. Yike, D. G. Dearborn, R. L. Narvy, and G. B. Naff (Mar. 1994). "Unique C1 inhibitor dysfunction in a kindred without angioedema. I. A mutant C1 INH that inhibits C1-s but not C1-r." In: *Journal of immunology (Baltimore, Md. : 1950)* 152.6, pp. 3199–209 (cit. on p. 86).

Wuillemin, W. A., M. Minnema, J. C. Meijers, D. Roem, A. J. Eerenberg, J. H. Nuijens, H. ten Cate, and C. E. Hack (Mar. 1995). "Inactivation of factor XIa in human plasma assessed by measuring factor XIa-protease inhibitor complexes: major role for C1-inhibitor." In: *Blood* 85.6, pp. 1517–26 (cit. on p. 84).

Xiao, J., H. Xia, J. Zhou, I. F. Azmi, B. A. Davies, D. J. Katzmann, and Z. Xu (Jan. 2008). "Structural basis of Vta1 function in the multivesicular body sorting pathway." In: *Developmental cell* 14.1, pp. 37–49. DOI: 10.1016/j.devcel.2007.10.013 (cit. on p. 95).

Yue, P., Z. Li, and J. Moult (Oct. 2005). "Loss of protein structure stability as a major causative factor in monogenic disease." In: *Journal of molecular biology* 353.2, pp. 459–73. DOI: 10.1016/j.jmb.2005.08.020 (cit. on p. 6).

Zahedi, R., J. J. Bissler, A. E. Davis, C. Andreadis, and J. J. Wisnieski (Mar. 1995). "Unique C1 inhibitor dysfunction in a kindred without angioedema. II. Identification of an Ala443–>Val substitution and functional analysis of the recombinant mutant protein." In: *The Journal of clinical investigation* 95.3, pp. 1299–305. DOI: 10.1172/JCI117780 (cit. on pp. 86, 87).

Zahedi, R., J. J. Wisnieski, and A. E. Davis (July 1997). "Role of the P2 residue of complement 1 inhibitor (Ala443) in determination of target protease specificity: inhibition of complement and contact system proteases." In: *Journal of immunology (Baltimore, Md. : 1950)* 159.2, pp. 983–8 (cit. on p. 86).

Zhang, K. Y. J., K. D. Cowtan, and P. Main (1997). "Combining constraints for electron-density modification". In: *Macromolecular Crystallography Part B*. Ed. by R. M. S. Charles W. Carter Jr. Vol. 277. Methods in Enzymology. Academic Press. Chap. 4, pp. 53–64. DOI: http://dx.doi.org/10.1016/S0076-6879(97)77006-X (cit. on pp. 27, 28).

Zuraw, B. L., M. Cicardi, R. J. Levy, J. H. Nuijens, A. Relan, S. Visscher, G. Haase, L. Kaufman, and C. E. Hack (Oct. 2010). "Recombinant human C1-inhibitor for the treatment of acute angioedema attacks in patients with hereditary angioedema." In: *The Journal of allergy and clinical immunology* 126.4, 821–827.e14. DOI: 10.1016/j.jaci.2010.07.021 (cit. on pp. 7, 87).

Nederlandse samenvatting

Eiwitten zijn de moleculaire raderen van het leven en spelen een centrale rol in vrijwel alle cellulaire processen. De vergisting van wort tot bier, fotosynthese in planten, de vermeerdering van erfelijk materiaal, de doorgave van stimuli in zenuwcellen en de verdediging tegen ziekteverwekkers zijn slechts enkele van de processen waarbij verschillende eiwitten tezamen een bijdrage leveren in zeer complexe, doch goed georganiseerde systemen.

De rol van een enkel eiwit in zo'n systeem kan vaak begrepen worden vanuit de driedimensionale structuur. Deze kan inzicht geven in het verloop van de chemische reacties en moleculaire interacties waaraan het eiwit zijn functie ontleent. Daarbij kan de globale vouwing van het eiwit van belang zijn, maar ook de plek en reactiviteit van de individuele aminozuren, de bouwstenen waaruit eiwitten bestaan. Verreweg de meerderheid van de eiwitstructuren wordt bepaald middels *röntgendiffractie*. Hierbij wordt een kristal met een zeer smalle monochromatische röntgenbundel vanuit verschillende hoeken bestraald. Uit de resulterende verstrooiingspatronen kan middels fouriertransformatie de elektronendichtheid in het kristal verkregen worden, waaruit vervolgens de molecuulstructuur, ook wel kristalstructuur genoemd, afgeleid kan worden. Behalve fundamenteel inzicht in de werking en de ontstaansgeschiedenis van het leven, kan kennis van de kristalstructuur van een eiwit dat betrokken is bij een ziekteproces bijdragen tot ontwikkeling van nieuwe medicijnen. Kennis en gerichte aanpassing van de structuur van eiwitten die gebruikt worden in industriële toepassingen

kunnen kostenbesparingen en nieuwe producten opleveren.

De organisatie van chromatine is essentieel in alle organismen. Niet alleen vanwege haar rol in het efficiënt verpakken van het DNA zodat het in de cel of celkern past, maar ook door de betrokkenheid van chromatineorganisatie bij allerlei cellulaire processen. Door bijvoorbeeld lokaal het DNA toegankelijker of juist meer compact te maken kan de expressie van een gen beïnvloed worden. Chromatine-eiwitten kunnen het DNA samentrekken door het DNA te buigen of door een DNA streng op twee verschillende plekken te binden en zo lussen te vormen. Andere chromatine-eiwitten maken het DNA meer compact door als een spoel met DNA omwikkeld te worden. Hoofdstuk 4 beschrijft de structuur van het Sso10a2 eiwit uit de Archaeon *Sulfolobus solfataricus*. Tezamen met de homologe eiwitten Sso10a en Sso10a3 speelt Sso10a2 mogelijk een rol in chromatinedynamiek en -organisatie in *S. solfataricus*. Sso10a2 vormt een uitermate stabiele dimeer waarbij de lange C-terminale α-helix een belangrijke rol speelt. Daarnaast bevat de Sso10a2 monomeer een DNA-bindend domein dat ook in andere chromatine-eiwitten is geïdentificeerd. Ten opzichte van de Sso10a structuur zijn een aantal van de secundaire structuurelementen in het DNA-bindende domein van de Sso10a2 monomeer verschoven, wat mogelijk van invloed is op de binding van DNA.

Ruconest® is een recombinant C1INH eiwit dat recentelijk is goedgekeurd voor behandeling van patiënten met hereditair angio-oedeem. C1INH behoort tot de familie van serpineproteaseremmers en reguleert ontstekingsprocessen in het bloed door het voorkomen van spontane activatie van het complementsysteem. Genetische C1INH deficiëntie veroorzaakt hereditair angio-oedeem. Deze ziekte gaat gepaard met zwellingen die levensbedreigend kunnen zijn wanneer ze voorkomen in de bovenste luchtwegen. Hoofdstuk 5 beschrijft de succesvolle kristallisatie van twee varianten van het recombinante C1INH eiwit. Van beiden is aangetoond dat ze hun remmende activiteit *in vitro* behouden. In tegenstelling tot de eerder opgeloste *latente* structuur van C1INH zijn beide varianten gekristalliseerd met de glycosilering intact. Mogelijkerwijs verschaft de structuur meer inzicht in de *actief–latent* transitie welke voor deze ser-

pine nog niet volledig is opgehelderd.

Om de kristalstructuur middels fouriertransformatie uit de diffractiepatronen te berekenen zijn zowel de intensiteit van de verstrooide röntgenbundels als hun onderlinge fases nodig. De intensiteit is gemakkelijk te meten met behulp van een röntgendiffractometer, echter de faserelatie kan niet direct experimenteel bepaald worden. Dit probleem is in de kristallografie bekend als het *faseprobleem*. Hoofdstuk 1 geeft een overzicht van de verschillende methoden die in de loop der jaren ontwikkeld zijn om het faseprobleem op te lossen. Hierbij zijn grofweg twee verschillende benaderingen te onderscheiden. Een techniek die de afgelopen jaren erg populair is geworden, maakt gebruik van een gelijkende reeds bekende eiwitstructuur om een eerste schatting van de fases van de onbekende structuur te verkrijgen. De tweede groep methoden maakt gebruik van informatie uit het diffractie-experiment om een deel van de molecuulstructuur en vervolgens de fase-informatie voor het gehele molecuul te achterhalen. Het verkrijgen van de fase-informatie op deze manier wordt ook wel experimentele fasebepaling genoemd.

Om de structuurbepaling te vereenvoudigen en sneller te maken zijn veel van de taken in de structuuranalyse geautomatiseerd. Het programma Crank koppelt verschillende programma's die elk afzonderlijk een enkele taak uitvoeren in de structuuroplossing middels experimentele fasebepaling. Hierdoor is het mogelijk voor een eiwit van gemiddelde grootte met verstrooiingspatronen van goede kwaliteit binnen enkele uren een vrijwel volledig model te genereren. In hoofdstuk 2 komen de vele verbeteringen in Crank aan bod die bij tests met meer dan honderd verschillende datasets tot structuren leidden die aanmerkelijk meer volledig zijn dan bij een vorige versie van Crank. Daarnaast zijn enkele structuren opgelost met de nieuwe versie die voorheen nog niet automatisch opgelost konden worden.

Veel van de verbeteringen in Crank hebben hun oorsprong in nieuwe geavanceerde waarschijnlijkheidsfuncties waarmee meerdere bronnen van (experimentele) informatie tegelijkertijd beschouwd kunnen worden en welke de correlaties tussen verschillende bronnen van informatie

in acht nemen. Daarnaast kan ook de experimentele foutmarge beter worden gemodelleerd met dit type functies. Hoofdstuk 3 beschrijft de toepassing van de nieuwe *SAD-DM* waarschijnlijkheidsfunctie die faseinformatie na experimentele fasebepaling combineert met fases verkregen vanuit empirisch verbeterde elektronendichtheid.

Vaak is de elektronendichtheid in eerste instantie van onvoldoende kwaliteit om een model van de eiwitstructuur in te bouwen. Door gebruik te maken van kennis over de typische verdeling van de elektronendichtheid in een kristal kan deze verbeterd worden. Zo varieert de elektronendichtheid bijvoorbeeld sterk op plaatsen waar eiwit zit en is deze uniform in de tussenruimte tussen verschillende eiwitmoleculen in het kristal. De verbeterde fases worden verkregen door inverse fouriertransformatie van de verbeterde elektronendichtheid en gecombineerd met de oorspronkelijke experimenteel bepaalde fases. Dit proces wordt herhaald totdat de elektronendichtheid niet langer noemenswaardig verbetert. De nieuwe *SAD-DM* functie voor fasecombinatie is een aanmerkelijke vooruitgang in vergelijking met het meest gebruikte algoritme σ_A. Dit komt onder andere doordat de *SAD-DM* functie in tegenstelling tot σ_A er bij de fasecombinatie niet vanuit gaat dat de oorspronkelijke en verbeterde fases onafhankelijk van elkaar zijn.

Het verkrijgen van een goed verstrooiend eiwitkristal is een proces dat maanden tot jaren in beslag kan nemen. Echter een goed verstrooiend kristal is geen garantie dat de structuur opgelost kan worden. In veel gevallen bevatten de diffractiepatronen hiertoe niet genoeg informatie. Dankzij de geavanceerde waarschijnlijkheidsfuncties zoals beschreven in hoofdstuk 2 en 3 kan de beschikbare data beter gebruikt worden. Zo kunnen structuren opgelost worden die met bestaande software niet of moeilijk op te lossen zijn.

Curriculum vitae

Education and experience

2013 – present	Scientist at PRISNA bv, Leiden *Development of methods for isolation and analysis of phytochemicals*
2008 – 2013	PhD research at the Department of Biophysical Structural Chemistry, Leiden University, Leiden *Structural characterization of several proteins and development of software in the area of X-ray protein crystallography*
2005 – 2008	MSc in Functional Genomics, *cum laude* from Delft University of Technology, Delft *Majored in bioinformatics, structural biology and molecular genetics*
2004 – 2006	Teaching Assistant at the Leiden Institute of Chemistry, Leiden University, Leiden *Supervision of labs in recombinant DNA technology, biochemistry and molecular cell biology*
2000 – 2005	BSc in Life Science & Technology from Delft University of Technology, Delft *Fundamental and applied subjects from several disciplines within the field of life sciences*

Presentations

October 2012	Joint Poster presentation with R.P.C Driessen at the 3rd Leiden Institute of Chemistry Reedijk Symposium, Leiden, The Netherlands titled *"Differences in binding modes of crenarchaeal Sso10a chromatin proteins"*
August 2011	Poster presentation at the IUCr Crystallographic Computing School in Oviedo, Spain and the 22nd Congress of the International Union of Crystallography in Madrid, Spain titled *"Advances in the* CRANK *software suite for automated crystal structure solution"*
October 2010	Poster presentation at the 1st Leiden Institute of Chemistry Reedijk Symposium, Leiden, The Netherlands titled *"*MULTICOMB *Providing better electron density maps for improved automatic crystal tracing"*
August 2009	Poster presentation at the EMBL workshop Diffraction Data Collection Using Synchrotron Radiation in Berlin, Germany and the 25th European Crystallographic Meeting in Istanbul, Turkey titled *"Recent advances in* CRANK*"*
July 2009	Oral presentation at the EMBO Practical Course on High Throughput Protein Production and Crystallization in Oxford, United Kingdom titled *"Construct design and preliminary expression analysis of Apoptin interacting protein 3"*
March 2009	Poster presentation at the annual meeting for crystallographers in Lunteren, The Netherlands titled *"*CRANK*, more than the sum of its parts"*

Publications

Liu, Z., F. Galli, W.-J. Waterreus, E. M. Meulenbroek, R. I. Koning, G. E. M. Lamers, R. C. L. Olsthoorn, N. S. Pannu, T. H. Oosterkamp, A. J. Koster, R. T. Dame, and J. P. Abrahams (Apr. 2012). "Single-walled carbon nanotubes as scaffolds to concentrate DNA for the study of DNA-protein interactions." In: *Chemphyschem : a European journal of chemical physics and physical chemistry* 13.6, pp. 1569–75. DOI: 10.1002/cphc.201100896.

Pannu, N. S., W.-J. Waterreus, P. Skubák, I. Sikharulidze, J. P. Abrahams, and R. A. G. de Graaff (Apr. 2011). "Recent advances in the CRANK software suite for experimental phasing." In: *Acta Crystallographica, Section D: Biological Crystallography* 67.Pt 4, pp. 331–7. DOI: 10.1107/S0907444910052224 (cit. on p. 37).

Ruhaak, L. R., C. Huhn, W.-J. Waterreus, A. R. de Boer, C. Neusüss, C. H. Hokke, A. M. Deelder, and M. Wuhrer (Aug. 2008). "Hydrophilic interaction chromatography-based high-throughput sample preparation method for N-glycan analysis from total human plasma glycoproteins." In: *Analytical chemistry* 80.15, pp. 6119–26. DOI: 10.1021/ac800630x.

Skubák, P., W.-J. Waterreus, and N. S. Pannu (June 2010). "Multivariate phase combination improves automated crystallographic model building". In: *Acta Crystallographica, Section D: Biological Crystallography* 66.7, pp. 783–788. DOI: 10.1107/S0907444910014642 (cit. on pp. 38, 42, 51, 54, 70).