Crystallographic Analysis of Pathological Crystals, Periplasmic Domain of Ligand-free CitA Sensor Kinase and PDI-related Chaperones

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August Universität zu Göttingen

vorgelegt von
Madhumati Sevvana
aus Visakhapatnam, India

Göttingen, 2006

D7

Referent: Prof. George M. Sheldrick
Koreferentin: Prof. Dr. Isabel Usón

Tag der mündlichen Prüfung: 4th July 2006

Acknowledgements

Many people helped me during the course of my work as friends, teachers and colleagues. First of all, I am grateful to Prof. George M. Sheldrick for his enthusiastic supervision and unfailing support during my thesis work. He is the best teacher and advisor I could ever wish for and is a great source of knowledge and inspiration for his students and colleagues. I would especially like to thank him for the great lab environment and for the right amount of guidance to explore and pursue my research goals.

I am indebted to Prof. Dr. Isabel Usón for accepting to be co-referent of my thesis. I would like to particularly acknowledge her for teaching me the basics of MR and her advice during structure solution of ERp29. My greatest gratitude to Dr. Regine Herbst-Irmer, from whom I learned dealing with pathological cases like twinning and disorder in small molecule crystallography. I would like to thank her for valuable discussions and suggestions regarding twinned Glucose Isomerase, Insulin and CitA structures. I am very thankful to Dr. David M. Ferrari for a wonderful collaboration during structural and functional analysis of Wind and ERp29 proteins. I would like to thank him for providing me with the protein samples and helping me in purifying Wind mutants. I would specially like to thank Dr. Stefan Becker for his undwindling support in CitA project.

I would like to thank Dr. Tim Grüne, Dr. Stefan Becker, Prof. Dr. Isabel Usón, Dr. Regine Herbst-Irmer, Dr. Stephan Rühl and Kathrin Meindl for their continuous support during the course of my thesis work. I am very thankful to Helmut Dehnhardt for all his technical assistance and keeping the diffractometers in a working state. I thank Christine Schlicker, Marianna biadene, Roland Pfoh, Burkhard heisen, Dr. Ina Dix and Aritra Pal for a wonderful environment in the lab. I thank Dr. Gábor Bunkóczi, Dr. Judit Debreczeni, Dr. Jose Antonio Cuesta Seijo, Dr. Qingjun Ma, Dr. Chaoshe Guo and Dr. Kathrin Barnewitz for their guidance and timely help. I would like to thank Prof. Dr. N. Siva kumar for his valuable help on countless occasions.

My warm thanks to my husband, Kiran Ayodhya for his support, encouragement and his constant interest in my work. I am very grateful to my father Dr. Veera Narayana Rao Sevvana, my mother Dr. Jayalakshmi Savaram, my sister Swati Sevvana and my brother Dr. Mohan Rao Sevvana for their endless encouragement, support and personal strength through the years. The strength of my family has been a tremendous boost to me.

karmaṇy evādhikāras te

mā phaleṣu kadācana

mā karma-phala-hetur bhūr

mā te saṅgo 'stv akarmaṇi

- Bhagavad Gitā -

Abbreviations

Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
BI	Bovine Insulin
Cys	Cysteine
GI	Glucose Isomerase
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
His	Histidine
HPKs	Histidine protein kinases
Ile	Isoleucine
Leu	Leucine
Lys	Lysine
Met	Methionine
NMT	Non-merohedral twin
PEG	Polethylene glycol
Phe	Phenylalanine
Pro	Proline
Ser	Serine
TCA	Tricarboxylic acid
TCSs	Two component systems
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
Val	Valine

Molecular Graphics

All three dimensional figures were drawn with *CHIMERA* [92], *MOLSCRIPT* [64], *BOBSCRIPT* [26] and rendered with *RASTER3D* [81] or drawn and rendered in *PYMOL* [19].

Contents

1 Methods in Crystallography **11**

1.1 Crystallization methods 11

1.2 X-ray data collection and processing 13

1.2.1 Cryocrystallography 13

1.2.2 Post-crystallization improvement of crystal quality 14

1.2.3 Data collection strategies 15

1.2.4 Data processing 16

1.2.5 Data quality indicators 19

1.3 Macromolecular Phasing methods 20

1.3.1 Molecular replacement 20

1.3.2 Experimental phasing 23

1.4 Refinement 25

2 Analysis of Non-Merohedral Twins **28**

2.1 Introduction 28

2.2 Aim of present work 30

2.3 Methods 32

2.3.1 Data Collection 32

2.3.2 Twin indexing 32

2.3.3 Multicomponent data integration 33

2.3.4 Multicomponent data scaling 34

2.3.5 Phasing and model building 35

2.3.6 Twin Refinement and Validation 35

2.4 Results 36

2.4.1 Twinned Bovine Insulin 36

2.4.2 Twinned Glucose Isomerase 42

2.5 Discussion and Future Perspectives 45

3 Sensor Kinase CitA 48

3.1 Introduction 48
3.1.1 CitA sensor kinase from *Klebsiella pneumoniae* 50
3.1.2 Structure of ligand-bound periplasmic domain of CitA 54
3.2 Aim of Present Work 55
3.3 Materials and Methods 56
3.3.1 Crystallization & data collection 56
3.3.2 Data processing and detection of twinning 56
3.3.3 Structure solution, bias minimization and model building 62
3.3.4 Refinement and structure validation 64
3.3.5 Inverse fouriers using *SHELXE* 66
3.4 Results 67
3.4.1 Structure of the ligand-free periplasmic domain of CitA 67
3.4.2 Biologically relevant tetramer 69
3.4.3 Citrate-bound and citrate-free forms 72
3.5 Discussion and future perspectives 73

4 PDI-related Chaperones 77

4.1 Introduction 77
4.1.1 Windbeutel and dorso-ventral patterning in *Drosophila* 78
4.1.2 Crystal structure of wild-type His-Wind 79
4.1.3 Chaperone function of PDIs 81
4.1.4 Close relative: human ERp29 82
4.2 Aim of Present Work 83
4.3 Materials and Methods 84
4.3.1 Expression, purification & crystallization 84
4.3.2 Data Collection and Processing 86
4.3.3 Structure Solution 86
4.3.4 Model Bias Minimization, Refinement & Validation 89
4.3.5 Hydrophobic Molecular Interaction Field 91
4.4 Results and Discussion 91
4.4.1 Structure of Wind-His and its mutants 91
4.4.2 Superposition of His-Wind and its mutants on Wind-His 92

4.4.3 The dimer interface 93
4.4.4 Opening of the CTGC loop in Y53F 93
4.4.5 Substrate binding site and hydrophobic MIF 96
4.4.6 Putative substrate binding site 98
4.4.7 Dimerization in Wind and human ERp29 99
4.5 Future perspectives 101

Abstract

The current work reports on developing methods to elucidate crystal structure of non-merohedrally twinned protein crystals and analysis of structures of different proteins that mediate signal transduction.

Non-merohedral twinning is one among the several pathological cases encountered rarely during crystal structure analysis. Processing X-ray data from non-merohedrally twinned small molecule crystals has been a routine method for several years. The existing small-molecular methods were successfully extended to solve two twinned test protein structures, namely Bovine Insulin (BI-Zwilling) and Glucose Isomerase (GI-Drilling), using Single Wavelength Anomalous Scattering (SAD) with in-house data.

Bacterial response to external stimuli is mediated by several signal transduction systems, which are vital for their survival. These systems are termed two component systems or histidine protein kinases in prokaryotes. Two component systems offer themselves as new drug targets, because they have not been identified in mammalian systems or/and animal kingdom. In the present study the crystal structure of merohedrally twinned ligand-free periplasmic domain of the citrate sensor, CitA from *Klebsiella pneumoniae* was determined. *Klebsiella pneumoniae* is the major cause of pneumonia, nosocomial infections, sepsis and urinary tract infections. Comparison of structures of the ligand-free and ligand-bound CitA domains, shed light on some mechanistic aspects of the citrate sensor signal transduction system.

Wind, a product of *windbeutel* gene and a member of the PDI-related chaperone family, is one of the several genes required for dorso-ventral polarization in the developing embryo of *Drosophila melanogaster*. Dorso-ventral polarization requires communication between somatic cell derived follicle cells and germ-line derived oocyte, which occurs through a cascade of signal transduction steps called the Gurken-EGFR pathway. A previously determined crystal structure of Wind, suggested a homodimer with the dimerization interface along the N-terminal b-domain, which has the characteristic thioredoxin fold. Pipe, a product of *pipe*, was identified to be the potential interacting partner of the Wind dimer and a putative substrate binding site on Wind was characterized. Crystal structures of several non-functional mutants of Wind and a Wind-peptide complex indicated Wind-Pipe interaction to be, mainly due to the aromatic and/or hydrophobic behaviour of the substrate binding site and conservation of the dimerization interface. Crystal structure of human ERp29, a close relative of Wind, revealed a similar three dimensional architecture and conservation of the dimerization interface.

Chapter 1

Methods in Crystallography

1.1 Crystallization methods

Macromolecular crystallization is a multiparametric process [21]. Intrinsic physicochemical parameters that affect crystallization are temperature, pH, time, ionic strength and purity of chemicals, density and viscosity effects, pressure, electric and magnetic fields *etc*. There are several methods to crystallize biological macromolecules, all of which aim at bringing the solution of macromolecules to a supersaturation state. The most important parameter affecting crystallization is the purity of the sample. A sample purity of more than 95% is desired to obtain best results. Crystallization by vapour diffusion and dialysis are the most common methods used, although batch crystallization and interface diffusion methods are not uncommon.

Establishing a solubility diagram (phase diagram) is one way of quantifying the influence of crystallization parameters on the solubility of macromolecules. The determination of solubility of a macromolecule depends on various parameters, a two-dimensional solubility diagram is a representation of its solubility as a function of one parameter, all other parameters being constant. The diagram represented in Figure 1.1a, comprises the following zones:

- The solubility curve S divides the undersaturated and supersaturated zones. In an experiment where crystallizing agent and biological macromolecule concentration correspond to conditions of curve S, the saturated protein solution is in equilibrium with the crystallized protein. This corresponds to the situation at the end of the process of crystal growth from a supersaturated solution, or of dissolution of crystals in an undersaturated solution.

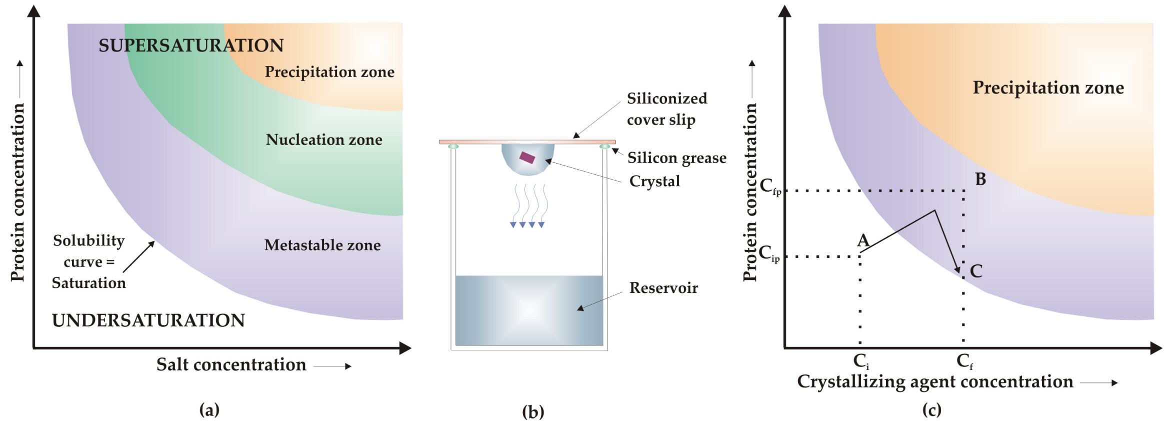


Figure 1.1: (a) Schematic representation of a two dimensional solubility diagram, illustarted by the change of protein concentration with respect to crystallizing agent concentration. (b) Crystallization by hanging drop vapour diffusion. (c) Schematic solubility diagram and correlation between macromolecule and precipitating agent concentrations in crystallizing experiments using vapour diffusion setup with crystallization. C_i and C_{ip} are the initial crystallizing agent concentration and protein concentration respectively and C_f and C_{fp} are the final crystallizing agent and protein concentration respectively.

- Below the solubility curve (S) the solution is undersaturated and the biological macromolecule will never crystallize.
- Above the solubility curve (S), the concentration of the biological macromolecule is higher than the concentration at equilibrium for a given salt concentration. This corresponds to the supersaturation zone. Depending upon the kinetics to reach equilibrium and the level of supersaturation, this region may itself be subdivided into three zones. (1) The precipitation zone, where excess protein immedietly separates from the solution in an amorphous state. (2) The nucleation zone, where excess of biological macromolecule separates under a crystalline form. (3) In the metastable zone: a supersaturated solution may not nucleate for a long period of time, unless the solution is mechanically shocked or a seed crystal introduced. This zone corresponds ideally to the growth of crystals without nucleation of new crystals.

The principle of vapour diffusion crystallization is indicated in Figure 1.1b. A droplet containing the biological macromolecule to crystallize with buffer, crystallizing agent, and additives, is equilibrated against a reservoir containing a solution of crystallizing agent at a higher concentration than the droplet. Equilibration proceeds by diffusion of the volatile species until vapour pressure in the droplet equals the one of the reservoir. If equilibration occurs by water exchange (from the drop to the reservoir), it leads to a droplet volume change. Consequently, the concentration of all constituents in the drop will change. For drop with a higher vapour pressure than water, the exchange occurs from drop to reservoir. The same principle applies to hanging drop, sitting drop and sandwich drops. In a classical case where the concentration of crystallizing agent in the reservoir is twice the one in the drop, the protein will start to concentrate from an undersaturated state A (at concentration C_i) to reach a supersaturated state B (at concentration C_f) with both protein and crystallizing concentrations increasing by a factor of two. A hypothetical case leading to crystals is describeed in the phase diagram shown in Figure 1.1c.

In batch method of crystallization the macromolecule to be crystallized is mixed with crystallizing agent at a similar concentration to that of biological macromolecule such that supersaturation is instantaneously reached. Because one starts from supersaturation, nucleation tends to be too large. However in some cases fairly large crystals can be obtained when working close to the metastable region. Samples are dispensed and incubated under oil, thus preventing evaporation and uncontrolled concentration changes of the components in the micro-droplets.

The key to the iterative process of crystallization is to observe the results of the crystallization trials, and to correlate the results back to the cocktail of chemicals that was used in that experiment. One would need to predict from these observations where a crystallization optimum would be reached. Crystallization is a time dependent event, so each trial must be observed many times, over a long time span.

1.2 X-ray data collection and processing

1.2.1 Cryocrystallography

Cryogenic techniques are routinely used in many laboratories for data collection on in-house facilities as well as on synchrotron sources. Collecting data at low temperatures greatly reduces X-ray induced radiation damage to macromolecular crystals [37]. Data collection at low temperatures can be achieved either by using Nitrogen gas at 100 K or Helium at 30 K [38]. Macromolecular crystals are more susceptible to radiation damage when compared to small molecules because of large amounts of solvent molecules and the potential for diffusion within the sample. Radiation damage to macromolecular crystals has increased drastically with the present day synchrotron sources. Cryogenic techniques reduce the diffusion of atoms within the sample. Atomic motion is reduced at cryogenic temperatures and this, depending upon the relative degree of dynamic and static disorder in crystals of a particular macromolecule, can render higher-resolution data. Background X-ray scattering is also reduced, when compared to data collected from crystals mounted in capillaries.

1.2.2 Post-crystallization improvement of crystal quality

In the past when one could obtain crystals but these crystals did not diffract to a reasonable resolution limit, several strategies to enhance crystal quality were used. For example strategies like searching for new crystallization conditions, crystallizing the protein of interest from a different organism, crystallizing a different form of the protein by using proteases that produce smaller fragments, generating new constructs encoding a truncated form of the protein or mutating surface amino acids to enhance protein crystallization were used. All these methods cannot be used on crystals that have already been grown.

However before residing back, methods like postcrystallization soaking, cross linking, crystal annealing and controlled dehydration have been reported to dramatically improve diffraction resolution of protein crystals [47]. Crystal annealing is a rapid method to reduce flash cooling induced disorder that can increase diffraction quality of protein crystals. The method involves warming the flash cooled crystal to room temperature and flash cooling it again [45]. Three different protocols have been reported. Macromolecular crystal annealing (MCA) involves removing the crystal from the cold liquid nitrogen stream and soaking the crystal back in the cryoprotectant solution. After a few minutes of incubation, the crystal is recooled in the cryostream [45]. The flash annealing (FA) method involves blocking the cold-stream for 3s, thrice with intervals of 6s [132]. Annealing on the loop (AL), involves blocking the nitrogen stream, but in this case the length of time varies from crystal to crystal [44]. AL uses a single warming step.

Reduction of solvent content in protein crystals can produce more closely packed and better ordered crystals, extending the resolution of X-ray diffraction patterns [36]. Crystal dehydration emerges as the post-crystallization treatment that has produced the most remarkable improvements in the diffraction resolution of macromolecular crystals.

Several methods for crystal dehydration are described in Heras & Martin, 2005 [47]. One such example is crystal dehydration by transfering the crystal into a dehydrating solution, which is usually the mother liquor either with a higher concentration of precipitant or supplemented with cryoprotective solutions such as PEG400, PEG600, glycerol or MPD. One of the mechanisms of cryoprotection is to reduce protein solvation, so that cryoprotectants can act as dehydrants and *vice versa*. Post-crystallization soaking in higher ionic strength solutions, cryoprotectants or heavy-atom-containing solutions sometimes results in improvement in the crystal quality without dehydration. Chemical cross-linking using glutaraldehyde increases the robustness of the crystal against mechanical stress, reduces its solubility and can also improve crystal quality [75]. Cross-linking involves the reaction of lysine amines with aldehydes of the cross-linking molecule, thereby the improvement of diffraction quality depends on the number and position of lysine residues in the asymmetric unit.

1.2.3 Data collection strategies

Data collection is the final experimental step in the determination of a three dimensional structure using X-ray crystallography and is an important scientific process rather than a mere technicality [17]. Every care should be taken to collect the data with highest quality and least possible errors. Things that could go wrong are the instability of the incident beam intensity and wavelength (this happens mostly at synchrotron sources), synchronization of shutter opening/closing with the goniometer rotation and detector calibration, both for spatial distortion and for non-uniformity of response.

Collecting optimum X-ray diffraction data involves a number of choices and compromises, including choice of crystal, source, rotation range, exposure time and programs for integrating and scaling [27]. A strongly diffracting crystal would not need a powerful beam. A good combination would be collecting high resolution native data at the synchrotron and derivative data in-house. Low beam divergence is prefered but does not help much in a highly mosaic crystal. The beam size should not be significantly bigger than the crystal. Whereas, if the beam size is much smaller than the crystal, the relative intensity of a reflection is greatly affected. Lower mosaicity is always prefered than highly mosaic crystals. Low background improves the signal to noise ratio. This means that using as little cryoprotectant as possible improves the data quality a lot. Larger crystals might diffract farther, whereas smaller crystals would freeze properly. Data collected with very high redundancy to the maximum possible resolution and as complete data as possible is an additional advantage.

The rotation width per image should be set to resolve the longest cell axis, taking into account the reflection width [17]. Those data sets consisting of a relatively large rotation-angle increments, where each image contains a number of fully recorded reflections are termed thick sliced data sets [94]. Those with relatively small rotation angle increments, where each image contains predominantly partially recorded reflections are termed thin sliced data sets. A data set of thick sliced images usually has more fully recorded reflections, fewer partially recorded reflections, more spatial overlaps, higher X-ray background, more saturated pixels and a lower total number of images. In general a two dimensional integration of reflections *(i.e.*, reflections on each image are integrated separately and then the integrated contributions from individual intensities are summed up) is carried out using the programs *MOSFLM* [70] or *HKL2000* [88]. A thin sliced data set has no fully recorded reflections, fewer spatial overlaps, lower X-ray background, fewer saturated pixels and would cost more measurement time. A three-dimensional reflection integration is carried out in programs such as *XDS* [56] or *SAINT* (Bruker AXS, Madison, WI). The exposure time per image should be long enough to give reasonable statistics at highest resolution.

Most autoindexing programs offer a strategy option taking into consideration the cell, space group, mosaicity and exposure time into account, inorder to calculate the best possible strategy to obtain as complete data in as short time as possible.

1.2.4 Data processing

Autoindexing

The position **X** (x, y, z) of a reciprocal-lattice point can be given as

$$X = [\Phi][A]\mathbf{h} \tag{1.1}$$

The matrix [Φ] is a rotation matrix around the camera's spindle axis for a rotation of φ. The vector **h** represents the Miller indices (h, k, l) and [A] defines the reciprocal unit-cell dimensions and the orientation of the crystal lattice with respect to the camera axes when φ=0. Thus,

$$[A] = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix} \tag{1.2}$$

where a^*_x, a^*_y and a^*_z are the components of the crystal a^* axis with respect to the orthogonal camera axis. When an oscillation image is recorded, the position of a reciprocal-lattice point is moved from point x_1 to x_2, corresponding to a rotaion of the crystal from φ_1 to φ_2. The recorded position of the reflection on the detector corresponds to the point x when it intersects with a point on the Ewald sphere somewhere between x_1 and x_2. But the actual value of φ cannot be determined and lies somewhere in between both the values.

A reflection recorded at position (X, Y) on a flat detector normal to the X-ray beam, at a distance D from the crystal, corresponds to

$$x = \frac{X}{\lambda(X^2 + Y^2 + D^2)^{\frac{1}{2}}} \tag{1.3}$$

$$y = \frac{Y}{\lambda(X^2 + Y^2 + D^2)^{\frac{1}{2}}} \tag{1.4}$$

$$z = \frac{D}{\lambda(X^2 + Y^2 + D^2)^{\frac{1}{2}}} \tag{1.5}$$

where λ is the X-ray wavelength.

If an approximate [A] matrix is available, the Miller indices of an observed peak at (X, Y) can be roughly determined using 1.5 and 1.1, where [104]

$$\mathbf{h} = [A]^{-1}[\Phi]^{-1}X \tag{1.6}$$

Integration

Data integration refers to the process of obtaining estimates of diffracted intensities and their standard deviations from the raw images recorded by an X-ray detector. There are two distinct procedures to determine the integrated intensities: summation integration and profile fitting. Summation integration involves simply adding pixel values for all pixels lying within the area of a spot, and then subtracting the estimated background contribution to the same pixels. Profile fitting assumes that the actual spot shape or profile is known in two or three dimensions and the intensity is derived by finding the scale factor that, when applied to the known profile, gives the best fit to the observed spot profile. Profile fitting requires two separate steps: the determination of standard profiles and the evaluation of profile-fitting intensities.

Scaling

The measured diffraction intensities are not all on the same scale because they are affected by a number of physical factors from the experiment *viz.* factors that are dependent on the primary beam and the crystal rotation axis, direction of the diffracted beam and detector, most of which are difficult to measure directly. The process of data reduction uses the redundancy of multiple measurements of symmetry related reflections to put all observations on a common scale by fitting a scaling model which reflects the experiment [28].

Factors related to the incident X-ray beam are variations in the rotation rate, rapid fluctuations in incident beam intensity or errors in synchronization of the shutter cause systematic errors that are impossile or difficult to model. Correctable factors are slow variation in incident beam intensity, change in illuminated volume, if the beam is smaller than the crystal and absorption in the primary beam. The most difficult systematic error related to diffracted beam and the crystal is radiation damage. The relative B-factor is a correction for average radiation damage. Zero-dose extrapolation works, when there are many observations for each reflection well spaced out in time. The detector should be properly calibrated for spatial distortion and senstivity of response as well as for any defective regions and should be stable. The detector corrections are usually difficult to be extracted from the diffraction data and the integration program has to be told about shadows of the beam stop, shadow of the cryostream etc.

There are several scaling models to model the correction as a function of rotation and the direction of the diffracted beam. The simplest model applies a different scale factor for each image, but the scale does not usually vary sharply from one image to another, so a smooth function is more appropriate. The other traditional component of the scaling model is a relative B factor, $\exp(-2B\sin^2\theta/\lambda^2)$, where B is a function of time. This provides a resolution dependent radiation-damage correction. Absorption in the secondary beam direction is best parameterized as coefficients of real spherical harmonics, either in the rotating crystal or in the diffractometer frame.

1.2.5 Data quality indicators

The quality or the information content of a crystal structure is highly dependent on the quality or the information content of the underlying diffraction data [127]. When averaged intensities/amplitudes are available, the quality of the data can be assessed by looking at the resolution, completeness, mean I/σI (signal to noise ratio) and Wilson plot appearance. The resolution of the data set is defined as the minimum distance at which two features in the corresponding electron density map can be resolved. Missing data or low data completeness leads to a deterioration of the model parameters.

Most important information is obtained from the merging statistics prior to merging equivalent reflections. The first criterion is the redundancy (N) of the data. Since X-ray data measurements are based on counting statistics, the average measurements should become more accurate as more and more measurements are made. A highly redundant data set will therefore be of higher quality than a data set where every reflection has been measured only once.

The most frequently reported descriptor of data quality is the conventional merging R factor R_{merge}. But R_{merge} is intrinsically dependent on the redundancy of the data and is calculated as

$$R_{merge} = \frac{\sum_{hkl}\sum_i |I_i(hkl) - \overline{I(hkl)}|}{\sum_{hkl}\sum_i I_i(hkl)} \tag{1.7}$$

Lower redundancy will always lead to a lower R_{merge}, but at the same time give less accurate data. This dependence of R_{merge} on the redundancy can be remedied by the introduction of redundancy independent merging R factor $R_{r.i.m}$, which is calculated as

$$R_{r.i.m} = \frac{\sum_{hkl}[N/(N-1)]^{1/2}\sum_i |I_i(hkl) - \overline{I(hkl)}|}{\sum_{hkl}\sum_i I_i(hkl)} \tag{1.8}$$

$R_{r.i.m}$ takes into account how often a given reflection has been measured, and could therefore be used as a substitute for the conventional R_{merge}.

Another R factor that can be calculated is the so called precision indicating R factor $R_{p.i.m}$, which describes the precision of the averaged measurement. Since in the course of structure determination and refinement, averaged intensities or amplitudes are normally used, this R factor should deliver the most information when it comes to predicting the performance of a given data set in the structure determination. $R_{p.i.m}$ is calculated as

$$R_{p.i.m} = \frac{\sum_{hkl}[1/(N-1)]^{1/2}\sum_i |I_i(hkl) - \overline{I(hkl)}|}{\sum_{hkl}\sum_i I_i(hkl)} \tag{1.9}$$

1.3 Macromolecular Phasing methods

The electron density in a crystal can be obtained by calculating the Fourier summation,

$$\rho(xyz) = \frac{1}{V}\sum_{hkl} |F(hkl)| exp[-2\pi i(hx + ky + lz) + i\alpha(hkl)] \tag{1.10}$$

in which F|(*hkl*)| is the structure factor amplitude of reflection (*hkl*), including the temperature factor, and α(*hkl*) is the phase angle. *x*, *y* and *z* are coordinates in the unit cell. From the diffraction pattern the values of I(*hkl*) are obtained after applying the correction factors like Lorentz polarization correction etc. Because I(*hkl*) = |F(*hkl*)|^2 the amplitudes |F(*hkl*)| can be found. Unfortunately no information is available on the phase angles. In principle four techniques exist for solving the phase problem in Protein X-ray crystallography:

- The isomorphous replacement method, which requires the attachment of heavy atoms to the protein molecules in the crystal.
- The anomalous diffraction method, which depends on the presence of strong anomalously scattering atoms in the protein crystal.
- The molecular replacement method for which the similarity of the unknown structure to an already known structure is a prerequisite.
- Direct methods, which attempt to derive the structure factor phases, electron density or atomic coordinates by mathematical means from a single set of X-ray intensities.

Methods of solving the phase problem using molecular replacement and single wavelength anomalous scattering will be described.

1.3.1 Molecular replacement

The procedure of obtaining phases by molecular replacement (MR) relies on the fact that proteins, homologous in their amino acid sequence, have a very similar folding of their polypeptide chain. The problem is to transfer the known protein molecular structure from its crystalline arrangement to the crystal of the protein for which the structure is not yet known. The solution is the MR method, which was initiated in the pioneering studies by Rossmann and Blow (1962). Placement of the molecule in the target unit cell requires its proper orientation and precise position, which involves rotation and translation. In the rotation step the spatial orientation of the known and unknown molecule with respect to each other is determined while in the next step, the translation needed to superimpose the now correctly oriented molecule onto the other molecule is calculated. If a crystal structure has more than one protein molecule or a number of equal subunits in the asymmetric unit, then their relative position can also be determined by MR. This noncrystallographic symmetry is useful information in the process of improving protein phase angles by noncrystallographic symmetry averaging [20].

The basic principle of the MR method can be understood by regarding the Patterson function of a protein crystal structure. The patterson map is a vector map: vectors between atoms in the real structure show up as vectors from the origin to maxima in the Patteron map. If the pairs of atoms belong to the same molecule, then the corresponding vectors are relatively short and their end-points are found not far from the origin in the Patteron map; they are called *self-Patterson* vectors. If there were no intermolecular vectors (*cross-patterson* vectors), this inner region of the Patterson map would be equal for the same molecule in different crystal structures, apart from a rotation difference. For homologous molecules it is not exactly equal but very similar. Therefore, the *self-Patterson* vectors can supply us with the rotational relationship between the known and the unknown molecular structures. From the *cross-Patterson* vectors the translation required for moving the molecules to their correct position can be derived. The principle of separating the Patterson vectors into these two groups can be used for orientation and translation determination.

The rotation function $R(\theta_1\theta_2\theta_3)$ can be defined as a match between two Patterson functions , such that

$$R(\theta_1\theta_2\theta_3) = \int_U P_1(x)P_2(x')dx \tag{1.11}$$

where $R(\theta_1\theta_2\theta_3)$ represents the rotation function when the Patterson $P_1(x)$ (of the known structure) is rotated by the angles θ_1, θ_2 and θ_3 from an arbitrarily defined original orientation and then superimposed onto another Patterson $P_2(x')$ (of unknown protein). The superimposed Pattersons are then integrated over the the volume U. A large value of R indicates a good match between the two Pattersons [105]. Here a *cross-rotation* function determines the orientation of the known molecule with respect to the unknown molecule in each of the asymmetric units of the crystal under investigation. In the presence of NCS, a *self-rotation* function, the self-vectors for two NCS related molecules fulfill the symmetry between the molecules, i.e., maxima found in the rotation function calculated using the Patterson function of the measured dataset and its rotated image correspond to the number and type of symmetry operators of the molecules connected by NCS.

Having the orientation of the search model in the asymmetric unit, the next step is to calculate the translation required to overlap the model onto the unknown structure. One straightforward technique to determine the position of the model is the calculation of the translation function that uses the cross Patterson vectors of the model structure and the observed Patterson function,

$$T(\mathbf{t}) = \int_v P_{1,2}(\mathbf{u}, \mathbf{t}) \times P(\mathbf{u}) d\mathbf{u} \tag{1.12}$$

where $P_{1,2}(\mathbf{u},\mathbf{t})$ is the set of cross Patterson vectors of two molecules related by crystallographic symmetry and $P(\mathbf{u})$ is the observed Patterson function calculated from the measured data.

Bias minimization

Model bias is a significant problem in macromolecular crystallography because it can lead to misinterpretation of electron-density maps, even when the maps are relatively accurate overall [6, 101, 100, 50, 61]. It can occur in many stages of crystallographic analysis and is particularly important in molecular replacement, model building, ligand-binding and conformation-change studies and structure validation. Many methods like $\sigma_A$weighted maps [101], composite omit maps [50, 12] and prime and switch phasing [122], have been developed for reducing model bias in electron density map calculations.

1.3.2 Experimental phasing

SAD phasing

The X-rays diffracted by a non-anomalous scatterer arises from the electrons surrounding the atomic nucleus. This scattering is represented by the atomic form factor, $f^0(\theta)$. If X-rays can excite those electrons which are able to jump from a lower to a higher energy shell, an auxillary resonant anomalous signal is observed. The anomalous signal can be described by an additional complex form factor, f'+if''. Both f' and f'' depend on the wavelength of the experiment performed. The full atomic form factor (Figure 1.2a) is

$$f(\theta, \lambda) = f^0(\theta) + f'(\lambda) + if''(\lambda) \tag{1.13}$$

The anomalous effect increases with increasing wavelength upto the point that corresponds to the absorption edge of the particular element, where it diminishes abruptly. The anomalous signal obtained from light atoms is usually negligibly small, but careful data aquisition can allow phasing of proteins on the anomalous differences of for example intrinsically present sulfurs in proteins.

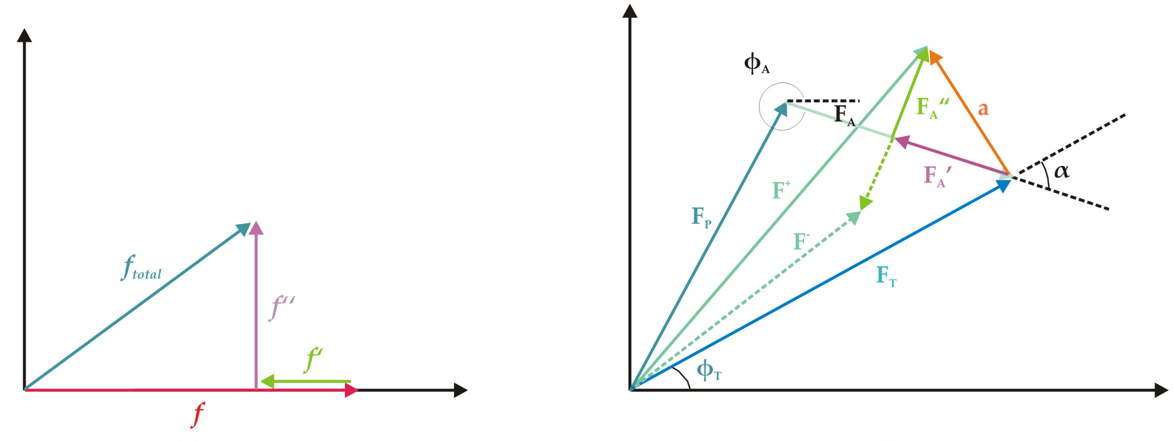
The total non-anomalous scattering factor $\mathbf{F}_\mathrm{T}$ is a sum of $\mathbf{F}_\mathrm{P}$, the structure factor of the protein and $\mathbf{F}_\mathrm{A}$, the structure factor of the anomalous scatterers ignoring their anomalous effects as can be seen in Figure 1.2b. $\mathbf{F}'_\mathrm{A}$ represents the dispersive signal and is applied in opposite direction as $\mathbf{F}_\mathrm{A}$. The vector **a** represents the anomalous contribution of the heavy atoms:

$$\frac{f'}{f''}F_A + \frac{if''}{f'}F_A = F'_A + F''_A = a \tag{1.14}$$

As a consequence of anomalous scattering the structure factors of Friedel mates $\mathbf{F}^+$ and $\mathbf{F}^-$ are no longer equal in length and $\varphi^+ \neq \varphi^-$. From the construction shown in Figure 1.2b the following relations can be drawn. Applying the cosine theorem to the triangle determined by $\mathbf{F}_\mathrm{T}$, $\mathbf{F}^+$ and **a**,

$$|F^+|^2 = |F_T|^2 + |a|^2 - 2|F_T||a|cos(\varphi_T + 180 - \varphi_a) \tag{1.15}$$

can be derived and by expressing the length of **a** with $|\mathbf{F}_\mathrm{A}|$, the length of $\mathbf{F}^+$or $\mathbf{F}^-$, i.e.



(a)

(b)

Figure 1.2: (a) Vector representation of the atomic form factor of an anomalous scatterer. (b) Argand diagram in the presence of anomalous scatterers. F_T, the total non-anomalous scattering is constructed as the sum of F_p (structure factor of protein) and F_A (non-anomalous contribution of heavy atoms). The complex conjugate of F^- is shown.

the measured intensities can be written in the general form:

$$|F^+|^2 = |F_T|^2 + \frac{f'^2 + f''^2}{f^2}|F_A|^2 + \frac{2f'}{f}|F_T||F_A|cos\alpha + \frac{2f''}{f}|F_T||F_A|sin\alpha \tag{1.16}$$

where α=φ_T-φ_A. This equation gives the basis of SAD and MAD experiments that aim at the determination of $\mathbf{F}_T$ from the heavy atom structure factor $|\mathbf{F}_A|$ and the angle α between $\mathbf{F}_A$ and $\mathbf{F}_T$. The phase of the heavy atom structure factor $\mathbf{F}_A$ can be deduced from the heavy atom substructure that is usually solved using $|\mathbf{F}_A|$ values calculated from equation 1.16.

Density modification techniques

The phase information obtained from experimental measurements on macromolecules using isomorphous replacement or anomalous diffraction is usually not sufficient for constructing an electron density map useful for model building and interpretation. Many density modification techniques have been developed in recent years for improving the quality of electron density maps by incorporation of prior knowledge about the features expected in these maps. Most powerful of these methods are solvent flattening, histogram matching, phase extension, molecular replacement, entropy maximization and iterative model building. The fundamental basis of density modification methods is that there are many possible sets of structure factor amplitudes and phases that are all reasonably probable based on the limited experimental data, and those structure factors that lead to maps that are most consistent with both the experimental data and the prior knowledge are the most likely overall.

Phase improvement by density modification involves the alternate applications of two processes: (a) Map modification, where the electron density is modified in accordance with chemical knowledge about the nature of protein structures. This gives rise to a modified map. (b) Phase combination, where the modified phases are weighted on the basis of how well the modified magnitudes match to their observed values, and then perform a weighted combination with the experimental phases, which is performed in reciprocal space. Knowledge about expected values of the electron density in part or all of the unit cell can be a very strong constraint on the crystallographic structure factors, which thereby improves the electron density. For example, prior knowledge about electron density often consists of the identification of a region where the electron density is flat owing to the disordered solvent. Real-space information of this kind has generally been used to improve the quality of crystallographic phases obtained by other means, such as single wavelength anomalous dispersion.

1.4 Refinement

Refinement is the process of adjusting the model to find a closer agreement between the calculated and observed structure factors. The agreement index between calculated and observed structure factors is usually represented by an R-factor calculated as

$$R = \frac{\sum_{hkl} ||F_{obs}| - k|F_{calc}||}{\sum_{hkl} |F_{obs}|} \tag{1.17}$$

The adjustment of model consists of changing the positional parameters and the temperature factors for all atoms in the structure, except the hydrogen atoms. Thus, in a medium resolution data set, the parameters to be refined are three positional parameters (x, y and z) and one isotropic temperature factor parameter (B). Constraints and restraints are means, which may be necessary in cases where parameters should be fixed to certain values. A constraint is an exact mathematical relationship which affects one or more parameters in such a way that not all the parameters can be freely and independently refined. Constraints reduce the number of refined parameters and thereby increase the data to parameter ratio. A restraint is an approximate target value for a particular parameter or a function of some parameters. It provides an extra piece of information about the structure. It thus increases the number of observations rather than reducing the number of parameters.

Least-Squares refinement

The refinement technique is based on the principle of least squares. In least squares the observations have fixed values and the parameters are varied such that the calculated values approach the observations as closely as possible in the refinement. Refinement by least squares is an iterative process. In each step the parameters to be refined change, or rather should change, toward their final value without reaching them right away. Usually a great many cycles are carried out before the refinement converges to the final parameter set. In least squares refinement (unrestrained refinement) the function to be minimized is

$$Q = \sum_{hkl} w(hkl)(|F_{obs}(hkl)| - |F_{calc}(hkl)|)^2 \tag{1.18}$$

The summation is over all crystallographically independent reflections and w is the weight given to an observation.

Maximum-Likelihood refinement

The Bayesian view of maximum likelihood crystallographic refinement is that the prior probability comes from chemistry and the likelihood comes from the X-ray diffraction experiment. The probability function for refinement (here called $P_{\text{refinement}}$) is thus, by Bayes' theorem, the product of the prior probability (here called $P_{\text{chemistry}}$) and the likelihood (here called P_{Xray})

$$P(model; experiment) = P(model) \times P(experiment; model)$$

$$P_{refinement} = P_{chemistry} \times P_{Xray}$$

P(model;experiment) reflects the knowledge about model after experiment. *P*(model) is what is known before the experiment, for example bond lengths, bond angles, chiral volumes etc. P(experiment;model) is the behaviour of the experiment if the model would be known. In other words conditional probability distribution of experimental data if coordinates with errors are known.

For example, a carbon-oxygen double bond is known to be 1.23 Å long. So, if the electron density for a structure showed no density 1.23 Å from a particular peptide carbon, but a strong of density 2 Å away from it, prior knowledge of the carbon-oxygen double-bond length means that it would be extremely likely that the density 2 Å away was due to noise or some other features of the structure. However, if the O atom had been moved into this density during rebuilding, a refinement program would use Bayes' theorem to restrain the bond length to 1.23 Å and produce the more likely structure [78].

Simulated Annealing

Crystallographic refinement can be formulated as a chemically-restrained non-linear optimization problem. The goal is to optimize the simultaneous agreement of an atomic model with observed data and known chemical information. This target function depends on several atomic parameters (coordinates, B-values, occupancies). The large number of adjustable parameters gives rise to a very complicated target function. This results in the multiple minima problem: there are many local minima in addition to global minima. A technique which is good at overcoming local minima is simulated annealing. Annealing is a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase, which is then slowly cooled so that all particles arrange themselves in the lowest energy state. Simulated annealing is the computational simulation of the annealing process. Simulated annealing increases the probability of finding a more optimal solution than gradient-descent methods because motion against the gradient is allowed. The likelihood of this uphill motion is determined by the temperature. Local conformational changes lower potential energy, hence increasing kinetic energy. This energy is removed by temperature control. Once in a lower energy position (i.e. sidechain in density) it is unlikely to jump over the local energy barrier back to a wrong position.

Chapter 2

Analysis of Non-Merohedral Twins

2.1 Introduction

Twins are classified as regular aggregates consisting of crystals of the same species arranged together in some defined mutual orientation [40]. A twin can be defined by a twin law (a symmetry operation defining the orientation of a domain/s with respect to other domain/s) and twinning fraction (defining fractional contribution of each domain). As a result of such a symmetry operation, which usually does not belong to the point group of the crystal system, the diffraction pattern from different domains is either rotated, reflected or inverted depending on nature of the twin law and the corresponding intensities are weighted according to the twinning fraction [91]. As described in Herbst-Irmer & Sheldrick (1998) [48]; twins can be broadly classified into merohedral, pseudo-merohedral, reticular-merohedral and non-merohedral (NMT). In summary the properties of different kinds of twinning are as follows:

- In merohedral twins, the twin law is a symmetry operator of the crystal system, but not of the point group of the crystal. Therefore the reciprocal lattices of different domains superimpose exactly and twinning is not directly detectable by looking at the diffraction pattern. These can be further classified as racemic (twin operator belongs to Laue group but not to point group) and holohedral (twin operator does not belong to the Laue group) twins.

If J_1 and J_2 are untwinned intensities of a pair of reflections related by twinning operation, then the individual intensities measured from a twinned crystal with twinning fraction α can be calculated as $I_1 = \alpha J_1 + (1 - \alpha)J_2$ and $I_2 = (1 - \alpha)J_1 + \alpha J_2$. When the twinning fraction is equal to 0.5, the intensities will merge perfectly in the higher symmetry point group. When the twinning fraction is less than 0.5, merging in higher symmetry point group produces R_{int} values which are not very big compared to the correct point group. The most important warning sign from such a twinned crystal is that the reflection intensities do not obey Wilson's statistics. For example in a diffraction pattern from an untwinned sample, one can observe a very low percentage of very weak reflections and also a smaller number of strong reflections. But, in case of a merohedral twin, where there is exact overlap of reflections, the probability that the twin operation brings two very weak or very strong reflections together is less. As a result the number of very weak or very strong reflections decrease drastically and have a sharper distribution around the mean value than those resulting from Wilson's statistics. An ideal example of diffraction pattern from a merohedral twin is shown in Figure 2.1b. Detection, structure solution and refinement of such a twin will be further described in Chapter 3.

- In pseudo-merohedral twins, the twin operator belongs to a higher crystal system than the structure. This usually occurs, when the metric symmetry is higher than the symmetry of the structure. For example in monoclinic structures where β equals 90°.
- Obverse/reverse twinning in rhombohedral space group is a typical example of twinning by reticular merohedry. The reciprocal lattices from different domains superimpose exactly, but the systematically absent reflections for one domain are present for the other domain and *vice versa*. One third of reflections are still absent. There are both overlapped and non-overlapped reflections.
- The most important difference between NMT and other types of twinning is that in NMT the twinning operator is an arbitrary operator. This means that the reciprocal lattices of the two/more different lattices do not superimpose exactly and the twinning can be detected in most cases by observing the diffraction pattern.

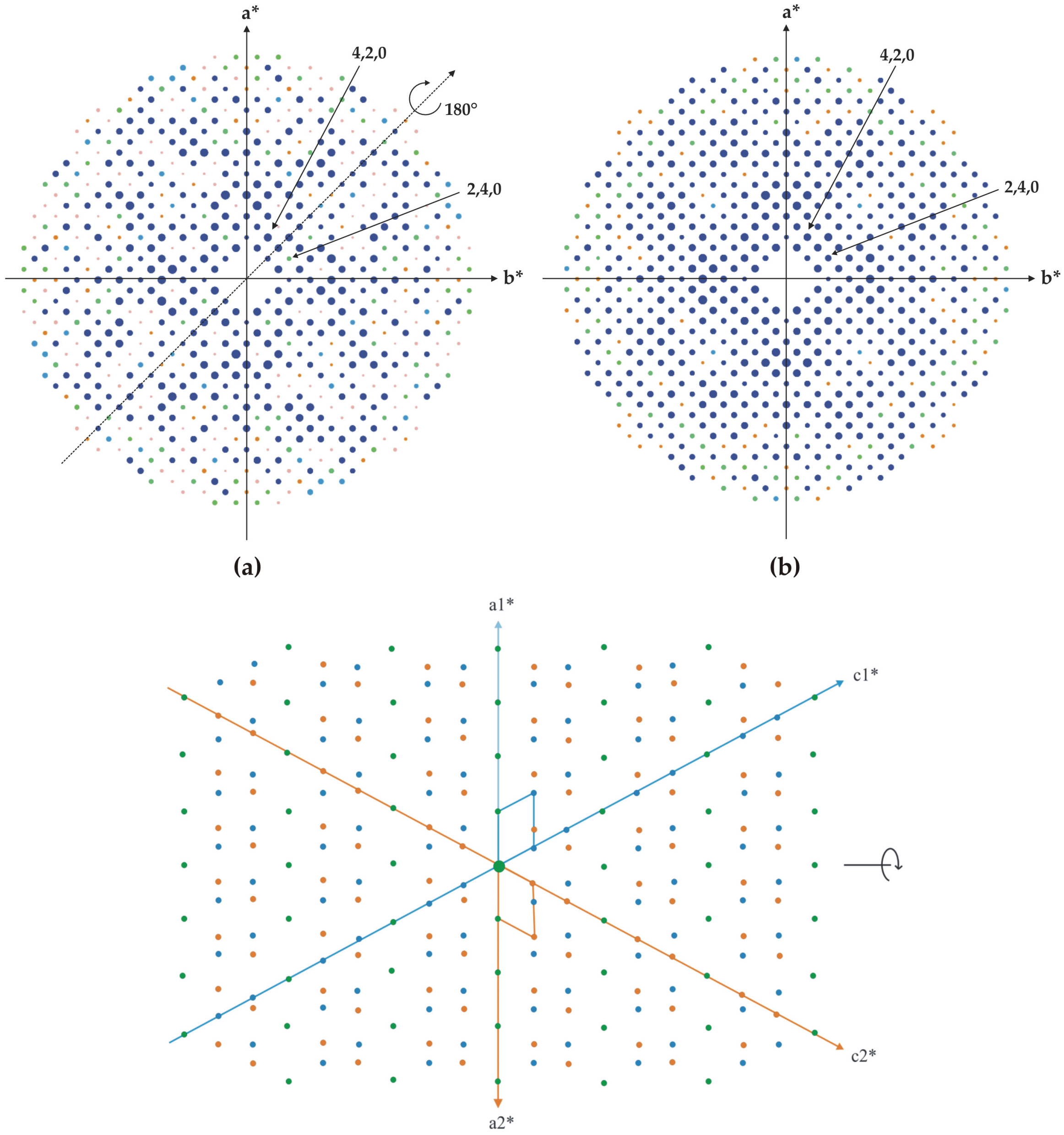
NMT can occur in any space group, where the twinning operation can be a 2-, 3-, 4-, or even a many fold rotation producing a twin (Zwilling in German), triplet (Drilling in German), quadruplet or a multiple-twinned crystal (Mehrling in German). As a result of such a rotation, three kinds of reflections can be observed in the diffraction pattern: exactly overlapping (EOR), partially overlapping (POR) and non-overlapping reflections

(NOR) as seen in Figure 2.1c. Here, one can see an easy case of twinning axis, in which every reflection, where $l = 2n$ is exactly overlapping and only EOR and NOR reflections are seen. Most common observation is that one finds a major domain with a fractional contribution of 80-90% and one or more minor domain/s.

Cell determination using automatic software is problematic and sometimes one finds a super-cell fitting a large number of reflections but with a larger cell volume than expected. Data collection and processing requires special software taking the overlapping reflections (EOR, POR & NOR) into account. Phasing requires de-twinned data containing contributions only from the major domain and for refinement to converge, contributions from all the domains should be taken into account. The methods leading to a successful crystal structure determination from such a twinned crystal will be described in this chapter.

2.2 Aim of present work

With the advent of high throughput crystallography, the number of crystal structures being solved each year has taken a giant stride. Not all the attempts to solve macromolecular structures, despite having good crystals and X-ray data, are successful, crystal twinning being one of several reasons. Twinning is certainly not a problem always resolvable by routine crystal structure determination packages/softwares and requires crystallographer's intervention at least during some stages of structure determination. Processing X-ray data from twinned small-molecule crystals has been a routine method for several years, but there was a significant backlash towards extending and improving these methods to be suitably applicable to macromolecules until recently [131, 18]. In the present work, it has been attempted to optimize data processing, phasing and refinement of non-merohedrally twinned (NMT) macromolecular crystals. Crystal structure solution of two such twinned test protein structures namely, Bovine Insulin (BI-Zwilling), from here on referred to as BI-Zwilling and Glucose Isomerase (GI-Drilling), from here on referred to as GI-Drilling, using Single Wavelength Anomalous Scattering (SAD) with in-house data will be described.



(a) (b)

(c)

Figure 2.1: (a) Diffraction pattern from an untwinned tetragonal lattice. (b) Diffraction pattern from a twinned tetragonal lattice. The twinning operator is a 2-fold rotation about 110 plane in (a), with a twinning fraction of 0.33. As a result one can see that the reflection 2,4,0 and 4,2,0 have similar intensity distribution. (c) Reciprocal lattice of a two-component non-merohedral twin. First component is colored in blue and second component in orange. The non-merohedral twin is formed by a 2-fold rotation around C* axis as a result every reflection, where l = 2n is exactly overlapped. The exactly overlapped reflections are colored in green.

2.3 Methods

2.3.1 Data Collection

Both data sets were collected at 100 K using ω scans with Bruker rotating-anode generator, at Cu Kα wavelength, equipped with Osmic focussing mirrors and a Bruker *SMART* 6000 4K detector. The data were collected in low, medium and high resolution passes at a detector distance of 10 or 18 cm in thin-slice mode to minimize artificial overlap of the reflections because of detector geometry. A minimum of three runs for each of low, medium and high resolution passes were collected at different φ angles to obtain complete and multiple observations of data in order to maximize the weak anomalous signal from sulfur (in the case of BI-Zwilling) and manganese (in the case of GI-Drilling) at Cu Kα wavelength. It was important to collect data, as precise as possible avoiding ice rings, pin in the beam etc. so that the only problem encountered during data processing was caused by twinning. Lower mosaicity of both crystals was an additional advantage in solving the current problem.

2.3.2 Twin indexing

NMT can be identified very easily during data collection from visual inspection of frames and if one has a vast collection of crystals that might not be twinned, one should avoid collecting data from NMT crystals, which would in any case lead to bad data statistics compared to untwinned crystals. But sometimes such twinning is common and unavoidable especially in plate like crystals due to biophysical properties of the sample itself. In such cases indexing of crystals using automatic indexing softwares is often not possible and special programs like *GEMINI [118], DIRAX [22]* and *CELL_NOW [116]* should be used. The program *CELL_NOW* was used to obtain unit cell and orientation matrix of two or more domains constituting the twinned crystal.

In the program *CELL_NOW*, multiple random real-space starting vectors **d** with lengths between user-input limits $\boldsymbol{d}_{min}$ and $\boldsymbol{d}_{max}$ are refined by iterative linear least squares refinement, and then for each vector, the function

$$Q = \sum_{i=1}^{n} W_i (T_i - M_i)^2 \tag{2.1}$$

is minimized, which is a sum of squares of the differences between T_i ($T_i = d_x.X +$

$d_y.Y$ + $d_z.Z$) and M_i (the nearest integer to T_i) for each reflection with the weighting scheme:

$$w_i = I^{\frac{1}{2}} P^2 [P^2 + (M_i - T_i)^2] \tag{2.2}$$

where P^2 is the *precision factor*, which is usually 0.01, and I is the intensity of reflection i. T_i=$\boldsymbol{dX}_i$, where $\boldsymbol{X_i}$=$\boldsymbol{S_i}$-$\boldsymbol{S_0}$ ($\boldsymbol{S_i}$ is the observed scattering factor). This strongly downweights reflections that do not fit well [(M_i-T_i)^2] >> P^2] and so concentrates the refinement on the better-fitting reflections, which in general will belong to the same twin domain. The reflections that do not fit the first twin domain are then used to search for a second domain. After locating the first domain the program rotates the same unit cell and tries to dock it using non-fitting reflections. Since the cell is only re-oriented and not re-determined, minor domains can also be located in this way. *CELL_NOW* prints out the two/more orientation matrices, the unit cell, the twin law, the number of reflections fitting each domain of the twinned crystal and also gives an indication of the lattice centering of the crystal.

In order to get the best possible orientation matrices for all the domains it is important to threshold reflections spanning a wide range of frames separated by a few degrees.

2.3.3 Multicomponent data integration

Once one is able to index a NMT crystal, data-integration poses a non-trivial problem. One can ignore twinning and integrate using only one/major orientation matrix, wherein all the available data processing programs and refinement programs can be used. But, this ofcourse leads to poorer results and the degree of poorer results depends on the relative amounts of twinning fraction. One can omit all the overlapping reflections which could lead to omission of most of the data. Alternatively, one can omit all the partially overlapping reflections (POR) and split the exactly overlapping reflections (EOR) by integration of the main domain and then splitting the EOR to prepare a *hkl* file with both the components. All data processing programs can possibly be used and *SHELXL* [113] can be used for refinement. This could also lead to omission of most of the data and thereby low completeness. The best strategy so far used in *SAINT* (Bruker AXS, Madison, WI) or *EvalCCD* [23] is to integrate the data using the information from all orientation matrices.

SAINT uses a profile based twin pairing method to determine when spots start to overlap. In case of twin overlaps, *SAINT* estimates the intensities of individual spot components and writes them to the output reflection file but the intensity esd's are still equal for each spot components and correspond to esd of the aggregate intensity sum. *SAINT* takes two/more orientation matrices from *CELL_NOW* and produces a multi-component file with contribution from both/all domains. After this stage we have overlapping reflections (EOR and POR) that are crudely de-twinned and from now on will be referenced as OR (overlapping reflections) and non-overlapping reflections (NOR).

To extract the best possible anomalous signal from the data it was important to integrate the low, medium and high resolution scans separately and to repeat the integration with refined orientation matrices especially for medium and high resolution scans. It was also necessary to constrain the unit cell of the second and third components to the first one for better refinement convergence.

2.3.4 Multicomponent data scaling

In case of NMT, scaling, absorption correction and merging of reflections from crudely de-twinned intensities from *SAINT* is done using a special version of *SADABS* called *TWINABS* [115]. One can choose to use the same scaling model or different scaling models for different components of the twin. In the present examples using different scaling models for different components produced better results because of different behaviour of crystals in the beam, for the simple reason that it is physically impossible to centre both crystals perfectly on the goniometer head. *TWINABS* reads in the multi-component raw intensities from *SAINT* and after scaling the intensities, produces a merged or an unmerged (Friedel mates are not merged) *HKLF4* file ($h\ k\ l$, F_o^2, $\sigma(F_o^2)$) containing contributions from first/major domain, which is crudely de-twinned for structure solution and a merged/unmerged *HKLF5* file ($h\ k\ l$, F_o^2, $\sigma(F_o^2)$, +/-m, where m is the m^{th} component of n components) containing contributions from all components for structure refinement. Reflections with contribution from the major domain only have been used for refinement, meaning in *HKLF5* file, we have reflections where m is either 1 for NOR reflections or -2 followed by 1 for OR in case of a two component twin.

2.3.5 Phasing and model building

Both substructures were solved using dual-space recycling methods using the detwinned data in *SHELXD* [109]. The normalized difference structure factors were calculated using *XPREP* (Bruker AXS, Madison, WI) from *HKLF4* file prepared by *TWINABS*. Density modification was carried out using *SHELXE* [114]. The new free-lunch algorithm in *SHELXE*, which improves the electron density map by extrapolating the moduli and phases of non-measured reflections beyond and behind the experimental limit [13], produced better maps both for BI and GI twins. At this stage, any macromolecular experimental phasing programs can be used for substructure solution and density modification, once a crudely de-twinned data set is available. The map from *SHELXE* was traced using *ARP/wARP* [93] in both cases. Experimental phasing would work occasionally, where large number of reflections are unaffected by twinning, otherwise such structures can be solved by molecular replacement methods upon availability of a good search model (when the search model is at least 35% similar).

2.3.6 Twin Refinement and Validation

For refinement of structures, a *HKLF5* format hkl file was used containing the contribution from both/all components. The structures were refined with *SHELXL* applying adequate bond angle, bond length, chiral volume restraints and constraints wherever necessary. The twin refinement method of Pratt *et al.*, (1971) [99] and Jameson (1982) [53] has been implemented in *SHELXL* as described in Herbst-Irmer & Sheldrick (1998) [48]. F_c^2 is calculated as

$$F_c^2 = (o.s.f)^2 \sum_{m=1}^{n} k_m F_{cm}^2 \tag{2.3}$$

where $o.s.f$ is the overall scale factor, k_m is the fractional contribution of twin domain m, F_{cm} is the calculated structure factor of the twin domain m and n is the number of twin domains. The sum of the fractional contributions k_m must be unity, so (n-1) of them can be refined, which are set by the *BASF* parameter and k_1 is calculated by

$$k_1 = 1 - \sum_{m=2}^{n} k_m \tag{2.4}$$

The refined *BASF* parameter gives the fractional contribution of each component.

Since the present version of *SHELXL* cannot deal both with *HKLF5* format file and calculate R_{free} at the same time, the *HKLF5* data was divided into a 5% R_{free} set calculated in thin shells and a second data set containing 95% of data for refinement using a *PERL* script, where care was taken to include all the twin related and symmetry related reflections in the same set. It is forseen that future versions of *TWINABS* would offer an opportunity to generate the R_{free} flags in a much more elegant way.

At the end of each refinement, *SHELXL* was rerun with the R_{free} data set and zero cycles of conjugate-gradient least squares refinement to calculate the free-R value. Each refinement was followed by thorough check of the geometry of the structure and additional disordered components using $2\boldsymbol{F_o}$-$\boldsymbol{F_c}$ and $\boldsymbol{F_o}$-$\boldsymbol{F_c}$ maps in the program *COOT* [25]. It should be noted that the $\boldsymbol{F}_c$ coefficients calculated in *SHELXL* are de-twinned but without σ_{A} weighting and can be read by any graphics program. These refined models were used as reference structures to calculate mean phase error and map correlation coefficients using the method of Lunin and Woolfson (1993) [74], which is implemented in a pre-release version of *SHELXPRO* [113].

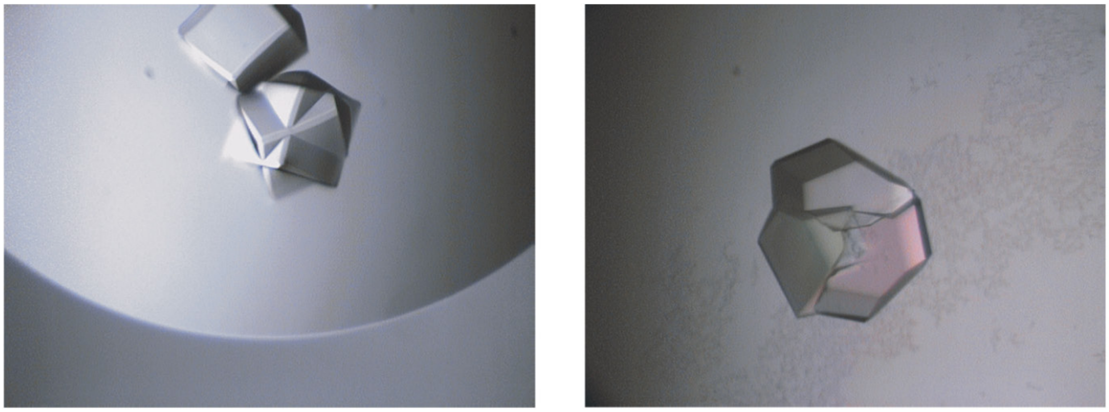
The structures were validated with *PROCHECK* [67]. All residues lie within the allowed regions of the Ramachandran plot.

2.4 Results

In the present study two well known standard test proteins, namely, Insulin and Glucose Isomerase, that can be easily crystallized have been used. Both crystals were processed as twinned data sets and phased using anomalous scattering.

2.4.1 Twinned Bovine Insulin

Bovine Insulin is a hormone with 51 amino acids and 3 disulfide bonds. Taking advantage of the six sulfurs in the asymmetric unit and cubic symmetry, it was possible to phase the protein using sulfur-SAD though the crystals were twinned. BI (e.g. Sigma, cat. No. I5500) was dissolved in 0.02 M $\mathrm{Na_2HPO_4}$ and 0.01 M $\mathrm{Na_3EDTA}$ to a final concentration of 30 mg/ml. BI-Zwilling was crystallized by hanging-drop vapour diffusion method [80] at 20 °C by equilibrating against a reservoir containing 0.2 M $\mathrm{Na_2HPO_4/Na_3PO_4}$ pH 10.0 and 0.01 M $\mathrm{Na_3EDTA}$. Cubic crystals grew in about an hour, and most of the crystals were inter-grown due to higher concentration of protein as in



(a) (b)

Figure 2.2: Crystals of (a) Bovine Insulin (BI-Zwilling) and (b) Glucose isomerasae (GI-Drilling) crystallized by hanging drop vapour diffusion method at 20 °C.

Figure 2.2a.

The data was collected from the twinned crystal flash frozen using 30% glycerol as cryoprotectant under liquid nitrogen stream and with a detector distance of 10 cm. The twin law from *CELL_NOW* and data collection statistics are described in Table 2.1 and Table 2.2. The low, medium and high resolution data were integrated separately using *SAINT* and scaled using *TWINABS*. Two different scaling models were used for scaling the different components of the crystal in *TWINABS*. The behaviour of the different components of the crystal in the beam can be observed in the plots of normalized scale factors against resolution in Figure 2.3a. But when compared to GI twin, the two components of the BI twin behave similar in the beam because the centers of both domains more or less coincide.

The substructure was solved using *SHELXD* with a correlation coefficient of 40.16 and 25.24 for all and weak data respectively and found all six sulfurs. Figure 2.4b shows the anomalous map around the six sulfurs contoured at 3σ. 20 cycles of density modification with *SHELXE* gave interpretable electron-density maps with a mean phase error of 20° and are similar to final refined maps as can be seen in Figure 2.5a & b . *ARP/wARP* traced 96.9% of the main chain and side chain atoms without any difficulty. Until this stage the detwinned *HKLF4* data was used. The structure refined to a final R-factor of 18.9% and R_{free} of 21.4% using the twin refinement protocol in *SHELXL*. All non-hydrogen atoms were refined anisotropically and waters were modelled using

Table 2.1: Twin law, Twinning axis and degree of Twin rotation axis for BI-Zwilling and GI-Drilling determined by the program *CELL_NOW*. BI-Zwilling is a 90° two component rotation twin and GI-Drilling is a 120° rotation three component twin.

	BI-Zwilling			**GI-Drilling**		
Twin Law						
1st to 2nd Component	0.459	-0.625	0.631	-0.248	-0.066	0.924
	0.824	0.036	-0.565	-0.861	0.373	-0.198
	0.330	0.780	0.532	-0.364	-1.012	-0.159
1st to 3rd Component				-0.558	0.215	-0.780
		-		-0.496	0.674	0.467
				0.677	0.784	-0.254
1st to 2nd Component						
Rotated by	89.2°			121.6°		
about real axis	0.927	0.208	1.000	-0.519	1.000	-0.585
1st to 3rd Component						
Rotated by		-		124.7°		
about real axis				-0.123	1.000	0.449

Table 2.2: Data statistics for BI-Zwilling and GI-Drilling. The GI-Drilling data set has a very low completeness of 81.7%. From the statistics showing the number of unique reflections present in each component, the BI-Zwilling looks like an equal component twin, whereas the third component slightly dominates in GI-Drilling. [1]Values for redundancy and completeness are from XPREP for first component only and values in parenthesis are for the outer resolution shell.[2]Values reported for the first, second or third components are from *TWINABS*.

Data collection statistics	**BI-Zwilling**	**GI-Drilling**
Wavelength (Å)	1.54178	1.54178
X-ray source	Cu Kα (in-house)	Cu Kα (in-house)
Detector	Bruker SMART6000	Bruker SMART6000
Space group	$I2_13$	I222
Resolution (Å)[1]	1.55 (1.64 - 1.55)	1.6 (1.7 - 1.6)

Cell parameters	a = b = c = 78.023 Å
	$\alpha = \beta = \gamma = 90°$
		a = 92.850 Å, b = 98.788 Å,
		c = 103.730 Å
		$\alpha = \beta = \gamma = 90°$

Reflections (unique)[2]		
First component	207931 (11398)	220819 (47708)
Second component	207911 (11403)	221371 (48870)
Third component	-	221749 (52029)
Redundancy[1]	19.44 (6.14)	5.06 (1.95)
Completeness (%)[1]	97.5 (84)	81.7 (61.8)
Mean I/σ(I)[2]		
First component	9.8	8.6
Second component	7.8	5.4
Third component	-	7.5
R_{int} (%)[2]		
First component	2.87	3.83
Second component	3.31	6.62
Third component	-	4.19

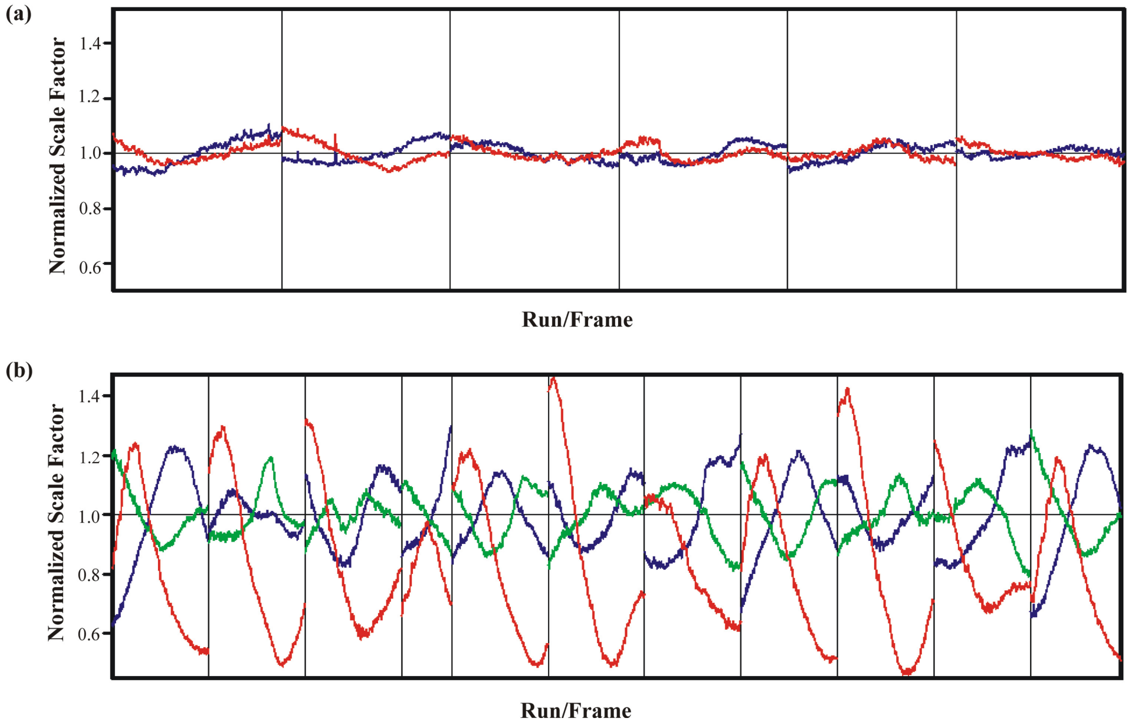


Figure 2.3: Normalized scale factor against run/frame number from *TWINABS* showing the behaviour of different domains in the beam for (a) BI-Zwilling (domain 1 is colored in blue and domain 2 in red) (b) GI-Drilling (domain 1 is colored in blue, domain 2 in red and domain 3 in green.

Table 2.3: Refinement statistics for HKLF5 and HKLF4 format files of BI-Zwilling and GI-Drilling. [1]Five percent of unique reflections in thin shells have been taken as test set. Final working R-factor and total R-factor for BI-Zwilling are quite different for all data and unique data because of Twin-pairing errors. The refined BASF value for BI-Zwilling, suggests a 0.27% twinning fraction, which does not agree well with the numbers deduced from mean I/σI values in Table 2.2, that indicate a fractional contribution of 0.44% as described in Table 2.2. This is most probably an effect of Twin-pairing errors. The R-factors for GI-Drilling are much better at a resolution of 1.6 Å, probably because of the low data completeness.

Refinement statistics	**BI-Zwilling**		**GI-Drilling**	
	HKLF5	HKLF4	HKLF5	HKLF4
Final working R-factor (%)				
$F_o > 4\sigma(F_o)$	18.93	14.64	12.33	16.28
For all data (unique data)	20.42 (15.71)	15.35	15.23 (16.19)	18.8
Final Free R-factor (%)[1]				
$F_o > 4\sigma(F_o)$	21.39	18.43	15.17	19.93
For all data (unique data)	22.68 (19.55)	19.92	17.77 (19.32)	22.69
Final total R-factor (%)				
$F_o > 4\sigma(F_o)$	18.98	14.73	12.38	16.38
For all data (unique data)	20.44 (15.71)	15.46	15.25 (16.24)	18.98
BASF (%) (Twinning Fraction)	27.06	-	20.21 & 44.06	-
Solvent content (%)	58.09	-	49.16	-
Mean B value (Å^2)				
Main chain atoms	18.64	18.369	13.58	13.63
Side chain atoms and solvent	29.94	28.910	21.15	20.92
Number of Protein atoms	389	389	3011	3011
Number of solvent atoms	89	89	420	420

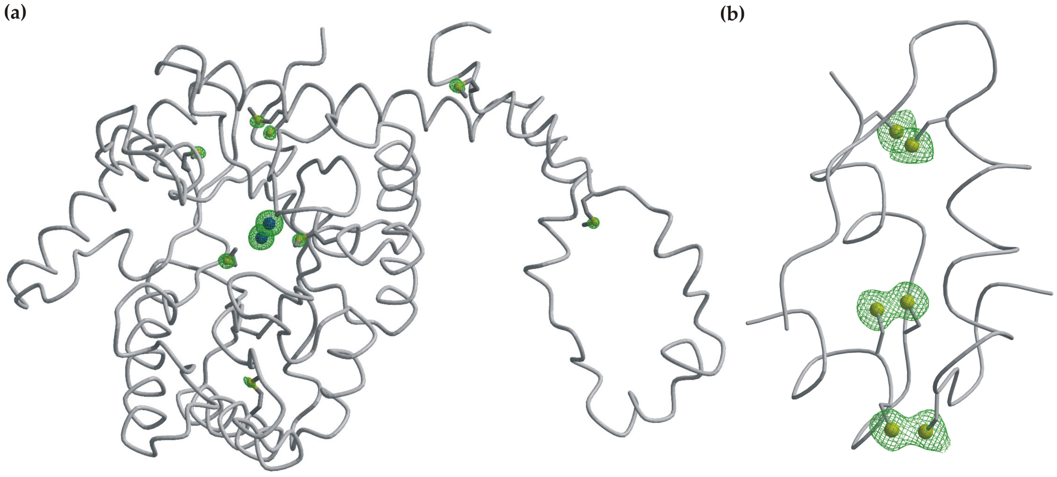


Figure 2.4: Anomalous map from *SHELXE* contoured at 3σ in (a) GI-Drilling, contoured around the 1.5 Mn atoms and some Methionine sulfurs and in (b) BI-Zwilling contoured around the disulfide sulfurs.

ARP/wARP and *COOT*. The refinement statistics are tabulated in Table 2.3. The fractional contribution of each component from the refined BASF values are 73% and 27% respectively. Though the crystal looks like a fifty percent twin, because of the cubic symmetry and non-integer twin law there were many unique reflections, where the same reflection from the first component or its symmetry related one was overlapping with different second component reflections but, when taken into account the fractional contribution of each unique reflection only 27% of the total intensities belong to the second component.

The structure was validated using *PROCHECK*, where 90.5% of residues are in the most favoured regions and 9.5% of residues are in additionally allowed regions of the Ramachandran plot.

2.4.2 Twinned Glucose Isomerase

The active form of Glucose Isomerase has 385 amino acids out of which 8 methionine's, one Mg and one Mn ion at the active site were utilised for anomalous phasing in-house. Glucose Isomerase (Hampton, cat. No. HR7-100) was dialysed against 5 m*M* Tris-HCl buffer, pH 7.5, 10 m*M* $\mathrm{MnCl_2}$ and 5 m*M* $\mathrm{MgCl_2}$. GI was then concentrated to a final

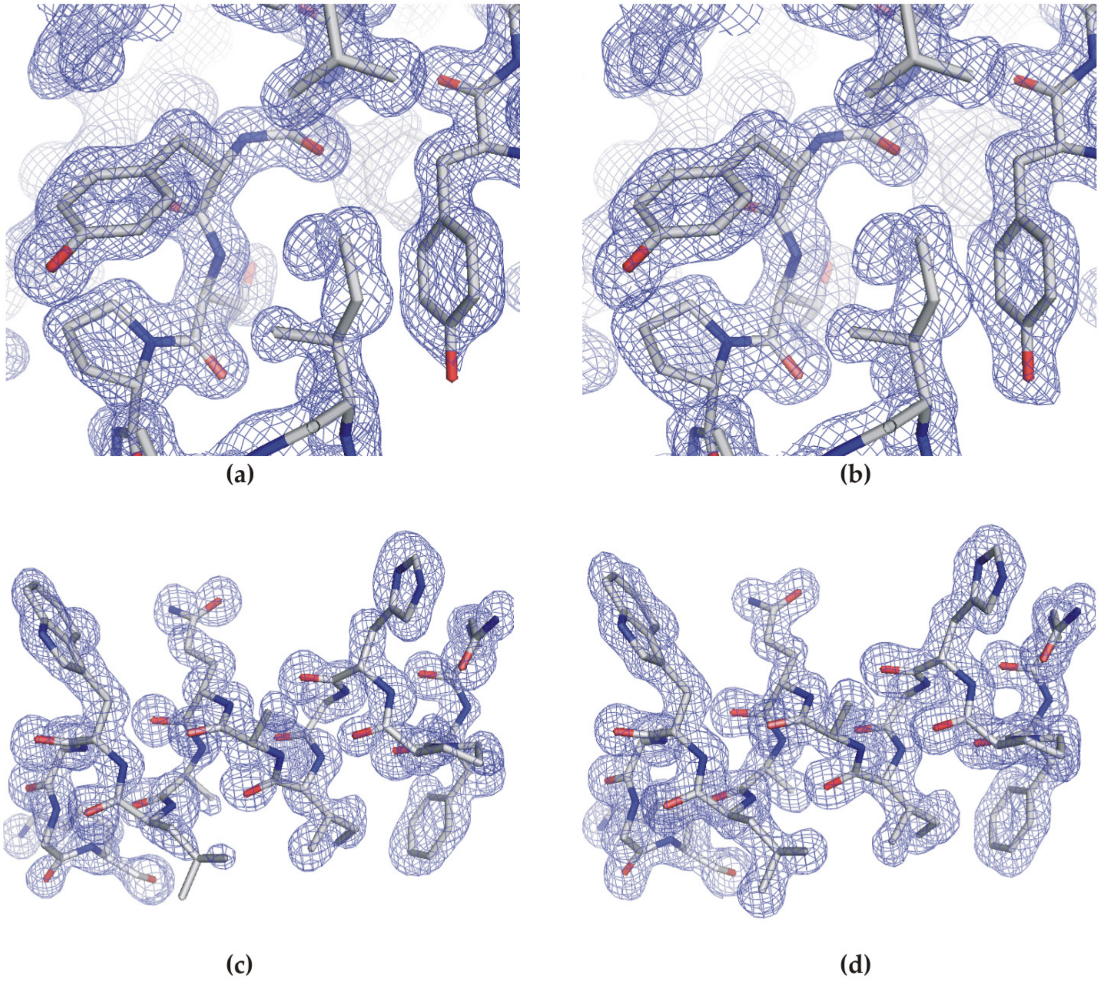


Figure 2.5: (a) *SHELXE* map calculated without the free lunch algorithm and at a resolution of 1.55 Å, contoured at 1σ in BI-Zwilling (b) Final anisotropically refined map from *SHELXL* contoured at 1σ in BI-Zwilling (c) *SHELXE* map calculated using the new free lunch algorithm with data expanded to 1.3 Å and contoured at 1σ in GI-Drilling. Maps from *SHELXE* were quite fragmented, because of less complete and non-optimal data and improved a lot after density modification coupled with model building in *ARP/wARP* (d) Final isotropically refined map from *SHELXL* contoured at 1σ in GI-Drilling

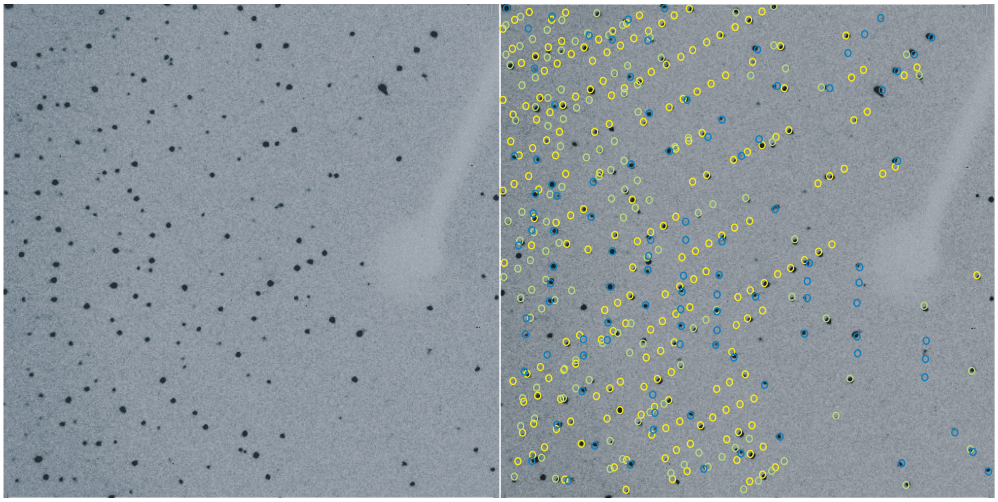


Figure 2.6: An example image of GI-Drilling. On the left is an image taken at $2\theta = 40°$ and a detector distance of 18 cm. On the right is the image overlaid with the three orientation matrices after indexing using *CELL_NOW*, the first domain is colored in green, second in yellow and the third domain in blue.

concentration of 20 mg/ml and crystallized by hanging drop vapour diffusion method by equilibrating against a reservoir containing 0.05 m*M* Tris-HCl buffer pH 7.5, 0.1 *M* $\mathrm{MnCl_2}$ and 14% MPD. Crystals grew in about 2 days as in Figure 2.2b. 25% MPD was used as cryoprotectant and data collected at 100 K with a detector distance of 18 cm because of the long cell axis.

As evident from the crystal in Figure 2.2b, GI crystal was a three-component 120° rotation twin. The twin law and data collection statistics are described in Table 2.1 and Table 2.2. The twin law and the three orientation matrices were determined by *CELL_NOW* and the data integrated using *SAINT*. Figure 2.6 shows an example image of GI-Drilling overlaid with the three orientation matrices from *CELL_NOW*. *TWINABS* plots in Figure 2.3 indicate that the quality of the data from the second component is not as good as from the first or the third component. Normal output of *TWINABS* diagnostic plots has been modified to show the behaviour of all three components in the beam.

The data was truncated to 3.5 Å for substructure solution in *SHELXD* and the structure solved with a correlation coefficient of 23.84 and 20.47 for all and weak data respectively (Figure 2.4a). 500 cycles of *SHELXE* gave interpretable maps with a mean phase error of 30° (Figure 2.5c & d). The map from *SHELXE* was used to trace 98.03% of the total residues with *ARp/wARP*. The structure was refined isotropically using *SHELXL* with a final R-factor of 12.33% and R_{free} of 15.17%. The R-values are comparitively better for this resolution range because of the lower data completeness as seen in Table 2.3. Waters were modelled using *ARP/wARP* and *COOT*. Refinement statistics are in Table 2.3. The final refined BASF value suggests the fractional contribution of each component of the rotation twin to be 44.06%, 20.21% and 35.73% respectively. The refinement suggested that one of the Mn sites was partly occupied with Mg. Both these atoms were restrained to have similar isotropic displacement parameters and x, y, z parameters.

The structure was validated using *PROCHECK*, where 91.5% residues are in the most favoured regions, 7.9% residues are in additionally allowed regions and 0.6% of residues are in generously allowed regions of the Ramachandran plot.

2.5 Discussion and Future Perspectives

SAD phasing: Success *vs* Failure

Upon post-mortem analysis of these data sets for possible explanations that lead to successful experimental phasing, at least half of Friedel pairs in each resolution shell (Figure 2.7a and 2.7c) were unaffected by twinning. This is a major difference to merohedral twinning, where nearly all reflections are affected by twinning and structure solution gets tricky sometimes. The percentage of reflections that were overlapping are more at high resolution than at low resolution at least in the case of GI. Figure 2.7a and 2.7c show the distribution of number of Friedel pairs in each resolution shell and an assesment of anomalous signal in these resolution shells for BI-Zwilling and GI-Drilling respectively.

However it is sometimes possible that when a reflection is distributed in more than a few frames or when a reflection is not overlapped but its symmetry equivalent is overlapped (this situation does not usually occur when the twin law has integer values and the symmetry is lower than cubic), the program *SAINT* assumes and integrates the reflection to be overlapped (containing two/more components) and also as a non-overlapped single-component. These errors, introduced during data integration are called “Twin-pairing errors”. In the histograms presented in Figure 2.7a and 2.7c, these errors have not been removed. The effect of these “Twin-pairing errors” can be seen more in the case of BI-Zwilling because of the cubic symmetry. One can also see that in the histograms presented in Figure 2.7a and 2.7c, the sum total of the overlapped

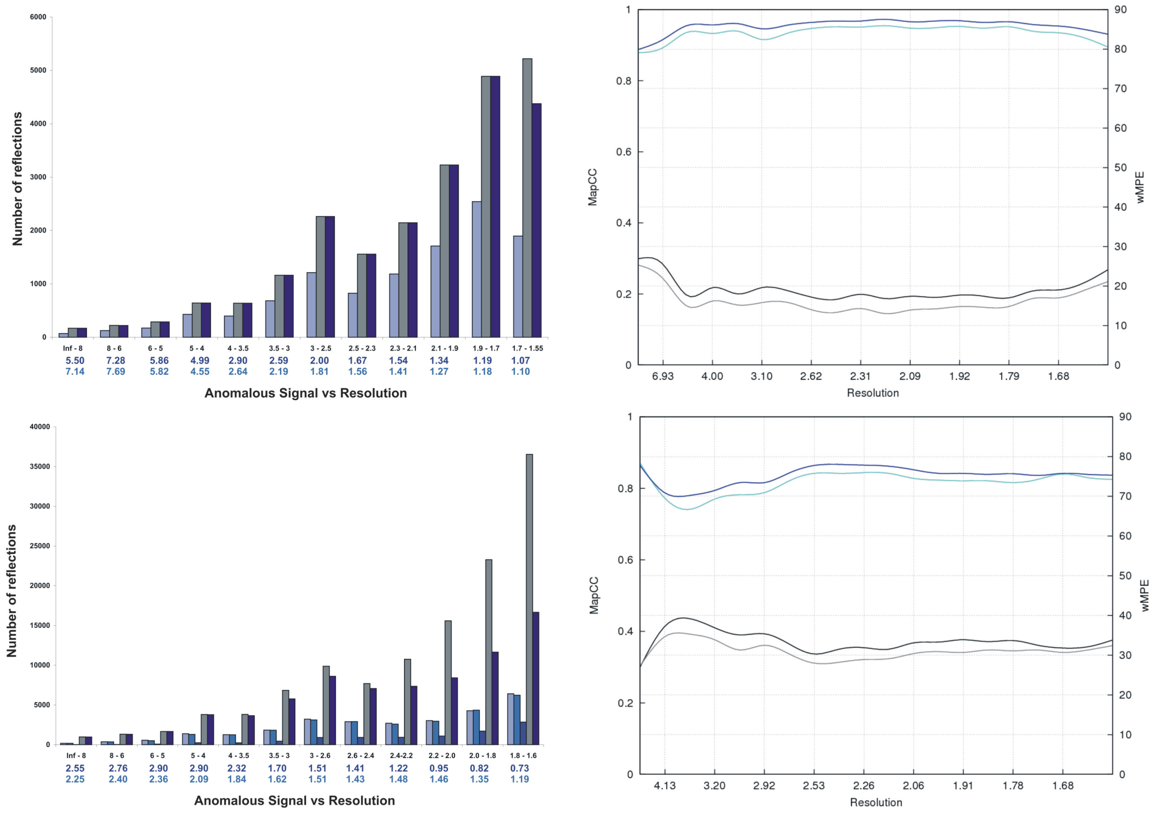


Figure 2.7: Number of overlapped and non-overlapped Friedel pairs in resolution shells in (a) BI-Zwilling and (c) GI-Drilling. In both plots grey bars indicate the theoretical number of Friedel pairs in that particular resolution shell and navy blue bars to the right of grey bars indicate non-overlapped reflections from the first domain. The light blue bars to the left of grey bars are the Friedel pairs (a) of the first component overlapping with the second component and in (c) of the first component overlapping with the second, third and both components. Below the histograms (a) and (c) is an indication of the amount of anomalous signal in each resolution shell. The first line is based on input sigmas and the second on variance of F^+ and F^-. The plots (b) and (d) show map correlation coefficients (MapCC) plotted on y1-axis in blue (*SHELXE* with free-lunch algorithm) and light blue lines (*SHELXE*) and weighted mean phase error (wMPE) plotted on y2-axis in grey (*SHELXE* with free-lunch algorithm) and black lines (*SHELXE*) against resolution for (b) BI-Zwilling and (c) GI-Drilling respectively.

and non-overlapped reflections do not add up to the theoretical number of Friedel pairs expected in that resolution shell, mainly because of these errors. There is a debate on whether to remove or keep these data during structure refinement. During the refinement of current structures, these errors have not been removed. In BI-Zwilling as seen in Table2, the difference between the R-values for all data (20.44%) and unique data (15.71%) is large also because of these errors.

One can see the corresponding results from these data during structure solution in Figure 2.7b and 2.7d, where the mean phase error and map correlations coefficient between maps from *SHELXE* and after final refinement are compared.

It can be summarized from these results that data processing, structure solution and refinement of NMT crystals is not ruled out even for large macromolecular crystals. But at present, these methods are limited to data collected on Bruker detectors and Bruker data processing softwares. As quoted recently in Lebedev *et al*, (2006) [68], the number of structures in the Protein Data Bank that might be twinned, where twinning was ignored during structure solution and refinement is large. Successfull identification and refinement of data from non-merohedrally twinned crystals would require data processing programs to process such twin data collected on other detectors.

Chapter 3

Sensor Kinase CitA

3.1 Introduction

Chemotaxis is a mechanism by which bacteria efficiently and rapidly respond to changes in the chemical composition of their environment, approaching chemically favourable environments and avoiding unfavourable ones [10]. This response to external stimuli, mediated by several signal transduction systems, plays a major role for their survival. The two-component regulatory systems (TCSs), also referred to as the histidyl-aspartyl phosphotransferase systems or histidine protein kinases (HPKs), constitute widespread signal transduction devices in prokaryotes. These mechanisms that are involved in sensing and adapting to changes in chemical and physical parameters in the environment, such as different ions, temperature, pH, oxygen pressure, osmolarity, chemicals and contact with host cells, are regulated mainly through changes in gene expression [59].

Drug discovery aims at studying novel mechanisms to attenuate bacterial virulence [5]. In this context, a detailed understanding of the signal processing pathways in bacterial TCSs is important, primarily because these systems are until now not identified in mammals (and the whole animal kingdom), although some eukaryotic histidyl-aspartyl systems are reported in *Arabidopsis thaliana* (as receptors for the hormone ethylene), *Dictyostelium* [123] and yeast (in osmosensing) [98]. Secondly, interfering with a pathogen's HPKs may expose it to destruction by host immune system rather than being directly toxic, which might also not evoke resistance in the target strain [129]. Thus HPKs offer a new set of targets for novel classes of antimicrobial drugs.

Two component systems (TCSs)

The TCSs are typically composed of a membrane-located sensor with histidine kinase (HK) activity and a cytoplasmic response regulator (RR) [5, 119, 29]. Generally, stimuli detected by these systems are transformed into a cellular signal by autophosphorylation of the cytoplasmic domain of sensory proteins at a conserved histidine residue. The phosphorylated histidine of these sensor proteins is the source for phosphorylation of an aspartic acid residue in the so-called 'receiver domain', which is the amino terminal end of the response regulator. A carboxy terminal DNA-binding domain allows the response regulator to function as a transcriptional regulator. Phosphorylation of the regulatory proteins induces a conformational change, which alters their DNA-binding properties, rendering them competent to regulate gene expression. Some TCSs are characterized by a complex phosphorelay between two histidine and two aspartic acid residues present in four signalling domains [128]. These can either be independent proteins or be integrated into a multidomain TCS in various combinations .

Therefore the chemistry underlying the basic two-component phospho transfer signal transduction pathway involves three phosphotransfer reactions and two phosphoprotein intermediates [119]:

- Autophosphorylation: HK-His + ATP <=> HK-His~P + ADP
- Phosphotransfer: HK-His~P + RR-Asp <=> HK-His + RR-Asp~P
- Dephosphorylation: RR-Asp~P + H_2O <=> RR-Asp + P_i

The γ-phosphoryl group in ATP is first transferred to a conserved histidine (His) side chain of the histidine kinase (HK). Most of the histidine kinases function as dimers or higher oligomers. Autophosphorylation is a bimolecular reaction between homodimers, in which one HK monomer catalyzes the autophosphorylation of histidine in the second monomer and *vice versa*. The response regulator (RR) then catalyzes the transfer of this phosphoryl group from the phospho-His residue to a conserved aspartic acid (Asp) side chain within its own regulatory domain. Finally, the phosphoryl group is transferred from the phospho-Asp residue to water in a hydrolysis reaction. All three reactions require divalent metal ions, presumably Mg^{+2} *in vivo*.

Extracellular PAS domains in cellular responses

Transmembrane sensor kinases or HPKs contain an extracytoplasmic sensor domain (recent crystal structures suggest a PAS fold in few of these domains), flanked by two transmembrane helices, a cytoplasmic signal transfer PAS domain and a histidine kinase domain as in Figure 3.1. PAS is an acronym formed from the names of proteins in which imperfect repeat sequences were first recognized: the *Drosophila* period clock protein (PER), vertebrate aryl hydrocarbon receptor nuclear translocator (ARNT), and *Drosophila* single-minded protein (SIM). PAS domains are combined with a variety of regulatory modules in multidomain proteins from all three kingdoms of life: *Bacteria*, *Archaea* and *Eukarya*. These include histidine and serine/threonine kinases, chemoreceptors, photoreceptors, circadian clock proteins, voltage-activated ion channels, cyclic nucleotide phosphodiesterases and regulators of responses to hypoxia and embryonal development of the central nervous system [120]. In members of *Bacteria* and *Archaea*, PAS domains are found almost exclusively in TCSs. Thus a variety of cellular responses to changes in environmental and intracellular conditions are controlled *via* PAS-containing receptors, transducers and regulators.

Environmental signals are detected either directly or indirectly by the N-terminal periplasmic sensing domain, for example, a PAS fold identified in citrate sensor CitA of *Klebsiella pneumoniae* [102] and *E. Coli* fumarate sensor DcuS. These diverse sensing domains although similar in their three dimensional structure, share little primary sequence similarity, thus supporting the idea that they have been designed for specific ligand stimulus interactions. There are some examples of binding modes of these ligands [102], but the transfer of stimulus from the sensing domain into the cytoplasmic domain/s still remains elusive.

3.1.1 CitA sensor kinase from *Klebsiella pneumoniae*

The two-component regulatory CitA/CitB is essential for induction of the citrate fermentation genes in *Klebsiella pneumoniae* [57]. It was shown that *citA* and *citB* mutants were no longer able to grow under anoxic conditions when citrate was sole carbon and energy source [7, 8, 9]. CitA represents a membrane bound sensor kinase consisting: (1) a periplasmic domain (CitAp) flanked by two transmembrane helices, (2) a linker domain, where a PAS (Per-Arnt-Sim) fold has been predicted and (3) the kinase domain composed of the phosphorylation subdomain and the ATP-binding subdomain

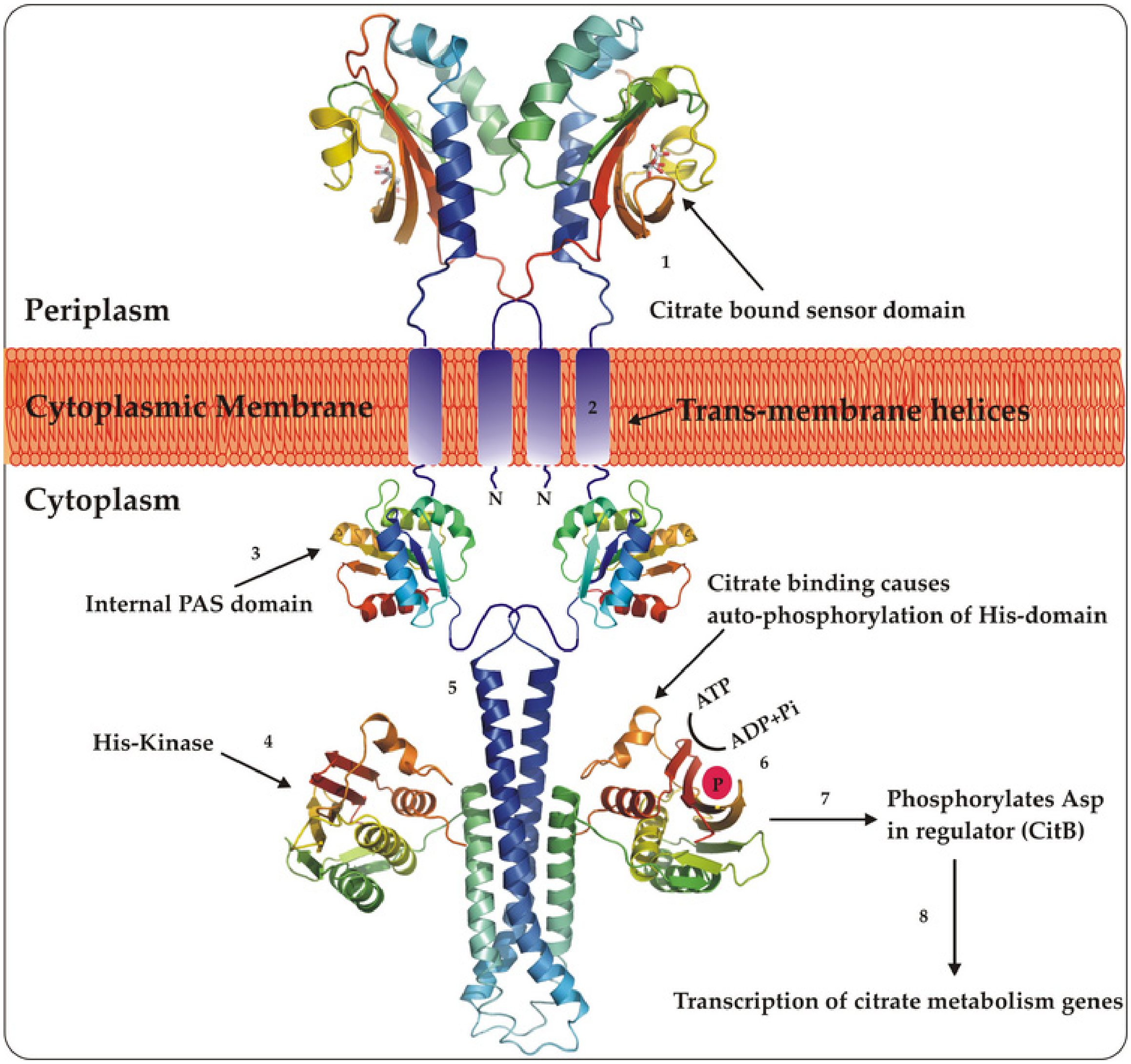


Figure 3.1: A virtual model of CitA histidine kinase TCSs in *Klebsiella pneumoniae*. Two component systems consist of a periplasmic sensor domain (ribbon model of citrate bound CitA, PDB ID: 1P0Z), internal PAS domain (PDB ID: 1D5W) and histidine kinase domain (PDB ID: 2C2A). (1) External stimuli is sensed by, for example binding of the ligand to sensor domain (2) ligand-binding transfers a signal *via* the two transmembrane helices to (3) internal PAS domain (4 & 5) signal transfer causes autophosphorylation of His in histidine kinase domain (7) which then phosphorylates Asp in response regulator (8) leading to transcription of citrate metabolism genes.

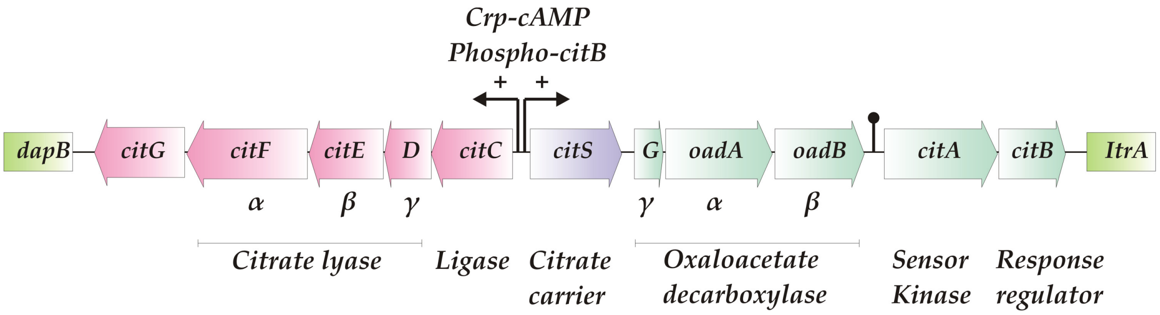


Figure 3.2: Organization of the citrate fermentation genes in *Klebsiella pneumoniae*. The 13 kb DNA region encompassing genes involved in citrate fermentation and the corresponding proteins are shown in the upper part of the figure. Activation of transcription at the promoters in front of *citC* and *citS* by phospho-CitB and Crp-cAMP is indicated. The pin between *oadB* and *citA* represents a stem-loop structure that presumbly is responsible for partial termination of transcription starting from the *citS* promoter (Figure reproduced from Bott, 1997 [9]).

as shown in Figure 3.1. The transcriptional activator CitB is composed of the conserved N-terminal receiver domain and a C-terminal domain with a helix-turn-helix motif. The citrate fermentation genes form a cluster on the chromosome, composed of *citC* and *citS* operon as illustrated in Figure 3.2. The *citC* operon encodes citrate lyase ligase, the γ-, β- and α-subunits of citrate lyase and a protein presumably involved in the synthesis of 2′-(5″-phosphoribosyl)-3′-dephospho-CoA prosthetic group of citrate lyase [7]. The *citS* operon encodes the Na^+ pump oxaloacetate citrate carrier CitS and the γ-, β- and α-subunits of the Na^+ pump oxaloacetate decarboxylase.

Citrate metabolism under anaerobic conditions

Aerobic citrate metabolism always occurs *via* the tricarboxylic acid cycle (TCA cycle) whereas for anaerobic catabolism of this tricarboxylic acid a different fermentation pathway is used in *Klebsiella pneumoniae* [9]. This involves oxaloacetate decarboxylase rather than reactions of the TCA cycle and is strictly dependent on sodium ions. Expression of the citrate fermentation genes has to be carefully regulated because the synthesis of citrate lyase and oxaloacetate decarboxylase under inappropriate conditions would affect the normal functioning of the citric acid cycle. This might lead either to a futile cycle of citrate synthesis or deprivation of oxaloacetate. On the other hand, oxaloacetate decarboxylase could perturb the Na^+ balance across the cytoplasmic membrane. This aspect is of tremendous importance as described by Bott *et al*, 1995 [8]. They indicated that the expression of the genes of the *citC* and *citS* operons are highly dependent upon the extracellular concentration of citrate and Na$^+$ ions.

Citrate sensing by the bacterium

It has been proposed that the periplasmic sensor domain of CitA (CitAp) recognizes the presence of citrate under anaerobic conditions, resulting in a conformational change, in at least parts of the periplasmic domain. This conformational change is then transferred to the histidine-kinase domain *via* two transmembrane helices and the internal PAS domain. The molecular mechanism by which this signal is transferred from periplasm to the cytoplasm is not clear until now. Biochemical experiments showed that citrate binding to the sensor domain induces autophosphorylation of a conserved histidine residue in the kinase domain. This results in transfer of the phosphate group to an aspartic acid residue in CitB. It is predicted that phosphorylation of CitB, again induces a conformational change resulting in its binding to the intergenic region of the *citC-citS* operon. The citAB genes are the promoter-distal genes of the *citS* operon and are positively autoregulated [8]. The periplasmic sensor domain of CitA (CitAp) binds citrate in a 1:1 stoichiometry with the highest affinity at a pH of 5.7. It does not bind or has less affinity to other tri- and dicarboxylic acids such as isocitrate and tricarballylate etc [57].

Two component systems related to CitA/CitB

There is a whole family of sensor histidine kinases [57], which share similar topology and domain organization with CitA. Some examples of proteins most closely related to CitA of *K. pneumonia* are the CitA/CitB TCS's and sensor kinase DcuS (C_4-dicarboxylate uptake) from *Escherischia coli*. There have been evidences showing that *E.coli* CitA/CitB is not only involved in citrate metabolism but also in the regulation of plasmid inheritance [58]. The DcuS protein together with its response regulator, DcuR, is responsible for the C_4-dicarboxylate-dependent induction of several proteins [41] and is involved in catabolism of several C_4-dicarboxylates [54, 2], fumarate respiration [117, 55] etc. The periplasmic domain of DcuS functions as a receptor for several C_4-dicarboxylates, including succinate, fumarate, malate, tartrate, aspartate and maleate. The structure of the ligand-free periplasmic domain of DcuS, determined by NMR spectroscopy reveals a PAS-fold similar to the CitA periplasmic domain structure [62, 90]. Other relatives include CitS, YdbF and YufL from *Bacillus subtilis*, DctB from *Rhizobium leguminosarum*,

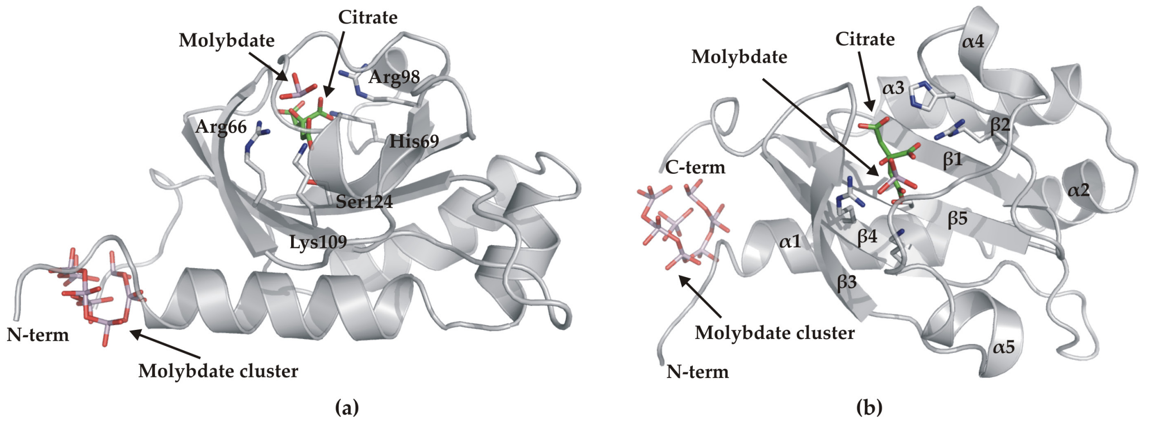


Figure 3.3: Structure of ligand-bound CitAp sensor kinase domain (PDB ID: 1P0Z). (a) Side view, highlighting the molybdate cluster bound at N-terminus, bound citrate and molybdate in the citrate-binding site and residues involved in H-bonding interactions with citrate and molybdate. (b) Top view showing the N- and C-terminus and the secondary structure assignment.

Dcts from *Rhodobacter capsulatus* and Tcp from *Salmonella typhimurium* [57].

All these *citA*-like genes were found to be associated with response regulators that are similar both in amino terminal receiver domain and carboxy terminal DNA binding domain to *K. pneumoniae* CitB, thus suggesting that all the members of this family co-evolved from a common ancestral TCS.

3.1.2 Structure of ligand-bound periplasmic domain of CitA

The structure of the ligand-bound CitAp periplasmic domain has been determined by Reinelt *et al*, 2003 [102] to a resolution of 1.6 Å. In the ligand-bound state, the CitAp domain forms a mixed α/β-structure with a central five stranded antiparallel β-sheet containing residues, 56-61 (β1), 66-68 (β2), 94-100 (β3), 104-110 (β4) and 121-128 (β5) as shown in Figure 3.3. This sheet is flanked on one side by the N-terminus, formed by two long helices α1 (10-24) and α3 (38-51) connected by a third small helix α2 (27-34). The other side of the β-sheet is packed against the long loop region (major loop, residues 63-92) that connects strands β2 and β3 of the sheet and contains a short α-helix (α4, residues 72-76) and a 3_{10}-helix (α5, residues 85-90). The overall fold of the structure resembles a PAS fold. The whole structure as can be seen in Figure 3.3, encloses a concave pocket, which forms the binding site of citrate.

In the published structure, a molybdate (MoO_3) was modelled near the citrate binding site. In this complex, the molybdenum is octahedrally coordinated by the three oxo groups and a tridentate interaction with citrate. This molybdate appears to be a contaminant from the crystallization buffer. A second polymolybdate containing seven molybdenum centres was found binding near the N- and C-terminus. A sodium ion was also modelled (not shown in figure), which is octahedrally coordinated with three water molecules, the main chain carbonyl group of Pro111, and the hydroxyl moieties of Ser110 and Ser24, thereby linking the central β-sheet with the C-terminal end of the α1-helix.

CitAp ligand binding groove

The CitAp ligand binding groove is formed by the helix α4, flanked by the major loop residues namely, 69-71 and 77-84. The binding of citrate is primarily stabilized by hydrophobic interactions between citrate and the residues Tyr56 and Met79 and main chain H-bond interactions of Thr58, Ser101 and Leu102. The majority of citrate-protein interactions are formed by hydrogen-bonds originating from the side-chain atoms of Arg66, His69, Arg107, Lys109 and Ser124 (partially conserved in CitA subfamily). The residues, Arg66, His69 and Arg107 are highly conserved among the CitA subfamily of histidine kinases [39]. Hydrogen bonds to the oxygen atoms of MoO_3 are formed by Gly81, Arg107 and Arg98. Among these residues only Gly81 is partly conserved.

3.2 Aim of Present Work

Signal transduction systems function as intracellular information processing pathways that link external stimuli to specific adaptive responses in the cell [128]. Although there is a great diversity of stimuli and corresponding responses, a relatively small number of molecular strategies are used for signal processing. For example, in this study of *K. pneumoniae* CitAp sensor kinase protein, the signal (presence of citrate) is recognized and processed by a phosphorelay mechanism mediated by phosphorylation of His/Asp residues. The determination of structures of these domains might aid in understanding their function/s. Secondly a number of different infections are cause by *Klebsiella pneumoniae*, medically the most important species of this genus. *Klebsiella pneumoniae* causes community acquired pneumoniae, nosocomial infections (hospital acquired infections), sepsis and is the major cause of urinary tract infections [106]. *Klebsiella* infections are encountered far more often now than in the past. This is probably due to bacteriums's antibiotic resistance properties. So, a thorough understanding of the regulation of various metabolic processes in this species is of great medical importance.

The structure of ligand-bound CitAp [102] sensor kinase domain shed light on the ligand binding mechanism and ligand binding stoichiometry in such systems. The major task of the current work was to determine the structure of the ligand-free state of CitAp sensor kinase domain, to look for any conformational differences between ligand-bound and ligand-free state and thereby gain insight into the mechanistic aspects of signal transfer from periplasm *via* the transmembrane helices to the cytoplasmic portion of the protein.

3.3 Materials and Methods

3.3.1 Crystallization & data collection

Both native and selenomethionine derivatized constructs of CitAp ligand binding domain were expressed and purified by the group of Dr. Stefan Becker, Max Planck Institute for Biophysical Chemistry, Goettingen. Initial trials to identify the crystallization conditions of CitAp were carried out using a Robot. CitAp was crystallized by the hanging drop vapour diffusion method [80] by equilibrating a drop containing a mixture of 2 μl of 15 mg/ml protein in 10 mM HEPES, pH 7.5 and 1 μl of reservoir with a reservoir containing 20 mM HEPES, pH 7.5, 0.63 M NaH_2PO_4 and 0.63 M KH_2PO_4. The crystals were flash frozen using 30% glycerol (equilibrated with reservoir containing 10% glycerol for 2 minutes and then quickly soaked in a reservoir solution containing 30% glycerol) as cryoprotectant under liquid nitrogen stream and data were collected to a resolution of 2.0 Å for native and around 2.8 Å for selenomethionine derivatized crystals. The diffraction pattern was very anisotropic as can be seen in Figure 3.4. The native data were collected to a completeness of 97% using the strategy computed by the program *BEST* [97], taking into account the anisotropy and high mosaicity of the data.

3.3.2 Data processing and detection of twinning

All data were processed with *HKL2000* [88] and the space group and data statistics determined by *XPREP* (Bruker AXS, Madison, WI). The data collection statistics are summarized in Table 3.2. Analysis of data suggested varying amounts of twin fractions in both native and derivative crystals as can be seen from the statistics in Table

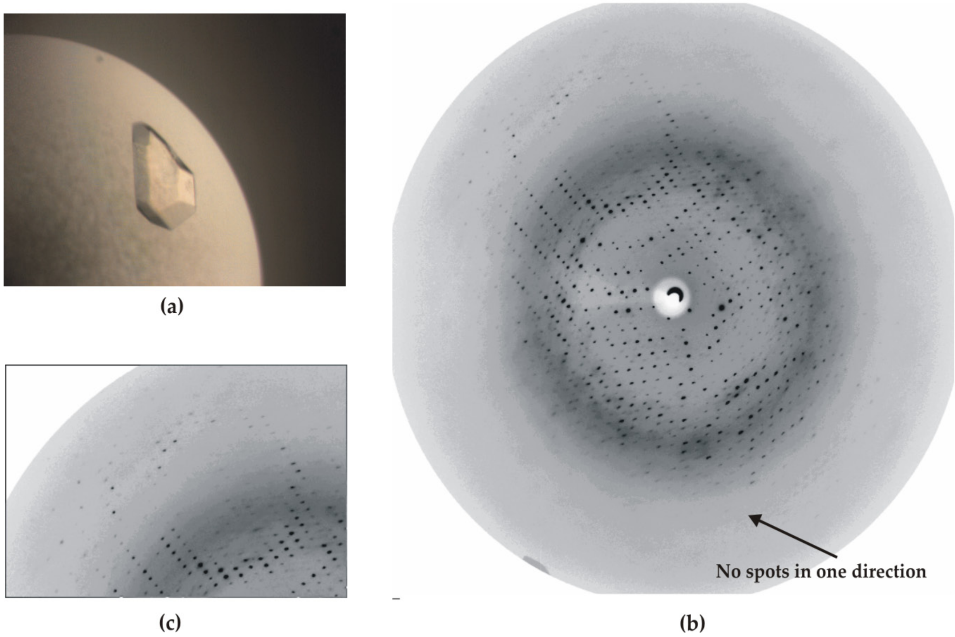


Figure 3.4: (a) Tetragonal crystal of native CitAp sensor kinase domain (b) Anisotropic diffraction pattern (c) Zoomed in image, showing diffraction spots extending in one direction.

3.1 . From Table 3.1, the values for $|E^2\text{-}1|$ of the native and derivative data sets, are far above the expeceted values of 0.968 (centrosymmmetric structures) and 0.763 (non-centrosymmetric structures), before correcting for anisotropy. In ideal cases twinning decreases the $|E^2\text{-}1|$ values, but here because of the presence of NCS and anisotropic data, the values are far above the expected values (as will be seen later with different values of twin fraction obtained after refinement and twin tests). Twin tests carried out using the Yeates & Fam twinning server at UCLA [133] suggested a twin fraction of about 0.2. There were some difficulties initially to decide if the space group was I4 or I422, because twinning in I4 along the 110 axis would emulate the space group I422. Several other tests for twinning were carried out using *DETWIN* [15] as explained below (a summary of the description of basic theories underlying each twin test has been excerpted from Dauter, 2003 [18] in order to describe the behaviour of data in the plots shown in Figure 3.5):

Table 3.1: Test for merohedral twinning in *XPREP*, twin law 010 100 00-1. [1]Native data merged in I4 after correcting for anisotropy in *TRUNCATE*. [2]Native data merged in I4, before correcting for anisotropy. [3]Twin fraction.

Data set	**Point group**	**R_{int}**	**$\|E^2-1\|$**	**BASF [3]**
Native[1]	4/m	0.000	0.530	0.311
	4/mmm	0.165	0.521	
Native[2]	4/m	0.000	1.018	0.139
	4/mmm	0.333	0.962	
Se-Peak	4/m	0.087	1.063	0.249
	4/mmm	0.199	1.043	
Se-Herm	4/m	0.108	1.041	0.144
	4/mmm	0.251	1.012	

Britton plot of 'negative intensities'

Intensities measured from a twinned crystal are a combination of intensities from the twin law related domains weighted by their relative volumes

$$I_1 = \alpha J_1 + (1-\alpha) J_2 \tag{3.1}$$

$$I_2 = (1-\alpha) J_1 + \alpha J_2 \tag{3.2}$$

where I_1 and I_2 are measured intensities from a pair of twin related reflections, J_1 and J_2 are the corresponding detwinned intensities and α, the twinning fraction. If α is smaller than 0.5, these equations can be solved to get the set of detwinned intensities J_1 and J_2:

$$J_1 = \frac{(1-\alpha) I_2 - \alpha I_1}{(1-2\alpha)} \tag{3.3}$$

$$J_2 = \frac{(1-\alpha) I_1 - \alpha I_2}{(1-2\alpha)} \tag{3.4}$$

According to Britton [11], the estimation of a higher value of α results in an estimation of too negative true intensities. The 'Britton plot', giving the number of negative intensity estimation as a function of twin fraction, α in the detwinning procedure has two linear asymptotes, one for $\alpha < \alpha_{opt}$ and another for $\alpha > \alpha_{opt}$. The point at which these two lines meet gives the estimated value of twin fraction. In Figure 3.5c, the asymptotic

Table 3.2: Data collection statistics of native, peak and high energy remote data sets of CitAp sensor kinase domain. Values in parenthesis are for the outer resolution shell. Both peak and high energy remote data sets were collected on highly mosaic crystals as can be seen from the statistics for mosaicity.

Data statistics	**CitAp-Native**	**CitAp-Peak**	**CitAp-Herm**
Wavelength (Å)	1.05	0.97910	0.9793
X-ray source	BW6	BW6	BW6
Detector	MARCCD	MARCCD	MARCCD
Space group	I4	I4	I4
a (Å)	64.163	63.973	64.050
b (Å)	64.163	63.973	64.050
c (Å)	144.746	145.512	145.511
$\alpha = \beta = \gamma$ (°)	90	90	90
Mosaicity (°)	1.5	2.2	2.2
Resolution (Å)	2.0 (2.10-2.00)	2.85 (2.95-2.85)	2.82 (2.95-2.82)
Reflections	19151	6248	11097
Redundancy	5.49 (1.60)	12.12 (2.99)	1.38 (0.56)
Completeness (%)	97.0 (84.1)	91.1 (40)	79.6 (34.6)
Mean I/σ(I)	16.70 (2.77)	22.32 (5.76)	7.81 (2.52)
R_{int}(%)	4.77 (14.58)	8.66 (18.13)	4.80 (14.58)

straight lines for CitAp native data extrapolate to a twin fraction of about 0.2.

Murray-Rust F_1/F_2 plot

The twinning equations can be rearranged to give the ratio J_2/J_1,

$$\frac{J_2}{J_1} = \frac{I_2 - \alpha(I_1 + I_2)}{I_1 - \alpha(I_1 + I_2)} \tag{3.5}$$

since both J_1 and J_2 are positive, neither numerator nor denominator can be negative, which can be expressed in terms of structure factor amplitudes by

$$\sqrt{\frac{\alpha}{1-\alpha}} < \frac{F_2}{F_1} < \sqrt{\frac{1-\alpha}{\alpha}} \tag{3.6}$$

This property was proposed by Murray-Rust [85] for estimation of the twin fraction. All points in the graph between F_1 *vs* F_2 should lie between the two limiting straight lines corresponding to $[\alpha/(1-\alpha)]^{1/2}$ and $[(1-\alpha)/\alpha]^{1/2}$. The slope of the line bounding the two points can be used to estimate the twin fraction. From Figure 3.5d, again the approximately calculated slopes of 4 and 0.5 indicate a twin fraction of 0.2.

Yeates *S(H)* plot

This test proposed by Yeates [130, 131] is based on the behaviour of the ratio of the difference to the sum of intensities of reflections related by the twin law,

$$H = \frac{|J_1 - J_2|}{(J_1 + J_2)} \tag{3.7}$$

The cumulative *S(H)* distributions for non-centrosymmetric reflections is

$$S(H) = \frac{H}{1 - 2\alpha} \tag{3.8}$$

Thus the dependence of the cumulative distribution of this parameter *S(H)* is linear in *H* for non-centrosymmetric structures. The slope of *S(H)* plot is $1/(1-2\alpha)$ and depends on twinning fraction as can be seen from a value between 0.2 and 0.25 in Figure 3.5a.

But all these statistical tests deliver some biased results when the asymmetric unit has translational Non Crystallographic Symmetry (NCS) elements, i.e. when there is more than one molecule per asymmetric unit or when the data are anisotropic. In the present structure translational NCS is present and the data are very anisotropic. The

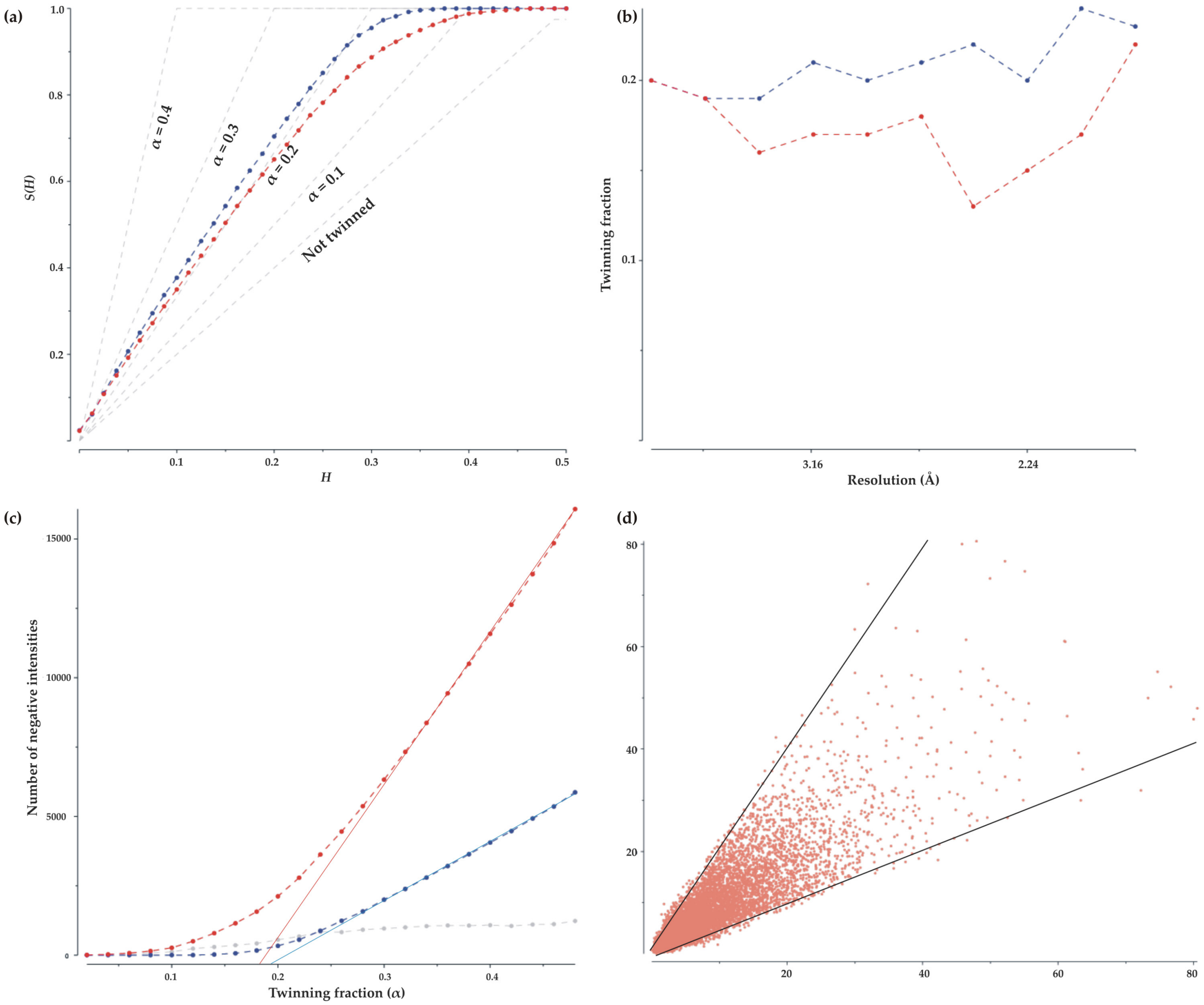


Figure 3.5: (a) Yeates *S(H)* plot indicating a twin fraction of about 0.2 (b) Twin fraction as a function of resolution calculated from CCP4 *DETWIN* (c) Britton plot of number of negative intensities resulting from detwinning procedure (d) Murray-Rust F_1/F_2 plot, the slopes of the lines enclosing the sector indicate a twin fraction of 0.2. In all graphs blue lines are for all data and red lines denote data, where intensities are higher than 4σ.

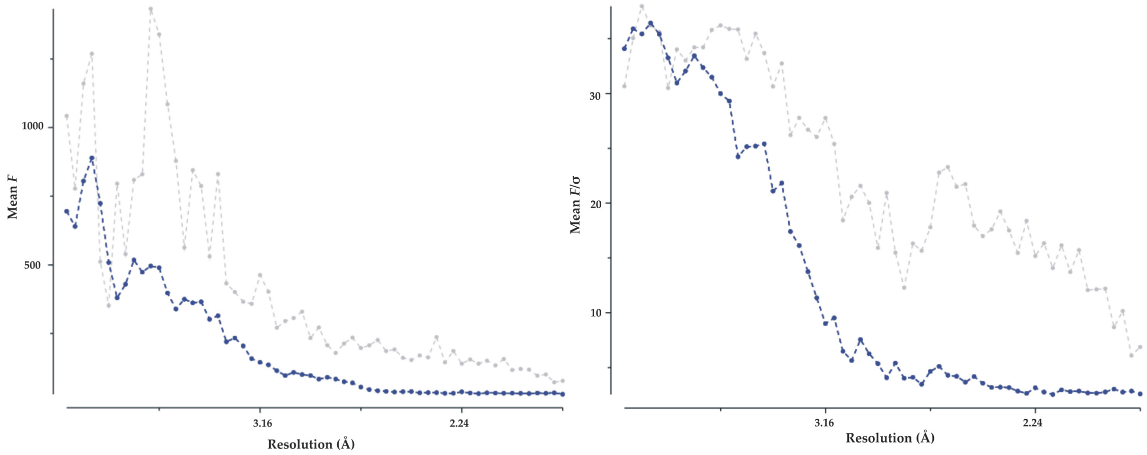


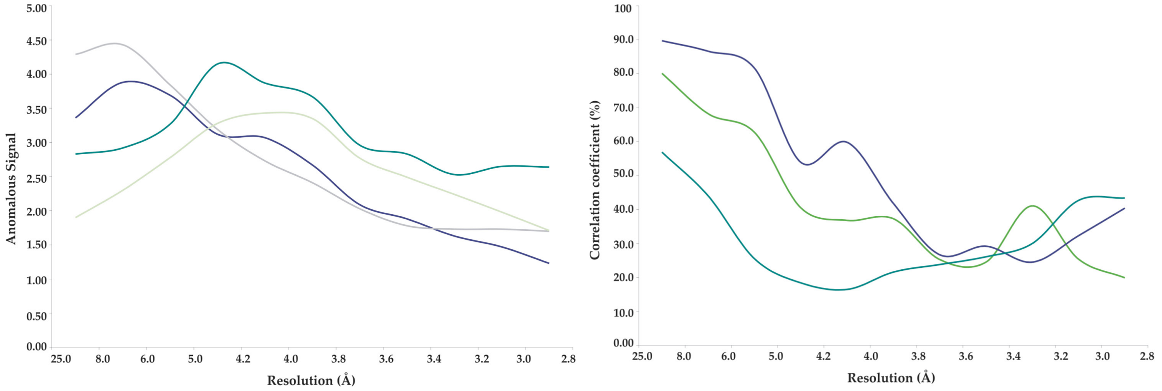
Figure 3.6: Falloff of mean F and mean F/σ values as a function of $(\sin\theta/\lambda)^2$ in three orthogonal directions indicated by grey, blue and red lines (in this case both red and blue lines overlap because of tetragonal space group). The different falloff rates in two directions indicate the degree of anisotropy.

anisotropy of the data can be seen from the analysis of degree of anisotropy in all three directions (Figure 3.6) using the program *FALLOFF* incorporated in *TRUNCATE* [35]. Thus, the values for twin fraction obtained using these statistical tests and the one after final refinement do not coincide. The native data corrected for anisotropy and without any detwinning was used for structure solution using molecular replacement.

3.3.3 Structure solution, bias minimization and model building

Both peak and high energy remote data sets were noisy, partly because of the even higher mosaicity of 2.2 when compared to native data sets. Plots of anomalous signal *vs* resolution and correlation coefficients of signed anomalos differences against resolution are presented in Figure 3.7a & b. Trials to solve the substructure by truncating the data to a resolution of 3.9 Å were not successful.

Therefore, the structure was solved by molecular replacement using *PHASER* [79]. The citrate bound form of CitAp domain (PDB ID: 1P0Z) was used as search model to look for two molecules in the asymmetric unit, as indicated by a Matthews coefficient of 2.48 for two molecules in the *asu*. The minor loop and major loop residues are involved in ligand binding and differences in this region were expected upon ligand binding/unbinding, and so were removed from the search model. To minimize bias, in-



(a)

(b)

Figure 3.7: (a) Anomalous signal plotted against resolution for peak data (blue and grey) and high energy remote (green and light green) (b) Signed anomalous correlation coefficients plotted against each resolution shell (light green), peak (blue), high energy remote (dark green).

fluenced by the search model, residues in the N-terminus (5-10), C-terminus (128-135), minor loop (99-104) and major loop (68-90) were chopped out from the search model. The structure solved with a Z-score of 34 and a log-likelihood gain of 1237.62.

Bias minimization & model building

After an initial rigid body refinement in *REFMAC* [86], positive density (in $m\boldsymbol{F_o}$-$D\boldsymbol{F_c}$ maps) appeared near the N-terminus of both monomers and was modelled as two small helical regions. At this point the R and $\mathrm{R_{free}}$ values before modeling the N-terminus were 42.3 and 43.8 respectively. A σA-weighted map calculated from this model by the program *SIGMAA* [101] was input to *RESOLVE* [121, 122] to generate bias-minimized maps using prime and switch phasing. 75% of the model was rebuilt by automated model building in *RESOLVE* and model correction by hand from these bias minimized maps. The R and $\mathrm{R_{free}}$ values at this point were 34.78 and 39.61 respectively. A composite omit map was calculated using simulated annealing in *CNS* [12]. Visible residues of the sequence near both minor and major loops were docked to the MR model using the composite omit map calculated in *CNS*. Before starting with refinement in *CNS*, $\mathrm{R_{free}}$ reflections were assigned using a special protocol in *CNS* taking the twin law into account, such that pairs of twin related reflections are in the same set. An initial refinement using torsion angle dynamics was carried out taking into account a twin fraction of 0.2, firstly to reduce the bias towards the previous test set and secondly to improve the geometry of the model. The model was refined using the twin refinement protocol in *CNS* by simulated annealing and energy minimization protocol.

In CNS, twinning is directly incorporated into the refinement target. The least squares crystallographic target used is:

$$TARGET = \left\{F_o - k\sqrt{(1-\alpha)(F_c)^2 + \alpha(F_c')^2}\right\}^2 \tag{3.9}$$

where α is the twinning fraction and F'_c denotes application of the twinning operator. The scale factor k between observed and calculated data remains fixed during a refinement cycle and is only recalculated when the value for the weight between crystallographic and geometric terms is updated. The twin fraction α also remains fixed during refinement. The structure was further refined with *SHELXL* because the twin fraction cannot be refined in *CNS*. The R and R_{free} values at this stage were 28.15 and 33.31 respectively.

3.3.4 Refinement and structure validation

The model was refined using the twin refinement protocol in *SHELXL* as explained in Chapter 2. Similar to the refinement of non-merohedral twins, *SHELXL* instructions with an initial twin fraction (BASF) of 0.2 along with the twin law (TWIN 010 100 00-1) were used for merohedral twin refinement [48]. Refinement statistics are tabulated in Table 3.3. Tighter restraints were applied to isotropic B-values. The structure was refined alternating with model building in real space using $2\boldsymbol{F_o}$-$\boldsymbol{F_c}$ and $\boldsymbol{F_o}$-$\boldsymbol{F_c}$ maps in *COOT* [25]. Waters were modelled using *ARP/wARP* [93] and *COOT*.

Occupancy refinement of loops in *SHELXL*

Occupancy of minor and major loops were refined using free variables (FVAR) in combination with SUMP instruction. The SUMP instruction was used such that the sum of the differences of occupancies of the neighbouring amino acids is zero with a standard deviation of 0.05. This would result in a smoother distribution and a gradual increase in the B-values as the loops approach the solvent region and get very mobile. A bumpy curve of the occupancies (B-values) of the residues plotted against residue number in loop regions can be seen as in Figure 3.8. The bumpy nature of the curve can be explained by the stabilization of the residues in some regions of the loop, either by interactions with residues in the protein itself or interactions with symmetry equivalent molecules. This

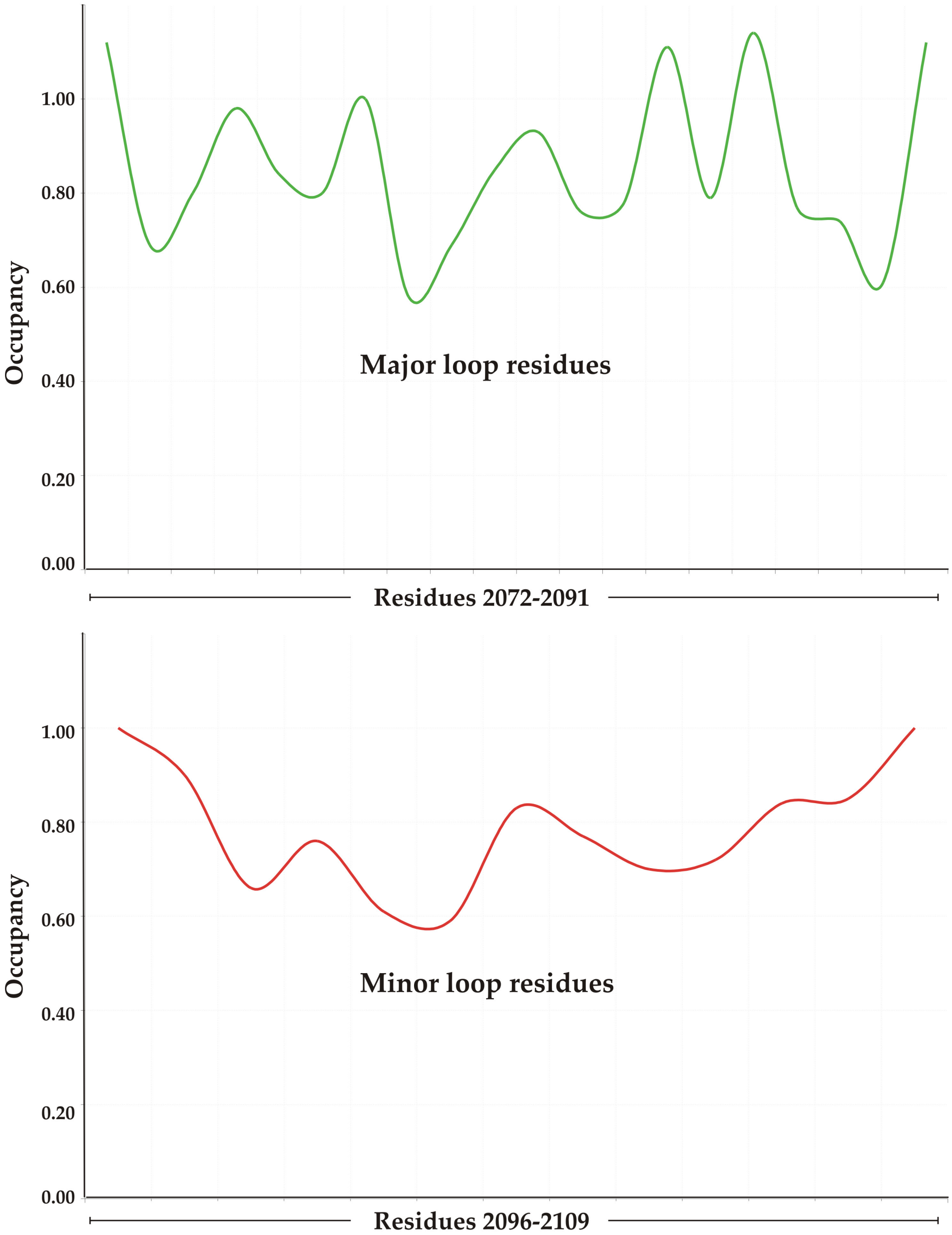


Figure 3.8: Occupancy of each of the residues constituting the major and minor loops in ligand free CitAp sensor kinase. Some of the residues in the loop regions are stabilized by hydrogen bond and hydrophobic interactions with residues in the protein itself or symmetry related molecules thus reflecting the bumpy nature of the occupancy of these loops.

Table 3.3: Refinement statistics of CitAp sensor kinase domain. 5% of reflections in thin shells were taken as test data set. The final refined BASF value indicates a twin fraction of 0.42.

Refinement statistics	**CitAp-native**	**Refinement statistics**	
Final R-factor (%)		**Mean B value** (Å^2)	
$F_o > 4\sigma(F_o)$	19.49	**Monomer 1**	
For all data (unique)	19.61	Main chain atoms	39.510
		Side chain atoms	37.736
Final Free R-factor (%)		**Monomer 2**	
$F_o > 4\sigma(F_o)$	26.81	Main chain atoms	40.951
For all data (unique)	26.91	Side chain atoms	41.048
		Solvent	44.37
R.m.s deviations		**No. of protein atoms**	1783
Bond lengths (Å)	0.006	**No. of solvent atoms**	106
Bond angles (Å)	0.026	**BASF**	42.04

novel approach of refining the occupancies of individual residues, should be a better way of defining the mobility/stability of loops in proteins, and has more physical sense (as it is a more direct reflection of the measured data) rather than describing it by higher or lower B-values of the corresponding residues.

The final model was validated with *PROCHECK* [67]. 60.8% of residues lie in the allowed regions, 37.73% in additionally allowed regions, 8.5% of residues in generously allowed regions and 1.9 % (4 residues) in disallowed regions of the Ramachandran plot.

3.3.5 Inverse fouriers using *SHELXE*

The data from Se-derivatized crystals were noisy, primarily because of the high mosaicity. For this reason, phases from molecular replacement solution and normalized difference structure factors from the Se-data set (peak data set only) were used to back calculate the positions (re-refine positions) and phases of seleniums using inverse fourier transformation in *SHELXE* [114]. Six selenium positions were identified as can be seen in Figure 3.9. Phase extension trials using selenium substructure and native data produced fragmented maps (data not shown), probably because of the incomplete data as can be seen in Table 3.2. These data would require careful processing for further density modification trials.

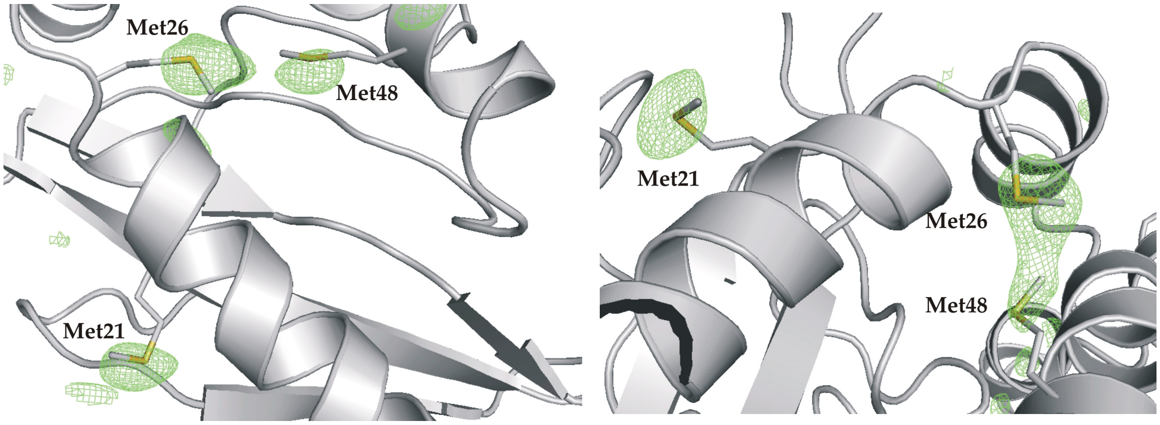


Figure 3.9: Anomalous map contoured around Selenium sites in Met21, Met26 and Met48 at 3σ in both monomer A and monomer B.

3.4 Results

3.4.1 Structure of the ligand-free periplasmic domain of CitA

The structure of the ligand-free periplasmic domain of CitA has been determined to a resolution of 2.0 Å. CitAp crystallized as a dimer in the asymmetric unit in tetragonal space group, I4. Both monomers in the asymmetric unit differ from each other, in that parts of sequences could not be traced in each of them. Figure 3.11 shows a superposition of monomer A (monA) on monomer B (monB). In monA, the N-terminus could be traced from residue 1 and the C-terminus upto residue 131, whereas in monB, the N-terminus could only be traced from residue 3 and the C-terminus upto residue 128. The residues involved in citrate binding in the citrate bound structure lie in two major helices, namely major loop residues (63-92) and minor loop residues(96-105). In both monA and monB, the minor loop could be traced completely whereas in monA, a part of the major loop residues (76-85) could not be traced, as can be seen in Figure 3.11 & Figure 3.10.

Ligand-free CitAp has a similar fold (a mixed α/β structure) when compared to ligand-bound structure. The central β-sheet is formed by β1 (55-62), β2 (64-70), β3 (94-100), β4 (103-113) and β5 (120-130). Quite similar to the ligand-bound structure, this central β sheet is flanked on one side by two α helices, α1 (10-24) and α3 (38-52) joined by another helix α2 (27-36). In contrast to the ligand-bound structure, the N-terminus is helical (denoted as α' (3-8)) in both monomers. Major differences lie on the ligand

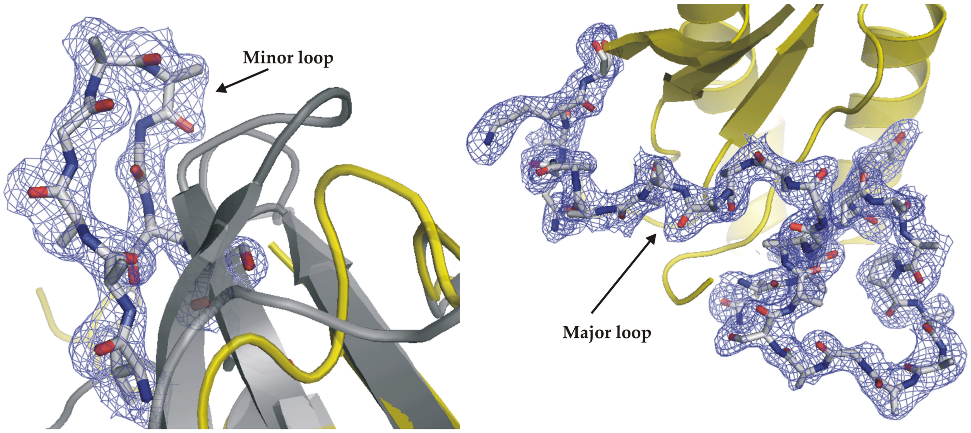


Figure 3.10: Final refined map from *SHELXL* after occupancy refinement contoured at 1σ around residues constituting (a) minor loop (b) major loop in monB.

binding side of the central β sheet, where both helix $\alpha 4$ and $\alpha 5$, present in the citrate bound structure are lost and a small β strand (β' (78-80)) appears.

Least squares superposition of the main chain atoms of monA on monB gave a r.m.s.d, (root-mean-square-deviation) of 1.07, which indicates that both monomers are different in the very mobile loop regions of the structure. NCS restraints were not used during refinement or model building (while calculating composite omit maps) for the same reason. In the published structure of the ligand-bound CitAp, two kinds of dimers were characterized namely, EG-type dimers and GJ-type dimers (named after the chain ID) as shown in Figure 3.12b & c. These dimer interfaces bury a total of 1797.0 Å^2 and 1414.6 Å^2 of accesible surface for EG and GJ-type dimers respectively. The EG-type dimers appear to have stronger interactions, where most of the hydrogen bond interactions lie between 2.5 to 3.0 Å mediated by Glu6-Ser93′, Gly82-Lys99′, Asp85-Lys99′, Glu86-Ser104′ and Arg98-Gly81′. Whereas, hydrogen bond interactions in GJ-type dimers are comparitively weaker, lying between 3.0 to 3.4 Å and are mediated by Glu11-His134′, Arg15-Phe51′, Gln22-Ser50′, Phe51-Gln11′, Pro47-Gln22′, Phe51-Phe51′, Phe51-Gln22′ and Glu133-Gln11′.

In ligand-free CitAp structure the dimer in the asymmetric unit has a buried accessible surface area (ASA) of 637.6 Å^2 (since the two monomers are different, the buried ASA, calculated considering two molecules of monA is 1005.1 Å^2 and calculated considering two molecules of monB is 270.1 Å^2) due to dimerization. The dimer interface interactions are mediated by hydrogen bonds between main chain and side chain

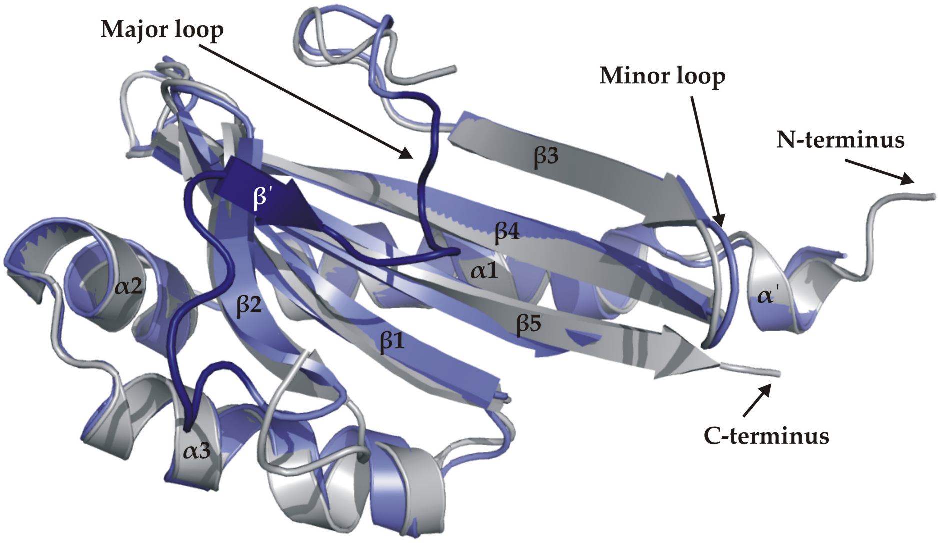
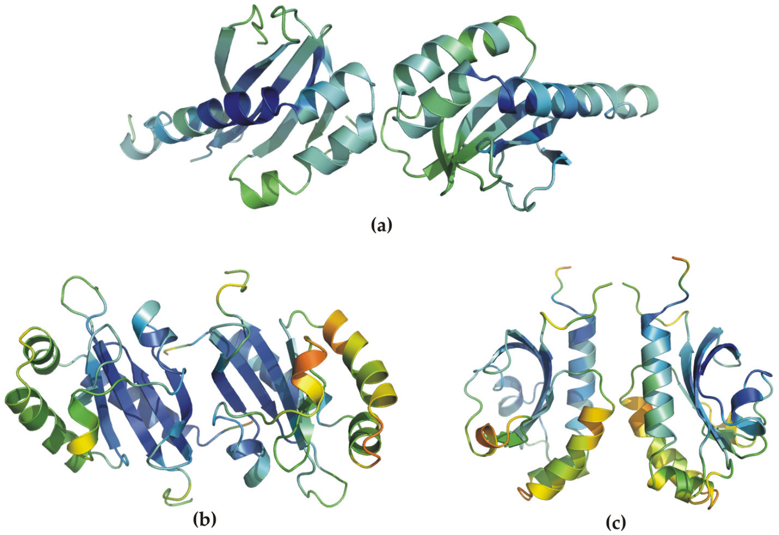


Figure 3.11: Superposition of monomer A (grey) on monomer B (lightblue). Secondary structure elements, N- and C-terminus are labelled. Major loop (dark blue) missing in monomer A.

atoms of residues Lys35-Arg36′, Lys35-Asp37′, Lys35-Lys35′ and Lys35-Ala39′ and various hydrophobic interactions. Since a large number of residues in the major loop region of monA could not be modelled and less than 800 Å^2 of accesible surface area (ASA) is buried, this dimerization interface is not conclusive [96]. An alternative oligomerization interface with symmetry equivalent molecules has been characterized and is discussed in detail in the next section.

3.4.2 Biologically relevant tetramer

Each of the monomers (monA and monB) in the asymmetric unit, forms a tetramer with symmetry equivalent molecules upon application of crystallographic 4-fold rotation axis as can be seen in Figure 3.13a & b. Each of the tetramers has a buried ASA of 6402.4 Å^2 and 4211.4 Å^2 for monA and monB respectively. The tetramer interactions are mediated by hydrogen bond and hydrophobic interactions (lying between 2.6-3.5 Å) between residues Gln11-Gln14′, Gln22-Thr128′, Pro27-Thr55′, Glu28-Arg49′, Phe51-Phe51′, Asp53-Gln22′ and Thr55-Gln22′ in monB tetramer and interactions (lying between 2.35-3.5 Å) between residues Glu5-Gln14′, Gln11-Gln14′, Met21-Glu30′, Gln22-



(a)

(b)

(c)

Figure 3.12: (a) Ligand-free CitAp dimer in the asymmetric unit with a buried ASA of 637.6 Å^2 (b) EG-type dimer in ligand-bound CitAp structure with a buried ASA of 1797.0 Å^2 (c) GJ-type dimer in ligand-bound CitAp structure with a buried ASA of 1414.6 Å^2. All dimers are colored according to B-values. In the ligand-free structure tighter restraints were applied to B-values and so the distribution of B-values cannot be compared between different kinds of dimers.

Asp53′, Gln22-Thr55′, Gln22-Thr128′, Pro27-Thr55′, Glu28-Asn71′, Phe51-Phe51′, Asn71-Glu28′, Ser93-Gln131′, Ser110-Glu130′, Pro111-Glu130′ and Tyr127-Gln22′ in monA tetramer.

As can be seen in Figure 3.13, a hydrophobic core formed by four phenylalanines (Phe51), with an interaction radius of about 3.3 Å is present. This kind of hydrophobic core plays a major role in the stability of the tetramer. This interaction was also identified to be a major stabilizing interaction in GJ-type dimer of the citrate-bound structure. The published crystal structure of a histidine kinase from *Thermotoga maritima* (PDB ID: 2C2A), also seems to have this kind of hydrophobic core stabilizing the histidine kinase dimer formed with symmetry equivalents (Figure3.14). But, so far all the hypothesis about sensor kinases indicate that the functionally active/inactive state is dimeric. The dimeric nature of CitAp is further confirmed by gel filtration and dynamic light scattering experiments (data not shown). Whether this interaction that we see in the crystal is forced by crystal packing or if the citrate unbound form is tetrameric under physiological conditions would require further investigation.

The surface electrostatic potential of the tetramer was calculated and the potential contoured between 4 kT/e and -4 kT/e. The top view or the periplasmic face (Figure 3.15a) and the bottom view or the N/C-terminus (Figure 3.15c) do not show any charac-

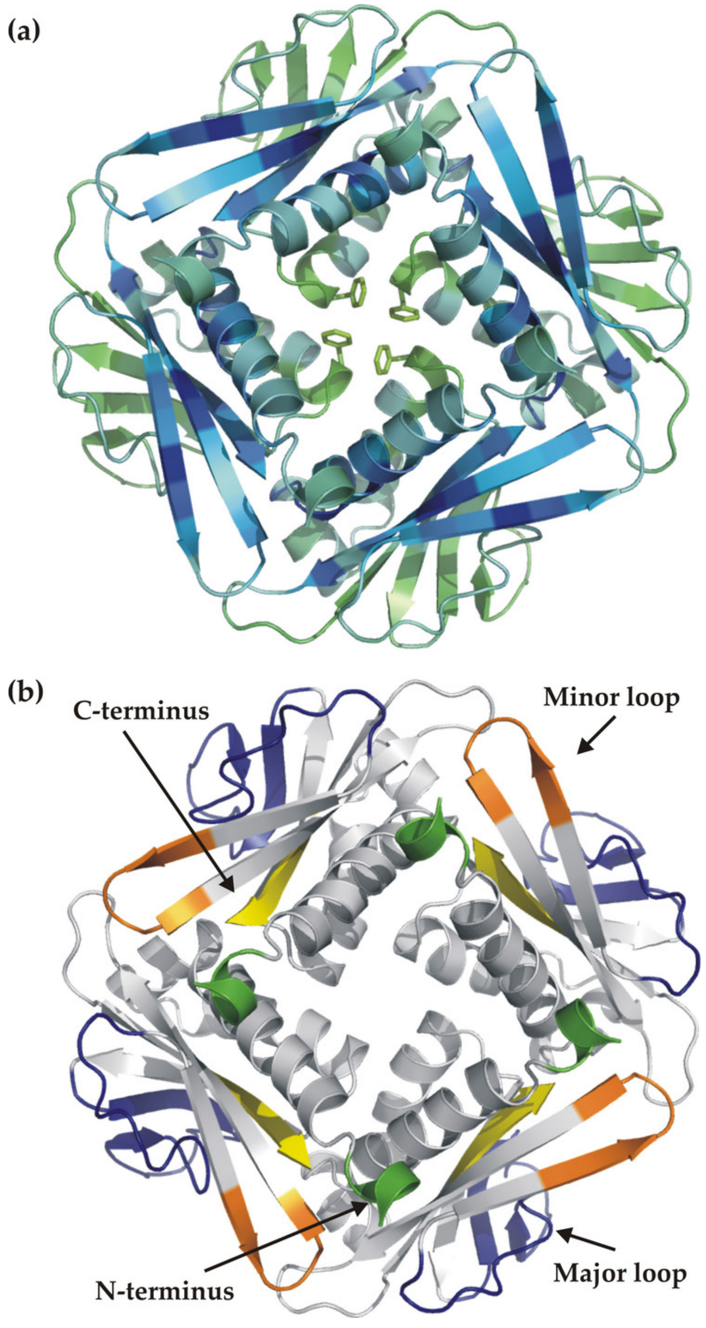


Figure 3.13: CitAp tetramer formed with symmetry equivalent molecules upon application of crystallographic 4-fold axis (a) N- and C-terminal face of CitAp highlighting the four Phe51's in the central core, colored according to B-values (b) Periplasmic face of CitAp, the N-terminal residues are colored green, C-terminal residues are colored yellow, the major loop residues are colored blue and the minor loop residues are colored orange.

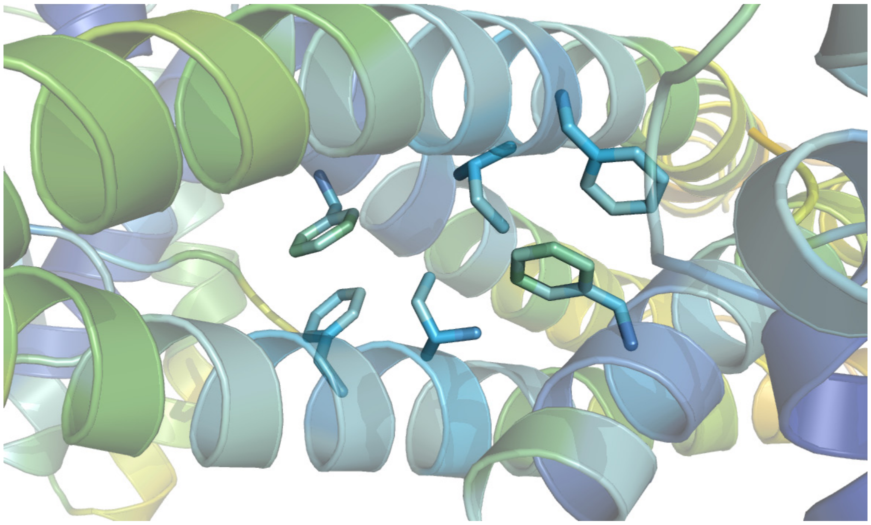


Figure 3.14: The hydrophobic core formed by isoleucine (Ile255) and phenylalanines (Phe254 and Phe312) in the published crystal structure of histidine kinase of a sensor kinase (PDB ID: 2C2A).

teristic features. The citrate binding site is highlighted on the side view of the tetramer surface, which is more basic (Figure 3.15b).

3.4.3 Citrate-bound and citrate-free forms

Least squares superposition of main chain atoms of monA and monB on citrate-bound CitA (PDB ID: 1POZ) gave a r.m.s.d. (root-mean-square-deviation) of 1.562 and 1.237 respectively. As shown in Figure 3.16b, one can observe major differences both in minor loop and major loop regions in the ligand-free form of the CitAp domain. The minor loop has moved away from the citrate binding site by an angle of about 45°. The major loop has lost both the helices α4 and α5 and a small β-strand is found in this loop region. Part of the major loop is displaced into the citrate binding site as can be seen in Figure 3.16a, where citrate is colored in violet and the displaced loops of the ligand-free form colored in yellow. As also described earlier, a part of the N-terminus has turned helical. In the ligand-bound form, the N-terminus is more an undefined loop, but this could be because of the binding of the polymolybdate near this region.

Not all residues involved in interactions with citrate were traced in both monomers. In monA, both main chain and side chains of residues Tyr56, Thr58, Ser101, Arg66, His69, Arg107, Lys109 and Ser 124 were traced, residues Leu102, His69 and Met79 could

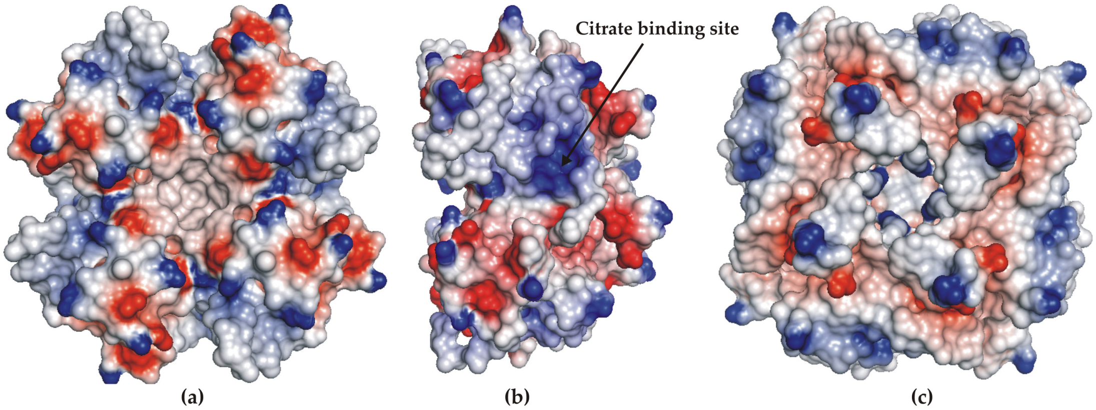


Figure 3.15: Electrostatic potential contoured between -4 kT/e (red) and 4 kT/e (blue). White represents contouring at 0 kT/e. (a) Periplasmic side of sensor kinase domain (b) Side view of tetramer showing the citrate binding site (c) N-terminal side of sensor kinase domain.

only be modelled till the β-carbon atom. In monB, both main chain and side chains of residues Tyr56, Thr58, Ser101, Arg107, Lys109 and Ser124 were traced, the residues Met79, Leu102, Arg66 and His69 could only be modelled till the β-carbon atom. In Figure 3.17, the relative orientation of the side chain traced amino acids in the ligand-free form (grey sticks) is compared with respect to the orientations in the ligand-bound form (green sticks). Major differences can be observed in orientations of Ser101, Arg66 and Lys109.

3.5 Discussion and future perspectives

Molecular breathing upon ligand binding

From the ligand-free crystal structure of CitA, major conformational changes in the secondary structure elements have been observed when compared to the ligand-bound state. To recapitulate, the minor loop moves away from the citrate binding site by approximately 45°, the major loop lost both its helices and parts of the major loop are displaced into the citrate binding site. Parts of the minor loop residues facing the C-terminal side form hydrogen bonds with the residues at the C-terminus, which in turn is involved in hydrogen bonding and hydrophobic interactions with N-terminal residues in the same molecule.

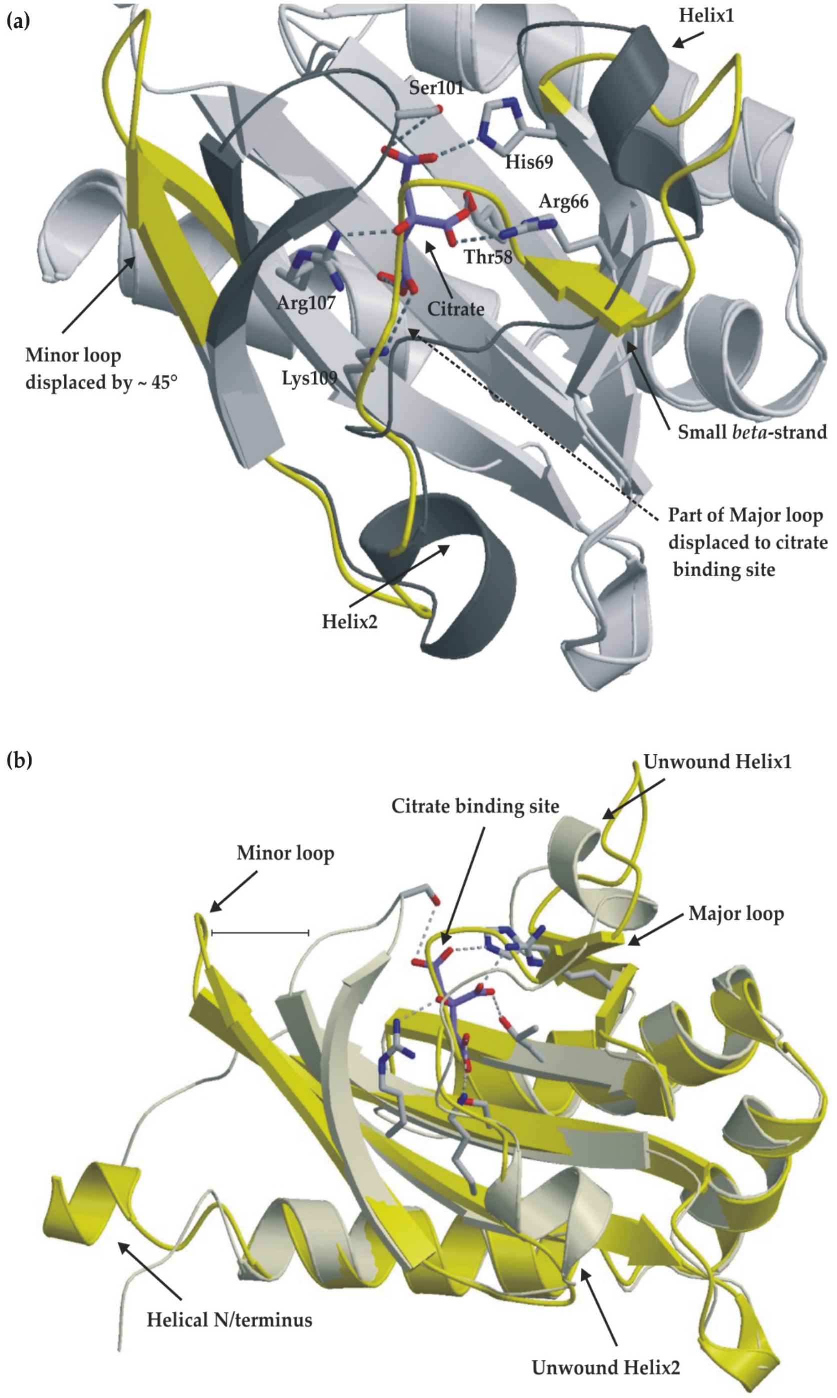


Figure 3.16: (a) A closer look at the citrate binding site of the closed and open conformations (b) Superposition of monB colored in yellow (ligand-free conformation) on 1P0Z colored in grey (ligand-bound conformation).

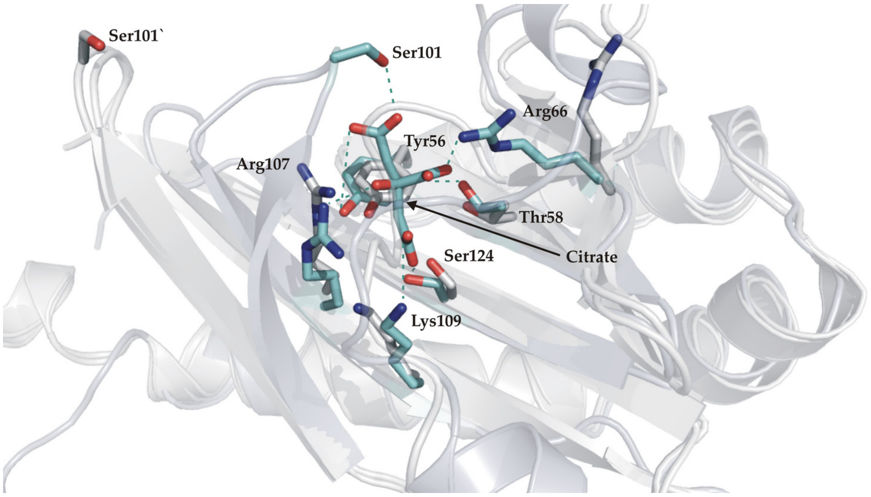


Figure 3.17: Active site residues in ligand-bound (green sticks) and ligand-free (grey sticks) forms of CitAp sensor kinase.

Another major difference between the citrate bound and citrate unbound form is seen in the N-terminal region, where residues in the amino terminus of the ligand-free form is helical and in the ligand-bound form is forming a loop. Thus it might be postulated that upon ligand binding the minor loop, which is nearer to the C-terminus in ligand-free form and makes hydrogen bonds with some of the C-terminal residues, moves away from the C-terminal region. Thereby parts of the helical N-terminus would lose their secondary structure. This conformational change in the N-terminal transmembrane helices might then be sensed by the internal PAS domain *via* the transmembrane helices. The question of how these subtle conformational changes result in a major signal transduction event within the cytoplasm may be answered by comparing the behaviour of the aspartate receptor, where the coupled enzyme can detect small changes in the receptor conformation by a piston mechanism of transmembrane signalling [89].

Stoichiometry of HPKs

As can be seen in the tetrameric structure of the ligand-free CitAp (Figure 3.13), most of the tetrameric interactions are mediated by the residues in the minor loop. These residues move in a circular swinging fashion around the circumference of the tetramer. The asymmetric unit has a dimer, with a buried ASA of approximately 637.6 Å^2. Each of the monomers forming this dimer in turn forms a tetramer, lying in different planes as exemplified in Figure 3.18. It can be hypothesized that this kind of higher oligomeric

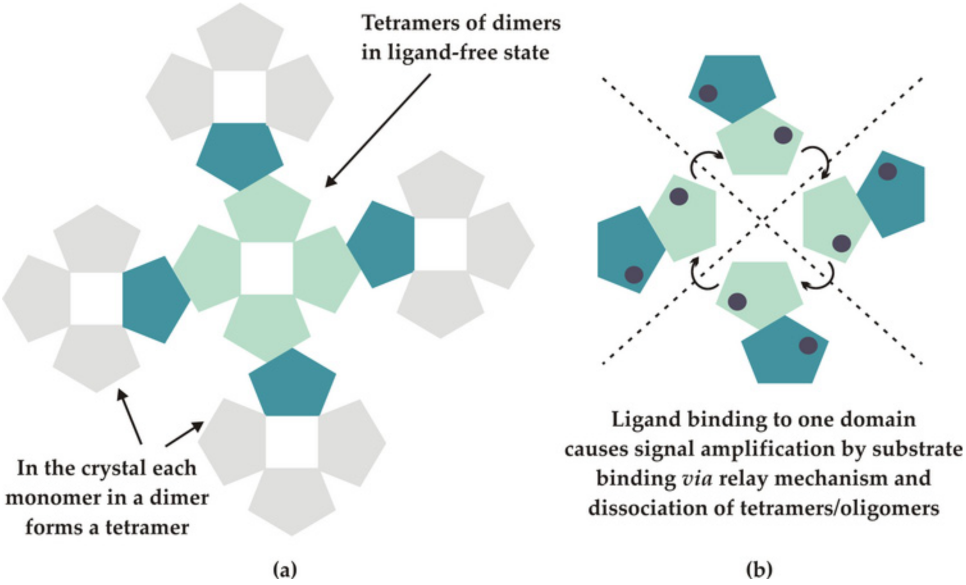

Tetramers of dimers
in ligand-free state
In the crystal each
monomer in a dimer
forms a tetramer
(a)
Ligand binding to one domain
causes signal amplification by substrate
binding *via* relay mechanism and
dissociation of tetramers/oligomers
(b)

Figure 3.18: (a) Tetramer of dimers formed by the crystal packing interactions in CitAp ligand-free structure. (b) Dissociation of the tetramers upon ligand binding.

state in the ligand-free form might be required for efficient signal transfer upon ligand binding, which might occur cooperatively by movement of the minor loop towards the ligand binding site, thereby forming individual dimers. This kind of cooperativity might aid in signal amplification and faster reaction of bacterial cells to environmental signals [30].

Although these conformational changes and oligomeric state could account for the initialization of signal transduction mechanisms, there is a need for the determination of the three dimensional structure of the internal cytoplasmic PAS domain of the CitA sensor kinase to bridge the gap between signal sensing (by periplasmic domain) and signal processing (by histidine kinase domain). Moreover the structure of ligand-bound CitA contained large amounts of molybdate both near the citrate binding site and at the N-terminus, which in turn might not give a logical view of the ligand-bound sensor under physiological conditions, because most of the biochemical experiments confirm that apart from Na^+ ions no other ions are required for signal transduction by CitA sensor kinase. This would also urge a need for redetermination of the ligand-bound CitAp structure without heavy metals in the crystallization conditions and the structure of the whole transmembrane sensor kinase.

Chapter 4

PDI-related Chaperones

4.1 Introduction

Endoplasmic reticulum (ER) is the site of synthesis of various soluble proteins (fully translocated across the ER membrane and released into the ER lumen), transmembrane proteins (partly translocated and embedded in the membrane) and lipids and is the major storage organelle for Ca^{+2}. These proteins and lipids are secreted either to the cell exterior or other organelles like the ER itself, the Golgi apparatus, lysosomes, endosomes, secretory vesicles, and the plasma membrane. Thus, it offers an optimal folding environment for all newly synthesized proteins by a variety of chaperones and folding factors like calnexin/calreticulin, the Hsp70 homologue Immunoglobulin heavy chain binding protein (BiP), the protein disulfide isomerase (PDI), and PDI-related family of redox enzymes and chaperones. The ER sends many of its proteins and lipids to the Golgi apparatus. The Golgi apparatus dispatches these proteins and lipids to a variety of destinations. Misfolded proteins during synthesis on ER are exported from the ER and degraded in the cytosol by the proteasome.

PDI and PDI-related proteins

PDI is a member of the thioredoxin superfamily. They are known for their various redox and chaperone activities, calcium homeostasis and regulation of protein export from the ER for degradation [32]. In this context, PDI is a catalyst of the rate limiting steps of disulfide bond formation, isomerization and reduction within the ER [33]. They are usually homodimers and contain a C-terminal -KDEL- retrieval sequence [33]. Thioredoxin is a small ubiquitous protein of approx, 12 kDa. They are characterized by the presence of the -Cys-X-X-Cys- thioredoxin box and a thioredoxin fold, *viz* β-α-β-α-β-α-β-β-α. Many proteins catalysing the redox reactions involving dithiols *in vivo* belong to this family [32].

Based on the domain structure, the PDI's can be divided into classical PDI-like, containing redox active or inactive thioredoxin domains and unique PDI-like members containing other domains in addition to thioredoxin domains. Some examples of classical PDI-like proteins include PDI-α, PDI, PDIp, ERp57, PDIr, ERp72, ERp5, ERp18, ERp46, PDI-β and calsequesterin. The PDI-unique family consists of PDI-D group of proteins. The PDI-D group of proteins is characterized by the presence of one or two N-terminal domains about 100-120 residues in length called the thioredoxin domain (b-domain) followed by an all-α-helical C-terminal domain about 110 residues in length termed the D-domain [32]. Some examples of PDI-D family members are Wind protein from *Drosophila*, human ERp29, rat ERp29 *etc*. The three dimensional structural analysis of functional and nonfunctional mutants of *Drosophila* Wind and preliminary structural analysis of human ERp29, belonging to this family are studied in the current research.

4.1.1 Windbeutel and dorso-ventral patterning in *Drosophila*

Wind, a product of *windbeutel* gene is one of the several genes required for dorso-ventral patterning in the developing embryo of *Drosophila melanogaster*. Dorso-ventral patterning requires communication between the germ-line derived Oocyte and the somatically derived follicle cells of the ovary, which occurs through a cascade of signal transduction steps called the Gurken-EGFR pathway as shown in Figure 4.1.

Initially, the Oocyte nucleus moves to the anterio-dorsal part of the cell and the *gurken* mRNA is synthesized between the oocyte and follicle cells. The product of *gurken*, Gurken [87], accumulates around the oocyte nucleus, which is then secreted to the follicle cells. These follicle cells then differentiate to a dorsal morphology. The signal is received by the follicle cells *via* Torpedo, the homologue of the human epidermal growth factor receptor (EGFR) [125]. Torpedo (EGFR) is expressed in all follicle cells, however, it is only activated in the dorsal follicle cells receiving the Gurken signal. This signal transduction leads to two major consequences; a change in the properties of follicle cells that prevents them from acquiring ventral fate and regulates the second pathway, especially by restricting the expression of the *pipe* gene (in the second pathway) in ventral follicle cells.

The second pathway, which sends signal from the follicle cells to the embryo, requires the action of a dorsal group that include 11 genes and ultimately leads to the specification of the dorso-ventral axis of the embryo. This process is realized *via* a proteolytic cascade [83], which results in the formation of a nuclear gradient of the transcription factor Dorsal. Dorsal mRNA is supplied by the mother but protein gradient is generated after fertilization.

Windbeutel, *pipe* and *nudel* are the only three genes expressed by mother follicle cells. *Windbeutel* encodes a putative ER resident protein Wind, which is required to localize Pipe to the Golgi apparatus [63, 111]. Wind belongs to the Protein Disulfide Isomerasae (PDI) related protein sub-family (PDI-D) as described before. The gene *pipe* encodes the protein, Pipe. Pipe is a Golgi-resident type-II transmembrane protein related to the vertebrate glycosaminoglycan-modifying enzyme, heparan sulfate 2-O-sulfotransferase and is critical factor in determining the dorso-ventral polarization [111]. The gene *nudel* encodes Nudel, which has an extracellular matrix domain and a serine protease domain [51]. Nudel specifies the site of generation of Spätzle ligand, after fertilization of the Oocyte. Pipe is spatially restricted to ventral follicle cells, whereas Wind and Nudel are not. The expression of Pipe in dorsal follicle cells is inhibited by EGFR. In the Golgi, Pipe modifies an yet unidentified protein called factor (x), which in complex with Nudel is the key step in establishment of dorso-ventral polarity. Other proteins involved in signal transduction cascade leading to dorso-ventral polarization in *Drosophila* are described in Figure 4.1.

4.1.2 Crystal structure of wild-type His-Wind

The first crystal structure of an eukaryotic PDI-related protein, Wind was determined by Ma *et al*, 2003 [76]. The protein Wind consists of an N-terminal b-domain (118 residues) and a C-terminal D-domain (107 residues) connected by a flexible linker (11 residues). Wind crystallized as a homodimer in the asymmetric unit with the dimerization interface along the b-domains. The b-domain has the characteristic thioredoxin fold (with the order of secondary structure elements $\beta1$-$\alpha1$-$\beta2$-$\alpha2$-$\beta3$-$\alpha3$-$\beta4$-$\beta5$-$\alpha4$) whereas the D-domain consists of an antiparallel all α-helical secondary structure arrangement. The function of the D-domain remains unclear, although its C-terminal sequence, KDEL [84], most probably confers ER retention.

The dimer is formed by a head-to-tail arrangement of the b-domains. The dimer interface is nearly symmetrical and consists of residues to either side of $\beta1$ strand (residues

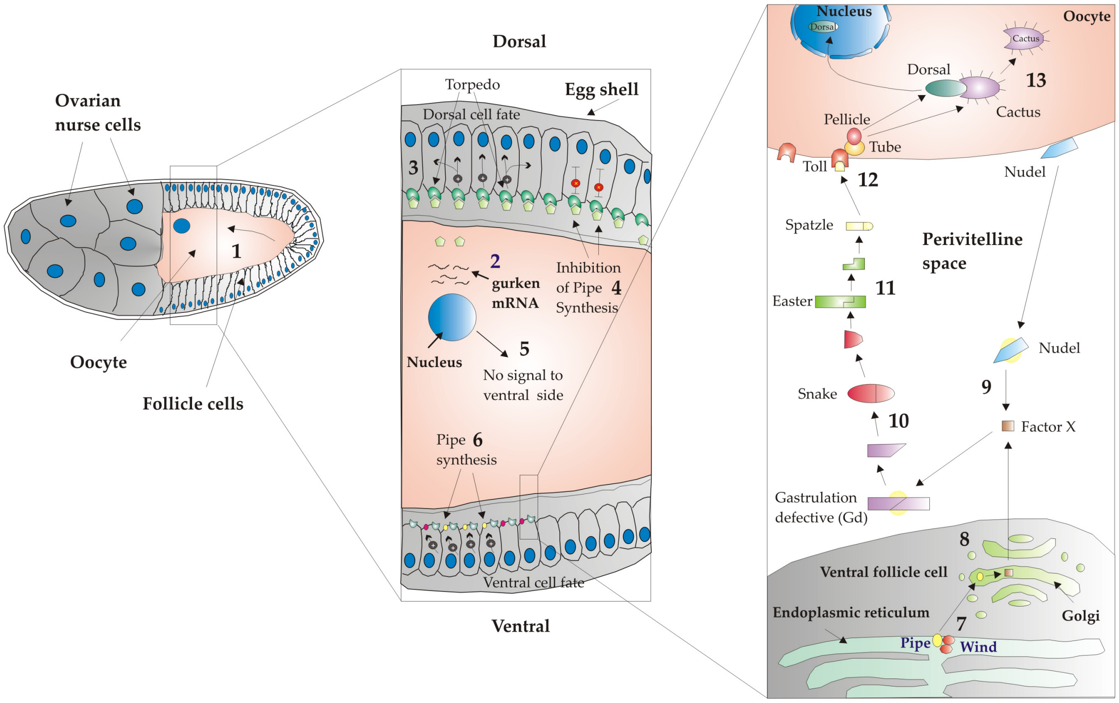


Figure 4.1: Signal transduction cascade during *Drosophila* dorso-ventral polarization (figure reproduced from molecular biology course of Prof. Dr. Fritz Aberger). (1) Oocyte nucleus travels to anterior dorsal side of the Oocyte synthesizing *gurken* mRNA, which remains between the nucleus and follicle cells. (2) *gurken* messages are translated. The Gurken protein is received by Torpedo protein during mid-oogenesis. (3) Torpedo signal causes follicle cells to differentiate to a dorsal morphology. (4) Synthesis of Pipe is inhibited in dorsal follicle cells. (5) Gurken protein does not diffuse to ventral side. (6) Ventral follicle cells synthesize Pipe protein. (7) Wind is indispensable for Pipe to translocate from the ER to the Golgi. (8) In ventral follicle cells, Pipe completes the modification of the unknown factor (x). (9) Nudel and factor (x) interact to split the Gastrulation-deficient (Gd) protein. (10) The activated Gd protein splits the Snake protein and the activated Snake protein cleaves the Easter protein. (11) The activated Easter protein splits Spätzle and activated Spätzle binds to Toll receptor protein. (12) Toll activation activates Tube and Pelle, which phosphorylates the Cactus protein. Cactus is degraded, releasing it from Dorsal. (13) Dorsal protein enters the nucleus and ventralizes the cell.

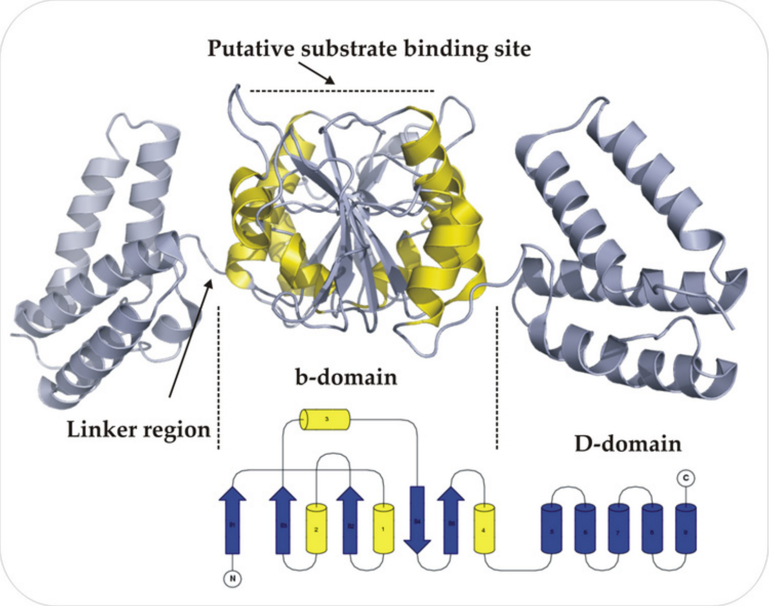


Figure 4.2: (a) Crystal structure of wild-type His-Wind dimer (1OVN). (b) Topology plot showing the b-domain containing thioredoxin fold and the all α-helical D-domain. The putative substrate binding site is also highlighted. The dimerization mode is along the N-terminal b-domains.

24-34), residues within and after $\alpha 1$ (residues 37-43) and $\alpha 2$ helices (residues 70-75). Important dimer interface interactions include multiple contacts between the backbone of the conserved Gly26 and side chain atoms of the conserved Asp31 and similar Leu33 residues, interactions between a side chain oxygen of Asp31 and backbone atoms of Thr25, and hydrophobic interactions between the Leu33 and Thr25 residues as well as interactions between the similar Val28 and Ser34 residues. The side chain of Arg41 is involved in a series of hydrophobic contacts with Phe42 and Pro43 as well as interactions between backbone atoms of Lys74 and the conserved Asp75. Dimerization creates a deep hydrophilic cleft and the CTGC motif lies at the floor near the entrance to the dimer cleft. From various biochemical experiments Wind dimerization seems to be important for its substrate processing efficiency [4].

4.1.3 Chaperone function of PDIs

Molecular chaperones were originally defined as proteins that assist in the self-assembly of other polypeptide chains but are generally not part of the functional units [24]. The chaperone function of a protein relies on its ability to interact non-covalently with specific peptide sequences or epitopes in substrate proteins [34]. However, until recently no concise data on peptide specificity, peptide binding sites, or the molecular basis for the observed selectivity of these proteins were available. Data on such inherently weak interactions would require the three dimensional structure of the protein and the interaction partners concerned. Thus much could be learned from the study of peptide binding and chaperone activity of a naturally occuring PDI family protein lacking redox properties but having a clearly defined substrate that can be studied both *in vivo* and *in vitro*. The CTGC motif in Wind seems to be redox inactive, so it has been postulated that the substrate processing activity of Wind is solely dependent on its chaperoning property. The potential interacting partner of Wind has been identified as Pipe and a putative substrate binding site has been mapped on the surface of Wind by various biochemical and genetic studies [4]. Three dimensional structures of complexes of Wind with Pipe or with Pipe epitopes might cast light on chaperone activity of this family of proteins.

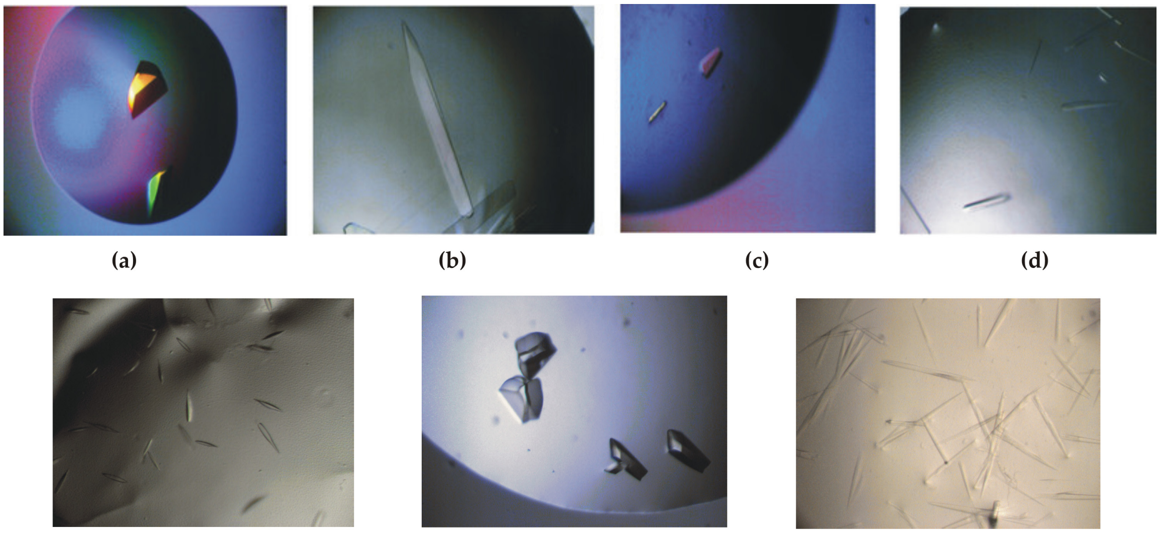
4.1.4 Close relative: human ERp29

Human ERp29 [31] is also an ER-lumenal soluble protein homologous to rat ERp29 and *Drosophila* Wind. Deduction of the amino acid sequence revealed a lack of -CXXC-motif, indicating that ERp29 is a PDI-like protein lacking oxidoreductase and disulfide isomerase activities. Structure predictions reveal a similar three dimensional architecture like that of the Wind protein. The NMR structure of rat ERp29 determined by Mkrtchian and colleagues [71], revealed a similar fold for the b-domain and an all-helical, but slightly different arrangement of the D-domains. Recently the ERp29 gene has been shown to be activated upon the treatment of rat thyroid epithelia by the thyroid stimulating hormone [66], which may be implicated, along with other ER chaperones, in the maturation and/or secretion of thyroglobulin. Co-localization of ERp29 with thyroglobulin in the putative intracellular transport structures has suggested its role in folding and export of secretory proteins [107].

4.2 Aim of Present Work

The structure and mode of binding of endoplasmic reticulum PDI-related proteins to their substrates is currently a focus of intense research. The crystal structure of wild-type Wind has been described by Ma *et al.*, 2003 [76]; the asymmetric unit consists of a dimer with the dimerization contact surface along the N-terminal b-domain. Based on this structure, a series of mutational studies have been carried out to map substrate binding site/s on the surface of Wind, and a putative peptide binding site in the Wind b-domain has been characterized with the help of *in vitro* binding assays [4]. Within the Wind dimer, a surface tyrosine cluster formed by Tyr53, Tyr55 and Tyr86 is important for substrate binding. Mutations at these sites (Y53S, Y55K, Y55S, Y86Q and Y86L) completely abrogate Pipe processing efficiency. The Y55F mutant did not show any negative effect on Pipe processing, which further suggests that the aromatic/hydrophobic behaviour of the inner Tyr53/Tyr55 pair plays a major role in Wind-Pipe interaction. In order to assess whether structural differences are responsible for these observations, crystal structures of the Wind Y53S, Y53F and Y55K mutants were determined and some hydrophobic molecular interaction field (MIF) calculations were performed to look at the surface hydrophobic behaviour in these mutants.

Another question addressed in this work was to clarify the dimerization interface in Wind, because a previously reported NMR structure of a homologous protein from rat, ERp29, led to a completely different interpretation of the dimerization mode [71] and the arrangement of helices in the D-domain were in disagreement with the crystal structure. Original crystal structure of wild-type Wind (PDB ID: 1OVN) has been determined with His_6-tag on N-terminus (His-Wind). To rule out the possibility that the His_6-tag at N-terminus, which is close to the dimerization region, was responsible for any alteration in the dimerization mode, the crystal structure of wild-type Wind with His_6-tag on the C-terminus rather than on the N-terminus, was also determined (Wind-His). Attempts to solve crystal structure of human ERp29 and the D-domain of human ERp29 alone will also be described.



(a) (b) (c) (d) (e) (f) (g)

Figure 4.3: Crystals of (a) Wind-His (b) Y53S (c) Y53F (d) Y55K (e) human ERp29 (f) Wind-peptide complex (g) D-domain of human ERp29 grown by hanging drop vapour diffusion method.

4.3 Materials and Methods

4.3.1 Expression, purification & crystallization

Wild-type Wind-His protein (with His_6-tag on the C-terminus) and the His-Wind mutants Y53S, Y53F and Y55k (with His_6-tags on N-terminus) were expressed and purified as described in Ma *et al.*, 2003 [76]. Human ERp29 and D-domain of human ERp29 were expressed and purified using a modified protocol of the one described in Ferrari *et al.*, 1998 [31].

Crystals of Wind-His, His-Wind mutants: Y53S, Y53F, Y55K, Wind-peptide complex, human ERp29 and D-domain of human ERp29 were obtained using hanging-drop vapour diffusion method [80]. Crystallization and cryoprotectant conditions are described in Table 4.1. Almost all crystals appeared in 2-4 days except for Y55K crystals which took 7-12 days to grow (Figure 4.3).

Table 4.1: Crystallization and cryoprotectant conditions of His-Wind, Y53S, Y53F, Y55K, Wind-peptide complex, human ERp29 and D-domain of human ERp29

Protein	Reservoir	Protein concentration	Cryoprotectant
Wind-His	0.1 *M* MES pH 5.8-6.6, 20 °C 50 m*M* NaCl 16-20% PEG 400	15 mg/ml	Reservoir + 25% PEG 400 10% Glycerol
Y53S	0.1 *M* MES pH 5.8-6.6, 20 °C 50 m*M* NaCl 16-20% PEG 400 5% Glycerol	10 mg/ml	Reservoir + 25% PEG 400 10% Glycerol
Y53F	0.1 *M* MES pH 5.8-6.6, 20 °C 50 m*M* LiCl 16-20% PEG 400	18 mg/ml	Reservoir + 25% PEG 400 10% Glycerol
Y55K	0.1 *M* Tris-HCl pH 8.2-8.8, 4 °C 25 m*M* $MgCl_2$ 10% PEG 4000 5% Glycerol	5 mg/ml	Reservoir + 20% Glycerol
Wind-pept	0.1 *M* MES pH 5.6-6.4, 20 °C 50 m*M* NaCl 5% DMSO 10% PEG 400	15 mg/ml + equimolar peptide	Reservoir + 25% PEG 400 10% Glycerol
ERp29	100 m*M* Na-acetate pH 4.6-5.2, 4 °C 450m*M* $(NH_4)_2SO_4$ 300 m*M* PEG 2000 MME	12 mg/ml	Reservoir + 20% Glycerol
D-domain	2.5 *M* Na-Malonate pH 7.5, 4 °C 30 m*M* $CaCl_2$	8 mg/ml	Reservoir

4.3.2 Data Collection and Processing

The crystals of each of the mutants, human ERp29, Wind-peptide complex and D-domain of human ERp29 were soaked in a suitable cryoprotectant solution as in 4.1 and mounted in a loop in a cold liquid nitrogen stream. The data sets were collected in two passes, a high resolution and a low resolution pass. Crystals of human ERp29 were initially equilibrated with the reservoir solution containing 5% and 10% glycerol for 2 minutes each, before transferring to a cryoprotectant solution containing 15% glycerol, as the crystals were sensitive when soaked directly into a solution containing 15% glycerol. The mosaicity of all crystals was on an average 1.5° or higher, probably because of the very mobile D-domains. The data collection statistics are summarized in Table 4.2 and Table 4.3. All data were processed with *HKL2000* [88] and the space group and data statistics determined by *XPREP* (Bruker AXS, Madison, WI). The D-domain of ERp29 diffracted to a resolution of 3.8 Å and could not be processed any further because of very weak diffraction power.

4.3.3 Structure Solution

The Wind-His, Y53S and Y53F structures were solved by molecular replacement using *EPMR* [60] with search models taken from the His-Wind structure (PDB ID: 1OVN). Monomers consisting of one b- and D-domain were used except for the solution of the Y55K, for which individual b- and D-domains were employed. Solutions were found with correlation coefficients of 49.6 %, 52.5 %, and 51.5 % for Wind-His, Y53S and Y53F respectively. The Y55K mutant proved to be more problematic and so *PHASER* [79] was employed. First the two b-domains were located, then a search was made for the D-domains. It was only possible to locate the second D-domain with a much lower log-likelihood gain and a very low Z-score. Though there was a solution for both the D-domains, the second D-domain had to be discarded, because of very high B-values and no traceble electron density in this region during refinement.

The structure of human ERp29 was solved by molecular replacement using brute force methods in *PHASER* [79] by first doing a rotational search and then doing a translational search around this solution, where individual b and D-domains of *Drosophila* Wind (PDB ID: 1OVN) were used as search models, similar to Y55K solution. A mixed model from the sequence alignment of ERp29 and Wind search model was prepared using the FFAS server [52] similar to the method proposed by Schwarzenbacher *et al* [110].

Table 4.2: Data collection statistics of Wind-His, Y53S and Y53F mutants. Values in parenthesis are for the outer resolution shell.

Data statistics	**Wind-His**	**Y53S**	**Y53F**
Wavelength Å	0.8976	0.8976	0.8976
X-ray source	BESSY-BL2	BESSY-BL2	BESSY-BL2
Detector	MARip345	MARip345	MARip345
Space group	C2	C2	C2
a (Å)	107.78	108.47	109.44
b (Å)	50.36	50.55	51.71
c (Å)	98.67	98.9	100.65
β (°)	112.19	112.06	112.70
Mosaicity (°)	1.8	1.2	1.5
Resolution (Å)	2.35 (2.45-2.35)	1.75 (1.85-1.75)	2.28 (2.37-2.28)
Reflections	19292	47523	22333
Redundancy	4.11 (3.00)	3.86 (2.96)	3.43 (3.0)
Completeness (%)	98.28 (99.2)	99.1 (95.9)	98.3 (89.9)
Mean I/σ(I)	20.05 (9.46)	19.32 (3.96)	15.19 (4.55)
$\mathbf{R_{int}}$(%)	4.05 (26.2)	4.36 (24.32)	4.99 (24.32)

Table 4.3: Data collection statistics of Y55K, human ERp29 and Wind-peptide complex. Values in parenthesis are for the outer resolution shell.

Data statistics	**Y55K**	**Wind-peptide**	**human ERp29**
Wavelength Å	1.0	1.0	1.0
X-ray source	DESY-X13	BESSY-BL1	BESSY-BL1
Detector	MARCCD	MARCCD	MARCCD
Space group	$P2_1$	C2	P2
a (Å)	35.42	107.97	58.257
b (Å)	119.09	50.35	68.144
c (Å)	64.03	98.30	70.509
β (°)	112.20	112.20	107.41
Mosaicity (°)	1.9	1.4	1.6
Resolution (Å)	2.25 (2.34-2.25)	2.0	2.95 (2.99-2.95)
Reflections	22938	31372	10742
Redundancy	4.67 (2.62)	5.83	6.59 (3.44)
Completeness (%)	97.8 (91.2)	98.7	95.7 (74.9)
Mean I/σ(I)	16.83 (3.51)	24.15	16.02 (3.75)
$\mathbf{R_{int}}$(%)	6.5 (26.6)	3.93	6.76 (26.71)

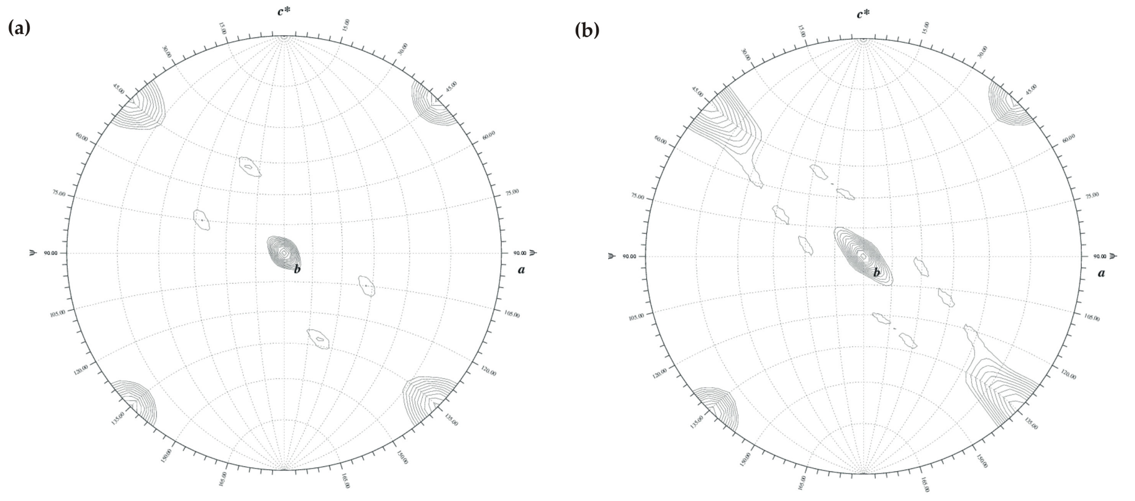


Figure 4.4: Self rotation function at κ = 180° calculated with (a) experimental data and (b) solution using data between 10.0 Å and 3.5 Å resolution and a 20 Å radius of integration.

Attempts to solve the structure using individual b and D-domains from NMR structures of rat ERp29 (PDB ID: 1G7D & 1G7E) as search models failed.

In the beginning, the two b-domains were located in the asymmetric unit and then the D-domains were located by searching around the already located b-domains. Self rotation plots (Figure 4.4) from the data and solution were calculated using *GLRF* [124] and compared to identify the correct solution. Ideal data from *PHASER* solution was calculated using *SHELXPRO* [113] and *XPREP* (Bruker AXS, Madison, WI). As in Figure 4.4, at κ = 180° a clear 2-fold non-crystallographic axes between the a and c^* axes can be seen, in other words there are two molecules in the asymmetric unit as can be seen from the peaks at ψ = 45° and ψ = 90°, both for experimental data and solution.

4.3.4 Model Bias Minimization, Refinement & Validation

For Wind-His, Y53S, Y53F and Y55K, Prime and Switch phasing in *RESOLVE* was used to minimize model bias [122, 121]. All the structures were refined against $\boldsymbol{F}$ using TLS refinement (four independent TLS domains for two b-domains and two D-domains were defined) and NCS restraints in *REFMAC* [86], alternating with model building in real space using $2m\boldsymbol{F}_o$-$D\boldsymbol{F}_c$ and $m\boldsymbol{F}_o$-$D\boldsymbol{F}_c$ in *COOT* [25]. Waters were modelled using *ARP/wARP* and *COOT*. The refinement statistics are summarized in Table 4.4.

Different methods have been attempted to minimize model bias in case of human

Table 4.4: Refinement statistics of Wind-His, Y53S, Y53F and Y55K. 5% of reflections in thin shells were taken as test data set.

Refinement statistics	**Wind-His**	**Y53S**	**Y53F**	**Y55K**	**Wind-pep**
Final R-factor (%)					
Working Set	22.60	21.92	21.92	22.27	22.68
Working + Test set	22.95	22.15	22.24	22.63	22.85
Final Free R-factor (%)	28.34	26.75	28.06	29.25	27.02
R.m.s deviations					
Bond lengths (Å)	0.025	0.021	0.031	0.033	0.028
Bond angles (°)	2.057	1.784	2.377	2.745	2.120
Mean B value ($Å^2$)					
Monomer 1					
Main chain atoms	15.707	25.125	30.770	49.889	39.64
Side chain atoms	15.313	28.175	29.646	49.065	41.23
Monomer 2					
Main chain atoms	25.458	41.263	41.375	47.438	41.22
Side chain atoms	22.696	37.593	38.636	46.207	46.81
Solvent	15.338	30.284	32.510	40.634	40.27
No. of protein atoms	3287	3306	3255	2699	3575
No. of solvent atoms	108	167	98	55	135
PDB entry code	2C0E	2C0G	2C0F	2C1Y	-

ERp29 from the molecular replacement solution because of poor resolution and weak data. After an initial rigid body refinement in *REFMAC* [86], a σA-weighted map (Map-1) calculated by the program *SIGMAA* [101] and Prime and Switch phasing (Map-2) in *RESOLVE* [121, 122] were used to rebuild parts of the model. A composite omit map was calculated using simulated annealing (Map-3) in *CNS* [12]. Visible residues of the sequence were docked to the MR model by comparing maps 2 and 3 and using this model a 2-fold NCS averaged map was calculated using the program *DM* [16]. The model was again corrected using the map from *DM* and the structure was refined using simulated annealing and energy minimization protocol in *CNS*. Owing to weak and poor data statistics, the structure could only be refined to an R-factor of 28% and R_{free} of 33%.

The final models were validated with *PROCHECK* [67]. All residues lie in the allowed regions of the Ramachandran plot. The coordinates and structure factors for Wind-His, Y53S, Y53F and Y55k have been deposited in the Protein Data Bank (entry codes 2C0E, 2C0F, 2C0G and 1C1Y).

4.3.5 Hydrophobic Molecular Interaction Field

The hydrophobic molecular interaction fields at the surfaces of the different mutants were analyzed using a DRY probe in the program package *GRID* [42]. The hydrophobic probe finds favourable locations that interact with other molecule/s on surface of a protein in aqueous environment [72]. All *GRID* calculations were performed for the whole volume of the protein using a grid spacing of 1.0 Å.

4.4 Results and Discussion

4.4.1 Structure of Wind-His and its mutants

The overall fold of all three mutants is similar to the wild-type structure. The mutants crystallize as homodimers in the asymmetric unit, each monomer consisting of two domains, the N-terminal b-domain (118 residues) and the C-termianl D-domain (107 residues), connected by a flexible linker of 11 residues. The b-domain adopts an α/β fold with the order of secondary structure elements β1-α1-β2-α2-β3-α3-β4-β5-α4. The strands of the β-sheet form a central core surrounded by the four α-helices. This fold, characteristic of protein-disulfide isomerases, is called the thioredoxin fold. The D-domain has a five-helix fold with all the helices in anti-parallel arrangement. Both the N-terminal His_6-tag in the case of Y53S, Y53F and Y55K and the C-terminal His_6-tag of Wind-His were not visible in the density. The second D-domain was completely absent in Y55K. The overall fold of human ERp29 also resembles Wind. human ERp29 crystallized as two independent monomers in the asymmetric unit, about 6 Å apart from each other and the biological dimer is formed with symmetry equivalent molecule, upon application of a crystallographic 2-fold axis.

4.4.2 Superposition of His-Wind and its mutants on Wind-His

Least squares superposition of main chain atoms of all the mutants (Y53S, Y53F, Y55K) and wild-type His-Wind on Wind-His gave a r.m.s.d. (root-mean-square-deviation) of 0.247, 0.272, 0.365 and 0.434 for His-Wind, Y53S, Y53F and Y55K respectively. The program *ESCET* [108] was used to determine conformationally invariant regions between different pairs of mutants. The C.S.I (conformational similarity index) relative to His-Wind, calculated using *ESCET*, is in the range of 90% for all the structures except for Y55K mutant which has a C.S.I. of around 70%. In all the structures, the D-domains show the least significant variations, which may in part be a consequence of their higher B-values. There are some minor differences in the loop regions of the b-domain as illustrated in Figure 4.5 . All the structures show significant variations in the conformations of some side-chains. Compared with other structures, the D-domain of Y55K exhibits an appreciably different orientation relative to the b-domain, with a rotation of about 45° around Gly145. With such an orientation of the D-domain with respect to the b-domain, residue Cys149 in the linker region is closer to Tyr143 with a $C\alpha$-$C\alpha$ distance between them of about 4.44 Å compared to about 6.9 Å in the wild-type and other mutant structures. Cys149 is also closer to Tyr194 (3.65 Å) and Asn154 (3.94 Å) thus stabilizing the movement of the D-domain. The significance of this if any with a recent observation [49] in a Wind homologous protein ERp29 (from rat), where residue Cys125 (homologous to Cys149 in Wind) plays a key structural role in providing stability to the C-domain has yet to be clarified.

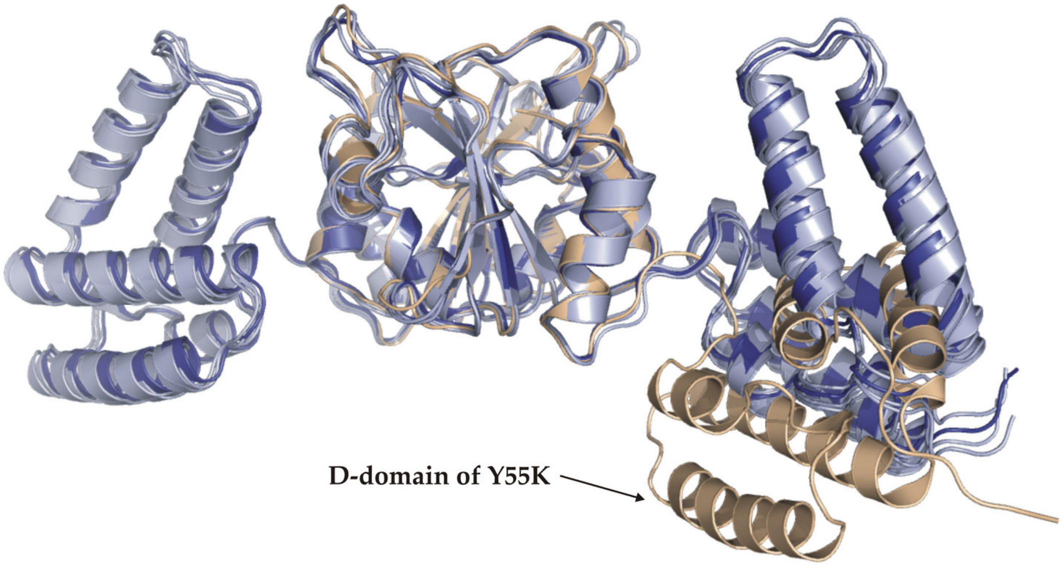


Figure 4.5: Superposition of Wind-His with Y53S, Y53F (all colored in light blue) with His-Wind (dark blue) and Y55K (wheat). The D-domain of Y55K exhibits an appreciably different orientation relative to the b-domain when compared to other mutants.

4.4.3 The dimer interface

The Wind dimer is formed by a head-to-tail arrangement of b-domains with the participation of residues constituting $\beta 1$ strand, residues within and after $\alpha 1$ and $\alpha 2$ helices, mediated through a variety of hydrophobic and H-bonding interactions shown in Figure 4.6 . The conserved residues Gly26, Val28, Gln31 and Arg41 contribute to the dimer interface (Figure 4.7a). In wild-type Wind-His and all mutants, the H-bonding distances in the region of the dimer interface agree within experimental error with those in His-Wind. The similarities in dimerization modes of Wind-His and His-Wind also indicate that the N-terminal His$_6$-tag in His-Wind does not interfere with dimerization.

4.4.4 Opening of the CTGC loop in Y53F

In the crystal structure of Y53F, the CTGC loop in both monomers has lost its disulfide bond, as shown in Figure 4.8. As a result of a change of approximately 180° in ψ(Thr25), Cys27 is about 4.2 Å from the carboxyl oxygen of Glu32. The distance between Tyr53 and cys27 is about 11.8 Å, which rules out the possibility of the mutation affecting this opening of the cysteine loop. Whether this opening has a physiological reason or is due to radiation damage is unknown. The diffraction data did not show any other indication

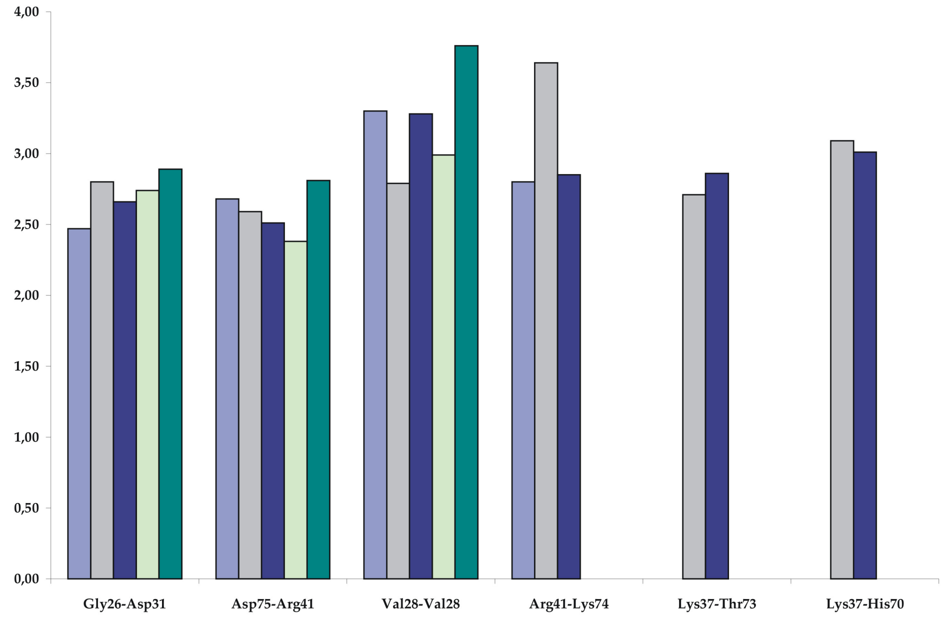


Figure 4.6: Dimer interface interactions in His-Wind (light blue), Wind-His (light grey), Y53S (dark blue), Y53F (light green) and Y55K (dark green). The conserved amino acid interactions namely Gly26(N)-Asp31(OD2), Asp75(OD1)-Arg41(NH2), Val28(CG1)-Val28(CG1) have minor differences in the distances between corresponding amino acids, showing the conservation of dimerization interface in all mutants. Some additional interactions in other mutants are also depicted.

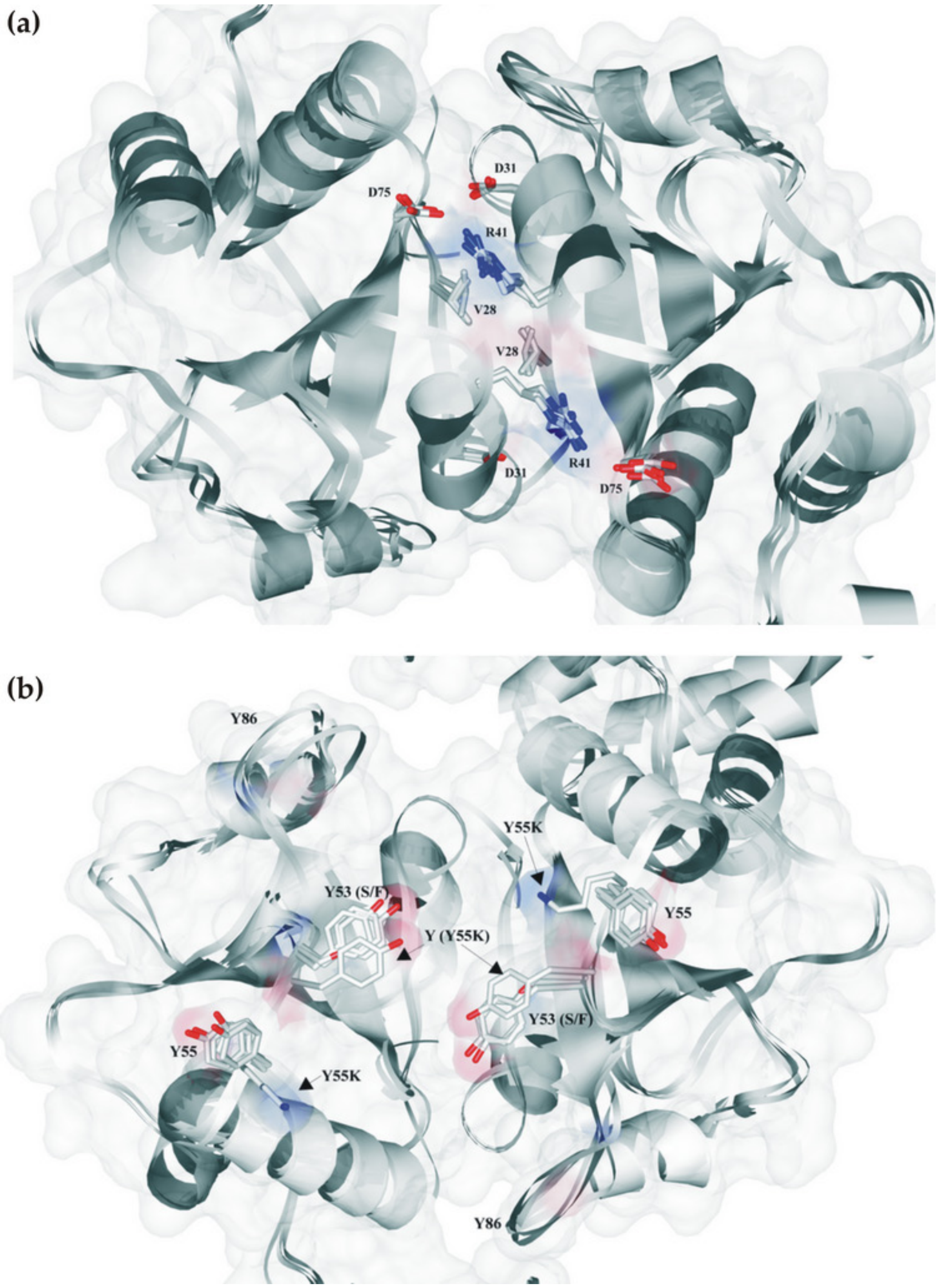


Figure 4.7: (a) Superposition of the five structures viewed approximately down the non-crystallographic twofold axis of the dimer, with the key residues in the structurally strongly conserved Wind-His dimer interface highlighted. (b) Superposition of the the residues at putative substrate binding site in Wind-His, His-Wind, and the Y53F, Y53S and Y55K mutants. The orientation of the residues Tyr53 and Lys55 in the Y55K mutant are slightly different to those in the other four structures. The binding site stretches over both monomers and in the wild-type presents a non-polar surface composed of aromatic residues at the opposite end of the non-crystallographic twofold axis of the dimer to that shown in (a).

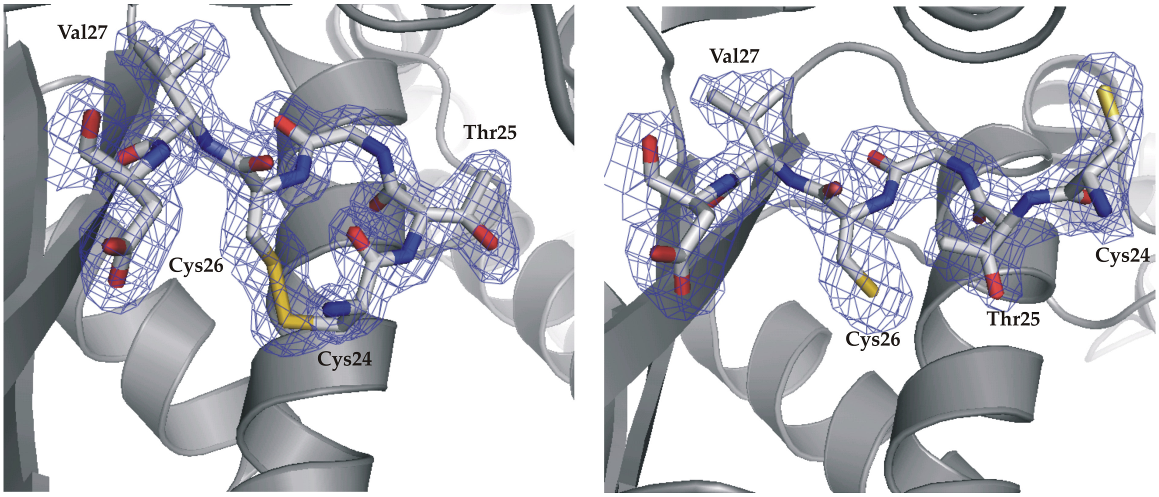


Figure 4.8: (a) Intact disulfide bond in Wind-His (His-Wind, Y53S and Y55K are similar) and (b) the open disulfide bridge in Y53F, which is accompanied by an approximately 180° rotation about ψ(Thr25).

of radiation damage.

4.4.5 Substrate binding site and hydrophobic MIF

The substrate binding site containing Tyr53, Tyr55 and Tyr86 residues was examined for possible structural differences amongst the mutants Y53S, Y53F and Y55K. This region is close to pseudo-twofold axis of the dimer so it is possible that the sites in the two monomers combine to make a single Pipe binding site. The mutants Y53S and Y53F are conformationally very similar, but in Y55K the lysine side-chain points upwards 90° away from the plane of tyrosine ring (Figure 4.7b). It is possible that this results in a tighter association of the substrate (Pipe) with Wind dimer leading to a stronger complex, accounting for the observations in *in vivo* assays with the Wind-Y55K mutant and Pipe, where Pipe translocation from Golgi was abrogated.

To analyze the hypothesis that the Wind-Pipe interaction is predominantly hydrophobic involving aromatic interactions, the relative hydrophobic potentials on the surface shown in Figure 4.9 were estimated using a DRY (hydrophobic) probe in the program *GRID* [42]. The net charge on each of the dimers as calculated using GRIN [42] are -7.38, -9.05, -11.23 and -13.21 for Wind-His, Y53F, Y53S and Y55K respectively, so the two inactive mutants are appreciably more polar. In both wild-type Wind-His and Y53F mutant, the substrate binding site has a much larger hydrophobic interaction area at the

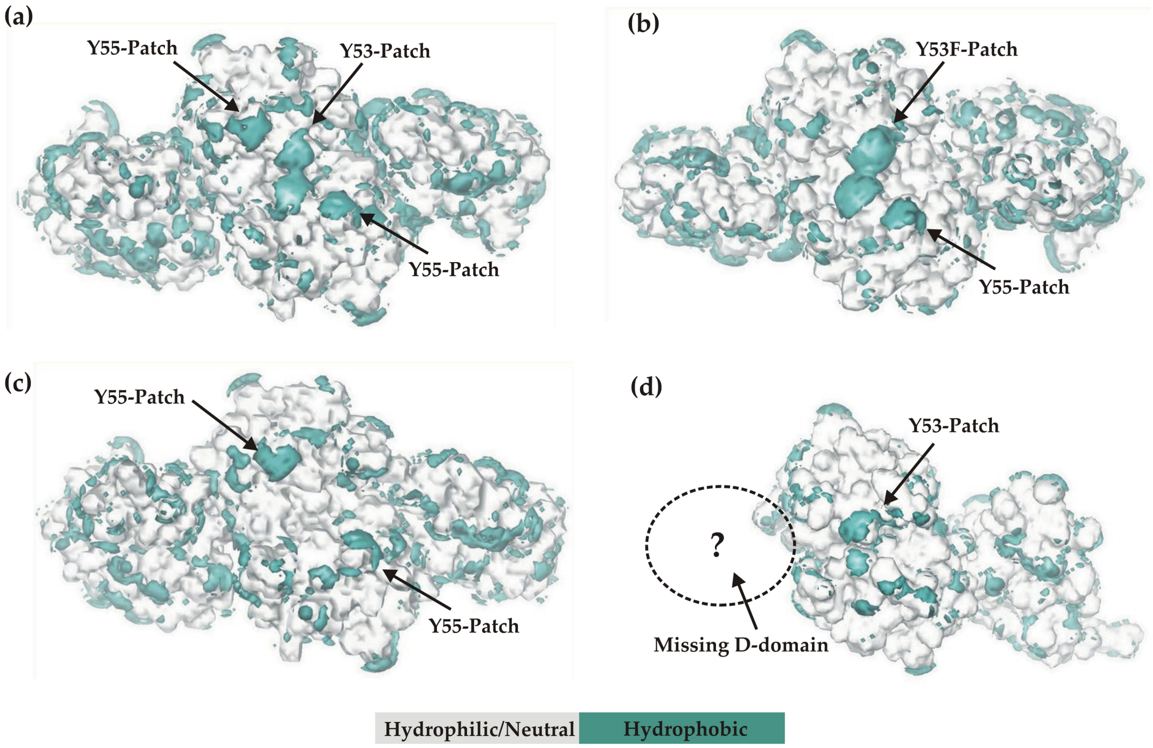


Figure 4.9: Hydrophobic molecular interaction field of (a) Wind-His dimer (b) Y53F dimer(c) Y53S dimer and (d) Y55K dimers, contoured at -0.170 Kcal/mole, highlighting the putative substrate binding surface. Wind-His and Y53F are more hydrophobic whereas Y55K and Y53S are more hydrophilic.

"tyrosine-triad" than in the inactive mutants Y53S and Y55K.

4.4.6 Putative substrate binding site

First of all, to postulate the mechanism of substrate binding, all interactions near the putative substrate binding site with the symmetry equivalent molecules were compared in Wind-His, which weakly binds the substrate, Pipe and Y55K, which strongly binds the substrate Pipe. In Y55K, Tyr53 of one of the b-domains forms a H-bond with Arg218 and the second Tyr53 forms a H-bond with Tyr44 of a symmetry equivalent molecule. One of the Lys55 residues forms a H-bond with Ser137 of a symmetry equivalent molecule and the second Lys55 is involved in hydrophobic interactions with a symmetry equivalent Phe42. In wild-type Wind-His, except for one tyrosine (Tyr55) that forms a H-bond with symmetry equivalent Arg237, all other tyrosines are involved in hydrophobic interactions. These interactions give us an indication of the substrate binding mechanism in Y55K, which binds much more strongly than wild-type Wind-His to Pipe or to peptides based on parts of Pipe sequence.

Secondly, in parallel to biochemical experiments performed by Barnewitz *et al*, 2004 [4] several trials were made to co-crystallize complexes of Wind-His and its mutants with peptides mimicking substrate binding epitopes in Pipe sequence [4] and determine the crystal structure. Several different conditions were tried to prepare these complexes and co-crystallizing at different pH gradients (5.0-8.0) using MES, Tris-HCl and HEPES buffers, with different salt concentrations (0.2-2.0 *M*) and different chloride salts at different temperatures (4 °C, 15 °C and 20 °C) for Wind-His, Y53S and Y55K mutants with four different kinds of 13-mer peptides (NEIEFYQFSRQRL, VRRNFTNEIEFYQ, RFFEGVRDIYATS and VEGIGDHRRQSLF) that mimick epitopes on Pipe binding Wind. Most of the times, though crystals were obtained, there was no density identified near the peptide binding site corresponding to the co-crystallized peptides. Only in the case of Wind-His co-crystallized with the peptide sequence NEIEFYQFSRQRL, it was possible to model fragments of two peptides binding near the putative Pipe-binding site as can be seen in Figure 4.10 . It was assumed that probably these peptide sequences would need secondary structural elements hanging on both sides of the binding -FY-motif and so cocrystallization trials with B-chain of Insulin, which also has this motif at the N-terminus were carried out, which again resulted in no complexes. Although one can have a glimpse of Pipe binding mechanism near the putative Pipe-binding site on Wind, such kind of weak hydrophobic interactions forming protein-peptide complexes

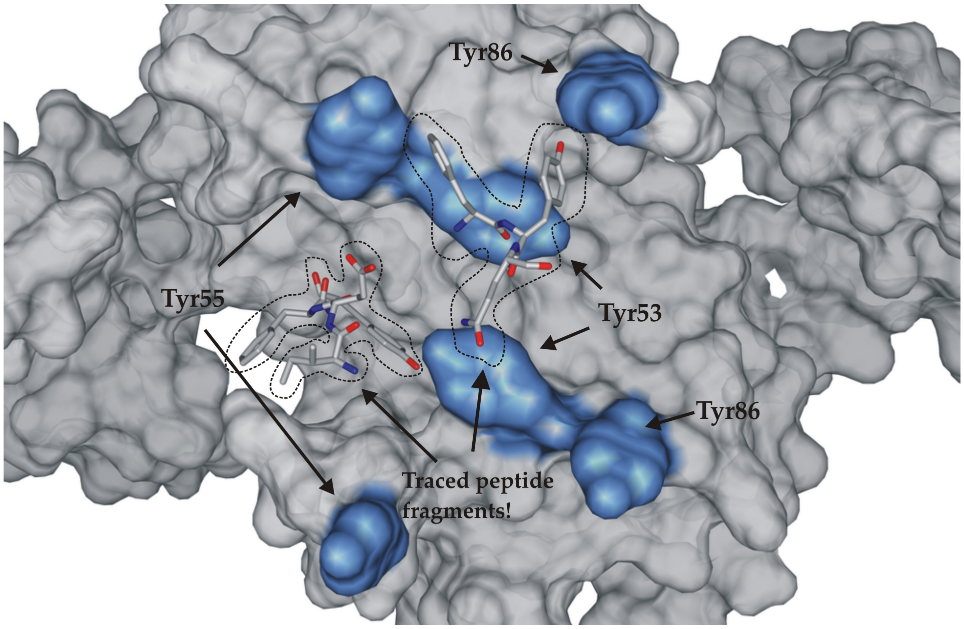


Figure 4.10: Surface plot of Wind-His near the putative peptide bindng site, showing the traced peptides, which has been co-crystallized and refined to an occupancy of 40%.

are difficult to capture in crystal structures. Further trials need to be made with longer peptides and more biochemical experiments need to be done to identify exact binding kinetics of Wind-Pipe interactions in order to get a clear picture of these interactions in co-crystal structures.

4.4.7 Dimerization in Wind and human ERp29

The b-domain of Wind includes a homodimerization surface encompassing residues to either side of the β1 strand, residues within and after α1 and α2 helices. Here important stabilizing interactions are mediated by Gly26, Val28, Asp31 and Arg41. Dimerization is important for the function of Wind and mutants that cannot dimerize are functionally inactive. In contrast, for rat ERp29, it has been suggested that the dimerization site comprises residues Asp71 (Wind Glu60), Phe118, Arg122, Asp123 (Wind Ile108, Lys112, Gly113) and Trp144 (Phe 135 in Wind). As shown in Figure 4.11, the proposed dimerization surfaces are on different faces of the thioredoxin fold. The proposed dimerization interface in rat ERp29 [71] is unsatisfactory because critical dimer interface residues in Wind and interface surface complementarity were conserved in the rat ERp29 structure but remain unaccounted for and the not unambiguous interpretation of the structural

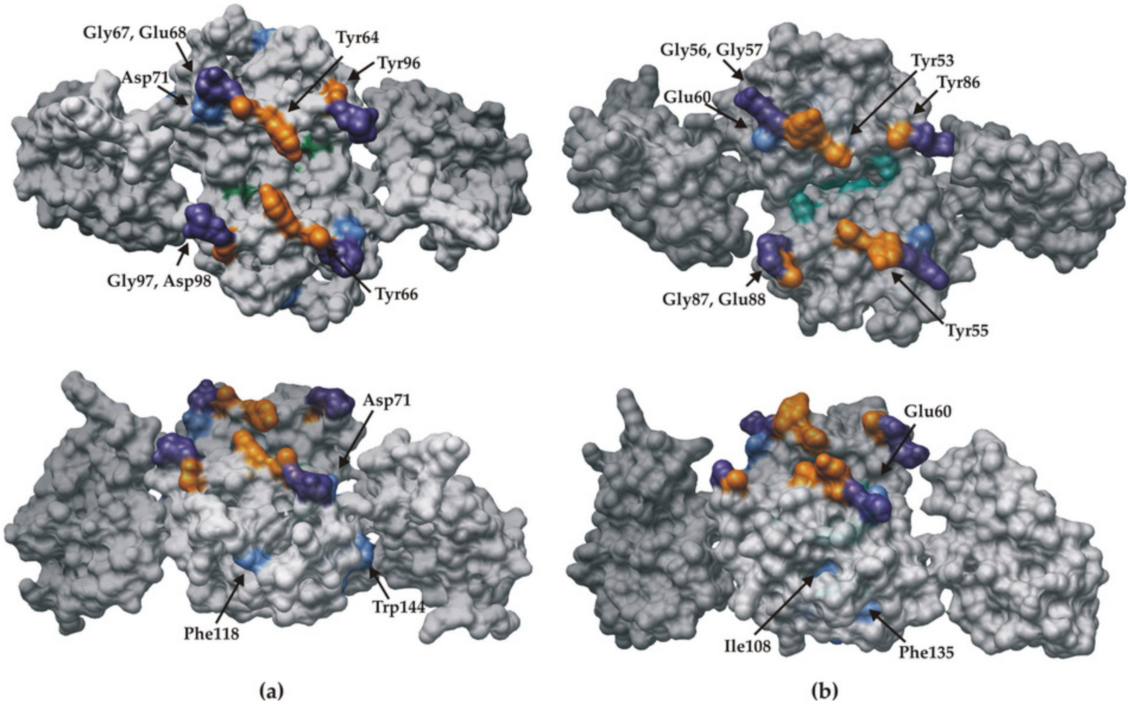


Figure 4.11: Variant modes of ERp29 dimerization compared to Wind. (a) Top view and side view of 2.95 Å crystal structure of human ERp29. Residues corresponding to the peptide binding site tyrosine cluster in Wind (orange) and required dimer interface residues as indicated in work (green) are indicated. Applied to this image, residues corresponding to suggested dimer interface residues in rat ERp29 (light blue) and multimerization residues (darkblue) are indicated. (b) Top view and side view of Wind 1.9 Å crystal structure, with tyrosine cluster of peptide binding site (orange), and dimer interface residues (green). Applied to this images, residues corresponding to suggested dimer interface residues in rat ERp29 (light blue) and multimerization residues (dark blue). The dimerization site suggested for rat ERp29 is clearly distal to the actual dimerization interface of human ERp29 and *Drosophila* Wind b-domains. The overall crystal structure of human ERp29 thus closely parallels that of Wind.

data, acknowledged by the authors themselves.

Analysis of the preliminary crystal structure of human ERp29 at the N-terminal dimerization interface, where most residues could be modelled, shows important contributions to the dimer interface from residues Gly37, Val28, Leu39, Asp42 and Lys52 (Figure 4.11a &b). Furthermore, there is no other evidence of an alternative dimerization interface in the crystal including any of the residues Asp71, Phe118, Arg122, Asp123 and Trp144 of ERp29, which were suggested to be involved in dimerization of rat ERp29, nor any of the residues (Gly67, Glu68, Gly97 and Asp98) suggested to be involved in multimerization. The conservation of the dimerization interface residues is further supported by mutagenisis experiments as are described in Lippert *et al*, 2006 [73].

The crystal structure of human ERp29 indicates a molecule somewhat more compact than Wind, with dimensions of 51.5 x 29.4 x 92.3 $Å^3$ compared to 55.2 x 34.5 x 101.4 $Å^3$ for Wind. A more compact form was also indicated for rat ERp29 on the basis of hydrodynamic properties of the protein [49]. The crystal structure of human ERp29 was solved to 2.95 Å using individual b and D-domains of Wind as search models to remove any bias introduced on the mode of dimerization. It has been observed that critical dimerization interface residues are conserved in both proteins. Thus it is most likely that this model based on the reported crystal structures holds for all PDI-Dβ proteins.

4.5 Future perspectives

The current work has given an indirect evidence of the mechanism of substrate binding/interaction to PDI-related protein, Wind. But there is a need for the determination of crystal structure of its interacting partner, Pipe, a transmembrane protein, and the structure of Pipe in complex with its interacting partner Wind, in order to understand the molecular details of binding mechanism. This model might help in understanding the finer details of chaperone function of various proteins in general in the endoplasmic reticulum.

Bibliography

[1] Abergel, C. (2004). *Acta cryst*. **D60**, 1413-1416.

[2] Abo-Amer, A. E., Munn, J., Jackson, K., Aktas, M., Golby, P., Kelly, D. J. & Andrews, S. C. (2004). *J. Bacteriol*. **186(6)**, 1879-1889.

[3] Adams, J. A., Jewell, D., Jorgensen, K., Mickley, M. & Newman, J. M. (2002). *JALA*. **7(6)**, 36-40.

[4] Barnewitz, K., Guo, C., Sevvana, M., Ma, Q., Sheldrick, G. M., Söling, H.-D. & Ferrari D. M. (2004). *J. Biol. Chem*. **279**, 39829-39837.

[5] Beier, D & Gross, R. (2006). *Current opinion in Microbiology*. **9**, 143-152.

[6] Bhat, T. N. (1988). *J. Appl. Cryst*. **21**, 279-281.

[7] Bott, M. & Dimroth, P. (1994). *Mol. Microbiol*. **14**, 347-356.

[8] Bott, M., Meyer, M. & Dimroth, P. (1995). *Mol. Microbiol*. **18**, 533-546.

[9] Bott, M. (1997). *Arch. Microbiol*. **167**, 78-88.

[10] Bren, A & Eisenbach, M. (2000). *J. Bact*. **182(24)**, 6865-6873.

[11] Britton, D. (1972). *Acta Cryst*. **A28**, 296-297.

[12] Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst*. **D54**, 905-921.

[13] Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005). *Acta Cryst*. **D61**, 556-565.

[14] Clegg, W., Blake, A. J., Gould, R. O. & Main, P. (2001). Editor. Crystal structure analysis. Oxford science publications.

[15] Collaborative computational project, No 4. (1994). *Acta cryst*. **D50**, 760-763.

[16] Cowtan, K. (1994). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*. **31**, 34-38.

[17] Dauter, Z. (1999). *Acta Cryst*. **D55**, 1703-1717.

[18] Dauter, Z. (2003). *Acta Cryst*. **D59**, 2004-2016.

[19] DeLano, W. L. (2003). The *PYMOL* molecular graphics system. *DeLano Scientific LLC*, San Carlos, CA, USA.

[20] Drenth, J. (1994). Editor. Principles of protein X-ray crystallography. Springer advanced texts in chemistry.

[21] Ducruix, A. & Giege, A. (1992). Crystallization of nucleic acids and proteins. Oxford university press.

[22] Duisenberg, A. J. M. (1992). *J. Appl. Cryst*. **25**, 92-96.

[23] Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *J. Appl. Cryst*. **36**, 220-229.

[24] Ellis, R. J., van der Vies, S. M. & Hemmingsen, S. M. (1989). **55**, 145-153.

[25] Emsley, P. & Cowtan, K. (2004). *Acta Cryst*. **D60**, 2126-2132.

[26] Esnouf, R. M. (1999). *Acta Cryst*. **D55**, 938-940.

[27] Evans, P. R. (1999). *Acta Cryst*. **D55**, 1771-1772.

[28] Evans, P. (2006). *Acta Cryst*. **D62**, 72-82.

[29] Falke, J. J., Bass, R. B., Butler, S. L., Chervitz, S. A. & Danielson, M. A. (1997). *Annu. Rev. Cell Dev. Biol.* **13**, 457-512.

[30] Falke, J. J. (2002). *PNAS*. **99(10)**, 6530-6532.

[31] Ferrari, D. M., Nguyen Van, P., Kratzin, H. D. & Söling, H. -D. (1998). *Eur. J. Biochem*. **255**, 570-579.

[32] Ferrari, D. M. & Söling, H.-D. (1999). *Biochem. J.* **339**, 1-10.

[33] Freedman, R. B., Hirst, T. R. and Tuite, M. F. (1994). *Trends Biochem. Sci.* **19**, 331-336.

[34] Freedman, R. B., Klappa, P., and Ruddock, L. W. (2002). *EMBO Rep.* **3**, 136-140.

[35] French, G. S. & Wilson, K. S. (1978). *Acta Cryst.* **A34**. 517.

[36] Frey, M. (1994). *Acta Cryst.* **D50**, 663-666.

[37] Garman, E. F. & Schneider, T. R. (1997). *J. Appl. Cryst.* **30**, 211-237.

[38] Garman, E. F. (1999). *Acta Cryst.* **D55**, 1641-1653.

[39] Gerharz, T., Reinelt, S., Kaspar, S., Scapozza, L. & Bott, M. (2003). *Biochemistry.* **42**, 5917-5924.

[40] Giacovazzo, C. (2002). Editor. *Fundamentals of Crystallography*. Oxford University Press.

[41] Golby, P., Davies, S., Kelly, J. R., Guest, J. R. & Andrews, S. C. (1999). *J. Bacteriol.* **181(4)**. 1238-1248.

[42] Goodford, P. J. (1985). *J. Med. Chem.* **28**, 849-857.

[43] Gu, Y. Z., Hogenesch, J. B. & Bradfield, C. A. (2000). *Annu. Rev. Pharmacol. Toxicol.* **40**, 519-561.

[44] Harp, J. M., Hanson, B. L., Timm, D. E. & Bunick, G. J. (1999). *Acta Cryst.* **D55**, 1329-1334.

[45] Harp, J. M., Timm, D. E. & Bunick, G. J. (1998). *Acta Cryst.* **D54**, 622-628.

[46] Hashimoto, C., Hudson, K. L. & Anderson, K. V. (1988). *Cell.* **52(2)**, 269-279.

[47] Heras, B. & Martin, J. L. (2005). *Acta Cryst.* **D61**, 1173-1180.

[48] Herbst-Irmer, R. & Sheldrick, G. M. (1998). *Acta Cryst.* **B54**, 443-449.

[49] Hermann, V. M., Cutfield, J. F. & Hubbard, M. J. (2005). *J. Biol. Chem.* **280**, 13529-13537.

[50] Hodel, A., Kim, S. -H. & brünger, A. T. (1992). *Acta Cryst.* **A48**, 851-858.

[51] Hong, C. C. & Hashimato, C. (1995). *Cell*. **82(5)**, 785-794.

[52] http://ffas.ljcfr.edu/ffas-cgi/cgi/ffas.pl

[53] Jameson, G. B. (1982). *Acta Cryst*. **A32**, 239-244.

[54] Janausch, I. G., Zientz, E., Tran, Q. H., Kröger, A. & Unden, G. (2002). *Biochim. Biophys. Acta*. **1553**, 39-56.

[55] Janausch, I. G., Garcia-Moreno, I. & Unden, G. (2002). *J. Biol. Chem*. **277(42)**, 39809-39814.

[56] Kabsch, W. (1988). *J. Appl. Cryst*. **21**, 916-924.

[57] Kaspar, S., Perozzo, R., Reinelt, M. M., Pfister, K., Scapozza, L. & Bott, M. (1999). *Mol. Microbiol*. **33(4)**, 858-872.

[58] Kaspar, S. & Bott, M. (2002), *Arch Microbiol*. **177**, 313-321.

[59] Khorchid, A., Inouye, M., Ikura, M. (2005). *Biochem. J*. **385**, 255-264.

[60] Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst*. **D55**, 484-491.

[61] Kleywegt, G. J. (2000). *Acta Cryst*. **D56**, 249-265.

[62] Kneuper, H., Janausch, I. G., Vijayan, V., Zweckstetter, M., Bock, V., Griesinger, C. & Unden. G. (2005). *J. Biol. Chem*. **280(21)**, 20596-20603.

[63] Konsolaki, M. & Schüpbach, T. (1997). *Genes and Development*. **12**, 120-131.

[64] Kraulis, P. J. (1991). *J. Appl. Cryst*. **24**, 946-950.

[65] Krissinel, E. & Henrick, K. (2004). *Acta Cryst*. **D60**, 2256-2268.

[66] Kwon, O. Y., Park, S., Lee, W., You, K. H., Kim, H. & Shong, M. (2000). FEBS Lett. 475, 27-30.

[67] Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst*. **26**, 283-291.

[68] Lebedev, A. A., Vagin, A. A., Murshudov, G. N. (2006). *Acta Cryst*. **D62**, 83-95.

[69] Lee, b. & Richards, F. M. (1971). *J. Mol. Biol*. **55**, 379-400.

[70] Leslie, A. G. W. (1999). *Acta Cryst*. **D55**, 1696-1702.

[71] Liepinsh, E., Baryshev, M., Sharipo, A., Ingelman-Sundberg, M., Otting, G. & Mkrtchian, S. (2001). *Structure*. **9**, 457-471.

[72] Liljefors, T. (1998). *Perspectives in Drug Discovery and Design*. **9**, 3-17.

[73] Lippert, U., Sevvana, M. & Ferrari, D. M. (2006). *Manuscript submitted to JBC*.

[74] Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst*. **D49**, 530-533.

[75] Lusty, C. L. (1999). *J. Appl. Cryst*. **32**, 106-112.

[76] Ma, Q., Guo, C., Barnewitz, K., Sheldrick, G. M., Söling, H.-D., Uson, I. & Ferrari, D. M. (2003). *J. Biol. Chem*. **278**, 44600-44607.

[77] Massa, W. (2004). Editor. Crystal structure determination. Springer-Verlag Berlin.

[78] McCoy, A. J. (2004). *Acta Cryst*. **D60**, 2169-2183.

[79] McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst*. **D61**, 458-464.

[80] McPherson, A. (1992). *J. Crystal. Growth*. **122**, 161-167.

[81] Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol*. **277**, 505-524.

[82] Morisato, D. & Anderson, K. V. (1994). *Cell*. **76(4)**, 677-688.

[83] Morisato, D. & Anderson, K. V. (1995). *Annu. Rev. Genet*. **29**, 371-399.

[84] Munro, S. & Pelham, H. R. (1987). *Cell*. **48**, 899-907.

[85] Murray-Rust, P. (1973). *Acta Cryst*. **B29**, 2559-2566.

[86] Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst*. **D53**, 240-255.

[87] Neuman-Silberg, F. & Schüpbach, T. (1993). *Cell*. **8;75(1)**, 165-174.

[88] Otwinowski, Z. & Minor, W. (1997). *Methods Enymol*. **276**, 307-326.

[89] Ottemann, K. M., Xiao, W., Shin, Y. K. & Koshland, D. E. (1999). *Science*. **285**, 1751-1754.

[90] Pappalardo, L., Janausch, I. G., Vijayan, V., Zientz, E., Junker, J., Peti, W., Zweckstetter, M., Unden, G. & Griesinger, C. (2003). *J. Biol. Chem.* **178**, 39185-39188.

[91] Parsons, S. (2003). *Acta Cryst.* **D59**, 1995-2003.

[92] Petterson, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605-1612.

[93] Perrakis, A., Morris, R. J., Lamzin, V. S. (1999). *Nat. Struct. Biol.* **6**, 458-463.

[94] Pflugrath, J. W. (1999). *Acta Cryst.* **D55**, 1718-1725.

[95] Phillips, W. C., Stanton, M., Stewart, A., Qian, H., Ingersoll, C. & Sweet, R. M. (2000). *J. Appl. Cryst.* **33**, 243-251.

[96] Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). *Proteins.* **41**, 47-57.

[97] Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* **D59**, 1145-1153.

[98] Posas, F., Wurgler-Murphy, S. M., Maeda, T., Witten, E. A., Thai, T. C. & Saito, H. (1996). *Cell.* **86**, 865-875.

[99] Pratt, C. S., Coyle, B. A. & Ibers, J. A. (1971). *J. Chem. Soc.* 2146-2151.

[100] Ramachandran, G. N. & Srinivasan, R. (1961). *Nature (London)*, **190**, 159-161.

[101] Read, R. J. (1986). *Acta Cryst.* **A42**, 140-149.

[102] Reinelt, S., Hofmann, E., Gerharz, T., Bott, M. & Madden, D. R. (2003). *J. Biol. Chem.* **278(40)**. 39189-39196.

[103] Robinson, V. L., Buckler, D. R. & Stock, A. M. (2000). *Nat. Struct. Biol.* **7(8)**, 626-633.

[104] Rossmann, M. G. & Arnold, E. (2001). Editor. International tables for crystallography, Volume F. Kluwer academic publishers.

[105] Rossmann, M. G. (2001). *Acta Cryst.* **D57**, 1360-1366.

[106] Sahly, H. & Podschun, R. (1997). *Clinical & Diagnostic Laboratory Immunology.* **4(4)**, 393-399.

[107] Sargsyan, E., Baryshev, M., Szekely, L., Sharipo, A. & Mkrtchian, S. (2002). *JBC.* **277(10)**, 17009-17015.

[108] Schneider, T. R. (2002). *Acta Cryst*. **D58**, 195-208.

[109] Schneider, T. R. & Sheldrick, G. M. (2002). *Acta cryst*. **D58**, 1772-1779.

[110] Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. (2004). *Acta Cryst*. **D60**, 1229-1236.

[111] Sen, J., Goltz, J. S., Konsolaki, M., Schüpbach, T. & stein, D. (2000). *Development*. **127**, 5541-5550.

[112] Segeev, P., Streit, A., Heller, A. & Steinmann-Zwicky, M. (2001). *Dev. Dyn*. **220(2)**, 122-132.

[113] Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol*, **277**, 319-341.

[114] Sheldrick, G. M. (2002). *Z. Kristallogr*. **217**, 644-650.

[115] Sheldrick, G. M. (2002). *TWINABS*, University of Göttingen.

[116] Sheldrick, G. M. (2003). *CELL_NOW*, University of Göttingen.

[117] Six, S., Andrews, S. C., Unden, G. & Guest, J. R. (1994). *J. Bacteriol*. **176**, 6470-6478.

[118] Sparks, R. (1999). *GEMINI*, Bruker AXS Inc. Madison.

[119] Stock, A. M., Robinson, V. L. & Goudreau, P. N. (2000). *Annu. Rev. Biochem*. **69**, 183-215.

[120] Taylor, B. L. & Zhulin, I. B. (1999). *Micro. Mol. Biol. Rev*. **63(2)**. 479-506.

[121] Terwilliger, T. C. (2001). *Acta Cryst*. **D57**, 1763-1775.

[122] Terwilliger, T. C. (2004). *Acta Cryst*. **D60**, 2144-2149.

[123] Thomason, P. & Kay, R. (2000). *J. Cell. Sci*. **113**, 3141-3150.

[124] Tong, L. & Rossmann M. G. (1997). *Methods Enzymol*, **276**, 594-611.

[125] Wadsworth, S. C., Vincent, W. S. 3rd. & Bilodeau-Wentworth, D. (1985). *Nature*. **314(6007)**, 178-80.

[126] Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). *Prot. Eng*. **8**, 127-134.

[127] Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130-135.

[128] West, A. H. & Stock, A. M. (2001). *Trends Biochem Sci.* **26**, 369-376.

[129] Wolanin, P. M., Thomason, P. A. & Stock, J. B. (2002). *Genome Biology*. **3(10)**, 3013.1-3013.8.

[130] Yeates, T. O. (1988). *Acta Cryst*. **A44**, 142-144.

[131] Yeates, T. O. (1997). *Methods Enzymol*. **276**, 344-358.

[132] Yeh, J. I. & hol, W. G. (1998). *Acta Cryst*. **D54**, 479-480.

[133] http://www.doe-mbi.ucla.edu/Services/Twinning/

Publikationen

1. Maskey, R. P., **Sevvana, M.**, Usón, I., Hemke, E. & Laatsch, H. *Gutingimycin: a highly complex metabolite from a marine streptomycete*. (2004). *Angew Chem Int Ed Engl*. 43(10), 1281-1283.

2. Magyar, A., Wolfling, J., Kubas, M., Cuesta Seijo, J. A., **Sevvana, M.**, Herbst-Irmer, R., Forgo, P. & Schneider, G. *Synthesis of novel steroid-tetrahydroquinoline hybrid molecules and D-homosteroids by intramolecular cyclization reactions*. (2004). *Steroids*. 69(5), 301-312.

3. Barnewitz, K., Guo, C., **Sevvana, M.**, Ma, Q., Sheldrick, G. M., Söling, H. -D. & Ferrari, D. M. *Mapping of a substrate binding site in the protein disulfide isomerase-related chaperone Wind based on protein function and crystal structure*. (2004). *J. Biol. Chem*. 279(38), 39829-39837.

4. Than, N, N., Fotso, S., **Sevvana, M.**, Sheldrick, G. M., Fiebig, H. H., Kelter, G. & Laatsch, H. *Sesquiterpene Lactones from Elephantopus scaber*. (2005). *Z. Naturforsch*. 60b, 200-204.

5. Dalai, S., Limbach, M., Zhao, L., Tamm, M., **Sevvana, M.**, Sokolov, V. V. & de Meijere, A. *Highly diasteroselective sequential Michael-Aldol reactions of Methyl-2-chloro-2-cyclopropylideneacetate with Grignard reagents and aldehydes*. (2006). *Synthesis*. 471-479.

6. **Sevvana, M.**, Biadene, M., Ma, Q., Guo, C., Söling, H. -D, Sheldrick, G. M. & Ferrari, D. M. *Structural elucidation of the PDI-related chaperone Wind with the help of mutants*. (2006). *Acta Cryst*. D62, 589-594.

7. Maskey, R. P., Fotso, S., **Sevvana, M.**, Usón, I., Grün-Wollny, I. & Laatsch. *Kettapeptin: Isolation, structure elucidation and activity of a new hexadepsipeptide antibiotic from a terrestrial Streptomyces* sp. (Accepted by *J. Antibiot*, 2006).

8. Lippert, U*., **Sevvana, M***. & Ferrari, D. M. *On the structure and function of protein disulfide isomerase related PDI-D proteins*. (Submitted to *J. Biol. Chem*, 2006). *These authors contributed equally to the work.

9. **Sevvana, M.**, Herbst-Irmer, R. & Sheldrick, G. M. *SAD phasing and refinement of non-merohedrally twinned macromolecular crystals*. (Manuscript under preparation to be submitted to *Acta Cryst A*).

10. **Sevvana, M.**, Vinesh, V., Zweckstetter, M., Reinelt, S., Madden, DR., Sheldrick, G. M., Bott, M., Griesinger, C. & Becker, S. Ligand-induced Switch mechanism regulates Signal Transduction in Sensor Histidine Kinase CitA (Manuscript under preparation).

Konferenzbeiträge

1. **Sevvana, M**. *Towards a novel covalent DNA-Intercalator from the structure of antimetabolite, Gutingimycin*. Oral presentation at 6th Heart of Europe biocrystallography meeting, September 2003.

2. **Sevvana, M**. *Solving non-merohedrally twinned crystals using single wavelength anomalous scattering*. Oral presentation at 37th course of International school of crystallography at Erice, Sicily, May 2005.

3. **Sevvana, M**. Herbst-irmer, R. & Sheldrick, G. M. *Solving non-merohedrally twinned crystals using single wavelength anomalous scattering*. Poster presentation at 37th course of International school of crystallography at Erice, Sicily, May 2005.

4. **Sevvana, M.**, Ma, Q., Barnewitz, K., Guo, C., Söling, H. -D., Ferrari, D. M. & Sheldrick, G. M. *Structural and functional analysis of PDI-related proteins*. (2005). *Acta Cryst*. A61, C250. Poster at XX congress of the International union of Crystallography, Florence, August 2005.

5. Dix, I., **Sevvana, M.**, Bunkoczi, G., Debreczeni, J. Sheldrick, G. M. *The EU BIOXHIT standard test crystal*. (2005). *Acta Cryst*. A61, C147. Poster at XX congress of the International union of Crystallography, Florence, August 2005.

Lebenslauf

Persönlische daten

Name	Madhumati Sevvana
Geburtsdatum	29th July 1978
Geburtsort	Visakhapatnam
Staatsangehörigkeit	Indisch
Familienstand	Verheiratet

Schulbildung

1991-1993	Central board for Secondary Education, India
1993-1996	Board of Intermediate Education, India, mit abschluss abitur

Hochschulstudium

1996-1999	**B.Sc in Chemistry, Biochemistry and Biotechnology** Andhra University, India
1999-2001	**M.Sc in Biochemistry** University of Hyderabad, India
2001-2002	**M.Sc in Molecular Biology** Max-Planck Research School, Göttingen
2002-2003	Diplomarbeit am Lehrstuhl für Strukturchemie zum Thema: "Crystal structure analysis of biologically active natural products"

Promotion

2003-2006	Dissertation am Lehrstuhl für Strukturchemie der Georg-August-Universität Göttingen im Arbeitskreis von Prof. George M. Sheldrick zum Thema: "Crystallographic analysis of Pathological Crystals, Periplasmic Domain of ligand-free CitA Sensor Kinase cand PDI-related Chaperones."