Crystal Structure of Wind, a PDI-Related Protein Required for *Drosophila melanogaster* Dorsal-Ventral Development

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Georg-August-Universität zu Göttingen

vorgelegt von

Qingjun Ma

aus Pingyi County, P.R. China

Göttingen 2003

D 7

Referent: Prof. G.M. Sheldrick, Ph.D.

Korreferent: Prof. Dr. R. Ficner

Tag der mündlichen Prüfung: 2. Juli 2003

Acknowledgements

The present thesis was carried out at the Department of Structural Chemistry, University of Göttingen, under the supervision of Professor Ph.D. George M. Sheldrick.

In the first place, I would like to thank Professor Ph.D. George M. Sheldrick for his interest, constant guidance, motivation and invaluable criticisms and for the freedom he offered me in the work.

Priv. Doz. Dr. Isabel Usón deserves a great many thanks for her tutorship on this work, and for ready discussion, support, ideas and critical reading of manuscripts.

My warm thanks to Professor Dr. Hans-Dieter Söling, Dr. David M. Ferrari and Mr. Chaoshe Guo at the Max-Planck-Institute for Biophysical Chemistry, for their collaborations and providing me good proteins to do this work. Mr. Chaoshe Guo collected many literatures helping me write this thesis.

I thank Professor Dr. Ralf Ficner at the Department of Molecular Structural Biology, University of Göttingen, for accepting the job as co-referee.

I thank Dr. Regine Herbst-Irmer and Dr. Thomas R. Schneider for their professional helps during my study.

Thanks to all my present and former colleagues for the pleasant working atmosphere at the lab. I specially want to thank Dr. Thomas Pape, who took much time to read and correct my thesis, and helped me a lot in computer techniques; Dr. Ralph Krätzner, who gave me many warmhearted helps in my study and life; Miss Ilke Müller gave good tips for experiment; Miss. Eftichia Alexopoulos, Mr. Fabio Dall'Antonia, Mr. Gabor Bunkoczi and Mrs. Judit Debreczeni for their constant help throughout my work.

I additionally want to thank Mr. Helmut Dehnhardt and Dr. Mathias Noltemeyer for their helps in X-ray techniques. Thanks to the staff at DESY, Hamburg, especially Dr. Ehmke Pohl, providing the synchrotron beamline for diffraction experiment.

I thank Professor Dr. med. Kurt von Figura, Dr. Bernhard Schmidt, Dr. Thomas Dierks, Mr Qinghua Fang and Mr. Jianhe Peng for their helps when I worked at the Department of Biochemistry II, University of Göttingen.

Last but not least, I want to thank my wife and my parents for their support and affection. And thanks to all my friends for giving me a lot of general support all the time.

Abbreviations

apoLp-III	apolipophorin III
ASA	accessible surface area
CC	correlation coefficient
CCD	charge coupled device
Cryo-EM	cryo-elctron microscopy
DESY	Deutsches Elektronen Synchrotron
DV	dorsal-ventral
E.coli	*Escherichia coli*
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
EMBL	European Molecular Biology Laboratory
Eqn.	equation
ER	endoplasmic reticulum
ERGIC	ER-Golgi intermediate compartment
ERp	endoplasmic reticulum protein
FOM	figure of merit
GD	Gastrulation-defective
GPX	glutathione peroxidase
GRX	glutaredoxin
GRP94	glucose-regulated protein 94
GSH	glutathione (reduced)
GSSG	glutathione (oxidized)
GST	glutathione S-tansferase
HEPES	N-2-Hydroxyethylpiperazine-N'-2-ethanesulfonic acid
Hsps	heat-shock proteins
kDa	Kilo Dalton
LDLR	low density lipoprotein-receptor
LS	least-squares
M	molar, mol/l
MAD	multi-wavelength anomalous diffraction
MALS	multiangle light scattering
MES	2-(N-Morpholino)-ethanesulfonic acid
ML	maximum likelihood
MIR	multiple isomorphous replacement
MR	molecular replacement
MW	molecular weight
NCS	non-crystallographic symmetry
NMR	nuclear magnetic resonance
PCR	polymerase chain reaction
PDB	Protein Data Bank
PDI	protein disulfide isomerase
PDI-D	PDI-related protein PDI-D
PDI-Dα	PDI-related protein PDI-Dα, redox active
PDI-Dβ	PDI-related protein PDI-Dβ, redox inactive
PEG	polyethylene glycol
PPIase	protein prolyl isomerase
PSMF	Patterson superposition minimum function
QC	quality control
RAP	receptor-associated protein
RER	rough endoplasmic reticulum

RIP	radiation-damage introduced phasing
SAD	single-wavelength anomalous diffraction
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SER	smooth endoplasmic reticulum
SIR	single isomorphous replacement
SIRAS	single isomorphous replacement anomalous scattering
SP	signal peptide
Tris	Tris-(hydroxymethyl)aminomethane
TRX	thioredoxin
TRX-like	thioredoxin-like
v/v	volume/volume

One and three letter symbols for the for amino acids:

A	Ala	alanine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
Y	Tyr	tyrosine

Table of contents

1. Introduction

1.1. Biological Background

1.1.1. The endoplasmic reticulum and the Golgi apparatus

The endoplasmic reticulum (ER) is a cellular organelle, enclosed by a single continuous phospholipid bilayer membrane, accounting for more than 10% of the cell volume (Fig. 1-1). The enclosed 'sac' is called the ER lumen, the internal space of the ER. The ER membrane typically makes up more than half of the total membrane in the cell and is located between the nucleus and the cytosol and specifically the Golgi apparatus. There are two basic types of ER: the rough endoplasmic reticulum (RER) and the smooth endoplasmic reticulum (SER).

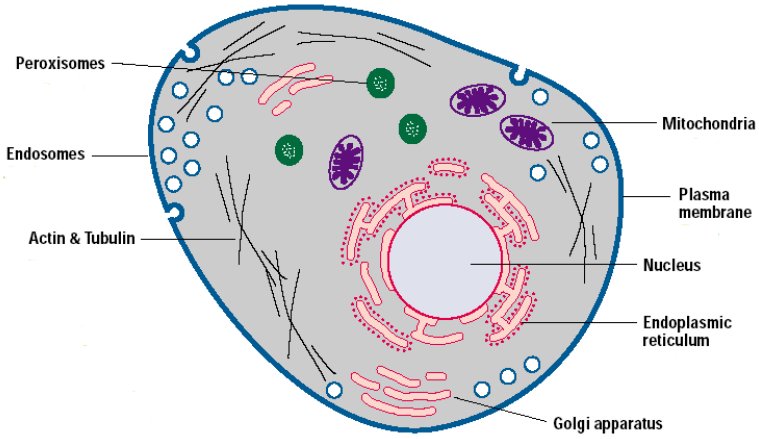


Figure 1-1 A simple diagram of the cell.

The RER is associated with ribosomes, thus having a studded appearance of the cytoplasmic face of its membrane. The membranes tend to be in 'sheets' or flatten sacs called cisternae and are connected to outer nuclear membrane. The major functions of RER are: Synthesis and segregation of proteins; Addition of N-linked oligosaccharides (core glycosylation); Protein folding and oligomerization; Lipid synthesis: Phospholipids and cholesterol.

The SER lacks ribosomes. It consists of intertwined tubules and is connected to RER. Its functions include: lipid synthesis, fat metabolism, detoxifications (e.g., barbiturates, alcohol), glycogen metabolisms, and synthesis of steroids, cholesterol, triglycerides, steroid hormone production. SER is also connected with calcium storage in muscle cells, which has an affect on muscular contraction and hence on body movements.

The Golgi apparatus consists of stacks (5-20) of flattened saccules, similar to hollow pancakes. Its inner (*cis*) face is directed toward the ER while the outer (*trans*) face is directed outwards. The Golgi apparatus can be described as the post office in the cell. Its principal role is to package molecules for transport to the cell surface and vacuoles. It also does proline hydroxylation, O-linked glycosylation, N-linked glycan modification.

1.1.2. Quality control in the ER

Generally, proteins function correctly only if matured and localized in the right cellular compartment. So, proteins are exported from the ER to their destinations after the synthesis. This "export" process is strictly controlled.

The process of ‘quality control’ (QC) in the ER involves a variety of mechanisms that collectively ensure that only correctly folded, assembled and modified proteins are transported along the secretory pathway. In contrast, nonnative proteins are retained and eventually targeted for degradation (Ellgard *et al.*,1999; Ellgaard and Helenius, 2001). Such a quality control process includes: proper folding, addition of carbohydrates, disulfide bond formation, prolyl *cis-trans* isomerization, proteolytic cleavages and assembly into multimeric proteins etc.

Those quality control mechanisms that apply to all proteins expressed in the ER have been termed ‘primary quality control’. No specific signals or amino-acid sequence motifs are needed for primary QC. The molecular chaperones and foldases used in primary QC are abundant in the ER. Chaperone (Ellis, 1987) is a protein that catalyzes the correct folding of newly synthesized or denatured proteins into their native conformations, such as the heat-shock proteins (Hsps) (Hartl, 1996), GroEL, BiP (Munro and Pelham, 1986), glucose-regulated protein 94 (GRP94), calnexin (Bergeron et al., 1994), calreticulin (Meldolesi et al., 1996) and others. Foldases assist in the correct folding of polypeptides, including protein prolyl isomerase (PPIase) (Fischer, 1994) and protein disulfide isomerase (PDI), which accelerate rate-limiting prolyl *cis-trans* and disulfide bond isomerization reactions, respectively. The chaperons and foldases are not distinct from each other. Not only may PPIase and PDI proteins act as chaperones, but also chaperones may have catalytic properties. These chaperones and/or foldases have the capacity to recognize properties common to

nonnative proteins such as exposed hydrophobic areas. Even minor deviations from the native conformation, because of incomplete folding or misfolding, lead to a protein being bound by one or more of these factors and therefore to its retention in the ER. In fact, there are various severe diseases deriving from the endogenous proteins containing mutations or defects that affect folding and lead to protein accumulation in the ER.

In order to be secreted, many proteins must fulfill criteria beyond those that are imposed by the primary QC system. The term secondary QC refers to various selective mechanisms that regulate the export of individual protein species or protein families (Ellgaard et al., 1999). It comprises a rapidly growing list of protein specific factors. Each of the factors involved has its own specific recognition mechanism and many of these factors interact with the folded cargo proteins or late folding intermediates. According to the acting mechanism, these assistant proteins are roughly sorted into three classes (Ellgaard and Helenius, 2003). Those proteins that are needed to fold and assemble specific proteins as 'outfitters', those needed to accompany proteins out of the ER as 'escorts' and those needed to provide signals for intracellular transport as 'guides'. The group of outfitters includes specialized chaperones and enzymes such as Nina A, a peptidyl-prolyl *cis/trans* isomerase that ensures the transport competence of specific rhodopsins in *Drosophila melanogaster* (Stamnes *et al.*, 1991). A well-known escort is the receptor-associated protein (RAP), which binds to members of the low density lipoprotein-receptor (LDLR) family in the ER and escorts them to the Golgi complex to protect them from premature ligand-binding in the early secretory pathway (Bu, 2001). A lectin, known as ER-Golgi intermediate compartment (ERGIC)-53 provides a example for ‘guide’, which cycles between the ER and the Golgi complex and seems to act as a transport receptor for certain proteins that carry high-mannose N-linked glycans (Appenzeller *et al.*, 1999).

1.1.3. PDI and PDI family

The lumen of the ER offers an oxidizing environment with glutathione being the main mediator. The ratio of the reduced glutathione (GSH) to the oxidized glutathione (GSSG) in the ER is 1-3:1, for comparison the ratio in the cytosol is normally 30-100:1 (Hwang *et al*., 1992). This ratio in the ER is strikingly similar to the optimal conditions for *in vitro* protein folding and formation of disulfide bonds. In fact, Protein folding in the ER is often associated with the formation of native disulfide bonds, and this is facilitated by an enzyme, called protein disulfide isomerase (PDI) (Freedman et al., 1994).

PDI (EC 5.3.4.1) is a member of the thioredoxin superfamily and is highly abundant in the lumen of the ER. PDI comprises four thioredoxin-like (TRX-like) structural domains (Fig. 1-2), **a**, **b**, **b'** and **a',** a linker region between **b'** and **a'**, an N-terminal signal peptide (SP) to permit translocation of the protein into the ER and a C-terminal **c** domain that is rich in acidic amino acids and contains the KDEL ER retention signal (Munro and Pelham, 1987). The **a** and **a'** domains contain a –CGHC- motif and are redox active, while **b** and **b'** lack such a motif and are redox inactive, which may be important for peptide binding (Klappa, *et al*., 1998).

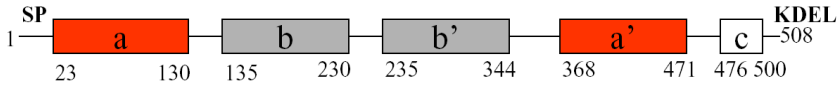


Figure 1-2 The composition of PDI.

PDI is usually isolated as a homodimer, although monomers and homotetramers are also known to occur. PDI is a multi-functional protein that catalyzes the formation and isomerization of disulfide bonds during protein folding. PDI exhibits chaperone-like activity as well, which is independent of its redox/isomerase activities. It even facilitates folding for some proteins that contain no disulfide bonds (Wang and Tsou, 1993).

Human and yeast cells both contain several PDI homologues in the ER**.** The PDI homologues are characterized by the presence of one or more domains with sequence homology to thioredoxin (TRX), a signal sequence and a (K/H)DEL or similar ER localization signal**.** They are different in number and organization of their thioredoxin

domains. The common members of the PDI family are summarized in table 1-1. PDI itself works in the primary quality control level. Nevertheless, other PDI-related proteins may work in the primary and/or second control level.

protein	MW(kD)	domain structure	redox-active-site sequence
PDI	55	a-b-b'-a'-c	-CGHC-
P5	46	a^0-a-b-c	-CGHC-
ERp72	71	c-a^0-a-b-b'-a'	-CGHC-
ERp57	54	a-b-b'-a'	-CGHC-
PDIp	55	a-b-b'-a'	-CGHC-/-CTHC-
PDIR	57	b-a^0-a-a'	-CSMC-/-CGHC-/-CPHC-
ERp28(PDI-Dβ)	26	b-D	NONE
Dd-PDI(PDI-Dα)	38	a^0-a-D	-CGHC-

Table 1-1 Summary of PDI-related proteins.

1.1.4. Thioredoxin fold

The PDI-related proteins mainly consist of several TRX-like domains, except for the PDI-D subfamily, which contains a unique D-domain. The thioredoxin fold (Martin, 1995) is a ubiquitous structural motif adapted by many proteins with various functions. It is defined in the SCOP database as: *core: 3 layers, a/b/a; mixed beta-sheet of 4 strands, order 4312; strand 3 is antiparallel to the rest*. In detail, the fold includes an N-terminal βαβ motif and a C-terminal ββα motif connected by a third helix. The beta-strands in the N-terminal motif run parallel while those from the C-terminal motif run antiparallel. The alpha helices of the N- and C-terminal motifs line up parallel on one side of the sheet. The alpha helix connecting the N- and C-terminal motifs is located on the opposite side of the beta-sheet to the other two helices and is perpendicular to them. It has been identified in the three-dimensional structures of proteins from six classes: thioredoxin (TRX), glutaredoxin (GRX), glutathione S-transferase (GST), DsbA, glutathione peroxidase (GPX) and PDI-related protein (Fig. 1-3). The thioredoxin domain in each protein is not homologous in amino acid sequence, but the spatial structure is very similar.

The thioredoxin fold comprises about 80 residues, but each of the proteins containing it has inserts in addition to the fold. There are some points in the thioredoxin fold, where additional structure elements can be inserted without disrupting the overall fold (Fig. 1-4a).

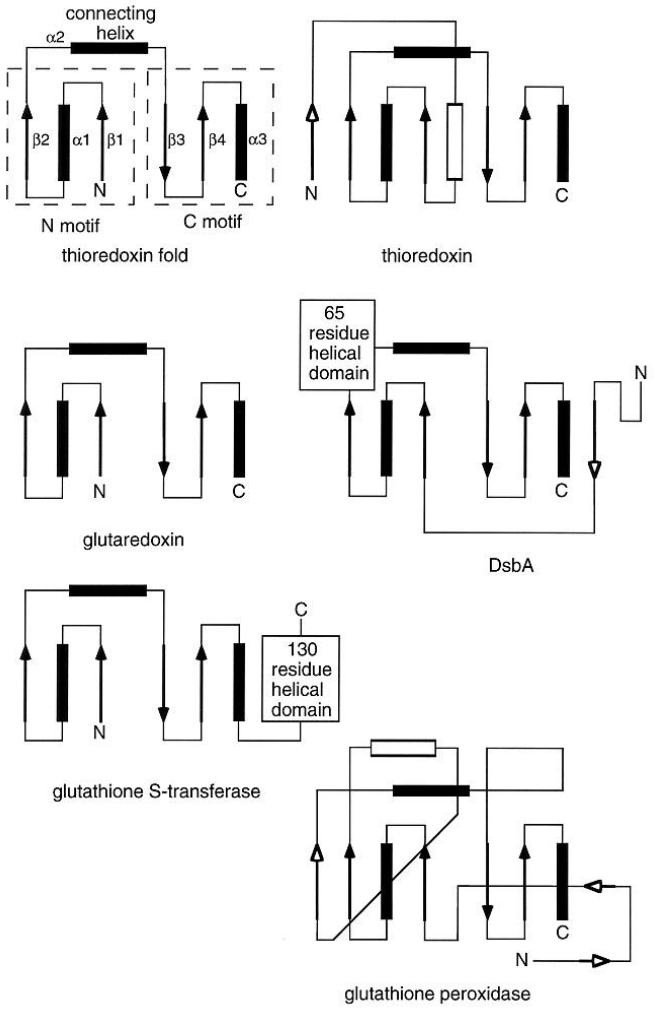
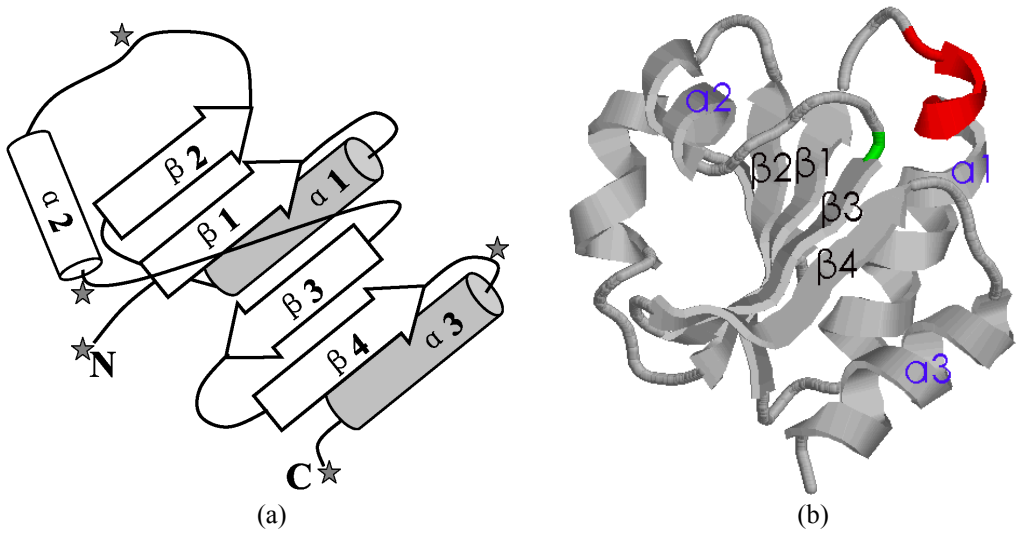


Figure 1-3 The architecture of some TRX-like proteins. β-sheet strands are drawn as arrows and α-helices as rectangles. The secondary structure elements forming the thioredoxin fold are shown in black. (Martin, 1995)



(a) (b)

Figure 1-4 Thioredoxin fold. (a). The insertion points in thioredoxin fold. (b). A ribbon diagram of TRX. Only the secondary structure elements forming the thioredoxin fold are marked using the same nomenclature in thioredoxin fold.

There are some other features for the thioredoxin domains (Fig. 1-4b).

(1) For the proteins that have redox/isomerase activities, the redox-active CXXC group is located at the N terminus of the helix α1 of the thioredoxin fold. This region is conserved in both sequence and structure.

(2) Generally, a *cis*-proline is localized between α2 and β3 of the thioredoxin fold. This region is called *cis*-Pro loop and is essential for maintaining the local substructure.

(3) There seems to be a common substrate binding site in the thioredoxin domain. It is located in a region on the tips of the beta sheet, where the CXXC and/or *cis*-proline are localized.

1.1.5. Dorsal-ventral polarization in *Drosophila* embryo development

The embryo of *Drosophila melanogaster* (Fig. 1-5) is initially symmetrical. However, normal *Drosophila* embryo development needs the correct polarization, i.e., left-right, anterior-posterior, dorsal-ventral (DV) polarization. The establishment of the DV axis occurs during oogenesis and is a result of communication between the germ-line-derived oocyte and the somatically derived follicle cells of the ovary (Fig. 1-6).

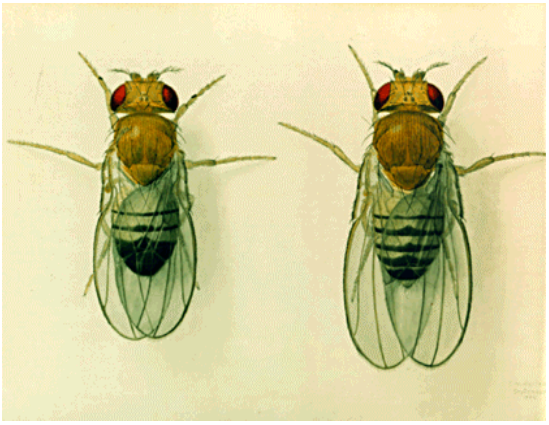


Figure 1-5 *Drosophila melanogaster*

The DV polarization process is launched by the communication from oocyte to the follicle cells, called the Gurken-EGFR pathway. Initially, the oocyte nucleus moves to the anterior dorsal part of the cell. It synthesizes the *gurken* mRNA between the oocyte and follicle cells. Then the dorsalizing signal, Gurken (Neuman-Silberberg and Schupbach, 1993), a growth factor homologous to epidermal growth factor (EGF), accumulates around the oocyte nucleus and then is secreted to the follicle cells, which differentiates to a dorsal morphology later. This signal is received by the follicle cells via the homologue of the human epidermal growth factor receptor (EGFR, Wadsworth *et al.,* 1985). EGFR is expressed in all follicle cells, and is only activated in the dorsal follicle cells receiving the Gurken signal.

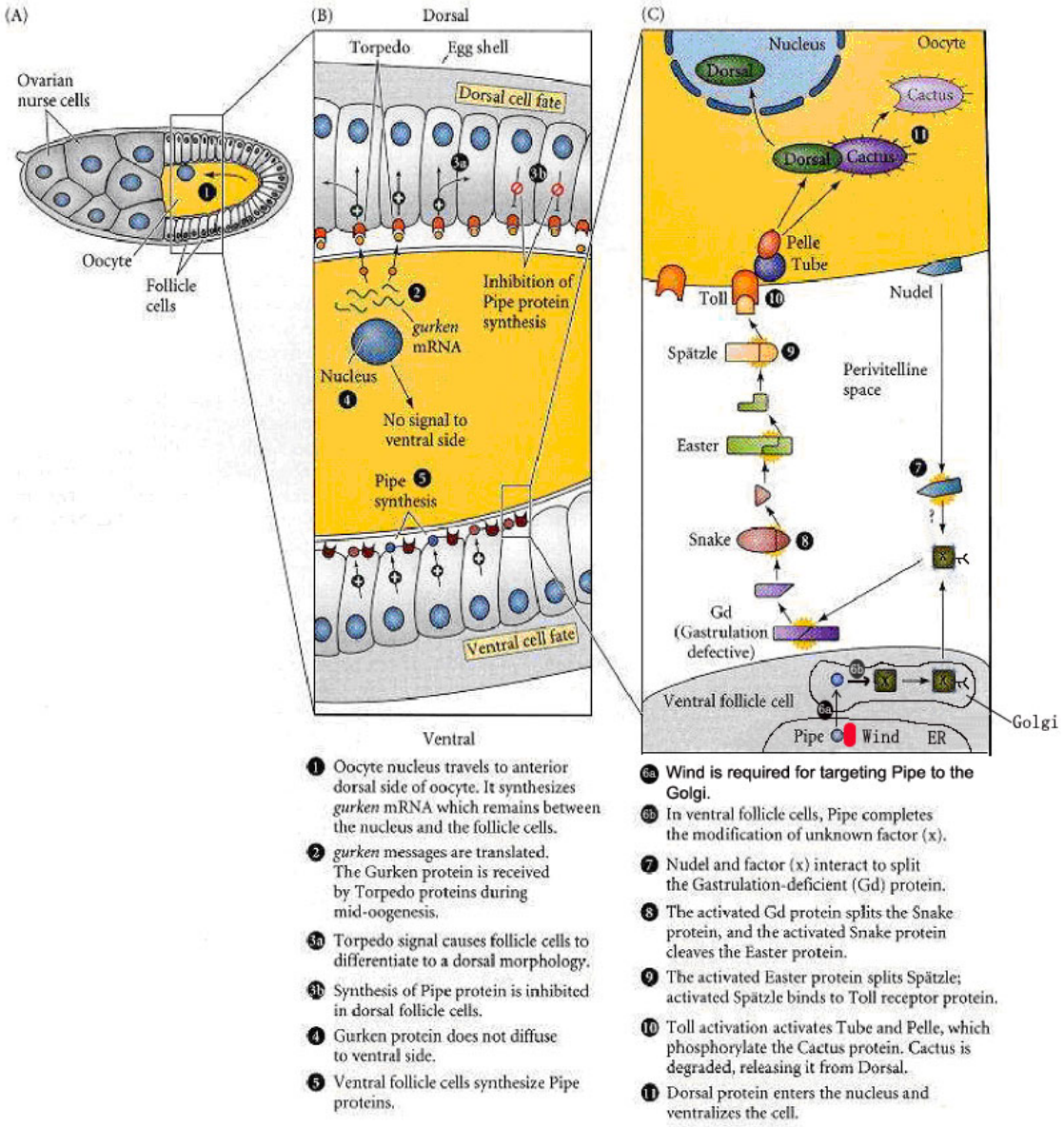


Figure 1-6 The dorsal-ventral polarization process in *Drosophila* embryo (Adapted from Molecular Biological course of Fritz Aberger and modified)

Then follows the second communication, from the follicle cells to the oocyte, which ultimately leads to the specification of the dorsal-ventral axis of the embryo. In a word, this process is realized via a proteolytic cascade (Morisato and Anderson, 1995), which results in formation of the nuclear gradient of the transcription factor Dorsal. Dorsal mRNA is supplied by the mother but protein gradient is generated after fertilization. Nevertheless, this process is quite complex. Besides Dorsal, a dorsal group of at least 11 other genes and their expressed proteins are involved.

Wind (or windbeutel), *pipe* and *nudel* are the only three genes expressed by mother follicle cells and work in the immediate downstream after the Gurken-EGFR pathway. The gene *wind* encodes a putative ER resident PDI-related protein, Wind, which is required for localizing Pipe to the Golgi apparatus (Konsolaki and Schüpbach, 1998; Sen *et al.,* 2000). The gene *pipe* encodes the protein Pipe, a homologue of mammalian 2-O-sulfotransferase (2-OST) (Sergeev *et al.,* 2001). Pipe is the critical factor to define the DV symmetry (Sen *et al.*, 1998). The gene *nudel* encodes a modular protein with an extracellular matrix domain and a serine protease domain (Hong and Hashimoto, 1995). It has been suggested that Nudel is secreted by the follicle cells and may possibly be incorporated in the vitelline membrane, thus specifying the site of generation of the active Spätzle ligand, after fertilization of the oocyte. Wind and Nudel are not spatially restricted to the ventral follicle and expressed in all follicle cells. However, the expression of Pipe is negatively regulated by EGFR. Thus Pipe is spatially restricted to the ventral follicle cells and its expression in dorsal follicle cells is inhibited by EGFR that is only activated in the dorsal side (Sen *et al.,* 1998). In the Golgi apparatus, Pipe modifies an as yet unidentified proteoglycan x, which combines Nudel to trigger a protease cascade leading to the DV polarization of the embryo. So, Pipe plays a pivotal role in the process that defines the DV axis of the embryo and that its spatially regulated activity may provide the link between the establishment of DV polarity in the follicle cells and the transmission of DV patterning information to the developing egg and future embryo (Sen, 1998).

Gastrulation-defective (Gd), Snake, Easter are expressed in the oocyte and are all serine proteases. They are released into the perivitelline space. Gd is cleaved by Nudel-x and is activated. Subsequently, Gd cleaves Snake, and Snake cleaves Easter. At the end of the protease cascade, Easter cleaves Spätzle (Morisato and Anderson, 1994). This reaction apparently occurs shortly after fertilization and only on the ventral side of the embryo. Cleaved Sp**ä**tzle is the ligand for Toll (Hashimoto *et al.*, 1988), a receptor on the egg membrane. The uniformly distributed Toll is thus only activated in the ventral side. Dorsal is held in the egg cytoplasm by Cactus. An entire pathway is then "designed" to separate Cactus from Dorsal in the ventral region. Toll signaling activates Pelle (a protein kinase) and tube (function unknown yet). Pelle phosphorylates Cactus. Phosphorylated Cactus is degraded and

Dorsal is free to enter the nucleus. This regulatory process leads to the high nucleus gradient of Dorsal in the ventral side and low gradient in the dorsal side. Fate of the cells is determined by Dorsal gradient in the nucleus. Genes that have a low affinity Dorsal binding region are activated. Genes that have a high affinity Dorsal binding region are inhibited in conjunction with other enhancer regions. Target genes also influence one another. All of these finally lead to the formation of dorsal and ventral tissues.

1.1.6. Wind

Wind, encoded by the gene *wind*, is a PDI-related ER resident protein and is required for *Drosophila* embryo DV polarization.

In structure, Wind and its homologous proteins human ERp28 (Ferrari *et al.*, 1998), rat ERp29 (Liepinsh *et al.*, 2001) etc. belong to the PDI-D subfamily (Ferrari and Söling, 1999). Members of PDI-D subfamily are either redox-active (PDI-Dα) or redox-inactive (PDI-Dβ), but all are characterized by a C-terminal alpha-helical domain of about 110 amino acids (termed the D-domain), the function of which is unknown yet. These proteins are the only known PDI members to display a domain not related to TRX.

The complete Wind has 257 residues, including a putative 21-residued signal peptide at the N-terminus, and a KEEL ER retention signal in the C-terminus. A second KEEL near the C-terminus seems to be another retention signal (Fig. 1-7).

MMHILVTLLLVAIHSIPTTWAVTCTGCVDLDELSFEKTVERFPYSVVKFDIAYPYGEKHE
AFTAFSKSAHKATKDLLIATVGVKDYGELENKALGDRYKVDDKNFPSIFLFKGNADEYVQ
LPSHVDVTLDNLKAFVSANTPLYIGRDGCIKEFNEVLKNYANIPDAEQLKLIEKLQAKQE
QLTDPEQQQNARAYLIYMRKIHEVGYDFLEEETKRLLRLKAGKVTEAKKEELLRKLNILE
VFRVHKVTKTAPEKEEL

Figure 1-7 The complete sequence of Wind. The putative signal peptide is colored gray; the CTGC motif is colored cyan; the KEEL ER retention signal is colored yellow.

The mature wind (signal peptide cleaved) has a theoretical pI of 5.86 and Mw of 27100.98 Dalton. There are three cysteines in Wind. The two near the N-terminus form a CXXC motif in sequence. From secondary structure prediction, the protein contains two

distinct domains: one TRX-like domain (called b-domain) at N-terminus and one alpha-helical D-domain at C-terminus.

Pipe is the key patterning protein in *Drosophila* DV development. It works properly only in the Golgi apparatus. It has been proved at the genetic level that Wind is required for the correct localization of Pipe to the Golgi. Wind deficient female flies show an aberrant distribution of Pipe protein and the embryos have the dorsal fate (Sen *et al.*, 2000). When expressed in COS-7 cells, Pipe is retained in the ER in a presumably inactive form in the absence of Wind. On simultaneous over-expression of Wind, a clear redistribution of Pipe to the Golgi was observed. Wind is considered to work as a putative chaperon here. So far, it seems to act specifically on Pipe, so it could be a factor involved in the second QC process of ER. Most likely, Wind functions as either an 'outfitter' or an 'escort' for Pipe transport to the Golgi (Sen *et al.*, 2000). The relationship of Pipe and Wind may be analogous that of receptor associated protein (RAP) and low-density lipoprotein (LDL) receptor (Bu, 2001). A physical interaction between RAP and LDL receptor prevents aggregation and premature ligand binding in the ER, with RAP escorting LDL receptor to the Golgi. Here, the complex dissociates and RAP is retrieved to the ER by the KDEL receptor (Pelham, 1990). Wind may work as a folding catalyst or chaperone for Pipe folding, a likely prerequisite for the migration of Pipe from the ER to the Golgi, allowing only spatially and temporally appropriate oligosaccharide modification by Pipe.

1.2. Crystallographic background

Nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography and cryo-electron microscopy (Cryo-EM) are three techniques that can provide 3-D information for macromolecular structures. The structure of Wind was determined by X-ray crystallography. Some basic crystallography knowledge is underlined here, mainly focusing on protein crystallography.

1.2.1. Sample preparation for crystallization

Generally, protein samples can be prepared from plant, animal or microbes directly. However, a modern method is using gene recombination techniques. First, the target gene is

selected and amplified by polymerase chain reaction (PCR). These gene fragments are cloned into a proper vector, and then transferred to *E. coli*, insect cell or cell-line for expression. Large amounts of proteins can usually be obtained by this method in a short time. Then these proteins are purified with various chromatography techniques, such as affinity chromatography, size-exclusion chromatography, ion-exchange chromatography, etc. The purified and concentrated (e.g. 5-20mg/ml) protein samples can then be used for crystallization.

1.2.2. Crystallization

Protein crystallization occurs when the concentration of protein in solution is greater than its limit of solubility and so the protein is in a supersaturated state. It is a multifactor process, affected by protein purity and concentration, temperature, pH, precipitants, additives and so on. These parameters have to be determined by trial-and-errors. The purity of a protein is the most important requisite. The purer, the better. There are three stages of crystallization: nucleation, growth, and cessation of growth (Fig. 1-8). The solute concentration for crystal growth is normally lower than that for nucleation. The commonly used crystallization methods include: vapor diffusion (Fig. 1-9), dialysis, microbatch, seeding etc.

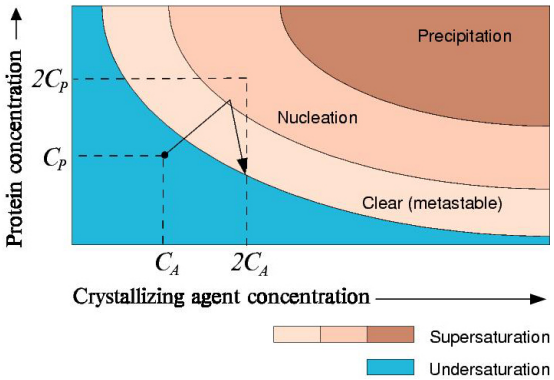


Figure 1-8 Phase diagram of crystallization for a typical protein. Crystal form nuclei at the Nucleation zone, then grow in the metastable zone.
(Adapted from http://perch.cimr.cam.ac.uk/Course/)

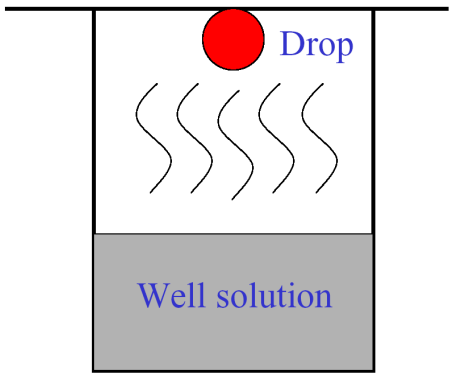


Figure 1-9 hanging drop vapor diffusion method. Vapor diffusion takes place in a sealed container.

1.2.3. X-ray diffraction

X-rays are electromagnetic waves with typical photon energies in the range of 100 eV - 100 keV. Generally, only specific wavelength X-rays in the range of 0.7 to 2.5 Å are used in protein crystallography. Because the wavelength of X-rays is comparable to the interatomic distances, they are ideally suited for probing the structural arrangement of atoms in the crystal. When a crystal is placed in the path of an X-ray beam its atoms act, owing to the forced vibrations of the electrons, as secondary sources emitting X-rays to each direction. The frequency and wavelength of these emitted rays are identical to those of the incident beam. Because a crystal is constructed of atoms or molecules arranged in a regular spatial pattern, only in certain directions the individual scattered wavelets may recombine in phase to produce a strong reinforced but deviated beam (Fig 1-10, 1-11). This is called diffraction. The diffraction angles are determined by the crystal lattice, while the amplitudes and phases of the diffracted waves are determined by the structure of the cell content.

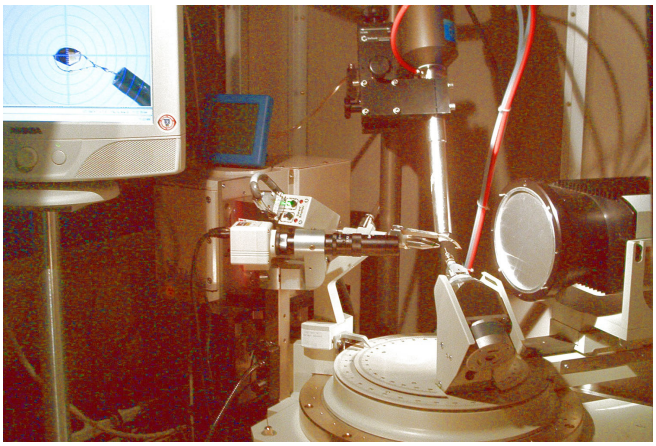


Figure 1-10 The smart-6000 in Göttingen. The crystal is held in a loop.

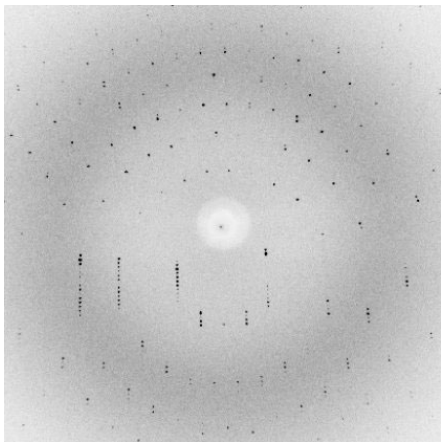


Figure 1-11 The diffraction from a real crystal.

The diffractions can be simply regarded as the reflections by lattice planes in the crystal, so they are also called reflections in crystallography. The scattering from the crystal unit cell is described with structure factor $\mathbf{F}_{hkl}$. $\mathbf{F}_{hkl}$ is a vector, consisting of both amplitude $|\mathbf{F}_{hkl}|$ and phase θ. $\mathbf{F}_{hkl} = |\mathbf{F}_{hkl}|\exp(i\theta)$. If ρ_{xyz} (where x, y, z are fractional coordinates) is the electron density in the unit cell, then $\mathbf{F}_{hkl}$ is the Fourier transform of ρ_{xyz}:

$$\mathbf{F}(hkl) = V\int_x\int_y\int_z \rho(xyz)\exp[2\pi i(hx+ky+lz)] \qquad (1\text{-}1a)$$

When the atomic scattering factor $\mathbf{f}_j$ is used, $\mathbf{F}_{hkl}$ can also be represented as:

$$\mathbf{F}(hkl) = \sum_j^N \mathbf{f}_j \exp[2\pi i(hx_j+ky_j+lz_j)] \qquad (1\text{-}1b)$$

And ρ_{xyz} is the inverse Fourier transform of $\mathbf{F}_{hkl}$:

$$\rho(xyz) = \frac{1}{V}\sum_h\sum_k\sum_l \mathbf{F}(hkl)\exp[-2\pi i(hx+ky+lz)] \qquad (1\text{-}2)$$

1.2.4. Phasing

From the Eqn. 1-2, once the amplitudes and phases of the structure factors are known, the electron density in the cell can be calculated. During a standard protein crystallography experiment only the intensities of the diffracted X-ray beams are recorded, from which the amplitudes can be obtained. Unfortunately, the relative phases of these wavelets, crucial for reconstructing the image of the molecule, are lost. This is called "phase problem", which is the central problems in crystallography. Many methods have been developed to deduce the phases for the reflections, including single isomorphous replacement (SIR), mutiple isomorphous replacement (MIR), single-wavelength anomalous dispersion (SAD), multi-wavelength anomalous dispersion (MAD), SIR including anomalous scattering (SIRAS), statistically based direct methods, molecular replacement (MR) methods, etc.

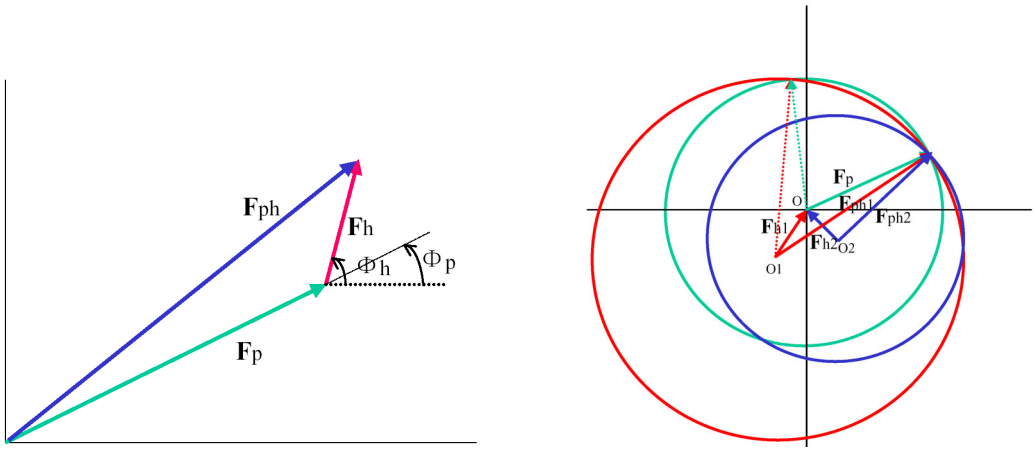
1.2.4.1. SIR and MIR

One traditional method to solve the phase problem is isomorphous replacement method. The idea is that introducing heavy atoms into the crystal structure leads to changes in the diffraction intensities. Then, phase information could be extracted from the intensity differences if the native and derivative crystals are isomorphous.

If only one heavy atom derivative is available, SIR method can be used to solve the phase problem. Applying the cosine law in the phase triangle (Fig. 1-12a), we get

$$|\mathbf{F}_{ph}|^2 = |\mathbf{F}_p|^2 + |\mathbf{F}_h|^2 + 2|\mathbf{F}_p||\mathbf{F}_h|\cos\alpha \qquad (1\text{-}3)$$

where, $\alpha = \Phi_p - \Phi_h$.



(a) (b)

Figure 1-12 Phasing by Isomorphous replacement methods. (a). Harker construction for SIR phasing.(b). Harker construction for MIR phasing.

There are two possible values for α, which is called two value ambiguity. However, it is still possible to solve the phase problem combing SIR with the powerful density modification methods.

Once a second heavy atom derivative is available, we can draw another phase triangle (Fig. 1-12b). Then the value of α can be determined uniquely. This is MIR method.

Another novel method is called radiation-damage introduced phasing (RIP). Due to the radiation damage, the diffraction waves are different before and after radiation. The theory can be thought similar to SIR phasing.

1.2.4.2. SAD and MAD

The normal dispersion occurs only if the free electron model is provided, where the atomic scattering factor is real and proportional to the number of electrons of the atom. The reflections obey the Friedel law: $|\mathbf{F}_{hkl}|=|\mathbf{F}_{\overline{hkl}}|$, and $\phi_{hkl}=-\phi_{\overline{hkl}}$. However, electrons are not really free but bound to the atom nucleus. When X-ray energy is near the atom absorption edges, the anomalous dispersion will occur. And the atomic scattering factor becomes a complex number: $f = f_0 + f' + if''$, where f_0 is the normal scattering factor, f' and f'' are the real and imaginary dispersion corrections, respectively. A direct result is that the amplitudes of the Friedel pairs are not necessarily identical. The differences between them can be used for phasing with the SAD and MAD methods.

The MAD method provides the information to extract all variable values for phasing (Fig. 1-13a):

$$|\mathbf{F}^{\pm}|^2 = |\mathbf{F}_T|^2 + a|\mathbf{F}_A|^2 + b|\mathbf{F}_T||\mathbf{F}_A|\cos\alpha \pm c|\mathbf{F}_T||\mathbf{F}_A|\sin\alpha \qquad (1\text{-}4)$$

where, $a = (f''^2+f^2)/f_0^2$, $b = 2f'/f_0$, $c = 2f''/f_0$, $\alpha = \Phi_T-\Phi_A$. a, b and c are different for each wavelength. Provided that $|\mathbf{F}^{\pm}|$ has been measured at two or more wavelengths we can extract $|\mathbf{F}_T|$, $|\mathbf{F}_A|$ and α for each reflection.

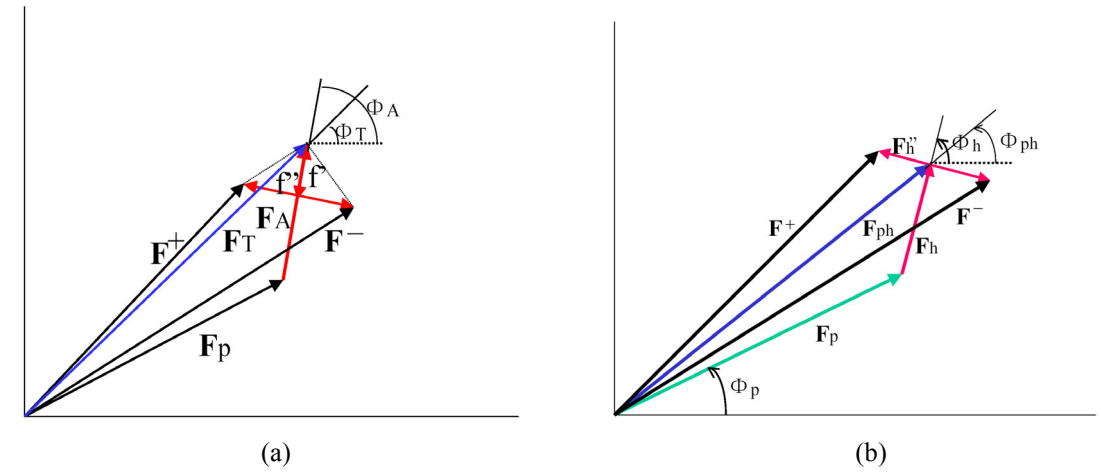


Figure 1-13 Phasing by the anomalous signals. (a). Harker construction for MAD phasing. (b). Harker construction for SAD phasing.

If only one wavelength is available, the phase problem may be solved by the SAD method (Fig. 1-13b). In the SAD method, we have some approximations and use $\mathbf{F}_{ph}$ instead of $\mathbf{F}_T$:

$$|\mathbf{F}^{\pm}|^2 = |\mathbf{F}_{ph}|^2 + |\mathbf{F}_h''|^2 \pm 2|\mathbf{F}_{ph}||\mathbf{F}_h''|\sin\alpha \qquad (1\text{-}5)$$

where, $\alpha = \Phi_{ph}-\Phi_h$. There are two possible values for **α**, which gives a two ambiguity problem as SIR method. Similarly, this phase ambiguity can also be solved combining the SAD and density modification method.

The SAD and MAD methods become very popular recently, since many heavy atoms have absorption edges within the normally used X-ray wavelengths for crystallography. The advantage over the SIR or MIR method is that only one crystal is needed.

1.2.4.3. SIRAS

We can also combine the isomorphous replacement and anomalous method together to better solve the phase problem, such as SIRAS (Fig. 1-13b).

In addition to the relationship derived from the anomalous signals (Eqn. 1-5), we have another relationship derived from the isomorphous replacement:

$$|\mathbf{F}_p|^2 = |\mathbf{F}_{ph}|^2 + |\mathbf{F}_h|^2 - 2|\mathbf{F}_{ph}||\mathbf{F}_h|\cos\alpha \qquad (1\text{-}6)$$

where $\alpha = \Phi_{ph} - \Phi_h$. From the two formula, the value of α will be uniquely determined.

In all the above cases, we assume ideal conditions. While in practice, the phase triangle is not closed resulting from experiment errors. The phases should be described by probability density function.

1.2.4.4. Direct methods and dual-space recycling

Direct method is to derive the structure factor phases directly from the observed amplitudes through probabilistic relationships. In general, the phase and the amplitude of a wave are independent quantities. However, in the X-ray diffraction by the crystal, it is possible to relate these two quantities, taking into account the important properties of the electron density function : (1) it is everywhere positive $\rho(r) \geq 0$ (positivety) (2) it is composed of discrete atoms (atomicity). These two conditions impose the constraints on the phases. Only certain values of the phases are consistent with both conditions.

Since the atomic scattering factors drop with increasing diffraction angles, reflections measured at different angles are not directly comparable. The normalized structure factors $\boldsymbol{E_h}$ (Hauptman & Karle, 1953) are used in the direct methods, which remove the effect of diffraction angles on the reflection intensities:

$$\boldsymbol{E_h} = \frac{\boldsymbol{F_h}}{(\varepsilon \sum_j |\boldsymbol{f}_{j\boldsymbol{h}}|^2)^{1/2}} \qquad (1\text{-}7)$$

where $\boldsymbol{F_h}$ are the crystal structure factors; ε is a statistical factor; $\boldsymbol{f}_{j\boldsymbol{h}}$ is the atomic scattering factor of atom j at index $\boldsymbol{h}$.

There are several important phase relations in direct methods, such as the triplet phase relation, the positive quartet relation and the negative quartet relation. From these phase

relations, conventional direct methods can efficiently solve the structures containing fewer than approximately 100 independent non-H atoms. So they are widely used to solve the small molecular structures. However, it is not able to solve the macromolecular structures that usually have more than 1000 non-H atoms. Nowadays, the macromolecular structures can be successfully solved with an improved direct-method procedure, known as *Shake-and-Bake* (Miller *et al.*, 1993) or dual-space recycling. The distinctive feature of this procedure is the repeated and unconditional alternation of reciprocal-space phase refinement (*Shaking*) with a complementary real-space process that seeks to improve phases by applying constraints (*Baking*). However, the application of dual-space recycling method in *ab initio* protein structure determination needs atomic resolution data (1.2 Å or better), which are less available in protein crystallography. In the same time, it is also very effective to solve the substructure of heavy atoms or anomalous centers, and does not require high resolution data (3 Å is enough).

1.2.4.5. MR method

When homologous structures or identical structures in a different crystal form are known, Molecular replacement (MR) method (Rossmann & Blow, 1962) can be used to solve the phase problem quickly. The idea is simple: the similar structures will give similar diffractions if they have the similar orientation and location in the same cell. The problem is to find the right orientation defined by three rotation angles, and the right location defined by three translation parameters for the search model. This 6-dimensional calculation is applicable in modern computers. However, it can usually be separated into two 3-dimensional calculations: rotation and translation. A rotation function can be computed to find the three rotation angles, and then the oriented model can be placed in the cell with the translation function. As a rule of thumb, the MR method will be probably rather straightforward if the search model is fairly complete and shares at least 40% sequence identity with the unknown structure. However, the phases from MR suffer from lingering model bias.

1.2.5. Substructure determination

There is a hidden phase problem for isomorphous replacement and anomalous methods, which is the location of heavy atoms or anomalous centers to calculate the reference structure factor. The key to solve this problem is to get a good estimate for F_h or F_A.

(1). In the SIR and MIR case (Fig. 1-12a,b), we have $F_h \approx | F_{ph}-F_p |$.

(2). In the SAD case (Fig. 1-13b), we have $F_h'' \approx 1/2|F^+-F^-|$, and $F_h=kF_h''$, where k is a constant.

(3). In the SIRAS case, it is possible to obtain a much better estimate of the heavy atom scattering by combining both isomophous replacement difference and anomalous difference:

$$F_h^2 = F_p^2 + F_{ph}^2 - 2F_pF_{ph}\{1-wk(F^+-F^-)/2F_p^2\}^{1/2} \qquad (1\text{-}8)$$

(4). In the MAD case, F_A can be exactly calculated from Eqn. 1-4.

Once the estimates of F_h or F_A have been obtained, the location of heavy atoms or anomalous centers could be determined by Patterson and/or direct methods as we do for small molecule structure determination.

All differences mentioned above are small, and highly accurate data are required for successful substructure determination and subsequent phasing. Both substructure enantiomorphs will fit the F_h or F_A, so their hand must be checked with the derived protein phases. In the isomorphous replacement methods, the wrong enantiomorph of the substructure produces a wrong-handed protein map; in the anomalous dispersion methods, the wrong enantiomorph of substructure produces a meaningless map.

1.2.6. Phase improvement by density modification and phase combination

Normally, the initial phases obtained from the above methods are noisy, the calculated electron density maps are difficult to interpret or even uninterpretable. However, we have some prior knowledge of the crystal structure, which can impose constraints on the phases and thus improve them to generate an easily interpretable electron density map. This process is generally called density modification. For instance, we know that a large part of the protein crystal is composed of bulk solvent, thus the electron density and its variance are different in the protein and solvent region. In fact, this is the idea of solvent flattening. The phases from different sources can be combined together to improve phases as well, which is called phase

combination. The commonly used methods for phase improvement are summarized in the table 1-2 (Zhang *et al.*, 2001).

Constraints	Use	Effectiveness and limitation
(1) Solvent flatness	Solvent flattening	Works best at medium resolution. Relatively resolution insensitive. Good for phase refinement. Weak on phase extension
(2) Ideal electron-density distribution	Histogram matching	Works at a wide range of resolutions. More effective at higher resolution. Very effective for phase extension.
(3) Equal molecules	Molecular averaging	Works better at low to medium resolution. Its phasing power increases with the number of molecules in the asymmetric unit.
(4) Protein backbone connectivity	Skeletonization	Requires near atomic resolution to work.
(5) Local shape of electron density	Sayre's equation	The equation is exact at atomic resolution. It can be used at non-atomic resolution by choosing an appropriate shape function. Its phasing power increase quickly with resolution. Very powerful for phase extension.
(6) Atomicity	Atomization	If the initial map is good enough, iteration could lead to a final model
(7) Structure-factor amplitudes	Sim weighting	Can be used to estimate the reliability of the calculated phases after density modification. It assumes the random distribution of errors that cause the discrepancy between the calculated and observed structure-factor amplitudes.
(8) Experimental phases	Phase combination	This can be used to filter out the incorrect component of the estimated phases. Most phase combination procedures assume independence between the calculated and observed phases.

Table 1-2 The commonly used phase improvement methods.

1.2.7. Model building and refinement

From the X-ray diffractions, we actually get the electron density distributions in the unit cell. An interpretable electron density map can usually be produced after density modification and phase combination. Now we can put atoms in their corresponding electron densities to obtain a molecular model. This process is called model building. In the past, model was built manually. Recently there are programs available for autobuilding, but in many difficult cases, extensive human intervention is still needed.

The initially built model is then optimized to improve the agreement between the observed and calculated data. This process is called refinement. There are two popular criteria to measure the agreement: least-squares (LS) and maximum likelihood (ML).

In the LS methods, the measure of agreement is the L_2 norm of the residuals, which is

simply the sum of the squares of the differences between the observed and calculated data: $L_2(x)=\sum_i w_i[y_i - f_i(x)]^2$ where w_i is the weight of observation y_i (intensity or amplitude) and $f_i(x)$ is the calculated value of observation *i* given the parameter x. The refinement target is to minimize the residuals.

The ML methods evaluate likelihood of the observations given the model. Then, the likelihood of a model given a set of observations is the product of the probabilities of all of the observations given the model. It can be formulated as: $L=\prod_i P_a(\mathbf{F}_i;\mathbf{F}_{i,c})$ where $P_a(\mathbf{F}_i;\mathbf{F}_{i,c})$ is the conditional probability distribution of the structure factor $\mathbf{F}_i$ given the model structure factor $\mathbf{F}_{i,c}$. It can be written in a logarithm form: $\log L=\sum_i \log P_a(\mathbf{F}_i;\mathbf{F}_{i,c})$.

The refinement target is to maximize the likelihood of the model. The ML refinement takes into account the errors in both the model and the observations, which makes it particularly useful when the model is incomplete, to say, in the early stages of the refinement.

Once the criterion is decided, the parameters of the model including atomic coordinates, thermal factors and sometimes the occupancies are optimized to approach the target. Since protein crystals contain 30-70% amorphous solvent, a bulk solvent model is also refined especially when low resolution data are used. More often, protein crystals do not diffract to very high resolution, leading to a very poor (about 1~3) data/parameter ratio. Therefore, the model is not well determined by the X-ray data considering experimental errors. The solution is to incorporate prior knowledge such as molecular geometry into the refinement in the forms of restraints and/or constraints. Restraints can be treated as additional observations, while constraints decrease the number of parameters and both increase the data/parameter ratio. To avoid the model overfitting the data, a part (usually 5-10%) of reflections are set aside from the refinement as a cross validation to monitor the refinement process (Brunger, 1993).

After refinement, the model quality should be evaluated by electron density, crystallographic residual, molecular geometry and even biological sense etc. In principle, the high quality of the refined model is possible only if high quality data are available. Good data give good model. High quality of the data means: very high degree of completeness, high redundancy, high I/σ(I) ratio, high resolution, etc. High quality data are also extremely important for successful substructure determination and the subsequent phasing process.

1.3. Aim of this work

Wind plays a very import role in the DV polarization of the *Drosophila* embryo development. It is required for the correct location of the patterning protein Pipe to the Golgi. This has been proved at the genetic level, while little is known at the molecular level. A 3-dimensional structure is crucial to understand the function and mechanism. Here I report the crystal structure of Wind at 1.9 Å resolution with detailed analyses of the structure. Many features, such as oligomerization state, surface electrostatic potential/hydrophobicity, and a possible substrate binding site are revealed, which greatly contributes to understanding the structure-function-relationship of Wind and also facilitates biochemical studies. Wind is also the first complete crystal structure of a PDI-related protein in the ER. It provides a model to study other PDI-related protein structures as well.

2. Materials and methods

2.1. Cloning, expression and purification

Drosophila windbeutel cDNA encoding the mature Wind was amplified by PCR from a lambda-ZAP cDNA library and ligated into the BamHI/SacI sites of pQE-30, generating an N-terminal extension including a 6xHis-tag (MRGSHHHHHHGS).

Wind protein was expressed in *E. coli* XL1-Blue cells by induction of an OD_{600} = 0.7 culture for 3 hours at 37°C with 1mM IPTG. The recombinant protein was harvested by brief sonication of lysozyme-treated cells in pH 8.0-adjusted phosphate buffered saline including 0.2mM Pefabloc protease inhibitor, followed by addition of triton-X 100 to 0.1% (v/v) and gel filtration over a Talon nickel affinity column. Bound protein was washed with 4 bed volumes each of wash buffer (20mM Tris-Cl, 150 mM NaCl, 0.1% triton-X 100) and salt wash buffer (20mM Tris-Cl, 350 mM NaCl, 0.1% Triton-X 100), then washed again with 4 bed volumes wash buffer and eluted in 4 bed volumes elution buffer (20 mM Tris-Cl, pH 8.0, 300 mM NaCl, 100 mM imidazol, 0.05% Triton-X 100). The eluted protein was dialyzed extensively against dialysis buffer (10mM Hepes, pH7.5, 50mM NaCl, 0.01% (v/v) β-mercaptoethanol), concentrated to 20-25 mg/ml, and stored at 4°C.

The above protein cloning, expression, and purification were done by Guo *et al.*. at the Max-Planck-Institute for Biophysical Chemistry, Göttingen.

2.2. Sample quality

Protein purity was verified by a 12.5% silver stained SDS-PAGE gel which is sensitive to a *ng* level (Fig. 2-1). The gel showed one main band of 26 kD which corresponds to the protein Wind, and a very weak band of lower molecular weight (MW) which is a contaminant. The purity was estimated to be above 99%, which is very good for subsequent crystallization experiments. The protein samples were distributed among 50 or 100 µl aliquots and stored at –85°C in the freezer.

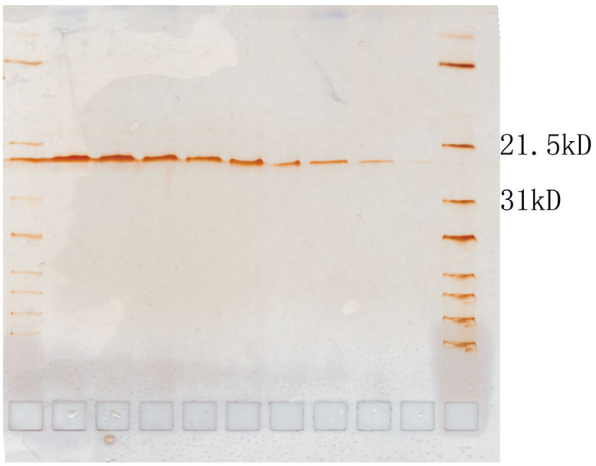


Figure 2-1 SDS-PAGE gel of Wind. Lanes in the middle show Wind at a concentration gradient. lanes at either side show molecular weight markers.

2.3. Crystallization

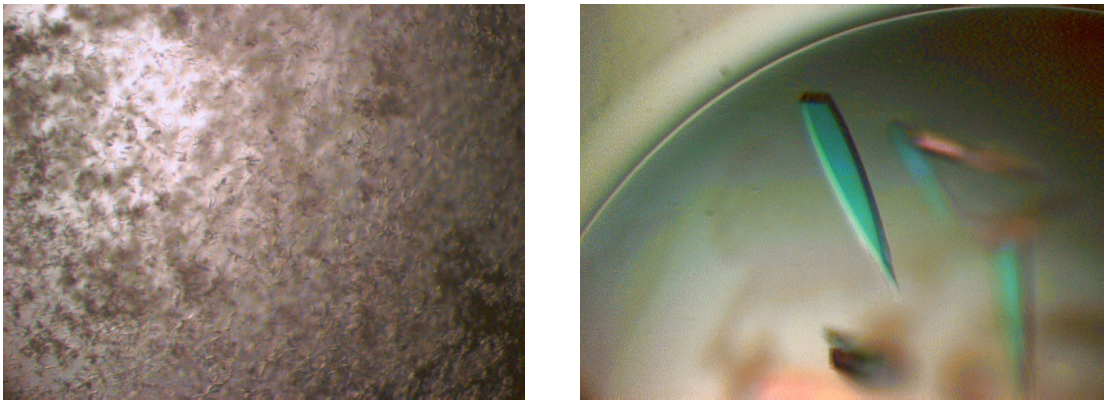
2.3.1. Screening

The initial screening was performed with vapor diffusion method using Hampton Crystal Screens 1 & 2, Hampton PEG/ION screen and Emerald Wizard I & II at 4°C and 20°C. All of the above products use the sparse matrix method, which is based on the successful conditions reported in literatures. The initial protein solution (23mg/ml) was too concentrated (precipitation occurred in most conditions of a test screen) and was diluted 4 times: To one volume sample, 1 volume of [0.1M Hepes, 50 mM NaCl] and 2 volumes ddH_2O were added. The crystallization drops were set up by mixing 4μl protein solution with 2μl well solution. Crystals formed under the following conditions.

(A). 0.1M CsCl, 0.1M MES pH6.5, 30% Jeffamine M-600 (#24 of Hampton Screen 2), 4°C. Hundreds of small plate crystals appeared after 1 day (Fig. 2-2a).

(B) 0.2M $Mg(NO_3)_2$, 20% PEG 3350 (#16 of Hampton PEG/ION), 4°C. The crystals were single and of adequate size, but took a few months to form.

(C) 15% Ethanol, 0.1M Tris pH7.0 (#42 of Emerald Wizard I), 4°C. The crystals appeared after one week and the shape was similar to those in (A).



(a) (b)

Figure 2-2 Crystal pictures of Wind. (a) Crystals from initial screen (#24 of Hampton Screen 2). (b) Crystals after optimization.

2.3.2. Optimization

All the three initial conditions were optimized and condition (A) gave the best results (Fig. 2-2b). Several parameters were tuned to grow good crystals for diffraction.

(1) Protein concentration: The protein concentration in the screening worked fine. Little was changed on this parameter.

(2) Temperature: Crystals grew under several temperatures, such as 4, 15, 20, 25°C. Generally, the lower the temperature, the more crystal nuclei and thus, the smaller final size. The nucleation process is very difficult to control under low temperature even adjusting other parameters. The crystallization seems sensitive to temperature variations and no crystals appeared in a quiet and temperature-uncontrolled room.

(3) pH: Crystals grew well from pH 5.8 to 6.5 with 0.1M MES as buffer material.

(4) Buffer: Besides MES, crystals also appeared in other buffers such as imidazol, ADA, cacodylate etc. But the crystals looked much worse in other buffer systems.

(5) Precipitant: Low molecular weighted PEGs, such as PEG600, PEG400, PEG300 of 14%-25% concentration worked well. However, they had different effects on crystallization rate, and PEG300 gave the largest crystals.

(6) Salts: Both CsCl and NaCl were successfully used as additives. But big crystals preferred CsCl. The crystallization is sensitive to the salt concentration which should be lower than 0.2M. No crystals grew at higher concentrations.

(7) Other factors: Crystals preferred to grow from a material surface, e.g. sticking to the

slide surface. Smearing a thin layer of silicon grease on the slide surface helped to grow better crystals and diminish mechanical damages to the crystals when mounting.

After optimizing all possible factors, crystals used for the native data collection were grown by the hanging drop vapor diffusion method at 20 °C by mixing 6μl 5.8mg/ml Wind in 5mM Hepes pH7.5, 25mM NaCl, 0.0025%(v/v) β-mercaptoethanol with 3μl reservoir solution containing 0.1M MES pH6.1, 0.1M CsCl, 2mM $CaCl_2$, 16% PEG 300. However, some other slightly different conditions were used in other diffraction experiments as well. The MES buffer was made by mixing MES-H solution and MES-Na solution to obtain the required pH. The crystallization behavior around this condition was: white heavy precipitates formed immediately after setting up the drop, then within half an hour, the precipitates transformed to oil covering the drop surface. After one day, the ship-like crystals grew from the oily drop surface.

2.3.3. Cryo solutions

Several cryo-protectants were optimized for best results (table 2-1) (Rodgers, 2001). Finally, either [0.1M MES pH6.1, 0.1M CsCl, 20% PEG300, 10% Glycerol] or [0.1M MES pH6.1, 0.1M NaCl, 20% PEG300, 10% Glycerol] was used as cryo-solution.

Cryo-protectant	suggested concentration
DMSO	2-20%
Erythritol	>50%(w/v)
Ethylene glycol	15-45%
Glycerol	15-45%
Inositol	20-50%
MPD	25-40%
PEG 200-600	30-50%
Raffinose	>50%(w/v)
Sucrose	> 50%(w/v)
$(NH_4)_2SO4$	50% saturation
2-propanol	>70%
Xylitol	20-50%

Table 2-1 Commonly used cryo-protectants.

2.3.4. Heavy atom derivatives

Since Wind contains 3 cysteines and Hg^{2+} ions have a good affinity to sulfur atoms, Hg compounds are good candidates for heavy atom derivatization. Anyway, several commonly

used heavy atom compounds such as K_2PtCl_4, $HgCl_2$, PbAc, ZnI_2 and some lanthanide compounds etc. had been tried (table 2-2). Finally, $HgCl_2$ proved to be successful. The Hg derivative was prepared by growing crystals with the well solution of 0.1M MES pH6.1, 0.08M NaCl, 2mM CaCl, 16% PEG300, and soaking them in 0.1M MES pH6.1, 0.1M NaCl, 20% PEG300, 10% Glycerol, 0.05mM $HgCl_2$ for 2 weeks. The long soaking time is not necessary and could be shortened to 3 days. However, the very low concentration (0.05mM) of $HgCl_2$ is critical for success. When higher concentration was used, the derived crystals were somewhat crashed and diffracted badly.

Used times	Compound
287	Potassium tetrachlorophlatinate(II)
111	Potassium dicyanoaurate(I)
103	Uranyl acetate
101	Mercury(II) acetate
98	Mercury(II) chloride
85	Ethylmercurythiosalicylate (EMTS)
82	Potassium tetraiodomercurate(II)
81	para-Chloromercuriobenzenesulfonate(PCMBS)
75	Trimethyllead(IV) acetate
73	Potassium pentafluorooxyuranate(VI)
73	Phosphatotris(ethylmercury)
61	Potassium tetranitritoplatinum(II)
60	Uranyl nitrate
58	Potassium tetracyanoplatinate(II)
57	Dichlorodiammineplatinum(II)
51	Potassium hexachloroplatinate(IV)
51	Methylmercury chloride
44	Potassium tetrachloroaurate(III)
42	para-Chloromercurybenzoate(PCMB)
39	Lead(II) acetate

Table 2-2 The 23 most commonly used heavy-atom compounds reagents. The first column gives the number of times the reagent has appeared in the heavy-atom data bank. (Carvin *et al.*,2001)

2.4. X-ray data collection and processing

All the diffraction experiments were carried out at low temperature (100K). The crystals were soaked in the cryo solution for a short time (seconds or minutes), then plunged directly into liquid nitrogen (rather than flash cooling in the nitrogen gas).

The native data were collected on a Mar CCD detector at the synchrotron beamline X11 at the EMBL Hamburg outstation/DESY. The dose mode was used in collection. High resolution data were collected first with a high dose and fine oscillation angle, while subsequently the low resolution data were collected with low dose and wide oscillation angle. The radiation damage became apparent in the latter stage of the high resolution data collection,

judged by the resolution decreasing. Hg derivative data were collected on Mar345 image plate *in house*. In order to overcome the non-isomorphism problem between the synchrotron and the *in house* data, one native dataset was also collected *in house*.

The X-ray data were indexed & integrated with DENZO, and scaled with SCALEPACK (Otwinowski and Minor, 1997), respectively. Several later frames in the high resolution dataset suffering from radiation damage were discarded. The reflections of symmetry equivalents and Friedel pairs were not merged. XPREP (Sheldrick, 2001) was used to merge reflections and to prepare data files for the subsequent procedures. 5% of the reflections were flagged in thin shells for cross validation in the refinement. Truncate (CCP4, 1994) in the CCP4 package was used to calculate amplitudes from intensities. The data statistics are summarized in table 2-3.

There was a cell-scaling problem in the synchrotron data. It is not unusual due to the inaccurate wavelength calibration for the synchrotron beamline. The cell parameters were calibrated by the WHAT IF program (Vriend, 1990) during model evaluation process, by comparing the bond lengths of the refined model against standard values. The statistics (table 2-3) for the synchrotron dataset were re-analyzed afterwards with the corrected cell parameters. If not mentioned, the statistics in this thesis are corresponding to the corrected cell parameters. However, this problem was only detected after doing refinement. So the original cell parameters from SCALEPACK (108.415 50.823 99.526 90.00 112.58 90.00) were used in substructure determination, phasing, density modification, and the early stages of refinement. In the latter stages of refinement, the cell parameters were calibrated and updated in the next cycle. The corrected cell parameters are not obviously different from the old ones, and no serious problems were found for this reason. Nevertheless, the cell parameters affect the resolution statistics. The model was finally refined to 1.9Å resolution corresponding to the corrected cell.

Table 2-3. Statistics of the data collection and structure solution and refinement.

Data collection			
	native 1	native 2	Hg-derivative
Wavelength(Å)	0.811	1.5418	1.5418
X-ray Source	X11 beamline, EMBL/DESY	Cu rotating anode	Cu rotating anode
Detector	Mar CCD	Mar345	Mar345
Spacegroup	C2	C2	C2
Cell parameters	a=106.678	a=108.107	a=107.998
(Å or °)	b=50.358	b=50.802	b=50.935
	c=98.616	c=99.030	c=99.122
	β=112.84	β=112.45	β=112.77
Mosaicity(°)	0.703	0.76	1.125
Resolution limits	37.64-1.90	19.91-2.69	53.23-2.99
(Å)	(2.00-1.90)	(2.80-2.69)	(2.80-2.69)
Reflections(unique)	38218	13890	10142
Completeness(%)	99.8(99.6)	99.0(93.7)	99.1(93.8)
Mean I/σ(I)	13.96(4.77)	21.19(3.64)	19.29(6.45)
Redundancy	4.50(3.79)	7.09(4.52)	6.75(5.13)
Rint(%)[1]	5.65(27.59)	7.12(44.08)	6.13(23.99)
Rsigma(%)[2]	4.47(19.59)	3.88(29.00)	4.28(15.50)
Refinement statistics			
Reflections	38216		
Rwork/Rfree[3](%)	21.58/25.57		
Protein atoms	3295		
Solvent atoms	152wat,1 Cesium		
Mean B value (Å^2)	37.25		
R.m.s.d bond lengths(Å)	0.023		
R.m.s.d bond angles (°)	2.13		

Values in the parenthesis correspond to the highest resolution shell.

[1]Rint=$\Sigma|F_o^2 - F_o^2(mean)|/ \Sigma[F_o^2]$

[2]Rsigma=$\Sigma[\sigma(F_o^2)]/ \Sigma[F_o^2]$

[3]5% reflections are selected in thin shells as test dataset.

Rwork=$\Sigma||F_o|-|F_c||/\Sigma|F_o|$ for working dataset and Rfree=$\Sigma||F_o|-|F_c||/\Sigma|F_o|$ for test dataset.

The solvent content was estimated by Matthews method (Matthews, 1968) (Table 2-4).

Molecular numbers /asymmtric unit	Matthew coeffecient Vm (Å^3/Da)	Solvent content (%)
1	4.3	71.1
2	2.1	42.1
3	1.4	13.2

Table 2-4 Solvent content estimates with Matthews method.

Generally, protein crystals contain 30-70% solvent. From the calculation we know that most likely there are 2 molecules in one asymmetric unit with a solvent content ~42%.

2.5. Substructure determination

The Hg substructure was solved and refined by a combination of difference Patterson analysis and dual-space recycling method using the SIRAS information from the native and derivative datasets collected in house.

SIR or SAD data tend to be very noise because they are based on small differences between the observed structure factors. If data are used to a higher resolution than there are significant dispersive or anomalous differences, the effect will be to add noise. Since dual-space recycling methods are based on normalized structure factors, which emphasize the high resolution data, they are particularly sensitive to this. So it is critical to truncate the resolution to a good value in order to use dual-space recycling methods for substructure determination.

For SAD, if the correlation coefficient of anomalous differences from two quality-comparable crystals or from one crystal in two orientations is available, the data should be truncated to the resolution where it drops to 25-30%. Or, the data could be truncated where either the ratio of anomalous difference to its standard deviation drops to about 1.3 or the anomalous difference itself drops under about 1.3 electrons.

For SIR, the data could be truncated to where the Rmerge of native and derivative data is about 15~25%, which indicates significant disperse differences and good isomorphism.

The dispersive differences between the native and Hg derivative data, and the anomalous differences between the Bijvoet pairs of the Hg derivative data were analyzed with XPREP (table 2-5).

The anomalous signals disappeared in the noise at 4.2Å resolution where the ΔF is 1.3e. The dispersion signals between the native and derivative data are good up to 3.1 Å resolution with the Rmerge (or Rint) of 22.2% after local scaling.

Table 2-5 Data analysis of dispersive and anomalous signals.

```
Anomalous signal/noise ratios (1.0 is random).  The first line is based on
input sigmas, the second on variances of F+ and F- (if not already averaged):
  Inf - 8.0 - 6.0 - 5.0 - 4.6 - 4.4 - 4.2 - 4.0 - 3.8 - 3.6 - 3.4 - 3.2 - 3.0 Å
    1.18  1.37  1.22  1.16  1.19  1.26  1.09  1.10  1.06  1.19  1.20  1.14
    2.22  1.96  1.57  1.40  1.35  1.37  1.17  1.10  1.04  1.15  1.15  1.08
 35.1  Neighbors used on average for F+/F- local scaling
 Rint(anom) =   0.0574  before and   0.0585  after local scaling
 34.5  Neighbors used on average for local scaling to native/remote dataset
 Rint =   0.2656  before and   0.2220  after local scaling
```

The F_h and α (phase difference between F_{ph} and F_h) were prepared by XPREP. In fact, α is not used in substructure determination but in phasing.

A Patterson map is calculated with the F_h^2 to 4.2 Å resolution (Fig. 2-3). There are eight clear peaks on the Harker section y=0.5, so 4 Hg sites were generally expected for C2 spacegroup. But it was proved there are only three Hg sites and this will be explained later.

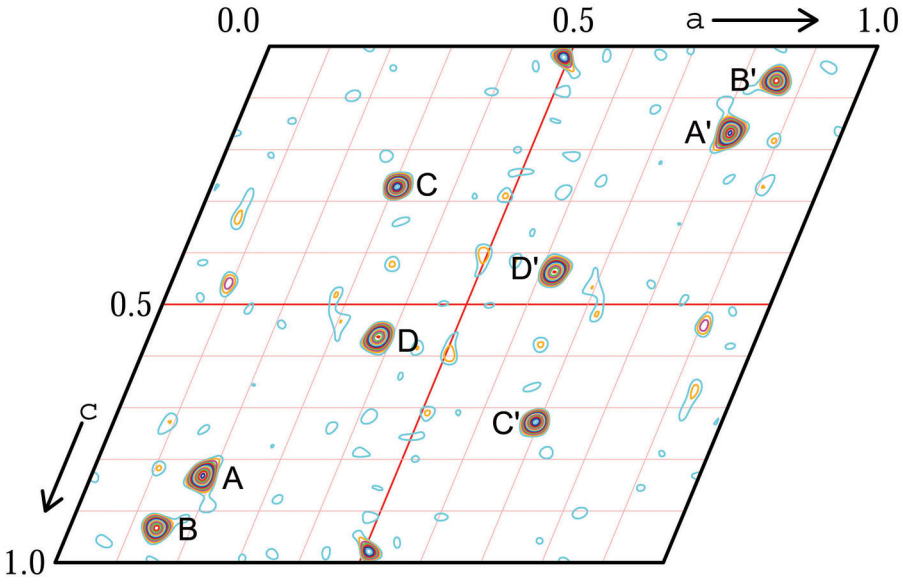


Figure 2-3 Patterson map (section y = 0.5) of SIRAS data truncated to 4.2 Å.

The substructure was determined and refined with SHELXD (Sheldrick *et al.*, 2001; Schneider & Sheldrick, 2002). SHELXD combines Patterson analysis and direct method to solve the substructure. The starting phases are generated from the Patterson-seeds instead of random atoms. Then followed the dual-space (*shake-and-bake*) recycling process: phases are expanded from the ~40% most reliable ones using the tangent formula in reciprocal space; atom peaks are picked in real space to impose constraints on the phases.

The occupancies of the atoms are refined after the peak-search in the final dual-space cycles as well. Potential solutions are identified by high values of the correlation coefficient CC between E_o and E_c (Fujinaga & Read, 1987).

$$CC = \frac{100[\sum(wE_oE_c)\sum w - \sum(wE_o)\sum(wE_c)]}{\{[\sum(wE_o^2)\sum w - (\sum wE_o)^2]\cdot[\sum(wE_c^2)\sum w - (\sum wE_c)^2]\}^{1/2}} \qquad (2\text{-}1)$$

Principally, attention should be paid to the three following aspects.

(1). The resolution should be properly truncated.

(2). The number of sites requested should be within about 20% of the true value so that the occupancy refinement works well.

(3). In the case of data from a soaking experiment, the rejections of sites on special positions should be switched off.

In practice, several parameters were tested. After trials, the SIRAS data were truncated to 3.2 Å (SHEL 999 3.2) and 3 sites (FIND 3) were requested. The solution was straightforward. The best job gave 5 possible sites with the CC(all)/CC(weak) of 27%/20%. A sharp drop in the occupancy takes place between the third and the fourth site. Checking the cross table, there are only three convincing Hg sites.

The substructure seemed to be solved. However, both substructure enantiomorphs will give equally good agreement with the observed F_h values, so either this substructure or its enantiomorph is correct, which will be judged in the later phasing process.

```
REM TRY   285   CC 26.94   CC(weak) 19.88   TIME    696 SECS
TITL mar-hg30-siras in C2
CELL 1.54178 108.4150  50.8230  99.5260 90.000 112.581 90.000
ZERR   4.00    0.0217   0.0102   0.0199  0.000   0.030  0.000
LATT -7
SYMM -X, Y, -Z
SFAC Hg
UNIT   12
HG01   1  0.159103  0.171814  0.081655   1.0000  0.2
HG02   1  0.320938 -0.234932  0.465726   0.7367  0.2
HG03   1  0.461685 -0.323334  0.354558   0.4531  0.2
HG04   1  0.315071 -0.430801  0.448568   0.2701  0.2
HG05   1  0.203148  0.171501 -0.010104   0.1316  0.2
HKLF 3

END
```

(a)

```
Minimum distances (top row, 0 if special position) and PSMF (bottom row)
 Peak     x       y       z      self   cross-vectors
 99.9 0.1591  0.1718  0.0817   32.0
                               61.0
 73.6 0.3209 -0.2349  0.4657   31.8   41.0
                               57.0   55.3
 45.3 0.4617 -0.3233  0.3546   26.9   40.2   22.5
                                7.1   53.8   17.8
-----------------------------------------------
 27.0 0.3151 -0.4308  0.4486   32.6   39.4   10.1   22.0
                                3.1   27.8    0.0    7.5
 13.1 0.2031  0.1715 -0.0101   27.2   11.8   44.7   32.0  43.2
                                0.0   11.1    0.0    0.0   0.0
```

(b)

Table 2-6 The substructure solution. (a) The best substructure given by SHELXD. There is a sharp drop between the third and the fourth site. (b) The corresponding cross table of this substructure. The possible wrong peaks are characterized with 0.0 for PSMF value.

2.6. Phasing and density modification

2.6.1. Phasing and initial density modification with SHELXE

The structure was solved by SIRAS method. The phase calculation/extension and initial density modification were performed with SHELXE (Sheldrick, 2002).

SHELXE is a quick and robust phasing program. Three steps are used for phasing.

(1) Calculate protein phases by adding α to the substructure phases with weights that depend on the normalized structure factors E_h. In the SAD and SIR cases, the centroid phases are calculated (α = 90 or 270˚ for SAD and α = 0 or 180˚ for SIR, respectively). In the MAD

and SIRAS cases, α is not restricted and can be any value between 0 and 360°.

(2) The two-fold ambiguity of phases in SAD/SIR is solved by the low density elimination method.

(3) If the heavy atoms are present in the native, their σ_A-weighted direct estimates of protein phases are included.

All the steps produce independent phase estimates and so can be combined using σ_A-weights (Read, 1986).

An optional density modification procedure is immediately following phasing in SHELXE. SHELXE uses the *sphere of influence* method to improve phases, which incorporates some general chemical information. The variance of the density is calculated for each pixel in the map on a spherical surface of radius 2.42Å, which corresponds to the dominant 1,3-interatomic distance in all macromolecular structures. The pixels with the highest variances are assigned to the protein region and the others to the solvent region. The density ρ in the solvent region is flipped ($\rho'=-\gamma\rho$ where γ is about one). In the protein region ρ is replaced by $[\rho^4/g^2\sigma^2(\rho) +\rho^2]^{1/2}$(with g usually 1.0) if positive and by zero if negative. A fuzzy solvent boundary is used, and ρ in the fuzzy region is set to a weighted sum of the two treatments.

In each phase refinement cycle, two figures of merit (FOM), *contrast* and *connectivity* are calculated. The *contrast* is defined as the variance of Variance averaged over all pixels and the *connectivity* is defined as the fraction of adjacent pixels that either both in the solvent or both in the protein regions. The absolute values of the *contrast* and *connectivity* vary with the solvent content and resolution of the data, they both tend to be higher for correct maps. The density-modified phases are normally combined with the original phases to avoid the overestimation of FOM.

The first three Hg sites in the substructure were kept to calculate the reference phases. Initial protein phases were obtained by adding α to these phases. Then followed 200 cycles of density modification. In order to determine the correct enantiomorph, two SHELXE jobs were set up: "shelxe syn55 mar-hg30-siras -s0.55 -m200 -b", and "shelxe syn55 mar-hg30-siras -s0.55 -m200 -b -i " for the enantiomorph. The correct enantiomorph was distinguished from the wrong one quickly after a few cycles of density modification, by the much higher FOM

values of both *contrast* and *connectivity*. This can also be checked with the resulting maps. In SIRAS phasing, the correct heavy atom substructure will lead to the correct structure and the other will give a meaningless map. In fact, not the substructure from the previous SHELXD job, but its enantiomorph proved to be correct. For the solvent content values from 35-70% were tested. Here 55% solvent content (-s0.55) was used which gave a bit better FOM than others.

shelxe syn55 mar-hg30-siras -s0.55 -m200 -b Overall	**shelxe syn55 mar-hg30-siras -s0.55 -m200 -b -i**
CC between Eobs (from delF) and Ecalc (from heavy atoms) = 21.47%	Overall CC between Eobs (from delF) and Ecalc (from heavy atoms) = 21.47%
<wt> = 0.047 for initial phases	<wt> = 0.047 for initial phases
<wt> = 0.078, Contrast = 0.055, Connect. = 0.761 for dens.mod. cycle 1	<wt> = 0.082, Contrast = 0.078, Connect. = 0.812 for dens.mod. cycle 1
<wt> = 0.084, Contrast = 0.109, Connect. = 0.809 for dens.mod. cycle 2	<wt> = 0.090, Contrast = 0.203, Connect. = 0.865 for dens.mod. cycle 2
<wt> = 0.089, Contrast = 0.101, Connect. = 0.807 for dens.mod. cycle 3	<wt> = 0.094, Contrast = 0.241, Connect. = 0.874 for dens.mod. cycle 3
<wt> = 0.093, Contrast = 0.094, Connect. = 0.805 for dens.mod. cycle 4	<wt> = 0.097, Contrast = 0.283, Connect. = 0.882 for dens.mod. cycle 4
STOPPED	. . .
	<wt> = 0.300, Contrast = 0.581, Connect. = 0.894 for dens.mod. cycle 199
	<wt> = 0.300, Contrast = 0.582, Connect. = 0.894 for dens.mod. cycle 200
	Mean weight and estimated mapCC as a function of resolution
	d inf - 4.03 - 3.19 - 2.78 - 2.52 - 2.34 - 2.20 - 2.09 - 2.00 - 1.92 - 1.85
	<wt> 0.716 0.648 0.586 0.559 0.577 0.561 0.534 0.498 0.449 0.388
	<mapCC> 0.911 0.868 0.835 0.799 0.823 0.810 0.769 0.733 0.713 0.684
	N 4310 4314 4360 4378 4300 4355 4307 4247 4481 3979
	Final weighted contrast = 0.582 and connectivity = 0.900
	Site x y z h(sig) near old near new
	1 0.8411 0.8293 0.9176 45.3 1/0.10 1/39.42 3/40.46 3/40.49 2/40.88
	2 0.6807 1.2342 0.5344 35.5 2/0.18 3/22.60 3/23.90 1/40.88 3/41.59
	3 0.5373 1.3214 0.6439 19.9 3/0.18 2/22.60 2/23.90 1/40.46 1/40.49

Table 2-7 The SHELXE jobs for both substructure enantiomorphs. The job on the left indicated the wrong enantiomorph which was given by the previous SHELXD job.

The three heavy atom sites were also confirmed (by the option -b). These three Hg^{2+} are bound to the cysteines of the protein (Fig. 3-13).

The Patterson map was checked again. It is found that, A&A' are the Harker vectors of site **1**; B&B' are the Harker vectors of site **2**; C&C' are the cross vectors of **1**&**3**; D&D' are the cross vectors of **1** & **3**'s symmetry equivalent (-0.5373 1.3214 -0.6439). These two cross vectors happen to be on the Harker section, while the Harker vectors of site **3** are not obviously seen in the Patterson map due to its low occupancy.

2.6.2. Further density modification with DM

The phases output from SHELXE were further refined with DM (CCP4, 1994) using solvent flattening and histogram matching. 300 cycles were run with 55% solvent content. The reflections used for calculations were extended slowly from low (6 Å) to high resolution. The refined phases were combined with the starting ones. In the resultant electron density map, some regions were found to be related by a 2-fold non-crystallographic symmetry (NCS). A small fragment such as one helix or beta stand, or simply the b-domain of ERp29 NMR model (Liepinsh *et al.*, 2001), was fitted into these related regions, respectively, with XFIT (McRee, 1999). Least-squares fits of these fragments or domains gave the symmetry operator for the NCS. This operator was used for another DM job, in which the molecular averaging method was performed in addition. The SHELXE phases were still used as input. The improper NCS averaging was used and the monomer mask was generated automatically with the program. 200 cycles were run with 50% solvent content. Afterwards, 500 additional cycles were run with the refined NCS operator, with 48% solvent content. The resolution was slowly expanded and phases were combined as previously described. Afterwards, the density-modified phases from different strategies were evaluated with SHELXPRO, using the phases from the refined model as reference (Fig. 2-4a, b, c). These three phase sets are comparable and the one after NCS averaging is slightly better but is still worse than expected. Other strategies with different settings for the parameters such as solvent content, starting phases and the number of running cycles etc. were tested as well. The resulting maps from all the above refined phases were not significantly different from each other and all contributed to the subsequent model building process.

2.7. Model building

All the electron density maps were easily interpretable, although the electron densitiy for most side chains were not clearly seen. The molecular boundary was clear and the secondary elements such as alpha helices and beta strands were recognized. The b-domain and D-domain were distinguishable. However, the electron density of the D-domain was only clear for one monomer, while that for the other monomer was too weak and incomplete at all.

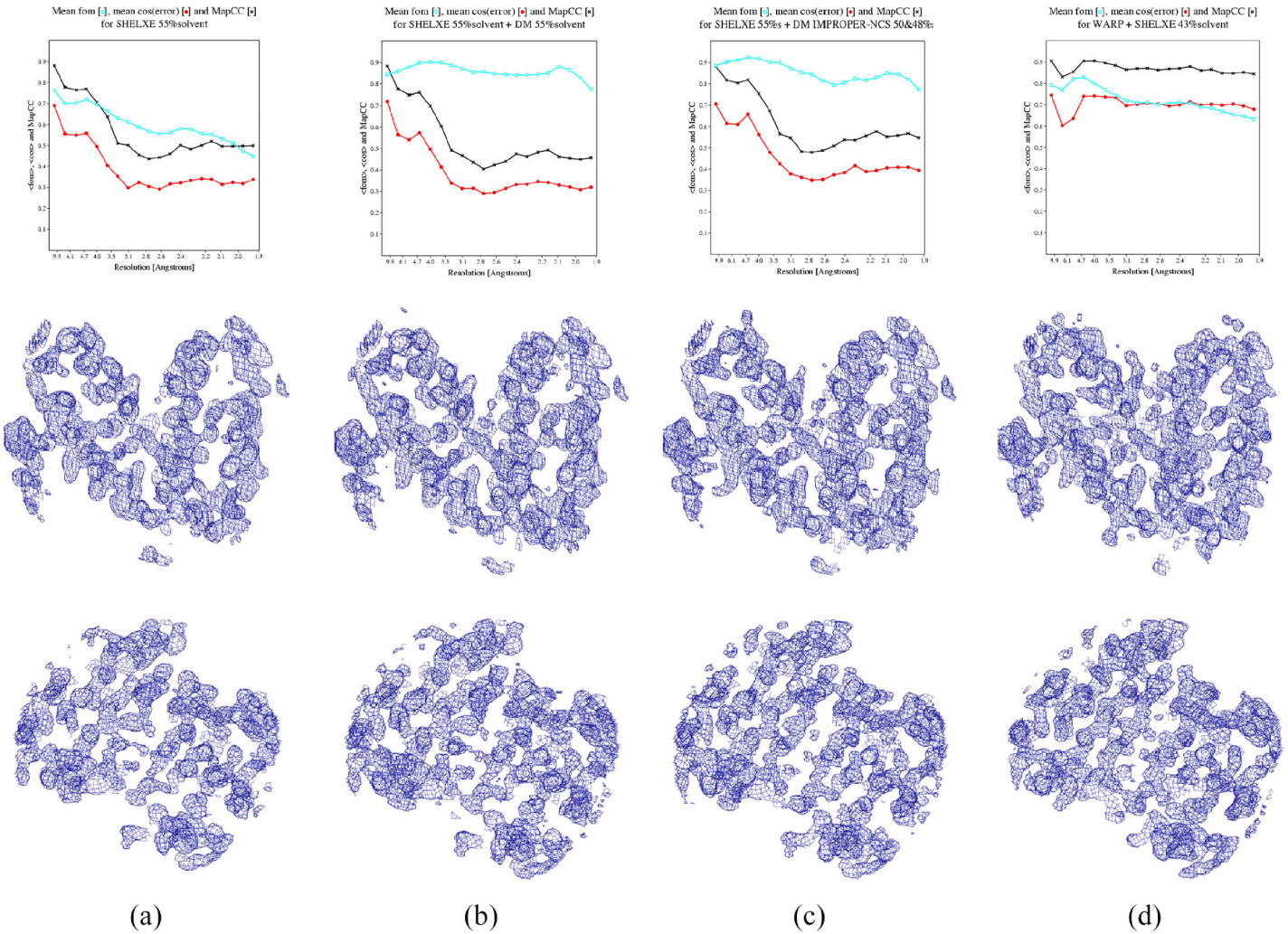


Figure 2-4 Phases from different strategies.

In columns
(a) Phases from a SHELXE job.
(b) Phases from DM, without NCS averaging.
(c) Phases from DM, including NCS averaging.

In rows
(1) Phase evaluation by SHELXPRO, using the refined model to calculate the reference phases.
(2) σ_A-weighted mF_o map contoured at 1σ level in the region of the D-domain in monomer A.
(3) σ_A-weighted mF_o map contoured at 1σ level in the region of b-domains.

The autotracing with ARP/wARP (Perrakis *et al.*, 1999) failed given the experimental maps directly, while MAID (Levitt, 2001), which works better at medium resolution could autotrace several fragments (totally about 100 residues). On this basis, more residues were traced by hand, with the help of the ERp29 structure, which is Wind's homologue and solved by NMR method (Liepinsh *et al.*, 2001). This NMR model fitted roughly in the b-domain, but deviated quite a lot in the D-domain. In fact, the ERp29 structure was once used as a search model for MR method but was not successful.

The main chain of this partially built model was extended by ARP/wARP. The keyword of "keepmodel" was used. Some new residues were traced. In the meantime, the output phases of ARP/wARP were refined with SHELXE (e.g., shelxe 2fofc.phi -m10 -s0.43). And a new map was calculated from these refined phases. Upon this new map some new residues were traced with MAID and/or by hand. Afterwards, all the available structure elements were combined together to provide a better partial model, which was extended with ARP/wARP again. The above processes were iterated several cycles.

In the first several cycles, the residues output from ARP/wARP were even less than the input. However, with the iteration going on, the output turned better than input. Finally, about 360 residues were output from ARP/wARP, from which one almost complete monomer could be constructed for the main chain. The corresponding phases were refined with SHELXE and generated a high quality map (Figure 2-4d). And the electron density of many side chains turned clear.

Some side chains were added automatically with MAID, others were added by hand with XFIT according to the sequence. The side chains were truncated for those residues that didn't have clear electron density for side chain atoms. However, all the residues including the truncated ones kept their correct nomenclature to have the restraints properly. Combining all the structure elements, an almost complete dimer model was built and used for later refinement.

2.8. Model refinement

The refinement was carried out with REFMAC5 (Murshudov *et al.*, 1997) using a maximum likelihood target on amplitudes, which is optimized by CGMAT method. The model was refined to 1.9 Å resolution and the 5% previously selected reflections were not included in refinement but as a cross validation dataset to calculate R_{free} (Brunger, 1993). The atomic coordinates and isotropic B factors were refined with proper restraints. The bulk solvent model was used. An overall anisotropic B factor was refined as well. NCS restraints were tested but not used.

During each cycle of refinement, the model was adjusted with XFIT based on the σ_A-weighted 2mFo-DFc and mFo-DFc maps (Read, 1986). The Fourier coefficients to calculate these maps were output by REFMAC5. In addition, experimental maps were also used to minimize the model bias. Only those atoms that were clearly defined by the electron density were included in the model. Solvent molecules (water) were added with XFIT where the peaks were above 2σ in the σ_A-weighted mFo-DFc map and were kept if their B factors were not bigger than 55 $Å^2$ after refinement. One cesium atom was found for its high electron density. It was distinguished from calcium atom by the bond geometry. This cesium was confirmed by the refinement against the Hg derivative data (CsCl was not used in the solutions for this crystal, however, $CaCl_2$ was still used), where this large electron density disappeared. Refinement statistics are summarized in table 2-3. The coordinates and structure factors have been deposited in the Brookhaven Protein Data Bank with the entry codes 1OVN.

2.9. Structure analyses

2.9.1. Molecular geometry

Main chain torsion angles were analyzed with PROCHECK (Laskowski *et al.*, 1993). A detailed check was also performed on WHAT IF (Vriend, 1990) server (http://www.cmbi.kun.nl/gv/servers/WIWWWI/). Secondary structure elements were identified with WHAT IF using the DSSP method (Kabsch and Sander, 1983).

2.9.2. Surface electrostatic potentials

The surface electron potentials were calculated with the program MEAD. In MEAD (Bashford, 1997), the protein molecule is assumed to be an object of low dielectric constant (typically 1.0-4.0) with embedded charges, while the solvent is modeled as a high dielectric continuum (typically 80). The locations of the embedded charges, the shape of the dielectric boundary and the boundary of the electrolytic region are determined by the coordinates and radii of the atoms from the 3-D model. The electrostatic potential Φ(r) is usually governed by the linearized Poisson-Boltzmann equation:

$$\nabla\varepsilon(r)\nabla\phi(r) - \kappa^2(r)\varepsilon(r)\phi(r) = -4\pi\rho(r) \qquad (2\text{-}2)$$

where ρ is the charge distribution, ε is the dielectric constant which takes on different values in the molecular interior or exterior, and κ is a parameter that represents the effect of mobile ions in solution. In MEAD, a finite difference method is usually used to solve this equation.

The electrostatic potential calculation requires the complete protein structure model. The missing side chain atoms including hydrogens were re-constructed according to the geometry with the "generate.inp" script in CNS (Brunger *et al.*, 1998). A few of the missing residues at the N-terminus and C-terminus were not constructed for their uncertain positions. Then the charges and radii of atoms were assigned with "assign_parse.pl" in MEAD. The cysteines in disulfide were named CSS and treated differently from free ones. The histidines were treated as neutral ones and named as HID. The electrostatic potentials were calculated with the program POTENTIAL in MEAD package, setting $\varepsilon_{protein}$ = 4, *ionstrength* = 0 and defaults for other parameters.

2.9.3. Surface hydrophobic potentials

The surface hydrophobic potentials were calculated with the drug discovery package GRID (Goodford, 1985). GRID combines hydrophobic effects, hydrogen-bonding interactions and induction/dispersion interactions together to calculate interaction energy between the hydrophobic probe and protein, which is better than simply defining the hydrophobic surface based on atom or residue type. The hydrophobic probe "DRY" was used to calculate the hydrophobic potentials on the protein surface. The complete model as used in the electrostatic

potential calculations was also required for a proper hydrophobic potential calculation. In addition, the model was supposed to have a slight flexibility (MOVE=1).

2.9.4. Other analyses

Crystallographic contacts were analyzed with CCP4 (CCP4, 1994). And the intermolecular contacts were analyzed with the protein interaction server (http://www.biochem.ucl.ac.uk/bsm/PP/server/) (Jones and Thornton, 1996).

The surface was calculated with MSMS (Sanner *et al.*, 1995) with a probe of 1.2Å radius.

The B factors of Wind were analyzed by BAVERAGE (CCP4, 1994) in CCP4 package.

The alignment based on sequence was done with CLUSTERW (Thompson *et al.*., 1994). While the alignment based on 3-D structure was performed with SSM (http://www.ebi.ac.uk/msd-srv/ssm/) or DALI (http://www.ebi.ac.uk/dali/) **(**Holm and Sander, 1993**)**, which gave the superposing matrixes for the matching molecules as well. LSQKAB (CCP4, 1994) and XFIT were also used for molecular least-squares superposition. Domain motion was analyzed with the DynDom program (Hayward and Lee, 2002).

The molecules were represented with XFIT (McRee, 1999), DINO (http://www.dino3d.org), MOLSCRIPT (Kraulis, 1991), RASMOL (Sayle and Bissell, 1992) and RASTER3D (Merrit and Murphy, 1994).

F2MTZ and MTZ2VARIOUS (CCP4, 1994) were used to convert data formats. MAPMAN (Kleywegt and Jones, 1996) was used to convert map formats. DINO was used to convert coordinate formats.

3. Results

3.1. Structure quality

The refined model has a well-restrained geometry with crystallographic R = 21.6%, Rfree = 25.6% (Table 2-3). One asymmetric unit contains two monomers: A & B. The residues in each monomer are numbered according to the complete protein sequence of Wind. Residues 24 to 252 have been modeled for monomer A and 23 to 248 for monomer B. Some residues at both termini (22, 253-257) and the N-terminal His_6-tag (MRGSHHHHHHGS) could not be seen in the electron density map. Most residues in the model have good electron density, while some residues in loop regions are not well defined and show unclear electron density (Fig 3-1 (a), (b) and (c)).

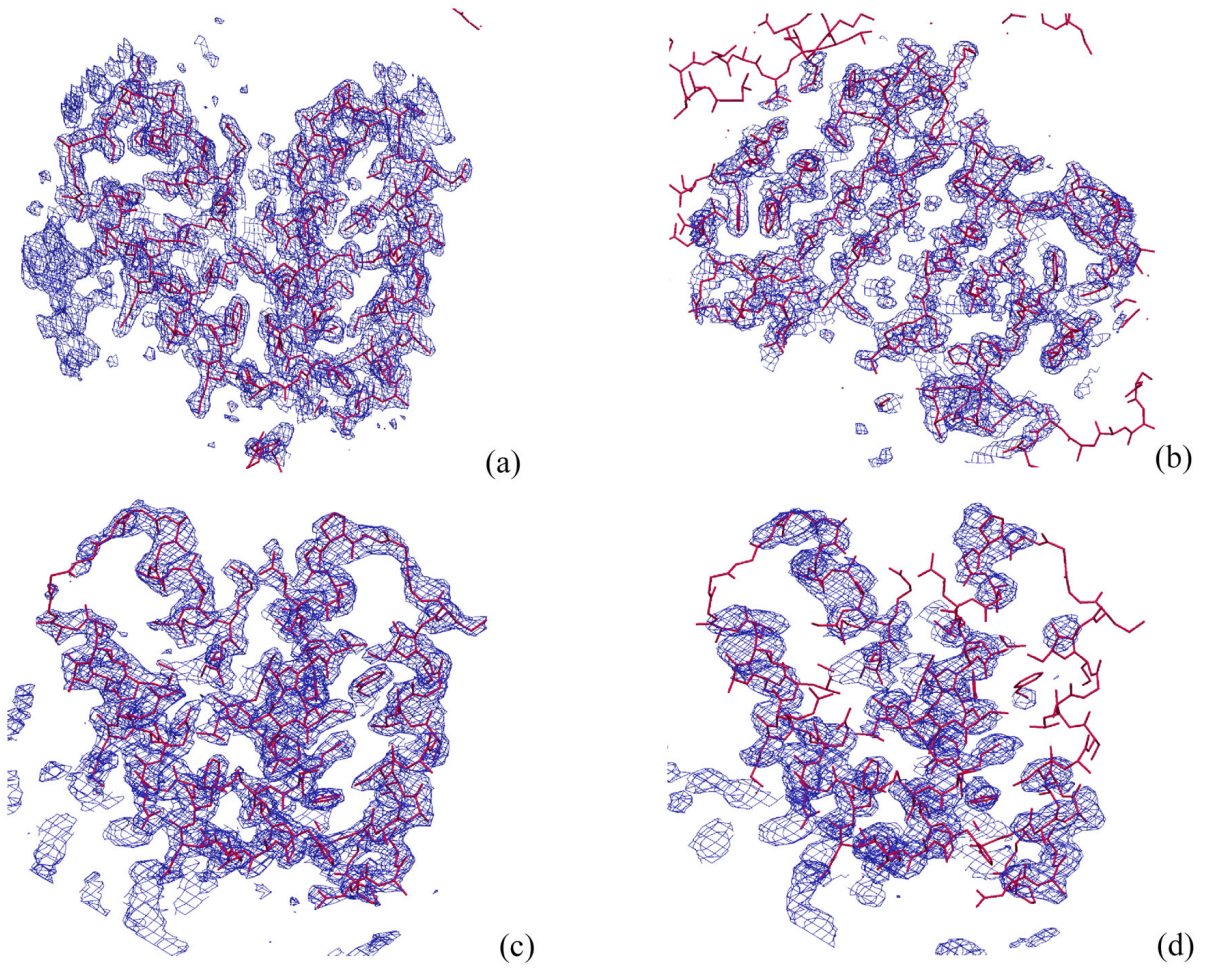


Figure 3-1 Electron density for Wind in some regions. (a), (b) and (c) show σ_A-weighted $2mF_o$-F_c electron density map of the refined model, while (d) shows the σ_A-weighted mF_o experimental electron density map calculated from SHELXE phases. (a).The D-domain of monomer A. (b). The b-domains, in the middle is the dimer interface. (c). The D-domain of monomer B. (d). The D-domain of monomer B in the same orientation of (c).

Some sidechain atoms lacking clear electron density are absent from the model. The electron density of the D-domain of monomer B is rather weak even for main chain, which is consistent with what has been seen in the experimental maps (Fig 3-1(d)).

The torsion angles were not restrained during the refinement, so they could be used as good monitors for the model quality. The molecular geometry of the model is excellent. In the Ramachandran plot (Fig. 3-2), 386 residues (~92.6%) are within the most favored region, 27 residues (~6.5%) lie in additionally allowed region, and 4 residues (~1.0%) lie in generally allowed regions. These 4 residues are located in flexible loop regions that are less well defined by the diffraction data. No residue lies in a disallowed region.

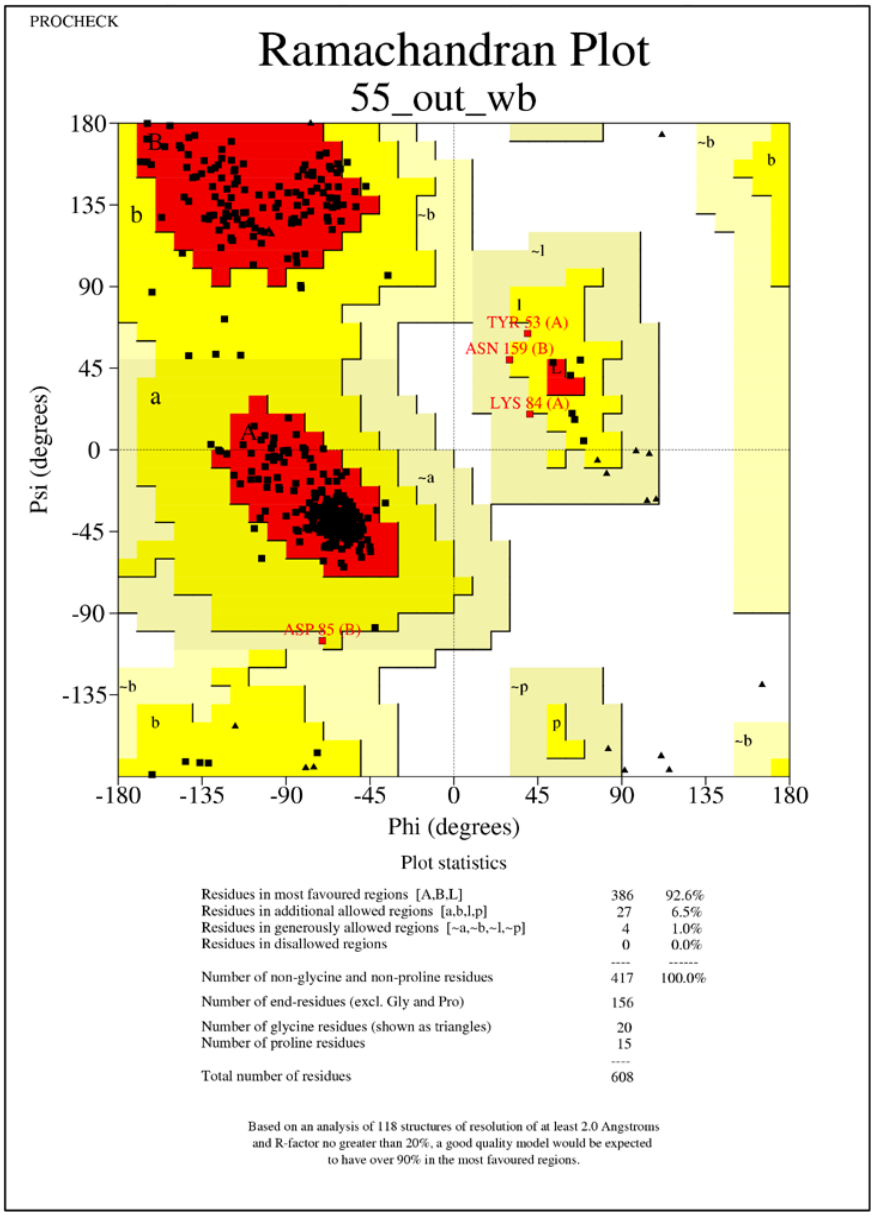


Figure 3-2 Ramachandran plot of the Wind structure.

Most of the detailed checks with WHAT IF score well. Nevertheless, a few atoms violate the antibumping restraints, which is not uncommon for a flexible protein molecule since the crystal structure is an average picture of both, time and space.

3.2. Overall structure

3.2.1. Monomer structure

Wind crystallizes as a homodimer, consisting of monomer A and B. The secondary structure elements of each monomer were analyzed with the WHAT IF program (Fig. 3-3). The structures of the two monomers differ slightly, as a result of a different environment in the crystal.

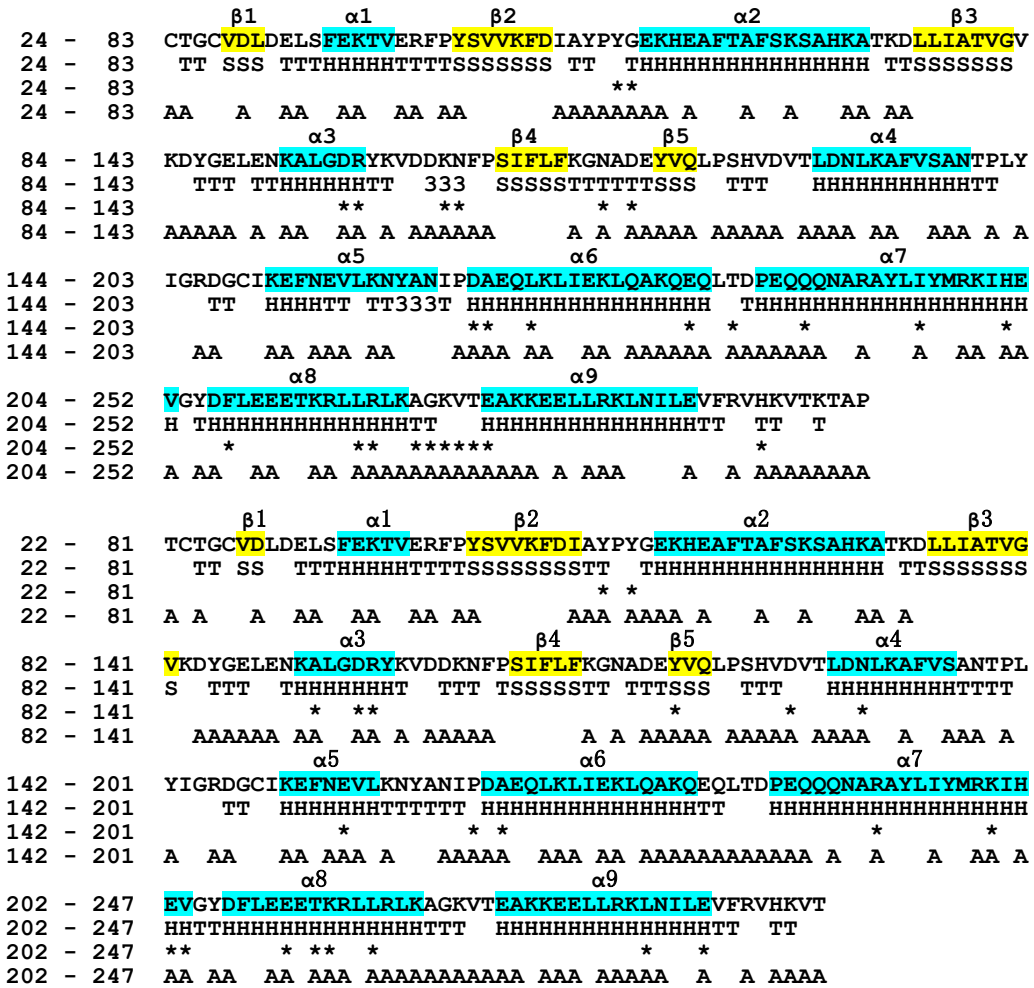


Figure 3-3 Secondary structure, symmetry and accessibility. The first block (residues 24-257) corresponds to monomer A, and the second block (residues 23-247) corresponds to monomer B. The type of secondary structure is marked at the top. Residues in the beta strands are shadowed yellow, and those in the helices are shadowed blue. Residues involved in symmetry contacts are labeled with an asterisk. Residues that are clearly solvent accessible are labeled with a capital A.

The secondary structure elements of Wind are arranged as β1-α1-β2-α2-β3-α3-β4- β5-α4-α5-α6-α7-α8-α9 (Fig. 3-4).

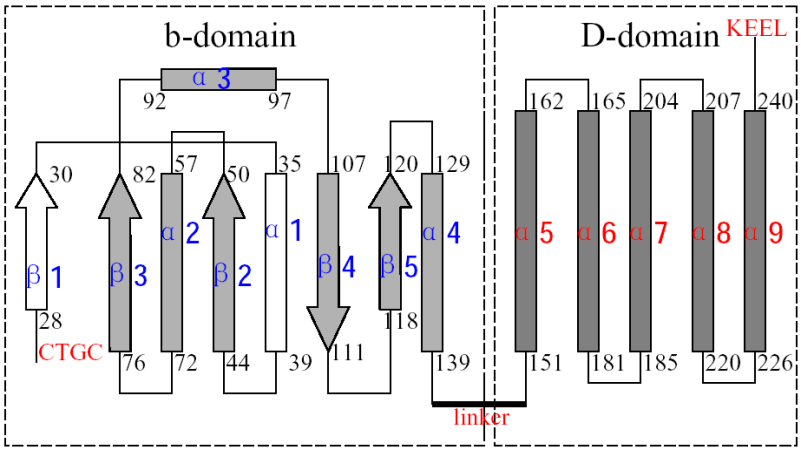


Figure 3-4 The topology of the Wind monomer. The start and end position of secondary structure elements are marked using residue number. The CTGC motif and the KEEL ER retention sequence are marked in their corresponding positions. In the b-domain region, the elements forming the thioredoxin fold are colored in gray.

Each monomer is composed of two distinct domains connected by a flexible short liker of 11 residues (140-150) (Fig. 3-5). At the N-terminus is a TRX-like b-domain of about 117 residues (23-139) from β1 to α4. The structure of b-domain is very similar to that of TRX: five beta strands form the core beta sheet with β4 antiparallel to the other strands, surrounded by four alpha helices. At the C-terminus is an alpha helical D-domain of about 102 residues (151-252) from α5 to α9. This domain is unique to the PDI-D subfamily. The five helices are up-and-down arranged. α5 is distorted and bends slightly at the middle part of the helix, thus looking more like two small helices connected by a short loop. In monomer B, the second part of this helix even looks like a loop. α6 and α7 are antiparallel and lie in one plane, thus form a rectangle. Similarly, α8 and α9 together form another rectangle plane. The two rectangle planes are roughly parallel but are rotated to each other by about 40˚ around an axis that is perpendicular to both planes. The whole D-domain is rather flat, thus many residues in this domain are solvent accessible (Fig. 3-3). The missing C-terminal tail (including the KEEL retention sequence) is probably disordered as a flexible loop pointing outside.

3.2.2. Dimer structure

The crystal structure of wind clearly shows a homodimer with dimensions of about 106 x 53 x 37 $Å^3$, formed by a head-to-tail arrangement of the N-terminal b-domains with no participation by residues within the D-domain (Fig. 3-5). The D-domains are separated by b-domains and lie in opposite sides.

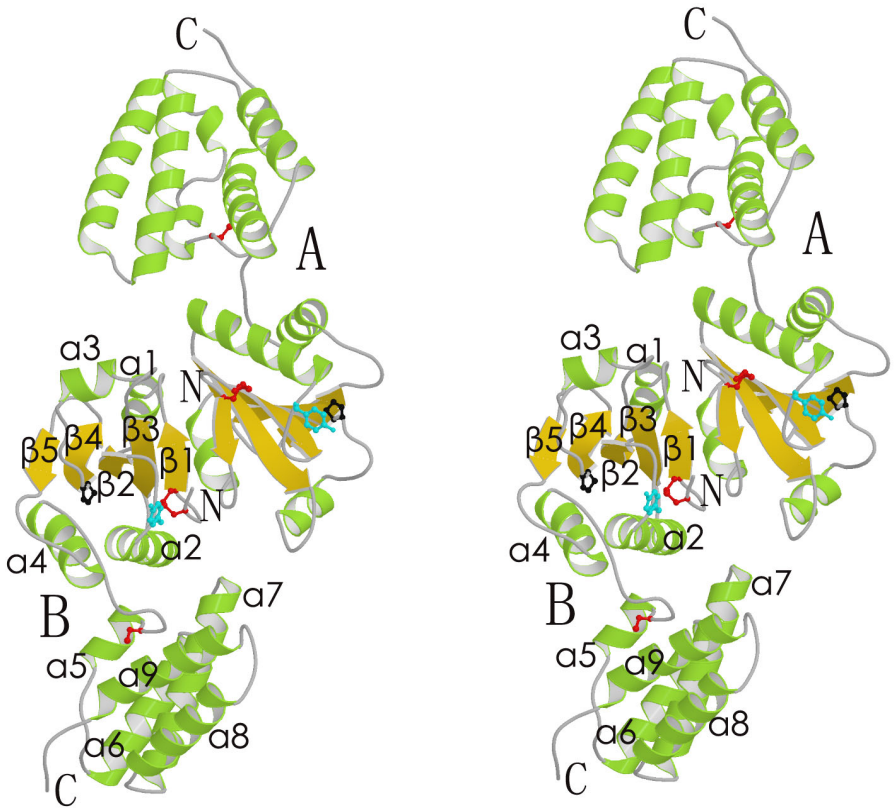


Figure 3-5 Stereo view of the structure of Wind dimer. The two monomers are marked with big "A" and "B", respectively. The N- and C-terminal are marked with "N" and "C", respectively. The secondary structure elements in monomer B are marked. The cysteines are presented as a red stick model. The Tyr55 is presented as a cyan stick model. The *cis*-P106 is presented as a black stick model. Beta strands are colored in yellow, and helices in green.

The dimer shows a loose 2-fold non-crystallographic symmetry for the b-domains of the two monomers. For each monomer, a solvent accessible area of about 758 $Å^2$ is buried in the interface, which corresponds to about 6.3% of the surface of the monomer (Table 3-1).

The interface is formed mainly by the residues on either side of β1 (residues 24 - 34), residues within and after α1 (residues 37 - 43) and residues within and between α2 and β3 (residues 67 - 77) via hydrophobic interactions as well as hydrogen bonding, and van der Waals interactions (Table 3-2).

Protein Interface Parameter	Value	
	monomer A	**monomer B**
Interface Accessible Surface Area	766.44	748.99
% Interface Accessible Surface Area	6.26	6.34
Planarity	1.36	1.32
Length & Width	30.66 & 19.90	28.99 & 19.35
Length/Width Ratio	0.67	0.72
Interface Residue Segments	2	2
% Polar Atoms in Interface	30.20	31.01
% Non-Polar Atoms in Interface	69.80	68.90
Secondary Structure	Alpha	Beta
Hydrogen Bonds	5	5
Salt Bridges	0	0
Disulfide Bonds	0	0
Gap Volume	7739.10	7739.10
Gap Volume Index	5.11	5.11
Bridging Water Molecules	8	8

Table 3-1 Information for the dimer interface.

monomer A

Residue	Interface ASA ($Å^2$)/ % Interface ASA	H-Bonds
CYS 24	1.93 / 0.26	.
THR 25	30.67 / 4.10	.
GLY 26	35.86 / 4.79	1
CYS 27	8.27 /1.11	.
VAL 28	63.74 / 8.52	.
ASP 29	22.29 / 2.98	.
LEU 30	2.52 / 0.34	.
ASP 31	37.81 / 5.05	1
LEU 33	100.10 / 13.38	.
SER 34	37.08 / 4.96	.
LYS 37	32.32 / 4.32	.
THR 38	29.91 / 4.00	.
ARG 41	130.93 / 17.50	2
PHE 42	63.04 / 8.43	.
PRO 43	21.79 / 2.91	.
LYS 67	10.55 / 1.41	.
HIS 70	64.11 / 8.57	.
THR 73	9.42 /1.26	.
LYS 74	35.42 / 4.74	1
ASP 75	4.24 / 0.57	.
LEU 76	6.11 / 0.82	.

(a)

monomer B

Residue	Interface ASA($Å^2$)/ % Interface ASA	H-Bonds
CYS 24	1.58 / 0.21	.
THR 25	24.16 / 3.16	.
GLY 26	39.54 / 5.17	1
CYS 27	5.92 / 0.77	.
VAL 28	65.08 / 8.50	.
ASP 29	23.10 / 3.02	.
LEU 30	3.45 / 0.45	.
ASP 31	37.88 / 4.95	1
LEU 33	90.29 / 11.79	.
SER 34	35.60 / 4.65	.
LYS 37	97.69 / 12.76	.
THR 38	30.61 / 4.00	.
ARG 41	129.62 / 16.93	1
PHE 42	56.91 / 7.43	.
PRO 43	16.17 / 2.11	.
LYS 67	12.38 / 1.62	.
HIS 70	51.24 / 6.69	.
LYS 71	3.92 / 0.51	.
LYS 74	26.92 / 3.52	1
ASP 75	11.70 / 1.53	1
LEU 77	1.80 / 0.24	.

(b)

Table 3-2 Residues in the dimer interface. (a). Residues in monomer A. (b). Residues in monomer B. ASA: accessible surface area. The ASA of Lys37 shows an abnormal difference between the two chains, which is because some side chain atoms are not included in the refined model.

The strands β1 of the two monomers are roughly antiparallel, but do not have obvious direct interactions to form a continuous beta sheet. Instead they interact indirectly through a hydrogen bonding network mediated by 5 water molecules (Fig. 3-6).

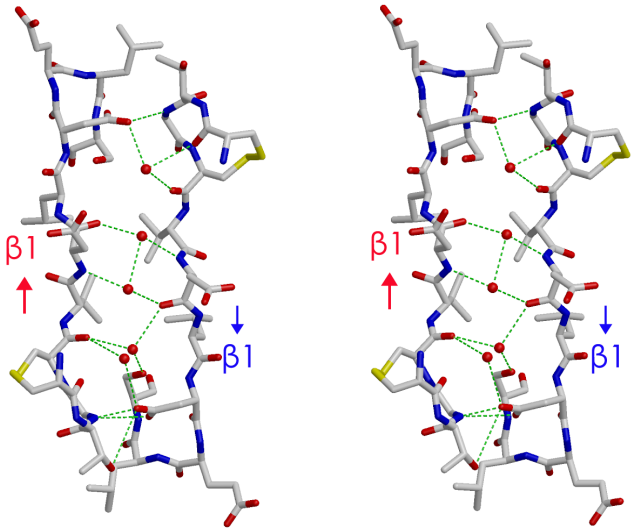


Figure 3-6 Stereo view of the hydrogen bonding network at the dimer interface. Strand β1 of monomer A (marked with β1) interacts with that of monomer B (marked with β1) via a hydrogen bonding network mediated by 5 water molecules. The direction (from N- to C-terminus) of the strands is shown with a red arrow for monomer A, and a blue arrow for monomer B.

Dimerization creates a deep hydrophilic cleft with approximate dimensions of 11 x 11 x 27 $Å^3$, between the b-domains of the monomers (Fig. 3-11b1). This cleft (hereafter referred to as dimer cleft), is flanked by residues from the loop between β2 and α2 and residues from the end of β3 and the following loop, with residues within and around β1 (residues 24-34) at its base (Fig. 3-5).

3.3. Temperature factors

The temperature factor (B-factor) in crystallography describes the atomic displacement related to the atomic positional coordinates. The displacement may be caused by the atom's dynamic vibration or static positional variance in space. It reflects the variability of the atomic coordinates in both time and space. The larger the B-factor is, the higher the flexibility would be. For protein structure, the B-factors of main chain are usually analyzed to study the molecular flexibility. Generally, main chain are buried in protein, so the B-factors of main chain may reflect the intrinsic flexibility of the structure. Thus, only the B-factors of main chain are analyzed for Wind structure (Fig. 3-7).

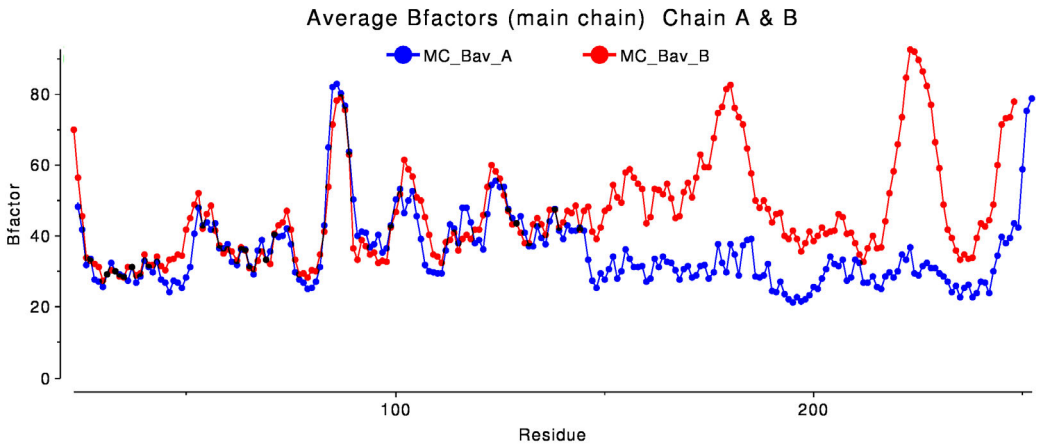


Figure 3-7 B-factor plot for the main chain of Wind. Red: monomer A; Blue: monomer B.

In the b-domain, the B-factors for main chain are comparable for both monomers (with a mean value of 39.3 and 41.0 $Å^2$ for monomer A and B, respectively). The flexible parts are located around turns and loops connecting secondary elements (Fig 3-8). The loop between β3 and α3 (84-90) shows particularly high flexibility in both monomers. It overlooks the dimer cleft and points to the D-domain of the second monomer.

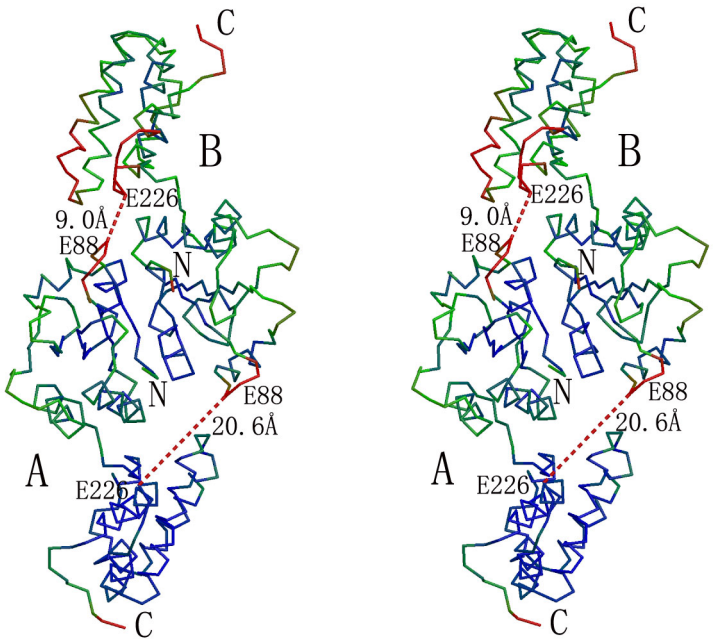


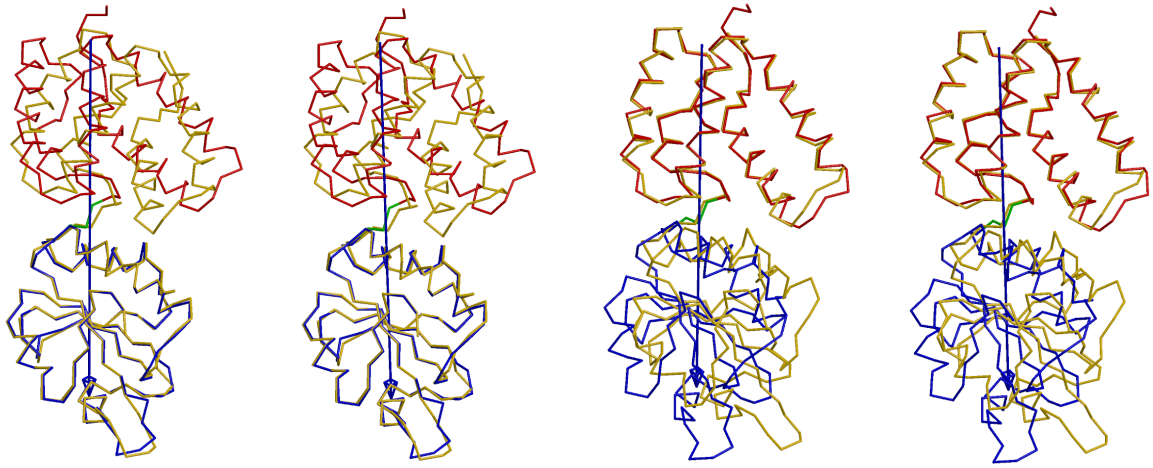
Figure 3-8 Stereo view of Cα trace of Wind dimer. The two monomers are marked with big "A" and "B", respectively. The N- and C-terminus are marked with "N" and "C", respectively. The residues are colored according to their B-factors. For B-factors: less than 25,colored in blue; 25-50, color ramped from blue to green; 50-75, ramped from green to red; bigger than 75, in red. The distance between E226 and E88 is marked.

In the D-domain, the B-factors for main chain of monomer B are much higher than those in monomer A (with a mean value of 31.2 and 53.3 $Å^2$ for monomer A and B, respectively). The C-terminus has a high flexibility in both monomers. In monomer B, the flexible parts are additionally located around the turn between α6 and α7, and that between α8 and α9, including some regions in helices. These flexible parts point to the b-domain of the other monomer. The very high B-factors correspond to the weak electron density in these regions. The big differences of B-factors (in the D-domain) between monomer A and B result from their different environment in the crystal (Fig. 3-3). The D-domain of monomer A has tighter crystallographic contacts and thus lies in a more compact environment than that of monomer B. Therefore, the motion of the D-domain in monomer A is highly constrained by the crystal contacts. We can imagine that the D-domain may be more flexible in solution.

The structure is defined by B-factors as well as coordinates. Here, in spite of the coordinates, the B-factors of the D-domains of monomer A and B are so different that we can say the D-domain of two monomers have a different structure in the crystal. As seen, the electron density of the D-domain is distinct for two monomers. The idea of NCS restraints are that the structures will be similar if they have the same chemical constitutions, so averaging would make a better model. Here, NCS restraints may lose their validity. Tests were done to decide using or not using NCS restraints in refinement. After comparison, the NCS restraints were not applied.

3.4. Comparison of the two monomers

A superposition of Cα atoms of the individual domains of the two monomers shows excellent correlation of their 3D structures with mean r.m.s.d. values for Cα coordinates of 0.61 Å and 0.92 Å for the b-domains and the D-domains respectively (Fig. 3-9). A mean r.m.s.d. value of 26.9 $Å^2$ in B-factors between corresponding Cα atoms of the D-domains (compared to the value of 5.6 $Å^2$ between the b-domains) indicates considerable mobility within the D-domain. This has been shown by the B-factor analysis (Fig 3-7).



(a) (b)

Figure 3-9 Stereo view of the superposition of the two monomers. (a). Superposition of the b-domains. (b). Superposition of the D-domains. Monomer B is colored in yellow. In monomer A, the b-domain is colored in blue; the D-domain is colored in red; the primary bending residues (Y143, I144 & G145) in the linker region are colored in green. The rotation axis is presented as a long blue arrow.

Interestingly, the two domains of both monomers could not be superimposed simultaneously, reflecting an ~32º relative rotational displacement of the D-domains which could be traced to the flexible linker region connecting the N- and C-terminal domains, and primarily to the main chain bond rotations around the residues, Y143, I144 and G154. Detailed information about the divergence in the linker region is listed in table 3-3.

rotation angle around the axis (º)	translation along the axis (Å)	bending residues	residue i	residue i+1	distance of hinge axis to residue i in monomer A/B(Å)	changes in psi(i) /phi(i+1) (º)	angles of psi(i) axis to the hinge axis in monomer A/B (º)	progress (%)
32.2	0	Y143 -G145	L142	Y143	7.0/6.7	2.3/-2.5	116.4/111.9	21.1
			Y143	I144	5.7/5.3	-1.0/-4.3	63.6 /61.6	10.0
			I144	G145	2.3/2.1	-12.0/1.7	45.5 /54.7	46.1
			G145	R146	3.5/3.8	5.3/4.7	103.6 /100.3	22.8

Table 3-3 Details of the linker region. The values in the table show how the D-domain rotated from the position in monomer B to that in monomer A (Figure 3-9a).

One consequence of this displacement is that, in the crystal, the D-domain rotates in a manner such that the distance between the tip of the D-domain of monomer B to the dimer cleft is only 9.0 Å (between Cα of E226 of monomer B and E88 of monomer A), while this distance is 20.6 Å (between Ca of E226 of monomer A, E88 of monomer B) for monomer A (Fig. 3-8).

Such a domain motion may be induced by the crystal packing, which is very important to form a proper crystal lattice. However, there is no such constraints in solution. We can imagine that the D-domain may rotate freely in solution. The rotation angle limit is not studied yet.

3.5. The conserved residues on the protein surface

As a consequence of convergent evolution, those residues that are important for both, structure and function, are conserved for all members of one protein family. The identical residues in the four members of PDI-Dβ family (Wind, rat ERp29, human ERp28 and mouse ERp29) are marked in the sequence alignment (Fig. 3-10). Most of them are important for the structural fold, while others may be important for function.

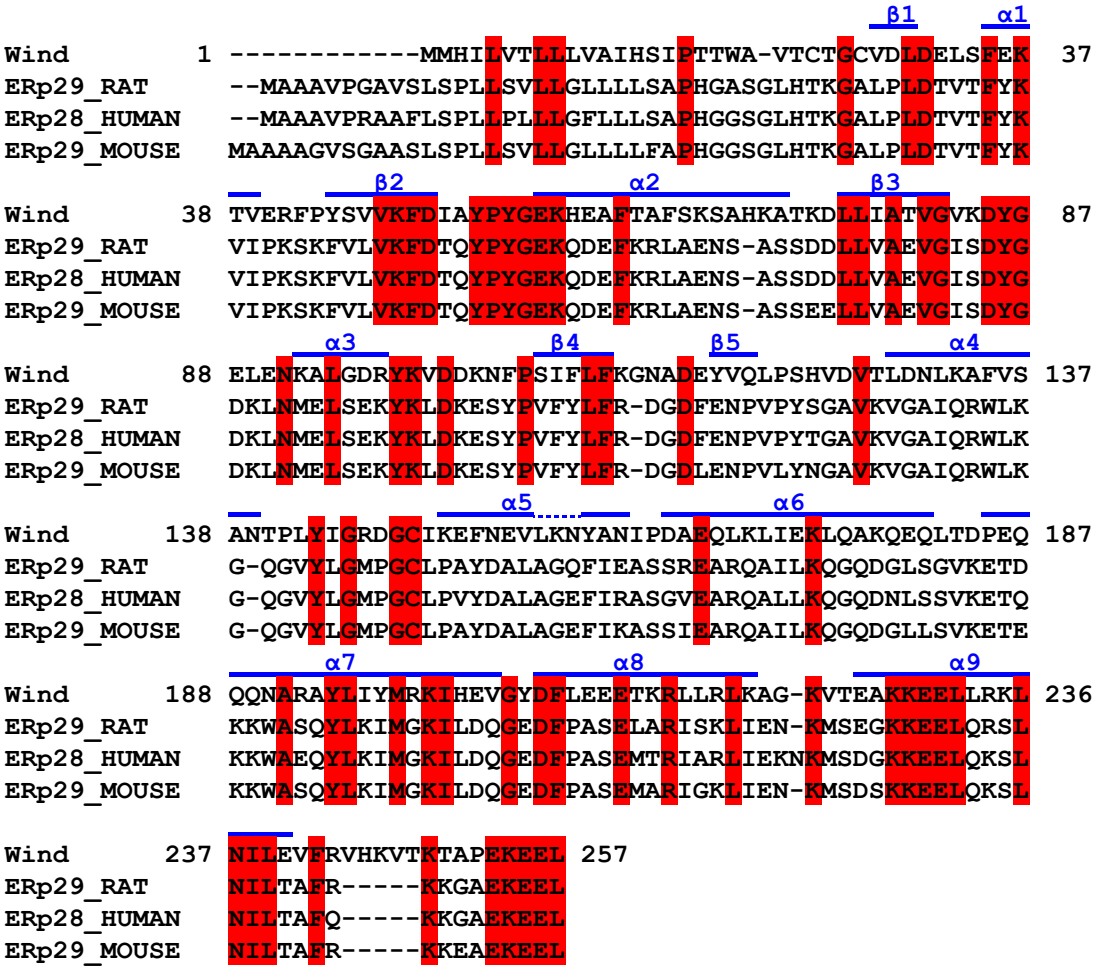


Figure 3-10 Sequence alignment of PDI-Dβ proteins. The type of secondary structure is shown at the top, and the residue range is marked with a blue bar. The identical residues among all four proteins are marked in red.

Since protein interactions tend to occur on the protein surface, the functional residues tend to appear on the surface and form a cluster. Here, these conserved residues in PDI-Dβ proteins show a pronounced pattern on the surface (Fig. 3-11a1, a2).

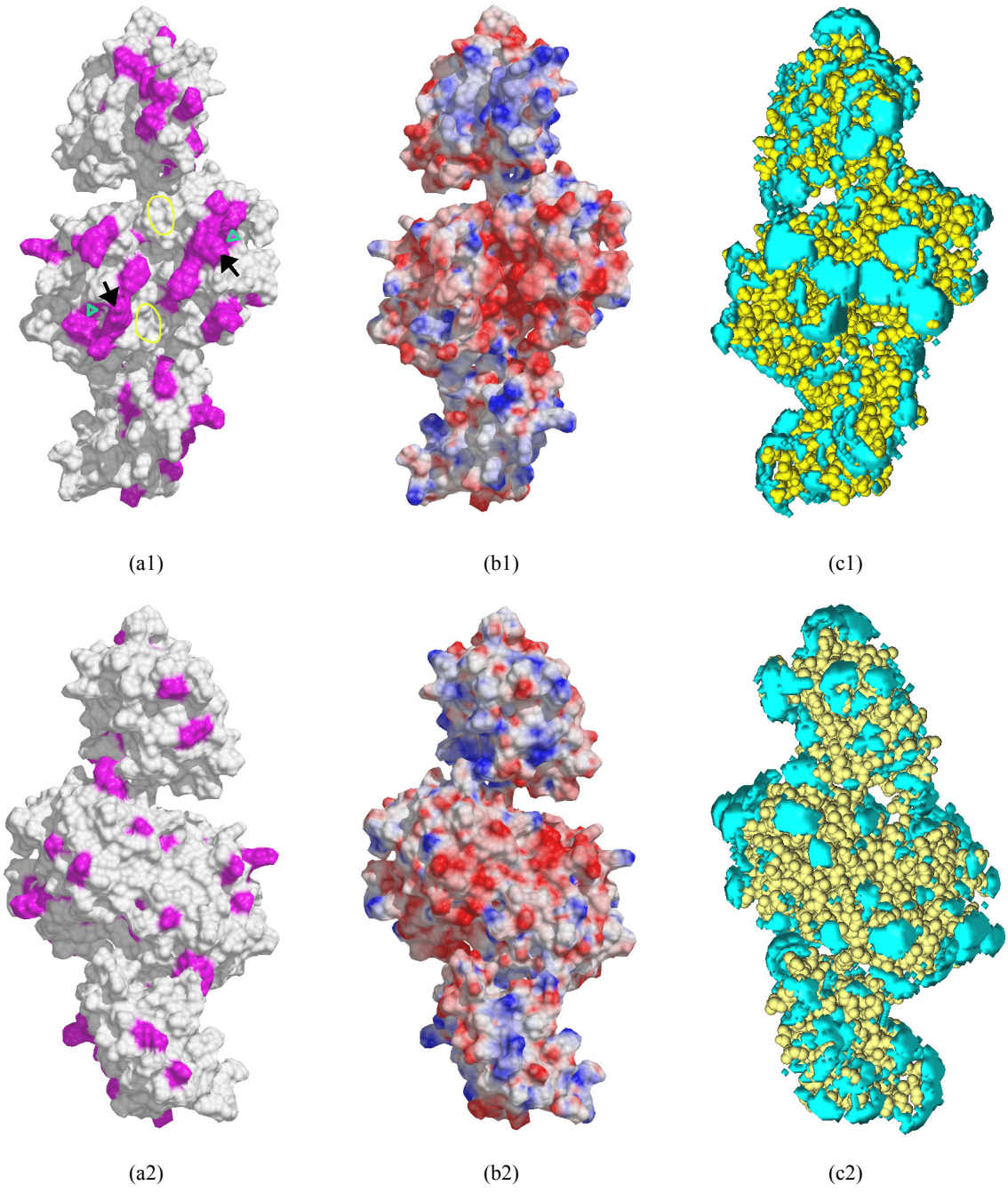


Figure 3-11 The surface features of Wind dimer. (a1). Conserved surface in the side with the dimer cleft. The orientation is the same as in Fig. 3-5.The residues identical in PDI-Dβ proteins (Fig. 3-10) are colored in magenta. The position of CTGC motif is marked with a yellow circle; P106 with a small green triangle; and Y55 with a black arrow. (a2). Conserved surface. Color as in (a1). The dimer is rotated 180° compared to (a1). (b1). Surface electrostatic potentials. Orientation as in (a1). Negative potential is colored in red, positive potential in blue. (b2). Surface electrostatic potentials. Orientation as in (a2). Colors as in (b1). (c1). Surface hydrophobic potentials. Hydrophobic patches are shown in cyan. Orientation as in (a1). (c2). Surface hydrophobic potentials. Colors as in (c1); Orientation as in (a2).

Most of them are distributed on the side with the dimer cleft. Particularly, the conserved residues, D50 in β2, Y53 to G56 in $loop_{β2-α2}$ (meaning the loop connecting β2 and α2), E57 and K58 in α2, and P106 preceding β4 together form a cluster on the surface of each b-domain. Such clusters lie in the bank areas of the dimer cleft. Except for D31, no other surface residues on the base of the dimer cleft are conserved. The conserved surface residues in the D-domain are located mainly in α8 and α9.

3.6. The electrostatic potentials on the surface

The b-domains show a slightly negative potential, while the D-domains show a slightly positive potential (Fig 3-11b1, b2). This is consistent with the theoretical pI value of 5.1 and 8.1, for the b-domain and D-domain, respectively. Very obviously, the dimer cleft shows a negative potential. The residues D29, D31, E32 around β1 and E88, E90 in the $loop_{β3-α3}$ from each monomer significantly contribute to this negative potential. Only D31 is completely conserved in other PDI-Dβ proteins. The main chain oxygen atoms of β1 from both monomers pointing towards the cleft also contribute to this negative potential. Moreover, no positive residues lie in this region to counteract this negative potential.

3.7. The hydrophobic patches on the surface

No large hydrophobic patches are found on the dimer surface. However, some small patches on top of the dimer cleft add up to form a relatively large hydrophobic area (Fig. 3-11c1,c2). Residues I51, A52, Y53 and Y55 in the $loop_{β2-α2}$, Y86 in the $loop_{β3-α3}$ and F105 in the $loop_{α3-β4}$ from each monomer contribute significantly to this relatively extensive hydrophobic surface area. Nevertheless, the dimer cleft is nearly completely hydrophilic.

3.8. The cysteines and the CTGC motif

Since Wind is a PDI-related protein, cysteine residues may play an important role in its function. Wind contains three cysteines in each monomer. One cysteine (C149) is located in a β-turn of the linker region and has a free thiol which is solvent accessible. The two cysteines (C24 and C27) in the N-terminal CTGC motif form a disulfide bond in both monomers (Fig. 3-12). Both two sulfur atoms in the CTGC motif are buried inside the protein and are not

solvent accessible. Such a CTGC tetrapeptide, which does not occur in PDI-Dβ homologues from other species, is located within the dimer cleft about 10 Å away from the classical CXXC site and about 20 Å from the CTGC motif of the other monomer.

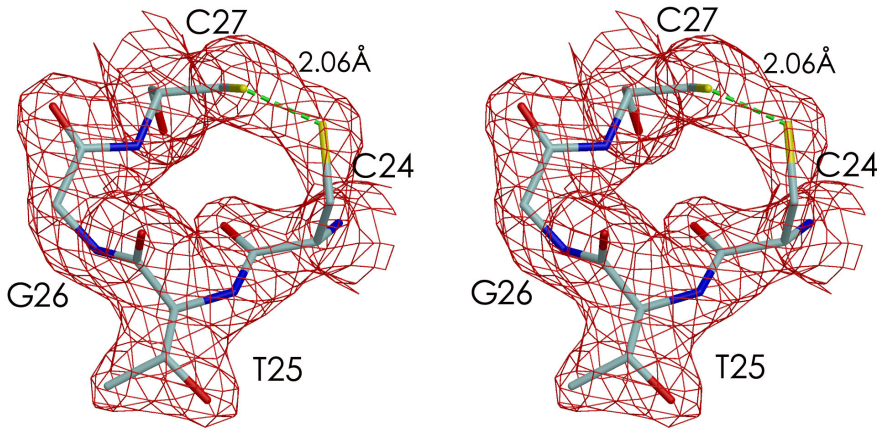


Figure 3-12 Stereo view of the CTGC motif of monomer B. The residue names are marked at their corresponding positions. The distance between two sulfur atoms is shown. Their electron density (red) is shown (contoured at 1σ level in σ_A-weighted 2mFo-DFc map).

In the Hg derivative, Hg^{2+} ions are not only bound to the free cysteines in the linker region but also bound to the CTGC motif of monomer B. But no observable Hg^{2+} is bound to the CTGC motif in monomer A. The reason for this asymmetric Hg^{2+} binding is not clear. Hg^{2+} usually has a high affinity to free thiol or thiolate and rarely binds to disulfide. This may indicate that the disulfide bond in the CTGC motif of monomer B could be partially opened in the Hg derivative crystal.

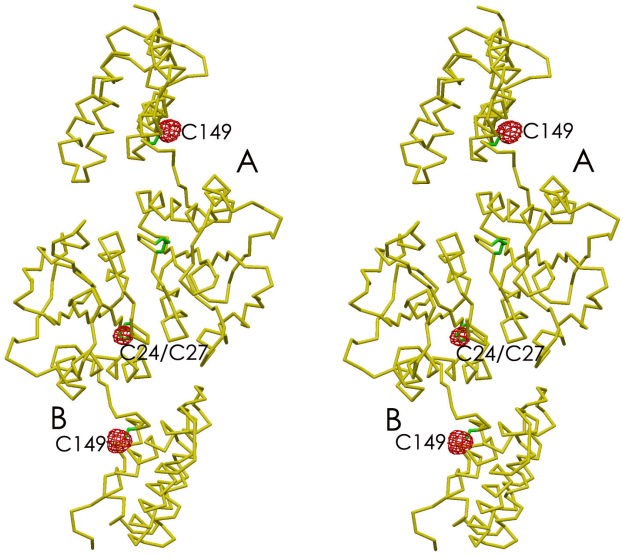


Figure 3-13 Stereo view the Hg^{2+} binding sites. The Cα trace of the dimer is shown in yellow. The two monomers are marked with big "A" and "B", respectively. The cysteines are labeled and represented in green using stick model. Binding Hg^{2+} ions are presented with their electron density (red), contoured at 5σ level in σ_A-weighted experimental map.

3.9. The *cis*-proline

As many other TRX-like proteins, each monomer of Wind contains a proline (P106) in the less common *cis* conformation (Fig. 3-14), located in the loop preceding β4, which is conventionally called *cis*-Pro loop.

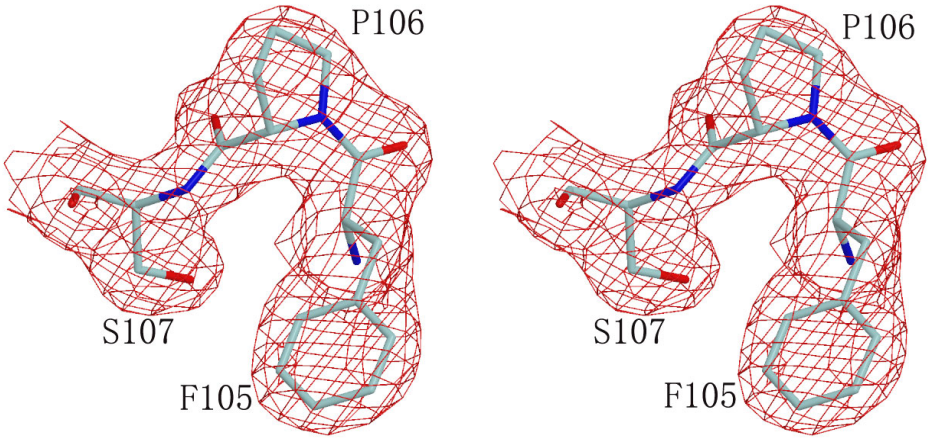


Figure 3-14 Stereo view of the *cis*-Pro106 of monomer B. The residues are labeled. Their electron density (red) is shown (contoured at 1σ level in σ_A-weighted 2mFo-DFc map).

P106 is located just before β4 and the *cis* peptide bond between F105 and P106 introduces a sharp turn, leading the subsequent β4 into the protein interior to form the structure core (Fig. 3-5). The corresponding *cis*-proline in other TRX-like proteins, such as TRX, is not only very important to maintain the local structure in this region but also involved in their activities (Qin *et al.*, 1996). This proline is conserved in PDI-Dβ proteins. Surprisingly, the corresponding proline reported in the NMR structure of ERp29 is *trans* (Liepinsh *et al.*, 2001).

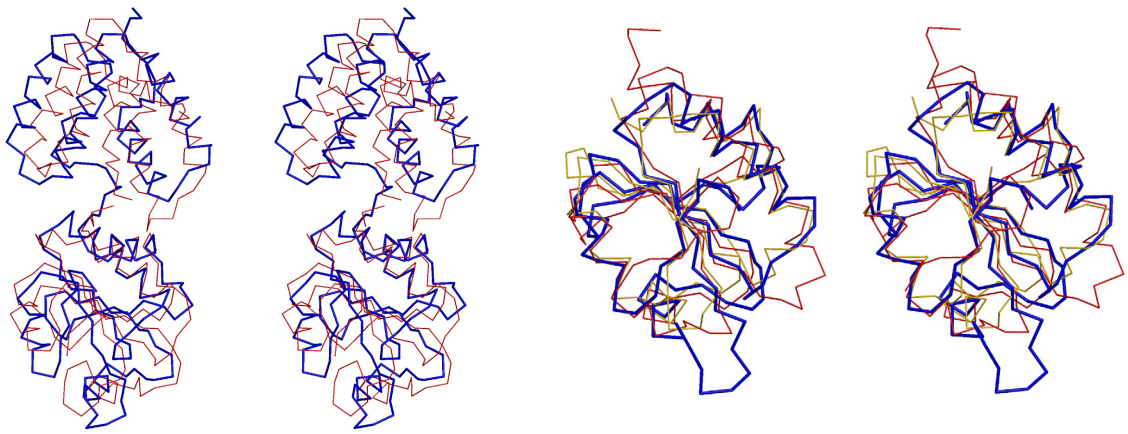
3.10. Comparison of Wind and other PDI-related proteins

A structure based alignment of Wind and other PDI-related proteins (ERp29, PDI a and b domains) is made for the TRX-like b-domain (Fig. 3-15). Their secondary structures can be aligned roughly well, but there is a conspicuous pentapetide (DYGEL) present in Wind, but absent from both PDI a and b domain. This pentapeptide loop (D85-L89) of unknown function is located at the turn between β3 and α3. This charged loop is surface exposed and overlooks the dimer cleft. It is partially conserved in the PDI-Dβ proteins (Fig 3-10).

Figure 3-15 Alignment of wind with rat ERp29, PDI a and b domains, based on structure. The type of secondary structure is marked at the top, residues within the elements are shadowed yellow. The partially conserved pentapeptide insert found in PDI-Dβ proteins is colored in blue. The position of Y55 is marked in magenta. The CTGC motif is colored in red.

The b-domain and D-domain of ERp29 are superposed to the corresponding domains of Wind, respectively (Fig. 3-16a). In spite of differences in detail, the b-domains fit fairly well with an r.m.s.d. value of 2.6 Å for all α carbon (Cα) atoms. However, the D-domains are so different in their helical orientations that they can not be superimposed, although their folds are similar.

Superposition of the structures of the TRX-like domains of Wind and PDI renders r.m.s.d. (Cα) values of ~1.9Å for Wind and PDI a domain, and ~1.7Å for Wind and PDI b domain, respectively (Fig 3-16b). The additional pentapeptide loop in Wind lacks the compartment in both domains of PDI.



(a) (b)

Figure 3-16 Stereo view of the superpositions of Wind and other PDI-related proteins. (a). Superposition of Wind and ERp29. The Cα trace of Wind is colored in blue, ERp29 colored in red. The b-domain and the D-domain of ERp29 are superposed onto the corresponding part of Wind, respectively. (b). Superposition of Wind b-domain and PDI a- & b domain, respectively. Wind colored in blue, PDI a domain in red, PDI b domain in gold. The protruding loop in Wind (at the bottom of the map) is the additional pentapeptide in PDI-Dß proteins.

4. Discussion

4.1. The dimer

4.1.1. Wind exists as a dimer both in the crystal and in solution

The crystal structure of Wind shows a clear dimer and the TRX-like b-domain acts as the homodimerization module. The buried area in the interface for each monomer of ~758 $Å^2$ is much higher than the normal crystal packing contact (mostly below 300 $Å^2$) and is well within the range of the buried surface areas observed for protein homodimers (368 $Å^2$ to 4746 $Å^2$) (Jones and Thornton, 1996). However, the buried area is much smaller than the theoretical value of about 1860 $Å^2$ for a protein of MW = 27 kD and is only about 6.3% of the total protein area, which indicates a rather weak homodimer. Such a dimer may easily dissociate into two monomers. Does Wind really exist as a dimer in solution?

Dimerization of Wind in solution is supported by multiangle light scattering (MALS) measurements following size exclusion chromatography , in vitro and in vivo cross-linking and native gel analysis of the purified recombinant, His6-tagged protein, which suggest a dimer of 2x27 kDa apparent molecular mass (Ma *et al.*, 2003). The protein sample used for crystallization is also recombinant His_6-tagged Wind. Nevertheless, the His_6-tag is not involved in the interface, thus it is unlikely that dimerization is induced by the His_6-tag.

The homologue of Wind in rat, ERp29 has been proved to exist as a dimer as well as a monomer or high oligomer (Mkrtchian *et al.*, 1998). In fact, many other TRX-like proteins such as TRX, DsbA, calsequestrin, PDI and so on, can exist as a dimer. So, the dimer appears to be a usual oligomeric state for TRX-like proteins. Of course, these proteins also exist as monomers or form other oligomers.

So, we may safely draw the conclusion that, Wind exists as a dimer, both in the crystal and in solution.

4.1.2. Dimerization yields a significant dimer cleft

The dimerization of Wind yields a significant dimer cleft, which has some conspicuous features.

(1). The dimension of this cleft is large enough to hold a small peptide (Fig 3-11). The

main chain oxygen atoms of β1 from each chain point towards the cleft, providing potential hydrogen bond acceptors.

(2). This cleft is hydrophilic and negatively charged, which may provide an electrostatic mechanism to bind the substrate (Fig 3-11b1). If this were true, it would be analogous to the mechanism suggested for the interaction of the sarcoplasmic reticulum protein calsequestrin with the membrane protein Junctin (Zhang *et al.*, 1997). Indeed, in Pipe, Wind's putative substrate, a region predicted to be within the luminal domain (S96-N139) shows a high content of basic residues (pI = 11.3), distinct from other, more C-terminally located sequences with net negative charge (Ma *et al.*, 2003).

(3). Although Wind lacks large hydrophobic patches that are normally present in chaperons, there is a considerable hydrophobic band spanning the cleft (Fig. 3-11c1). This may provide a mechanism to bind the substrate by hydrophobic interactions as other chaperons do.

(4). Many conserved surface residues in the PDI-Dβ family are located in the bank areas of the cleft (Fig. 3-11a1). Particularly, a conserved loop flanking the cleft is required for the function of Wind (Ma *et al.*, 2003).

(5). The possible substrate binding sites lie in the bank areas of the cleft (see 4.3.).

(6). The CXXC motifs are located at the bottom of this cleft and guard the entrance (see 4.2.).

(7). An additional pentapeptide loop of unknown function that is unique in PDI-Dβ proteins, overlooks the cleft.

So dimerization seems to be biologically significant as is reflected by the above characteristics around the dimer cleft.

4.1.3. The interface might not be conserved in PDI-related proteins

In Wind, the thioredoxin fold presents a suitable module for homodimerization. Since the structures of PDI-Dβ proteins are similar (at least for the TRX-like domain), may we derive the dimer interface of other PDI-Dβ proteins from that of Wind? Firstly, we will investigate which residues contribute most to the dimer interface in Wind. If the buried area represents the interaction intensity, the most important residues involved in the interface will be V28,

L33, R41, F42 and H70 (Table 3-2). Checking the residues in the corresponding positions of other three PDI-Dβ proteins, most of them are not conserved (Fig. 3-10). Thus, it is suspicious to conclude that other PDI-Dβ proteins form an interface in the same position. In fact, the interface of the rat ERp29 dimer has been studied by NMR (Liepinsh *et al.*, 2001). Compared to the interface of Wind, it appears to be located on the opposite side of the TRX-like domain and is mainly composed by residues around β5, $loop_{β5-α5}$ and α5.

Similarly, these important interface residues are not observed in the a and b domains in PDI, either (Fig. 3-15). We can not discuss the interface in the a' and b' domains because we lack a proper structure based alignment. So, We are not able to infer the interface of PDI from the Wind interface.

4.2. The CTGC is neither redox-active nor required for Pipe location

Wind contains one CTGC motif, missing in its high-level eukaroytic homologues. The location of this motif is far away from the classical redox-active CXXC positions in other TRX-like proteins. However, other locations also allow for redox-active CXXC sites, as was demonstrated for CXXC containing proteins that lack a thioredoxin fold (Langenback & Sottile, 1999; O'Neill *et al.*, 2000). The question arises whether the CTGC motif in Wind has a redox activity. The biochemical experiments address this question (Ma *et al.*, 2003).

The N-terminal His_6-tagged Wind is not able to reduce the disulfide bonds in insulin *in vitro*. From the structure, we have learned that the N-terminal His_6-tag should be located within or near the dimer cleft, although it is not directly seen in the electron density map. Thus the CTGC motif, which is at the bottom of the dimer cleft, may be blocked by the His_6-tag. However, the C-terminal his_6-tag Wind, in which the CTGC motif should be well exposed, shows the same result, indicating that Wind lacks a redox-active site.

Furthermore, Wind could not complement the PDI-deficient yeast, where the complementation is due solely to a general redox function without specific substrate binding interactions, indicating that Wind lacks a general catalytic redox/isomerase activity *in vivo*.

Mutation of CTGC to STGS does not result in defective localization of Pipe to the Golgi, indicating that the CTGC motif is not required to direct pipe to the Golgi. Wind may perform its function by a redox/isomerase independent chaperone activity. Since no CXXC motif is

present in Wind homologues, PDI-Dβ proteins may only share the similarity with PDI in their chaperone function.

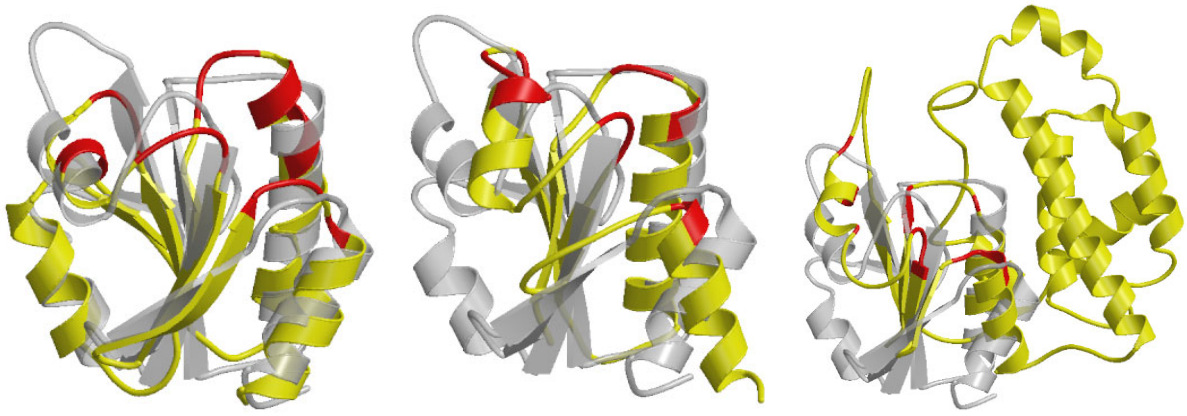
4.3. A proposed substrate binding site on the b-domain

TRX-like proteins such as TRX, GRX, GST and DsbA have a putative binding site around the CXXC and/or cis-Pro loop (Martin, 1995). Some protein-substrate-complex structures are available for TRX (Qin *et al.*, 1995; Qin *et al.*, 1996), GRX (Berardi and Bushweller, 1999) and GST (Xiao *et al.*, 1996). The features of the binding sites in these proteins are summarized in table 4-1.

the complex structures	residues in binding site	features
TRX-peptide1 (PDB 1CQH)	31, 32, 34, 59, 60, 66, 67, 71-75, 90-92	polar atoms in the interface: 32% non-polar atoms in the interface: 68% disulfide bonds: 1 hydrogen bonds: 1
TRX-peptide2 (PDB 1MDI)	29-32, 34, 35, 37, 38, 58-61, 63, 66, 67, 71-75, 90-92, 94	polar atoms in the interface: 32% non-polar atoms in the interface: 68% disulfide bonds: 1 hydrogen bonds: 2
GRX-peptide (PDB 1QFN)	7, 10, 11, 13, 16, 20, 38, 39, 42-46, 57-60, 72, 73, 76, 80	polar atoms in the interface: 43% non-polar atoms in the interface: 57% disulfide bonds: 1 hydrogen bonds: 5
GST-glutathione (PDB 6GST)	6, 7, 12, 42, 45, 49, 58-61, 71-73, 104	polar atoms in the interface: 69% non-polar atoms in the interface: 31% disulfide bonds: 0 hydrogen bonds: 10

Table 4-1 Substrate binding sites in TRX, GRX and GST revealed by their protein-substrate-complex structures.

The residues involved in the substrate binding are mainly located, named as the corresponding parts in the Wind structure, in $loop_{\beta2\text{-}\alpha2}$, $loop_{\beta3\text{-}\alpha3}$, $loop_{\alpha3\text{-}\beta4}$, $loop_{\beta5\text{-}\alpha4}$ and small parts in their neighboring secondary structure elements (Fig 4-1).



(a) (b) (c)

Figure 4-1 The substrate binding regions of TRX-like proteins revealed by their complex structures. (a) TRX-peptide2 (PDB 1MDI). (b). GRX-peptide (PDB 1QFN) (3) GST-glutathione (PDB 6GST). The residues involved in substrate binding are colored in red. The structure of Wind is superposed on and colored in shallow gray.

GRX and TRX form an intermolecular disulfide bond with the substrate. But the hydrophobic interactions and van der Waals interactions also contribute decisively to the binding. The redox-active DsbA is postulated to share the same binding site specified by a hydrophobic/uncharged patch (Guddat *et al.*, 1997). GST lacks the CXXC motif and there is no disulfide bond involved in binding. Instead, it binds the substrate mainly via extensive hydrogen bonds. A tyrosine at the end of β1 plays an important role for catalysis (Liu *et al.*, 1992). Thus, the CXXC motif is not required for binding. And the binding does not show a preference for a specific interaction type. All the above proteins are enzymes and have their catalytic center in the substrate binding region. We know, the binding region and catalytic region are usually different for enzymes. So the catalytic center may not be necessary for substrate binding. This region may also provide a substrate binding site for those TRX-like proteins that lack catalytic activities, but act as chaperons, such as Wind.

In Wind, such a region exists in the vicinity of the dimer cleft and is mainly composed of $loop_{\beta2\text{-}\alpha2}$, $loop_{\beta3\text{-}\alpha3}$, $loop_{\alpha3\text{-}\beta4}$ and $loop_{\beta5\text{-}\alpha4}$, with the *cis*-P106 in the center (Fig. 4-2). Wind lacks the CXXC motif there, which is not essential for binding as discussed.

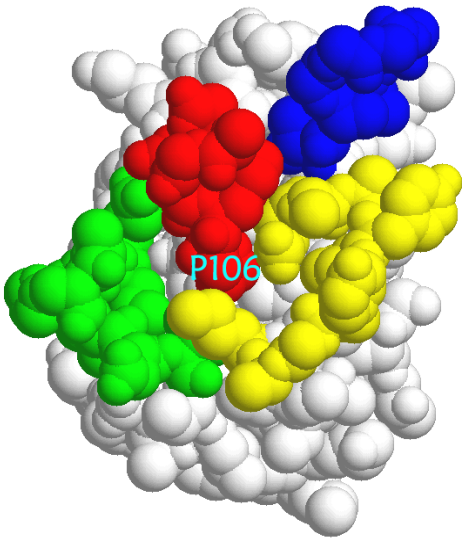


Figure 4-2 A proposed binding site of Wind. Yellow: $loop_{\beta2-\alpha2}$; Blue: $loop_{\beta3-\alpha3}$; Red: $loop_{\alpha3-\beta4}$, Green: $loop_{\beta5-\alpha4}$. The cis-P106 is marked.

.

Mutation studies (Ma *et al.*, 2003) have proved the importance of this region for function of Wind. Y55, a residue in $loop_{\beta2-\alpha2}$, is exposed to the surface at the verge of the dimer cleft, and its mutation to lysine should not affect the overall structure of the protein (Fig. 3-5, Fig. 3-11a1). However, Pipe failed to exit the ER when co-expressed with Wind Y55K mutant, although this mutant is highly expressed and retained in the ER. So, this region seems to provide a substrate binding site in Wind. This region in Wind contains a mixture of hydrophobic and hydrophilic patches (Fig. 3-11b1, c1), so we may expect the binding to be a combination of hydrophobic interactions, hydrogen bonds interactions and van der Waals interactions.

Comparing Wind with the above example proteins, the binding areas show a diverse morphology (Fig. 4-1), which may be responsible for substrate specificity.

In PDI, the b' domain, which lacks the CXXC motif and is thus redox inactive, is essential and sufficient for small peptide binding, but not sufficient for big peptide binding (Klappa *et al.*, 1998). This is consistent with the chaperon activity being independent of the redox/isomerization activity of PDI (Puig *et al.*, 1994). As discussed, the binding region in the TRX-like domain is rather conserved, we infer that the PDI b' domain adapts a similar region for peptide binding.

4.4. Both the b-domain and the D-domain are required for function

All of the above events take place in the TRX-like b-domains of Wind. Simply from the destroyed function of Y55K mutant, we know that the b-domain is required. However, Wind consists of two distinct domains: the TRX-like b-domain and the helical D-domain. What is the function of the D-domain? Is the b-domain alone sufficient to relocate Pipe to the Golgi?

Pipe failed to be located into the Golgi in the cells co-expressing the Wind b-domain alone (Wind-N) as well as in cells co-expressing an identical construct fused to a KEEL retrieval sequence (Wind-N-KEEL). The Wind-N construct is poorly retained in the ER due to the lack of a retrieval sequence, while the Wind-N-KEEL construct is efficiently retained. This indicates that the b-domain itself is insufficient to locate Pipe into the Golgi, even if it has an ER retention signal. Thus, not only the KEEL retrieval signal but also other parts in the D-domain are required for the function of Wind.

Furthermore, when the D-domain of Wind is substituted with that of mouse ERp28, the fusion protein retains the capacity to locate Pipe into the Golgi. Even the efficiency of relocation does not differ significantly from that promoted by wild-type Wind. This indicates that, although the D-domain is required, it does not provide the specificity. But the full-length ERp28 fails to promote location of Pipe to the Golgi, indicating that the b-domain of Wind can not be replaced with that of ERp28. Thus, the b-domain not only is required, but also might provide the specificity. The postulated binding region in the b-domain may define this specificity. This specificity is unlikely provided by the conserved residues in this site. Some residues, involved in the postulated binding region but not conserved in PDI-Dβ proteins, such as I51, A52 and F105 may be good candidates for the specificity. This problem could be well addressed if the mutations of these residues are available.

4.5. The D-domain

Since the D-domain of Wind and that of mouse ERp28 are exchangeable, it indicates that, they may have similar structures. But we have learned (see 3.10.) that the conformation of the D-domain of ERp29 is different from that of Wind (since mouse ERp28 is more than 95% homologous to rat ERp29 in sequence, we guess they should have similar 3-D structures). If Wind adapts the same conformation as ERp29, some hydrophobic residues would be exposed

on the surface, which is unlikely to occur, in general. As shown, the D-domain of Wind has an inherent flexibility revealed by the mainchain B-factors. We speculate that the conformation changes in Wind D-domain could occur so that more hydrophobic residues could be exposed to bind the substrate. Indeed, a structural rearrangement in similar helix-bundle proteins has been observed. For instance, apolipophorin III (apoLp-III) is organized as an antiparallel five-helix bundle in the lipid free state with the hydrophobic faces of the helices oriented towards the protein interior, whereas the hydrophilic faces directed to the aqueous environment. But when it interacts with lipoprotein surfaces, apoLp-III displays a preferred opening of the helix bundle so that the hydrophobic faces of the amphipathic helices are exposed, thereby facilitating binding to the lipid surface (Narayanaswami *et al.*, 1999). Similarly, the structure rearrangement of the D-domain of Wind may also occur induced by substrate binding. In the same time, the hydrophobic residues could be re-buried, which would provide a mechanism to release the cargo.

Another obvious feature of the D-domain is its movement owing to the linker flexibility. We may compare the relationship of the D-domains and b-domains in the dimer to that of the arms and the body. The D-domains may swing freely on each side of the b-domains, thereby improving the efficiency to catch the substrate. In the crystal, this movement is induced by the packing contacts. In solution, the movement may be induced by substrate binding: once the substrate is present, it can induce the arms (D-domains) to embrace and draw it to the body (b-domains). Similar domain movement is also observed in DsbA (Guddat *et al.*, 1998) and DsbC (Mccarthy *et al.*, 2000) structures, which may be important for their functions.

As discussed, the D-domain is required for the function of Wind not merely due to its ER retention signal. We infer that the D-domain is involved in substrate binding as well as the b-domain. This mimics the case in PDI: in PDI, the b' domain is essential but not sufficient for efficient binding of larger peptides, thus contributions from additional domains are required (Klappa *et al.*, 1998). Then, where is the possible binding site in the D-domain?

Based on the fact that the D-domain of Wind can be efficiently replaced with that of ERp28, we deduce the substrate-binding site in the D-domain may be conserved in the PDI-Dβ proteins. In the D-domain, the conserved residues in PDI-Dβ proteins are mainly located in the last two helices, α8 and α9. Some of them are exposed and form a considerable

"conserved cluster" on the protein surface (See 3.5.), which may be a good candidate for the binding area in the D-domain. We should do some mutagenesis for residues in this region to determine whether this region is required for binding or not.

4.6. The flexible linker region contains a free cysteine

The linker region is also very interesting. It is flexible, leading to an observable domain displacement in the crystal structure. The flexibility is not equally present in the whole region, but mainly occurs around three residues, Y143, I144 and G145, which act as a hinge. As we have discussed, the domain movement could make biological sense. We are interested in whether the flexibility of the linker region is essential for the function of Wind or not. From the structure, residue mutations in the linker region should not destroy the overall structure. So site directed mutagenesis may provide a good approach. I144 and G145 have relatively short side chains, thus less hindering the domain movement. We may mutate them to residues with long side chain, such as tryptophan or arginine. We may expect the domain movement should be less free in the mutant than that in wild-type Wind. Then the function can be easily detected by a Pipe location experiment. It is also a good way to simply knock out these three mobile residues from the linker region. Thus the domain movement should be well blocked. The function of Wind can be checked as above. Then, we may draw a conclusion whether the flexibility in the linker region is important or not.

Except for its flexibility, which provides the structural basis for domain movement, the linker region contains a free cysteine (C149). This cysteine is solvent accessible and may have a high reaction activity. We have two indications for this. Firstly, in Hg derivative, the Hg^{2+} ion is bound to this cysteine. Secondly, when the crystallization was performed as described but with GSSG in protein solution, not crystals but precipitation were observed. While in the same experiment with GSH, small crystals were still available. Although crystallization is affected by many factors, we may still speculate that: the free thiol of this cysteine is oxidized by GSSG and form a disulfide bond with it, thus the domain movement is hampered so that the proper crystal packing could not take place. If this cysteine is really highly active, we may infer that, it could be involved in the substrate binding by forming an intermolecular disulfide bond with the substrate. Similarly, mutating this cysteine to other residues provides a good

way to check its function in Wind.

4.7. Unusual solubility pattern of Wind

In the purification and crystallization, Wind showed a salt-in solubility. In order to check it, a solubility study was done. Some Wind crystals grew in [0.1M MES pH6.1, 0.1M CsCl, 16% PEG300] drops. Adding [0.1M MES pH6.1, 0.1M CsCl, 20% PEG300, 0.5/0.8/1M NaCl] to these drops results in re-dissolving rather than crashing the crystals. However, the precipitates (may be denatured protein) in the drop could not be dissolved again. $CaCl_2$ showed the same effect but with higher efficiency. Even 0.2M $CaCl_2$ could lead to a total dissolving of the crystals. How the salt greatly increases the solubility of Wind even in the presence of high concentrations of precipitant (20% PEG300) is not clear.

5. Conclusions

Wind is a key factor to trigger dorsal-ventral patterning in the *Drosophila* embryo development. Genetic experiments indicate that Wind is required for targeting the dorsal-ventral patterning protein, Pipe, from the ER to the Golgi. So far the mechanism is unknown at the molecular level. Here, we have determined the three-dimensional structure of Wind with X-ray crystallography to 1.9 Å resolution. This crystal structure and the site directed mutagenesis studies based on it greatly promote our insight into the relationship of the structure and function of Wind.

Wind is a PDI-related ER resident protein consisting of two distinct domains: an N-terminal thioredoxin-like b-domain and a C-terminal helical D-domain, connected by a short flexible linker. Due to the flexibility of the linker, the orientation of the two domains could vary, which might be important for substrate capture.

Wind exists as a homodimer both in the crystal and in solution. The dimer is formed by a head-to-tail arrangement of the b-domains without the participation of the D-domains. The dimerization yields a deep dimer cleft, which is large enough to hold a small peptide. This dimer cleft is negatively charged, with a considerable hydrophobic band spanning the top. And the surface residues in the bank areas of this cleft are rather conserved in PDI-Dβ family.

A possible substrate binding site in the b-domain is proposed, where *cis*-P106 is located in the center. This site is located in the vicinity of the negatively charged dimer cleft and is consistent with the putative binding site indicated in other thioredoxin-like proteins. The mutant Y55K completely abrogates Pipe transport (Ma *et al.*, 2003), indicating that, this site is required for the function of Wind. We infer that the dimer cleft and this proposed binding site are both involved in substrate binding by a mechanism combining hydrogen-bond interactions, electrostatic interactions, hydrophobic interactions and van der Waals interactions.

A near N-terminal CTGC motif is located at the bottom of the dimer cleft. Biochemical experiments (Ma *et al.*, 2003) indicate that it lacks the general redox activity and is not required for the location of Pipe. Wind might perform its function by a redox/isomerase independent chaperone activity.

The b-domain is essential but not sufficient for function and the D-domain is also

required. Except for providing a KEEL retrieval signal, the D-domain might also provide an additional substrate-binding site.

The linker region may be important as well. It provides not only the structural basis for the domain movement but also a free cysteine, which has the potential to form an intermolecular disulfide bond with the substrate.

In this work, we have analyzed the structure of Wind in detail and postulated some hypotheses for its function and reaction mechanism. In the meantime, we have also proposed some promising candidates for further site directed mutagenesis studies, which would provide more information on the function of Wind. Nevertheless, the mechanism of how Wind acts in the cell is not really clear so far. There are still many open questions. Is the direct substrate of Wind really Pipe? Is the substrate a peptide or protein? Does it function alone or with other proteins together in a complex state? In order to answer these questions, the structure of a Wind-substrate-complex is necessary. This will be the object of further study.

References

Appenzeller, C., Andersson, H., Kappeler, F. and Hauri, H.P. The lectin ERGIC-53 is cargo transport receptor for glycoproteins. Nature cell Biol. 1, 330-334 (1999).

Bashford, D. An object-oriented programming suite for electrostatic effects in biological molecules. in Scientific Computing in Object-Oriented Parallel Environment, Vol. 1343 (eds. Ishikawa, Y., Oldehoft, R.R., Reynders, J.V.W. & Tholburn, M.). 233-240 (Springer, Berlin, 1997).

Berardi, M. J., Bushweller, J. H.: Binding Specificity and Mechanistic Insight Into Glutaredoxin-Catalized Protein Disulfide Reduction J. Mol. Biol. 292, 151 (1999).

Bergeron, J.J., Brenner, M.B., Thomas, D.Y. and Williams, D.B. Calnexin: a membrane-bound chaperone of the endoplasmic reticulum. Trends Biochem Sci. 19(3),124-128 (1994).

Brunger, A.T. Assessment of phase accuracy by cross validation: the free *R* value. Methods and applications. Acta Cryst. D49, 24-36 (1993).

Brunger, A.T. et al. Crystallography and NMR system(CNS): A new software system for macromolecular structrue determination. Acta Cryst. D54, 905-921 (1998).

Bu,G. The roles of receptor-associated protein(RAP) as a molecular chaperone for members of the LDL receptor family. Int. Rev. Cytol. 209, 79-116 (2001).

Carvin, D., Islam S.A., Sternberg M.J.E and Blundell T.L. The preparation of heavy-atom derivatives of protein crystals for use in multiple isomophous replacement and anomalous scattering. *in*: InternationalTables for Crystallography (Volume F: Crystallography of Biological macromolecules. Editors: M.G. Rossmann and E. Arnold,2001. Cordrecht: Kluwer Academic Publishers, The Netherlands). 247-255 (2001).

Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. Acta Cryst. D 50, 760-763 (1994).

Dissing, M., Giordano, H. and Delotto, R.. Autoproteolysis and feedback in a protease cascade directing Drosophila dorsal-ventral cell fate. EMBO. J. 20, 2387-2393 (2001)

Ellgaard,L. and Helenius,A..quality control: towards an understanding at the molecular level. Current Opinion in Cell Biology, 13, 431–437 (2001).

Ellgaard,L. and Helenius,A. Quality control in the endoplasmic reticulum. Nat Rev Mol Cell Biol. 4(3), 181-91(2003)

Ellgaard,L., Molinari,M. and Helenius,A. Setting the standards: quality control in the secretory pathway. Science. 286, 1882-1888 (1999).

Ellis,J. Proteins as molecular chaperons. Nature 328, 378-379 (1987).

Ferrari, D.M., Nguyen van, P., Kratzin, H.D. and Söling H.-D. ERp28, a human endoplasmic-reticulum-lumenal protein, is a member of the protein disulfide isomerase family but lacks a CXXC thioredoxin-box motif. Eur J Biochem. 255(3), 570-579(1998).

Ferrari, D.M. and Söling H.-D. The protein disulphide-isomerase family: unravelling a string of folds. Biochem. J. 339, 1-10 (1999).

Fischer, G. Peptidyl-prolyl cis/trans isomerases and their effectors. Angew. Chem. Int. Ed. Engl. 33, 1415-1436 (1994).

Freedman, R.B., Hirst, T.R. and Tuite, M.F. Protein disulfide isomerase: Building bridges in protein folding. Trends Biochem. Sci. 19, 331-336 (1994).

Fujinaga, M. and Read, R.J. Experiences with a new translation-function program J. Appl. Cryst. 20, 517-521 (1987).

Goodford, P.J. A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. J. Med. Chem., 28, 849-857 (1985).

Guddat, L.W., Bardwell, J.C.A and Martin, J.L. Crystal structures of reduced and oxidized DsbA: investigation of domain motion and thiolate stabilization. Structure. 6(6), 757-67 (1998).

Guddat, L.W., Bardwell, J.C.A, Zander, T. and Martin, J.L. The uncharged surface featrues surrounding the active site of Escherichial coli DsbA are conserved and are implicated in peptide binding. Protein Science. 6, 1148-1156 (1997).

Haebel, P. W., Goldstone, D., Katzen, F., Beckwith, J., Metcalf, P.: The Disulfide Bond Isomerase Dsbc is Activated by an Immunoglobulin-Fold Thiol Oxidoreductase: Crystal Structure of the Dsbc-Dsbdalpha Complex. EMBO J. 21, 4774-4784 (2002)

Hartl, F.U. Molecular chaperons in cellular protein folding. Nature. 381, 571-579 (1996).

Hashimoto C., Hudson K. L., Anderson K. V..The Toll gene of Drosophila, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. Cell. 52(2), 269-279 (1988).

Hayward S., Lee R. A.. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. J Mol Graph Model, 21(3), 181-183 (2002).

Holm, L. and Sander, C. and Protein Structure Comparison by alignment of distance matrices. J. Mol. Biol. 233, 123-138 (1993).

Hong, C.C. and Hashimoto,C. An unusual mosaic protein with a protease domain, encoded by the the nudel gene, is involved in defining embryonic dorsoventral polarity in Drosophila. Cell. 82, 785-794 (1995).

Hwang, C., Sinskey, A.J. and lodish, H.F. Oxidized redox state of glutathione in the endoplasmic reticulum. Science. 257, 1496-1502 (1992)

Ikawa, M., Wada, I., Kominami, K., Watanabe, D., Toshimori, K., Nishimune, Y. and Okabe, M.. The putative chaperone calmegin is required for sperm feritlity. Nature. 387, 607-611 (1997)

Jones, S. and Thornton,J.M. Principles of protein-protein interactions. Proc. Natl.Acad. Sci.(USA) 93,13-20 (1996).

Kabsch,W. and Sander,C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577-2637 (1983).

Klappa, P., Ruddock, L. W., Darby, N. J. and Freedman, R. B. The b' domain provides the principal peptide-binding site of protein disulfide isomerase but all domains contribute to binding of misfolded proteins. EMBO J. 17, 927-35 (1998).

Kleywegt, G.J. and Jones, T. A. xdlMAPMAN and xdlDATAMAN – programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection datasets. Acta Cryst. D52, 826-828 (1996)

Konsolaki, M. and Schüpbach, T. *Windbeutel*, a gene required for dorsoventral patterning in Drosophila, encodes a protein that has homologies to vertebrate proteins of the endoplasmic reticulum. Genes & Development. 12, 120-131 (1998).

Kraulis, P.J. MolScript -- a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24, 946-950 (1991).

Langenback, K.J. and Sottile, J. Identification of protein-disulfide isomerase activity in fibronectin. J. Biol.Chem. 274, 7032-7038 (1999).

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283-291 (1993).

Levitt, D. G.. A New Software Routine that Automates the Fitting of Protein X-Ray Crystallographic Electron Density Maps. Acta Cryst. D57, 1013-1019 (2001).

Liepinsh, E., Baryshev, M., Sharipo, A.,Ingelman-Sundberg, M., Otting, G. and Mkrtchian,S. Thioredoxin fold as homodimerization module in the putative chaperone ERp29: NMR structures of the domains and experimental model of the 51 kDa dimer. Structure. 9, 457–471 (2001).

Liu, S., Zhang, P.,Ji,X.,Johnson,W.W.,Gilliland,G.L. and Armstrong, R.N.. Contribution of Tyrosine 6 to the Catalytic Mechanism of isoenzyme 3-3 of glutathione S-transferase. J. Bio. Chem. 267, 4296-4299 (1992).

Ma, Q., Guo, C., Barnewitz, K., Sheldrick, G. M., Soeling, H.-D., Uson, I., Ferrari, D. M. Crystal structure and functional analysis of *Drosophila* Wind – a protein disulfide isomerase-related protein. J. Biol. Chem. 278, 44600-44607 (2003).

Martin, J.L. Thioredoxin - a fold for all reasons. Structure. 3, 245-250 (1995).

Martin, J.L., Bardwell, J.C.A. and Kuriyan, J. Crystal structure of the DsbA protein required for disulphide bond formation in vivo. Nature. 365, 464-468 (1993).

Matthews, B.W. Solvent content of Protein Crystals. J. Mol. Biol. 33, 491-497(1968).

McCarthy AA, Haebel PW, Torronen A, Rybin V, Baker EN, Metcalf P.Crystal structure of the protein disulfide bond isomerase, DsbC, from Escherichia coli.Nat Struct Biol. 7(3), 196-199 (2000).

McRee, D. E. XtalView/Xfit -- A versatile program for manipulating atomic coordinates and electron density. J. Struct. Biol. 125, 156-165 (1999).

Meldolesi, J., Krause, K.H. and Michalak, M. Calreticulin: How may functions in how many cellular compartments? (The second International Calreticulin Workshop; Como, Italy; April 29-30, 1996. Cell Calcium. 20, 83-86 (1996).

Merrit, E.A. & Murphy, M.E.P. Raster3D version 2.0 -- a program for photorealistic molecular graphics. Acta Crystallogr. D50, 869-873 (1994).

Miller,R., DeTitta, G.T., Jones, R., Langs, DA., Weeks, C.M., Hauptman, H.A.On the application of the minimal principle to solve unknown structures. Science. 259(5100), 1430-1433 (1993).

Mkrtchian, S., Baryshev, M., Matvijenko, O., Sharipo, A., Sandalova, T., Schneider, G., Ingelman-Sundberg, M. and Mkrtchiana, S. Oligo-merization properties of ERp29, an endoplasmic reticulum stress protein. FEBS Lett. 431, 322–326 (1998).

Morisato,D. and Anderson, K.V. The spatzle gene encodes a component of the extracellular signaling pathway establishing the dorsal-ventral pattern of the Drosophila embryo. Cell. 76(4), 677-688 (1994).

Morisato,D. and Anderson, K.V. Signaling pathways that establish the dorsal-ventral pattern of the Drosophila embryo. Annu Rev Genet. 29, 371-399 (1995).

Munro, s. and Pelham, H.R. An Hsp70-like protein in the ER:identity with the 78kD glucose-regulated protein and immunoglobulin heavy chain binding protein. Cell. 46, 291-300 (1986).

Munro, s. and Pelham, H.R. A C-terminal signal provents secretion of luminal ER proteins. Cell. 48, 899-907(1987)

Murshudov, G.N., A.A., Vagin & E.J., Dodson. Refinement of Macromolecular Structures by the Maximum-likelihood Method. Acta Cryst. D53, 240-255 (1997).

Narayanaswami, V., Wang, J., Schieve, D., Kay,C. M. and Ryan, R. O. A molecular trigger of lipid binding-induced opening of a helix bundle exchangeable apolipoprotein. PNAS. 96, 4366-4371(1999).

Neuman-Silberberg, F.S. and Schupbach, T. The Drosophila dorsoventral patterning gene gurken produces a dorsally localized RNA and encodes a TGF alpha-like protein. Cell. 75(1), 165-174 (1993).

O'Neill, S., Robinson, A., Deering, A., Ryan, M., Fitzgerald, D. J. and Moran N.The platelet Integrin alpha IIbbeta 3 has an endogenous thiol isomerase activity. J. Biol. Chem. 275, 36984-36990 (2000).

Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. 276, 307-326 (1997).

Pelham,H.R.B. The retention signal for soluble proteins of the endoplasmic reticulum. Trends. Biochem. Sci. 15, 483-486 (1990).

Perrakis, A., Morris, R. & Lamzin, V.S. Automated protein model building combined with iterative structure refinement. Nature Struct. Biol. 6, 458-463 (1999).

Puig, A., Lyles, M.M., Noiva, R. and Gilbert, H.F. The role of the thiol/disulfide centers and peptide binding site in the chaperone and anti-chaperone activities of protein disulfide isomerase. J Biol Chem. 269(29), 19128-19135 (1994).

Qin, J., Clore, G.M., Kennedy, W.P., Hutch, J.R. and Gronenborn, A.M.. Solution structure of human thioredoxin in a mixed disulphide intermediate complex with its target peptide from the transcription factor NF-kB. Structure 3, 289-297 (1995).

Read, R.J. Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst. A42, 140-149 (1986).

Ren, B., Tibbelin, G., Pascale, D.D., Rossi, M., Bartolucci, S. and Ladenstein, R.. A protein disulfide oxidoreductase from the archaeon Pyrococcus furiosus contains two thioredoxin fold units. Nature Struct. Biol. 5, 602-611 (1998).

Rodgers, D.W. Cryocrystallography techniques and devices. *in*: InternationalTables for Crystallography (Volume F:Crystallography of Biological macromolecules. Editors: M.G. Rossmann and E. Arnold, 2001. Cordrecht: Kluwer Academic Publishers, The Netherlands). 202-208 (2001).

Rossmann, M.G. & Blow, D.M. The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst. 15, 24-31(1962)

Sanner, M., Olson, A. J. and Spehner,J. C. Fast and robust computation of molecular surface. Proc. 11th ACM Symp. Comp. Geom, C6-C7 (1995).

Sayle, R. and Bissell, A. RasMol: A Program for Fast Realistic Rendering of Molecular Structures with Shadows, in Proceedings of the 10th Eurographics UK '92 Conference, University of Edinburgh, Scotland (1992).

Schultz, L.W., Chivers, P.T. and Raines, R.T. The CXXC motif: crystal structure of an active-site variant of Escherichia coli thioredoxin. ACTA. Cryst. D55,1533-1538 (1999).

Schneider,T.R. and Sheldrick, G.M. Substructure solution with SHELXD. Acta Cryst. D58, 1772-1779 (2002).

Sen,J., Goltz, J.S., Stevens, L. and Stein, D. Spatially restricted expression of pipe in the Drosophila egg chamber defines embryonic dorsal-ventral polarity. Cell. 95(4), 471-481 (1998).

Sen, J., Goltz, J.S., Konsolaki, M., Schupbach, T. and Stein D. *Windbeutel* is required for function and correct subcellular localization of the Drosophila patterning protein Pipe. Development. 127(24), 5541-5550 (2000).

Sergeev,P., Streit, A., Heller, A. and Steinmann-Zwicky, M. The Drosophila dorsoventral determinant PIPE contains ten copies of a variable domain homologous to mammalian heparan sulfate 2-sulfotransferase. Dev Dyn. 220(2), 122-132 (2001).

Sheldrick, G. M. XPREP, program for reciprocal space exploration. Version 6.12. Bruker Nonius Inc., Madison, Wisconsin, USA (2001).

Sheldrick, G. M. Macromolecular phasing with SHELXE. Z. Kristallogr. 217, 644-650 (2002).

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, M. & Usón, I. Ab initio phasing. International Tables for Crystallography, Vol. F, edited by E. Arnold & M. G. Rossmann, 333-351 (2001).

Stamnes, M.A., Shieh, B. H., Chuman, L., Harris, G. L. and Zuker, C. S. The cyclophilin homolog ninaA is a tissue-specific integral membrane protein required for the proper synthesis of a subset of Drosophila rhodopsins. Cell. 65, 219-227 (1991).

Thompson, J. D., Higgins, D.G., and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22, 4673-4680 (1994)

Vriend, G. WHAT IF: A molecular modeling and drug design program. J. Mol. Graph. 8, 52-56 (1990).

Wadsworth, S. C., Vincent III, W. S. and Bilodeau-Wentworth, D. A Drosophila genomic sequence with homogogy to the human epidermal growth factor receptor. Nature. 314, 178-180 (1985).

Wang, C. C. and Tsou, C. L. Protein disulfide isomerase is both an enzyme and a chaperone. FASEB J. 7, 1515-1517 (1993).

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G. and Weiner, P.. A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc. 106, 765-784 (1984).

Xiao, G., Liu, S., Ji, X., Johnson, W. W., Chen, J., Parsons, J. F., Stevens, W. J., Gilliland, G. L., Armstrong, R. N.: First-sphere and second-sphere electrostatic effects in the active site of a class mu gluthathione transferase. Biochemistry. 35, 4753-4765 (1996)

Zhang, L., Kelly, J., Schmeisser, G., Kobayashi, Y. M. and Jones, L. R. Complex formation between junctin, triadin, calsequestrin and the ryanodine receptor. The Journal of Biological Chemistry. 272, 23389-23397 (1997).

Zhang, K. Y. J., Cowtan, K. D. and Main, P. Phase improvement by iterative density modification. *in*: InternationalTables for Crystallography (Volume F: Crystallography of Biological macromolecules. Editors: M.G. Rossmann and E. Arnold, 2001. Cordrecht: Kluwer Academic Publishers, The Netherlands). 311-324 (2001).

Publications

Ma, Q., Guo, C., Barnewitz, K., Sheldrick, G. M., Soeling, H.-D., Uson, I., Ferrari, D. M. Crystal structure and functional analysis of *Drosophila* Wind – a protein disulfide isomerase-related protein. The Journal of Biological Chemistry. 278, 44600-44607 (2003).

Bai G., **Ma Q.**, Roesky H. W., Vidovic D., Herbst-Irmer R.. New synthetic route for organic polyoxometallic clusters: synthetic and structural investigations on the first dumb-bell shaped polyoxozirconium hydroxide with the [Zr-9(mu(5)-O)(2)(mu(3)-O)(4)(mu-O)(4)(mu-OH)(8)] core structure. Chemical Communications. 7, 898-899 (2003).

Debreczeni J. E., Bunkoczi G., **Ma Q.**, Blaser H., Sheldrick G. M.. In-house measurement of the sulfur anomalous signal and its use for phasing. ACTA Crystallographica Section D-Biological Crystallography. 59, 688-696 (2003).

Ding Y., **Ma Q.**, Roesky H. W., Uson I., Noltemeyer M., Schmidt H. G. Syntheses, structures and properties of [{HC(CMeNAr)(2)}Ge(E)X] (Ar=2,6-iPr(2)C(6)H(3); E = S, Se; X = F, Cl). Dalton Transactions. 6, 1094-1098 (2003).

Ding Y., **Ma Q.**, Roesky H. W., Herbst-Irmer R., Uson I., Noltemeyer M., Schmidt H. G.. Synthesis, structures, and reactivity of alkylgermanium(II) compounds containing a diketiminato ligand. Organometallics. 21, 5216-5220 (2002).

Ding Y., **Ma Q.,** Uson I., Roesky H. W., Noltemeyer M., Schmidt H. G.. Synthesis and structures of [{HC(CMeNAr)(2)}Ge(S)X] (Ar=2,6-iPr(2)C(6)H(3), X = F, Cl, Me): Structurally characterized examples with a formal double bond between group 14 and 16 elements bearing a halide. Journal of the American Chemical Society. 124, 8542-8543 (2002).

Curriculum Vitae

Name:		Qingjun Ma
Date of birth:		July 8th, 1975
Place of birth:		Shandong Province, P. R. China
Marital status:		Married
Nationality:		P. R. China
Education:		
	1982-1987	Primary school, Pingyi, China
	1987-1990	Middle school, Pingyi, China
	1990-1993	High school, Pingyi, China
Studies:		
	1993-1997	Nankai University, Tianjin, China
	1997	Bachelor of Science, in Microbiology
	1997-2000	Institute of Biophysics, Chinese Academy of Sciences
	2000	Master of Science, in Molecular Biology
Dissertation:		
	2000-2003	Department of Structural Chemistry, University of Göttingen