

Protein Crystallography

Group leader:	Victor Lamzin
Staff scientist:	Andrea Schmidt
Postdoctoral fellows:	Olga Kirillova, Gerrit Langer*, Tilo Strutz*
PhD student:	Petrus Zwart*
Technicians:	Venkataraman Parthasarathy*, Babu Pothineni*, Katja Schirwitz
Visitors:	Serge Cohen, Zbigniew Dauter, Francisco Fernandez Perez*, Christian Jelsch, Mattheos Kakaris, Joergen Koepke, Olga V. Koroleva, Richard J. Morris, Peter Østergaard, Tatiana V. Pegasova, Anastassis Perrakis, Denis V. Rebrikov, Wojciech Rypniewski, Jozef Sevcic, Elena V. Stepanova, Clemens Vonrhein

*Indicates part of the year only

Outline

A major component of the research activity of the group is the development of underlying methodology for crystal structure determination of biological macromolecules. These technology developments have now been given the central emphasis in the whole process of high-throughput structure determination and encompass all steps from crystal production, model building to high-accuracy structure interpretation.

An on-going development, the Automated Refinement Procedure (ARP/wARP suite) for refinement and building of protein structures in its present state may already be in a position to promote progress in automating the steps of deriving an essentially complete structural model from the X-ray data. The time required for building a protein structure can be shortened from several man-days or even man-months to a few CPU hours. New features such as automatic recognition of secondary structure elements, ligands and the application of non-crystallographic symmetry are currently under way.

The second research direction of the group is accurate structural enzymology: the investigation of macromolecular structures at atomic or ultra-high resolution. For many biological or biotechnological applications a good accuracy of the structural model is required. Protein models determined at extremely high resolution provide fine structural details and features relevant to the function which may remain hidden at lower resolution. A respectable amount of these structures is coming from measurements at the EMBL Hamburg beamlines. The rapidly growing number of protein structures at very high resolution, specifically developed methods as well as new combinations of protein crystallography with complementary techniques is astonishing and are about to establish an entirely new field in structural biology.

The improvement of diffraction properties of protein crystals forms the third research direction. Most often the availability of suitable crystals is the limiting step on the way to high quality protein structures. We investigate a possible use of specifically targeted engineering as a means to modify the surface properties and hence crystallisation tendency of biological macromolecules in order to make the production of high-quality crystals suitable for higher throughput.

Breaking good resolutions with ARP/wARP

(Richard J. Morris, EBI; Serge Cohen, Francisco J. Fernandez, Mattheos Kakaris, Anastassis Perrakis, NKI Amsterdam, NL; and Clemens Vonrhein, GlobalPhasing Cambridge, UK; and Petrus H. Zwart, Olga Kirillova, Victor S. Lamzin, EMBL-Hamburg)

Once initial phase information in crystallography experiment has been made available, it is desired to construct a chemically sensible model of the macromolecule, which represents the experimental electron density distribution with a set of labelled atoms and their corresponding coordinates. For any Structural Genomics project or other high-throughput structure determination initiative the traditionally time consuming and labour intensive step of model building has to be made fast, reliable and highly automated. ARP/wARP (Perrakis *et al.*, 1999; Lamzin *et al.*, 2001; Morris *et al.*, 2002) has successfully tackled this problem. The previous ARP/wARP version 5.1 required data to 2.0 Å (in some cases 2.3 Å was sufficient). Version 6.0 (released in July 2002) can now successfully build models building with diffraction data extending to about 2.5 Å, thereby increasing the applicability of the program.

An initial electron density map at 2.5 Å does not allow all atoms to be placed with confidence and accuracy. A decision-making process during the model building becomes necessary to "guess" the most likely interpretation based only on local bonding geometry of originally non-bonded atoms within valid bonded distance limits. In such situations, one has to rely on the experience of a crystallographer and interactive graphics software to find the most plausible solution or to attempt to formulate some heuristic that mimics this process. As described in Morris *et al.* (2002), a density-weighted match between found and expected protein Ca geometry is computed and the best set of highest scoring main chain fragments is sought. The search technique is a modified depth limited search algorithm (Russel and Norvig, 1995). An implementation of this method has been shown to cope better with inaccurate free atom positions and forms a core of the novel ARP/wARP tracing algorithm.

In the following the on-going developments are briefly outlined.

1. NCS may be used to deliver more complete models by extending polypeptide fragments generated during the main-chain tracing step. This could be particularly valuable in cases of similar conformation of the NCS related fragments but different quality of the electron density resulting from *e.g.* poor phases, model bias or disorder.
2. Identification of secondary structural elements in an electron density map based on prior knowledge of their motifs and stereo-chemistry should considerably enhance model building in general and, particularly, provide the extension to the lower (*e.g.* 3.0 Å) resolution of the X-ray data. We currently exploit a discriminant analysis pattern recognition technique for location of helical fragments. A helical structural motif is parameterised by a set of small overlapping fragments, which fulfil a number of stereochemical conditions including interatomic distances, valence and dihedral angles.
3. Using an approach that resembles the Conditional Dynamics proposed by Scheres and Gros (2001) and the formalisms used in the coordinate error estimation procedure, a good prediction of the possible chemical nature of a particular area in the unit cell can be obtained (Zwart and Lamzin, 2003). Although the procedure is being designed for the automatic recognition and building of bound ligands and small molecular fragments, its extension to an interpretation of parts of protein structure is trivial.

Distance distributions and electron-density characteristics of protein models and the influence of positional errors on the Debye effects

(Petrus H. Zwart, Victor S. Lamzin, EMBL-Hamburg)

Estimation of the map quality is essential for automated model building in macromolecular crystal structure analysis. Decisions on the placement of a structural element based on density criteria should ideally reflect both the prior information of the expected electron density for a particular structural element as well as a measure of how well the placement of such a fragment represents the observed density with a given error. Although a number of error and map-quality estimation procedures are available, it is worthwhile investigating whether other straightforward estimates can be obtained or used in conjunction with existing methods.

The real-space approach to this problem presents an analytical expression for the distribution of an interatomic distance resulting from a known error-free distance and a Gaussian perturbation of the atomic coordinates. This is used to estimate the coordinate error on the basis of known geometric features of protein models via the nearest neighbour or the radial distance distribution and describes the dependence of the map correlation on the positional error of the protein model, the resolution of the X-ray data and the overall atomic displacement parameter. This can be used in the decision

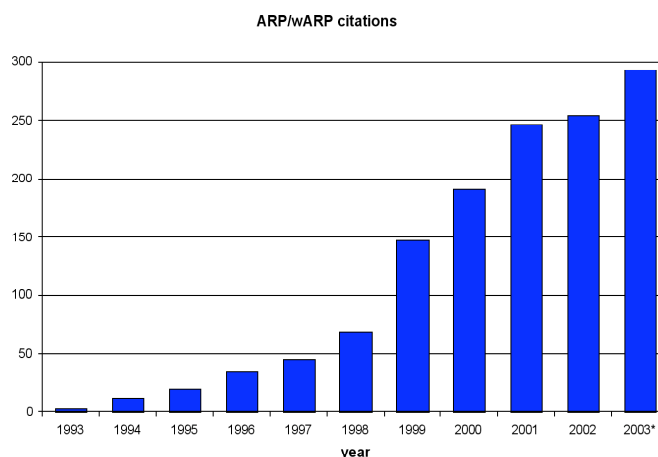


Figure 1. Citations on the use of ARP/wARP.

making process during automated model-building, where estimation of the map quality is essential.

Let a pair of atoms (x_i, x_j) separated by a target distance d_{jk}^{tar} undergo a random Gaussian perturbation of the positional parameters. Let the variance of the displacement in the x , y and z directions to be equal to σ_j^2 for atom j and σ_k^2 for atom k . Assuming that the errors of the positional parameters are independent, it can be shown (Arfken and Weber, 1995; Abramovicz and Stegun, 1974) that the observed interatomic distance after the perturbation, d_{jk}^{obs} , is distributed according to

$$f(d_{jk}^{obs}) = \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma_k^2)}} e^{-\frac{(d_{jk}^{obs} - d_{jk}^{tar})^2}{2(\sigma_j^2 + \sigma_k^2)}} \frac{d_{jk}^{obs}}{d_{jk}^{tar}} \left(1 - e^{-\frac{2d_{jk}^{tar} d_{jk}^{obs}}{\sigma_j^2 + \sigma_k^2}} \right)$$

This expression becomes identical to the Maxwell distribution (Weisstein, 1999) for $d_{jk}^{obs} = 0$. For this reason, we denote it as the non-central Maxwell distribution, as it describes the distribution of a vector length of a spherical three-dimensional Gaussian centered on a vector with given length d_{jk}^{tar} . The expected r.m.s.d. between an error-free and perturbed structure given a Gaussian error model with variances in each direction equal to σ_m^2 can be shown to be equal to the square root of the second raw moment of the Maxwell distribution: $E[\text{rmsd}]_{d_{jk}^{obs}=0} = 3^{1/2} \sigma_m$. The presented analytical distribution of an interatomic distance given the target distance and a Gaussian error model serves as an essential component in modelling of distance distributions in proteins.

The reciprocal space approach relates a Gaussian perturbation of the atomic positional parameters and the average squared structure factor amplitude. Using an error-dependent radial distance distribution of an atomic protein model, it can be shown that the Debye effects diminish exponentially as a function of increasing posi-

tional errors (Figure 2). These relations can be used to estimate the quality of an atomic model and the corresponding phases. The limiting case of equal atoms with an infinitely large coordinate error results in the classical Wilson model. A key point here is that the presented error-estimation method is rather sensitive to the quality of the low-resolution part of the data set used. This is ascribed to the fact that the Debye effects at high resolution diminish faster than those at lower resolution.

A main result was the derivation of the atomic shape parameter ω and its relation to the nominal resolution of X-ray data and its main characteristics, the overall temperature factor intensity falloff (the so called Wilson plot B factor). The atomic shape factor has been estimated using the formulations outlined above and 69 selected structures from PDB for which experimental X-ray data were available. Least-squares fitting of the parameters resulted in the following expression:

$$\omega^2 = (0.078 d_{\min} + 0.043 B_{\text{wil}}^{1/2} + 0.322)^2$$

The first two coefficients are the weights to the contributions of the nominal resolution and the average atomic displacement factor on the blurring of the electron density. The third coefficient can be seen as modelling the average width of an atom at rest at infinite resolution. Using the empirically derived quantity ω as a classifier for an X-ray data set to choose appropriate density templates during chain tracing and the construction of tailor made decision boundaries as a function of map quality and data-set characteristics is currently being implemented in the latest version of ARP/wARP and awaits thorough testing to validate the results.

Trypsin revisited: crystallography at (sub) atomic resolution and quantum chemistry revealing details of catalysis

(Christian Jelsch, Nancy University, FR; and Wojtek Rypniewski, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland; Andrea Schmidt, Victor S. Lamzin, EMBL-Hamburg)

Atomic resolution crystallography ($< 1.2 \text{ \AA}$) and ultra-high resolution crystallography ($< 0.8 \text{ \AA}$) are powerful tools for the investigation of proteins. Fine electronic detail can be visualised and yield valuable information on protein function. The assessment and description of intra- and intermolecular contacts reaches a degree of accuracy which was considered unattainable in the past. In addition, information extracted from the anisotropic temperature factors gives insight into the mobility and the presence of directional movement in the protein.

Fusarium oxysporum trypsin is a serine protease cleaving peptides at the C terminus of arginines or lysines. It shows considerable autoproteolysis even in the crystallisation solution (Rypniewski *et al.*, 2001). A tri-peptide fragment was always present in the active site of the atomic resolution and ultra-high resolution structures and could only be substituted by covalently bind-

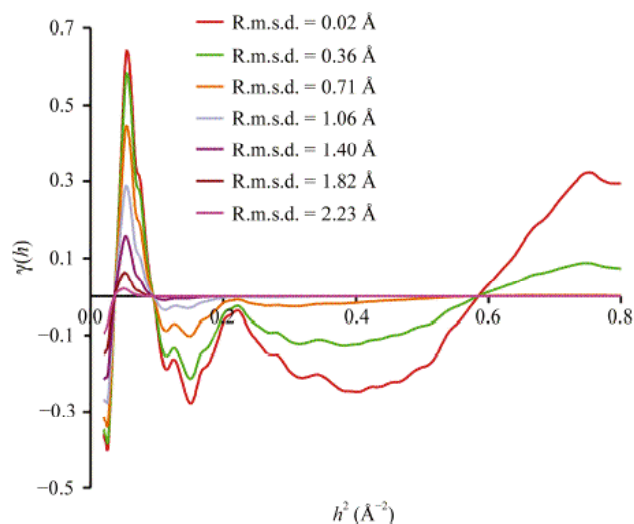


Figure 2. The effect of a coordinate error on the Debye effects calculated for lysozyme.

ing inhibitors. Structures of the native enzyme at pH 4 and pH 5, two inhibitors (PMSF and DFP) and borate as an additive were determined. Data to atomic and ultra-high resolution were collected on beamlines X11 and X13 at EMBL Hamburg.

The electron density in the active site was interpreted as a tri-peptide gly-ala-arg with the arginine binding to the P1 site, in close contact with Ser195 of the catalytic triad. Two water molecules nearby are in favourable position for involvement in catalysis. The structures showed significant deviations from standard geometry and valence state (Figure 3). In order to explain these deviations and to assess the electronic situation, quantum chemical calculations were carried out on the native and inhibited structures which indicated that a true reaction intermediate of considerable covalent character was trapped in the active site. These calculations also indicated that both water molecules were involved in proteolysis, W1 as the nucleophile, W2 as the “activator” via an extremely short H-bond to the substrate carbonyl oxygen. W2 showed enormous capacity to take up excess electrons, the charge of which was mediated by the contacts to the oxyanion hole.

The final structure indicated that the molecule was very dynamic, showing different conformations for the “loaded” and “empty” states. The occupancy of the substrate was strongly correlated with the degree of disorder and the ratio of occupancies between the individual conformations of the protein. Binding of inhibitors triggered a change in the anisotropy of the active site residues indicating a tendency for directional movement of the active site residues in the inhibitor structures (Figure 4).

Multipole refinement with MoPro (Guillot *et al.*, 2001) was applied to the native structures containing the autoproteolysis fragment. These results directly showed interesting electronic details in the protein active site and confirmed the results of the quantum

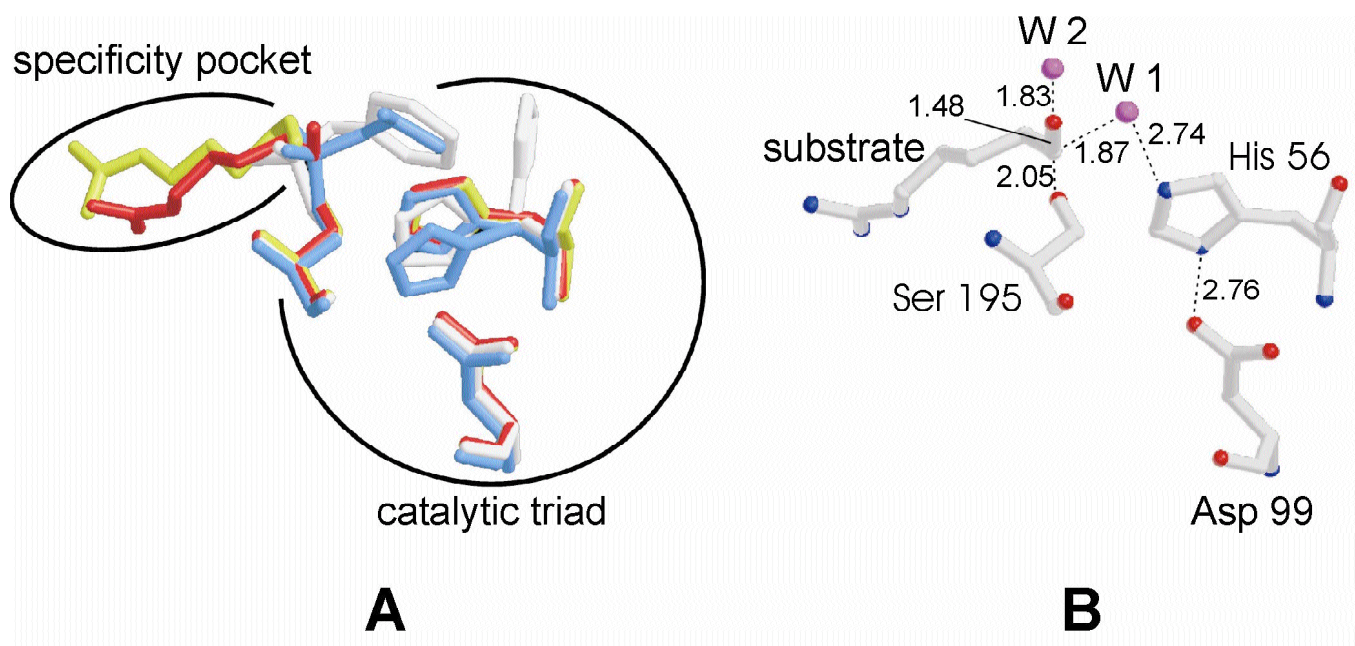


Figure 3. A: Overlay of the residues constituting the active site in trypsin showing the catalytic residues and the substrates/inhibitors; B: The active site with severe deviations from the “standard” stereochemistry.

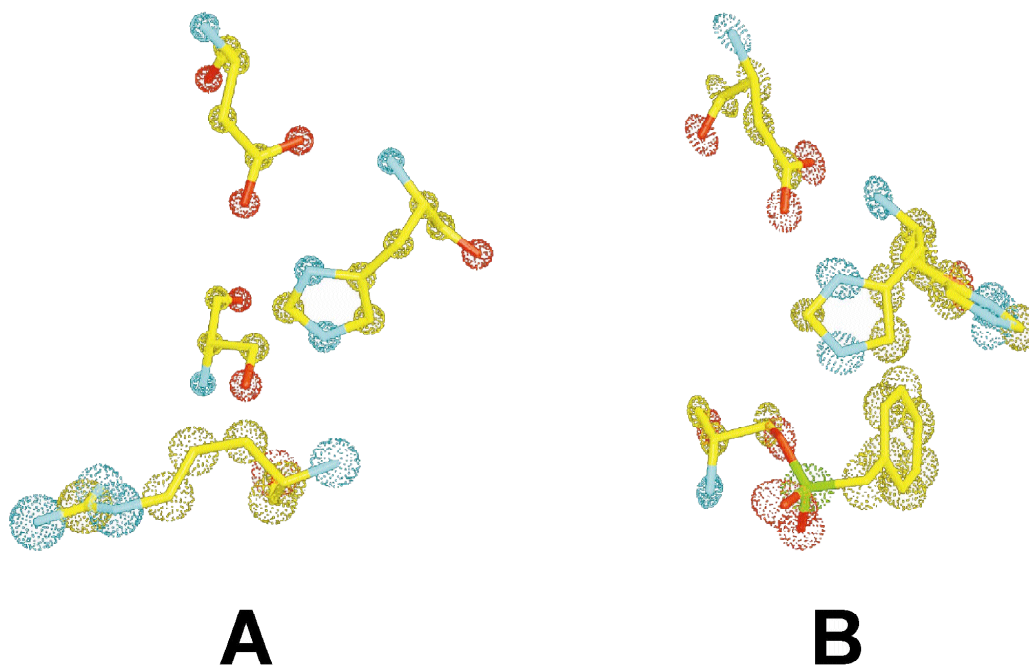


Figure 4. Thermal ellipsoids (30% probability) for the catalytic triad and the substrate (A) and inhibitor PMSF (B) show higher anisotropy for the covalently bound inhibitor.

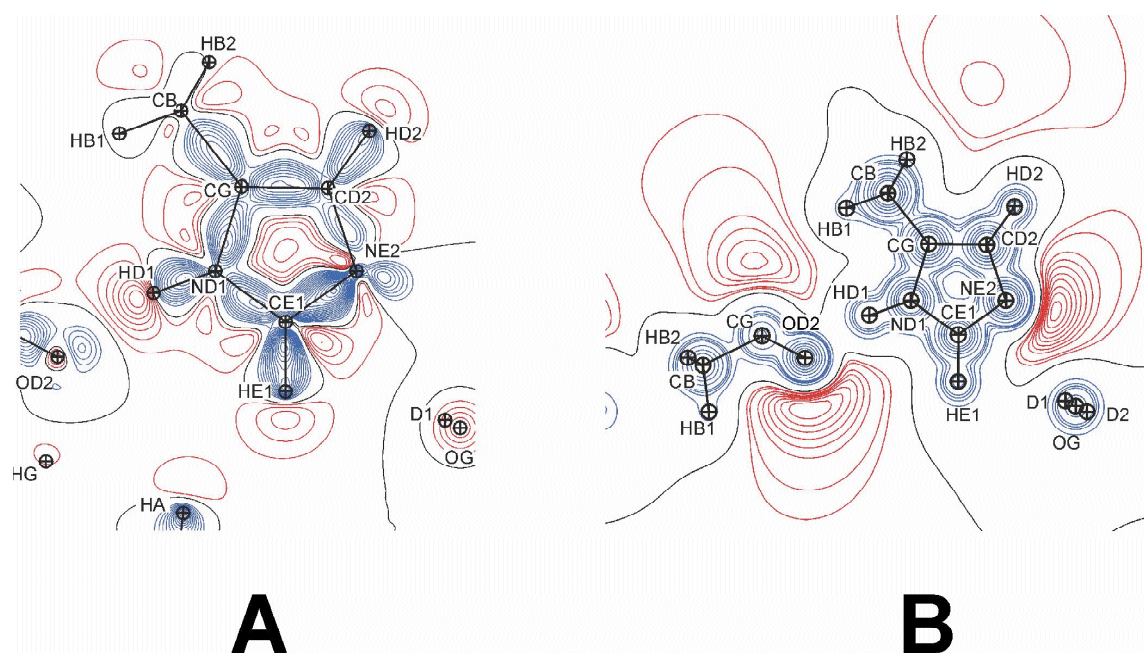


Figure 5. A: deformation density (deviations from spherical representation) of His 56 of the catalytic triad showing its non-protonated state. B: electrostatic potential map showing the charge distribution around His 56. His 56 is a strong nucleophilic activator for water W1.

chemical calculations: a covalent interaction between the de-protonated Ser 195 and the substrate carbonyl, and the de-protonated nature of His 56 of the catalytic triad were observed. A map of the electrostatic potential in the active site showed strong negative charge on His 56 indicating the nucleophilic character necessary for the activation of the catalytic water molecule W1 (Figure 5).

Grow better crystals: Modelling crystal contacts by site-directed mutagenesis and assessment of crystallisation results

(Petrus H. Zwart, Katja Schirwitz, Victor S. Lamzin, EMBL-Hamburg)

A need for availability of crystals that diffract to very high resolution has lead us to a recent launch of a project aimed at investigation on the rational mutagenesis of surface residues for improvement of X-ray crystal diffraction. It is hypothesised that surface residues with high conformational entropy, specifically lysines and glutamates, impede protein crystallisation. Mutating these to alanines often results in more successful crystallisation, particularly when clusters of lysines are disrupted.

We have chosen formate dehydrogenases as a model system as they show variable sequences and crystallisation tendencies. The various crystal forms displayed a wide range of diffraction power and contain different number of large clusters of positively charged residues. We have isolated the genomic DNA from *Candida boidinii* (ATCC 48180) (DSMZ, Braunschweig) and identified the formate dehydrogenase (fdh) gene as identical to EMBL Nucleotide Sequence Database accession no. AJ011046. We amplified the fdh gene using flanking

primers and cloned it into an *E. coli* expression vector. The integrity of the recombinant DNA was confirmed by DNA-sequencing (MWG-Biotech). *Candida boidinii* formate dehydrogenase (Cbfdh) was expressed in *E. coli*. Based on the purification procedure described for native fdh isolated from *Candida boidinii* and *Candida methylca* (Slusarczyk *et al.*, 2000 and Allen *et al.*, 1995, respectively) a routine purification protocol for the recombinant Cbfdh was established in our lab yielding to 10mg pure and homogenous recombinant Cbfdh from one liter *E. coli* – suspension. Spectrophotometric determination of formate dehydrogenase activity showed specific activity of 1.5 U/mg at 22°C for the recombinant Cbfdh. We established an Isothermal Calorimetry assay to distinguish the activity of different sites involved in catalysis. Dynamic Light Scattering proved the monodispersity of the protein solution (Polydispersity-index < 0.1). Interestingly we observed some variations on the apparent molecular weight of the protein dependant on the conditions. Right now it is unclear whether this effect is based on temperature changes or potential secondary modification of the protein. A Small Angle X-ray scattering experiment is planned to investigate the precise oligomerisation state of the recombinant Cbfdh. Mass spectrometry will be carried out to check for phosphorylated sites as recently found in fdh of potato tuber mitochondria (Bykova *et al.*, 2003) and indicated by three protein band on an Isoelectric focussing PAGE.

Various disorder prediction programs identified some regions of assumed disorder in the Cbfdh protein. Combining these predictions with the modeling of the Cbfdh protein on the *Pseudomonas* sp.101 high-resolution X-ray structure (Labrou *et al.*, 2001) was used to identify amino-acids potentially problematic for crystallisation. We will monitor the influence on the crystallisation of replacing these amino-acids with alanines. In a

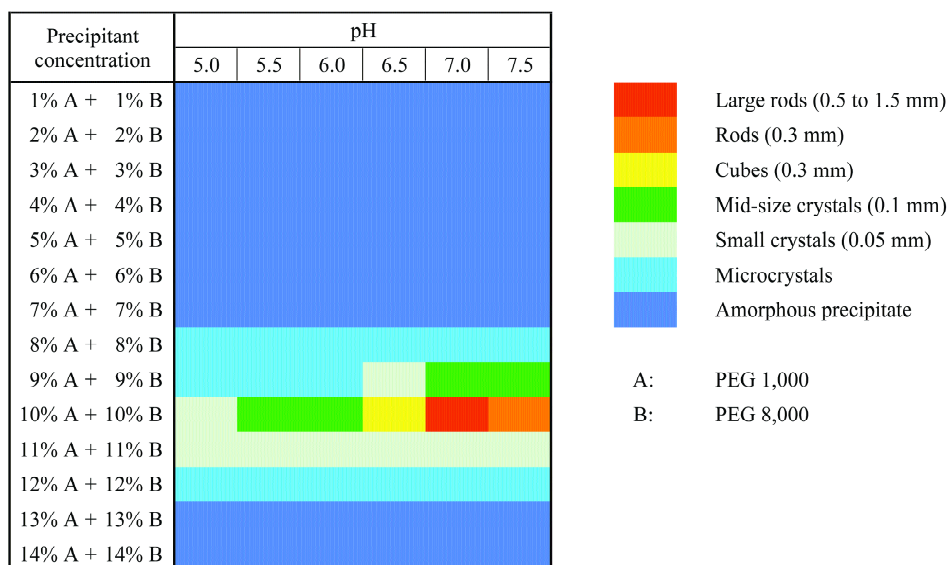


Figure 6. Optimisation of crystallisation conditions.

long-term prospective we expect to proceed towards developing a mechanism for the prediction of amino acids that interfere with the formation of intermolecular contacts needed for nucleation and proper crystal growth.

We intend to link the rational mutagenesis studies with more general studies of protein crystallisation. In our collaborative work on crystallisation of the four-copper

laccase from *Coriolus hirsutus* (Pegasova *et al.*, 2003) we arrived at a neat graphical representation of the results of crystallisation trials. Although we have not carried out a systematic full factorial analysis for optimization of the crystal growth as described, for example, by Carter and Yin (1994), we investigated the effect of the pH of the buffer and the concentration of the precipitant (Figure 6).

Publications during the year

Morris, R.J., Perrakis, A. & Lamzin, V.S. (2003) ARP/wARP and automatic interpretation of protein electron density maps. In *Methods Enzymol.* Carter, C.W. & Sweet, R.M. (Eds.) 374, 229-244

Pegasova, T.V., Zwart, P., Koroleva, O.V., Stepanova, E.V., Rebrikov, D.V. & Lamzin, V.S. (2003). Crystallization and preliminary X-ray analysis of a four-copper laccase from *Coriolus hirsutus*. *Acta Cryst.*, D59, 1459-1461

Schmidt, A., Jelsch, C., Ostergaard, P., Rypniewski, W. & Lamzin, V.S. (2003). Trypsin revisited: crystallography at (sub) atomic resolution and quantum chemistry revealing details of catalysis. *J. Biol. Chem.*, 278, 43357-44336

Terwilliger, T.C., Park, M.S., Waldo, G.S., Berendzen, J., Hung, L.W., Kim, C.Y., Smith, C.V., Sacchettini, J.C., Bellinzoni, M., Bossi, R., *et al.*, (2003). The TB structural genomics consortium: a resource for Mycobacterium tuberculosis biology. *Tuberculosis (Edinb)*, 83, 223-249

Zwart, P.H. & Lamzin, V.S. (2003). Distance distributions and electron-density characteristics of protein models. *Acta Cryst.*, D59, 2104-2113

Other references

Abramovicz, M. & Stegun, I.A. (1974). *Handbook of mathematical functions.* Dover Press, New York, NY.

Allen, S.J. & Holbrook, J.J. (1995). Isolation, sequence and overexpression of the gene encoding NAD-dependent formate dehydrogenase from the methylotrophic yeast *Candida methylolica*. *Gene*, 162, 99-104

Arfken, G.B. & Weer, H.J. (1995). *Mathematical methods for physicists.* Academic Press, San Diego, CA

Bykova, N.V., Stensballe, A., Egsgaard, H., Jensen, O.N. & Moller, I.M. (2003). Phosphorylation of formate dehydrogenase in potato tuber mitochondria. *J. Biol. Chem.*, 278, 26021-26030

Carter, Jnr. C. W. & Yin. Y. (1994) Quantitative analysis in the characterization and optimization of protein crystal growth *Acta Cryst.* D50, 572-590

Guillot, B., Viry, L., Guillot, R., Lecomte, C., and Jelsch, C. (2001). Refinement of proteins at subatomic resolution with MOPRO. *J. Appl. Cryst.*, 34, 214-223

Labrou, N.E. & Rigden, D.J. (2001). Active-site characterization of *Candida boidinii* formate dehydrogenase. *Biochem. J.*, 354, 455-463

Lamzin, V.S., Perrakis, A., & Wilson, K.S. (2001). *International tables for crystallography.* In "Crystallography of Biological Macromolecules." M. Rossmann & E. Arnold (eds). Kluwer Academic Publishers, Dordrecht, The Netherlands, 720-722

Morris, R.J., Perrakis, A. & Lamzin, V.S. (2002). ARP/wARP's model-building algorithms. I. The main chain. *Acta Cryst.* D58, 968-975

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

Russel, S. & Norvig, P. (1995). In *Artificial Intelligence*. pp. 77-80, Prentice Hall, ISBN 0-13-360124-2.

Rypniewski, W.R., Ostergaard, P.R., Norregaard-Madsen, M., Dauter, M. & Wilson, K.S. (2001). *Fusarium oxysporum* trypsin at atomic resolution at 100 and 283 K: a study of ligand binding. *Acta Cryst. D57*, 8-19

Scheres, S.H.W. & Gros, P. (2001). Conditional optimization: a new formalism for protein structure refinement *Acta Cryst.*, D57, 1820-1828

Slusarczyk, H., Felber, S., Kula, M.R. & Pohl, M. (2000). Stabilization of NAD-dependent formate dehydrogenase from *Candida boidinii* by site-directed mutagenesis of cysteine residues. *Eur. J. Biochem.*, 267(5), 1280-1289

Weisstein, E. (1999). *CRC concise encyclopedia of mathematics*. Chapman & Hall, New York, NY