The crystallization expert system Xtaldb, and its application to the structure of the 5'-nucleotidase YfbR and other proteins

Matthew David Zimmerman
Johnstown, PA

B.S. in Biochemistry, Juniata College, Huntingdon, PA, 1998

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Structural, Computational Biology and Biophysics Program

University of Virginia
January, 2008

Dr. Wladek Minor W. Minor

Dr. Edward Egelman

Dr. Robert Nakamoto Robert K Nakamoto

Dr. Fraydoon Rastinejad Fraydoon Rastinejad

Dr. Michael Wiener Michael Wiener

Abstract

Growing crystals is a critical step in the process of determining the 3-D structure of macromolecules by X-ray crystallography. While significant progress has been made in analyzing quantitatively crystallization experiments, in general, crystallization of biological macromolecules remains more of an art than a science. This work presents Xtaldb, an expert system for the quantitative analysis of crystallization experiments. Xtaldb provides tools for efficiently designing crystallization screens, tracks in a semi-automatic manner most of the parameters of the experiments, and provides sophisticated search, analysis, and data graphing tools. The algorithm used to produce balanced random screens is the fastest and most robust available. Xtaldb was tested on a set of six novel proteins that had failed to produce diffraction-quality crystals in a high-throughput structure determination pipeline. Of them, five yielded crystals diffracting to 3.5 Å or better and three 3-D structures were elucidated. One of the three proteins was YfbR, a member of the HD-domain phosphohydrolase superfamily. Structural analysis of the 3-D structure and biochemical work confirmed phosphohydrolase activity. Further studies by a collaborator demonstrated that YfbR is a 5'-deoxynucleotidase. Xtaldb was used to produce two crystals of catalytically inactive YfbR mutants in the presence of metal cofactor and the substrates TMP or dAMP. The structures of the complexes explained the mechanism of the unique pattern of substrate selectivity, further supported by computational docking studies. The complex structures suggested a plausible atomic mechanism of catalysis for the enzyme, the first proposed mechanism for an HD-domain phosphohydrolase based directly from enzyme-substrate complex structures.

Table of Contents

Acknowledgements

Thanks first of all to Wladek Minor. I appreciate the opportunity to be your advisee, though I apologize for not picking up Polish faster. I am also grateful to Maks Chruszcz, Igor Shumilin, Marek Grabowski, and Jim Spencer for their mentorship. (I would like to particularly thank Maks for reading a draft of this manuscript.) Thanks to Marcin Cymborowski, Heping Zheng, Janusz Petkowski, Kasia Koclega, Wojtek Potrzebowski, Katya Filippova, Olga Kirillova, John Raynor, Irek Plis, Michal Bujacz, Piotr Lasota, Brady Lauback, Amit Phull, and to the other members of the Minor lab for all of your work and for making the fourth floor of Jordan Hall a fun place to be.

Thanks to Michael Wiener. I first learned how to be a crystallographer under your guidance. A few years ago, you handled a difficult situation with remarkable grace, and I am very grateful for that. Much thanks to Arun Mohanty, David Chimento, Michael Purdy, Chad Simmons, Ann Marie Stanley, and James Vergis for your friendship and advice, as well as the other members of the Wiener lab. Thanks also go to David Cooper, and John Chapman. I also thank the other members of my committee, past and present: Fraydoon Rastinejad, Robert Nakamoto, Edward Egelman, and Eduardo Perozo.

I thank Andrzej Joachimiak, Qizhi Zhang, and the staff of the Structural Biology Center and the Biology department at Argonne National Laboratory for help in data collection and Aled Edwards, Alexei Savchenko, Alexander Yakunin, Marina Kudritska, and Michael Proudfoot at the University of Toronto for providing the original clones and the initial protein production protocols, and designing and running the general phosphohydrolase screens. Some of the results shown in this report are derived from work performed at Argonne National Laboratory, at the Structural Biology Center of

APS. Argonne is operated by University of Chicago Argonne, LLC, for the U.S. Department of Energy, Office of Biological and Environmental Research under contract DE-AC02-06CH11357. The work described was supported by NIH PSI grants GM62414 and GM074942.

Thanks to all of my wonderful friends, both those from before grad school and those I met in Charlottesville: David S., Joel, Heather J., David J., Selena, Lindsay, Gail, Connie, Brantley, Sarah C., Tim, Virginia, Sara P., Jessie, Meredith, Kim, Lisa F., Lisa H., Jo, Todd, Alex, Andrew, Brian V., David V., Jimbo, Grant, Katie, Geoff, Aaron, Will, Brian L., Caroline, Sasha, Missy—well, there are too many of you to count. Particular thanks to everyone I got to know at the Wesley Foundation at U.Va., especially Alex Joyner and Deborah Lewis, and at Wesley Memorial U.M.C., especially Gary Robbins, Elizabeth Foss, Sylvia Milner, Renita Sheesley-Banks, Todd Shelton, and all of the wonderful members of the choir. You kept me sane through all of this. I particularly want to thank Bonnie. I didn’t get to know you until relatively late in this process, and I know things didn’t turn out the way we had hoped, but I don’t know if I would have finished without your support in the last several months.

And finally, I thank my family, particular my parents Rev. Dennis and Debra Zimmerman, my grandparents Albert and Jane Beachem and Merlin and Judy Zimmerman, and my sister's full house: Suzanne, Ben, Ellie, Claire, Abbey, and Micah Smedberg. Throughout this very long journey (nearly a decade!), you have been my emotional anchor. I couldn't have done this without you.

Index of Figures

Index of Tables

Acronyms

Acronym	*Full name*
AMP	Riboadenosine-5’-monophosphate
API	Abstract programming interface
AU	Asymmetric unit
BMCD	Biological Macromolecular Crystallization Database
BME	β-Mercaptoethanol
CSV	Comma-separated values
dAMP	2’-Deoxyriboadenosine-5’-monophosphate
HEPES	4-(2-Hydroxyethyl)-1-piperazineethanesulfonic acid
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K_M	Michaelis constant
LB	Lysogeny broth
LIMS	Laboratory information management system
MAD	Multiple-wavelength anomalous diffraction
MES	4-Morpholinoethanesulfonic acid
MIR	Multiple-wavelength isomorphous replacement
NCS	Non-crystallographic symmetry
NDSB	Non-detergent sulfobetaines
NDSB 195	Dimethylethylammonium propane sulfonate
NDSB 201	3-(1-Pyridino)-1-propane sulfonate
NDSB 256	Dimethylbenzylammonium propane sulfonate
NTA	Nitrilotriacetic acid
OA	Orthogonal array
PDB	Protein Data Bank
PDE	Eukaryotic 3’,5’-cyclic-phosphodiesterase
PEG	Polyethylene glycol
P_i	Inorganic phosphate
pI	Isoelectric point
PMSF	Phenylmethylsulfonyl fluoride
pNPP	4-nitrophenylphosphate
pNPP-Na	4-nitrophenylphosphate disodium salt
RDBMS	Relational database management systems
SAD	Single-wavelength anomalous diffraction
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SeMet	Selenomethionine
SIR	Single-wavelength isomorphous replacement
SQL	Structured Query Language
TB	Terrific Broth
TEV	Tobacco etch virus
TLS	Translation/libration/screw paramters
TMP	2’-Deoxyribothymidine-5'-monophosphate
V_{max}	Michaelis-Menten maximum catalysis velocity
XML	Extensible Markup Language

1 Introduction

1.1 Macromolecular X-ray crystallography

Crystallography of biological macromolecules is a technique for experimentally determining the three-dimensional structure of the atoms of the proteins or nucleic acids or both under study. Electrons of atoms in biological macromolecules predominantly scatter X-rays elastically, meaning when they are shot with collimated, monochromatic X-ray radiation, they emit radiation of the same wavelength.

The monochromatic waves emitted from different atoms constructively or destructively interfere with one another in different directions. If there is no long- range order of atoms in space, the pattern of interference is too weak to be seen. However, if the atoms are arranged in an ordered manner, regular interference occurs and the resulting pattern of diffraction may be measured with the appropriate equipment. Waves of wavelength λ of two identical objects separated by a distance d will maximally reinforce at specific diffraction angles θ given by Bragg's law,

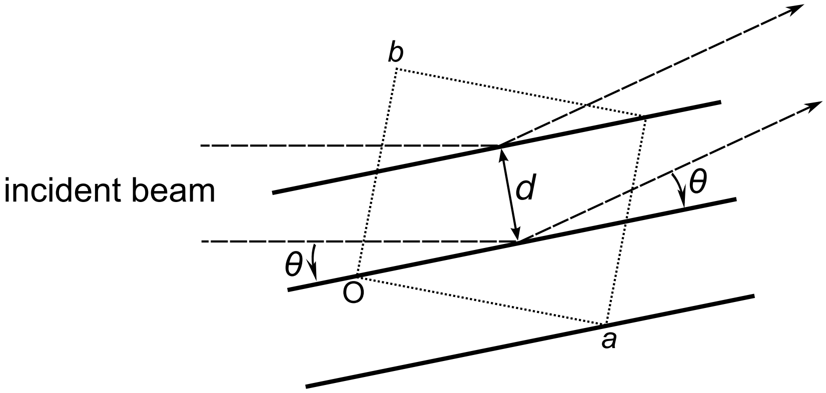
$$2d\sin\theta = n\lambda$$

(where n is an integer ≥ 0). Bragg's law also applies to objects in regular lattices in two and three dimensions. When the atoms are arranged in a crystalline lattice, the repeating subunit (or unit cell) is a block of length a, b, and c, respectively, along each of its sides.

One can imagine an infinite number of sets of parallel planes subdividing the lattice. As the crystal lattice becomes very large (i.e., contains many unit cells in all directions), only those planes of slope that evenly divide the lattice by integer fractions of the unit cell dimensions will contain perfectly repeating patterns of electron density, and

Figure 1-1: Bragg diffraction planes.

The solid lines represent the Bragg diffraction planes corresponding to (1,2,0) passing through the rectangular unit cell shown. The incident beam strikes the planes and are "reflected" in a particular direction θ *given by the plane spacing* d *and the wavelength of the incident beam.*



only those planes produce measurable diffraction beams. Thus the continuous pattern of diffraction becomes essentially discrete. Each set of planes can be represented by a triplet of integers (*h*, *k*, *l*), where each index represents the number of times the planes subdivide each unit cell. For example, Fig. 1-1 illustrates Bragg's law for the (1,2,0) planes of a typical crystal. Note too that the indices can be negative as well, indicating slope in a dimension in an opposite direction. As *h*, *k*, and *l* increase, the spacing (or resolution) *d* decreases, and as *d* decreases, the diffraction angle *θ* increases.

The mathematical construct representing the phase and amplitude of the wave "reflected" from a particular set of Bragg planes is called a *structure factor*, written as **F**(*hkl*) (the commas between *h*, *k*, and *l* are omitted for conciseness). A structure factor vector with amplitude |**F**| and phase *φ* is typically expressed as a complex number using Euler's formula:

$$\mathbf{F} = |\mathbf{F}|\cos\phi + i|\mathbf{F}|\sin\phi = |\mathbf{F}|\exp(i\phi).$$

Each structure factor can be represented as a sum of the electron scattering factors f_j, which is an atom-dependant quantity that describes how strongly each atom *j* scatters electrons as a function of atomic number (the number of electrons per atom), scattering angle and thermal motion, each multiplied by a vector representing the position of each atom and the reflection planes. The structure factor **F**(*hkl*) can be calculated as a function of each of the *n* atoms in the unit cell,

$$\mathbf{F}(hkl) = \sum_{j=1}^{n} f_j \exp\left(2\pi i \mathbf{r}_j \cdot \mathbf{S}(hkl)\right),$$

where $\mathbf{r}_j$ is a vector representing the position of atom *j* in the unit cell, and **S**(*hkl*) is the normalized vector perpendicular to the (*h*,*k*,*l*) diffraction planes.

The structure factors and the density $\rho(x, y, z)$ of electrons in the unit cell (of volume *V*) are related by the Fourier transform, written as

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l \mathbf{F}(hkl)\exp(-2\pi i(hx+ky+lz)).$$

However, it's important to note that **F**(*hkl*) is an imaginary vector quantity, with both an amplitude and a phase φ. The amplitudes of the structure factor vectors can be accurately determined by an X-ray diffraction experiment. The most popular is the so-called oscillation method. The protein crystal is mounted on a goniostat and rotated in a high-intensity X-ray beam. For each frame of data (usually a 0.5° to 1.0° wedge), reflections are collected on a two-dimensional X-ray detector. The intensity of each diffraction spot is proportional to the square of the structure factor amplitude. The fundamental problem is that the phases of each structure factor cannot be measured directly, making accurate phase determination one of the major obstacles in crystallography.

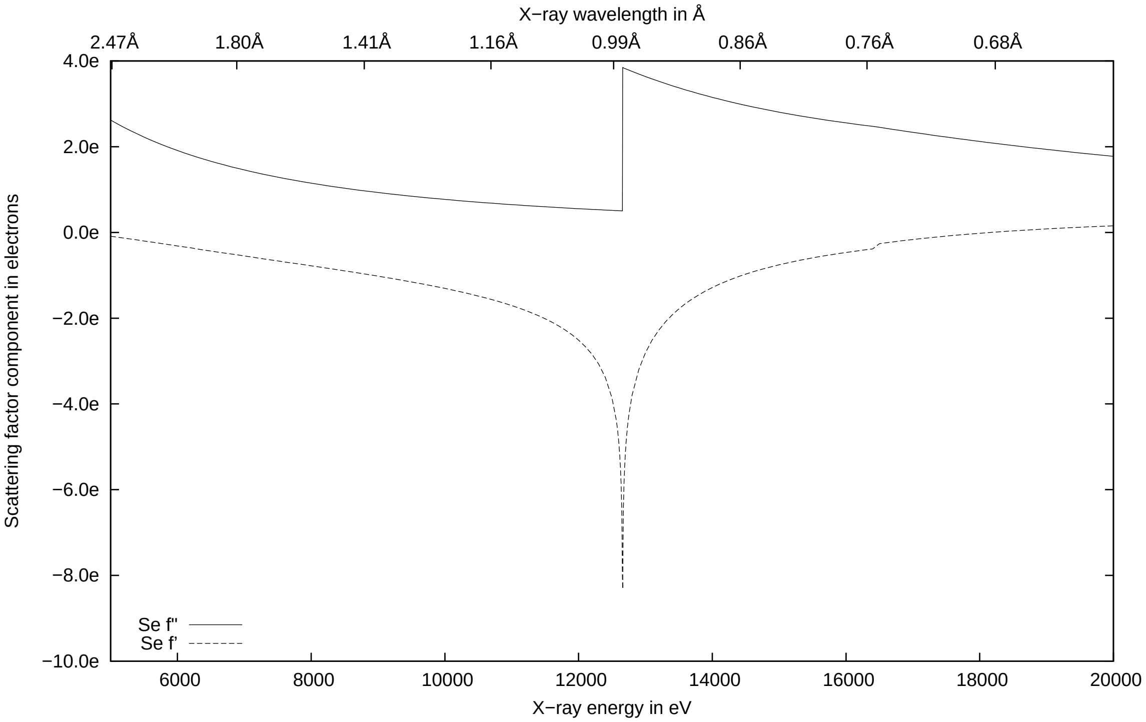
A number of different techniques for determining the phase information have been developed. If the number of atoms is small, the number of strong reflections is large, and the collected data are of atomic resolution (< 1.2 Å), the phases can even be determined directly from relationships between sets of reflections (direct methods; see Appendix A). Other techniques for producing phase information include isomorphous replacement, anomalous diffraction, and molecular replacement. In single-wavelength isomorphous replacement (SIR), phase information is derived from comparison of two sets of diffraction amplitudes collected on two different crystals. Usually one of the two crystals is untreated (native) and the other treated to contain high-molecular weight atoms, usually metals or other co-factors (derivative). Multiple-wavelength isomorphous replacement (MIR) is employed to improve phase information, where two or more derivative crystals are used. The main problem with the isomorphous replacement technique is imperfect isomorphism between crystals.

Anomalous diffraction works analogously, save that two sets of amplitudes are determined from a single data collection experiment so the data are perfectly isomorphous (in the absence of radiation damage). The anomalous diffraction method (Hendrickson, 1981) uses the fact that in practice the electron scattering factor for a given atom is a complex number containing three components, given by

$$f = f_o + f' + if'',$$

where f_o is the elastic scattering component, and f' and f'' are the anomalous scattering components, the former in phase and the latter rotated 90º out of phase (due to the *i* in the term). The anomalous components are wavelength-dependant, and are significant only near the resonance frequencies for excitation of elections for the element (see Fig. 1-2). Only selected elements, many of them metal ions or other high-molecular weight atoms, have detectable anomalous scattering components at the X-ray wavelengths commonly used for macromolecular crystallography. One method for introducing ordered

Figure 1-2: Coefficients for anomalous scattering factors of Se.
The theoretical amplitudes for the anomalous scattering factor components f' and f'' as a function of X-ray wavelength. The sharp jump in f'' near 0.99 Å is usually referred to as the absorption edge. (Data for the figure by Ethan Merritt at the University of Washington, taken from http://skuld.bmsc.washington.edu/scatter/.*)*



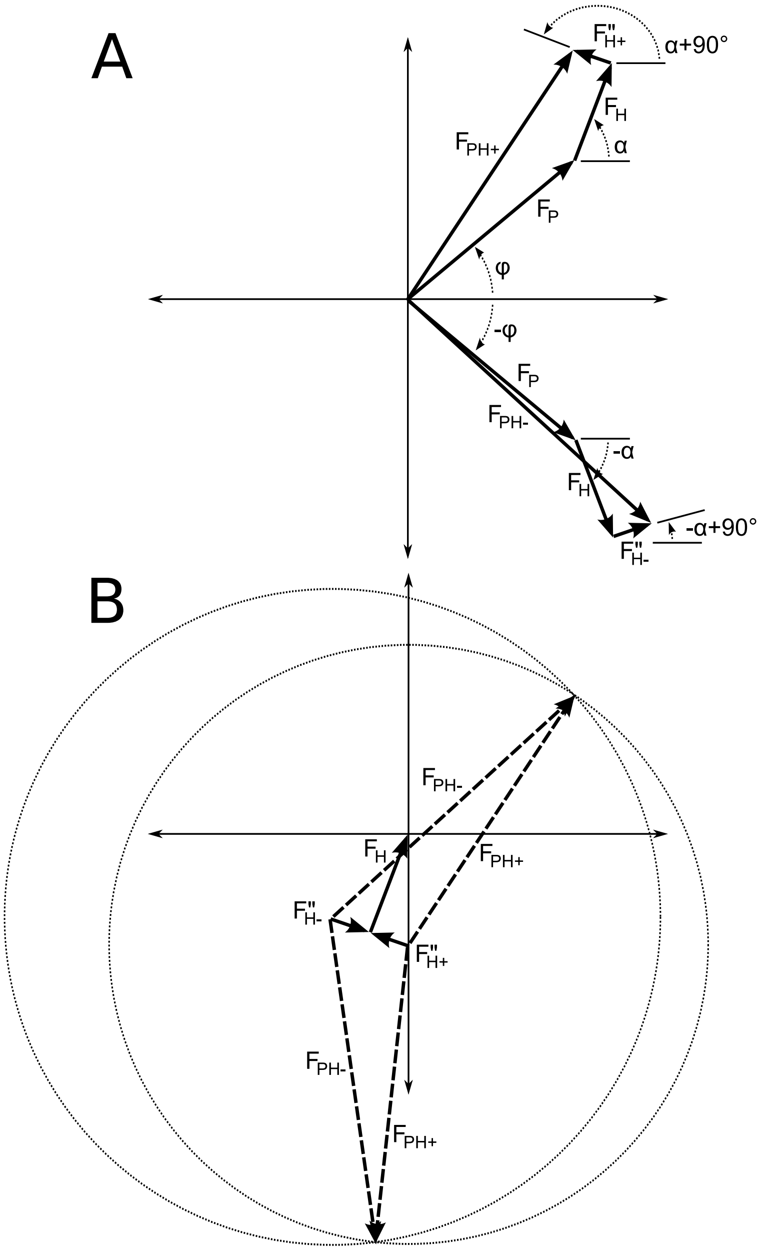
anomalous scattering atoms is to produce modified protein with selenomethionine, where the S atom in methionine is replaced with Se (Hendrickson *et al.*, 1989; Doublie, 1997), though S can be used for anomalous phasing (Hendrickson, 1981).

In purely elastic scattering, the structure factor amplitudes and phases for the pair of reflections (*h*,*k*,*l*) and (-*h*,-*k*,-*l*) differ only in the sign of the phase. However, if there are ordered anomalously scattering atoms, the contribution from those atoms produce a difference in the magnitude and the phases of the two structure factors. This is represented graphically in Fig. 1-3A, where the vectors for the **F**(*h*,*k*,*l*) and the **F**(-*h*,-*k*,-*l*) structure factors are shown (written as $\mathbf{F}_{PH+}$ and $\mathbf{F}_{PH-}$, for protein + heavy atoms). $\mathbf{F}_P$ is the elastic scattering from the protein alone, $\mathbf{F}_H$ the in-phase scattering component from the heavy atoms alone, and $\mathbf{F}_{H+}$ and $\mathbf{F}_{H-}$ the out-of-phase scattering components from the heavy atoms. If the positions of the heavy atoms in the unit cell are known, the phases and amplitudes of $\mathbf{F}_{H}$, $\mathbf{F}_{H+}$, and $\mathbf{F}_{H-}$ can be calculated, as shown in Fig. 1-3B. (There are a number of techniques for determining heavy atom positions from anomalous data, one of which is described in Chapter 4.) Since the amplitudes of $\mathbf{F}_{PH+}$ and $\mathbf{F}_{PH-}$ but not their phases are known, by drawing two circles of the appropriate radii, two estimates for the proper phase angles can be determined, as shown in Fig. 1-3B. This method of phase solution is known as single-wavelength anomalous diffraction (SAD). In SAD there is an ambiguity where the correct phase cannot be determined directly. Traditionally, additional sources of phase information, from additional isomorphous replacements or from another dataset collected at a different wavelength (or wavelengths) with different anomalous scattering components are used (multiple-wavelength anomalous diffraction; MAD).

The phase ambiguity in SAD can be resolved by density modification techniques. By taking the centroid phase between the two determined phases, weighted by a figure of merit (a measure of phase probability), an initial electron density map can be generated by the Fourier transform (as described on page 11). This map, while incorrect, does contain experimental phase information and can be further refined by using some *a priori* assumptions about the expected characteristics of the electron density for a biological macromolecule. One such assumption is that some regions of the map should contain essentially no ordered density (the solvent), and the remaining regions should contain density corresponding to the distributions observed in similar, previously solved

Figure 1-3: Phase diagram for SAD.

(A) Representation of the structure factors $\boldsymbol{F}_{PH+}$*, for reflection (h,k,l) and* $\boldsymbol{F}_{PH-}$*, for reflection (-h, -k, -l), and the components of which they are composed. In both figures, the horizontal axis is the real component of the complex number, and the vertical axis is the imaginary component. (B) A Harker construction (Harker, 1956) for SAD phasing, used to determine the proper phases for* $\boldsymbol{F}_{PH+}$ *and* $\boldsymbol{F}_{PH-}$*. The angles for the (-h,-k,-l) reflection are negated so they may be compared with (h,k.l) directly.*



structures. The techniques of solvent flattening and histogram matching iteratively modify the density map according to that assumption. Similarly, a reasonable assumption is that the electron density of a macromolecule should have high connectivity, which can be improved by various techniques, such as iterative skeletonization (Cowtan, 1994). For accurate SAD data, these methods are generally able to produce an interpretable map.

In general, anomalous diffraction requires more accurate and precise diffraction amplitude data than isomorphous replacement. Improvement in data collection and processing tools in recent years have increased the popularity of anomalous diffraction in general and SAD in particular (Dauter *et al.*, 2002; Minor *et al.*, 2006).

Another approach for phasing is molecular replacement, used in the special case where the 3-D structure to be solved is known to be structurally similar to a 3-D structure that has already been determined (the search model). Given the set of atoms ($\mathbf{r}_1$, …, $\mathbf{r}_\mathrm{n}$) in the search model, a set of calculated structure factors $\mathbf{F}_c$ can be calculated by summing the atomic scattering factors (using the expression for $\mathbf{F}(hkl)$ shown on page 11). A target function is defined which measures the difference between the amplitudes of the experimentally measured structure factors $\mathbf{F}_o(hkl)$ and $\mathbf{F}_c(hkl)$, and the search model is rotated and translated in the unit cell to minimize this target function. Molecular replacement may even be used in situations where the search structure and the target data have different spacegroups and unit cell parameters. However, it is limited in effectiveness to cases where the search and target structures are similar, and the technique can introduce bias into the model for the target as the initial phases come entirely from the search model.

Once an initial set of phases is determined by one of the methods described above, a 3-D electron density map is generated. The next step is to build the molecular model, which is the interpretation of the electron density map. In the case of proteins, the polypeptide backbone of recognizable secondary structure elements are built first. Once sections of backbone are built, the density corresponding to the sidechains of each amino acid residue can be used to dock the known polypeptide sequence into the electron density. A number of programs have been developed to automate the process of initial map interpretation and model building; two of them are described in Chapter 4. A number of 3-D graphical applications have been developed for the purposes of building a molecular model into electron density, such as O (Jones *et al.*, 1991) or COOT (Emsley & Cowtan, 2004).

Each atom in the model typically contains a correction for thermal motion, where the corrected electron scattering factor f_j for atom j is given by

$$f_j^2 = \exp\left(-2B_j \frac{\sin^2\theta}{\lambda^2}\right)\left(f_j^o\right)^2 .$$

B_j is a quantity representing the thermal motion of atom j, and f_j^o is the scattering factor for a stationary molecule. Due to the limited resolution of data in macromolecular crystallography, B is usually modeled isotropically (assuming uniform vibration in all directions) using a single parameter per atom, proportional to the mean of the square of the atomic vibrations u:

$$B = 8\pi^2\overline{u^2} .$$

There are alternative methods of modeling B. One is an anisotropic model, which uses six parameters to model the thermal vibration of the atom as an ellipsoid rather than a sphere, but requires very high resolution data due to the significant increase in the number of parameters. Another involves adding translation/libration/screw (TLS) parameters, where the macromolecule is subdivided into rigid units, and the overall torsional motion of each unit is refined and added as a component of the B-factor for each atom (Schomaker & Trueblood, 1968; Schomaker & Trueblood, 1998).

When an initial model has been built, it is refined against the collected structure factor amplitude data. One of the most effective methods of computational refinement is the maximum-likelihood method, as implemented in the program REFMAC (Murshudov *et al.*, 1997). The conditional probability distribution of a set of model parameters $\mathbf{x}$ (which includes the atom positions, B values, etc.) given the set of observed structure factor amplitudes $|\mathbf{F}_\mathrm{o}|$ is written $P(\mathbf{x}, |\mathbf{F}_\mathrm{o}|)$. ($P(\mathbf{A};\mathbf{B})$ is defined as the conditional probability of **A** given a known **B**.) By using Bayes's method (and assuming the errors in each structure factor are independent) this conditional probability can be expressed as

$$P\left(\mathbf{x};\left|\mathbf{F}_o\right|\right)\propto p(\mathbf{x})\prod_{h,k,l}P\left(\left|\mathbf{F}_o(hkl)\right|;\mathbf{F}_c(hkl)\right),$$

where $p(\mathbf{x})$ is the prior probability (in other words, the known prior information) for the model parameters $\mathbf{x}$ (Murshudov *et al.*, 1997). The $p(\mathbf{x})$ term is implemented in terms of stereochemistry (constraints on the bond lengths, angles, *etc.* of the atoms), as well as other constraints such as non-crystallographic symmetry (NCS), where two or more copies of a model in an asymmetric unit are constrained to have similar model parameters. As in molecular replacement, the set of $\mathbf{F}_\mathrm{c}$ are calculated from the model. Rather than maximizing this function, this is calculated as the minimization of the log-likelihood function *LLK*, derived by taking the negative logarithm of both sides, as

$$LLK=-\log P\left(\mathbf{x};\left|\mathbf{F}_o\right|\right)\propto -\log p(\mathbf{x})-\sum_{h,k.l}P\left(\left|\mathbf{F}_o(hkl)\right|;\mathbf{F}_c(hkl)\right).$$

Thus there are two terms in the function to be minimized, one representing the agreement of the model with stereochemistry, and one representing the fit of the calculated structure factors with those observed. Typically a weighting term is used to balance the relative importance of each term during the minimization process. The minimization is performed iteratively by standard gradient minimization methods (Press *et al.*, 1992).

After a cycle of refinement, the refinement model is loaded back into a graphical application, and the model can be compared to the electron density map. For this purpose, σ_A-weighted $2F_o$-F_c (which gives a continuous electron density in which to fit the model) and F_o-F_c (which shows the differences in the maps generated for the observed and calculated structure factors) maps are usually used. σ_A is a coefficient for the overall error in the structure factor distribution, which can be estimated from a relatively small set of reflections (Read, 1997). The maps are calculated by a Fourier transform using the model phases and amplitudes equal to $2m|\mathbf{F}_o|$-$D|\mathbf{F}_c|$ (or $m|\mathbf{F}_o|$-$D|\mathbf{F}_c|$ for F_o-F_c) where m and D are parameters derived from σ_A (Read, 1997). While this introduces bias into the map by using phases and part of the amplitudes from the model (though σ_A is intended to minimize this bias), generally the $2F_o$-F_c map is less noisy and easier to interpret. This new map is used to extend and improve the model, and then the whole cycle is repeated.

To monitor the convergence of the refinement, the correlation coefficient R is defined (Brunger, 1997), as

$$R = \frac{\sum_{h,k,l} \mathbf{w}(hkl)\left\| \mathbf{F}_o(hkl) \right| - k\left| \mathbf{F}_c(hkl) \right\|}{\sum_{h,k,l} \left| \mathbf{F}_o(hkl) \right|}.$$

$\mathbf{w}(hkl)$ is a weight for each structure factor, and k is a scaling factor. As the model improves and $|\mathbf{F}_c|$ converges toward $|\mathbf{F}_o|$, this correlation coefficient decreases in value. However, since the model does have some bias since it is built and refined against a map that contains data from the previous model, incorrect models can be refined to artificially low R values. The bias is also related to the relatively low data-to-parameters ratio in macromolecular crystallography, due to the limited resolution of the data. To counteract this bias, Brunger suggested an application of an established statistical method known as cross-validation (Brunger, 1992). The observed structure factor amplitudes are partitioned into two sets; a small test set containing about 5% of reflections, and the working set. Only the working reflections are used in the refinement and building process, and the test set is explicitly excluded. Since the model built does not contain any information derived from the test set, the correlation coefficient R_{free} (calculated identically to R save that only reflections in the test set are used) should report agreement between the model and the observed data free of bias (Brunger, 1992). Generally, R_{free} is larger than R, but like R, R_{free} should decrease with subsequent cycles of refinement.

Other methods of structure validation monitor the stereochemical properties of the model. The Ramachandran plot graphs the polypeptide backbone torsion angles φ and ψ, where some regions of the graph are geometrically favorable or allowed and others are forbidden (Ramachandran & Sasisekharan, 1968). Analogous methods exist for DNA. Some tools have been developed, such as PROCHECK (Laskowski *et al.*, 1993), that monitor bond lengths, angles, and torsion angles for statistical outliers as compared to the distributions found in previously solved structures. The Molprobity program (Lovell *et al.*, 2003) adds hydrogens to a molecular model (which are usually omitted in macromolecular crystallographic refinement) and then detects regions of the map where interatomic clashes are seen. Molprobity also detects bad rotamers—infrequently seen conformations of amino acid sidechains—and deviations in C_α-C_β geometry. Once a 3-D macromolecular structure is solved, refined to convergence, and validated, it is deposited in the Protein Data Bank (Berman *et al.*, 2002) and assigned a four-character identifier.

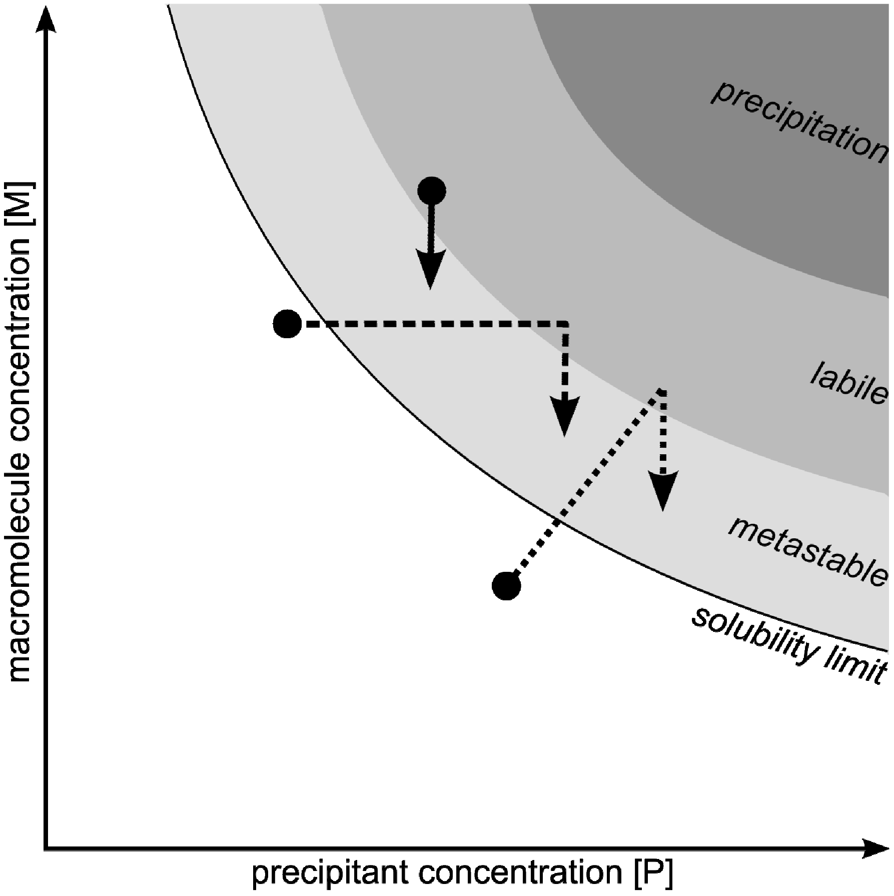
1.2 Biomacromolecular crystallization

Protein crystallization is a dynamic process that depends upon a large number of parameters, of both the biological and biochemical properties of the macromolecule and of the physical and chemical properties of the crystallization experiment (Giegé & Ducruix, 1992). To produce crystals of macromolecules for diffraction, sufficient quantities of soluble macromolecule(s) must be obtained and purified to near homogeneity (usually tens of micrograms to milligrams). Crystals used for these diffraction studies must be large (usually >50 μm in all three dimensions) and possess long- range order in order to measure diffraction amplitudes with sufficient accuracy—for example, better than 1% for SAD experiments. Because in general there are no methods to predict crystallization conditions *a priori*, crystallization remains a difficult step (particularly for “high-hanging fruit” like membrane proteins). Thus studies into the quantitative and statistical analysis of prior crystallization experiments have increased (Rupp & Wang, 2004).

The process of macromolecular crystallization requires that the protein or nucleic acid be supersaturated in aqueous solution, under conditions where nucleation and growth of crystals is both kinetically and thermodynamically favorable. Typically, such states are achieved by mixing highly concentrated macromolecule solutions with high concentrations of chemicals known as precipitants. Figure 1-4 shows a theoretical phase diagram for a macromolecule mixed with precipitant. Below the solubility limit for the macromolecule, no phase transition to crystals is observed. The phase above the solubility limit can be roughly divided into different regions which vary in kinetic

Figure 1-4: Idealized phase diagram of crystallization experiments.

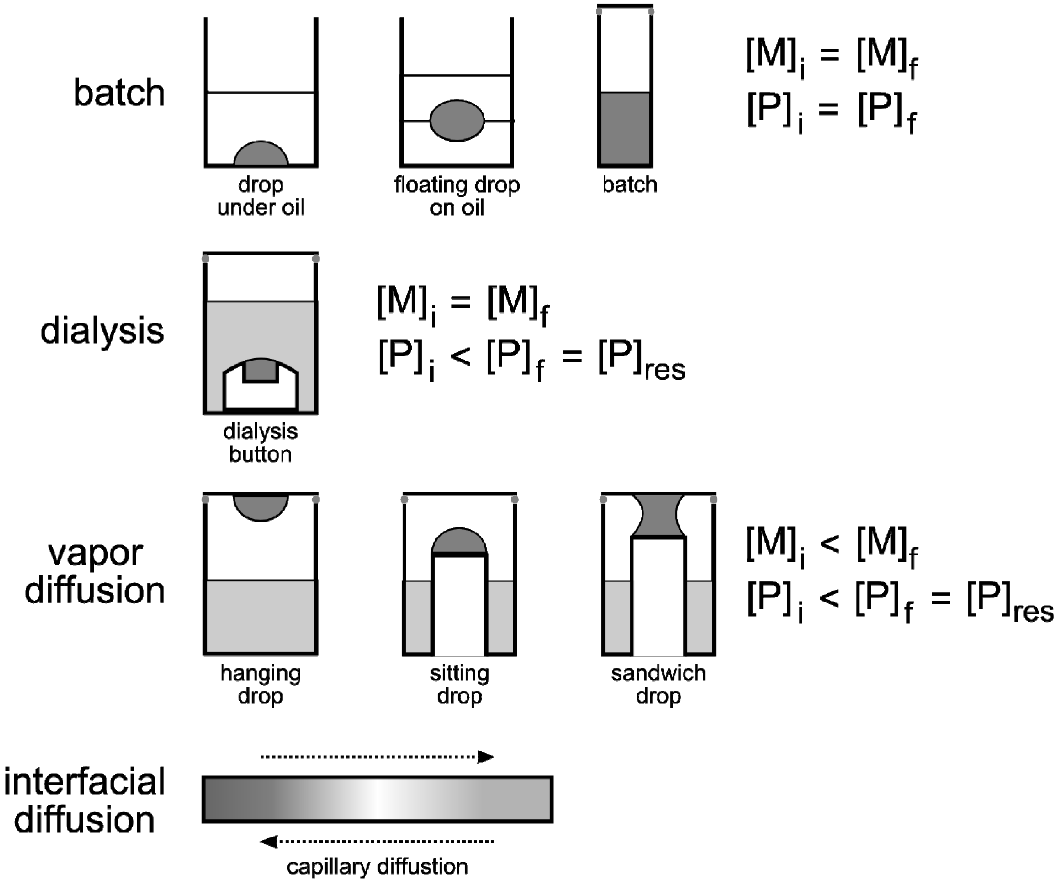
The horizontal axis represents the concentration of a precipitant (salt, polymer, etc.) and the vertical axis represents the concentration of the macromolecule to crystallize. The solid curve represents the limit of solubility for the protein as a function of precipitant concentration, and the different supersaturated regions above the solubility limit are marked in shades of gray. The different arrowed lines represent idealized successful batch (solid line), dialysis (dashed line), and vapor diffusion (dotted line) crystallizations. For each line, the dot represents the initial state of the experiment, and the downward arrow represents the drop of macromolecule concentration as crystals are formed.


precipitation
labile
metastable
solubility limit
macromolecule concentration [M]
precipitant concentration [P]

behavior. (The term "region" is used to emphasize that these regions are not formal thermodynamic phases). Just above the solubility limit is the metastable region, where the macromolecule is not yet concentrated enough to spontaneously form crystal nuclei, but will crystallize in the presence of existing nuclei (or other nucleation materials). Above

Figure 1-5: A survey of different crystallization methods.

In all of the diagrams, macromolecule solutions are shown in dark gray, and precipitant solutions are shown in light gray. For batch experiments, the concentrations of macromolecule and precipitant stay constant throughout the experiment, while for dialysis and vapor diffusion, the initial concentrations of macromolecule $[M]_i$ or precipitant $[P]_i$ or both may be lower than the final equilibrium concentrations ($[M]_f$ and $[P]_f$). In interfacial diffusion methods, a gradient of macromolecule or precipitant or both is formed in a variety of media.



the metastable region is the labile region, where crystal nuclei form fairly readily. Far into the labile region, amorphous precipitation is usually more favorable kinetically. An ideal crystallization experiment should begin (or initially reach) the labile region so that crystal nuclei may form, but then move into the metastable region so that the crystallizing protein is predominantly ordered into growing crystals rather than forming new nuclei.

There are a wide variety of crystallization methods that have been developed to traverse this macromolecule phase space, some of which are summarized in Fig. 1-5. The simplest crystallization experiment is batch crystallization, where the precipitant and macromolecule solution are simply mixed together and then sealed to prevent evaporation. While this is the most straightforward design for such an experiment, the concentration of macromolecule [M] and precipitant [P] in the solution do not change (apart from the possible drop in [M] due to crystal or precipitation formation—see Fig. 1-4). This does not allow the same opportunity to traverse the macromolecule's phase space that other techniques allow. Batch solutions can be simply sealed in a small container (such as a well of a 96- or 192-well plate), or drops of batch solutions may be placed under solutions of oil. A technique has also been developed to float batch solutions between oil solutions of different densities, which removes the influence of the container on the crystallization process (Chayen, 1996).

One method to increase the ability to traverse a macromolecule's phase space is to separate the macromolecule and the precipitant into different containers in a non-equilibrium state and then allow them to slowly equilibrate. In this case, the initial state of the macromolecule may even be below its solubility limit, and as the system slowly equilibrates, the macromolecule solution moves into the metastable region. One such technique is dialysis (Fig. 1-5), where a macromolecule solution is placed in a dialysis bag or button, separated by a porous membrane from the precipitant. The dialysis membrane has pores of a size that allow the precipitant to permeate through but not the macromolecule. Therefore the concentration of macromolecule remains constant but the precipitant concentration increases (Fig 1-4). Since dialysis is labor-intensive and difficult to scale to small volumes and to many experiments, the method is not frequently used.

A far more popular crystallization method is vapor diffusion (Fig. 1-5). A small crystallization drop containing a mixture of both the macromolecule and the precipitant solution, and a much larger reservoir containing only the precipitant, are sealed into an airtight container. In this setup, the concentration of the precipitant in the drop is lower than in the reservoir (or “well”). Over a period lasting hours to days, the system equilibrates as water diffuses through the vapor phase from the drop to the reservoir (Mikol *et al.*, 1990). This has the significant advantage that while the macromolecule solution is traversing crystallization space (Fig 1-4), it does so slowly, which kinetically favors crystal formation over precipitation. As well, vapor-diffusion methods are well-suited both for miniaturization and high-throughput experiments, particularly on 96-well crystallization plates. Multiple techniques for vapor diffusion are used, including hanging-drop, sitting-drop, and sandwich-drop setups (Fig. 1-5).

In some ways, interfacial diffusion methods (Fig. 1-5) are similar to vapor diffusion, but in this circumstance, one or both of the macromolecule and precipitant solutions are allowed to diffuse into a solid or liquid substrate. These methods cause a gradient of the solution or solutions to be formed, which allows multiple points of the phase diagram to be sampled simultaneously. Several different interfacial diffusion methods have been developed, such as capillary diffusion (Phillips, 1985), free-interface diffusion (Sauter *et al.*, 2001), and microfluidic chips (van der Woerd *et al.*, 2003).

The two-dimensional phase space diagram is an extreme oversimplification; macromolecular solubility depends upon many parameters, which are physical, biological, or chemical. A list of approximately 70 significant experimental parameters

Table 1-1: Parameters that may play a role in successful crystallization.

Physical parameters
Macromolecule identity
Macromolecule type
Macromolecule molecular weight(s)
Macromolecule isoelectric point(s)
Macromolecule sequence(s)
Putative flexible domains (tags, etc)

Cloning & expression
Time of expression
Person/robot who expressed molecule
Expression vector
Expression method
Expression species

Purification
Time of purification
Person/robot who purified molecule
Macromolecule purification method(s)
Macromolecule purity (%)
Macromolecular alterations (proteolysis, tag cleavage, etc.)
Solubility
Storage buffer(s) and pH
Storage temperature

Crystallization
Time of crystallization
Person/robot who set up crystallization
Drop volume (μL)
Well volume (μL)
Crystallization method
Plate type / material
Macromolecule concentration (M)
Complex component ratio
Precipitant(s) identity
Precipitant(s) concentration (M)
Additives / ligands identity
Additives / ligands concentration (M)
Temperature
Humidity
Initial screen or optimization?

Observation
Times of observation
Person/robot who made each observation
Entity / entities in drop
Number of entities
Amount of precipitation
Drop contaminants
Degree of crystal twinning or clustering
Crystal size (height, width, depth)
Crystal shape
Crystal color
Is crystal birefringent?
Is crystal macromolecular (i.e. not salt)?

Harvesting & freezing
Time(s) drop is opened
Reason drop is opened (seeding, harvesting, etc)
Seeding method
Crystal used for seeding
Identity of heavy atom in soak
Concentration of heavy atom in soak
Time of harvesting
Person/robot who harvested crystal
Harvesting method
Number of soaking/harvesting steps
Cryosolution component(s) identity
Cryosolution component(s) concentration
Behavior of crystal in cryosolution

Data collection
Time of diffraction
Person/robot who collected diffraction data
Space group
Diffraction limit
Crystal mosaicity
Unit cell parameters
Diffraction wavelength(s)
Anomalous signal

that can affect the production of well-diffracting crystals are listed in Table 1-1. Thus the parameter space that must be explored is not two-dimensional but highly multi-dimensional, where only small regions of this space correspond to combinations of parameters that yield diffracting crystals.

The process of sampling crystallization space is often divided, somewhat arbitrarily, into two phases: crystal screening and optimization. Crystal screening is carried out when little or no information is known about the best conditions for producing crystals of a given macromolecule. These initial screens attempt to efficiently sample a large range of possible conditions, recognizing the highly-dimensional nature of crystallization space. Many of these crystal screens, in the form of sets of precipitant cocktails, are commercially available (Jancarik & Kim, 1991; Stura *et al.*, 1992; Cudney *et al.*, 1994; Scott *et al.*, 1995). Once an initial "hit" is determined, the parameters of the hit are typically systematically varied to optimize that condition. Unlike initial crystal screening, where efficient mechanisms for sampling crystallization space are at least occasionally used, optimizations are frequently performed exhaustively, despite the fact that there may still be many parameters to be searched around the initial hit. The same efficient sampling techniques could be applied to optimization.

1.3 Crystallization space screening methods

Since the crystallization space is highly multi-dimensional, efficient techniques for intelligently sampling this space are needed, and several have been developed. These screening methods can be generalized in the terms of design of experiment (DOE) theory, as originally formalized by R. A. Fisher (Fisher, 1951) and applied to protein crystallization by C. W. Carter, Jr. and co-workers (Carter & Yin, 1994; Carter, 1999). A "basis set" of *n* factors is chosen, where each factor represents a parameter in the crystallization experiment, such as temperature, precipitant identity, precipitant concentration, drop volume, additive identity, pH, etc. Each factor has 2 or more discrete levels, where the number of levels for each parameter is given by the set (l_1, ..., l_n) with each $l_i \geq 2$. In cases where the factor can adopt a continuous range of values, such as precipitant concentration, a set of value steps are chosen, though some random screening algorithms use a continuous range for picking values of those parameters.

The simplest, and least efficient, type of experiment design is the full factorial, where every possible combination of factors is tested. Thus the number of experiments *E* required is given by:

$$E = \prod_{i=1}^{n} l_i \ .$$

24-well grid screening is an example of a full factorial design, with two factors of 4 and 6 levels, respectively. However, when the number of levels and/or number of parameters increases, the geometric increase in the number of experiments required makes full factorial designs unfeasible. Consider the example of 8 precipitants at 4 concentration levels, 4 organic additives at 3 concentration levels, and 5 different buffers, a typical basis set of reagents for a 48 or 96 condition commercial screen. A full factorial design would require 8×4×4×3×5 = 1920 experiments, which is prohibitive given the typically small amounts of purified, highly-concentrated protein available for crystallization screening. There were attempts to build automatic systems designed to screen >100,000 crystallization conditions per day (Abola *et al.*, 2000), but no further reports have confirmed if that level of output has been reached.

A number of screening methods have been developed to sample search space more efficiently, such as footprint (Stura *et al.*, 1992), random (Shieh *et al.*, 1995), sparse matrix (Jancarik & Kim, 1991), incomplete factorial (Carter & Carter, 1979), and response surface screens (Carter, 1997). Both theoretical models and experimental data show that random screens, where all of the parameters of the basis set tested are varied randomly, produces successful crystallization in fewer trials than grid and footprint screens, where one or two of the parameters involved are varied systematically (Segelke, 2001). Biased random screening methods, due to the proliferation of commercially available kits, have become the *de facto* standard for crystallization screening, but grid or full-factorial methods are far more widely used for crystal optimization, as it is considerably easier to design such experiments by hand or spreadsheet applications. Generally, the use of efficient methods of sampling parameter space such as random designs or response surfaces have been not widely reported for crystal optimization, likely due to the relative difficulty in designing and analyzing such experiments.

Random crystallization experiment designs may be subdivided into three different types: purely random, biased, and balanced. Given a basis set of parameters or “factors”, such as precipitant identity, concentration, temperature, protein batch, *etc*., purely random designs randomly choose combinations of factors until the requested number of experimental runs is found. A few publicly available crystallization tools generate purely random designs, such as CRYSTOOL (Segelke, 2001) and XtalBase (Meining, 2006). These programs deviate from purely random designs only in that both programs check for insoluble combinations of chemicals. Biased random designs, such as the “sparse matrix” screen (Jancarik & Kim, 1991), randomly select combinations of factors, but bias the selection toward conditions that have produced successful crystallizations in the past. Virtually all widely used commercial designs for initial screening follow this model (Jancarik & Kim, 1991; Stura *et al.*, 1992; Cudney *et al.*, 1994; Scott *et al.*, 1995), but many of them are often out-of-date, as they are based on the state of the PDB from many years ago. Algorithmic approaches have been proposed to use Bayesian statistics to generate such biased random screens (Hennessy *et al.*, 2000).

Balanced random screens also randomly select experiments, but place an additional constraint that the design be "balanced" (i.e., each possible level for each factor is represented an equal number of times). For example, given three factors **a**, **b**, and **c**, where each factor has three levels 0, 1, and 2, the experimental design shown below is balanced both with respect to each factor and to each pair of factors:

a	**b**	**c**
0	0	1
0	1	0
0	2	2
1	0	0
1	1	2
1	2	1
2	0	2
2	1	1
2	2	0

The factors are listed in columns, and each row represents an individual condition or "run". For all combinations of columns—(**a**,**b**), (**b**,**c**), and (**a**,**c**)—each possible combination of factors—(0, 0), (0,1), (0,2) (1,0), (1,1), (1,2), (2,0), (2,1), and (2,2)—occurs an equal number of times. Note that a full factorial design would require 27 runs. In DOE theory, such an experimental design is known as an orthogonal array of strength two (OA). Balanced designs are much more suited for statistical analysis by linear regression than purely random or biased designs (Carter, 1999). Using a predefined 64-condition OA for crystallization experiments has been previously suggested (Kingston *et al.*, 1994).

While it is trivial to generate a random experimental design where each individual factor is balanced, generating screens where all possible binary combinations of factors are balanced is far more difficult. While it may be impossible to generate an OA for a given set of factors and levels within a specified number of runs, creating an array that is as balanced as possible is a computationally difficult problem. A few programs have been developed to generate balanced random screens for crystallography: INFAC (Carter *et al.*, 1988), DESIGN (Sedzik, 1994), and SAmBA (Audic *et al.*, 1997). Of these three, only one (SAmBA) is still available online. The incomplete factorial method, as implemented by INFAC, and DESIGN informally address this problem by making sure that each combination of levels is represented at least once for each pair of factors, but often such designs are sub-optimally balanced (Audic *et al.*, 1997). SAmBA uses more sophisticated backtracking and simulated annealing algorithms to create balanced designs (Audic *et al.*, 1997).

1.4 Statistical and empirical analysis of crystallization

Quantitative analysis of macromolecular crystallization may be addressed in two different ways, using either “micro” or “macro” approaches. The “micro” approach addresses the fundamental thermodynamic and kinetic properties of crystallization experiments. These properties have been under study for many years, and most, but not all, of these mechanisms are fairly well understood (Durbin & Feher, 1996). However, this theoretical approach often proves less than useful in practical situations, as measuring fundamental thermodynamic and kinetic quantities of proteins often take up more time and resources than simply setting up best-guess crystallization conditions. One possible exception to this rule is the observed correlation of the solution property known as the second virial coefficient with the formation of macromolecular crystals (George & Wilson, 1994).

In contrast, the “macro” approach takes a top-down, empirical view of protein crystallization. Rather than pursuing the molecular mechanisms of crystallization, data from prior experiments are observed and then statistical trends are calculated. The earliest such analyses took successful crystallization conditions from the PDB or the literature and used it to produce some statistical analyses. However, this approach suffers from a “tip-of-the-iceberg” effect, as the possibly hundreds or thousands of unsuccessful experiments that produced the one successful outcome are not reported.

The first effort to create such a global database of crystal conditions was the Biological Macromolecular Crystallization Database (BMCD) of Gilliland and coworkers (Gilliland, 1988; Gilliland *et al.*, 1994; Gilliland *et al.*, 1996). The crystal entities in the

BMCD were entered into the database by extracting the appropriate crystallization conditions from the literature. Each crystal entry contains basic information on the macromolecule crystallized, the crystallization methods, the identities and concentrations of the chemical precipitants and additives used, and basic diffraction information.

Several papers described quantitative analysis of the BMCD. Samudzi and coworkers measured the distribution of various physical parameters of the proteins crystallized and found that (1) most macromolecular molecular weights were less than 100 kDa, (2) most macromolecules crystallized at concentrations of 0.01 to 2.0 mM, and (3) the most common pH for crystallization was approximately 7.0 (Samudzi *et al.*, 1992). The cluster analysis of the BMCD (a method that locates "natural" groupings of data within the given parameter space) showed that the data partitioned most significantly into eight groups, which showed significant asymmetry with respect to macromolecule type, crystallization method, and precipitant chemical(s) used (Samudzi *et al.*, 1992). At the time this analysis was performed (1992), there were 1025 crystal conditions in the BMCD. The same authors repeated these analyses in 1998 (Farr *et al.*, 1998) using a later edition of the BMCD, which contained more crystallization conditions (approximately 2300) and was significantly more complete (>90%). After cluster analysis, the data best fit into 25 categories, and the authors used this information to generate a list of general recommendations for crystallization experiment design based upon the properties of the macromolecule (or complex) to be screened (Farr *et al.*, 1998).

The analysis of Hennessy and coworkers took a different approach, as they divided the BMCD macromolecules into a taxonomic hierarchy and applied a Bayesian method to generate a probability of crystallization for a given set of parameters (hierarchical class, pH, temperature, component concentrations, etc.), which can be used to weight or bias the components of the basis set in the design of random screens (Hennessy *et al.*, 2000). Formal Bayesian analysis requires data from crystallization failures as well as successes. Because the BMCD does not contain information on failed crystallization experiments, Hennessy and coworkers approximate this by considering diffraction resolutions lower than 2.5-3.5 Å as failure (Hennessy *et al.*, 2000). 25% of PDB deposits have resolutions of 2.5 Å or lower.

The crystallization database system XtalBase (Meining, 2006) was used to import BMCD data. Each component of the crystallization solutions in the BMCD were imported into bins of a given range of concentration or pH, and a statistical measure was calculated that approximated the probability of crystallization given each possible combination of a pair of components (Meining, 2006).

Now, in the adolescence of high-throughput crystallization efforts (structural genomics), large groups are beginning to collect large sets of data on crystallization that include failed experiments as well as the successful ones. Some of these sets of data are large enough for broad-scale analysis through “data mining”. The results of these broad-scale analyses have focused on determining the most effective solutions in given pre-mixed crystallization reagent screens, and properties of macromolecules that best serve as predictors of ability to crystallize. It is interesting to note that the rates of progression from purified proteins to crystallization and from crystallization to diffraction are similar for different structural genomics centers (O'Toole *et al.*, 2004).

The Joint Center for Structural Genomics (JCSG) reported several results from a multi-tiered crystallization screen of the proteome of *Thermatoga maritima*. In the first tier, 539 proteins were successfully expressed and purified (out of a predicted 1877 ORFs), and initial crystallization conditions were identified for 465, using a set of 480 conditions from a set of 10 commercially available screens: Wizard I/II, and Cryo I/II from Emerald Biosciences, and Crystal Screen I/II/Cryo, PEG/Ion Screen, Grid Screen Ammonium Sulfate/PEG 6000/MPD and Grid Screen PEG/LiCl from Hampton Research (Lesley *et al.*, 2002). This initial screen determined that many proteins crystallized in multiple conditions, as half of the proteins screened crystallized in 5 or more conditions, and four crystallized in greater than 100 (Page *et al.*, 2003). A core set of 67 screen conditions from the 480 were identified that were capable of crystallizing 99% of the targets that produced crystals (Page *et al.*, 2003). The results of these experiments were eventually used to produce a new premixed sparse-matrix screen (Newman *et al.*, 2005).

A similar analysis was performed by the Aled Edwards group at the University of Toronto (Kimber *et al.*, 2003). A set of 755 different bacterial proteins were screened with the standard 48-condition sparse-matrix Crystal Screen I (Jancarik & Kim, 1991), and of them, crystals were observed for 45% of the targets (Kimber *et al.*, 2003). By clustering the results, it was determined that 24 of the 48 conditions were capable of crystallizing 94% of all of the proteins that crystallized, and 6 of the conditions were capable of crystallizing 60% (Kimber *et al.*, 2003).

The JCSG data set was also analyzed in terms of the properties of the proteins that were successfully crystallized, as measured by the relative rate of success, defined as the number of proteins that crystallized as a function of all proteins in the proteome (Canaves *et al.*, 2004). For a number of protein parameters, there were sharp drop-offs where the relative rate of success dipped below 15%, resulting in ranges of protein parameters where crystallization was more probable. Polypeptides with 80 to 560 amino acid residues, isoelectric points between 4.5 and 9.5, and a percentage of charged residues below 45% were significantly more likely to crystallize (Canaves *et al.*, 2004).

Given the pace and means by which crystallization trial information is currently generated, using at least semi-automatic systems for tracking such data is essential (Lorber, 2001). In the high-throughput structural biology environment, the traditional written notebook- or spreadsheet-based approach is impractical. The computerized, database-driven laboratory information management systems (LIMS) to store, annotate, and analyze the data developed mainly for high throughput laboratories can also be used to analyze the data from small scale crystallization experiments.

Some of the earliest efforts for using computers to annotate protein crystallization experiments described in the literature used spreadsheet applications (Hannick *et al.*, 1992; Hassell *et al.*, 1994). Spreadsheets are still used fairly widely today, particularly for optimizations, due to the simplicity and flexibility of their interface. However, this same flexibility makes searching or mining spreadsheets for information difficult. While the data are in computer-readable form, they are not structured and thus their syntax and semantics are not defined. An ideal system for storing protein crystallization information must allow the researcher a great deal of flexibility in preparing and analyzing crystallization experiments but yet store the data of those experiments in a highly structured way to make them amenable to broad-scale searching and analysis. One critical flaw in most LIMS systems is an over-reliance on user input, coupled with insufficient data analysis tools.

Relational database management systems or RDBMS (Codd, 1960) provide a flexible and efficient mechanism for storing many data and the complex relationships between them, and have become the *de facto* standard for modern database systems. Data in RDBMS are manipulated and searched using the Structured Query Language (SQL).

In contrast, there are two fundamental approaches to LIMS interface design. One uses web interfaces, with Internet browsers as the platform for the client for inputting, reporting and sometimes managing data. Such interfaces can be easily installed and distributed, since the client requires little more software than a standard Web browser. However, such interfaces are limited in complexity by the constraints of Hypertext Markup Language (HTML) and Javascript, though the rapidly-maturing AJAX (asynchronous JavaScript with XML) set of technologies do help address these problems to some degree. The vast majority of LIMS designed for structural biology use a web-based approach (Bertone *et al.*, 2001; Haebel *et al.*, 2001; Harris & Jones, 2002; Manjasetty *et al.*, 2003; Goh *et al.*, 2003; Morris *et al.*, 2005; Prilusky *et al.*, 2005; Meining, 2006). However, this type of interface makes communication with hardware attached to the client machine very difficult.

Another approach is to develop stand-alone programs for the native platform of the client computer. Such clients are more difficult to install and to keep up-to-date, since upgrades of the client must be distributed to each computer. The development time and effort for such systems can also be longer, or may be restricted to particular architectures. Java-based clients solve these problems to some extent, by distributing the code through a web browser, at the expense of ensuring a Java run-time environment is installed with each browser (Elkin & Hogle, 2001; Zolnai *et al.*, 2003; Fulton *et al.*, 2004; Amin *et al.*, 2006). However, the resulting systems can be much more sophisticated and user-friendly, and are capable of extensive communication with external hardware.

2 The Xtaldb protein crystallization expert system

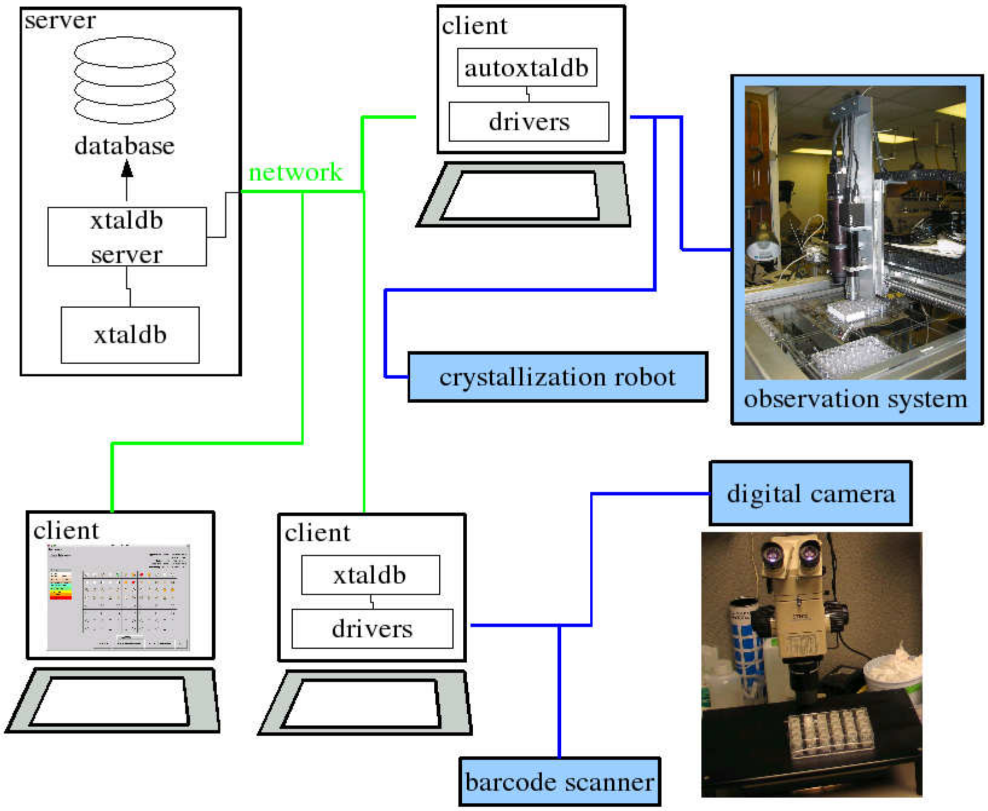
2.1 Crystal experiment design and tracking

Conclusions based on analyses of crystallization information are highly dependant on the set of experimental data used to generate those conclusions. Due to the sheer number of parameters involved in protein crystallization, concerns have been raised about the validity of applying crystallization knowledge obtained through data-mining from one research group to another (Rupp, 2003). Each research group may use different equipment and protocols for protein expression and purification. The difference in source material may dramatically affect crystallization, and ultimately, data from one's own laboratory is the most reliable source of information for data-mining to design the most successful crystallization experiments.

Additionally, as discussed in Chapter 1, the typical kinds of crystallization experiments designed by hand, such as grid screens, are relatively inefficient in terms of searching crystallization space. Some tools exist to generate more efficient random screens, but the screens generated by most of these tools are not suitable for quantitative statistical analysis by linear regression models (Carter & Yin, 1994; Carter, 1999), and as a consequence, such methods have never been frequently used. To design crystallization experiments for statistical analysis, collect detailed information about crystallization experiments, and generate detailed analyses of the results of those experiments, the Xtaldb system has been developed.

Xtaldb is a client-server based system, where multiple clients through a network connect to a central server, as illustrated in Fig. 2-1. The central server contains a

Figure 2-1: Overall architecture of the Xtaldb system.



database and an image server that store all of the experimental information collected by each of the clients. Each client provides the interface for designing plates, observing drops, running searches and analyses, and communicates with the hardware for semi-automated or automated data acquisition. Because of this client-server architecture, multiple experimenters in multiple research groups can conduct and track crystallization experiments simultaneously and access data collected by others. The system uses a permissions-based security model to prevent one experimenter from modifying crystallization data entered by another, or from accessing data collected by another research group. Xtaldb is a component of the LabDB laboratory information management system (LIMS; see Section 2.5), which is designed to keep track of an experimental pipeline from cloning to structure solution.

Xtaldb is the first system that allows not only detailed analysis of all crystallization parameters but also takes into account that chemicals and solutions can change with time due to evaporation, oxidation, etc. The stock solutions and preparations of macromolecules used to prepare a crystallization plate represent physical entities. For example, if a stock solution of 4 M NaCl was used to prepare crystallization wells, the system tracks when that stock was prepared, who prepared it, and the specific reagent bottle and lot of NaCl the stock was prepared from. These solutions and protein preparations can be tracked back to their sources, either the purification process used to generate the protein or the lot of chemical used to prepare the stock solutions. Two modules of the LabDB LIMS system, described in Section 2.5, interface with laboratory equipment to collect this information in a semi-automatic way.

In practice, crystallization experiments are performed by designing plates that contain mixtures of protein and solution, using the techniques described in Section 1.2. In Xtaldb, there are two modes for designing experiments. The first is a manual mode with a spreadsheet-like interface. The plate is prepared by selecting each stock solution, then selecting the wells to which that stock will be added, at a set concentration. The system has sophisticated validity checks. For example, the client checks if a requested concentration is too high to be produced by dilution of the corresponding stock solution and if necessary reports the error. The system then uses the concentrations of the stock solutions to generate a pipetting guide, calculating the volumes of each of the stock solutions and of water needed to mix the wells and drops specified. A plate layout may be saved to a local file in XML format and reused on other plates. Finally, a barcode may be assigned to a plate for quick subsequent identification.

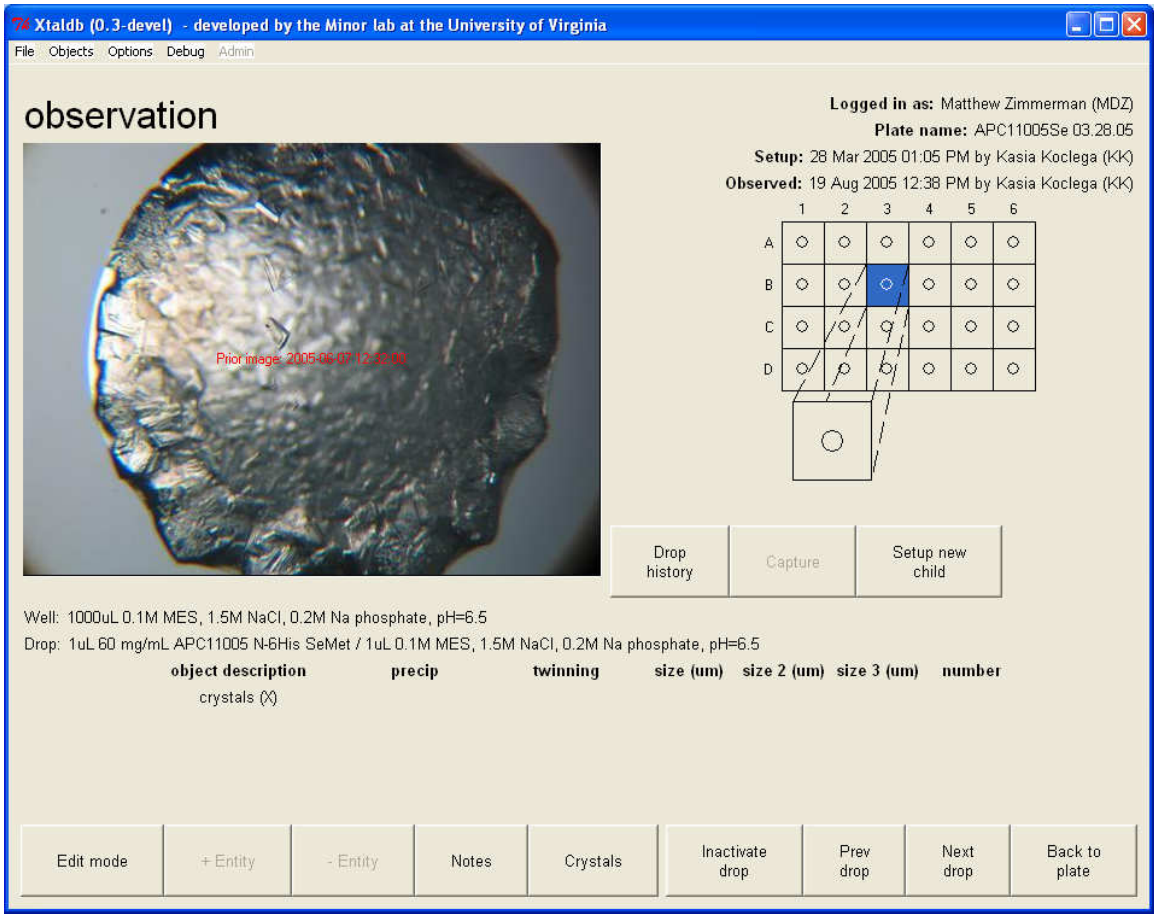
The spreadsheet-like client interface is designed to be very flexible, to permit as many different types of crystallization experiments as possible. The system handles both 24- and 96-well plate formats, can handle multiple crystallization methods (vapor-diffusion and batch), and can handle an arbitrary number of drops per well, up to 9. Each individual drop of a plate to be designed can be treated independently of the others.

The system also keeps track of a library of pre-mixed commercial and custom crystallization screens. Many commercial crystallization screens are already entered into the database, and a tool is provided for importing new commercial or custom pre-mixed screen into the system from comma-separated values (CSV) text or Excel spreadsheet files. A component of the spreadsheet-like manual interface allows conditions from a pre-mixed screen to be added to a range of wells and drops.

The second mode for designing plates in Xtaldb uses a series of “wizards” to simplify preparing standard premixed screens or balanced random screens on a set of plates. In the premixed screen wizard, some parameters like the screen, plate type, and well volume are selected, and the common layout of drops to be set in each well is chosen. As in the manual mode, the premixed screen wizard verifies the information entered and then generates the appropriate number of plates for all of the conditions in the premixed crystallization screen.

The wizard for generating random balanced screens is similar. Along with the number of drops and the name and type of the plates, the parameters (factors) to be varied and the levels for each parameter are selected. Parameters to be chosen are not limited to reagents and their concentrations, but may also include other parameters of the

Figure 2-2: The drop observation window.



Logged in as: Matthew Zimmerman (MDZ)
Plate name: APC11005Se 03.28.05
Setup: 28 Mar 2005 01:05 PM by Kasia Koclega (KK)
Observed: 19 Aug 2005 12:38 PM by Kasia Koclega (KK)

Well: 1000uL 0.1M MES, 1.5M NaCl, 0.2M Na phosphate, pH=6.5

Drop: 1uL 60 mg/mL APC11005 N-6His SeMet / 1uL 0.1M MES, 1.5M NaCl, 0.2M Na phosphate, pH=6.5

object description	precip	twinning	size (um)	size 2 (um)	size 3 (um)	number
crystals (X)						

experiment: the macromolecule concentration, pH, PEG molecular weight, and drop volumes or ratios or both. As before, the system validates the input. If it is valid, the Xtaldb client generates the balanced random screen on one or more plates and generates a pipetting guide for each plate.

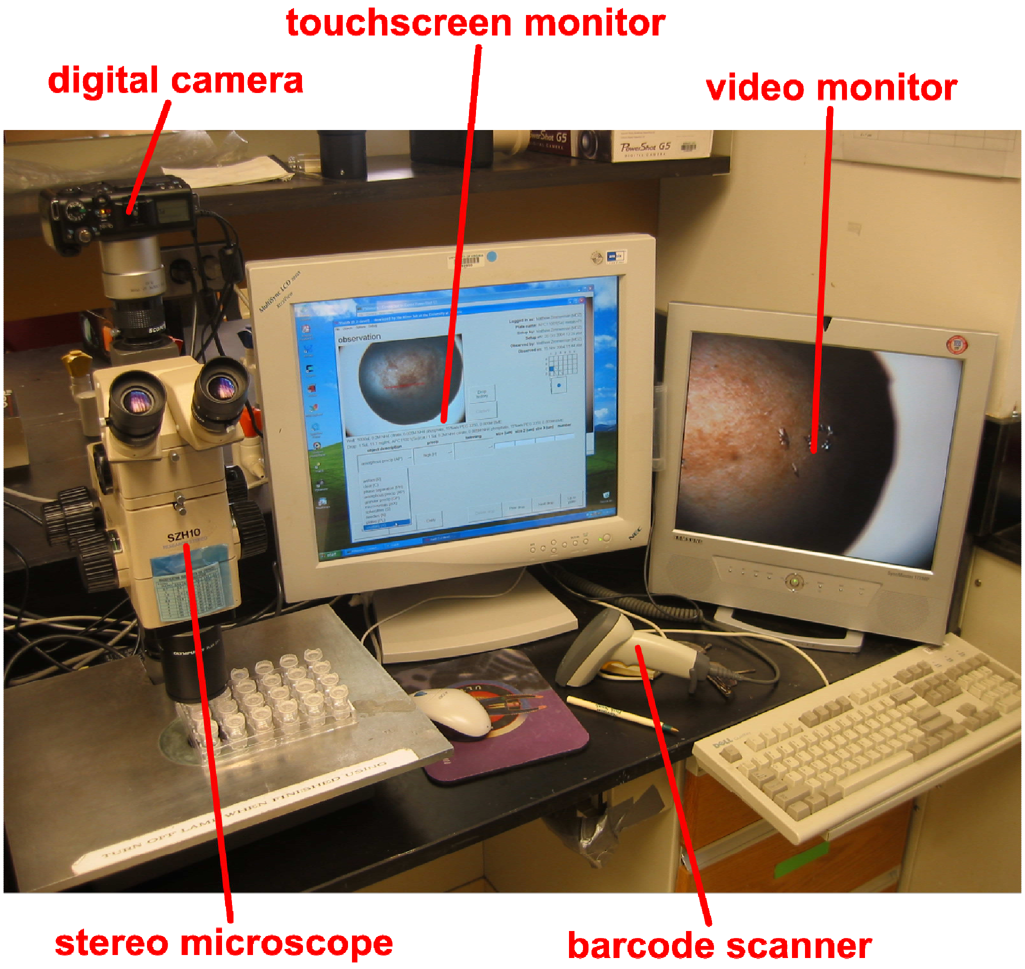
Xtaldb utilizes an algorithm for generating balanced experimental designs recently described by Xu (Xu, 2002), though the program’s implementation has been modified to permit numbers of levels not evenly divisible into the number of runs. The algorithm builds the experimental design one column (factor) at a time, and iteratively swaps elements in that column until a statistical optimality criterion is maximized (Xu, 2002). The Xtaldb implementation is faster and more robust than any other reported program for generating balanced random screens for crystallization (see Section 2.4 for more details), and is written in the C programming language for purposes of speed.

Plates may be edited after they are created, using the manual spreadsheet-like interface, even if they are created with one of the wizards. If a reagent is spilled or a drop is otherwise damaged during the process of setting up a plate, the edit utility allows modification of the plate after it has been entered into the database to reflect the unexpected changes. Once a plate has been observed, however, no more changes can be made to the plate layout. A general system for importing plate designs and pre-made screens in text, CSV, and XML formats is also under development.

Once a plate has been set up, it is observed one or more times. Fig. 2-2 shows the window for manually annotating the observation of a drop. By clicking a button, an image of the drop is automatically captured. The contents of drops are annotated manually as consisting of one or more entities, which range from precipitation to crystals, by selecting down the appropriate description in a pull-down box. More than one entity may be set for the same drop. Additionally, other orthogonal characteristics of the entities may be described, such as amount of precipitation or crystal size.

The system accesses various types of equipment through modules that utilize abstract programming interfaces (API). In the prototype installation of the system at the University of Virginia, several pieces of hardware were connected to a client system to automate certain aspects of data acquisition. Figure 2-3 shows a layout of a client system installed in the Rastinejad/Minor laboratory at the University of Virginia. A barcode scanner is used to authenticate and identify experimenters and quickly access particular

Figure 2-3: The layout of a typical installation of Xtaldb.



plates, which are labeled with unique barcodes. A touch screen monitor, located close to the stereo microscope, is used to quickly annotate observations. A digital camera mounted on the microscope records images interactively, and the current drop image on the monitor may be compared directly with previous images. The library for communicating with digital cameras is modular and allows multiple types of digital cameras to be used. In this prototype workstation, a Canon Powershot G5 camera is mounted on an Olympus SZX-10 stereo microscope (Fig. 2-3).

A modified version of the Xtaldb client called Autoxtaldb is provided for communication with various pieces of laboratory automation. Autoxtaldb provides a noninteractive, command-line-based API to the expert system that may be utilized by modular drivers to communicate with other forms of automation. Additionally, to facilitate communication with automation and other data management systems, Xtaldb provides capabilities for exporting data into different formats. The system can export information in CSV or XML, and through add-on modules, is capable of generating command scripts to drive pipetting robots (such as TTP Labtech's Mosquito system).

When crystals are harvested from crystallization drops for X-ray diffraction analysis, an interface is provided to record the harvested crystals in the database. Each crystal is identified by name, and information stored about the crystals can include information about the size, mounting method, cryo-protectant solution, harvest and diffraction time, who performed the experiments, and free-form notes. These crystals harvested can be interfaced with the HKLdb interface of HKL-3000 (Minor *et al*., 2006), which takes the list produced with Xtaldb and associates information about data collection, indexing, scaling, and structure solution with each crystal in the database. Diffraction information may also be entered by hand in situations where HKLdb is not present. Thus in terms of analysis, information about the diffraction characteristics of crystals may also be used in analyzing crystallization trials, by identifying conditions that produce visually good but poorly diffracting crystal forms.

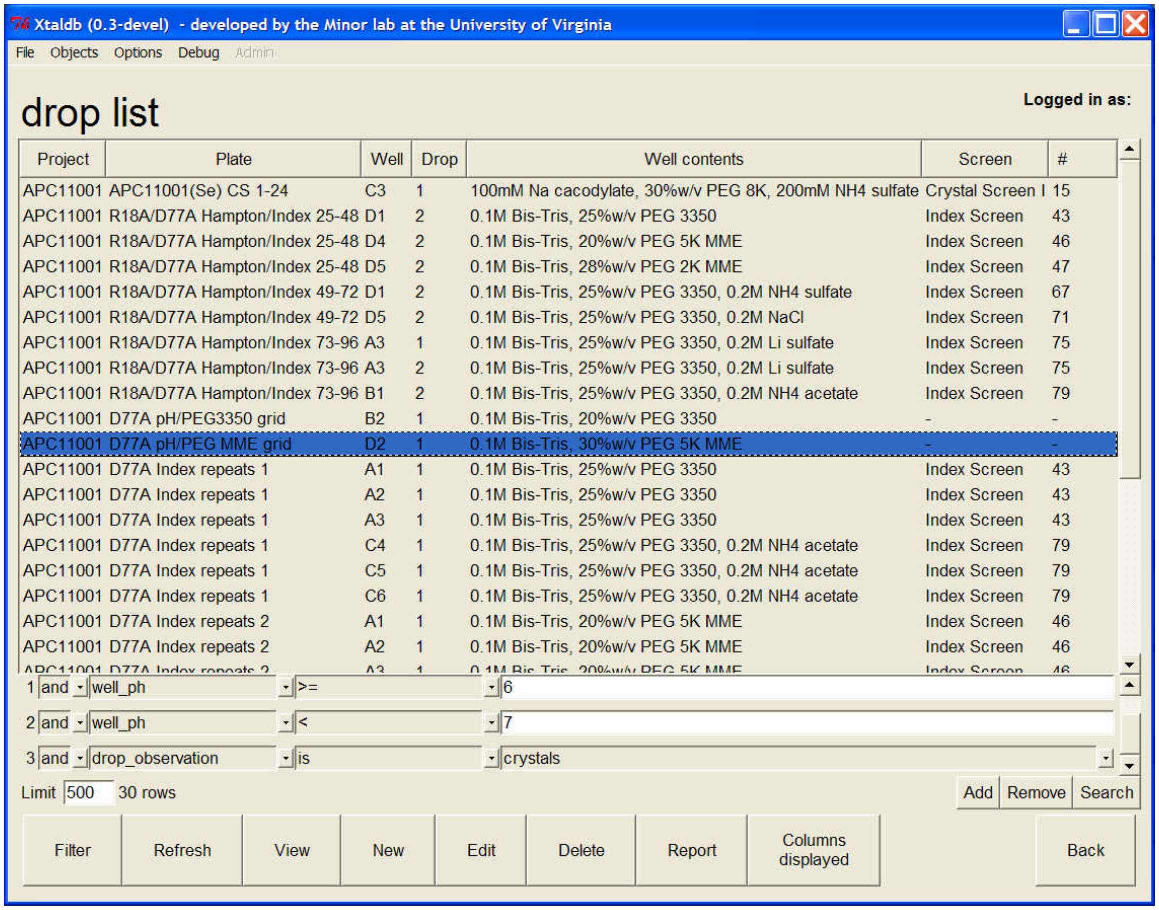
2.2 Data analysis

There are two primary tools in Xtaldb for analyzing the database for information: a search dialog and a data mining console. In both tools, the search criteria entered into the system are translated into SQL queries that are passed to the database system, though neither requires detailed knowledge of SQL.

The first tool is a dialog for searching all plates, drops, and crystals, which is shown in Fig. 2-4. Search conditions, such as project name, plate name, well contents, drop observation, etc., are selected, and can be combined by the Boolean operators “AND” and “OR” to build up more complex searches. Virtually all of the recorded parameters of the plates, drops, and crystals can be used as search conditions. Those plates, drops, or crystals that match the search criteria are returned in a columnar list, which can be sorted by each column. Clicking on a particular search result will link the user to the page describing that result. This dialog does not require any knowledge of the SQL language, as the query is completely generated by the inputs to the interface. However, the results reported are limited to simple lists of data that match the given specified condition.

The second tool for analysis is the data mining console, which is illustrated in Fig. 2-5. On the console, the SQL queries are specified directly, though a large set of editable SQL templates are available. Using the templates does not require deep knowledge of SQL, just some common sense. This tool gives almost full flexibility, as much more complex analyses of the data can be generated as it is possible to construct any SQL query. However, in that case, more detailed knowledge of the SQL language is required.

Figure 2-4: The drop search dialog.



drop list

Logged in as:

Project	Plate	Well	Drop	Well contents	Screen	#
APC11001	APC11001(Se) CS 1-24	C3	1	100mM Na cacodylate, 30%w/v PEG 8K, 200mM NH4 sulfate	Crystal Screen I	15
APC11001	R18A/D77A Hampton/Index 25-48	D1	2	0.1M Bis-Tris, 25%w/v PEG 3350	Index Screen	43
APC11001	R18A/D77A Hampton/Index 25-48	D4	2	0.1M Bis-Tris, 20%w/v PEG 5K MME	Index Screen	46
APC11001	R18A/D77A Hampton/Index 25-48	D5	2	0.1M Bis-Tris, 28%w/v PEG 2K MME	Index Screen	47
APC11001	R18A/D77A Hampton/Index 49-72	D1	2	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M NH4 sulfate	Index Screen	67
APC11001	R18A/D77A Hampton/Index 49-72	D5	2	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M NaCl	Index Screen	71
APC11001	R18A/D77A Hampton/Index 73-96	A3	1	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M Li sulfate	Index Screen	75
APC11001	R18A/D77A Hampton/Index 73-96	A3	2	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M Li sulfate	Index Screen	75
APC11001	R18A/D77A Hampton/Index 73-96	B1	2	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M NH4 acetate	Index Screen	79
APC11001	D77A pH/PEG3350 grid	B2	1	0.1M Bis-Tris, 20%w/v PEG 3350	-	-
APC11001	D77A pH/PEG MME grid	D2	1	0.1M Bis-Tris, 30%w/v PEG 5K MME	-	-
APC11001	D77A Index repeats 1	A1	1	0.1M Bis-Tris, 25%w/v PEG 3350	Index Screen	43
APC11001	D77A Index repeats 1	A2	1	0.1M Bis-Tris, 25%w/v PEG 3350	Index Screen	43
APC11001	D77A Index repeats 1	A3	1	0.1M Bis-Tris, 25%w/v PEG 3350	Index Screen	43
APC11001	D77A Index repeats 1	C4	1	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M NH4 acetate	Index Screen	79
APC11001	D77A Index repeats 1	C5	1	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M NH4 acetate	Index Screen	79
APC11001	D77A Index repeats 1	C6	1	0.1M Bis-Tris, 25%w/v PEG 3350, 0.2M NH4 acetate	Index Screen	79
APC11001	D77A Index repeats 2	A1	1	0.1M Bis-Tris, 20%w/v PEG 5K MME	Index Screen	46
APC11001	D77A Index repeats 2	A2	1	0.1M Bis-Tris, 20%w/v PEG 5K MME	Index Screen	46
APC11001	D77A Index repeats 2	A3	1	0.1M Bis-Tris, 20%w/v PEG 5K MME	Index Screen	46

1 and well_ph >= 6

2 and well_ph < 7

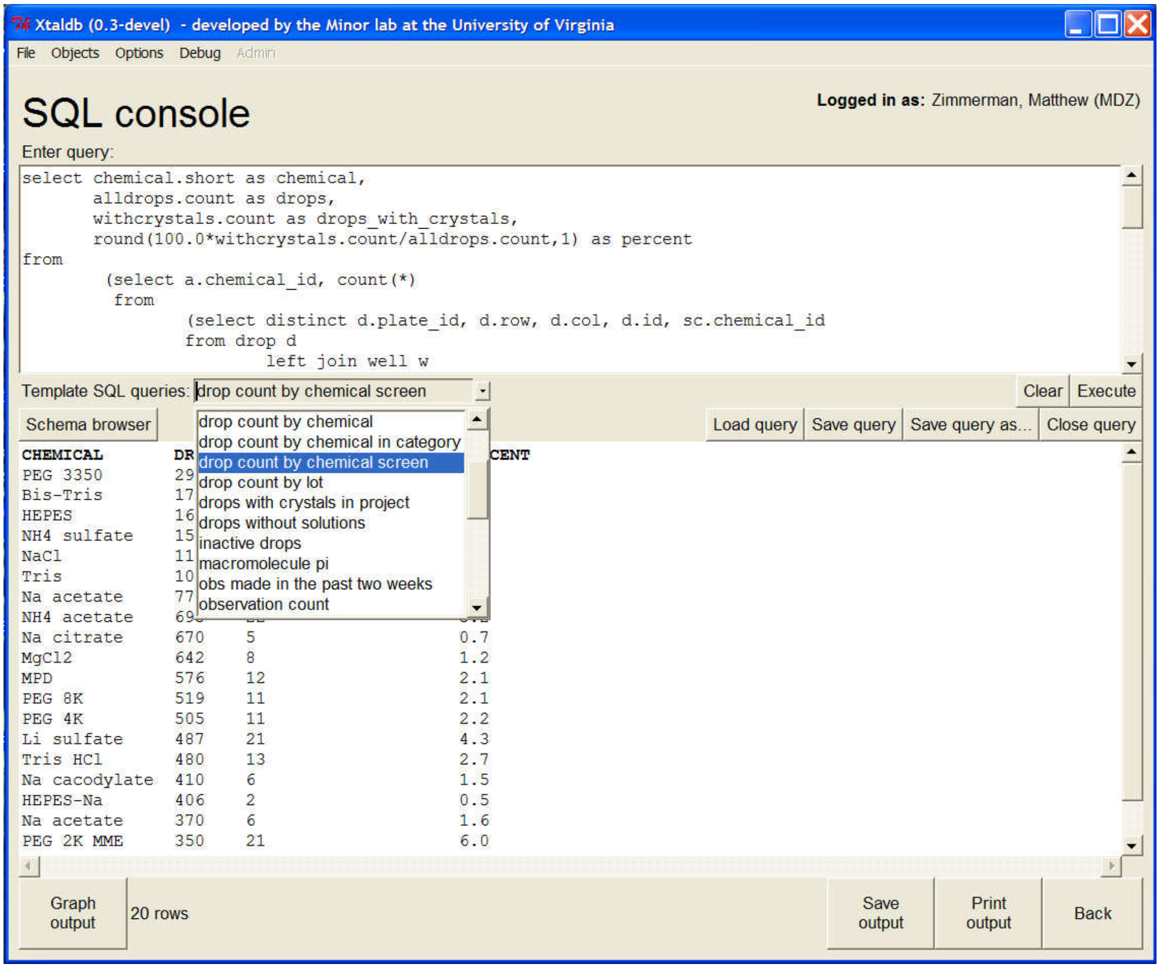
3 and drop_observation is crystals

Limit 500 30 rows

Fig. 2-5 shows an example template query, which generates the list of most frequently used chemicals, counts of all of the drops that contain that chemical with observed crystals, and divides that number by the total number of drops that contain that chemical to measure the success rate by reagent. Similar templates are provided to count the relative success for each commercial screen component, to list the plates and observations created in the past two weeks, to generate a list of crystallization pH as a function of pI, and to perform many other tasks.

Information found using the data mining console of Xtaldb may be presented in graphical form by automatic generation of bar and scatter graphs. This is implemented through an interface to the Gnuplot plotting package. Data from random screen

Figure 2-5: The data analysis window.



Xtaldb (0.3-devel) - developed by the Minor lab at the University of Virginia
File
Objects
Options
Debug
Admin
SQL console
Logged in as: Zimmerman, Matthew (MDZ)
Enter query:
Template SQL queries: drop count by chemical screen
drop count by chemical
drop count by chemical in category
drop count by chemical screen
drop count by lot
drops with crystals in project
drops without solutions
inactive drops
macromolecule pi
obs made in the past two weeks
observation count
Clear
Execute
Schema browser
Load query
Save query
Save query as...
Close query
Graph output
20 rows
Save output
Print output
Back

```
select chemical.short as chemical,
       alldrops.count as drops,
       withcrystals.count as drops_with_crystals,
       round(100.0*withcrystals.count/alldrops.count,1) as percent
from
        (select a.chemical_id, count(*)
         from
                (select distinct d.plate_id, d.row, d.col, d.id, sc.chemical_id
                from drop d
                        left join well w
```

```
CHEMICAL        DR                            CENT
PEG 3350        29
Bis-Tris        17
HEPES           16
NH4 sulfate     15
NaCl            11
Tris            10
Na acetate      77
NH4 acetate     69
Na citrate      670    5                      0.7
MgCl2           642    8                      1.2
MPD             576    12                     2.1
PEG 8K          519    11                     2.1
PEG 4K          505    11                     2.2
Li sulfate      487    21                     4.3
Tris HCl        480    13                     2.7
Na cacodylate   410    6                      1.5
HEPES-Na        406    2                      0.5
Na acetate      370    6                      1.6
PEG 2K MME      350    21                     6.0
```

experiments can be exported in text-based formats (raw text, CSV, and XML) that may be imported into other statistical analysis programs.

The ultimate criterion of success for a crystallization experiment is structure elucidation using the crystal produced by that experiment. However, such a criterion is not very well suited for purposes of data mining, as once such a crystallization experiment is obtained, no further data analysis is needed (at least for that project!). Analysis of crystallization experiments requires introducing multiple criteria that may be used to identify the success of a given crystallization experiment. Drops may be visually annotated as containing crystals, the drop may be assigned a numerical quality score, or more detailed information about the observed crystals in drops may be listed, such as size, shape, birefringence, etc. Crystals may be harvested from a drop, and diffraction information, such as diffraction limits, mosaicity, etc., may be measured. The numerical quality score as a simplest approximation may serve at first glance to be a reasonable measure of the overall quality of the drop, but it is inherently subjective, and if it is recorded at the time that the crystallization drop is observed, may not reflect the quality of crystals that are harvested and shot. Other measures, such as more detailed parameters of crystal morphology and diffraction behavior are better in terms of giving a more detailed description but are more difficult to represent and search in the database, and may be less likely to be consistently recorded. In reality, all of these parameters should be recorded and considered when doing analysis of crystallization experiments.

2.3 Implementation details

The Xtaldb client is a standalone, modular, and object-oriented Perl/Tk program. The client itself is cross-platform, running on Windows, Linux, and Mac OSX, though some of the hardware driver modules are operating system dependant. Binary distributions of the client contain a Perl interpreter and all of the libraries to run the program without installing additional software. The Gnuplot graph-plotting program is optional, and only required for use of the graphing features as described above. The server depends on two open-source programs; the relational database server PostgreSQL (version 7.3 or later) and Apache HTTPD (version 2 or later). The server itself is composed of a database server, an image server, and a small set of accessory scripts. The prototype installation of the server is installed on Linux Fedora Core 6, with PostgreSQL 8.1 and Apache 2.2, but server installations have also been successfully tested on other Linux distributions and on Windows XP.

As described in Chapter 1, relatively few crystallization database systems use a native client interface. While this in some cases makes the process of updating the software more difficult, the choice of using a native program allows the use of more extensive hardware applications and the use of native software APIs for hardware applications. As an example, Xtaldb can directly interact with a Canon digital camera for direct image capture and upload to the database, which would be difficult to impossible for web-based interfaces. The use of a native client also permits complicated calculations to be performed on the client side, reducing the computational load on the server. This simplifies considerably the processor requirements for the system's server, which is only required to have the network bandwidth to serve many clients and disk space to hold the data and images.

2.4 Balanced random screen design algorithms

As described in Chapter 1, three other programs have been described that generate balanced random screen designs for crystallization: INFAC, DESIGN, and SAmBA. To compare the performance of the balanced random screen design algorithm implemented in Xtaldb to these programs, identical sets of factors and levels were passed to all four programs, and the resulting designs were compared.

To measure to the degree of “balance” or efficiency of each generated design, the measures D and A_2 were used (Xu, 2002). The D parameter measures the efficiency of an experimental design by taking the determinant of the correlation matrix for the model matrix, weighted by the number of degrees of freedom. The value of D for any design is $D \leq 1$, where values closer to 1 represent a design with fewer unbalanced factors or binary combinations of factors. $D = 1$ if and only if the design is an orthogonal array. The A_2 criterion measures the sum of squares of the off-diagonal elements of the model correlation matrix, so that $A_2 = 0$ if and only if the design is an orthogonal array, and the value increases as the design becomes more imbalanced.

Using these measures, identical designs generated by INFAC, DESIGN, and Xtaldb are shown in Table 2-1. For virtually all of the designs tested, Xtaldb produced more efficient experimental designs than the other two programs. In addition, Xtaldb generated consistently efficient designs regardless of the number of factors or runs, while the designs made by INFAC or DESIGN are much more inconsistent in their level of efficiency. This suggests that in addition to being less efficient, both programs employ less robust algorithms.

Table 2-1: Comparison of Xtaldb versus INFAC and DESIGN.

Balanced screens generated by the algorithm implemented in Xtaldb are compared to experimental designs published in the literature generated by the programs INFAC and DESIGN. Source or executable code for both programs are no longer available online. N is the number of runs in each experiment, and levels shows the number of levels for each of the factors of designs. For example, (4,3,3) describes an experiment with three factors with 4, 3, and 3 levels respectively. D and A_2 are calculated as described in the text. References: 1-(Abergel et al., 1991), 2-(Audic et al., 1997), 3-(Carter et al., 1988), 4-(Sedzik, 1994).

				Published design		*Xtaldb*	
N	*Levels*	*Program*	*Ref.*	*D*	A_2	*D*	A_2
12	(4,3,3)	INFAC	1	0.982	0.125	0.982	0.125
12	(3,3,3,2)	INFAC	1	0.912	0.542	0.942	0.375
16	(4,4,3,2)	INFAC	2	0.956	0.386	0.969	0.261
20	(3,3,3,3,3,2)	INFAC	3	0.748	2.280	0.972	0.302
24	(3,3,3,3,2,2)	INFAC	1	0.840	1.272	0.981	0.188
24	(4,3,3,2,2,2)	INFAC	1	0.880	1.101	0.997	0.031
24	(6,3,3,2,2,2)	INFAC	1	0.753	2.472	0.976	0.281
36	(9,4,4,4,3,3,2)	DESIGN	4	0.836	2.780	0.980	0.435
48	(6,6,5,2,2)	INFAC	1	0.929	1.002	0.979	0.318
48	(6,4,3,3,2,2)	INFAC	1	0.930	0.941	0.995	0.070
88	(11,8,8,8,4,4,4,4,2)	DESIGN	4	0.736	9.690	0.977	0.985

The third balanced random screen program, SAmBA, was compared to Xtaldb using the same *D* and A_2 measures for a set of experimental designs as shown in Table 2-2. SAmBA implements two algorithms for generating experimental designs, one using simulated annealing and one using a backtracking algorithm. SAmBA produced experimental designs that are as efficient, or in some cases slightly more efficient, than those of Xtaldb. However, the SAmBA program had some serious issues. Beyond trivial experimental designs, the backtracking algorithm of SAmBA was far too slow for practical use. The simulated annealing algorithm performed much more quickly, but was not robust. If it was unable to find a suitably minimal design for a given number of experiments within a certain number of cycles, it added a run and repeated the process. This resulted in experimental designs with 10 to 100 more runs than requested, and

Table 2-2: Comparison of Xtaldb versus SAmBA.

D and A_2 are calculated as in the text. Time measures the time for a typical execution of the C version of SAmBA (with the simulated annealing algorithm, using code downloaded from http://www.igs.cnrs-mrs.fr/samba/), and the C executable implementing screen design in Xtaldb. Both programs were compiled and run on a 3.2 MHz Intel Pentium 4 system with 1 GB RAM, running Fedora Core Linux 6.

[a-d] The designs output by SAmBA for these parameters had 36, 43, 53, and 139 runs respectively.

[e] The SAmBA algorithm did not converge within an extended period of time (> 6 hours).

		SAmBA			*Xtaldb*		
N	*Levels*	*D*	*A_2*	*Time*	*D*	*A_2*	*Time*
12	(4,3,3)	0.982	0.124	<0.1 s	0.982	0.125	<0.1 s
12	(3,3,3,2)	0.953	0.375	<0.1 s	0.942	0.375	<0.1 s
16	(4,4,3,2)	0.969	0.261	<0.1 s	0.969	0.261	<0.1 s
20	(4,4,3,3,2)	0.954	0.557	<0.1 s	0.943	0.560	<0.1 s
20	(5,4,4,3,2)	0.963	0.474	0.2 s	0.961	0.486	<0.1 s
20	(3,3,3,3,3,2)	0.978	0.320	0.1 s	0.972	0.302	<0.1 s
24	(3,3,3,3,2,2)	0.980	0.188	0.1 s	0.981	0.188	<0.1 s
24	(4,3,3,2,2,2)	0.997	0.031	0.1 s	0.997	0.031	<0.1 s
24	(4,4,3,3,3,2)	0.984	0.205	18 s	0.983	0.205	<0.1 s
24[a]	(6,3,3,2,2,2)	1.000[a]	0.000[a]	0.2 s	0.976	0.281	<0.1 s
30[b]	(6,5,5,4,3)	0.972[b]	0.466[b]	5m 56s	0.967	0.560	0.2 s
36	(6,5,4,3,2)	0.977	0.336	0.2 s	0.976	0.336	0.2 s
36[c]	(6,5,5,4,3,2)	0.983[c]	0.299[c]	6m 32s	0.963	0.660	0.4 s
36[d]	(9,4,4,4,3,3,2)	0.997[d]	0.054[d]	18m 57s	0.980	0.435	0.7 s
48	(6,6,5,2,2)	0.979	0.318	0.2 s	0.979	0.318	0.6 s
48	(8,6,4,3,2)	0.992	0.142	0.4 s	0.992	0.142	0.8 s
48	(6,4,3,3,2,2)	0.995	0.070	0.3 s	0.995	0.070	0.7 s
57	(6,6,5,4,2)	0.985	0.250	0.4 s	0.981	0.336	1.7 s
88	(11,8,8,8,4,4,4,4,2)	n/a[e]	n/a[e]	n/a[e]	0.977	0.985	32 s
168	(14,12,4,3,2)	0.999	0.031	8.4 s	0.999	0.031	1m 21s

computation times of several minutes (Table 2-2). It was difficult to predict for which sets of input such pathological behavior would be seen. For one particularly large input, SAmBA failed to produce a suitable design even after several hours of computation. In contrast, Xtaldb consistently produced experimental designs in seconds, and the computation time was roughly proportional to the number of factors and the number of experiments (Table 2-2).

2.5 The LabDB LIMS

Xtaldb is a component of a larger crystallization laboratory information management system (LIMS) called LabDB, which is currently being developed at the University of Virginia. The LabDB system consists of four components: Wetlab for chemicals, chemical bottles, and solutions; PepDB for protein cloning, expression and purification; Xtaldb for crystallization and crystal harvesting; and a plug-in module for HKL-3000 for diffraction data collection and structure solution called HKLdb (Figure 2-6). Each component interacts directly with appropriate laboratory hardware to eliminate human beings as much as possible from the process of data harvesting, and each is optimized for the task for which it is designed.

All four components store the data they use in a central database, and as a result share common organizational information, such as projects, laboratories, user accounts, passwords, and identification barcodes (Fig. 2-7). Information collected by one module is accessible within another. The overall system database is very large, containing more

Figure 2-6: Overall schematic of the LabDB LIMS system.

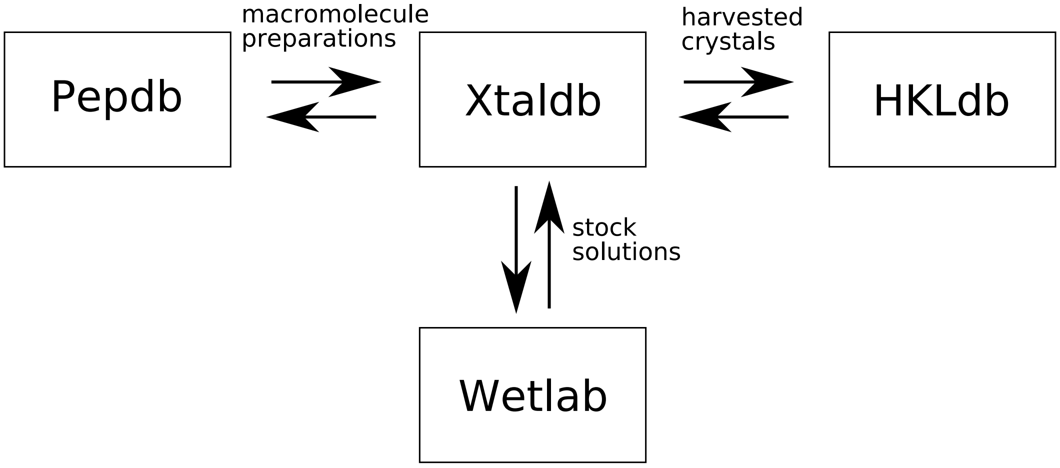
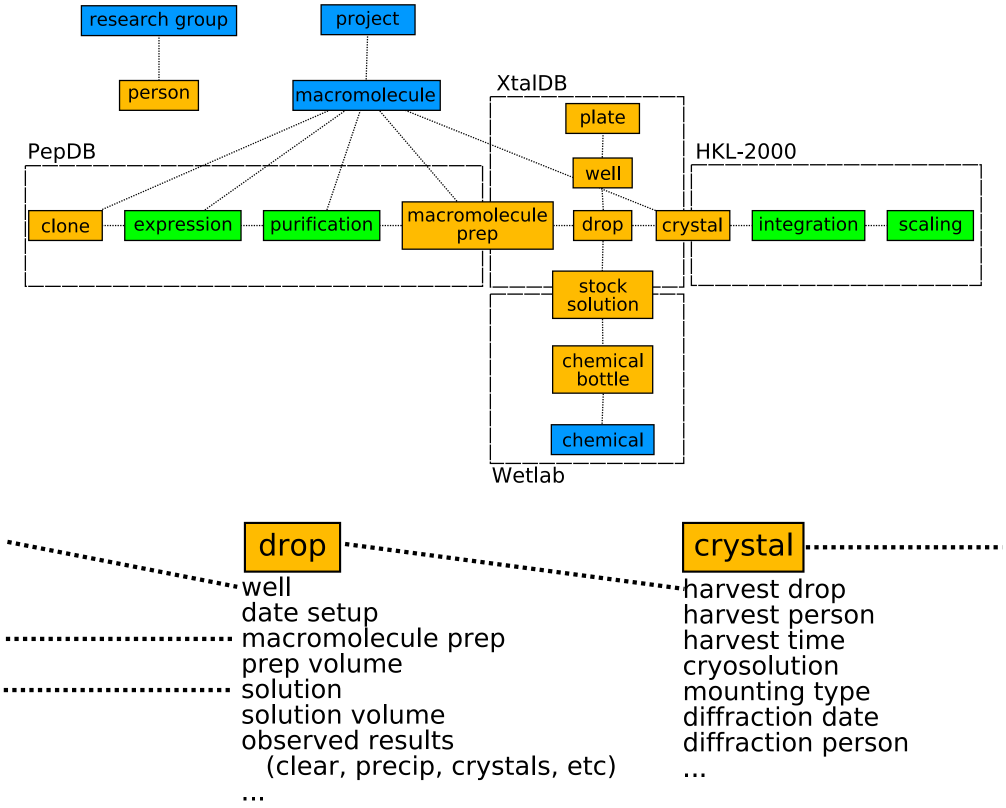


Figure 2-7: Simplified diagram of the database schema.

Each box represents a table with a list of objects, where each object is either a physical entity (orange), process (green), or concept (blue). The thin lines represent links between objects in the database. Each interface uses a specific set of tables in the database, labeled with dashed boxes. All interfaces make use of the research group, person, project, and macromolecule tables. Each object has a list of "attributes", which describe the different properties of each object.



than 70 tables separated into separate namespaces for each component.

The Wetlab component tracks laboratory chemicals, bottles, and solutions in the LabDB database. It maintains an inventory of chemicals and bottles, labeling with barcodes, using a hand-held personal digital assistant (PDA). It interfaces with standard commercial off-the-shelf laboratory hardware, and provides a mechanism to semi-automatically create stock solutions of reagents and set their pH. Each solution is labeled with a barcode and logged into the central database, as well as when and by whom the solution was made. Each stock solution is traceable back to the chemical bottle and lot from which it was made.

The Wetlab system consists of two parts. The first is a server application that communicates with the hardware, interfacing through the network to a Radiometer Analytical pH meter, a Mettler-Toledo balance, and a Zebra barcode printer through a port server. The code in the server that drives each piece of equipment is modular, and any hardware exporting data via the RS-232 serial protocol could be used. In addition, a WebDAQ unit drives a set of off-the-shelf analog fluid pumps that pump water, NaOH, HCl and acetic acid. The second is a Visual C++ client running on an iPAQ PDA outfitted with a Socket barcode scanner. The client communicates with the server through a wireless network. Researchers manage the chemical bottle and solution inventory, create solutions, and control the laboratory hardware with the PDA.

Solutions may be prepared to a set molarity, molality, or percentage weight/weight. The PDA is used to enter the concentration and volume desired, then scans the barcode of the reagent bottle. The system looks up the molecular weight of the compound in the database and calculates the amount of chemical to add. The experimenter weighs out the reagent on the balance until the calculated amount is reached. The system reads the true amount of reagent measured and adjusts the final volume of the solution accordingly, either by informing the user for molar solutions or by pumping the appropriate volume of water for molal or weight/weight solutions. Finally, a detailed label is printed with a unique barcode to identify the solutions. Buffer solutions may be set to the appropriate pH by invoking the pH suite with the PDA. The fluid pumps add concentrated acid or base as appropriate until the setpoint is reached. The system then records the actual pH in the database and prints out a modified label. When new chemical bottles arrive in the lab, the researcher inputs their information into the database with the PDA. The database records the current weight or volume of each reagent and automatically decrements the appropriate amount when a solution is made. Wetlab was developed by the author, Piotr Lasota, Wojtek Potrzebowski, and others.

Pepdb is a PHP-based component that stores information about protein cloning, expression, and purification. Researchers use a Web browser to input data into forms as each step is completed. Detailed information about each clone is stored. Pepdb imports data from an AKTA FPLC protein purification system, using an interface written by Heping Zheng.

HKLdb is a plug-in module for HKL-3000 (Minor *et al.*, 2006) that links harvested crystals tracked with the Xtaldb system to the diffraction, scaling, structure solution, and (limited) structure refinement data collected on that crystal. Like the other components, users log in upon starting the HKL-3000 program and select a project and crystal, and subsequent steps carried out in the program are recorded in the database. HKLdb is being developed by Marcin Cymborowski, Wladek Minor, and Maksymilian Chruszcz.

3 Applications of Xtaldb

3.1 Test set

New crystallization methods and techniques are very often tested on readily available and easily crystallized proteins, such as lysozyme, glucose isomerase, or thaumatin (for recent examples, see Dunlop & Hazes, 2005; Hansen *et al.*, 2006; Khurshid & Chayen, 2006, among many others). However, this approach is not too informative as these model proteins usually crystallize very readily; it is relatively difficult to find conditions in which lysozyme does *not* crystallize (Chayen & Saridakis, 2001). Using diffraction data as a method of comparing crystal quality of such proteins is usually meaningless as handling of crystals (harvesting, freezing, mounting, etc.) introduces more variability than the crystallization methods.

For that reason, we used a different approach to test the Xtaldb system. A set of

Table 3-1: Initial test proteins screened with Xtaldb.

	Organism	MW (kDa)	Function	Structure? (PDB id)	Diffracting crystals?	Reso. (Å)
YfbR	*E. coli*	22.7	5'-deoxynucleotidase	Yes (2PAQ)	Yes	2.1
RcsA	*E. coli*	23.5	Activator of colanic acid capsular polysaccharide synthesis	No	Yes	3.0
YgiC	*E. coli*	45.0	Glutathionylspermidine synthase/amidase homolog	No	Yes	3.9
TM0549	*T. maritima*	19.4	Acetohydroxyacid synthase regulatory subunit homolog	Yes (2FGC)	Yes	2.3
TM0913	*T. maritima*	29.8	MazG homologue	No	Yes	2.6
TM1030	*T. maritima*	23.8	TetR-family transcriptional regulator	Yes (1Z77)	Yes	2.0

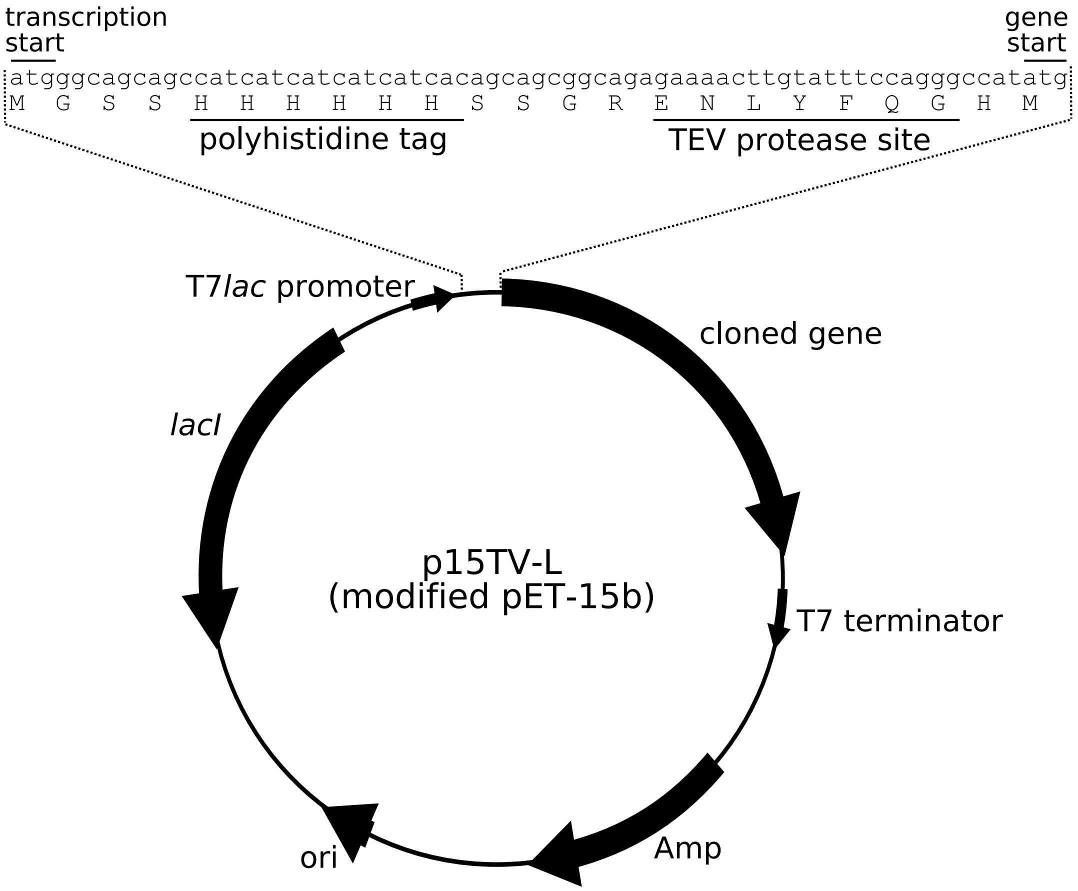
novel bacterial proteins that failed to produce diffraction quality crystals in the pipeline of the Midwest Center for Structural Genomics (MCSG) were purified, screened and optimized using the *Xtaldb* system. The proteins chosen for analysis are listed in Table 3-1. None of the proteins had a known 3-D structure, so that experimental phases would have to be determined (though preference was made for proteins containing sufficient methionine for SeMet phasing). Additionally, most of the proteins chosen were identified by sequence to be members of protein families with some annotated biochemical information but no known 3-D structure. Clones of each protein were obtained from the lab of Aled Edwards at the University of Toronto (described in Section 3.6 below).

3.2 Protein expression

The T7 RNA polymerase expression system was used to produce protein in the bacterium *E. coli* for purification (Studier, 1991; Studier *et al.*, 1990). All of the targets obtained were cloned into the p15TV-L vector, a modification of the multi-copy pET-15b vector (Novagen), as shown in Fig. 3-1. Like other pET-based vectors, the p15TV-L vector places the target protein gene under the control of a promoter recognized only by the RNA polymerase from bacteriophage T7. When the vector is transformed into a typical *E. coli* strain, the native RNA polymerase does not produce the cloned gene as it does not recognize the T7 promoter (Studier *et al.*, 1990). Because this decreases the constitutive, uninduced level of protein expression and thus provides improved protection from toxic DNA constructs to the expression cells, this approach is widely used for protein production in *E. coli*.

To introduce T7 RNA polymerase for expression into system, derivatives of the *E. coli* strain BL21(DE3) are used. BL21(DE3) contains a lysogen of the λ-derived bacteriophage DE3 inserted into the chromosome. The DE3 lysogen contains the gene *lacI*, encoding for the *lac* repressor, and the gene for T7 RNA polymerase under the control of the *lacUV5* promoter (Studier *et al.*, 1990). In the absence of lactose (or one of its analogs), the *lac* repressor produced by *lacI* binds to *lacUV5* and prevents expression of T7 RNA polymerase. Lactose analogs such as isopropyl-β-D-thiogalactopyranoside (IPTG) inactivate the *lac* repressor, inducing expression of T7 RNA polymerase and in turn expressing the target gene (Studier *et al.*, 1990). In p15TV-L (and pET-15b), a second copy of the *lacI* gene and a *lac* repressor binding site at the T7 promoter is

Figure 3-1: Schematic diagram of a gene cloned into the p15TV-L vector.
The region of the plasmid sequence that encodes the N-terminal tag of the expressed protein differs from the standard pET-15b vector is shown in expanded view. "ori" is the origin of plasmid replication.



thought to prevent constitutive expression of the target protein due to the “leaky” nature of *lac*UV5. A gene conferring resistance to the antibiotic ampicillin (*Amp*) allows selection for cells containing the plasmid (Fig. 3-1).

p15TV-L alters the target protein upon translation by appending 23 residues to the N-terminus of the expressed polypeptide. This tag contains 6 sequential histidines, which serve as an affinity tag for immobilized metal affinity chromatography (IMAC; described below), connected by a cleavage site recognized by tobacco etch virus (TEV) protease

(Fig. 3-1). As the N-terminal tag is lengthy and may disrupt formation of crystal lattices, cleavage with TEV protease provides the ability to remove all but two residues of the tag. For more than one test protein, removal of the affinity tag proved necessary for crystallization (see below).

To produce native protein, the cloned vectors were transformed into the *E. coli* Rosetta(DE3) or Rosetta2(DE3) expression strains (Novagen). Both strains are derivatives of BL21(DE3) and both contain extra chloramphenicol-resistant plasmids with genes encoding tRNAs for codons rarely seen in *E. coli* genes. The presence of the rarest codons in the transcript of genes, particularly when the rare codons are found in sequential clusters, has been shown to negatively affect both the translation accuracy and amount of expression of heterologous proteins in *E. coli* (Kane, 1995; Rosenberg *et al.*, 1993). Co-expressing tRNAs for rare codons has been shown to improve expression of heterologous proteins (Brinkmann *et al.*, 1989; Del Tito *et al.*, 1995). Rosetta(DE3) cells contain genes expressing tRNAs for the codons AUA, AGG, AGA, CUA, CCA, and GGA. Rosetta2(DE3) adds the gene expressing tRNA for codon CGG (Novagen).

The strain of cells used to produce protein incorporated with selenomethionine (SeMet) was B834(DE3)pLysS. B834(DE3) is a derivative of BL21(DE3) with mutations rendering it unable to synthesize methionine (i.e., it is a methionine auxotroph). The pLysS vector constitutively expresses a low level of T7 lysozyme in BL21(DE3) cells. T7 lysozyme selectively inhibits the activity of T7 RNA polymerase, which prevents basal expression of target proteins in pET-derived vectors due to the leaky nature of the *lacUV5* promoter, allowing the cloning of relatively toxic target proteins into the cells (Studier, 1991). When the pET system is induced with IPTG, the level of expression of

Table 3-2: Buffers used in expression, purification, and crystallization.

Name	*Contents*	*pH*
Lysogeny broth (LB) medium (Bertani, 1951)	1 %w/v bacto-tryptone 0.5 %w/v yeast extract 1 %w/v NaCl	~7
Terrific Broth (TB) medium (Tartof & Hobbs, 1987)	1.2 %w/v bacto-tryptone 2.4 %w/v yeast extract 0.4 %w/v glycerol 2.0 %w/v glucose 17 mM KH_2PO_4 72 mM K_2HPO_4	7.8
M9 minimal Medium	8 g/L Na_2HPO_4 4 g/L KH_2PO_4 5 g/L NaCl 5 g/L NH_4Cl 0.4 %w/v glucose 1 mg/L thiamine Trace metals: 1 mM $MgSO_4$, 1 mM $CaCl_2$, 30 μM $FeCl_3$, 61 μM $ZnCl_2$, 1.6 μM boric acid, 760 nM $CuCl_2$, 420 nM $CoCl_2$, 80 nM $MnCl_2$	7.4
Resuspension buffer	500 mM NaCl 50 mM HEPES 5 %v/v glycerol	7.5
Binding buffer	Resuspension buffer + 5 mM imidazole	7.5
Wash buffer	Resuspension buffer + 30 mM imidazole	7.5
Elution buffer	Resuspension buffer + 250 mM imidazole	7.5
TEV protease dialysis buffer	500 mM NaCl 50 mM HEPES 5 %v/v glycerol	7.5
Crystallization buffer	500 mM NaCl 10 mM HEPES	7.5

T7 RNA polymerase overwhelms the constitutive level of T7 lysozyme produced by the pLysS vector, preventing much change to the overall yield of induced protein expression (Studier, 1991).

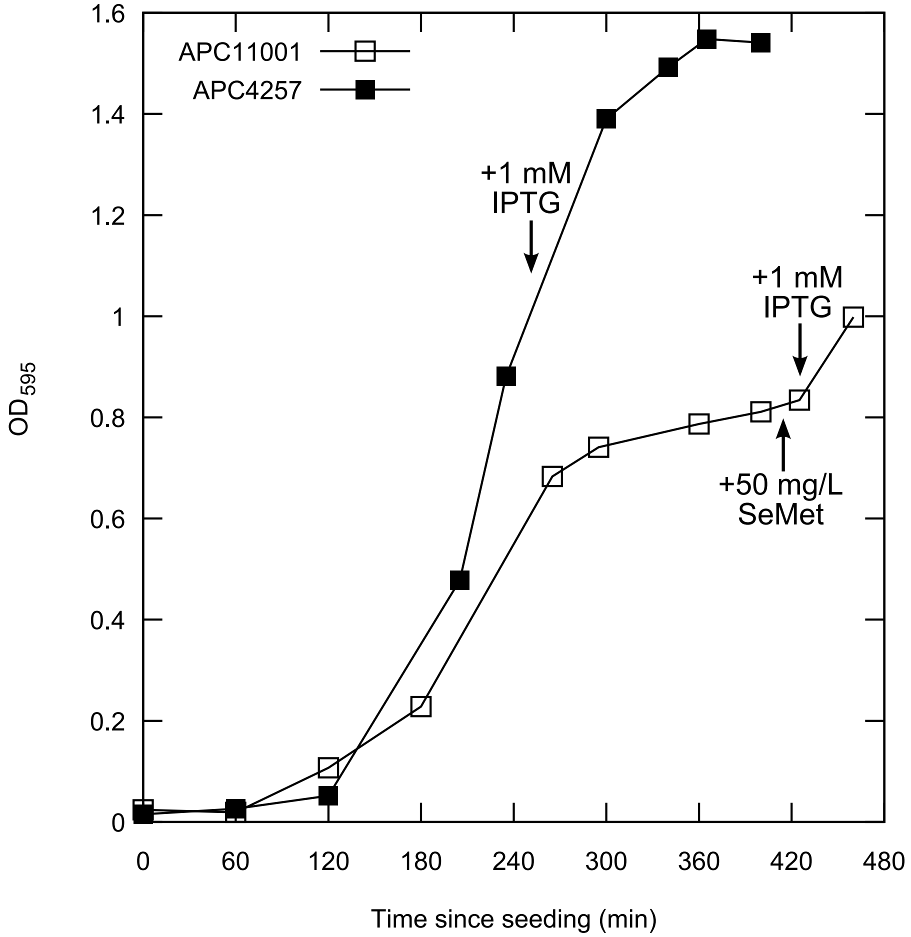
10-25 mL cultures of either the Rosetta(DE3) or Rosetta2(DE3) expression strains transformed with the target genes cloned into p15TV-L vectors were grown with shaking in baffled flasks in LB media for 3-4 hours at 37ºC. These cultures were then used to seed 1L baffled flasks containing sterile rich media (either LB or TB medium; see Table 3-2) and the appropriate antibiotics for selection (50 mg/L ampicillin and 34 mg/L chloramphenicol). The cultures were allowed to grow to stationary phase (average OD_{595} was 1.5 for LB, 2.0 for TB), and were then induced by adding IPTG to a final concentration of 1 mM.

To produce protein incorporated with SeMet, B834(DE3)pLysS cells (transformed with the target genes cloned into p15TV-L vectors) were grown overnight in small cultures of 10 mL of M9 minimal media (Table 3-2) with 50 mg/L methionine. The small cultures were used to seed 1L baffled flasks of M9 minimal media with 8 mg/L methionine, grown with shaking at 37ºC. When the cultures reached a stationary phase, usually around an OD_{595} of 0.6-0.8, SeMet was added to a final concentration of 50 mg/L. After 15 minutes, the SeMet cultures were induced by adding 1 mM IPTG.

Typical growth curves for a 1L native Rosetta(DE3) culture and for a 1L SeMet B834(DE3)pLysS culture are shown in Fig. 3-2. For both the native and Se-Met expressions, after induction the cultures were grown at 16ºC overnight (10-16 hours), and were harvested by centrifugation. To monitor growth of *E. coli* cultures, small samples were taken at regular intervals and monitored by optical density at 595 nm (OD_{595}).

Figure 3-2: Typical growth of 1L expression cultures.

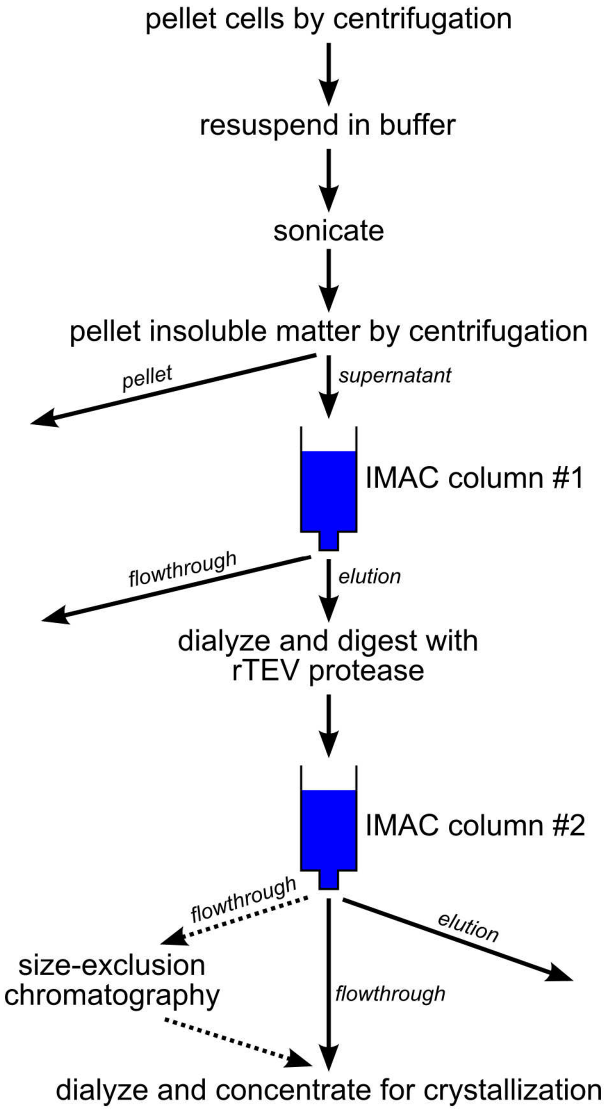
Native protein is in closed squares and SeMet protein is in open squares. OD$_{595}$ is the optical density at 595 nm, blanked versus *sterile medium. APC11001 was grown in B834(DE3)pLysS cells in M9 minimal medium containing 8 mg/L methionine. APC4257 was grown in Rosetta(DE3) cells in LB medium. The points where SeMet and IPTG were added to the cultures are marked by arrows.*



3.3 Protein purification

The overall purification pipeline is summarized in Fig. 3-3. The cell pellets were resuspended in resuspension buffer (Table 3-2), and the resuspended cells were lysed by ultrasonication, which uses sonic waves to disrupt the membranes of the cells. The lysis was performed in the presence of a protease inhibitor cocktail (final concentration each of

Figure 3-3: Schematic diagram of the protein purification process.


pellet cells by centrifugation
resuspend in buffer
sonicate
pellet insoluble matter by centrifugation
pellet
supernatant
IMAC column #1
flowthrough
elution
dialyze and digest with rTEV protease
IMAC column #2
flowthrough
elution
size-exclusion chromatography
flowthrough
dialyze and concentrate for crystallization

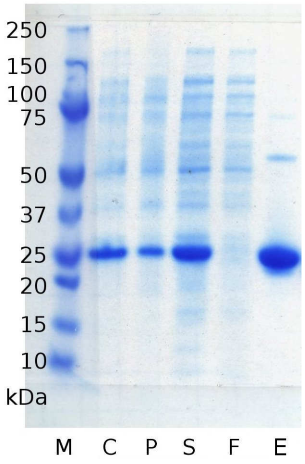
1 mM benzamidine, 1 mM PMSF), which prevents degradation of the expressed protein by proteases released upon cell lysis. The insoluble components of the resulting lysate were pelleted by centrifugation at 15000 rpm for 15-30 minutes in a Sorvall SS-34 rotor. In virtually all cases, the purified proteins remained soluble after this centrifugation step, though when TM0549 was prepared in the presence of SeMet, most of the protein pelleted in the insoluble lysis fraction, and had to be purified by a refolding protocol (see below).

Protein purification is typically monitored by SDS-PAGE, a technique for separating denatured proteins (approximately) by molecular weight using an electrical potential across a matrix of polyacrylamide. A set of protein markers of known molecular weight are used to calibrate the migration of each protein. An SDS-PAGE gel for a typical purification of wild-type YfbR is shown in Fig. 3-4. A significant fraction of the protein in the cells is YfbR, and while some amount of the protein is lost in the insoluble pellet fraction, an amount sufficient for purification was retained in the flowthrough.

The first chromatography step purified protein by means of immobilized metal affinity chromatography (IMAC). In IMAC divalent cations such as Ni^{2+}, Co^{2+}, or Zn^{2+} are bound to a stationary substrate, usually polymeric solids like agarose packed into a chromatography column (Porath *et al.*, 1975). Certain amino acids residues are capable of chelating these divalent cations much more effectively than others, and long strings of six (or greater) histidine residues, which can be specifically added to the N- or C-termini of recombinant proteins, bind to Ni^{2+} with high affinity (Hochuli *et al.*, 1988). The histidines coordinate the metal cation, which is also coordinated to a compound like nitrilotriacetic acid (NTA) covalently linked to the solid substrate (Hochuli *et al.*, 1988).

Figure 3-4: SDS-PAGE of a typical purification experiment.

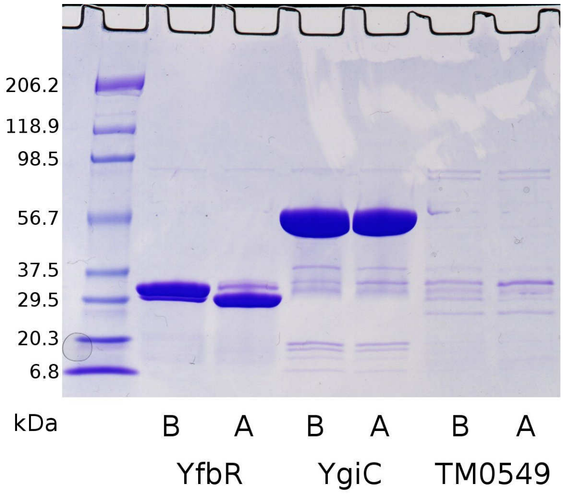
An SDS-PAGE gel of a typical purification of YfbR is shown below. The lanes are as labeled: M – molecular markers, C – cells before lysis, P – pellet after lysis, S – supernatant after lysis, F – first IMAC flowthrough, E – first IMAC elution. The molecular weight of each marker protein is shown in kDa along the left hand side of the image.



To release the bound protein, imidazole (the sidechain moiety of histidine) is applied to the column and competes with the histidine for metal cation binding, thus liberating the bound protein. Originally such purifications were performed under denaturing conditions (Hochuli *et al.*, 1988), but by using a lower concentration of imidazole during the binding and wash steps, binding of lower-specificity proteins can be inhibited and allow purification of 95% or better (Janknecht *et al.*, 1991).

The supernatants were applied by gravity flow to 5 mL (packed volume) columns of Ni-NTA resin (QIAGEN) affinity resin equilibrated with binding buffer (Table 3-2). The resins were washed with 10-20 column volumes of wash buffer (with 30 mM imidazole), then purified protein was eluted with binding buffer (with 250 mM imidazole). All buffers used to purify SeMet-substituted proteins also contained 5 mM β-mercaptoethanol (BME). In the typical purification shown in Fig. 3-4, virtually all of the wild-type YfbR protein bound to the affinity resin, and after elution was fairly (but not

Figure 3-5: SDS-PAGE of typical TEV protease digestions.
An SDS-PAGE gel of SeMet-incorporated versions of three test proteins, before (B) and after (A) 48 hours of digestion at 4º C with recombinant TEV protease. The overall level of expression of TM0549 was very low (as discussed below). The molecular weights of the marker proteins in kDa are shown along the left hand side of the image.



completely) pure.

After the initial IMAC step, the polyhistidine affinity tag was cleaved by digestion with recombinant TEV protease (Kapust *et al.*, 2001) at 4ºC for 2-4 days during dialysis into binding buffer. The cleavage was confirmed by SDS-PAGE. Examples of TEV protease cleavage of three different test proteins are shown in Fig. 3-5. Some of the proteins could not be cleaved by the protease, and in those cases the elutions from the first affinity column protein was used in further crystallization experiments. For example, in Fig. 3-5, YfbR shows clear lysis of the tag (as indicated in a decrease in protein size after the reaction), while YgiC does not change in size.

Those proteins that were cleaved were dialyzed back into binding buffer and applied to a second Ni-NTA column prepared as before, a technique sometimes referred to as subtractive purification. After cleavage of the tag, the protein should pass through in the flowthrough of the second IMAC column, which the impurities that bound to the last column will once again bind as before. In most cases the majority of the cleaved protein passed through the column, and this flowthrough was dialyzed into crystallization buffer and prepared for crystallization.

In the case of YfbR, however, the cleaved protein bound again to the affinity resin. The cleavage of the tag was confirmed by mass spectroscopy, so this result was unexpected. Using different metal affinity matrices such as TALON Superflow (BD Biosciences) resin, which uses Co^{2+} as the metal instead of Ni^{2+}, resulted in similar results. The mystery was eventually solved when it was determined that the protein was a likely metal-binding protein (discussed further in Chapter 4). Even though the tag was removed, the inherent ability of the protein to bind Ni^{2+} and Co^{2+} prevented purification of YfbR by subtractive IMAC methods.

For YfbR, the elution from the second IMAC column was further purified by size-exclusion chromatography, which (roughly) separates proteins by molecular diameter. In the case of YfbR, a HiPrep 16/60 Superdex 200 column operated with an AKFA FPLC system (GE Healthcare) was used. Eventually preparations of all proteins were further purified by size-exclusion, as the technique is capable of separating the monodisperse and aggregating populations of a given protein in solution.

3.4 Crystallization

After purification, each protein was dialyzed into 10 mM HEPES pH 7.5 and 0.5 M NaCl, and concentrated to 8-50 mg/mL. Protein was screened for crystallization in a series of commercial screens: Hampton Research's Crystal Screen I, Crystal Screen II, PEG/Ion, and Index Screen, and promising hits were optimized. Other commercial screens from deCODE Genetics and QIAGEN were used if hits were not found in the initial set. A variety of optimization techniques were used, such as including small concentrations of organic additives or putative cofactors, varying drop sizes and ratios, sampling multiple temperatures (4º, 20º, 37º, 50º C), and in one case, re-purifying protein using a refolding protocol.

After optimization with Xtaldb, all six proteins in the initial test set subsequently produced diffracting crystals, with five diffracting to 3.5 Å or better. Selenomethionine-substituted protein was produced and crystallized for each of those five proteins, and of them, the structures of three could be solved by MAD or SAD methods. Each of these proteins were subsequently co-crystallized with putative ligands or cofactors, and in total resulted in 7 PDB deposits. As the Xtaldb system has matured, all crystallization experiments set up in the Minor laboratory at the University of Virginia are now tracked with the system. The use of Xtaldb on other projects in the Minor laboratory produced three other new crystal structures: mouse apolipoprotein A-1 binding protein (PDB id 2DG2), mouse sperm c-type lysozyme-like protein 1 (2DOI), and *P. aeruginosa* protein PA5185 (2AV9). These proteins yielded 9 additional PDB deposits.

Crystals of SeMet-substituted wild-type YfbR (see Chapter 4) were crystallized by hanging-drop vapor diffusion, where the well contained a solution of 10% w/v PEG 3350, 0.3 M NH_4 citrate, and 5 mM BME, and the drop contained 2 μL of protein solution mixed with 2 μL of well solution. Cleavage of the N-terminal poly-histidine tag proved to be necessary, as none of the conditions with tags attached produced hits. Crystals were harvested, briefly soaked in a 4:1 mixture of well solution and glycerol, and flash-frozen in liquid N_2 for diffraction at 103 K.

YfbR alanine mutants (see Chapter 4) were screened both in commercial screens and in conditions that produced crystals for the wild-type protein. Well-diffracting crystals of the E72A mutant were produced by hanging-drop vapor diffusion, as described above, in 0.1 M NH_4 citrate, 5% w/v PEG 3350, 0.2-0.8% w/v NDSB 256, and 1% v/v glycerol. Crystals were transferred into a soaking buffer containing 5 mM $CoCl_2$, 0.15 M Na acetate pH 4.5, 5% w/v PEG 3350, and either 8.3 mM 2'-deoxyriboadenosine-5'-monophosphate (dAMP) or 7.5 mM thymidine-5'-monophosphate (TMP), and incubated for 10-18 hours. The crystals were then briefly transferred into a 3:1 mixture of soaking buffer and PEG 400 and flash-frozen in liquid N_2.

An initial hit for native TM0549 was obtained from Hampton Research’s Crystal Screen I commercial screen, condition #44 (0.2 M Mg formate). rTEV protease proved ineffective in cleaving the tag off of the protein, and TM0549 was not treated with the enzyme in subsequent purifications. After optimization, some native TM0549 crystals based on this initial condition diffracted fairly well (to a resolution of 2.5 Å). However, it was found that Se-Met-incorporated protein precipitated during the standard purification process. Some efforts were made to prepare heavy-metal soaks for either SIR/MIR or MAD, but failed. Eventually, protein purification was carried out under denaturing conditions, by taking the insoluble protein fraction from the cell fractionation spin (see Fig. 3-3) and dissolving in binding buffer plus 6 M guanidine hydrochloride (Petkowski *et al.*, 2007).

The crystallization of TM0549 was performed in the presence of a relatively high concentration (0.5 M) of L-arginine, which was determined by means of a screen of refolding agents (Petkowski *et al.*, 2007). While the use of L-arginine in preventing aggregation in protein refolding is well known (Tsumoto *et al.*, 2004), the use of the amino acid as an additive for crystallization has not been widely reported. The refolded Se-Met incorporated protein was re-screened in a number of premixed commercial screens, though the initial hit of Mg formate proved to be necessary. The optimized vapor diffusion well contained 2% v/v glycerol, 0.4% w/v NDSB 201, and 150 mM Mg formate, and the purified protein solution in the drop was 1 uL 7.7 mg/mL in 0.5 M L-arginine pH=8.5 and 50 mM NaCl mixed with 1 uL well solution. 50×50×50 μL crystals appeared after about 24 hours. The structure was solved by Se-Met SAD at by Janusz Petkowski (Petkowski *et al.*, 2007).

TM1030 could be purified in both native and Se-Met incorporated forms without refolding. An initial hit was found in Hampton Research's Index screen condition #95; 0.1 M KCSN and 30% w/v polyethylene glycol 2000 monomethyl ether (PEG 2K MME). Cleavage of the N-terminal polyhistidine tag proved necessary, as uncleaved protein failed to crystallize. After optimization, the final crystallization conditions had a well solution of 50 mM Na citrate pH=4.5, 30% w/v PEG 2K MME, 0.1 KCSN, with 1 μL protein at 7.2 mg/mL mixed with 1μL of well conditions. The structure of TM1030 was solved by Kasia Koclega and others at 2.0 Å resolution by means of an unusual dual-

Table 3-3: Different crystal forms of PA5185.

Spacegroup	*Reso (Å)*	*Solvent content (%)*	*Molecules in AU*	*Additive*
C2	2.4	44	12	$(NH_4)_2SO_4$
$P3_221$	3.1	61	2	HEPES
$P222_1$	2.5	55	6	NDSB 201
C222	1.9	54	3	MES

crystal “MAD” experiment (Koclega *et al.*, 2007). Subsequently, a crystal of TM1030 crystallized with the addition of 3% w/v dextrose yielded a structure at 1.75 Å resolution, and another crystallized with the addition of 10 mM NH_4 sulfate in a different crystal form (spacegroup C2 instead of $P2_12_12$).

By using different additives, the *P. aeruginosa* protein PA5185, which was not part of the initial screen but was later analyzed with Xtaldb, crystallized in 4 different crystal forms, as shown in Table 3-3. The initial crystal form crystallized to 2.4 Å and contained a sulfate group. When the structure was superimposed on a homolog, the sulfate group was found to superimpose on a phosphate group of the ligand in the homolog structure. Thus the set of additives were chosen due to the fact that they contained sulfate or sulfate-like groups. These different crystal forms differ significantly in resolution, ranging from 1.9 Å to 3.1 Å, with widely different spacegroups and percentage of solvent content. This resulted in a significant improvement of resolution to 1.9 Å over the original 2.4 Å.

3.5 Quantitative analysis of crystallization experiments

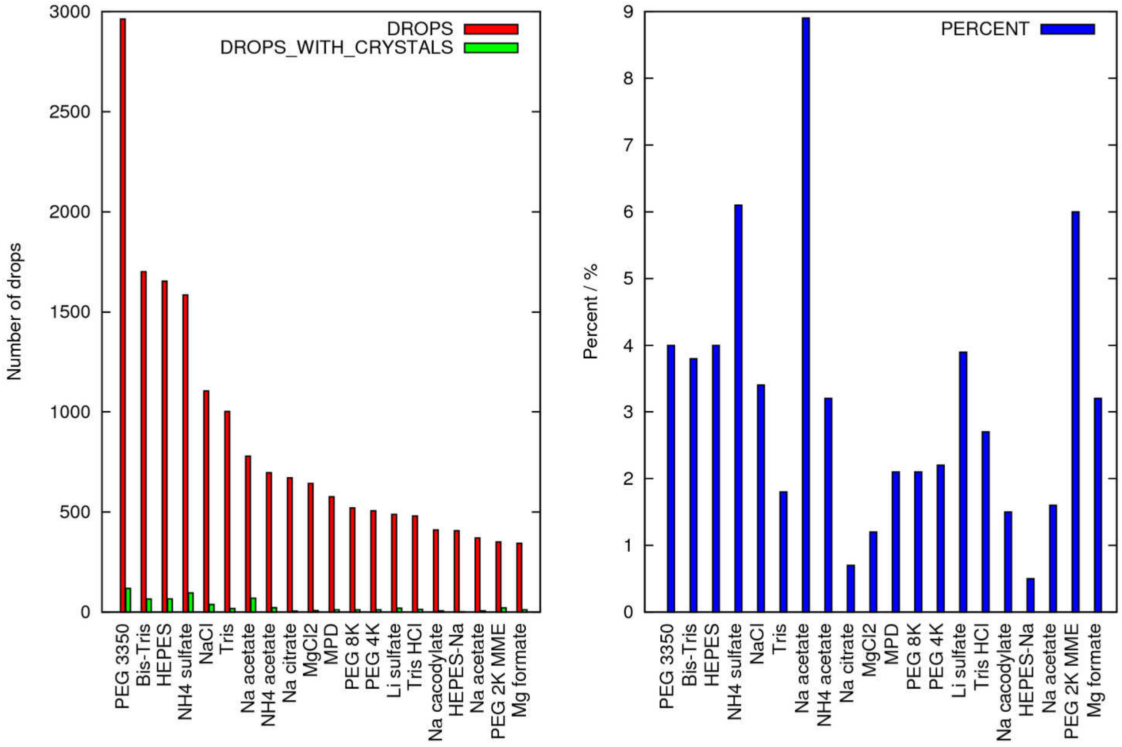
The Xtaldb system is now routinely used for all crystallization setups in the Minor laboratory. At the time of writing, 44343 crystallization drops on 1549 plates, set up using solutions of 23 different proteins, are tracked with the system. While the statistical significance of results drawn with this data set do not compare with the larger data sets collected by other groups (see Chapter 1), this set of crystallization experiments has proven useful in designing template queries for dynamically analyzing data.

To choose a basis set of additives for the design of random screens, it is helpful to know the most “successful” reagents used in the past. In other words, if a particular reagent is present relatively frequently in crystallization drops containing crystals, Bayesian reasoning suggests that future experiments should be biased toward the used of that reagent in the future. This logic influenced the choice of reagents used in most commercial screens, and reagents such as PEG (Radaev *et al.*, 2006), ammonium sulfate (Gilliland, 1988; Peat *et al.*, 2005), and sodium malonate (McPherson, 2001) have been found frequently in successful crystallization experiments.

Fig. 3-6 shows the twenty most frequently used reagents in crystallization screens tracked with the Xtaldb system. Optimizations were explicitly excluded to prevent possible bias due to replicated conditions (see below). As commercial crystallization screens were predominantly used, it is not surprising that PEG 3350 and ammonium sulfate are highly represented in the list of overall usage frequencies. (One of the screens used was Hampton Research’s PEG/Ion screen, where all 48 conditions contain PEG 3350.) Despite the fact that PEG 3350 is used almost twice as frequently as the second

Figure 3-6: Most frequently used reagents and relative rates of success.

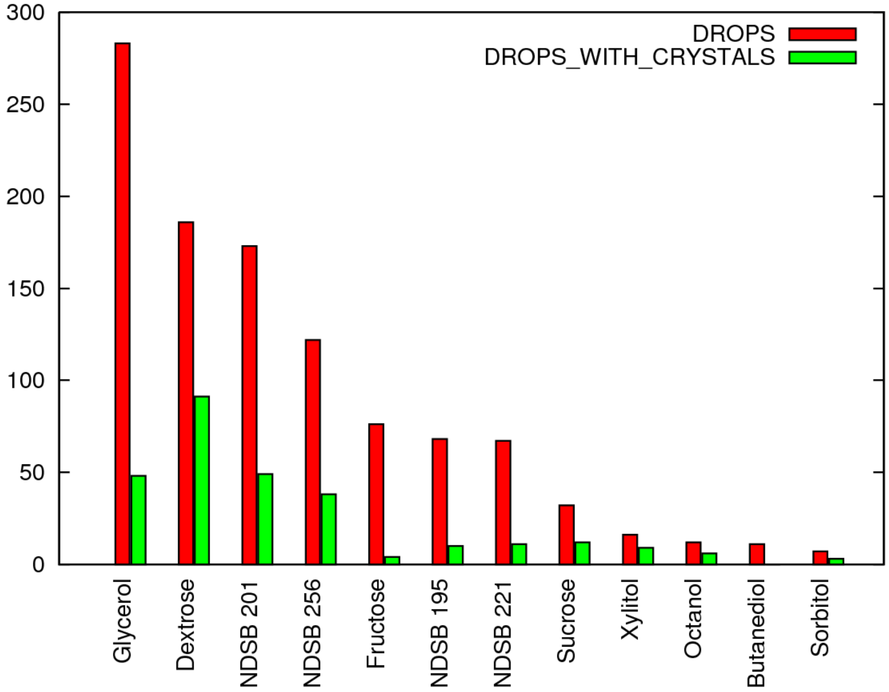
In the left plot, the twenty most frequently used reagents in the prototype Xtaldb database, where the red bars show the total number of drops, and the green bars the number of drops annotated to contain crystals. The analysis was limited to screening experiments, not optimization, to avoid bias. The right plot shows the relative rate of success for the same twenty reagents by percentage. The figure was generated with Xtaldb.



most frequent chemical, with our set of proteins, ammonium sulfate, sodium acetate, and PEG 2K MME were more frequently seen in successful crystallizations (Fig. 3-6).

The use of additives has also proven successful to some extent within the test set, as shown in Fig. 3-7. Either wild-type or ligand-bound forms of all three of the proteins with structures solved were crystallized in the presence of non-detergent sulfobetaines (NDSB), a set of zwitterionic ammonium propane sulfonate derivatives with favorable solubilization and protein stabilization properties (Vuillard *et al.*, 1996). Of the 170 drops set up containing NDSB 201, 49 are recorded to contain crystals. We obtained similar

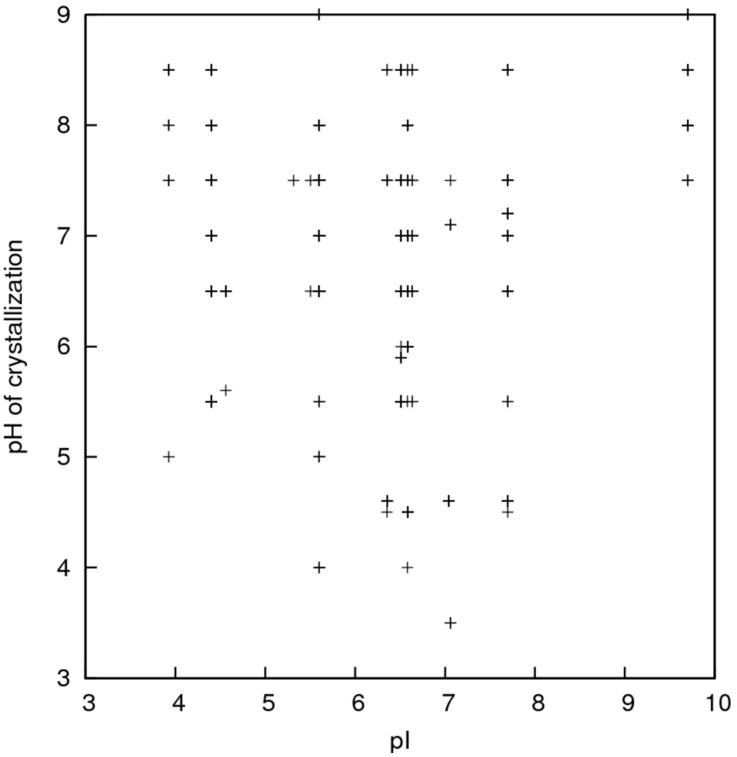
Figure 3-7: Additive frequency and success rates for test proteins.
The nine compounds below represent the most frequently used additives in the crystallization Xtaldb test set, where the red bars indicate the total number of drops set up containing the reagent, and the green bars the number of drops annotated to contain crystals. The figure was generated with Xtaldb.



results with NDSB 256 (38 of 118) and NDSB 195 (10 out of 55). Other additives such as sugars (dextrose, sucrose, xylitol) and small organics (glycerol, octanol) showed similarly high rates of successful crystallization. It should be noted that such data could be biased as successful crystallization conditions are replicated for the purpose of producing sufficient crystals for diffraction, and there is no simple mechanism for determining which crystallization experiments are duplicates.

One particular topic of debate in the field of empirical crystallization analysis is whether there is an observable relationship between the isoelectric point (pI; the pH at which a given protein is electrically neutral) of a particular protein and the pH levels at

Figure 3-8: Relation between pI and pH of crystallization.
Each data point represents a crystallization drop in the Xtaldb annotated to contain crystals, where the x value is the calculated pI for the protein and the y value is the pH of crystallization. Drops that did not contain buffers are omitted from this plot. The figure was generated using Xtaldb.



which it crystallizes. Most studies have suggested there is no relationship (Page *et al.*, 2003), while others have claimed that there is a correlation (Kantardjieff & Rupp, 2004) though the methodology used to determine that correlation has been questioned (Huber & Kobe, 2004). Fig. 3-8 plots the pH of crystallization drops, as a function of calculated pI for the protein in that drop, for all of the experiments annotated to contain crystals in the prototype Xtaldb system. While one protein with a relatively high pI (> 9) does appear to preferentially produce crystals in the pH range of 8-9, in general, there is no observable correlation between pI and crystallization pH.

All of these analyses were done in real time using built-in template SQL queries in the console dialog of Xtaldb.

3.6 Structures solved with Xtaldb

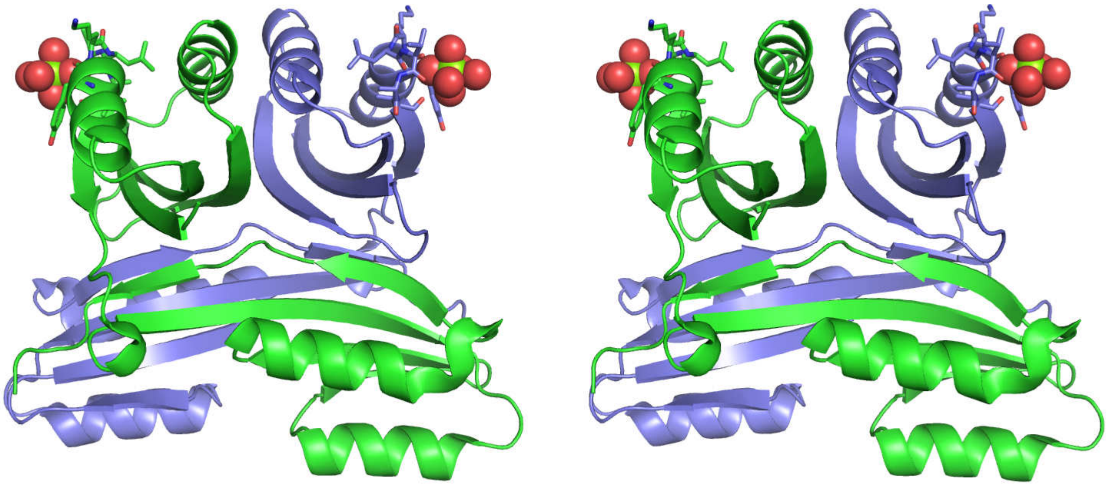
All three of the structures solved with Xtaldb to date subsequently led to full biochemical and structural studies, so many derivative crystals were grown with Xtaldb. All three structures also led to full papers describing structure/function relationships. One example is YfbR, which was established by sequence analysis to be a member of a widely distributed metal-dependant phosphohydrolase family. This annotation guided the structural and biochemical analyses of this protein that described in the next chapter.

The crystal structure of TM0549 is shown in Fig. 3-9. TM0549 from *Thermotoga maritima* is similar by sequence to the so-called small regulatory subunit of the acetohydroxy acid synthase isoform III of *E. coli* (Kaplun *et al.*, 2006). Upon determination of the crystal structure, the necessity of including Mg formate in the crystallization conditions became obvious, as each monomer contained an ordered $Mg(H_2O)_6^{2+}$ ion that formed major crystal contacts, shown as space-filling spheres in Fig. 3-9 (Petkowski *et al.*, 2007). It is unclear if this metal ion is necessary for the physiological functioning of the protein, as the *E. coli* ortholog was also crystallized in Mg^{2+} but did not contain an ordered ion in the electron density (Kaplun *et al.*, 2006).

Acetohydroxy acid synthase catalyzes two similar reactions, one catalyzing the condensation of 2 pyruvate molecules to 2-acetolactate, and the second catalyzing the condensation of pyruvate and 2-ketobutyrate to 2-aceto-2-hydroxybutyrate, which are precursors for the synthesis of leucine and valine. The small regulatory subunit, which contains an ACT domain and a ferrodoxin-like domain, does not catalyze the reaction, but it is both required for proper assembly of the functional enzyme, and its ACT domain

Figure 3-9: The crystal structure of TM0549.

The putative biological dimer is shown in a stereo view, with one monomer in green and the other in blue. Mg^{2+} ions and their coordinating waters are shown as space-filling spheres.



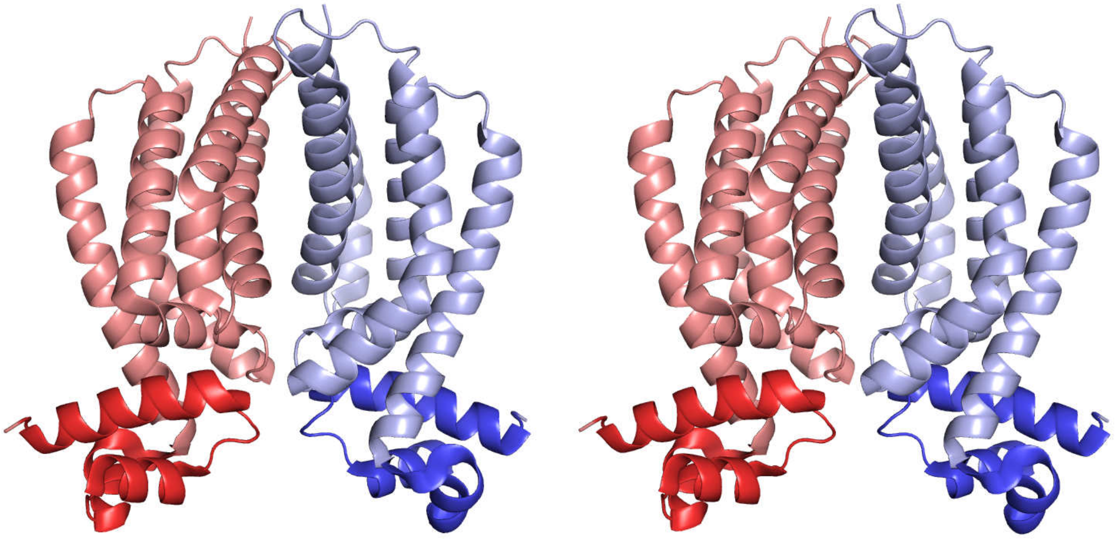
binds valine, which is a negative regulator of enzyme function (Mendel *et al.*, 2003).

The crystal structure of TM1030 is shown in Fig. 3-10. The protein has been determined by both sequence and structural similarity analysis to DNA-binding transcriptional regulators of the TetR family. The three helices at the N-terminus, shown in darker colors in Fig. 3-10 are highly conserved and contain the helix-turn-helix (HTH) DNA binding motif. The larger, C-terminal ligand-binding domain is less conserved among members of the family.

Shortly after the structure of TM1030 was deposited (PDB code 1Z77), a structure of the same protein crystallized in different conditions was deposited by the Joint Center for Structural Genomics (1ZKG). This structure contained electron density for a large, unidentified linear ligand (most likely PEG), and had significant conformational changes, presumably caused by the binding of the unknown ligand (Premkumar *et al.*, 2007). Most notably, the distance between the DNA-binding HTH

Figure 3-10: The crystal structure of TM1030.

The (putative) biological dimer is shown in stereo representation, with one monomer in red and one in blue. Within each monomer, the N-terminal DNA-binding domain is shown in a darker color, and the C-terminal ligand-binding domain in a lighter color.



motifs in the two subunits of the dimer increased from 30 Å to 50 Å, larger than the distance between two major grooves in DNA, and it has been postulated that this conformational change prevents DNA binding (Koclega *et al.*, 2007).

TM1030 is structurally similar to the multidrug resistance protein EthR from *M. tuberculosis*, which was solved in complex with the ligand hexadecyl octanoate (Frenois *et al.*, 2004). The EthR protein was identified as a component in the mechanism of ethionamide resistance, a drug for the treatment of multidrug-resistant *M. tuberculosis* infection (Frenois *et al.*, 2004). While the unknown ligand in the other structure of TM1030 binds in a different conformation than the aliphatic molecule in EthR, it is of similar size and binds in a similar location. It is postulated that TM1030 has a similar function to EthR, binding a long, possibly aliphatic molecule, and disrupting the probable DNA-binding function of the protein (Koclega *et al.*, 2007).

4 Structural analysis of the 5'-nucleotidase YfbR

4.1 HD-domain phosphohydrolases

The first test target to have its structure solved was YfbR, a 199 amino-acid, 22.7 kDa protein from *E. coli*. By sequence it was known that YfbR belonged to a large superfamily of proteins known as the HD-domain metal-dependant phosphohydrolase family. The HD-domain, characterized by a divalent-cation-binding **H...HD...D** motif, is a widely conserved catalytic domain found in nearly 5000 proteins in bacteria, archaea, and eukaryotes. Enzymes containing HD-domains act as broad-substrate-range phosphohydrolases that can catalyze both metal-dependent and -independent phosphomonoesterase and phosphodiesterase reactions (Aravind & Koonin, 1998). They have diverse functions associated with nucleic acid and nucleotide metabolism and signal transduction (Seto *et al.*, 1988; Aravind & Koonin, 1998; Yakunin *et al.*, 2004).

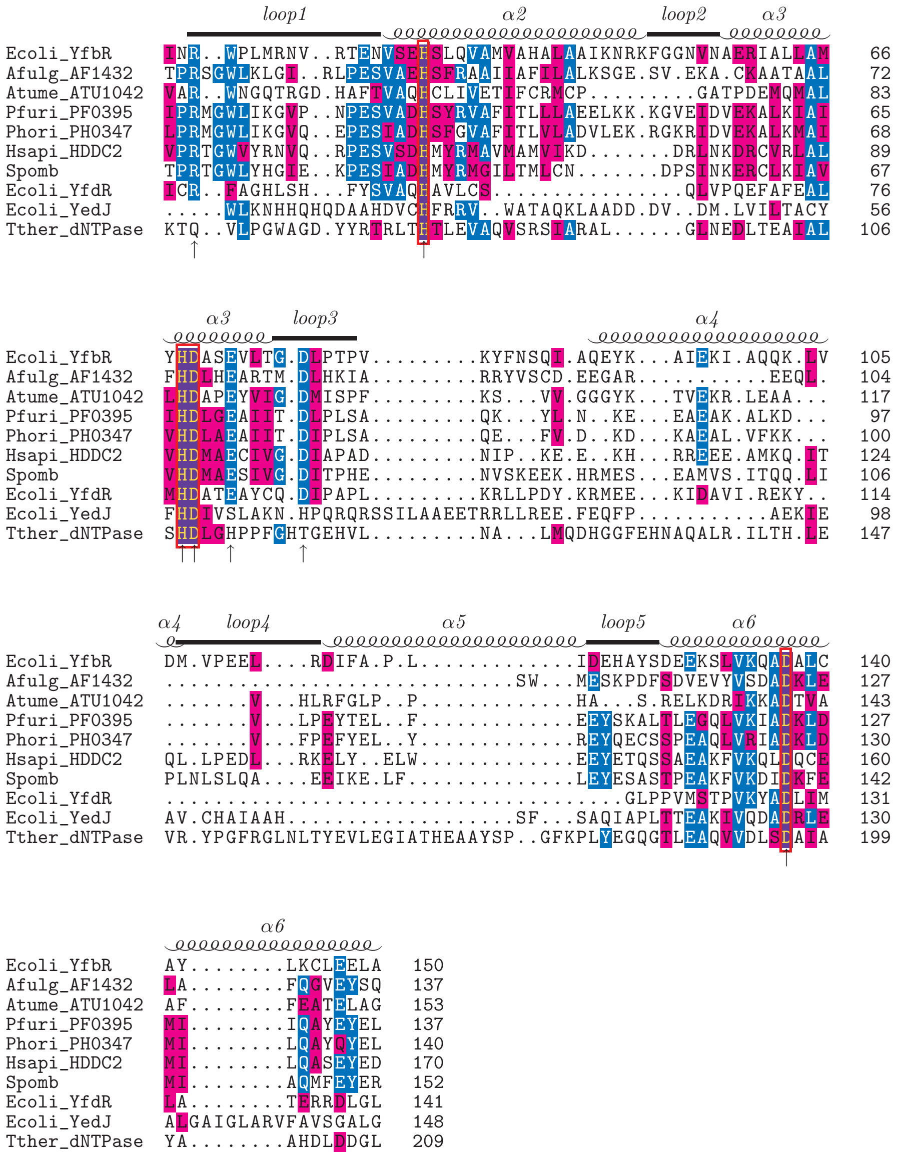
In each sequenced genome, there are seven to twenty HD domain proteins encoded as stand-alone proteins or fused to nucleotidyltransferase or helicase domains (Aravind & Koonin, 1998). To date, only four HD domain proteins have been characterized biochemically: dGTPase (Seto *et al.*, 1988), RelA/SpoT (An *et al.*, 1979), and tRNA nucleotidyltransferase (Yakunin *et al.*, 2004) from *E. coli*, and the *Thermus thermophilus* dNTPase (Kondo *et al.*, 2004), and of them only two have structures elucidated: the catalytic fragment of the RelA/SpoT homolog from *Streptococcus equisimilis* (PDB id 1VJ7; Hogg *et al.*, 2004) and *T. thermophilus* dNTPase (2DQB; Kondo *et al.*, 2007). The bifunctional RelA/SpoT catalyzes both the synthesis and hydrolysis of (p)ppGpp, which is produced in large quantities in the so-called “stringent response" in bacteria, a mechanism by which the organisms conserve energy during nutrient starvation (Chatterji & Ojha, 2001). Microbial dNTPases control the intracellular pool of dNTPs and hydrolyze them to deoxynucleoside and inorganic triphosphate (Seto *et al.*, 1988; Kondo *et al.*, 2004).

HD-domain proteins are also related to the catalytic domain of class I eukaryotic phosphodiesterases (PDEs), which hydrolyze the 3'-5' cyclic phosphate bond of the intracellular second messengers cAMP and cGMP (Bender & Beavo, 2006). PDEs which contain the core **H...HD...D** motif but are distinct from the HD domain superfamily in that they contain additional conserved regions. Due to their pharmacological relevance, several structures of the catalytic domains of PDE4 and PDE5 have been solved in complex with substrate analogs (Huai *et al.*, 2003; Xu *et al.*, 2004; Zhang *et al.*, 2004). The catalytic domains of PDE4 and PDE5 contain a Zn^{2+} and an unknown divalent metal cation (presumably Mg^{2+}), both of which are thought to coordinate a bridging hydroxide ion which serves as the nucleophile in a single-step catalysis mechanism, and a conserved His residue protonates the leaving O3' group (Houslay & Adams, 2003; Xiong *et al.*, 2006). However, apart from the four residues of the HD motif itself, the conserved residues in the PDE superfamily identified to play a role in substrate recognition or catalysis are not conserved in non-PDE members of the HD-domain superfamily.

E. coli YfbR has over 100 orthologs found in bacteria, archaea, and eukaryotes. The human genome encodes one YfbR ortholog, the uncharacterized HD-domain-containing protein HDDC2 (28% sequence identity), which might represent a novel intracellular 5`-nucleotidase in humans. In addition to YfbR, the *E. coli* genome encodes two more stand-alone HD domain proteins, YfdR (178 amino acids) and YedJ (231

Figure 4-1: Sequence alignment of YfbR with homologs.

HD-domain proteins of known structure (AF1432, ATU1042, PF0395, PH0347, T. thermophilus *dNTPase), paralogs (E. coli YfdR, YedJ) and orthologs from humans (HDDC2) and* S. pombe *are shown. Residues identical to the consensus are in blue and similar residues are shown in purple. The secondary structure of YfbR is shown above the alignment. The signature HD motif is boxed in red, and alanine mutation sites that reduced or abolished catalytic activity (see text) are marked with arrows. The sequence alignment figure was created with TEXSHADE (Beitz, 2000).*



```
Ecoli_YfbR     INR..WPLMRNV..RTENVSEHSLQVAMVAHALAAIKNRKFGGNVNAERIALLAM   66
Afulg_AF1432   TPRSGWLKLGI..RLPESVAEHSFRAAIIAFILALKSGE.SV.EKA.CKAATAAL   72
Atume_ATU1042  VAR..WNGQTRGD.HAFTVAQHCLIVETIFCRMCP.........GATPDEMQMAL   83
Pfuri_PF0395   IPRMGWLIKGVP..NPESVADHSYRVAFITLLLAEELKK.KGVEIDVEKALKIAI   65
Phori_PH0347   LPRMGWLIKGVQ..EPESIADHSFGVAFITLVLADVLEK.RGKRIDVEKALKMAI   68
Hsapi_HDDC2    VPRTGWVYRNVQ..RPESVSDHMYRMAVMAMVIKD.......DRLNKDRCVRLAL   89
Spomb          TPRTGWLYHGIE..KPESIADHMYRMGILTMLCN.......DPSINKERCLKIAV   67
Ecoli_YfdR     ICR..FAGHLSH...FYSVAQHAVLCS...............QLVPQEFAFEAL   76
Ecoli_YedJ     .....WLKNHHQHQDAAHDVCHFRRV..WATAQKLAADD.DV..DM.LVILTACY   56
Tther_dNTPase  KTQ..VLPGWAGD.YYRTRLTHTLEVAQVSRSIARAL......GLNEDLTEAIAL  106

Ecoli_YfbR     YHDASEVLTG.DLPTPV.........KYFNSQI.AQEYK...AIEKI.AQQK.LV  105
Afulg_AF1432   FHDLHEARTM.DLHKIA.........RRYVSCD.EEGAR...........EEQL.  104
Atume_ATU1042  LHDAPEYVIG.DMISPF.........KS...VV.GGGYK...TVEKR.LEAA...  117
Pfuri_PF0395   IHDLGEAIIT.DLPLSA.........QK...YL.N..KE...EAEAK.ALKD...   97
Phori_PH0347   VHDLAEAIIT.DIPLSA.........QE...FV.D..KD...KAEAL.VFKK...  100
Hsapi_HDDC2    VHDMAECIVG.DIAPAD.........NIP..KE.E..KH...RREEE.AMKQ.IT  124
Spomb          VHDMAESIVG.DITPHE.........NVSKEEK.HRMES...EAMVS.ITQQ.LI  106
Ecoli_YfdR     MHDATEAYCQ.DIPAPL.........KRLLPDY.KRMEE...KIDAVI.REKY..  114
Ecoli_YedJ     FHDIVSLAKN.HPQRQRSSILAAEETRRLLREE.FEQFP..........AEKIE   98
Tther_dNTPase  SHDLGHPPFGHTGEHVL.........NA...LMQDHGGFEHNAQALR.ILTH.LE  147

Ecoli_YfbR     DM.VPEEL....RDIFA.P.L.............IDEHAYSDEEKSLVKQADALC  140
Afulg_AF1432   ..........................SW...MESKPDFSDVEVYVSDADKLE  127
Atume_ATU1042  .......V...HLRFGLP..P.............HA...S.RELKDRIKKADTVA  143
Pfuri_PF0395   .......V...LPEYTEL..F.............EEYSKALTLEGQLVKIADKLD  127
Phori_PH0347   .......V...FPEFYEL..Y.............REYQECSSPEAQLVRIADKLD  130
Hsapi_HDDC2    QL.LPEDL...RKELY..ELW.............EEYETQSSAEAKFVKQLDQCE  160
Spomb          PLNLSLQA....EEIKE.LF..............LEYESASTPEAKFVKDIDKFE  142
Ecoli_YfdR     ....................................GLPPVMSTPVKYADLIM  131
Ecoli_YedJ     AV.CHAIAAH..................SF...SAQIAPLTTEAKIVQDADRLE  130
Tther_dNTPase  VR.YPGFRGLNLTYEVLEGIATHEAAYSP..GFKPLYEGQGTLEAQVVDLSDAIA  199

Ecoli_YfbR     AY........LKCLEELA  150
Afulg_AF1432   LA........FQGVEYSQ  137
Atume_ATU1042  AF........FEATELAG  153
Pfuri_PF0395   MI........IQAYEYEL  137
Phori_PH0347   MI........LQAYQYEL  140
Hsapi_HDDC2    MI........LQASEYED  170
Spomb          MI........AQMFEYER  152
Ecoli_YfdR     LA........TERRDLGL  141
Ecoli_YedJ     ALGAIGLARVFAVSGALG  148
Tther_dNTPase  YA........AHDLDDGL  209
```

amino acids). These uncharacterized proteins share low sequence similarity with YfbR (18.8% and 15.3% sequence identity, respectively). A sequence alignment of several stand-alone HD domain proteins is shown in Fig. 4-1,created by inputting the set of HD-domain proteins along with the 31 sequences of the COG1896 domain (Tatusov *et al.*, 2003) to the 3DCOFFEE web server (Poirot *et al.*, 2004), which generates alignments using both sequence and structural information. The alignment reveals the conservation of the predicted metal-coordinating HD domain motif **H…HD…D**.

4.2 Wild-type structure solution

Diffraction data on a SeMet-substituted wild-type YfbR crystal were collected at the peak (0.9793 Å), inflection (0.9795 Å), and “high energy remote” (0.9755 Å) wavelengths for Se fluorescence at beamline 19-ID of the Structural Biology Center (SBC-CAT) at the Advanced Photon Source (APS) (Rosenbaum *et al.*, 2006). The data

Table 4-1: Crystallographic data collection and refinement statistics.
Each of the crystal structures of YfbR described in this thesis are shown. Statistics for the highest resolution shell for each structure are shown in parentheses. The redundancy was calculated by grouping Friedel pairs of reflections together.

Data collection	*wild-type YfbR*	*E72A-Co-TMP*	*E72A-Co-dAMP*
Space group	R3	R3	R3

Unit cell parameters	a=b=137.0 Å
	c=56.3 Å
		a=b=136.1 Å
		c=55.4 Å
			a=b=135.6 Å
			c=54.9 Å

Wavelength (Å)	0.9793	0.9793	0.9793
Resolution range (Å)	50-1.95 (2.02-1.95)	50-2.1 (2.14-2.10)	50-2.1 (2.18-2.10)
I/σI	12.1 (1.7)	22.9 (2.8)	28.2 (3.2)
R_{merge}	0.102 (0.580)	0.081 (0.423)	0.063 (0.563)
Unique reflections	28633	22238	21959
Completeness (%)	98.2 (87.8)	100.0 (100.0)	99.9 (100.0)
Redundancy	5.0 (3.6)	5.1 (5.1)	4.5 (4.4)
Solvent content (%)	44	43	42
Solved by	SeMet SAD	MR	MR
Refinement			
Resolution range (Å)	50-2.1	40.4-2.1	40.1-2.1
No. protein atoms/AU	2793	2783	2777
No. waters/AU	126	42	53
No. substrate atoms/AU	0	51	53
R/R_{free} (%)	18.7/23.7	19.6/24.7	19.6/26.0
Mean B-factor (Å^2)	26.6	35.0	38.3
RMSD bond length (Å)	0.022	0.020	0.022
RMSD angles (°)	1.81	1.84	1.91
Ramachandran favored (%)	96.7	96.4	98.8
Ramachandran allowed (%)	3.3	3.6	1.2
Ramachandran outliers (%)	0.0	0.0	0.0
PDB id	2PAQ	2PAR	2PAU

were indexed, integrated, and scaled with HKL-2000 (Otwinowski & Minor, 1997) in space group R3. The frames of data collected diffracted to about 2.0 Å with absorption correction, but were highly mosaic (1.4-1.7º). The data collection parameters are listed in Table 4-1.

Se sites were determined using the program SHELXD (Schneider & Sheldrick, 2002). SHELXD was given the expected number of heavy atom sites (14 in this case), generated an initial set of Se sites consistent with the Patterson correlation of the data, then conducted cycles of dual-space refinement (the so-called "shake-and-bake" algorithm). Calculated phases are generated from the set of model sites, and refined with the collected structure factors; then the refined phases are used to generate a new density map, where the peaks are used as the model sites for the next cycle (Schneider & Sheldrick, 2002). Correlation coefficients measure the agreement of the calculated structures factors with those measured. For wild-type YfbR, the correlation coefficients for the Se sites using data from all three wavelengths (MAD) were significantly worse than those for only the peak wavelength (SAD), using a resolution cutoff of 2.7 Å. A set of 12 Se sites with good occupancy were found.

At this point, the Se substructure was used in two different pipelines to generate electron density maps and initial models. In the first pipeline, an alpha version of HKL-3000 (at the time called "HKL-2000_ph"; Minor *et al.*, 2006) was used. To determine the proper hand of the sites found, HKL-3000 used SHELXE (Sheldrick, 2002) to generate phases with both the original sites and those sites inverted. The sites with the proper hand (i.e., which set of sites gave higher contrast and connectivity of the resulting maps) were then passed into MLPHARE (Otwinowski, 1991), which further refined the heavy atom positions and then solved for the native phases by SAD phase solution. Because unambiguous phases cannot be determined with SAD data alone (see Section 1.1), centroid phases with their figures of merit were output, and then passed into the density modification program DM (Cowtan, 1994), which was able to generate an interpretable map. This map was input into RESOLVE for automatic model building. RESOLVE identifies regions of the map that appear to form α-helices or β-strands, then uses template fragments from known structures to build model atoms into those regions (Terwilliger, 2002). In wild-type YfbR, RESOLVE built 155 residues out of the expected 398, 31 of them with the proper sequence.

In the second pipeline, the SHELXD Se atom sites were passed into SHARP/autoSHARP, which uses the SHARP program for phasing (Bricogne *et al.*, 2003) and SOLOMON (Abrahams & Leslie, 1996) for density modification. This pipeline generated an electron density map using SAD phasing methods similar to those described above for MLPHARE and DM, *albeit* with a different likelihood function (Bricogne *et al.*, 2003). The resulting map was passed into the ARP/wARP suite for automatic model building (Perrakis *et al.*, 1999). ARP/wARP uses the “warpNtrace” algorithm, which builds nonbonded “free” atoms into the electron density, then traces patterns matching protein stereochemistry. The hybrid model is refined and then the tracing step is repeated, followed by sequence docking (Perrakis *et al.*, 1999). The ARP/wARP program traced 151 out of 398 residues but was not able to dock any sequence information.

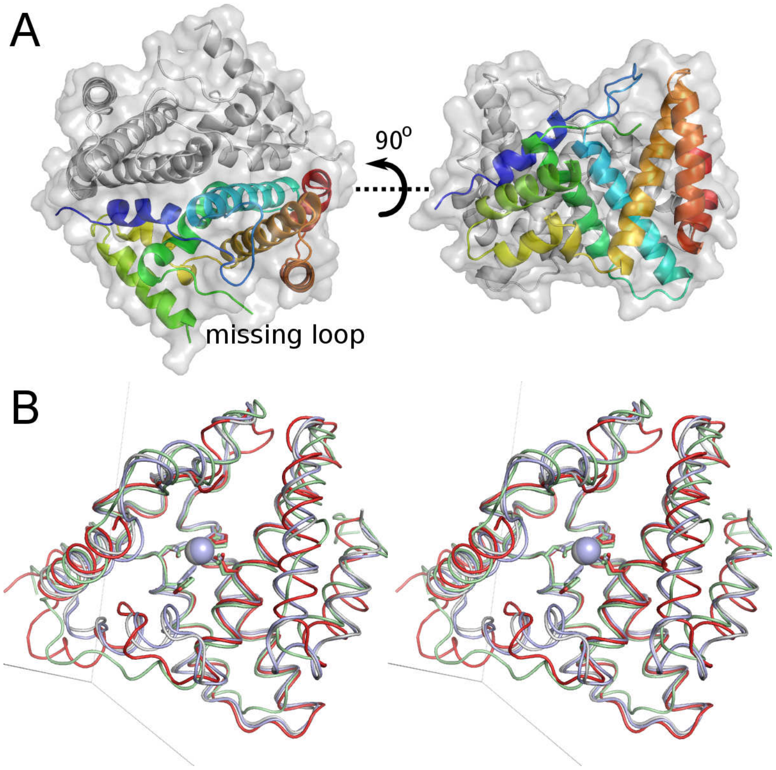
The two different automatic building algorithms (RESOLVE and ARP/wARP) complimented each other surprisingly well, in general building similar secondary structure fragments in either offset or overlapping regions. By superimposing both molecules in a molecular graphics program, as well as taking advantage of the two-fold symmetry of the dimer observed in the structure, a dimer model with complete sequence could be obtained. This model had 175 out of 199 residues for each polypeptide (350 residues in total).

The combined model was refined by cycles of simulated annealing and restrained maximum-likelihood refinement in CNS 1.1 (Brunger *et al.*, 1998) and manual rebuilding in O (Jones *et al.*, 1991). Two-fold NCS constraints were used during the maximum-likelihood refinement. The model was validated with the PROCHECK and Molprobity programs, both of which analyze model geometry and Ramachandran plot statistics. Refinement and Ramachandran statistics are shown in Table 4-1. When refinement was complete, the structure was deposited to the PDB with id 1WPH. Later, when mutant YfbR structures were determined and refined using the programs REFMAC5 (Murshudov *et al.*, 1997) from the CCP4 package (CCP4, 1994), and COOT (see below), it was observed that the R and R_{free} statistics for the wild-type model were significantly worse than those for the mutants. The wild-type structure was improved by maximum-likelihood restrained refinement in REFMAC5 and manual rebuilding with COOT (Emsley & Cowtan, 2004). TLS restraints (Winn *et al.*, 2001) were used in refinement. The TLS model consisted of 12 groups (6 segments per chain) as identified by the TLS Motion Determination server (Painter & Merritt, 2006), and a new, slightly corrected model was submitted to the PDB with id 2PAQ.

The globular structure of YfbR, shown in Fig. 4-2(a), consists of 8 α-helices per monomer, connected by extended loops. Two polypeptide chains are found in the

Figure 4-2: Wild-type YfbR crystal structures.

(A) Cartoon representation of the presumed dimer of wild-type YfbR, viewed in two different orientations. One of the two polypeptides of the dimer is colored by primary sequence, with the N-terminus in blue and the C-terminus in red. The missing loop between helices α3 and α4 is indicated. (B) Stereo view of a monomer of wild-type YfbR (red) superimposed with the structures of the orthologs AF1432 (PDB id 1ynb; green), PF0395 (PDB id 1xx7; blue), and PH0347 (PDB id 2cqz; gray) superimposed. The sidechains of the HD motif are shown in stick representation. The structures of PF0395 and PH0347 each contain a Zn^{2+} in their respective HD metal-binding motifs, which are shown as a blue sphere and gray sphere, respectively.



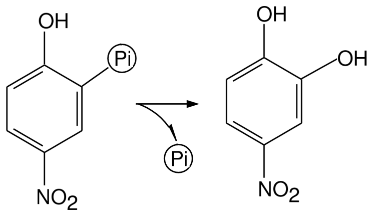
asymmetric unit. In both chains, residues 82-90 in the loop between helices α3 and α4 and residues 188-199 of the C-terminus of the protein were disordered and not included in the structure. In addition, the side chains of several residues, particularly in helix α4, were also not modeled. Analysis of the crystal packing as well as size exclusion chromatography results (see Chapter 3) suggested that the dimer seen in the asymmetric unit is also the biological unit, where the dimer interface surface area is approximately 2100 Å^2.

The architecture of the four residues (H33, H68, D69, and D137) that compose the cation binding site is very similar to that found in the structure of the catalytic N-terminal fragment of the *S. equisimilis* RelA (Hogg *et al.*, 2004), save that the imidazole group of H68 is oriented away from the cation binding site. No metal is observed in the divalent cation binding site. The crystallization condition that produced this crystal contained a significant concentration of NH4 citrate (0.1-0.4 M), and the citrate appeared to be necessary to produce crystals large enough for diffraction. Therefore, it was hypothesized that citrate acts as a chelating agent, preventing binding of divalent cations.

A search for structural homologs identified four stand-alone HD domain proteins whose structures were deposited in the PDB following solution of the structure of YfbR: AF1432 from *Archaeglobus fulgidus* (PDB id 1YNB), PF0395 from *Pyrococcus furiosus* (1XX7), PH0347 from *Pyrococcus horikoshii* (2CQZ), and ATU1052 from *Agrobacterium tumefaciens* (2GZ4). Although these proteins share relatively low sequence similarity with YfbR (26% to 32% sequence identity), their crystal structures indicate strong structural similarity (Z-scores of 10.7 to 18.4, r.m.s.d. 2.3 to 2.7 Å). The structure-based alignments and all statistics were calculated using DaliLite (Holm & Park, 2000). As shown in Figure 4-2B, the structures of AF1432, PF0395, and PH0347 were particularly similar to YfbR in the regions corresponding to helices α2, α3, and α6 (r.m.s.d. 2.3 to 2.5 Å). These helices contain the predicted catalytic HD motif suggesting that these enzymes use the same catalytic mechanism.

4.3 Biochemical analysis of YfbR

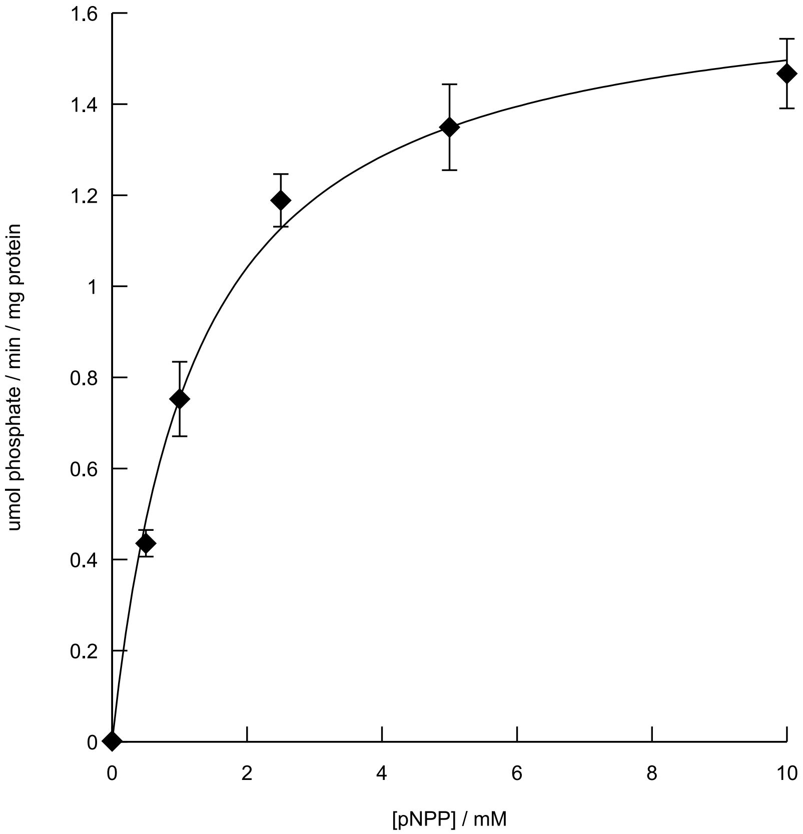
To determine the function of YfbR, the protein was first assayed in the presence of the reporter compound *p*-nitrophenol phosphate (pNPP; see Appendix A for the small molecule structure). pNPP is colorless, but the 2-hydroxy-4-nitrophenol product produced when the phosphate is cleaved is yellow, and the rate of production of phosphate may be measured by absorbance at 410 nm.



The commercially available enzyme alkaline phosphatase was used as a positive control. Purified YfbR was mixed into a series of 200 μL reactions on a 96-well microplate, where each reaction contained 5 mM $CoCl_2$, 1.38 μg protein (as determined by Bradford assay), 50 mM HEPES pH=7.5, and 0-10 mM pNPP, and allowed to react for 187 minutes at 37°C. There were seven duplicates at each pNPP concentration. When the reaction had completed, each well's absorbance at 410 nm was measured in a microplate spectrophotometer. The results of the reaction course are plotted in Fig. 4-3. The results were fit to a single-site Michaelis-Menten kinetics model, where V_{max} is the maximal rate of catalysis, and K_m is the the concentration of phosphate where the rate of catalysis is half of V_{max} (the Michaelis constant). The results fit well to the model, with $K_m = 1.23 \pm 0.13$ mM and $V_{max} = 1.680 \pm 0.053$ μmol/min/mg protein (though it is possible that the V_{max} is underestimated due to the long incubation time of the reaction).

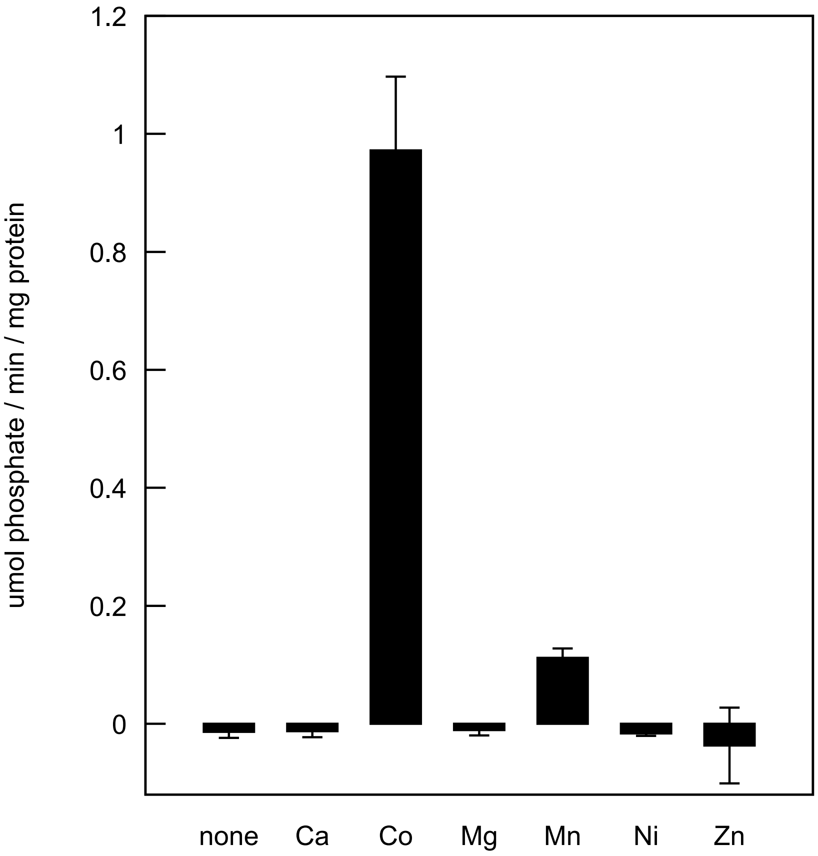
The metal dependence of phosphohydrolytic activity of YfbR was also measured.

Figure 4-3: Kinetic parameters for hydrolysis of pNPP by YfbR.
The hydrolysis was measured as free phosphate released per min per mg of YfbR as a function of substrate concentration. The solid curve represents a single-state Michaelis-Menten fit of the data. The error bars represent the standard error (n=7).



200 µL reactions were set up on a microplate, each containing 30 mM pNPP, 1.7 µg protein, 50 mM HEPES pH=7.5, and 5 mM of one of $CaCl_2$, $CoCl_2$, $CuCl_2$, $MgCl_2$, $MnCl_2$, $NiCl_2$, or $ZnCl_2$ (n=5 for each divalent cation). After incubating for 152 min at 37ºC, the plate was scanned for absorbance at 410 nm, and the results are shown in Fig. 4-4 (the $CuCl_2$ reaction precipitated and was omitted). A metal is strictly required for activity of the enzyme, with the highest activity in the presence of Co^{2+} , followed by Mn^{2+}. However, since Co^{2+} is very rare *in vivo*, it is postulated that Mn^{2+} serves as the primary cofactor in living cells.

Figure 4-4: Strict metal dependence of phosphohydrolysis by YfbR.
Rate of phosphate liberation from YfbR was measured by the breakdown of 30 mM pNPP in the presence of 5 mM of the divalent cation (as the chloride salt), 1.7 µg YfbR and 50 mM HEPES pH=7.5. Error bars represent the standard error (n=5).



Shortly after these experiments were conducted, a paper was published by Alexander Yakunin and coworkers at the University of Toronto (Proudfoot *et al.*, 2004). Their work confirmed both that YfbR had phosphohydrolase activity, and the pattern of strict metal selectivity observed here (Co^{2+} > Mn^{2+} > Cu^{2+}). More interestingly, by using a generalized screen of natural phosphatase substrates, they identified the likely substrates for YfbR. Specifically, 90+ endogenous phosphate-containing compounds (nucleotides and phosphorylated sugars, amino acids, and organic acids) were screened in microplates in 160 µl reactions with $CoCl_2$ and incubated for one hour at 37°C. The reactions were halted by the addition of 40 μl of the Malachite Green reagent (Baykov *et al.*, 1988) and the level of P_i released was measured by absorbance at 630 nm.

YfbR is strictly specific to 5'-deoxyribonucleotides (dNMPs), and does not show activity for 5'-ribonucleotides (NMPs), 3'-deoxyribonucleotides (3'-dNMPs), or nucleotide di- (NDPs) or triphosphates (NTPs). NDPs and NTPs also competitively inhibit the activity of the enzyme. However, the enzyme appears not to discriminate among particular nucleotide bases, as there was less than a 2-fold difference in activity (K_m = 12-47 μM, V_{max} = 0.37-0.71 mM) among the six biological 5'-deoxynucleotides tested (Proudfoot *et al.*, 2004). This is the first nucleotidase with this particular pattern of specificity. The 5'-NT activity of YfbR is strictly dependant on the presence of divalent metal cation and has a slightly alkaline pH optimum of 8.0 (Proudfoot *et al.*, 2004).

5'-nucleotidases (5'-NTs; EC 3.1.3.5) form the catabolic component of the set of enzymes that regulate the intracellular pool of nucleotides and 2'-deoxynucleotides for DNA and RNA synthesis (Bianchi *et al.*, 1986). 5'-nucleotidases have also been shown to play a role in the production of extracellular adenosine for cell signaling in mammalian cells (Zimmermann, 1992; Hunsucker *et al.*, 2005), in nucleotide scavenging pathways in mammals (Hunsucker *et al.*, 2005), and the phosphate-starvation response of certain bacteria (Rittmann *et al.*, 2005).

Mammalian 5'-nucleotidases have been relatively well-characterized, both biochemically and structurally (Bianchi & Spychala, 2003; Hunsucker *et al.*, 2005). Seven classes of 5'-NTs have been identified: five located in the cytosol (cN-IA, cN-IB, cN-II, cN-III, and cdN), one found in the mitochondrial matrix (mdN), and one anchored to the outer surface of the plasma membrane (eN). While the classes vary in patterns of substrate specificity, all of the intracellular 5'-NTs belong to the α/β-fold haloalkanoic acid dehydrogenase (HAD) superfamily, and structures have been solved for cN-II (PDB id 2J2C), cN-III (Bitto *et al.*, 2006), and mdN (Rinaldo-Matthis *et al.*, 2002). Based on structural studies, intracellular 5'-NTs are thought to share a common mechanism, where the nucleotide phosphate binds a coordinated divalent metal ion (usually Mg^{2+}) and a conserved Asp makes an in-line nucleophilic attack on the bound phosphate, forming a phosphoaspartyl intermediate. This phosphoenzyme intermediate is then liberated by a second nucleophilic attack by water (Bitto *et al.*, 2006; Himo *et al.*, 2005; Wallden *et al.*, 2005). In contrast, eN belongs to the calcineurin superfamily of binuclear metallophosphatases, which includes the purple acid and Ser/Thr protein phosphatases. The structure of UshA, a periplasmic homolog of eN from *E. coli* has been solved (Knofel & Strater, 1999). Unlike the intracellular 5'-NTs, eN contains two metal ions in the catalytic site, and is proposed to have a single-step catalytic mechanism, where the attacking nucleophile is a metal-bound water or hydroxide ion (Knofel & Strater, 2001).

Much less is known about prokaryotic 5'-NTs. A number of membrane-associated, periplasmic and extracellular bacterial 5'-NTs have been identified and purified, but few have been extensively characterized, such as UshA from *E. coli*, NucA from *Haemophilus influenzae* (Zagursky *et al.*, 2000), or HppA from *Helicobacter pylori* (Reilly & Calcutt, 2004). Recently, using a systematic general enzymatic screen against a large set of purified bacterial proteins (Kuznetsova *et al.*, 2005), three of the first intracellular bacterial 5'-NTs were identified in *E. coli*: SurE, YjjG, and YfbR. SurE is a 5'(3')-nucleotidase and a member of a conserved domain found in prokaryotes and eukaryotes. SurE shows catalytic activity (K_m = 0.10-0.37 mM, V_{max} = 10-22 μmol/min/mg) for both purine and pyrimidine ribonucleotides and deoxyribonucleotides, can utilize a number of different divalent metals for activity ($Mn^{2+} > Co^{2+} > Ni^{2+} > Mg^{2+}$), and has an optimum pH of 7.0 (Proudfoot *et al.*, 2004). YjjG (like the eukaryotic intracellular 5'-NTs) is a member of the HAD superfamily, and displays relatively high activity (K_m = 0.51-0.77 mM, V_{max} = 46-73 μmol/min/mg) for the 5'-nucleotides UMP, dUMP, and TMP, with much lower activity for a variety of other mono- and diphosphate nucleotides. Divalent metal cation is required for activity ($Mg^{2+} > Mn^{2+} > Co^{2+}$), and the pH optimum (pH 7.5) is nearly neutral (Proudfoot *et al.*, 2004).

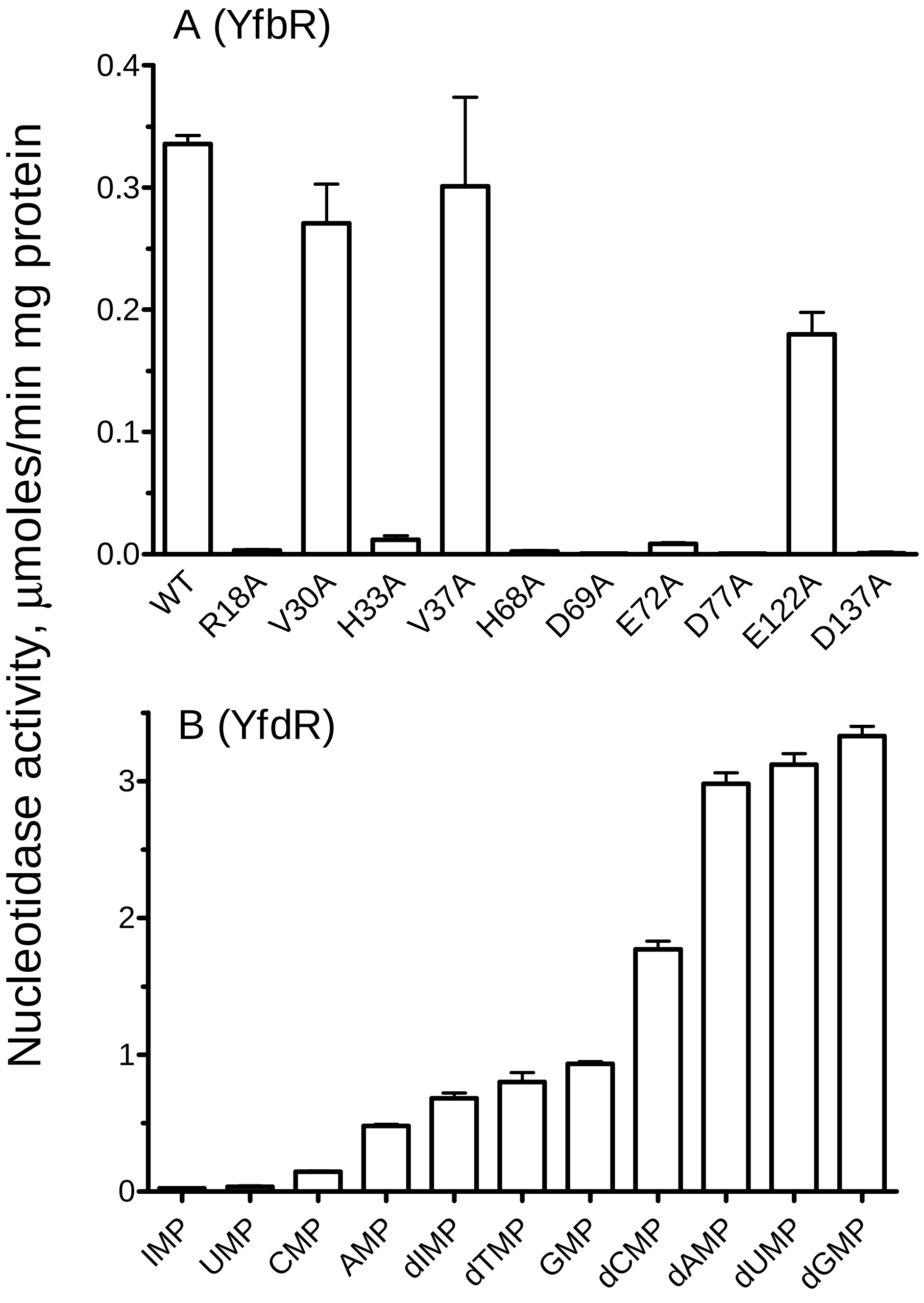
4.4 Alanine scanning mutagenesis

To determine which residues play a role in substrate recognition or catalytic activity or both, twelve conserved residues located near the predicted active site were identified (R18, W19, V30, H33, V37, H68, D69, E72, D77, E122, and D137; marked by arrows in Fig. 4-1). This part of the research was done in collaboration with Michael Proudfoot and Alexander Yakunin from the University of Toronto. The QuikChange^{TM} site-directed mutagenesis kit (Stratagene) was used to mutate the selected amino acid residues of YfbR to Ala, the proteins were over-expressed and purified using previously described protocols (Gonzalez *et al.*, 2006), and the nucleotidase activity of mutant YfbR was compared to that of the wild-type using dAMP as a substrate (Fig. 4-5). The V37A mutant showed wild-type activity and substrate affinity, V30A and E122A exhibited reduced activity and affinity, and seven mutants (R18A, H33A, H68A, D69A, E72A, D77A, and D137A) had greatly reduced or negligible enzymatic activity, demonstrating that they are important for activity. Four of these residues (H33, H68, D68, and D137) comprise the metal-binding motif of HD domain proteins.

Additionally, general phosphohydrolase assays of YfbR homologs YfdR and YedJ from *E. coli* were performed by Michael Proudfoot, and of AF1432 from *A. fulgidus* by the author. Each protein was over-expressed, purified and tested for phosphohydrolase activity against a range of phosphatase and phosphodiesterase substrates (Kuznetsova *et al.*, 2005). In these assays, both AF1432 and YedJ exhibited no significant activity, whereas YfdR was found to be another *E. coli* 5'-nucleotidase with a substrate range broader than that of YfbR (Fig. 4-5).

Figure 4-5:Nucleotidase activity of mutants of YfbR and of YfdR.

(A) Nucleotidase activity of YfbR alanine scanning mutants was measured in the presence of 0.5 mM $CoCl_2$ and 0.1 mM dAMP. (B) Nucleotidase activity of YfdR for selected natural substrates in the presence of 0.5 mM $CoCl_2$. (Figure produced by Michael Proudfoot and Alexander Yakunin, and reproduced with permission.)



The general screens were performed essentially as previously described (Kuznetsova *et al.*, 2005). Briefly, purified protein was screened for phosphatase activity using the general phosphatase substrate *p*-nitrophenyl phosphate (*p*NPP) in 200 μL reactions. The reactions contained 50 mM HEPES-K pH 7.5, 5 mM $MgCl_2$, 1 mM $MnCl_2$, 0.5 mM $NiCl_2$, 20 mM *p*NPP and 10μg of purified protein. Reactions were incubated for one hour at 37°C and then activity was measured by absorbance at 410 nm. The YfbR mutant profile was studied by testing the mutants for activity against dAMP under saturating conditions for the wild-type protein. The 160 μL reaction mixtures contained 50 mM HEPES-K pH 8.0, 0.5 mM $CoCl_2$, 0.1mM dAMP, and 1 μg of a YfbR mutant (or wild-type). The reactions were incubated for 20 minutes, then 40 μL of the Malachite Green reagent was added and the free P_i was measured at 630 nm. The dAMP saturation curves for the YfbR mutants that showed activity were done with 160 μL reaction mixtures containing 50 mM HEPES-K pH 8.0, 0.5 mM $CoCl_2$, 3 μM – 1.25 mM dAMP, and 0.6 – 1.2 μg of a YfbR mutant or wild-type protein.

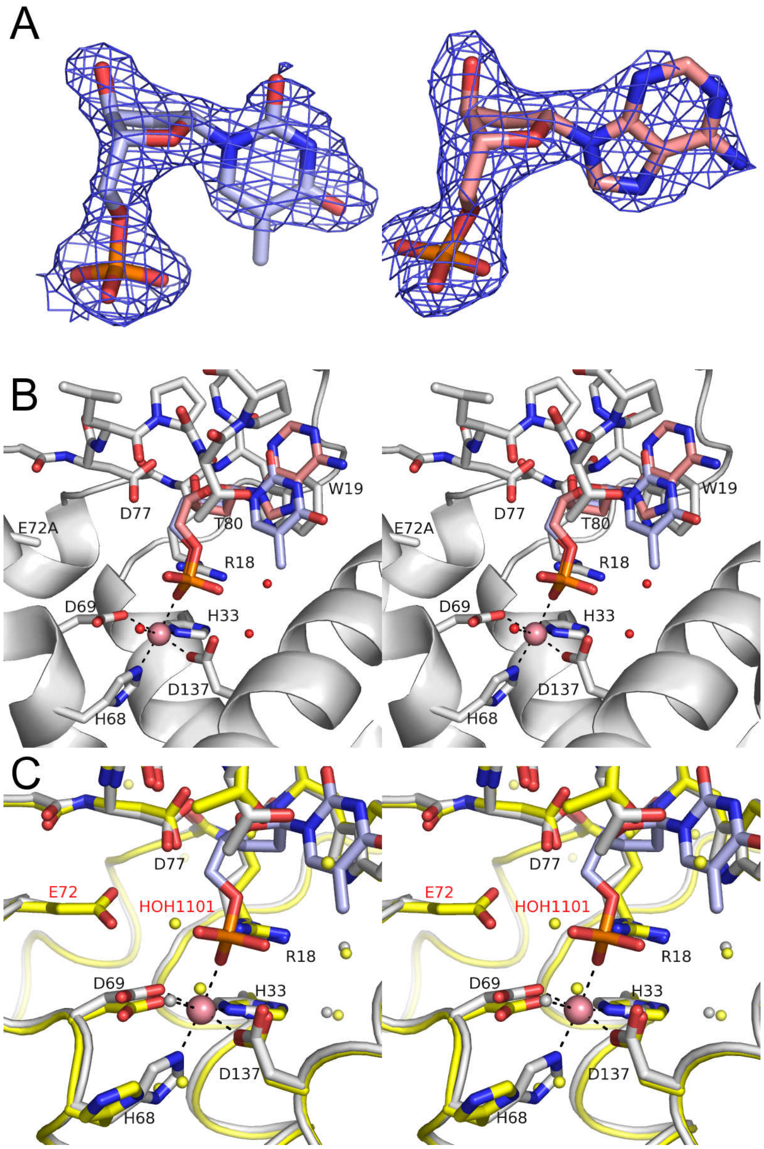
4.5 Mutant substrate-bound structures

Diffraction data on substrate-soaked YfbR E72A mutant crystals were also collected at APS beamline 19-ID. Reflection data were collected, indexed, integrated, and scaled using HKL-3000. Data for the dAMP-soaked crystal demonstrated merohedral twinning (Yeates, 1997). Twinning in general comes about due to the non-perfect nature of real-world crystals, where multiple crystal lattices are present. For most kinds of twinning, this is easy to detect, as multiple, distinct patterns of reflection spots may be seen on the oscillation image. However, if the different crystal lattice dimensions happen to coincide in all three dimensions, the reflections from both lattice patterns will also coincide, and the twinning is said to be merohedral (Yeates, 1997). In this case of the dAMP-soaked crystal, it was by the operation (*h*, -*h*-*k*, -*l*), with a twin fraction of 0.346 (meaning approximately 34.6% of the unit cells belong to one of the two lattice types). However, as long as the merohedral twinning is not perfect (i.e., with a twinning fraction significantly less than 0.5), this effect can be corrected by a simple mathematical transformation (Yeates, 1997). Corrected structure factors produced by the DETWIN program from the CCP4 distribution (CCP4, 1994) were used for structure solution and refinement.

Structures were solved using the molecular replacement mode of HKL-3000 (Minor *et al.*, 2006), which uses MOLREP (Vagin & Teplyakov, 1997). The wild-type structure, with waters removed and SeMet replaced by methionine, was used as a search model. The structures were iteratively improved by manual rebuilding with COOT, followed by maximum-likelihood restrained refinement with REFMAC5. 2-fold non-

Figure 4-6: YfbR nucleotide binding site.

(A) σ_A-weighted $2F_o$-F_c electron density maps for nucleotide substrate in the E72A-Co-TMP (blue) and E72A-Co-dAMP (red) structures. Maps are contoured at 1σ. (B) Stereo view of TMP (blue) and Co^{2+} (pink sphere) in the binding site of the E72A-Co-TMP structure (light gray). dAMP from the equivalent position in E72A-Co-dAMP is superimposed in pink. (C) Stereo view of the binding site in E72A-Co-TMP (light gray) superimposed on wild-type YfbR (yellow). The Co^{2+} and TMP substrates of E72A are represented as in (B). Water O atoms from E72A-Co-TMP and from wild-type YfbR are shown as gray and yellow spheres, respectively. Residue E72 and one of the bound waters (HOH1101) of wild-type YfbR are labeled in red.



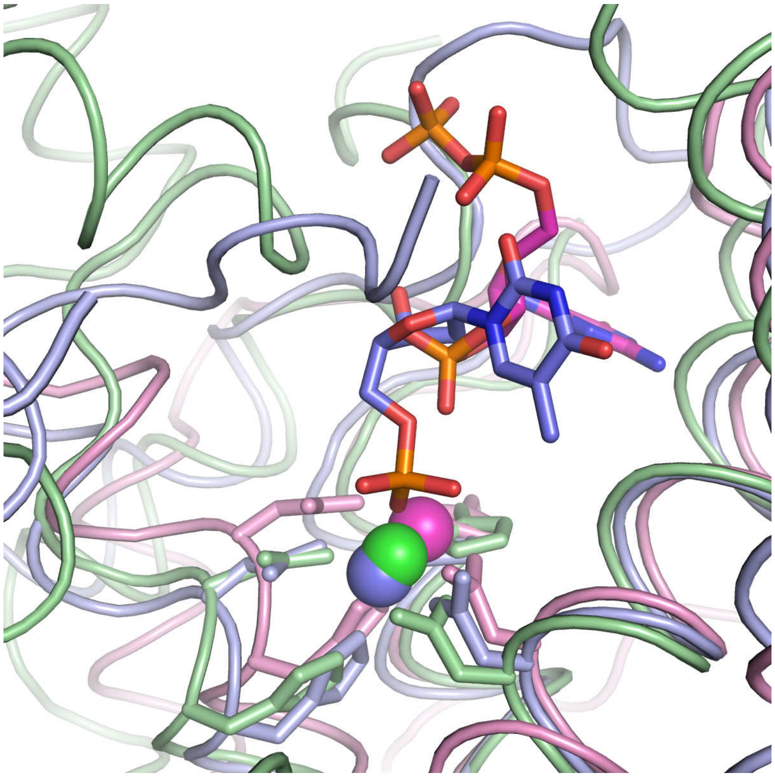
crystallography symmetry (NCS) and TLS restraints were used in refinement. The TLS model consisted of 14 groups, 7 per chain, identified as described above.

Seven catalytically inactive YfbR mutant proteins were passed through crystallization trials, and the E72A protein produced diffraction-quality crystals, in a crystal form similar to that of wild-type. Co-crystallizations of YfbR E72A with $CoCl_2$ and dNMPs failed, likely due to the chelating effect of the crystallization solution. However, by soaking crystals of the mutant for long periods of time (10-18 hours) with the substrates in a buffer where citrate was replaced by acetate, two datasets on two well-diffracting crystals were obtained with relatively high occupancy for both metal and nucleotide in the electron density. Two structures of YfbR E72A were solved by molecular replacement (using the structure of the wild type YfbR as the search model): one soaked with $CoCl_2$ and TMP (E72A-Co-TMP) and the other soaked with $CoCl_2$ and dAMP (E72A-Co-dAMP). The data collection and structure solution statistics of both structures are summarized in Table 4-1. The overall conformation of the substrate-soaked structures of YfbR E72A is very similar to the wild type YfbR, with an r.m.s.d. of 0.45 Å and 0.48 Å for E72A-Co-dAMP and E72A-Co-TMP, respectively. The occupancy of the nucleotide and Co^{2+} in each chain of E72A-Co-dAMP is 0.8 and 0.9, respectively, while there is essentially complete substrate occupancy in E72A-Co-TMP. $2F_o$-F_c maps are shown for both nucleotides in Fig. 4-6A.

In both structures, the metal-binding HD motif in each monomer contained a single strong peak of density in the $2F_o$-F_c maps (~13.5 σ for E72A-Co-TMP, ~12 σ for E72A-Co-dAMP), and residue H68 is reoriented to a position coordinating the metal cation in the binding site. An anomalous difference map calculated for a similar crystal

Figure 4-7: Superposition of three structures of HD-domain proteins.

Structures are shown with divalent cations and substrates (or natural inhibitors) bound. YfbR E72A complexed with Co^{2+} and TMP (this work) is shown in blue. S. equisimilis *RelA complexed with Mn^{2+} and guanosine-5'-diphosphate-2',3'-cyclic monophosphate (Hogg et al., 2004) is shown in magenta.* T. thermophilus *dNTPase complexed with Mg^{2+} (Kondo et al., 2007) is shown in green.*



soaked with $CoCl_2$ and diffracted at a wavelength of 1.28 Å confirmed there is a single metal ion per monomer, which is most likely Co^{2+}. In both structures, the metal ion is coordinated in a distorted octahedral configuration by the four residues of the HD motif (H33, H68, D69, and D137), a water molecule, and a phosphate oxygen of the bound nucleotide (Fig. 4-6B). The structures of *P. horikoshii* PH0347 and *P. furiosus* PF0395 each demonstrated the presence of a Ni^{2+} atom coordinated in an equivalent position by the four residues of the HD motif: H33, H67, D68, and D124 (the same numbering for both proteins).

The active states of the biochemically characterized HD domain enzymes *T. thermophilus* dNTPase (Kondo *et al.*, 2007) and *S. equisimilis* RelA bound with the natural inhibitor guanosine-5'-diphosphate-2',3'-cyclic monophosphate (Hogg *et al.*, 2004) are superimposed on the structure of E72A-Co-TMP in Fig. 4-7. In all three structures, the four residues of the HD motif, as well as the three α-helical fragments to which they belong, share a conserved structural architecture and bind divalent cations in equivalent positions. However, the structure of the residues involved in substrate recognition and catalysis are not conserved, either by sequence or in structure. Although all of the nucleotides seem to bind in the same general position relative to the metal ion, the ribose rings and nitrogenous bases all adopt significantly different orientations, and bind to non-conserved motifs. There is no substrate in the structure of dNTPase, but the residues predicted to be involved in substrate binding are not conserved (Kondo *et al.*, 2007).

4.6 Substrate binding and selectivity

In both monomers of both structures, nucleotide substrate is found adjacent to the metal binding site, and the nucleotides bind in similar configurations. The phosphate is coordinated to the Co^{2+} and also forms a hydrogen bond with the side-chain of conserved residue R18 (Fig. 4-6B). Mutation of this residue to alanine abolished activity of the enzyme, suggesting that the positively-charged R18 plays an important role in orienting the phosphate into the proper position. Additionally, the structures of both E72A-substrate complexes revealed that the YfbR active site can not accommodate more than one phosphate group of the substrate (Fig 4-6B). This explains why YfbR does not hydrolyze deoxyribonucleoside di- or triphosphates.

The 2'-deoxyribose rings of the nucleotides in E72A-Co-TMP and E72A-Co-dAMP all adopt C2'-endo configurations, the conformation preferred in B-DNA. Three H-bonds stabilize the ribose in this configuration: a strong hydrogen bond (2.5 Å) from the 3' hydroxyl to a carboxyl oxygen on D77, and two weaker hydrogen bonds from backbone nitrogens to the ribose: W19 N-H ... O3' and T80 N-H ... O4' (Fig. 4-6B). The D77A mutant was also catalytically inactive, likely due to the role D77 plays in binding the ribose of the nucleotide. In the structure of the human mitochondrial deoxyribonucleotidase dNT-2, the 3'-OH group of the deoxyribose is also coordinated by the hydrogen bonding to the side chain of an aspartate residue (D43) (Rinaldo-Matthis *et al.*, 2002). In contrast to this enzyme, YfbR does not hydrolyze deoxyribonucleoside 3'-monophosphates (Proudfoot *et al.*, 2004). In the active site of YfbR, the 3'-phosphate group would be positioned far away from the metal and the putative metal-coordinated catalytic water molecule making the reaction impossible. In addition, in both E72A-substrate structures, the 3'-O of ribose is located close to the side chain of P20 (3.8 – 4.0 Å), suggesting that YfbR will not bind these nucleotides.

The aromatic indole group of W19 makes substantial van der Waals contacts with the ribose and part of the nitrogenous base of the nucleotide. In particular, the plane of the indole is located approximately 3.6 Å from the 2' carbon atom of the deoxyribose (Fig. 4-6B). Due to the C2'-endo configuration of the deoxyribose, if there was a hydroxyl bound to the 2' carbon, the oxygen would be located approximately 2.5 Å from the plane of the indole group of W19.

This strongly suggests that the W19 side chain plays the major role in determining the selectivity of YfbR for 2'-deoxyribonucleotides over ribonucleotides. Unfortunately, the W19A mutant protein could not be purified, suggesting that a larger residue is required in this position for protein stability or solubility or both. The substrate selectivity of *E. coli* YfdR is similar to that of human deoxyribonucleotidases mdN and cdN (Fig. 4-5), and there is a Phe residue (F37) after the conserved R36 in its amino acid sequence (Fig. 4-1). The presence of a (slightly smaller) Phe side chain in the binding site of YfdR is the likely reason for the reduced selectivity of this protein toward deoxyribonucleotides.

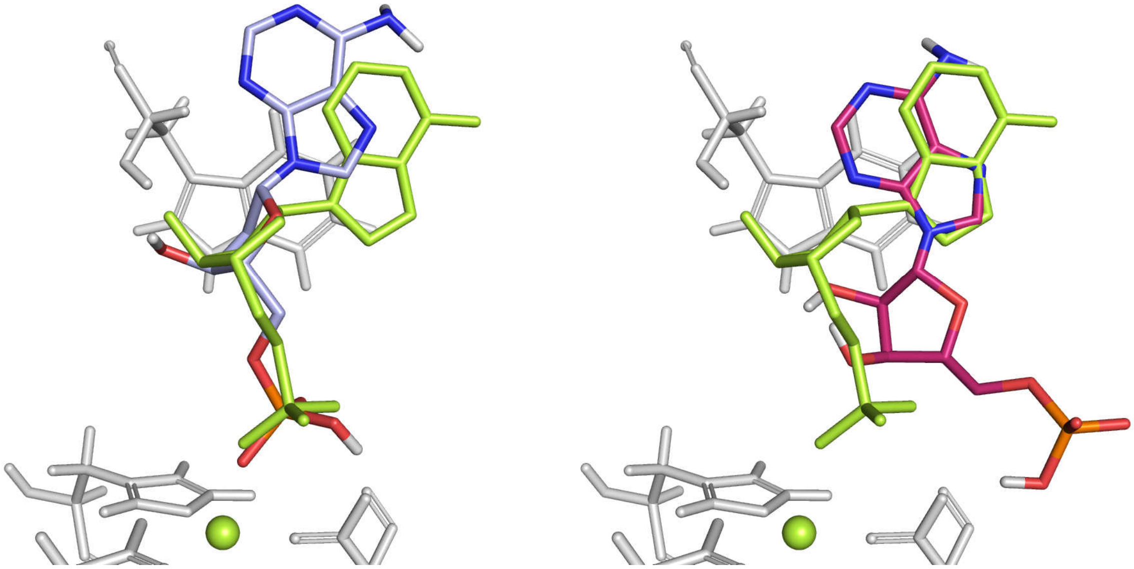
Human mitochondrial (mdN or dNT-2) and cytosolic (cdN or dNT-1) deoxyribonucleotidases are the only known 5'-nucleotidases that prefer the deoxyribo-form over the ribo-form of nucleoside 5`-monophosphates (Bianchi & Spychala, 2003; Wallden *et al.*, 2005). Structural studies of mdN revealed that its selectivity for deoxyribonucleotides is due to the presence of a hydrophobic pocket formed by F49, F102, and I133 near the 2'-carbon of the sugar (Wallden *et al.*, 2005). In the structures of the biochemically uncharacterized YfbR orthologs PF0395 and PH0347, the conserved tryptophan residue (W19 in PH0347) is positioned 3 Å further away from the substrate binding site due to the insertion of two amino acid residues after the conserved Arg (Fig. 4-1) suggesting that these proteins might have lower selectivity for deoxyribonucleotides. This architecture is also seen in the structure of AF1432, which had no measurable nucleotidase activity (see Section 4.4).

To further explore the ability of different nucleotides to bind to YfbR, a computational docking analysis was performed with the Autodock program (Morris *et al.*, 1998). Specifically, flexible models of dAMP and AMP were applied to the structure of YfbR E72A. The dAMP and PEG were manually removed from the structure of YfbR E72A bound with Co^{2+} and the nucleotide. The model was run through the REDUCE utility on the Molprobity (Lovell *et al.*, 2003) web server, which protonates the polypeptide chains. Separately, using the dAMP molecule in the B monomer of the E72A-Co-dAMP structure as a template, energy-minimized structures of protonated dAMP and AMP were generated. For both structures, one of the three oxygens on the phosphate was protonated.

The B monomer of the E72A-Co-dAMP structure and the nucleotide structures were prepared for docking analysis with Autodock (version 4) (Morris *et al.*, 1998). The dAMP and AMP structures were treated as flexible, with 7 and 8 torsions respectively. The E72-Co-dAMP monomer was treated as rigid. Appropriate docking parameters for Co^{2+} have not been determined, but a recent paper presents force-field parameters producing successful docking to Zn^{2+} and Mg^{2+} in AutoDock (Chen *et al.*, 2007).

Figure 4-8: Best docking solutions for dAMP and AMP.

Models of the best computational docking solutions for dAMP (left, in blue) and AMP (right, in red) to YfbR E72A. The metal ion and dAMP seen in the substrate-bound structure are shown in green. The metal-binding residues and W19 are shown as grey sticks.



Accordingly, the Co^{2+} ion was changed to a Mg^{2+} , and following the example of Chen and coworkers, a partial charge of -0.8*e* was used for the Mg^{2+} cation (Chen *et al.*, 2007). The precalculated grids for the protein structure were 91 by 91 by 91 elements, centered on the known nucleotide binding site, with the default grid spacing of 0.375 Å. The docking analyses for dAMP and AMP were done identically on the same grids, using the default Lamarckian genetic algorithm with default parameters.

A number of solutions for both nucleotides were obtained. The solutions with nucleotide conformations closest to the correct binding conformation observed in the soaked structures are shown in Fig. 4-8. The closest binding solution for dAMP (left) binds both the phosphate and ribose in essentially the proper mode and orientation, with only the nitrogenous base rotated about 30º out the position observed in the crystal structure. Autodock estimates a free energy of binding of -8.73 kcal/mol for this computational solution, with an RMSD of 1.5 Å to the known reference structure. None of the AMP solutions bind in the proper mode or orientation, and the closest solution (right) has an estimated free energy of binding of -6.54 kcal/mol, with an RMSD of 2.9 Å to the known reference structure.

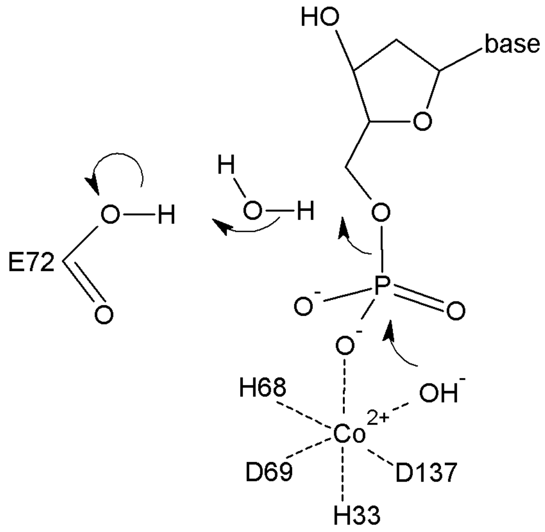
In both complex structures, the nucleotide bases of both TMP and dAMP adopt an *anti* conformation, and they are widely exposed to bulk solvent. No hydrogen bonds or other specific contacts are made to the N or O atoms on the edge of the aromatic base typically involved in nucleotide recognition. The binding patterns of both molecules (a purine and a pyrimidine) are very similar and for both nucleotides the polar atoms on the aromatic base are exposed to solvent. This lack of specific contacts for the nitrogenous bases of both nucleotides is consistent with the lack of overall specificity for particular 5'-deoxyribonucleotides observed for YfbR.

4.7 Catalytic mechanism of YfbR

Virtually all enzymatic substitution reactions of phosphates proceed via an in-line attack by a nucleophile, where the entering and leaving groups are located on opposite sides of the phosphorus atom (Frey, 1982). In the E72A-Co-TMP and E72A-Co-dAMP structures, the most likely candidate for the nucleotide is an activated water molecule coordinated to the Co^{2+}. This model is supported by a strict metal dependence of the YfbR nucleotidase activity demonstrated previously (Proudfoot *et al.*, 2004). While a hydroxide ion and a water molecule cannot be directly distinguished by electron density at 2.1 Å resolution, the shortened Co^{2+}—O bond distance of 1.9 Å for the enzyme suggests that the moiety is in hydroxide form. A metal-coordinated hydroxide ion is also proposed to be the nucleophile in the mechanism of catalysis by the eukaryotic phosphodiesterase PDE4 (Huai *et al.*, 2003; Xu *et al.*, 2004).

The proton for the leaving 5'-O atom of the deoxynucleoside may come from the side chain of the conserved E72, which is essential for YfbR activity (Fig. 4-5). This glutamate is conserved in the HD domain proteins from clusters 2 (YGK1 family) and 3 (RelA/SpoT family) (Aravind & Koonin, 1998). Since the carboxylate group of E72 is located ~6 Å from the nucleotide binding site, and cannot interact directly with the substrate, we propose that the water molecule (seen in the free-state crystal structure and marked as HOH1101 in Fig. 4-6C) acts as a bridge for proton transfer from E72 to the leaving group O atom. There are no other ordered water molecules or residue side chains capable of donating a proton to the leaving oxygen in the active site of YfbR. Thus, in the proposed mechanism of catalysis by YfbR (shown in Fig. 4-9), a hydroxide ion bound to

Figure 4-9: Proposed reaction mechanism for 5'-deoxynucleotidase activity.



the Co^{2+} atom makes a nucleophilic attack on the phosphate of the bound nucleotide, and residue E72, mediated through a water molecule, donates a proton to the dephosphorylated O5' group of the deoxyribose.

Since the phosphate conformation is rotated slightly away from the optimal position for nucleophilic attack by a Co^{2+}-bound hydroxide ion, the possibility that another nucleophile is involved in the reaction cannot be excluded. For example, the oxygen of the carboxylate group of D137 not coordinated to the metal could serve as the nucleophile, forming a covalent phosphoaspartyl intermediate. The phosphate then would be liberated by a second substitution reaction with water. Such a two-step mechanism has been proposed for human deoxynucleotidase (Knofel & Strater, 2001; Himo *et al.*, 2005). Alternatively, the water bound to E72 could serve as the nucleophile rather than as the proton donor, as it was suggested for the conserved E81 in the *S. equisimilis* RelA (Hogg *et al.*, 2004). However, in YfbR this mechanism would require a kinetically slow pseudorotation of the hypervalent phosphorus intermediate (Westheimer, 1968), which is rarely observed in phosphohydrolases or phosphotransferases. Additional functional and structural experiments with active state analogs are necessary to determine the mechanism conclusively.

5 Conclusions

Has Xtaldb, the crystallization expert system for the design, tracking, and analysis of crystallization experiments described above, been a success? Xtaldb is now in practical, everyday use at the University of Virginia. Out of an original set of uncharacterized proteins that failed to yield diffraction-quality crystals, 87% have SeMet crystals that diffract to 3.5 Å or better, and the 3-D structures of 50% have been solved. The system has been used to track >40,000 crystallization drops of 23 different proteins, and its use has resulted in 16 PDB deposits and three full structural papers, with more to come.

Xtaldb provides many searching and graphical tools for real-time quantitative analysis of crystallization experiments. The system tracks images and annotations of crystallization drops, is capable of representing many different kinds of crystallization experiments, and links that information to preparations of protein and chemical reagent information. Xtaldb is integrated with several different kinds of hardware for accurate capture of crystallization data. The system also actively tracks information from up- and down-stream in the macromolecular crystallization pipeline, by serving as the crystallization component of the LabDB protein crystallography LIMS system.

Xtaldb also provides tools for designing experiments that search crystallization space efficiently. The system provides a mechanism for generating efficient random screens. The module for producing random screens produces well-balanced designs suitable for linear regression analysis and is, to the author’s knowledge, the fastest and most robust program available for the task.

It is significant that the initial set of test proteins for Xtaldb were more-or-less uncharacterized bacterial proteins that failed to yield diffracting crystals in a high-throughput crystallography pipeline. Since as in many structure determination projects, the proteins were over-expressed in a recombinant expression system, and purified by various chromatography techniques, this protein preparation information was included into the database for analysis. This information helped suggest the different techniques for optimization that were then employed to produce the crystals we obtained. Most importantly, the structures of the uncharacterized proteins solved resulted in more extensive explorations of structure-function relationships. The role of TM0549, as a putative regulatory subunit for acetohydroxyacid synthase was explored (Petkowski *et al.*, 2007). Stuctural studies of TM1030 led to a plausible mechanism of DNA-binding for this member of the TetR transcriptional regulator family (Koclega *et al.*, 2007). Finally, YfbR, a member of the large, widely-conserved HD-domain phosphohydrolase superfamily, was analyzed in detail.

The wild-type structure of YfbR was solved, and then the expected phosphohydrolase activity of the enzyme was confirmed. A collaborator determined that YfbR was a 5’-deoxyribonucleotidase with a novel pattern of selectivity, and generated catalytically inactive alanine substitution mutants. These mutants were screened for crystallization, and crystals of the YfbR E72A mutant soaked with Co^{2+} and either TMP or dAMP were solved, with divalent cation and substrate nucleotide ordered in the structure. The binding mode of the substrate of the enzyme agrees with the pattern of selectivity observed in the biochemical assays. Bioinformatic analysis of homologs and a computation docking analysis with the substrate dAMP and the non-substrate AMP provide further confirmation of the selectivity of the novel nucleotide binding site. Alignment of the wild-type structure with the substrate-bound mutants strongly suggests a plausible mechanism for phosphohydrolysis. This mechanism differs from the one suggested for PDE, and is the first mechanism for a HD-domain phosphohydrolase derived directly from a substrate-bound crystal structure.

The substrate binding mode demonstrated here differs significantly from those proposed for other members of the HD-domain phosphohydrolase superfamily. The known substrates of the HD-domain hydrolases include such various nucleotides as 5`- and 2`-deoxyribo- and ribonucleoside monophosphates, (p)ppGpp, (d)NMPs, and 2`,3`-cNMPs (Seto *et al.*, 1988; Kondo *et al.*, 2004; Hogg *et al.*, 2004; Proudfoot *et al.*, 2004). Thus, the members of the HD-domain superfamily have developed different modes of substrate binding while retaining a similar structural fold and using the same metal binding motif to catalyze a phosphohydrolase reaction. In summary, structure-function analysis of YfbR resulted in observation of plausible molecular mechanisms for both the novel mechanism of substrate specificity and for catalysis of the enzyme.

In the field of quantitative analysis of crystallization experiments, no clear consensus has been reached about what general conclusions may be drawn from the macromolecular crystallization process, such as is illustrated by the debate over the correlation of crystallization pH to pI. Crystallization of biological macromolecules is a complicated process, and as some have noted, there is no silver bullet. Rupp and Wang relate a quote by locomotive engineer Karl Gölsdorf (1861–1916): “There is no single place on a steam engine where you can save a ton, but a thousand places where you can save a kilogram (Rupp & Wang, 2004).” In many cases, the conclusions drawn by quantitative analysis of crystallization experiments will be localized and specific to the particular crystallization problem at hand.

However, there are significant questions in the global analysis of crystallization experiments that Xtaldb should be able to address in the future. One is the process of determining which parameters of a crystallization experiment have the most significant effect on the experimental outcome, or which methods for crystallization are the most successful. Crystallographers are routinely warned that many possibly trivial parameters can affect crystallization (see Section 1.2), which might make it appear that the act of producing macromolecular crystals is difficult to impossible. Yet the nearly 40,000 X-ray diffraction structures in the PDB are clear evidence that despite the apparent complexity of the macromolecular crystallization problem, in many cases it can be practically solved. In truth, it is likely that many of the parameters of a crystallization experiment have only trivial effects on the outcome compared to others. We simply don’t know in any objective way which parameters belong in which category, or which crystallization methods are most effective.

A necessary prerequisite for conducting such an analysis is collecting as much information as possible about crystallization experiments, in a controlled (and preferably semi-automated) manner. Xtaldb is uniquely suited for this task, due both to the tools it provides for designing and observing experiments as well as its interfaces to laboratory hardware. As more experiments are designed and results are collected with the system, the analyses made with the system will increase in statistical significance. In the future, the set of complete, or nearly complete, data harvested with the system will allow robust analysis of the relative importance of the parameters involved in protein crystallization, using methods like factor analysis. Such an analysis should significantly increase the global success rate of crystallization by determining quantitatively, rather than qualitatively, which methods of producing macromolecular crystals are most effective.

Appendix A: Small-molecule structures of HD-domain protein and PDE ligands

A.1 Introduction

Small-molecule crystallography is mathematically a more complicated problem than macromolecular crystallography. However, the limited number of atoms in the unit cell and the higher diffraction limit (usually around 0.7 Å) makes the solution of small molecule structures a straightforward process, since the data to parameter ratio is much larger. In general, small-molecule structures can be solved by “direct methods”, which essentially solves the phase problem present in crystallography, by allowing structure solution from a single diffraction amplitude data set. Direct methods represent one of the most significant advances in the history of crystallography, and Herbert A. Hauptman and Jerome Karle were awarded the Nobel Prize in Chemistry in 1985 for developing the mathematical theory of direct methods.

Direct methods are techniques for calculating phases *ab initio* from structure factor amplitude data. They rely upon known inequalities between groups of three (or more) structure factors, treated probabilistically, as constraints to refine a set of randomly generated phases toward the correct ones (Giacovazzo, 2001). Only the strongest reflections should be used, as the inequalities are closest to equalities as the reflections increase in intensity. However, the method requires that the diffraction limit be of atomic resolution or better (≤ 1.2 Å) and that the number of non-hydrogen atoms to be found to be 600 or less (Sheldrick *et al.*, 2006), though there has been some success with molecules with as many as 2000 non-hydrogen atoms (Frazão *et al.*, 1999).

Because of the larger data to parameter ratio, the atoms in a small-molecule structure can be modeled with anisotropic temperature factors. Unlike isotropic temperature factors, which assume that the thermal motion of the atom is uniform in all directions (see Chapter 1), the anisotropic model models the motion as an ellipsoid, which more accurately reflects the vibration of the bonded atom. The ellipsoid is represented by six parameters, three radii and three orientation angles, as opposed to a single parameter for the isotropic model. This increase in the number of parameters is permitted due to the atomic resolution of the data and the much larger data-to-parameter ratio as compared to macromolecular crystallography. The atomic resolution of the data also allows the direct modeling of hydrogen atoms, which are usually omitted in macromolecular crystallography, though the signal for the atoms is weak due to the very low electron scattering factor for hydrogen.

Small-molecule structures are typically reported with three values for model-to-data fit, as generated by the crystallographic refinement program SHELXL (Sheldrick & Schneider, 1997). The first is a conventional *R* value, analogous to the *R* and R_{free} used in macromolecular refinement:

$$R_1 = \frac{\sum \left\| F_o \right| - \left| F_c \right\|}{\sum \left| F_o \right|} .$$

This value is reported for all reflections and for only those reflections where $|F_o| > 4\sigma$. However, a weighted R factor, which makes use of intensities ($|F|^2$) instead of structure factor amplitudes ($|F|$), is also reported,

$$wR_2 = \left(\frac{\sum w(F_o^2 - F_c^2)^2}{\sum w(F_o^2)^2} \right)^{\frac{1}{2}},$$

as well as a goodness-of-fit parameter *S*, defined as

$$S = \left(\frac{\sum w(F_o^2 - F_c^2)^2}{n - p} \right)^{\frac{1}{2}},$$

where *n* is the number of reflections and *p* the number of parameters (Sheldrick & Schneider, 1997). Like the conventional R_1 value, the fit improves as wR_2 decreases. The ideal value for the goodness-of-fit parameter *S* is unity. The weighting parameter *w* in both equations is given by:

$$w = \frac{1}{\sigma^2(F_o^2) + \left(a \frac{2F_c^2 + F_o^2}{3} \right)^2 + b \frac{2F_c^2 + F_o^2}{3}}.$$

The parameters *a* and *b* are automatically refined by SHELXL to flatten the analysis of variance (Sheldrick & Schneider, 1997).

A.2 Methods and results

The production of small-molecule crystals is significantly simpler than crystals of biological macromolecules, and is performed by simple evaporation. Two ligands of HD-domain proteins, disodium 4-nitrophenylphosphate (pNPP-Na) and zardaverine, were obtained from commercial sources. pNPP-Na was dissolved into a solution of 70% ethanol in water, and the solvent was then allowed to evaporate at room temperature in a fume hood for approximately 1 week. The evaporation flask was wrapped in aluminum foil to prevent light-induced breakdown of the compound. A 200×200×150 μm crystal was attached with a small bead of glue to a pin and mounted on the goniometer of a Rigaku R-AXIS RAPID diffractometer, using Mo K_α radiation for data collection. Zardaverine was produced by dissolving the compound in methanol, which was allowed to evaporate at room temperature in a fume hood. A 90×190×510 μm crystalline plate was obtained, which was attached to a pin and mounted on the goniometer of the Rigaku R-AXIS RAPID diffractometer, which at the time had a sealed tube producing Cu K_α radiation.

Diffraction data were collected, indexed, integrated, and scaled with a version of HKL-3000 (Minor *et al.*, 2006), called HKL-3000_sm, which is integrated with SHELX-97 for small-molecule structure solution. After the data are scaled, the structure data is passed to the structure solution program SHELXS (Sheldrick, 1997), which uses direct methods to generate an initial density map. HKL-3000_sm provides a three-dimensional OpenGL interface to dynamically add, reassign, or delete atoms, followed by a cycle of refinement with the crystallographic refinement program SHELXL (Sheldrick &

Table A-1: Small molecule X-ray data collection and refinement parameters.

	pNPP-Na	*Zardaverine*
Data collection		
Radiation type	Mo Kα	Cu Kα
Wavelength (λ)	0.709 Å	1.54 Å
Spacegroup	$P2_1/c$	$P2_1/n$

Unit cell parameters	a=19.007(1) Å,
	b=11.982(1) Å,
	c=13.492(1) Å,
	β=103.708(1)°
		a=7.239(1) Å,
		b=15.838(1) Å,
		c=10.099(1) Å,
		β=91.991(6)°

Diffraction angle (θ) range	2.0° – 40.3°	5.2° – 72.3°
Resolution range	0.55 Å – 10 Å	0.80 Å – 8.5 Å
Refinement		
Reflections collected	130337	41088
Num. of reflections (n)	18300	2228
Num. of reflections ($I > 2\sigma$)	12628	2027
R_1	0.048	0.035
wR_2	0.146	0.093
S	1.10	1.04
Model parameters (p)	397	213
Weighting parameters	a=0.0727, b=0.9368	a=0.0585, b=0.1893

Schneider, 1997). The data collection and refinement statistics are shown in Table A-1.

Both model used anisotrophic temperature factors, which are represented in the figures as displacement ellipsoids, where the ellipsoid encloses the 50% probability level for the average position of the atom. All figures of models with displacement ellipsoids were generated using the program ORTEP-3 (Burnett & Johnson, 1996; Farrugia, 1997). The models also have hydrogen atoms, and in both cases some electron density was seen for the hydrogens. However, due to the weakness of the hydrogen electron density, for purposes of refinement the hydrogens were simply allowed to “ride” on their bonded atoms at fixed ideal distances. The hydrogens were refined with isotropic temperature factors. In the case of pNPP-Na, which did have ordered solvent molecules, the positions of the riding hydrogens were determined by optimizing the H-bonding network with the CALC-OH program (Nardelli, 1999).

A.3 Crystal structure of pNPP-Na

pNPP has been used for decades as a marker of phosphatase activity, as described in Chapter 4. pNPP was originally employed clinically in the detection of alkaline phosphatase activity in blood serum (Bessey *et al.*, 1946) and is used today as a general indicator of phosphohydrolase activity. The anion is a component of a set of general enzymatic screens against libraries of proteins of unknown function (Kuznetsova *et al.*, 2005), and was used to originally determine the function of YfbR as a phosphohydrolase (Chapter 4).

The anion is commercially available in the dianion form as a hexahydrate disodium salt. Jones and coworkers reported the structure using

Figure A-1: Contents of asymmetric unit of the crystal structure of pNPP. Non-hydrogen atoms are shown as displacement ellipsoids at the 50% probability level. H atoms are shown as spheres of arbitrary radius. Na-O contacts are shown as dashed lines.

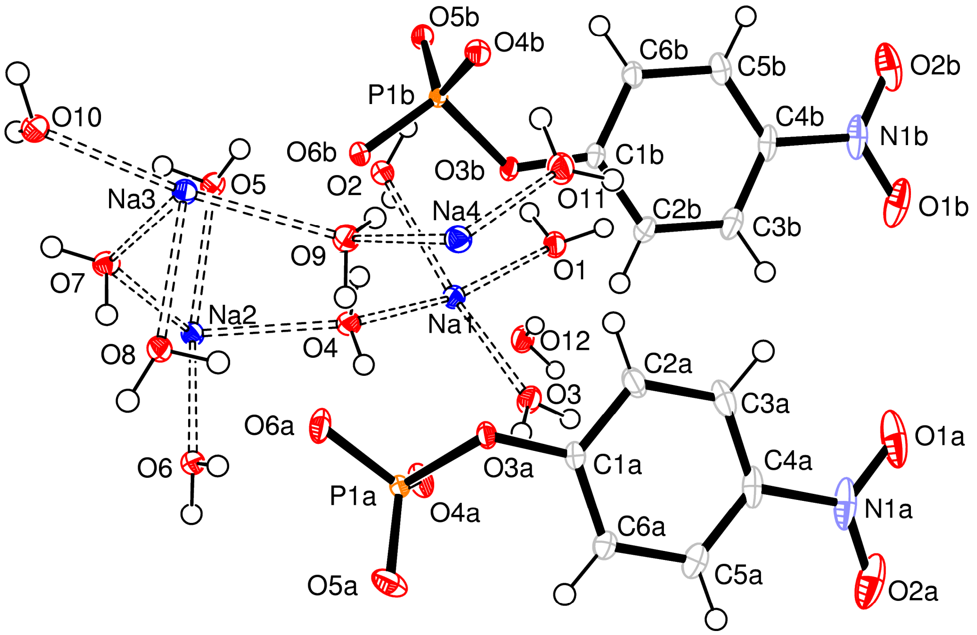
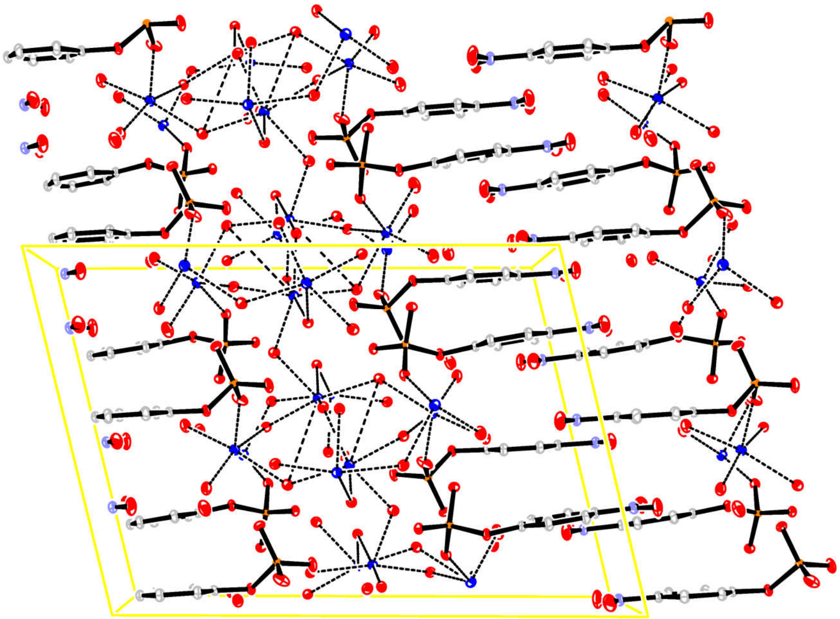


Figure A-2: Crystal packing in the structure of pNPP.

The unit cell is shown as a yellow box. Na-water contacts are shown as dashed lines. Hydrogens are omitted for clarity.



bis(cyclohexylammonium) as the countercation, but were unable to crystallize the disodium salt (Jones *et al.*, 1984b). The structure of the disodium 4-nitrophenylphosphate hexahydrate salt was solved at 103 K to a resolution of 0.55 Å (Fig. A-1). The structure of the sodium salt with a different hydration state was later reported (Kuczek *et al.*, 2007). The data collection and refinement statistics are shown in Table A-1. There are two 4-nitrophenylphosphate anions in the asymmetric unit of the crystal structure. While the 4-nitrophenyl groups of both anions are nearly identical in structure, the phosphate groups differ in conformation, in both the C2—C1—O3—P1 torsion angle [anion A: 137.96 (9)º; anion B: 158.52 (9)º] and the C1—O3—P1—O4 torsion angle [for A, 49.26 (10)º; for B, 43.42 (10)º].

In the crystal structure of pNPP-Na, alternating non-polar and hydrophilic layers are observed (Fig. A-2). The non-polar regions consist of tightly packed nitrophenyl groups, while the hydrophilic layers contain the phosphate ions, sodium ions and their bound water molecules.

The crystal packing for pNPP-Na differs from that of the bis(cyclohexylammonium) salt (Jones *et al.*, 1984b). In the latter, the cyclohexylammonium cations interdigitate between the nitrophenyl groups, while in the former the aromatic rings stack directly. Jones and coworkers report that the length of the CO—P bond in their structure of the 4-nitrophenylphosphate dianion is significantly longer [1.664 (5) Å] than those of other dianion alkylphosphates, with a mean of 1.614 (4) Å (Jones et al., 1984a). The measured CO—P distances in the sodium salt structure [P1A—O3A = 1.6461 (8) Å and P1B—O3B = 1.6519 (8) Å] are not significantly different from that found in the bis(cyclohexylammonium) salt structure. Our measure of the mean CO—P—O$^-$ angle of 105.2º in (I) is significantly smaller than both the mean of 114.0º for the bis(cyclohexylammonium) salt and the reported mean of 112.8º for dianion alkylphosphates (Jones *et al.*, 1984b). Finally, the bis(cyclohexylammonium) salt shows significantly shorter lengths for the remaining three P—O bonds compared with the means (1.514, 1.519 and 1.510 Å) for 22 other dianion alkylphosphates (Jones *et al.*, 1984a). In contrast, our values for these bonds do not differ significantly from the reported mean values. While the solved structure follows the trend for CO—P bond lengths as a function of R—OH pKa as explored by Jones and coworkers, the pKa of the 4-nitrophenyl group alone is not sufficient to explain the differences in P—O$^-$ bond length or CO—P—O$^-$ bond angles. Neighboring cations and water molecules also influence the geometry of the phosphate group.

A.4 Crystal structure of zardaverine

6-(4-difluoromethoxy-3-methoxyphenyl)-3(2*H*)-pyridazinone (zardaverine) selectively inhibits PDE3 and PDE4 (Rabe *et al.*, 1994).. Zardaverine, along with other dialkoxyphenol-containing compounds, have been examined as possible treatments for acute respiratory failure (Rabe *et al.*, 1994; Schermuly *et al.*, 2003; Schmidt *et al.*, 2000). Zardaverine has also been used as a template compound to computationally generate a library of PDE4 inhibitors (Krier *et al.*, 2005).

Two protein structures of zardaverine complexed with PDE4 type D have been reported, one structure solved to a resolution of 2.9 Å (Lee *et al.*, 2002), and the other to 1.54 Å (Card *et al.*, 2004). In the higher resolution structure of PDE4D-zardaverine, the

Figure A-3: The asymmetric unit of the crystal structure of zardaverine.
Non-hydrogen atoms are shown as displacement ellipsoids at the 50% probability level. H atoms are shown as spheres of arbitrary radius.

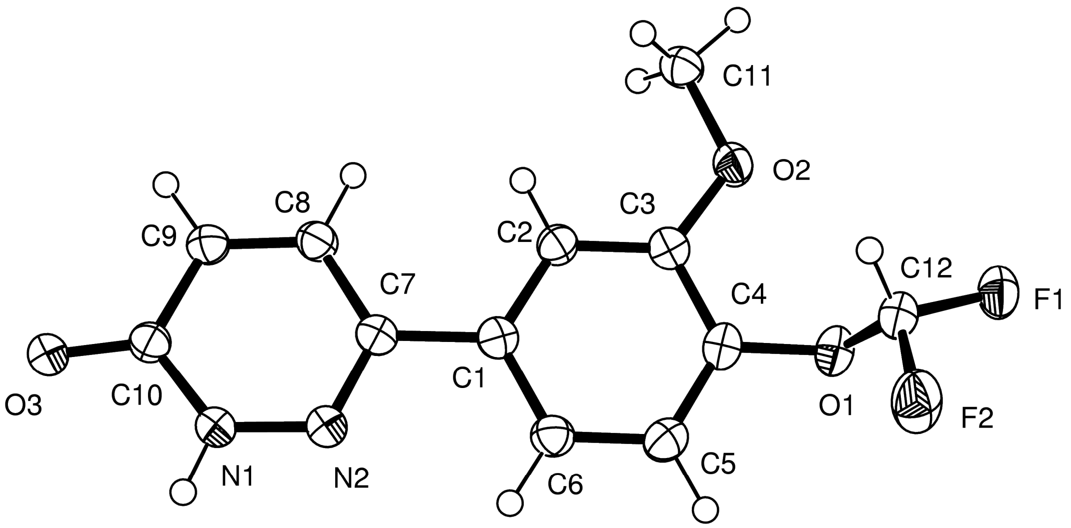
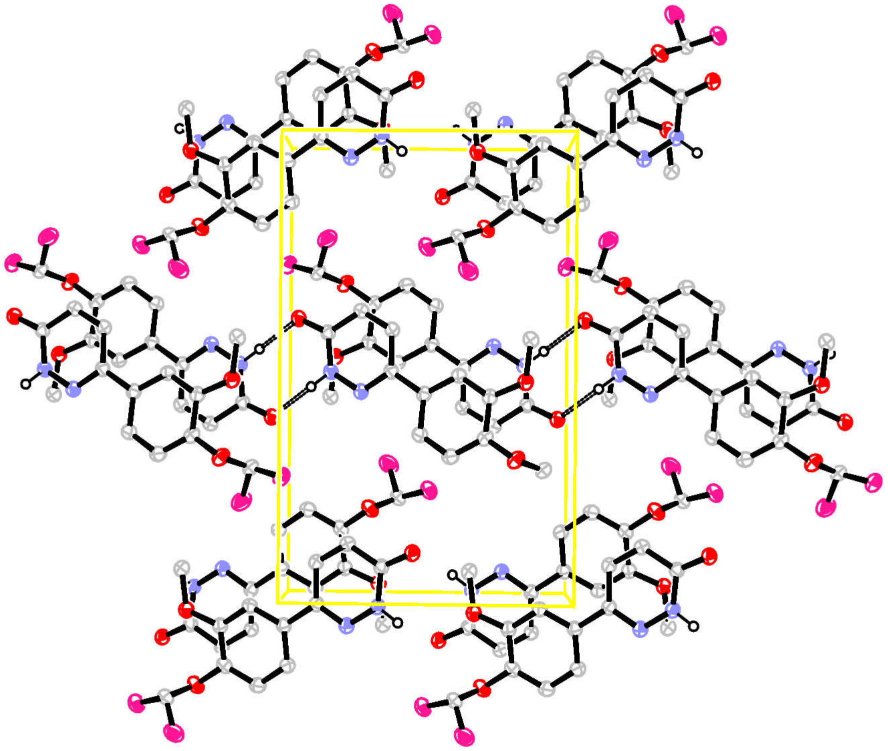


Figure A-4: Crystal packing of zardaverine.

The hydrogen bonds involved in dimer formation are shown as dashed lines.



compound is found in three different conformations: A, B, and C, with occupancies of 47%, 33%, and 20%, respectively. In all three conformers, the two rings are not co-planar. The torsion angles along C6—C1—C7—N2 are 39.6° for A, 146.6° for B, and 165.1° for C. The A and B conformers are essentially equivalent except that the pyridazinone ring is flipped along the C1—C7 bond. This is possible because the pyridazinone ring is not anchored by H bonds to the protein.

The small-molecule structure of zardaverine was solved at 103 K, which is shown in Fig. A-3. The data collection and refinement parameters are shown in Table A-1. In contrast to the zardaverine models in the PDE4D structures, in the small-molecule structure, the N1—H1···O3 (*x*, *y*+1, *z*+1) H-bond stabilizes the pyridazinone ring in a single conformation.The crystal packing ensures that the two rings of the compound are essentially co-planar, with a C6—C1—C7—N2 torsion angle of 4.9(2)°, as the molecules are packed in nearly planar layers, as shown in Figure A-4.

In the small-molecule structure, dimers of zardaverine are observed. The H1 atom (which is the only hydrogen in the structure capable of strong hydrogen bonding) forms a H bond with O3(*x*, *y*+1, *z*+1) of a symmetry-related molecule. This hydrogen bond is nearly ideal with an H1···O3 distance of 1.887 Å and an N1—H1···O3 angle of 174.04°. The N1—H1···O3 H-bonds are responsible for dimer formation. The O3 atom is also involved in a short contact interaction with hydrogen atom H12 (*x*+1, *y*+1, *z*+1), where the O3···H12 distance is 2.23(1) Å, and the O3···H12—C12(*x*+1, *y*+1, *z*+1) angle is 163(1)°.

References

Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C. & Fontecilla-Camps, J. C. (1991). *J Biol Chem* **266**, 20131-20138.

Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. (2000). *Nat Struct Biol* **7 Suppl**, 973-977.

Abrahams, J. P. & Leslie, A. G. (1996). *Acta Cryst D* **52**, 30-42.

Amin, A. A., Faux, N. G., Fenalti, G., Williams, G., Bernadou, A., Daglish, B., Keefe, K., Middleton, S., Rae, J., Tetis, K., Law, R. H., Fulton, K. F., Rossjohn, J., Whisstock, J. C. & Buckle, A. M. (2006). *Proteins* **62**, 4-7.

An, G., Justesen, J., Watson, R. J. & Friesen, J. D. (1979). *J Bacteriol* **137**, 1100-1110.

Aravind, L. & Koonin, E. V. (1998). *Trends Biochem Sci* **23**, 469-472.

Audic, S., Lopez, F., Claverie, J. M., Poirot, O. & Abergel, C. (1997). *Proteins* **29**, 252-257.

Baykov, A. A., Evtushenko, O. A. & Avaeva, S. M. (1988). *Anal Biochem* **171**, 266-270.

Beitz, E. (2000). *Bioinformatics* **16**, 135-139.

Bender, A. T. & Beavo, J. A. (2006). *Pharmacol Rev* **58**, 488-520.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst D* **58**, 899-907.

Bertani, G. (1951). *J Bacteriol* **62**, 293-300.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A. M., Arrowsmith, C. H., Montelione, G. T. & Gerstein, M. (2001). *Nucleic Acids Res* **29**, 2884-2898.

Bessey, O. A., Lowry, O. H. & Brock, M. J. (1946). *J Biol Chem* **164**, 321-329.

Bianchi, V., Pontis, E. & Reichard, P. (1986). *Proc Natl Acad Sci U S A* **83**, 986-990.

Bianchi, V. & Spychala, J. (2003). *J Biol Chem* **278**, 46195-46198.

Bitto, E., Bingman, C. A., Wesenberg, G. E., McCoy, J. G. & Phillips, G. N., Jr. (2006). *J Biol Chem* **281**, 20521-20529.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003). *Acta Cryst D* **59**, 2023-2030.

Brinkmann, U., Mattes, R. E. & Buckel, P. (1989). *Gene* **85**, 109-114.

Brunger, A. T. (1992). *Nature* **335**, 472--475.

Brunger, A. T. (1997). *Methods Enzymol* **277**, 366-396.

Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst D* **54**, 905-921.

Burnett, M. N. & Johnson, C. K. (1996). Report ORNL-6895. Oak Ridge National Laboratory.

Canaves, J. M., Page, R., Wilson, I. A. & Stevens, R. C. (2004). *J Mol Biol* **344**, 977-991.

Card, G. L., England, B. P., Suzuki, Y., Fong, D., Powell, B., Lee, B., Luu, C., Tabrizizad, M., Gillette, S., Ibrahim, P. N., Artis, D. R., Bollag, G., Milburn, M. V., Kim, S. H., Schlessinger, J. & Zhang, K. Y. (2004). *Structure* **12**, 2233-2247.

Carter, C. W., Jr. (1997). *Methods Enzymol.* **276**, 75-99.

Carter, C. W., Jr. (1999). *Crystallization of nucleic acids and proteins: a practical approach*, edited by A. Ducruix & R. Giegé, pp. 75-120. Oxford: Oxford University Press.

Carter, C. W., Jr., Baldwin, E. T. & Frick, L. (1988). *J Cryst Growth* **90**, 60-73.

Carter, C. W., Jr. & Carter, C. W. (1979). *J Biol Chem* **254**, 12219-12223.

Carter, C. W., Jr. & Yin, Y. (1994). *Acta Cryst D* **50**, 572-590.

CCP4 (1994). *Acta Cryst D* **50**, 760-763.

Chatterji, D. & Ojha, A. K. (2001). *Curr Opin Microbiol* **4**, 160-165.

Chayen, N. E. (1996). *Protein Eng* **9**, 927-929.

Chayen, N. E. & Saridakis, E. (2001). *J Cryst Growth* **232**, 262-264.

Chen, D., Menche, G., Power, T. D., Sower, L., Peterson, J. W. & Schein, C. H. (2007). *Proteins* **67**, 593-605.

Codd, E. F. (1960). *Communications of the ACM* **13**, 377-387.

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

Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. (1994). *Acta Cryst D* **50**, 414-423.

Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst D* **58**.

Del Tito, B. J., Jr., Ward, J. M., Hodgson, J., Gershater, C. J., Edwards, H., Wysocki, L. A., Watson, F. A., Sathe, G. & Kane, J. F. (1995). *J Bacteriol* **177**, 7086-7091.

Doublie, S. (1997). *Methods Enzymol* **276**, 523-530.

Dunlop, K. V. & Hazes, B. (2005). *Acta Cryst D* **61**, 1041-1048.

Durbin, S. D. & Feher, G. (1996). *Annu Rev Phys Chem* **47**, 171-204.

Elkin, C. D. & Hogle, J. M. (2001). *J Cryst Growth* **232**, 563-572.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst D* **60**, 2126-2132.

Farr, J., R. G., Perryman, A. L. & Samudzi, C. T. (1998). *J Cryst Growth* **183**, 653-668.

Farrugia, L. J. (1997). *J Appl Cryst* **30**, 565.

Fisher, R. A. (1951). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Frazão, C., Sieker, L., Sheldrick, G., Lamzin, V., LeGall, J. & Carrondo, M. A. (1999). *J Biol Inorg Chem* **4**, 162-165.

Frenois, F., Engohang-Ndong, J., Locht, C., Baulard, A. R. & Villeret, V. (2004). *Mol Cell* **16**, 301-307.

Frey, P. A. (1982). *Tetrahedron* **38**, 1541-1567.

Fulton, K. F., Ervine, S., Faux, N., Forster, R., Jodun, R. A., Ly, W., Robilliard, L., Sonsini, J., Whelan, D., Whisstock, J. C. & Buckle, A. M. (2004). *Acta Cryst D* **60**, 1691-1693.

George, A. & Wilson, W. W. (1994). *Acta Cryst D* **50**, 361-365.

Giacovazzo, C. (2001). *International Tables of Crystallography* **B**, 210-234.

Giegé, R. & Ducruix, A. (1992). *Crystallization of Nucleic Acids and Proteins: A Practical Approach*, edited by A. Ducruix & R. Giegé, pp. 1-18. Oxford, Great Britain: Oxford University Press.

Gilliland, G. L. (1988). *J Cryst Growth* **90**, 51-59.

Gilliland, G. L., Tung, M., Blakeslee, D. M. & Ladner, J. (1994). *Acta Cryst D* **50**, 408-413.

Gilliland, G. L., Tung, M. & Ladner, J. (1996). *J Res Natl Inst Stand Technol* **101**, 309-320.

Goh, C. S., Lan, N., Echols, N., Douglas, S. M., Milburn, D., Bertone, P., Xiao, R., Ma, L. C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G. T. & Gerstein, M. (2003). *Nucleic Acids Res* **31**, 2833-2838.

Gonzalez, C. F., Proudfoot, M., Brown, G., Korniyenko, Y., Mori, H., Savchenko, A. V. & Yakunin, A. F. (2006). *J Biol Chem* **281**, 14514-14522.

Haebel, P. W., Arcus, V. L., Baker, E. N. & Metcalf, P. (2001). *Acta Cryst D* **57**, 1341-1343.

Hannick, L. I., Perozzo, M. A., Schultz, L. W. & Ward, K. B. (1992). *J Cryst Growth* **122**, 303-305.

Hansen, C. L., Classen, S., Berger, J. M. & Quake, S. R. (2006). *J Am Chem Soc* **128**, 3142-3143.

Harker, D. (1956). *Acta Cryst* **9**, 1-9.

Harris, M. & Jones, T. A. (2002). *Acta Cryst D* **58**, 1889-1891.

Hassell, A. M., Harrocks, T. J., Dashman, E. H. & Mistry, A. (1994). *Acta Cryst D* **50**, 459-465.

Hendrickson, W. A. (1981). *Science* **254**, 51-58.

Hendrickson, W. A., Pahler, A., Smith, J. L., Satow, Y., Merritt, E. A. & Phizackerley, R. P. (1989). *Proc Natl Acad Sci USA* **86**, 2190-2194.

Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P. A. & Rosenberg, J. M. (2000). *Acta Cryst D* **56 (Pt 7)**, 817-827.

Himo, F., Guo, J. D., Rinaldo-Matthis, A. & Nordlund, P. (2005). *J Phys Chem B* **109**, 20004-20008.

Hochuli, E., Bannwarth, W., Dobeli, H., Gentz, R. & Stuber, D. (1988). *Bio-Technology* **6**, 1321-1325.

Hogg, T., Mechold, U., Malke, H., Cashel, M. & Hilgenfeld, R. (2004). *Cell* **117**, 57-68.

Holm, L. & Park, J. (2000). *Bioinformatics* **16**, 566-567.

Houslay, M. D. & Adams, D. R. (2003). *Biochem J* **370**, 1-18.

Huai, Q., Colicelli, J. & Ke, H. (2003). *Biochemistry* **42**, 13220-13226.

Huber, T. & Kobe, B. (2004). *Bioinformatics* **20**, 2169-2170; author reply 2171-2164.

Hunsucker, S. A., Mitchell, B. S. & Spychala, J. (2005). *Pharmacol Ther* **107**, 1-30.

Jancarik, J. & Kim, S.-H. (1991). *J Appl Cryst* **24**, 409-411.

Janknecht, R., de Martynoff, G., Lou, J., Hipskind, R. A., Nordheim, A. & Stunnenberg, H. G. (1991). *Proc Natl Acad Sci U S A* **88**, 8972-8976.

Jones, P. G., Sheldrick, G. M., Kirby, A. J. & Abell, K. W. Y. (1984a). *Acta Cryst C* **40**, 547-549.

Jones, P. G., Sheldrick, G. M., Kirby, A. J. & Abell, K. W. Y. (1984b). *Acta Cryst C* **40**, 550-552.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst D* **47**, 110-119.

Kane, J. F. (1995). *Curr Opin Biotechnol* **6**, 494-500.

Kantardjieff, K. A. & Rupp, B. (2004). *Bioinformatics* **20**, 2162-2168.

Kaplun, A., Vyazmensky, M., Zherdev, Y., Belenky, I., Slutzker, A., Mendel, S., Barak, Z., Chipman, D. M. & Shaanan, B. (2006). *J Mol Biol* **357**, 951-963.

Kapust, R. B., Tozser, J., Fox, J. D., Anderson, D. E., Cherry, S., Copeland, T. D. & Waugh, D. S. (2001). *Protein Eng* **14**, 993-1000.

Khurshid, S. & Chayen, N. E. (2006). *Ann N Y Acad Sci* **1077**, 208-213.

Kimber, M. S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C. H., Vedadi, M., Gerstein, M. & Edwards, A. M. (2003). *Proteins* **51**, 562-568.

Kingston, R. L., Baker, H. M. & Baker, E. N. (1994). *Acta Cryst D* **50**, 429-440.

Knofel, T. & Strater, N. (1999). *Nat Struct Biol* **6**, 448-453.

Knofel, T. & Strater, N. (2001). *J Mol Biol* **309**, 239-254.

Koclega, K. D., Chruszcz, M., Zimmerman, M. D., Cymborowski, M., Evdokimova, E. & Minor, W. (2007). *J Struct Biol* **159**, 424-432.

Kondo, N., Kuramitsu, S. & Masui, R. (2004). *J Biochem (Tokyo)* **136**, 221-231.

Kondo, N., Nakagawa, N., Ebihara, A., Chen, L., Liu, Z. J., Wang, B. C., Yokoyama, S., Kuramitsu, S. & Masui, R. (2007). *Acta Cryst D* **63**, 230-239.

Krier, M., Araujo-Junior, J. X., Schmitt, M., Duranton, J., Justiano-Basaran, H., Lugnier, C., Bourguignon, J. J. & Rognan, D. (2005). *J Med Chem* **48**, 3816-3822.

Kuczek, M., Bryndal, I., Malinowska, J. & Lis, T. (2007). *Acta Cryst E* **63**, m889-m891.

Kuznetsova, E., Proudfoot, M., Sanders, S. A., Reinking, J., Savchenko, A., Arrowsmith, C. H., Edwards, A. M. & Yakunin, A. F. (2005). *FEMS Microbiol Rev* **29**, 263-279.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J Appl Cryst* **26**, 281-283.

Lee, M. E., Markowitz, J., Lee, J. O. & Lee, H. (2002). *FEBS Lett* **530**, 53-58.

Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L. S., Miller, M. D., McPhillips, T. M., Miller, M. A., Scheibe, D., Canaves, J. M., Guda, C., Jaroszewski, L., Selby, T. L., Elsliger, M. A., Wooley, J., Taylor, S. S., Hodgson, K. O., Wilson, I. A., Schultz, P. G. & Stevens, R. C. (2002). *Proc Natl Acad Sci USA* **99**, 11664-11669.

Lorber, B. (2001). *Acta Cryst D* **57**, 479.

Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins* **50**, 437-450.

Manjasetty, B. A., Hoppner, K., Mueller, U. & Heinemann, U. (2003). *J Struct Funct Genomics* **4**, 121-127.

McPherson, A. (2001). *Protein Science* **10**, 418-422.

Meining, W. (2006). *J Appl Cryst* **39**, 759-766.

Mendel, S., Vinogradov, M., Vyazmensky, M., Chipman, D. M. & Barak, Z. (2003). *J Mol Biol* **325**, 275-284.

Mikol, V., Rodeau, J.-L. & Gieg\'e, R. (1990). *Anal Biochem* **186**, 332-339.

Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst D* **62**, 859-866.

Morris, C., Wood, P., Griffiths, S. L., Wilson, K. S. & Ashton, A. W. (2005). *Proteins* **58**, 285-289.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998). *J Comp Chem* **19**, 1639-1662.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst D* **53**, 240-255.

Nardelli, M. (1999). *J Appl Cryst* **32**, 563-571.

Newman, J., Egan, D., Walter, T. S., Meged, R., Berry, I., Ben Jelloul, M., Sussman, J. L., Stuart, D. I. & Perrakis, A. (2005). *Acta Cryst D* **61**, 1426-1431.

O'Toole, N., Grabowski, M., Otwinowski, Z., Minor, W. & Cygler, M. (2004). *Proteins* **56**, 201-210.

Otwinowski, Z. (1991). *CCP4, SERC Daresbury Laboratory, Warrington, UK*.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol* **276**, 307-326.

Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., Stevens, R. C. & Lesley, S. A. (2003). *Acta Cryst D* **59**, 1028-1037.

Painter, J. & Merritt, E. A. (2006). *J Appl Cryst* **39**, 109-111.

Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Acta Cryst D* **61**, 1662-1669.

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

Petkowski, J. J., Chruszcz, M., Zimmerman, M. D., Zheng, H., Skarina, T., Onopriyenko, O., Cymborowski, M. T., Koclega, K. D., Savchenko, A., Edwards, A. & Minor, W. (2007). *Protein Sci* **16**, 1360-1367.

Phillips, G. N., Jr. (1985). *Methods Enzymol* **114**, 128-131.

Poirot, O., Suhre, K., Abergel, C., O'Toole, E. & Notredame, C. (2004). *Nucleic Acids Res* **32**, W37-40.

Porath, J., Carlsson, J., Olsson, I. & Belfrage, G. (1975). *Nature* **258**, 598-599.

Premkumar, L., Rife, C. L., Sri Krishna, S., McMullan, D., Miller, M. D., Abdubek, P., Ambing, E., Astakhova, T., Axelrod, H. L., Canaves, J. M., Carlton, D., Chiu, H. J., Clayton, T., DiDonato, M., Duan, L., Elsliger, M. A., Feuerhelm, J., Floyd, R., Grzechnik, S. K., Hale, J., Hampton, E., Han, G. W., Haugen, J., Jaroszewski, L., Jin, K. K., Klock, H. E., Knuth, M. W., Koesema, E., Kovarik, J. S., Kreusch, A., Levin, I., McPhillips, T. M., Morse, A. T., Nigoghossian, E., Okach, L., Oommachen, S., Paulsen, J., Quijano, K., Reyes, R., Rezezadeh, F., Rodionov, D., Schwarzenbacher, R., Spraggon, G., van den Bedem, H., White, A., Wolf, G., Xu, Q., Hodgson, K. O., Wooley, J., Deacon, A. M., Godzik, A., Lesley, S. A. & Wilson, I. A. (2007). *Proteins* **68**, 418-424.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C*, 2nd Ed. ed. Cambridge: Cambridge University Press.

Prilusky, J., Oueillet, E., Ulryck, N., Pajon, A., Bernauer, J., Krimm, I., Quevillon-Cheruel, S., Leulliot, N., Graille, M., Liger, D., Tresaugues, L., Sussman, J. L., Janin, J., van Tilbeurgh, H. & Poupon, A. (2005). *Acta Cryst D* **61**, 671-678.

Proudfoot, M., Kuznetsova, E., Brown, G., Rao, N. N., Kitagawa, M., Mori, H., Savchenko, A. & Yakunin, A. F. (2004). *J Biol Chem* **279**, 54687-54694.

Rabe, K. F., Tenor, H., Dent, G., Schudt, C., Nakashima, M. & Magnussen, H. (1994). *Am J Physiol* **266**, L536-543.

Radaev, S., Li, S. & Sun, P. D. (2006). *Acta Cryst D* **62**, 605-612.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Adv. Protein Chem.* **23**, 283-483.

Read, R. J. (1997). *Methods Enzymol* **277**, 110-128.

Reilly, T. J. & Calcutt, M. J. (2004). *Protein Expr Purif* **33**, 48-56.

Rinaldo-Matthis, A., Rampazzo, C., Reichard, P., Bianchi, V. & Nordlund, P. (2002). *Nat Struct Biol* **9**, 779-787.

Rittmann, D., Sorger-Herrmann, U. & Wendisch, V. F. (2005). *Appl Environ Microbiol* **71**, 4339-4344.

Rosenbaum, G., Alkire, R. W., Evans, G., Rotella, F. J., Lazarski, K., Zhang, R. G., Ginell, S. L., Duke, N., Naday, I., Lazarz, J., Molitsky, M. J., Keefe, L., Gonczy, J., Rock, L., Sanishvili, R., Walsh, M. A., Westbrook, E. & Joachimiak, A. (2006). *J Synchrotron Radiat* **13**, 30-45.

Rosenberg, A. H., Goldman, E., Dunn, J. J., Studier, F. W. & Zubay, G. (1993). *J Bacteriol* **175**, 716-722.

Rupp, B. (2003). *J Struct Biol* **142**, 162-169.

Rupp, B. & Wang, J. (2004). *Methods* **34**, 390-407.

Samudzi, C. T., Fivash, M. J. & Rosenberg, J. M. (1992). *J Cryst Growth* **123**, 47-58.

Sauter, C., Otalora, F., Gavira, J.-A., Vidal, O., Giege, R. & Garcia-Ruiz, J. M. (2001). *Acta Cryst D* **57**, 1119-1126.

Schermuly, R. T., Leuchte, H., Ghofrani, H. A., Weissmann, N., Rose, F., Kohstall, M., Olschewski, H., Schudt, C., Grimminger, F., Seeger, W. & Walmrath, D. (2003). *Eur Respir J* **22**, 342-347.

Schmidt, D. T., Watson, N., Dent, G., Ruhlmann, E., Branscheid, D., Magnussen, H. & Rabe, K. F. (2000). *Br J Pharmacol* **131**, 1607-1618.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst D* **58**, 1772-1779.

Schomaker, V. & Trueblood, K. N. (1968). *Acta Cryst B* **24**, 63-76.

Schomaker, V. & Trueblood, K. N. (1998). *Acta Cryst B* **54**, 507-514.

Scott, W. G., Finch, J. T., Grenfell, R., Fogg, J., Smith, T., Gait, M. J. & Klug, A. (1995). *J Mol Biol* **250**, 327-332.

Sedzik, J. (1994). *Arch Biochem Biophys* **308**, 342-348.

Segelke, B. W. (2001). *J Cryst Growth* **232**, 553-562.

Seto, D., Bhatnagar, S. K. & Bessman, M. J. (1988). *J Biol Chem* **263**, 1494-1499.

Sheldrick, G. M. (1997). *SHELXS-97: Program for the Solution of Crystal Structures.*

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

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2006). *International Tables of Crystallography* **F**, 333-345.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol* **277**, 319-343.

Shieh, H. S., Stallings, W. C., Stevens, A. M. & Stegeman, R. A. (1995). *Acta Cryst D* **51**, 305-310.

Studier, F. W. (1991). *J Mol Biol* **219**, 37-44.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). *Methods Enzymol* **185**, 60-89.

Stura, E. A., Nemerow, G. R. & Wilson, I. A. (1992). *J Cryst Growth* **122**, 273-285.

Tartof, K. D. & Hobbs, C. A. (1987). *Bethesda Res. Lab Focus* **9**, 12-14.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. & Natale, D. A. (2003). *BMC Bioinformatics* **4**, 41.

Terwilliger, T. C. (2002). *Acta Cryst D* **58**, 1937-1940.

Tsumoto, K., Umetsu, M., Kumagai, I., Ejima, D., Philo, J. S. & Arakawa, T. (2004). *Biotechnol Prog* **20**, 1301-1308.

Vagin, A. & Teplyakov, A. (1997). *J Appl Cryst* **30**, 1022-1025.

van der Woerd, M., Ferree, D. & Pusey, M. (2003). *J Struct Biol* **142**, 180-187.

Vuillard, L., Baalbaki, B., Lehmann, M., Norager, S., Legrand, P. & Roth, M. (1996). *J Cryst Growth* **168**, 150-154.

Wallden, K., Ruzzenente, B., Rinaldo-Matthis, A., Bianchi, V. & Nordlund, P. (2005). *Structure* **13**, 1081-1088.

Westheimer, F. H. (1968). *Acc Chem Res* **1**, 70-78.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst D* **57**, 122-133.

Xiong, Y., Lu, H. T., Li, Y., Yang, G. F. & Zhan, C. G. (2006). *Biophys J* **91**, 1858-1867.

Xu, H. Q. (2002). *Technometrics* **44**, 356-368.

Xu, R. X., Rocque, W. J., Lambert, M. H., Vanderwall, D. E., Luther, M. A. & Nolte, R. T. (2004). *J Mol Biol* **337**, 355-365.

Yakunin, A. F., Proudfoot, M., Kuznetsova, E., Savchenko, A., Brown, G., Arrowsmith, C. H. & Edwards, A. M. (2004). *J Biol Chem* **279**, 36819-36827.

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

Zagursky, R. J., Ooi, P., Jones, K. F., Fiske, M. J., Smith, R. P. & Green, B. A. (2000). *Infect Immun* **68**, 2525-2534.

Zhang, K. Y., Card, G. L., Suzuki, Y., Artis, D. R., Fong, D., Gillette, S., Hsieh, D., Neiman, J., West, B. L., Zhang, C., Milburn, M. V., Kim, S. H., Schlessinger, J. & Bollag, G. (2004). *Mol Cell* **15**, 279-286.

Zimmermann, H. (1992). *Biochem J* **285 (Pt 2)**, 345-365.

Zolnai, Z., Lee, P. T., Li, J., Chapman, M. R., Newman, C. S., Phillips, G. N., Jr., Rayment, I., Ulrich, E. L., Volkman, B. F. & Markley, J. L. (2003). *J Struct Funct Genomics* **4**, 11-23.