METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Structure-Based Drug Discovery

Edited by

Leslie W. Tari

Trius Therapeutics, San Diego, CA, USA

Humana Press

Editor
Leslie W. Tari
Trius Therapeutics
San Diego, CA, USA
ltari@triusrx.com

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-61779-519-0 e-ISBN 978-1-61779-520-6
DOI 10.1007/978-1-61779-520-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011944430

Printed on acid-free paper

Humana Press is part of Springer Science+Business Media (www.springer.com)

Preface

The potential utility of atomic resolution structures of protein drug targets in drug discovery has long been acknowledged. Without structure, medicinal chemists must rely on the costly, time-consuming endeavor of screening large libraries of compounds for hits, and are often forced to live with high molecular weight, non-ligand-efficient inhibitor scaffolds that must be blindly decorated with thousands of groups to generate SAR, improve potency and properties. With knowledge of the shape and chemical composition of the ligand-binding pocket of the drug target, the de novo design of ligand efficient inhibitor scaffolds is enabled. Also, iterative-structure-guided ligand optimization can be used to rationally improve early leads in a few steps rather than with thousands of analogs. However, despite its promise, structure-based drug design (SBDD) did not live up to expectations in its early days: only a limited range of protein targets were tractable to crystallographic studies, crystal structures took months or years to solve, and limitations in computing power and unrealistic expectations of the capabilities of molecular modeling methods reduced the scope and effectiveness of SBDD.

The last decade has seen the confluence of several enabling technologies that have allowed protein crystallographic methods to live up to their true potential. Off-the-shelf systems exist that allow the rapid cloning, and recombinant expression and isolation of large quantities of protein in a wide range of prokaryotic or eukaryotic hosts. Low-cost nanovolume liquid-handling robotic systems are available for the automated screening of vast arrays of diverse solution conditions to find crystallization conditions for a protein target using minimal quantities of protein. Latest generation synchrotron radiation sources allow for the collection of high-resolution X-ray diffraction data on microcrystals in minutes. Continuing improvements in computing power and advances in crystallographic software have made it possible to go from X-ray dataset to refined crystal structure in less than an hour on a laptop computer. Taken together, these advances have made it possible to tackle difficult biological targets with a high probability of success: intact bacterial ribosomes have been structurally elucidated, as well as eukaryotic trans-membrane proteins like the potassium channel and GPCRs. Of additional importance is the impact the above mentioned advances have had on the throughput of crystallographic structure determinations: it is now possible for medicinal chemists to have access to structural information on their latest small molecule candidates bound to the therapeutic target within days of compound synthesis, allowing structure-guided ligand optimization to occur in "real time." Also, using fragment screening, crystal structures of hundreds of small molecule cores complexed with the protein target can be utilized to construct novel inhibitor scaffolds.

The goal of this book is to provide scientists interested in adding SBDD to their arsenal of drug discovery methods with a practical guide to the methods used to generate crystal structures of biological macromolecules, how to leverage the structural information to design new inhibitor classes de novo, and how to iteratively optimize hits and convert them to leads. Where possible, specific protocols are described. Some examples highlighting the utility of structural biology in the discovery and development of small molecule and protein therapeutic agents are provided in the later chapters.

I am deeply grateful to all contributors who agreed to share their experiences in the development and application of methodologies that support SBDD. I believe their patience and hard work will be rewarded by the impact this volume has on scientists involved in drug discovery. I would like to extend special thanks to John Walker for his guidance, inspiration and patience in the preparation of this volume. Also, I am grateful to Les Tari Sr. for his critical evaluation of this volume and sharp editorial eye.

San Diego, CA, USA *Leslie W. Tari*

Contents

Contributors

Juan Carlos Almagro • *Centocor R&D Inc., Radnor, PA, USA*
John Badger • *Zenobia Therapeutics, San Diego, CA, USA*
Daniel C. Bensen • *Trius Therapeutics, San Diego, CA, USA*
Mark L. Cunningham • *Trius Therapeutics, San Diego, CA, USA*
Judit É. Debreczeni • *Structure and Biophysics, Discovery Sciences, AstraZeneca, Alderley Park, Macclesfield, UK*
Suzanne C. Edavettal • *Centocor R&D Inc., San Diego, CA, USA*
Paul Emsley • *Department of Biochemistry, University of Oxford, Oxford, UK*
Andrew D. Ferguson • *Discovery Sciences, AstraZeneca Pharmaceuticals, Waltham, MA, USA*
John Finn • *Trius Therapeutics, San Diego, CA, USA*
Gary L. Gilliland • *Centocor R&D Inc., Radnor, PA, USA*
Isaac D. Hoffman • *Takeda San Diego, San Diego, CA, USA*
Michael J. Hunter • *Centocor R&D Inc., San Diego, CA, USA*
Andy Jennings • *Takeda San Diego, San Diego, CA, USA*
Felice C. Lightstone • *Lawrence Livermore National Laboratory, Physical and Life Sciences Directorate, Livermore, CA, USA*
Jinquan Luo • *Centocor R&D Inc., Radnor, PA, USA*
Shawn P. Maddaford • *NeurAxonInc, Mississauga, ON, Canada L5K 1B3*
Toan B. Nguyen • *Lawrence Livermore National Laboratory, Physical and Life Sciences Directorate, Livermore, CA, USA*
Gyorgy Snell • *Takeda San Diego, San Diego, CA, USA*
Ronald V. Swanson • *Centocor R&D Inc., San Diego, CA, USA*
Leslie W. Tari • *Trius Therapeutics, San Diego, CA, USA*
Omid Vafa • *Centocor R&D Inc., Radnor, PA, USA*
Sergio E. Wong • *Lawrence Livermore National Laboratory, Physical and Life Sciences Directorate, Livermore, CA, USA*

Chapter 1

The Utility of Structural Biology in Drug Discovery

Leslie W. Tari

Abstract

Access to detailed three-dimensional structural information on protein drug targets can streamline many aspects of drug discovery, from target selection and target product profile determination, to the discovery of novel molecular scaffolds that form the basis of potential drugs, to lead optimization. The information content of X-ray crystal structures, as well as the utility of structural methods in supporting the different phases of the drug discovery process, are described in this chapter.

Key words: X-ray crystallography, Structure-based drug design, Fragment screening, Structural bioinformatics, Lead optimization

1. Introduction

The discovery of new drugs is a time and labor-intensive process. On average, the discovery of a new drug requires the preparation and evaluation of approximately 10,000 compounds over 12 years at a cost of more than $350 million (1). Once in the marketplace, many drugs fail to recover their development costs (as many as 30%, according to data from the 1980s (2)), and many others are ultimately withdrawn from the market. These facts coupled with limits on patent lifetime, escalating global competition, and increasingly stringent government regulations for drug approval have demanded more efficient and accelerated approaches to drug discovery. Conventional "brute force" methods of lead discovery via high-throughput screening (HTS) of proprietary synthetic, combinatorial, or natural product libraries, while effective in many cases, are expensive and have limitations; they require access to large compound libraries (sometimes over 1,000,000 compounds), often yield hits with high molecular weight, poor ligand efficiency,

Leslie W. Tari (ed.), *Structure-Based Drug Discovery*, Methods in Molecular Biology, vol. 841,
DOI 10.1007/978-1-61779-520-6_1,

limited or no potential for optimization, and provide no information to guide ligand optimization.

Advances in crystallographic methods, computational power, molecular biology, and recombinant protein expression systems over the last 30 years have provided researchers with rapid and reliable access to three-dimensional structural information on a wide variety of protein drug targets. Structural information on protein–ligand complexes can eliminate much of the complexity involved in the discovery and optimization of prospective drug leads. Indeed, structure-guided drug design efforts have led to the discovery of high profile drugs in multiple therapeutic areas, including the peptidomimetic HIV protease inhibitors for the treatment of HIV, the neuraminidase inhibitor Tamiflu™ for the treatment of influenza, the carbonic anhydrase inhibitor dorzolamide for the treatment of glaucoma, and the thrombin inhibitor ximelagatran, an oral anticoagulant (3). Access to structural information on the target of interest can streamline all aspects of drug discovery, from target selection to lead discovery and optimization, using methods that are summarized in this chapter.

2. The Information Content of Protein Crystal Structures

Protein crystals, like any crystalline substance, are regular, three dimensionally periodic arrays of identical molecules or molecular complexes (see Fig. 1). A common misconception regarding protein crystal structures is that they are not representative of the protein in solution due to the influence of extensive intermolecular interactions present in the crystalline state. The idea that protein crystal structures are heavily biased by "solid state" artifacts arises from inaccurate comparisons made between protein crystals and crystals of small molecular weight compounds. Crystals of small molecules and proteins differ in ways that extend beyond the properties of their component molecules. Small-molecule crystals typically only comprise the small molecule, while protein crystals contain 25–90% solvent by volume, depending on the protein. The remaining volume in protein crystals is occupied by protein molecules, and is analogous to an ordered gel with large interstitial spaces between protein molecules. By comparison, the number of contacts made in relation to the molecular mass of the protein in protein crystals is smaller by orders of magnitude than it is for small-molecule crystals. This causes the mechanical stability and integrity of protein crystals to be much worse than it is for crystals of small molecules. The high solvent content and tenuous thermodynamic stability of protein crystals complicate the subsequent steps in X-ray diffraction experiments, since these properties result in crystal handling difficulties, susceptibility to temperature changes

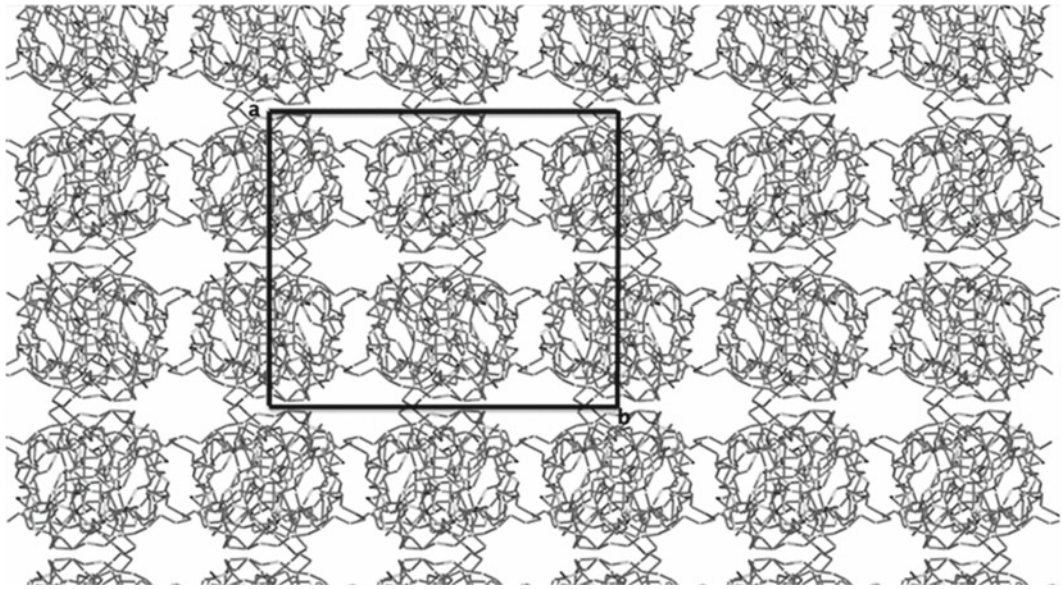

Fig. 1. A view of crystal packing in a *Haemophilus influenzae* dihydrofolate reductase crystal. Boundaries for a single unit cell within the crystal are shown. The view is perpendicular to the *c*-axis of the unit cell. The unit cell is the fundamental building block of the crystal, a translationally periodic substance comprising trillions of unit cells that extend in three dimensions. The unit cell is an arbitrary construction that describes the smallest "box" with the highest metric symmetry.

and dehydration, weaker diffraction, and greater sensitivity to radiation damage. However, the key role played by solvent in protein crystallization is a double-edged sword; while it adversely affects diffraction, it is the very element that makes protein crystal structures valuable. The high solvent content of protein crystals is essential for maintaining the structures of the macromolecules in their solution states. Therefore, to a large extent, proteins in crystals possess the structural, enzymatic, and functional properties of their counterparts in solution. Protein crystal structures must be regarded with care, however. In the hands of the uninformed, the danger exists that crystallographic structural data will be misinterpreted, or overreaching conclusions drawn. An understanding of the parameters derived from crystallographic experiments is essential if structural information from crystallographic experiments is to be used effectively to support drug discovery.

X-ray crystallography and light microscopy share the same basic principle; electromagnetic radiation scattered by the object to be imaged is recombined and focused by a lens to reform the image of the object. Theoretically, the resolving power of any imaging technique is equal to one half of the wavelength of the radiation used for imaging. To resolve the atomic details of protein structures, crystallographic experiments involve the exposure of protein crystals to high-energy monochromatic X-rays (wavelengths on the order of 1 Å). Imaging using X-rays is complicated by the fact

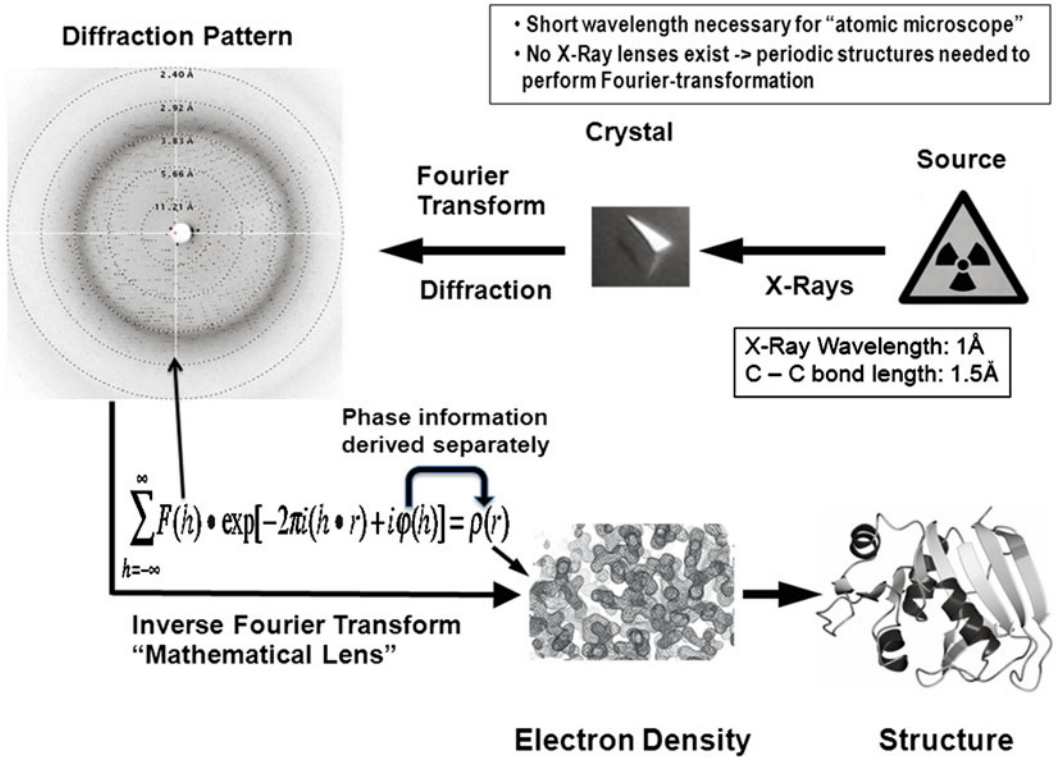

Fig. 2. A schematic outlining the steps in a crystallographic structure determination. Crystals are systematically exposed to monochromatic X-rays in multiple orientations, and the diffraction patterns are captured with electronic detectors. Since crystals are three-dimensionally periodic substances, the diffraction pattern comprises a series of spots rather than a continuous function. Each spot represents a family of diffracted waves that map to discrete spatial periodicities in the unit cell of the crystal. The diffraction pattern is a summation of waves of electromagnetic radiation and can thus be described by a Fourier series, and the diffraction pattern and disposition of the atomic contents of the unit cell are related mathematically by a Fourier transform. An image of the atomic contents of the unit cell of the crystal is derived by applying a mathematical lens (inverse Fourier transform, equation shown on the *lower left*) to the diffracted X-rays. The image reconstruction process is complicated by the fact that only intensities of the diffracted X-rays are measurable ($F(h)$ terms in the equation shown), but not the relative phase shifts between each family of diffracted waves. The missing information is referred to as the crystallographic phase problem. The missing phases are obtained using other experimental or computational methods described in the text. Since the diffraction of X-rays is caused by the interaction of the X-rays with electrons, the resulting image obtained in a crystallographic experiment is of the electron density distribution in the unit cell of the crystal. Interactive model building software is used to build the final atomic model into electron density.

that X-rays interact very weakly with matter, so that no lenses exist which are able to reconstruct the image from the scattered X-rays. Hence, the scattered X-rays from crystals must be captured with electronic detectors and the function of a lens must be simulated mathematically. A schematic describing the steps involved in the solution of a crystal structure is shown in Fig. 2.

Mathematical reconstruction of the structure of the atomic contents of the crystal is complicated by the fact that one of the two key pieces of information describing the diffracted X-ray waves, the relative phase shifts between the different families of diffracted

waves, cannot directly be measured (see Fig. 2). Three methods are commonly employed to overcome the phase problem, as summarized below.

(a) *Molecular replacement.* When an approximate model of the unknown crystal structure is available, it can be used to overcome the phase problem. The principle is simple; the model is first oriented and then positioned in the unit cell of the target crystal structure using rotation and translation functions. The correctly oriented model is subsequently used to calculate approximate phases and electron density maps. Alternate cycles of interactive correction and rebuilding of the model into electron density and model refinement are used to improve the quality of the phases and to transform the model structure into the real structure. The success of molecular replacement depends critically on two factors: the fraction of the asymmetric unit for which suitable models exist, and the r.m.s. deviation (after optimal superposition) between the model and target structures. Generally, r.m.s. deviation increases with decreasing sequence identity, or in cases where the target structure undergoes significant conformational changes with respect to the model structure (e.g., movement of protein domains). In the latter case, the model structure can be separated into individual fragments that are sequentially oriented and positioned in the unit cell. Newer maximum-likelihood molecular replacement algorithms, such as those implemented in the program Phaser (4) are more discriminating, and have been successful in solving difficult molecular replacement problems that were previously intractable.

(b) *Isomorphous replacement methods.* This is a classical approach used to solve protein structures with unknown folds. Crystals are soaked in multiple solutions containing salts of heavy atoms such as Hg, Pt, Pb, Au, etc., until conditions are found where a small number of heavy atoms incorporate in well-defined positions on the crystallized protein molecule (without altering the structure of the underlying protein). By analyzing the differences in the intensities of diffraction patterns from the native and heavy atom derivatized protein crystals, it is possible to determine the locations of the heavy atoms in the unit cell and to use the scattering "signal" from the heavy atoms to calculate phases and an electron density map (reviewed in refs. (5–7)).

(c) *Anomalous scattering methods.* For heavier elements, some inner shell electrons have absorption edges in the range of the X-ray wavelengths used in diffraction experiments. The heavy atoms in the protein crystal cause absorption of the impinging radiation, and impart small phase shifts on the radiation scattered from the crystal. This phenomenon is used to determine

the positions of the heavy atoms in the unit cell, and subsequently to extract phase information to allow electron density map generation. Anomalous scattering can be used to supplement the phase information obtained from isomorphous heavy atom derivatives, or to independently obtain complete phase information. A very powerful de novo phase determination method utilizes anomalous scattering from proteins that are homogeneously labeled with selenomethionine (incorporated during recombinant expression of the protein in *Escherichia coli*), a derivatized selenium-containing amino acid. Independent diffraction experiments are carried out (on the same crystal, if possible) at multiple X-ray wavelengths on the high and low energy sides of the selenium absorption edge that maximize the anomalous diffraction signal. This method requires a tunable X-ray source, which is present only at synchrotrons (reviewed in refs. (5–7)).

X-ray diffraction is caused by the interaction of the electric field vector of monochromatic X-rays with electrons in a protein crystal. These details, coupled with the fact that crystals are made up of three-dimensionally periodic lattices of molecules, have several important consequences (for excellent reviews see refs. (5–7)): (1) X-ray diffraction experiments generate three-dimensional images of the electron density distribution of the molecular components of the crystal. So heavier atoms generate a proportionally stronger signal, and hydrogen atoms are generally not discernable in protein crystal structures. (2) The short wavelength radiation used in X-ray diffraction experiments allows for the resolution of macromolecular structures at an exquisite level of detail (typical protein crystal structures are determined at resolutions between 1.5 and 3.0 Å resolution). (3) In a crystallographic experiment, the structure of the molecular contents of the unique portion of a crystal (called the asymmetric unit of the unit cell, which is the microscopic building block of the crystal) are obtained, and the resulting crystal can be built by the application of crystallographic symmetry operators to the contents of the asymmetric unit, as shown in Fig. 1. Since the diffraction signal from a crystal arises from constructive interference from trillions of crystallographic asymmetric units, the resulting crystal structure comprises a time- and space-averaged picture of the contents of the copies of asymmetric units that are sampled. Hence, components of the asymmetric unit with a large degree of random spatial heterogeneity, i.e., disordered protein loops or side chains and the bulk solvent occupying the spaces between protein molecules, fade into the background and cannot be modeled. However, in cases where a molecular component of a crystal, such as a protein side chain, occupies a finite number of distinct, low energy conformations in different asymmetric units, it is possible to simultaneously characterize each alternative conformation.

Examination of the equation relating diffracted X-rays to the crystal structure provides insight into the structural parameters that are modeled in a crystallographic experiment (see Eq. 1).

$$F_{hkl} = \sum_{j=1}^{N} f_j e^{-(B\sin^2\theta)/\lambda^2} e^{2\pi i(hx+ky+lz)}. \quad (1)$$

Equation 1 is one of the explicit forms of the structure factor equation (8). Each F_{hkl} term represents a unique family of diffracted X-ray waves from the crystal (diffracted waves from crystals constructively interfere to form patterns of spots, as shown in Fig. 2, which can each be assigned integer indices h, k, and l), which correspond to discrete spatial periodicities in the crystal lattice. The intensity and phase of each family of diffracted waves is derived via a summation of the scattering contributions from all of the atoms in the asymmetric unit of the crystal. The second exponential term in Eq. 1 computes the net phase shift relative to an arbitrary origin of the scattered wave with index h, k, l due to the relative positions of the individual atoms in the unit cell (with fractional coordinates x, y and z). The f_j term corresponds to the scattering factor for each atom in the summation, and is directly proportional to the number of electrons in the atom in question. The first exponential $B\sin^2\theta/\lambda^2$ term (θ is the angle of the scattered radiation with respect to the source X-ray beam, and λ is the wavelength of the X-rays) accounts for the reduction in the intensity of the scattered radiation with scattering angle due to interference between scattered waves from different parts of the electron cloud surrounding each atom. X-ray scattering is attenuated further by smearing of the electron clouds surrounding each atom due to thermal motion of the atoms. Atomic thermal motion is modeled using the extra B term in the structure factor equation. As a first approximation it is assumed that the thermal motion of atoms is isotropic (spherically symmetric), with $B = 8\pi^2\mu^2$, where μ is the root mean square amplitude of atomic vibration. Using the calculation above, for a B-factor of 15 Å^2, the displacement of an atom from its equilibrium position is approximately 0.44 Å, and it is as much as 0.87 Å for a B-factor of 60 Å^2. Thus, analysis of B-factors is very important during any structural analysis to provide insight into the dynamics and structural integrity of different regions of a protein molecule. However, one must exercise caution before interpreting B-factors too quantitatively. In addition to measuring dynamic disorder caused by temperature dependent vibration of atoms, the B-factor is also influenced by subtle structural differences between protein molecules in different unit cells throughout the crystal (which spatially smears the atom positions), steric constraints from intermolecular lattice contacts, and certain systematic experimental errors, such as absorption of the X-ray beam during X-ray data collection. Advanced mathematical models can be used to provide more

detailed information on atomic thermal motions. For example, the relative motions of entire protein domains can be characterized using TLS refinement (9). Also, when high-quality X-ray data are available from crystals that diffract to high resolution (typically better than 1.2 Å, rare in protein structure determinations), the isotropic thermal correction can be replaced by a tensor, which corrects not only for the extent of thermal motion of the atoms but also for spatial anisotropy in their motions (10).

Based on the mathematical description of X-ray diffraction provided above, four parameters are optimized in a single crystal X-ray diffraction experiment for each atom in a protein crystal structure: the *x*, *y*, and *z* coordinates of each atom and the *B*-factor describing the thermal motion of each atom. The quality of resulting electron density maps and the accuracy of refined parameters in protein crystal structures are largely dependent on the resolution of the X-ray diffraction data (equivalent to the pixel size of electron density sections). Examples of the effects of diffraction resolution on electron density map quality are shown in Fig. 3. The model is generally manually built (or refit) into electron density by a crystallographer, using two types of electron density maps, $|2F_o - F_c|\alpha_c$ maps, and $|F_o - F_c|\alpha_c$ difference maps, described below.

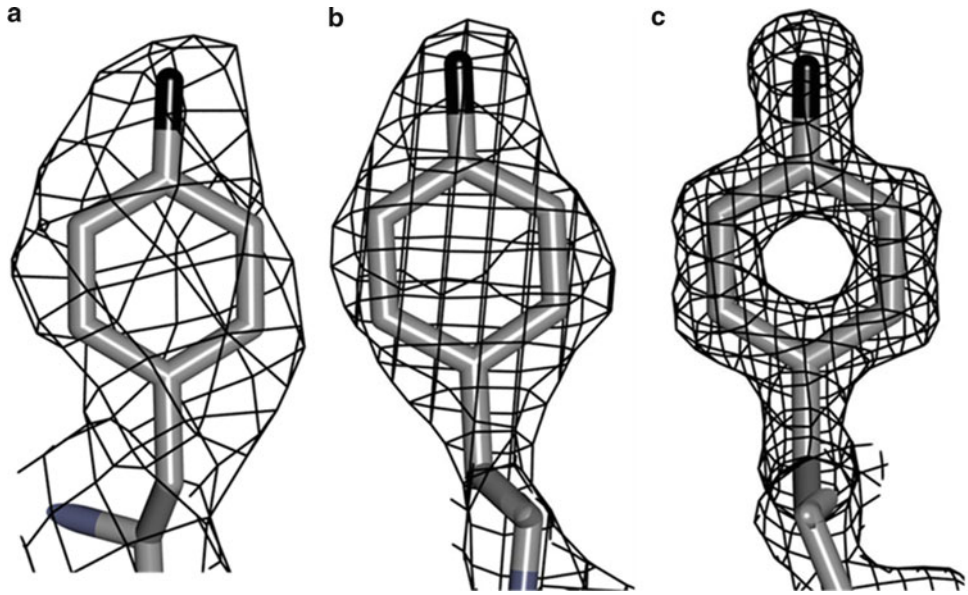

Fig. 3. Representative electron density maps contoured around tyrosine residues (using $|2F_o - F_c|\alpha_c$ coefficients) from three refined crystal structures: (**a**) A 2.8 Å resolution structure of *Francisella tularensis* topoisomerase IV, (**b**) A 2.2 Å structure of *Escherichia coli* topoisomerase IV, and (**c**) A 1.4 Å structure of *Enterococcus faecalis* DNA gyrase B (all from D. Bensen and L. Tari, unpublished results). The electron density maps were contoured using the electron density visualization software COOT (*see* ref. (11), Chapter 6). At better than 3.0 Å resolution, amino-acid side chains can be recognized with the help of protein sequence information, while at better than 2.5 Å resolution solvent molecules can be observed and added to the structural model with some confidence. As the resolution improves to better than 2.0 Å resolution, fitting of individual atoms may be possible and most of the amino-acid side chains can be readily assigned even in the absence of sequence information.

$|F_o - F_c|\alpha_c$ maps. $|F_o - F_c|\alpha_c$ maps, or difference maps, are generated by subtracting the calculated structure factor amplitudes (F_c, from the best current model structure) from the observed structure factor amplitudes (F_o), using phase information (α_c) calculated from the available model structure. To a good approximation, this operation is equivalent to subtracting the electron density calculated from the model from the "real" electron density in the crystal. What is left behind is the electron density for ordered components of the crystal structure that have not been accounted for by the model, or that have not been modeled correctly. Features that are present in the true structure that have not been accounted for in the model structure appear as positive peaks, while atoms that have been incorrectly placed in the model structure (i.e., that do not exist in the real structure) appear as holes or negative peaks. These maps are used to fix improperly modeled side-chains and/or entire polypeptide chains, as well to fit substrates, inhibitors, and ordered solvent molecules into the structure. A special type of difference map called an omit map can be used to confirm the presence of important features in a protein structure. An omit map is calculated by removing the feature of interest (say, an inhibitor) from the model, refining the structure in the absence of that feature, and calculating a new difference map. If the feature of interest is still observed in a difference density map, then it is real, and not an artifact caused by model bias present in the calculated phases. An example of a difference map is shown in Fig. 4.

$|2F_o - F_c|\alpha_c$ maps. $|2F_o - F_c|\alpha_c$ maps are the maps most commonly used for model fitting. They are used instead of $|F_o|\alpha_c$ maps, which suffer from model bias, and tend to show only electron density that is associated with the model. As described above, $|F_o - F_c|\alpha_c$ maps reveal everything in the $|F_o|\alpha_c$ map that has not been modeled. The $|2F_o - F_c|\alpha_c$ map essentially superposes an $|F_o|\alpha_c$ map over an $|F_o - F_c|\alpha_c$ difference map, so that it simultaneously shows both the electron density for the model and the electron density for features that have not been accounted for by the model. Several weighting schemes are used to further diminish the effects of model bias, including figure-of-merit and σ_A weighting schemes (reviewed in refs. (5–7)). An example of a $|2F_o - F_c|\alpha_c$ electron density map is shown in Fig. 4.

In addition to providing a more detailed picture of the electron density, higher resolution X-ray data correlates with a greater number of experimental observations to support structure refinement. For a typical protein structure from a crystal with a solvent content of about 50%, the number of experimental observations and refinement parameters will be about the same at 2.8 Å resolution. The paucity of experimental data compared with the number of parameters that need to be defined make least squares model optimization methods intractable. Additionally, at resolutions lower than 2.8 Å, individual atomic *B*-factors have a very limited

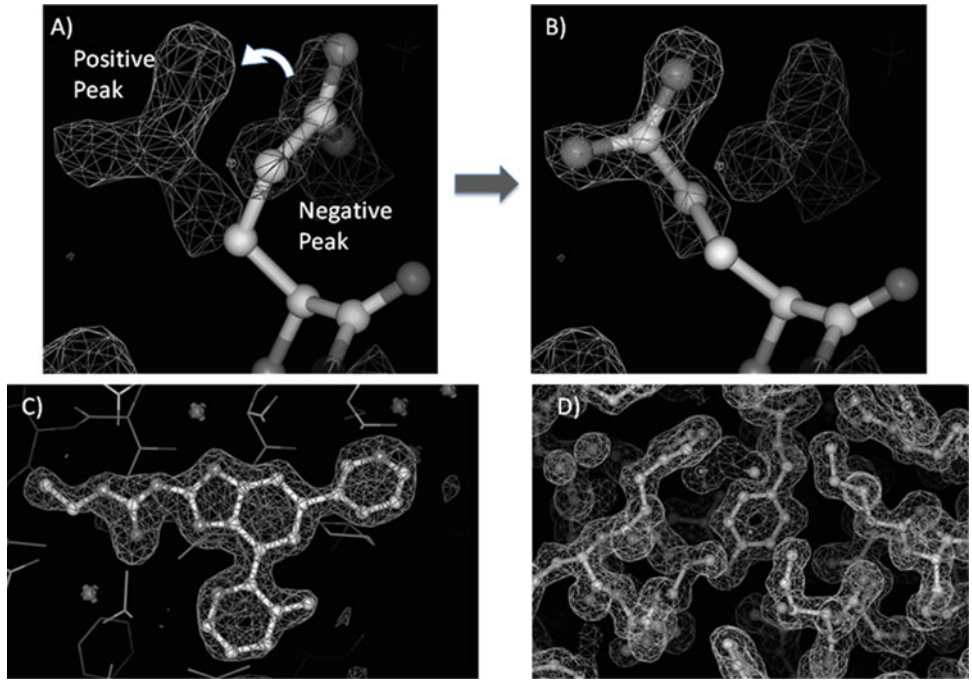

Fig. 4. Examples of $|F_o - F_c|\alpha_c$ and $|2F_o - F_c|\alpha_c$ electron density maps. The electron density maps in all panels are drawn as thin chicken-wire representations. In (**a**) an $|F_o - F_c|\alpha_c$ map contoured at 3σ is used to fit an incorrectly modeled glutamic acid side chain in an *E. faecalis* GyrB crystal structure. In the model structure, part of the side chain is in a negative electron density peak, while a positive difference density peak on the left-hand side of the figure reveals the correct position for the side chain from the experimental data. The correctly positioned glutamic acid side chain is shown in (**b**). In (**c**), an $|F_o - F_c|\alpha_c$ difference electron density map contoured at 3.5σ was used to fit a small-molecule inhibitor into the substrate-binding pocket of *E. faecalis* gyrase B. The difference map was calculated in the absence of the inhibitor, indicating that the difference density shown arises entirely from the experimental X-ray data. Panel (**d**) shows a representative section of a $|2F_o - F_c|\alpha_c$ electron density map contoured at 1σ for an *E. faecalis* GyrB crystal structure. The map displays electron density for both regions of the model that have been correctly fit, as well as regions that have not been accounted for by the model. Because it comprises a superposition of an $|F_o|\alpha_c$ map and a $|F_o - F_c|\alpha_c$ map, $|2F_o - F_c|\alpha_c$ maps are less subject to the effects of model bias than $|F_o|\alpha_c$ maps. During model fitting, crystallographers generally utilize $|2F_o - F_c|\alpha_c$ and $|F_o - F_c|\alpha_c$ simultaneously to trace the polypeptide chain and correct errors in the existing model.

physical meaning. The problem of statistical under determination is overcome by augmenting the X-ray diffraction data with structural parameters of proteins and peptides derived from small-molecule crystallography and spectroscopic data. The resulting function that is minimized in a crystallographic structure refinement incorporates the experimental X-ray data and a molecular mechanics function (which restrains bond lengths, angles, stereochemistry, planarity of peptide bonds and aromatic groups, etc. to reasonable values). The quality of structures refined in this fashion is excellent, even for structures determined at modest resolutions. Properly refined protein crystal structures generated from carefully measured X-ray data yield atomic positions that are precise to within one fifth to one tenth of the stated experimental resolution. Once a structure is fully refined, multiple criteria are used to judge the quality of the model, as described below.

R-factor. The *R-factor* is the averaged error (in percent) between the observed structure-factor amplitudes (the experimentally measured F_{hkl} values) and the calculated structure-factor amplitudes ($F_{hklcalc}$) from the refined model of the contents of the crystal. The ultimate value of the *R-factor* in a well-refined structure depends on a number of variables, including the proportion of the contents of the unit cell that can be correctly modeled, the relative weights assigned to the molecular mechanics restraints vs. the experimental X-ray data during refinement, the experimental resolution of the diffraction experiment and the accuracy and overall quality of the measured experimental X-ray intensities. In protein structures with numerous dynamically disordered loops or domains that cannot be modeled, the *R-factor* will not converge to low values. However, as a general rule of thumb a correctly refined protein structure should have an *R-factor* around 20%.

Free R-factor (R_{free}). The function that is minimized during a protein structure refinement is extremely complex, with multiple false minima. Hence, when not used with care, modern refinement algorithms can converge on convincing *R-factors* for incorrect structures. The R_{free} (12) statistic is an extremely simple and powerful independent validation tool used in modern protein structure refinement. The R_{free} function is identical to the *R-factor*; the only difference is that it is calculated using a small (5–10%) randomly sampled subset of the X-ray diffraction data that is excluded from structure refinement throughout the refinement process. In a correctly refined structure, R_{free} will track with the *R-factor* to within 5–10%. For incorrect structures, R_{free} will remain at a value near the limit observed for random atomic models fit to an X-ray dataset (~57%). In addition to R_{free}, the geometric quality of the refined protein structure should be used to evaluate the model. The averaged bond lengths and angles of the final model should not deviate much from ideal values (r.m.s. deviations from ideality should be within 0.02 Å for bond lengths and 3° for bond angles), and the majority of the protein residues should possess "allowed" combinations of ϕ, ψ main-chain dihedral angles. It is important to note that protein folding can force some residues into disallowed ϕ, ψ values, which can have important functional significance (13). All residues in disallowed regions must be carefully checked to ensure that they are well described by experimental electron density.

Identification and refinement of ordered solvent molecules becomes more reliable when data are available to at least 2.5 Å resolution. Even then, before a water molecule is used in mechanistic or computational analysis, it is always wise to check its *B*-factor and to see if there exists at least one hydrogen bond to hold the water to the protein or a nearby solvent molecule.

Unless the structure has been determined at very high resolution, electron density and refinement do not discriminate between the oxygen and nitrogen atoms of asparagines and glutamines, or

the alternative conformations of histidine side chains. In a detailed structural analysis, it is always necessary to check alternative conformations of Asn, Gln, or His side chains to decide which one makes more sense chemically (i.e., by analyzing available H-bonding networks). Also, great care has to be exercised when fitting dynamically disordered protein side chains that are not fully described by electron density. The crystallographer knows they are present from the amino-acid sequence, and incorporates them in conformations commonly observed for that side chain from databases of high-resolution structures. The final refined conformation of the side chain must ultimately be decided using the crystallographer's knowledge of chemistry and side-chain conformational preferences, in conjunction with the refinement program's force field. In many structures, entire loops or even domains are too disordered to show any observable electron density. In such cases, the offending loops/domains are not included in the final model. When analyzing crystal structures, an additional point of caution that must be noted regarding potential artifacts that can arise from contacts between adjacent molecules in a crystal lattice. In the ideal scenario, the protein of interest crystallizes in a lattice that leaves the active-site/receptor pocket solvent exposed, with no lattice contacts preventing the motion of functionally important mobile structural elements surrounding the drug-binding site (i.e., the lattice should not impede ligand-induced conformational changes in the protein). However, protein crystallization does not allow for control of lattice contacts, and the ideal situation does not always exist. Hence, before a new protein crystal form is nominated as a potential candidate for supporting structure-based drug design, a careful analysis of the crystal lattice contacts between neighboring molecules related by crystallographic or noncrystallographic symmetry must be carried out to assess the steric accessibility of the receptor pocket and the solvent space around it, as well as the nature and quantity of lattice contacts in the vicinity. This sort of analysis is particularly important if the crystals are produced for the purpose of ligand soaking experiments to support fragment screening or high throughput structure determination. If multiple crystal forms are available, the crystal forms that approach the ideal criteria should be chosen. Cocrystallization experiments usually circumvent problems related to lattice constraints, since the protein and ligand are mixed in solution, allowing the system to reach a low energy conformational state before crystallization occurs. Additional important parameters to consider when analyzing crystal structures are the solution conditions used in crystallization. Some proteins undergo significant structural changes in different solution conditions. A classic example is ribonuclease A, which undergoes large, pH-dependent conformational changes that have been characterized crystallographically (14).

3. Using Structure in Target Selection and Product Profile Development

In addition to supporting lead discovery and lead optimization, structural information can be used at a very early stage in a drug discovery program to evaluate the viability of a protein as a drug target. Does the protein possess a binding pocket with suitable properties for potent inhibitor development? In a large, structurally related protein family, such as eukaryotic protein kinases, is it possible to develop selective inhibitors against a kinase of interest? More generally, what are the prospects for the development of specific inhibitors against a protein target while avoiding off-target binding? In an antibiotic program, do the protein orthologs encompassed by the proposed target product profile possess sufficient structural homology to allow for the development of a small-molecule agent with the desired spectrum? Careful analysis of the structures of the protein target(s) of interest coupled with structural bioinformatics and molecular modeling can be used to address questions such as those posed above. Such an analysis is important to expose liabilities in target selection or the proposed drug product profile early in a drug discovery program, before a substantial investment of time, money and manpower has been made to pursue a flawed hypothesis.

For example, in the antibacterial arena, the emergence of genomics and proteomics has profoundly changed the approach used for the identification of new targets essential for the survival of bacteria (15). To highlight how this information is used to facilitate target selection, the analysis that led to the selection of bacterial topoisomerases as prospective drug targets at the author's company is summarized below. To pursue a drug discovery program, we sought essential bacterial targets with the following properties: (1) Novel proteins that are not targets of marketed antibiotics, to avoid issues of cross-resistance with existing antibiotics. (2) Targets possessing recessed ligand-binding pockets with mixed polar/lipophilic character, the potential for solvent sheltered "anchoring interactions" and no closely related human counterparts. (3) A high degree of sequence/structure conservation in the ligand-binding pockets of the protein target(s) across bacterial species commonly implicated in bacterial infections. (4) If possible, the option to inhibit multiple bacterial targets with a single therapeutic agent to minimize the threat of resistance emergence. A detailed structural bioinformatics analysis of proteins in several key bacterial pathways revealed the bacterial topoisomerases DNA gyrase and topoisomerase IV as prospective drug targets that met the criteria listed above. DNA gyrase is a type II topoisomerase that plays an essential role in bacterial DNA replication with no direct mammalian counterpart. The enzyme catalyzes the introduction of negative supercoils into DNA using the free energy of

ATP hydrolysis (16). DNA gyrase consists of two subunits, GyrA and GyrB that form a functional heterodimer A_2B_2. GyrA is involved in DNA cleavage and religation, while the GyrB domain contains the ATP-binding site and mediates the passage of the uncut DNA strand through the strand that is cleaved by GyrA (16). A closely related bacterial enzyme from the topoisomerase II family is topoisomerase IV (topo IV), which also forms a heterodimer C_2E_2 consisting of two ParC subunits and two ParE subunits (17). Despite possessing a high degree of sequence identity with DNA gyrase, topo IV is involved in different aspects of DNA replication than gyrase. The two topoisomerase complexes are well established drug targets. Fluoroquinolone antibiotics, such as ciprofloxacin, exert their antimicrobial activity via inhibition of the GyrA and ParC subunits (18). However, no commercial antibiotics have yet reached the market which target the ATP binding domains of the respective topoisomerase complexes (GyrB and ParE), despite the fact that GyrB and/or ParE inhibition has been shown to effectively kill bacteria (19). A sequence alignment of the ATP-binding domains of DNA gyrase and topo IV from key pathogens involved in community acquired pneumonia mapped on to the crystal structure of one of the enzymes (see Fig. 5), suggests that the development

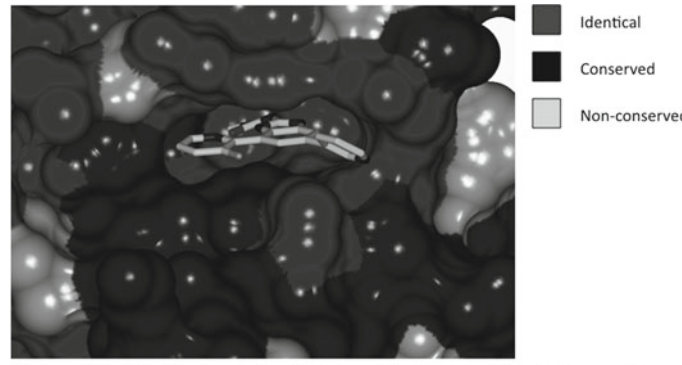

Fig. 5. A solvent accessible surface representation of the ATP-binding pocket of GyrB from the crystal structure of *E. faecalis* GyrB complexed with a benzimidazole inhibitor (D. Bensen and L. Tari, unpublished results). The surface is colored by the degree of sequence conservation observed in the underlying residues for GyrB and ParE enzymes from the major pathogens implicated in community acquired pneumonia. Amino-acid sequences for the relevant proteins were extracted from the KEGG database (20) and sequence alignments were performed with CLUSTALW (21). The high degree of overall sequence conservation (not shown) and the remarkable degree of sequence conservation in the ATP-binding pockets of the selected GyrB and ParE orthologs suggest that the geometries and compositions of the active sites of the enzymes from the different pathogens possess sufficient similarity to allow for the development of dual targeting, broad spectrum inhibitors. Subsequent generation of homology models and crystal structures of several of the orthologs listed on the figure confirmed this hypothesis.

of broad spectrum, dual-targeting inhibitors against these enzymes is feasible. As the above example demonstrates, structural bioinformatics can be an important component in the target selection process and drug product profile determination early in a drug discovery program.

4. Using Crystallographic Methods to Initiate a Drug Discovery Program

The likelihood of success in a small-molecule drug discovery program is greatly enhanced by the availability of multiple molecular scaffolds that bind to and elicit the desired effects on the protein target, while offering prospects for optimization into drug leads. However, the discovery of viable molecular scaffolds for SBDD and medicinal chemistry optimization is not trivial. HTS, when successful, often delivers hits with high molecular weights and poor potential for optimization. The probability of a small-molecule ligand matching the shape and chemistry of a protein target decreases as the complexity and size of the ligand increases, since there exists a greater chance that some part of the ligand will possess features that do not complement those of the protein target. Theoretically, the probability that a small molecule will bind to a protein target decreases exponentially with increasing ligand complexity (22). Thus, there is an advantage to screening for hits using less complex, lower molecular weight compounds (called fragments, with molecular weights ranging from 100 to 250 Da), which interact with only a small number of sites on the protein and possess a greater chance of achieving favorable steric and chemical complimentarity with the protein target. However, the advantage of screening with fragments is offset by the fact that fragments generally bind with much lower affinities than the larger compounds typically screened in HTS. Most biophysical techniques perform poorly at detecting weak binding, limiting their utility in screening fragment libraries. X-ray crystallography, however, is an extremely sensitive technique, capable of detecting compounds with binding constants in the low millimolar range. The extension of crystallographic methods into the high-throughput realm over the past decade has led to the adoption of crystallographic fragment screening in many industrial and academic centers as a drug discovery tool. In this section, the two flavors of crystallographic fragment screening are reviewed: random fragment screening and pharmacophore-based fragment screening.

4.1. Random and Pharmacophore-Based Fragment Screening

The basic premise of crystallographic fragment screening is simple. A protein target is screened against a small library (typically <1,000 molecules) of structurally diverse, highly soluble low molecular weight compounds. The library is screened in one of two ways: pregrown