Top Curr Chem (2014) 336: 67–98
DOI: 10.1007/128_2012_326

Published online: 20 May 2012

The Role of Protein Structural Analysis in the Next Generation Sequencing Era

Wyatt W. Yue, D. Sean Froese, and Paul E. Brennan

Abstract Proteins are macromolecules that serve a cell's myriad processes and functions in all living organisms via dynamic interactions with other proteins, small molecules and cellular components. Genetic variations in the protein-encoding regions of the human genome account for >85% of all known Mendelian diseases, and play an influential role in shaping complex polygenic diseases. Proteins also serve as the predominant target class for the design of small molecule drugs to modulate their activity. Knowledge of the shape and form of proteins, by means of their three-dimensional structures, is therefore instrumental to understanding their roles in disease and their potentials for drug development. In this chapter we outline, with the wide readership of non-structural biologists in mind, the various experimental and computational methods available for protein structure determination. We summarize how the wealth of structure information, contributed to a large extent by the technological advances in structure determination to date, serves as a useful tool to decipher the molecular basis of genetic variations for disease characterization and diagnosis, particularly in the emerging era of genomic medicine, and becomes an integral component in the modern day approach towards rational drug development.

Keywords Drug development · Genetic diseases · Misfolding · Missense mutations · Mutation analysis · Protein structures · Structure based drug design

Contents

W.W. Yue (✉), D.S. Froese, and P.E. Brennan
Structural Genomics Consortium, University of Oxford, Old Road Campus Research Building, Oxford OX3 7DQ, UK
e-mail: wyatt.yue@sgc.ox.ac.uk

1 Structural Biology in the Post-Genomics Era

1.1 Introduction

The past decade has seen an explosion of genome sequences, thanks to the many advances in sequencing technology. These global sequencing efforts have provided us with genetic blueprints for a myriad of organisms in all kingdoms of life. The approach to biomedical research therefore has undergone a radical and dramatic transformation in the post-genomics era. In the emerging era of genomic medicine, it is now possible to sequence completely 3×10^9 base pairs in the human genome for individual patients. We are now tasked with the annotation and description of the plethora of genomic data with regards to biological functions. Although the protein-coding genomic space (exome) is small, where protein-coding exons account for only 1% of the human genome, it represents a majority of the targets for drug development, and 85% of Mendelian diseases are caused by genetic variations in the exomic space. A protein is not merely an "alphabetical" sequence of amino acids, but a macromolecule with three-dimensional (3D) shape and form, capable of performing specialized biological functions in the cell via dynamic interactions with other proteins, small ligands and cellular components. These functional properties depend on a protein's three-dimensional structure, and the field of structural biology is instrumental in directing research towards an understanding of protein function and disease. A large amount of resources have now been put in place, at the disposition of the broad community of non-structural biologists in biomedical research, to exploit the wealth of protein structure information.

In this chapter we aim to provide a brief overview of the current status in protein structure determination, and summarize how protein structure analysis is integral to two active and growing areas of biomedical research, namely understanding genetic variations at a protein level to help disease diagnosis and guiding the development

of small molecule therapeutics. Due to the broad subject matter, it is beyond the scope of this chapter to provide an extensive discussion of all significant developments in the ever expanding applications of structural biology. We, however, refer the interested reader to some excellent articles in the relevant sections for more in-depth reviews. We also apologize to all those colleagues whose important work could not be cited, or was cited indirectly, because of space consideration and reference limits.

1.2 Methods of Obtaining Structural Information

1.2.1 Experimental Approaches

As of December 2011, the Protein Structure Database (PDB) contained ~77,700 protein structures in the public domain (http://www.pdb.org). These 3D structures are experimentally derived by methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (EM). Among them, X-ray crystallography is the dominant structure provider (Fig. 1a), contributing ~87% of total PDB entries. Since the first protein crystal structure in the 1960s (that of myoglobin [1]), the field of protein crystallography has made tremendous technological advances in all stages of the structure determination process. Examples of such development include the use of heterologous systems (e.g. bacteria, baculovirus-infected insect cells, yeast) to recombinantly express proteins in milligram quantities [2], use of fusion tags and automated chromatography platforms to purify proteins, use of robotics in performing nanolitre-scale crystallization experiments [3], improvement of synchrotron and in-house X-ray sources that reduces data collection time and extends resolution limits [4]; and software development to accelerate the *in silico* data processing steps [5]. At present high-resolution crystal structures can often be determined within days of obtaining diffraction-grade crystals.

A second method of structure determination, solution NMR, analyzes resonance assignment derived from short-range inter-proton distances in a protein (Fig. 1b). Compared to crystallography, which requires the protein in a crystalline state, solution NMR benefits from studying the protein in its native form, allowing the observation of protein conformational dynamics and flexibility [6]. NMR provides an alternative route to structure determination, especially for proteins difficult to crystallize, contributing ~11% of total PDB entries. It is also very informative in mapping ligand binding residues, by titration of the ligand onto the protein and analyzing chemical shifts in a heteronuclear single-quantum correlation (HSQC) spectrum. However, solution NMR consumes a considerable amount of radioactive isotope-labelled protein sample and time in the resonance assignment. There is also a size limit for proteins amenable to solution NMR measurement (<30 kDa), although this limit will continue to be pushed back by technological improvements [7].

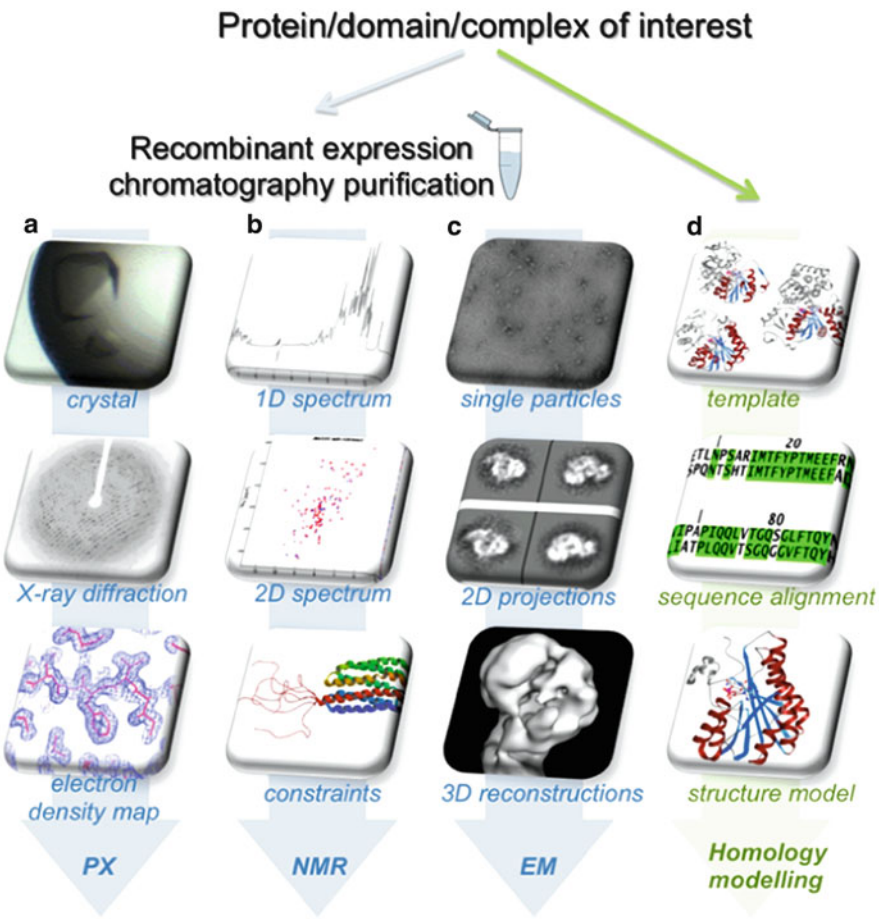

Fig. 1 Methods of protein structure determination. Experimentally protein structures can be determined by protein crystallography (PX), nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (EM). These methods take advantage of recombinant technology that facilitates the heterologous expression and purification of protein domains or complexes. Structure models can also be constructed by homology modeling, if a structure template homologous to the protein of interest is available

Electron microscopy (EM) can determine macromolecular structures at medium to low atomic resolution, using single particle analysis where individual protein molecules are imaged in solution and the 3D structure is reconstructed by back projection of the 2D images (Fig. 1c) [8]. EM is useful in studying multiprotein supramolecular complexes, particularly when combined with crystallography studies of the protein components. This allows the fitting of the individual proteins, of which crystal structures were determined, into the molecular envelope of the intact complex as determined by EM, to understand their relative orientations within the complex. However the use of EM in small molecule development and understanding genetic variations is currently limited by the data resolution restraint.

1.2.2 Homology Modelling

Of the ~30,000 or so gene products predicted for the human genome, only around 15% have been structurally characterized by the experimental methods outlined above. For the many remaining proteins in human and other organisms, computational modeling continues to bridge the gap between known sequences and available structures. The method of comparative or homology modeling allows a structural model to be constructed for a target protein based on its similarity to one or more known structures [9], on the premise that proteins sharing similar sequences fold into similar 3D structures [10]. Today, a number of modeling programs are available (e.g. MODELLER, salilab.org/modeller), some developed as online servers where a sequence-to-structure process can be performed simply by a few clicks. Their popular usage can in part be reflected by the number of available protein models in online repositories such as SWISS-MODEL (swissmodel.expasy.org), Modbase (modbase.compbio.ucsf.edu) and Protein Model Portal (www.proteinmodelportal.org). Due to its popularity, homology modeling has played an influential role in functional annotation and drug discovery for many protein families, e.g. kinases and GPCRs (see examples in [11]).

Common to all modeling tools and servers is an overall four-step procedure (Fig. 1d): (1) given the sequence of the target protein, homologues with known 3D structures are identified; (2) a sequence alignment between the target protein and homologues assigns residue correspondence between sequences; (3) the alignment guides the model building of the target protein, using the homologue structure as template; (4) finally, the constructed model is subjected to refinement and validation of its stereo-chemical properties. In general, the accuracy of homology models depends heavily on the suitability of the template, with higher sequence homology between target and template resulting in less positional errors (as measured by root-mean-square deviations, rmsd, between their corresponding main-chain atoms). In practice, sequence identity cut-offs between 40% [12] and 70% [13] have been used to produce reliable models for understanding protein function and drug discovery (at a 60% identity level rmsd is usually <1 Å). Models derived from lower identity templates ($<30\%$) often have higher main-chain and side-chain errors due to a poor quality sequence alignment with too many position gaps [14].

2 Protein Structure Analysis in Understanding Genetic Variations

2.1 Studying Diseases in the Next Generation Sequencing Era

The recent advent of next generation sequencing (NGS; also known as massively parallel sequencing) has progressed from the time- and cost-consuming Sanger sequencing models to much quicker and cheaper methods [15], and revolutionized

our approaches to study the relationship between genotype and disease [16]. Making particular impact has been the use of exome sequencing (i.e. all exons in a genome) to investigate the genetic bases of rare Mendelian disorders with low and sporadic incidence in the population [17]. Its success stems partly from not being technically limited by small patient sample size, a major hurdle with conventional methods of disease gene discovery such as linkage analysis and homozygosity mapping. Today, exome sequencing has led to the discovery of new pathogenic variants and candidate genes for a number of genetic disorders (e.g. Miller syndrome [18], Freeman–Sheldon syndrome [19], Kabuki syndrome [20]), and has also offered opportunities to study complex polygenic diseases (e.g. diabetes, Alzheimer's and heart disease) where susceptibility is affected by multiple genes with complex inheritance patterns.

NGS has therefore accelerated the rate of identifying variants in the human genome. An increasing emphasis is now placed on the effects of these variations on health and disease, although sieving through this huge volume of variant data is a laborious task. Most genetic variations occur at the single nucleotide level, represented as either single nucleotide polymorphisms (SNPs) if they have an incidence of $>1\%$ in the genome [21], or as rare variants with $<1\%$ occurrence. Rare variants, like SNPs, can be pathogenic (i.e. disease linked; often termed conveniently as mutations) or benign (i.e. not disease linked). Of particular importance to disease diagnostics are those SNPs and rare variants that lead to amino acid substitutions (missense variants) for two reasons. First, the contribution of missense variations to disease is much higher than the summation of all other variant types (e.g. frameshifts, insertions, deletions, splicing, nonsense), with 60–75% of Mendelian disorders caused by amino acid substitutions [22, 23]. Second, while the consequences of most nonsense, frameshifts and insertions/deletions are self-evident (e.g. resulting in truncated proteins), the effects of missense variations on protein function and stability are more subtle and difficult to predict. Structural information at the protein level is therefore needed to understand fully their molecular effects.

2.2 *Structural Characterization of Missense Variations*

While traditionally not a front-line method of analysis, protein structural information has increasingly been incorporated into bioinformatics and *in silico* methods to characterize missense variants and predict their pathogenicity at the molecular level. In the following subsections we outline several approaches of structure-guided investigation of missense variations and the lessons learnt from these studies.

2.2.1 Bioinformatics Predictors

Following the identification of genetic variants, the next indispensable step is to discriminate between pathogenic and benign variations. The sheer volume of

genomic data, however, makes it too time-consuming and expensive to characterize every missense variant experimentally. To this end, numerous bioinformatics methods have been developed over the past decade to predict their molecular effects, and thus help prioritize a set of variants to be studied functionally. A number of excellent reviews on the available computational tools have been published recently ([24–26] and references therein for programs described below). Many prediction tools are implemented as online servers, taking an input sequence, and applying various algorithms to sort and score mutations by their pathogenicity. Structure-based algorithms, which identify a structural match to the input sequence and analyze the contributions of the variant amino acid to protein structural properties such as electrostatics, inter-residue contacts, and steric effects [27], are increasingly incorporated into prediction servers. They serve as complementary approaches to the sequence-comparison programs (e.g. SIFT, Panther and PhD-SNP) that are based on the premise that disease-causing mutations are generally concentrated at conserved amino acids with critical roles in protein structure and function [28]. Nowadays, many prediction methods combine both structural information and sequence conservation to improve their prediction performance and accuracy (e.g. nsSNP Analyzer, PolyPhen-1/2, SNAP, SNP&GO and SNPs3D). An emerging trend is to utilize multiple sets of prediction programs and servers to increase confidence in interpreting the predictions, since different algorithms use different information and have their own strengths and weaknesses. Currently there is an urgent need for standard classification as well as unbiased and statistically-relevant comparisons among the various programs, an active area of bioinformatics research [29, 30].

2.2.2 *In Silico* Structural Analysis

Protein structure analysis offers a promising avenue to study the molecular consequences of missense variants, by revealing the atomic environment surrounding the mutation site *in silico*. The most direct approach is by experimental structure determination of the protein in its mutant form if it can be expressed recombinantly and purified. This, however, often proves difficult, in part due to unstable conformations of the mutant proteins that lead to their intracellular degradation. Indeed, three-quarters of disease-associated missense mutations are postulated to destabilize the protein as their primary functional defect [12, 31]. Therefore, although thousands of proteins involved in different biological pathways and functions have been structurally characterized, only a very small proportion of these structures represent proteins inclusive of a disease associated mutation. Recent structure examples falling into this category include the ryanodine receptors RyR1 and RyR2 [32], FGFR2 tyrosine kinase domain [33] and glycogenin GYG1 [34]. As structural determination of mutant proteins often proves intractable, the alternative is to "model" the missense variation onto the wild-type structural environment, by fitting the new amino acid side-chain into the substitution site. The modeled amino acid is inspected visually using molecular graphics software,

such as PyMOL (Schrödinger, LLC.), Swiss-PDB viewer (Swiss Institute of Bioinformatics, Basel) and ICM (Molsoft, La Jolla), with particular attention paid to identifying the most acceptable side-chain conformation, from a library of allowed side-chain rotamers, that results in minimal steric clashes. This structure model is then subjected to refinement and energy minimization to yield an overall stabilized conformation.

The available mutant model, either from experimental methods or mutation modeling, can then be analyzed *in silico* to assess the impact of the amino acid substitution on a number of structural properties. These include possible changes in secondary structure elements, solvent accessibility, packing of neighbouring atoms and inter-atomic/inter-protein contacts, many of which can be examined using online tools ([24] and references therein). The *in silico* observations allow hypotheses about the molecular nature of the mutational defects to be made, and subsequently tested using a variety of biophysical and biochemical assay methods. Oligomeric state of the protein, for example, may be assessed by native gel electrophoresis, size-exclusion chromatography (SEC), analytical ultracentrifugation, or dynamic light scattering [35, 36]. Secondary and tertiary structure contents may be assessed by far-UV circular dichroism (CD) [37]. Protein unfolding may be monitored by chemical or thermal denaturing detected with far-UV CD or fluorescence [38]. Functional interactions with protein partners can be determined using co-immunoprecipitation followed by Western blot and SEC [39]; thermodynamics of protein binding to ligand or peptide can be determined via isothermal calorimetry (ITC) or surface plasmon resonance (SPR) [40]. Enzymatic catalysis and Michaelis–Menton kinetics can be measured if an assay specific to the protein of interest is available [33, 41]. The above list is non-exhaustive, as there are many options to examine every aspect of a protein's functional properties in the laboratory. Regardless of the approach(es) chosen, however, it is important to compare the results obtained with the mutant protein against that of wild-type before interpretations are made. It is also important to complement in vitro observations with in vivo studies to comprehend fully the physiological consequences, for example by introducing the variants into the relevant cell lines or genetically engineered animal models.

2.3 The Structural "Rule-Book" Governing Missense Variations

In the following section we review examples illustrating how a structural analysis of the atomic environment surrounding the variant residues, complemented with biochemical and biophysical studies, can be used to attribute deleterious phenotypes to different molecular effects. Together, these examples allow us to formulate a set of "structural rules" to help predict the likely deleterious effects of a missense variation, and can serve as an important toolkit for clinicians and geneticists who need to assess the disease relevance for any newly-identified variations.

2.3.1 Disrupting Protein Fold and Architecture

Phenylketonuria as a Paradigm of Misfolding Diseases

Computational analysis of disease causing variations predicts that ~75% of mutations lead to protein destabilization, while only 7% directly affect biochemical function, suggesting that, for many monogenic diseases, a change in protein stability is the major contributor to disease pathology [12, 31, 42] and giving rise to the concept of misfolding diseases [43]. A classic example of a misfolding disease is phenylketonuria (PKU; OMIM 261600) caused by destabilizing mutations in phenylalanine hydroxylase (PAH). PKU is the most common inborn error of amino acid metabolism (incidence of ~1 in 15,000) with more than 500 deleterious mutations reported, 60% of which are missense mutations scattered across the polypeptide [44]. The majority of mutations result in enzyme forms with reduced stability and a propensity to aggregate, resulting in protein degradation and turnover [45]. To understand how these mutations lead to a misfolded state, the available crystal structures have served as excellent tools to scrutinize the atomic environment of the missense mutation sites [46, 47] and to correlate between genotypes and phenotypes [48–50].

A common cause of destabilizing mutations is a structural perturbation to the protein core by a number of molecular mechanisms, depending on the nature of the original wild type and mutant residues. (1) Mutation of a large buried residue to a small one will create an unfavourable solvent cavity within the core, with larger cavities resulting in greater destabilization [51]. In PAH, mutations of buried phenylalanines (F39L, F55L, F372L), valines (V177A, V190A, V245A) and leucines (L255V, L348V) to smaller residues are commonly found (Fig. 2a). (2) The reverse of the above is also true. Mutations of small residues to large ones require the protein to accommodate bulky side-chains by disturbing the surrounding packing and secondary structure arrangements. Examples in PAH include a number of alanine-to-valine substitutions (A47V, A246V, A259V, A403V) (Fig. 2b). (3) Mutations of non-polar residues within a hydrophobic environment to polar residues may also destabilize a protein because of the thermodynamic penalties incurred on the unbonded polar group. These include mutations of isoleucine to serine (I94S) or threonine (I164T, I174T) or mutations of leucine to serine (L48S, L255S) (Fig. 2c). (4) Finally, mutations of polar and charged side-chains to hydrophobic ones may remove important stabilizing contacts (e.g. electrostatic or hydrogen-bonding interactions). This is especially true of arginines, such as Arg241 (R241C, R241H) and Arg252 (R252G, R252Q, R252W) in PAH (Fig. 2d).

In contrast to core residues, very few protein destabilizing mutations reside on the protein surface, as they can often be substituted with little effect [51]. However, there are exceptions if the mutation disrupts a hydrogen bond or electrostatic interaction at the surface (e.g. D84Y, R176L, R413P mutations in PAH) (Fig. 2e), or if the mutation affects the functional oligomeric state. To this end, PAH forms a tetramer, and mutations that interfere with its tetramerization, e.g. the

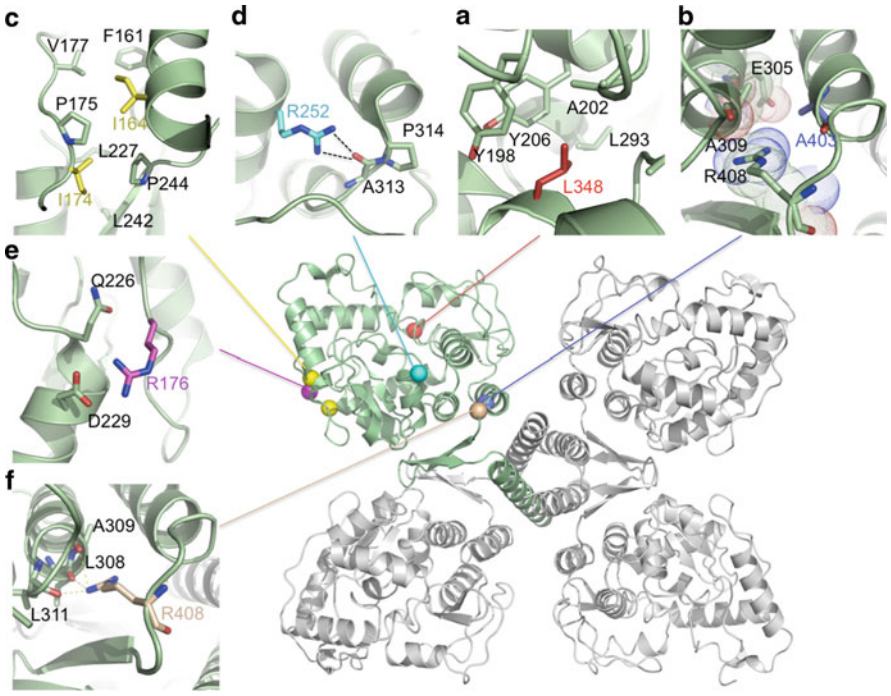

Fig. 2 Structure of human phenylalanine hydroxylase PAH. The tetrameric architecture of PAH (PDB code 2PAH) is shown with one of its monomer subunits coloured in *green*. Six regions of the PAH monomer are highlighted in panels **a**–**f** to illustrate the different molecular mechanisms that can govern a destabilizing missense mutation. These include (**a**) mutation of larger to smaller residues; (**b**) mutation of smaller to larger residues; (**c**) mutation of nonpolar to polar residues in hydrophobic core; (**d**) mutations of polar to nonpolar residues; (**e**) mutation of surface polar residues; and (**f**) mutations of residues involved in the oligomerization interface

single most common PKU mutation, R408W, which results in the loss of an inter-subunit hydrogen-bond (Fig. 2f), causes improper oligomeric assembly and hence reduces stability [47].

"Special" Residues: Glycine, Proline and Cysteine

Amino acids such as glycine, proline and cysteine often impart certain structural constraints on the protein, and their substitutions can be deleterious. Proline, with its cyclic side-chain, restricts the protein backbone conformations. Therefore, mutations to proline often distort the native backbone conformation, and interrupt the α-helix or β-sheet in which the mutated amino acid resides. The L166P mutation in the DJ-1 protein, located in the middle of helix α7 in its crystal structure (Fig. 3a), is one of the most deleterious missense mutations linked with early onset Parkinson's disease. A combination of NMR, CD and molecular dynamics studies

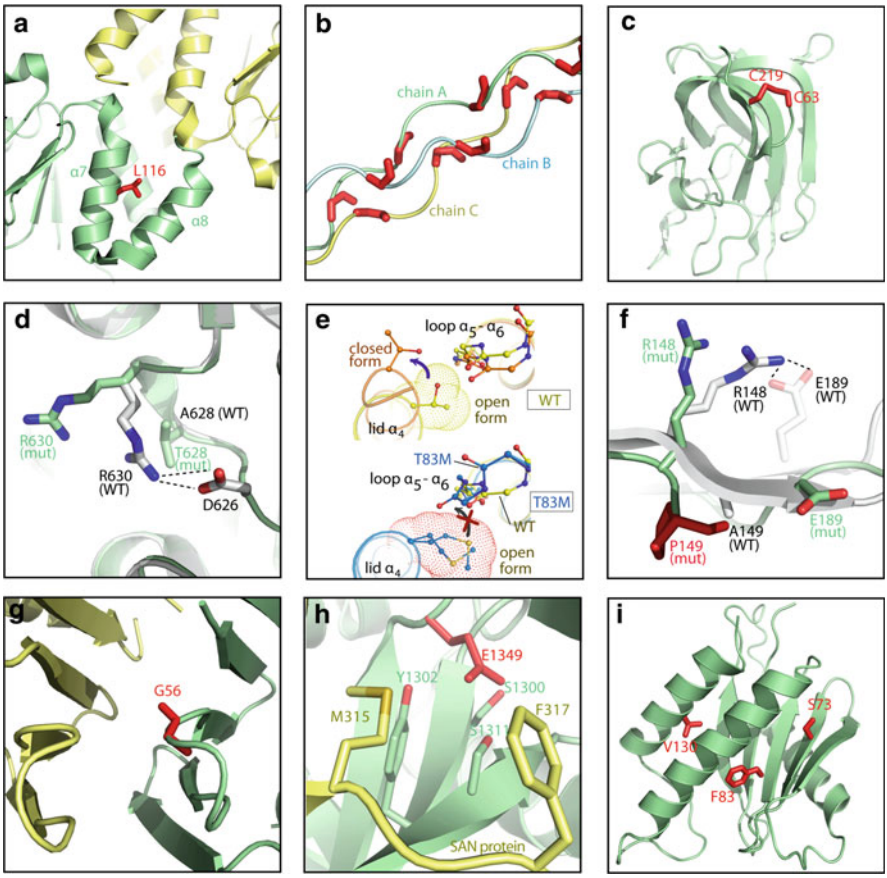

Fig. 3 Defective protein functions due to missense mutations. Where applicable, the site of mutation described in the text is coloured *red*. (**a**) Human DJ-1 protein (PDB code 1PDV). Two monomeric subunits (*green, yellow*) are shown. (**b**) Collagen-like peptide (1CAG) in a triple helix conformation. Glycine residues are shown in *sticks*. (**c**) Structure of Factor VIII C2 domain (1IQD) that is homologous to retinoschisin highlights the highly-conserved disulphide bond (Cys63–Cys219 in retinoschisin). (**d**) Structure of FGRF2 tyrosine kinase domain in wild-type (1GJO, *white*) and A628T mutant (3B2T, *green*). (**e**) Structures of human glycogenin-1 show that the conformational movement of lid α4 in the wild-type (3T7O, *top*) is forbidden in the T83M mutant protein (3RMW, *below*). (**f**) Aldolase B in the wild-type (1QO5, *white*) and A149P mutant protein (1XDM, *green*). (**g**) Spermine synthase (3C6K). The G56S mutation is located at the dimer interface (*yellow*, *green*). (**h**) Myosin MyoVIIa (*green*) in complex with SAN protein (*yellow*) (3PVL). (**i**) SDELIN protein (1H3Q) with the site of missense mutations disrupting transcription factor binding shown in *red sticks*

have shown that the L166P substitution causes DJ-1 to lose α-helical content and leads to global structural destabilization. Since helices α7 and α8 engage in numerous inter-molecular contacts, the mutant is also incapable of functional dimer formation [37].

With the absence of a side-chain, glycine is the smallest of all amino acids and possesses conformational properties and freedoms inaccessible to other amino acids. Therefore, substitutions from glycine can be debilitating to protein stability and folding. The major structural component of skin, bone and tendons is type I collagen, where two α1 and one α2 protein chains are tightly packed in a heterotrimer. The intermolecular interface is mediated by many Gly-x-y sequence repeats from the three chains, forming a triple helix conformation that is essential to collagen structure and function (Fig. 3b). Many missense mutations substituting a single glycine to larger residues are known to cause the brittle bone disease osteogenesis imperfecta (OMIM 166200), with disease severity dependent upon the size of the mutant amino acid [52]. Another example involves a Gly-to-Asp mutation at the hairpin turn of glycogen phosphorylase, which causes glycogen storage disorder type VI [53].

Cysteine is unique among amino acids in its ability to form inter-residue disulphide bonds that are often critical to maintaining the protein fold. Therefore substitution of a cysteine involved in disulphide bond formation, or to a cysteine that yields an unnatural disulphide bond, may disrupt protein structure. Retinoschisin (RS), a photoreceptor and bipolar cell secreted protein, forms a large disulphide-linked multisubunit complex. At least 25% of the >125 known RS mutations result in the loss or gain of a cysteine and cause X-linked juvenile retinoschisis (OMIM 312700). A combined biochemical and modeling study showed that among the disease causing mutations, C142W and C219R resulted in the breakage of intra-subunit disulphide bonds (Cys110–Cys142 and Cys63–Cys219, respectively) (Fig. 3c), while C59S and C223R abolished an inter-subunit disulphide bond (Cys59–Cys223) [54], hence providing a molecular explanation to how these mutations lead to misfolded protein, defective subunit assembly and aberrant subcellular localization.

2.3.2 Disrupting Protein Functions

While less prevalent than destabilizing mutations, an amino acid substitution can lead to the specific loss, or diminishing, of a protein functional property, such as catalysis, protein–protein interactions, and oligomerization. A number of recent structure examples that are complemented with functional studies are described below.

Affecting Enzyme Catalysis

Mutations in the tyrosine kinase domain (e.g. A628T) of fibroblast growth factor receptor 2 (FGRF2) cause lacrimo-auriculo-dento-digital syndrome (OMIM 149730). Ala628 is a highly conserved residue in the active site catalytic loop. The crystal structure of $FGFR2_{A628T}$ mutant protein reveals that substitution of Ala628 to a more polar and bulky threonine residue alters the configuration of key residues in

the active site that are involved in tyrosine substrate binding [33]. For example, the side-chain of Arg630 has been shifted 160° away (Fig. 3d) and cannot coordinate with the substrate. This observation is supported by activity assays showing weakened substrate binding and severely impaired tyrosine kinase activity [33].

A new form of glycogen storage disorder (GSD15; OMIM 613507) has recently been identified with genetic defects in glycogenin (GYG1), a glycosyltransferase that catalyzes the initiation of glycogen synthesis. The complete structural snapshots of GYG1 along its catalytic cycle have been provided by X-ray crystallography and show a substantial "lid" movement that closes the active site for catalysis [34]. The disease-linked mutation T83M incorporates a bulky Met side-chain into the mobile "lid" region and prevents the essential movement, as revealed in the mutant protein structure (Fig. 3e). As a result, the glycosyltransferase activity of $GYG1_{T83M}$ is completely abolished.

Disruption of Quaternary Structure

Hereditary fructose intolerance (OMIM 229600) is caused by mutations in aldolase B, the most prevalent being A149P. The mutant protein structure shows that the A149P substitution disrupts the β-strand element at the mutation site, abolishes a salt-bridge at the adjacent Glu148 residue (Fig. 3f) and also produces a distal effect causing disorder in the 110–129 loop at the dimer–dimer interface [55]. This offers an explanation as to why the mutant protein exists as a solution dimer, and cannot form the homotetramer essential for its catalysis [35]. This study also nicely elucidates the long-range structural perturbations caused by a single amino acid substitution, an observation which would not have been elucidated by modeling a mutant side-chain onto the wild-type structure.

Genetic defects in spermine synthase (SMS), an enzyme converting spermidine to spermine, cause the X-linked disorder Snyder–Robinson Syndrome (OMIM 309583). The crystal structure of SMS reveals that the protein is a homodimer, with the G56S disease mutation lying close to the dimeric interface (Fig. 3g). Any side-chain incorporated at this position is postulated to protrude towards the opposite subunit and disrupt dimer stability, a hypothesis supported by native gel analysis showing the absence of dimer formation in the mutant protein [36].

Disruption of Protein–Protein Interaction

Mutations in the myosin protein MyoVIIa, part of a complex network of proteins in the stereocilia of the inner ear, cause syndromic deaf-blindness (OMIM 276900). A recent structural determination of the MyTH4-FERM tandem domain of MyoVIIa in complex with its protein binding partner Sans reveals that the Glu1349 mutation site on MyoVIIa forms direct interaction with Sans (Fig. 3h), and as a result a single E1349K substitution is responsible for a 20-fold reduction in binding affinity towards Sans, as measured by ITC [40].

Four missense mutations on the SDELIN protein, a subunit of the endoplasmic reticulum Transport Protein Particle complex, are known to cause the X-linked rare bone disorder spondyloepiphyseal dysplasia tarda (OMIM 313400). Three of these mutations (S73L, F83S and V130D) are located in a hydrophobic pocket (Fig. 3i) that is proposed to function as a binding site for transcription factors such as MBP1, PITX1 and SF1, on the basis of the SDELIN crystal structure. Yeast two-hybrid studies have confirmed that these three mutations indeed resulted in a loss of protein–protein interactions [56].

2.3.3 Hot Spot Regions

In addition to visualizing the atomic environment of individual mutation sites, as detailed in Sects. 2.3.1 and 2.3.2, structure analysis can also be employed at the whole protein level, for instance, to map all known variations onto the protein 3D structure and identify "hot spot" regions that harbour a high frequency of missense variations. Hot spot mapping can provide insight into phenotype–genotype relationship of mutations in a 3D structural context, and assist in disease diagnosis, for example, by focusing screening efforts on selected mutation-prone regions instead of over an entire gene, most of which may harbour no known mutations. Hot spot mapping can also help generate new conclusions about protein functions and evolutionary mechanisms such as mutability and selection pressure of different mutations by illustrating which regions of a protein can tolerate amino acid variations and which regions are intolerant. A classic example of hot spot identification is with the most commonly mutated cancer gene, TP53 (p53). In p53, an overwhelming majority of its somatic missense mutations are clustered into a loop-sheet-helix region of the DNA-binding domain (Fig. 4a) [57]. These mutations generally disrupt the DNA binding interface and hence mutant proteins are defective in sequence-specific DNA binding [58].

More recent examples of hot spot mapping can also be found in the literature. Mutations on the ryanodine receptors RyR1 and RyR2 (cf. Sect. 2.2.2) that lead to skeletal muscle disorders are concentrated in a highly basic loop (Fig. 4b) and have been found not to affect protein stability, but rather to disrupt the protein–protein or domain–domain interface [32, 59]. In another example, 15 missense mutation sites on dyskerin, the catalytic subunit of the Box H/ACA ribonucleoprotein particles, have been identified to cause a bone marrow failure called X-linked dyskeratosis congenita (OMIM 305000). The recently determined structure of the yeast homologue Cbf5 reveals that these mutations are all located in a 32-residue N-terminal extension (Fig. 4c) that forms an additional layer to the well-characterized RNA-binding PUA fold, a structural feature not found in archaea, and may function in protein–protein binding [60]. Within our group, we have mapped 55 known missense mutations causing fumarate hydratase deficiency (OMIM 606812) onto its human protein structure [61] and identified two hot spot regions, one clustering around the active site and the other affecting intra- and inter-subunit interactions. To aid further investigation by interested doctors/researchers, the online version of

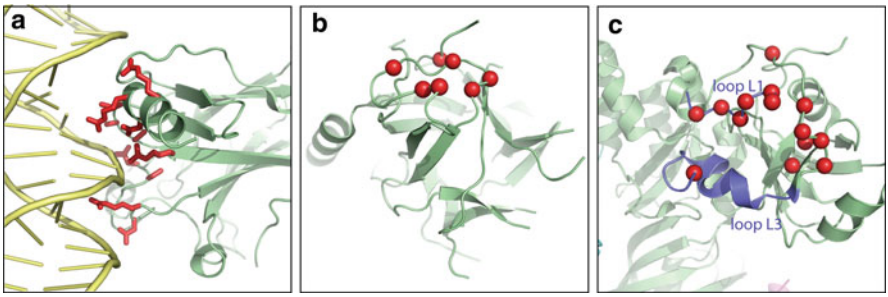

Fig. 4 Structure mapping of mutation hot spots. Mutation sites are shown in either *red sticks* or *spheres*. (**a**) p53 central domain in complex with DNA (PDB code 1TSR); (**b**) ryanodine receptor type 1 N-terminal domain (3HSM); and (**c**) yeast Cbf5 structure that is homologous to human dyskerin (3U28). The eukaryote-unique N-terminal extension (loops 1 and 3) is coloured *blue*

this article is accompanied with a web-based molecular viewer, allowing the reader to navigate the hot spot regions and each individual mutation along the protein landscape, in an interactive manner [62].

2.3.4 Lessons from Large-Scale Structural "Catalogues"

Taking advantage of the rapidly growing genomic and structural data, there are now efforts being made to catalogue missense variants on a large scale using vast protein datasets. Some of these efforts have focused on disease relevant protein families, such as kinases. For example, Lahiry et al. [63] used structural analysis of kinase mutations to correlate their locations on the protein to disease states. They observed that: (1) neutral mutations/polymorphisms, those that did not tend to cause disease, generally clustered in the C-terminal regions of the catalytic core, a region thought to have a basic structural role; (2) germline disease causing mutations, which cause metabolic disorders or loss-of-function developmental disorders, tended to cluster in the catalytic core in sites involved in regulation and substrate binding, as well as in protein–protein and allosteric interactions; (3) cancer causing somatic mutations were concentrated around the ATP binding and catalytic residues, directly influencing catalysis and resulting in the activation of oncogenes or deactivation of tumour suppressors.

Other studies have employed large datasets of missense variations spanning different protein families in order to detect any trends, consensus or "rules" which dictate whether certain types of amino acid changes will result in neutral polymorphisms or pathogenic mutations [64, 65]. These large-scale studies have generally arrived at the conclusion that pathogenic variants are more likely located in solvent-buried core regions, conserved residue positions, residues that contribute hydrogen bonds and those that alter more dramatically the physico-chemical properties of amino acids [64, 65]. Khan et al. [66] also looked at the distribution and frequency of pathogenic variations and found that arginine and glycine are the

most mutated residue types, while overall mutability (i.e. the likelihood of being introduced in missense variations) is highest for cysteine and tryptophan. Using similar approaches, Hurst et al. [65] found that mutations of glycine, cysteine and tryptophan were more likely to be pathogenic than others, confirming results from previous small-scale studies. They also provided online access to their large database of structurally-mapped missense variations (www.bioinf.org.uk/saap/db). Taken together, these proteome-wide structure-based mutation analyses will continue to help us formulate better rules for our prediction of whether an uncharacterized amino acid variation will be pathogenic or not, thereby improving disease diagnostics in the future.

3 Protein Structure Analysis in Drug Development

3.1 Structural Biology and Target-Centric Drug Development

The opportunities presented from the post-genome era have also transformed rapidly the field of drug development. We are now made aware of the unprecedented number of potential therapeutic proteins, estimated in one study as reaching 10% of the predicted coding regions in the human genome (i.e. ~3,000 proteins) [67]. On the other hand, the current FDA-approved drugs target only a small number (~300) of human proteins or proteins from other pathogenic organisms [68–70]. This has made a fundamental impact in the direction of biomedical research in steering towards a more target-centric approach to bridge this gap. The main focus in this approach is to identify therapeutically-relevant drug targets that meet the double criteria of being disease-linked i.e. it has a causative role in the onset and/or progression of a disease, and being druggable i.e. it can be bound and modulated by a small molecule.

At the same time, the current field of drug development is facing tremendous challenges with ever-increasing research and development costs (reaching in some estimates up to $2 billion per drug [71]) and high attrition rate along the entire pipeline [72], where many potential projects fail through the early stages of hit identification and optimization to lead. As a result, the pharmaceutical industry is under continuous pressure to look for novel, high confidence disease targets and alternative drug design approaches. Amenable to the target-centric approach while having potentials in addressing some of the challenges in the pharmaceutical industry, the field of structural biology has been playing an increasing role in drug development, particularly at the early stages (“drug discovery”). Today, using structures to identify new lead compounds and as a basis for rational drug design is an integral part of many a drug development project.

3.2 Early Structural Applications in Lead Optimization

Before the technological advances in the past decade that have made protein structure determination faster and more cost-effective, the use of structure information in drug development in the 1980s and early 1990s has been confined to the lead optimization stage, in directing the chemical alterations of initial compound hits to improve their affinity, potency and selectivity. In this process, the protein structure of interest is determined in complexes with lead compounds identified from a high throughput screening (HTS) campaign, accomplished either by co-crystallizing the protein solution pre-incubated with the lead molecule, or by soaking pre-formed crystals of the apo protein with the ligand solution. The determined structure of the protein–ligand complex reveals the modes of interaction between the protein and ligand at the atomic level, e.g. short-range interactions such as hydrogen bonds, salt bridges, and hydrophobic contacts, the distances between the various interacting groups and atoms, and the presence of water molecules at the protein–ligand interaction site. This information is used to guide further iterative rounds of chemistry optimization and protein–ligand structure determination to establish a structure–activity relationship.

The first marketed drug developed via this structure-based approach was captopril, an inhibitor for angiotensin converting enzyme for the treatment of hypertension and congestive heart failure. This drug was designed in the mid-1970s on the basis of the homologous carboxypeptidase A protein which had been structurally characterized at the time [73]. Today, structure-based design approaches have delivered drugs to the market for a wide range of diseases, including retroviral [74, 75], glaucoma [76], influenza [77, 78] as well as cancer [79, 80] (Fig. 5). With advances, particularly in crystallography, the timeframe of protein structure determination is now sufficiently short to be amenable for many other stages of the drug discovery pipeline. As a result, the tools of structural analysis that were traditionally used in lead optimization are now being exploited to assist the processes of target identification, assessment of target druggability, and hit identification (Fig. 6), as outlined below.

3.3 Use of Structures in Target Identification

An early consideration in the target-centric approach of drug discovery is to identify and prioritize therapeutically-important proteins in the genome. Obtaining structural information at this stage, in the apo- or relevant liganded states of the protein, is an important milestone in target identification. High resolution atomic structures of many therapeutic targets are now available in the public domain, including kinases (e.g. AMPK [81]), viral proteins (influenza polymerase [82]), cytochrome P450 [83], metabolic enzymes (acetyl-CoA carboxylase [84]) and G-protein coupled receptors (β1-adrenergic receptor [85]). This unprecedented wealth of structure information helps establishing sequence–structure–function relationship and assessing potential ligand-binding capabilities, and is now part of the essential

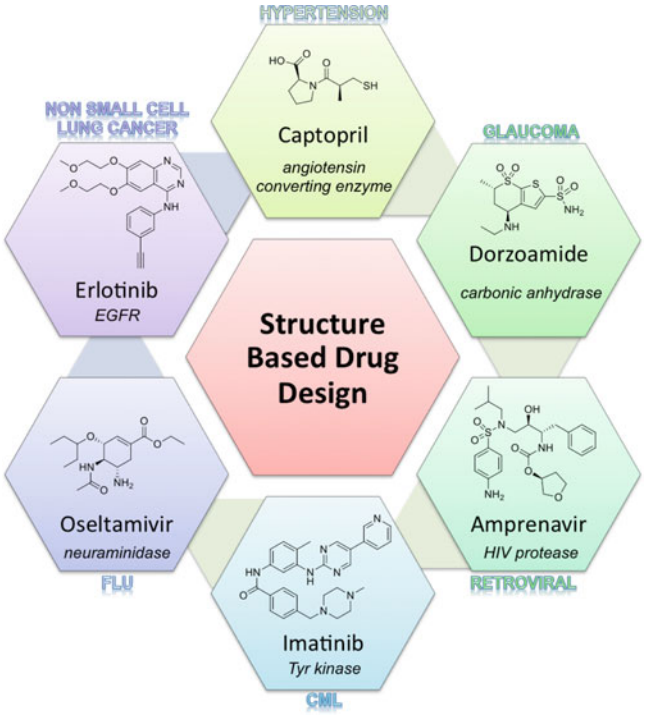

Fig. 5 Examples of structure-based drug design. FDA-approved drugs that have been derived from structure-based approaches. For each drug, its generic name, chemical structure, protein targeted and disease area applied is shown (references in the main text). *CML* chronic myeloid leukaemia; *EGFR* epidermal growth factor receptor

toolkit to complement in vivo target validation experiments (e.g. RNA interference screens, animal models, gene knockouts). The increase in available structures in the PDB also spurs the development of computational methods combining sequence and structural information to probe biological functions [86].

It is with the technological advances in structural biology that the field of "structural genomics" (SG) was born, to determine systematically 3D structures of proteins encoded in a genome primarily by crystallography and NMR. The overall objectives are to provide a structure coverage of the "protein universe" [87], to help define protein functions that cannot be predicted from sequences alone [88], and to facilitate the discovery, as well as selection, of genomic targets for drug therapy [89]. A number of large-scale SG efforts have emerged over the past 10 years, including RIKEN in Japan (www.riken.co.jp), Structure Proteomics in Europe (SPINE, www.spineurope.org), the Structural Genomics Consortium based in UK, Sweden and Canada (SGC, www.thesgc.com), as well as the Protein Structure Initiative in USA (PSI; www.nigms.nih.gov/psi). While sharing similar high-throughput methodologies and open access policy to their data, these SG initiatives differ in their scope and criteria for their target selection. A number of SG initiatives (e.g. PSI, RIKEN) aim to explore the novel protein folds that cannot

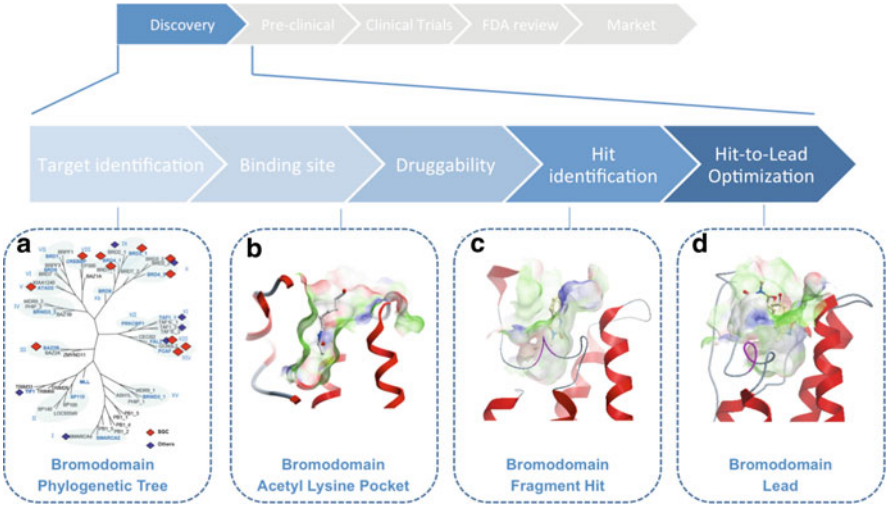

Fig. 6 Modern day structure-guided drug discovery. Protein structure analysis is nowadays incorporated into all early stages of drug development, including (**a**) target identification; (**b**) assessing binding site druggability; (**c**) hit identification, and (**d**) lead optimization. Example shown is from the chemical probe program at the Structural Genomics Consortium to develop small molecule binders for the family of histone-binding bromodomains (cf. [93] in main text)

be predicted from sequence [90], and subsequently leverage structure completeness of a genome by homology modeling of the remaining homologous proteins. Other SG programs take a biology-driven avenue, placing the emphasis more on medical relevance. For example, the Tuberculosis Structural Genomics Consortium [91] adopts an organism-based approach focusing on the obligate human pathogen *Mycobacterium tuberculosis*. To date nearly 10% of all proteins from the pathogen have been structurally characterized [92], which unravel a number of previously unannotated proteins as potential anti-tuberculosis targets. The human proteome-focused SGC studies protein families with therapeutic importance such as kinases, phosphatases and metabolic enzymes [93, 94]. These studies reveal structure–-function relationships between family members with regards to active site and substrate specificity (Fig. 6a), and emphasize their application to develop member/family-specific chemical probes and inhibitors [95].

3.4 Use of Structures in Assessing Druggability

3.4.1 Binding Site Detection

With the structural information of potential therapeutic targets made available, the next step in drug discovery is the identification of binding sites that are receptive to small molecule binding (Fig. 6b). The large repertoire of protein–ligand complexes

in the PDB has provided a structural view of a ligand binding site to be a small pocket or invagination on the protein, accessible to the surface exterior, where ligands can fit to mediate a biological function. This pocket should harbour amino acid side-chains that contribute to hydrogen bonds and hydrophobic contacts. Based on these concepts, a number of pocket identification software have been developed to detect binding sites on the protein structure, adopting two general approaches (see [96] and references therein). The geometry-based methods (e.g. SURFNET, LIGSITE) look for geometrically-complex regions on the protein as natural binding sites tend to be concave surface invaginations. The probe/energy-based methods (e.g. GRID, AutoLigand, ICM) calculate the interaction energy between a probe molecule and protein at different point locations to define regions with favourable interaction energies.

A thorough understanding of the binding pocket space helps not only to assess its potential for drug binding but also to annotate functionally under-characterized proteins (i.e. de-orphanization). For example, delineating residues involved at the ligand binding site can stimulate site-directed mutagenesis experiments to probe their catalytic or regulatory roles. Structural characterization of binding sites also reveals the ligand-induced conformational changes on the protein target, which can range from small side-chain adjustment to whole-domain rearrangement [97]. Binding pockets are therefore not static sites as revealed in a structural snapshot, but dynamic regions important for the protein function. The conformational plasticity of the ligand binding sites needs to be addressed during structural analysis and drug design.

3.4.2 Druggability Index

The next step following pocket detection is an evaluation of whether it has the shape and chemical complementarity to accommodate high-affinity, drug-like molecules. This likelihood prediction of drug binding ("druggability") is crucial to target selection in drug discovery, with the hope of screening out unlikely candidates at an early stage. The emerging concept of protein druggability [98] is an extension to the "drug-likeness" rule-of-five for small molecules that attributes good oral bioavailability of drug compounds to certain favourable physico-chemical parameters [99]. Research groups are developing tools to predict druggability and quantify it in a "druggability index" using different structure-based metrics. Some correlate druggability with hit rates obtained from NMR screening of small fragments [100], whereas others base their predictions on binding affinity calculations [101] or on comparison of binding sites between different proteins/families that bind the same ligand to identify hot spot residues [102]. Druggability indices are especially useful in identifying non-native small molecule binding sites such as between protein–protein interaction surfaces [103].

3.5 Use of Structures in Hit Identification

A therapeutic protein that satisfies the criteria of disease linkage and druggability can enter the pipeline of a drug discovery program to identify hit compounds that bind the target and exert an effect. Traditionally this has been achieved by HTS [104]. In this approach a vast library collection of physically available compounds, accumulated by large pharmaceuticals over many years of research, isolated from natural sources or synthesized from combinatorial chemistry, is experimentally tested on the protein target using a high-density assay that measures either binding to or biochemical modulation of a protein. The aim is to identify compounds with IC_{50} values better than, e.g. 10 μM for further hit-to-lead optimization. The power of HTS relies on the implementation of a robust and sensitive assay and the interrogation of a vast compound collection, both requirements consuming significant resources in materials, time and manpower. Its success in generating hits also depends upon target classes, robustness of the assay and propensity to deliver false positives [105]. With these challenges under consideration, novel approaches continue to be explored as complement to the HTS method in hit discovery. In particular, *in silico* methods exploiting structural information of the binding pocket space are being widely explored. To this end, three structure-based approaches, namely virtual screening, de novo design and fragment-based screening, are gaining promise and are nowadays incorporated into almost every drug discovery project (Fig. 6c).

3.5.1 Virtual Screening

Virtual screening (VS) is often considered as the computational alternative to the classic HTS, hence its alias "virtual HTS" (see [106] and references therein). VS interrogates large chemical libraries *in silico*, often available as public compound databases, to predict their binding mode and affinity towards the protein structure. The prediction is based on docking calculations and generally involves two steps. First, every compound in the library is individually placed onto the protein pocket to generate different conformations and orientations ("poses") by sampling through the pocket space, taking into account ligand and protein flexibility at the pocket. Second, the binding modes between target and the ligand in its different poses are evaluated by a scoring function, and subsequently ranked to identify binding hits from the highest-scoring ligands and poses.

A rigorous scoring function is crucial to a VS campaign, so that it allows proper enrichment of true compound hits among the top ranking scores. Many scoring functions are developed, taking into account the interaction energies between ligand and protein ("force field-based") or statistical observations from experimentally derived protein–ligand structures with the basic premise that true hits share common protein–ligand interactions ("knowledge-based"). Nowadays a variety of

docking software is available (e.g. DOCK, GOLD, AutoDock; see [106, 107] and references therein), each incorporated with different scoring functions. Current challenges in the docking tools include the need to improve scoring function accuracy, to take into account the various protonation, tautomerization and ionization states of compounds, and to predict ligand-induced protein conformations [107].

The strength of the structure-based VS approach is attributable to its capability to screen large databases (e.g. millions) of compounds with minimal computational power, more quickly and less expensively than HTS. Successful VS examples in hit identification include the development of EGFR inhibitors towards cancer cells [108], cysteine protease inhibitors of the SARS virus [109], and dihydroorotate dehydrogenase (DHODH) inhibitors towards rheumatoid arthritis [110]. The approaches of VS and HTS, with their mechanistic parallels, can also complement each other and have been applied side-by-side on the same drug development project [111] to facilitate hit identification.

3.5.2 De Novo Ligand Design

Structural knowledge of the binding pocket space can also guide the building of novel lead compounds from scratch [112, 113]. This de novo approach of drug design is not constrained by the known chemical structures from existing compounds, opening up the possibility of developing novel chemotypes [114]. The most common strategy is receptor/target-based de novo design, using a priori structural information of the target protein and its binding pocket. In this strategy, small building blocks (known as seeds or fragments) are positioned onto key interaction regions within the pocket, either by computational docking (as in Sect. 3.5.1), or recently by experimental methods such as crystallography and NMR (see Sect. 3.5.3). Each fragment can then be extended towards the neighbouring available space to build a lead compound that matches the binding pocket sterically and electrostatically ("growing" approach). Alternatively, multiple fragments bound independently at different but proximal regions of the pocket can be assembled into a lead compound using linker scaffolds ("linking" approach). A number of de novo drug design projects have yielded potential compound hits. For example, Heikkila et al. [115] exploited a species-specific hydrophobic ligand pocket on the *Plasmodium falciparum* DHODH protein to design potent parasite-specific compounds with an IC_{50} value of 43 μM. Ni et al. [116] developed inhibitors for the peptidylprolylisomerase cyclophilin A with IC_{50} values of 31.6 nM, with potentials as immunosuppressive agents. An important caveat of de novo design is that it often generates complex ligands with poor synthetic accessibility and pharmacokinetic properties. This is being addressed by software development to place emphasis on generating drug-like, synthetically-possible compounds.

3.5.3 Structure Based Fragment Screening

The X-ray and NMR methods of structure determination have also played a crucial role in a paradigm fragment-based screening approach [117]. Its premise is to screen experimentally hundreds to thousands of small compounds (usually between 100–300 Da in size) in order to identify low-affinity fragments (K_d in high μM range) that bind to different regions of the binding pocket, as a starting point for hit optimization. The subsequent optimization of fragment hits into a single hit compound can be rationalized, as in the de novo method, by the "growing" and "linking" processes. The concept of starting with small fragments is an appealing alternative to the conventional HTS attempts, with a number of merits. Fragments with their relatively small sizes and low complexity have been shown to provide higher hit rates than larger drug-like compounds from conventional screens [118], and can be optimized more efficiently. Fragments also allow a broader, more efficient sampling of the chemical space using a much smaller set of compounds (e.g. 100 fragments are equivalent to a 1,000,000 combinatory library) [114].

The relative weak binding of fragments (e.g. ~100 μM to 10 mM against target protein), which may be missed by a conventional HTS assay, can be experimentally determined by crystallography, NMR and other biophysical methods such as surface plasmon resonance [119, 120]. With inherently higher hit rates and the likelihood of multiple binding modes for a fragment hit, it is necessary to have its binding mode characterized from crystallography or NMR to allow hit-to-lead compound design. In particular, crystallography with its low-cost, high-throughput implementation is well attuned to fragment-based screening, allowing fast structure determination of protein-fragment complexes (Fig. 6d). A recent survey showed that 15 selective and potent inhibitors generated from fragment-based screening entered the phase I or II clinical trials. Examples include inhibitors for matrix metalloproteinase [121], aurora kinase [122], cyclin-dependent kinase 2 [123] and peroxisome proliferator-activated receptor [124]. An excellent update on fragment screening success examples across industry and academia was recently published [125].

4 Conclusion, Challenges and Future Perspectives

Over the past decade, the field of protein structural biology has responded to the challenging demands presented in the post-sequencing era by two revolutionary accomplishments. It has attained technological advances in the methods of structure determination in order to streamline the gene-to-structure process in a parallel, automated and miniaturized platform. Protein structures are now being solved by numerous academic and industrial research groups worldwide, on a daily or weekly basis. Structural biology has also broadened its scientific impact, successfully transforming itself from mere providers of structure information into an essential toolkit for molecular geneticists in the characterization and understanding of

diseases, and for medicinal chemists to assist all stages of the drug discovery process. With its continuing scientific contribution and technical improvements, structural biology is more ready now than ever to offer promise in some of the biological areas that have so far proven difficult (Sects. 4.1 and 4.2), and to open up new exciting avenues for its applications (Sect. 4.3).

4.1 Studying Protein–Protein Interactions

A myriad of cellular processes are mediated by protein–protein interactions (e.g. in signaling, metabolism, cellular structure and transport), often requiring the formation of multiprotein macromolecular machineries. A mechanistic understanding of these biological processes therefore requires an examination of the protein complexes at the molecular level. Experimental methods such as X-ray crystallography, NMR and EM are now being used to complement biochemical and biophysical methods such as yeast two-hybrid, immuno-precipitation and fluorescence resonance energy transfer, to understand these interactions better. However, complex structure determination remains challenging as compared to its single protein counterpart, and often requires systematic mapping and delineation of the interacting region to obtain co-purified and co-crystallized complexes [126, 127]. This substantial investment in time and effort is reflected by the number of protein complex structures in the PDB being only one-sixth of single protein structures. In the absence of co-crystal structures, *in silico* methods serve as a promising alternative to generate complex structure models by protein–protein docking and homology modeling, and will continue to attract considerable attention and research due to their comparative ease of use [128]. The identification of druggable protein–protein interactions that participate in diseases also represents an exciting avenue in drug discovery. Targeting a protein–protein interface for small molecule modulation is often considered less tractable than conventional single protein targets, due to the large interacting surface and less pocket-like features. Nevertheless, over the years a number of protein–protein interaction inhibitors have been developed, assisted by available structural information of both protein–protein complexes and of individual proteins (e.g. interaction partners of interleukin IL-2, B-cell lymphoma 2 Bcl-X_L and human papilloma virus transcription factor E2; [129] and references therein), and some are now entering clinical trials.

4.2 The High Hanging-Fruits of Membrane Protein Structures

In addition to multiprotein complexes, many classes of disease-associated and therapeutically important proteins remain refractory to the current methods of structure determination. Particularly in mind are the integral membrane proteins, such as the family of G-protein coupled receptors (GPCR) that are predicted targets

for ~30–50% of marketed drugs [68], and hence a major focus in pharmaceutical research. However, due to intrinsic difficulties with membrane protein crystallization, understanding GPCR structure and function has largely been achieved by homology modeling approaches. Recent structural breakthroughs, e.g. in the use of heterologous expression systems and in engineering mutations to stabilize proteins for crystallization [130], have brought the current number of available GPCR structures to six, an important increase, yet still in stark contrast to the total number of GPCRs predicted in the human genome (>900). Nevertheless, the structure determination over the past few years of a few highly-relevant GPCR drug targets (e.g. β1- and β2-adrenergic receptors [85, 131], A2A adenosine receptor [132], chemokine receptor CXCR4 [133] and dopamine D3 receptor [134]) has provided hope for structure-based methods to be applied routinely in GPCR drug discovery. These new structures offer promising opportunities for *in silico* compound screening and docking, and provide a diversity of available templates for homology modeling which, until the day that routine membrane protein crystallization has arrived, will continue to play a key role in leveraging structural coverage for this protein family.

4.3 Combining Mutation Analysis and Drug Design: Pharmacological Chaperones

An excellent example of combining the structural applications in mutation analysis and small molecule design is found in the emerging field of pharmacological chaperone therapy (PCT), a paradigm approach to treat inherited diseases that affect enzyme stability and function, such as phenylketonuria (cf. Sect. 2.3.1) and lysosomal storage disorders [135]. PCT involves the use of small molecules, often active site inhibitors or substrate mimics of the native protein, to stabilize mutant enzymes suffering from folding and trafficking defects. A great deal of ground work and proof-of-principle studies has incorporated structural information in order to establish the molecular basis of disease mutations and to identify those chaperone-responsive mutations with potential for PCT. To this end, a small-molecule screening effort to identify stabilizing therapeutic agents to treat PKU has already yielded two promising compounds for PAH stabilization [136]. Recently, crystal structures for a number of lysosomal hydrolases (e.g. β-hexosaminidase B [137] and acid β-glucosidase [138]) have been determined in complexes with pharmacological chaperones identified from chemical screening, to provide atomic insights into their modes of stabilization. The structure determination itself of these lysosomal enzymes is no small feat due to their heavily-glycosylated nature. The stage is now set for a systematic, structure-assisted approach in developing the next generation of chaperone compounds into clinical applications. The current work additionally reveals the potential of PCT as a general strategy to treat a wide range of rare genetic diseases, many of which are being unraveled by the year, and illustrates

how structural biology has suitably positioned itself within the translational approach from bench to clinic.

Acknowledgements We thank Laura Spagnolo (University of Edinburgh) for the picture contribution in Fig. 3. The Structural Genomics Consortium is a registered charity (number 1097737) that receives funds from the Canadian Institutes for Health Research, the Canada Foundation for Innovation, Genome Canada, GlaxoSmithKline, Lilly Canada, the Novartis Research Foundation, Pfizer, Abbott, Takeda, the Ontario Ministry of Research and Innovation and the Wellcome Trust.

References

1. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181(4610):662–666
2. Savitsky P, Bray J, Cooper CD, Marsden BD, Mahajan P, Burgess-Brown NA et al (2010) High-throughput production of human proteins for crystallization: the SGC experience. J Struct Biol 172(1):3–13
3. Page R, Stevens RC (2004) Crystallization data mining in structural genomics: using positive and negative results to optimize protein crystallization screens. Methods 34(3):373–389
4. Joachimiak A (2009) High-throughput crystallography for structural genomics. Curr Opin Struct Biol 19(5):573–584
5. Manjasetty BA, Turnbull AP, Panjikar S, Bussow K, Chance MR (2008) Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. Proteomics 8(4):612–625
6. Pellecchia M, Sem DS, Wuthrich K (2002) NMR in drug discovery. Nat Rev Drug Discov 1(3):211–219
7. Billeter M, Wagner G, Wuthrich K (2008) Solution NMR structure determination of proteins revisited. J Biomol NMR 42(3):155–158
8. Frank J (2002) Single-particle imaging of macromolecules by cryo-electron microscopy. Annu Rev Biophys Biomol Struct 31:303–319
9. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325
10. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826
11. Cavasotto CN, Phatak SS (2009) Homology modeling in drug discovery: current trends and applications. Drug Discov Today 14(13–14):676–683
12. Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353(2):459–473
13. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E et al (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum Mutat 23(5):464–470
14. Tramontano A, Morea V (2003) Assessment of homology-based predictions in CASP5. Proteins 53(Suppl 6):352–368
15. Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11(1):31–46
16. Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010) Massively parallel sequencing and rare disease. Hum Mol Genet 19(R2):R119–R124
17. Ku CS, Naidoo N, Pawitan Y (2011) Revisiting Mendelian disorders through exome sequencing. Hum Genet 129(4):351–370

18. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM et al (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42(1):30–35
19. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461(7261):272–276
20. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI et al (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42(9):790–793
21. International HapMap Consortium (2003) The International HapMap Project. Nature 426 (6968):789–796
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(Database issue):D514–D517
23. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS et al (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21(6):577–581
24. Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30(5):703–714
25. Jordan DM, Ramensky VE, Sunyaev SR (2010) Human allelic variation: perspective from protein function, structure, and evolution. Curr Opin Struct Biol 20(3):342–350
26. Karchin R (2009) Next generation tools for the annotation of human SNPs. Brief Bioinform 10(1):35–52
27. Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16(5):198–200
28. Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10(21):2319–2328
29. Hicks S, Wheeler DA, Plon SE, Kimmel M (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32(6):661–668
30. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30(8):1237–1244
31. Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356(5):1263–1274
32. Lobo PA, Van Petegem F (2009) Crystal structures of the N-terminal domains of cardiac and skeletal muscle ryanodine receptors: insights into disease mutations. Structure 17(11):1505–1514
33. Lew ED, Bae JH, Rohmann E, Wollnik B, Schlessinger J (2007) Structural basis for reduced FGFR2 activity in LADD syndrome: implications for FGFR autoinhibition and activation. Proc Natl Acad Sci USA 104(50):19802–19807
34. Chaikuad A, Froese DS, Berridge G, von Delft F, Oppermann U, Yue WW (2011) Conformational plasticity of glycogenin and its maltosaccharide substrate during glycogen biogenesis. Proc Natl Acad Sci USA 108(52):21028–21033
35. Malay AD, Procious SL, Tolan DR (2002) The temperature dependence of activity and structure for the most prevalent mutant aldolase B associated with hereditary fructose intolerance. Arch Biochem Biophys 408(2):295–304
36. Zhang Z, Norris J, Schwartz C, Alexov E (2011) In silico and in vitro investigations of the mutability of disease-causing missense mutation sites in spermine synthase. PLoS One 6(5): e20373
37. Anderson PC, Daggett V (2008) Molecular basis for the structural instability of human DJ-1 induced by the L166P mutation associated with Parkinson's disease. Biochemistry 47 (36):9380–9393
38. Niesen FH, Berglund H, Vedadi M (2007) The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat Protoc 2(9):2212–2221
39. Froese DS, Kochan G, Muniz JR, Wu X, Gileadi C, Ugochukwu E et al (2010) Structures of the human GTPase MMAA and vitamin B12-dependent methylmalonyl-CoA mutase and insight into their complex formation. J Biol Chem 285(49):38204–38213

40. Wu L, Pan L, Wei Z, Zhang M (2011) Structure of MyTH4-FERM domains in myosin VIIa tail bound to cargo. Science 331(6018):757–760
41. Bridwell-Rabb J, Winn AM, Barondeau DP (2011) Structure-function analysis of Friedreich's ataxia mutants reveals determinants of frataxin binding and activation of the Fe-S assembly complex. Biochemistry 50(33):7265–7274
42. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17(4):263–270
43. Gregersen N, Bross P, Vang S, Christensen JH (2006) Protein misfolding and human disease. Annu Rev Genomics Hum Genet 7:103–124
44. Mitchell JJ, Trakadis YJ, Scriver CR (2011) Phenylalanine hydroxylase deficiency. Genet Med 13(8):697–707
45. Dobson CM (2004) Principles of protein folding, misfolding and aggregation. Semin Cell Dev Biol 15(1):3–16
46. Jennings IG, Cotton RG, Kobe B (2000) Structural interpretation of mutations in phenylalanine hydroxylase protein aids in identifying genotype-phenotype correlations in phenylketonuria. Eur J Hum Genet 8(9):683–696
47. Erlandsen H, Stevens RC (1999) The structural basis of phenylketonuria. Mol Genet Metab 68(2):103–125
48. Dobrowolski SF, Pey AL, Koch R, Levy H, Ellingson CC, Naylor EW et al (2009) Biochemical characterization of mutant phenylalanine hydroxylase enzymes and correlation with clinical presentation in hyperphenylalaninaemic patients. J Inherit Metab Dis 32(1):10–21
49. Pey AL, Stricher F, Serrano L, Martinez A (2007) Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. Am J Hum Genet 81(5):1006–1024
50. Gersting SW, Kemter KF, Staudigl M, Messing DD, Danecka MK, Lagler FB et al (2008) Loss of function in phenylketonuria is caused by impaired molecular motions and conformational instability. Am J Hum Genet 83(1):5–17
51. Matthews BW (1993) Structural and genetic analysis of protein stability. Annu Rev Biochem 62:139–160
52. Xiao J, Madhan B, Li Y, Brodsky B, Baum J (2011) Osteogenesis imperfecta model peptides: incorporation of residues replacing Gly within a triple helix achieved by renucleation and local flexibility. Biophys J 101(2):449–458
53. Tang NL, Hui J, Young E, Worthington V, To KF, Cheung KL et al (2003) A novel mutation (G233D) in the glycogen phosphorylase gene in a patient with hepatic glycogen storage disease and residual enzyme activity. Mol Genet Metab 79(2):142–145
54. Wu WW, Molday RS (2003) Defective discoidin domain structure, subunit assembly, and endoplasmic reticulum processing of retinoschisin are primary mechanisms responsible for X-linked retinoschisis. J Biol Chem 278(30):28139–28146
55. Malay AD, Allen KN, Tolan DR (2005) Structure of the thermolabile mutant aldolase B, A149P: molecular basis of hereditary fructose intolerance. J Mol Biol 347(1):135–144
56. Jeyabalan J, Nesbit MA, Galvanovskis J, Callaghan R, Rorsman P, Thakker RV (2010) SEDLIN forms homodimers: characterisation of SEDLIN mutations and their interactions with transcription factors MBP1, PITX1 and SF1. PLoS One 5(5):e10646
57. Joerger AC, Fersht AR (2007) Structural biology of the tumor suppressor p53 and cancer-associated mutants. Adv Cancer Res 97:1–23
58. Cho Y, Gorina S, Jeffrey PD, Pavletich NP (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science 265(5170):346–355
59. Amador FJ, Liu S, Ishiyama N, Plevin MJ, Wilson A, MacLennan DH et al (2009) Crystal structure of type I ryanodine receptor amino-terminal beta-trefoil domain reveals a disease-associated mutation "hot spot" loop. Proc Natl Acad Sci USA 106(27):11040–11044
60. Li S, Duan J, Li D, Yang B, Dong M, Ye K (2011) Reconstitution and structural analysis of the yeast box H/ACA RNA-guided pseudouridine synthase. Genes Dev 25(22):2409–2421
61. Picaud S, Kavanagh KL, Yue WW, Lee WH, Muller-Knapp S, Gileadi O et al (2011) Structural basis of fumarate hydratase deficiency. J Inherit Metab Dis 34(3):671–676

62. Lee WH, Yue WW, Raush E, Totrov M, Abagyan R, Oppermann U et al (2011) Interactive JIMD articles using the iSee concept: turning a new page on structural biology data. J Inherit Metab Dis 34(3):565–567
63. Lahiry P, Torkamani A, Schork NJ, Hegele RA (2010) Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat Rev Genet 11(1):60–74
64. Gong S, Blundell TL (2010) Structural and functional restraints on the occurrence of single amino acid variations in human proteins. PLoS One 5(2):e9186
65. Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, Martin AC (2009) The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 30(4):616–624
66. Khan S, Vihinen M (2007) Spectrum of disease-causing mutations in protein secondary structures. BMC Struct Biol 7:56
67. Hopkins AL, Groom CR (2002) The druggable genome. Nat Rev Drug Discov 1(9):727–730
68. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5(12):993–996
69. Imming P, Sinning C, Meyer A (2006) Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov 5(10):821–834
70. Rask-Andersen M, Almen MS, Schioth HB (2011) Trends in the exploitation of novel drug targets. Nat Rev Drug Discov 10(8):579–590
71. Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D (2011) The cost of drug development: a systematic review. Health Policy 100(1):4–17
72. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR et al (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9(3):203–214. doi:10.1038/nrd3078
73. Hassall CH, Krohn A, Moody CJ, Thomas WA (1982) The design of a new group of angiotensin-converting enzyme inhibitors. FEBS Lett 147(2):175–179
74. Lapatto R, Blundell T, Hemmings A, Overington J, Wilderspin A, Wood S et al (1989) X-Ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. Nature 342(6247):299–302
75. Miller M, Schneider J, Sathyanarayana BK, Toth MV, Marshall GR, Clawson L et al (1989) Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. Science 246(4934):1149–1152
76. Supuran CT, Scozzafava A, Casini A (2003) Carbonic anhydrase inhibitors. Med Res Rev 23 (2):146–189
77. von Itzstein M, Wu W-Y, Kok GB, Pegg MS, Dyason JC, Jin B et al (1993) Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature 363(6428):418–423. doi:10.1038/363418a0
78. Lew W, Chen X, Kim CU (2000) Discovery and development of GS 4104 (oseltamivir) an orally active influenza neuraminidase inhibitor. Curr Med Chem 7(6):663–672
79. Tokarski JS, Newitt JA, Chang CY, Cheng JD, Wittekind M, Kiefer SE et al (2006) The structure of Dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. Cancer Res 66(11):5790–5797
80. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science 289(5486):1938–1942
81. Chen L, Jiao ZH, Zheng LS, Zhang YY, Xie ST, Wang ZX et al (2009) Structural insight into the autoinhibition mechanism of AMP-activated protein kinase. Nature 459(7250):1146–1149
82. Yuan P, Bartlam M, Lou Z, Chen S, Zhou J, He X et al (2009) Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. Nature 458(7240):909–913
83. Williams PA, Cosme J, Vinkovic DM, Ward A, Angove HC, Day PJ et al (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. Science 305(5684):683–686
84. Zhang H, Tweel B, Li J, Tong L (2004) Crystal structure of the carboxyltransferase domain of acetyl-coenzyme A carboxylase in complex with CP-640186. Structure 12(9):1683–1691
85. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R et al (2008) Structure of a beta1-adrenergic G-protein-coupled receptor. Nature 454 (7203):486–491

86. Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S et al (2003) Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol 334(3):387–401
87. Grabowski M, Chruszcz M, Zimmerman MD, Kirillova O, Minor W (2009) Benefits of structural genomics for drug discovery research. Infect Disord Drug Targets 9(5):459–474
88. Shin DH, Hou J, Chandonia JM, Das D, Choi IG, Kim R et al (2007) Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center. J Struct Funct Genomics 8(2–3):99–105
89. Weigelt J (2010) Structural genomics-impact on biomedicine and drug discovery. Exp Cell Res 316(8):1332–1338
90. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D et al (2009) PSI-2: structural genomics to cover protein domain family space. Structure 17(6):869–881
91. Chim N, Habel JE, Johnston JM, Krieger I, Miallau L, Sankaranarayanan R et al (2011) The TB structural genomics consortium: a decade of progress. Tuberculosis (Edinb) 91 (2):155–172
92. Ehebauer MT, Wilmanns M (2011) The progress made in determining the *Mycobacterium tuberculosis* structural proteome. Proteomics 11(15):3128–3133
93. Yue WW, Oppermann U (2011) High-throughput structural biology of metabolic enzymes and its impact on human diseases. J Inherit Metab Dis 34(3):575–581
94. Edwards A (2009) Large-scale structural biology of the human proteome. Annu Rev Biochem 78:541–568
95. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O et al (2010) Selective inhibition of BET bromodomains. Nature 468(7327):1067–1073
96. Perot S, Sperandio O, Miteva MA, Camproux AC, Villoutreix BO (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. Drug Discov Today 15(15–16):656–667
97. Hammes-Schiffer S, Benkovic SJ (2006) Relating protein motion to catalysis. Annu Rev Biochem 75:519–541
98. Keller TH, Pichota A, Yin Z (2006) A practical view of 'druggability'. Curr Opin Chem Biol 10(4):357–361
99. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1–3):3–26
100. Hajduk PJ, Huth JR, Fesik SW (2005) Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 48(7):2518–2525
101. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR et al (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25(1):71–75
102. Ciulli A, Williams G, Smith AG, Blundell TL, Abell C (2006) Probing hot spots at protein-ligand binding sites: a fragment-based approach using biophysical methods. J Med Chem 49(16):4992–5000
103. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450(7172):1001–1009. doi:10.1038/nature06526
104. Bleicher KH, Bohm HJ, Muller K, Alanine AI (2003) Hit and lead generation: beyond high-throughput screening. Nat Rev Drug Discov 2(5):369–378
105. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T et al (2011) Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov 10(3):188–195. doi:10.1038/nrd3368
106. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 11(13–14):580–594
107. Kalyaanamoorthy S, Chen YP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today 16(17–18):831–839

108. Cavasotto CN, Ortiz MA, Abagyan RA, Piedrafita FJ (2006) In silico identification of novel EGFR inhibitors with antiproliferative activity against cancer cells. Bioorg Med Chem Lett 16(7):1969–1974
109. Dooley AJ, Shindo N, Taggart B, Park JG, Pang YP (2006) From genome to drug lead: identification of a small-molecule inhibitor of the SARS virus. Bioorg Med Chem Lett 16 (4):830–833
110. McLean LR, Zhang Y, Degnen W, Peppard J, Cabel D, Zou C et al (2010) Discovery of novel inhibitors for DHODH via virtual screening and X-ray crystallographic structures. Bioorg Med Chem Lett 20(6):1981–1984
111. Ferreira RS, Simeonov A, Jadhav A, Eidam O, Mott BT, Keiser MJ et al (2010) Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors. J Med Chem 53(13):4891–4905
112. Schneider G, Hartenfeller M, Reutlinger M, Tanrikulu Y, Proschak E, Schneider P (2009) Voyages to the (un)known: adaptive design of bioactive compounds. Trends Biotechnol 27(1):18–26
113. Hartenfeller M, Schneider G (2011) De novo drug design. Methods Mol Biol 672:299–323
114. Fink T, Bruggesser H, Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 Daltons. Angew Chem Int Ed Engl 44(10):1504–1508
115. Heikkila T, Thirumalairajan S, Davies M, Parsons MR, McConkey AG, Fishwick CW et al (2006) The first de novo designed inhibitors of *Plasmodium falciparum* dihydroorotate dehydrogenase. Bioorg Med Chem Lett 16(1):88–92
116. Ni S, Yuan Y, Huang J, Mao X, Lv M, Zhu J et al (2009) Discovering potent small molecule inhibitors of cyclophilin A using de novo drug design approach. J Med Chem 52(17):5295–5298
117. Blundell TL, Jhoti H, Abell C (2002) High-throughput crystallography for lead discovery in drug design. Nat Rev Drug Discov 1(1):45–54
118. Hann MM, Leach AR, Harper G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Comput Sci 41(3):856–864
119. Hartshorn MJ, Murray CW, Cleasby A, Frederickson M, Tickle IJ, Jhoti H (2005) Fragment-based lead discovery using X-ray crystallography. J Med Chem 48(2):403–413
120. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science 274(5292):1531–1534
121. Wada CK, Holms JH, Curtin ML, Dai Y, Florjancic AS, Garland RB et al (2002) Phenoxyphenyl sulfone N-formylhydroxylamines (retrohydroxamates) as potent, selective, orally bioavailable matrix metalloproteinase inhibitors. J Med Chem 45(1):219–232
122. Howard S, Berdini V, Boulstridge JA, Carr MG, Cross DM, Curry J et al (2009) Fragment-based discovery of the pyrazol-4-yl urea (AT9283), a multitargeted kinase inhibitor with potent aurora kinase activity. J Med Chem 52(2):379–388
123. Wyatt PG, Woodhead AJ, Berdini V, Boulstridge JA, Carr MG, Cross DM et al (2008) Identification of N-(4-piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a novel cyclin dependent kinase inhibitor using fragment-based X-ray crystallography and structure based drug design. J Med Chem 51(16):4986–4999
124. Artis DR, Lin JJ, Zhang C, Wang W, Mehra U, Perreault M et al (2009) Scaffold-based discovery of indeglitazar, a PPAR pan-active anti-diabetic agent. Proc Natl Acad Sci USA 106(1):262–267
125. de Kloe GE, Bailey D, Leurs R, de Esch IJ (2009) Transforming fragments into candidates: small becomes big in medicinal chemistry. Drug Discov Today 14(13–14):630–646
126. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 103(21):8060–8065
127. Brooun A, Foster SA, Chrencik JE, Chien EY, Kolatkar AR, Streiff M et al (2007) Remedial strategies in structural proteomics: expression, purification, and crystallization of the Vav1/Rac1 complex. Protein Expr Purif 53(1):51–62

128. Mukherjee S, Zhang Y (2011) Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19(7):955–966
129. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450(7172):1001–1009
130. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS et al (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. Science 318(5854):1266–1273
131. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS et al (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318(5854):1258–1265
132. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR et al (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science 322(5905):1211–1217
133. Wu B, Chien EY, Mol CD, Fenalti G, Liu W, Katritch V et al (2010) Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. Science 330 (6007):1066–1071
134. Chien EY, Liu W, Zhao Q, Katritch V, Han GW, Hanson MA et al (2010) Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. Science 330 (6007):1091–1095
135. Chaudhuri TK, Paul S (2006) Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J 273(7):1331–1349
136. Pey AL, Ying M, Cremades N, Velazquez-Campoy A, Scherer T, Thony B et al (2008) Identification of pharmacological chaperones as potential therapeutic agents to treat phenylketonuria. J Clin Invest 118(8):2858–2867
137. Bateman KS, Cherney MM, Mahuran DJ, Tropak M, James MN (2011) Crystal structure of beta-hexosaminidase B in complex with pyrimethamine, a potential pharmacological chaperone. J Med Chem 54(5):1421–1429
138. Lieberman RL, Wustman BA, Huertas P, Powe AC Jr, Pine CW, Khanna R et al (2007) Structure of acid beta-glucosidase with pharmacological chaperone provides insight into Gaucher disease. Nat Chem Biol 3(2):101–107