Bioinformatics and variability in drug response: a protein structural perspective

Jennifer L. Lahti1, Grace W. Tang1, Emidio Capriotti1,2, Tianyun Liu3 and Russ B. Altman1,3,*

1Department of Bioengineering, Stanford University, Stanford, CA, USA
2Department of Mathematics and Computer Science, University of Balearic Islands, Palma de Mallorca, Spain
3Department of Genetics, Stanford University, Stanford, CA, USA

Marketed drugs frequently perform worse in clinical practice than in the clinical trials on which their approval is based. Many therapeutic compounds are ineffective for a large subpopulation of patients to whom they are prescribed; worse, a significant fraction of patients experience adverse effects more severe than anticipated. The unacceptable risk–benefit profile for many drugs mandates a paradigm shift towards personalized medicine. However, prior to adoption of patient-specific approaches, it is useful to understand the molecular details underlying variable drug response among diverse patient populations. Over the past decade, progress in structural genomics led to an explosion of available three-dimensional structures of drug target proteins while efforts in pharmacogenetics offered insights into polymorphisms correlated with differential therapeutic outcomes. Together these advances provide the opportunity to examine how altered protein structures arising from genetic differences affect protein–drug interactions and, ultimately, drug response. In this review, we first summarize structural characteristics of protein targets and common mechanisms of drug interactions. Next, we describe the impact of coding mutations on protein structures and drug response. Finally, we highlight tools for analysing protein structures and protein–drug interactions and discuss their application for understanding altered drug responses associated with protein structural variants.

Keywords: bioinformatics; protein structure; drug response; pharmacogenetics

1. INTRODUCTION

Population-level statistics irrefutably demonstrate the benefits of pharmaceutical innovation over the past century, which has seen the introduction of antibiotics, statins and cancer therapeutics. Rapid advances in the fields of genomics, proteomics and biotechnology have fuelled the drug discovery process. Yearly from 1982 to 2010, an average of 18 drugs were approved for human use by the US Food and Drug Administration (FDA), with approximately four acting on novel target structures [1].

Yet, in spite of this historical success, the pharmaceutical industry continues to face exceptional challenges. Over the past decade, escalating investments in basic and clinical research have not seen equal returns. Instead, both the developmental rate of new molecular entities and the approval rate of new drugs have dropped by roughly 50 per cent [2,3]. During clinical development, efficacy and safety concerns contribute equally to the attrition of candidate drugs [4]. Even marketed drugs display limited efficacy, with studies showing them to be effective for only 30–60% of the patients to whom they are prescribed [5,6]. Furthermore, for many drugs whose therapeutic windows are narrow and the consequences of adverse events are life-threatening, up to one-third of patients develop unacceptable toxicity [7]. Consequently, a significant number of marketed drugs have poor risk–benefit ratios for diverse patient populations. This occurrence has been termed the ‘efficacy–effectiveness gap’ and is, ultimately, a result of variability in patient–drug responses [8].

The observation that patients are neither equally responsive to beneficial drug effects nor equally susceptible to adverse events motivates the call for a paradigm shift from population-level to patient-specific medicine [6,8]. To address this directive, two cooperative aims have been proposed, (i) determine the detailed molecular mechanisms of drug action and (ii) understand the
effect of genetic variants on patient–drug response. The former aim is the focus of the field of pharmacodynamics (reviewed in [9]), and the latter the focus of pharmacogenetics (reviewed in [10,11]). At the intersection lies the challenge of understanding how genetic differences between individuals can translate into structural alterations in protein drug targets and ultimately into variable patient–drug response.

The genetic basis for inter-individual drug response variability has been studied extensively over the last 50 years [12]. While there are numerous behavioural and environmental factors that contribute to patient–drug response, genetic factors also often have a key, if not a dominant, role [7]. Specifically, genetic variants affect gene expression, mRNA processing and stability, and protein structure. Each of these variations can have functionally significant consequences for drug response [12]. Moreover, genetic polymorphisms are observed in all of the principle effectors of therapeutic response, drug transporters, drug-metabolizing enzymes and drug targets. Gaining a detailed understanding of the underlying mechanisms of phenotypic variability in drug response at the protein level is a key factor in the establishment of personalized medicine [13].

Traditionally, the roles of genetic variations in proteins were investigated using sequence analysis tools to predict the tolerability of a given amino acid substitution and its probable effect on protein function [14]. Yet, interpreting the effect of a mutation within the three-dimensional context of the protein structure offers more information [15]. Analysis of three-dimensional structures can provide valuable insight into the mechanisms of drug–target interaction and the relationships between mutations and differential therapeutic responses [16]. Such detailed structural analysis of protein–drug interactions was not always feasible in the past, but structural genomics initiatives have resulted in an explosion of high-resolution structures of known and potential drug target proteins.

In this review, we discuss the relationship between structural protein variations and differences in patient–drug response. The scope is not limited to, but is strongly focused on, the effect of mutations on structures of primary drug targets of human origin. We begin with an overview of protein targets and common mechanisms of drug interactions. Next, we describe the impact of structural variations on drug response. Finally, tools and databases developed for analysing protein structures and protein–drug interactions are presented and their potential applications for gaining insight into protein structural variants displaying altered drug response are discussed.

2. SMALL-MOLECULE DRUGS AND THEIR PROTEIN TARGETS

2.1. Properties of small-molecule drugs

Structures of therapeutic agents are highly diverse and range from small-molecule compounds, to antibodies, to whole cells [17]. This structural diversity allows them to specifically interact with and modulate the function of their diverse targets. Although therapeutic biologics have seen increased development during the past decade, small-molecule drugs still account for over two-thirds of new molecular entities approved by the US FDA [18]. Thus, this review focuses on small molecular therapeutics (typically 200–550 Da [19]) and the structural mutations in their protein targets that can lead to altered drug response. Yet, it is important to note that such mutations can affect the behaviour of all therapeutic agent classes.

Small-molecule drugs must meet several criteria, including having reasonable solubility and stability levels in aqueous media, appropriate structural and physicochemical features to specifically interact with their targets, and satisfactory pharmacokinetic profiles for clinical use (for recent reviews see [20–22]). Guidelines for evaluating the drug-likeness of a molecule (for example, Lipinski’s Rule of Five [23] and its derivatives [24]) have been widely adopted to aid the development process. However, applying these guidelines warrants caution, as the guidelines assume that the target of interest requires a compound whose molecular properties are similar to those of the average drug. In addition, applying drug-like screening criteria to compounds in early stages of development can be disadvantageous because the molecular properties of lead compounds undergo extensive optimization before clinical introduction. On average, the optimization process increases a compound’s molecular weight and complexity [25]; thus, an initial compound with drug-like properties would probably lie outside of the desired physicochemical space after development is complete.

Frequently, the universal application of drug-likeness guidelines without regard for the structure of the intended protein target is detrimental. Consideration of the three-dimensional target structure can provide a more accurate understanding of a drug’s requisite properties [26,27]. In short, fine-tuning the molecular properties of drugs to the structures of their protein targets promotes binding interactions of high affinity and specificity.

2.2. Three-dimensional structures of protein targets

Ideal therapeutic targets share several features, involvement in a biologically relevant pathway, functional and structural characterization, and druggability [28]. A druggable protein is one possessing structural characteristics that favour interactions with drug-like compounds and whose function can be modulated through such interactions. Inference of druggability historically relied on sequence homology of the protein of interest to known drug targets [29]. However, protein families lacking homology to drug targets have yielded novel targets and not all members of a protein family are equally druggable [30].

Instead, three-dimensional structures can provide information more relevant to protein druggability. In a seminal paper, Cheng et al. [30] applied knowledge derived from biophysical principles and protein structure to accurately predict protein druggability and drug binding affinities. Several subsequent studies then extracted structural and physicochemical descriptors associated with druggability from known protein–ligand complex structures [31–33]. A recent comparison of two protein structure datasets, one comprising drug targets and the
MECHANISMS OF DRUG ACTIVITY

3.1. Molecular recognition between small molecules and proteins

Binding events, like the formation of protein–drug complexes, are governed by enthalpic and entropic contributions. The former stem from stabilizing interactions with the formation of hydrogen bonds and salt bridges, and the latter involve penalties from the loss of conformational freedom of the protein and drug. The balance between enthalpic and entropic contributions determine the free energy of the interaction and, thereby, the favorability of the binding event at equilibrium. The enthalpy of molecular recognition between a protein and a small molecule depends on two key components, shape complementarity and physicochemical complementarity. Shape complementarity permits the protein and small molecule to achieve sufficient proximity and contact surface area to form stabilizing interactions, while physicochemical complementarity determines the nature of these interactions. Amino acid mutations occurring in target proteins have the ability to disrupt both shape complementarity as well as physicochemical compatibility. Their impact on drug binding, and therefore drug response, depends on the nature of the mutation and its three-dimensional structural context, as discussed in §4.
3.1.1. Models of molecular recognition

Molecular complementarity was first thought of as a lock-and-key fit, where the small molecule (key) possessed perfect shape and physicochemical complementarity to the protein (lock). However, the predominance of structural rearrangements, both minor and major [51], upon ligand binding are better explained by two recently adopted models—induced fit and conformational selection [52,53].

The induced fit model attributes protein structural changes to the binding event of the ligand. The recently solved complex of human neutrophil elastase (HNE) with a dihydropyrimidone inhibitor exemplifies an induced fit binding mechanism, since protein structural rearrangements near the inhibitor differ from the conformations of both ligand-free HNE and other HNE-inhibitor complexes [54].

Conversely, in the conformational selection theory the ligand selects the most complementary conformation from an ensemble of equilibrium structures. This mode of interaction has been implicated in the binding selectivity of imatinib to tyrosine kinases (see §4.3.1.1) [55]. Molecular dynamics (MD) studies reveal that kinases with high imatinib affinity spend more time in conformations compatible with drug binding compared with kinases with low imatinib affinity.

Yet other studies suggest that real interactions reflect a mixture of the two binding models [55–57]. For example, recent in silico work demonstrated that, contrary to previous assumptions, ligand binding to the lysine-, arginine-, ornithine-binding protein proceeds through two stages, initial complex formation by conformational selection followed by ligand-induced transition to the observed bound state [57].

Differences between unbound (apo state) and bound (holo state) protein structures add a transient dimension to molecular complementarity. This has important implications for the use of structural data in detecting and understanding protein–drug interactions. Namely, available protein structures may not display a conformation compatible with small-molecule binding. These non-binding structures can result from protein structural preferences or random chance and crystallization artefacts [58]. Caution must therefore be used when relying on structural data to assess whether a protein binds a drug and the mode of binding. MD simulation (see §5.5) is a useful approach for generating protein structural ensembles and capturing alternative conformations that may be relevant for protein function or small-molecule binding [59,60]. Protein dynamics information enhances structure-based binding site prediction methods and reduces their dependence on experimentally determined target protein structures.

3.1.2. Shape complementarity

Shape complementarity refers to the geometric fit, or steric fitness, of a small molecule and its surrounding protein environment. The importance of shape complementarity in protein–drug interactions is discussed in a recent review by Kortagere et al. [61]. Numerous protein–ligand complexes in the PDB depicting close intermolecular contacts between small molecules and proteins illustrate the importance of shape complementarity for binding. For example, the recently solved complex of the first bromodomain of human Brd4, a transcription factor complex protein and therapeutic target of interest in cancer, with a highly potent and specific inhibitor depicts excellent shape complementarity (figure 1a and b) [62].

However, high shape complementarity is not always a requisite for drug binding and is often incomplete or imperfect. This is evident from the crystal structure of the human Cdc34 ubiquitin-conjugating enzyme in complex with a novel allosteric inhibitor showing that the ligand does not fully interact with the protein target [63]. The shape complementarity of this complex is both imperfect, since water-mediated binding interactions introduce unoccupied space between binding partners (figure 1c), and incomplete, as there is a solvent-exposed carboxylic acid group on the ligand (figure 1d).

Shape complementarity dictates whether a ligand is sufficiently close to a protein to form favourable interactions and is therefore a critical determinant of binding. As such, shape complementarity has been applied in virtual screening of drug discovery approaches [64]. Yet, it is important to note that geometric compatibility does not always indicate physicochemical compatibility. Thus, in the following section we describe the role of physicochemical complementarity in molecular recognition.

Numerous shape-matching technologies and protein pocket predictors can be used to gain insights into protein–drug interactions (see §5.3). Ebalunode and Zheng published a recent review on shape complementary methods [65].

3.1.3. Physicochemical complementarity

Physicochemical complementarity refers to non-covalent interactions holding proteins and ligands in a complex. These interactions can involve long-range ionic bonds or weaker short-range interactions including hydrogen bonds, van der Waals forces and hydrophobic packing. Electrostatic complementarity, accounting for both ionic and electrostatic interactions, is one of the most important forces governing protein–ligand complex formation, affecting binding affinity as well as the rate of protein–ligand association (reviewed in [66]). For example, in pancreatic endoplasmic reticulum kinase, electrostatic complementarity to an aspartate in the binding site strongly influences inhibitor affinity, with a lack of complementarity translating to weaker affinity [67]. Likewise, electrostatic steering of a small molecule into the proper orientation for binding enhances complex formation. Protein–drug interactions are further affected by pH owing to the titratable groups (weak acid or weak base) found in many drugs that alter the ionization state of the molecule.

As previously discussed in §2.1, many drugs are hydrophobic in nature and have a partition coefficient greater than one, indicating a preference for solubilizing into octanol versus water [23]. Such compounds are energetically unfavourable in the aqueous compartments of the body, thereby producing a driving force
for protein binding. This tendency is responsible for much of the non-specific interactions of hydrophobic drugs. Non-specific interactions are therefore primarily driven by physicochemical compatibility between the compound and protein.

While the effects of both the hydrophobic effect and hydrogen bonds on protein–ligand binding events are well documented, the molecular underpinnings of these phenomena remain the focus of investigation. Of particular importance to understanding protein–ligand interactions, the mechanism of hydrophobicity [68] and the definition of hydrogen bonds [69] have been recently updated.

3.2. Modes of drug binding

Drugs and other chemical compounds interact with proteins in diverse ways. The nature of the interaction may be reversible (e.g. the binding of a competitive antagonist) or irreversible (e.g. the covalent modification of a protein from a suicide inhibitor). The location at which drugs interact with proteins can also vary. Drug binding at the protein’s orthosteric site or allosteric site has different implications (figure 2), which we discuss further in the following sections.

3.2.1. Orthosteric binding sites

Classical drug development approaches predominantly focused on targeting the protein orthosteric site (also known as the active site for enzymes). Endogenous ligands bind at the orthosteric site to elicit a biological response. Thus, a popular mechanism of drug action is to occupy the orthosteric site, thereby blocking endogenous ligand binding and modulating protein function (figure 2).

As we noted previously, protein kinases constitute a large protein family of strong pharmaceutical interest. There are two classes of kinase inhibitors. Type 1 inhibitors exert their effects by blocking adenosine triphosphate (ATP) binding to the catalytic kinase domain. Type 2 inhibitors also bind in the active site, but block kinase activity by stabilizing the inactive protein conformation (see §4.3.2). Imatinib, a popular drug used for treating chronic myelogenous leukaemia, is a type 2 inhibitor that binds to the deregulated tyrosine kinase, Bcr-Abl, stabilizing it in an inactive conformation [71].

Because the sequence and structure of active sites are often highly conserved among protein families, cross-reactivity is a significant problem for drugs targeting orthosteric sites. For example, many kinase inhibitors have low selectivity profiles and bind to a
Allosteric binding sites
Allosteric drugs interact with their protein targets at sites spatially distinct from the protein orthosteric site (figure 2). The binding event induces protein conformational rearrangements that lead to altered activity. Allosteric modulators produce a change in affinity or efficacy for the endogenous ligand, while allosteric agonists or antagonists alter the activation state of the protein itself [76].

Targeting protein allosteric sites has received significant attention recently because of the benefits of allosteric modulation versus orthosteric modulation [70,77]. First, in the event of drug overdose, allosteric drugs are likely to pose less heath risk because their effects saturate once full occupancy of targeted sites is reached. Next, allosteric compounds are less likely than orthosteric molecules to desensitize their targets and therefore have decreased tendency for acquired drug tolerance [78,79]. Most importantly, allosteric drugs have enabled highly selective targeting of protein family subtypes. In contrast to orthosteric sites, which are generally highly conserved, allosteric sites have much greater sequence and structural diversity. However, protein allosteric sites are often challenging to locate, characterize and target [80].

The shift towards allosteric therapeutics is evident in the rhodopsin-like GPCR protein family. Numerous drugs, including atenolol (an anti-hypertension drug) and salbutamol (an anti-asthma drug) bind to GPCR orthosteric sites in order to alter the receptor activity. However, like protein kinases, the orthosteric sites of GPCRs are highly conserved [81], leading to problems with off-target activity and thus motivating a move towards development of allosteric compounds to enable targeting of specific GPCR subtypes [70]. GPCR allosteric sites are more diverse, offering more degrees of chemical freedom in developing allosteric drugs compared with orthosteric drugs [82]. Cinacalcet is the first example of an FDA-approved allosteric GPCR modulator, and functions by increasing the sensitivity of its receptor to calcium in the treatment of hyperparathyroidism [70].

4. PROTEIN VARIANTS WITH ALTERED THREE-DIMENSIONAL STRUCTURES AND DRUG RESPONSES

Based on genome sequencing of individuals from different populations, it is estimated that each person’s proteome contains roughly 10,000–11,000 mutations compared to the reference proteome [85]. A subset of these mutations (those resulting in premature stop codons, splice-site disruptions, and frame shifts) probably has severe functional consequences, yielding approximately 250–300 loss-of-function protein variants per individual [85]. However, for the great majority of mutations, it is difficult to predict a priori what their effect will be on the resultant protein’s structure and function.

4.1. Classes of mutations
Single nucleotide polymorphisms (SNPs) fall either within non-coding (including promoter, operator, enhancer and transcription factor binding regions) or coding regions of DNA. Because of redundancy in the genetic code, some mutations within coding regions (synonymous mutations) do not change the encoded protein sequence. On the other hand, non-synonymous SNPs produce either polypeptide sequences that have an amino acid substitution (missense mutations) or are truncated (nonsense mutations). Phenotypes resulting from mutations are generally thought of in a protein structural context, where an amino acid substitution or deletion leads to altered protein structure and function. However, some mutations exert their effects via changes to the mRNA that can lead to altered mRNA splicing, folding or stability, and therefore altered protein product [86]. Here we focus our
discussion specifically on the structural effects of protein missense mutations, since over half of all known human disease-associated mutations are missense SNPs \cite{87}.

Missense mutations can drastically alter protein structure and function, resulting in inter-patient variability in drug response. There are varied structural mechanisms through which missense mutations exert their effects, such as altering the physicochemical or geometric properties of a protein binding pocket, modifying structure dynamics (i.e. conferring or restricting flexibility) or disrupting folding and stability. A missense mutation occurring in the target protein may have a pharmacodynamic effect, while one occurring in a protein involved in drug absorption, distribution, metabolism or excretion may alter drug pharmacokinetics. Protein variants of the former group often impact response to a specific drug class while those of the latter group often affect a diverse array of drug classes.

In the following sections, we describe the mechanisms by which missense mutations alter the three-dimensional protein structure and thereby change drug pharmacodynamics and/or pharmacokinetics. Elucidating the structural underpinning of these effects is important for understanding the mechanisms of drug response and for predicting the clinical implication of novel genetic variants.

4.2. Effects of missense mutations on protein structure

Early approaches for studying the effect of genetic variation on protein function were sequence-based. They relied on the hypothesis that functionally relevant residues would exhibit higher sequence conservation, as mutation would probably be deleterious \cite{88}. Accordingly, several sequence-based approaches harnessing evolutionary conservation information were developed to predict deleterious missense SNPs \cite{89–94}

Although amino acid conservation is a useful metric for identifying functionally important residues, its scope is inherently limited to one-dimensional sequence space. It is much more informative to consider the effect of a mutation within the three-dimensional context of the protein structure. Three-dimensional context of the mutation can reveal the nature of the local environment (solvent-exposed or buried), the proximal interacting residues that are not necessarily contiguous in primary sequence, and the relative position of the mutation to binding or active sites. Furthermore, while SNPs occurring in highly conserved functional sites may directly disrupt protein activity, it does not follow that SNPs in regions under little or no selective pressure are tolerable. Such mutations still have the potential to affect protein folding, dynamics, stability and activity. These effects manifest through changes to the protein structure; thus to accurately predict the effect of a missense mutation, it is necessary to study its structural context.

Recent algorithms for predicting the effects of mutations on protein function \cite{15,95–97} and stability \cite{98–101} are beginning to include protein structure as an input feature. However, high-resolution three-dimensional structures are available only for a subset of known proteins. When an experimentally determined protein structure is unavailable, structure prediction techniques (see \S 5.2) can provide a model from which to extract structural context information for a mutation.

4.3. Pharmacodynamic effects of missense mutations

Structural variants of target proteins with differential drug response compared with their wild-type counterparts separate into two broad categories, some have mutations affecting the binding site that directly alter drug interaction, while others have mutations distal to the binding region that give rise to long-range structural perturbations or altered protein conformations. Regardless of the mechanism responsible for variable drug outcome, detailed knowledge of the three-dimensional target protein structure is critical for understanding drug–target interactions. Furthermore, once the structures of disease-relevant variants are elucidated, they can be specifically targeted by novel therapeutic strategies for improved outcome. In the following case studies, we briefly discuss the consequences of missense mutations on three-dimensional target protein structures and drug response.

4.3.1. Protein variants with binding site mutations

4.3.1.1. Missense mutations altering drug binding: kinase ‘gatekeeper’ residue substitution. As previously discussed, kinases are important cellular signalling proteins whose aberrant expression and activation is widely implicated in cancer. Kinases are therefore among the most pursued classes of drug targets, with several ATP-competitive inhibitors approved for clinical use. However, the efficacy of these agents is often limited by the subsequent emergence of drug resistance \cite{102}. Such resistance often develops through the acquisition of mutations that abrogate inhibitor binding. The most widely observed of these mutations occur at the ‘gatekeeper’ residue, whose sidechain bulk controls accessibility of the hydrophobic ATP binding pocket \cite{102}. Interestingly, kinase gatekeeper mutations confer drug resistance through two distinct structural mechanisms, (i) sterically blocking binding of the drug to the active site and (ii) decreasing the apparent drug potency by increasing the binding site affinity for ATP.

Patients with the Bcr-Abl oncoprotein frequently acquire mutations in the Abl kinase domain after treatment with ATP-competitive inhibitors, resulting in drug resistance. In particular, the T315I Abl gatekeeper mutation accounts for approximately 20 per cent of clinically observed drug resistance to imatinib, the current gold-standard treatment for Bcr-Abl positive leukaemias \cite{103}. Examination of the crystal structure of the Abl T315I mutant compared with that of the wild-type kinase revealed that gatekeeper residue replacement with isoleucine sterically blocks binding of imatinib in the active site, resulting in drug resistance \cite{104} (figure 3). Structural knowledge of the Abl gatekeeper variant has brought about the development of inhibitors that target alternate Abl kinase druggable pockets or are capable of accommodating the T315I mutation \cite{106–108}.
Figure 3. Mutation of kinase gatekeeper residue confers drug resistance. (a) Overlay of wild-type Abl kinase (light grey cartoon) with bound imatinib (sticks) and T315I variant (dark grey cartoon) shows equivalent global structures. (b) Wild-type Abl binding pocket (yellow mesh) has a threonine gatekeeper residue (grey stick and semitransparent surface) bound to imatinib (yellow sticks). (c) Wild-type Abl binding pocket (yellow mesh) and imatinib (yellow sticks) overlaid on the T315I variant structure shows an inability of the mutant to accommodate the drug owing to protrusion of the isoleucine gatekeeper residue (red stick and semitransparent surface) that sterically prevents drug binding. (Wild-type Abl kinase, PDB 2HYY [105]; Abl kinase T315I, PDB 2Z60 [104].) (Online version in colour.)

Likewise, the epidermal growth factor receptor (EGFR) tyrosine kinase gatekeeper mutation, T790M, is also associated with clinical drug resistance. This mutation typically arises in patients possessing the oncogenic L858R mutation, accounting for approximately half of all clinically observed resistance to gefitinib and erlotinib [109,110]. Prior to attaining the crystal structure of the EGFR T790M/L858R variant, the gatekeeper mutation was proposed to sterically block binding of inhibitors in the active site [111]. More recently, Yun and colleagues solved the crystal structure of the variant and demonstrated that EGFR T790M/L858R is structurally capable of accommodating inhibitors in its kinase active site [112]. In fact, the observed drug resistance of EGFR T790M mutants is owing to increased binding affinity for ATP. Novel inhibitors specifically targeting the EGFR T790M variant are currently in clinical trials [113].

4.3.1.2. Missense mutations altering drug effect: androgen receptor binding pocket expansion. The androgen receptor (AR) is a nuclear hormone receptor essential to normal male development and the maintenance of male-specific organs. Altered AR signalling is implicated in multiple malignancies, including prostate cancer. While the incidence of mutations in the AR ligand-binding domain is low in primary tumours, it increases over the course of treatment with antiandrogens such as bicalutamide and flutamide [114]. Of particular concern are mutations in the AR ligand-binding domain that convert these therapeutic antagonists to partial agonists [115].

One such ligand-binding domain mutation observed in AR-dependent malignancies is T877A. Sack and colleagues first reported the crystal structure of the AR T877A variant and found that the alanine substitution increases the binding pocket volume, allowing bulkier ligands like the antiandrogens to bind to and activate the receptor [116]. Further structural investigations by Bohl et al. [117–119] showed that the AR T877A mutation (figure 4a and b) as well as a similar W741L mutation (figure 4c and d) expand the ligand-binding pocket and alter its physicochemical properties, leading to the conversion of potent antagonist drugs to agonists. Using the crystal structures of AR mutants with expanded binding pockets, drug discovery via structure-based approaches have led to second-generation antiandrogens that are currently undergoing clinical trials [115,122].

4.3.2. Protein variants with non-binding site mutations

4.3.2.1. Missense mutations altering protein conformation: KIT kinase shifted conformational equilibrium. Approximately 85 per cent of patients with gastrointestinal stromal tumours have activating mutations in KIT receptor tyrosine kinase [123]. Imatinib is an effective first-line treatment, but half of patients on this therapy acquire further KIT mutations conferring drug resistance within two years [124]. A portion of KIT variants with mutations at D816, located in the activation loop of the catalytic domain, are also resistant to second-line treatment with sunitinib [125].

Structural studies by Mol et al. revealed that KIT populates a structural ensemble ranging from an inactive autoinhibited state to an activated conformation [126,127]. Recently, Gajiwala and colleagues found that the D816H/V mutations both shift the conformational equilibrium of KIT variants towards the active form [128]. Yet, both imatinib and sunitinib bind exclusively to the inactive conformation which is less populated in KIT D816 variants, resulting in abrogated clinical efficacy. Similar mechanisms of resistance have been reported with other receptor tyrosine kinases, motivating drug development efforts focused specifically on targeting kinases in the active conformation [129,130].

4.3.2.2. Missense mutations affecting protein stability: p53 decreased thermal stability. Inactivation of the p53 tumour suppressor is an almost universal feature of human cancers [131]. Typically, p53 tumour
suppressor functions as a critical barrier to tumour development by binding to DNA and regulating cell cycle progression and apoptosis [132]. Hence, restoration of p53 activity has been the focus of intensive cancer therapeutic efforts [131]. p53 has an extremely limited half-life owing to both its low thermal stability ($T_m \approx 44^\circ C$) and targeted ubiquitination by HDM2, its negative regulator [132,133]. Blocking the p53 binding site on HDM2 is sufficient to reactivate the p53 response in cells and induce rapid tumour regression [131,132]. The crystal structure of p53 bound to the mouse homologue of HDM2 (MDM2) showed the presence of a deep druggable pocket at the protein–protein interface and inspired the structure-based development of numerous small-molecule HDM2 inhibitors [43,131,132,134]. Some of these HDM2 inhibitors (e.g. Nutlin-3) have shown promise in preclinical studies and are currently in early clinical trials [135–138].

While HDM2 inhibitors are promising therapeutics for patients with wild-type p53, approximately 50 per cent of human cancers have mutations in the p53 DNA-binding domain that make HDM2-targeted drug treatment ineffective [139,140]. For example, crystallographic studies on p53 Y220C, a common oncogenic variant, revealed that this surface mutation connects two pre-existing clefts to form an extended solvent accessible crevice (figure 5), disrupts packing of the hydrophobic core, and drastically decreases thermodynamic stability [139]. As a result, p53 Y220C is too unstable to function at physiological temperature and is rapidly depleted by denaturation [142]. The extended cleft has been the focus of structure-based drug discovery efforts aimed at rescuing unstable p53 Y220C mutants [133,143].

4.4. Pharmacokinetic effects of missense mutations

In general, all drugs are slightly promiscuous and the effect they elicit depends on their interaction with numerous proteins throughout the body, not just the target protein. Specifically, the proteins involved in drug pharmacokinetics have an important role in determining clinical outcome. While the focus of this
review is primarily on understanding the mechanism by which structural perturbations in the primary drug target can change drug response, the same analysis can be used to gain insight into mutations in proteins involved in drug absorption, distribution, metabolism or excretion.

Cytochrome P450 (CYP) enzymes are responsible for the oxidative metabolism of environmental compounds, pollutants and drugs [144]. Their essential role in drug metabolism makes CYP enzymes of great pharmacokinetic importance. A large number of CYP variants linked to altered drug pharmacokinetics have been reported for several CYP family members, including CYP1A2 [145], CYP2B6 [146], CYP2C9 [147], CYP2C19 [145], CYP2D6 [148], CYP2J2 [149] and CYP3A5 [150]. As a family, CYP enzymes bind a remarkably broad range of ligands. Recent structural studies show that CYP structural flexibility is the major determinant of ligand binding promiscuity [151,152]. Thus, missense mutations that are peripheral to the CYP active site can exert long-range effects that disrupt the enzyme’s flexibility or binding pocket structure, resulting in altered drug binding and metabolism [153,154].

The CYP2C9 isoform metabolizes more than 100 drugs in current clinical use [155]. There are large interindividual variations in CYP2C9 activity and, thus, in clinical response to therapeutics metabolized by the enzyme. Specifically, 32 marketed drugs exhibit CYP2C9 variant-dependent metabolism [155]. CYP2C9*3, a common variant carrying an I359L mutation, is correlated with decreased metabolism and clearance for multiple drugs including the anticoagulant warfarin [156]. Williams et al. first reported the crystal structure of apo-CYP2C9 and CYP2C9 in complex with warfarin [144]. Building on this crystallographic data, Sano et al. [157] computationally investigated the structural mechanisms underlying decreased CYP2C9*3 warfarin metabolism. Although the I359L mutation does not directly hinder drug binding, this substitution introduces long-range structural perturbations that result in expansion of the binding pocket volume and increased fluctuations in warfarin-coordinating residues (figure 6). These structural alterations collectively cause warfarin to bind in a region of the binding pocket that is more removed from the active site, leading to decreased enzymatic activity [157]. To account for this altered metabolism, a number of pharmacogenetic algorithms predict warfarin dosing based on patient CYP2C9 genotypes to achieve maximum efficacies with minimum toxicities [158–162].

5. COMPUTATIONAL TOOLS

Structure-based computational methods offer molecular insights into drug–protein relationships and the mechanisms by which missense mutations elicit differential drug responses. Below we discuss five major classes of bioinformatics tools, three-dimensional structure visualization, protein structure prediction, binding site detection and comparison, ligand docking and scoring, and MD. Although none of these tools were specifically developed for studying the structural effects of protein sequence mutations, they can be readily applied for this purpose [163]. For example, protein structure prediction tools have been used to model protein variants given a homologous protein structure [164,165]. Mutations mapped to the protein structure permit analysis of their structural and/or functional consequences using binding site prediction tools [166,167]. Ligand docking studies on protein structures with varied binding site residues have elucidated protein–ligand binding mechanisms [168] and selectivity [169]. In addition, simulation of protein MD and flexibility have provided insight into the effect of mutations on ligand binding [170,171]. Collectively, these bioinformatics tools offer great potential for understanding high-resolution structural details of protein variants that give rise to altered drug response.

It is important to note that biological information can be incorporated into numerous of the tools discussed below. Such prior information can focus research efforts and greatly enhance the accuracy and interpretation of the computational results. Evolutionary conservation information helps define and refine residue contact

Figure 5. Mutation of p53 surface residue decreases protein stability. (a) Overlay of wild-type p53 (light grey) and Y220C variant (dark grey) shows high structural similarity. (b) Mutation of Tyr220 (grey stick and semitransparent surface) to Cys (red stick and semitransparent surface) creates a cleft on the surface of mutant p53 (mesh) that destabilizes the protein structure. (Wild-type p53, PDB 1UOL [141]; p53 Y220C, PDB 2J1X [139].) (Online version in colour.)
information, aiding protein structure prediction [172] and both comparative [173] and free-modelling prediction techniques [174]. Such sequence information also contributes to binding pocket and binding residue identification [175–177]. Biophysical data, such as that derived from NMR experiments, can complement standard force fields and bias the sampling of protein conformational space towards regions consistent with experimental observations [178]. In this way, biophysical restraints have been used to assist protein structure predictions [179,180], MD simulations [181,182] and protein docking studies [183,184]. Furthermore, protein mutagenesis studies are routinely applied to both inform [185] and validate [186] docking experiments.

There are an overwhelming number of useful tools for each of the categories described. Highlighted below are examples of programs that are either widely adopted or recently developed. Numerous commercial software packages are also available for each of the categories of tools discussed below; however, we emphasize those methodologies that are freely available to academic researchers. For comprehensive tables of all available programs, refer to the review articles cited within the individual sections.

5.1. Protein structure visualization

There are several widely adopted software programs for visualizing protein structures, such as PyMOL (http://www.pymol.org) [187], UCSF Chimera (http://www.cgl.ucsf.edu/chimera) [188], and VMD (http://www.ks.uiuc.edu/Research/vmd) [189].

5.2. Protein structure prediction

The basic requirement for studying the structural mechanisms underlying variable drug response is a high-resolution protein structure. Three-dimensional structural data are increasingly available from the PDB; however, many drug target structures remain elusive owing to crystallization difficulties and protein size limitations of NMR. Moreover, while the structure of a wild-type target may be available, those of the variant proteins are often unknown. In such scenarios, computational methods can be used to predict three-dimensional protein models [164]. There are two main modelling approaches, comparative (or homology) methods that use structures of homologous proteins as starting templates, and free (or ab initio) methods that use knowledge-based algorithms or first principles [190]. Computational methods for predicting protein structure have been reviewed in detail [190–192], as have automated protein modelling servers [193].

5.2.1. Comparative modelling techniques

Comparative modelling is based on the observation that proteins with similar amino acid sequences have similar structures [194]. Thus, the three-dimensional structure of a protein (model) can be built based on the experimentally determined structure of a homologous protein (template). Comparative modelling methods have been extensively reviewed and evaluated [195–198]. The process of constructing a homology model for a protein of interest consists of the following four stages, template selection, alignment of target and template protein sequences, model generation and model evaluation and refinement [195].

There are four principle approaches for constructing homology models, spatial restraint, segment matching, multiple template and artificial evolution. MODELER [199] and other spatial restraint techniques extract geometric features (bond lengths, angles, etc.) from the template structure and construct a model by satisfying these restraints. Segment matching tools, including SegMod/ENCAD [200] and the Pfrag extension [201], divide the target protein into fragments, independently align each to a template and assemble the fragment models. In techniques such as SWISS-MODEL [202], multiple template alignments are used to identify conserved structural regions, which are modelled as rigid bodies, while variable regions are built up around them. In artificial evolution programs such as Nest [203], iterative modifications (substitutions, deletions, or insertions) and energy minimizations are
applied to the template structure to gradually build up the target protein structure.

Regardless of the modelling approach, model accuracy depends heavily on template selection and sequence alignment quality. Both factors affect the observed sequence similarity between target and template. When sequence similarity is high (more than 50%), homology modelling can reliably generate accurate, high-resolution predictions suitable for drug design, mutational analysis and binding site detection [190,192,204]. Models built using moderate (30–50%) sequence similarity are typically of lower resolution, but offer sufficient detail to assess druggability and generate hypotheses regarding sequence mutations [190,192,204]. For models generated from templates of low (less than 30%) sequence similarity, the resulting structures are generally speculative; however, accuracy can be improved by using multiple template structures, if available [201,204,205].

The ninth edition of the critical assessment of techniques for protein structure prediction (CASP) assessed the state of the art in comparative modelling techniques over a diverse pool of 116 target sequences [206]. Comparison of 61 665 template-based predictions to their experimentally determined structures demonstrated that the best-performing comparative modelling approaches are capable of accurately predicting both overall protein structure and local interactions. Notably, the methods underlying these best-performing technologies are distinct and have different strengths and weaknesses, suggesting that superior models may be attained by integrating different techniques. One commonly observed limitation of comparative modelling techniques is a poor correlation between their estimated and assessed model accuracy.

5.2.2. Free modelling techniques

In the absence of suitable templates for the protein of interest, model creation can be based on experimental data, such as interresidue distance and contact maps, along with secondary structure predictions and advanced force fields [174]. However, experimental contact data are not broadly available, thus requiring that structures be predicted from primary amino acid sequence alone.

Free modelling methods use knowledge-based potentials, physics-based potentials or a hybrid of the two to predict protein structure from first principles. Physics-based approaches, such as QUARK [207,208], perform protein folding using Monte Carlo optimization on physicochemical statistics potentials. These approaches may offer insight into the protein folding pathway, but more importantly require no a priori structural knowledge [209]. In contrast, knowledge-based methods, including ROSETTA [210] and I-TASSER [211], assemble structural fragments with local sequence similarity to the target using Monte Carlo simulation [208,212]. These fragment-assembly tools have been used to predict protein structures with high accuracy [212,213]. Currently, the computational complexity of free modelling methods limits both the model resolution and its applicability to larger protein sequences.

Although free modelling techniques are rapidly advancing, very high accuracy models are rare [214]. Recent CASP results indicate free modelling method performance is highly dependent on target length, with an apparent upper limit of 120 residues [214]. Similarly, accurate prediction of multi-domain structures remains a significant challenge [215]. Much room for improvement remains; evaluation of 16 971 free modelling predictions of 30 target sequences in CASP9 showed that even top-performing technologies produced a number of physically unrealistic models [215]. Despite these challenges, recent years have seen dramatic improvements in the prediction accuracy of free modelling techniques for short target sequences [214].

5.3. Drug binding site analysis

Understanding differential drug outcomes requires high-resolution structural knowledge of the binding site. Computational methods to identify and analyse drug binding sites can supplement experimentally derived structural knowledge. Such tools have multiple applications, including prediction of drug specificity, guidance of drug development and repurposing, and prediction or interpretation of drug response. In addition, binding site prediction and analysis tools offer insights into the effect of mutations on the druggability of the binding pocket [166,167].

5.3.1. Binding site prediction

Binding site detection is challenging because proteins frequently undergo large structural changes upon ligand binding (see §3.1.1). Structure-based computational algorithms for predicting binding sites are described in detail in recent reviews [216–218]. These tools can be divided into four main categories, structural similarity, geometric, energy-based and docking.

Structural similarity approaches, such as 3DLigandSite [219], compare a query structure with a binding site library extracted from protein–ligand complexes, select a subset of similar structures and superimpose the ligands onto the query structure to infer the binding site location. Geometric-based methods, including fpocket [220], are based on shape and assume that the binding site is located within a cavity; potential binding cavities are typically detected by placing or rolling spheres of fixed or variable radii along the protein structure. SiteHound and other energy-based algorithms work on the hypothesis that the energetic properties of binding sites differ from those of the surrounding protein surface [221,222]. Creation of interaction affinity maps between the protein surface and representative chemical probes reveal protein surface patches of high total interaction energy that represent probable ligand binding pockets. Finally, docking methods for predicting binding sites computationally dock libraries of drug-like fragments, as in FTMAP [223], or compounds, as in MolSite [224], to the protein structure. These tools are computational analogues to experimental approaches for studying protein druggability by NMR [40] or X-ray crystallography [225].

Pocket prediction programs vary widely in performance, largely depending on the input structure conformation.
(apo- or holo-protein) and number of returned sites considered for further analysis [38,226,227]. In addition to predicting their location, many of these tools also provide detailed characterization of binding sites suitable for use in comparative studies, as discussed in the following section.

5.3.2. Binding site structural comparison

Proteins lacking in overall sequence and structural homology can share binding site similarities and therefore bind to common ligands. Algorithms for identifying structurally similar binding sites have broad applications, including evaluating protein druggability and inferring protein–ligand interactions, both on- and off-target, as recently reviewed in [218]. Comparative studies of binding sites typically proceed through three steps. Positional and physicochemical properties of cavity residues are first reduced to simplified geometric patterns. These geometric patterns are then aligned to maximize the overlay of shared features. Finally, scoring metrics assess shared features of the final pattern alignments to quantify binding site similarity.

One aspect in which binding site comparison methods differ is in the identification of the optimal pattern alignment. Comparison of protein active site structures (CPASS) employs a straightforward alignment approach, exhaustively iterating translations and rotations of the query pattern to match a fixed target pattern [228]. More efficient methods group proximal pattern elements as triplets and optimally align these using geometric matching, as in ProSurfer [229], or geometric hashing, as in SiteEngines [230]. CavBase takes advantage of clique detection algorithms to align binding sites by representing patterns as graphs with pattern elements as nodes and element proximities as edges [231]. Other factors impacting the performance of ligand binding site comparison tools include the geometric pattern resolution and incorporation of binding site dynamics [232].

In contrast to geometric-based tools for binding site comparison, PocketFEATURE [233] characterizes protein pockets using the physicochemical microenvironments of the pocket residues. Comparison between two sets of pocket residue microenvironments allows identification of similar pockets that are likely to possess similar binding capabilities. Because the comparison uses only weak geometric restraints, this method is less reliant on the accuracy of crystallographic structures and has improved performance on dynamic binding sites.

A critical factor in binding site comparison methods is the correct identification of the binding sites themselves. Care must be taken to accurately identify entire binding cavities so that relevant comparisons can be gleaned. Furthermore, it is important to note that protein targets may bind a panel of small molecules in the same location and by various binding modes. Such heterogeneity can significantly alter the attributes of the binding site and therefore bias the results of comparison.

5.4. Docking and scoring technologies

Although protein structures are numerous, many protein targets have yet to be co-crystallized with their small-molecule ligands. In such cases, docking technologies (reviewed in [234–237]) allow exploration of the specific structural details of the protein–ligand interaction. The principle contribution of docking approaches towards understanding the molecular details of protein–ligand interactions is to determine the binding pose of a small-molecule ligand. Docking technologies also have important applications in drug discovery, where their aim is to distinguish between true and false positives in lead identification. Docking approaches begin with a docking stage, during which ligand orientations and conformations are sampled within the spatial constraints of the predicted protein binding site. Then, during the scoring stage, the best poses for each ligand are identified and ligands are rank-ordered.

There are numerous docking programs, but none achieve high accuracy across all protein targets. Current methods achieve, at best, 60 per cent accuracy for the ligand-binding conformation [238]. Selection of the best docking algorithm and scoring function is target- and ligand-dependent [238,239].

5.4.1. Docking

Docking a ligand into a protein binding site is a multivariate problem that models several degrees of freedom, including intermolecular translation and rotation and intramolecular conformational changes [239]. There are multiple docking approaches, interaction site matching, incremental construction, genetic algorithms and Monte Carlo searches [234,236]. Interaction site matching approaches, such as FRED [240], represent ligands and protein binding sites as pharmacophores and optimize their overlay to generate a docked ligand pose. FlexX [241] and other incremental construction programs rebuild the ligand in the protein binding site using libraries of preferred ligand fragment conformations. Genetic algorithm docking techniques, including AutoDock [242], mimic natural evolution where ligand conformations are encoded on ‘chromosomes’, diversified by genetic operators (crossovers, mutations and migrations), and subjected to natural selection using a fitness function. Monte Carlo-based docking approaches, like Glide [243], iteratively introduce random perturbations to the ligand pose and accept/reject them using a Monte Carlo criteria. It is important to note that many docking technologies incorporate multiple or blended approaches into their methods.

As discussed earlier in §3.1, protein–drug binding is a structurally dynamic process. While historical docking programs treated the protein and ligand as rigid bodies, recent shifts towards a dynamic representation of the system have improved docking accuracy. However, incorporation of protein flexibility exponentially expands the docking search space [244]. Docking methods address this challenge through varied means, considering a conformational ensemble of target structures (FlexX-Ensemble [245]), softening the binding site through reduced van der Waals penalties (ADAM [246]), and incorporating sidechain or backbone flexibility through rotamer libraries (AutoDock [247]). These methods capture both the induced fit and conformational selection models of protein–ligand binding.
by accounting for protein flexibility and structural ensembles, respectively.

In addition to protein and ligand flexibility, a number of other considerations also affect docking performance, such as protonation and tautomeric state of the ligand [248] as well as treatment of water molecules and solvent in the binding pocket [249]. These factors continue to limit docking performance today [237].

5.4.2. Scoring
Following docking, the resulting protein–ligand complexes are scored to identify the most biologically probable conformations and estimate their interaction strength. There are three types of scoring functions, force field-based, empirical and knowledge-based [235]. Force field scoring functions assess the physical atomic interactions of the system using van der Waals, electrostatics, and bond stretching/bending/torsional forces that are parameterized from experimental data and quantum mechanical calculations. More simply, empirical scoring functions evaluate ligand–protein complexes using energy terms (van der Waals energy, electrostatics, hydrogen bonding, desolvation, entropy, and hydrophobicity) weighted by fitting to binding affinity data of experimentally determined protein–ligand structures. Knowledge-based scoring functions derive pairwise atomic interaction potentials from experimentally determined protein–ligand complexes, with the assumption that frequently observed interactions are favourable.

The most stringent tests of a scoring function are rank ordering a series of related compounds and predicting their binding affinities [250]. Yet, both goals remain elusive using current scoring algorithms [237,239]. Recent alternative approaches addressing this limitation include use of machine-learning scoring functions [251], machine-learning scoring functions incorporating geometric descriptors [252] and consensus scoring with multiple scoring functions [237]. These approaches occasionally perform well for a particular compound series or target but are not universally applicable. Importantly, as the content of structural databases grows, the performance of machine-learning scoring functions is expected to further improve.

5.5. Molecular dynamic simulation
Protein flexibility and dynamics are fundamental to protein–drug interactions (see §3.1.1, 4.3.2.1 and 4.4). Protein dynamics range in scale from rearrangement of binding pocket sidechains to large coordinated movements of entire protein domains [51]. MD simulation is a popular tool for studying the conformational space accessible to proteins and protein–ligand complexes [253–255]. It has been used to refine experimental or modelled protein structures [256], reveal transient binding sites [257], examine the stability and strength of docked protein–ligand conformations [244], aid drug discovery [258,259], and explore altered drug binding profiles of protein variants [169,260,261].

In general, MD simulations depict the physical movements of atoms and molecules as they interact over time. This is accomplished by iteratively calculating the instantaneous forces present in the system (typically protein, ligand, solvent and often a lipid bilayer) and the resultant movements [253]. Forces between atoms and the potential energy of the system are defined by force fields, which contain energy functions with parameters derived from experimental or quantum mechanical studies. Common force fields developed specifically for the simulation of proteins include OPLS-AA [262], CHARMM [263], and AMBER [264]; each contain inherent biases that affect the sampled conformational space [265]. Widely adopted MD software include GROMACS [266], AMBER [267] and NAMD [268]. Thorough reviews on MD simulation methods were recently published [253,254,258].

Chief practical and technical considerations for MD studies are simulation length, system size and system resolution. Ideally, an MD simulation would provide a continuous, atomic-level view of system interactions over a long timescale. Recent methodological progress towards this ideal include the use of graphics processing units [269] and distributed computing [270]. These advances facilitate the simulation of relatively slow processes, such as the large conformational movements of kinases upon drug binding, and of large macromolecular systems, such as solvated GPCR structures embedded in a lipid bilayer [254]. Of exceptional note, Shaw and colleagues recently performed very long (approaching the millisecond and beyond) MD simulations on an all-atom protein system using a special-purpose machine [271]. However, simulation length poses an ongoing challenge, as many important biomolecular motions are slower than even the longest simulation timescales accessible today [272]. Short timescale MD simulations can be misleading, as they may not fully capture the dynamic nature of a protein–ligand system.

Given the limitations discussed above, MD simulation is of principle use when small conformational changes are expected. Normal mode analysis (NMA) is a powerful tool for exploring large conformational changes in protein structures, which are often important for ligand binding events [273–275]. NMA tools, such as The Elastic Network Model (eNémo) webserver [276], use a few low-frequency motions to describe rearrangements of protein domains and other types of large-amplitude MD. In contrast to many MD simulations, NMA offers the potential to extract essential dynamic information for global movements of large protein systems; however, coarse-grained models must often be used in such cases and the resulting studies therefore suffer from lower accuracy and specificity at the local scale [277].

6. DATABASES
There is currently no single database suitable for a thorough examination of the interplay between coding mutations, protein structure and drug response. Here, we briefly highlight several databases dedicated to individual components of the effect of genetic variation on
Table 2. Databases for studying three-dimensional protein structures, genetic variants and drug response.

<table>
<thead>
<tr>
<th>database</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>three-dimensional protein structures</td>
<td></td>
</tr>
<tr>
<td>CSA [278]</td>
<td><a href="http://www.ebi.ac.uk/thornton-srv/databases/CSA">http://www.ebi.ac.uk/thornton-srv/databases/CSA</a></td>
</tr>
<tr>
<td>DCD [41]</td>
<td><a href="http://fpocket.sourceforge.net/dcd">http://fpocket.sourceforge.net/dcd</a></td>
</tr>
<tr>
<td>ModBase [280]</td>
<td><a href="http://modbase.compbio.ucsf.edu">http://modbase.compbio.ucsf.edu</a></td>
</tr>
<tr>
<td>PDB [34]</td>
<td><a href="http://www.pdb.org">http://www.pdb.org</a></td>
</tr>
<tr>
<td>SitesBase [283]</td>
<td><a href="http://www.modelling.leeds.ac.uk/sb">http://www.modelling.leeds.ac.uk/sb</a></td>
</tr>
<tr>
<td>three-dimensional protein–ligand structures and interactions</td>
<td></td>
</tr>
<tr>
<td>CPASS [228]</td>
<td><a href="http://cpass.unl.edu">http://cpass.unl.edu</a></td>
</tr>
<tr>
<td>Relibase [287]</td>
<td><a href="http://relibase.rutgers.edu">http://relibase.rutgers.edu</a></td>
</tr>
<tr>
<td>genetic variants &amp; disease</td>
<td></td>
</tr>
<tr>
<td>genetic variants and three-dimensional protein structures</td>
<td></td>
</tr>
<tr>
<td>VaD [296]</td>
<td><a href="http://vandd.org">http://vandd.org</a></td>
</tr>
<tr>
<td>genetic variants and drug response</td>
<td></td>
</tr>
</tbody>
</table>

Drug response from the perspective of protein structure (summarized in table 2).

6.1. Databases of three-dimensional protein structures

The most comprehensive protein structure repository is the PDB, which currently holds approximately 73,000 protein structures (roughly 25% are of human origin) [34]. The PDB is well integrated into the network of bioinformatics tools and contains links to external resources for protein and small-molecule entities, integrated software packages for protein structure comparison and small-molecule similarity searching, and protein structural and functional annotations derived from other databases. Multiple secondary databases then extract and organize the experimentally determined structural data from the PDB using different criteria, sc-PDB collects structural examples of drug binding sites and includes analyses of the binding cavities and ligand chemical structures [299]. SitesBase annotates and compares ligand binding site structural similarities [283]. Druggable Cavity Directory (DCD) is a manually annotated repository of binding sites scored for druggability [41]. FireDB contains structures, ligands and annotated functionally important binding site residues [279]. Catalytic Site Atlas (CSA) annotates enzyme active sites, specifically catalytic residues, as three-dimensional structural templates for structures derived from the PDB [278]. In addition, there are a handful of repositories for protein models, including ModBase [280], Protein Model Portal [281] and SWISS-MODEL Repository [284].

6.2. Databases of protein–ligand interactions

Several databases integrate known protein–ligand interactions with a variety of external data sources. DrugBank is a catalogue of small molecules and integrates target protein structures when available [48]. Other databases, such as Binding MOAD [285] and PDBbind [286], link PDB structures with experimental binding data. Relibase offers tools for comparing ligand binding sites, analysing ligand similarity and searching for binding partners [287]. Similarly, CPASS database contains ligand-defined protein active sites from structures in the PDB [228]. Furthermore, the Structural Biology Knowledgebase (SBBK) [288] and Therapeutic Targets Database (TTD) [289] enhance the study of protein–ligand interactions with information regarding three-dimensional protein structures, ligands, pathways and diseases.

6.3. Databases linking genetic variants and disease

The most comprehensive SNP database is dbSNP, with approximately 20 million validated SNP entries [290]. Online Mendelian Inheritance in Man (OMIM) links genetic disorders with their causative genes [292]. SwissVar is a curated set of annotated missense SNPs linked with protein functional changes and possible disease
association [294]. SNP and copy number annotation database (SCAN) annotates SNPs according to physical chromosomal location or effect on gene expression [293]. The catalogue of published genome-wide association studies (GWAS) is a collection of manually curated SNP-trait and SNP-disease associations from nearly 1000 published GWAS [291].

6.4. Databases linking genetic variants and protein structure

The LS-SNP database maps human coding mutations onto protein structures and assesses positional residue conservation patterns within the protein superfAMILY to predict the mutation’s structural and functional impact [295]. Variations and Drugs (VnD) database is a structure-centric database of disease-related protein variants and drugs [296]. Protein family-specific databases also exist. Most notably, SuperCYP database reports manually curated data regarding the effect of SNPs on CYP enzyme structure, activity and drug metabolism [297].

6.5. Databases linking genetic variants and drug response

Parsing the relationship between genetic polymorphisms and drug outcome is complex. A patient’s clinical response depends on pharmacodynamic and pharmacokinetic interactions, both of which can be altered by genetic variation and disease state.

PharmGKB is a manually curated knowledgebase of the impact of genetic variation on drug response [298]. It collects information on genes, drugs and diseases and emphasizes the clinical interpretation of the genetic variants, including information on drug dosing and genetic tests. PharmGKB documents approximately 500 genetic variants that significantly affect drug response. Of these variants, 70 per cent alter pharmacodynamic mechanisms, 10 per cent affect pharmacokinetic mechanisms and 10 per cent disrupt both pharmacodynamics and pharmacokinetics. PharmGKB identifies a subpopulation of genetic variants affecting drug response as “very important pharmacogenomic (VIP)” genes. Missense mutations found in these VIP genes are mapped to representative homologous protein structures (where available) and associated drugs in table 3.

7. OUTLOOK

The interplay between genetic variation and protein structure forms the basis of interindividual variability in drug response. As such, this complex relationship is of increasing importance in bioinformatics and drug development. Three-dimensional protein structures are today frequently used in drug development practices, from selection of a therapeutic target, to determination of a molecular mechanism of action, to identification of a target patient population. As a result, great strides have been made in understanding the structural details governing drug–target interactions for recently approved therapeutic agents. Similarly, this information can be harnessed to predict the impact of missense mutations on drug response. In clinical practice protein structural insights along with a patient’s genetic profile can be leveraged to improve treatment outcome, as evidenced by the results of gene-based treatment trials [7].

Although the focused study of the interaction of a drug with its target protein can offer insights into treatment outcome, complete understanding of drug response requires a systems pharmacology perspective. For instance, multivariate systems biology approaches that consider multiple proteins, signalling pathways, cell types and tissues can accurately recapitulate and predict therapeutic responses [310–313]. Recently, there has been a call for the inclusion of interpatient variability into these models [312]. This strategy holds tremendous promise for improving drug performance in clinical settings, but will first require extensive structural and mechanistic knowledge. Thus, there is an urgent need for high-throughput methods to systematically analyze drug–target interactions from a structural perspective.

Although the rate of structure deposition to the PDB is ever-increasing [314], deposition of pharmaceutically relevant protein structures remains low, limiting our understanding of variable drug response for many targets. Challenges to attaining crystal structures for protein targets include obtaining requisite large quantities of purified protein and inability to prepare crystals of target proteins in biologically relevant conformations [315]. As a result, three-dimensional structures are unavailable for many therapeutically important proteins. For example, though one-third of clinical drugs target these proteins [1], only 13 human rhodopsin-like GPCR (Pfam PF00001) structures are currently available in the PDB. Furthermore, these structures represent a highly biased sampling of rhodopsin conformations, namely inactive and apoprotein forms, precluding structural understanding of GPCR activation upon agonist binding. Only recently did Standfuss et al. [316] release an agonist-bound rhodopsin structure. Barriers to structural availability are gradually being overcome by structural genomics technology advances such as improved protein production platforms, automated high-throughput crystallization and data collection systems, and advanced software for structure determination [16]. Additionally, recently launched initiatives focused on structural determination of biomedically important proteins, including the protein structure initiative, RIKEN Structural Genomics/Proteomics Initiative and Structural Genomics Consortium, have collectively contributed roughly 10 500 unique protein structures to the PDB [314].

Compounding the problem of incomplete and biased structural data for therapeutic targets, this information is not well linked to experimental and clinical information regarding genetic variants and drug response. Current practice for studying variable drug response in the context of target protein structures with genetic variants requires at least partial manual curation. In the future, we anticipate that a standardized lexicon, improved text-mining algorithms and high-throughput structural analysis or prediction methods will lead to integrated and highly annotated databases. Different
Table 3. PharmGKB missense variants affecting drug response. Column 1 contains the protein encoded by the VIP gene in parentheses. A representative protein structure from the PDB is found in column 2. Missense mutations in the gene are listed by reference SNP ID in column 3 and by amino acid substitution in column 4. Column 5 categorizes the variant’s effect as pharmacodynamic (PD) or pharmacokinetic (PK). Associated drugs are listed in column 6.

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB ID</th>
<th>rsID</th>
<th>mut</th>
<th>effect</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-glycoprotein (ABCB1)</td>
<td>3G60[300]</td>
<td>rs2032582</td>
<td>S893T</td>
<td>PD, PK</td>
<td>doxorubicin, paclitaxel</td>
</tr>
<tr>
<td>β-1 adrenergic Receptor (ADRB1)</td>
<td>2Y00[301]</td>
<td>rs2229107</td>
<td>S1141T</td>
<td>PK</td>
<td>phenytoin, phenobarbital, verapamil</td>
</tr>
<tr>
<td>β-2 adrenergic receptor (ADRB2)</td>
<td>2R4R[302]</td>
<td>rs1042713</td>
<td>R16G</td>
<td>PD</td>
<td>fleurouracil, metoprolol, salmeterol</td>
</tr>
<tr>
<td>catechol O-methyltransferase (COMT)</td>
<td>3BWM[303]</td>
<td>rs4680</td>
<td>V158M</td>
<td>PD</td>
<td>antipsychotics</td>
</tr>
<tr>
<td>cytochrome P450 2A6 (CYP2A6)</td>
<td>1Z10[304]</td>
<td>rs1801272</td>
<td>L160H</td>
<td>PK</td>
<td>nicotine, anthracyclines, capesitabine, cyclosporine</td>
</tr>
<tr>
<td>cytochrome P450 2B6 (CYP2B6)</td>
<td>3IBD[305]</td>
<td>rs28399468</td>
<td>R485L</td>
<td>PK</td>
<td>cyclophosphamide, enfluramizne, doxorubicin, efavirenz, furoxenadine</td>
</tr>
<tr>
<td>cytochrome P450 2C9 (CYP2C9)</td>
<td>1OG2[144]</td>
<td>rs8192709</td>
<td>R22C</td>
<td>PK</td>
<td>losartan, phenytoin, sulfonamides and urea deriv., valproic acid, warfarin</td>
</tr>
<tr>
<td>cytochrome P450 2D6 (CYP2D6)</td>
<td>2F9Q[306]</td>
<td>rs1065852</td>
<td>P34S</td>
<td>PD</td>
<td>tamoxifen, fenofibrate</td>
</tr>
<tr>
<td>dihydropyrimidine dehydrogenase (DPYD)</td>
<td>1GT8[307]</td>
<td>rs1065852</td>
<td>P34S</td>
<td>PD</td>
<td>tamoxifen, fenofibrate</td>
</tr>
<tr>
<td>glutathione S-transferase (GSTP1)</td>
<td>2A2R[308]</td>
<td>rs1695</td>
<td>I105V</td>
<td>PD, PK</td>
<td>cisplatin, cyclophosphamide, doxorubicin, fluorouracil</td>
</tr>
</tbody>
</table>

(Continued.)
information sources must then be integrated into a comprehensive structural database that allows selection of an annotated missense SNP, mapping of the mutation onto a protein structure, evaluation of the interaction between mutated residues and drug-like ligands, and prediction of the effect of novel SNPs on drug activity or metabolism. Such a database will play a critical role in establishing protein structure information as a key tool in pharmacogenetics studies.

Another area in which technological advances are needed is that of high-throughput structural analysis of mutant protein targets and their interactions with therapeutic compounds. Such tools would build upon existing methods that explore the relationship between

<table>
<thead>
<tr>
<th>protein</th>
<th>PDB ID</th>
<th>rsID</th>
<th>mut</th>
<th>effect</th>
<th>drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>inward-rectifying potassium channel 11 (B2RC52)</td>
<td>—</td>
<td>rs5219</td>
<td>K23E</td>
<td>PD</td>
<td>glibenclamide, metformin, repaglinide, sulfonamides &amp; urea deriv., capecitabine, cisplatin, fluorouracil, leucovorin, mercaptopurine, methotrexate, nitrous oxide, oxaliplatin, pemetrexed, carboplatin, cyclophosphamide, dactinomycin, doxorubicin, fluorouracil, leucovorin, mercaptopurine, methotrexate, nitrous oxide, oxaliplatin, pemetrexed, vincristine</td>
</tr>
<tr>
<td>methylenetetrahydrofolate reductase (MTHFR)</td>
<td>—</td>
<td>rs1801131</td>
<td>E429A</td>
<td>PD</td>
<td></td>
</tr>
<tr>
<td>folate transporter 1 (SLC19A1)</td>
<td>—</td>
<td>rs1051266</td>
<td>H27R</td>
<td>PD, PK</td>
<td>leucovorin, mercaptopurine, methotrexate, prednisone, fluuvastatin, pravastatin, atorvastatin, cerivastatin, HMG-CoA reductase inhib.</td>
</tr>
<tr>
<td>solute carrier organic anion transporter 1B1 (SLCO1B1)</td>
<td>—</td>
<td>rs11045819</td>
<td>P155T</td>
<td>PD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs2306283</td>
<td>N130D</td>
<td>PK</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs4149056</td>
<td>V174A</td>
<td>PD, PK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thiopurine S-methyltransferase (TPMT)</td>
<td>2BZG [309]</td>
<td>rs1142345</td>
<td>Y240C</td>
<td>PD, PK</td>
<td>mercaptopurine, s-adenosylmethionine, cisplatin, mercaptopurine, s-adenosylmethionine</td>
</tr>
<tr>
<td></td>
<td>rs1800460</td>
<td>A154T</td>
<td>PD, PK</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
protein sequence variants, three-dimensional structure and drug response, as discussed previously in this review, and would offer a high-resolution structural understanding of genome-wide pharmacogenetic data. Developments in this area would take advantage of the expansive amount of information anticipated from next-generation sequencing techniques [317] and high-throughput chemical proteomics techniques [318]. Furthermore, the resulting structural technologies could interface with larger systems biology and network pharmacology studies to provide broad insight into variability in drug response among diverse patient populations.

Importantly, structural details of target proteins and their corresponding mutant isoforms can also be applied to inform drug design strategies, as recently reviewed [319–324]. For example, detailed knowledge of a target protein’s three-dimensional structure facilitates the development of more selective drugs [169,325,326] and second-generation therapeutics for treating patients with unresponsive disease [327]. In addition, comparative analysis of target protein structures of currently marketed drugs allows for efficient drug repurposing to accelerate drug development for treating rare and orphan diseases [328–330]. Finally, structural knowledge of a drug target is critical for the intelligent combination of therapeutics acting on a single target in order to improve clinical efficacy and reduce the emergence of drug-resistant variants [331,332].

There is immense opportunity in translating information regarding the genetic differences between individuals and the three-dimensional structures of protein targets into drug development and clinical practice. Although the mechanism of many drugs is still not fully understood, we expect that the increasing number of detailed structural studies revealing drug–target interactions will lead to maximized likelihood of patient response, reduced drug resistance, and fewer adverse events. We believe that through the application of our expanding knowledge in this area it will be possible to address the ‘efficacy–effectiveness gap’ of currently marketed drugs and improve clinical outcomes.

J.L.L., G.W.T., T.L. and R.B.A. acknowledge support from the SIMBIOS National Center for Physics-based Simulation of Biological Structures (GM-61374) and NIH (LM-05652). G.W.T. also acknowledges support from the Stanford Bio-X Bioengineering Graduate Fellowship. E.C. acknowledges support from the Marie Curie International Outgoing Fellowship program (PIOF-GA-2009-237225) funded by the European Commission’s FP7 grant. R.B.A. also acknowledges the NIH/NIGMS Pharmacogenomics Research Network and Database and the PharmGKB resource (GM-61374).

REFERENCES

15. Capriotti, E. & Altman, R. B. 2011 Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics 12 (Suppl. 4), S3. (doi:10.1186/1471-2105-12-S4-S3)


to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22, 2729–2734. (doi:10.1093/bioinformatics/btl423)


106 Zhang, J., Yang, P. L. & Gray, N. S. 2009 Targeting MCT-10-0069)


129 Pan, B. S. et al. 2010 MK-2461, a novel multitargeted kinase inhibitor, preferentially inhibits the activated c-Met receptor. Cancer Res. 70, 1524–1533. (doi:10.1158/0008-5472.CAN-09-2541)


179 Bertini, I., Fragnai, G., Giachetti, A., Luchinat, C., Malletta, M., Parigi, G. & Yeo, K. J. 2005 Combining in silico tools and NMR data to validate protein–ligand


DeLano, W. L. 2002 The PyMOL Molecular Graphics System, version 1.2r3pre. Schrodinger, LLC.


Review. Bioinformatics and drug response  J. L. Lahti et al.  1437


J. R. Soc. Interface (2012)