Briefings in Bioinformatics, 20(6), 2019, 2167–2184

doi: 10.1093/bib/bby078
Advance Access Publication Date: 31 August 2018
Review article

Binding site matching in rational drug design: algorithms and applications

Misagh Naderi*, Jeffrey Mitchell Lemoine*, Rajiv Gandhi Govindaraj, Omar Zade Kana, Wei Pan Feinstein and Michal Brylinski

Corresponding author: Michal Brylinski, Department of Biological Sciences and Center for Computation & Technology, Louisiana State University, Baton Rouge, LA 70803, USA. Tel.: (225) 578-2791; Fax: (225) 578-2597; E-mail: michal@brylinski.org
*Contributed equally.

Abstract

Interactions between proteins and small molecules are critical for biological functions. These interactions often occur in small cavities within protein structures, known as ligand-binding pockets. Understanding the physicochemical qualities of binding pockets is essential to improve not only our basic knowledge of biological systems, but also drug development procedures. In order to quantify similarities among pockets in terms of their geometries and chemical properties, either bound ligands can be compared to one another or binding sites can be matched directly. Both perspectives routinely take advantage of computational methods including various techniques to represent and compare small molecules as well as local protein structures. In this review, we survey 12 tools widely used to match pockets. These methods are divided into five categories based on the algorithm implemented to construct binding-site alignments. In addition to the comprehensive analysis of their algorithms, test sets and the performance of each method are described. We also discuss general pharmacological applications of computational pocket matching in drug repurposing, polypharmacology and side effects. Reflecting on the importance of these techniques in drug discovery, in the end, we elaborate on the development of more accurate meta-predictors, the incorporation of protein flexibility and the integration of powerful artificial intelligence technologies such as deep learning.

Key words: pocket alignment; pocket matching; drug repositioning; drug side effects; off-targets; polypharmacology

Introduction

Computer-aided systems biological approaches have invigorated interest in exploiting the natural promiscuity of drugs in order to repurpose known drugs, elucidate and develop drugs targeting complex pathways and discover relationships between remotely related proteins. These endeavors have far-reaching consequences across a multitude of disciplines. Physiological responses and mechanisms of actions of therapies with multiple bioactive compounds lend themselves nicely to the philosophy of 'one drug, multiple targets' [1]. Research on cancer [2] and neurodegenerative diseases [3] has much to gain from insights

Misagh Naderi is a research associate in the Department of Biological Sciences at LSU. He recently graduated with a PhD in Biochemistry. He also holds an M.Sc. in Chemical Engineering and an M.Sc. in Pathobiological Sciences from LSU.
Jeffrey Mitchell Lemoine is an undergraduate student pursuing a dual B.S. degree in Biochemistry and Computer Science at LSU.
Rajiv Gandhi Govindaraj is a postdoctoral researcher in the Department of Biological Sciences at LSU. He holds a PhD in Computational Biology from Ajou University.
Omar Zade Kana is a research associate in the Department of Biological Sciences at LSU. He recently graduated with a B.S. in Biochemistry.
Wei Pan Feinstein is an IT consultant in the High-Performance Computing at LSU. She holds a PhD in Biomedical Sciences from the University of South Alabama and an M.Sc. in Computer Science from the University of Alabama.
Michal Brylinski is an associate professor in the Department of Biological Sciences and the Center for Computation & Technology at LSU. He holds a PhD in Chemistry from Jagiellonian University and a Pharm.D. from Wroclaw Medical University.
Submitted: 15 June 2018; **Received (in revised form):** 18 July 2018

into the human interactome. As such, it is imperative to scrutinize the intricate networks of drug–protein interactions at a systems level. Among many approaches to study disease-related biological networks [4], the physicochemical characterization of drug-binding pockets in macromolecular structures holds a significant promise to facilitate drug development efforts contributing to the understanding of protein molecular functions [5, 6].

Pocket matching algorithms assess similarities between pairs of binding sites. Binding sites are considered similar if they function in the same way and/or bind the same ligand. Under a classical, 'one drug, one target' view of pharmacology, there would be no similar drug-binding sites. Nonetheless, it became evident that the concept of one drug acting on a single receptor is inaccurate [7]. Modern pharmacology recognizes drugs as acting on biological systems, rather than exclusively on their intended targets. While it is possible for a drug to solely interact with its primary target, the number of target proteins per drug is likely quite high. It was initially estimated to be 6.3 on average [8]; however, due to a high degree of incompleteness of the available experimental data, a recent study suggested that the number of off-targets is substantially higher [9]. Accordingly, the drug promiscuity is the principle phenomenon behind drug repurposing, side effects and polypharmacology. Assuming that the molecular target for a given drug is known, a set of other proteins this compound may bind to can be inferred by pocket matching algorithms because similar binding sites are expected to bind similar ligands. Therefore, accurate pocket matching algorithms are invaluable to drug research and development programs, particularly those employing the network pharmacology paradigm.

In this communication, we review a number of methods to compare ligand-binding pockets in proteins with respect to underlying algorithms, the representation of molecular structures and performance assessments. We also describe their major applications in modern drug development, including drug repurposing, the design of multi-target drugs and the analysis of side effects. Finally, we discuss future directions in research focused on the characterization, classification and comparison of drug-binding pockets.

Ligand-binding pockets in proteins

Proteins routinely perform their biological functions by interacting with a variety of cellular molecules, such as other proteins, small organic compounds, nucleic acids and lipids. In contrast to large and mostly planar protein-binding interfaces [10], cavities and pockets in protein structures typically are locations binding small molecules [11]. These concave shapes allow ligands, such as endogenous compounds and pharmaceutical drugs, to form multiple, non-covalent interactions predominantly with the side chains of binding pocket residues. Ligand-binding pockets vary widely in size, most within the volume range of 100–1000 Å^3 [12]. In the following sections, we briefly review two areas of research on ligand-binding sites relevant to pocket matching in rational drug design, the detection of pockets in protein structures and the representation of their geometry and physicochemical properties.

Identification of pockets

Traditionally, the position of a ligand in the protein structure can be determined through experimental observation of protein–ligand binding events facilitated by X-ray crystallography and nuclear magnetic resonance (NMR). The advantage of these experimental techniques is that they provide high-resolution structures of ligands bound to target proteins, which can then directly be used to analyze and compare drug–target interactions with interaction pattern descriptors [13]. Nevertheless, due to the protein size, instability, low yield and other factors, many experimental protein–ligand data, particularly for novel targets and drugs, remain elusive. Although site-directed mutagenesis [14] and structure–activity relationship by NMR [15] can be used to locate binding pockets, X-ray crystallography or other biophysical studies are generally required to determine the exact binding modes of drugs. Recently, protein structures containing binding pockets intractable to other experimental techniques have been characterized with cryo-electron microscopy (cryo-EM) [16]. However, the application of cryo-EM to relatively small proteins remains challenging because of a low signal-to-noise ratio.

As an alternative, a number of computational methods have been developed to detect potential drug-binding sites based on the surface geometry, physicochemical properties, energetics and evolutionary information [17]. Geometry-based approaches often annotate the largest cavities in protein structures as putative binding sites, although other factors can also inform the detection of binding pockets. For instance, LIGSITE [18], SURFNET [19] and CASTp [20] employ purely geometrical characteristics, whereas Fpocket [21] considers additional physicochemical properties. Energy-based methods, such as Q-SiteFinder [22] and SiteHound [23], identify binding sites by modeling binding potentials and energies. Finally, LIGSITEcsc [24], ConCavity [25] and *e*FindSite [26] integrate structural and evolutionary information to detect ligand-binding pockets. Because of the rapidly growing structural and sequence data for numerous protein families, structure and evolution-based techniques are currently considered the most accurate methods to infer ligand-binding sites.

Pocket representation

Many pocket matching methods construct binding site alignments by finding equivalent atoms, feature points or residues between a given pair of pockets. Figure 1 shows various representations of a protein structure. In the center, an adenosine triphosphate (ATP) molecule bound to protein kinase C iota type, PKC-iota from human [Protein Data Bank (PDB)-ID: 3a8w, chain B] [27], is shown along with four selected binding residues, G252, Y256, A272 and D387. Some algorithms, e.g. Pocket Alignment in Relation to Identification of Substrates (PARIS) [28], describe binding pockets as sets of non-hydrogen atoms in the 3D space (Figure 1A). Individual atoms are typically assigned vectors of features corresponding to atomic coordinates and additional information, such as the atom type, partial charge and amino acid type. Other techniques rely on a reduced representation of ligand-binding sites. For instance, PocketFEATURE defines local regions in the structure with certain properties according to the FEATURE system [29], referred to as microenvironments. Figure 1B shows functional centers of selected microenvironments, non-polar (hypothetical Cβ of G252 and Cβ of A272), aromatic (Cγ, Cδ1, Cδ2, Cε1, Cε2 and Cζ of Y256) and negatively charged (Oδ1, Cγ and Oδ2 of D387). Compared to fine-grained models, representing amino acid residues with effective points reduces the complexity and improves computational performance. Another example is PocketAlign employing four different models of protein structures [30], one of which uses the backbone atoms, N, Cα, C and O, and places a single effective point at the side-chain centroid (Figure 1C). Finally, many

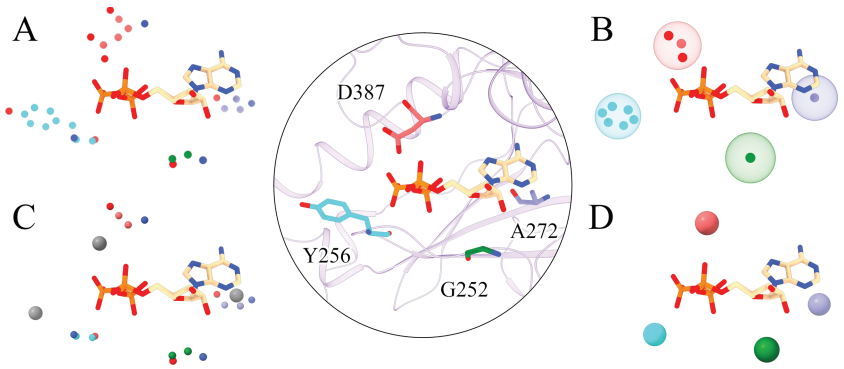

Figure 1. Various pocket representation methods. An ATP molecule colored by atom type (C – peach, N – blue, O – red, P – orange) bound to the pocket of protein kinase C iota type is shown in the center with four selected binding residues, G252, Y256, A272 and D387 represented as green, cyan, blue and red sticks, respectively. (**A**) All-atom pocket representation, in which individual atoms are depicted by small spheres colored by atom type (N – blue, O – red, C – green/cyan/blue/red). (**B**) Sets of functional centers of selected microenvironments, hypothetical Cβ (green), Cβ (blue), aromatic (cyan) and negatively charged (red). Functional centers are represented by small solid spheres, whereas large, transparent spheres correspond to the local microenvironments. (**C**) A mixed representation of pocket residues employing backbone atoms (small spheres colored as in **A**) and the side-chain centroids (large, gray spheres). (**D**) A coarse-grained model based on Cα atoms, which are depicted by solid spheres colored according to residues in the center image.

coarse-grained approaches, e.g. SMAP [31] and *e*MatchSite [32], utilize only Cα atoms of binding residues (Figure 1D). In general, approaches based on the reduced representation of protein structures are less sensitive to side-chain distortions in computer-generated protein models and low-resolution experimental structures, thus these techniques are particularly suitable to analyze large and heterogeneous datasets.

Pocket matching and alignment

In this communication, we review 12 programs to match ligand-binding sites. These tools are divided into five groups based on the alignment algorithm: I – clique-based methods, II – methods solving the assignment problem, III – methods combining the clique detection and the assignment algorithm, IV – methods employing the geometric hashing and sorting and V – methods employing the rotational and translational search. Table 1 shows the group assignment together with the major characteristics of individual programs, including the protein and pocket representation, datasets used to evaluate the performance, software availability and links. All tools selected for this study are freely available to the academic community as stand-alone software and webservers.

Key concepts in pocket matching are illustrated for PKC-iota and another ATP-binding protein, N5-carboxyaminoimidazole ribonucleotide synthetase, purK from *Escherichia coli* (PDB ID: 3eth, chain B) [33], whose sequence identity to PKC-iota is only 20.5%. Although purK is structurally unrelated to PKC-iota with a Template Modeling (TM)-score [34] of 0.30 and a Cα root-mean-square deviation (RMSD) of 4.81 Å over 118 aligned residues, both proteins have a similar ATP-binding pocket (Figure 2). The Szymkiewicz–Simpson overlap coefficient (SSC) was introduced to quantify the similarity of ligand-binding environments [35]. A large-scale analysis demonstrated that the median SSC of 0.30 for similar pocket pairs decreases to 0.05 for pairs of dissimilar pockets. The SSC calculated for binding sites in PKC-iota and purK is 0.31, indicating that both pockets indeed create a similar binding environment. For the purpose of demonstration, we selected four residues in each target; G252, Y256, A272 and D387 in PKC-iota (Figure 2A), and G125, L189, N237 and H244 in purK (Figure 2B). The correct alignment of binding sites constructed by superposing ATP molecules bound to both targets is shown in Figure 2C. Furthermore, various alignment algorithms are illustrated in Figure 3 for the PKC-iota/purK example to help explicate commonalities within each group and the essential features of individual methods.

The performance of pocket matching algorithms is frequently evaluated with the receiver operating characteristic (ROC) analysis. A ROC plot is a true positive rate (TPR) plotted against a false positive rate (FPR), defined as

$$TPR = \frac{TP}{TP + FN} \quad \text{and} \quad FPR = \frac{FP}{FP + TN},$$

where *TP* is the number of true positives, i.e. pairs of pockets binding the same ligand correctly classified as similar, and *TN* is the number of true negatives, i.e. pairs of pockets binding different ligands correctly classified as dissimilar. *FP* is the number of false positives, or over-predictions, i.e. pairs of pockets binding different ligands classified as similar, and *FN* is the number of false negatives, or under-predictions, i.e. pairs of pockets binding the same ligand classified as dissimilar. The corresponding area under the ROC curve (AUC) is a useful metric to measure the performance of pocket matching. A perfect classifier yields an AUC close to 1, whereas an AUC of about 0.5 indicates that a classifier is no better than random. Note that AUC values calculated for different datasets may not be directly comparable on account of varying ratios of similar and dissimilar pocket pairs, different classes of binding ligands and potential global similarities between target proteins frequently leading to an overestimated performance [35].

Group I: clique-based methods

In order to compare binding sites, protein structures can be transformed into graphs whose vertices represent structural or physicochemical features and edges signify distances or bonds. For instance, Figure 3A shows the graph representations of pockets in PKC-iota (on the left) and purK (on the right) with vertices corresponding to binding residues and connections indicating spatially close residues. Constructing a pocket alignment can then be reformulated as a graph-based similarity problem. Although graph-matching problems do not belong to a particular complexity class [36], the graph isomorphism is proven to be

Table 1. Selected programs to match ligand-binding sites in proteins. A total of 12 methods are assigned to five groups (I–V) based on the alignment algorithm. Protein and pocket structure representations employed by individual programs are provided along with datasets used to evaluate their performance, the availability and links to stand-alone software and webservers. The last column gives the primary citation for each method

Group	Program	Representation	Datasets	Availability	URL	Citation
I: *Clique detection*	SMAP	Cα atoms	Adenine-binding[g]	S, W	http://compsci.hunter.cuny.edu/~leixie/smap/smap.html	[30]
	ProBiS	Functional groups[a]	Non-redundant PDB[h]	S, W	http://probis.cmm.ki.si	[37]
	IsoMIF	Molecular interaction fields[b]	Adenine-binding[g], steroid-binding[i], Kahraman[j], Homogenous[k], PDBbind[l], sc-PDB[m]	S	http://biophys.umontreal.ca/nrg/NRG/IsoMIF.html	[38]
II: *Assignment problem*	*e*MatchSite	Cα atoms	Adenine-binding[g], steroid-binding[i], Kahraman[j], Homogenous[k], TOUGH-M1[n]	S	https://github.com/michal-brylinski/ematchsite	[31]
	APoc	Cα atoms, Cα–Cβ vectors	Calcium-binding[o], APoc[p], TOUGH-M1[n]	S	http://pwp.gatech.edu/cssb/apoc/	[61]
III: *Clique and assignment*	G-LoSA	Chemical feature points[c]	APoc[p], TOUGH-M1[n]	S	https://compbio.lehigh.edu/GLoSA	[70]
	BSAlign	Cα atoms, Cα–Cβ vectors	ATP-binding[r]	S	http://www.aungz.com/BSAlign/	[71]
IV: *Geometric hashing and sorting*	PocketAlign	N, Cα, C, O, Cβ and side-chain centroid	Pockets selected from different datasets[s]	S, W	http://proline.physics.iisc.ernet.in/pocketalign/	[29]
	PocketFEATURE	Microenvironments[d]	Adenine-binding[g], FAD-binding[t]	S	https://simtk.org/projects/pocketfeature	[76]
	SiteEngine	Pseudo-centers[e]	ASTRAL database[u], serine proteases[v], TOUGH-M1[n]	S, W	http://bioinfo3d.cs.tau.ac.il/SiteEngine/	[68]
V: *Rotational and translational search*	PARIS	Non-hydrogen atoms	Kahraman[j], Homogenous[k]	S	http://projects.cbio.mines-paristech.fr/paris/paris.html	[27]
	SiteAlign	Fingerprints[f]	sc-PDB[w]	S	http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.html	[85]

Representation: [a]Hydrogen-bond donors and acceptors, mixed acceptor/donor groups, aromatic and aliphatic. [b]Hydrophobic, aromatic, hydrogen-bond donors and acceptors and positively and negatively charged groups. [c]Hydrogen-bond donors and acceptors, hydroxyl groups, positively and negatively charged atoms, aromatic rings and hydrophobic aliphatic groups. [d]Non-polar, polar, positively and negatively charged and aromatic groups. [e]Hydrogen-bond donors and acceptors, mixed acceptor/donor groups, hydrophobic aliphatic and aromatic groups. [f]Topological: the distance of Cβ from the pocket center, the side chain orientation and size; chemical: hydrogen bond donors and acceptors, aliphatic and aromatic groups and the type of charge.
Datasets: [g]247 proteins complexed with adenosine diphosphate (ADP), adenosine triphosphate (ATP), flavin adenine dinucleotide (FAD), nicotinamide adenine dinucleotide (NAD), S-adenosyl-L-homocysteine (SAH) and S-adenosylmethionine (SAM) and 101 control proteins believed not to bind ligands containing an adenine moiety. [h]~23 000 structures. [i]8 proteins complexed with 17β-estradiol (EST), estradiol-17β-hemisuccinate (HE7) and equilenin (EQU), and 1854 control proteins binding small molecules whose size is comparable to that of steroids but have different chemical structures. [j]100 proteins complex with adenosine monophosphate (AMP), 3-β-hydroxy-5-androsten-17-one (AND), adenosine triphosphate (ATP), estradiol (EST), flavin-adenine dinucleotide (FAD), flavin mononucleotide (FMN), α-D-glucose (GLC), heme (HEM) and nicotinamide adenine dinucleotide (NAD) and phosphate (PO4). [k]Homogeneous dataset comprises proteins complexed with the following ligands: pentaethylene glycol (1PE), β-octylglucoside (BOG), glutathione (GSH), lauryl dimethylamine-N-oxide (LDA), 2-lysine(3-hydroxy-2-methyl-5-phosphonooxymethyl-pyridin-4-ylmethane (LLP), palmitic acid (PLM), 4'-deoxy-4'-aminopyridoxal-5'-phosphate (PMP), S-adenosylmethionine (SAM), sucrose (SUC) and uridine monophosphate (U5P). [l]3446 structures. [m]8077 structures. [n]7524 non-redundant complexes containing 1266 different ligands. [o]1024 calcium-binding sites. [p]2090 subject and 21 660 control proteins. [r]126 ATP-binding proteins. [s]53 pocket pairs. [t]249 FAD-binding and 6709 non-FAD-binding proteins. [u]4375 structures. [v]24 structures. [w]6415 structures.
Availability: S, standalone software; W, webserver.

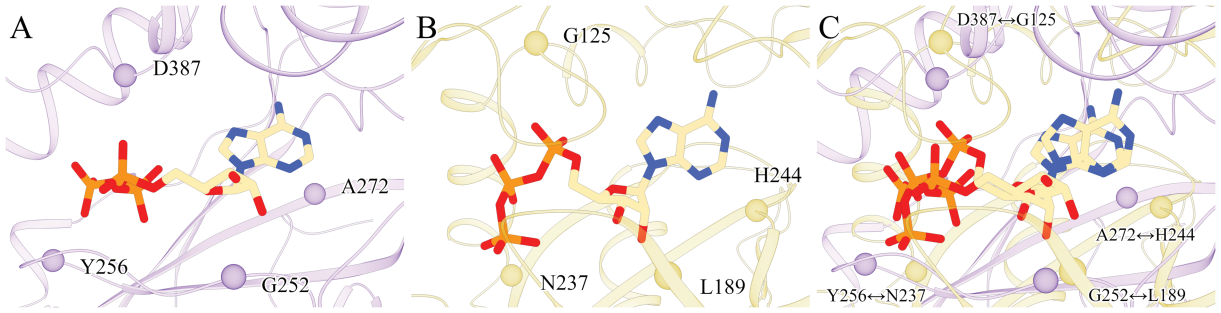

Figure 2. Similar binding sites in two globally unrelated proteins. ATP molecules bound to (**A**) human protein kinase C iota type, PKC-iota (purple) and (**B**) N5-carboxyaminoimidazole ribonucleotide synthetase, purK from *E. coli* (gold). The Cα atoms of four selected pocket residues in each structure are represented by solid spheres and labeled. (**C**) The superposition of PKC-iota and purK based on ATP molecules. Equivalent residue pairs indicated by double-headed arrows form the local alignment of two ATP-binding sites.

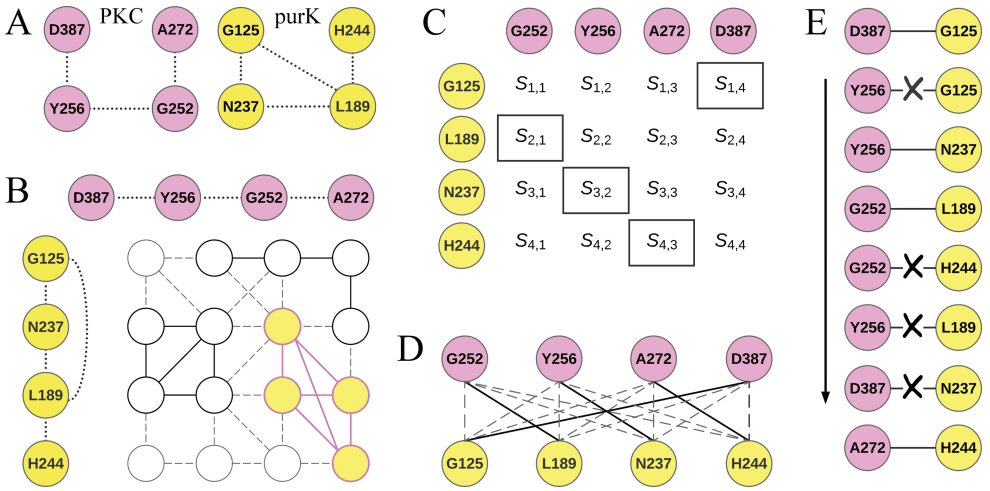

Figure 3. Different algorithms to align ligand-binding sites. Three types of techniques are presented, (**A** and **B**) the clique detection, (**C** and **D**) the assignment method and (**E**) the geometric sorting, for a pair of ATP-binding sites in PKC-iota (purple) and purK (gold) whose molecular structures are shown in Figure 2. (**A**) Graph representations of both pockets with binding residues depicted as vertices connected by dotted edges indicating close positions in the structure. (**B**) A modular product of the two graphs displayed on the top (PKC-iota) and the left side (purK). Three instances of a four-node subgraph are highlighted by thick solid lines with the maximum clique colored purple and gold. (**C**) An assignment matrix constructed for binding sites in PKC-iota (rows) and purK (columns) populated with pairwise residue-based scores. The optimal alignment obtained by solving the linear sum assignment problem (LSAP) is marked by solid boxes. (**D**) A bipartite graph showing all the possible one-to-one alignments between PKC-iota and purK binding residues with the optimal alignment found in **C** marked by solid lines. (**E**) Sorting of PKC-iota/purK residue pairs according to the assigned scores. The pocket alignment is constructed by eliminating those pairs having a residue that appears in a higher-ranked pair.

computationally tractable in a polynomial time [37]. In general, graphs whose nodes and edges are in one-to-one correspondence are considered isomorphic. Multiple approaches such as common-induced subgraphs and clique-based algorithms were developed to solve the graph isomorphism problem. In the context of binding site matching, the maximum common subgraph (MCS) between two graphs representing the target and query proteins corresponds to the largest set of equivalent binding residues. The MCS between two graphs can be found by determining the maximum clique in the association graph, which is the modular product of the two graphs. Within a given graph, the largest maximum clique is a complete subgraph that cannot be extended by including more nodes, such as the four-node subgraph colored purple and gold in Figure 3B. The maximum clique in the association graph can be traced back to nodes and edges in the original graphs. In other words, those initial nodes forming the maximum clique in the product graph correspond the largest set of residues in two binding sites that can be aligned. Three programs, SMAP [31], ProBiS [38] and IsoMIF [39], are assigned to this group.

SMAP represents amino acids with Delaunay tessellation of Cα atoms [31], similar to its predecessor, SOIPPA (the Sequence Order Independent Profile-Profile Alignment) [40]. Pocket structures are further characterized with a geometric potential, used to discriminate between possible ligand-binding sites and other functional sites merely based on the physical shape. The geometric potential depends not only on the global structure of a protein but also the environment surrounding each residue. Since SMAP scans the entire structure for ligand-binding sites, it does not require the predicted binding pocket information. This feature also allows SMAP to recognize binding sites located at protein interfaces in those proteins composed of multiple polypeptide chains. Since the tessellated structure can be considered as a graph with vertices corresponding to Cα atoms, SMAP employs a maximum-weight common subgraph (MWCS) algorithm to align a pair of protein binding sites. In order to compute the MWCS, two graphs representing binding pockets are merged if their nodes share similar features including geometric potential, surface normal orientation and their physical distance. SMAP defines a profile distance based on the amino acid

frequency, which is used as a weight associated with the newly merged graph vertices. Subsequently, a branch-and-bound algorithm is applied to the weighted and merged graph to identify a set of maximal weight alignments. This technique provides a solution for the maximum clique problem and, therefore, finds the largest number of adjacent pair vertices. It is important to note that the resulting alignments are fully sequence order-independent.

The scoring function implemented in SMAP to evaluate the constructed alignments is the sum of profile distances weighted by normal vectors and distances between aligned residues. The statistical significance is then assigned by a kernel density estimator calibrated against pairwise functional site alignments between the target protein and proteins randomly selected from the PDB [41]. The performance of SMAP was evaluated on a large dataset of adenine-binding sites. Although ATP and nicotinamide adenine dinucleotide (NAD)-binding sites may not have the same overall shape, both of them would contain a similar sub-pocket accommodating the adenine moiety. The benchmark set includes 247 adenine-binding proteins, whereas the control set comprises 101 non-redundant protein chains with diverse folds. Encouragingly, the algorithm effectively identifies known sequence and structural homologs within the same superfamily according to the Structural Classification of Proteins (SCOP) [42] with a TPR of 0.75 at an FPR of 0.05. This performance represents a significant improvement over Position-Specific Iterated Basic Local Alignment Search Tool [43] and Combinatorial Extension [44], whose TPR values at the same FPR are 0.55 and 0.60, respectively. SMAP is available online as a webserver, SMAP-WS. For a query protein, SMAP-WS performs a pocket similarity search and returns a list of hits sorted by the similarity score along with additional information such as the PDB-IDs and the biological descriptions of matched proteins.

ProBiS (Protein Binding Sites) detects the similarity among binding sites in protein structures having different folds [38]. Pocket matching in ProBiS involves comparing the geometrical and physicochemical properties of binding pockets, and it is conducted at the level of amino acid functional groups. Specifically, proteins are represented as graphs generated for surface residues identified with the Molecular Surface (MS) walk program [45]. Vertices are points in space corresponding to various functional groups of protein residues accountable for interactions with other molecules. Each vertex is labeled with certain physicochemical properties, including hydrogen bond acceptor, hydrogen bond donor, mixed acceptor/donor and aromatic and aliphatic attributes [46]. Two proteins are compared by constructing a product graph retaining only those edges whose lengths in individual graphs representing protein structures differ by less than 2 Å [47]. Subsequently, possible binding site similarities are examined by applying the maximum clique algorithm, where the maximum clique corresponds to the largest similarity between two compared protein graphs in terms of the number of vertices in the product graph. Each maximum clique is equivalent to a single, local structural alignment between two compared proteins. Finally, the constructed alignments are scored with a function combining surface vector angles, surface patch RMSD, surface patch size and expectation values. In absence of prior knowledge of experimental or predicted binding site, ProBiS employs evolutionary conservation method to identify potential binding site pockets.

The performance of ProBiS was evaluated on a set of 10 pairs of proteins adopting different folds, yet having a similar binding site and function [48]. Encouragingly, the mean RMSD calculated over equivalent binding residues for ProBiS alignments is only 4.8 Å. For comparison, other alignment tools produce alignments whose mean RMSD values are much higher, 22.1 Å for DaliLite [49], 9.5 Å for MolLoc [50] and 10.7 Å for MultiBind [51]. In large-scale applications, ProBiS can run in the one-against-all mode with a query protein compared to a non-redundant database of thousands of single-chain structures obtained from the PDB. This database is subject to weekly updates and currently comprises 42 270 structures. It should be noted that this approach is suitable only for protein structures determined by crystallography or NMR.

IsoMIF is an algorithm to investigate molecular interactions between drugs and their targets by comparing binding sites across protein families [39]. Cavities in IsoMIF are detected in the absence of bound ligands by a purely geometric method GetCleft [52]. Next, the physiochemical properties of cavity residues are mapped onto molecular interaction fields (MIFs) with a distance-dependent exponential function. Specifically, MIFs are computed with the following six chemical probes: hydrophobic, aromatic, hydrogen bond donor and acceptor and positively and negatively charged groups. The chemical and geometrical similarities between binding sites are measured by the graph-matching clique detection algorithm [53]. Finally, the MIF similarity score is calculated as the Tanimoto coefficient over matched probes in the largest clique.

IsoMIF has been validated against a number of widely used datasets, Kahraman [54], Homogenous [28], Steroid [32] and SOIPPA [40]. It was demonstrated to have an outstanding performance with the average AUC of 0.82 ± 0.04 across all datasets tested. Its performance was closely followed by *e*MatchSite whose mean AUC is 0.80 ± 0.15. For comparison, SiteEngine and PocketMatch have considerably lower mean AUC values of 0.73 ± 0.16 and 0.60 ± 0.10, respectively. In addition, IsoMIF was subject to the extended validation against the PDBbind Refined [55] and sc-PDB [56] datasets. The mean AUC for all ligands is 0.93 for PDBbind and 0.87 for sc-PDB, whereas the mean enrichment factors 10, assessing the capability to identify those proteins binding the same ligand as the query, are 8.08 for PDBbind and 6.40 for sc-PDB. Even at the lowest level of pairwise sequence identity threshold of 15%, the AUC is 0.79 for both datasets and the enrichment values are 4.97 for PDBbind and 4.46 for sc-PDB. In addition to a high prediction accuracy, the advantage of IsoMIF is that it detects MIFs similarities among protein pockets, which can directly be used to create a pharmacophore model for structure-based drug design.

Group II: methods solving the assignment problem

Bipartite graphs are frequently employed to align ligand-binding sites in proteins. Here, binding residues from two pockets form two independent and disjoint sets of vertices. Edges connect pairs of vertices across the two sets according to the fitness score between two residues with weights assigned by a scoring function. The problem can be presented as a matrix whose rows and columns are sets of vertices representing binding sites and the elements are weights associated with the edges. For example, Figure 3C shows a matrix of all possible connections between binding residues in PKC-iota (horizontal purple nodes) and purK (vertical gold nodes), each assigned a score S. In the bipartite graph, an optimal alignment between two binding sites can be determined by solving a linear sum assignment problem (LSAP), i.e. selecting a set of edges between the pairs of vertices to maximize the sum of their scores. Assuming that the boxed

edges $S_{1,4}$, $S_{2,1}$, $S_{3,2}$ and $S_{4,3}$ in Figure 3C yield the highest possible total score, this solution to the LSAP creates the optimal alignment between binding sites in PKC-iota and purK shown in Figure 3D.

Although early algorithms to solve the LSAP date back to 1940s [57], the Hungarian algorithm reported in mid-1950s was the first method strongly polynomial in time complexity [58, 59]. In the following decades, numerous algorithms were devised to solve the LSAP with varying time complexities [60]. Currently, different applications of the Hungarian method implementing the shortest paths techniques are among the most widely used to solve the LSAP. In this section, we discuss two pocket alignment programs solving the assignment problem, *e*MatchSite [32, 61] and Alignment of Pockets (APoc) [62].

***e*MatchSite** was developed to compare and align binding pockets in a fully sequence order-independent manner [32, 61], building upon the progress made in evolution/structure-based ligand-binding site prediction with FINDSITE [63, 64], FINDSITE^{LHM} [65] and *e*FindSite [26, 66]. In this approach, pockets are identified by *e*FindSite for input apo-protein structures or computer-generated models. In order to compare a pair of binding sites, a set of seven scores are first calculated for each residue at the predicted binding pockets, including sequence and secondary structure profiles, hydrophobicity, spatial placement compared to neighboring residues and interactions with ligands. Based on these residue-level scores, Cα–Cα distances of all-against-all binding residues belonging to two pockets are estimated with a Support Vector Regression algorithm [67]. Subsequently, the Hungarian algorithm [58, 59] is employed to construct a local alignment by identifying an optimal set of residue pairs to yield the smallest sum of Cα–Cα distances. Finally, a matching score, *e*MS-score, is computed with machine learning for a given pair aligned pockets, considering the Cα-RMSD computed over superposed equivalent residues, the averaged residue-level scores, the physicochemical properties of putative binding ligands and the geometric hashing of binding sites.

*e*MatchSite offers a remarkably high tolerance to structure distortions in protein models; for instance, the accuracy of aligning adenine-binding sites in weakly homologous protein models is only 4–9% lower than that obtained for experimental structures. Furthermore, the performance of *e*MatchSite was compared to that of SOIPPA [40], PocketMatch [68], SiteEngine [69] and sup-CK [28] against a number of datasets comprising not only experimental structures but also various quality protein models. *e*MatchSite outperforms other methods across most datasets, particularly when protein models are used. For example, its performance in recognizing similar binding sites is 6% and 13% higher than that of SiteEngine against high- and moderate-quality protein models, respectively. *e*MatchSite is available to academic community as a webserver and a stand-alone software distribution.

APoc attempts to build sequence order-independent pocket alignments in a three-step process [62]. In the first step, APoc constructs an initial, sequential alignment based on gapless alignments, secondary structure comparison, fragment alignments and local contact pattern alignments. Subsequently, dynamic programming is applied in the second step to optimize the initial alignments between a pair of pockets. In the third step, the optimized sequential alignment is passed down to an iterative procedure, which searches for a non-sequential alignment between nonadjacent residue pairs by solving the equivalent LSAP with the shortest augmenting path algorithm [70]. APoc evaluates the constructed alignments with the pocket similarity score (PS-score) that considers the backbone geometry, the side-chain orientation and the chemical similarity of the aligned binding residues. A non-sequential alignment is accepted only if it yields a better PS-score than the sequential alignment. APoc also assigns a statistical significance to the PS-score based on the comparison of millions randomly selected pocket pairs.

Although APoc is claimed to construct sequence order-independent alignments, a recent study shows that non-sequential alignments by APoc seldom produce PS-score values that are better than those obtained for sequential alignments [35]. Consequently, APoc requires target proteins to have similar global structures in order to generate statistically significant alignments of their binding pockets. The performance of APoc was compared with that of SiteEngine against a random sample of 2000 pairs of pockets selected from the APoc dataset [62]. At an FPR of 0.05, APoc achieves a TPR of as high as 62%, whereas the TPR for SiteEngine is only 17%, demonstrating that APoc offers considerably better performance than SiteEngine. Nonetheless, an independent study reveals that this high performance of APoc is likely overestimated by using a biased validation dataset; the actual accuracy of APoc is notably lower when a high-quality, unbiased dataset is employed [35]. Specifically, using pockets detected by LIGSITE [24], the sensitivity of APoc at an FPR of 0.05 is as high as 87.4% for globally similar target pairs, yet it is only 37.9% for pairs of globally dissimilar proteins binding similar ligands. Therefore, in contrast to other pocket matching algorithms, APoc may not be suitable to investigate drug-binding pocket similarity across the protein fold space.

Group III: methods combining clique detection and assignment algorithm

Two programs combining the clique detection and the assignment method, Graph-based Local Structure Alignment (G-LoSA) [71] and BSAlign [72], are discussed in this group. G-LoSA generates initial alignments as multiple maximum clique solutions, which are subsequently refined with the LSAP algorithm. BSAlign first employs the maximum clique detection to determine the MCS and then constructs residue alignments by finding the maximum number of matching vertex pairs with the Hungarian method.

G-LoSA is a feature point-based algorithm developed to detect similar pockets and construct local structure alignments [71]. G-LoSA represents protein structures with sets of Chemical Feature (CF) points calculated for amino acid residues. CFs include hydrogen bond donors and acceptors, hydroxyl groups, positively and negatively charged atoms, aromatic rings and aliphatic hydrophobic groups. Two different search models, an iterative maximum clique search and a fragment superimposition algorithm, are implemented in G-LoSA to solve the LSAP. All possible sequence order-independent alignments of a pair of binding sites constructed based on CFs are evaluated by a size-independent structure similarity score (GA-Score). GA-score ranges from 0 to 1 with the average value across random local structure pairs of 0.49.

G-LoSA was demonstrated to consistently outperform APoc against diverse benchmarking datasets. For instance, the AUC for G-LoSA for calcium-binding sites is as high as 0.98, whereas the AUC for APoc is only 0.46 [71]. In addition, the performance of G-LoSA was assessed for the detection of a local structure conservation in entire proteins. Comparing a query binding site

against the entire structure of a target protein is more challenging because it has a higher chance to generate false positive results than matching two binding sites of a comparable size. In these benchmarking calculations, using G-LoSA yields an AUC of 0.86, whereas the AUC for APoc is 0.78. A recent study employing an independent dataset also shows that, in contrast to APoc, G-LoSA offers a fairly reliable performance against diverse types of local structures [35].

BSAlign matches functional sites in query and target proteins by employing a graph isomorphism algorithm to find MCSs [72]. Query binding sites are identified by selecting protein residues within 5 Å from a bound ligand in the experimental complex structure. Target proteins and query binding sites are represented as graphs encoding the geometrical and physicochemical properties of amino acids. Specifically, nodes are protein residues and edges connect those residues whose Cα atoms are within a distance of 15 Å. Furthermore, nodes are assigned a solvent accessibility, physicochemical properties, and the secondary structure information, whereas edges are ascribed the distance between Cα atoms and the angle between Cα–Cβ vectors of the connected residues. BSAlign calculates the similarity between two graphs, representing a pair of binding sites, as the size of the MCS, which is identified by finding the maximum clique in the edge product of two graphs [73]. Cliques are computed with Cliquer a branch-and-bound maximum clique detection algorithm [74] and then mapped back to the pairs of edges by the Hungarian assignment method [58]. In order to refine the resulting alignment, those residue pairs that are misaligned are iteratively removed [75]. The largest set of aligned residue pairs with the lowest RMSD is reported as the final alignment between binding sites. Since BSAlign scans the entire target protein to find substructures similar to the query binding site, neither known nor predicted target binding site information is required. BSAlign is reported to be 14 times faster than SiteEngine in searching a dataset of 126 target proteins with a query ATP-binding site.

Group IV: methods employing geometric hashing and sorting

Geometric hashing is another technique often used as the principal constituent of programs to match ligand-binding sites in proteins. This concept was originally developed in computer vision to rapidly identify sets of geometric attributes across large databases [76]. Geometric hashing is an indexing-based approach extracting geometric patterns from objects, which are then stored in a hash table. Similar objects can be identified simply by searching for occurrences of the same patterns. Major advantages of geometric hashing include partial matching, the recognition of objects that have undergone transformations, a high computational efficiency, and a low time complexity. A conceptually related technique is geometric sorting illustrated in Figure 3E. This method employs a list of all residue pairs between two binding sites sorted in a descending order by the weight assigned with a scoring function. Here, the alignment is constructed by iteratively adding residue pairs starting from the top and excluding those positions that are already paired, such as G125 in the second pair because it is paired with D387 based on the top-ranked pair. Note that using this heuristic approach significantly reduces the complexity by avoiding a costly backtracking algorithm. In addition to PocketAlign [30], PocketFEATURE [77] and SiteEngine [69] assigned to this group, geometric hashing and sorting are employed by many other programs to find ligand-binding sites on protein surfaces [78], quantify the similarity between binding pockets [68] and classify functional sites in proteins [79].

PocketAlign defines ligand-binding sites as sets of protein residues within 4 Å from any ligand atom [30]. The algorithm encodes shape descriptors in the form of geometric perspectives (GPs), initially employing Cα atoms, but more detailed representations of proteins are used as well. Specifically, four different schemes are utilized to represent pocket residues: backbone atoms (N, Cα, C, O) and the centroid of the side-chain (CNTR), backbone atoms only, side-chain atom (Cβ) and CNTR and CNTR only. The GP for a given binding residue is a list of one-against-all distances sorted in a descending order. A pair of GPs is compared by counting the number of common distance elements, defined as those values differing by less than 0.5 Å. For a pair of binding sites, all-against-all comparison of the corresponding GPs is conducted in order to compute the geometric perspective score (GPS) matrix. Furthermore, the scoring function also includes the BLOSUM 62 amino acid substitution matrix [80] to account for the sequence similarity between two binding sites.

PocketAlign requires the alignment process to be initiated with seed mappings generated by sorting the GPS scores while preserving the corresponding pair information. The alignment construction starts with the highest score incrementally adding additional pairs to define a frame. Since multiple combinations of pairs can be selected for the single alignment, a frame may have many mappings. Once a new pair is added to a mapping, the RMSD value is calculated by the least square superposition according to the Kabsch algorithm [81]. The second highest GPS element serves as the starting point for another frame and so on. This greedy approach continues for the entire set to incrementally combine residue pairings avoiding the costly backtracking algorithm. Candidate alignments are evaluated by the Q scoring function developed to obtain the maximum number of matched pairs leading to a minimum RMSD. The entire process is repeated for four representation schemes and the best mapping with the highest Q value among all frames and schemes is selected as the final binding site alignment. Additionally, PocketAlign conveniently outputs PyMOL scripts to facilitate the visual inspection of the constructed alignments.

PocketFEATURE measures pocket similarity by comparing microenvironments characterized using the FEATURE system [77]. FEATURE microenvironments are centered around 22 predefined functional atoms of a residue, e.g. the gamma carbon of Asp, and are described by 80 physicochemical properties and 6 concentric spherical shells. Given a microenvironment pair that belongs to the same functional center group, e.g. aromatic, positive-charged, etc., a normalized Tanimoto similarity coefficient (TC), referred to as the microenvironment similarity score, is calculated. Microenvironments between two binding sites are aligned if their TC resides under a certain threshold. The sum of all TC values for aligned microenvironment pairs represents the overall similarity between a pair of binding sites and is termed the binding site similarity score. PocketFEATURE has little reliance on geometric properties of the ligand or the pocket because it does not impose rigid geometric matching criteria on the microenvironments. In fact, the only geometric requirement is that the matching microenvironments be present within the pocket of interest.

Two benchmark sets were employed to assess the accuracy of PocketFEATURE. The first dataset is the SOIPPA set of 247 sites from non-redundant protein structures known to bind an adenine-containing ligand and a control set of 101 cavities

from non-redundant protein structures believed not to bind an adenine-containing moiety [40]. The AUC for PocketFEATURE against this dataset is 0.85. Furthermore, it was demonstrated to outperform other algorithms to match adenine-binding sites. For instance, at a specificity of 95%, the sensitivities for PocketFEATURE and SOIPPA are 40% and less than 30%, respectively. The second dataset comprises 6958 druggable binding sites derived from sc-PDB [82], including 249 flavin adenine dinucleotide (FAD)-binding proteins and 6709 non-FAD-binding proteins. Here, the AUC for PocketFEATURE is as high as 0.85 with nearly 65% of FAD-binding sites correctly identified at 95% specificity. FAD-binding proteins, which bind to either butterfly or elongated confirmation of FAD, are a particular example of the geometric independence of PocketFEATURE. When querying for binding sites similar to those binding the butterfly conformation of FAD, PocketFEATURE correctly recognizes binding sites in proteins complexed with the elongated conformation, and vice versa. These results demonstrate that PocketFEATURE allows for the flexibility of both ligand and pocket geometries.

SiteEngine is a surface-based algorithm developed to identify similar functional sites on the surface of proteins having no sequence or fold similarities [69]. This method employs a new, low-resolution surface representation with chemically important surface points. Specifically, each amino acid in a protein structure is described by a physicochemical pseudo-center representing a certain interaction type, such as the hydrogen-bond donor, hydrogen-bond acceptor, mixed donor/acceptor, hydrophobic aliphatic and aromatic contacts. The solvent accessible surface of a protein is rendered by the Connolly smooth molecular surface algorithm [83]. The overlap between a pair of protein surfaces is quantified by a heuristic algorithm based on computationally efficient geometric hashing and the matching of triangles of physicochemical pseudo-centers with hierarchical scoring schemes. SiteEngine employs the Match score as a scoring function. When a query binding site is compared to the target binding site, the score of the best solution is normalized by that obtained from matching the query to itself. Since all features in the query–query self-pair match, this score represents the maximal possible match, and the query–target pair will never exceed that score. Consequently, the Match score adopts values in the 0–100% range.

The performance of SiteEngine was evaluated in three types of applications. The first scenario considering searching the database of complete protein structures with a binding site is illustrated by two searches conducted with the estradiol-binding site of the sex hormone binding globulin and with the adenine-binding site of the cAMP-dependent protein kinase. Encouragingly, in both cases, the highest-ranking solutions contained unrelated proteins that perform the same function as the query site. For instance, searching the ASTRAL database of protein structures [84] with the estradiol-binding site identifies estrogen sulfotransferase and tropinone reductase. The adenine-binding site is matched by SiteEngine not only to a number of kinases with the same fold as the query, but also to those proteins having different global structures, such as replication factor C and D-Ala-D-Ala ligase. As an additional test, SiteEngine was employed to classify binding patterns for serine proteases. Here, the similarity of the corresponding functional groups created by catalytic triads was correctly recognized and meaningful binding site alignments were constructed.

The second application type is the searching of a database of binding sites with a binding site. These calculations are more focused because only those regions known to function as binding sites are considered. As an example, SiteEngine was employed to infer the function of hypothetical proteins MJ0577 and MJ0226. The ATP-binding site from MJ0577 is correctly matched to the AMP-binding sites of electron transfer flavoprotein subunits belonging to the same SCOP superfamily [85] as the query protein, and the ADP-binding site of arsenite-translocating ATPase ArsA having a different fold than the query. Further, SiteEngine correctly aligned the ANP-binding site of the hypothetical protein MJ0226 to binding sites in autoinducer-2 production protein LuxS, a tandem phosphatase domain of RPTP LAR, and adenylate kinase complexed with the substrate-mimicking inhibitor Ap5A.

The third application type considers searching the database of binding sites with a complete protein structure. Since these calculations are generally less focused and may lead to alignments involving regions without functional significance, it is recommended to first identify potential binding pockets with cavity detection methods and then conduct a more focused search. These examples demonstrate that SiteEngine is a valuable tool for a wide range of applications, from the identification of secondary binding sites of drugs that may lead to side effects to the recognition of molecular functions of hypothetical proteins obtained from structural genomics projects.

Group V: methods employing rotational and translational search

The last group includes two programs, PARIS [28] and SiteAlign [86]. These methods treat binding pockets as rigid objects either represented by non-hydrogen atoms (Figure 1A) or projected onto a sphere. A rotational and translational search is performed in order to maximize the overlap between a pair of objects according to a scoring system. Because a systematic search over all rotational and translational degrees of freedom in the 3D space is prohibitively expensive, heuristics are typically employed. Common techniques to reduce the wall time required to align a pair of binding sites include conducting a discrete, low-resolution search and limiting initial rotations to the alignment of the principal axes. Subsequently, the conformational space can be explored around these initial solutions in order to define the best pocket superposition.

PARIS represents a pocket as a 3D cloud of points with fixed relative positions, where each point is an atom bearing labels [28]. These labels have customizable degrees of importance, i.e. weights, which can be either discrete or continuous and represent chemical and structural properties of that atom. Examples of labels are the partial charge, parent residue, hydrogen donor and acceptor, hydrophobicity, hydrophilicity, and secondary structure. PARIS aligns binding sites according to similarities between atom densities in 3D space, rather than any pairwise matching of atoms or residues. The similarity measure for the clouds of points employs a Gaussian kernel with the radial basis function representing vector similarity. With one pocket fixed in place, the optimal translation and rotation of the other pocket are found using a gradient ascent algorithm starting from multiple initial configurations. The optimized relative position of the two pockets results in a similarity score, referred to as the sup-CK, and the superposition is made according to the maximum sup-CK.

PARIS was benchmarked against three datasets. The Kahraman dataset comprises 100 protein crystal structures in complex with 1 of 10 ligands [54]. Additional protein structures complexed with one of the 10 ligands were then added to

prepare the extended Kahraman dataset consisting of the total of 972 structures. These new proteins were filtered at a pairwise sequence identity of 30% in order to avoid any bias by including close homologs. Because these Kahraman datasets comprise ligands of different sizes and chemical properties, the Homogeneous dataset was compiled to include 100 structures of proteins in complex with 10 ligands of a similar size. Using this dataset allows to evaluate the performance of binding site matching algorithms against those pockets binding ligands of similar sizes. Sup-CK achieves a remarkable performance against the Kahraman dataset with an AUC of 0.86. For comparison, a method based on the spherical harmonic decomposition [54] has an AUC of 0.77, whereas MultiBind [51] has an AUC of 0.71. Further, the performance of sup-CK does not deteriorate when the extended Kahraman dataset including 872 new pockets is employed, consistently remaining above those of other methods. Although the Homogeneous dataset is more challenging compared to the Kahraman datasets, sup-CK still outperforms other algorithms when pockets binding ligands of similar sizes are used, for instance, the AUC is 0.71 for sup-CK and 0.69 for MultiBind.

Since PARIS defines pockets as 3D clouds of points with parameterizable weights, the similarity measure can set the degree of importance of a particular label. Therefore, it may be sequence order-independent if the residue label is given no weight. In fact, PARIS could be used to assess the similarity between any clouds of points, e.g. molecules, whole proteins, etc., in addition to protein binding sites. This feature makes PARIS a great tool to expose function similarities even in the absence of any detectable global sequence and structure similarity between proteins.

SiteAlign is a fingerprint-based method to measure distances between druggable protein cavities [86]. It maps binding site properties onto a discretized sphere placed at the center of the ligand binding site. First, a 1 Å radius sphere is placed at the center of mass of Cα atoms of cavity-lining residues, defined as those amino acids within a distance of 6.5 Å from any ligand heavy atom. The sphere is then discretized into a geodesic polyhedron having 80 triangular faces, which approximate the shape of a sphere. Subsequently, a geometric vector from the Cα atoms of each pocket residue to the cavity center of mass is constructed in order to project binding residues onto the polyhedron faces. In the case of multiple vectors passing through a single face, a residue maximizing the similarity score between two binding sites is chosen for that face. Finally, the faces are assigned eight descriptors, three of which are topological and five are chemical. Topological descriptors include a 'fuzzy' distance from the Cβ atom to the pocket center of mass, the side-chain orientation, which can be either inward or outward with respect to the pocket center, and a 'fuzzy' size of the side-chain. Here, the term 'fuzzy' refers to using bins rather than exact values. For example, the topological descriptor for a distance is discretized into a series of 30 bins, each of 0.5 Å length. Because such an approach allows for certain flexibility in the binding site representation accounting for the inexactness of 3D structures, SiteAlign can utilize pockets identified in low-quality experimental structures and computer-generated protein models.

Five chemical descriptors consist of the hydrogen bond donor and acceptor counts, the aliphatic and aromatic character and the type of charge. The final representation of a binding site is a sphere with 80 faces, each associated with a feature vector of eight descriptors corresponding to the residue passing through that face; the feature vector is a zero vector if no residues pass through. The similarity between two binding sites is assessed with a normalized sum of the Manhattan similarity between vectors. Specifically, 1 minus the normalized Manhattan distance gives the similarity between two triangles and the overall similarity score between binding sites is computed as the normalized sum of all pairwise scores. In addition, SiteAlign constructs the structural alignment of two pockets by systematically rotating and translating one sphere around the other in order to maximize the overall similarity score.

The performance of SiteAlign was evaluated against a dataset of 376 pairs of related binding sites compiled from the sc-PDB repository of druggable binding sites [82]. Here, only enzymes co-crystallized with a drug-like ligand were employed and pairs of matching sites were defined based on the Enzyme Commission (EC) numbers [87]. Further, selecting only the lowest and the highest molecular weight ligands for a given EC number ensured the sufficient diversity of binding sites. Encouragingly, as many as 75% of alignments constructed by SiteAlign are true positives with a good score and a correct alignment, whereas only 1.6% are false positives with a good score but an incorrect alignment. The remaining cases correspond to true negatives having a few residues in common and accommodating different ligands at different binding modes.

In addition, the capability of SiteAlign to compute the cross-similarity of members of a protein family was tested against serine proteases. This particular family was selected because of a large number of protein-inhibitor structures available, the diversity of folds and functionally important characteristics of catalytic sites. A binding site for a bovine trypsin inhibitor was compared against 6415 ligand-binding sites in the sc-PDB dataset including 357 serine proteases. A vast majority of similar binding sites according to SiteAlign come from serine endopeptidases. Furthermore, not only proteases with the same fold and the substrate cleavage specificity were detected, but also those having different folds and cleavage preferences. Finally, SiteAlign was also applied to predict binding sites for ligands with different promiscuity levels. The sc-PDB dataset was queried with binding sites for 4-hydroxytamoxifen (4-OHT) in three primary targets, estrogen receptor (ER) subtypes α and β [88] and estrogen-related receptor (ERR) γ subtype [89]. Not only binding sites for 4-OHT in ER and ERR structures were correctly identified, but also in other proteins such as p38 MAP kinase, a known target of 4-OHT [90]. Overall, SiteAlign was demonstrated to efficiently detect the local similarity among active sites, discriminate between protein subfamilies and identify targets of promiscuous drugs.

Applications

Programs to match binding sites in proteins have a number of applications. In this communication, we focus on three applications of particular importance to modern drug development schematically shown in Figure 4, drug repurposing (Figure 4A), polypharmacology (Figure 4B) and the analysis of side effects (Figure 4C).

Drug repurposing

Drug repurposing, or repositioning, is an effort to extend a purpose of a drug beyond its intended target(s) by identifying an additional set of targets, namely, off-targets. Repurposing approved drugs significantly reduces research costs, accelerates the development and improves success rates by leveraging the existing toxicity, efficacy and safety data. A therapeutic benefit of drug repositioning is exemplified by mesocarb and tolimidone.

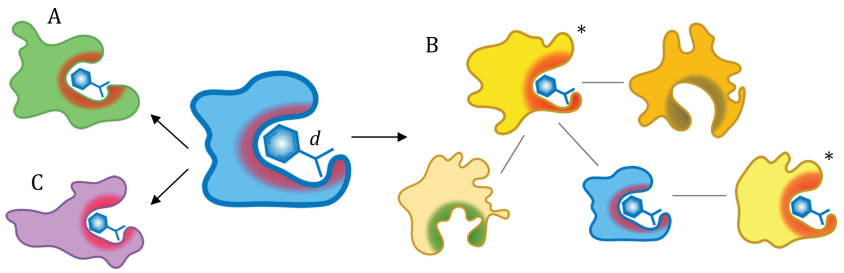

Figure 4. Diagram of selected applications of pocket matching algorithms. A blue structure in the center is the primary target for drug *d*. Binding sites in other proteins that are similar to the *d*-binding pocket in the primary target are colored red. **(A)** Repurposing of *d* to another protein colored green implicated in a different disease than the primary target is associated with. **(B)** An illustration of the concept of pocket-based polypharmacology. If the primary target is part of a disease-related pathway involving other proteins with similar binding sites, marked with asterisks, then *d* or its derivatives are candidates for the polypharmacological action on that pathway. **(C)** An analysis of drug side effects caused by off-target binding. A purple protein identified to have a similar pocket to that in the blue structure is a potential off-target for *d*.

The former was developed in the former Soviet Union in the 1970s as a psychomotor stimulant for the treatment of several neurological conditions including attention deficit disorder and alcoholism [91]. Later, mesocarb was identified as an inhibitor of motor deficits in a mouse model of Parkinson's disease [92], and it is currently considered a candidate for repurposing to treat this condition. Another instance is tolimidone, a compound originally developed for gastric ulcer that was terminated because of the limited efficacy in phase II clinical studies. However, it was found to lower blood glucose levels with a glycemic control response better than that of approved anti-diabetic drugs [93]. Encouragingly, tolimidone shows positive clinical results in a phase II trial for type II diabetes. Furthermore, delamanid, an approved drug for multi-drug resistant tuberculosis, was found to be a potent inhibitor of Leishmania donovani both *in vitro* and *in vivo*, suggesting its alternate use as a much-needed oral therapy for visceral leishmaniasis [94]. Notwithstanding these successful cases, drug repurposing often relies on a successful development of appropriate screening assays for various diseases and may not provide a detailed molecular basis of the drug mechanism of action.

A clear necessity for a rational approach to find alternative indications for existing therapeutics has stimulated the development of computational methods for drug repositioning [95]. Current algorithms fall into two distinct categories, disease- and drug-based techniques. The former methods target particular conditions, e.g. cancer, genetic and infectious diseases, utilizing heterogeneous data ranging from genome to phenome in order to identify repurposable drugs [96]. Network analysis [97], gene-expression signatures [98], literature mining [99] and genome-wide association studies [100] are frequently employed in disease-based drug repositioning. Contrastingly, drug-based techniques initiate the discovery from the chemical perspective, considering similarities among drug molecules [101], target binding sites [102] and side effects [103]. Ample resources available to drug developers, such as PubChem [104], ChEMBL [105], DrugBank [106] and BindingDB [107], provide comprehensive biological information on the drug chemical structure, physicochemical properties, binding affinity, molecular and cellular activity, as well as absorption, distribution, metabolism, excretion and toxicity profiles. These data are routinely employed to compare drugs by quantifying the similarity of their chemical structures, mechanisms of action and side effects.

Various computational repositioning strategies have been used with appreciable success in the past years. Nonetheless, those methods employing chemical profiles as well as indirect features such as pathway information, side-effect similarities and social media [108] do not exploit binding site similarities, which appear to be a more important component to drug promiscuity than ligand properties alone [109]. Repositioning based on the binding-site similarity, or pocket-based repositioning, relies on the idea that a drug, and perhaps chemically similar derivatives, may bind to a site that is similar to its intended target. This concept is illustrated in Figure 4A. The blue protein shown in the center is the primary target for a drug *d*, and the green structure, which may be implicated in a different disease state, has a similar binding site (highlighted in red). Should this be the case, the drug *d* can potentially be repositioned to treat another condition by targeting the green protein.

The process of pocket-based repositioning can be broken into three phases, pocket superposition, ligand repositioning, and pocket-ligand refinement. First, the pockets to be superposed are selected from a set of well-matched pocket pairs generated by a pocket matching algorithm. For example, pockets in vantenib-bound protein-tyrosine kinase 6 (PTK6) and GTPase KRas (KRAS) were paired with a high matching score and then superposed according to their local alignment [110]. Vantenib originally bound to the PTK6 pocket, now occurs inside the KRAS pocket; namely, the PTK6 ligand has been repositioned to the KRAS pocket. The geometry of the repositioned ligand might not be optimal since the orientation of vantenib has maintained the binding specifics from the PTK6 pocket. On that account, the new ligand-pocket complex, vantenib-bound KRAS, is further refined in order to optimize molecular interactions with the new target site. A major advantage to this structure-based procedure is that a proposed conformation of the ligand within the binding site is provided, leading the way for drug structure optimization.

Polypharmacology

Drugs are often developed to interfere with one cognate target, but often these drugs have many unintended non-canonical interactions. This drug promiscuity makes it difficult to develop therapies for diseases with complex pathogenic pathways because drug cocktails frequently increase the risk of dosage problems, inhibiting drug–drug interactions and toxicity [5, 111]. Drug discovery and synthesis is also an incredibly expensive and labor-intensive process that often yields drugs that can have problematic side effects [112]. To combat these problems, and to take advantage of the drug promiscuity, an emerging paradigm of drug development, polypharmacology, has been gaining traction in recent years. Instead of developing drugs with one specific target in mind, polypharmacology takes the entire set of possible drug–protein interactions into account to be

utilized against intricate disease networks [1]. Polypharmacology certainly presents its own set of challenges. The set of all possible drug interactions can be difficult to ascertain through *in vivo* and *in vitro* methods rendering the discovery and optimization of drugs to multiple targets incredibly time and resource exhaustive [111, 113].

In recent years, kinase inhibitors have been approved by the US Food and Drug Administration for cancers because of their anti-tumor activity. Although the detailed mechanisms of some of these drugs have not been fully elucidated, their modes of action are likely propagated by polypharmacological phenomena. For instance, a systems biological approach revealed that anaplastic lymphoma kinase inhibitor, ceritinib, has multiple non-canonical targets related to anaplastic lung lymphoma. Further, paclitaxel was identified to synergize with ceritinib to attack lung cancer targets such as auto-phosphorylating focal adhesion kinase [111]. Another example is a new class of drug compounds called di-2-pyridylketone thiosemicarbozones (DpTs) binding to copper and iron ions and impeding three specific areas in cancer progression, metastasis, tumor growth and drug resistance [114, 115]. DpTs have a unique and intriguing polypharmacological profile [116]. To interfere with metastasis, DpTs upregulate *N*-myc downstream-regulated gene 1 protein shown to inhibit cell signaling pathways for cellular locomotion and proliferation. These drugs also interfere with the tumor growth through reducing the DNA production and arresting the cell cycle. Finally, DpTs are able to overcome drug resistance mechanisms in cancer by inducing autophagy in cells expressing high levels of *P*-glycoprotein.

In the area of herbal medicine, systems biology approaches can be employed to elucidate the mechanism of action for specific herbal treatments. Indeed, traditional medicine has its roots in the polypharmacological paradigm and the promiscuity of natural products [117]; however, much of what is known is based on empirical evidence. Since there are many bioactive chemicals at work in herbal mixtures, the entire set of possible reactions needs to be accounted for to determine their mechanism of action [118]. Further, the analysis of sedative hypnotic effects also lends themselves to the polypharmacological paradigm. The 'one target, one drug' philosophy works poorly in complex systems such as the central nervous system (CNS) and, more specifically, in the study of sleeping and sedative medication, where bioactivity against multiple targets holds significant relevance. A number of traditional sedatives like benzodiazepines often have side effects of addiction and residual drowsiness, while other sedatives like melatonin receptor agonists often prove less effective than its contemporaries. Systems analysis of the mechanism of sleep was recently employed to find novel, more efficacious drugs for sleep disorders [119].

Computational modeling and pocket matching algorithms have taken a center stage to expand the space of the rational design of therapeutic agents [1]. A specific way to identify polypharmacological phenomena is to compare the 3D structures of drug-binding pockets in proteins. This approach is schematically presented in Figure 4B. Here, the blue structure shown in the center is part of a disease-related pathway or a sub-network containing four other yellowish proteins. Two of these targets, marked by asterisks, have similar binding sites to that in the blue structure (highlighted in red), opening up a possibility to developed polypharmacological agents capable to simultaneously target all three proteins. It is important to note that since binding pockets are subjected to higher evolutionary pressures and thus they tend to be more conserved across proteomes, pocket matching may reveal novel relationships between proteins, their molecular function, and ligand-binding profiles. Indeed, optimal conditions were observed for multi-target combinations greatly expanding the space of opportunities for the rational design of drug polypharmacology [5].

Analysis of side effects

Drug side effects are the collection of unintended physiological activities in response to the correct administration of a drug. Although these activities may alleviate separate conditions, as in drug repurposing, undesirable effects are usually referred to as side effects or adverse effects. Predicting drug side effects is a complicated task because the possible drug interaction network could be vast and idiosyncratic [120]. More specifically, side effects may arise from drug–drug interactions, off-target binding, metabolic pathway inference and unknown sources. Approaches predicting side effects often leverage chemical similarity as well as the target protein and pathway information. In a chemical-similarity approach, the more chemically similar two drugs are, the more likely they share the same side effects [121]. This technique requires side effects to be known for at least one member of a group of chemically similar compounds. Further, a target-based approach correlates the target pathway and protein information with side effects [122].

A pocket-based prediction of side effect utilizes pocket matching algorithms to construct a potential off-target network for a particular target. For instance, Figure 4C presents a purple protein with a binding site similar to that in the blue protein shown in the center. This information on a potential off-target can be used to elucidate the side effects of those drugs targeting the blue protein. As an example of a pocket-based inference of drug side effects, SMAP was employed to construct an off-target network for fatty acid amide hydrolase (FAAH) [123]. FAAH is an integral membrane enzyme and a potential therapeutic target for the treatment of pain and CNS disorders. Unfortunately, the development of FAAH inhibitors has been complicated by the unintended binding of these drugs to off-targets and has even resulted in the death of a patient during clinical trials for the experimental drug BIA 10-2474. The binding site in FAAH was found to be similar to a pocket at the dimer interface of an *N*-methyl-D-aspartate (NMDA) receptor. After the computational verification of FAAH and NMDA binding site similarity, the BIA 10-2474-NMDA complex analysis together with the phenomic examination of NMDA receptor function strongly suggested that the NMDA receptor is an off-target for BIA 10-2474. This case study shows that pocket matching algorithms can effectively assess the possible mechanisms of drug side effects. This information, even if not always precise, reduces the scale of side-effect investigations from entire proteomes down to small clusters of similar pockets, which is invaluable to the drug development process.

Future directions

Considering the importance of ligand-binding site alignments to modern drug development, a dynamic growth of this field of research can certainly be expected. We anticipate that several types of methods will become available in the near future, meta-predictors, approaches considering pocket dynamics and those employing deep learning. Below, we discuss successful applications of these techniques in bioinformatics, highlighting their remarkable potential to advance binding pocket matching.

Meta-predictors

Meta-predictors operate by considering outputs from a variety of individual algorithms under the assumption that combined predictions are more accurate than those produced by a single method. For instance, protein meta-threading is a widely employed technique to identify suitable templates for the prediction of protein tertiary structures and spatial constraints. Consensus models generated by LOMETS [124] from the top predictions of nine component-threading algorithms are at least 7% more accurate than the best individual methods. Similarly, *e*Thread integrates 10 state-of-the-art threading/fold recognition algorithms and extensively uses various machine learning techniques to carry out fully automated template-based protein structure modeling [125]. @TOME-2 is another method employing meta-threading to detect template proteins, which are then used in structure modeling [126]. Since meta-threading effectively detects many facets of protein molecular function, it has been successfully used to select template proteins to model interactions between proteins and ligands, metal ions, inorganic clusters, other proteins and nucleic acids [127].

Another group of meta-techniques in structural bioinformatics relevant to the topic of this review employ multiple algorithms to detect ligand-binding sites. An example is metaPocket combining four methods, LIGSITEcsc [24], PASS [128], Q-SiteFinder [22] and SURFNET [19], to improve the success rate of pocket prediction [129]. Indeed, benchmarking calculations conducted against both bound and unbound structures demonstrate that metaPocket improves the success rate from approximately 70–75% at the top-ranked prediction. Further, COACH is a consensus approach selecting ligand-binding templates from the BioLiP database [130] with two comparative methods, TM-SITE and S-SITE [131]. Subsequently, ligand-binding predictions are combined with those obtained from other state-of-the-art tools, COFACTOR [132], FINDSITE [65] and ConCavity [25]. By integrating multiple programs, COACH increases the accuracy of binding residue prediction by 15% over the best individual methods.

Recently, proof-of-concept of a meta-approach to compare binding pockets has been reported [35]. Here, direct methods to align binding sites are combined with an indirect technique to quantify the pocket similarity with structure-based virtual screening. It is important to note that the indirect comparison of pockets by means of virtual screening is methodologically orthogonal to direct techniques employing local binding site alignments. Encouragingly, integrating the results from alignment-based tools, APoc [62], G-LoSA [71] and SiteEngine [69], with those obtained by two molecular docking programs, AutoDock Vina [133] and rDock [134], into a meta-predictor improves the performance of existing methods to detect similar binding sites in unrelated proteins by 5–10%. These results provide a solid rationale for including structure-based virtual screening as part of protocols detecting similar ligand-binding sites in unrelated proteins [135].

Pocket dynamics

Proteins are highly dynamic systems often displaying a significant conformational heterogeneity [136, 137]. Particularly, the plasticity of binding sites is pivotal for the interaction specificity of many proteins. Pocket dynamics range from relatively small fluctuations of binding residue side chains to the appearance/disappearance of sub-pockets, or even the formation of completely new pockets [138]. A special case is a class of intrinsically disordered proteins, i.e. biologically active molecules without stable structure, which are important drug targets because of their involvement in numerous human diseases [139]. These phenomena create the necessity to consider protein internal motion and intrinsic disorder in matching ligand-binding pockets. Although taking full account of the structural flexibility renders a number of practical challenges in molecular modeling [140], protein dynamics can successfully be exploited to discover new bioactive molecules [141], as well as analyze the drug-target complementarity [142] and binding selectivity [143]. We expect that important future developments in binding site matching will include new computational tools capable of accounting for the conformational flexibility of proteins.

Deep learning

Although machine learning has been used to compare ligand-binding sites in proteins [32, 61], applications employing deep learning are yet to be developed. Deep learning is attracting a significant attention due to a number of successful applications in image processing [144], speech recognition [145], natural language research [146], decision-making [147] and even self-driving vehicles [148]. Deep learning algorithms are essentially biologically inspired networks mimicking neural connections and learning process in the human brain, combined with advanced machine learning techniques. In a nutshell, large amount of data is fed into a deep learning framework, in which highly complicated training models are then generated, trained and evaluated. Properly trained models have capabilities to make highly accurate decisions even for previously unseen input. A number of deep learning architectures are currently available, including Generative Adversarial Networks [7], AlexNet [144], ZF Net [149] and GoogleNet [150].

Not surprisingly, deep learning approaches hold a great promise for applications in biology and biomedicine. For instance, a convolutional neural network was employed to study the regulatory code of accessible genome [151]. Trained to learn the functional activity of DNA sequences based on genomic data from 164 cell types, this model achieves a higher prediction accuracy than previous methods. Another example is DL-Pro, a deep learning-based classifier to assess the quality of computer-generated protein models [19]. DL-Pro was demonstrated to outperform state-of-the-art scoring functions, DOPE [152], DFIRE [153] and OPUS-Ca [154], on targets selected from the Critical Assessment of protein Structure Prediction [155]. The first application of a deep learning algorithm, the stacked autoencoder [156], to predict protein–protein interactions has been recently reported [157]. Not only the best cross-validated model achieves a phenomenal accuracy of 97%, but also the performance of the autoencoder is superior to those by other methods against several external datasets.

Traditional machine learning requires a manual feature engineering to select task-specific handcrafted features that are subsequently used to train the model and classify the data. In contrast, deep learning employs a set of techniques to allow a learning system to automatically discover representations needed for the efficient classification from the raw data. Since extracting raw features over hand-crafted optimization is particularly beneficial for highly complex problems, we expect that novel approaches employing deep learning algorithms are going to outperform conventional methods to match ligand-binding site in proteins.

Conclusions

The study of similarity of ligand-binding sites across the protein structure space is of paramount importance to modern drug discovery, especially in the context of polypharmacology, drug repurposing, and the development of biopharmaceuticals with improved safety profiles. There has been a tremendous progress in the development of novel techniques to compare binding sites. Many approaches implement graph-based algorithms, such as clique and the maximum common subgraph detection, as well as techniques solving the assignment problem, such as the Hungarian and the shortest augmenting path algorithms. Other methods employ geometric hashing and sorting, as well as the rotational and translational search. Most of these approaches were devised to conduct sequence order-independent binding site matching, which is required to investigate pocket similarity across the protein fold space. Nonetheless, many tools require high-quality experimental structures of target proteins, limiting their usability in large-scale pharmacological applications.

Protein and pocket structures are represented by a variety of models ranging from fine-grained systems considering individual atoms, through various types of sub-residual functional groups and interaction points, to coarse-grained systems representing individual residues by a single effective point at their Cα atoms or the center of mass. Other commonly used representations include interaction fields, molecular surface, and projections on geometrical objects. Scoring functions often employ potentials widely used to model drug–protein interactions, such as hydrogen bond donors and acceptors, positively and negatively charged groups, aromatic contacts and hydrophobic and hydrophilic attributes. In addition, some methods incorporate evolutionary terms including sequence and secondary structure profiles, and amino acid substitution matrices, as well as purely geometrical features, e.g. physical Cα–Cα distances and angles between Cα–Cβ vectors of binding site residues.

A variety of datasets are available to benchmark the performance of binding site matching methods. Pairs of similar and dissimilar pockets are typically defined based on the similarity of binding ligands, binding environments and molecular function performed by the target proteins. Focused datasets were compiled to evaluate the identification of particular types of pockets, e.g. adenine-, steroid-, calcium- and FAD-binding. Other datasets comprise several different ligands having certain properties, for example, all ligands in the homogeneous set are of a similar size. Finally, large and representative collections of complex structures, such as PDBbind, sc-PDB, and TOUGH-M1, contain a variety of different ligands binding to structurally diverse protein targets. These ample resources can be utilized to thoroughly evaluate the performance of existing binding site matching programs and assist in the development of novel algorithms.

Considering the significance of ligand-binding site comparison in rational drug discovery, we anticipate a continuous growth of this research area with foreseeable advances in algorithm development. Particularly, meta-predictors combining orthogonal methodologies, deep learning-based methods, and techniques considering pocket dynamics will likely become available in the near future.

Key Points

- Numerous techniques comparing binding sites are available to support rational drug design.
- Important applications of computational pocket matching are drug repurposing, polypharmacology, and the analysis of side effects.
- A variety of datasets are available to benchmark the performance of binding site matching.
- Meta-predictors, techniques incorporating protein flexibility and deep learning methods will likely become available in the near future.

Funding

National Institute of General Medical Sciences of the National Institutes of Health (R35GM119524).

References

1. Reddy AS, Zhang S. Polypharmacology: drug discovery for the future. *Expert Rev Clin Pharmacol* 2013;**6**:41–7.
2. Peng X, Wang F, Li L, *et al*. Exploring a structural protein-drug interactome for new therapeutics in lung cancer. *Mol Biosyst* 2014;**10**:581–91.
3. Geerts H, Hofmann-Apitius M, Anastasio TJ, *et al*. Knowledge-driven computational modeling in Alzheimer's disease research: current state and future trends. *Alzheimers Dement* 2017;**13**:1292–302.
4. Cho DY, Kim YA, Przytycka TM. Chapter 5: network biology approach to complex diseases. *PLoS Comput Biol* 2012;**8**: e1002820.
5. Duran-Frigola M, Siragusa L, Ruppin E, *et al*. Detecting similar binding pockets to enable systems polypharmacology. *PLoS Comput Biol* 2017;**13**:e1005522.
6. Coleman RG, Sharp KA. Protein pockets: inventory, shape, and comparison. *J Chem Inf Model* 2010;**50**:589–603.
7. Medina-Franco JL, Giulianotti MA, Welmaker GS, *et al*. Shifting from the single to the multitarget paradigm in drug discovery. *Drug Discov Today* 2013;**18**:495–501.
8. Mestres J, Gregori-Puigjane E, Valverde S, *et al*. Data completeness—the Achilles heel of drug-target networks. *Nat Biotechnol* 2008;**26**:983–4.
9. Wang C, Hu G, Wang K, *et al*. PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinformatics* 2016;**32**:579–586.
10. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 1996;**93**:13–20.
11. Weisel M, Proschak E, Kriegl JM, *et al*. Form follows function: shape analysis of protein cavities for receptor-based drug design. *Proteomics* 2009;**9**:451–9.
12. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* 1998;**7**:1884–97.
13. Desaphy J, Raimbaud E, Ducrot P, *et al*. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J Chem Inf Model* 2013;**53**:623–37.
14. Gao ZG, Chen A, Barak D, *et al*. Identification by site-directed mutagenesis of residues involved in ligand recognition and activation of the human A3 adenosine receptor. *J Biol Chem* 2002;**277**:19056–63.
15. Shuker SB, Hajduk PJ, Meadows RP, *et al*. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 1996;**274**:1531–4.

16. Boland A, Chang L, Barford D. The potential of cryo-electron microscopy for structure-based drug design. *Essays Biochem* 2017;**61**:543–60.
17. Leis S, Schneider S, Zacharias M. In silico prediction of binding sites on proteins. *Curr Med Chem* 2010;**17**:1550–62.
18. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 1997;**15**:359–63 389.
19. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 1995;**13**:323–30, 307-328.
20. Binkowski TA, Naghibzadeh S, Liang J. CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res* 2003;**31**:3352–5.
21. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 2009;**10**:168.
22. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 2005;**21**:1908–16.
23. Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res* 2009;**37**:W413–6.
24. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* 2006;**6**:19.
25. Capra JA, Laskowski RA, Thornton JM, *et al*. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol* 2009;**5**: e1000585.
26. Brylinski M, Feinstein WP. eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *J Comput Aided Mol Des* 2013;**27**:551–67.
27. Takimura T, Kamata K, Fukasawa K, *et al*. Structures of the PKC-iota kinase domain in its ATP-bound and apo forms reveal defined structures of residues 533-551 in the C-terminal tail and their roles in ATP binding. *Acta Crystallogr D Biol Crystallogr* 2010;**66**:577–83.
28. Hoffmann B, Zaslavskiy M, Vert JP, *et al*. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics* 2010;**11**:99.
29. Wei L, Altman RB, Chang JT. Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences. *Pac Symp Biocomput* 1997; 465–76.
30. Yeturu K, Chandra N. PocketAlign a novel algorithm for aligning binding sites in protein structures. *J Chem Inf Model* 2011;**51**:1725–36.
31. Ren J, Xie L, Li WW, *et al*. SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. *Nucleic Acids Res* 2010;**38**:W441–4.
32. Brylinski M. eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models. *PLoS Comput Biol* 2014;**10**:e1003829.
33. Thoden JB, Holden HM, Firestine SM. Structural analysis of the active site geometry of N5-carboxyaminoimidazole ribonucleotide synthetase from Escherichia coli. *Biochemistry* 2008;**47**:13346–53.
34. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;**57**:702–10.
35. Govindaraj RG, Brylinski M. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinformatics* 2018;**19**:91.
36. Köbler J, Schöning U, Torán J. Graph isomorphism is low for PP. *Comput Complexity* 1992;**2**:301–30.
37. Muzychuk M. A solution of the isomorphism problem for circulant graphs. *Proc London Math Soc* 2004;**88**:1–41.
38. Konc J, Janezic D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* 2010;**26**:1160–8.
39. Chartier M, Najmanovich R. Detection of binding site molecular interaction field similarities. *J Chem Inf Model* 2015;**55**:1600–15.
40. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc Natl Acad Sci USA* 2008;**105**:5441–6.
41. Berman HM, Battistuz T, Bhat TN, *et al*. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 2002;**58**:899–907.
42. Hubbard TJ, Murzin AG, Brenner SE, *et al*. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 1997;**25**:236–9.
43. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
44. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 1998;**11**:739–47.
45. Konc J, Hodošček M, Janežič D. Molecular surface walk. *Croat Chem Acta* 2006;**79**:237–41.
46. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* 2002;**323**:387–406.
47. Konc J, Janezic D. Protein-protein binding-sites prediction by protein surface structure conservation. *J Chem Inf Model* 2007;**47**:940–4.
48. Russell RB. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol* 1998;**279**:1211–27.
49. Holm L, Kaariainen S, Rosenstrom P, *et al*. Searching protein structure databases with DaliLite v.3. *Bioinformatics* 2008;**24**:2780–1.
50. Angaran S, Bock ME, Garutti C, *et al*. MolLoc: a web tool for the local structural alignment of molecular surfaces. *Nucleic Acids Res* 2009;**37**:W565–70.
51. Shulman-Peleg A, Shatsky M, Nussinov R, *et al*. MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res* 2008;**36**:W260–4.
52. Gaudreault F, Morency LP, Najmanovich RJ. NRGsuite: a PyMOL plugin to perform docking simulations in real time using FlexAID. *Bioinformatics* 2015;**31**:3856–8.
53. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. *Commun ACM* 1973;**16**:575–7.
54. Kahraman A, Morris RJ, Laskowski RA, *et al*. Shape variation in protein binding pockets and their ligands. *J Mol Biol* 2007;**368**:283–301.
55. Liu Z, Li Y, Han L, *et al*. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 2015;**31**:405–12.
56. Meslamani J, Rognan D, Kellenberger E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinformatics* 2011;**27**: 1324–6.

57. Easterfield TE. A combinatorial algorithm. *J London Math Soc* 1946;**21**:219–26.
58. Kuhn HW. The Hungarian method for the assignment problem. *Naval Res Logist Q* 1955;**2**:83–97.
59. Munkres J. Algorithms for the assignment and transportation problems. *J Soc Ind Appl Mathematics* 1957;**5**:32–8.
60. Burkard R, Dell'Amico M, Martello S. *Assignment Problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2009.
61. Brylinski M. Local alignment of ligand binding sites in proteins for polypharmacology and drug repositioning. *Methods Mol Biol* 1611;**2017**:109–22.
62. Gao M, Skolnick J. APoc: large-scale identification of similar protein pockets. *Bioinformatics* 2013;**29**:597–604.
63. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci USA* 2008;**105**:129–34.
64. Skolnick J, Brylinski M. FINDSITE: a combined evolution/structure-based approach to protein function prediction. *Brief Bioinform* 2009;**10**:378–91.
65. Brylinski M, Skolnick J. FINDSITE: a threading-based approach to ligand homology modeling. *PLoS Comput Biol* 2009;**5**:e1000405.
66. Feinstein WP, Brylinski M. eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models. *Mol Inform* 2014;**33**:135–50.
67. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:27.
68. Yeturu K, Chandra N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics* 2008;**9**:543.
69. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. *J Mol Biol* 2004;**339**:607–33.
70. Derigs U. The shortest augmenting path method for solving assignment problems - motivation and computational experience. *Ann Oper Res* 1985;**4**:57–102.
71. Lee HS, Im W. G-LoSA for prediction of protein-ligand binding sites and structures. *Methods Mol Biol* 1611;**2017**:97–108.
72. Aung Z, Tong JC. BSAlign: a rapid graph-based algorithm for detecting ligand-binding sites in protein structures. *Genome Inform* 2008;**21**:65–76.
73. Koch I, Lengauer T, Wanke E. An algorithm for finding maximal common subtopologies in a set of protein structures. *J Comput Biol* 1996;**3**:289–306.
74. Ostergard PRJ. A fast algorithm for the maximum clique problem. *Discrete Appl Math* 2002;**120**:195–205.
75. Alexandrov NN, Fischer D. Analysis of topological and non-topological structural similarities in the PDB: new examples with old structures. *Proteins* 1996;**25**:354–65.
76. Wolfson HJ, Rigoutsos I. Geometric hashing: an overview. *IEEE Comput Sci Eng* 1997;**11**:263–78.
77. Liu T, Altman RB. Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding. *PLoS Comput Biol* 2011;**7**:e1002326.
78. Saberi Fathi SM, Tuszynski JA. A simple method for finding a protein's ligand-binding pockets. *BMC Struct Biol* 2014;**14**:18.
79. Brakoulias A, Jackson RM. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* 2004;**56**:250–60.
80. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;**89**:10915–9.
81. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr D Biol Crystallogr* 1976;**A32**:922–3.
82. Kellenberger E, Muller P, Schalon C, *et al*. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J Chem Inf Model* 2006;**46**:717–27.
83. Connolly M. Analytical molecular surface calculation. *J Appl Crystallogr* 1983;**16**:548–58.
84. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 2000;**28**:254–6.
85. Murzin AG, Brenner SE, Hubbard T, *et al*. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;**247**:536–40.
86. Schalon C, Surgand JS, Kellenberger E, *et al*. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* 2008;**71**:1755–78.
87. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;**28**:304–5.
88. Borgna JL, Rochefort H. High-affinity binding to the estrogen receptor of [3H]4-hydroxytamoxifen, an active antiestrogen metabolite. *Mol Cell Endocrinol* 1980;**20**:71–85.
89. Coward P, Lee D, Hull MV, *et al*. 4-Hydroxytamoxifen binds to and deactivates the estrogen-related receptor gamma. *Proc Natl Acad Sci USA* 2001;**98**:8880–4.
90. Seval Y, Cakmak H, Kayisli UA, *et al*. Estrogen-mediated regulation of p38 mitogen-activated protein kinase in human endometrium. *J Clin Endocrinol Metab* 2006;**91**:2349–57.
91. Zapletalek M, Hubsch T, Kindernayova H. Clinical experience with sydnocarb in neuroses and psychoses. *Act Nerv Super* 1975;**17**:235–6.
92. Erdo SL, Kiss B, Rosdy B. Inhibition of dopamine uptake by a new psychostimulant mesocarb (Sydnocarb). *Pol J Pharmacol Pharm* 1981;**33**:141–7.
93. Saporito MS, Ochman AR, Lipinski CA, *et al*. MLR-1023 is a potent and selective allosteric activator of Lyn kinase in vitro that improves glucose tolerance in vivo. *J Pharmacol Exp Ther* 2012;**342**:15–22.
94. Patterson S, Wyllie S, Norval S, *et al*. The anti-tubercular drug delamanid as a potential oral treatment for visceral leishmaniasis. *Elife* 2016;**5**.
95. Li J, Zheng S, Chen B, *et al*. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016;**17**:2–12.
96. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 2011;**12**:303–11.
97. Guney E, Menche J, Vidal M, *et al*. Network-based in silico drug efficacy screening. *Nat Commun* 2016;**7**:10331.
98. Lamb J, Crawford ED, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–35.
99. Andronis C, Sharma A, Virvilis V, *et al*. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform* 2011;**12**:357–68.
100. Sanseau P, Agarwal P, Barnes MR, *et al*. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* 2012;**30**:317–20.
101. Keiser MJ, Setola V, Irwin JJ, *et al*. Predicting new molecular targets for known drugs. *Nature* 2009;**462**:175–81.
102. Ehrt C, Brinkjost T, Koch O. Impact of binding site comparisons on medicinal chemistry and rational molecular design. *J Med Chem* 2016;**59**:4121–51.

103. Campillos M, Kuhn M, Gavin AC, *et al*. Drug target identification using side-effect similarity. *Science* 2008; **321**:263–6.
104. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem Substance and Compound databases. *Nucleic Acids Res* 2016;**44**: D1202–13.
105. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
106. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
107. Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen* 2001;**4**:719–25.
108. Rastegar-Mojarad M, Liu H, Nambisan P. Using social media data to identify potential candidates for drug repurposing: a feasibility study. *JMIR Res Protoc* 2016;**5**:e121.
109. Kellenberger E, Schalon C, Rognan D. How to measure the similarity between protein-ligand binding sites? *Curr Comput Aided Drug Des* 2008;**4**:209–20.
110. Brylinski M, Naderi M, Govindaraj RG, *et al*. eRepo-ORP: exploring the opportunity space to combat orphan diseases with existing drugs. *J Mol Biol* 2018;**430**:2266–73.
111. Kuenzi BM, Remsing Rix LL, Stewart PA, *et al*. Polypharmacology-based ceritinib repurposing using integrated functional proteomics. *Nat Chem Biol* 2017; **13**:1222–31.
112. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov* 2004;**3**:417–29.
113. Apsel B, Blair JA, Gonzalez B, *et al*. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat Chem Biol* 2008;**4**:691–9.
114. Whitnall M, Howard J, Ponka P, *et al*. A class of iron chelators with a wide spectrum of potent antitumor activity that overcomes resistance to chemotherapeutics. *Proc Natl Acad Sci USA* 2006;**103**:14901–6.
115. Yuan J, Lovejoy DB, Richardson DR. Novel di-2-pyridyl-derived iron chelators with marked and selective antitumor activity: in vitro and in vivo assessment. *Blood* 2004;**104**:1450–8.
116. Jansson PJ, Kalinowski DS, Lane DJ, *et al*. The renaissance of polypharmacology in the development of anti-cancer therapeutics: inhibition of the 'Triad of Death' in cancer by Di-2-pyridylketone thiosemicarbazones. *Pharmacol Res* 2015;**100**:255–60.
117. Fang J, Liu C, Wang Q, *et al*. In silico polypharmacology of natural products. *Brief Bioinform* 2017.
118. Li P, Chen J, Zhang W, *et al*. Transcriptome inference and systems approaches to polypharmacology and drug discovery in herbal medicine. *J Ethnopharmacol* 2017;**195**:127–36.
119. Drakakis G, Wafford KA, Brewerton SC, *et al*. Polypharmacological in silico bioactivity profiling and experimental validation uncovers sedative-hypnotic effects of approved and experimental drugs in rat. *ACS Chem Biol* 2017;**12**: 1593–602.
120. Berger SI, Iyengar R. Role of systems pharmacology in understanding drug adverse events. *Wiley Interdiscip Rev Syst Biol Med* 2011;**3**:129–35.
121. Vilar S, Ryan PB, Madigan D, *et al*. Similarity-based modeling applied to signal detection in pharmacovigilance. *CPT Pharmacometrics Syst Pharmacol* 2014;**3**:e137.
122. Mizutani S, Pauwels E, Stoven V, *et al*. Relating drug-protein interaction network with drug side effects. *Bioinformatics* 2012;**28**:i522–8.
123. Dider S, Ji J, Zhao Z, *et al*. Molecular mechanisms involved in the side effects of fatty acid amide hydrolase inhibitors: a structural phenomics approach to proteome-wide cellular off-target deconvolution and disease association. *NPJ Syst Biol Appl* 2016;**2**:16023.
124. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res* 2007;**35**: 3375–82.
125. Brylinski M, Lingam D. eThread: a highly optimized machine learning-based approach to meta-threading and the modeling of protein tertiary structures. *PLoS One* 2012;**7**: e50200.
126. Pons JL, Labesse G. @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. *Nucleic Acids Res* 2009;**37**:W485–91.
127. Brylinski M. Unleashing the power of meta-threading for evolution/structure-based function inference of proteins. *Front Genet* 2013;**4**:118.
128. Brady GP Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* 2000;**14**:383–401.
129. Huang B. MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS* 2009; **13**:325–30.
130. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013;**41**:D1096–103.
131. Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 2013;**29**:2588–95.
132. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res* 2012;**40**:W471–7.
133. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455–61.
134. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, *et al*. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput Biol* 2014;**10**: e1003571.
135. Govindaraj RG, Naderi M, Singha M, *et al*. Large-scale computational drug repositioning to find treatments for rare diseases. *NPJ Syst Biol Appl* 2018;**4**:13.
136. Luchinat C. Exploring the conformational heterogeneity of biomolecules: theory and experiments. *Phys Chem Chem Phys* 2016;**18**:5684–5.
137. Mittermaier A, Kay LE. New tools provide new insights in NMR studies of protein dynamics. *Science* 2006;**312**:224–8.
138. Stank A, Kokh DB, Fuller JC, *et al*. Protein binding pocket dynamics. *Acc Chem Res* 2016;**49**:809–15.
139. Uversky VN. Intrinsically disordered proteins and novel strategies for drug discovery. *Expert Opin Drug Discov* 2012; **7**:475–88.
140. Tuffery P, Derreumaux P. Flexibility and binding affinity in protein–ligand, protein–protein and multi-component protein interactions: limitations of current computational approaches. *J R Soc Interface* 2012;**9**:20–33.
141. Kunze J, Todoroff N, Schneider P, *et al*. Targeting dynamic pockets of HIV-1 protease by structure-based computa-

tional screening for allosteric inhibitors. *J Chem Inf Model* 2014;**54**:987–91.
142. De Vivo M, Cavalli A. Recent advances in dynamic docking for drug discovery. *WIREs Comput Mol Sci* 2017;**7**:e1320.
143. Shaw VS, Mohammadiarani H, Vashisth H, *et al*. Differential protein dynamics of regulators of G-protein signaling: role in specificity of small-molecule inhibitors. *J Am Chem Soc* 2018;**140**:3454–60.
144. Krizhevsky A, Sutskever I, Hinton GE. *ImageNet Classification With Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems*, **Vol. 1**. Lake Tahoe, Nevada: Curran Associates Inc., 2012, 1097–105.
145. Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Proc Mag* 2012;**29**:82–97.
146. Lipton ZC, Berkowitz J, Elkan C. *A critical review of recurrent neural networks for sequence learning*. arXiv:1506.00019 [cs.LG].
147. Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016;**529**:484–9.
148. Chen Z, Huang X. End-to-end learning for lane keeping of self-driving cars. In: 2017 IEEE Intelligent Vehicles Symposium (IV). 2017, Los Angeles, CA, pp. 1856–60.
149. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B *et al*. (eds). *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Cham: Springer International Publishing, 2014, 818–33.
150. Szegedy C, Liu W, Jia Y, *et al*. *Going deeper with convolutions*. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, p. 1–9.
151. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;**26**:990–9.
152. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* 2006; **15**:2507–24.
153. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002;**11**:2714–26.
154. Wu Y, Lu M, Chen M, *et al*. OPUS-Ca: a knowledge-based potential function requiring only Calpha positions. *Protein Sci* 2007;**16**:1449–63.
155. Moult J, Pedersen JT, Judson R, *et al*. A large-scale experiment to assess protein structure prediction methods. *Proteins* 1995;**23**:ii–v.
156. Bengio Y. Learning deep architectures for AI. *Foundations Trends Mach Learn* 2009;**2**:1–127.
157. Sun T, Zhou B, Lai L, *et al*. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics* 2017;**18**:277.