Local Alignment of Ligand Binding Sites in Proteins for Polypharmacology and Drug Repositioning

Michal Brylinski

Abstract

The administration of drugs is a key strategy in pharmacotherapy to treat diseases. Drugs are typically developed to modulate the function of specific proteins, which are directly associated with particular disease states. Nonetheless, recent studies suggest that protein-drug interactions are rather promiscuous and the majority of pharmaceuticals exhibit activity against multiple, often unrelated proteins. Certainly, the lack of selectivity often leads to drug side effects; on the other hand, these polypharmacological attributes can be used to develop drugs acting on multiple targets within a unique disease pathway, as well as to identify new targets for existing drugs, which is known as drug repositioning. To support drug development and repurposing, we developed *e*MatchSite, a new approach to detect those binding sites having the capability to bind similar compounds. *e*MatchSite is available as a standalone software and a webserver at http://www.brylinski.org/ematchsite.

Keywords *e*MatchSite, Drug development, Drug repositioning, Polypharmacology, Computer-aided drug discovery, Binding site alignment, Sequence order-independent alignment

1 Introduction

Network analysis of interactions between proteins and small organic compounds is broadly applicable throughout the drug development process in both biology and chemistry. The classical picture of selective ligand binding has been challenged by experimental and computational studies, which strongly suggest that the space of protein-ligand interactions is dense and highly connected [1]. Several independent studies were conducted to estimate the promiscuity of protein–ligand interactions. For example, a large-scale across-target activity analysis carried out for 189,807 active compounds from PubChem [2] shows that the majority (62%) of them exhibit activity against multiple, often unrelated targets [3]. Another study investigating a set of 3138 compounds against 79 targets reported that 47% of the compounds can be classified as "promiscuous" and 24% as "highly promiscuous" with multiple

targets hit at an IC_{50} of <10 μM [4]. Furthermore, a thorough survey of a network of 5215 protein-ligand interactions connecting 829 compounds with 557 targets estimated that the average number of target proteins per ligand is 6.3 [5]. Although this notable binding promiscuity may complicate drug development, it also creates appealing opportunities for polypharmacology and drug repurposing.

Classical algorithms detecting relationships between proteins widely used in bioinformatics cannot be applied to investigate drug cross-reactivity because many compounds bind to multiple proteins that are totally unrelated to each other at the global sequence and structure levels [6, 7]. Therefore, a comprehensive analysis of the protein-drug interaction space requires a different set of tools. A direct comparison of binding sites is capable of describing ligand binding at the molecular level to provide useful insights into the compound mode of action [8]. Most algorithms for binding site matching fall into two categories: alignment-free and alignment-based methods. Geometric hashing is a typical example of the alignment-free approach; it measures the overall similarity of two binding sites, however, without providing any structural information on putative ligand binding modes and molecular interactions with target proteins [9]. In contrast, methods based on binding site alignments elucidate why two sites are similar, identify the sets of atoms/residues that contribute to the similarity, and describe putative ligand binding modes. SuMo (Surfing the Molecules) was one of the first approaches to use a residue-independent stereochemical group description combined with a fast, graph-based algorithm to compare protein structures and substructures [10]. Another method, SiteEngine, matches low-resolution protein surfaces constructed by converting triangles of physicochemical properties into a discrete set of chemically important points [11]. Finally, SOIPPA performs sequence order-independent profile-profile alignments of binding pockets using a coarse-grained representation of protein structures [12].

Despite encouraging progress in the development of sequence order-independent algorithms for ligand binding site alignment, many of these approaches perform well only against high-quality binding sites extracted from experimental protein structures. This insufficient accuracy hinders the reconstruction of protein-drug interaction networks across proteomes; thus, it is imperative to develop new approaches insensitive to structural deformation in ligand binding regions of protein models. To mitigate this issue, we developed *e*MatchSite, a new algorithm that performs sequence order-independent local binding site alignments using computer-generated protein models [13]. A key feature responsible for its high performance is the extensive use of evolutionary information that can be extracted even from weakly homologous templates complexed with ligands. In addition, *e*MatchSite provides a

calibrated significance score to identify those pockets capable of binding chemically similar ligands regardless of any global sequence and structure similarities between the target proteins. Benchmarking calculations demonstrate that *e*MatchSite outperforms other algorithms constructing sequence order-independent alignments of ligand binding sites. Importantly, *e*MatchSite maintains its high prediction accuracy against protein models; therefore, it opens up the possibility of investigating drug-protein interactions for complete proteomes with prospective systems-level applications in polypharmacology and rational drug repositioning.

2   Materials

2.1   Input Data

Input data for *e*MatchSite consist of two protein structures in the Protein Data Bank (PDB) format, whose ligand binding sites were annotated by *e*FindSite (*see* **Note 1**). *e*FindSite is a ligand binding site prediction and virtual screening algorithm that detects common ligand binding sites in a set of evolutionarily related proteins identified by meta-threading [14, 15]. In order to perform binding site annotation with *e*FindSite, users can employ either its standalone version or webserver located at http://www.brylinski.org/efindsite (*see* **Note 2**). It is noteworthy that both *e*FindSite and *e*MatchSite work well not only with experimentally solved structures, but also with computer-generated protein models (*see* **Note 3**).

2.2   Programs Used

*e*MatchSite is written in C++ and requires the following libraries: zlib (www.zlib.net), gzstream (www.cs.unc.edu/Research/compgeom/gzstream), and libsvm (www.csie.ntu.edu.tw/~cjlin/libsvm). In addition, *e*MatchSite requires a compound library, which is available at http://www.brylinski.org/content/ematchsite-standalone-package. Below, we describe options for running the standalone version of *e*MatchSite.

2.2.1   Input Options

-i input_file, where input_file is a single text file providing the location of all data files required by *e*MatchSite. Each line should contain only one keyword followed by a space and the path to the input file. Lines starting with # are ignored. List of keywords (A—first protein, B—second protein):

- structureA and structureB—target structures in the PDB format
- profilesA and profilesB—sequence profiles
- secstrA and secstrB—secondary structure profiles
- pocketsA and pocketsB—*e*FindSite pockets
- numberA and numberB—*e*FindSite pocket numbers (default 1)

- `alignmentsA` and `alignmentsB`—*e*FindSite alignments
- `ligandsA` and `ligandsB`—*e*FindSite ligands

Alternatively, users can specify the path to individual data files from command line. For example, the following arguments can be used to provide the location of target structures:

`-structureA first_pdb`, where `first_pdb` is the first protein in the PDB format,

`-structureB second_pdb`, where `second_pdb` is the second protein in the PDB format.

The path to other input data can be specified in a similar way.

2.2.2 Output Options   `-o output_name`, where `output_name` is a PDB file containing all results from *e*MatchSite.

2.2.3 Virtual Screening Options   `-m score_func`, where `score_func` is a scoring function to perform ligand-based virtual screening. Currently, implemented functions include single and combined scores (*see* [15] for details):

Single scoring functions:

`tst`—classical Tanimoto coefficient for Daylight fingerprints,

`tsa`—average Tanimoto coefficient for Daylight fingerprints,

`tsc`—continuous Tanimoto coefficient for Daylight fingerprints,

`tmt`—classical Tanimoto coefficient for MACCS fingerprints,

`tma`—average Tanimoto coefficient for MACCS fingerprints,

`tmc`—continuous Tanimoto coefficient for MACCS fingerprints.

Combined scoring functions:

`sum`—data fusion using the sum rule (default),

`max`—data fusion using the max rule,

`min`—data fusion using the min rule,

`svm`—machine learning using Support Vector Machines.

2.2.4 Output Files   *e*MatchSite outputs a single file that contains (1) numerical scores for the constructed alignment of binding sites, (2) aligned residue pairs with the corresponding Cα-Cα distances, (3) transformation matrices to superpose the second protein onto the first protein, and (4) the coordinates of the second protein upon the superposition of two binding sites.

2.3 Web Sites   The webserver available at http://www.brylinski.org/ematchsite provides a convenient interface to run *e*MatchSite. In addition, the website provides a standalone package that can be installed locally for high-throughput computations, benchmarking datasets

and results for an easy comparison with other algorithms constructing local alignments of binding sites, as well as a detailed manual and tutorial to help run *e*MatchSite.

3 Methods

3.1 Web Submission Form

The submission form for the *e*MatchSite webserver requires two target pockets annotated by *e*FindSite. Fig. 1 shows that users need to provide a 10-digit *e*FindSite ticket for each target and specify the pocket number if multiple pockets are identified in the target structure. In addition, each submission requires a unique "Job ID" and, optionally, an email address where the results will be sent. If a user prefers not to use email notifications, the automatically generated 10-digit ticket for each submission can be used to check the job status using the "Job Tracking" form in the right sidebar at http://www.brylinski.org.

3.2 Result Page

Results generated by the *e*MatchSite webserver are arranged into several sections, as shown in Fig. 2. The first section (Fig. 2a) gives the general information including the "Job ID," the "*e*MatchSite ticket" that can be used to retrieve the results within a month from the submission date, and the identifiers of protein targets. Further, three numerical scores for the local alignment of binding sites are provided, including the alignment score ranging from 0 to 1, the

This server is running **eMatchSite v1.0**. Click here to see some sample results.

Email:

michal@brylinski.org

Optional, check FAQ if you prefer not to use email notifications.

Job ID: *

1a0iA2aotA

3-10 alphanumerical characters.

eFindSite ticket (A): *

F1MauhwObn

Provide *e*FindSite ticket for the first protein.

eFindSite ticket (B): *

F86i3ItlsH

Provide *e*FindSite ticket for the second protein.

eFindSite pocket (A): *

1 ⇕

*e*FindSite pocket number (first protein).

eFindSite pocket (B): *

1 ⇕

*e*FindSite pocket number (second protein).

Submit!

After successful submission **do not** click the 'Back' button in your web browser.

Fig. 1 Web submission form for the *e*MatchSite server. Fields marked with *red asterisks* are mandatory
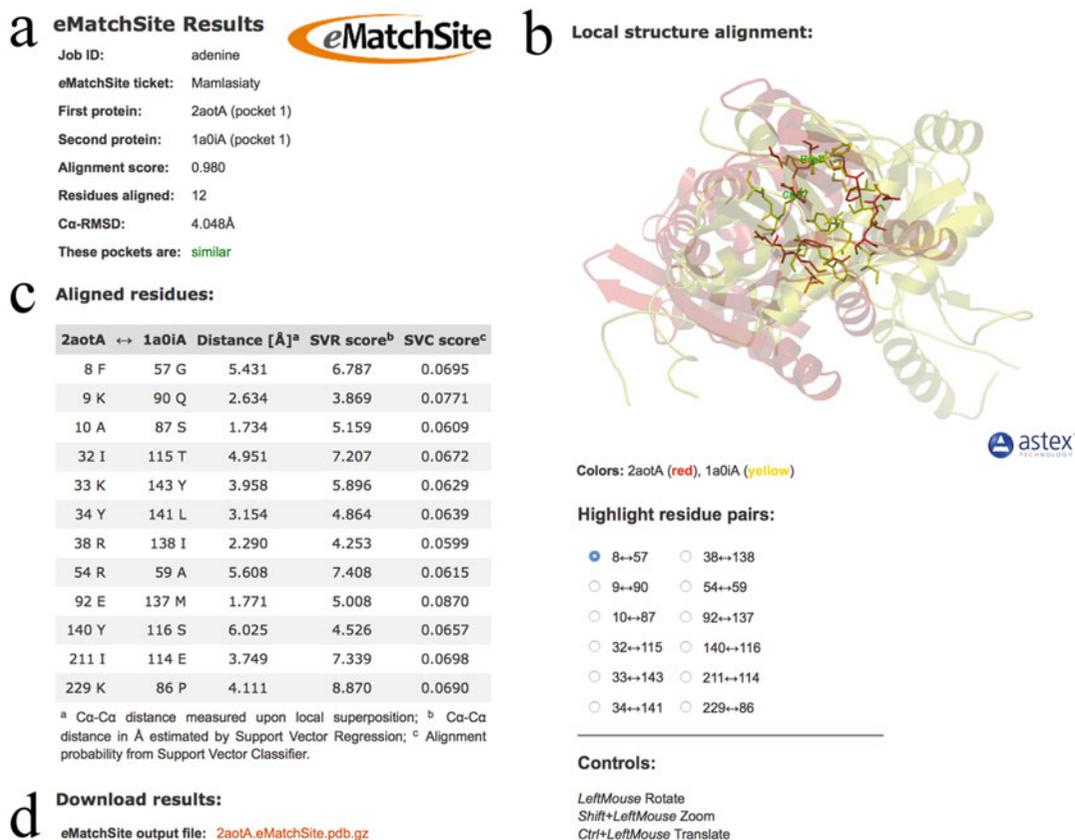
Fig. 2 Result page for the *e*MatchSite webserver. (**a**) Job information and numerical scores for pocket similarity, (**b**) AstexViewer applet showing the superposition of target pockets according to the constructed sequence order-independent alignment, (**c**) a list of aligned residue pairs and the corresponding numerical scores, and (**d**) the download section

number of aligned residues, and the Cα-RMSD calculated upon the superposition of target pockets. Based on the alignment score, the similarity of the pair of target pockets is determined. The superposition of target binding sites is visualized using the AstexViewer web applet [16] (Fig. 2b). Here, protein structures are displayed as transparent cartoons, whereas binding residues are shown as solid sticks. Moreover, individual aligned residue pairs can be highlighted and labeled using radio buttons. The next section contains a table showing the local alignment of binding sites (Fig. 2c). For each aligned residue pair, the Cα-Cα distance measured upon the local superposition, as well as the Cα-Cα distance and the probability score estimated by machine learning are listed. Finally, the last section provides a download link to the output file generated by *e*MatchSite (Fig. 2d).

Note that the results are kept on the server for 1 month only, after which all data associated with a particular submission will be deleted. However, we keep one sample job for each webserver, so

that users can quickly find out whether the webservers offer a desired functionality. These sample results can be accessed anytime either by clicking on links provided in the submission web forms, e. g., "This server is running *eMatchSite v1.0*. Click here to see some sample results," or by using 10-digit tickets, "Futermylok" for *e*FindSite and "Mamlasiaty" for *e*MatchSite, in the "Job Tracking" form in the right sidebar.

3.3 Case Studies   We selected a couple of illustrative examples to demonstrate how *e*MatchSite detects those pockets binding similar ligands in non-homologous proteins by constructing the sequence-order independent alignments of their binding sites. The primary target is benzoylformate decarboxylase (BFD) from *Pseudomonas putida* complexed with thiamin-2-thiazolone diphosphate (PDB-ID: 1yno) [17]. BFD belongs to the family of enzymes dependent on thiamine diphosphate and catalyzes the conversion of benzoylformate to benzaldehyde and carbon dioxide. Thiamin-2-thiazolone diphosphate (ThTDP) is a potent inhibitor of several thiamin-dependent enzymes that initiate the catalyzed reactions by forming a covalent adduct between the substrate and thiamin diphosphate (ThDP) through the C2 atom of the thiazolium ring [18]. In ThTDP, the proton on C2 is replaced with an oxygen atom to effectively inactivate thiamin-dependent enzymes. ThTDP binds to its target enzymes with an essentially identical binding mode as ThDP, but at a 10–1000 higher affinity compared to ThDP [19–21].

In addition to the primary target BFD, we selected two off-targets known to bind ThTDP, oxalyl-coenzyme A decarboxylase from *Oxalobacter formigenes* (OXC, PDB-ID: 2c31) [22] and the dehydrogenase/decarboxylase component of the human branched-chain α-ketoacid dehydrogenase complex (hE1b, PDB-ID: 2bff) [23]. OXC plays an important role in the catabolism of the highly toxic compound oxalate and it is structurally similar to BFD with a TM-score [24] of 0.85 despite a low sequence identity of 25% (Table 1, Global similarity). hE1b, which catalyzes the decarboxylation of branched-chain α-ketoacids derived from the amino acids leucine, isoleucine, and valine, shares neither sequence nor structure similarity with BFD (Table 1, Global

Table 1
Global and local structure similarity between the target BFD and off-targets OXC and hE1b

Target/off-target	Global similarity		Local similarity	
	Sequence identity (%)	TM-score	Pocket RMSD (Å)	Ligand RMSD (Å)
BFD/OXC	25	0.85	3.18	0.88
BFD/hE1b	25	0.35	2.72	0.89
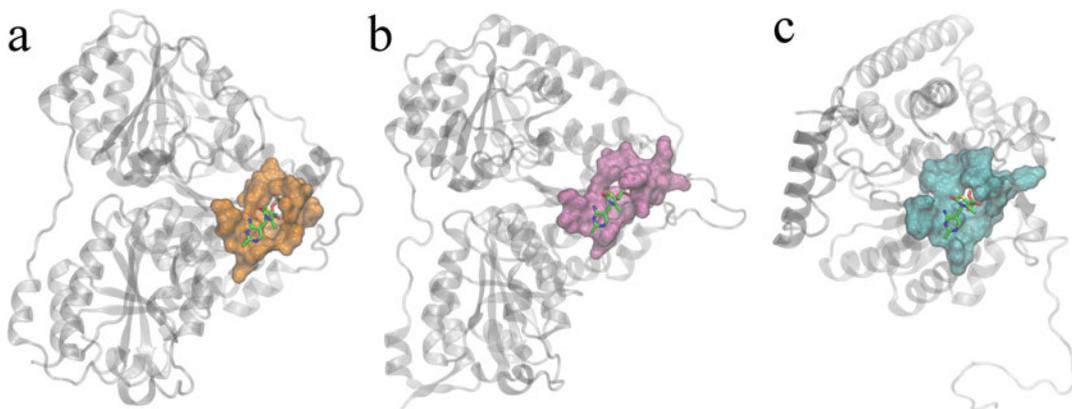
Fig. 3 Ligand binding pockets annotated with *e*FindSite. Target structures are shown as transparent *gray cartoons*, whereas binding residues are rendered as the molecular surface for (**a**) BFD, (**b**) OXC, and (**c**) hE1b. Binding sites in BFD, OXC, and hE1b are colored in *orange*, *pink*, and *cyan*, respectively. ThTDP ligands bound to the target structures are shown as *sticks colored* by the atom type (carbon—*green*, nitrogen—*blue*, oxygen—*red*, sulfur—*yellow*, phosphorus—*tan*)

similarity). Both off-targets represent a challenge to local binding site alignment algorithms due to their low sequence homology with the primary target, BFD.

*e*MatchSite requires binding sites and residues to be annotated by *e*FindSite; therefore, *e*FindSite webserver at http://www.brylinski.org/efindsite was used to identify binding sites in the crystal structures of BFD, OXC, and hE1b. The results are shown in Fig. 3. A binding site for ThTDP in the primary target BFD was identified with a 95% confidence; the Matthews correlation coefficient (MCC) calculated over binding residues is as high as 0.89 (Fig. 3a). The prediction confidence for off-targets OXC and hE1b is 94% and 93%, respectively. MCC calculated for binding residues identified in OXC is 0.78 (Fig. 3b) and 0.88 for hE1b (Fig. 3c). Note that bound ThTDP ligands are shown in Fig. 3 only to assess the accuracy of binding pocket prediction with *e*FindSite, which detects and annotates binding sites in ligand-free protein structures [14, 15].

Despite a low homology between the primary target and off-targets, binding sites in both OXC and hE1b are correctly recognized by *e*MatchSite as highly similar to the ThTDP-binding site in BFD, indicated by a confidence of 93% and 94%, respectively. Using data reported by *e*MatchSite, we can analyze how these high similarity scores were calculated. *e*MatchSite constructs sequence order-independent alignments using machine learning with Support Vector Regression techniques (SVR). Specifically, it assigns SVR scores to all possible combinations of binding residues in the first (target) and the second (off-target) protein. Then, it applies the Kuhn-Munkres algorithm [25, 26] to identify a unique set of residue pairs that give the shortest overall distance between their Cα

atoms. The Kuhn-Munkres algorithm, also known as the Hungarian method, belongs to the complexity class P [27], efficiently solving combinatorial assignment problems in polynomial time. In eMatchSite, this algorithm produces sequence order-independent alignments whose sum of Cα-Cα distances is guaranteed to be the smallest among all possible alignment combinations.

Clearly, a high correlation between Cα–Cα distances estimated by machine learning (SVR scores) and real distances calculated upon the superposition of bound ThTDP molecules is a critical factor to produce correct alignments. Encouragingly, light gray circles in Fig. 4 demonstrate that the machine learning model implemented in eMatchSite accurately predicts Cα-Cα distances; the Pearson correlation coefficients for BFD/OXC (Fig. 4a) and BFD/hE1b (Fig. 4b) are as high as 0.98 and 0.94, respectively. As a consequence, the unique sets of residue pairs selected by the Kuhn-Munkres algorithm to yield the shortest overall Cα-Cα distance actually correspond to the reference alignments constructed by superposing ThTDP molecules bound to the primary target and off-targets (dark gray triangles in Fig. 4). Bear in mind that ThTDP ligands bound to BFD, OXC, and hE1b are used only to validate alignments generated by eMatchSite that employs binding pockets annotated by eFindSite in ligand-free target structures.

Geometrical and physicochemical matching of binding sites in eMatchSite is supported by a chemical correlation, which was originally devised to study the inhibitor cross-reactivity within the human kinome [28]. In essence, a fingerprint-based virtual screening is performed against two pockets using a nonredundant subset of the ZINC library [29] comprising 23,659 molecules.
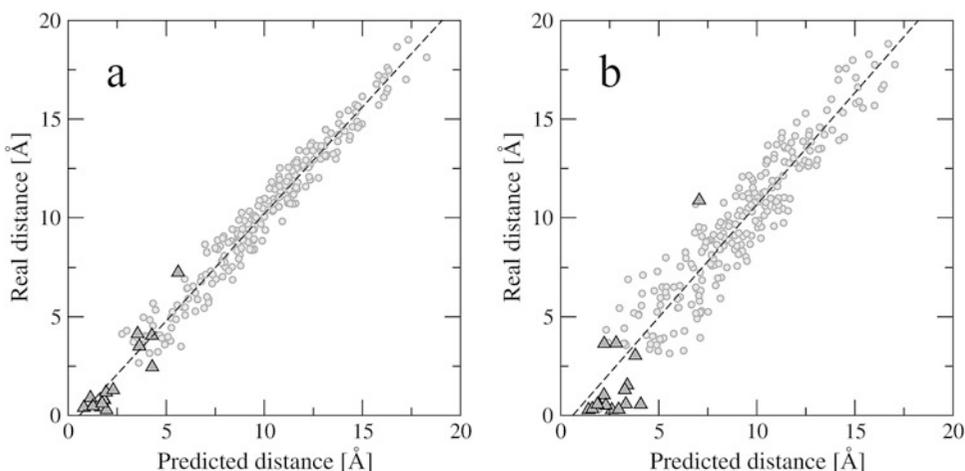
Fig. 4 Accuracy of the prediction of inter-residue distances by eMatchSite. Correlation between real distances upon the superposition of ThTDP ligands and those predicted by eMatchSite for target protein structures is shown for (**a**) BFD/OXC and (**b**) BFD/hE1b pairs. *Dark gray triangles* show residue pairs from the reference alignment
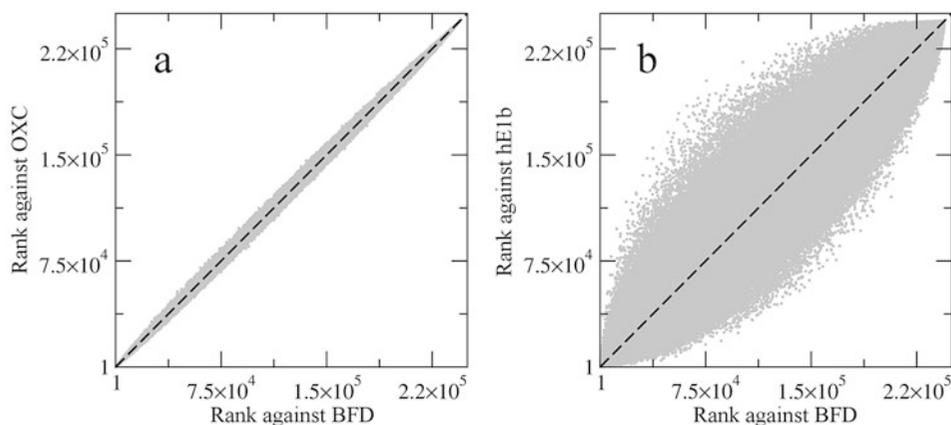
Fig. 5 Chemical correlation between ThTDP binding sites by *e*MatchSite. Rank correlation is plotted for a nonredundant subset of the ZINC library ranked using fingerprint-based virtual screening against a pair of target binding sites, (**a**) BFD/OXC and (**b**) BFD/hE1b

Subsequently, the Kendall τ rank correlation coefficient [30] is calculated for the ranked compounds under the assumption that virtual screening should yield a similar ranking for those pockets binding similar compounds. Indeed, Fig. 5 shows a high chemical correlation between binding sites in the primary and off-targets selected for this case study. The Kendall τ rank correlation coefficient is 0.98 between BFD and OXC (Fig. 5a), and 0.81 between BFD and hE1b (Fig. 5b).

Sequence order-independent alignments constructed by *e*MatchSite for BFD/OXC and BFD/hE1b are reported in Table 2. Sixteen residues are involved in the alignment between BFD and OXC, and 15 residues in that between BFD and hE1b; in both cases, the distances between the aligned Cα atoms upon the superposition of binding sites are fairly short. Binding pockets in off-targets superposed onto the pocket in the primary target structure are shown in Fig. 6. Table 1 (Local similarity) reports the RMSD calculated over Cα atoms of 3.18 Å for BFD/OXC (Fig. 6a) and 2.72 Å for BFD/hE1b (Fig. 6b). The accuracy of alignments constructed by *e*MatchSite can be evaluated by an RMSD calculated over the non-hydrogen atoms of ThTDP molecules upon the superposition of binding residues. Encouragingly, Table 1 (Local similarity) shows that the ligand RMSD is as low as 0.88 Å for BFD/OXC (Fig. 6c) and 0.89 Å for BFD/hE1b (Fig. 6b). Overall, these two case studies demonstrate that *e*MatchSite effectively recognizes binding site similarity and constructs biologically correct sequence order-independent alignments for pockets inferred by *e*FindSite from ligand-free protein structures.

Table 2
Sequence order-independent alignments constructed by *e*MatchSite for BFD/OXC and BFD/hE1b

BFD	OXC		hE1b	
Position	Position	Distance (Å)	Position	Distance (Å)
S375	G394	3.23	–	–
T376	A395	1.07	Y108	3.67
S377	N396	1.52	R109	3.19
G400	G420	2.03	S157	0.43
G401	V421	1.97	P158	0.79
L402	M422	0.61	L159	0.86
G426	G445	0.94	G187	1.19
D427	D446	1.32	E188	1.07
G428	S447	1.38	G189	0.90
S429	A448	0.95	A190	0.71
Y432	F451	0.48	E193	0.62
T456	G479	3.48	Y219	1.57
Y457	G475	2.47	A220	1.34
G458	K478	2.35	N217	5.47
A459	Y477	1.75	H286	3.08
L460	I476	4.31	I221	2.83

Distances between the aligned Cα atoms of BFD and either OXC or hE1b are measured upon the superposition of binding sites

4   Notes

1. Although the current version of *e*MatchSite requires ligand binding sites to be annotated by *e*FindSite, we are working on other prediction protocols to compare binding pockets detected by purely geometrical methods. Nonetheless, *e*Find-Site was demonstrated to outperform many other algorithms in large-scale benchmarking calculations; therefore, the combination of *e*FindSite/*e*MatchSite works best in detecting similar binding sites across large datasets of protein structures. Since *e*FindSite typically detects more than one site for the majority of proteins, users may want to specify the binding site of interest if they plan to run *e*MatchSite for a handful of targets. In large-scale applications, the top-ranked binding sites should be used by default because *e*FindSite ranks the best pocket at rank 1 in about 80% of the cases.
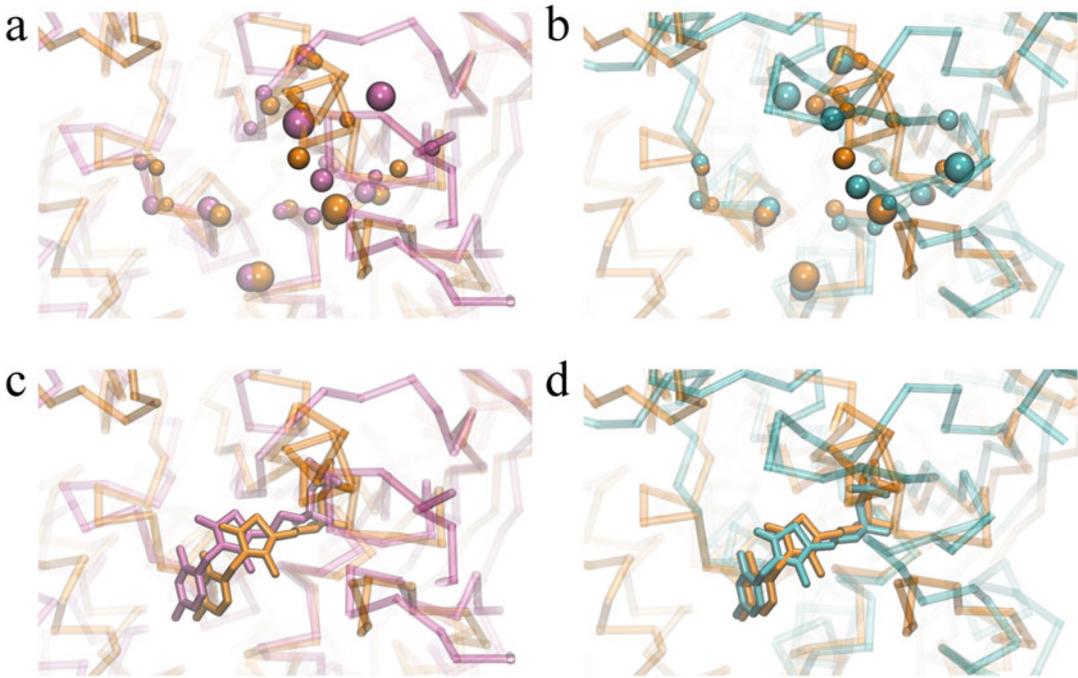
Fig. 6 Sequence order-independent alignments of ThTDP binding sites by *e*MatchSite. (**a**, **b**) Protein structures are superposed according to the local alignment of their binding sites with the Cα atoms of binding residues shown as *solid balls*. (**c**, **d**) Relative orientation of ThTDP ligands upon the local alignment of target binding sites. (**a**, **c**) BFD/OXC and (**b**, **d**) BFD/hE1b. BFD, OXC, and hE1b are colored in *orange*, *pink*, and *cyan*, respectively

2. Web portals for *e*FindSite and *e*MatchSite are intended to study selected proteins and their binding site similarities. However, both tools also have standalone versions that can be installed locally for high-throughput computations across large datasets of protein structures. Our website provides open source codes, the required template libraries, as well as detailed installation instructions and manuals to help users deploy *e*FindSite and *e*MatchSite on their computing systems. It is noteworthy that in addition to a serial code, *e*FindSite was ported to parallel accelerators in order to accelerate binding site annotations using heterogeneous computing systems [31, 32].

3. Both *e*FindSite and *e*MatchSite have been designed to work not only with experimental ligand-bound (holo) and ligand-free (apo) target structures, but also with computer-generated protein models. Compared to crystal structures, the accuracy of *e*FindSite predicting binding residues in high- and moderate-quality structure models decreases only by 4.2% and 9.9%, respectively [14]. Similarly, *e*MatchSite also maintains its capability to construct highly accurate alignments when protein models are used. Depending on the model quality, the

percentage of correctly aligned binding sites is only 4–9% lower than those aligned using crystal structures [13]. On that account, binding site similarities can be effectively detected using homology models generated across proteomes by contemporary protein structure prediction techniques.

Acknowledgments

References

1. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24:805–815

2. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH (2010) An overview of the PubChem BioAssay resource. Nucleic Acids Res 38:D255–D266

3. Li Q, Cheng T, Wang Y, Bryant SH (2010) PubChem as a public resource for drug discovery. Drug Discov Today 15:1052–1057

4. Azzaoui K, Hamon J, Faller B, Whitebread S, Jacoby E, Bender A, Jenkins JL, Urban L (2007) Modeling promiscuity based on in vitro safety pharmacology profiling data. ChemMedChem 2:874–880

5. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV (2008) Data completeness—the Achilles heel of drug-target networks. Nat Biotechnol 26:983–984

6. Benson SC, Pershadsingh HA, Ho CI, Chittiboyina A, Desai P, Pravenec M, Qi N, Wang J, Avery MA, Kurtz TW (2004) Identification of telmisartan as a unique angiotensin II receptor antagonist with selective PPARgamma-modulating activity. Hypertension 43:993–1002

7. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem 47:550–557

8. Kellenberger E, Schalon C, Rognan D (2008) How to measure the similarity between protein-ligand binding sites? Curr Comput Aided Drug Des 4:209–220

9. Yeturu K, Chandra N (2008) PocketMatch: a new algorithm to compare binding sites in protein structures. BMC Bioinformatics 9:543

10. Jambon M, Imberty A, Deleage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 52:137–145

11. Shulman-Peleg A, Nussinov R, Wolfson HJ (2004) Recognition of functional sites in protein structures. J Mol Biol 339:607–633

12. Xie L, Bourne PE (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci U S A 105:5441–5446

13. Brylinski M (2014) eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models. PLoS Comput Biol 10:e1003829

14. Brylinski M, Feinstein WP (2013) eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. J Comput Aided Mol Des 27:551–567

15. Feinstein WP, Brylinski M (2014) eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models. Mol Inform 33:135–150

16. Hartshorn MJ (2002) AstexViewer: a visualisation aid for structure-based drug design. J Comput Aided Mol Des 16:871–881

17. Polovnikova ES, McLeish MJ, Sergienko EA, Burgner JT, Anderson NL, Bera AK, Jordan F,

Kenyon GL, Hasson MS (2003) Structural and kinetic analysis of catalysis by a thiamin diphosphate-dependent enzyme, benzoylformate decarboxylase. Biochemistry 42:1820–1830

18. Arjunan P, Chandrasekhar K, Sax M, Brunskill A, Nemeria N, Jordan F, Furey W (2004) Structural determinants of enzyme binding affinity: the E1 component of pyruvate dehydrogenase from *Escherichia coli* in complex with the inhibitor thiamin thiazolone diphosphate. Biochemistry 43:2405–2411

19. Kluger R, Gish G, Kauffman G (1984) Interaction of thiamin diphosphate and thiamin thiazolone diphosphate with wheat germ pyruvate decarboxylase. J Biol Chem 259:8960–8965

20. O'Brien TA, Gennis RB (1980) Studies of the thiamin pyrophosphate binding site of *Escherichia coli* pyruvate oxidase. Evidence for an essential tryptophan residue. J Biol Chem 255:3302–3307

21. Nemeria N, Yan Y, Zhang Z, Brown AM, Arjunan P, Furey W, Guest JR, Jordan F (2001) Inhibition of the *Escherichia coli* pyruvate dehydrogenase complex E1 subunit and its tyrosine 177 variants by thiamin 2-thiazolone and thiamin 2-thiothiazolone diphosphates. Evidence for reversible tight-binding inhibition. J Biol Chem 276:45969–45978

22. Berthold CL, Moussatche P, Richards NG, Lindqvist Y (2005) Structural basis for activation of the thiamin diphosphate-dependent enzyme oxalyl-CoA decarboxylase by adenosine diphosphate. J Biol Chem 280:41645–41654

23. Machius M, Wynn RM, Chuang JL, Li J, Kluger R, Yu D, Tomchick DR, Brautigam CA, Chuang DT (2006) A versatile conformational switch regulates reactivity in human branched-chain alpha-ketoacid dehydrogenase. Structure 14:287–298

24. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710

25. Kuhn HW (1955) The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2:83–97

26. Munkres J (1957) Algorithms for the assignment and transportation problems. J Soc Ind Appl Math 5:32–38

27. Sipser M (2004) Time Complexity. In: Sipser M (ed) Introduction to the theory of computation. Thompson Learning, Inc., Cambridge, MA

28. Brylinski M, Skolnick J (2010) Cross-reactivity virtual profiling of the human kinome by X-react(KIN): a chemical systems biology approach. Mol Pharm 7:2324–2333

29. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182

30. Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

31. Feinstein W, Moreno J, Jarrell M, Brylinski M (2015) Accelerating the pace of protein functional annotation with Intel Xeon phi coprocessors. IEEE Trans Nanobioscience 14:429–439

32. Feinstein WP, Brylinski M (2015) Accelerated structural bioinformatics for drug discovery. In: Jeffers J, Reinders J (eds) High performance parallelism pearls volume two. Morgan Kaufmann, Waltham, pp 55–72