J Struct Funct Genomics (2011) 12:97–107
DOI 10.1007/s10969-011-9109-z

Accuracy of functional surfaces on comparatively modeled protein structures

Jieling Zhao · Joe Dundas · Sema Kachalo · Zheng Ouyang · Jie Liang

Received: 24 November 2010 / Accepted: 20 April 2011 / Published online: 4 May 2011

Abstract Identification and characterization of protein functional surfaces are important for predicting protein function, understanding enzyme mechanism, and docking small compounds to proteins. As the rapid speed of accumulation of protein sequence information far exceeds that of structures, constructing accurate models of protein functional surfaces and identify their key elements become increasingly important. A promising approach is to build comparative models from sequences using known structural templates such as those obtained from structural genome projects. Here we assess how well this approach works in modeling binding surfaces. By systematically building three-dimensional comparative models of proteins using MODELLER, we determine how well functional surfaces can be accurately reproduced. We use an alpha shape based pocket algorithm to compute all pockets on the modeled structures, and conduct a large-scale computation of similarity measurements (pocket RMSD and fraction of functional atoms captured) for 26,590 modeled enzyme protein structures. Overall, we find that when the sequence fragment of the binding surfaces has more than 45% identity to that of the template protein, the modeled surfaces have on average an RMSD of 0.5 Å, and contain 48% or more of the binding surface atoms, with nearly all of the important atoms in the signatures of binding pockets captured.

Keywords Protein binding surface · Comparative model · Signatures of binding pockets · Amylase

Abbreviation

PDB	Protein data bank
EC	Enzyme commission
CASTp	Computed atlas of surface topography of proteins
pRMSD	Pocket root mean square deviation
pvSOAR	Pocket and void surface patterns of amino acid residues
SOLAR	Signature of local active regions

Introduction

Proteins carry out their biological functions through binding interactions with other molecules. Knowledge of protein functional surfaces can provide important insight into how molecules interact. Identifying and characterizing binding sites on protein surfaces is an important task for structure-based function inference. Its success will aid in understanding of protein function, mechanism of enzyme reactions [1–4], characterization of differences of molecular interactions in natural and disease states [5], and designing of novel therapeutic compounds [6].

A variety of methods for binding site prediction have been developed [7–32]. One successful approach is through a database search for proteins that are similar to the query protein in certain aspects, such as sequence, fold,

J. Zhao (✉) · J. Dundas · S. Kachalo · Z. Ouyang · J. Liang
Department of Bioengineering, University of Illinois at Chicago, 851 S. Morgan Street, Room 218, MC-063, Chicago, IL 60607, USA
e-mail: jzhao31@uic.edu

J. Dundas
e-mail: jdunda1@uic.edu

S. Kachalo
e-mail: sema.kachalo@intel.com

Z. Ouyang
e-mail: zouyan1@uic.edu

J. Liang
e-mail: jliang@uic.edu

and physicochemical properties, and infer the functions of the query protein by transferring functional annotations from matched proteins [33–36]. However, as overall protein sequence or fold similarity often do not lead to functional similarity [37, 38], a promising approach for function prediction is to search proteins with similar local binding surfaces [18–25, 28–32, 39]. Fetrow et al. developed a method using active-site profiles to identify residues located in the spatial environment around the active site [18]. Babbitt et al. developed a method for establishing patterns of conservation that characterize individual superfamilies [40]. Binkowski et al. developed a method that matches binding surfaces by both sequence composition and geometric shape [22, 23, 25], which has been shown to be effective in predicting functions of proteins whose structures were solved in structural genomics project [23]. Tseng et al. extended this approach by incorporating evolutionary selection pressure due to biological function for assessment of binding surface similarity. Using a continuous time Markov model for residue substitution and an explicit phylogenetic tree, these authors were able to predict the functions of a large number of enzymes with accuracy [41]. Dundas et al. developed a method called SOLAR based on sequence order independent alignment of protein binding surfaces [42, 43] and successfully captured biologically important structural features.

A major limitation of these methods is the lack of high-resolution protein structures for detailed modeling of protein function surfaces, as current speed in deciphering protein sequences far exceeds that of protein structures. To infer protein functions from sequences and to build accurate structural models of functional surfaces, a promising approach is to derive model protein structures by template-based comparative modeling [44]. Comparative modeling can generate a large number of protein structures, often with high accuracy when the sequence identity between template and target proteins is $\geq$30% [45–49].

However, to our best knowledge, there has not been a systematic study on how well functional surfaces are reproduced in comparatively modeled protein structures. In this study, we evaluate on a large scale the effectiveness of comparative modeled protein structures in reconstruction of the binding surfaces of enzymes with diverse biological functions. We further develop criteria under which comparative models can be expected to produce accurate functional surfaces. Our paper is organized as follows. We first describe the data set, evaluation criteria, and overall computational procedures in the Methods section. This is followed by detailed description of our findings in the Results section. We conclude with remarks and discussion.

Materials and methods

Dataset

We first compiled a set of protein structures consisting of all enzyme proteins in the Protein Data Bank (PDB) [50, 51]. We then selected a subset of structures with the criteria of choosing the ones with reported *R*-values [52], with known enzyme commission (E.C.) numbers and with explicitly annotated binding site residues in the PDB file. This leads to a set of 13,166 protein structures, representing 231 different enzyme families, each with a distinct E.C. number. A different set of proteins consisting of 131 proteins in the α-amylase family and 31 proteins in the β-amylase family was constructed for detailed analysis of binding surface signatures.

Construction of modeled protein structures

For each of the 231 enzyme families, we carry out pairwise Smith-Waterman local sequence alignment [53] for all ordered pairs of proteins within the enzyme family. We select a subset of pairs of aligned proteins using the following criteria: First, only pairs with sequence identity more than 30% and less than 90% are selected. Second, only pairs with no more than three consecutively unaligned residues at either end of the sequence fragments for both binding pockets are selected. This restriction is necessary to prevent the generation of unrealistic structures using the MODELLER package. This results in a set of 26,590 pairs of proteins.

Next, for each ordered pair, a comparative model of protein structure is constructed for the second protein using the first protein as the structural template. MODELLER is a widely used tool for comparative modeling of protein structures. The main-chain atoms of core regions are obtained by superposing the core segment with the core structural template whose sequence is closest. Loops are generated by scanning a database of all known protein structures to identify the structurally variable regions that fit the anchor core regions and have a compatible sequence. Side chains are modeled based on their intrinsic conformational preferences and on the conformation of equivalent side chains in the template structures. Details of the MODELLER software can be found in reference [54]. We use the MODELLER package with default setting of parameters [44, 54]. As in many studies of comparative modeling of protein structures [55–60], we used default parameter values of MODELLER in our study. This choice is also necessary for the large scale study carried out.

Identification of functional surface

For each protein pair , we use the CASTp server [61] to identify their functional surface, which is taken as the

surface pocket containing annotated binding site residues. If there are more than one such pockets on a protein structure, only the largest one is selected, as the largest pocket often correspond to enzyme binding site [20, 29].

Identity of sequence fragments of binding pocket

Next, the fragment sequence identity formed by residues lining on the wall of each surface pocket is calculated:

$$s(\mathcal{A},\mathcal{N}) = \frac{|\mathcal{A}|}{|\mathcal{N}|},$$

where $\mathcal{A}$ is the set of aligned residues in the template pocket that are identical to corresponding residues of query protein sequence, $\mathcal{N}$ is the set of all residues in template pocket. Here $|\mathcal{A}|$ and $|\mathcal{N}|$ denote the number of elements in set $\mathcal{A}$ and $\mathcal{N}$, respectively.

Pocket-RMSD (*pRMSD*) measure

To assess how well the functional surface is modeled structurally, we measure the similarity between binding pockets on the modeled structure and on the real structure, as all proteins selected have known structures in the PDB. The first measure we use is the *pocket-RMSD* (*pRMSD*).

We compare atoms in the binding pocket of the real protein structure to the corresponding atoms in the modeled structure. Denote the set of pocket atoms on the binding pocket of the real structure as $\mathcal{P} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ and the corresponding set of modeled atoms $\mathcal{P}'$ as $\mathcal{P}' = (\boldsymbol{a}'_1, \boldsymbol{a}'_2, \ldots, \boldsymbol{a}'_n)$, where $\boldsymbol{a}_i \in \mathbb{R}^3$ and $\boldsymbol{a}'_i \in \mathbb{R}^3$ are the coordinates of the i-the atom in the real binding pocket and in the modeled structure, respectively, and n is the total number of atoms in the real binding pocket. The *pocket-RMSD* is calculated as:

$$pRMSD = \sqrt{\frac{\sum_{i=1}^{n} \|\boldsymbol{a}_i - \boldsymbol{R}\boldsymbol{a}'_i\|^2}{n}},$$

where $\boldsymbol{R}$ is the optimum rigid body rotation matrix which can be found using singular value decomposition (Fig. 1a). [62] that generates the least *pRMSD* value after translating the center of mass of each pocket to the origin.

Measure of recall of pocket atoms

To assess how complete the modeled binding surface is capturing the full binding surface on the real protein structure, we use the measure of *recall of pocket atoms*, namely the ratio of atoms in the real functional pocket that are captured in the modeled binding pocket. For each modeled structure, the recall r of real pocket atoms is calculated as

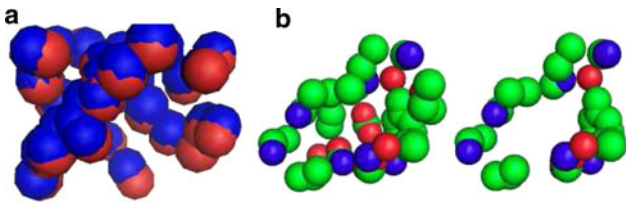

Fig. 1 Quantitative measures for assessing quality of modeled protein binding surfaces. **a** The *pRMSD* measures overall shape similarity of two binding surfaces. Atoms in *red* are functional surface atoms from real protein structure, and atoms in *blue* are corresponding atoms in modeled protein structure. After superimposing the centers of mass of the two binding pockets and rotating one surface pocket optimally, the point *RMSD* between corresponding atoms of real and modeled binding surfaces provides a measure of how the two surfaces differ in overall shape. **b** The *recall of pocket atoms* measures the fraction of atoms in the binding surface of the real structure (*left*) that are captured in the modeled binding surface (*right*). In this example, the recall of the real pocket atoms is 76%, as 41 out 54 atoms in the real binding surface are found in the modeled binding surface

$$r = \frac{|M \bigcap \mathcal{R}|}{|\mathcal{R}|},$$

where $\mathcal{R}$ is the set of atoms on the real functional surface, and $\mathcal{M}$ is the set of atoms from the modeled structure that are from the functional surface on the template structure (Fig. 1b).

Signatures of enzyme binding surfaces

Proteins often experience conformational changes and have complex binding activities. As a result, structures of the same enzyme family often have binding surfaces adopting different shapes. To determine structurally preserved and varied regions of the binding surface across an enzyme family, we compute the *signature* pockets of the enzyme binding surfaces using a method called SOLAR (Signature Of Local Active Regions, [43]).

Briefly, the signature pocket is calculated by first carrying out an all versus all sequence-order-independent pairwise structural alignment of the binding pockets from protein structures of the same enzyme activity (Fig. 2a). The structural alignments are then clustered hierarchically (Fig. 2b), and the resulting clustering tree is used to recursively combine two structurally aligned surfaces into one new surface pocket, representative of the original two binding surfaces (Fig. 2c). This procedure is carried out recursively until a pre-determined threshold of similarity is reached. By recording the *preservation ratio* for each atom in the final surface pocket, we can assess how conserved a specific atom is among all protein structures of the same enzyme function that were used to construct the signature pocket. Signature pockets consisting of those atoms above certain preservation ratio threshold from such analysis have shown to uncover atoms and residues that are important for

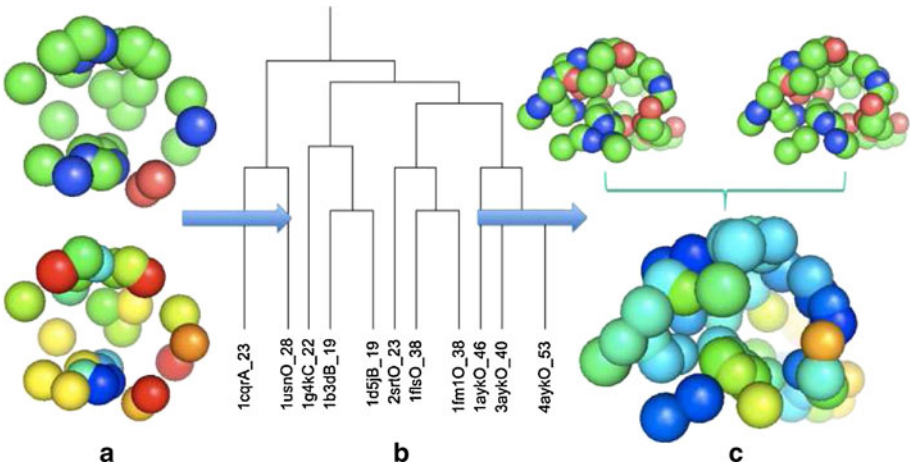

Fig. 2 Construction of signatures of binding surface pocket. Enzymes from the metalloendopeptidase family (E.C.number: 3.4.24) are used for illustration. **a** Pairs of functional surface pockets from different structures of this enzyme function are aligned structurally in a sequence-order-independent fashion. **b** Their surface pockets are then organized into a hierarchical tree based on measured pairwise similarity. **c** The hierarchical tree is used as a guide, and surface pockets of member proteins of the clade are combined recursively into a signature pocket. More details can be found in reference [43]

biological functions. Details of SOLAR is published in reference [43].

Measure of signature recall of real pocket atoms

To evaluate how well the modeled functional surface retains biologically important atoms, we calculate the *signature recall* r_s as:

$$r_s = \frac{|\mathcal{M} \bigcap \mathcal{S}|}{|\mathcal{S}|},$$

where $\mathcal{S}$ is the set of atoms in the signature pocket derived from real structures of the enzyme family, and $\mathcal{M}$ is the set of atoms from the modeled structure that are mapped from the functional surface on the template structure.

Results

Protein binding surfaces are much more conserved than the full sequences

Our set of proteins have diverse distribution of sequence identity between template and query proteins (Fig. 3a). This diverse distribution allows for a comprehensive and accurate assessment of the accuracy in constructed protein binding surfaces when using template-based comparative model.

Residues participating in the formation of binding pockets come from different regions in the primary sequences upon folding. These residues provide the necessary micro environment for biochemical reactions to occur, and often experience strong selection pressure. When these residues are concatenated, the resulting shorter sequence fragments have overall much higher sequence identity between protein pairs compared to that of the full sequences (Fig. 3b). The average identity for aligned pocket sequence fragments is 72%, which is much higher than the average identity of 52% for aligned backbone sequences. This indicates that there are strong similarity relationship between the functional surfaces of proteins. This is consistent with earlier findings [22, 23, 41, 63].

Shapes of binding surfaces in modeled protein structures

To evaluate how well modeled binding surfaces preserve the overall shapes of binding surfaces on real structures, we measure the pocket-RMSD *pRMSD* between them. The *pRMSD* value for each pair of real structure and modeled structure of 26,590 proteins are summarized in Fig. 4a, in which the distributions of *pRMSD* values at different intervals of identity of sequence fragments of binding pocket are shown.

We find that comparative modeling tools generally produce high quality models of binding surfaces. The overall average *pRMSD* of functional surface, which on average contain about 127 atoms, is less than 0.5 Å.

Sequence identity of binding pocket and recall of pocket atoms

We use the recall of pocket atoms *r* to measure the fraction of atoms in the real binding surfaces that appear in the corresponding binding surface pocket in the modeled protein structure. A high recall value indicates close to full identification of the functional surface atoms, and a low

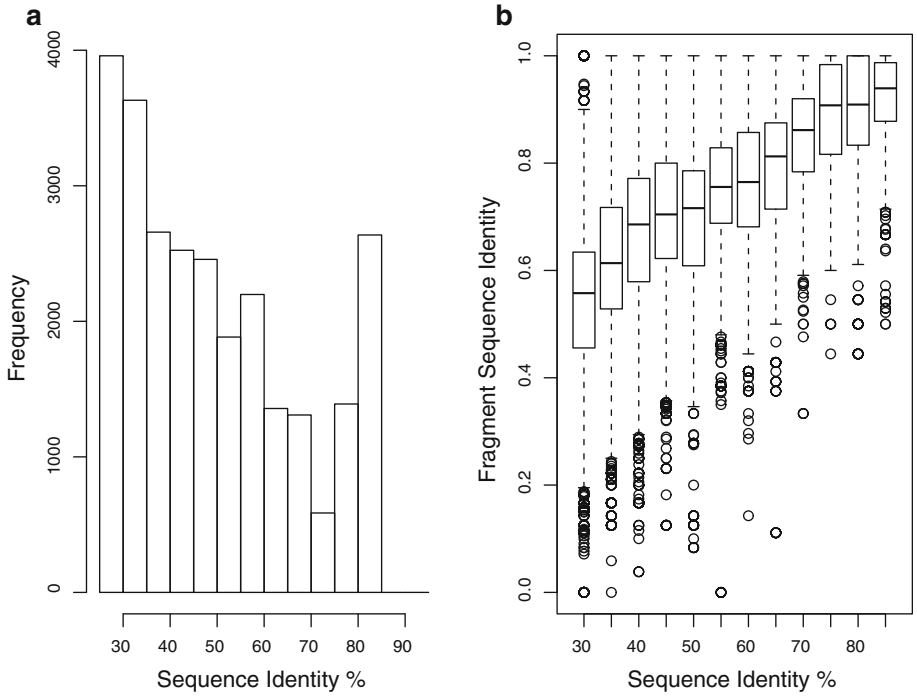

Fig. 3 Binding surfaces are generally more conserved compared to the full protein sequences. **a** Sequence identities between pairs of structural template and query proteins. Only those with >30% of sequence identity are used in this study. **b** The distribution of fragment sequence identity against overall identity of the full sequences. Here we use box plots to summarize the characteristics of the distribution. The *top edge* of the *box* represents the 75th percentile of the distribution; *center line* represents the median; the *bottom edge* the 25th percentile; The average identity for aligned pocket sequence fragments is 72%, which is much higher than the average identity of 52% of aligned backbone sequences

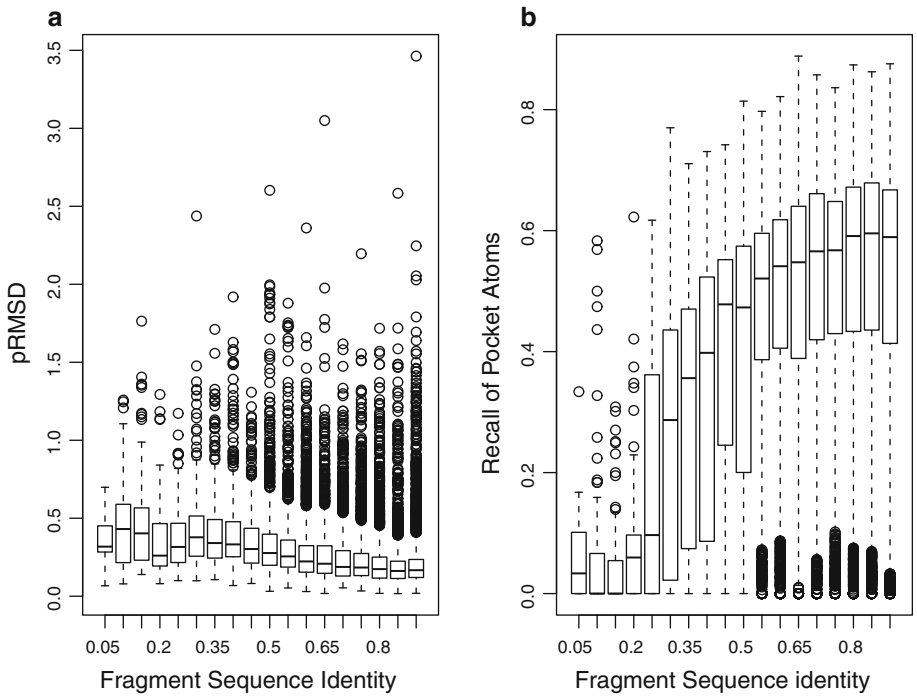

Fig. 4 Assessment of the shape and completeness of modeled functional surface pockets constructed by template-based comparative modeling method. **a** The distribution of *pRMSD* values between modeled and real binding surfaces for a set of 25,960 enzyme structures. *Box plot* of *pRMSD* are shown at different intervals of fragment sequence identity. Overall, modeled binding surfaces have very small *pRMSD* (on average 0.26 Å). **b** The distribution of the recall, namely, fractions of real pocket atoms captured in the modeled binding surfaces (recall) of the same set of modeled enzyme structures. The average recall of real pocket atoms is 46%

recall value indicates that many binding surface atoms are missing in the modeled binding surface pocket.

The results of 26,590 proteins are summarized in Fig. 4b. The overall recall of real pocket atoms is 46%. However, for pairs whose sequence fragments identity is less than 45%, the average recall of real pocket atoms is only 25%. This suggests that if the fragment sequence identity between the unknown protein and the template protein is $\geq$45%, comparative model can recapture a large portion of the true binding surfaces.

For a set of 4,848 protein pairs, we have examined results of modeled binding surfaces using alternatively either protein as the template. As proteins are subject to conformational fluctuation, binding surface in different

PDB structures may have differences. For example, there may be different usage of residues to form the binding pockets. Overall, we find that when protein structures with larger binding surfaces are used as templates, modeled binding surfaces have slightly better *pRMSD* and better recall of pocket atoms (data is not shown).

Effects of *R*-value on quality of modeled binding surface

We examine whether the *R*-value of the template structure affects the quality of the modeled binding surface. *R*-value measures how well the simulated diffraction pattern of the solved structure matches the experimentally-observed diffraction pattern [52]. A perfect fit has an *R*-value of 0, and a complete random fit has an *R* value of 0.63, while typical *R*-values are around 0.20. Here the average *R*-value of the template structures used to generate the 26,590 models is 0.19. The overall distribution of *R*-value is shown in Fig. 5a. We examined specifically whether the quality of modeled surfaces for the structures whose *R*-values are in the lowest 5% (with $R < 0.14$), the top 5% (with $R > 0.24$), and the middle 90% is different. We compare the distributions of both *pRMSD* and recall of pocket atoms for structures modeled using templates from these three sets. By plotting the quantiles from different sets against each other to obtain the Q-Q plot, we graphically compare the corresponding two distributions. If the two distributions being compared are similar, points in the Q–Q plot will approximately lie on the diagonal [64].

The Q–Q plots of pRMSD and recall for the lowest 5% and the middle 90%, for the top 5% and the middle 90% are summarized in Fig. 5b–e, respectively. In all four cases, the points approximately lie on the diagonals, suggesting that structures with different *R*-values do not have substantially different distribution in *pRMSD* and recall. That is, different *R*-values of template structures do not significantly affect the quality of modeled binding surfaces.

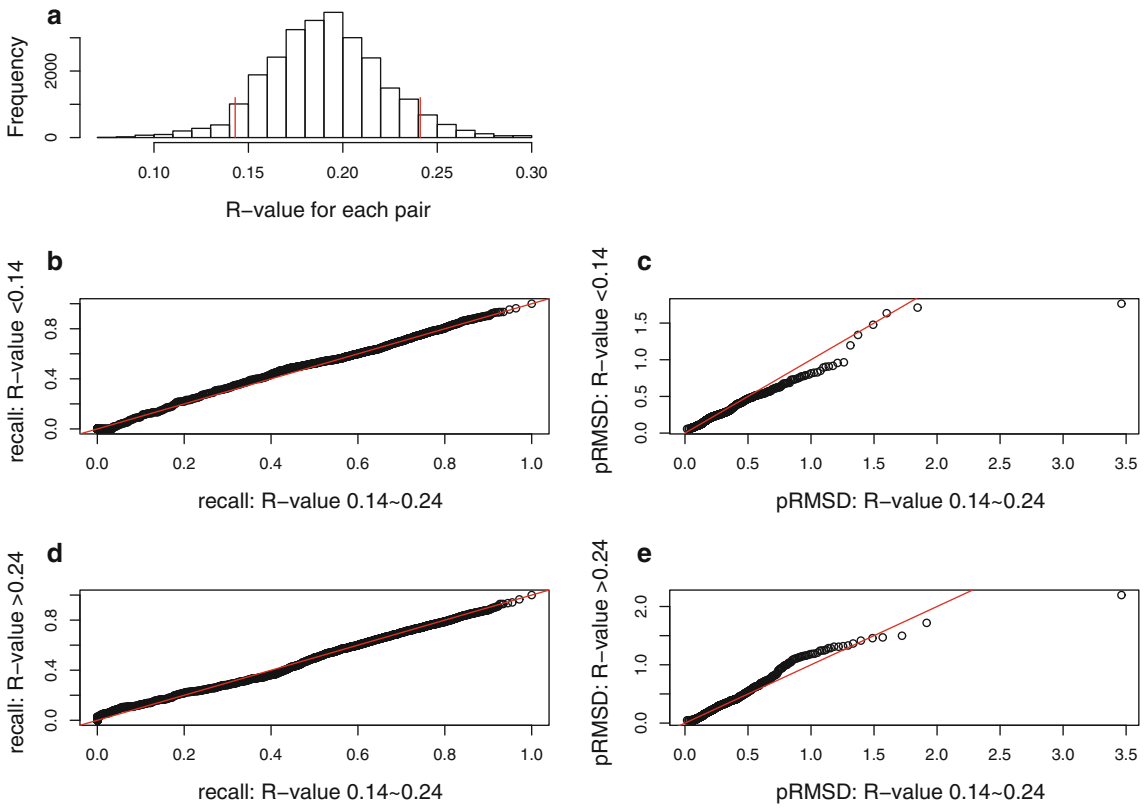

Fig. 5 *R*-value of template structure and quality of modeled binding surfaces. **a** The distribution of *R*-values of template protein structures used in the modeling of a total of 25,960 proteins. The mean *R*-value is 0.19. The *two red vertical lines* at $R = 0.14$ and $R = 0.24$ divide the samples into the lowest 5%, *middle* 90%, and *top* 5%. Q–Q plots of distributions of **b** values of recall of pocket atoms and **c** pRMSD values between those with *R*-value below 0.14 and those with *R*-value between 0.14 and 0.24. Q–Q plots of distributions of **d** recall of pocket atoms and **e** pRMSD between those with *R*-value above 0.24 and those with *R*-value between 0.14 and 0.24. Overall, these plots show that these distributions are alike, indicating that the *R*-value of a template structure has very limited influence on the quality of the modeled binding surface

Sequence identity of binding pockets and recall of signature atoms

Although in some cases comparative modeled protein structures have poor recall, namely, only a fraction of binding surface atoms are modeled correctly, we examine whether the key elements in a binding surfaces are modeled. As signatures of binding pockets provide information about spatially preserved atoms, they provide a map of key components that define the binding pocket.

Here we examine how well comparative models capture key atoms in the signature of the binding pocket for the enzyme class of amylase. Amylase plays an important role in breaking starch down into sugar. This enzyme family contains members whose sequences have diverged significantly, with members sharing often less than 25% sequence identity.

We have built comparative models for 162 sequences of amylase, including 131 α-amylase and 31 β-amylase using template structures. Overall, we have 3,406 modeled structures based on aligned pairs with the criteria of sequence identity between sequences of template and modeled structures must range from 30 to 90%, with the same gap restriction described earlier. We then construct two sets of signature pockets, one set from real amylase structures, and another set from modeled structures of the same set of amylase sequences.

The signature pockets of real amylase structures are shown in the left panel of Fig. 6, along with their clustering tree, and the signature pockets of the modeled amylase structures and their tree are shown in the right panel of Fig. 6. Overall, we find that the signature pockets derived from modeled amylase structures are structurally very similar to signature pockets derived from real protein structures. Although there may be missing binding surface atoms in modeled structures, the key elements in signatures are often modeled accurately, with overall shape preserved.

This is also demonstrated by recall analysis. We calculated the fraction of atoms appearing in the signature of real amylase binding pockets that are faithfully captured in modeled amylase binding pockets. The value of recall of all atoms and signature pocket atoms of modeled amylase binding pockets were plotted against identity of sequence fragments of binding pockets in Fig. 7a, b, respectively. The average recall of atoms of the full binding pocket is 77%, and this recall is improved to 94% when only signature atoms are considered. For those amylase with identity ≥45% between sequence fragments of binding pocket of modeled protein and template protein, the recall is also improved from 79 to 96%. Our results indicate that template-based models accurately model the structurally preserved and functionally important atoms of the binding pocket.

Signature binding pockets are well preserved in modeled binding surfaces

Because an enzyme family can bind to a diverse set of substrates and because similar or identical substrates may adopt different configurations, enzyme binding surfaces

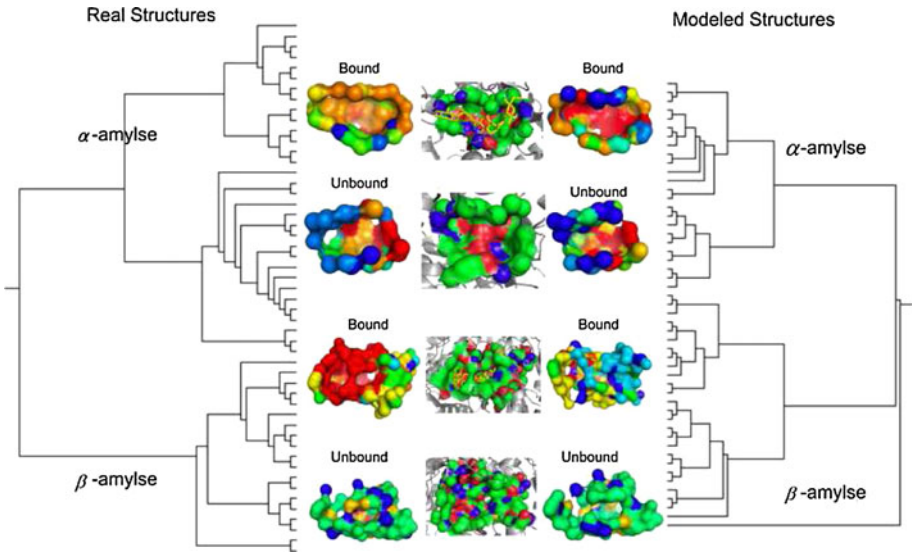

Fig. 6 The signatures of the binding surfaces of amylase derived from real structures (*left*) and from modeled structures (*right*). Binding surfaces on α- and β-amylase have their own basis set of signatures, and each requires two signature pockets. In each case, the two signatures correspond to conformations of the binding pocket with and without substrate bound. The clustering trees that were used to generate signatures are also shown. Signature pockets derived from modeled amylase structures are very similar structurally to signature pockets derived from real protein structures

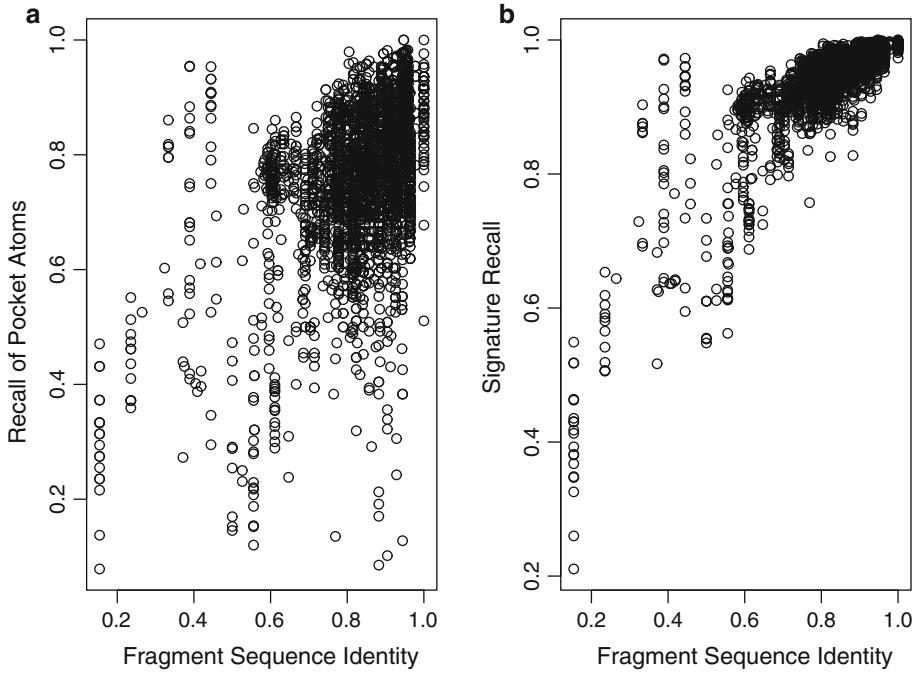

Fig. 7 Capturing biologically important signature atoms in modeled binding pockets of amylases. **a** The fraction of atoms on real functional surface that are captured in modeled protein surface (recall) is plotted against fragment sequence identity for 3,406 modeled amylase protein structures. The average recall is 77%. **b** The fraction of signature atoms on real functional surface that are captured in modeled protein surface (signature recall) is plotted. The overall average recall of signature atoms is 94%, much higher than the recall when all pocket atoms are considered (77%). For pairs whose identity of sequence fragments of binding pocket is greater than 45%, the average recall is 96%, which is much higher than that of recall of real pocket atoms (79%)

may experience significant differences in overall shape to accommodate these substrates. The SOLAR method can automatically identify multiple signatures that are required to account for the diverse substrate-binding surface conformations. These different signatures form the *basis set* of binding surfaces of an enzyme family. For the set of amylase structures, the basis set consists of two signature pockets for α-amylase and another two signature pockets for β-amylase. In either case, the two signature pockets represent conformations of the binding pocket with and without substrate bound (Fig. 6).

We found that modeled amylase binding surfaces also preserve such detailed information. The basis set of signature pockets derived from binding pockets of real amylase structures is very similar to the basis set derived from modeled amylase structures (Fig. 6). For example, there are two signatures for the binding surfaces with bound and unbound substrate for either α-amylase or β-amylase, respectively. All these details are fully captured in signatures derived from modeled binding surfaces, with two signatures for bound and unbound conformations for both α- and β-amylases. Overall, we found that signature binding pockets are well preserved in modeled protein structures and the important structural elements of the binding surfaces are accurately modeled.

Correlation of quality of local binding surface and global structure

The overall distribution of global RMSD of C_α atoms between modeled and real structures is shown in Fig. 8a. The distribution of pRMSD between modeled binding surface and real binding surface is shown in Fig. 8b. Their correlation is not strong (Fig. 8c, $R^2 = 0.22$). Majority of the modeled structures have good overall structure (low global RMSD) and their binding pockets are modeled well (low pocket RMSD). Nevertheless, 25 out of the 25,960 modeled proteins have poor global RMSD values (global RMSD $>$ 3.0 Å) but excellent pocket RMSD (pocket RMSD $<$ 0.5 Å). Detailed examination revealed that often this reflects accurately modeled secondary structures and loop regions near the binding site, with a number of poorly modeled loop regions in other parts of the protein, which results in overall poor global RMSD value. An example is shown in Fig. 8d.

Discussion

Identification and characterization of functional surfaces of enzymes are important for gaining insight into their

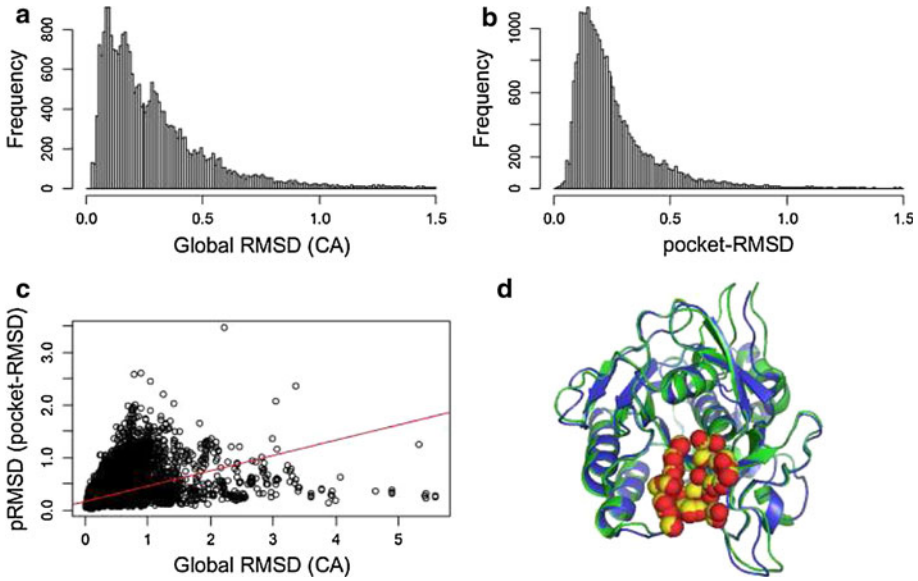

Fig. 8 Relationship between overall quality of modeled structures and quality of modeled binding surfaces. **a** The distribution of global C_α RMSD of 25,960 modeled proteins. Its mean value is 0.32 Å. **b** The distribution of pocket-RMSD, with a mean of 0.26 Å. **c** The correlation of pocket-RMSD and global RMSD (CA) of each of the 25,960 modeled protein structure. Overall, the correlation is not strong ($R^2 = 0.22$). **d** An example of predicted structure of peptide deformylase with binding surface modeled well but whose overall structure is in poorer quality. Here the PDB structure of `1dui` was used as the template, and a structure was built using the sequence of the structure of `2sec`. Compared to the true structure of `2sec` (*blue*), the modeled structure has a number of loops modeled with large deviation from the real structure, with an overall global cRMSD of 3.9 Å. However, the region near the binding site is modeled very accurately, and the modeled binding surface (*yellow atoms*) compared to the true binding surface (*red atoms*) has a small pRMSD of 0.14 Å

mechanism, and can lead to accurate prediction of protein functions through similarity comparison of local binding surfaces. However, a significant limitation for studying protein functional surfaces is the availability of high resolution protein structures. A promising approach towards solving this problem is to analyze protein functions using accurately modeled protein structures derived from structural templates such as those obtained from structural genomic projects.

Our study based on the amylase family suggest that modeled protein functional surfaces can capture a majority of the biologically important atoms, with overall 94% of the signature atoms included, far higher than the fraction of 77% when all binding pocket atoms are considered. Furthermore, signature pockets derived from modeled functional pockets are very similar to those derived from real functional pockets. Since amylase includes many proteins sequences with significant sequence divergence, we believe that similar pattern likely exists in other enzyme families.

In order to cover key conformations of the binding surfaces, signatures of binding pockets require the availability of many protein structures. As the signatures derived from modeled amylase structures are shown to be very similar to signatures derived from real structures, structures obtained through a template based modeling method therefore may alleviate the problem of lack of protein structures.

Conclusion

We have carried out a systematic study to assess the quality of binding surfaces on modeled protein structures. Our study is based on the analysis of 26,590 modeled protein structures from 231 enzyme families obtained from template-based comparative modeling method.

We found that functional surfaces on modeled protein structures are generally predicted accurately, with average pRMSD values <0.5 Å, suggesting that modeled binding surfaces have similar shape to real binding surfaces. This occurs even when the pocket sequence fragments of the modeled protein and the template protein have low sequence identity. In addition, we found that a significant fraction of the true binding surface atoms are captured in binding surfaces on modeled structures. When the fragment sequence identity is ≥45%, modeled binding surfaces contain on average over 77% of atoms of the true binding surfaces on real protein structures. Further analysis based on the amylase family suggested that when restricted to only the key elements forming signatures of the binding pockets, over 94% of the signature atoms are present in the binding pockets on modeled protein structure. Our results showed that surfaces generated by comparative models are often very accurate and informative, and can be used with some confidence to gain further insight into the enzyme mechanism and enzyme function. Furthermore, we also

showed that binding surfaces generated by comparative modeling can be used to construct template signatures of binding pockets.

Acknowledgments This work is supported by grants from NIH (GM079804, GM081682, GM086145, GM055876-13) and NSF (DMS-0800257).

References

1. Kinoshita K, Nakamura H (2003) Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. Protein Sci 12:1589–1595
2. Fersht A, Matouschek A, Serrano L (1992) The folding of an enzyme: I. theory of protein engineering analysis of stability and pathway of protein folding. J Mol Biol 224:771–782
3. Bartlett G, Porter C, Borkakoti N, Thornton M (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324:105–121
4. Putnam C, Arvai A, Bourne Y, Tainer J (2000) Active and inhibited human catalase structures: ligand and nadph binding and catalytic mechanism. J Mol Biol 296:295–309
5. Virkamaki A, Ueki K, Kahn C (1999) Protein-protein interaction in insulin signaling and the molecular mechanisms of insulin resistance. J Clin Invest 103(7):931–943
6. Ofran Y, Punta M, Schneider R, Rost B (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov Today 10:1475–1482
7. Henrich S, Salo-Ahen O, Huang B, Rippmann F, Cruciani G, Wade R (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. J Mol Recogn 23:209–219
8. Elcock A (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312:885–896
9. Ota M, Kinoshita K, Nishikawa K (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 327:1053–1064
10. Chelliah V, Chen L, Blundell T, Lovell S (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J Mol Biol 342:1487–1504
11. Cheng G, Qian B, Samudrala R, Baker D (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 33:5861–5867
12. Morita M, Nakamura S, Shimizu K (2008) Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. Proteins 73:468–479
13. Boobbyer D, Goodford P, McWhinnie P, Wade R (1989) New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. J Med Chem 32(5):1083–1094
14. Landon M, Lancia D, Yu J, Thiel S, Vajda S (2007) Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. J Med Chem 50(6):1231–1240
15. Vajada S, Guarnieri F (2006) Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Di De 9:354–362
16. Clark M, Guarnieri F, Shkurko I, Wiseman J (2006) Grand canonical monte carlo simulation of ligand-protein binding. J Chem Inf Model 46:231–242
17. Wade R, Goodford P (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. J Med Chem 36(1):148–156
18. Cammer S, Hoffman B, Speir J, Canady M, Nelson M, Knutson S, Gallina M, Baxter S, Fetrow J (2003) Structure-based active site profiles for genome analysis and sub-family classification. J Mol Biol 334(3):387–401
19. Brylinski M, Skolnick J (2007) A threading-based method (findsite) for ligand-binding site prediction and functional annotation. PNAS 105:129–134
20. Laskowski R (1995) Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13:323–330
21. Laurie A, Jackson R (2005) Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21:1908–1916
22. Binkowski A, Adamian L, Liang J (2003) Inferrring functional relationship of proteins from local sequence and spatial surface patterns. J Mol Biol 332:505–526
23. Binkowski A, Joachimiak A, Liang J (2005) Protein surface analysis for function annotation in high-throughput structural genomics pipeline. Protein Sci 14:2972–2981
24. Tseng Y, Liang J (2007) Predicting enzyme functional surfaces and locating key residues automatically from structures. Ann Biomed Eng 35(6):1037–1042
25. Tseng Y, Dundas J, Liang J (2009) Predicting protein function and binding profiles via matching of local evolutionary and geometric surface patterns. J Mol Biol 387:451–464
26. Levitt D, Banaszak J (1992) Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10:229–234
27. Hendlich M, Rippmann F, Barnickel G (1997) Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15:359–363
28. Huang B, Schroeder M (2006) Ligsite^{csc}: predicting ligand binding sites using the connolly surface and degree of conservation. BMC Struct Biol 6:19
29. Liang J, Edelsbrunner H, Woodward C (1995) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 7: 1884–1897
30. Brady G, Stouten P (2000) Fast prediction and visualization of protein binding pockets with pass. J Comput Aid Mol Des 14:383–401
31. Weisel M, Proschak E, Schneider G (2007) Pocketpicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J 1:7
32. Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinform 10:168
33. Kinoshita K, Nakamura H (2009) Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci 14:711–718
34. Loewenstein Y, Raimondo D, Redfern O, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A (2009) Protein function annotation by homology-based inference. Genome Biol 10:207
35. uncker A, Jensen L, Pierleoni A, Bernsel A, Tress M, Bork P, Heijne G, Valencia A, Ouzounis C, Casadio R, Brunak S (2009) Sequence-based feature prediction and annotation of proteins. Genome Biol 10:206
36. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nature 8:995–1005
37. Russell R, Sasieni P, Sternberg J (1998) Supersites within superfolds. Binding site similarity in absence of homology. J Mol Biol 282:903–918

38. Todd A, Orengo C, Thornton J (2001) Evolution of function in protein superfamilies from a structural perspective. J Mol Biol 307:1113–1143
39. Chen B, Honig B (2010) Vasp: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity. PLoS Comput Biol 6:1–11
40. Chiang R, Sali A, Babbitt P (2008) Evolutionarily conserved substrate substructures for automated annotation of enzyme superfamilies. PLoS Comput Biol 4:1–11
41. Tseng Y, Li W (2009) Identification of protein functional surfaces by the concept of a split pocket. Proteins 76:959–976
42. Liang J, Tseng Y, Dundas J, Binkowski A, Joachimiak A, Ouyang Z, Adamian L (2008) Predicting and characterizing protein functions through matching geometric and evolutionary patterns of binding surfaces. Adv Protein Chem 75:107–141
43. Dundas J, Adamian L, Liang J (2011) Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and nad binding proteins. J Mol Biol 406:713–729
44. Sali A, Blundell T (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234:779–815
45. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94
46. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Ann Rev Biophys Biomol Struct 29:291–325
47. Eramian D, Eswar N, Shen M, Sali A (2008) How well can the accuracy of comparative protein structure models be predicted?. Protein Sci 17:1881–1893
48. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294(5540):93–96
49. Fiser A (2009) Comparative protein structure modelling. Springer, Berlin, vol 3, pp 57–90
50. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide protein data bank. Nat Struct Biol 10:980–980
51. Berman H, Henrick K, Nakamura H, Markley J (2006) The worldwide protein data bank(wwpdb): ensuring a single, uniform archive of pdb data. Nucleic Acids Res 35:D301–D303
52. Kleywegt G, Jones A (1997) Model building and refinement practice. Methods Enzymol 277:208–230
53. Smith T, Waterman M (2006) Comparison of biosequences. Adv Appl Math 2:482–489
54. Eramian D, Marti-Renom M, Webb B, Madhusudhan M, Eswar N, Shen M, Pieper U, Sali A (2007) Comparative protein structure modeling with modeller. Curr Protoc Protein Sci 50:2.9.1–2.9.31
55. Li M, Wang B (2007) Homology modeling and examination of the effect of the d92e mutation on the h5n1 nonstructural protein ns1 effector domain. J Mol Model 13:1237–1244
56. Zheng Z, Zuo Z, Liu Z, Tsai K, Liu A, Zou GL (2005) Construction of a 3d model of nattokinase, a novel fibrinolytic enzyme from bacillus natto a novel nucleophilic catalytic mechanism for nattokinase. J Mol Graph Model 23:373–380
57. Kiss R, Kovari Z, Keseru G (2004) Homology modelling and binding site mapping of the human histamine h1 receptor. Eur J Med Chem 39:959–967
58. Gabdoulline R, Stein M, Wade R (2007) Apipsa: relating enzymatic kinetic parameters and interaction fields. BMC Bioinform 8:373–388
59. Bateman A, Finn R, Sims P, Wiedmer T, Biegert A, Soding J (2009) Phospholipid scramblases and tubby-like proteins belong to a new superfamily of membrane tethered transcription factors. Bioinformatics 25:159–162
60. Whalen K, Starczak V, Nelson D, Goldstone J, Hahn M (2010) Cytochrome p450 diversity and induction by gorgonian allelochemicals in the marine gastropod cyphoma gibbosum. BMC Bioinform 10:24–38
61. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34:W116–W118
62. Umeyama S (1991) Least-squares estimation of transformation parameters between two point patterns. IEEE Trans 13:376–380
63. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–882
64. Wilk M, Gnanadesikan R (1968) Probability plotting methods for the analysis of data. Biometrika 55:1–17