A Method for Localizing Ligand Binding Pockets in Protein Structures

Fabian Glaser,[1*] Richard J. Morris,[1] Rafael J. Najmanovich,[1] Roman A. Laskowski,[1] and Janet M. Thornton[1–3]
[1]*European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom*
[2]*Department of Biochemistry and Molecular Biology, University College, London, United Kingdom*
[3]*Department of Crystallography, Birkbeck College, London, United Kingdom*

ABSTRACT **The accurate identification of ligand binding sites in protein structures can be valuable in determining protein function. Once the binding site is known, it becomes easier to perform in silico and experimental procedures that may allow the ligand type and the protein function to be determined. For example, binding pocket shape analysis relies heavily on the correct localization of the ligand binding site. We have developed SURFNET-ConSurf, a modular, two-stage method for identifying the location and shape of potential ligand binding pockets in protein structures. In the first stage, the SURFNET program identifies clefts in the protein surface that are potential binding sites. In the second stage, these clefts are trimmed in size by cutting away regions distant from highly conserved residues, as defined by the ConSurf-HSSP database. The largest clefts that remain tend to be those where ligands bind. To test the approach, we analyzed a nonredundant set of 244 protein structures from the PDB and found that SURFNET-ConSurf identifies a ligand binding pocket in 75% of them. The trimming procedure reduces the original cleft volumes by 30% on average, while still encompassing an average 87% of the ligand volume. From the analysis of the results we conclude that for those cases in which the ligands are found in large, highly conserved clefts, the combined SURFNET-ConSurf method gives pockets that are a better match to the ligand shape and location. We also show that this approach works better for enzymes than for nonenzyme proteins. Proteins 2006;62:479–488.**

Key words: SURFNET; ConSurf-HSSP; residue conservation; ligand identification; ligand shape; protein clefts; functional sites

INTRODUCTION

The prediction of protein function from 3D structure is becoming increasingly important, as more and more protein structures of unknown function are solved by the various structural genomics projects (SGP).[1–4] In this context there have been many attempts to use structural information, particularly features relating to the functional sites of proteins,[4–9] as a measure of determining function. Since the shape, size, and chemical composition of the protein surface dictate the type of interaction the protein can make with its cognate ligand or other interacting partner (DNA, a second protein, etc.),[10] many studies have concentrated on the analysis of those properties using a variety of computational methods.[1,11] The overall objective is to predict as accurately as possible the ligand location (binding site), ligand type (cognate ligand, cofactor, etc.), and ultimately the protein function, based on the structure of that protein. This presents a number of important difficulties. For example, in most cases the binding site pocket is considerably larger than the ligand itself. Furthermore, in many enzymes the binding pocket is occupied by more than one molecule (e.g., substrate and cofactor).[8]

Previous approaches have aimed to shed light on these issues, mainly by identifying the approximate binding-site region or by describing and comparing binding-site properties (as an indirect way of comparing function). For example Mason et al.[12] dissect protein pockets into small binding volumes coupled with their physico-chemical characteristics (i.e., hydrophobicity, charge, etc.) and then try to map them to potential binding partners. Silberstein et al.[10] developed a method to identify substrate binding sites in proteins using computational solvent mapping. This method identifies pockets on the protein surface in which organic molecules tend to cluster, taking into account the free energy of binding of those probes to the protein residues. Regions, in which several different types of solvent probes bind, are predicted to be ligand-binding sites. Schmitt et al.[13] used a different approach to detect similarities between a query binding site and other similar preprocessed protein cavities stored in the Cavbase data-

Grant sponsor: the Structural Genomics Consortium (SGC); Grant sponsor: the European Molecular Biology Laboratory (EMBL). Grant sponsor: Structural Proteomics in Europe (SPINE); Grant number: QLG2-CT-2002-00988 (EU contract number). Grant sponsor: the BioSapiens network of excellence; Grant number: LHSG-CT-2003-503265.

Richard J. Morris's present address is John Innes Centre, Norwich Research Park, Colney, NR4 7HU Norwich, United Kingdom

*Correspondence to: Fabian Glaser, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom. E-mail: fabian@ebi.ac.uk

Received 27 April 2005; Revised 14 July 2005; Accepted 15 August 2005

Published online 22 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20769

base using cleft descriptors. Brady and Stouten[14] developed PASS, a purely geometrical method in which the protein is covered with spherical probes. These probes are then filtered and are given weights proportionally to the number of their neighboring spheres and the extent to which they are buried. Among the probes with the highest weights, the "active-site points" are determined, which represent the center of potential binding sites. Many other procedures have been developed.[8,12,15–19]

In this work we present SURFNET-ConSurf, a new method that identifies the accurate location and shape of ligand binding sites in proteins. SURFNET-ConSurf combines a pure geometrical method that identifies surface pockets together with an evolutionary method that estimates the degree of conservation of the amino acids surrounding these pockets. SURFNET-ConSurf has two modular and consecutive stages: First, the largest clefts on the protein surface[15] are identified by SURFNET.[20] Second, precalculated evolutionary rate estimates from the ConSurf-HSSP database[21] are used to trim each cleft by cutting away regions distant from highly conserved residues. In this way, the initial cleft volumes predicted by SURFNET are reduced to those regions in which the ligands are most likely to be found. The method is modular and may be implemented using any other cleft identification or conservation algorithms.

SURFNET-ConSurf improves on previous approaches by combining two known measures of "functionality" in proteins: cleft volume and residue conservation. Used separately, these methods have disadvantages: SURFNET generally identifies the main functional site as one of the largest clefts on the protein surface. However, the predicted cleft volume is in many cases much larger than the ligand that occupies it.[15] Thus, the location and shape of the ligand inside the binding cleft are not predicted accurately. On the other hand, the ConSurf-HSSP database provides a measure of the sequence evolutionary conservation for each protein site, thus suggesting the residues involved in ligand binding, but it does not include geometric features of the binding site (e.g., shape, volume, etc.). Additionally, in most cases several conserved patches can be identified, and it is not possible to predict which of them it is an actual ligand-binding site using conservation alone.[16,17,21]

In this paper, we show that combining conservation and surface pocket prediction allows a more accurate identification of ligand binding sites by significantly reducing the initial volume of pockets predicted by SURFNET (see Results). The accurate identification of the ligand binding site and shape is very important for functional assignment and in different fields such as docking, de novo drug design, etc.[22] To test our method, we analyzed a large nonredundant set of protein structures from the PDB for which the biologically-active quaternary structure is predicted. We have found that when applying SURFNET-ConSurf, one of the largest four clefts encompasses at least 20% of a ligand volume in 75% of the PDB structures under study. Additionally, SURFNET-ConSurf reduces the volumes of ligand binding pockets by more than 30% on average, to give clefts that are a better match to the ligand shape and observed location.

METHODS

The SURFNET-ConSurf method

The first stage of the SURFNET-ConSurf method consists of running the SURFNET program[20] version 1.6.2 with default parameters. The SURFNET algorithm identifies the clefts on the surface of a protein by placing a sphere between all pairs of atoms such that the sphere just touches each atom and is between some predefined minimum and maximum radius. Each sphere is progressively reduced in size if any other atoms intersect it until it either intersects with no further atoms, in which case it is retained, or its radius drops below the minimum size, in which case it is discarded. Once the clefts on the surface have been filled by spheres it is possible to cluster the spheres into separate regions (i.e., individual protein clefts) and calculate a volume for each cleft. Previous works have shown that the largest pockets on a protein surface are the most likely to be sites where small molecules and other ligands bind to the protein.[15,20,23] In this work, only the largest four surface clefts are considered.

The second part of SURFNET-ConSurf involves discarding the spheres that are distant from highly conserved residues, resulting in a trimming of the clusters representing each cleft. Residue conservation scores are obtained from the ConSurf-HSSP database version 1.0,[21] which provides estimates for the rate of evolution of each amino acid in a PDB structure. The scores are calculated using HSSP's multiple sequence alignments (MSA)[24] as the input for the Rate4Site algorithm[9] (for a detailed description of the ConSurf-HSSP methodology[21] and the conservation scale refer to the website http://consurf-hssp.tau.ac.il and references therein). During the trimming procedure, those spheres that are within a certain distance of any atom of a highly conserved residue are kept, and those that are not are discarded from the sphere list, thus reducing and reshaping the original cleft volumes. Two parameters govern this procedure: the maximum allowed distance between the atom and the sphere and the minimum conservation score cutoff (see "The Distance and Conservation Cutoffs" section in Results).

The trimming stage may break a cluster up into two or more separate clusters of SURFNET spheres [see Figs. 4(c), 5(f)]. In this work we sum the volumes of the new cluster(s) obtained from the original cleft as the new trimmed volume of the cleft after filtering. This subject may be worth further investigation, since the different subpockets may have biological significance, such as being the center of contact of different parts of the ligand to the protein, etc.

The Nonredundant PDB Data Set

We obtained our data set of nonredundant protein structures from the "NRPDB NCBI VAST" (NNV) resource (http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html). This resource clusters all PDB chains by sequence similarity, and chooses a representative for each group, based on an assessment of the quality of the structure. We worked with the nrpdb.060804 version of NNV, which contains 51,685 protein chains divided into 3702 clusters, obtained by clustering all chains using a BLAST *p*-value cutoff of 10e-7.[25] To obtain our data set, we took each representative chain and considered its "parent" PDB entry. If the entry contained at least one noncovalently bound ligand of at least seven atoms in size, and its "REMARK 350" records (optional annotation in the PDB header) indicated that the structure corresponds to or includes the biological unit of that protein, it was retained.

The requirement that the ligand be at least seven atoms in size was to exclude small ligands that happen to be in the liquor solution used in the crystallization process (e.g., PO_4, SO_4, glycerol etc.),[26,27] but they do not naturally bind to the protein in vivo. In some cases these ligands bind in the real substrate binding site, but we have no easy way of differentiating these from nonspecific interactions. Additionally, we removed all ligands forming one or more covalent bonds to their protein host. For example, N-acetyl-D-glucosamine (NAG), a common glycosylation factor.[28] Covalently bound ligands are often found in shallow nonspecific sites on the protein's surface, and therefore are not identifiable using our approach. Ideally, we would like to test the method only on those ligands that are directly involved in the biological activity of the protein in vivo. In this sense, the data set used here is merely an approximation to an ideal data set.

For the entries that do not have information about their quaternary structure in the PDB file, we could have used a multimer prediction algorithms such as PITA[29] or PQS[30] to obtain the complete biological unit. However, a comparison of the predictions made by these three methods (REMARK 350 annotation, PITA, and PQS) on a test set of 30 randomly chosen entries (see underlined entries in the Supplementary Table I found online at http://www.ebi.ac.uk/thornton-srv/databases/SURFNET-CONSURF), showed that there is generally no agreement between them. In only nine of the cases did the three methods agree (i.e., propose the same quaternary structure), in another 13 cases only two of them agree and in three cases the three of them disagree. In five other cases the results are not available for one or more of the methods. In this work, therefore, we chose to use only the annotated PDB files (i.e., those for which "REMARK 350" appears on the PDB header).

SURFNET-ConSurf currently has some technical limitations that further reduce the size of our final data set. For example, the maximum protein size that can be handled by the SURFNET program is one that fits a perfect cube of approximately ~115 Å per side. Thus, very large structures having many protein chains failed to complete the SURFNET calculation. In other cases the conservation scores from the ConSurf-HSSP database are not available, either because the HSSP file is not available, or the sequence has few homologs, or the ConSurf-HSSP calculation fails.[9,17,21] The net result of the filters described above and these program limitations, was a nonredundant data set of 244 PDB structures (see Table SI).

The Calibration Factor

In order to assess whether a given SURFNET-ConSurf prediction was a "success," we defined a "calibration factor" according to the following three empirical criteria: (1) the percentage of cleft reduction is larger than 25%, (2) the percentage of ligand volume lost by the reduction in pocket size is less than 15%, and (3) the percentage of the ligand volume included in the trimmed cleft is larger than 15%. If when applying SURFNET-ConSurf these three conditions are met, the calibration factor for an entry is set to "1" (i.e., success), otherwise the calibration factor is set to "0" (i.e., failure).

The "average calibration factor" (ACF) can then be calculated for any given test set of protein structures, where larger ACF values indicate more successful predictions.

RESULTS

The Distance and Conservation Cutoffs

Prior to using SURFNET-ConSurf we empirically optimized the two main parameters that are required for the method: the distance and conservation cutoffs used to filter the spheres produced by SURFNET. To this end we randomly chose a test set of 30 entries (half enzymes and half nonenzymes) from the initial nonredundant data set (see underlined entries in Table SI). Then, for each entry, we ran SURFNET-ConSurf for each of 189 combinations of sphere-atom distance and conservation cutoffs, using steps of 0.1 Å for the distance (between 3 and 5 Å), and steps of 1 for the ConSurf-HSSP conservation score cutoff [between level 1, no conservation and 9, maximum conservation; see the conservation scale on Figure 5(c)].

Figure 1 shows a plot of the calculated ACF values for each one of the 189 conditions. High ACF values indicate success: that is, ConSurf-HSSP scores have reduced the initial size of the cleft predicted by SURFNET, while retaining at least part of the ligand and not losing too much of it in the trimming process. Figure 1 shows that a combined effect of the atom–sphere distance and conservation cutoffs yields the highest ACF values in a large region (light orange and pink colors, ACF values between 0.24 to 0.3) between a distance of 3.0–4.2 Å, while the maximum is found with a distance of 3.5 Å and a ConSurf-HSSP conservation cutoff of 8.

In the orange and pink regions of Figure 1 (0.24–0.27 and 0.27–0.30 ACF value ranges, respectively), the high ACF values for very short distances (i.e., less than 3.2 Å occur at low conservation score cutoffs, suggesting that residue conservation has little effect here. In fact, what is happening is that the short distance cutoff is merely selecting the spheres close to the protein's surface and discarding the others, irrespective of residue conservation. This results in a large reduction in the number of spheres retained and, because ligands binding to proteins will tend to be close to the protein surface, the reduced cleft will encompass a significant part of the ligand, hence recording a "success." For longer sphere–atoms distances, high ACF values occur at high conservation cutoffs only (7, 8, and 9), as here it is principally the conservation filtering that is

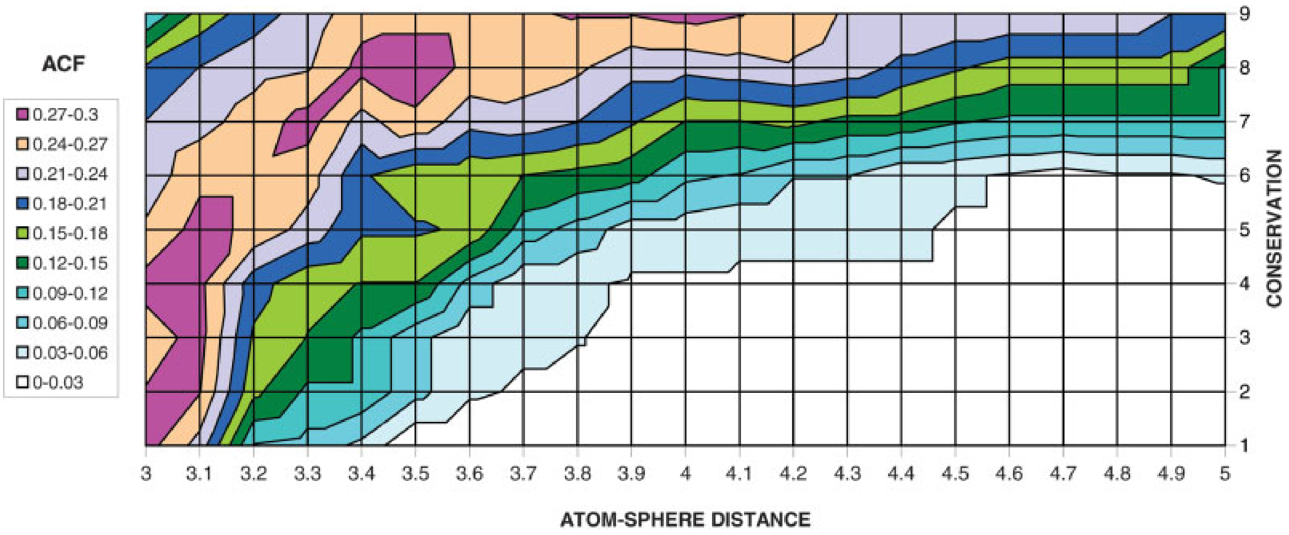

Fig. 1. The 3D-plot of the average calibration factor (ACF) for a subset of 30 entries (see underlined entries in Supplementary Table S.I) as a function of the two main parameters of the SURFNET-ConSurf method: atom–sphere maximum distance (ATOM–SPHERE DISTANCE axis) and conservation cutoff (CONSERVATION axis). The color-coding scale is from white for the lowest ACF values (bad performance), to orange and pink for the highest ACF values (good performance). The highest calibration factor is 0.2882, obtained with a distance of 3.5 Å and a ConSurf cutoff of 8.

reducing the number of spheres. Outside the orange and pink regions, the "distance effect" on the ACF value disappears quickly after the 3.1 Å, while the effect of the conservation is felt for almost any distance, giving increasingly high ACF values over almost any cutoff above 4. This "long range" effect of the conservation means that the volume reduction of the original SURFNET cleft occurs mainly in the regions where the ligand is not found, and explains why the use of conservation results in a significant reduction of the binding site, while still keeping most of the ligand (see below). Above distances of 3.8 Å and below conservation cutoffs of 4, the ACF values are very low, indicating that the cleft reduction is negligible.

Figure 1 demonstrates that the conservation cutoff helps to reduce the cleft size, while still keeping at least part of the ligand volume. The optimal combination of parameters, used for the rest of this work, is a distance cutoff of 3.5 Å and a conservation of 8 in the ConSurf-HSSP scale.

The Analysis of a Nonredundant Data Set of Protein Structures

Our data set comprised 244 nonredundant PDB entries of recorded biological quaternary structure, 112 of them enzymes (45.9%), 129 nonenzymes (52.9%), and three "hypothetical" (1.2%) proteins, according to PDBsum[31] and Uniprot[32] (enzymes are defined as proteins having an EC number in either database). These PDB entries contained 464 ligands of a minimum of seven atoms in size, not covalently bound to the protein. The ligands form a very heterogeneous set, including sugars, cofactors, substrate analogs, peptides, etc. The data set included the following distribution of multimers: 154 monomers, 76 dimers, five trimers, seven tetramers, and two pentamers.

SURFNET-ConSurf was run on the whole data set using the optimal parameters described above. Figure 2 shows the raw data for the ligand volume percentage encompassed by a cleft following the SURFNET stage (□) and after the full SURFNET-ConSurf run (●) for all 464 ligands in the data set. It also shows the cleft reduction percentage in each case (◇). The data shows that in most cases, when SURFNET identifies a ligand, the conservation filtering stage loses some of its volume, but in general, the loss is proportional to the percentage of ligand volume identified by SURFNET and the overall correlation is good. In very few cases in which the ligand is mostly included in one of the four SURFNET clefts the ligand is completely lost during the conservation filtering stage (see Discussion). Finally, the cleft reduction does not correlate with the ligand volume included in the cleft, and most values range between 20% and 60%.

In analyzing the data, the criterion we used to consider that a ligand binding site has been correctly "identified" is that at least 20% of the ligand's volume is included in one of the trimmed biggest four pockets produced by SURFNET-ConSurf. Using this criterion, the analysis of the data set (see Table SI; Fig. 2) shows that the method identifies correctly at least one ligand binding pocket in 75% of the PDB entries in the data set (183 from 244), and 57.5% of all ligands (267 from 464) [see Fig. 3(a)]. For the 267 cases in which the ligand site is successfully identified we can also calculate the percentage of the ligand volume included in the binding site pockets [Table SI; Fig. 3(b)]. The average ligand volume encompassed by a cleft at the initial SURFNET run is 94.6%, and drops to 87.1% after filtering by conservation, yielding an average loss of 7.5% of the ligand volume in this process [see Fig. 3(b)]. Finally, applying SURFNET-ConSurf reduces the original SURFNET cleft volumes by 30.2% on average [see Fig. 3(b)], although this reduction is much higher in many specific cases (the maximum reduction is 78.7%, see Table SI; Fig. 2).

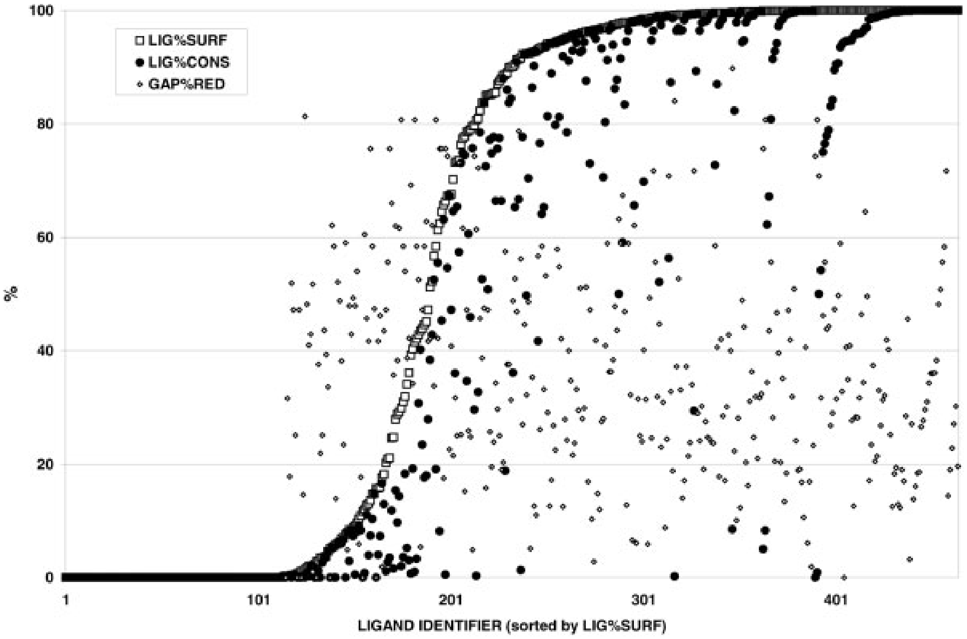

Fig. 2. LIG%SURF: The distribution of the percentage of ligand volume included within SURFNET pockets (□) for each ligand in the database. LIG%CONS: As LIG%SURF, for pockets trimmed by ConSurf-HSSP (●). GAP%RED: The reduction in cleft volume (◇). The data presented is sorted by LIG%SURF.

Separating the proteins in our data set into two subgroups containing enzymes and nonenzyme proteins respectively, the results differ markedly (see Fig. 3): Among the enzymes group, SURFNET-ConSurf identifies correctly at least one ligand binding pocket in 84.8% of the entries (95 from 112), while among the nonenzyme group the percentage is only 66.7% (86 from 129). In terms of the percentage of ligands successfully located, the success rate for enzymes is also better, since SURFNET-ConSurf locates 67.6% of the enzyme ligands (140 from 207) while only 49.0% are correctly located in the nonenzymes group (124 from 253). The average percentage of the ligand volume included in binding site pockets is 97.5% for enzymes and 91.8% for the nonenzymes at the initial SURFNET run, and drops to 91.9% for enzymes and 82.0% for nonenzymes after the ConSurf filtering stage, yielding a loss of 5.6% for enzymes and 9.8% for nonenzymes in the process [see Fig. 3(b)]. Finally, applying SURFNET-ConSurf reduces the original SURFNET cleft volumes similarly in enzymes and nonenzymes [31.2% and 29.4% respectively, see Fig. 3(b)].

These results show that although we can apply the SURFNET-ConSurf method to identify functional sites in both enzymes and nonenzyme proteins, the rate of success is higher for the first group. A more detailed analysis of the difference between the conservation and ligand identification for enzymes and nonenzymes is currently in progress (de Miguel et al., in preparation).

The Clefts

We can also analyze the data in terms of the clefts volume, shape, etc. Overall, there are a total of 276 original SURFNET clefts which include at least some portion of a ligand, and from these, 226 trimmed SURFNET pockets that include at least 20% of the volume of a valid ligand. This means that 82% of the trimmed clefts are correctly identified as ligand binding pockets.

Table SI contains several features describing the shape and size of the binding sites which are also shown in Figure 4(a–d). The graphs are based on the analysis of the 226 binding pockets which include at least 20% of the volume of a valid ligand. Figure 4(a) shows the distribution of clefts containing located ligands. Of the 226 ligand binding pockets, 55.3% (125) are found in the largest cleft, 25.2% (57) in the second largest cleft, 12.4% (28) in the third, and 7.1% (16) in the fourth largest cleft. Notably, the first and second pockets (i.e., clefts 1 and 2) include more than 80% of the binding sites. Figure 4(b) shows the volume distribution of SURFNET clefts before and after the trimming step (VOLSURF, VOLCONS columns in Table SI, respectively). This figure demonstrate that the majority (179 of 226, 79%) of the original SURFNET clefts have a volume under 5000 Å^3 and that the most common volumes cleft are between 1000–3000 Å^3, which together include 97 clefts (43%). After the trimming procedure, the distribution moves slightly to the left, as expected, so that 91% (206 from 226) of the trimmed SURFNET pockets now have a volume under 5000 Å^3 and 23% less than 1000 Å^3. Figure 4(c) shows the distribution of the numbers of clusters following the ConSurf-HSSP trimming procedure (NoCLUST column in Table SI). The great majority of ligand pockets remain in one trimmed pocket (62%) or are split into 2 or 3 smaller clusters (21% and 8% respectively, see i.e., Fig. 5), probably indicating that the majority of the conserved residues on a ligand binding pocket are distributed uniformly around the ligand binding pocket, or clustered in a few large regions.

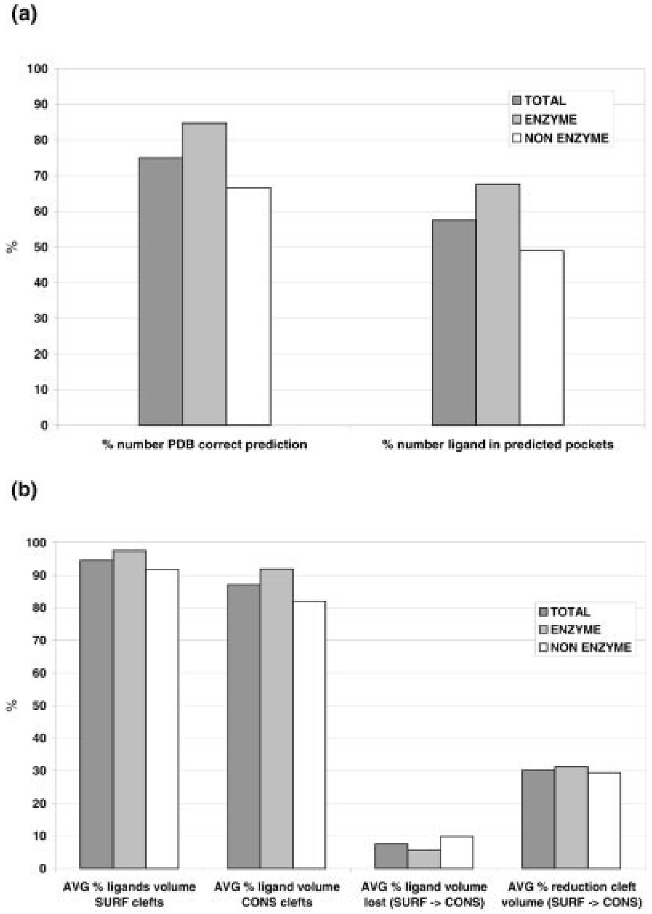

Fig. 3. The data set analysis, for three different groups: TOTAL: statistics calculated using all the PDB structures and their ligands in the data set. ENZYME: as TOTAL for the enzymes in the data set. NON ENZYME: as TOTAL for the non-enzyme proteins in the data set. **a**: Success values. "% number PDB correct prediction" is the percentage of PDB entries including at least one ligand in one of the four biggest trimmed SURFNET-ConSurf clefts. "% number ligand in predicted pockets" is the percentage of ligands out of 464 for which 20% or more of their total volume are included in one of the four biggest trimmed volumes. **b**: Volume reduction values. "AVG % ligands volume SURF clefts" is the average percentage of ligands volume included in the four biggest clefts produced by the SURFNET program. "AVG % ligand volume CONS clefts" is the average percentage of ligands volume included in the four biggest trimmed clefts. "AVG % ligand volume lost (SURF → CONS)" is the average ligand volume lost during the trimming procedure. "AVG % reduction cleft volume (SURF → CONS)" is the average volume cleft reduction of clefts including a ligand during the trimming procedure.

It is more difficult to estimate the shape of the clefts, since there is not a simple measure that will allow us to compare the shape of each pocket. We have used the ratio between the first and second principal axes to make some distinction between the different binding pockets. Ratios closer to 1.0 indicate a more sphere-like binding pocket. Pockets that are more "elongated" will have a larger ratio. Figure 4(d) shows that the distribution of this ratio before and after the trimming step barely changes (PRINSURF, PRINCONS columns on Table SI, respectively). Both PRINSURF and PRINCONS have an average ratio of 1.6, indicating a slight tendency towards elongated clefts (i.e., deep or wide clefts as opposed to rounded clefts).

Overall, the SURFNET-ConSurf results for the 244 nonredundant PDB structures shows that when a ligand is found in one of the four largest clefts on the protein surface, most of its volume remains within the reduced pocket. The results also clearly show that the conservation filtering stage (using ConSurf-HSSP scores) is able to trim the volume of the cleft found by SURFNET in most cases (see Fig. 1 and 2 and Table SI), leading to the reduction of 30% of its original volume on average.

A Specific Example

It is interesting to examine how closely the trimmed cleft volume represents the actual ligand shape for a specific example, since identifying the shape of the binding portion of the cleft could help predict the type of ligand that the protein binds. The example we present here is PDB entry 1ics (Table SI; Fig. 5). The largest cleft is reduced by 64.5% as a result of the ConSurf-HSSP filtration process, while the trimmed cleft volume still includes 96.9% of the ligand [see Fig. 5(f)]. But does this volume reduction produce a new pocket that represents the ligand shape and size more accurately?

PDB 1ics holds the structure of 12-oxophytodienoate reductase 1 (OPR1) from tomato.[33] This protein is involved in the biosynthesis of jasmonic acid (JA), catalyzing the NADPH-dependent reduction of the cyclopentenone 12-oxophytodienoate (OPDA) to 3-oxo-2(2′9[Z]-pentenyl)-cyclopentane-1-octanoate (OPC-8:0). Figure 5(a–e) presents the front and back surface of a monomer of OPR1 in complex with its cofactor, a flavin mononucleotide (FMN).[33] Figure 5(a,b) shows the four largest pockets found on the surface of this monomer by SURFNET. The largest pocket [in yellow mesh in Fig. 5(a)] includes the binding site of FMN, which accommodates also the substrate (see below). But this pocket is clearly much larger than the actual binding site pocket (its volume being 10.7 times the cofactor volume), thus it does not define the shape and location of the cofactor very well. Figure 5(c,d) shows the OPR1 protein color-coded by conservation scores from the ConSurf-HSSP database. The predicted conservation around the ligand seems to agree well with experimental data, since most residues known to form critical hydrogen contacts with FMN through their side chains are highly conserved (e.g., Arg239, Pro35, Thr37, and Gln 110), as well as the most important active site residues involved in substrate binding and enzymatic activity (i.e., Gln 110, Trp112, Tyr192, His187, and His190).[33] However, several other regions are found to be conserved [maroon-colored patch 2 in Fig. 5(c) and patches 3 and 4 in Fig. 5(d)], and it is not possible to differentiate between the binding site cleft and other functional sites using conservation alone.

Figure 5(e) shows the trimmed volume of the biggest cleft obtained from the SURFNET-ConSurf calculation (yellow mesh) and the protein as in Figure 5(c), while Figure 5(f,1) shows a superposition of the original SURFNET biggest pocket (transparent pink surface) and the trimmed volume obtained after the filtering stage [green mesh, isolated from Fig. 5(e)]. It is evident from Figure 5(e;f,1) that the reduced cleft is much smaller than the

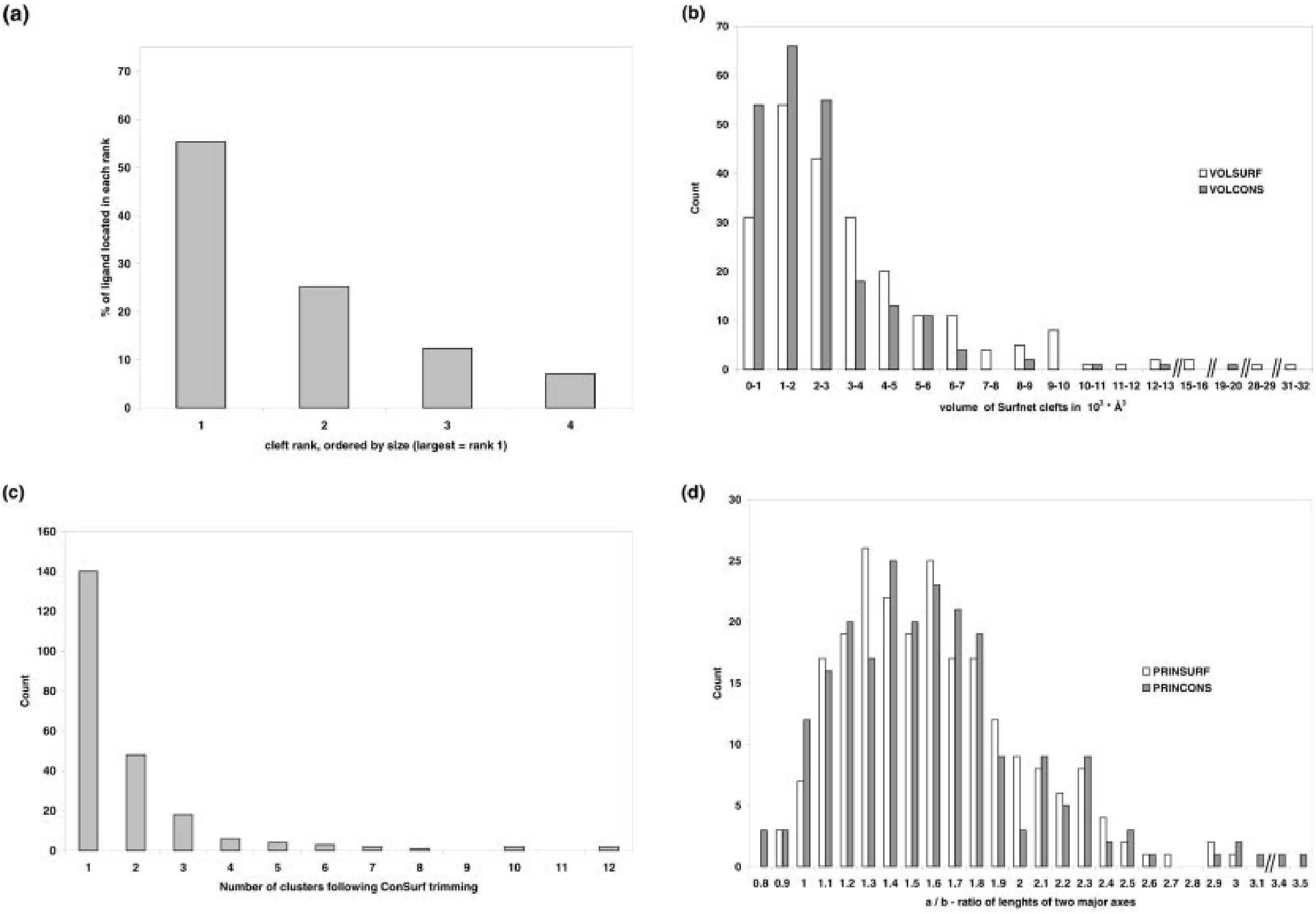

Fig. 4. The cleft analysis. **a**: The distribution of the number of ligands located in each of the largest four trimmed clefts. **b**: The distribution of volumes of the Surfnet clefts, before (VOLSURF) and after the trimming procedure (VOLCONS). **c**: The distribution of the numbers of clusters into which the original cleft is split following the ConSurf-HSSP trimming procedure. **d**: The distribution of the ratio between the two major cleft axes before and after the trimming step (PRINSURF, PRINCONS, respectively).

original, its volume being now 3.8 times the volume of the FMN cofactor, thus corresponding more closely to its shape and location. However, Figure 5(f,1) also shows that the trimmed volume still has significant extra space available, which may suggest the presence of an additional ligand (e.g., the substrate) in the same cleft.

To check this hypothesis we also ran SURFNET-ConSurf for PDB structure 1icq (not in the original data set), which corresponds to the same molecule (OPR1) crystallized with both the FMN cofactor and the substrate 9R,13R-OPDA. Figure 5(f,2) shows the volume of the main SURFNET cleft for this structure, including both the substrate and the cofactor (in yellow and magenta CPK representation, respectively), showing that the biggest central cluster of the trimmed volume (in green mesh) matches very well the sum of the volumes of both molecules (including 99.3% and 92.1% of the cofactor and substrate volumes, respectively), giving a much better localization, and explaining why we still had extra space left after the cleft reduction in Figure 5(f,1). Figure 5(f,2) also shows that the trimmed cleft misses a small portion of the substrate (less than 8%). The substrate portion lost during the conservation filtering stage is surrounded by the highly variable "hydrophobic tunnel" loops responsible for substrate specific recognition and binding,[33] and then it is "lost" by applying the ConSurf-HSSP conservation filtering.

This example shows that the reduction of the initial cleft volume produced by SURFNET can be biologically relevant, although it also demonstrates clearly one of the main limitations of SURFNET-ConSurf, which is not capable of identifying the parts of the ligands that are not surrounded by highly conserved residues. Also, in the case that more than one molecule is included in the trimmed volume, still remains the problem of differentiating between the shapes of each (see Discussion). In general, the shape and volume of the trimmed cleft could be modeled and compared with others clefts, thus obtaining information about the type of ligand present in both proteins. Finally, it is worth mentioning that in this example SURFNET-ConSurf performs particularly well, while in many other cases the reduced cleft volume does not enclose the whole ligand(s) (see Table SI), thus making it even more difficult the task of "guessing" which ligand is bound to the protein.

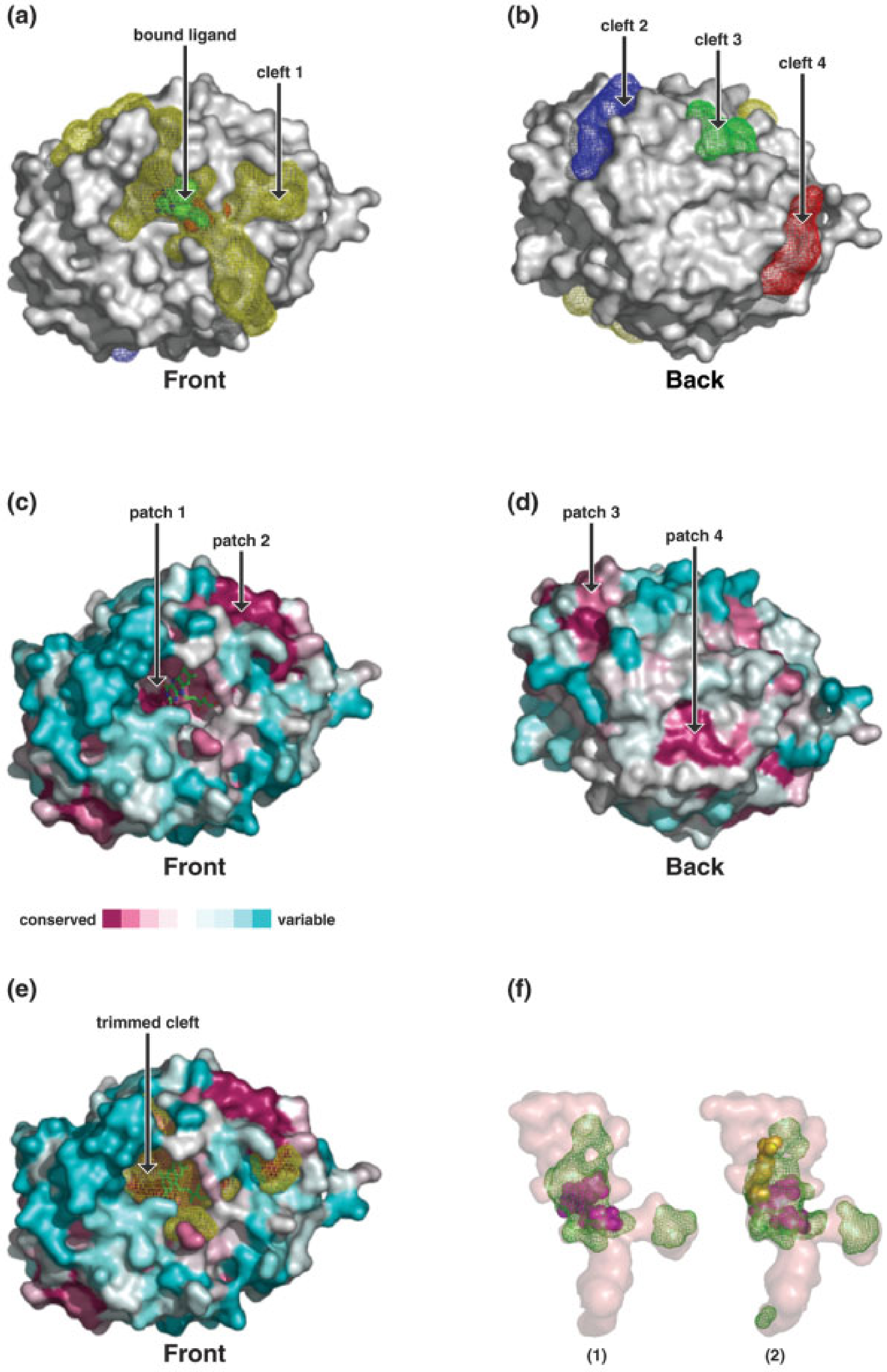

Fig. 5. The representations of the surface of the X-ray structure of 12-oxophytodienoate reductase 1, in complex with its cofactor [PDB 1ics, a to f(1)] and substrate [PDB 1icq, f(2)].[33] All figures except (f) show the surface representation of the protein chain. The figures in (c), (d), and (e) are color-coded by conservation using the scale under (c). In (a), (b), (e) and (f) the colored mesh represents the volume enclosing the SURFNET spheres on the cleft. **a**: The side of the protein showing the site where the FMN binds and cleft 1 (the largest cleft found by SURFNET). **b**: The back of the protein showing the other three major clefts. **c**: The front of the protein, color coded by conservation. The main two conserved patches on this face are indicated. **d**: As (c), but showing the back part of the protein. **e**: Front of protein with cleft 1 of figure (a) reduced in size on application of the conservation filter (see Methods). **f**: Frame (1): Volume of cleft 1 from (a) (transparent pink surface) and inside it volume of reduced cleft from (e) (in green mesh, two clusters). The FMN ligand is shown in magenta CPK representation. Frame (2): Like Frame (1) but for PDB 1icq. The Volume of cleft 1 (transparent pink surface) including the volume of reduced cleft (in green mesh, three separated clusters) for the same protein in complex with both the substrate (in orange CPK) and the FMN cofactor (in magenta CPK).

DISCUSSION

We present in this work SURFNET-ConSurf, a method for identifying the location and shape of ligand binding sites in proteins by combining two independent algorithms. First, the four largest clefts on the protein surface are calculated using SURFNET. Second, residue conservation as predicted by ConSurf-HSSP is used to trim those volumes, by keeping only those SURFNET spheres which are close to conserved residue atoms. SURFNET-ConSurf is applied in this work to a nonredundant data set of multimers, for which the biologically active quaternary structure is predicted. The use of the correct biological unit in this work is of prime importance since in many cases binding sites are found at the interface formed by more than one polypeptide chain.[10] Our method is able to reduce the size of the ligand binding site pocket volume by 30% on average, while retaining the majority of ligand atoms in the reduced volume. The analysis of binding-site clefts including ligands indicates that the majority are the first and second biggest clefts in the protein, having a volume of less than 5000 Å^3 and displaying an elongated rather than spherical shape.

One of the main limitations of SURFNET-ConSurf is that the evolutionary conservation scores provided by ConSurf-HSSP represent a "sequence-family average measure" of the physico-chemical determinants that makes a binding site specific towards a group of substrates. However, any given family member will have distinct physico-chemical features to make it specific toward its own substrate. Thus, this method cannot currently be used to accurately differentiate or compare members of the same family. One could, however, use a similar idea of excluding the identical parts of the binding pocket within a given family and focusing in this case on the variable residue regions. SURFNET-ConSurf will also fail to accurately identify the shape and volume of ligands in cases where the binding site includes more than one ligand molecule (i.e., cofactor and substrate in enzymes, see Results and Fig. 5), since the volume and shape of the binding site will, in the best case, represent a combination of the volume and shape of the ligands occupying it.

There are cases where SURFNET-ConSurf fails at either stage of the procedure (i.e., the ligand binding pocket is completely missed, see Table SI). The reasons for the failures are diverse: for example, the methodology fails when the ligand is mainly buried in the protein core (e.g., FS4 cluster ligand in PDB entry 1h2r) or when the ligand is found in a relatively shallow or small pocket on the protein surface. In both cases the binding site is generally not one of the four biggest volumes found by SURFNET. There are also a few cases in which the ligands that are identified by SURFNET but totally lost after the conservation filtering stage (see Fig. 1; Table SI). In these cases, the binding site residues have a low conservation, which can be caused by several reasons, like enzymes with homologs that perform different functions,[34] enzymes with more than one binding site, poor alignment or lack of data, etc.

Several improvements can be introduced to add or combine additional atom or residue properties, like hydrophobicity, charge, etc. in the filtering stage. In principle, it may be possible to cluster the spheres predicted by SURFNET in different local high-density clusters. The existence of more than one local cluster inside the same pocket may be indicative of a local higher degree of freedom accommodating more than one molecule. Initial trials have shown that this procedure could also further reduce the number of filtered spheres created by SURFNET. The ultimate goal of SURFNET-ConSurf and related methods is to predict the type of ligand bound to a specific binding site, thus enhancing the understanding of the biology of the protein and its family. One of the ways to achieve this goal may be the application of a shape algorithm to the reduced cleft volume, and then compare the shape of the binding pockets of a series of proteins, identifying those belonging to the same family, or ideally, those binding the same type of ligand.[35]

Although SURFNET-ConSurf does not yield the exact shape or type of the ligand bound to a protein, it improves over the existing methods and can become the basis of further improvements. Not only are the SURFNET clefts trimmed to regions where the ligands are most likely to be found, but also the smaller and nonfunctional pockets are greatly reduced. This means that SURFNET-ConSurf can be also seen as a better method for the correct identification of the binding site of a protein, than either SURFNET or ConSurf used independently. Finally, we are currently analyzing in detail the significant differences between the enzymes and nonenzymes protein subgroups. The differences seem to be due to a combination of a higher degree of conservation and larger binding sites of the enzyme group (de Miguel et al, in prep.). One potential outcome of this analysis could be to develop a different strategy for identifying enzyme and nonenzyme binding sites.[36]

As mentioned in the introduction, there are several problems in moving from here towards ligand identification based on the structure (i.e., shape and volume) of the binding site. For example, the predicted trimmed volume of the binding site is still much too large, in many cases being several times the volume of the ligand occupying it (data not shown). It is plausible to argue that although the ligand does not "use" all the volume available, the spare volume is necessary to "lead" the ligand into the correct location and exact binding position within the binding pocket. In some cases though, the larger cleft volume could be partially explained if several different molecules (not always present in the crystal) occupy the same binding site simultaneously (e.g., cofactor plus substrate). In those cases the volume of the ligand binding cleft will inevitably be larger than that of an individual ligand (see Fig. 5). This makes it difficult to fully identify the ligand type based only on the shape and the volume of the predicted binding cleft.

It should also be mentioned that in this work we are considering each and every ligand in the crystal structures, without manually checking that the ligands are "biologically relevant." Lastly, many protein families have evolved to include multiple members with different specificities. Therefore averaging over all homologs will cause confusion. Thus, different approaches will be needed to handle all these challenges. However, the main outcome of this work is that the simple combination of conservation and binding site volume provides a reasonable improvement towards this goal.

ACKNOWLEDGMENTS

F.G. and R.J.N. are funded by the Structural Genomics Consortium (SGC) and the European Molecular Biology Laboratory (EMBL). The SGC is a public private charitable partnership that receives funds from Canada Foundation for Innovation, Canadian Institutes for Health Research, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, the Ontario Innovation Trust, the Ontario Research and Development Challenge Fund and the Wellcome Trust. R.J.M. was supported by SPINE (EU contract number QLG2-CT-2002-00988).

We also acknowledge the BioSapiens network of excellence (contract number LHSG-CT-2003-503265). We wish to thank Jonathan Barker, Gareth Stockwell and Tom Funkhouser for useful comments and technical assistance.

REFERENCES

1. Goldsmith-Fischman S, Honig B. Structural genomics: computational methods for structure analysis. Protein Sci 2003;12:1813–1821.
2. Shapiro L, Harris T. Finding function through structural genomics. Curr Opin Biotechnol 2000;11:31–35.
3. Watson JD, Todd AE, Bray J, Laskowski RA, Edwards A, Joachimiak A, Orengo CA, Thornton JM. Target selection and determination of function in structural genomics. IUBMB Life 2003;55:249–255.
4. Stark A, Shkumatov A, Russell RB. Finding functional sites in structural genomics proteins. Structure (Camb) 2004;12:1405–1412.
5. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–1146.
6. Bell RE, Ben-Tal N. In silico identification of functional protein interfaces. Comp Funct Genomics 2003;4:420–423.
7. Innis CA, Anand AP, Sowdhamini R. Prediction of functional sites in proteins using conserved functional group analysis. J Mol Biol 2004;337:1053–1068.
8. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci US A 2004;101:14754–14759.
9. Pupko T, Bell RE, Mayrose I, Glaser F, Ben Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18 Suppl 1:S71–S77.
10. Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K, Vajda S. Identification of substrate binding sites in enzymes by computational solvent mapping. J Mol Biol 2003;332:1095–1113.
11. Sotriffer C, Klebe G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Farmaco 2002;57:243–251.
12. Mason K, Patel NM, Ledel A, Moallemi CC, Wintner EA. Mapping protein pockets through their potential small-molecule binding volumes: QSCD applied to biological protein structures. J Comput Aided Mol Des 2004;18:55–70.
13. 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.
14. Brady GP Jr, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14:383–401.
15. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5:2438–2452.
16. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
17. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–164.
18. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
19. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 2005.
20. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–328.
21. Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben Tal N. The ConSurf-HSSP database: The mapping of evolutionary conservation among homologs onto PDB structures. Proteins 2004;58:610–617.
22. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 2005.
23. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
24. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
25. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 2004;32: W20–W25.
26. Chayen NE, Saridakis E. Protein crystallization for genomics: towards high-throughput optimization techniques. Acta Crystallogr D Biol Crystallogr 2002;58:921–927.
27. Galkin O, Vekilov PG. Control of protein crystal nucleation around the metastable liquid-liquid phase boundary. Proc Natl Acad Sci USA 2000;97:6277–6281.
28. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA, Wormald MR. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. Glycobiology 2004;14:103–114.
29. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
30. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
31. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res 2005;33 Database Issue:D266–D268.
32. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33 Database Issue: D154–D159.
33. Breithaupt C, Strassner J, Breitinger U, Huber R, Macheroux P, Schaller A, Clausen T. X-ray structure of 12-oxophytodienoate reductase 1 provides structural insight into substrate binding and specificity within the family of OYE. Structure (Camb) 2001;9:419–429.
34. Argiriadi MA, Morisseau C, Goodrow MH, Dowdy DL, Hammock BD, Christianson DW. Binding of alkylurea inhibitors to epoxide hydrolase implicates active site tyrosines in substrate activation. J Biol Chem 2000;275:15265–15270.
35. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 2005;21(10):2347–2355. E-pub 22 Feb 2005.
36. Bate P, Warwicker J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. J Mol Biol 2004;340:263–276.