PROTEINS: Structure, Function, and Bioinformatics 58:610–617 (2005)

The ConSurf-HSSP Database: The Mapping of Evolutionary Conservation Among Homologs Onto PDB Structures

Fabian Glaser,[1†] Yossi Rosenberg,[1] Amit Kessel,[1] Tal Pupko,[2] and Nir Ben-Tal[1*]
[1]*Department of Biochemistry, Tel Aviv University, Tel Aviv, Israel*
[2]*Department of Cell Research and Immunology, Tel Aviv University, Tel Aviv, Israel*

ABSTRACT **The HSSP (Homology-Derived Secondary Structure of Proteins) database provides multiple sequence alignments (MSAs) for proteins of known three-dimensional (3D) structure in the Protein Data Bank (PDB). The database also contains an estimate of the degree of evolutionary conservation at each amino acid position. This estimate, which is based on the relative entropy, correlates with the functional importance of the position; evolutionarily conserved positions (i.e., positions with limited variability and low entropy) are occasionally important to maintain the 3D structure and biological function(s) of the protein. We recently developed the Rate4Site algorithm for scoring amino acid conservation based on their calculated evolutionary rate. This algorithm takes into account the phylogenetic relationships between the homologs and the stochastic nature of the evolutionary process. Here we present the ConSurf-HSSP database of Rate4Site estimates of the evolutionary rates of the amino acid positions, calculated using HSSP's MSAs. The database provides precalculated evolutionary rates for nearly all of the PDB. These rates are projected, using a color code, onto the protein structure, and can be viewed online using the ConSurf server interface. To exemplify the database, we analyzed in detail the conservation pattern obtained for pyruvate kinase and compared the results with those observed using the relative entropy scores of the HSSP database. It is reassuring to know that the main functional region of the enzyme is detectable using both conservation scores. Interestingly, the ConSurf-HSSP calculations mapped additional functionally important regions, which are moderately conserved and were overlooked by the original HSSP estimate. The ConSurf-HSSP database is available online (http://consurf-hssp.tau.ac.il). Proteins 2005;58:610–617.** © 2004 Wiley-Liss, Inc.

Key words: evolutionary rate; amino acid conservation; protein evolution; ConSurf; phylogeny; Rate4Site; pyruvate kinase

INTRODUCTION

The HSSP database of Homology-Derived Secondary Structure of Proteins[1] provides multiple sequence alignments (MSAs) of homologous proteins for each protein of known three-dimensional (3D) structure in the Protein Data Bank (PDB).[2] For each HSSP entry, the homologous proteins are added using a stringent threshold, corresponding to the minimal sequence identity required for considering two proteins as structural homologs (see Methods section).[1,3] The procedure guarantees that, by and large, the collected homologous proteins share a similar 3D-fold and related biological function(s). The HSSP database also includes a "variation entropy" score for each amino acid position in the protein, which is a measure of sequence variability. The variation entropy is an estimate of the Shannon information content,[4] and is often used to detect structurally and functionally important positions in the protein.[5–7] It is calculated based on the amino acid frequencies at each position within the homologous proteins; variable positions have high entropy, and conserved positions have low entropy. Evolutionarily constrained amino acid positions, which are commonly referred to as "evolutionarily conserved" (or simply "conserved"), are often important to maintain the 3D structure of the protein or its biological function(s). Thus, the identification of these positions enhances our understanding of the biological function of the protein.[8–11]

The variation entropy is a straightforward and clear definition of variability that is easy to calculate but has several drawbacks. First, it does not differentiate between moderate (e.g., Leu to Ile) and radical (e.g., Leu to Asp) replacements. Second, it does not take into account the sequence redundancy (i.e., the nonhomogenous nature of the MSA), and finally, it does not take into consideration the evolutionary relationships among the homologs in the MSA.[1,5,7,12] The Rate4Site algorithm,[13] which we recently introduced, provides a more accurate approach for scoring amino acid conservation. Rate4Site accepts as input a phylogenetic tree reconstructed from the MSA of the homologous sequences and provides a maximum likelihood estimate of the evolutionary rate of each amino acid

Grant sponsor: Research Career Development Award from the Israel Cancer Research Fund (to Nir Ben-Tal). Grant sponsor: Complexity Science Grant from the Yeshaia Horvitz Association (to Tal Pupko).

†Current address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

*Correspondence to: Nir Ben-Tal, Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel. E-mail: bental@ashtoret.tau.ac.il; Web: http://ashtoret.tau.ac.il/

Received 8 February 2004; Accepted 16 July 2004

Published online 21 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20305

position. The evolutionary relationships among the homologous proteins, as reflected in the tree topology and branch lengths, and the amino acid frequencies and replacement probabilities, are explicitly taken into account in the calculation. This makes Rate4Site more accurate and sensitive than other advanced methods that explicitly use phylogenetic trees[14,15] for the identification of the residues that compose functional regions.[11,13] Subsequently, we developed the web server ConSurf,[16] which implements this algorithm, enabling the identification of functionally important regions in proteins by mapping the level of evolutionary conservation of each position onto the 3D structure of a representative protein.

Our experience has been that the result of a Rate4Site calculation is highly sensitive to the quality of the MSA,[13] and that the automatic approach of producing an MSA used in ConSurf (i.e., a PSI-BLAST[17] search for homologs followed by CLUSTAL W[18] alignment) is not always reliable (e.g., true homologs are sometimes overlooked by the PSI-BLAST search, while nonhomologs are falsely detected). To alleviate this problem, we have developed ConSurf-HSSP, a web-available database that combines the ConSurf server with premade MSAs taken from the HSSP database. The ConSurf-HSSP database currently provides precalculated conservation scores for essentially the entire PDB.

METHODS

Relative Entropy Scores

Sequence variability can be estimated based on Shannon's information content.[1,5,7] Given the frequency of occurrence $f(r)_i$ of an amino acid of type r at position i in the alignment, the information content or entropy $[-\Sigma f(r)_i \ln f(r)_i]$ is a measure of the uniformity of the distribution $f(r)_i$. The relative entropy of the position (S_i) is obtained by multiplying the entropy by (100/ln 20). This normalization produces entropy scores between 0 and 100, corresponding to the most highly conserved and most variable residues[1,7]:

$$S_i = -(100/\ln 20)\,\Sigma f(r)_i \ln f(r)_i \qquad (1)$$

The HSSP Database

In order to produce each entry of the ConSurf-HSSP database, we used MSA files obtained from version 1.1 2001 of the HSSP database (http://www.cmbi.kun.nl/swift/hssp/).[1] This database contains a file for each PDB entry, with an MSA for each nonredundant protein chain.

The procedure for collecting "true" homologs for the HSSP alignments ensures some degree of structural homology with the target protein. An empirical homology threshold is used, which was determined by analyzing the alignment of many proteins of known structure in terms of the relation of their sequence and structural similarities, and their dependence on the alignment length. A protein is considered to be a structural homolog if the sequence similarity is equal to or greater than a homology threshold t(L), along a sequential region of length L. t(L) varies from about 25% (residues identity) for sequences of 80 residues or more, up to about 80% for sequences of 10 residues.[1] Each file also includes relative entropy based scores [Eq. (1)] of the variability of each position in the alignment, which is stored in the "SEQUENCE PROFILE AND ENTROPY" section.

The Evolutionary Rate

The ConSurf-HSSP database makes use of the Rate4Site algorithm[13] to calculate the rate of evolution at each amino acid position. In Rate4Site, the evolutionary conservation of a position is scored, based on its evolutionary rate; slowly evolving positions are evolutionarily conserved, while rapidly evolving positions are variable. Rate4Site, which is also suggested as the default method for scoring conservation in the ConSurf server,[16] closely models the evolutionary process by explicitly taking into account the phylogenetic relationships among the homologous sequences and thus overcomes problems due to uneven sampling in sequence space. Rate4Site also integrates into the calculation the frequencies of the naturally occurring amino acids and their replacement probabilities based on the JTT model.[19]

A Rate4Site calculation begins with the reconstruction of a phylogenetic tree based on the input MSA, using the Neighbor Joining (NJ) method.[20] It is important to emphasize that the tree is built to be consistent with the MSA only, without explicit use of the origin of the proteins. Thus, orthologs and paralogs are treated in the same manner, in that their mutual "evolutionary distances" are calculated based on the number of amino acid replacements. To increase the accuracy of the NJ algorithm, pairwise distances were computed using the maximum likelihood paradigm.[21,22] The evolutionary rate at each position is subsequently estimated, based on the tree topology and branch lengths. A more detailed explanation of the Rate4Site algorithm can be found in Pupko et al.[13] and in the OVERVIEW section of the ConSurf-HSSP database (http://consurf-hssp.tau.ac.il).

The ConSurf-HSSP Database

For each entry, the ConSurf-HSSP database stores a web interface produced by the ConSurf server, and the corresponding set of input and output files. The web interface provides links to the output files that were produced during the ConSurf/Rate4Site calculation, including a file in which the estimated evolutionary rates of each amino acid positon are recorded. Most important, it provides the means for visualization of the color-coded conservation scores on the selected polypeptide chain using Protein Explorer[23] (see the OVERVIEW page at http://consurf-hssp.tau.ac.il for more information).

The ConSurf-HSSP methodology requires a minimum of 5 homologous sequences for a ConSurf calculation. This cutoff reflects our experience that MSAs, including a smaller number of homologs, do not usually contain enough information to yield accurate conservation scores. This cutoff will ultimately be replaced by a measure of sequence variability.

The Color Code Scale

The color-coding scale used both in the ConSurf-HSSP database and relative entropy analysis was adopted from the ConSurf server. This is a discrete scale obtained by dividing the conserved and variable portions of the original conservation scores distribution into 4.5 identical intervals. Thus, a new scale of 9 equal-size categories of conservation is obtained. Applying this scaling procedure to both the ConSurf-HSSP and relative entropy scales yields a comparable color grades scale for both methods, which was used for coloring and grading conservation [see ConSurf-HSSP and RELENT columns in Table I and Fig. 1(a and b)]. In this work, residues with conservation grades of 9 or 8 were considered highly conserved; residues with grades 7 or 6 were considered conserved; residues with grade 5 were considered averagely conserved; residues with grades 4 or 3 were considered variable; and residues with grades 2 or 1 were considered highly variable.

Contact Projection

A relative exposure criterion was used for the identification of the residues that participate in protein interfaces [Fig. 1(c)]. The contact projection of a polypeptide chain (or domain) that is involved in an interprotein (or interdomain) interface was obtained by calculating the relative solvent accessibility of each of its residues (RSA_i) and their projection onto the protein structure. RSA_i was defined as

$$RSA_i = \Delta ASA_i / ASAm_i \quad (2)$$

$ASAm_i$ is an estimate of the maximum solvent accessibility of residue i in a protein environment, obtained by calculating the solvent-accessible surface area of the residue in an extended AX_iA tripeptide, where A is Ala and X_i marks the type of residue i. ΔASA_i is the difference in the Connolly solvent accessible surface area of residue i within the context of the protein upon burial at the interprotein or interdomain interface. The relative exposure (or contact) map was obtained by color-coding each residue by RSA_i, using a red-through-blue rainbow scheme indicating the maximum-through-minimum relative solvent accessibility scale [Fig. 1(c)]. The solvent accessible surface area was calculated using the SURFV program,[32] with a probe sphere of radius of 1.4 Å and default parameters.

Pyruvate Kinase

The pyruvate kinase structure of PDB entry 1a49 and the corresponding HSSP file of 188 homologous sequences (1a49.hssp) were used as the input to the ConSurf web server. The ConSurf output files can be accessed through the ConSurf-HSSP URL database by selecting the PDB ID 1a49 and chain A.

TABLE I. Key Functional Residues in Rabbit Muscle PK[24–31] and the Conservation Grades Assigned to These Residues

Function[a]	Residue[b]	ConSurf-HSSP[c]	RELENT[d]
Opening and closing active site	Pro116	9	9
	Lys223	8	7
	Phe243	9	9
Mg^{2+} binding	Glu271	9	9
	Asp295	9	9
K^+ binding	Thr113	9	7
	Ser76	9	9
	Asp112	9	9
	Asn74	9	9
	Ser242	9	9
Mg^{2+} or ATP binding	Lys206	9	8
	Arg119	9	9
	Asp177	9	7
ATP binding	His77	9	9
	Pro52	9	9
	Tyr82	5	4
	Arg72	9	9
	Asn74	9	9
	Arg119	9	9
	Lys366	7	3
Catalysis	Lys269	9	9
	Ser361	9	9
	Thr327	9	9
	Arg72	9	9
	Glu363	9	9
Allosteric regulation	Ser402	1	2
	Thr340	9	9
Putative FBP binding	Thr431	9	8
	Ser436	8	6
	Glu432	3	3
	Arg454	6	3
1–2 interface	Lys421	3	3
	Tyr443	6	5
	Glu409	5	5
1–3 interface	Asp177	9	7
	Arg341	9	9

[a]The reported functional role of each residue.
[b]The residue type and number.
[c]The ConSurf-HSSP conservation grades.
[d]The relative entropy conservation grades.

RESULTS

The ConSurf-HSSP Database

The current version of the ConSurf-HSSP database covers, in essence, the entire PDB. We experienced a few failures due to an insufficient number of homologous sequences (less than 5; see Methods section) in the HSSP alignment, since the PDB or HSSP entry included nonstandard characters or the HSSP entry was missing.

In the following section, we provide an in-depth analysis of the conservation pattern obtained for the structure of rabbit muscle pyruvate kinase[24] in order to demonstrate the capacity of the ConSurf-HSSP database. A comparison

of the ConSurf-HSSP calculations to the results obtained using HSSP's relative entropy follows.

Pyruvate Kinase

Pyruvate kinase (PK), which is found in all living organisms, catalyzes the conversion of phosphoenolpyruvate (PEP) to pyruvate, with the concomitant phosphorylation of adenosine diphosphate (ADP) to adenosine triphosphate (ATP) in the final step of glycolysis, requiring both magnesium and potassium ions for its activity.[24–26] There are 4 PK isoenzymes in mammals: M1 (muscles), M2 (kidney and lungs), L (liver), and R (blood cells). All of them are tetramers of identical subunits of about 500 amino acid residues, each of which is composed of 4 domains: A, B, C, and N [Fig. 1(a)]. The available data indicate that, while M1 is not allosteric, isoenzymes M2, L, and R show a complicated allosteric-regulation mechanism of PK's enzymatic activity,[33] which involves drastic intradomain and intersubunit rotational movements and a reaccommodation of each subunit within the tetramer.[27–30]

To validate the ConSurf-HSSP results, we looked at the conservation scores assigned to residues that are known to be functionally important. Since the conservation scores are calculated using the HSSP alignment, which includes many types of PK isoenzymes (mammals M1, M2, R, and L, and bacterial types I and II, etc.), the scores reflect the functional important characteristics common to all PK types. Table I presents a list of the 32 residues that are known to be functionally important based on experimental data.[25,28,29] These residues are involved in a variety of critical PK activities, such as opening and closing of the active site and coordination of the ligands. Twenty-four of these residues (75%) received high ConSurf-HSSP conservation grades of 8 or 9, as they should.

The active site and the main patch of conserved residues

We also examined the ConSurf-HSSP results obtained for PK by visual inspection of the pattern generated by projecting the conservation scores on the 3D structure of the enzyme. The asymmetric crystallographic units of the structure of rabbit muscle PK contain 2 tetramers, having a total of 8 identical subunits, each formed by 4 domains [Fig. 1(a)]: N (1–42), A (43–115 and 224–387), B (116–223), and C (388–530).[24] Evolutionary conservation was mapped onto the surface of subunit 1 of that structure. A large patch of highly conserved residues [Fig. 1(a)], referred to as the "main patch," and several smaller patches of conserved residues [Fig. 1(d)] were detected.

The main patch is a region composed of many highly conserved residues in the vicinity of the active site, some of which are exposed to the solvent and others that are buried in the protein core [Fig. 2(a)]. The "main patch" includes most of the known functional residues of PK [Table I; Figure 2(b)] and may be decomposed into 2 functional regions [Fig. 1(a–c)]: the "active site," which includes highly conserved residues that are directly involved in substrate binding and catalytic activity; and the "interface" region, which includes moderately conserved residues, most of which mediate the interaction between subunits 1 and 3 of the PK homotetramer. [It is noteworthy that the active-site residues play a dual role, since they are also partially involved in the intersubunit interface; Figure 1(c).] The difference in the conservation level of the "interface" and "active site" regions of the "main patch" presumably reflects differences in the evolutionary pressure. Figure 1(a) shows that the "interface" region is more evolutionarily tolerated, since it is formed mainly by amino acids that are less conserved than the active site (grade 8), thus suggesting that PK complex stability is not very sensitive to the detailed protein structure. In contrast, the highly conserved nature of the "active-site region" (most residues with grade 9) suggests that a very unique combination of amino acids is required to carry out the enzymatic activity.

Figure 2(a and b) shows a view of the "active site" through domain B. All the known functional residues in this region (except for Tyr82 and Lys366) are highly conserved. Interestingly, additional highly conserved residues can be detected in this region. Their high conservation score and proximity to the catalytic cleft suggest that they too are functionally important. Furthermore, Asp177 of domain B and Arg341 of domain A, both of which are in the "main patch," were assigned a high conservation grade of 9 (Table I). These 2 residues form a salt bridge near the opening of the active site. The formation and disruption of this salt bridge is related to a shift in the equilibrium between the active and inactive conformations of PK.[25]

Conformational changes

Each PK subunit can adopt several conformations, which are characterized by differences in the relative position of protein domains A, B, and C.[29] In PDB entry 1a49, 6 subunits (1, 3–7) adopt the closed conformation and 2 subunits (2 and 8) adopt an open or semiopen conformation.[24] The secondary patches of conserved residues marked as 1 and 2 in Figure 1(d) mediate interdomain and intersubunit interfaces in the PK structure, which are important to stabilize these conformations.

Patch 1 is the second largest conservation signal on PK's surface. It includes highly conserved functional residues, such as Pro116, Lys223, and Phe243 [Figs. 1(d) and (2a)], which are involved in the rotation of domain B relative to domain A.[25,29] This patch also includes a short sequence of moderately conserved residues (most of which were assigned a ConSurf-HSSP grade of 7), connecting between domains A and B, known as the "linker region."[29] Here again, the conservation analysis indicates that this patch can be subdivided into 2 regions that evolve at different rates, in agreement with the known biological role of both regions: the extremely conserved interdomain, interfacial region, which allows the closing and opening of the active site, and the flexible loop that connects domains A and B, which is moderately conserved [Fig. 1(d), patch 1].

Patch 2 contains only 4 residues: Arg254, Ser286, Asp287, and Lys321 [Fig. 1(d)]. Arg254 and the highly conserved Asp287 are close enough for their charged side-chains to interact electrostatically.

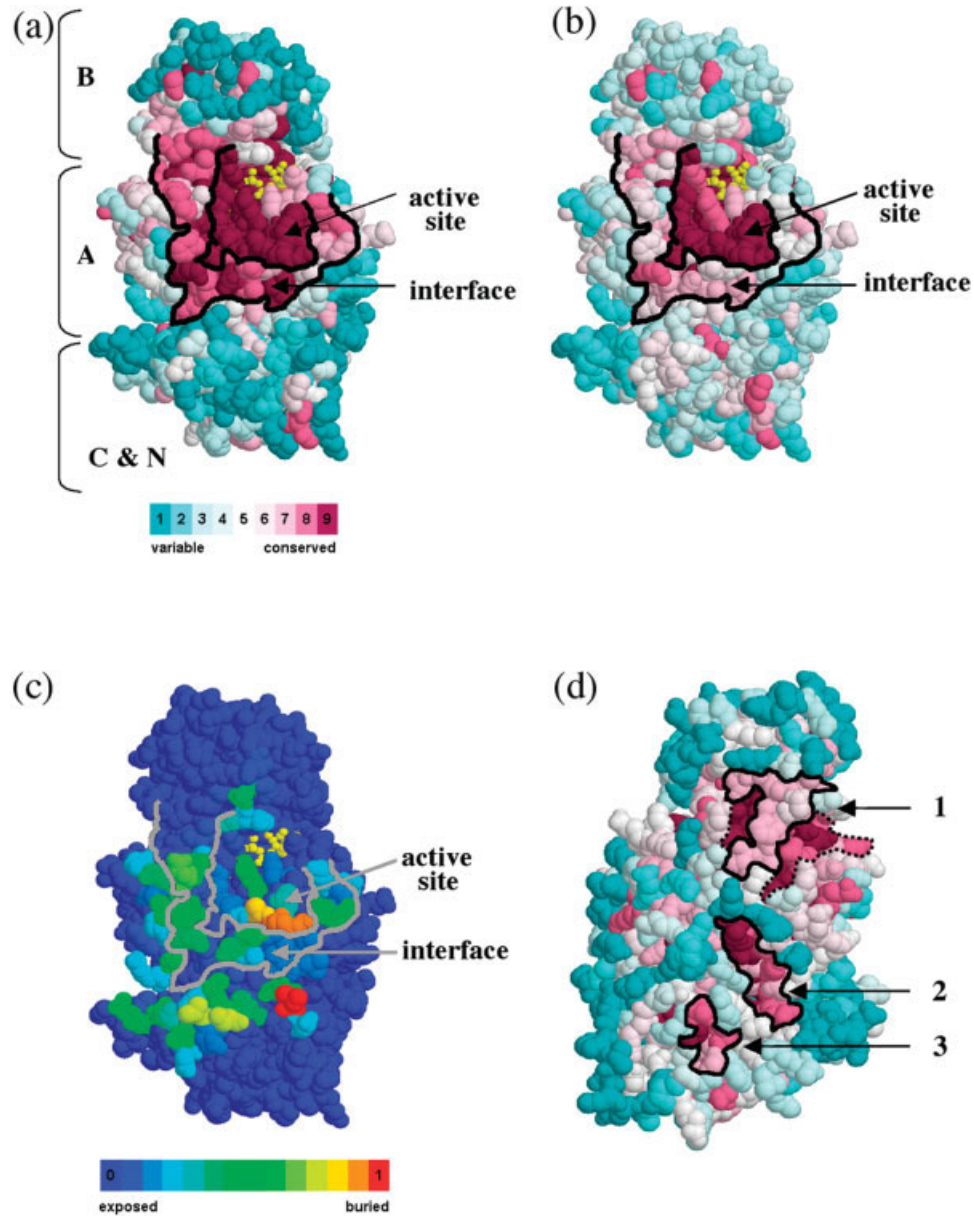
(a)
B
A
C & N
active site
interface
1 2 3 4 5 6 7 8 9
variable
conserved
(b)
active site
interface
(c)
active site
interface
0
1
exposed
buried
(d)
1
2
3

Fig. 1.

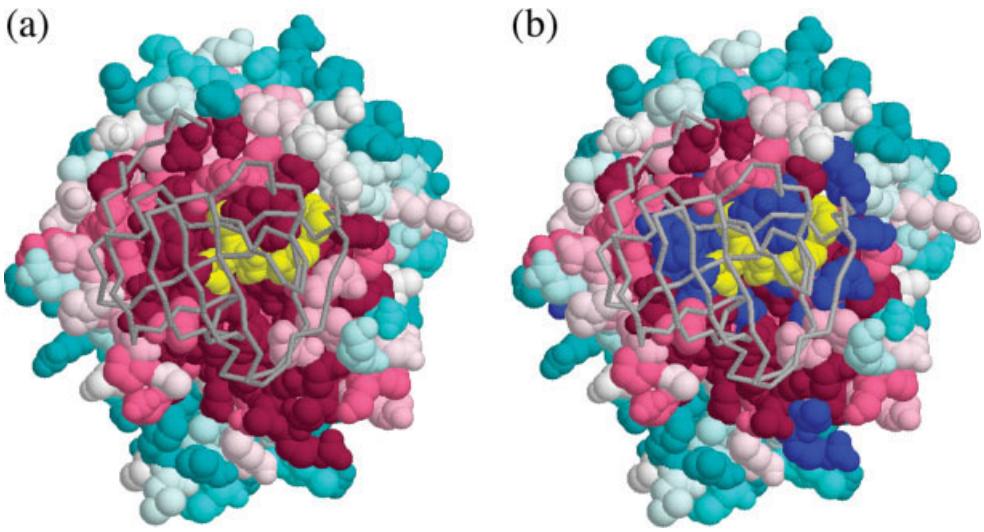
(a)
(b)

Fig. 2.

It has long been recognized that the interdomain interfaces in PK are critical for enzyme regulation.[31] The proposed "domain rotational" model for regulating PK's activity implies that the enzyme has to be equipped with structural elements that enable the coupling of the domain movement to conformational changes in the active site.[26] In agreement with this model, several mutagenesis experiments on the interface between domains A and C (the AC interface) have shown the existence of conserved interdomain salt bridges.[31] It has also been demonstrated that mutations on the AC interface are the most common cause of the hereditary, nonspherocytic hemolytic anemia.[26] The ConSurf-HSSP analysis of the A domain residues at the AB (Fig. 2) and AC (data not shown) interdomain interface regions shows a high degree of conservation in this region. These results reflect the evolutionary importance of the AC and AB interfaces, as predicted by the "domain rotational" model.

Fructose binding

Fructose 1,6-biphosphate (FBP) is one of the main allosteric regulators of mammalian PK, conferring different degrees of allostery, depending on sequence differences and regulatory patterns of the various PK isozymes.[25] Patch 3 is a small cluster of conserved residues that corresponds approximately to the putative binding region of FBP in the M1 rabbit muscle isoenzyme [Fig. 1(d), patch 3].[25,33] This patch is formed by residues Thr431, Ser436 (grades 9 and 8), Ser433, Gly519 (both with grade 7), and Arg454 (grade 6). Thr431, Ser436, and Arg454 are suspected to interact directly with 2 oxygen atoms in the 2 extremes of the FBP molecule, in the putative FBP binding site of the Ser402Pro M1 PK mutant.[25] It is important to note that although the M1 rabbit muscle PK is not allosterically activated, many of the homologs that are included in the HSSP alignment used here are, and this patch probably reflects the overall conservation of the FBP binding region throughout the whole PK family. Moreover, it has been shown that a single amino acid replacement is sufficient for the M1 isoenzyme to acquire allosteric properties.[25,33]

Scoring Conservation Using Rate4Site Versus Relative Entropy

To further characterize the ConSurf-HSSP database, we repeated the evolutionary conservation analysis of PK utilizing the relative entropy scores that are provided in the original HSSP database.[1] Table I provides a comparison of the conservation grades calculated using Rate4Site versus relative entropy for 32 of PK's amino acids that are known to be functionally important. Most of these residues are categorized as highly conserved by both methods; Rate4Site assigns 24 (75%) of them as highly conserved (grades 8 or 9), while relative entropy identified only 18 (56%).

Figure 1(a and b) shows the conservation pattern obtained using Rate4Site and the relative entropy in the vicinity of PK's active site. As previously mentioned, the "main patch" found in the ConSurf-HSSP analysis (using the Rate4Site algorithm) can be subdivided into 2 regions, the "active site" and the "interface." The active-site region obtained high conservation scores using both methods. However, the interface appears to be more highly conserved using Rate4Site's scores [Fig. 1(a)] compared to the relative entropy scores [Fig. 1(b)]. Overall, the ConSurf-HSSP results appear to correlate better with the contact projection obtained for the interface [Fig. 1(c)].

Further analysis of the relative entropy results showed virtually no conservation signals that correspond to patches 1, 2, and 3 [Fig. 1(d)] of the ConSurf-HSSP scores, which were obtained using Rate4Site, and for the inter-domain AB (Fig. 2) and AC (data not shown) interfaces patches.

DISCUSSION

Methods for 3D structure determination in proteins have improved tremendously over the past few years.[34,35] As a result, the PDB includes a significant number of proteins of known structure but as yet unknown or partially characterized function. The existence of these structures stimulated many research groups to develop various computational tools for the identification of the functional regions in proteins.[8,11,36–40]

Fig. 1. An analysis of the evolutionary conservation pattern of PK; subunit 1 of the Bis (Mg^{2+}-ATP-oxalate) PK complex (PDB ID: 1a49, chain A) was used. PK residues are presented in a space-filled model, and the substrate and ligands in the yellow ball-and-stick model. The same PK orientation was used in plates (**a**), (**b**), and (**c**); the orientation in plate (**d**) was obtained by rotating the enzyme by 150° to the right on the *y* axis relative to its orientation in the other plates. In (**a**), (**b**), and (**d**), PK was color-coded by conservation using the 9-level scale presented at the bottom of plate (**a**), with turquoise-through-burgundy indicating variable-through-conserved residues. (**a**) ConSurf-HSSP mapping of evolutionary conservation in and around PK's "main patch." The continuous black curves divide this patch into 2 regions: "active site" and "interface" (see text for details). The approximate location of the 4 domains (A, B, C, and N) is marked. (**b**) Mapping of the relative entropy in and around PK's "main patch." (**c**) Contact projection, estimated as the relative solvent accessible surface area of each residue [Eq. (2)], in and around PK's "main patch." A red-through-blue rainbow scheme indicating maximum-through-minimum relative solvent accessibility appears at the bottom. Again, the continuous gray curves divide the "main patch" into 2 regions: "active site" and "interface." (**d**) ConSurf-HSSP mapping of evolutionary conservation reveals 3 patches of conserved residues in addition to the "main patch" of plate (**a**). The patches are numbered and encircled by black curves. Patch 1 includes the linker region between domains A and B of subunit 1 (encircled by continuous black curve, most residues with grade 7) and several highly conserved residues that form PK's catalytic cleft (encircled by dashed black curves). Patch 2 includes polar and titratable residues (Arg254, Ser286, Asp287, and Lys321) that are evolutionarily conserved. Patch 3 includes part of the putative fructose 1,6-biphosphate binding region (residues Thr431, Ser436, Ser433, and Gly519).[25]

Fig. 2. ConSurf-HSSP mapping of evolutionary conservation at the interdomain interface between domains A and B. PK was rotated by 190° in the top-down direction of the horizontal axis relative to its orientation in Figure 1(a). (**a**) Domain A is presented using a space-filled model with the ConSurf-HSSP conservation grades as in Figure 1(a). Domain B is shown as a gray backbone model and substrate, and ligands are presented using the yellow space-filled model. (**b**) The same as in (**a**), but the functionally important residues Pro52, Asn74, Ser76, His77, Tyr82, Ser242, Phe243, Lys269, Asp295, Thr327, Thr340, Arg341, Ser361, Glu363, and Lys366 (some of which are partially buried), are colored blue.

In cases where a sufficient number of homologous proteins are available, evolutionary-based methods have proved to be very effective to this end.[5,7,9,12,36,41–45] For example, Valdar and Thornton[7] and Elcock and McCammon[5] used a straightforward criterion, based on the relative entropy, and correctly discriminated between biologically important and artificial interprotein interfaces observed in X-ray crystal structures in about 86% of a set of 76 proteins, which was compiled by Ponstingl et al.[46] Large-scale tests using variants of the Evolutionary Trace method which, like Rate4site, make explicit use of the phylogenetic relationships between the homologous proteins, were equally successful. For example, Aloy et al.[36] correctly identified 79% of the active sites in a set of 86 proteins; more recently, Madabushi et al.[47] and Yao et al.[48] applied the Evolutionary Trace method for the identification of various known functional interfaces, and reported a success rate of 90% or more for sets of 46 and 86 proteins, respectively. Innis et al.[49] reported a ~96% success rate in the prediction of functional sites in 470 cases tested, using a conserved functional group (CFG) clustering technique. This methodology relies on the examination of the extent and spatial distribution of functional group conservation to identify regions of a protein with functional significance.

Overall, these studies indicate that, with a few exceptions (see below), the major functional regions are highly conserved across homologous proteins and can therefore be easily detected. The challenge is then to detect functionally important amino acids that are not strictly conserved. This and other studies[11,13] demonstrated that the enhanced accuracy and sensitivity provided by the Rate4Site algorithm, used in the generation of the ConSurf-HSSP database, are particularly suitable for this task. For example, based on the different conservation scores assigned to the residues that compose PK's main patch, it was possible to differentiate between the highly conserved active site and the moderately conserved intersubunit interface of this enzyme [Fig. 1(a)].

The wealth of phylogenetic information and its unique contribution to the understanding of protein function have led us to produce the ConSurf-HSSP database of precalculated conservation scores for all the proteins in the PDB. The decision to use premade MSAs of homologous proteins from the HSSP database was critical to the establishment of ConSurf-HSSP. The experience from our laboratories and elsewhere has been that the conservation scores are very sensitive to the quality of the MSA used in the analysis. The ConSurf-HSSP database provides a combination of HSSP's high-quality MSAs and Rate4Site's accurate and sensitive method of scoring conservation, allowing the user an immediate access to the high-quality, precomputed conservation results. Users who are interested in providing their own MSA, rather than relying on the precomputed conservation scores, can use the ConSurf web server (http://consurf.tau.ac.il).[16]

Some rather important functional interfaces are not evolutionarily conserved; the hypervariable pathogen recognition regions in antibodies and major histocompatibility complex (MHC) molecules are excellent examples of this.[50] Such functional regions would be typically overlooked unless one aims at the most highly variable regions of the protein. PK provides another type of a nontrivial, functionally important region: the interface between the C domain of subunits 1 and 2 of PK (the 1–2 interface). This interface is functionally important in PK, being implicated in allosteric intersubunit communication,[25,28] yet it is only averagely conserved in the ConSurf-HSSP analysis (data not shown). The lack of conservation of the 1–2 interface is congruent with the fact that the allosteric regulation through this interface is lineage-specific. Differences along a short-sequence region forming the interface between subunits 1 and 2 seem to be responsible for the different allosteric properties of PK isozymes (e.g., 22 residues between M1 and M2).[51] Accordingly, the 1–2 interface is not conserved in the full ConSurf-HSSP analysis, although it is significantly conserved among a subset of 18 M1 isozyme homologues obtained from GenBank[52] (data not shown). In general, lineage-specific functional sites may be difficult to identify using the ConSurf-HSSP database; thus, complementary ConSurf analysis, using subsets of the homologous proteins corresponding to different clades in the phylogenetic tree, is recommended.

The ConSurf-HSSP database provides a fast and reliable source of functional information for hypothesis-driven studies using biochemical and mutagenesis analysis. In addition, it may also be used for high-throughput studies of all the proteins of known structure.

ACKNOWLEDGMENTS

We thank Sean O'Donoghue and Goran Neshich for helpful discussions. We are grateful to the Bioinformatics Service Unit of the George S. Wise Faculty of Life Sciences at Tel Aviv University for providing technical assistance and computational facilities.

REFERENCES

1. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
3. Dodge C, Schneider R, Sander C. The HSSP database of protein structure–sequence alignments and family profiles. Nucleic Acids Res 1998;26:313–315.
4. Shannon CE, Weaver W. The mathematical theory of communication. University of Illinois Press; 1963.
5. Elcock AH, McCammon JA. Identification of protein oligomerization states by analysis of interface conservation. Proc Natl Acad Sci USA 2001;98:2990–2994.
6. Valdar WS, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
7. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001;313:399–416.
8. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
9. Simon AL, Stone EA, Sidow A. Inference of functional regions in proteins by quantification of evolutionary constraints. Proc Natl Acad Sci USA 2002;99:2912–2917.
10. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–373.
11. Bell RE, Ben-Tal N. In silico identification of functional protein interfaces. Comp Funct Genom 2003;4:420–423.

12. Valdar WS. Scoring residue conservation. Proteins 2002;48:227–241.
13. 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.
14. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
15. 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.
16. 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.
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
18. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
19. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
20. Graur D, Li WH. Fundamentals of molecular evolution. 2nd edition. Sunderland, MA: Sinauer Associates; 1999.
21. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–376.
22. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996;266:418–427.
23. Martz E. Protein Explorer: easy yet powerful macromolecular visualization. Trends Biochem Sci 2002;27:107–109.
24. Larsen TM, Benning MM, Rayment I, Reed GH. Structure of the bis(Mg2+)-ATP-oxalate complex of the rabbit muscle pyruvate kinase at 2.1 Å resolution: ATP binding over a barrel. Biochemistry 1998;37:6247–6255.
25. Wooll JO, Friesen RH, White MA, Watowich SJ, Fox RO, Lee JC, Czerwinski EW. Structural and functional linkages between subunit interfaces in mammalian pyruvate kinase. J Mol Biol 2001;312:525–540.
26. Mattevi A, Bolognesi M, Valentini G. The allosteric regulation of pyruvate kinase. FEBS Lett 1996;389:15–19.
27. Friesen RH, Castellani RJ, Lee JC, Braun W. Allostery in rabbit pyruvate kinase: development of a strategy to elucidate the mechanism. Biochemistry 1998;37:15266–15276.
28. Friesen RH, Chin AJ, Ledman DW, Lee JC. Interfacial communications in recombinant rabbit kidney pyruvate kinase. Biochemistry 1998;37:2949–2960.
29. Larsen TM, Benning MM, Wesenberg GE, Rayment I, Reed GH. Ligand-induced domain movement in pyruvate kinase: structure of the enzyme from rabbit muscle with Mg2+, K+, and L-phospholactate at 2.7 Å resolution. Arch Biochem Biophys 1997; 345:199–206.
30. Fenton AW, Blair JB. Kinetic and allosteric consequences of mutations in the subunit and domain interfaces and the allosteric site of yeast pyruvate kinase. Arch Biochem Biophys 2002;397:28–39.
31. Valentini G, Chiarelli L, Fortin R, Speranza ML, Galizzi A, Mattevi A. The allosteric regulation of pyruvate kinase. J Biol Chem 2000;275:18145–152.
32. Sridharan S, Nicholls A, Honig B. A new vertex algorithm to calculate solvent accessible surface areas. FASEB J 1992;6:A174.
33. Munoz ME, Ponce E. Pyruvate kinase: current status of regulatory and functional properties. Comp Biochem Physiol B Biochem Mol Biol 2003;135:197–218.
34. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2:801–809.
35. Hurley JH, Anderson DE, Beach B, Canagarajah B, Ho YS, Jones E, Miller G, Misra S, Pearson M, Saidi L, Suer S, Trievel R, Tsujishita Y. Structural genomics and signaling domains. Trends Biochem Sci 2002;27:48–53.
36. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
37. Thornton JM. From genome to function. Science 2001;292:2095–2097.
38. Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 2002;47:219–227.
39. Sternberg MJ, Gabb HA, Jackson RM. Predictive docking of protein–protein and protein–DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
40. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Protein–protein interfaces: architectures and interactions in protein–protein interfaces and in protein cores: their similarities and differences. Crit Rev Biochem Mol Biol 1996;31:127–152.
41. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
42. Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. Nucleic Acids Res 1991;19:6565–6572.
43. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
44. Dean AM, Golding GB. Enzyme evolution explained (sort of). Pac Symp Biocomput 2000;6–17.
45. Blouin C, Boucher Y, Roger AJ. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res 2003;31:790–797.
46. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
47. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 2002;316:139–154.
48. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 2003;326:255–261.
49. Innis CA, Shi J, Blundell TL. Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. Protein Eng 2000;13:839–847.
50. Kern PS, Teng MK, Smolyar A, Liu JH, Liu J, Hussey RE, Spoerl R, Chang HC, Reinherz EL, Wang JH. Structural basis of CD8 coreceptor function revealed by crystallographic analysis of a murine CD8alphaalpha ectodomain fragment in complex with H-2Kb. Immunity 1998;9:519–530.
51. Ikeda Y, Tanaka T, Noguchi T. Conversion of non-allosteric pyruvate kinase isozyme into an allosteric enzyme by a single amino acid substitution. J Biol Chem 1997;272:20495–20501.
52. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23–27.