Protein–Protein Interfaces: Analysis of Amino Acid Conservation in Homodimers

William S. J. Valdar[1]* **and Janet M. Thornton**[1,2]

[1]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, London, United Kingdom*

[2]*Department of Crystallography, Birkbeck College, London, United Kingdom*

ABSTRACT **Evolutionary information derived from the large number of available protein sequences and structures could powerfully guide both analysis and prediction of protein–protein interfaces. To test the relevance of this information, we assess the conservation of residues at protein–protein interfaces compared with other residues on the protein surface. Six homodimer families are analyzed: alkaline phosphatase, enolase, glutathione S-transferase, copper-zinc superoxide dismutase, *Streptomyces* subtilisin inhibitor, and triose phosphate isomerase. For each family, random simulation is used to calculate the probability (*P* value) that the level of conservation observed at the interface occurred by chance. The results show that interface conservation is higher than expected by chance and usually statistically significant at the 5% level or better. The effect on the *P* values of using different definitions of the interface and of excluding active site residues is discussed. Proteins 2001; 42:108–124.** © **2000 Wiley-Liss, Inc.**

Key words: molecular recognition; oligomeric interfaces; residue conservation

INTRODUCTION

Protein-protein interactions are ubiquitous in biology. Transient associations between proteins underpin a broad range of biological processes, which includes hormone-receptor binding, protease inhibition, the action of antibody against antigen, signal transduction, correction of misfolding by chaperones, and enzyme allostery. Associations that are more permanent are essential for proteins whose stability or function is defined by a multimeric state. Such proteins range from those in grand assemblies, e.g., muscle fibres and viral capsids, to those in humbler ones, e.g., oligomeric enzymes and oxygen carriers.

Protein–protein interactions occur at the surface of a protein and are biophysical phenomena, governed by the shape, chemical complementarity, and flexibility of the molecules involved. Towards the common goal of understanding how proteins interact, a number of studies have characterized the properties of interfaces between polypeptide chains.

Most theoretical studies have examined the physical and chemical aspects of subunit interfaces in oligomers.[1–11] Oligomer interfaces tend to be planar, roughly circular in shape, variously segmented along the polypeptide chain, protruding, and by their depletion in charged groups and abundance of hydrophobic groups have an amino acid composition that lies between those of the interior and exterior of the protein.[1,6]

Assemblies involving proteins that must be independently stable before association, referred to here as "transient," have interfaces that differ from those in obligate complexes. These interfaces more closely resemble the protein exterior, containing a higher proportion of polar and charged groups than their oligomeric counterparts.[8,12] By the same token, salt-bridges and hydrogen bonding networks play more of a role in stabilizing these complexes than they do in oligomeric interfaces.[12]

Both types of interfaces are close-packed and exhibit a high degree of geometric and electrostatic complementarity.[6,12–14]

The consistency apparent in observations of oligomeric interfaces has led some groups to suggest the location of a putative interface may be predictable from protomer structure alone.[8,10,15,16] "Protomer" is used here to denote a component chain of a multimeric complex. Jones and Thornton[16] developed a predictive method in which, for each protomer in a dataset of dimers, they defined roughly circular patches on the molecular surface, then assessed and ranked each patch according to its chemical and physical properties. Because the properties that make a good interface depend on the type of complex, patches on protomers from homodimer, heterodimer and antigen–antibody complexes were ranked by different criteria. Their method proved most powerful when applied to simple homodimers and weakest when applied to transient dimers, mirroring the degrees of physico-chemical consistency observed for these types of complexes.

The geometric and electrostatic complementarity observed within interfaces has been the basis of many studies that dock two proteins of known structure (see Sternberg et al.[17] and references therein). These algorithms usually begin by treating the two proteins as rigid bodies that are docked to produce a tight complex. Putative complexes are then assessed and refined according to

Grant sponsor: BBSRC.

*Correspondence to: William S.J. Valdar, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK. E-mail: valdar@biochem.ucl.ac.uk

Received 7 July 2000; Accepted 14 September 2000

electrostatic or chemical criteria to predict the "best" complex.

In addition to theoretical analyses of crystal structures, a wealth of experimental studies has provided insights into protein-protein interactions (see references[18,19] and references therein; see also reference[20]). Structural analysis alone is often insufficient and sometimes misleading in determining which residues contribute most to the free energy of binding.[21] Thermodynamic studies in which the interface is systematically mutated reveal that the distribution of energetically important residues can be uneven across interfaces and concentrated in "hot spots" of binding energy.[22,23] There is also evidence that residues distant from the interface can play a critical role in stabilizing protein-protein interactions.[24] Such residues are believed to be energetically coupled with those directly involved in binding and allow binding energy to propagate through tertiary structure.[25]

If a protein's function is common within a homologous family and essential or advantageous for the survival of the host organism, the maintenance of that function describes the limits to which mutational variation in the sequence may be tolerated. So if a protein–protein interaction plays an important functional role, it is interesting to study how patterns of evolutionary conservation in the protomer sequences relate to the maintenance of this interaction. Analysis of conservation patterns in binding sites benefits from the fact that the residues involved are on the molecular surface and surface conservation is generally low. This potentially high signal-to-noise ratio arises because changes in surface residues do not generally influence folding and overall stability as much as changes in residues at the structural core, so any mutational intolerance that does exist can be detected more easily.

Several groups have explored patterns of conservation at binding sites in a systematic way using multiple alignments, and sometimes phylogenetic trees, of homologous sequences to map evolutionary information onto datasets of protein structures. Lichtarge et al.[26] classified residue positions in a multiple sequence alignment according to the size of the subfamily in which they were invariant. Mapping these classifications onto the structure of a representative protein, they found the relative conservation of binding site residues correlated well with observed patterns of relative binding energy. In a similar vein, Lockless and Ranganathan[25] showed evolutionary information could be used to predict energetic couplings between residues in a binding site and those distant from it. De Rinaldis et al.[27] developed a search tool that uses evolutionary information to compare binding motifs on protein surfaces. Given a query structure and its sequence alignment, their method maps amino acid substitution patterns of surface residues onto a three-dimensional grid. The grid, which corresponds to a coarse grain model of the structure, is filtered to remove unconserved positions and acts as a profile of the surface. For each protein in a database of structures, they then compare the amino acids on its surface with distributions in the profile and assess the similarity of any aligned residues using amino acid exchange probabilities from a mutation data matrix.

As mentioned above, a method to detect putative interfaces cannot rely on residues at the interface showing consistent physical and chemical properties. But if interface residues generally are conserved, then identifying an interface could be as straightforward as locating a conserved cluster of residues on the surface. Such a method would beg the question, however, since it relies on an unproven premise: that residues in interfaces are significantly more conserved than those on the rest of the surface.

Grishin and Phillips[28] tested a similar premise, that interface residues are significantly conserved with respect to all other residues in a protein, and concluded it to be false. They analyzed five oligomeric enzymes. For each enzyme sequence, they identified which positions corresponded to residues in the structural core, the active site, and the subunit interface. Then, for every pair of sequences in a multiple alignment of the oligomer, they compare the rate of evolution, which they define as the fractional sequence identity, at these positions with that over all positions in the protein. This comparison gives them a measure of how much slower mutations occur in active site, core, or interface positions than on average over the whole sequence. They found active site residues were by far the most conserved, evolving 50 times slower than average, whereas core and interface residues were only slightly conserved, evolving 2 and 1.5 times slower, respectively. Thus, although the interfaces were much less conserved than the active sites, they were still more conserved than the surface.

Grishin and Phillips's definition of conservation precludes substitutions of any kind.[28] But such a strict definition misses more subtle patterns of conservation: those in which substitutions conserve physico-chemical characteristics. Complete invariance at active sites positions is common because these motifs frequently rely on precise arrangements of specific amino acids. In contrast, residues at the structural core do tolerate mutations but within only a limited range.[29] A more sensitive measure of the rate of evolution, one that accounts for conservative substitutions, might have brought the scores for active site, core, and, possibly, interface positions closer together.

Herein, we investigate whether oligomer interfaces are significantly conserved with respect to the protein surface by studying in depth a small but strictly defined dataset of six homodimer families, each of which form two-chain complexes. To address this problem meaningfully, we determine the probability that a randomly chosen group of residues from the protein surface will be more or less conserved than the interface group. We estimate this probability for all six oligomer complexes in our dataset by simulation, performing a large number of trials to obtain the fraction of random selections that equal or better interface conservation. The trials are performed in two ways: "picking," in which groups of residues are chosen entirely at random from the surface, and "walking," in which randomly chosen groups may contain only residues that are structurally contiguous.

MATERIALS AND METHODS

The following protocol was followed to test whether the interface residues of a component chain in a protein-protein complex are significantly conserved with respect to all residues on the surface of that chain. First, functionally equivalent homologues of the protomer are identified and aligned multiply. Second, each position in the protomer is given a score that measures the degree to which it is conserved in evolution as inferred from the multiple alignment. Third, each residue in the protomer is classified according to the extent it lies in the surface and the extent it participates in the interface. Last, the average conservation score for residues in the interface is compared with the distribution of average conservation scores for the same number of surface residues in randomly selected groups. This comparison allows us to estimate the probability that a randomly selected group will have an equal or better average conservation than the interface, and hence assess whether the conservation of an interface is statistically significant.

To put this work in the context of previous analyses of protein–protein interfaces, the interface conservation of each protomer is also examined using "surface patches," after Jones and Thornton.[8]

Criteria for Dataset

Component chains from oligomer complexes were chosen to fulfill the following criteria. The protomer to be studied must form a stable, symmetric complex with one other protomer to which it is identical or nearly identical such that the oligomer is homodimeric and the conservation of only one chain need be considered. The complex must be shown by its associated literature to be essential to the stability and correct function of the protein. The full wild-type complex must be available as a structure determined by X-ray crystallography in either the Protein Data Bank (PDB)[30,31] or its derivative, the Protein Quaternary Structure File Server[32] (PQS; http://pqs.ebi.ac.uk/). Of all structures available for the complex, the structure chosen must have the best combination of the following properties: high resolution, inclusion of any bound cofactors that occur naturally; and, if applicable, the inclusion of a ligand similar in size and shape to that of the natural substrate. To enable the robust identification of a diverse set of homologues, the protomer should be represented in the CATH classification.[33] The protomer sequence must have non-fragment homologues in the SWISS-PROT protein sequence database[34] that are numerous (>10) and diverse (<70% mean pairwise sequence identity), and, by their annotation, share its function and multimeric state (see also Identification Alignment of Homologues). Applying these criteria gave rise to six homodimer families (see Table I).

Identification and Alignment of Homologues

Homologues for a given protomer are identified and aligned in two stages. At both stages, a homologue is included only if its annotation and associated references show unambiguously that it shares the protomer's function and precise multimeric state, and that it is a wild type protein and not a fragment. First, the sequence family of the protomer is identified in the CATH classification.[33] If there are at least three suitable representatives of that family, a multiple structural alignment of the protomer and these representatives is built using the CORA suite of programs.[35] Otherwise, a multiple sequence alignment from the ALIGN resource associated with the PRINTS database[36] or one from the Pfam database[37] is edited and used. This seed alignment is used to build a profile hidden Markov model (Eddy[38] and references therein), which, in turn, is used both to find more homologues in the SWISS-PROT sequence database and to align multiply the final set of homologues to the protomer sequence. This final alignment, referred to here as the "full alignment" for the family, is edited for high redundancy by removing the less well-characterized or shorter of any two sequences that are more than 90% identical. Searching for homologous structures in CATH was performed with reference to the CATH Dictionary of Homologous Superfamilies.[39] The construction of the profile and its application in searching for and aligning homologues were both performed with the HMMER2 software package (see http://hmmer.wustl.edu/). The filtering of large numbers of sequences by annotation was performed using the Sequence Retrieval System.[40] The removal of redundancy in the alignment was performed with the help of JalView.[41]

Scoring Residue Conservation

Each residue in the protomer of interest is assigned a numerical value *Cons* (ranging from 0 to 1) corresponding to the conservation of residue similarities at its position in the multiple sequence alignment. A value of 0 indicates the position is not conserved; a value of 1 indicates it is highly conserved. *Cons* uses amino acid exchange probabilities from a Dayhoff-like[42] mutation data matrix to assess the diversity of amino acids at an aligned position in a similar way to the sequence variability score of Sander and Schneider.[43] A weighted sum of all pairwise similarities between all residues present at the position is calculated and the function *Cons(i)* for position *i* in a given alignment is defined as

$$Cons(i) = \frac{\sum_{j}^{N} \sum_{k>j}^{N} W_j W_k Mut(s_j(i), s_k(i))}{\sum_{j}^{N} \sum_{k>j}^{N} W_j W_k}$$

where N is the number of sequences in the alignment; $s_j(i)$ and $s_k(i)$ are the amino acids at alignment position i of sequences s_j and s_k respectively; $Mut(a,b)$ measures the similarity between amino acids a and b as derived from a mutation data matrix m. The matrix used here is the Pairwise Exchange Table (PET91) of Jones and Thornton[44] in which $Mut(a,b)$ is related to $m(a,b)$ by the transformation

$$Mut(a, b) = \begin{cases} \dfrac{m(a, b) - \min(m)}{\max(m) - \min(m)}, & \text{if } a \neq \text{gap and } b \neq \text{gap} \\ 0, & \text{otherwise} \end{cases}$$

such that *Mut* takes values in the range [0,1] and all exchanges involving a gap score 0. W_j is the weight of sequence s_j, and is defined as the average evolutionary distance between s_j and all other sequences in the alignment:

$$W_j = \frac{\sum_{k \neq j}^{N} Dist(s_j, s_k)}{N - 1}$$

W_j takes values in the range [0,1]. W_j is small when sequence s_j is like all the other sequences and large when s_j is distant from them. *Dist*(s_j,s_k), the evolutionary distance between sequences s_j and s_k, is defined as

$$Dist(s_j, s_k) = 1 - \frac{\sum_{i \in Aligned_{jk}} Mut(s_j, s_k)}{n(Aligned_{jk})}$$

where $Aligned_{jk}$ is the set of all non-gap positions in s_j or s_k, and $n(Aligned_{jk})$ is the number of such positions.

Here the terms "conservation score" and "residue conservation" will be used to denote either the value returned by *Cons* for a given residue, or, when applied to a set of residues, the average value of *Cons* for all residues in that set. These definitions are vital because they underpin the whole analysis of conservation.

Interface Definition

Each residue in a protomer is assigned to one of the following disjoint sets: *Core, Exposed, Partially Buried*, or *Buried*. Qualitatively, *Core* residues are those in the structural core of the protein, *Exposed* residues are on the surface but do not participate in an interface, and *Partially Buried* and *Buried* describe residues that are on the surface of the protomer and participate in a multimer interface. Two further sets are referred to here: the *Surface* set, which is the union of the *Exposed, Partially Buried* and *Buried* sets, and contains all residues on the surface of the protomer; and the *Interface* set, which is used as a generic term for either the *Buried* set or the union of the *Buried* and *Partially Buried* sets. For clarity, *Buried* residues are said to form the "central zone" of the interface whereas *Partially Buried* residues form its "outer zone." Together, these sets define the "total interface."

The classes are assigned on the basis of solvent accessibility, which is calculated using NACCESS ,[45] an implementation of the Lee and Richards[46] algorithm, with a probe sphere of radius 1.4Å. A residue is deemed accessible if its relative accessible surface area (RSA) is $> 5\%$, a cut-off devised and optimized by Miller et al.[47] If a residue is accessible in the protomer it is in the *Surface* set, otherwise it is *Core*. If a residue in the *Surface* set loses RSA upon complexation it is in the *Interface* set, otherwise it is *Exposed*. If a residue in the *Interface* set is inaccessible (i.e., $< 5\%$ RSA) in the multimer complex, it is in the *Buried* set, else it is *Partially Buried*.

Ligand-Buried Residues

Ligand-buried residues are defined here as residues that become inaccessible in the protomer upon inclusion of ligand groups. Together they comprise the ligand-buried site for a protomer. Because residues that participate in an active or allosteric site (referred to here generically as a "binding site") are typically both accessible and highly conserved, the inclusion of ligand-buried residues, which are usually a subset of the binding site residues, in the *Surface* set will clearly affect any calculation that compares conservation of interface and surface. To investigate the effect of conserved ligand-buried residues, all tests described below are carried out twice, once with these residues included, or "unmasked," and once with them excluded, or "masked."

Patch Analysis of Interface Conservation

The conservation of the total interface of each unmasked protomer was examined using a variant of the "patch analysis" method of Jones and Thornton.[8] In the original procedure, a set of roughly circular overlapping patches, each covering as many residues as the interface, is defined on the surface of the protomer. Quantitative properties of patches and their constituent residues can then be described in terms of their distributions over all patches and related to the extent those patches overlap with the interface. Herein, the average conservation of residues in a patch is the only property considered and this quantity is termed the "patch score."

Testing the Significance of Interface Conservation

In order to assess the significance of conservation at a given interface the following null hypothesis, H_0, is tested: the average conservation of the *Interface* set is no higher than that obtained from an equal number of residues drawn without replacement from the *Surface* set by a random process. The negation of H_0 is the alternative hypothesis, H_1, which states that the *Interface* set has a higher average conservation than that of a set randomly selected in this way. A simulation experiment is performed to estimate the probability that H_0 is true. If the value of this probability (*P* value) falls below a certain threshold, customarily defined at 0.05, then H_0 is rejected in favour of H_1 and the conservation of the interface is considered statistically significant at the 5% level.

The *P* value expresses the probability that a selection of residues drawn from the surface by a random process will have an average conservation equal to or greater than that of the interface. This *P* value depends not only on the distribution of conservation over the surface and in the interface but also on the nature of the random process employed to make the selections. To ensure the test of H_0 is meaningful, it is a minimum requirement of any random process used that it is able to draw from the surface the set

TABLE I. Family Information for

Family					Representative promoter structure		
Name	Abbreviation	Description	References	Fold (class/architecture/topology)	Name	Resolution (in Ångstroms)	Heteroatom/ligand groups
Alkaline phosphatase	AP	Widely distributed non-specific phosphomonoesterase.	DuBose and Hartl, 1990;[56] Hullett et al., 1991;[57] Kim and Wyckoff, 1999;[58] Knowles, 1991[59]	Mainly beta/sandwich/immunoglobulin-like.	1alk chain A	2	1xMg 2xZn 1xPO4
Enolase	Enolase	Glycolytic enzyme catalyzes dehydration 2-phospho-D-glycerate (PGP) to phosphoenolpyruvate (PEP).	Babbitt and Gerlt, 1997;[60] Babbitt et al., 1996;[61] Larsen et al., 1996;[62] Zhang et al., 1997[63]	domain 1: alpha beta/2-layer sandwich/enolase-like; domain 2: alpha beta/barrel/TIM barrel.	1one chain A	1.8	1xPEP 1xMg
Glutathione S-transferase	GST	Catalyzes conjugation of glutathione to a variety of electrophilic substrates (including carcinogens and anti-cancer drugs) and makes the latter easier for the host to metabolise.	Board et al., 1995;[64] Board et al., 1997;[65] Neuefeind et al., 1997;[66] Rossjohn et al., 1997.[67]	domain 1: alpha beta/3-layer(aba) sandwich/glutaredoxin; domain 2: mainly alpha/non-bundle/glutathione S-transferase (subunit A, domain 2).	1 glq chain A	1.8	1xGTB (s-(p-nitrobenzyl) glutathione)
Copper, zinc superoxide dismutase	SOD	Neutralizes superoxide radicals. Consistent homodimers only cytoplasmic eukaryotic proteins.	Banci et al., 1998;[68] Bordo et al., 1994;[69] 1999;[70] Getzoff et al., 1989.[71]	mainly beta/sandwich/immunoglobulin-like.	1xso chain A	1.49	1xCu 1xZn
Streptomyces subtilisin inhibitor	SSI	Serine proteinase inhibitor that inhibits subtilisin strongly and other proteinases, including trypsin and chymotrypsin, to a lesser extent. Protomers inhibit one proteinase each to form E_2I_2 complex.	Hirono et al., 1984;[72] Kojima et al., 1993;[51] Laskowski and Kato, 1980;[48] Taguchi et al., 1997.[52]	alpha beta/layer sandwich/subtilisin inhibitor.	2sic chain I	1.8	1xStreptomyces subtilisin (2sic chain E)
Triose phosphate isomerase	TIM	Catalyzes interconversion of D-glyceraldehyde 3-phosphate and dihydroxy acetone phosphate.	Borchert et al., 1994;[73] Garza-Ramos et al., 1998;[74] Gopal et al., 1999;[75] Williams et al., 1999.[53]	alpha beta/barrel/TIM barrel	1tph chain 1	1.8	2-phosphoglycolohydroxamate

of residues corresponding to the interface. Two distinct random selection processes, "picking" and "walking," are employed here and these are described below.

For a given protomer surface, defined interface and random process, it is often computationally infeasible to enumerate all possible selections, compare the conservation of each to that of the interface, and hence evaluate P, the true P value. However, P can be estimated reliably enough by random sampling.

A trial is devised in which the random process selects *n(Interface)* residues from surface set and their average conservation is compared with that of the residues in the interface set. If t such trials are performed and each trial is independent of any other, then P can be estimated as

the Six Homodimer Families

Representative promoter structure			Seed alignment			Full alignment		
Data bank source	Organism	Primary citation	Source	Alignment method	Number of sequences	Number of sequences	Mean % seq id	Standard Deviation % seq id
PDB	Escherichia coil	Kim & Wyckoff, 1991.	Pfam seed alignment: alk_-phosphatase. Seed is edited.	Pfam: automatic.	6	11	45.6%	23.2%
PDB	Saccharomyces cerevisiae (baker's yeast)	Larsen et al., 1996.	PRINTS ALIGN resource alignment: enolase.	PRINTS: manual.	6	31	67.7%	12.1%
PDB	Mus musculus (house mouse) liver	Garcia-Saez et al., 1994.	SRS annotation search cross referenced with Pfam: gluts.	CORA structural alignment.	6	51	27.4%	18.7%
PDB	Xenopus laevis (African clawed frog)	Djinovic Carugo et al., 1996.	Selected structures from DHS, CATH level: 2.60.40.200.	CORA structural alignment.	4	34	59.4%	8.9%
PQS	bacillus amyloliquefaciens	Takeuchi et al., 1991.	Pfam seed alignment: SSI.	Pfam: automatic.	11	13	51.8%	10.8%
PDB	gallus gallus (chicken) muscle	Zhang et al., 1994.	Selected structures from DHS, CATH level 3.20.20.80.	CORA structural alignment.	9	55	43.2%	11.8%

$$\hat{p} = \frac{t_c}{t},$$

where t_c is the number of trials in which the selection was at least as conserved as the interface. The greater the number of trials, the more reliable the estimate, and when t is large the expected accuracy of $\hat{p}$ can be described formally by a confidence interval. A confidence interval is defined by a range, symmetric about the estimate, and an associated probability that P, the true value, is contained somewhere within this range. The "99% confidence interval" for the unknown value of P is thus the margin of error expected for $\hat{p}$ 99% of the time. This interval is given by $(\hat{p} - 2.58\sigma, \hat{p} + 2.58\sigma)$, where σ, the standard deviation of $\hat{p}$, is equal to $\sqrt{\hat{p}(1 - \hat{p})/t}$.

To ensure a high degree of accuracy, the number of trials performed for a given a estimate is constant at 10 million, resulting in a margin of error of at most ≃0.04% at least 99% of the time.

The *P* value for interface conservation is estimated stochastically as described above for each of the six protomers in the dataset. For each protomer, trials are performed under three variable conditions, giving rise to eight experiments per protomer. First, trials are performed using one of the two random selection processes, "picking" and "walking." Second, residues participating in an active or allosteric site are either included in the *Surface* set or masked out. Third, the *Interface* set is taken as either the *Buried* set (central zone) or the union *Partially Buried* ∪ *Buried* (total interface).

Picking: Unconstrained Selection of Residues

"Picking" is the first of the two random processes used here for selecting a group of residues from the *Surface* set of a protomer. For a given protomer, residues are drawn at random and without replacement from the *Surface* set until the number drawn is equal to *n(Interface)*, the number of residues in the *Interface* set. In picking, all selections occur with equal probability.

Walking: Structurally Constrained Selection of Residues

"Walking" is the second of the two random processes used here for selecting a group of residues from the *Surface* set of a protomer. Walking selects groups of residues from the surface of a protomer by successively stepping from one residue to any residue in contact with it chosen at random. A walk starts at any residue chosen from the entire *Surface* set. The walk is allowed to revisit residues any number of times, otherwise it could become trapped, but any particular residue is counted only once towards the final selection. The walk ends when the number of distinct residues visited is equal to the number of residues in the *Interface* set. In this scheme, two residues, A and B, are considered "in contact" if the distance between the van der Waals spheres of at least one of A's atoms and at least one of B's atoms is no more than 1Å. All walks are equiprobable but many walks may produce the same selection.

RESULTS

We investigated interface conservation in six homodimer families (abbreviations in parentheses): alkaline phosphatase (AP), enolase (Enolase), glutathione S-transferase (GST), copper-zinc superoxide dismutase (SOD), *Streptomyces* subtilisin inhibitor (SSI), and triose phosphate isomerase (TIM) (see Table I).

Table II lists the number of residues in the central zone and total interface of each protomer representative, and Figure 1 shows graphically the residues that make up the total interface. The total interface for a family was typically contiguous and compact, though not particularly circular. Only the total interface of SSI was non-contiguous, in which one residue, Pro37, was separated from the closest of the others by 3Å (Fig. 1a,c, SSI). An approximately linear relationship was observed both between the size of the central zone and total interface and between the size of the total interface and the number of surface residues. The largest total interface was that of AP and covered a third of that protomer's surface residues (Fig. 1a, AP; Table II). The smallest total interface, which covered barely a fifth of surface residues, belonged to GST (Fig. 1a, GST; Table II). The central zone was between 20% (GST) and ≃40% (AP) of the size of the total interface.

Ligand-buried residues were identified in AP, Enolase, SSI, and TIM (Table III, Fig. 1). The ligand-buried sites of AP and Enolase both comprise two highly conserved residues situated in pockets near the interface (Fig. 1a, AP, Enolase).

SSI's ligand-buried site is on the opposite side of the protomer to that of the interface (Table III, Fig. 1b, SSI) and consists of three residues, Met70, Cys71, and Pro72, which protrude into and block the active site of serine proteinase. These residues are centrally located within an inhibitory region SSI shares with other serine proteinase inhibitors (serpins). This region, the so-called "reactive site," stretches from Gly66 to Tyr75 and, between Met73 and Val74, contains the peptide bond used as bait for the catalytic triads of proteinase enzymes. In contrast with the other protomers, SSI's ligand-buried residues are not unanimously conserved.[48–51] Phylogenetic analysis by Taguchi et al.[52] suggests variability in this region may result from diversifying selection driven by the advantages of multi-specific inhibitors in the regulation of intrinsic proteases.

TIM's ligand-buried site was the largest of those studied and contains five highly conserved residues (Table III, Fig. 1a,c: TIM). TIM binds its substrate in a pocket created by the inside edge of its barrel topology. Although this pocket lies just outside the total interface, two ligand-buried residues, Asn11 and His95, belong to the central zone. This surprising observation results from the convoluted geometry of the interface in which the two component chains protrude so deeply into one another that one affects the solvent accessibility of residues that form the inside of the other's barrel.

Although GST and SOD are both enzymes and their representative structures included ligand groups (see Table I, Fig. 1a, GST, SOD; 1b, SOD), no ligand-buried residues were detected in either. This reflects the strict definition of the ligand-buried site (see Materials and Methods) in which ligand-buried residues must lose all accessibility upon binding ligands. For example, Lys13 of TIM, which is known to play an important role in catalysis[53] and is highly conserved, touches TIM's substrate. However, because it is not completely buried by the substrate it does not qualify here as a ligand-buried residue.

Defining which amino acid types are conserved in interfaces is complex and beyond the scope of this paper. Residues in the representative protomer of AP (Fig. 1c,d, AP) map directly to positions in AP's multiple alignment and so may host a number of amino acid types in varying proportions. Moreover, the notion of conservation as a continuous quantity suggests no obvious cutoff at which

TABLE II. *P* Values and Associated Information Calculated for the Six Homodimer Families

Family	Active site state	Number of surface residues	Mean surface *Cons*	S.D. surface *Cons*	Interface definition[a]	Number of interface residues	Mean interface *Cons*	Experiment type	*P* value for interface	Error (+/−)
AP	Unmasked	289	0.59	0.21	Central	37	0.71	Picking	6.86E-5	6.76E-6
								Walking	6.83E-3	6.72E-5
					Total	96	0.63	Picking	4.52E-3	5.47E-5
								Walking	7.02E-2*	2.08E-4
	Masked	287	0.58	0.21	Central	37	0.71	Picking	3.87E-5	5.08E-6
								Walking	3.12E-3	4.55E-5
					Total	96	0.63	Picking	2.36E-3	3.96E-5
								Walking	5.00E-2	1.78E-4
Enolase	Unmasked	263	0.72	0.22	Central	20	0.89	Picking	3.55E-5	4.86E-6
								Walking	3.74E-2	1.55E-4
					Total	52	0.82	Picking	3.02E-5	4.48E-6
								Walking	3.78E-2	1.56E-4
	Masked	261	0.72	0.22	Central	20	0.89	Picking	2.57E-5	4.14E-6
								Walking	3.18E-2	1.43E-4
					Total	52	0.82	Picking	2.18E-5	3.81E-6
								Walking	3.13E-2	1.42E-4
GST	Unmasked	158	0.45	0.16	Central	6	0.71	Picking	1.06E-4	8.40E-6
								Walking	1.77E-3	3.43E-5
					Total	30	0.52	Picking	3.72E-3	4.96E-5
								Walking	7.09E-2*	2.09E-4
	Masked	158	0.45	0.16	Central	6	0.71	Picking	1.10E-4	8.54E-6
								Walking	1.74E-3	3.40E-5
					Total	30	0.52	Picking	3.72E-3	4.96E-5
								Walking	7.11E-2*	2.10E-4
SOD	Unmasked	105	0.70	0.23	Central	5	0.93	Picking	8.73E-3	7.59E-5
								Walking	2.75E-2	1.33E-4
					Total	20	0.78	Picking	3.60E-2	1.52E-4
								Walking	2.02E-1	3.27E-4
	Masked	105	0.70	0.23	Central	5	0.93	Picking	8.73E-3	7.59E-5
								Walking	2.75E-2	1.34E-4
					Total	20	0.78	Picking	3.60E-2	1.52E-4
								Walking	2.02E-1*	3.27E-4
SSI	Unmasked	91	0.71	0.23	Central	7	0.85	Picking	4.50E-2	1.69E-4
								Walking	1.59E-1*	2.98E-4
					Total	27	0.84	Picking	1.50E-4	9.99E-6
								Walking	1.30E-2	9.25E-5
	Masked	88	0.71	0.22	Central	7	0.85	Picking	4.53E-2	1.70E-4
								Walking	1.64E-1*	3.02E-4
					Total	27	0.84	Picking	1.43E-4	9.75E-6
								Walking	1.35E-2	9.41E-5
TIM	Unmasked	168	0.59	0.22	Central	9	0.76	Picking	1.30E-2	9.25E-5
								Walking	1.65E-1*	3.03E-4
					Total	38	0.68	Picking	1.92E-3	3.57E-5
								Walking	1.71E-1*	3.07E-4
	Masked	163	0.58	0.21	Central	8	0.73	Picking	2.61E-2	1.30E-4
								Walking	1.63E-1*	3.02E-4
					Total	36	0.66	Picking	2.90E-3	4.39E-5
								Walking	1.43E-1*	2.86E-4

[a] Central = “central zone”, total = “total interface.”
*$P \geq 0.05$.

“conserved” residues could be distinguished. However, for completely conserved positions, i.e., those with a conservation score of 1, such an analysis is simple. By far the most common amino acid invariant at the interfaces of the six homodimers was glycine. This is probably because glycine does not have a side chain and so substituting it with an amino acid that does causes sterically unacceptable disruption of the interface. Arginine and valine were next most common but their numbers are low and so cannot be interpreted with confidence.

For each family, the statistical significance of residue conservation at the oligomeric interface was assessed by computing the probability (*P* value) that this conservation could have occurred by chance (see Materials and Meth-

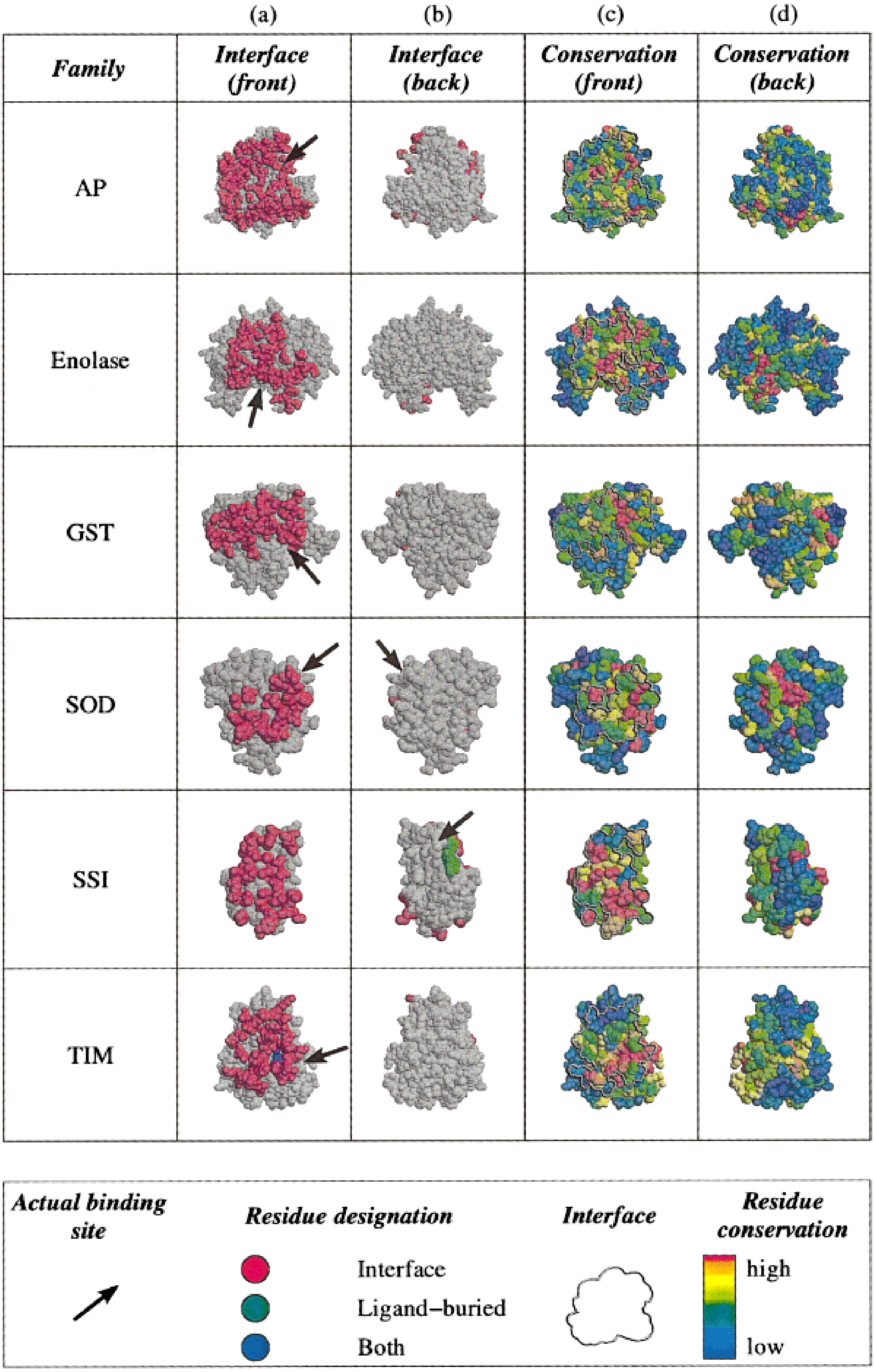

Fig. 1. A table to show the location of interface and ligand-buried residues (**a,b**), and residue conservation (**c,d**) for six families of homodimers. Protomer structures are elevated to show the interface head on in columns (a) and (c), and at a rotation of 180° about the y-axis in columns (b) and (d). In columns (a) and (b), ligand-buried residues and residues belonging to the total interface are indicated (see Materials and Methods for how these classes are defined). Arrows indicate the approximate location of the actual binding site as it is defined in the literature. Ligand-buried residues, typically a subset of residues in the actual binding site, are detected in AP, Enolase, SSI, and TIM. In the elevations presented here these residues are out of view for AP and Enolase, partially visible for TIM, and conspicuous in SSI. In columns c and d, each residue is coloured by the rank of its conservation score among all other conservation scores in the protomer. Rank *Cons* is used instead of absolute *Cons* so that dispersion of conservation over the surface can be more easily visualized. A steel wire effect delineates the perimeter of the total interface. The table shows conservation is not distributed uniformly on the surface but in clusters, and that the interface, although it includes both highly and poorly conserved residues, is on average more conserved than not. Atom coordinates were obtained from the PDB[30,31] and the PQS.[32]. Images were created using MOLSCRIPT[76] and Raster3D.[77]

TABLE III. Ligand-Buried Residues Detected in the Six Homodimer Families

Family	Ligand-buried residues
AP	Thr102, Asp327
Enolase	Ser39, Lys345
GST	None
SOD	None
SSI	Met70, Cys71, Pro72
TIM	Asn11, His95, Glu165, Gly210, Gly232

ods). If the *P* value was less than the predefined cutoff 0.05, the associated interface was considered significantly conserved. *P* value calculations were performed under three variable conditions (described in Materials and Methods), giving rise to eight distinct *P* values per family. In addition to the significance tests, patch analysis was performed on each family whereby the conservation of residues at a homodimeric interface is compared with that of roughly circular overlapping patches defined on the surface of a constituent protomer.

Conservation in Patches

The protomer dataset was analyzed using surface patches. For each family, patches containing as many residues as the total interface were defined on the surface of the representative protomer and the average residue conservation of each patch, i.e., its patch score, was calculated (see Materials and Methods). The number of patches defined, which related linearly to the size of the protomer, ranged from 87 in SSI to 235 in AP.

Figure 2 shows distributions of the patch scores for each family and reveals that, for all families, the average conservation of residues in the observed interface lies within the top quarter of the distribution. Specifically, the score of the interface coincides with the following percentiles: 77% (i.e., lying just within the top 23% of the distribution) (TIM), 80% (SOD), 82% (GST), 84% (Enolase), 91% (AP), and 92% (SSI). It is more meaningful to compare the interfaces of different families based on relative patch rank in this way than by absolute conservation score because the latter depends on the extent and diversity of the underlying multiple sequence alignments.

The mean of a patch score distribution tends towards the mean conservation of surface residues in the corresponding protomer. The higher moments (e.g., standard deviation, skewness, and kurtosis) depend not only on the shape of the distribution of conservation scores for individual residues but also on the patch size and how conservation is dispersed about the surface. As expected, larger patches tend to give narrower distributions. For example, the variance of residue conservation for AP is similar to that of the other families (see Table II) but, owing partly to the large number of residues in its total interface, its distribution of patch scores is markedly narrower (Fig. 2a). The less uniformly the extremes of residue conservation are dispersed over the surface of a protomer, the greater the difference between the highest and lowest patch scores. Dispersion, therefore, affects not only the width of the distribution, causing it to be spread out if residues with high and low conservation cluster in space, but also the skewness and kurtosis. For example, if residues with high and low conservation cluster heavily at opposite sides of a protomer, most patches will contain many more residues from one extreme than from the other, with few patches straddling both poles equally to achieve the mean score. The resulting distribution will have a sunken appearance (negative kurtosis) such as that seen for SSI (Fig. 2e). If the degree of concentration is greater at one pole than the other, the distribution will be correspondingly asymmetric (skewed) as seen for TIM (Fig. 2f).

Overlap of the Interface and Ligand-Buried Site

Patches are deemed to overlap with the interface if they contain at least half of the interface residues, and overlap with the ligand-buried site if they contain all the ligand-buried residues. Overlap is defined differently to take account of the difference in size between these two regions and how much of each a patch can reasonably cover, i.e., most patches that overlap some ligand-buried residues will overlap all of them whereas only one patch can cover all interface residues.

In fact, no patch overlapped any interface completely. The greatest percentage overlap achieved by a patch for a particular family ranged from 67% (AP and SSI) to 80% (SOD). Patches that overlapped the interface tended to be at least moderately and often highly conserved relative to other patches.

The stacked histogram for AP shows that patches that overlap with either the interface or the ligand-buried site occur throughout the distribution, but patches that overlap with both regions occur only among the higher ranks (Fig. 2a). Patches that overlap the interface score variably despite the apparent high percentile ranking of AP's true interface because any one patch owes at least a third of its score to residues outside the interface. Patches that overlap both interface and ligand-buried site score highly because they include not only conserved interface and ligand-buried residues but also some of the conserved binding site residues that surround the ligand-buried site. The narrowness and symmetry of the AP distribution are, as mentioned above, partly explained by the large size of each patch but also reflect the unclustered dispersion of high and low conservation over the surface observed in Figure 1c and d (AP).

The histogram for Enolase shows that overlaps with either the interface or the ligand-buried site occur almost exclusively at the top end of the distribution, with patches that overlap both taking the highest ranks (Fig.2b). Examining conservation at the surface of Enolase reveals a concentration of highly conserved residues around the ligand-buried site and in the region of the interface nearest to it (Fig. 1c,d, Enolase). As for AP, optimally scoring patches tend to be those that cover both regions. The width and positive skewness of the Enolase distribution reflects the clustering of high conservation at the surface in the absence of any poorly conserved clusters.

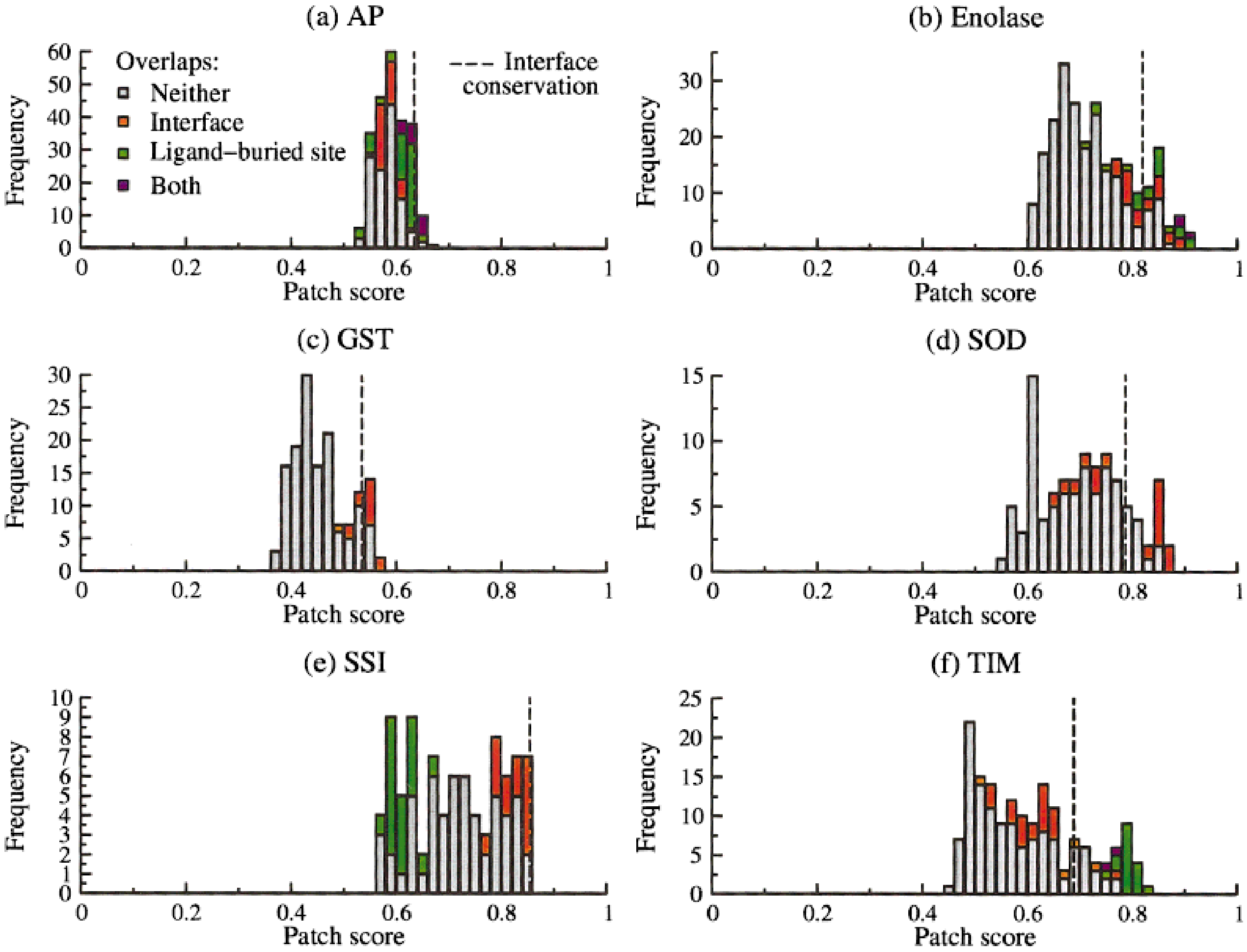

Fig. 2. Distributions of patch scores for six families of homodimers. The patch score, defined as the mean conservation score for all residues in a patch, is given in bins of width 0.02 along the x-axis. The number of patches that fall into a given patch score bin is presented on the y-axis. Stacking within histogram bars indicates the proportion of patches falling into a given bin that overlapped the interface, overlapped the ligand buried site, overlapped neither region and overlapped both (see Results for overlap criteria). A dashed line indicates the interface conservation, defined as the mean conservation score for all residues in the interface. The graphs show that patches overlapping the interface tend to score highly and that interface conservation consistently lies within the top quarter of the patch score distribution.

Patches that overlap the interface are found at only the top end of GST's patch score distribution and all the highest scoring patches have interface overlap (Fig. 2c). GST has smaller patches than Enolase, and, because the sequences that contribute to its alignment are more divergent, there is greater variance in the conservation of its surface residues (Table II). Yet GST has the narrower distribution. This is because the dispersion of conserved residues over the surface of GST is far less clustered than for Enolase, so its patch scores tend to deviate less from the distribution mean (see Fig. 1c,d, GST).

In SOD, overlap with the interface is split between the middle and top end of the distribution of patch scores (Fig. 2d). This dichotomy results from a slight clustering of poorly conserved residues on one side of the interface along with a slight clustering of highly conserved residues on the other (Fig. 1c,d, SOD). Patches that overlap the side of the interface near the unconserved cluster have moderate conservation whereas those that overlap the other side have high conservation.

The histogram for SSI shows a striking separation of interface overlap, which is confined to the upper end of the distribution, from ligand-buried site overlap, which is found exclusively at the lower end. This separation arises because conservation over the surface of SSI is polarized, with a majority of highly conserved residues around the interface and a majority of poorly conserved residues around the hypervariable ligand-buried site.

In TIM, there is a smooth progression from those patches that overlap the interface, which are moderately conserved, to patches that overlap the interface and the ligand-buried site, which are well conserved, to patches that overlap the ligand-buried site, which are confined to the highest ranks. The surface of TIM is marked by a gash of high conservation, which covers half the interface and spreads over and around the ligand-buried site (Fig. 1c, TIM). The remaining half of the interface is only moderately conserved and touches a nearby cluster of poor conservation. The progression described above is consistent with these observations and indicates that whereas patches that overlap the interface may score moderately, thanks to the intersection of the interface and the conserved gash, patches that cover the ligand-buried site and avoid the poorly conserved clusters score higher.

Significance of Interface Conservation

For a given protomer, the significance of interface conservation was assessed as follows. A set of residues equal in number to that of the interface was drawn at random from the surface 10 million times. The fraction of times this random set was at least as well conserved as the interface set was taken as the *P* value for the interface. If and only if the *P* value was less than the predefined cutoff 0.05, i.e., such that the probability of interface conservation being random was <5%, the interface was considered significantly conserved. Two distinct random processes, picking and walking, were used to draw residues from the surface. *P* values were generated using both processes for both definitions of the interface in each family. Moreover, for each combination, *P* values were estimated in both the absence and presence of ligand-buried residues (see Materials and Methods for details). *P* values, being a relative measure, transcend absolute residue conservation. The use of *P* values, therefore, allows the meaningful comparison of conservation across families whose alignments may differ in the extent of their sequence diversity. Results of the *P* value estimations are presented in Table II and Figures 3 and 4 distributions of conservation for random selection are shown in Figure 5.

The picking simulations showed that all the interfaces studied, regardless of which interface definition was used or whether or not ligand-buried residues were excluded, were significantly conserved. Enolase consistently gave the lowest, i.e., most significant, *P* values whereas the family with the highest *P* values depended on interface definition, SOD having the highest among total interfaces and SSI scraping just below the 5% cutoff among central zones (Table II, Fig. 3).

P values determined by walking were consistently higher than those determined by picking (Fig. 4a), and in some cases were outside the top 5% of conserved walks. However, in every family except TIM the interface was significantly conserved by at least one of its two definitions (Table II, Fig. 3).

The exclusion of ligand-buried residues (masking) affected *P* values by a small and usually negligible amount (Table II). Its most conspicuous effect was seen in the walking *P* values for AP, and in particular those corresponding to the total interface, where masking promoted conservation of the interface from just outside the top 7% of walks to barely within the top 5%. Masking made only a small difference because, in most cases, the residues of a protomer's ligand-buried site numbered far fewer than those of its interface. The effect of their presence or absence in a pick or walk was, therefore, small.

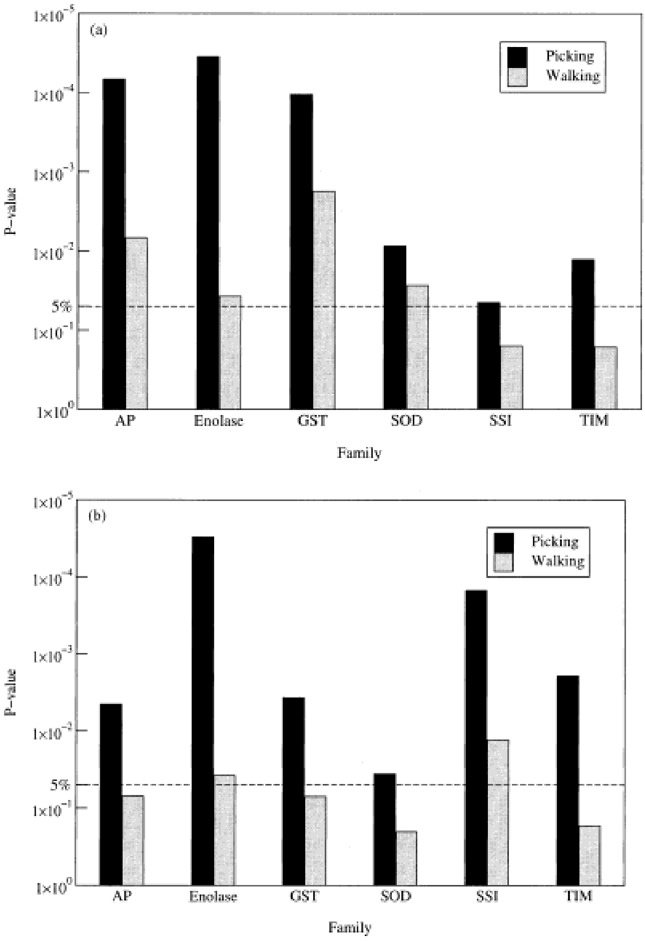

Fig. 3. *P* values for residue conservation of the interface obtained by picking and walking for six families of homodimers. Interfaces with *P* values smaller than 5%, i.e., above the dashed line, are considered significantly conserved. *P* values are shown for two interface definitions: (**a**) central zone and (**b**) total interface. The graphs show *P* values from picking are significant for both definitions of the interface in all families, and that *P* values from walking are higher and significant less often. Note that *P* values are shown for the unmasked simulations only, since masking made negligible difference to these results.

Although absolute *P* values varied between picking and walking in a protomer, the rank order between one definition of the interface and the other did not (Fig. 3). The central zone was unequivocally more conserved than the total interface in AP, GST, and SOD families. In Enolase and TIM, *P* values were similar between the two interface definitions. Only in SSI was the total interface clearly more conserved than the central zone.

Conservation in Picks

Figure 5 shows distributions of the conservation scores achieved by picks and walks. As with the patch distributions, the mean of a picking distribution estimates the mean conservation of surface residues. The higher moments relate only to the shape of the underlying distribution of conservation for individual surface residues and the number of residues chosen. They say nothing about how

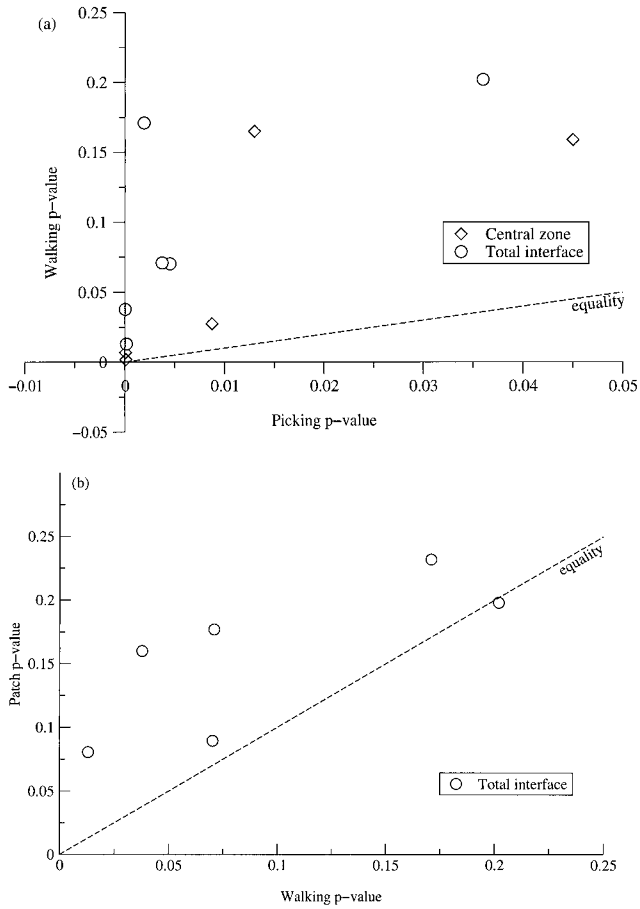

Fig. 4. Comparison of *P* values obtained by different methods. **a:** *P* values calculated by picking (x-axis) are plotted against *P* values calculated by walking (y-axis) for the same group of interface residues. The graph shows walking *P* values are consistently higher, i.e., denote less significant conservation, than those calculated by picking and that there is little correlation between *P* values from the two methods. **b:** *P* values computed by walking (x-axis) are plotted against those generated from the patch analysis (y-axis) for the same interface. The graph shows walking *P* values typically give lower, i.e., more significant, *P* values than from patch analysis and that *P* values from the two methods correlate reasonably. Note that *P* values are shown for the unmasked simulations only, since masking made negligible difference to these results.

conservation is dispersed over the surface since picking is blind to where residues are located in space.

Picks the size of the total interface fell into distributions that were narrower and more symmetric than picks the size of the central zone because they represent larger samples of the surface population and so show greater convergence toward the mean. Such properties are manifest in Figure 5a and b (iv): SOD's central zone, which comprises a mere five residues, has a wide and irregular picking distribution (Fig. 5a, iv) whereas its total interface, which comprises 20 residues, has a distribution that is symmetric and regular (Fig. 5b, iv). Owing to the greater spread of the distribution in Figure 5a (iv), the absolute level of interface conservation required to reach significance is far higher for the central zone than for the total interface. Despite this, the central zone, which has absolute interface conservation at 0.93, achieves a more significant *P* value than the total interface (Table II, Fig. 3), suggesting that evolutionary pressure to conserve residue type concentrates at the centre of the interface. In SSI, the converse is true: SSI's total interface achieves not only greater significance than its central zone but also a higher absolute conservation score (Table II). Tamura et al.[54] have demonstrated the importance of Val13, a central zone residue, to dimer formation and overall stability in SSI. The results presented here suggest residues in the outer zone also play crucial roles in this regard.

The picking results for the total interface complement the results of the patch analysis, both giving an indication of the relative conservation of the interface. Patch analysis did not provide *P* values as such, but the probabilities of a patch chosen at random being more conserved than the interface were 0.09 (AP), 0.16 (Enolase), 0.18 (GST), 0.20 (SOD), 0.08 (SSI), and 0.23 (TIM). None of these patch *P* values are less than 0.05 and they correlate poorly (with a Pearson's correlation coefficient of 0.34) with the *P* values generated by picking. Some differences are particularly striking. For example, TIM achieved significant *P* values by picking but relatively high, i.e., random, *P* values according to patch analysis. This is because picking is geometry-free and so escaped the effects of clustering that prevented TIM's interface from achieving a high patch score.

Conservation in Walks

The walking distribution for a particular protomer and its interface was always more spread out than the corresponding picking distribution, indicating that the dispersion of surface conservation in all families was more clustered than random to varying degrees (Fig. 5). This increase in distribution width marginalized the absolute conservation score of all the interfaces studied, pushing, in each case, a greater proportion of selections beyond the interface score. Thus the *P* values for walking were consistently higher that those for picking.

Figure 5 shows the walking distributions for each family. These are similar in shape to the distributions of patch scores shown in Figure 2 and corroborate inferences based on the patch data made above. The walking data give particularly strong support to these inferences because walking, unlike patch analysis, samples the space of every possible set of contiguous surface residues, which includes the interface set. The walking *P* values also correlate reasonably (with a Pearson's correlation coefficient of 0.77) with *P* values from the patch analysis, although they are significant more often (Fig. 4b).

For TIM, the similarities between walking and patch data are conspicuous (Fig. 2f, 5b, vi). TIM, the only protomer that failed to achieve a significant walking score for either definition of its interface, had walking distributions more positively skewed than those for any other family (Fig. 5a,b, vi). Most skewed of all was its distribution for the central zone (Fig. 5a, vi), where so many walks contained a majority of highly conserved residues that the resulting curve resembles an extreme value distribution. This resemblance, far from being

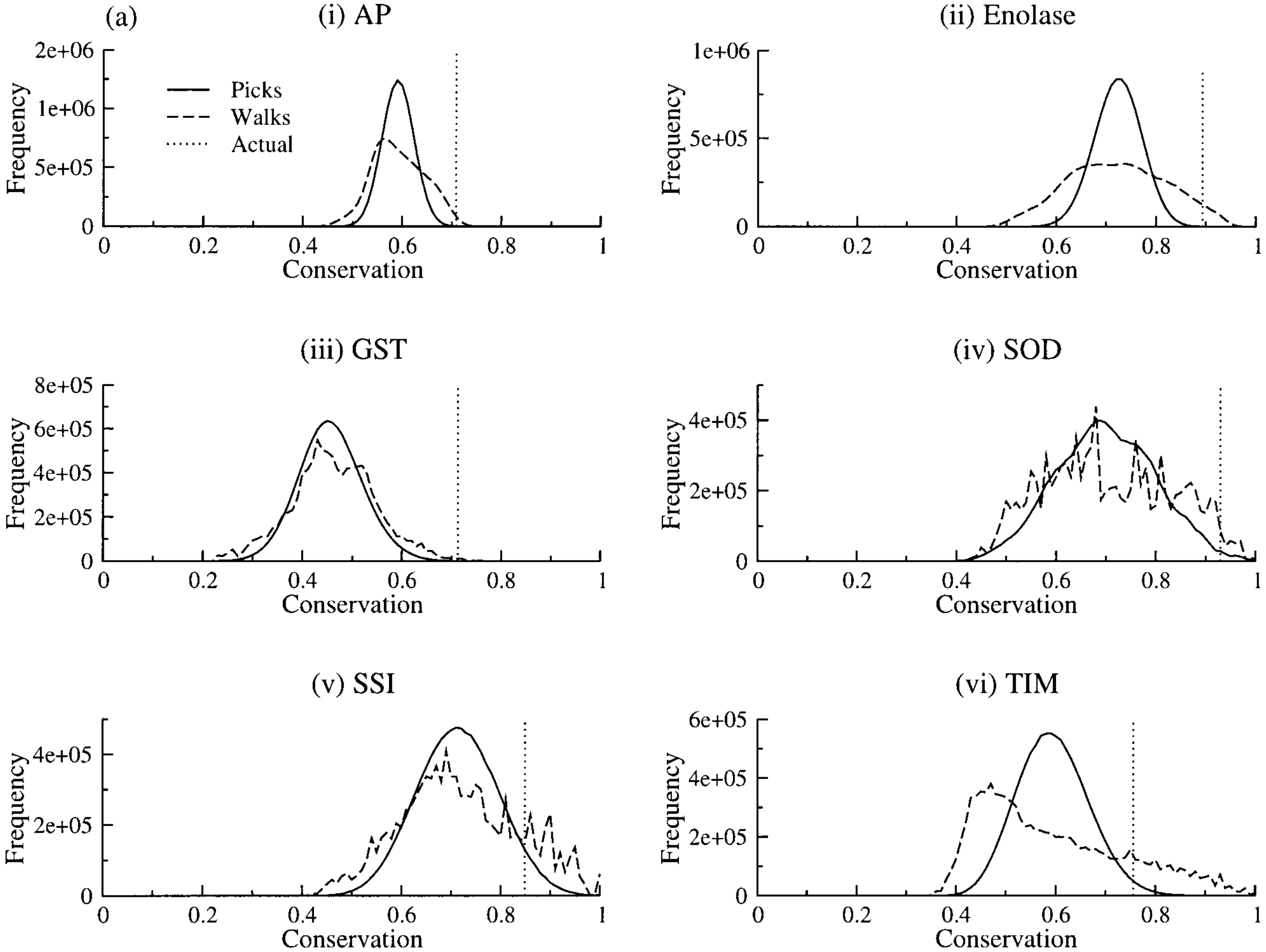

Fig. 5. Distributions of conservation for the picking and walking selection procedures. Each graph shows the distribution of conservation for 10 million picks (smooth line) and 10 million walks (dashed line) for one definition of the interface in one homodimer family. Graphs in (a) show distributions used to find p-values for the central zone; graphs in (b) show distributions used to find *P* values for total interface. In each graph, conservation, defined as the mean conservation score of residues in a selection, is presented in bins of width 0.01 along the x-axis. The number of selections that fall into a given bin is presented on the y-axis. A dotted line intersecting with the x-axis indicates the mean conservation of residues in the true interface. The *P* value determined by a given simulation is the fractional area of its curve that falls to the right of the dashed line. The graphs show picking tends to give regular, normal curves whereas walking gives irregular and often highly skewed curves. The implications of these findings is discussed in the Results section. Note that distributions are shown for the unmasked simulations only, since masking made negligible difference to these results.

coincidental, is a direct result of the surface composition of TIM, in which the gash of conservation (described above) relegates most walks that are outside it to the lower ranks of the distribution and promotes walks that overlap it to higher and higher ranks in diminishing numbers. TIM's walking results are thus consonant with its patch results and it is no surprise that its *P* values from both methods are the highest, i.e., most random, among all families.

DISCUSSION

The results show that the interfaces of all proteins studied here are more conserved than expected for a random distribution and that in most cases this conservation is statistically significant at the 5% level. In some cases, the selective pressure to remain invariant concentrates in the central zone; in others, conservation is about evenly matched across the complete interface; in one (SSI), selection against change may be strongest at the periphery of the interface.

Of the two methods used to select groups of surface residues, picking is the simplest and its results are the most straightforward to interpret. However, walking reveals more about the difficulties that would be inherent in predicting the location of interface using conservation alone. The strict definition of the ligand-buried site meant that many highly conserved residues that play a role in binding were ignored. If ligand-buried residues were defined as residues that lose merely some accessibility rather than those that become totally inaccessible on addition of ligand groups, the masked results would be different, probably giving lower *P* values in all cases except SSI,

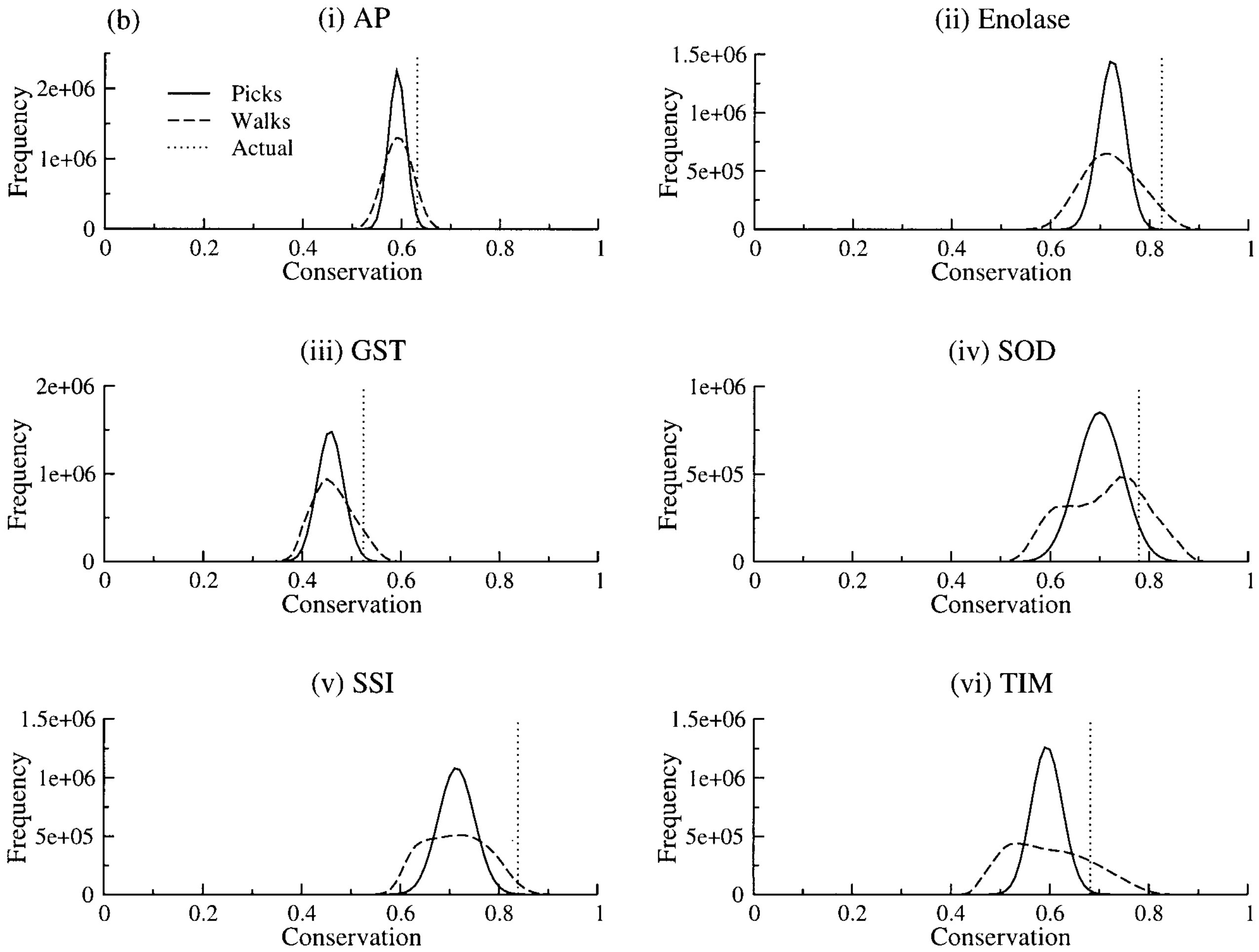

Figure 5. (Continued.)

where the *P* values would be higher. However, it was felt that to exclude so many residues from the analysis would be misleading.

The results suggest the analysis methods described here could be usefully applied to the problem of differentiating crystalline contacts from biologically relevant interfaces.[55] Proteins crystallize as multimers that may contain both biological contacts, which are subject to evolutionary constraints, and non-biological contacts, which are not. If family information is available, picking or walking *P* values could be used to detect interactions in a crystal structure that are biologically relevant.

In this analysis, we test whether conservation of an interaction is reflected in conservation of amino acid type at the site of that interaction. To ensure reliability, unusually stringent criteria were observed when compiling the dataset. For instance, it was compulsory that all sequences used in assessing conservation for a protomer share that protomer's multimeric states as explicitly recorded in their annotation. Further, in the interests of consistency, only homodimers were included: their annotation and nature of binding tend to be well documented, thus less likely to introduce confounding factors, than for other types of complexes. Such patterns may not be so distinct or be so readily detected for heterodimers or transient complexes. However, if the interfaces are functionally important, we expect them to be conserved. The challenge now is to use this information to help develop a method that can predict the location of an interface given only the structure of the protomer and its sequence alignment.

ACKNOWLEDGMENTS

William Valdar is funded by a BBSRC special studentship. We thank Dr. Roman Laskowski and Dr. Richard Mott for helpful discussions.

REFERENCES

1. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
2. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng 1988;2:101–113.
3. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
4. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990;265:16027–16030.

5. Korn AP, Burnett RM. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 1991;9:37–55.
6. Jones S, Thornton JM. Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 1995;63:31–65.
7. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
8. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
9. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 1996;260:604–620.
10. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: Characteristics and prediction. Proteins 1997;28:333–343.
11. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Eng 1997;10:999–1012.
12. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
13. Lawrence MC, Colman PM. Shape complementarity at protein–protein interfaces. J Mol Biol 1993;234:946–950.
14. McCoy AJ, Epa VC, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol 1997;268:570–584.
15. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.
16. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
17. Sternberg MJE, Gabb HA, Jackson RM. Predictive docking of protein-protein and protein-DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
18. Schreiber G, Fersht AR. Energetics of protein-protein interactions: analysis of the barnase–barstar interface by single mutations and double mutant cycles. J Mol Biol 1995;248:478–486.
19. Stites WE. Protein–protein interactions: interface structure, binding thermodynamics, and mutational analysis. Chem Rev 1997;97: 1233–1250.
20. Lakey JH, Raggett EM. Measuring protein-protein interactions. Curr Opin Struct Biol 1998;8:119–123.
21. Otzen DE, Fersht AR. Analysis of protein-protein interactions by mutagenesis: direct versus indirect effects. Protein Eng 1999;12: 41–45.
22. Clackson T, Wells JA. A hot-spot of binding-energy in a hormone-receptor interface. Science 1995;267:383–386.
23. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
24. Hedstrom L. Trypsin: A case study in the structural determinants of enzyme specificity. Biol Chem 1996;377:465–470.
25. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
26. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
27. De Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M. Three-dimensional profiles: A new tool to identify protein surface similarities. J Mol Biol 1998;284:1211–1221.
28. Grishin NV, Phillips MA. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. Protein Sci 1994;3:2455–2458.
29. Branden T, Tooze J. Introduction to protein structure. 2nd ed. New York and London: Garland; 1998. 410 p.
30. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
31. 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.
32. Henrick K, Thornton JM. PQS: A protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
33. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH - A hierarchic classification of protein domain structures. Structure 1997; 5:1093–1108.
34. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
35. Orengo CA. CORA - Topological fingerprints for protein structural families. Protein Sci 1999;8:699–715.
36. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P, Selley JN, Wright W. PRINTS prepares for the new millennium. Nucleic Acids Res 1999;27:220–225.
37. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL. The Pfam protein families database. Nucleic Acids Res 2000;28:263–266.
38. Eddy SR. Hidden Markov models. Curr Opin Struct Biol 1996;6: 361–365.
39. Bray JE, Todd AE, Pearl FMG, Thornton JM, Orengo CA. The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. Protein Eng 2000;13:153–165.
40. Etzold T, Ulyanov A, Argos P. SRS: Information retrieval system for molecular biology data banks. Methods Enzymol 1996;266:114–128.
41. Jalview: A java multiple sequence alignment viewer and editor. http://barton.ebi.ac.uk/. Clamp ME, Cuff JA, Barton GJ.
42. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins: Matrices for detecting distant relationships. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington DC: National Biomedical Research Foundation 1978; 5:345–358.
43. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
44. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
45. Hubbard SJ, Thornton JM. NACCESS [Computer Program]. Department of Biochemistry and Molecular Biology, University College London; 1993.
46. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
47. Miller S, Lesk AM, Janin J, Chothia C. The accessible surface-area and stability of oligomeric proteins. Nature 1987;328:834–836.
48. Laskowski M Jr, Kato I. Protein inhibitors of proteinases. Annu Rev Biochem 1980;49:593–626.
49. Hill RE, Hastie ND. Accelerated evolution in the reactive center regions of serine protease inhibitors. Nature 1987;326:96–99.
50. Takeuchi Y, Nonaka T, Nakamura KT, Kojima S, Miura K, Mitsui Y. Crystal-structure of an engineered subtilisin inhibitor complexes with bovine trypsin. Proc Natl Acad Sci USA 1992;89:4407–4411.
51. Kojima S, Kumagai I, Miura K. Requirement for a disulfide bridge near the reactive site of protease inhibitor SSI (*Streptomyces* subtilisin inhibitor) for its inhibitory-action. J Mol Biol 1993;230: 395–399.
52. Taguchi S, Kojima S, Terabe M, Kumazawa Y, Kohriyama H, Suzuki M, Miura K, Momose H. Molecular phylogenetic characterization of *Streptomyces* protease inhibitor family. J Mol Evol 1997;44:542–551.
53. Williams JC, Zeelen JP, Neubauer G, Vriend G, Backmann J, Michels PAM, Lambeir AM, Wierenga RK. Structural and mutagenesis studies of leishmania triosephosphate isomerase: A point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. Protein Eng 1999;12: 243–250.
54. Tamura A, Kojima S, Miura KI, Sturtevant JM. A thermodynamic study of mutant forms of *Streptomyces* subtilisin inhibitor. II. Replacements at the interface of dimer formation, Val13. J Mol Biol 1995;249:636–645.
55. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
56. DuBose RF, Hartl DL. The molecular evolution of bacterial alkaline phosphatase: Correlating variation among enteric bacteria to experimental manipulations of the protein. Mol Biol Evol 1990;7:547–577.
57. Hulett FM, Kim EE, Bookstein C, Kapp NV, Edwards CW, Wyckoff HW. Bacillus-subtilis alkaline phosphatase III and phosphatase IV: Cloning, sequencing, and comparisons of deduced amino acid sequence with *Escherichia coli* alkaline phosphatase three-dimensional structure. J Biol Chem 1991;266:1077–1084.
58. Kim EE, Wyckoff HW. Reaction mechanism of alkaline phosphatase based on crystal structures: Two-metal ion catalysis. J Mol Biol 1991;218:449–464.

59. Knowles JR. Enzyme catalysis: Not different, just better. Nature 1991;350:121–124.
60. Babbitt PC, Gerlt JA. Understanding enzyme superfamilies: chemistry as the fundamental determinant in the evolution of new catalytic activities. J Biol Chem 1997;272:30591–30594.
61. Babbitt PC, Hasson MS, Wedekind JE, Palmer DRJ, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids. Biochemistry 1996;35:16489–16501.
62. Larsen TM, Wedekind JE, Rayment I, Reed GH. A carboxylate oxygen of the substrate bridges the magnesium ions at the active site of enolase: structure of the yeast enzyme complexed with the equilibrium mixture of 2- phosphoglycerate and phosphoenolpyruvate at 1.8 angstrom resolution. Biochemistry 1996;35:4349–4358.
63. Zhang E, Brewer JM, Minor W, Carreira LA, Lebioda L. Mechanism of enolase: The crystal structure of asymmetric dimer enolase-2-phospho-D-glycerate/enolase-phosphoenolpyruvate at 2.0 angstrom resolution. Biochemistry 1997;36:12526–12534.
64. Board PG, Coggan M, Wilce MCJ, Parker MW. Evidence for an essential serine residue in the active-site of the Theta-class glutathione transferases. Biochem J 1995;311:247–250.
65. Board PG, Baker RT, Chelvanayagam G, Jermiin LS. Zeta, a novel class of glutathione transferases in a range of species from plants to humans. Biochem J 1997;328:929–935.
66. Neuefeind T, Huber R, Reinemer P, Knablein J, Prade L, Mann K, Bieseler B. Cloning, sequencing, crystallization and x-ray structure of glutathione S-transferase-III from Zea mays var. mutin: A leading enzyme in detoxification of maize herbicides. J Mol Biol 1997;274:577–587.
67. Rossjohn J, Feil SC, Wilce MCJ, Sexton JL, Spithill TW, Parker MW. Crystallization, structural determination and analysis of a novel parasite vaccine candidate: Fasciola hepatica glutathione S-transferase. J Mol Biol 1997;273:857–872.
68. Banci L, Benedetto M, Bertini I, Del Conte R, Piccioli M, Viezzoli MS. Solution structure of reduced monomeric Q133M2 copper, zinc superoxide dismutase (SOD). Why is SOD a dimeric enzyme? Biochemistry 1998;37:11780–11791.
69. Bordo D, Djinovic K, Bolognesi M. Conserved patterns in the Cu,Zn superoxide-dismutase family. J Mol Biol 1994;238:366–386.
70. Bordo D, Matak D, Djinovic-Carugo K, Rosano C, Pesce A, Bolognesi M, Stroppolo ME, Falconi M, Battistoni A, Desideri A. Evolutionary constraints for dimer formation in prokaryotic Cu,Zn superoxide dismutase. J Mol Biol 1999;285:283–296.
71. Getzoff ED, Tainer JA, Stempien MM, Bell GI, Hallewell RA. Evolution of CuZn superoxide-dismutase and the greek key β-barrel structural motif. Proteins 1989;5:322–336.
72. Hirono S, Akagawa H, Mitsui Y, Iitaka Y. Crystal structure at 2.6 angstrom resolution of the complex of subtilisin BPN' with *Streptomyces* subtilisin inhibitor. J Mol Biol 1984;178:389–413.
73. Borchert TV, Abagyan R, Jaenicke R, Wierenga RK. Design, creation, and characterization of a stable, monomeric triosephosphate isomerase. Proc Natl Acad Sci USA 1994;91:1515–1518.
74. Garza-Ramos G, Cabrera N, Saavedra-Lira E, DeGomez-Puyou MT, Ostoa-Saloma P, Perez-Montfort R, Gomez-Puyou A. Sulfhydryl reagent susceptibility in proteins with high sequence similarity triosephosphate isomerase from *Trypanosoma brucei, Trypanosoma cruzi* and *Leishmania mexicana*. Eur J Biochem 1998;253:684–691.
75. Gopal B, Ray SS, Gokhale RS, Balaram H, Murthy MRN, Balaram P. Cavity-creating mutation at the dimer interface of Plasmodium falciparum triosephosphate isomerase: Restoration of stability by disulfide cross-linking of subunits. Biochemistry 1999;38:478–486.
76. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallog 1991;24:946–950.
77. Merritt EA, Bacon DJ. Raster3D: Photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.