ARE BINDING RESIDUES CONSERVED?

CHRISTOS OUZOUNIS[1], CAROL PÉREZ-IRRATXETA[2], CHRIS SANDER[1], ALFONSO VALENCIA[2,c]

[1]The European Bioinformatics Institute, EMBL Outstation, Cambridge UK
[2]Protein Design Group, CNB-CSIC, Cantoblanco, Madrid, Spain

We present our attempt to quantify the evolutionary dynamics of functional residues in a representative set of protein structures and their homologous sequences. Using the log-odds formalism, the preference for all twenty amino acids to be conserved or participate in binding (or active) sites is examined. It appears that while there is a tendency for functional residues to be conserved, the two preference scales do not coincide. Remarkable differences between amino acid types emerge from this comparative study. The current approach is expected to lead towards a better understanding of functional site architecture in proteins.

1 Introduction

1.1 What is the role of conserved residues?

It is generally accepted that functional and structural constraints in proteins lead to the conservation of the chemical character of amino acid residues in polypeptide chains as observed in multiple sequence alignments of protein families [1-7]. Yet, despite the growth in sequence and structure information, it is still unknown how the conservation of particular residues affects structure and function in a quantitative way [2, 8-17] It is still rather unclear how particular residue types respond to evolutionary change, and which amino acids are most frequently involved in protein function [17, 18].

Some intuitive rules have been followed over the years: for example, conserved histidines or aspartates are first replaced by site-directed mutagenesis. Here, we attempt to answer the question of how to distinguish functionally important residues in relation with conservation, in a large representative set of protein families.

1.2 Which residues participate in protein function?

Multiple sequence alignment information has been successfully used in problems of structure prediction. Another important property of protein molecules, which is less liable to quantification – due to its loosely definable nature – is protein function. While structure is a global property of protein molecules and sustained over evolution, function is conferred to a protein by short regions of sequence, which in

[c] Corresponding author: valencia@cnb.uam.es

three dimensions form the active site or binding sites for small molecules or other macromolecules [19, 20].

Protein function can be defined as any property that represents substrate and coenzyme binding, enzymatic catalysis, regulation and effector interactions. A huge body of knowledge exists on the experimental characterization of amino acids in proteins [21]. These residues are often selected on the basis of their character and relative invariance during evolution.

Experimental analysis can be assumed to have been a biased choice of certain conserved residue types, ignoring a number of functional residues, which may or may not be conserved. Thus, even today we are not aware about the propensities of the twenty residue types to participate in function, and to what extent the common tenet that functional residues are conserved is true for different amino acids.

2 Methods

2.1 Definition of functional residues

Binding residues are defined as those in which at least one atom is in contact with a chemical compound included in the same entry. Contacts are defined as amino acid atoms at a distance less than 4 Å - or 2.55 Å when the chemical compound is a metal atom: the distance was set to a smaller radius for metal contacts to reflect the physical nature of these interactions [22]. All chemical compounds included in PDB files are counted with the exception of solvent molecules. These were identified by labels in the HETAM (heteroatom) records of the PDB files, e.g. "ACE", "BME", "BR", "CL", "DOD", "HOH", "MOH", "DIS", or by the presence of the same HETAM label more than three times, signifying the presence of a solvent molecule. A particular exception were those water molecules mediating contacts of a metal ion within a protein. Amino acids at a distance less than 4 Å from a water molecule and at the same time less than 2.55 Å from a metal ion were thus considered as binding.

2.2 Database construction

A representative set of protein structures was selected from the PDB-select list [23] which includes proteins with a maximal level of 25% sequence similarity. Three additional criteria were applied to exclude protein families with a biased sequence representation: first, only alignments with five or more proteins were considered; second, entries with at least three detected binding residues (positions) were used; and third, families with a number of invariant (absolutely conserved) residues smaller than 20% of the alignment length were discarded, where alignment length is defined as the number of positions occupied by more than 4 sequences. The resulted number of families was 140, containing 1,181,424 residues of which 4.7% (55,701) are defined as binding residues. An assumption in this analysis is that all residues aligned with the binding residue of the guide structure are also considered to be binding residues.

2.3 Preference parameters

Sequence variability was defined as the frequency of residue types found in each position. Gaps were not considered in this calculation. Variability is defined as:

Variability = int[(1-frequency aa type)*100]
with Conservation = 100-Variability.

Variability was divided in 7 intervals: (less or equal to) 0, 4, 10, 15, 50, 81, 100. These intervals were selected so that they contained approximately the same number of observations. At each position of the multiple sequence alignment, each residue is counted as a single observation. For example, Val found in four proteins at a given position of a protein family counts as four observations of Val. The frequency of the observations of two variables, i.e. residue type at a level of variability or conservation, was compared with the expected frequency if the two variables were independent. The tendency was expressed in log-odds (as in [24]), leading to a positive value if the pair observations were more frequent than expected and a negative value if less frequent than expected:

$$\log_2[(f[ij])/(f[i]*f[j])]$$

where f[i] and f[j] are the individual frequencies of two variables, e.g. variability and residue type, and f[ij] is the combined frequency of a pair observation of residue types and conservation levels.

3 Results

3.1 Which residues tend to be conserved?

We have calculated log-odds for the potential of residue types to remain conserved. We define the set of these values as the conservation potential (Table I). The levels of conservation are estimated as percentage of conserved residues per position. Interestingly, glicine and cysteine have the highest values, followed by several residues with positive values: proline, tryptophan, histidine, and surprisingly, aspartate (but not glutamate) and arginine (but not lysine) (Table I, Figure 1a).

Other amino acids have negative preference values to be conserved, strong for isoleucine, valine, alanine and weaker for methionine, glutamine, serine, lysine and threonine (Table I, Figure 1a). An interesting comparison is between residues of similar character: for example, arginine over lysine, or aspartate over glutamate. The aromatic amino acids in general tend to be conserved, while other hydrophobic residues do not (Figure 1a). It should be noted that no notion about functional residues is held at this point.

TABLE I. CONSERVATION AND BINDING POTENTIAL FOR THE TWENTY AMINO ACID TYPES.

	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**
CON *	-1.35	1.02	0.67	-0.38	0.29	1.42	0.77	-2.13	-0.91	-0.41
BIN **	-0.60	1.35	0.38	-0.51	0.26	-0.30	1.73	-1.08	0.18	-0.15

	M	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
CON *	-1.15	-0.34	0.86	-1.12	0.48	-1.11	-0.79	-1.48	0.84	0.31
BIN **	-0.38	0.10	-1.36	-0.56	0.23	0.29	-0.1	-0.61	-0.24	0.03

CON *, Conservation potentia. BIN **, Binding potential

One issue is how does this scale differ from the diagonal of amino acid substitution tables, where the frequency of exchange to the same residue type is expressed. In the present case, non-substitution counts take into consideration only conserved residues, while the diagonal of substitution tables is composed of two elements: conserved non-substitutions and non-conserved substitutions. Therefore, our derived conservation potential concerns only conserved residues in invariant positions. Thus, the scale differs in essence from the diagonal values of amino acid substitution tables (data not shown).

3.2 Which residues tend to be functional?

Using the same formalism, we have also calculated log-odds for the potential of residue types to be functional, based on contacts with bound heteroatoms. We call these values the binding potential (Table I). Two residue types, cysteine and histidine, have distinctly high values. Another five types, aspartate, serine, phenylalanine, arginine and lysine marginally positive preferences. Eleven types display negative preferences, with proline and isoleucine having the most negative values and valine, alanine, glutamine, glutamate and methionine having negative tendencies (Figure 1b). The comparison between residues of similar character is again instructive: two contrasting cases are the aspartate/glutamate and asparagine/glutamine pairs. It should be noted that no notion about conservation is held, in analogy with the conservation potential.

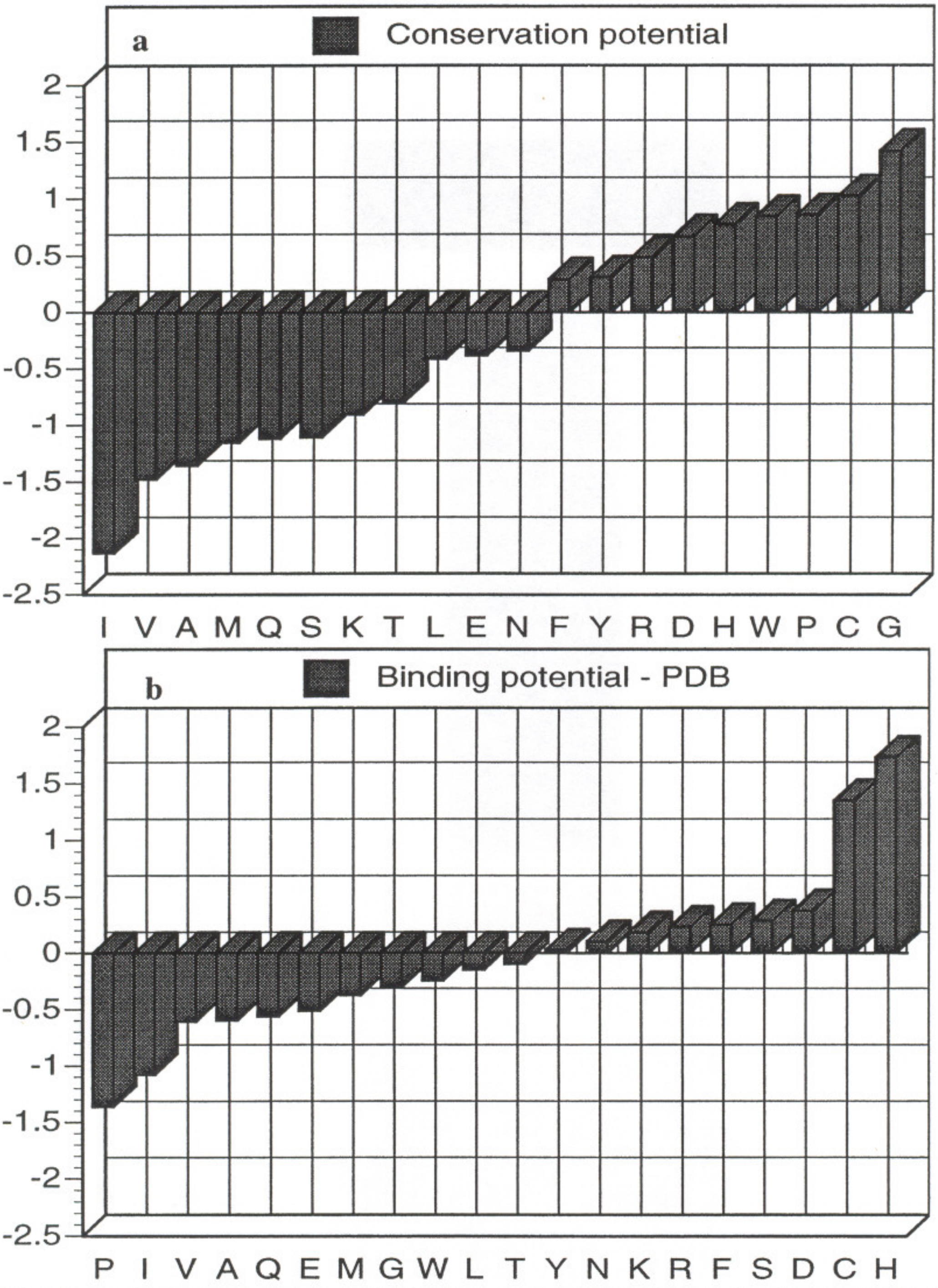

Figure 1: Preferences for (a, upper panel) conservation and (b, lower panel) and binding potentials for the 20 amino acids. Preferences are calculated as logodds of the conditional probability for the observation (conservation or binding) over the frequency of the particular residue type (see Methods). (1a): There are four groups of conservation, with the group [CG] having the highest tendency to be conserved, the group [PWHDR] having less such tendency, the group [TKSQM] having less tendency to be conserved, while the last group [IVA] being the most non-conserved residues. (1b): The same scale for binding preferences, where there are four groups that are identified: [CH] with a strong preference for binding heteroatoms, [DSFRK] with very weak preference, [MEQAV] with a tendency not to be part of a binding site, and finally [IP] that have a stronger negative preference.

Obviously, the two scales for the conservation and binding potentials differ. For example, tryptophan and glycine tend to be conserved, without participating in function, as defined here (contacting heteroatoms, binding). On the other hand, histidine has a higher tendency to be functional than being conserved. The only residue which has high values for both conservation and binding is cysteine. Thus, conservation does not seem to be the only factor determining binding potentials. We have dealt with this important issue in considerable detail, namely the question of how different residue types at various levels of variability affect preferences for binding sites

3.3 Different measures for binding potentials

The previous results are based on the analysis of binding sites in three-dimensional protein structures. It is conceivable that other sources of information could be used to increase or correct the assignment of residues as participants in active or binding sites. We have explored the use of information contained in protein databases and in particular SWISS-PROT, as an alternative to the definition of binding residues. We have extracted all residues which are labeled as "BIND/BINDING", "METAL", "ZN", "ACT/ACTIVE" in SWISS-PROT and calculated the function potentials as above. An important observation is that for the same set of proteins, the total number of residues labelled in SWISS-PROT is much smaller than the one found contacting heteroatoms.

When these are compared with the potential values derived by the contact definition (above), the main trends are similar, despite some exceptions and can be considered as supportive of our calculated binding sites (Figure 2). On the one hand, glicine, and to small extent aspartate and lysine are very frequently annotated as binding residues in SWISS-PROT (these cases, if annotated correctly, may represent underestimated observations using the structural data, e.g. coenzyme, substrate or other heteroatom missing). On the other hand, tryptophane, leucine, argnine, phenylalanine and tyrosine are rarely indicated as binding residues in SWISS-PROT (possibly representing underestimated observations from experiments).The annotations in the sequence databases are only in part derived from direct experimental approach and more frequently derived from intuitively defined sequence patterns by similarity. It appears that in annotations, experts avoid the definition of Pro and hydrophobic side chains as binding and tend to prefer glicine or aspartate. It is shown below that there are no objective reasons to support these decisions.

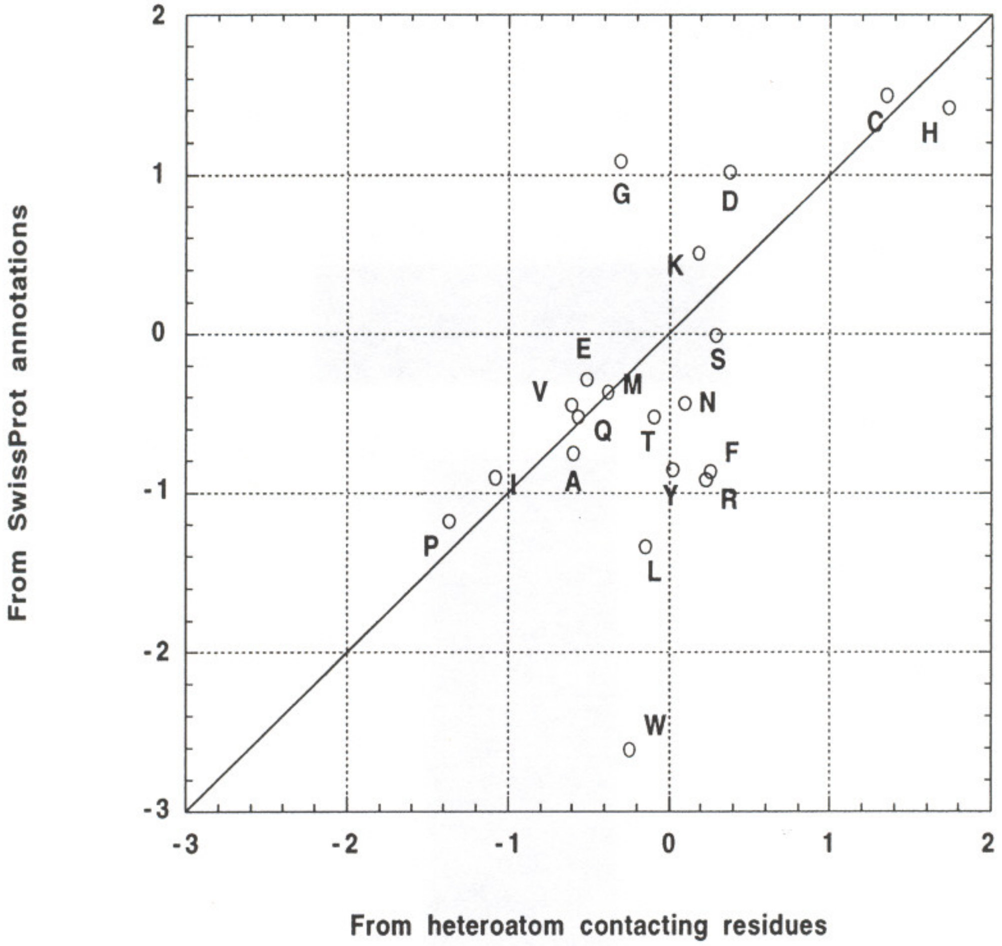

Figure 2: Binding potentials as derived from the current dataset using structural information and heteroatom contacts contrasted with the same scale derived from SWISS-PROT, for residues annotated as ACTIVE_SITE or BINDING_SITE (see Methods).

3.4 Which conserved residues are functional?

The scales for the function and conservation potentials shows that the two are not correlated (Figure 1). Since the two scales of preferences are different, it is instructive to ask the following question: which conserved residues are functional, and how does this correspond to their general tendency? Obtaining the same scale of values for completely conserved residues and comparing to the 'function potential' scale without any notion of conservation (Figure 3), results in two important observations: First, even if the order is generally preserved, there are significant differences. Extreme cases are isoleucine and valine, which, surprisingly, have a reinforced anti-tendency to be binding when they are conserved, and proline, glutamate, glicine and triptophane (among others), which show tendency to be in binding sites only when invariant. Second, the scales changes: while the binding preferences without the notion of conservation fluctuate between -1.5 and 2, when only conserved residues are considered there values vary between -2 to 4.5 (Figure 3). Therefore, conservation enhances the contrast mainly for preferences but also for *anti-preferences* . For example, tyrosine have no tendency to be involved in binding sites (Log-odd 0), while *conserved* tyrosines are more likely to be binding (Log-odd 2.2). Therefore, it is compelling that in such an analysis the binding potential should taken in consideration the conservation potential and not only invariant positions. This fact has led us to a derivation of a binding potential using various levels of conservation.

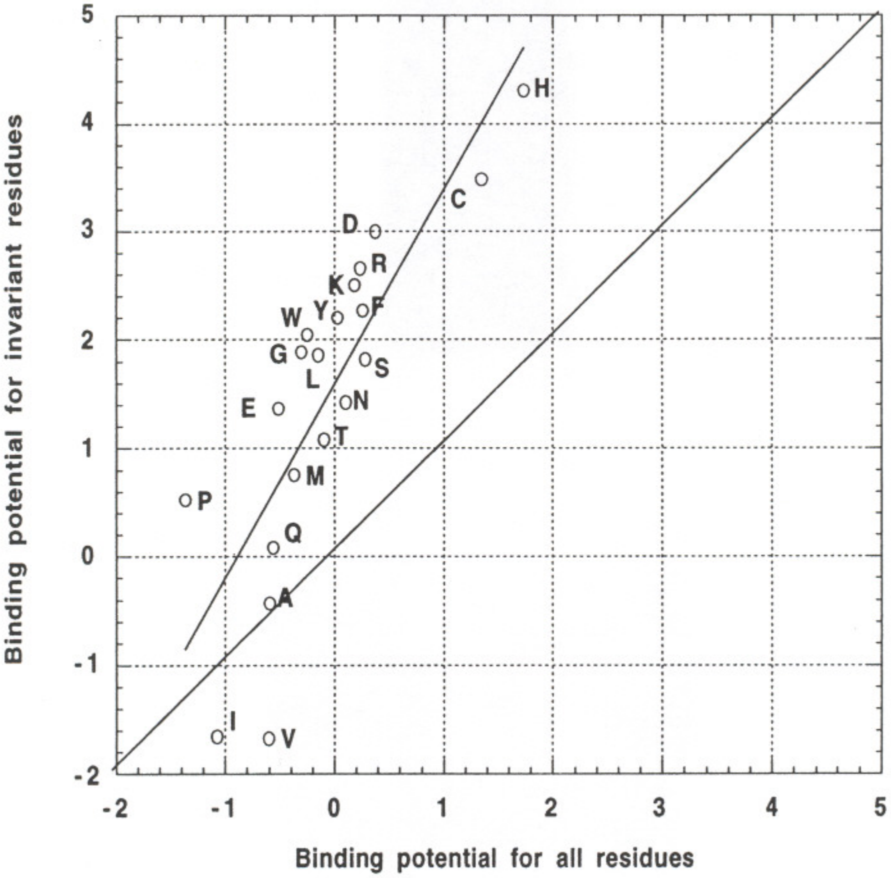

Figure 3: The relationship between the binding potentials for all (x-axis) or only conserved (y-axis) residues. Although there is some correspondence between the two scales, most residues have more tendency to be binding when completely conserved, while some (Ile and Val) have stronger anti-preferences if conserved.

3.5 Which functional residues are conserved?

The obvious extension of this approach is to derive function potentials not only for invariant residues, but for all different levels of conservation with regard to the residue types. We have derived the log-odds for the potential of residue types to be functional for seven (mutually exclusive) levels of conservation (see Methods). It is evident that different residues have variant behaviour with respect to their conservation levels (illustrated for two residues, Figure 4).

The contribution to the overall tendency for being part of a binding site strongly depends on the relative contribution of the seven classes towards this effect. For example, non-conserved asparagine tend to be parts of binding sites more frequently than non-conserved aspartate and conserved aspartate tend to participate in binding sites more than conserved asparagines (Figure 4). It can be generally expected that conserved residues have more extreme function potential values than variable residues, or in other words, that conservation enhances the preference or anti-preference for a residue to be in a binding site

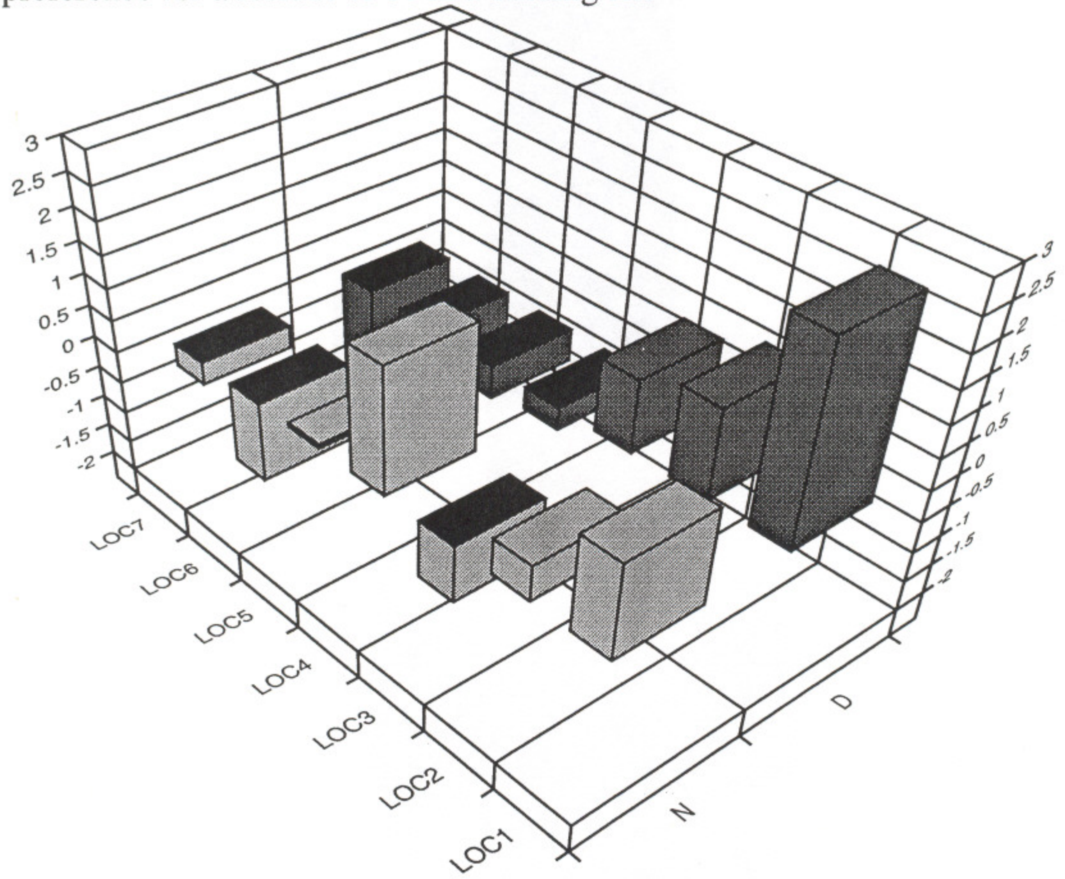

Figure 4: Illustration of the methodological approach to delineate the relationship between conservation and variability in two particular cases, for aspartate and asparagine. In the case of aspartate, there is a larger proportion of binding or conserved residues, while at the same time binding residues tend to be more conserved. In the case of asparagine, there are less binding or conserved residues, but interestingly the proportion of variable residues involved in binding has a significant contribution towards the total value (in particular class 4). LOC stands for Log-Odds at the different variability Classes (1-7).

3.6 Which residues are "regular"?

If there was a good correlation between conservation and function (as is generally assumed), we should observe monotonically decaying lines for different levels of variability (i.e. aspartate in Figure 4 and 5). In contrast, it is evident that some residues are 'well-behaved' with respect to conservation, i.e. their tendencies correspond to this expectation, while others do not (i.e asparagine in Figure 4 and 5). In some cases the deviations are sharp. For instance, in the case of tryptophane, glutamine and methionine there is a very sharp contrast between high and low conservation (Figure 5). Even more interesting are cases such as cysteine or glutamate: even when variable, they still display high preferences for binding. These fluctuations come from the fact that residues have different chances of functionally meaningful substitutions. Suggesting that it may be more difficult to replace an invariant arginine that forms part of an active or a binding site than an invariant asparagine.

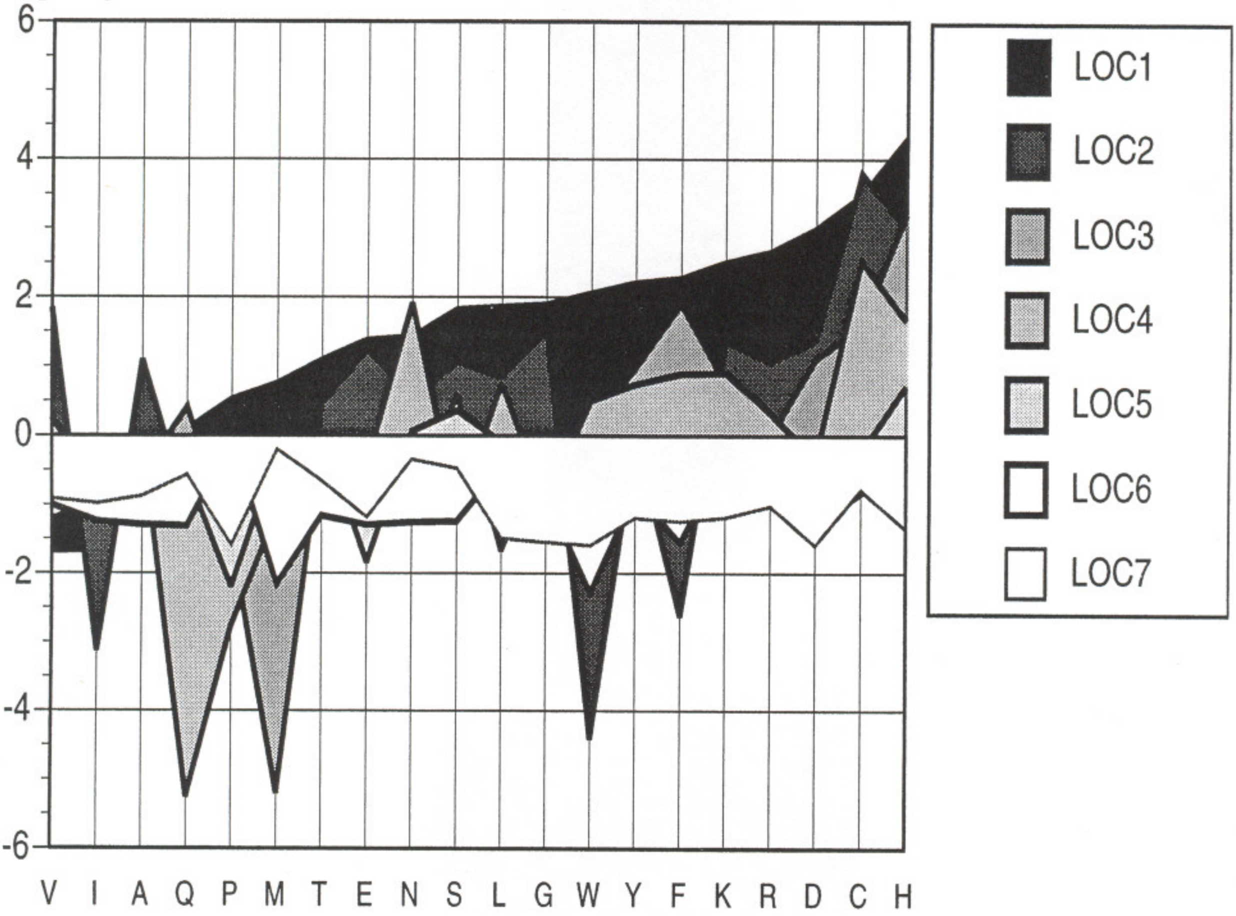

Figure 5: Binding site preferences for the 20 amino acid types at different levels of conservation, as defined in Methods (and previous figure). The upper line corresponds to completely conserved residues. Different amino acid types display a variable behaviour with respect to conservation/variability. While most residues have an 'expected' decay of the binding preference signal with decreasing level of conservation, for example aspartate, other residues have a significant tendency to participate in binding even when they are not highly conserved, for example glutamine, while yet others lose the tendency to be in binding sites as long as they are not conserved, for example tyrosine. LOC stands for Log-Odds Class (1-7), as above.

3.7 Enhancing information content

There are some obvious extensions to the current approach. The definition of functional residues depends on the integrity of the structural data, which is far from being perfect. In some instances, the substrate or the coenzyme may be absent in the crystal structure and some interactions in the active or binding site may not be accounted for. Ways of improving this situation may come from better experimental data and methods of predicting the active and binding sites from the structure in the absence of ligands. In addition, the limited data necessitates the use of only structural information at present, while in the future sources like sequence databases can in principle be combined with the existing approach. In particular the consideration of protein clusters (sub-families) should be addressed in the calculation of conservation, since part of the variability observed in binding sites is very likely originated by specific adaptations of different protein sub-families[14, 25].

4 Discussion

4.1 Protein function from a structural viewpoint

The analysis and prediction of functional residues in proteins may not only contribute to a better understanding of protein function and become a tool for computer-guided experimentation, but can also provide valuable insights into the problem of active and binding site architecture in proteins [2, 13, 15, 17, 21, 26-28].We are further analyzing the structural properties of the binding sites as defined by heteroatom contacts, for example accessibility, secondary structure location and neighbour properties (conservation, types). This may eventually lead to a better understanding of binding site architecture and its prediction.

4.2 Protein structure from a functional viewpoint

In the present report, we have undertaken the quantitative analysis of functional residues and demonstrated that amino acid residues have distinct preferences for binding coenzymes and substrates. Also, some residue types of very similar chemical character display asymmetries in their preferences towards function and conservation. In addition to providing a scale for these preferences, it will be possible to perform predictions for functional residues [17], given a multiple alignment (in preparation). These predictions may be more powerful than the simple traditional approach of combining intuitive assumptions and the trivial identification of conserved residues.

Acknowledgments

This work has been supported by grants from: CICYT Spain and Biotech.-EU.

References

1. D.S. Sigman and M.A.B. Brazier. (Academic Press, New York, 1980)
2. S.A. Benner and D. Gerloff. *Adv. Enzyme Regul.* **31**, 121-181 (1991)
3. S. Brenner. *Nature* **334**, 528-530 (1988)
4. B.S. Cooperman, A.A. Baykov and R. Lahti. *Trends Bioch. Sci.* **17**, 262-266 (1992)
5. N. Howell. *J. Mol. Evol.* **29**, 157-169 (1989)
6. P.K. Hwang and R.J. Fletterick. *Nature* **234**, 80-83 (1986)
7. E. Zuckerkandl and L. Pauling. in *Evolving Genes And Proteins* (eds. Bryson, V. & Vogel, H.J.) 97-166 (Academic Press, New York, 1965).
8. A.R. Poteete, D. Rennell and S.E. Bouvier. *Proteins* **13**, 38-40 (1992)
9. G. Casari, C. Sander and A. Valencia. *Nature Struct. Biol.* **2**, 171-178 (1995)
10. A.P. Golovanov*, et al. FEBS lett.* **375**, 162-166 (1995)
11. S. Ohno. *J. Mol. Evol.* **40**, 102-106 (1995)
12. A. Pastore and A.M. Lesk. *Proteins* **8**, 133-155 (1990)
13. C. Perez-Iratxeta, B. Rost and A. Valencia. *submitted* (1997)
14. G. Pujadas, F.M. Ramirez, R. Valero and J. Palau. *Proteins* **25**, 456-472 (1996)
15. R.B. Russell, J. Breed and G.J. Barton. *FEBS Lett.* **304**, 15-20 (1992)
16. W.R. Taylor. *J. Theor. Biol.* **119**, 205-218 (1986)
17. M.J. Zvelebil, G.J. Barton, W.R. Taylor and M.J.E. Sternberg. *J. Mol. Biol.* **195**, 957-961 (1987)
18. A.E. Gabrielian, V.S. Ivanov and A.T. Kozhich. *CABIOS* **6**, 1-2 (1990)
19. D.S. Goodsell and A.J. Olson. *Trends Bioch. Sci.* **18**, 65-68 (1993)
20. L. Pearl. *Nature* **362**, 24-24 (1993)
21. D. Ringe. *Curr. Opin. Struct. Biol.* **5**, 825-829 (1995)
22. M.M. Yamashita, L. Wesson, G. Eisenman and D. Eisenberg. *Proc. Natl. Acad. Sci. USA* **87**, 5648-5652 (1990)
23. U. Hobohm, M. Scharf, R. Schneider and C. Sander. *Protein Sci.* **1**, 409-417 (1992)
24. C. Ouzounis, C. Sander, M. Scharf and R. Schneider. *J. Mol. Biol.* **232**, 805-825 (1993)
25. A. Valencia, P. Chardin, A. Wittinghofer and C. Sander. *Biochemistry* **30**, 4637-4648 (1991)
26. T. Hubbard Tramontano, A., and the 1995 IRBM workshop team. *Folding &Design* **1**, R55-R63 (1996)
27. A. Valencia*, et al. Proteins.* **22**, 199-209 (1995)
28. B. Rost and C. Sander. *Current Opinion in Biotech.* **5**, 372-380 (1994)