doi:10.1016/j.jmb.2003.12.078 J. Mol. Biol. (2004) **336**, 1265–1282

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

A Family of Evolution–Entropy Hybrid Methods for Ranking Protein Residues by Importance

I. Mihalek, I. Reš and O. Lichtarge*

Department of Molecular and Human Genetics
Baylor College of Medicine
One Baylor Plaza T921
Houston, TX 77030, USA

**Corresponding author*

In order to identify the amino acids that determine protein structure and function it is useful to rank them by their relative importance. Previous approaches belong to two groups; those that rely on statistical inference, and those that focus on phylogenetic analysis. Here, we introduce a class of hybrid methods that combine evolutionary and entropic information from multiple sequence alignments. A detailed analysis in insulin receptor kinase domain and tests on proteins that are well-characterized experimentally show the hybrids' greater robustness with respect to the input choice of sequences, as well as improved sensitivity and specificity of prediction. This is a further step toward proteome scale analysis of protein structure and function.

Keywords: bioinformatics; evolutionary trace; information entropy; functional site; binding interaction

Introduction

When doing protein mutation studies, it is helpful to have an estimate of the relative importance of residues as a guide. In this way, priority can be given to the residues more likely to play a critical role in the protein function or structure. Here, we re-examine how to rank residues by importance, starting with a set of homologous sequences.

All of the alignment or evolutionary residue-scoring methods assume that the importance of a residue is reflected in its evolutionary conservation: the more important the residue, the sooner it becomes fixed in different evolutionary branches, and the more divergent are the branches between which it does vary. (As a working definition of "important residue", we might take "the residue that cannot be mutated without measurably affecting the protein structure or function.") There are various approaches to turning this observation into a quantitative prediction of relative residue importance: scoring strict conservation, property conservation, or entropy of a position, or, more elaborately, scoring conservation in related families (even if not across the families).

In one of the earliest attempts to quantify the conservation of a residue at certain alignment position, Zvelebil *et al.* converted the count of residues with a certain property (hydrophobicity, size, etc.) into a property called conservation number, which enabled them to distinguish poorly conserved loop regions from the rest of the protein structure.[1] For the current progress in this line of thought, see Valdar.[2]

Casari *et al.* proposed an interesting method in their 1995 work: consider the whole sequence as a vector in number-of-residue-types times length-dimensional space, and think of the alignment as a matrix of such vectors.[3] Its eigenvectors should then carry the information about residue preference for each subfamily represented in the alignment. In that way, the tree information is recovered from the analysis, rather than being its input. The information-theoretic school likes to place the root of its genealogical tree straight at the historical work of Shannon & Weaver, where the entropy of a finite state system was reinterpreted as a measure of its information content.[4] With the advance of computational methods in genetics, the idea resurfaced in connection with the information content and thermodynamics of DNA-binding sites.[5,6] In 1991 Shenkin *et al.* used entropy as a robust measure of variability of positions in immunoglobulin sequence, and noted that high variability of a position can be a result of its evolutionary neutrality.[7] More recently, entropy-based measures of position conservation have been used for systematic computational

Abbreviation used: ET, evolutionary trace.

E-mail address of the corresponding author: lichtarge@bcm.tmc.edu analysis of conservation profile in multiple sequence alignment,[8,9] and the approach was also elegantly extended to detect correlated mutations in a sequence.[10,11] Mirny & Shakhnovich[12] as well as Hannenhalli & Russell[13] introduced the notion of summing or averaging the site entropy over several related protein groups, but stopped short of iteratively applying this approach to the hierarchical division of sequences into groups induced by evolutionary tree, the idea that we propose here.

In a parallel and independent development, Lichtarge *et al.*,[14] as well as Livingstone & Barton,[15] pointed out that low variability is not equivalent to the conservation of a residue within a subgroup (or a subtree), and that the knowledge of such within-group conservation can be used successfully in estimation of the residue importance. A similar observation was made by Ptitsyn in the context of structurally important residues.[16] While Livingston & Barton incorporated and built onto the work of Zvelebil *et al.*, Lichtarge *et al.* took an all-or-none approach in considering the within-group conservation, but pointed out how to include the tree information in a systematic, iterative way. The method was named evolutionary trace (ET), and has since been shown capable of detecting protein interaction sites and directing protein mutation studies.[17–20] The present work grew out of the effort to make the ET more robust against deviations from the ideal family-tree picture, occurring in the actual protein evolution (and database-dependent research).

A comparative study of various methods mentioned, in terms of their capability to rank the residues, was, to the best of our knowledge, never performed (see del Sol Mesa *et al.*[11] for comparison of several methods' ability to pick residues physically close to functionally important residues). In good part, the reason is that it is impossible to obtain from the experiment equivalent and independent information, an experimental yardstick against which to measure the performance of various theoretical approaches. This would involve mutation of every single residue in the protein, or at least sizable portion thereof, presently not a feasible option. As an alternative, we construct by literature search a tentative key residue set for several well-investigated proteins, a task in which we are greatly assisted by Protein Mutant Database.[21] We then estimate the quality of a method by its capability to rank the members of the key set highly. In other words, taking this set as a "gold standard", we study the sensitivity–specificity performance of a method. The methods we focus on in this work are entropy, ET, and two hybrid methods. We propose a general way to construct a hybrid, illustrate the use of these methods using insulin receptor kinase domains as an example, and find that combinations of ET and entropic approaches are more robust against small irregularities in the input, and have increased prediction sensitivity and specificity.

Theory

In this section we want to lay out the framework for incorporating entropy into a tree-based residue ranking system (or *vice versa*). To this purpose, we need to define some terminology: node ordering in a binary tree, and hierarchical division of leaves

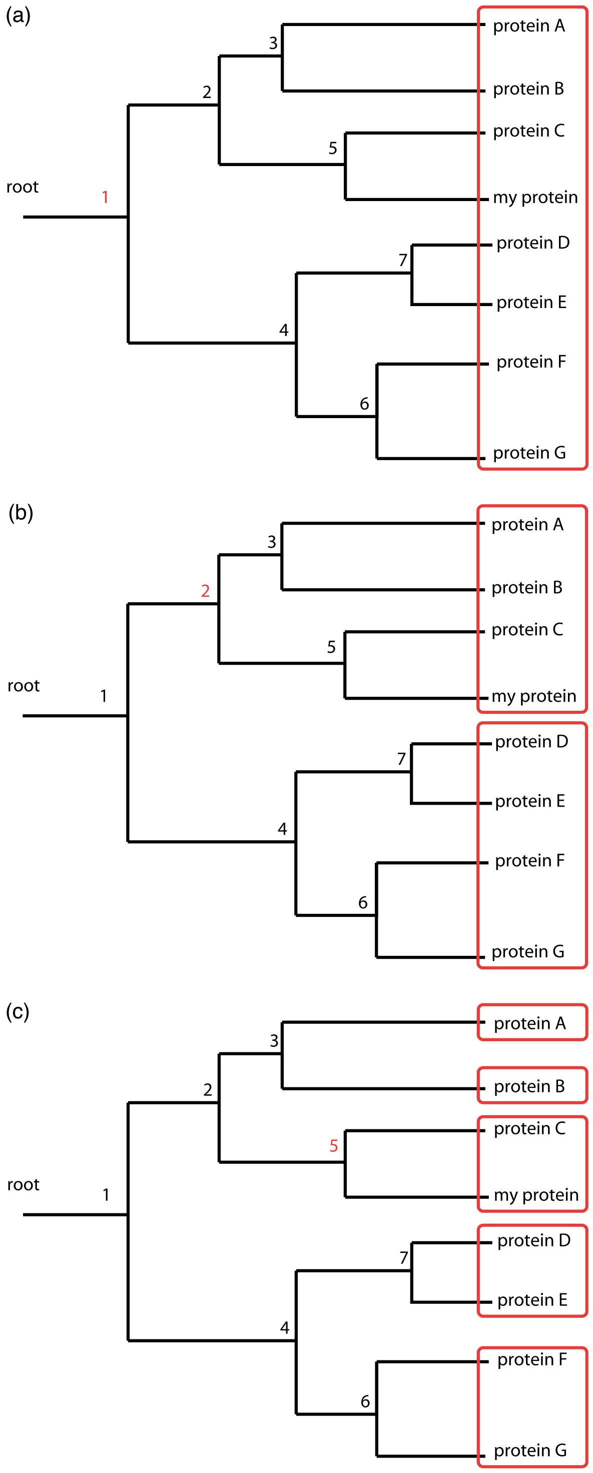

Figure 1. Node ordering and hierarchical division of leaves into groups. (a) Node $n = 1$ corresponds to all leaves belonging to the same group; (b) $n = 2$ to two groups; and (c) $n = 5$ to five groups.

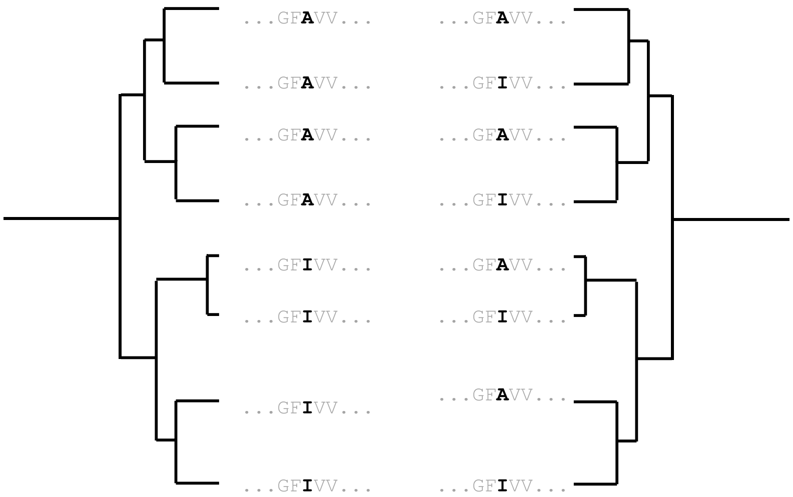

Figure 2. Conservation estimate using information entropy of a column *versus* evolutionary trace: the entropy method sees no difference between these two cases.

into groups induced by the tree. We express the existing methods (ET and entropy) in these terms, and review their complementary strengths. Next, we propose a straightforward combination of the two, and point out a way to think about this unification in quite general terms.

In the alignment-based trees the number of sequences in the multiple sequence alignment, N, equals the number of leaves in the tree. The nodes in the tree can then be numbered iteratively in an ordered way, following the UPGMA method:[22] starting with $n = N$ leaf nodes, replace the two nearest nodes by a new one. Assign to the new node the number $n - 1$, decrease n by 1, and repeat the iteration. Notice that in this way the nodes close to the leaf level get assigned big numbers, while the root is labeled as node $n = 1$. Implicit in this numbering scheme is the hierarchical division of leaves into groups; at each step in the iteration corresponding to the assignment of node number n, the leaf set is divided into n groups (Figure 1). Each group is labeled by a number $g = 1,\ldots,n$, the ordering of which is unimportant.

With the above conventions in mind, we can express the ET score for the residue at position i as follows:

$$r_i = 1 + \sum_{n=1}^{N-1} \times \begin{cases} 0 & \text{if position } i \text{ conserved within each group } g \\ 1 & \text{otherwise} \end{cases} \qquad (1)$$

(1 is here for historical reasons.) In other words, the summation stops at a division into groups such that the position i is conserved within each group. (Any further subdivision preserves that property.) Assigning r_i to each residue leads to a relative ranking scheme: given any two residues, the one with smaller r_i is considered more important. It is the ordering of alignment positions according to r_i that matters, rather than the values of r_i themselves. This is the property shared by all the methods we discuss.

In the approach relying directly on the entropy of each alignment column, each position is

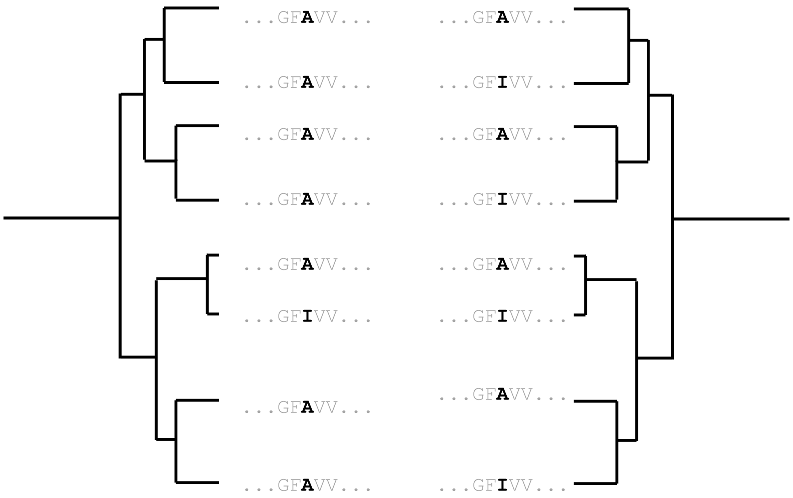

Figure 3. Conservation estimate using information entropy of a column *versus* evolutionary trace: to the evolutionary trace the two columns appear equally unimportant.

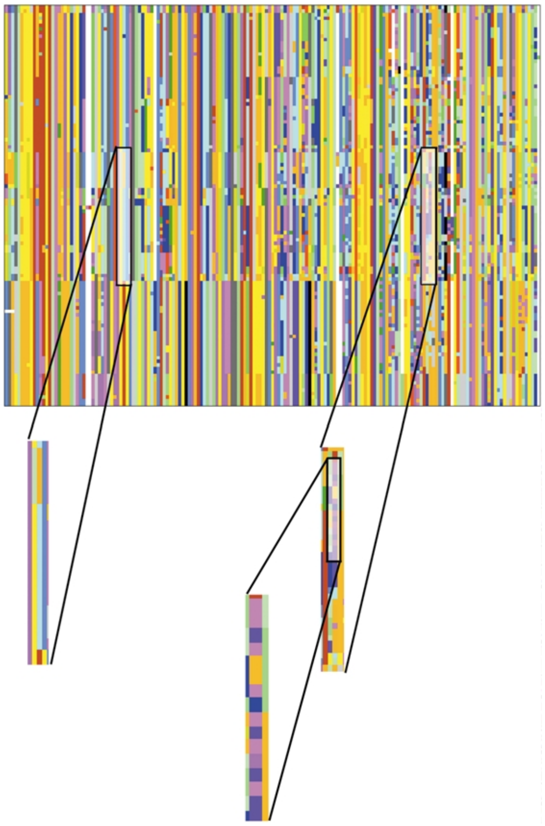

Figure 4. A look at the alignment columns with increasing resolution. The column on the right shows persistent variability in ever smaller groups of sequences. Equation (4) will assign a higher score to that column than to that on the left.

assigned the information entropy:

$$s_i = -\sum_{a=1}^{20} f_{ia} \ln f_{ia} \qquad (2)$$

and ranked accordingly. Here, f_{ia} stands for the frequency of the appearance of the amino acid of type a within the column i. (Contribution to the sum from the amino acid types that are not represented in the column is to be understood as 0.) The frequency is simply the count of times an amino acid type appears within a column, divided by the length of the column. The entropy can be thought of as a score that is assigned exclusively to $n = 1$ group in our tree-induced division into groups. An important property of this quantity is that it is equal to zero when the column is completely conserved, and maximal when every amino acid type is equally represented. Because of that, the best ranking positions are again those with the smallest s_i. These two methods may give different but complementary results, as we illustrate in the two examples in Figures 2 and 3. In both, a hypothetical column from an alignment is shown, together with the evolutionary tree for the proteins that the whole sequences belong to. The entropy cannot distinguish between the two columns in Figure 2, while the ET correctly recognizes that the column on the left can be split into two groups, such that this position is conserved within each.

In the opposite case, both columns in Figure 3 appear equally unimportant to ET, because somewhere close to the leaf level the type I gets aligned with A (so the sum in the expression for r_i, equation (1) does not stop until the leaf level is reached). The entropy, on the other hand, will be sensitive to the fact that the left column is almost conserved, while the one on the right is convincingly not conserved.

To get the advantage of both methods, it is straightforward to combine them into an expression for real-valued score:

$$\rho_i = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \left(-\sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \right) \qquad (3)$$

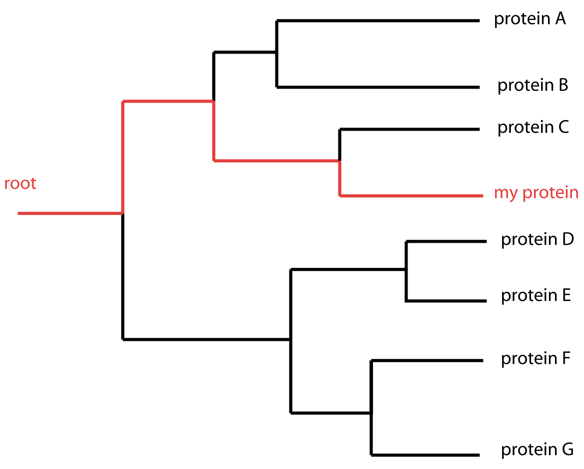

Figure 5. The path from the root to the query protein.

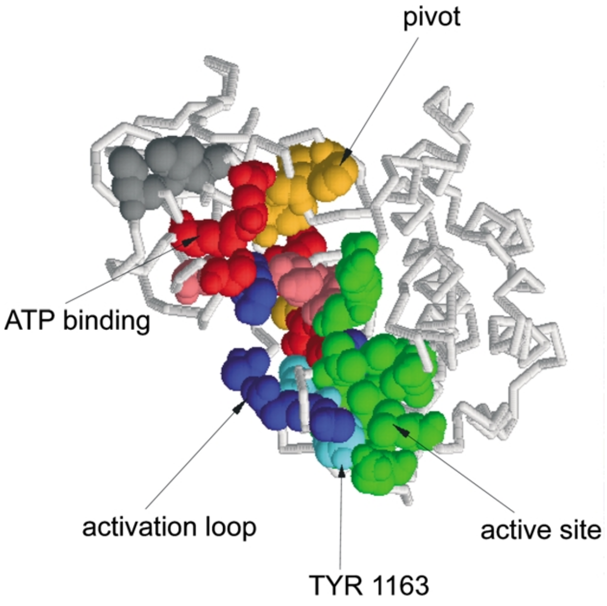

Figure 6. A map of key positions in the insulin receptor kinase. The structure: PDB entry 1irk.

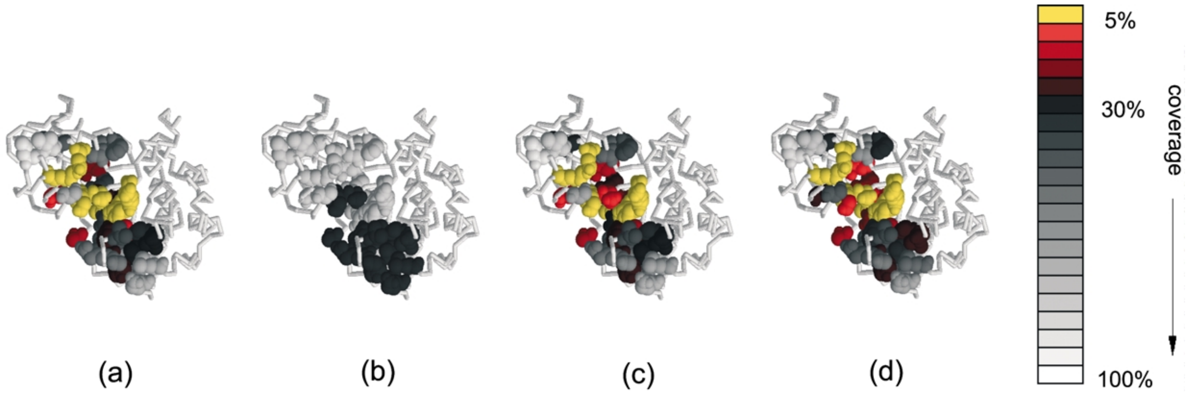

Figure 7. Success of the four methods in locating the key residues on insulin receptor kinase domain (1irk) using the raw sequence selection. (a) Column entropy; (b) integer-valued evolutionary trace; (c) zoom; and (d) real-valued evolutionary trace. The term coverage refers to the percentage of the residues selected together with the residues of interest (shown in spacefill representation). The residues colored yellow or bright red are detected as important with high specificity.

where f_{ia}^{g} stands for the frequency of the appearance of the amino acid of type a within the group g. In what follows, we refer to i as real-valued (evolutionary) trace score (and, for distinction, the method defined by equation (1) is called integer-valued ET).

This expression can be viewed as an extension of the ET method: a group g contributes 0 to the score if the residue at position i is conserved within the group, but it can contribute any real value between 0 and $\ln 20$ (maximum entropy for the system with 20 amino acid types) if the position is not conserved. Alternatively, the expression represents a series of tree-based corrections to the information entropy score s_i (which is recovered if we keep only the $n = 1$ term). $1/n$ factors out the dependence on the number of groups.

In writing down the expression for real-valued score i (equation (3)) we implicitly assumed that all groups of sequences at all tree subdivisions are equally important. However, this is an ideal case, which might not be true in practice: the groups containing evolutionarily distant sequences may carry information of less relevance to the particular protein we are aiming to analyze. This leads to generalization of equation (3):

$$\Re_i = 1 + \sum_{n=1}^{N-1} w_{\text{node}}(n) \sum_{g=1}^{n} w_{\text{group}}(g) \times \left(-\sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g} \right) \quad (4)$$

where $w_{\text{node}}(n)$ and $w_{\text{group}}(g)$ are weights assigned to a node and a group, respectively. Taking $w_{\text{node}}(n) = 1/n$ and $w_{\text{group}}(g) = 1$, we recover the expression for the real-valued ET i. The entropy corresponds to $w_{\text{node}}(n) = 1$, if $n = 1$ and 0 otherwise, and $w_{\text{group}}(g) = 1$.

This quantity is a mathematical translation of the observation that some positions remain variable no matter how fine the resolution at which we observe them, while some positions, overall variable, are seen as conserved as soon as we limit our consideration to the part of alignment corresponding to a single (sub)family. For illustration, we turn to Figure 4. It is a graphical representation of the alignment used in the analysis of P21Ras protein (1ctq in subsection the method comparison). The lines represent sequences, and the columns represent the alignment columns. Different amino acid types are assigned a different color. By increasing the resolution in the column on the left, we soon find ourselves looking at the stretches of the column with the same amino acid type. This column corresponds to the position in the protein sequence that is evolutionarily privileged and conserved within several large groups of sequences. On the contrary, in the column on the right, even under the twofold increase in the resolution, we still see a lot of variability; this position is clearly not under strong evolutionary pressure. Even though the number of types of amino acids appearing in both columns might be comparable, it is the fine-grained variability in the column on the right that makes the difference. Equation (4) attempts to

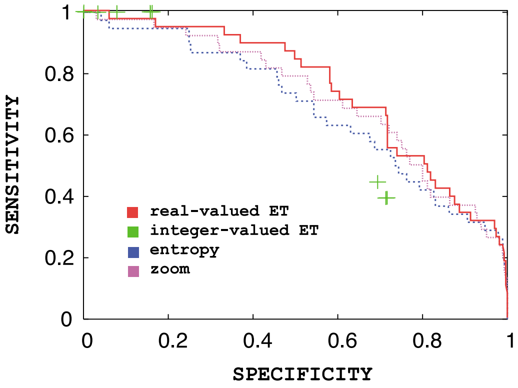

Figure 8. Sensitivity *versus* specificity of prediction for the four methods using the raw sequence selection.

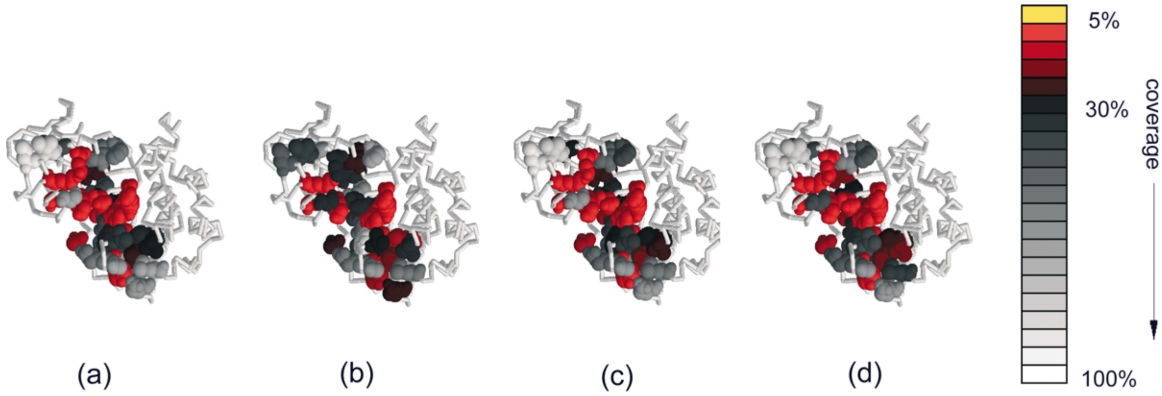

Figure 9. The same as Figure 7 using the pruned selection of sequences.

capture that notion by assigning that column a higher number (the higher values signaling less conservation and, by implication, less importance for the protein).

As a further illustration of the flexibility of this formulation, we introduce yet another scoring method that we call zoom, for reasons to be explained shortly. We use the re-weighting freedom to assign bigger weight to the part of the tree that contains the protein of interest. If the purpose of the analysis is to locate key residues in one particular sequence, we might focus on the part of the tree containing the path from the root to the leaf representing our protein (Figure 5):

$$w_{node}(n) = \begin{cases} 1 & \text{if } n \text{ on the path to the query protein} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Thus, to the nodes 1, 2 and 5 in Figure 5 we might assign the weight of 1, and the weight of 0 to all remaining nodes in the tree. The sum in over n in equation (4) would be limited, in this example, to the terms corresponding to $n = 1, 2, 5$. Each node remaining in the sum implies division of the sequence set into $n = 1, 2, 5$ groups. To fulfil our program of focusing on the part of the tree tracing the evolution of one particular protein, we choose to assign weight to each group decreasing with its distance from the group containing our protein. This ensures that information from the lineage leading directly to our protein contributes more than the information from the neighboring tree branches.

One possibility is the weight with Gaussian falloff:

$$w_{\text{group}}(g) = e^{-d_g^2/d_0^2} \tag{6}$$

with d_g the distance between the group g and the group containing the target protein, and d_0 the parameter that determines the sharpness of the falloff. In its most extreme version, the group weight is 1 only if the group contains the target protein:

$$w_{node}(n) = \begin{cases} 1 & \text{if } g \text{ contains to the query protein} \\ 0 & \text{otherwiswe} \end{cases} \tag{7}$$

This choice of weighting leads to the sum in equation (4) over only the terms that refer to the ever narrower group of sequences containing the

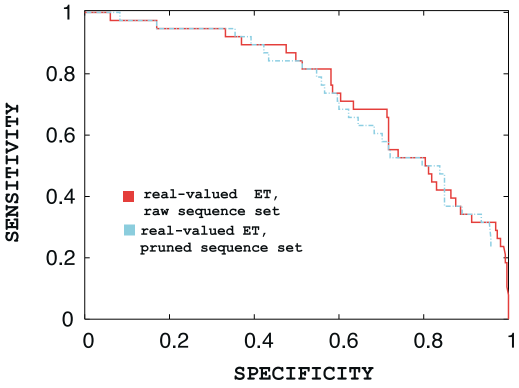

Figure 10. Comparison of the performance of the real-valued evolutionary trace on the two (raw and pruned) sequence selections: the pruned selection lacks the information needed to achieve the maximum specificity.

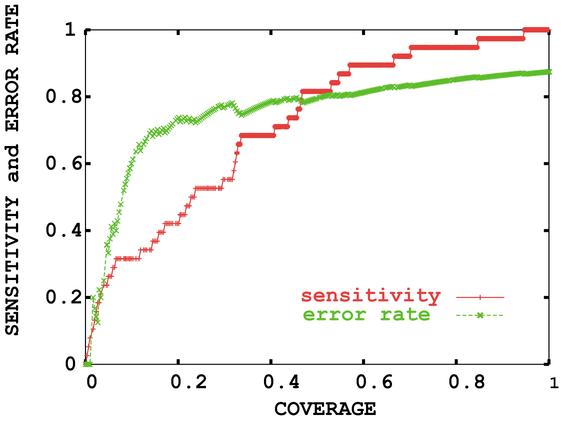

Figure 11. Specificity and error rate as a function of coverage for the real-valued evolutionary trace.

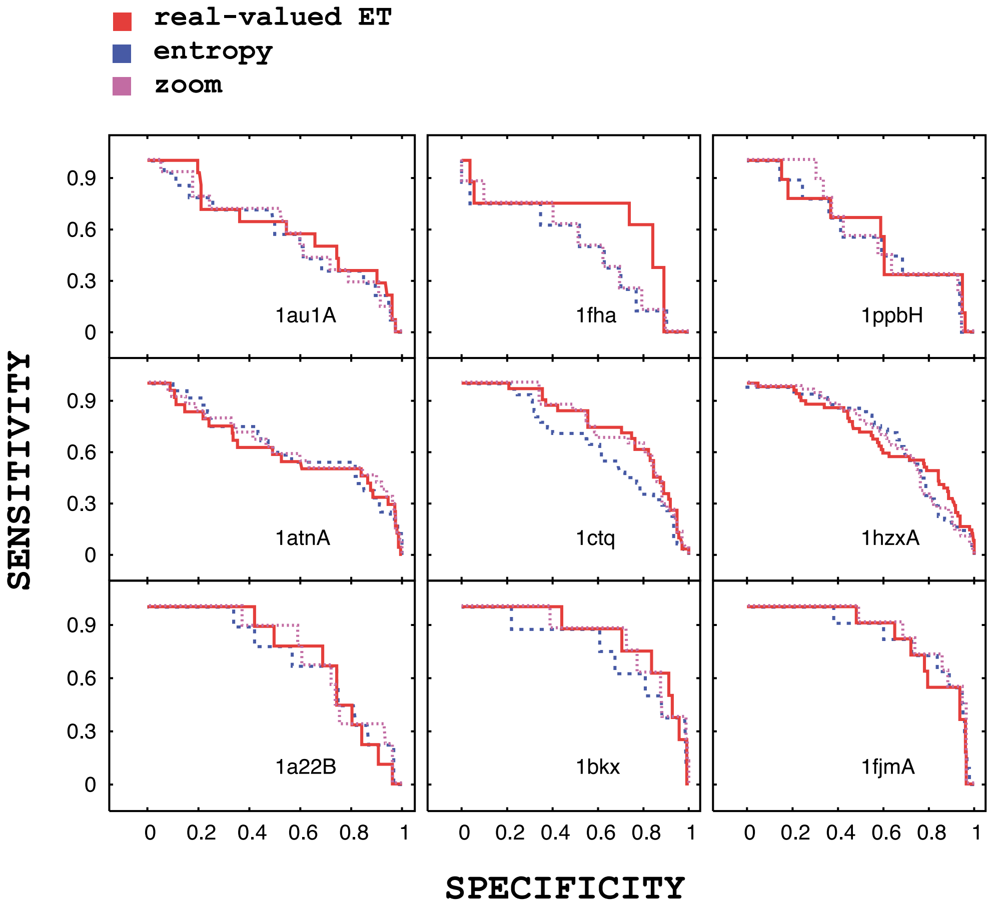

Figure 12. Sensitivity as a function of specificity for real-valued trace (red, continuous line), zoom (pink, dotted line), and plain entropy (blue, broken line) methods for scoring relative residue importance.

target protein. That is why we choose the shorthand name zoom. (If we replace entropy with absolute conservation requirement, as in equation (1), and keep this particular choice of weight, we effectively recover a method used by Aloy *et al*. in their 2001 study.[23]) Here, we present the results obtained by using the weight given in equation (6); that is, we do take the neighboring branches into account, but with a small relative weight. The motivation for this choice of $w_{\text{group}}(g)$ lies in the observation that the rates of mutation in protein residues may vary from one position to another and, for the same position, from one (sub)family to the next (see, for example, Patthy[24]). In the most extreme case, a residue may lose its function completely in one family, and there its evolutionary constraints are lifted. This position is then free to vary within one family, but not necessarily in another. If we rely too heavily on the neighboring family information, we might lose track of the residue's importance in our target protein

Results and Discussion

Case study: insulin receptor kinase

To illustrate the ideas laid out in the previous

Table 1. The number of cases each method outperforms the others, according to the four measures discussed in the text

	Raw				Pruned			
Input set residue scoring method	rvET	Zoom	Entropy	ivET	rvET	Zoom	Entropy	ivET
Max. *A*	3	3	1	0	0	0	0	1
Min. *d*	3	0	2	0	0	1	0	2
Max. *z*	4	1	1	0	1	0	0	2
Max. SHMS	2	3	0	0	0	1	1	2

A is maximum area under the specificity–sensitivity curve, *d* is the distance of closest approach to the (1,1) point on the specificity–sensitivity graph, *z* is the *z*-score of the key set overlap with respect to hypergeometric distribution, and SHMS is the specificity at half maximum sensitivity.

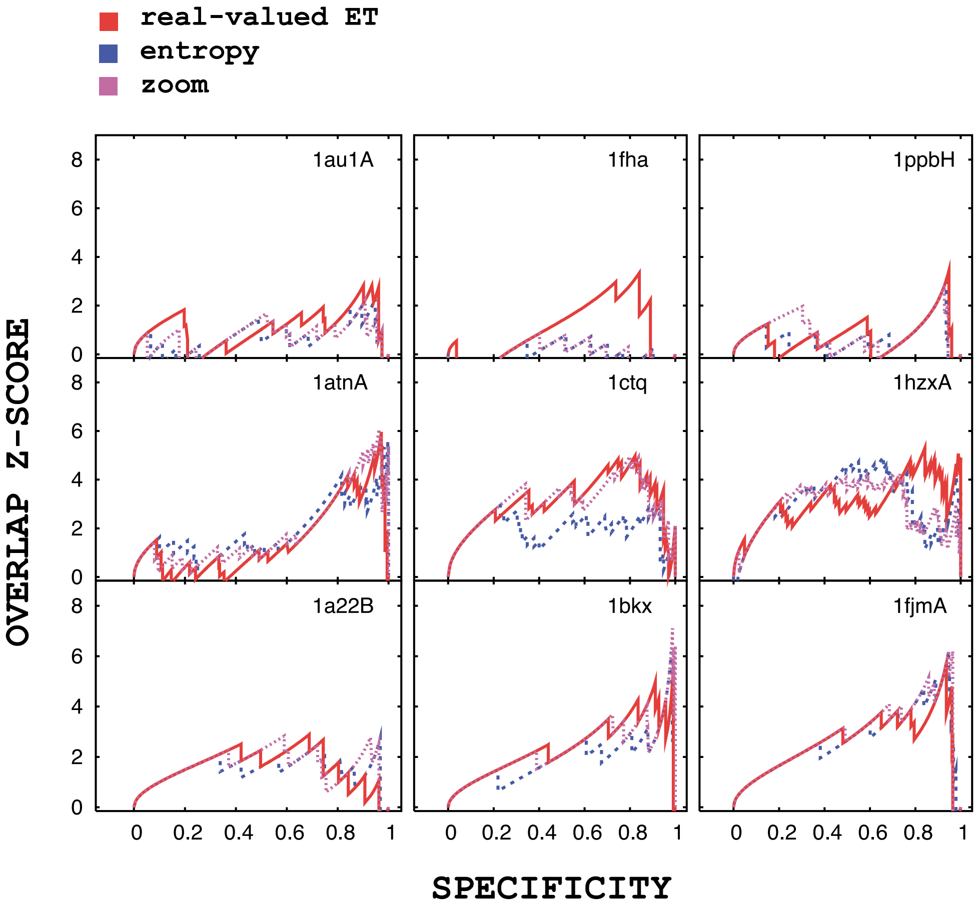

Figure 13. Hypergeometric *z*-score for the epitope overlap as a function of specificity. Overlap refers to the overlap of the residue selection with the set of the residues experimentally determined to be of critical importance for the protein. The coloring scheme is the same as in Figure 12.

section, we look more closely at one particular case: insulin receptor kinase domain. This protein is well studied, and in our analysis we rely on the work by Hubbard.[25]

Insulin receptors are transmembrane proteins that bind the insulin hormone. The binding leads to autophosphorylation of tyrosine residues in the activation loop of the protein. This results in enhancement of catalytic activity and creation of binding sites for downstream signaling proteins. We focus on the kinase domain of this protein. Its structure and residue enumeration can be found in the Protein Data Bank[26] under the code 1irk. The four key parts of the 1irk machinery are (i) the residues involved in ATP binding (red in Figure 6, the residues flanking ATP are colored pink; the ATP molecule itself is not included), (ii) active site (peptide-binding site; green), (iii) rotational pivot points and other residues involved in lobe closure (orange), which are important in conformational change between inactive and phosphorylated state, and (iv) activation loop (blue), which occupies the ATP-binding site in the inactive form with the three key tyrosine residues involved in autophosphorylation highlighted (cyan). The residues singled out by Hubbard[25] for their importance are shown in the spacefill representation (the complete list of these residues can be found in Materials and Methods). The rest of the protein is represented by the backbone.

We will suppose no prior knowledge about the protein function and compare the capability of the four methods to locate the key residues without incurring too many "false positives."

The starting point of the investigation is retrieval of sequences from the database: we accept all sequences from the NCBI Entrez non-redundant database with sufficient similarity to the 1irk sequence (in practical terms: with BLAST[27] *E*-value smaller than 0.05; this is the "raw" database return). This set of 304 sequences consists of kinases of very diverse provenance: insulin and insulin-like receptors, assorted growth factors, viral sequences, and gene products are all represented. Some of the sequences deposited in the database are mutants, and some are represented only by a fragment. The mutations (introduced in the laboratory or appearing in a viral version of the protein) will skew the distribution of the amino acid types appearing in certain position in the wild-type protein. The fragments, on the other

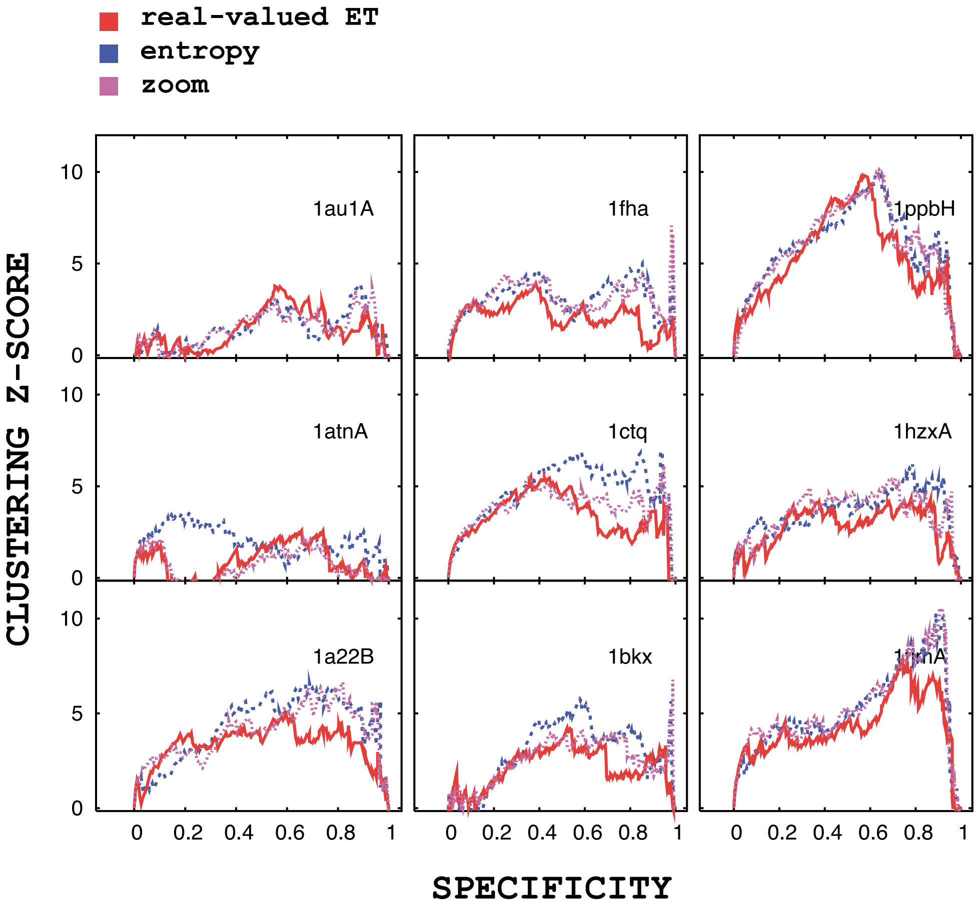

Figure 14. Clustering weight z-score as a function of specificity. The coloring scheme is the same as in Figure 12.

hand, will introduce false gaps, typically at both ends of the sequence. In this work, the gaps are treated as the 21st amino acid, so their impact is comparable to that of the mutants.

The integer-valued ET cannot handle such noisy input, and the alignment has to be pruned of the undesirable sequences. In doing this, we follow the automated protocol, described in Materials and Methods.

Once the alignment is available, we use one of the methods to create the sorted list of residues, according to their relative importance, the most important residues first.

In Figures 7 and 9 we reproduce the residues shown in Figure 6, but the coloring scheme now reflects the placement of each residue on the importance list. The placement is expressed in terms of coverage: the percentage of protein residues found equally high or higher on the list. In particular, the residues shown in yellow appear among the top 5% of the residues. The residues shown in black and various shades of gray appear at coverage of at least 30%, which is not a very informative result. The residues appearing between 5% and 30% of coverage are shown in red, the lighter colors corresponding to smaller coverage, and therefore to a more specific result.

The panels in Figure 7 show the results on the raw sequence selection, using (a) column entropy, (b) integer-valued ET, (c) zoom, and (d) real-valued ET. The column entropy successfully places most of the residues involving the proper functioning of the protein high on the list. The integer-valued trace is not expected to work in this case: the raw alignment abounds in cases like that depicted schematically in Figure 3 (the role of I in that Figure is frequently played by a gap), which breaks down the model within which the integer-valued ET works. The zoom and the real-valued method improve on the prediction using entropy, by moving several residues from the activation loop and the ATP-binding site to smaller percentage bin.

The improvement can be made more obvious in combined sensitivity *versus* specificity diagram for all four methods (Figure 8). Sensitivity is the percentage of key residues (as given by Hubbard[25]) that is found among the top n residues on the list, and specificity is the percentage of residues that are not singled out as important, and that can be found below the nth position. Each point on the graph is the specificity–sensitivity pair for one particular choice of n (see also discussion in the subsection the method comparison). For almost any choice of n (the exception being of marginal size), the zoom and the real-valued ET top the entropy in both sensitivity and specificity.

What is the practical implication of this result? The numerical values in Figure 8 can be restated

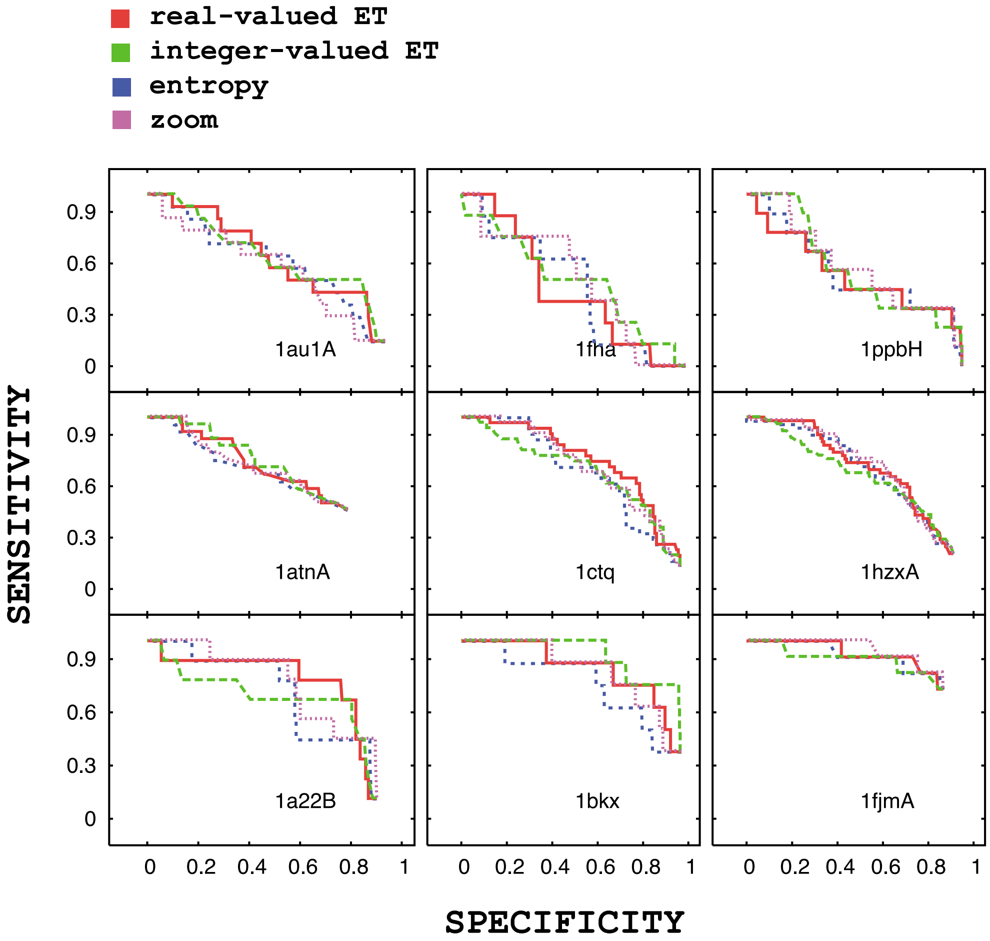

Figure 15. Comparison of performance: real-valued evolutionary trace on the raw sequence set, and integer-valued evolutionary trace on the pruned sequence set. For the protein set description, see Materials and Methods. Sensitivity–specificity curve.

as follows: suppose we want to do a mutation study on 1irk. By picking from the list created by the real-valued ET the top 2% of residues, which can at the same time be mapped onto the 1irk structure, we have selected six residues: E1047, H1130, N1137, G1082, R1131, and D1132. Five of them belong to the set of 38 residues pointed out by Hubbard,[25] i.e. we have detected 13% (5/38) of the important residues, with 17% (1/6) error rate. Even this, supposedly false positive, residue, H1130, might merit consideration: it is a highly conserved histidine residue, part of a catalytic loop, but not discussed explicitly in the reference work. At 3% coverage we have selected nine residues, seven important and two not, which equals specificity of 19% and error rate ("false alarm rate") of 22%. These two percentages climb to 26% and 33%, respectively, when 5% of residues are selected. But after this point the error rate starts growing much faster than the specificity (Figure 11); at 10% coverage we have discovered 32% of the key residues, but 60% of our guesses are false positives. The best trade-off between the two quantities is achieved for coverages smaller than or close to 5%. The residues that we pick at this point are residues belonging to the active site (R1131, D1132, R1136 and N1137), the ATP-binding site residues (K1030, D1150 and F1151), and the other protein kinase conserved residues (E1047 and G1082). In all, 40% of our selection is outside the set selected by Hubbard (H1130, A1028, L1123, A1134, M1120 and G1064).[25] Doing mutation studies on these residues is possibly a misplaced effort, but some of them merit further consideration: H1130, mentioned earlier; M1120, which is methionine everywhere except in a group of sevenless homologues from Drosophila and Anopheles, and certain mammalian and avian oncogenes where it is quite convincingly aligned with a cysteine residue, and which is within 4 Å from the catalytic site; and G1064, conserved in almost all sequences, possibly of structural importance. Selection of the top 23% of residues results in detection of 50% of key residues, with a false positive error of 72%. For comparison, at the same coverage (23%), the column entropy finds 42% of the highlighted residues, with 77% of selected residues being false positives.

On the pruned selection of sequences, all methods perform comparably well (Figure 9). The

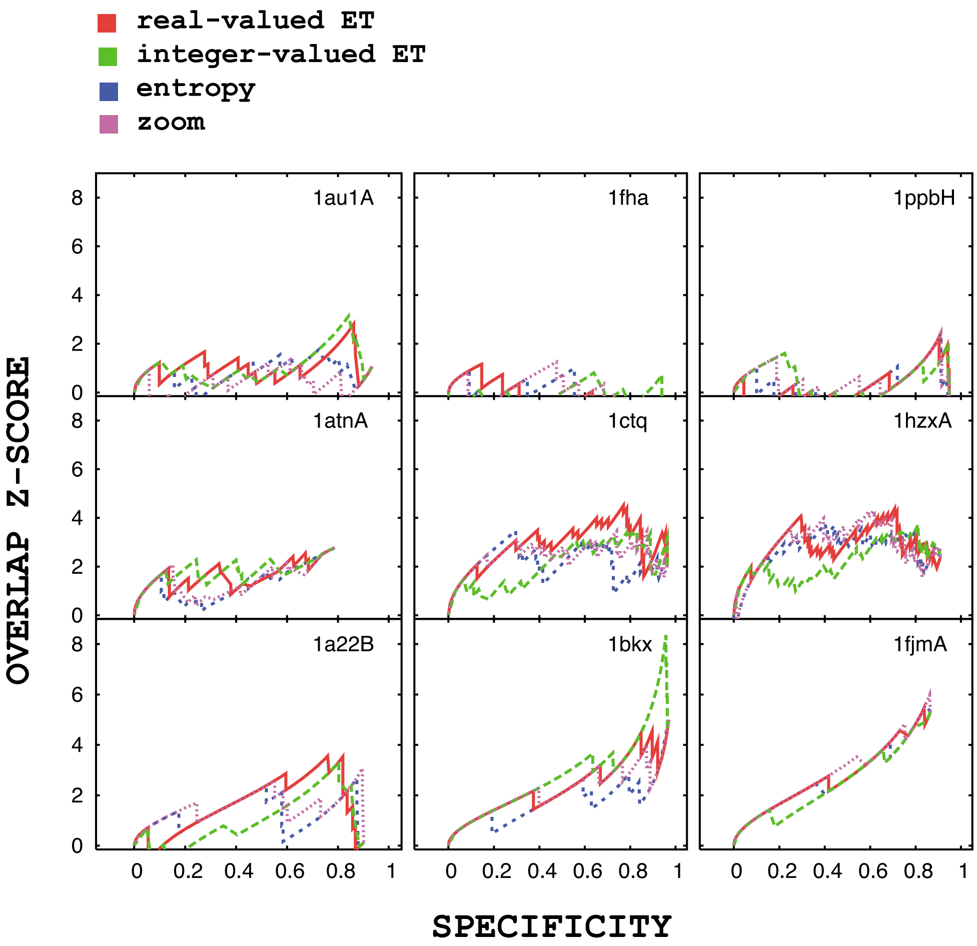

Figure 16. Hypergeometric z-score, geometric epitope: the same as Figure 18: hypergeometric z-score. "Epitope" refers to geometric epitope determined by closeness to the ligand.

success rate for the real-valued trace at 23% coverage is still 50%, with the error rate slightly decreased (70%). The other three methods perform somewhat worse in both indicators. The main difference, however, is that now the lowest coverage equals 7% (all the conserved residues being tied at the top of the list; which holds for the whole family of methods we are discussing): 41% of these residues are not in the selected set. The overall result of reducing the number of sequences is a disproportionate loss of specificity, compared with the gain in sensitivity. This is not surprising; we discard potentially useful information with each sequence. This result is shown also in Figure 10: the two sensitivity–specificity curves (for the real-valued ET) for the two sequence selection methods almost match, the main difference appearing in the lower right corner (high specificity) for which the analysis on the pruned sequence set lacks the data.

Multi-protein comparison of methods

To compare the performance of different methods, we interpret the ranked list of residues produced by each as a set of predictions of the most important residues. Thus, the x topmost residues constitute a predicted set of important residues of size x. The larger x, the greater our chance of selecting all of the important residues in the protein, but at the same time, the greater x increases our chance of incurring an excess baggage of false positives. This is expressed in a more orderly manner by using quantities termed sensitivity and specificity. The sensitivity is defined as the ratio of the number of important residues that our method finds correctly to the known total number of important residues (true identified positive/actual positive), while specificity is the number of unimportant residues predicted by the method divided by the number of residues known not to be important (true identified negative/actual negative). In the ideal case, we would be able to pick all of the important residues, and only the important residues; in this case both sensitivity and specificity would be equal to 1.

It is important to understand that we do not really know what the actual set of true positives is. As our best guess about the position of the key residues, we take the set of residues for which an importance indication can be found in the literature (as the importance indication we take a study

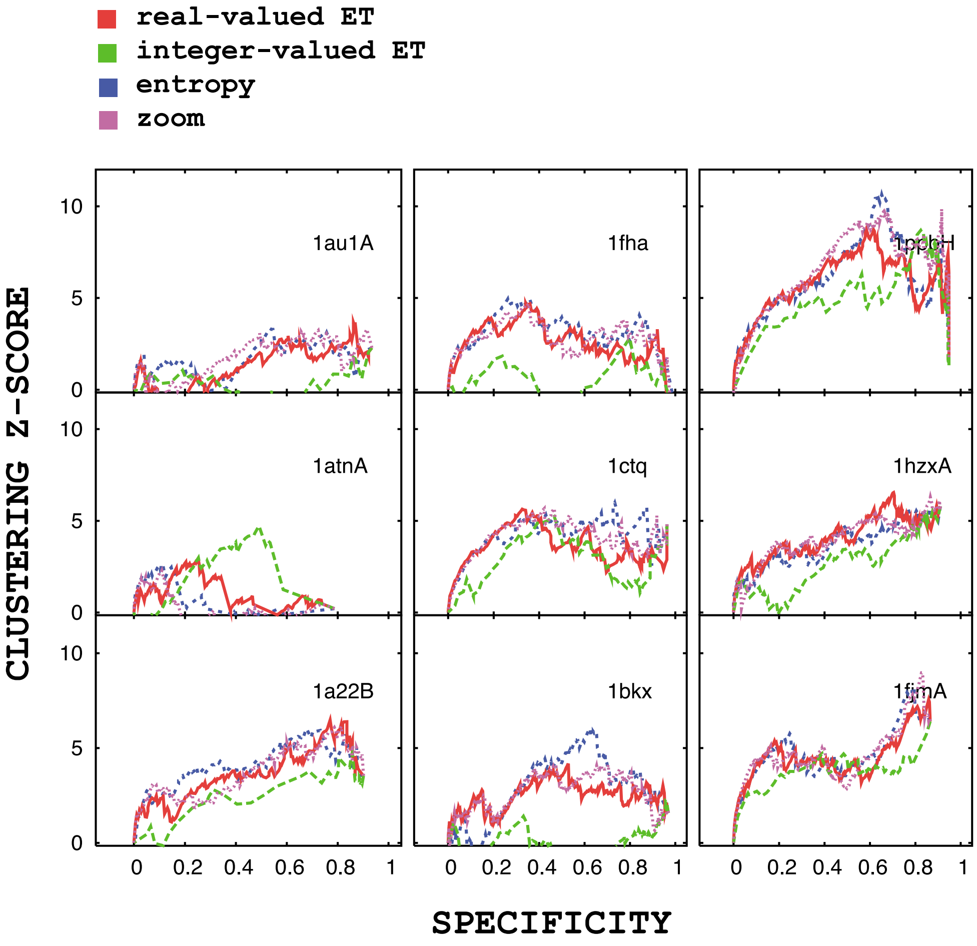

Figure 17. Similar to Figure 18: clustering z-score *versus* specificity.

showing significant structural, functional or health impact of the residue mutation; see Materials and Methods). Our choice of nine test-case proteins was dictated precisely by the possibility of finding a reasonable number of critical residues in the literature (see Materials and Methods).

With this in mind, we compare the specificity–sensitivity curves for the entropy (the blue line in Figure 12), the real-valued ET (red) and the zoom methods (pink). In each case, we start from the same multiple sequence alignment, and construct the ranked list of residues using each of the methods. For each selection of the x topmost residues, we calculate the associated sensitivity and specificity, and taking the two as point coordinates we plot the point on a graph. We start somewhere close to (1,0) point on the graph; our first selection of x residues is expected to have small sensitivity, but cannot degrade significantly the specificity (which then stays close to the value of 1). As we increase the number x, the sensitivity should grow faster than the specificity drops. Somewhere in this sequence of predictions we would like to get as close as possible to the (1,1) point in the graph, indicating that we have made the most of both sensitivity and specificity. Once the specificity becomes degraded seriously (that is, once we have included in our selection too many false positives), the predictions start losing their importance; the practical value of a method is reflected in its behavior on the right-hand side of the graph.

The sequence sets on which we base the predictions in this case were obtained by doing a simple BLAST search (see Materials and Methods) without further sequence selection ("pruning", colloquially). Integer-valued ET was not designed to cope with such a "raw" selection of sequences, and we discuss its predictions below. The two hybrid methods obviously perform at least as well as the entropy, and usually better (see Table 1).

There are cases like 1ppbH in Figure 12 when no method performs convincingly well. We must be aware that they are all subject to noise in the input: the sequence selection is a sampler of data available in the database, as much as of our ability to select them smartly. Similarly, the set of true positives is a reflection of our semi-automatic way of gathering information; a "key" residue that scores in the lowest percentile using any method is somewhat suspect. Finally, the reference mutations, although the best data available, still may give only a partial picture of the true importance of residues.

The statistical significance of the results in

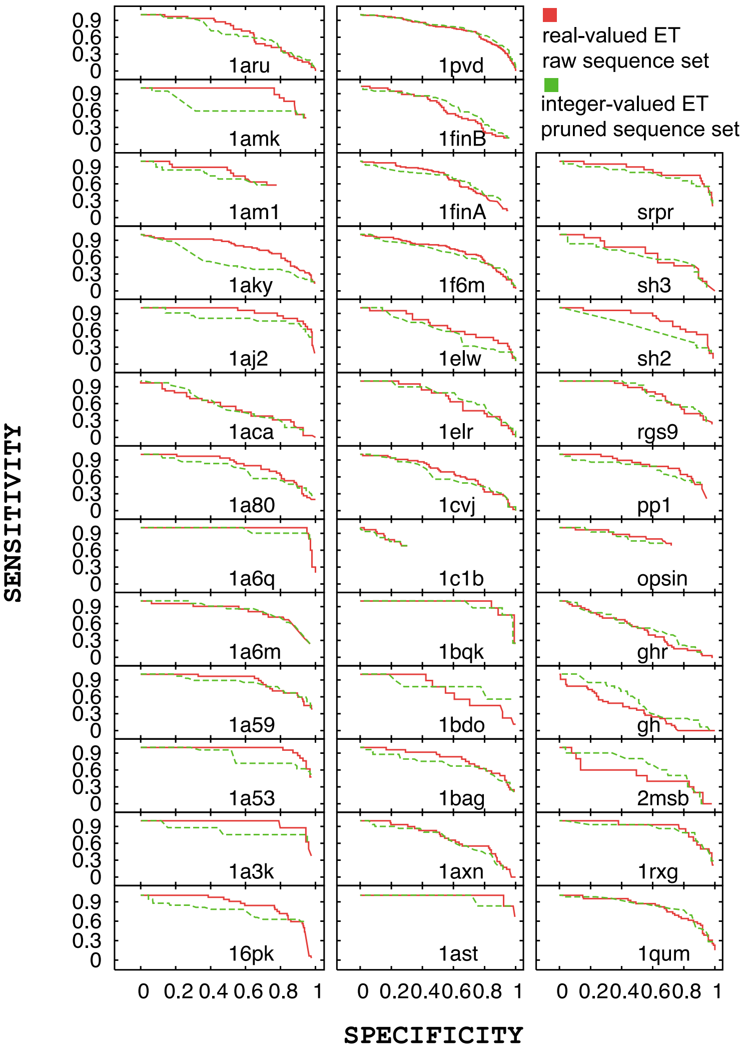

Figure 18. The analogue of Figure 12, this time with the "pruned" input set of sequences. The results for the integer-valued evolutionary trace: green broken line.

Figure 12 can be made clearer by comparing them with the prediction success one would obtain by picking the residues at random. The probability distribution for x_c correct guesses from the selection of x, given the size of the protein and the size of the key residue set, is the hypergeometric distribution. The expressions for the average and the standard deviation for that distribution are well-known quantities,[28] so we can calculate the z-score for our findings, where:

$$z\text{-score} = (x_c - \langle x_c \rangle)/x$$

The results are shown in Figure 13. The x-axis is again the specificity, to enable direct comparison with Figure 12. They reinforce our finding that hybrid methods outperform the entropy in the sense that the results, which are better than the random average by two to six standard deviations, comfortably top the entropy.

A possible measure of quality of prediction, completely independent of the epitope knowledge, is the on-structure clustering of the residue selection.[29,30] It is believed that the important protein residues tend to make non-random clusters when mapped on the three-dimensional structure of the folded protein. The quantity we use to measure this effect is termed selection clustering weight:[30]

$$w = \sum_{i<j}^{L} S(i)S(j)A(i,j)(j-i) \qquad (8)$$

where L is the length of the peptide chain, $i,j = 1,\ldots,L,S$ is the selection function, which assigns 1 to selected residues, and 0 otherwise (remember we are selecting x out of possible L residues), and A is the adjacency matrix: $A(i,j) = 1$ if the residues i and j are in contact on the folded chain, and 0 otherwise. We again compare this quantity with the average[30] in the case of random choice of residues, and show the associated z-score (Figure 14). The x-axis is again the specificity. Perhaps unexpectedly, the residues selected by entropy cluster at least as well as those obtained by using the alternative scoring schemes. Nevertheless, in all of the cases, the high z-scores assure us of the non-randomness of our prediction.

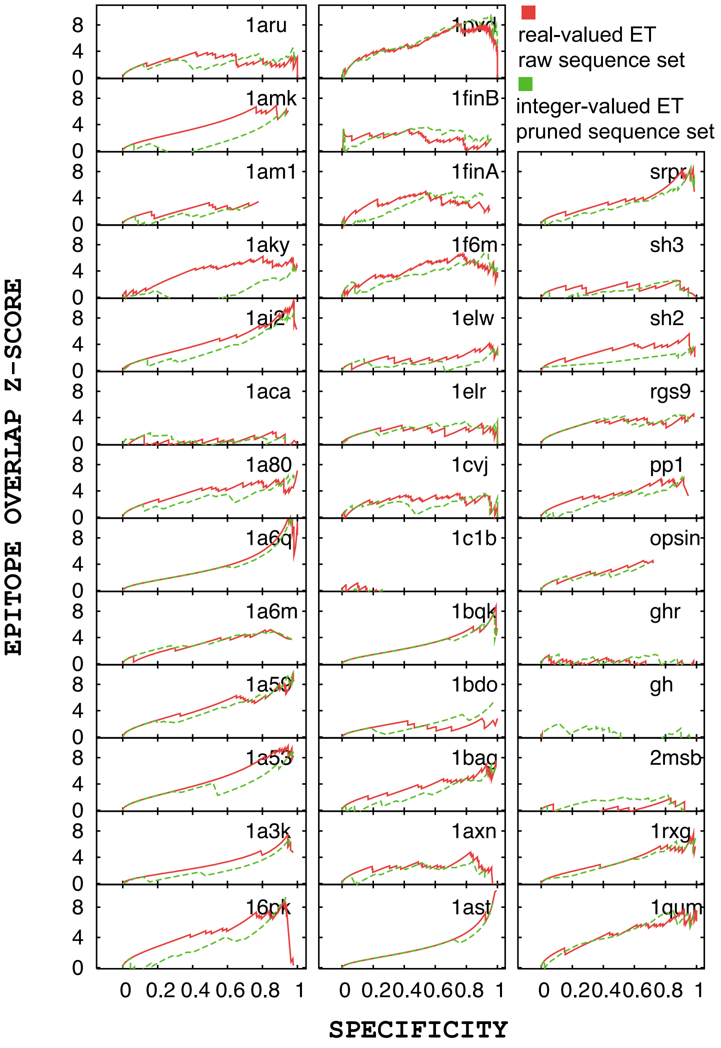

Figure 19. The analogue of Figure 13, on the pruned input set of sequences. The results for the integer-valued evolutionary trace: green broken line.

To prune or not to prune?

To perform our analysis (using any of the methods) we have to make an initial selection of sequences. Needless to say, we cannot start with a random selection, but rather with one that samples a protein family tree fairly. Usually, a database search with some reasonable homology cutoff will satisfy this first requirement. The input sequence set may be further reduced using some rejection criterion.

In particular, the integer-valued ET was not designed to work with raw return of a database search. Inconsistencies in the sequence selection, or isolated cases of atypical mutations, sequence fragments, and isoforms can degrade the quality of the (integer-valued) ET prediction significantly. To compare the integer-valued ET with the rest of the methods we must remove from the alignment various impostor sequences, the process that we refer to as pruning.

The power of the real-valued ET lies in the fact that we are at all able to discuss the possibility of not pruning the input set. To further illustrate the point, we resort to a protein set used previously by this group,[31] the set consisting of more proteins then our central test set, but with a somewhat less satisfactory choice of important residues, determined by their spatial proximity to the ligand (that information was available in each protein's PDB[26] file). The raw sequence set was obtained by doing a database search with a certain similarity cutoff, while the pruned set was obtained by visual inspection of the multiple sequence alignment and removal of sequences that struck the operator as being out of place, resulting in the selection used by Yao *et al*.[31] As is notable in Figures 18–20, while this method may lead to an improved prediction (cf. 1bdo), in many cases it is not so.

Turning back to our original data set, the analogues of Figures 12–14 for the pruned input set of sequences are given in Figures 15–17. The prediction can in this case be improved, as demonstrated by the cases of 1bkx and 1fjmA proteins (Figure 15), but the loss of information at the input may result in the loss of specificity, as can be seen by comparing the results for 1atnA, 1hzxA, 1a22B, 1bkX, and 1fjmA in Figures 12 and 15. There is obviously a trade-off involved; when

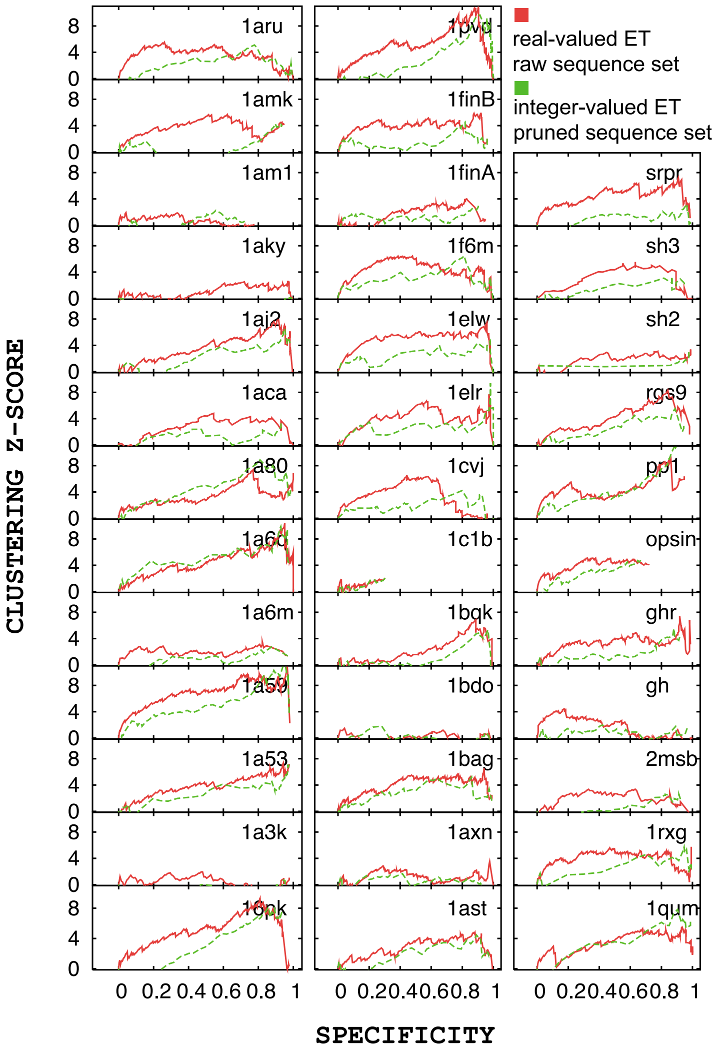

Figure 20. The analogue of Figure 14, on the pruned input set of sequences. The results for the integer-valued evolutionary trace: green broken line.

to stop the pruning may have to be decided on an individual basis.

What is the overall gain in using the hybrid methods?

To make that estimate we need some kind of a "measure of a measure." We use four measures: the area under the specificity–sensitivity curve (which we hope to be as big and as close to 1 as possible; *A* in Figure 21); the specificity at the point where half of the maximum sensitivity is achieved (SHMS); the distance of closest approach to the (1,1) point in the specificity–sensitivity diagram (which we hope to be as close to 0 as possible; *d* in Figure 21) and, finally, the maximum achieved *z*-score under comparison with the hypergeometric distribution. While the area under the specificity–sensitivity curve measures the overall performance of a method, SHMS indicates its average performance, and the last two quantities refer to its single best performance point.

Currently, we are not able to construct the test set of the size allowing for the statistical analysis of our findings (see Materials and Methods), but the breakdown for our small test set is given in Table 1, which lists the number of times each method (considered together with the input

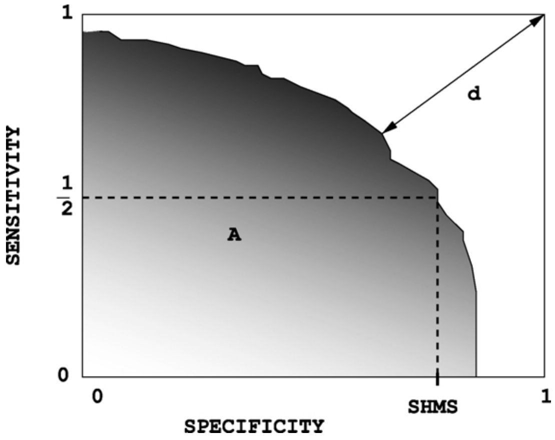

Figure 21. Quality of prediction measures: a diagram of measures of quality of prediction. *A* is the area under the curve (shaded), *d* is the distance of closest approach to the (1,1) point, and SHMS is specificity at half maximum sensitivity.

Table 2. The proteins used as the test set

PDB code	Name	Function	Organism	Fold	No. residues	No. mutations found
1a22B	Growth hormone receptor	Signaling	Human	Immunoglobulin-like β sandwich (all β)	192	9
1atnA	Actin	Motile	Rabbit	RibonucleaseH-like motif (α/β)	372	24
1au1A	Interferon β	Protective	Human	4-Helical cytokine (all α)	166	14
1bkx	CAMP-dependent kinase	Enzyme	Mouse	Kinase-like (α/β)	337	8
1ctq	P21 Ras	Signaling	Human	P-loop-containing nucleotide triphosphate hydrolase	166	31
1fha	Ferritin	Storage	Human		172	8
1fjmA	Serine/threonine phosphatase-1	Enzyme	Rabbit	Metallo-dependent phosphatase	294	11
1hzxA	Rhodopsin	Signaling	Cow	GPCR-A family	340	49
1ppbH	Thrombin	Protective	Human	Trypsin-like serine protease (all β)	259	9

sequence set) achieves the highest score according to our four measures. The results indicate that there is definitive advantage in using hybrid methods, while keeping the input sequence set reasonably large.

Conclusion

We discussed several methods of ranking protein residues by their importance for the protein as a whole. They are of necessity somewhat schematic. In particular, they do not systematically handle the varying evolution rates across the evolutionary tree (a problem that we attempted to address partially in defining the zoom method; equations (4)–(6)), and they ignore some important possibilities such as correlated mutations. We suggested a general way of integrating tree and variability information, laid out in equation (4). Our choice of weights in equation (4) is by no means unique, and can conceivably be improved by incorporating more detailed knowledge about the tree. Also, the set of amino acid types can be changed to reflect only a number of underlying chemical properties.[2] To make more definitive mutual comparison of the methods, more experimental backup can be wished for in determining bigger and possibly ranked sets of key protein residues.

We have noted that the column entropy is a very robust method in the face of small irregularities and inconsistencies in the input alignment, but is insensitive to evolutionary relations within the protein family represented by the alignment. The integer-valued ET captures these details perfectly, but it is this same input sensitivity that forces it to dispense with a good deal of potentially useful information. Our most successful take on uniting the two approaches is the real-valued ET, demonstrably on top of its class: capable of handling raw sequence input, it matches or tops the sensitivity–specificity performance of other alignment/evolution methods, and its robustness makes it potentially applicable on the proteomic scale.

A server with an implementation of real-valued ET method will be available†.

Materials and Methods

Key residues in the study case

The following residues were taken to be the key residues for the protein function (the numbers refer to the enumeration in the 1irk PDB entry): 1006, 1010, 1030, 1038, 1042, 1045, 1047, 1054,1061, 1077, 1079, 1082, 1083, 1085, 1089, 1092, 1131, 1136, 1137, 1139, 1150, 1151, 1152, 1155, 1158, 1162, 1163, 1132, 1164, 1166, 1171, 1172, 1173, 1176, 1181, 1215, 1216, 1219. For a full discussion, see Hubbard.[25]

Proteins in the test set

In constructing the test set we tried to put together as diverse a set of proteins as possible (Table 2). We were limited most strongly by the requirement that enough point mutations can be found in the literature that are annotated either as critical (for either structure or function of the protein) or as disease-related (see subsection Key residues below). In this work, we settled for at least eight residues determined as critical in at least one mutational study. Furthermore, we were looking for proteins for which an X-ray crystallographic structure is available, and for which more than 20 homologous but not nearly identical sequences could be found. The proteins satisfying all of the requirements are all mammalian, and mostly human, reflecting the current research bias toward proteins of medical importance (for which it is the easiest to find mutational studies).

Key residues in the test set

To find a reasonable independent estimate of residue importance for proteins in our test set, we relied on Protein Mutant Database‡.[21,32] Specifically, we used entries containing reference to single-point mutations in proteins at least 70% identical with our test cases, annotated as either causing a complete loss of protein

† http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/
‡ http://pmd.ddbj.nig.ac.jp/

activity ([0] in the notation used in the database) or a disease.

Auxiliary test set with a geometric epitope

In Results and Discussion, in Figures 18–20 we used a protein set described as a protein–ligand dataset by Yao *et al.*,[31] and the geometric epitope definition as described by them.

Sequence selection

To obtain a set of sequences homologous to the protein of interest, we did the BLAST[27] search against the NCBI Entrez non-redundant protein sequence database, and used all the sequences with the E-score <0.05. Such a set of sequences is referred to as raw in the main text.

The sequence sets referred to as pruned were obtained by removing from their raw counterpart the sequences that have less than 40% similarity and more than 98% identity with the query sequence, the sequences that are outliers in the sequence similarity tree, and those that have an out-standing number of gaps. Finally, we remove sequences with residues that persistently belong to a minority type within its group (see below). (i) Determining outliers: from the sequence identity tree (current implementation of ET uses unweighted pair group method to build it[22]) we remove the subtrees whose sibling node (the tree is binary) has 30 times more leaves in its subtree. (ii) Sequences with an outstanding number of gaps: for each rank (starting from the rank 1) consider the induced subset sizes. If the subset contains more than 15 sequences, remove the sequences with the number of gaps more than ten standard deviations away from the average number of gaps in the subset. If it contains 15 or less sequences, construct a mock sequence that has a gap in the places where the majority of sequences do, and any amino acid (X) otherwise. Then remove all the members of the subset whose gap pattern differs by 3% of the alignment length from the pattern in the mock sequence. (iii) Sequences with persistent minority residues: for all residues in a sequence calculate negative log frequency of its appearance in the corresponding column. From the alignment remove the sequences for which the sum of theses values deviates by two standard deviations from the average. Repeat iteratively for all the groups in the tree with more than 15 members.

The goal of the described method of pruning a raw set of sequences is to be reproducible, and yet mimic the effects of manual sequence selection.

Alignment

The multiple alignments were obtained using ClustalW1.7[33] in the quicktree mode, with all other parameters kept at their default values.

Alignment-based tree

We used UPGMA[22] tree, with 1 minus percentage sequence similarity as the sequence distance (see below).

Parameters of the ranking methods

The residue ranking methods we considered were parameter-free, except for the zoom method. Equation (6), describing the group weight that the zoom assigns, calls for one parameter, d_0. Any d in that equation is the distance between sequences. In our implementation it is equal to 1 minus percentage sequence similarity, a number that equals 0 when two sequences are 100% similar, and 0 if no two aligned positions are similar. Two amino acid residues are considered similar if their entry in the BLOSUM62[34] matrix is equal to or greater than the average off-diagonal entry in that matrix. The value used for d_0 was 0.05.

Acknowledgements

The authors give many special thanks to the team of creators and curators of the Protein Mutant Database, without which it would be practically impossible to construct our test set. Thanks go also to the members of the Lichtarge laboratory for critically reading the manuscript. This work was supported by NIH GM066099 and NSF DBI-0318415 grants. I.M. gratefully acknowledges partial support from the W.M. Keck Center for Computational and Structural Biology, and I.R. from the March of Dimes Foundation.

References

1. Zvelebil, M., Barton, G., Taylor, W. & Sternberg, M. (1997). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.
2. Valdar, W. (2002). Scoring residue conservation. *Proteins: Struct. Funct. Genet.* **48**, 227–241.
3. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178.
4. Shannon, C. & Weaver, W. (1949). *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.
5. Schneider, T., Stormo, G. & Gold, L. (1986). Information content of binding sites of nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.
6. Berg, O. & von Hippel, P. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750.
7. Shenkin, P., Erman, B. & Mastrandrea, L. (1991). Information-thoretical entropy as a measure of sequence variability. *Proteins: Struct. Funct. Genet.* **11**, 297–313.
8. Sunyaev, S., Eisenhaber, F., Rodhenkov, I., Eisenhaber, B., Tumanyan, V. & Kuznetsov, E. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.* **12**, 387–394.
9. Pei, J. & Grishin, N. (2001). Al2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
10. Atchley, R., Terhalle, W. & Dress, A. (1999). Positional dependence, cliques, and predictive motifs in the bhlh protein domain. *J. Mol. Evol.* **48**, 501–516.
11. del Sol Mesa, A., Pazos, F. & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.

12. Mirny, L. & Shakhnovich, E. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196.
13. Hannenhalli, S. & Russell, R. (2000). Analysis and prediction of func- tional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–67.
14. Lichtarge, O., Bourne, H. & Cohen, F. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
15. Livingstone, C. & Barton, G. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–775.
16. Ptitsyn, O. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278**, 655–666.
17. Lichtarge, O., Bourne, H. & Cohen, F. (1996). Evolutionarily conserved gαβγ binding surfaces support a model of the g protein–receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 1483–1488.
18. Sowa, M., He, W., Wensel, T. & Lichtarge, O. (2000). A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl Acad. Sci. USA*, **97**, 1483–1488.
19. Sowa, M., He, W., Slep, K., Kercher, K., Lichtarge, O. & Wensel, T. (2001). Prediction and confirmation of a site critical for effector regulation of rgs domain activity. *Nature Struct. Biol.* **8**, 234–237.
20. Onrust, R., Herzmark, P., Garcia, P., Lichtarge, O., Kingsley, C. & Bourne, H. (1997). Receptor and betagamma binding sites in the α subunit of the retinal G protein transducin. *Science*, **275**, 381–384.
21. Kawabata, T., Ota, M. & Nishikawa, K. (1999). The protein mutant database. *Nucl. Acids Res.* **27**, 355–357.
22. Waterman, M. (2000). *Introduction to Computational Biology*, Chapman & Hall/CRC, London/Boca Raton.
23. Aloy, P., Querol, E., Aviles, F. & Sternberg, M. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
24. Patthy, L. (1999). *Protein Evolution*, Blackwell Science, Oxford.
25. Hubbard, S. (1997). Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and atp analog. *EMBO J.* **16**, 5573–5581.
26. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
27. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and Psi-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
28. Casella, G. & Berger, R. (2001). *Statistical Inference*, Duxbury, Pacific Grove, California.
29. Madabushi, S., Yao, H., Marsh, M., Kristensen, D., Philippi, A., Sowa, M. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
30. Mihalek, I., Res, I., Yao, H. & Lichtarge, O. (2003). Combining inference from evolution and geometric probability in protein structure evaluation. *J. Mol. Biol.* **331**, 263–279.
31. Yao, H., Kristensen, D., Mihalek, I., Sowa, M., Shaw, C., Kavraki, L. & Lichtarge, O. (2002). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.
32. Nishikawa, K., Ishino, S., Takenaka, H., Norioka, N., Hirai, T., Yao, T. & Seto, Y. (1994). Constructing a protein mutant database. *Protein Eng.* **7**, 773.
33. Thompson, J., Higgins, D. & Gibson, T. (1994). Clustal W: improv-ing the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
34. Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Edited by F. E. Cohen

(Received 25 July 2003; received in revised form 5 December 2003; accepted 19 December 2003)