Identification of Functionally Conserved Residues With the Use of Entropy–Variability Plots

Laerte Oliveira,[1] Paulo B. Paiva,[1] Antonio C.M. Paiva,[1] and Gerrit Vriend[2]*
[1]*Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, Brazil*
[2]*Center of Molecular and Biomolecular Informatics, Katholieke Universiteit Nijmegan, The Netherlands*

ABSTRACT **We introduce sequence entropy–variability plots as a method of analyzing families of protein sequences, and demonstrate this for three well-known sequence families: globins, ras-like proteins, and serine-proteases. The location of an aligned residue position in the entropy–variability plot correlates with structural characteristics, and with known facts about the roles of individual amino acids in the function of these proteins. The large numbers of known sequences in these families allowed us to introduce new filtering methods for variability patterns. The results are discussed in terms of a simple evolutionary model for functional proteins. Proteins 2003;52:544–552.**

Key words: entropy–variability plots; globins; ras-like proteins; serine-proteases; recalcitrant residue positions; protein structure evolution

INTRODUCTION

Recent developments in sequencing whole genomes have led to a flood of sequence information. At the same time, developments in X-ray crystallography (synchrotron radiation, new software, faster production of X-ray–quality crystals, etc.) and NMR (stronger magnetic fields, new pulse sequences, etc.) have led to a rapid increase in the number of solved structures. Today, about 45 sequences enter the EMBL database every minute (http://www3.ebi.ac.uk/services/dbstats/), and about 7 new structures are deposited every day (http://www.rcsb.org/pdb/holdings.html), of which one or two are typically sequence-unique (http://www.cmbi.kun.nl/whatif/select/) (i.e., share less than 25% sequence identity with any previously deposited structure).

These data contain an enormous amount of information that can be extracted by techniques such as multiple-sequence alignment, pattern recognition, profile searches, correlated mutation analysis, and so on. A practical example is the HSSP project,[1,2] which aims to combine sequence and structure information. So-called HSSP-files hold the largest possible reliable multiple-sequence alignments for proteins of known structure. These files implicitly contain a lot of structural and functional information, and many programs are available in these files to visualize the sequence conservation and variability information (expressed as a Shannon entropy term). This allows researchers to find important (conserved) residues quickly, such as those in the active site. Alternatively, variable positions can be found in which mutations may be introduced without disturbing the fold and function of the protein.

Residue conservation has been evaluated in multiple-sequence alignments by means of variability (number of different amino acids found), Shannon entropy, and variance-based and score-matrix indices.[3–5] The patterns of conservation in proteins have been used for quality assessment and refinement of multiple-sequence alignments[4]; they have also been described as the fingerprints left by evolution in the structure.[6] Recent studies have shown that conserved residues are often clustered in certain regions of protein structures,[4] sometimes at "universally conserved positions,"[7] so called because they can form a motif characteristic of the fold. Sometimes, these positions are also found in the corresponding sequence segments of analogs, and their location often coincides with that of supersites.[8]

Many groups have used the identification of conservation patterns in proteins as a method to search for function. Some of these methods are based on energy calculations on proteins of known structure, and look for charge and shape complementarities in protein and ligand surfaces that are thought to interact.[9–14] Other groups have predicted functional motifs from an analysis of protein interaction surfaces using principal component analysis,[15] analysis of physicochemical descriptors to score protein–protein interactions,[16] search for motifs in Blocks databases,[17] or alignment of hinge regions.[18] Evolutionary trace analyses involve searching for conservation patterns in different branches of phylogenetic trees and mapping them onto three-dimensional (3D) structures to look for clusters of functionally important residues.[19–22]

Some of these methods require knowledge of the 3D structure of the protein, but most require no more than a multiple-sequence alignment. All methods based on sequence variability analysis use a single measure of variabil-

Grant sponsor: São Paulo State Research Foundation (FAPESP); Grant sponsor: Brazilian National Research Council (CNPq); Grant sponsor: Organon; Grant sponsor: Unilever

*Correspondence to: Gerrit Vriend, CMBI KUN, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands. E-mail: vriend@cmbi.kun.nl

Received 21 May 2002; Accepted 14 April 2003

ity, for which a Shannon-type entropy term is mostly used. Many of these methods are well suited to find functionally important residues of one kind or another, but despite the abundance of knowledge about protein families that can be obtained, methods do not yet exist to generate a comprehensive overview of the function of all residues relative both to each other and to the structure.

It seems obvious that a residue conserved in a sequence family must be involved in a function common to the family. A residue conserved only in subfamilies, on the other hand, is likely to have a functional role in those subfamilies only. A problem with this simple understanding is that we often have too few sequences to establish the significance of conservation. If a residue is conserved, the question is whether this is because it has a functional role, or the number of sequences is too small, or because the sequences do not vary enough. We propose to harvest the wealth of information in very wide multiple-sequence alignments. The number of conserved residues decreases as more sequences are aligned.[1,2] For the families we have studied, so many sequences are available that any observed sequence conservation pattern is clearly significant.

We observe that residues located at one specific position can differ from subfamily to subfamily, but if they are conserved within each of the subfamilies, they perform the same (or similar) function. For example, a calcium-binding ligand can be an aspartic acid in some subfamilies and an asparagine in others. This "conservation of the location of function" concept, combined with the use of very many sequences, can answer this question. If the residue at a position in the multiple-sequence alignment is conserved in one subfamily but variable in another, residues at this position do not have a function that is important for the whole family. We call such residue positions "recalcitrant" and present a simple qualitative algorithm to detect recalcitrant residue positions.

We have developed a sequence-analysis technique based on the combination of two commonly used sequence-variability measures. The first is variability, defined as the number of different amino acid types observed at each position. The second is Shannon entropy. Each residue position in the alignment is plotted on the entropy–variability diagram. Boxes in this plot appear to represent groups of residues that share a common structural or functional characteristic.

The two measures for variation and conservation (variability and entropy) are not new, but their combination is new. We find large functional differences between residues with similar entropy but different variability. Similarly, we find large functional differences between residues with similar variability but different entropy. The fine-tuning of the entropy by the variability, and of the variability by the entropy, allows us to draw many more conclusions about the role of individual residue positions than are possible using other techniques.

We have tested the method on three protein families for which very many sequences are available: globin chains,[23–27] ras-like proteins,[28–32] and serine-proteases.[33,34] We chose these three families because of the extensive literature about them, and because the role of almost every residue in these families is known. Clustered in the entropy–variability plots are positions related to the main function, to binding cofactors or regulatory ligands; positions in the core of the protein, either closer to or further away from the main functional site; and positions at the surface not associated with any known function. We provide a qualitative recipe for the division of the entropy–variability plots into boxes that correspond with these functions. Several aspects of the method are not yet fully optimized. We think, however, that the analysis of entropy–variability plots holds great promise for the near future, when thousands of sequences will become available for many protein families.

MATERIALS AND METHODS

We obtained sequences from GenBank[35] and TrEMBL,[36] and 3D coordinates of protein structures from the Protein Data Bank (PDB).[37] We performed multiple-sequence alignments as described before,[38] using the sequence manipulation options of the program WHAT IF.[39] The profile-driven multiple-sequence alignment procedure has two steps. First, we used profiles corresponding to the full length of the sequences to align groups of related sequences (the so-called sequence groups). The percentage of sequence identity between the consensus sequence of the profile and the individual sequences was typically around 90%. In the second step, we aligned the groups of aligned sequences (sequence groups) using only segments with higher than average sequence identity. In practice, these segments tend to correspond to regular secondary structure elements. We removed sequences if they were identical over the full length of the final alignment to a sequence already incorporated, or if they contained unidentified residues. Residue positions were not used if one or more sequences displayed a deletion at that position.

The Shannon entropy at position p in the multiple sequence alignment, Sp, is given by

$$Sp = -\sum_{i=1}^{20} f_{pi} * \ln(f_{pi}), \qquad (1)$$

in which i loops over the 20 amino acid types, f_{pi} is the weighted frequency of residue type i at alignment position p (see below for weight factors), and p loops over the length of the profile. Sp can range from 0.0 for fully conserved sequence positions to ln(20), when all 20 residue types are observed at a frequency of 0.05.

We define the variability at position p in the multiple-sequence alignment, V_p, as the number of different residue types observed at position p in at least 0.5% of all sequences (obviously, V_p varies from 1 to 20).

We introduced sequence weights to reduce the influence of sequences that are either too similar to each other, or too different from all others. The weight W_g for a sequence group g is defined as

A)

```
GROUP                          POSITION
                        10         20          30         40
                         |         |           |          |*
                 FKLVFLGEQSVGKTSLITRFMYDSF  GSDVIIMLVGNKTDLADKRQ
   1             FKLVFLGEQSVGKTSLITRFMYDSF  GSDVIIMLVGNKTDLADKRQ
                 FKLVFLGEQSVGKTSLITRFMYDSF  GSDVIIMLVGNKTDLADKRQ
                 FKLVFLGEQSVGKTSLITRFMYDSF  GSDVIIMLVGNKTDLSDKRQ

                 *
                 HKVIMVGSGGVGKSALTLQFMYDEF  EDKIPLLLVGNKSDLEDRRQ
                 HKVIMVGSGGVGKSALTLQFMYDEF  EDKIPLLLVGNKSDLEDRRQ
   2             LKVIMVGSGGVGKSALTLQFMYDEF  EDKIPLLLVGNKSDLEDRRQ
                 HKVIMVGSGGVGKSALTLQFMYDEF  EDKIPLLLVGNKSDLEDRRQ
                 HKVIMVGSGGVGKSALTLQFMYDEF  EDKIPLLLVGNKSDLEDRRQ

                         *                              ***
                 YKLVVVGARGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGAVGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGARGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGAGGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGARGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAGRTV
   3             YKLVVVGARGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGAGGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGAGGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGAKGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAARTV
                 YKLVVVGARGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLAAHTV
                 YKLVVVGAGGVGKSALTIQLIQNHF  SDDVPMVLVGNKCDLPARTV
```

B)

Position	Group 1 S	V	2 S	V	3 S	V	Av	P	H-Entropy
1	0.00	1	0.50	2	0.00	1	0.17	0.33	0.056
9	0.00	1	0.00	1	1.16	4	0.39	0.33	0.129
41	0.56	2	0.00	1	0.13	2	0.23	0.67	0.154
42	0.00	1	0.67	2	0.13	2	0.27	0.67	0.181
43	0.00	1	0.00	1	0.13	2	0.04	0.33	0.013

Fig. 1. Example of the calculation of H-entropy. (**A**) Part of the multiple-sequence alignment of ras-like proteins. Two fragments of sequences are shown for three groups. Positions that are variable in a group are labeled with an asterisk. (**B**) S and V stand for entropy and variability. Av indicates the average entropy at each position. Position refers to the position in the multiple-sequence alignment in (A). P is the fraction of groups in which entropy is larger than zero. H-entropy is given by the product: Av $\times$ P.

$$W_g = \left\{ [1 - S_g/\ln(20)] \Big/ \sum_{j=1}^{N} [1 - S_j/\ln(20)] \right\}, \quad (2)$$

in which S_x is the entropy value averaged over all sequence positions in sequence group x, ln(20) is the maximum possible entropy, and N is the number of sequence groups. The weights of individual sequences are derived by dividing the weight of their sequence group by the number of sequences in that group.

We used the following algorithm to detect recalcitrant residue positions:

1. Select all sequence groups that consist of more than two sequences.
2. Determine Sp for each position p in every group [Eq. (1)].
3. Determine for each position p the average Sp value over all groups.
4. Determine for each position p the fraction of groups with nonzero Sp.
5. Determine for each position p the H-entropy, which is defined as the product of the average entropy (step 4) with the fraction of groups having a nonzero entropy (step 5).
6. All residue positions p for which the H-entropy is larger than a cutoff value (normally 20% of the maximum H-entropy observed in the whole alignment) are called recalcitrant.

To illustrate the calculation of H-entropy, Figure 1(A) shows some ras-like sequence segments. Two segments contain 5 positions that are not entirely conserved in groups 1–3: 1, 9, 41, 42, and 43 (marked with an asterisk). Figure 1(B) shows how the H-entropy values were calculated. For clarity, we used no weighting procedures in this example.

RESULTS AND DISCUSSION

Aiming at a method for harvesting the wealth of information present in sequence (and structure) data, we analyzed conservation patterns in multiple-sequence alignments. Many of the results obtained in this study (alignments, list of file names, etc.) are too voluminous to print. They are available at http://www.gpcr.org/articles/.

TABLE I. Family Statistics: Three-Sequence Families

Protein Family	Groups	Seq	Pos
Globin chains	364	753	113
Ras-like proteins	335	562	152
Serine-proteases	176	301	173

Groups: the number of sequence groups (i.e., groups of sequences with more than 90% sequence identity to the group's profile); Seq: total number of sequences in the alignment; Pos: number of sequence positions that could be aligned.

Sequences and Alignment

To infer a major functional role from conservation, it is not enough that a residue is very conserved; absolute conservation is required. Examples are the heme-binding methionine in the cytochromes and the active-site serine in serine-proteases that are indeed absolutely conserved.[40,41] In a statistical sense, the bias that one may see with a small sequence set might lead to the false identification of a residue as being functionally conserved. On the other hand, not having enough sequence data might mean that a truly conserved residue is not identified, again, because of sample bias, or the like. On the other hand, high residue variability can be interpreted more directly, because it indicates one of three possibilities:

1. Variable positions are not crucial for any function.
2. They are involved in a function that differs from species to species.
3. The (local) alignment is incorrect.

Local errors occur routinely in multiple-sequence alignments at positions corresponding to loops in the 3D structure, whereas the sequence alignment of regular secondary structure elements is normally routine. If the structures are locally very different, the alignments are meaningless. Insertions and deletions in the multiple-sequence alignment are the best indicators of significant differences in the 3D structure. We therefore exclude from our analyses all positions in which an insertion or deletion is observed.

Local alignment errors are common when large families of sequences are aligned in a single run, especially if many pairs of sequences show less than 25% pairwise identity. Because we wanted to incorporate as many sequences as possible, we had to design an alignment method that could cope with sequence families with an average pairwise sequence identity as low as 20%. Our two-step alignment procedure produces better results for the three families used in this study (based on structure alignments) than can be obtained with standard alignment software. Recent studies on multiple-sequence alignments validated by structure alignments have led to similar conclusions.[42,43]

Table I shows some statistics about the alignments used in this study. Figure 2(A–C) indicates the consensus sequences.

Most multiple-sequence alignment techniques are based on an all-against-all pairwise sequence alignment. We intend to use entropy–variability analysis for large sequence families, such as the nuclear receptors (900 sequences) or G protein–coupled receptors (>2000 sequences). With such large numbers of sequences programs, such as CLUSTAL, based on an all-against-all pairwise sequence, alignment become inconveniently slow, and when these sequence families grow even further, the only solution is the use of an alignment method that consumes CPU time as a linear function of the number of sequences. There are several reasons for the choice of a two-step iterative profile procedure. The first reason is speed. When a number of sequences with average pairwise sequence identity of 90% or more are aligned against a dedicated profile, their alignment is virtually guaranteed to be correct. Use of these aligned sequences as one block in the main (iterative) alignment procedure saves CPU time. Second, recalcitrant residues can only be detected in groups of sequences that have a high sequence similarity, and that are certainly aligned correctly. Third, when new sequences come in, they only need to be aligned once against each of the profiles to determine where they belong in the hierarchy.

At present, the number of profiles is about half the number of sequences. This seems strange, because it means that for many profiles, there is just one sequence available. In practice, we see that each sequence database update leads to a larger sequence:profile ratio. Furthermore, we need this approach, because we need groups of highly similar sequences to detect the recalcitrant residues that would give rise to all kinds of artifacts, if they were used in the analyses. We classify the proteins in groups that have around 90% pairwise sequence identity. Other groups have developed methods to classify proteins based on the analysis of sequence databases, attaining high levels of accuracy.[44–46] Some of these methods use entropy to define subclasses of proteins similar to our method,[44] but all of them aim at the identification of all possible functional variants of the proteins to classify as many sequences as possible. Our classification cannot be compared directly to these methods, because we have a different goal. Our main intention was to classify the sequences into groups by using two criteria: (1) including as many and as different sequences as possible; (2) aligning them in highly similar groups first to detect recalcitrant residue positions. These groups of highly similar sequences do not necessarily need to agree with the perfect phylogenetic clustering. At about 90% sequence identity, the sequences are certainly homologous, and homology is all that is needed for the detection of recalcitrant residues.

Entropy–Variability Plots

Figure 3(A–C) shows the entropy–variability distributions. As expected, considering the heterogeneity of the sequences used, the entropy values are widely spread. No residue positions with variability larger than 14 were observed in the globin family, because all highly variable loops had at least one insertion or deletion in at least one family member (these loops do not contain residues for which any functional importance has been reported in the literature). To cluster the residue positions in five groups

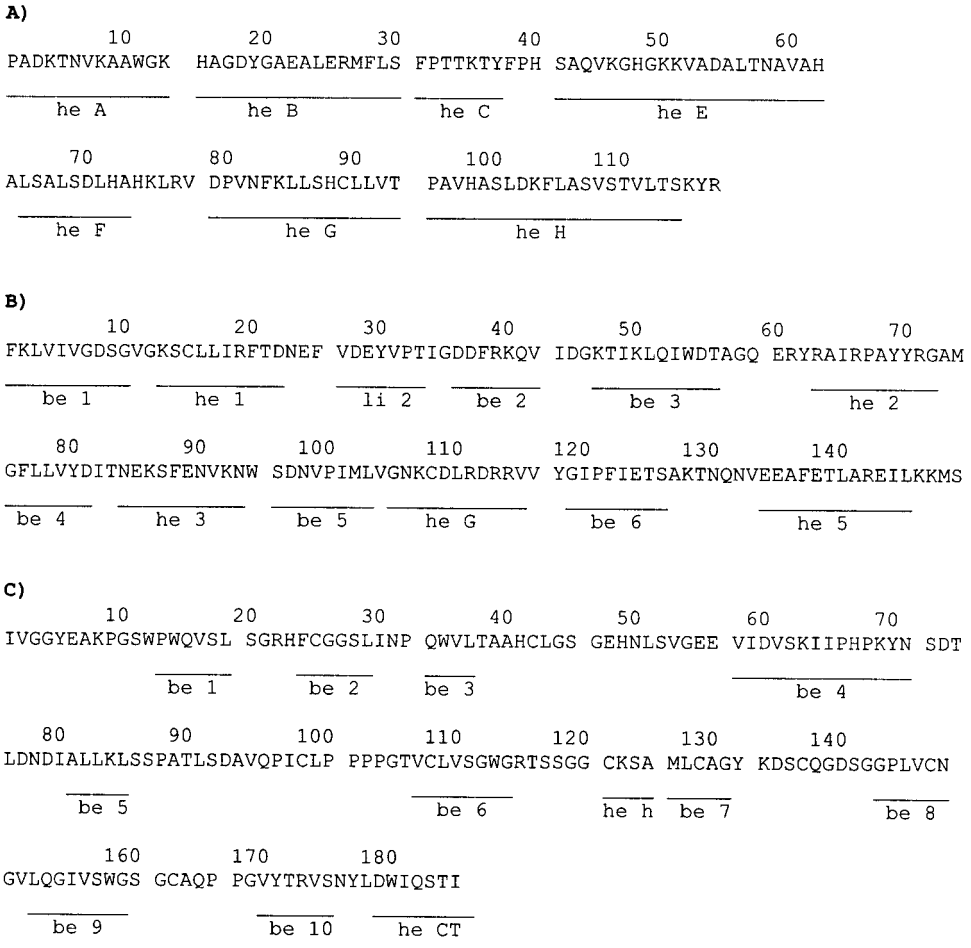

Fig. 2. Consensus sequences: (**A**) globin chains; (**B**) ras-like proteins; (**C**) serine-proteases. The underlined positions correspond to helices (he) and β-strands (be), with commonly used nomenclature. Li indicates a linker region. Secondary structure information was obtained from the PDB files for globins (2HHD), ras-like proteins (5P21), and serine proteases (2PTC).

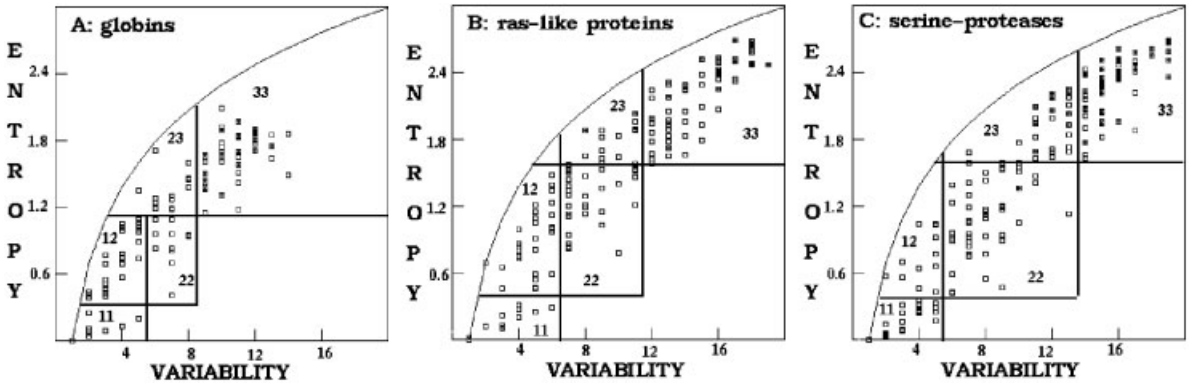

Fig. 3. Entropy–variability plots: (**A**) globins; (**B**) ras-like proteins; (**C**) serine-proteases. The curved line indicates the maximum entropy possible as function of the variability.

(labeled 11, 12, 22, 23, 33; Fig. 3) according to their entropy and variability, we used the following simple qualitative procedure:

1. The entropy axis was divided into three parts. The lower boundary was at entropy = 0.4, and the upper boundary was placed halfway between entropy = 0.4 and the maximal observed entropy.
2. The variability axis was divided into three parts. The lower boundary was at the highest variability in Box 11, and the upper boundary was at the highest variability in Box 22.

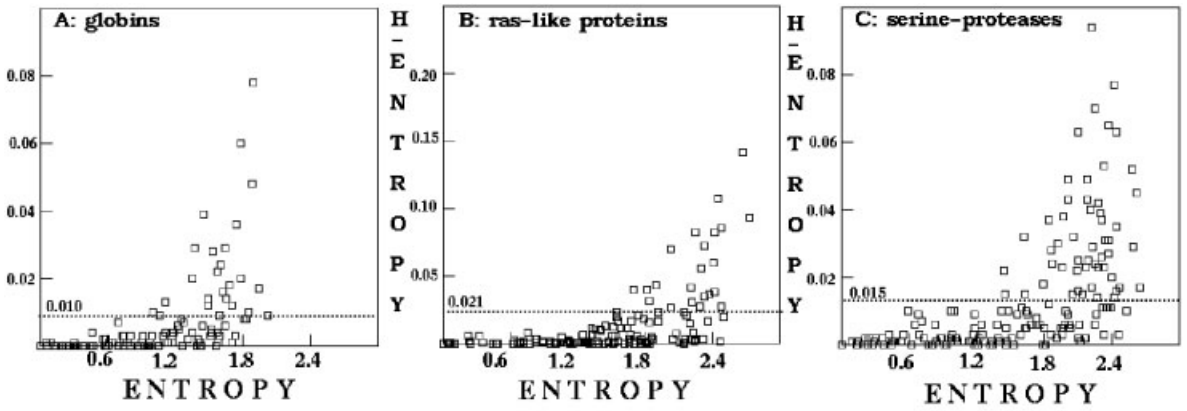

Fig. 4. H-entropy–entropy plots: (**A**) globins; (**B**) ras-like proteins; (**C**) serine-proteases. Dashed lines indicate the recalcitrance cutoff values. Note that plots have different vertical scales.

We selected the three classes of proteins, because the role of each of the amino acids was already known from years of study in hundreds of laboratories. We optimized the clustering algorithm, so that the residues of known function cluster optimally in the five boxes. This is not a quantitative process, but a qualitative one that gives the best results for these three well-studied molecules, and for the G protein–coupled receptors (GPCRs) discussed in the next article in this issue. Consequently, we cannot justify the choice of parameters. The parameters chosen were hand-optimized for the three classes we studied. We observed that the function of a residue near a boundary between two boxes tends to be a combination of the functions of residues in those flanking boxes. This means that the precise definition of the boxes is not too important. Despite the procedure for defining boxes in Figure 3(A–C) being rigorously defined by the two rules mentioned above, it is to be expected that future analyses of more families with more sequences will lead to a fine-tuning of the algorithm.

Residue-Position Function Correlates With Boxes

The mapping of the residue positions in the boxes in Figure 3(A–C) in the 3D structures shows the following correlations:

1. Box 11 contains residue positions with low entropy and low variability, which form the main functional site. These residue positions are involved in catalysis or signaling mechanisms. Most positions of key structural residues (e.g., Cys-Cys bridges) are also found in this box.
2. Box 12 contains positions of residues located in the core. They are adjacent to the positions of Box 11 and mainly form the first shell of positions around the main functional site.
3. Box 22 contains mainly positions of core residue, farther away from Box 11 positions. The residues in Box 22 positions are thought to have a structural role, but their location between the main functional site and the modulator site(s) suggests that they are also involved in communication between modulators and the main functional site. For several amino acids observed in this box, such a function has been experimentally established.
4. Box 23 contains most residue positions involved in interactions with the modulator(s). They can be located either at the surface or in the core of the proteins.
5. Box 33 contains residue positions located mainly at the surface of the proteins. The positions in Box 33 involved in modulator interaction are found mainly at the surface of the protein, in locations that suggest they are not involved in communication between the modulator and the main functional site. For some positions in this box, the alignment is doubtful, and most recalcitrant residue positions are observed in this box.

In summary, the boxes in Figure 3(A–C) allow us to identify the main functional site (normally called the active site), one or more modulator sites, and a protein core that connects these regions. The remaining residues are highly variable and either do not have any clear function, or have a function that is important only for a small subset of the sequences, and are thus not important to the family as a whole. We have tried to quantify these observations in terms of physicochemical properties as a function of the box number. A series of results given in the website are inconclusive. For example, the active site of the globins is buried, the active site of the serine proteases is half-buried, and the active site of the ras-like proteins is exposed. Therefore, a parameter such as solvent accessibility cannot explain the distribution over the boxes. The residues are really distributed over the boxes according to the signature that evolution has left in the entropy–variability combination.

H-Entropy

Figure 4 shows plots of H-entropy as a function of the Shannon entropy. The dashed lines indicate the cutoff values for recalcitrance. Cutoff values were 0.010, 0.021, and 0.015 for the globins, ras-like proteins, and serine-proteases, respectively.

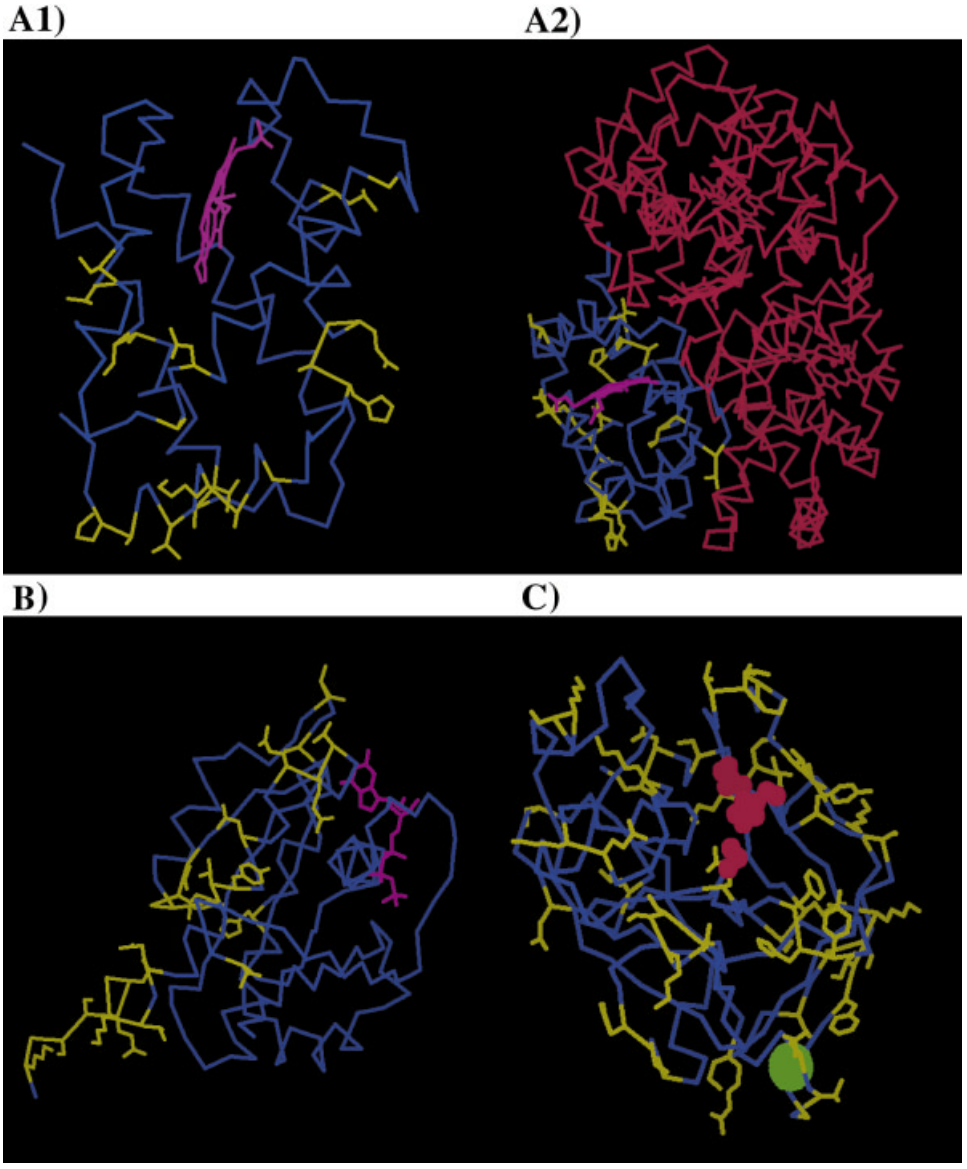

Fig. 5. Recalcitrant residue positions: (**A1**) globin monomer; (**A2**) globin tetramer ("other three monomers" drawn in red); (**B**) ras-like proteins; (**C**) serine-proteases. The haem group (A1 and A2), the GDP (B), and the catalytic triad (C) are shown in purple. The PDB files used are the same as in Figure 2.

In the globins, most recalcitrant residues are located at the side of the molecule opposite the heme-binding site, mainly in helices A, B, E, and H [Fig. 5(A)]. In the hemoglobin tetramer, the recalcitrant positions are mostly solvent-accessible and are not part of the α1-β1 or α1-β2 interfaces that modulate the cooperative events [Fig. 5(B)]. In the ras-like proteins, the recalcitrant positions are at the opposite side of the molecule to the nucleotide-binding site, in helix 3, helix G, helix 5, strand 6, and in the C-terminal segment (Ct) [Fig. 5(C)]. In the serine-proteases, the recalcitrant positions are located mainly in turns between β-strands and in the helices h and ct. These parts form the surfaces of the enzyme and are not part of the interface between the two domains (where the catalytic residues are located) [Fig. 5(D)].

In summary, the recalcitrant residues are not involved in any of the known functions of the three proteins studied (neither in ligand binding, nor in catalysis, multimerization, cooperativity, etc.). Thus, we can conclude that use of recalcitrance as a filter significantly improves the analyses of multiple-sequence alignments.

Residue Types Observed in the Entropy–Variability Sectors of Figure 3(A–C)

High frequencies of Gly, His, Leu, Arg, Thr, and Tyr are observed in the heme-binding site of globins. The nucleotide-binding site of ras-like proteins contains high frequencies of Ala, Asp, Glu, Phe, Gly, Lys, Gln, Ser, and Thr. The serine-protease catalytic center and its direct environment are formed mainly by Ala, Cys, Asp, Gly, His, Leu, Pro,

Ser, and Trp residues. There is an apparent preference for hydrophobic (Leu, Ile, Val, and Phe) and small (Ser, Asn, Gly, and Ala) residues in the core positions of the three proteins studied. No residue preferences are observed for the variable positions. These residue preferences agree with our understanding of the three classes of proteins. We find typical core residues in the core, typical heme- or nucleotide-binding residues in the heme- and nucleotide-binding pockets, and no residue preferences in the surface positions that have no special function.

CONCLUSIONS

Entropy–variability plots tell us about sequence–structure and sequence–function relationships. The recalcitrant positions are interesting. The residues found in these positions may often be conserved in many of the subfamilies of proteins used for the initial profile alignments, but this apparent conservation may be the result of small numbers of sequences and low variability within them, rather than being indicative of functional importance. Recalcitrant positions can be detected when very many sequences are available. Much experimental information is available for the three families of proteins analyzed in this study, and the individual function (or absence thereof) is known for nearly all residue positions. This allowed us to determine a qualitative algorithm for the determination of recalcitrant positions. This algorithm will undoubtedly be improved. With the current available data, however, this algorithm is the best we can do, and the results indicate that the detection of recalcitrant residues is worth the effort.

The entropy–variability plot can be a data-mining tool that promises success, if very large numbers of sufficiently variable sequences and structural information are available. Thus, it is a natural choice for genomic, proteomic, structomic, and so on, projects producing a flood of data and needing a data-mining tool to provide a maximum amount of information for multiple purposes.

ACKNOWLEDGMENTS

Our thanks to Florence Horn and David Thomas.

REFERENCES

1. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
2. Dodge C, Schneider R, Sander C. The HSSP database of protein structure–sequence alignments and family profiles. Nucleic Acids Res 1998;26:313–315.
3. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. Proteins 1991;11:297–313.
4. Pei J, Grishin NV. AL2CO: Calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700–712.
5. Mirny L, Shakhnovich E. Evolutionary conservation of the folding nucleus. J Mol Biol 2001;308:123–129.
6. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Zuckerkandl E, Pauling L, editors. Evolving genes and proteins. New York: Academic Press; 1965. p 97–166.
7. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–196.
8. Russell RB, Sasieni PD, Sternberg MJ. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
9. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule–ligand interactions. J Mol Biol 1982;161:269–288.
10. DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID, Venkataraghavan R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J Med Chem 1988;31:722–729.
11. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
12. Miranker A, Karplus M. Functionality maps of binding sites: A multiple copy simultaneous search method. Proteins 1991;11:29–34.
13. Lamb ML, Jorgensen WL. Computational approaches to molecular recognition. Curr Opin Chem Biol 1997;1:449–457.
14. Wang W, Donini O, Reyes CM, Kollman PA. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 2001;30:211–243.
15. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
16. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
17. Pietrokovski S, Henikoff JG, Henikoff S. The Blocks database—a system for protein classification. Nucleic Acids Res 1996;24:197–200.
18. Shatsky M, Nussinov R, Wolfson HJ. Flexible protein alignment and hinge detection. Proteins 2002;48:242–256.
19. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
20. Sali A, Overington JP, Johnson MS, Blundell TL. From comparisons of protein sequences and structures to protein modelling and design. Trends Biochem Sci 1990;15:235–240.
21. Innis CA, Shi J, Blundell TL. Evolutionary trace analysis of TGF-beta and related growth factors: Implications for site-directed mutagenesis. Protein Eng 2000;13:839–847.
22. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001; 307:1487–1502.
23. Nobbs CL, Watson HC, Kendrew JC. Structure of deoxymyoglobin: A crystallographic study. Nature 1966;209:339–341.
24. Watson, HC. The stereochemistry of the protein myoglobin. Prog Stereochem 1969;4:299–333.
25. Perutz MF. Sterochemistry of cooperative effects of haemoglobin. Nature 1970;228:726–739.
26. Perutz MF. The haemoglobin molecule. Proc R Soc 1969;173(B): 113–140.
27. Royer WE Jr, Hendrickson, WA, Chiancone E. Structural transitions upon ligand binding in a cooperative dimeric hemoglobin. Science 1990;249:518–521.
28. Pai EF, Kabsch W, Krengel U, Holmes KC, John J, Wittinghofer A. Structure of the guanine-nucleotide-binding domain of the Ha-ras oncogene product p21 in the triphosphate conformation. Nature 1989;341:209–214.
29. Pai EF, Krenge, U, Petsko GA, Goody RS, Kabsch W, Wittinghofer A. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 Å resolution: Implications for the mechanism of GTP hydrolysis. EMBO J 1990;9:2351–2359.
30. Takai Y, Sasaki T, Matozak T. Small GTP-binding proteins. Physiol Rev 2001;81:153–208.
31. Huang L, Hofer F, Martin GS, Kim SH. Structural basis for the interaction of Ras with RalGDS. Nat Struct Biol 1998;5:422–426.
32. Pacold ME, Suire S, Perisic O, Lara-Gonzalez S, Davis CT, Walker EH, Hawkins PT, Stephens L, Eccleston JF, Williams RL. Crystal structure and functional analysis of Ras binding to its effector phosphoinositide 3-kinase gamma. Cell 2000;103:931–943.
33. Ruhlmann A, Kukla D, Schwager P, Bartels K, Huber R. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor: Crystal structure determination and stereochemistry of the contact region. J Mol Biol 1973;77:417–436.
34. Huber R, Kukla D, Bode W, Schwager P, Bartels K, Deisenhofer J, Steigemann W. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor: II. Crystallographic refinement at 1.9 Å resolution. J Mol Biol 1974;89:73–101.

35. GenBank: http://www.ncbi.nlm.nih.gov/genbank
36. Swissprot/TrEMBL: http://www.ebi.ac.uk/swissprot
37. Protein Data Bank: http://www.rcsb.org/pdb
38. Oliveira L, Paiva AC, Vriend G. A common motif in G protein–coupled seven transmembrane helix receptors. J Comput-Aided Mol Des 1993;7:649–658.
39. Vriend G. WHAT IF: A molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
40. Calhoun MW, Lemieux LJ, Garcia-Horsman JA, Thomas JW, Alben JO, Gennis RB. The highly conserved methionine of subunit I of the heme-copper oxidases is not at the heme-copper dinuclear center: Mutagenesis of M110 in subunit I of cytochrome bo3-type ubiquinol oxidase from *Escherichia coli*. FEBS Lett 1995;368:523–525.
41. Dodson G, Wlodawer A. Catalytic triads and their relatives. Trends Biochem Sci 1998;23:347–352.
42. Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. Bioinformatics 2002;18:306–314.
43. Elofsson A. A study on protein sequence alignment quality. Proteins 2002;46:330–339.
44. Hannenhalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 2000;303:61–76.
45. Silverstein KA, Shoop E, Johnson JE, Retzel EF. MetaFam: A unified classification of protein families: I. Overview and statistics. Bioinformatics 2001;17:249–261.
46. Yona G, Linial N, Linial M. ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins 1999;37:360–378.