From Sequence to Structure and Function with Entropy-Variability Plots
Laerte Oliveira¹, Paulo B. Paiva¹, Antonio C.M. Paiva¹, Gerrit Vriend².
¹Escola Paulista de Medicina, UNIFESP, Sao Paulo, Brazil
²CMBI, KUN, The Netherlands.

Running Title: Entropy-Variability Plots

Correspondence to: Gerrit Vriend, CMBI KUN, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
Tel 31-24-3653391; Fax 31-24-3652977; Email "Vriend@CMBI.KUN.NL"

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

We introduce sequence entropy-variability plots as a method of analysing 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 versus 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.
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.) are leading to a rapid increase in the number of solved structures. As a result of this, today about forty five sequences enter the EMBL database every minute [http://www3.ebi.ac.uk/Services/DBStats/], and about seven 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 there are many programs to visualize the sequence conservation and variability (expressed as a Shannon entropy term) information available in these files. This allows researchers to quickly find important (=conserved) residues such as those in the active site. Alternatively, variable positions can be found where mutations are likely to be introducible 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, variance-based and score-matrix indices\(^3\)-\(^5\). The patterns of conservation in proteins have been used for the 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 often are clustered in certain regions of protein structures\(^4\), sometimes at "universally conserved positions"\(^7\), so called because they can form a motif that is 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 so-called super-sites\(^8\).

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 supposed to interact\textsuperscript{9-14}. Other groups have predicted functional motifs from an analysis of protein interaction surfaces using principal component analysis\textsuperscript{15}, analysis of physicochemical descriptors to score protein-protein interactions\textsuperscript{16}, search for motifs in Blocks databases\textsuperscript{17}, or alignment of hinge regions\textsuperscript{18}. Evolutionary trace analyses involve searching for conservation patterns in different branches of phylogenetic trees and mapping them onto 3D structures to look for clusters of functionally important residues\textsuperscript{15-22}.

Some of these methods require knowledge of the three dimensional structure of the protein, but most of them require no more than a multiple sequence alignment. All methods that are based on sequence variability analysis use one single measure of variability, for which mostly a Shannon-type entropy term is used. Many of these methods are very well suited to find functionally important residues of one kind or another, but despite the great deal of knowledge about protein families that can be obtained, methods do not exist yet to generate a comprehensive overview of the function of all residues relative to each other and relative 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 sub-families, on the other hand, is likely to have a functional role in those sub-families 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\textsuperscript{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 sub-family to sub-family, but if they are conserved within each of the sub-families, they perform the same (or similar) function. For example, a calcium binding ligand can be an aspartic acid in some sub-families 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 sub-family, but variable in another, residues at this position do not have a function that is important for the whole family. We will call such residue positions 'recalcitrant', and we will 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 versus 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. It is the fine-tuning of the entropy by the variability and the fine-tuning of the variability by the entropy that allows us to draw many more conclusions about the role of individual residue positions, than is possible using other techniques.

We have tested the method on three protein families for which very many sequences are available: globin chains, ras-like proteins, and serine-proteases. We chose these three families because there exists a very extensive literature about them and the role for almost every residue in these families is known. Positions related to the main function, positions related 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 cluster nicely in the entropy-variability plots. We provide a qualitative recipe for the division of the entropy-variability plots in boxes that correspond with these functions. Several aspects of the method are not yet fully optimised. 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.

MATERIAL AND METHODS

Sequences were obtained from GenBank and TrEMBL. 3-D coordinates of protein structures were obtained from the PDB. Multiple sequence alignments were performed as described before, using the sequence manipulation options of the program WHAT IF. The profile-driven multiple sequence alignment procedure had two steps. First, profiles corresponding to the full length of the sequences were used to align groups of related sequences (the so-called sequence groups). The percentage sequence identity between the consensus sequence of the profile and the individual sequences was typically around 90%. In the second step,
the groups of aligned sequences (sequence groups) were aligned using only segments with higher than average sequence identity. In practice, these segments tend to correspond to regular secondary structure elements (helix/strand). Sequences were removed 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, $S_p$, is given by:

$$S_p = - \sum_{i=1}^{20} f_{pi} \ln(f_{pi})$$  \hspace{1cm} (1)

in which $i$ loops over the 20 amino acid types; $f_{pi}$ is the weighed frequency of residue type $i$ at alignment position $p$ (see below for weight factors); and $p$ loops over the length of the profile. $S_p$ can range from 0.0 for fully conserved sequence positions till $\ln(20)$ when all twenty 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).

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

$$W_g = \frac{N}{\sum_{j=1}^{N} (1-S_j/\ln(20))}$$

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.

The following algorithm was used to detect recalcitrant residue positions:

1. Select all sequence groups that consist of more than two sequences.
2. Determine $S_p$ for each position $p$ in every group (eq 1).
3. Determine for each position $p$ the average $S_p$ value over all groups.
4. Determine for each position $p$ the fraction of groups with non-zero $S_p$.
5. Determine for each position $p$ the H-entropy. The H-entropy is defined as the product of the average entropy (step 4) with the fraction of groups
having a non-zero entropy (step 5).

6. All residue positions \( p \) for which the H-entropy is larger than a cut-off value (normally 20% of the maximum H-entropy observed in the whole alignment) are called recalcitrant.

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

**RESULTS AND DISCUSSION**

Aiming at a method for harvesting the wealth of information present in sequence (and structure) data, we analysed 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/](http://www.gpcr.org/articles/).

**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 haem-binding methionine in the cytochromes, and the active site serine in serine-proteases, that are indeed absolutely conserved\(^{40,41}\). However, the problem with this simple reasoning is that residue conservation is likely to be systematically found if the number of sequences is too small, or if the sequences are too closely related. On the other hand, high residue variability can be interpreted in a more direct way since 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 3-D structure. If the structures are locally very different, the alignments are not necessarily wrong but are meaningless. Insertions and deletions in the multiple sequence alignment are the best indicators of significant differences in the 3-D structure. We therefore exclude from our analyses all positions where 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 a less than 25% pair-wise 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 pair-wise sequence identity as low as 20%. Our two-step alignment procedure produces better results for the three families used in this study (conclusion based on structure alignments) than could be obtained using 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. Figures 2A-C show the consensus sequences.

Most multiple sequence alignment techniques are based on an all-against-all pair-wise 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, that are based on an all-against-all pair-wise sequence alignment, become unpleasantly slow, and when these sequence families grow even further, the only solution is the use of an alignment method such as iterative profile alignment that grows linearly in CPU time as 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 pair-wise sequence identity of 90% or more are aligned against a dedicated profile, their alignment is virtually guaranteed to be correct. Using 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 know 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 sequence database updates each time lead to a larger sequences/profiles ratio. Further, 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 artefacts if they were used in the analyses. We classify the proteins in groups that have around 90% pair-wise sequence identity. Other research groups have developed methods to classify proteins based on analysis of sequence databases, attaining high levels of accuracy.\(^{44-46}\) Some of these methods use entropy to define sub-classes of proteins similar to our method, but all of them aim at the identification of all possible functional variants of the proteins in order to classify as many sequences as possible. Our classification cannot be compared to these methods because we have a different goal. Our main scope was
to classify the sequences into groups using two criteria, 1) including as many and as different sequences as is 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 surely homologous, and homology is all that is needed for the detection of recalcitrant residues.

**Entropy-variability plots**

Figures 3A-C show 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 are 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 (labelled 11, 12, 22, 23, 33; see fig 3) according to their entropy and variability, we used the following simple qualitative procedure:

1. The entropy axis was divided in three parts. The lower separation was at entropy=0.4 and the higher separator was placed halfway entropy=0.4 and the maximal observed entropy.
2. The variability axis was divided in three parts. The lower separation was at the highest variability in box 11 and the higher separation was at the highest variability in box 22.

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 labs. We optimised the clustering recipe so that the residues of known function cluster optimally in the five boxes. This is not a quantitative process, but a qualitative process that gives the best results for these three well-studied molecules, and for the GPCRs mentioned in the next article. Consequently, we cannot give a justification for the choice of parameters. The parameters chosen were hand-optimised 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 very precise definition of the boxes is not too overly important. Despite that the procedure for defining boxes in Figures 3A-C is rigorously defined by the two rules mentioned above, it is to be expected that the future analyses of more families that contain more sequences will lead to a fine-tuning of the recipe. At present, this is the best we can do.

**Residue position function correlates with boxes**

The mapping of the residue positions in the boxes of Figures 3A-C in the three-
dimensional structures shows the following correlations (Figs. 4A-C):

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 signalling 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. They mainly form the first shell of positions around the main functional site.

3. Box 22 contains mainly positions of core residue, further 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 that are mainly located at the surface of the proteins. The box 33 positions that are involved in modulator interaction are mainly found at the surface of the protein in locations that suggest that 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 Figures 3A-C allow us to identify the main functional site (normally called 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 only is important for a small subset of the sequences, and thus not important to the family as a whole. We have tried to quantify these observations in terms of physico-chemical properties as function of the box number. A series of results, given at the website, are non-conclusive. 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 very exposed. Therefore, a parameter like solvent accessibility cannot explain the distribution over the boxes. The residues really are distributed over the boxes according to the signature that the evolution has left in the entropy-variability combination.

**H-entropy**

Figure 5 shows plots of H-entropy as function of the Shannon entropy. The dashed
lines indicate the cutoff values for recalcitrance. The cutoff values were 0.010, 0.021, and 0.015 for the globins, ras-like proteins, and serine-proteases, respectively.

In the globins, most recalcitrant residues are located at the opposite side of the molecule to the haem binding site mainly in the helices A, B, E and H (Fig. 6A). 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. 6B). 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. 6C). 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. 6D).

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, multimerisation, cooperativity, etc). Thus, we can conclude that using recalcitrance as a filter significantly improves the analyses of multiple sequence alignments.

Residue types observed in the entropy/variability sectors of Figures 3A-C
High frequencies of Gly, His, Leu, Arg, Thr and Tyr are observed in the haem-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 centre 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 haem- or nucleotide-binding residues in the haem- and nucleotide binding pockets and no residue preferences in the surface positions that have no special function.

Evolutionary considerations
Our observations agree with a very simple two-stage evolutionary model:
1. The ancestral and non-functional form of a protein gets a binding site, and starts performing a function.
2. This function is modulated by secondary binding sites (modulator sites). Once an active site exists, it will stay conserved. Most active sites have critical sequence and structure requirements, which, together with the constraints of the
protein fold, restrict the mutational speed and the range of possible residues of most of the protein's core. The modulation of this function should occur quite quickly on an evolutionary time scale. This high speed requires the involvement of surface residues. The core residues between the active site and the modulator sites are less variable near the conserved active site residues and more variable near the modulator sites. Obviously, the residue positions involved in modulator interaction are normally conserved between proteins that interact with the same modulator, but tend to differ between proteins that interact with different modulators.

Some surface residues that are not directly involved in modulation of the main function are also likely to have a functional role. However, these functions are of lesser importance and thus put weaker constraints on the sequence, or they are only important in certain species or certain cell types. Such residues may, for example, have a function related to:

1. DNA or RNA structural integrity
2. Replication or translation requirements
3. DNA or RNA modification requirements
4. Scarcity of certain amino acids
5. Thermostability requirements

CONCLUSIONS

Entropy-variability plots tell us a lot 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 sub-families 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 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 analysed 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. In due time this algorithm will undoubtedly be improved. With the presently available data, however, this algorithm is the best we can do, and the results indicate that the detection of recalcitrant residues is very much worth the effort.

Entropy-variability plots can be an automatic data-mining tool, which promises to be quite successful if very large numbers of sufficiently variable sequences and structural information are available. Thus, it is a natural choice for genomic, proteomic, structomic etc., projects, that are producing a flood of data
and need a flexible and automatic data-mining tool in order to provide as quickly as possible a maximum amount of information for multiple purposes.

ACKNOWLEDGMENTS
This work was supported by grants from the São Paulo State Research Foudation (FAPESP), the Brazilian National Research Council (CNPq), Organon and Unilever. The authors thank Florence Horn and David Thomas.

REFERENCES

27. Royer WE Jr, Hendrickson, WA, Chiancone E. Structural transitions upon
Structure of the guanine-nucleotide-binding domain of the Ha-ras oncogene
Refined crystal structure of the triphosphate conformation of H-ras p21 at
1.35 Å resolution: implications for the mechanism of GTP hydrolysis.
30. Takai Y, Sasaki T, Matozak T. Small GTP-binding proteins. Physiol Rev
31. Huang L, Hofer F, Martin GS, Kim SH. Structural basis for the interaction
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
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
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
36. Swissprot/TrEMBL: http://www/ebi.ac.uk/swissprot.
37. PDB: http://www.rcsb.org/pdb.
38. Oliveira L, Paiva AC, Vriend G. A common motif in G protein-coupled
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.


<table>
<thead>
<tr>
<th>Protein family</th>
<th>Groups</th>
<th>Seq</th>
<th>Pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>globin chains</td>
<td>364</td>
<td>753</td>
<td>113</td>
</tr>
<tr>
<td>ras_like proteins</td>
<td>335</td>
<td>562</td>
<td>152</td>
</tr>
<tr>
<td>serine-proteases</td>
<td>176</td>
<td>301</td>
<td>173</td>
</tr>
</tbody>
</table>

**Table I. Family statistics.** The three sequence families. 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.
**Figure 1.**

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 labelled with an asterisk.

<table>
<thead>
<tr>
<th>GROUP</th>
<th>POSITION</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B) S and V stand for entropy and variability. Av indicates the average entropy at each position. Positions refer 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 x P.

<table>
<thead>
<tr>
<th>Position</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Av</th>
<th>P</th>
<th>H-Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.50</td>
<td>1.16</td>
<td>0.39</td>
<td>0.33</td>
<td>0.129</td>
</tr>
<tr>
<td>9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
<td>0.23</td>
<td>0.67</td>
<td>0.154</td>
</tr>
<tr>
<td>41</td>
<td>0.56</td>
<td>0.00</td>
<td>0.67</td>
<td>0.27</td>
<td>0.67</td>
<td>0.181</td>
</tr>
<tr>
<td>42</td>
<td>0.00</td>
<td>0.67</td>
<td>0.13</td>
<td>0.04</td>
<td>0.33</td>
<td>0.013</td>
</tr>
<tr>
<td>43</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
<td>0.04</td>
<td>0.33</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Figure 2.

A)

<table>
<thead>
<tr>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDKTNKAAWGK</td>
<td>HAGDYGAEALERMFLS</td>
<td>FPTTKTYFPF</td>
<td>SAQVKGHGKKVADALTNAVAH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>he A</td>
<td>he B</td>
<td>he C</td>
<td>he E</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALSALSDLHAKLRLV</td>
<td>DPVNFKLShCllLVT</td>
<td>PAVHASLDKFLASVSTVLT SKYR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>he F</td>
<td>he G</td>
<td>he H</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B)

<table>
<thead>
<tr>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>FKLVIVGDSVGKSCllLIRFTDNEF</td>
<td>VDEYVPTIGDDFRKQV</td>
<td>IDGKTIKLQIQWDAGQ</td>
<td>ERYRAIRPAYYRGAM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be 1</td>
<td>he 1</td>
<td>1i 2</td>
<td>be 2</td>
<td>be 3</td>
<td>he 2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFLLYVDITNEKFENKKNW</td>
<td>SDNVPIMLVGKCDLRDRRVV</td>
<td>YGIPFIETS AKTNQNVVEAEFLAREILKMS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be 4</td>
<td>he 3</td>
<td>be 5</td>
<td>he G</td>
<td>be 6</td>
<td>he 5</td>
<td></td>
</tr>
</tbody>
</table>

C)

<table>
<thead>
<tr>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVGGYEAKPGSWPQVSL</td>
<td>SGRHFCCGSLINP</td>
<td>QWVLTAHACLGS</td>
<td>GEHNLSVGEES</td>
<td>VIDVSKIIPHPKYN</td>
<td>SDT</td>
<td></td>
</tr>
<tr>
<td>be 1</td>
<td>be 2</td>
<td>be 3</td>
<td></td>
<td></td>
<td>be 4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDNDIALKLSSPATLSDAVQICLP</td>
<td>PPPGTVCGLSGWGRGSSG</td>
<td>CKSA MLCA GY KDSCQGDSGGPLVCN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>be 5</td>
<td></td>
<td>be 6</td>
<td>he h</td>
<td>be 7</td>
<td>be 8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>160</th>
<th>170</th>
<th>180</th>
</tr>
</thead>
<tbody>
<tr>
<td>GVLQGIVSWGS</td>
<td>GCAQP</td>
<td>PGVYTRVSNYLDWIQSTI</td>
</tr>
<tr>
<td>be 9</td>
<td>be 10</td>
<td>he CT</td>
</tr>
</tbody>
</table>

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), using the commonly used nomenclature. Li indicates a linker region. Secondary structure information was obtained from the PDB11 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.
Fig. 4. Location in the 3D structure of residues in the five boxes of fig 3.
Stereo figures indicating the location of the boxes in the structures. A,B) globins. C,D) ras-like proteins. E,F) serine-
proteases. The PDB files used are the same as in figure 2. Dark blue positions, not indicated by a sphere, were
excluded from the analyses because of insertions or deletions at those positions, or because of being recalcitrant.
A) Globins. 11=red, 12=gold, 22=green; B) Globins. 23=red, 33=gold; C) Ras-like proteins. 11=red, 12=gold,
22=green; D) Ras-like proteins. 23=red, 33=gold; E) Serine proteases. 11=red, 12=gold, 22=green; F) Serine
proteases. 23=red, 33=gold. In A-D the haem and the GDB are shown in purple. In E and F the catalytic triad and
the calcium indicated by a small, light blue spheres.
Fig. 5. H-entropy-entropy plots.
A) globins. B) ras-like proteins. C) serine-proteases. Dashed lines indicate the recalcitrance cut-off values. Note that plots have different vertical scales.
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.