Correlated Mutation Analyses on Very Large Sequence Families

L. Oliveira,[b] A. C. M. Paiva,[b] and G. Vriend*[a]

The 'omics era' (the era of genomics, proteomics, and so forth) is marked by a flood of data that need to be interpreted to become useful information. Thanks to genome sequencing projects, large numbers of sequence families with more than a thousand members each are now available. Novel analytical techniques are needed to deal with this avalanche of sequence data. Sequence entropy is a measure of the information present in an alignment, whereas sequence variability represents the mutational flexibility at a particular position. Entropy versus variability plots can reveal the roles of groups of residues in the overall function of a protein. Such roles can be as part of the main active site, part of a modulator binding site, or transduction of a signal between those sites. Residues that are involved in a common function tend to stay conserved as a group, but when they mutate, they tend to mutate together. Correlated mutation analysis can detect groups of residue positions that show this behaviour. The combination of entropy, variability and correlation is a powerful tool to convert sequence data into useful information. This analysis can, for example, detect the key residues involved in cooperativity in globins, the switch regions in ras-like proteins and the calcium binding and signalling residues in serine proteases. We have extrapolated from these three classes of structurally and functionally well-described proteins to G-protein-coupled receptors (GPCRs). We can detect the residues in the main functional site in GPCRs that are responsible for G-protein coupling, the residues in the endogenous agonist binding site, and the residues in between that transduce the signal to and fro between these sites. The results are discussed in the light of a simple two-step evolutionary model for the development of functional proteins.

KEYWORDS:

G-protein-coupled receptors · molecular evolution · protein structures · sequence entropy · sequence variability

Introduction

Genomic and proteomic projects are producing massive amounts of data and it is likely that 'structomics', that is, structural studies, and metabolomic projects will also soon contribute to the avalanche of data. All these data are shouting for novel computational techniques of data handling, analysis, mining and other steps to convert them into useful information and to answer the questions that initiated the 'omics' projects. The rate of sequence output by genomic projects has grown to about a gigabase per month[1]. The extraction of useful information from the thousands and thousands of sequences that become available every day is presently the largest data analysis challenge of the bioinformatics community.

The rule that has guided most analyses of multiple sequences is that "conserved residues are important". The availability of thousands of sequences for a rapidly increasing number of sequence families has recently prompted us to add a second rule, which states that "very conserved residues are very important".[2] Indeed, most of the literature on multiple sequence alignment analyses concentrates on sequence conservation[3–7] rather than variability. These studies therefore concentrate mainly on the active site and on the distinction between surface and core residues. Schneider and Sander[5] analysed multiple sequence alignments and concluded that high sequence conservation correlates mainly with a functional role and much less with structural requirements. Wood and Pearson[4] analysed 36 protein families for which structures were available for diverse members. They concluded that differences in mutation sensitivity might reflect differences in the nature and extent of interactions between basic elements of secondary structure. Friedberg and Margalit[6] analysed residue conservation in structurally similar but sequence dissimilar proteins. They detected residue positions that are, "mutually persistently conserved (MPC)", that is, they are conserved within families but show differences between families. Mirny and Shakhnovich[3] observed this same phenomenon and called it, "conservation of conservation (CoC)". They conclude this pattern to be the result of evolutionary pressure on protein folding. Evolutionary trace (ET) methods[7–9] can extract functionally important residues by using multiple sequence alignments and phylogenetic trees. MPC, CoC and ET are special cases of correlated mutation analyses (CMA).[10] CMA can be used to convert sequence correlation patterns into three-dimensional structure information,[11] or it can be used to determine functionally important residues.[10, 12–15] A series of early CMA studies has already indicated that conservation of the location of functionally

[a] Prof. Dr. G. Vriend
CMBI, KUN
Toernooiveld 1, P.O. Box 9010
6500 GL Nijmegen (The Netherlands)
Fax: (+31) 24-3652-977
E-mail: vriend@cmbi.kun.nl

[b] Dr. L. Oliveira, Prof. Dr. A. C. M. Paiva
Escola Paulista de Medicina
UNIFESP, Sao Paulo (Brazil)

important residues is more important than conservation of the residue types at those positions.[10, 14] The use of the CMA technique for the detection of functionally important residues is based entirely on the rule that conserved residues are important and the fact that the location of function is more conserved than the residues performing that function. All these studies deal mainly with sequence conservation. Sequence variability has remained much less well studied.

Sequence Entropy and Variability

In two previous articles[2, 16] we introduced multiple sequence alignment analysis methods that can be used when many (hundreds or thousands) sufficiently variable sequences are aligned. The sequence entropy (S_p) at position p in a multiple sequence alignment is defined by Equation (1), in which i loops over the 20 amino acid types, and f_{pi} is the relative frequency of residue type i at alignment position p. The sequence variability (V_p) at position p is defined as the number of different residue types observed at position p in at least 0.5% of all sequences. Loosely speaking, S_p can be seen as a measure of information, and V_p as a measure of the freedom or chaos at position p in a multiple sequence alignment.

$$S_{\mathrm{p}} = -\Sigma_{i=1,20} f_{pi} \times \ln(f_{pi}) \qquad (1)$$

Alignment Quality Considerations

Two major problems are associated with the analysis of variability patterns in multiple sequence alignments. First, the alignments must be correct because wrong alignments clearly give rise to wrong ideas about the variability. Second, it must be clear whether lack of variability at certain alignment positions is a genuine characteristic resulting from evolutionary pressure, or the result of the use of too few or insufficiently divergent sequences.

To tackle the first problem, we use an iterative, profile-based alignment procedure that incorporates structure-superposition-derived alignments whenever possible. This iterative approach also helps to overcome the limitations of most publicly available sequence alignment programs, which cannot align thousands of sequences.

The second question is whether conservation is real or not. This becomes less of a problem when many sequences can be aligned. The conservation pattern of the MPC residue positions that was detected by Friedberg and Margalit,[6] for example, suggests a functionally important role for these sites. These residues are different between remote homologues (see ref. [17] for a definition) therefore their functional role is not as part of the main active site but is one that can or must be different between remote homologues, such as modulator binding, membrane insertion, multimerisation interface, and so on. This functional importance is inferred from the fact that conservation is observed at the same position in many or all homologous families. Therefore, if the conservation is not observed at this position in a series of families, a functional role cannot be inferred. If sequences are available for only a few families, this kind of conservation can easily be observed by chance alone but when many sequences from hundreds of families are aligned, a lack of functional importance will invariably lead to the observation of variability within at least some families. We introduced the concept of H entropy in a previous article[2] in order to define this condition and we provided a semiquantitative recipe for the automatic detection of such residue positions. These positions were called recalcitrant because analysis of an insufficiently large number of sequences will lead to incorrect conclusions about the functional role of certain residue positions. Recalcitrant residue positions are excluded from all analyses described here.

Entropy versus Variability Plots

Entropy versus variability plots can be divided into five boxes of residue positions that show a remarkable correlation between the box number and the function of the residues. These boxes are labelled 11, 12, 22, 23 and 33 in Figure 1. The following conclusions were drawn about the residue positions in these five boxes:

- Box 11 contains residue positions with low entropy and low variability that form the main functional site. These residue positions are involved in catalysis or signalling mechanisms. Some key conserved residues with a structural role (for example, Cys–Cys bridges) are found in this box too.
- Box 12 contains residues in the protein core. They are spatially close to the residues in box 11. They form the first shell of residues around the main functional site.
- Box 22 contains mainly core residues. The residues in box 22 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.
- Box 23 contains most residues involved in modulator or effector interactions.
- Box 33 contains residues 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.

The entropy–variability plots were analysed for four sequence families: globins, ras-like proteins, serine proteases, and G-protein-coupled receptors (GPCRs). We observed that signals are transduced from modulator sites to the main active site (or vice versa) along the path, modulator = box 23 → box 22 → box 12 → box 11 = main activity, in all four families.

Correlated Mutation Analyses

In this study we used CMA to refine the signalling path given above. The simple two-step evolutionary process suggested for the four aforementioned protein families is supported by this

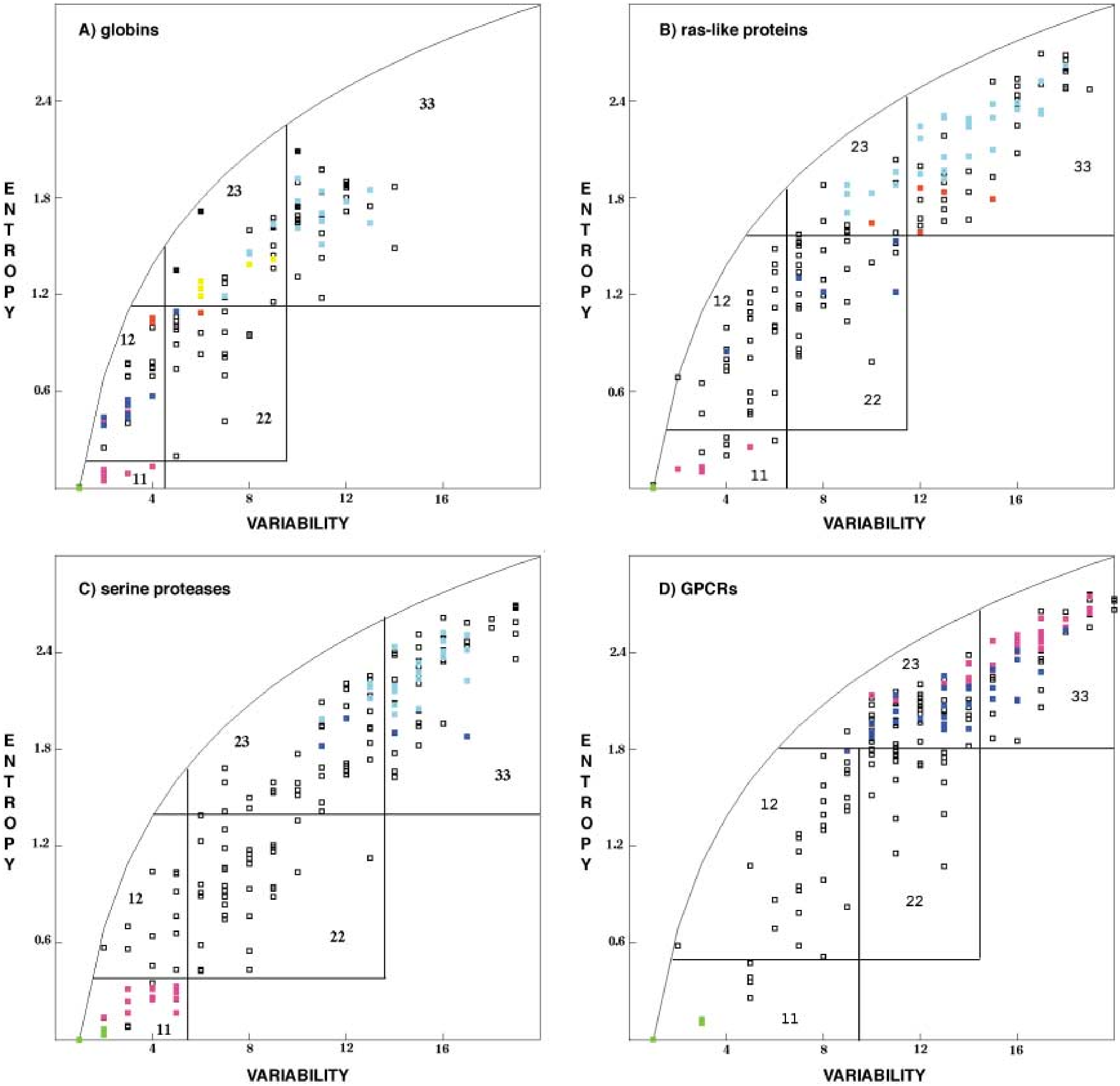

Figure 1. Correlated networks mapped on entropy – variability plots. A) Globins: green, network 1; pink, network 2; dark blue, network 3; red, network 4; yellow, network 5; black, network 6; light blue, network 7. B) Ras-like proteins: green, network 1; pink, network 2; dark blue, network 3; red, network 4; light blue, network 5. C) Serine proteases: green, network 1; pink, network 2; dark blue, network 3; light blue, network 4. D) GPCRs: green, network 1; pink, network 2; dark blue, network 3.

analysis. An explanation for the oft-observed conservation of the location of function is proposed.

Results and Discussion

A comparison of thousands of protein structures[18–20] leads one to observe that the location of function is much more conserved than the function itself.[3, 4, 6] It is beyond the scope of this article to analyse this phenomenon exhaustively. However, a short qualitative discussion of the entropy/variability/correlation analysis approach is given and a potential answer is proposed for the question of why the location of function is so conserved.

A two-step evolutionary model

We arrived at a two-step process for the evolution of functional proteins by using Occam's razor.[2] In the first step, the protein acquires its main function, which initially is uncontrolled and probably constitutive. Constitutive activity of any kind is unlikely to be advantageous so a regulatory mechanism must evolve quickly. The protein acquires such a regulation mechanism, which normally involves the binding of modulators, in the second step. The requirement that this modulation evolve quickly on an evolutionary time-scale makes the modulator site appear at a location where mutations can take place that have major consequences neither for the main function nor for the structural integrity of the protein. The best location for a modulator site is therefore at the surface of the protein. Clearly, the modulator site must be some distance from the main active site to avoid steric hindrance. This requirement is complicated by the fact that communication must take place between the sites. This signal transduction must proceed through the protein core as a result of the physical separation of these sites. The two steps of our simple evolutionary model are easy to imagine but we have no idea how the core evolves into a signal funnel that can transfer signals to and fro between a conserved main functional site and a variable modulator site.

Conservation of the location of function

The entropy versus variability analyses illustrate that there are layers of residue positions involved in a similar task. Box 11 in the entropy – variability plot contains the fully conserved active-site residue positions. Box 12 contains slightly less conserved residue positions that form a layer around the active site. Box 22 contains the intermediately conserved/variable core residue positions located between the residues of box 12 and the modulator-site residues of box 23. Box 23 contains the highly variable modulator interaction positions. The existence of this information funnel could explain why the location of function is much more conserved than any detail of the function itself. The generic functional protein possesses a main functional site and a modulator site. Naturally, both sites can mutate during evolution, but the evolutionary pressure at the level of protein function will make certain mutations more likely to survive than others. In the unlikely event that the main functional site mutate to acquire another function, the residues in boxes 11, 12 and to a lesser extent 22, would have to mutate too. This would make any remaining sequence similarity undetectable and the newly evolved protein would not be recognised as a member of the sequence family from which it evolved. On the other hand, mutations in the modulator site can lead to the acquisition of the capability to bind another modulator. These mutations involve residue positions in box 23. In order to make this mutated modulator site communicate with the main functional site, some residue positions in box 22 will have to mutate to communicate properly with the box-23 residue positions, and residues in box 12 might occasionally need to mutate in order to transduce signals to and fro between residue positions in boxes 11 and 22. It seems simpler for an evolving functional protein to acquire new modulator functionalities at an existing modulator site than at a new site because the box-23 residues of an existing modulator site are already correctly located for communication with the box-22 residues. A new location for a novel modulator would require an entirely new signalling path to evolve. This reasoning combined with the simple two-step model for the evolution of functional proteins might explain why the location of function is much more conserved than the function itself. This theory almost certainly holds for the evolution of modulator sites, but may also hold for the evolution of the main functional site. Application of these arguments to the main functional site might even provide and explanation for the fact that all active sites in proteins with a similar fold are located at the same location (for example, the active site of all TIM-barrel molecules is located at the C-terminal side of the central barrel).

Even though our hypothesis only includes one generic functional protein, it must be clear that the evolution of a protein cannot be considered separately from the environment that provides the pressure that makes one version of the protein more likely to survive than another. This hypothesis is independent of any discussion about convergent or divergent evolution. We argue only in terms of the relation between protein structure and characteristics of the variability patterns in a multiple sequence alignment, and analyse this relation in terms of the amount of evolutionary protein engineering that is needed to achieve that relation. Time may tell if our hypothesis is right or wrong; it remains a fact that the location of function is highly conserved. Proteins with a similar structure tend to have both their main active site and their modulator binding site(s) at similar positions.

Correlation analysis

It can be seen from the example in the Methods Section that a high correlation score between two residue positions can mean one of three things:

- Category 1: both residue positions are entirely conserved
- Category 2: both residue positions are variable and have the same variability and the same entropy
- Category 3: the mutational behaviour of both residue positions shows a perfect correlation, that is, the residues at both positions stay conserved in tandem or mutate in tandem.

Category 1 is a special case of category 3 and when very large numbers of sequences are aligned category 2 also becomes a special case of category 3. It can easily be proven that a correlation score of 1.0 can only be observed between two residue positions that have the same entropy and the same variability. Correlated mutation analyses were carried out on the same four classes of molecules as were used for variability/entropy analyses in two previous studies:[2, 16] globins, ras-like molecules, serine proteases and GPCRs. Figure 1 shows the entropy versus variability plots for these four sequence families. Colours indicate groups of residue positions with a high correlation score. A more detailed description of these networks is presented in the websites of previous studies.[2, 16] Figure 2 shows the location of these residue positions in the three-dimensional structures of the proteins. The networks are indicated in the consensus sequences in Figure 3.

In globins we can observe seven networks with a correlation score higher than 0.8. Ras-like proteins have five such networks, serine proteases four and GPCRs three. These networks were evaluated in the light of the three-dimensional protein structures and the compiled knowledge about the functional role of

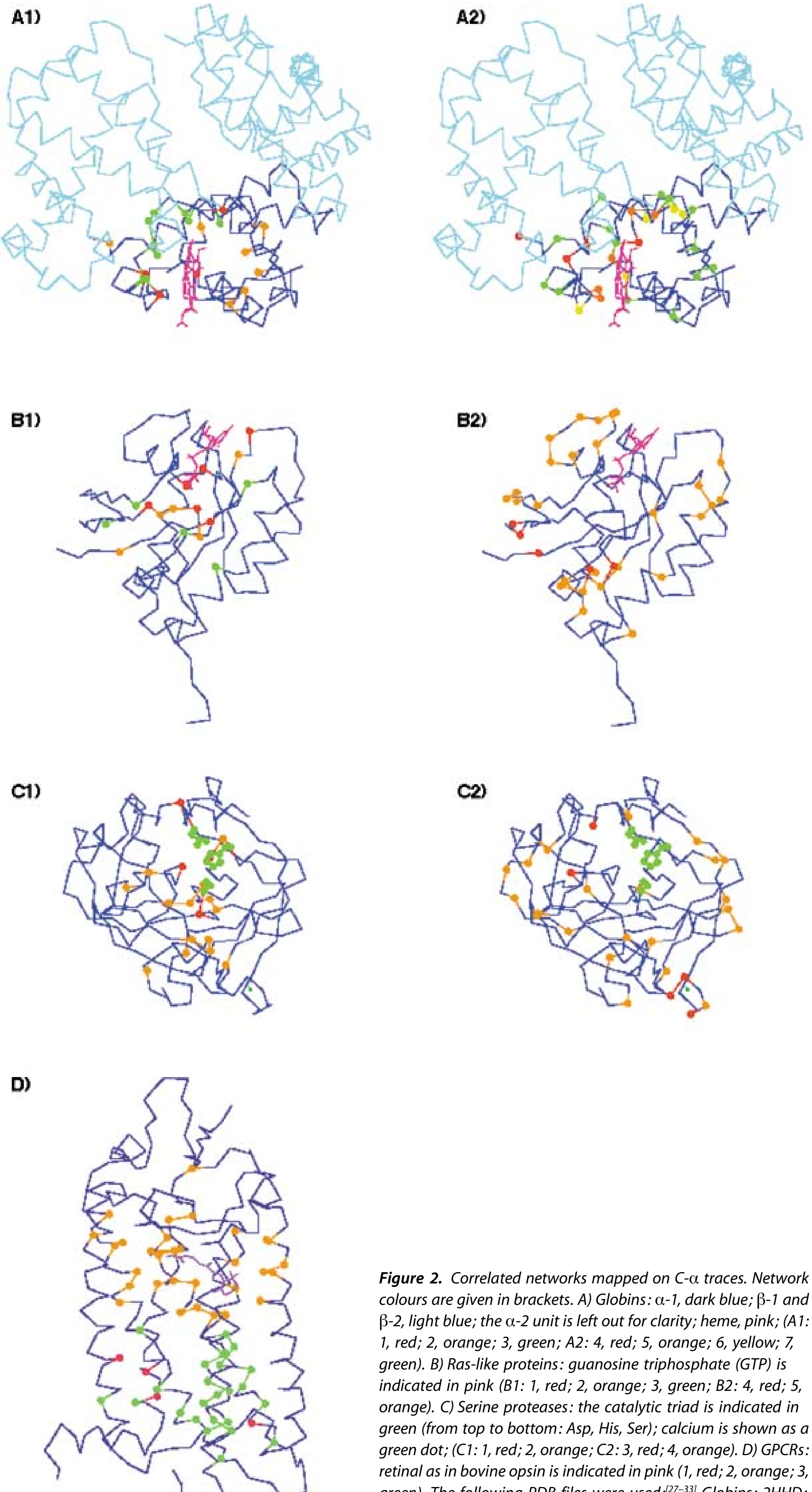

Figure 2. Correlated networks mapped on C-α traces. Network colours are given in brackets. A) Globins: α-1, dark blue; β-1 and β-2, light blue; the α-2 unit is left out for clarity; heme, pink; (A1: 1, red; 2, orange; 3, green; A2: 4, red; 5, orange; 6, yellow; 7, green). B) Ras-like proteins: guanosine triphosphate (GTP) is indicated in pink (B1: 1, red; 2, orange; 3, green; B2: 4, red; 5, orange). C) Serine proteases: the catalytic triad is indicated in green (from top to bottom: Asp, His, Ser); calcium is shown as a green dot; (C1: 1, red; 2, orange; C2: 3, red; 4, orange). D) GPCRs: retinal as in bovine opsin is indicated in pink (1, red; 2, orange; 3, green). The following PDB files were used:[27–33] Globins: 2HHD; ras-like proteins: 5P21; Serine proteases: 2PTC; GPCRs: 1F88.

individual residues. The detailed results are available from the GPCR database[21]. Residues in networks 1 and 2 in globins are involved in haem binding. Residues in networks 3, 4 and 5 are at or near the α1 – β2 interface and many of them are known to play a role in the cooperativity of globins. The function of network 6 is unclear. Network 7 seems to be involved in α1 – β1 interactions but the relation between residue position and cooperativity is less clear at the α1 – β1 interface than at the α1 – β2 interface. In the ras-like proteins network 1 is involved in GTP binding. The other networks play a role in the switch regions. In serine proteases network 1 contains the active-site residues. The function of network 2 is unclear, network 3 forms the calcium binding site and network 4 contains residues involved in calcium binding and in communication between the calcium site and the rest of the molecule. In GPCRs, the first network consists of the most conserved residue positions involved in G-protein coupling. Network 2 is involved in ligand binding and network 3 is formed by residues that are involved in G-protein coupling and activation. We do not find networks that consist mainly of core residues. The networks are mapped on the entropy versus variability plots in Figure 1. These plots illustrate even better than the three-dimensional coordinates that two types of networks exist. One type is related to the main activity (box 11 and 12) and the other type is related to modulator interaction (box 23 and 33). These two kinds of networks are easy to understand. Networks related to the main function give a high correlation score because they are highly conserved. The networks related to the modulator binding show a high correlation score because all modula-

A)

```
                    727  272     152475167   7   22              55 6137 1
AADKTNVKGIFAK   HAEEYGAEALDRMFTTYPQTKTYFPH SAQIKGHGKKVAAALVEAANH TLSKLSDLHA
_____________    _______________________   _____________________ __________
      he A          he B        he C                he E           he F

7 34333733  3576 7          67 6    7 43 7 24
QKLRVDPVNFKLLGQCFLVV  PEVHASLDKFLCAVGTVLTAKYR
   _______________    ____________________
        he G                   he H
```

B)

```
42    1    11    5  55 5  555 5   3 5 555     4  4 3  1221 23 45 455     5
FKLVIVGDSGVGKSCLLIRFTDNEF VDEYVPTIGDDFRKQV IDGKTIKLQIWDTAGQERYRAIRPAYYRGAM
__________  ___________   ________ _______    ___________      ___________
 be 1            he 1      li 2      be 2         be 3            he2

      3  555      53  5   5    2  1   5 5       5                   5 5 5
GFLLVYDITNEKSFENVKNW SDNVPIMLVGNKCDLRDRRVV YGIPFIETSAKTNQNVEEAFETLAREILKKMS
_______  __________  ________ ____________  _________      _____________
be 4       he 3        be 5     he G          be 6           he 5
```

C)

```
   2       42      44    22   4       121  4  24 33 4 3    4 4 4
IVGGYEAKPGSWPWQVSL SGRHFCGGSLINP QWVLTAAHCLGS GEHNLSVGEE VIDVSKIIPHPKYN SDT
            ______     ______    ____                    ______________
              be1      be2        be3                       be4

4  1           4    4  4   4 4 4  4   222         4      4       11222   4
LDNDIALLKLSSPATLSDAVQPICLP PPPGTVCLVSGWGRTSSGG CKSA MLCAGY KDSCQGDSGGPLVCN
     _____                      ________       ____ ______          ______
      be5                         be6          he h be 7             be 8

   32  1          2         3 1
GVLQGIVSWGS GCAQP PGVYTRVSNYLDWIQSTI
  ________          _______  _______
    be 9            be 10     he CT
```

D)

```
      2             1    3       3 3  3          2 22  22  2
ELQVLLFVLYLLIFLVGLLGNLLVILVIL    PMNFFLLNLAVADLLFSLTLPPKMLVNL
-----------------------------    ----------------------------
         Helix I                       Helix II

  2 22  22 222  33  33  3 1 3      33  33  3      2
GCKLQLFLFYLNGYASIFLLTAIAVDRYLAIC PRRAKLLIALIWVLSLLLS
-------------------------------- -------------------
         Helix III                     Helix IV

22  222  2                       3  33 33  33   2      2  2
AYVIVLFILGFLLPLLIILVSYG          ERKAIKSLLGVVKAFSTCWLPYHVVLLF
-----------------------          ----------------------------
         Helix V                       Helix VI

 2   22  22  2   3  1   13
SPDLAKLISLFLAYVNSCLNPII YA
-----------------------
         Helix VII
```

Figure 3. Consensus sequences of the four structure classes. A) Globins. B) Ras-like proteins. C) Serine proteases. D) GPCRs. The secondary structure elements are indicated according to the commonly used nomenclature for these four molecular classes. The numbers above the sequences indicate the network of correlated residue positions to which the residue beneath belongs.

tors bind to (a large subset of) the same group of residues. These residues are conserved within families of proteins that bind the same modulator but are different for families that bind different modulators. There are no networks related to signal transduction between the main function and the modulator binding sites. This result means that either there is no stringent evolutionary pressure on these core residues or the evolutionary pressure to transduce a signal between two sites while at the same time maintaining the structural integrity of the protein is too complex to be detected by CMA.

Conclusions

Residue positions with low variability and low entropy (boxes 11 and 12) are involved in the main activity of the protein. Residue positions with high entropy and medium variability (mainly box 23, but also a bit of box 33) can be involved in modulator binding. Both functionalities have multiple aspects. For example, the modulator has to be bound and the signal must then be generated and transduced. CMA can add detail to the entropy versus variability plot by revealing which groups of residue positions are involved in the same aspect of the main functional site or the modulator site. CMA does not show the function of certain residue positions, but it sorts residues into groups that are under similar evolutionary pressures and thereby provides an invaluable addition to entropy – variability analyses aimed at elucidation of the role of each amino acid in a protein.

Methods

Sequences and alignment: Sequences were obtained from GenBank,[22] TrEMBL,[23] and the GPCRDB[21]. Three-dimensional protein coordinates were obtained from the PDB.[24] Multiple sequence alignments were performed by using the WHAT IF program,[25] as described before.[2, 10] Lee et al. recently reviewed multiple sequence alignment programs that can align thousands of sequences.[26] They concluded that progressive multiple sequence alignment methods are fast but at the price of a loss of accuracy and gap scoring artefacts. Our alignment procedure is not meant to be automated but relies on manual inspection of three-dimensional structures and the incorporation of other forms of experimentally obtained information. Additionally, we concentrate only on that part of the fold shared by all proteins in the alignment and neglect areas where insertions and deletions occur.

Correlated mutation analysis: Correlated mutation analyses have been used to answer questions about the role of individual residues in protein structure or protein function. The question asked strongly influences how CMA should be used. Valine and isoleucine, serine and threonine, and so on, are very similar pairs of residues for studies of structural aspects. However, these residue pairs are very different when functional aspects are considered. Consequently, all structure-related CMA techniques use Dayhof-style residue similarity matrices.[11] When functional aspects of proteins are studied by CMA techniques, the identity matrix is normally used. We wanted to extract information about the functional role of residue positions from multiple sequence alignments and thus used an identity matrix for residue similarities; residues are either identical or they are different, but there is nothing in between. The correlation score between residue positions i and j is calculated by using the expression:

$$C(i,j) \;=\; W(i,j)\sum_{p=1}^{n-1}\sum_{q=p+1}^{n}\delta(i_p,i_q,j_p,j_q)$$

in which n is the number of sequences. p and q run over all pairs of positions in the sequences i and j. i_p, i_q, j_p and j_q form the residue pairs (p,q) in the sequences i and j, respectively. The delta function is 1 when $i_p = j_p$ and $i_q = j_q$ or when $i_p \mp j_p$ and $i_q \mp j_q$. The delta function is 0 when $i_p = j_p$ and $i_q \mp j_q$ or when $i_p \mp j_p$ and $i_q = j_q$. The weight factor $W(i,j)$ for the sequence pair i,j is given by:

$$W(i,j) \;=\; w_i \times w_j / \sum_{k=1}^{n-1}\sum_{l=k+1}^{n} w_k \times w_l$$

in which w_i and w_j are the sequence weight factors, as described previously.[2] Table 1 shows an example of four sequences of seven residues each that illustrates this scoring scheme. For simplicity, all sequence weights are kept at 1.0 in this example. The correlation score $C(1,2)$ between residue positions 1 and 2 in Table 1, for example, is given by:

Table 1. Example to explain correlation scores.[a]

Sequence	Residue position	i,j	$C(i,j)$	i,j	$C(i,j)$
	1234567	1,2	0.84	1,7	1.00
1	ASPWLRL	1,3	0.33	3,4	1.00
2	ASPWIDL	1,4	0.33	5,6	1.00
3	TTPWVKG	1,5	0.67	6,7	0.67
4	TGPWMEG	1,6	0.67	2,3	0.16

[a] Four hypothetical sequences of seven residues each are shown with the correlation scores $C(i,j)$ for 10 selected pairs of residue positions.

$$C(1,2) = \frac{2}{n \times (n-1)}\sum_{p=1}^{n-1}\sum_{q=p+1}^{n}\delta(1_p1_q,2_p,2_q)$$

$$= 1/6 \times (\delta(\mathrm{ASAS}) + \delta(\mathrm{ASTT}) + \delta(\mathrm{ASTG}) + \delta(\mathrm{ASTT}) + \delta(\mathrm{ASTG}) + \delta(\mathrm{TTTG})) = 1/6 \times (1+1+1+1+1+0) = 0.84$$

Similarly,

$$C(2,3) \;=\; 1/6 \times (\delta(\mathrm{SPSP}) + \delta(\mathrm{SPTP}) + \delta(\mathrm{SPGP}) + \delta(\mathrm{SPTP}) + \delta(\mathrm{SPGP}) + \delta(\mathrm{TPGP})) \;=\; 1/6 \times (1+0+0+0+0+0) \;=\; 0.16$$

A group of residue positions is called a network if all pairwise correlation scores are above a certain cut-off value. Typical cut-off values are in the range 0.7 – 0.9.

Recalcitrance: Residue positions that show limited variability within a subset of the subfamilies represented in the multiple sequence alignment are called recalcitrant. A precise definition of recalcitrance is given elsewhere.[2] Recalcitrant residue positions were not used in any of the analyses. These residues often show a high CMA score but are nevertheless not functionally important because a subset of the families shows some variability at that position, which indicates that the conservation is not the result of evolutionary pressure to keep a functional role.

Additional material: Much data too voluminous to put in print (multiple sequence alignments, tabulated variability and entropy values, the results of the CMA, file names, previous articles and articles submitted only recently, etc.) are available from the GPCRDB[21].

This work was supported by grants from the Sao Paulo State Research Foundation (FAPESP), the Brazilian National Research Council (CNPq), Organon and Unilever. The authors thank David J. Thomas for critically reading the manuscript.

[1] EMBL Nucleotide Sequence Database: http://www.ebi.ac.uk/embl/

[2] L. Oliveira, P. B. Paiva, A. C. M. Paiva, G. Vriend, *Proteins* **2002**, submitted; http://www.cmbi.kun.nl/7tm/articles/2002_1/index.html

[3] L. A. Mirny, E. I. Shakhnovich, *J. Mol. Biol.* **1999**, *291*, 177–196.

[4] T. C. Wood, W. R. Pearson, *J. Mol. Biol.* **1999**, *291*, 977–995.

[5] C. Sander, R. Schneider, *Nucleic Acids Res.* **1993**, *21*, 3105–3109.

[6] I. Friedberg, H. Margalit, *Protein Sci.* **2002**, *11*, 350–360.

[7] S. Madabushi, H. Yao, M. Marsh, D. M. Kristensen, A. Philippi, M. E.Sowa, O. Lichtarge, *J. Mol. Biol.* **2002**, *316*, 139–154.

[8] O. Lichtarge, H. R. Bourne, F. E. Cohen, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 7505–7511.

[9] O. Lichtarge, H. R. Bourne, F. E. Cohen, *J. Mol. Biol.* **1996**, *257*, 342–358.

[10] L. Oliveira, A. C. M. Paiva, G. Vriend, *J. Comput.-Aided Mol. Des.* **1993**, *7*, 649–658.

[11] U. Goebel, C. Sander, R. Schneider, A. Valencia, *Proteins* **1994**, *18*, 309–317.

[12] M. Singer, G. Shepherd, G. Vriend, *Chem. Senses* **1995**, *6*, 277–278.

[13] M. Singer, L. Oliveira, G. Vriend, G. Shepherd, *Recept. Channels* **1995**, *3*, 89–95.

[14] W. Kuipers, L. Oliveira, G. Vriend, A. P. Ijzerman, *Recept. Channels* **1997**, *5*, 159–174.

[15] F. Horn, R. Bywater, G. Krause, W. Kuipers, L. Oliveira, A. C. M. Paiva, C. Sander, G. Vriend, *Recept. Channels* **1998**, *5*, 305–314.

[16] L. Oliveira, P. B. Paiva, A. C. M. Paiva, G. Vriend, *Proteins* **2002**, submitted; http://www.cmbi.kun.nl/7tm/articles/2002_2/index.html

[17] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **1995**, *247*, 536–540.

[18] L. Conte, B. Ailey, T. J. Hubbard, S. E. Brenner, A. G. Murzin, C. Chothia, *Nucleic Acids Res.* **2000**, *28*, 257–259.

[19] S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, L. Holm, *Nucleic Acids Res.* **2001**, *29*, 55–57.

[20] J. E. Bray, A. E. Todd, F. M. Pearl, J. M. Thornton, C. A. Orengo, *Protein Eng.* **2000**, *13*, 153–165.

[21] a) GPCRDB web site: http://www.gpcr.org/articles/; b) F. Horn, J. Weare, M. W. Beukers, S. Horsch, A. Bairoch, W. Chen, O. Edvardsen, F. Campagne, G. Vriend, *Nucleic Acids Res.* **1998**, *26*, 275–279.

[22] GeneBank: http://www.ncbi.nlm.nih.gov/entrez/

[23] SwissProt/TrEMBL: http://www.ebi.ac.uk/swissprot/

[24] PDB: http://www.rcsb.org/pdb/

[25] G. Vriend, *J. Mol. Graphics* **1990**, *8*, 52–56.

[26] C. Lee, C. Grasso, M. F. Sharlow, *Bioinformatics* **2002**, *18*, 452–464.

[27] T. A. Nakayama, H. G. Khorana, *J. Biol. Chem.* **1991**, *266*, 4269–4275.

[28] J. C. Burnett, P. Botti, D. J. Abraham, G. E. Kellog, *Proteins* **2001**, *42*, 355–377.

[29] H. C. Watson, J. C. Kendrew, *Nature* **1966**, *209*, 339–341.

[30] E. F. Pai, W. Kabsch, U. Krengel, K. C. Holmes, J. John, A. Wittinghofer, *Nature* **1989**, *341*, 209–214.

[31] K. D. Cobett, T. Alber, *Trends Biochem. Sci.* **2001**, *26*, 710–716.

[32] A. Ruhlmann, D. Kukla, P. Schwager, K. Bartels, R. Huber, *J. Mol. Biol.* **1973**, *77*, 417–436.

[33] W. Bode, P. Schwager, *FEBS Lett.* **1975**, *56*, 139–143.

Received: June 3, 2002 [F 429]