BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i130–i137
DOI: 10.1093/bioinformatics/btg1017

Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins

*Andreas Heger and Liisa Holm**

Institute of Biotechnology, P.O. Box 56, 00014 University of Helsinki, Finland

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Evolutionary comparison leads to efficient functional characterisation of hypothetical proteins. Here, our goal is to map specific sequence patterns to putative functional classes. The evolutionary signal stands out most clearly in a maximally diverse set of homologues. This diversity, however, leads to a number of technical difficulties. The targeted patterns—as gleaned from structure comparisons—are too sparse for statistically significant signals of sequence similarity and accurate multiple sequence alignment.

Results: We address this problem by a fuzzy alignment model, which probabilistically assigns residues to structurally equivalent positions (attributes) of the proteins. We then apply multivariate analysis to the 'attributes x proteins' matrix. The dimensionality of the space is reduced using non-negative matrix factorization. The method is general, fully automatic and works without assumptions about pattern density, minimum support, explicit multiple alignments, phylogenetic trees, etc. We demonstrate the discovery of biologically meaningful patterns in an extremely diverse superfamily related to urease.

Contact: Liisa.Holm@Helsinki.fi

ABBREVIATIONS

ADA, adenosine deaminase; Blast, basic local alignment search tool; CDA, cytosine deaminase; DHO, dihydroorotase; ICA, independent component analysis; NMF, non-negative matrix factorization; PCA, principal components analysis; PSI-Blast, position-specific iterated Blast; PTE, phosphotriesterase; URE, urease

INTRODUCTION

It is a common observation that functional residues in proteins are conserved in evolution above the background rate of sequence divergence. Therefore sequence motifs are an attractive option to characterize the function of proteins computationally where this has not yet been done experimentally. Conserved patterns suggest positive selection, and it is natural to assume that they play a role in the respective proteins' functional role and/or evolutionary relationships.

Here, we study sequence-derived patterns in the urease superfamily (Holm and Sander, 1997). The urease superfamily is a challenge to many bioinformatics analysis methods because it unifies a large number of sequence families which have low sequence similarity (for example, Blast detects similarity only between 11% of all pairs in the superfamily). Structure alignment has revealed invariantly conserved residues in the active site. The signature motif of the active site involves residues at the ends of beta-strands 1, 5, 6 and 8 of a $(\text{beta/alpha})_8$-barrel fold. Unfortunately, available sequence alignment and motif discovery programs fail to deliver correct alignments of the complete active site between distantly related families in the superfamily (Heger *et al.*, 2003). A key difficulty is that the signature pattern is very sparse and embedded in high entropy sequence segments.

Efficient algorithms have been developed for the discovery and complete enumeration of patterns in unaligned sequences (Brazma *et al.*, 1995; Rigoutsos and Floratos, 1998), but disregarding alignments throws away important information. For example, the $H.H.*H.*H.*D$ signature pattern is not discriminative in sequence database searches (almost any histidine-rich protein will match). However, the significance of pattern instances in the urease superfamily is obvious in structural superimposition, since all residues come together in 3D. The classical definition of patterns is in the context of strings based on a small alphabet (e.g. 20 amino acids). Here, we incorporate the context of a protein fold to our patterns so that, for example, each histidine is distinct in the urease signature. This effectively multiplies the size of the alphabet by the number of structurally equivalent positions, and transforms the problem to one of multivariate analysis.

Proteins are represented in a generalized sequence space, where the attributes are putative structurally equivalent positions. Structural equivalence would normally be defined as columns in a multiple alignment (Casari *et al.*, 1995; Madabushi *et al.*, 2002), but in the present case we cannot determine a reliable multiple alignment. Therefore, we use the idea of 'fuzzy clustering' of residues to deal

*To whom correspondence should be addressed.

with these uncertainties. Sequence-derived alignments are more reliable the closer the sequence distance. Distantly related proteins can be aligned using many closely spaced intermediates as stepping stones. This is the idea of transitive alignment. We use a library of statistically significant pairwise alignments to derive transitive alignments between distant relatives. Taken together, all transitive alignments generate a filtered solution space of possible multiple alignments. The frequency at which a given residue pair occurs in the transitive alignments is used as an estimate for the probability of their structural equivalence.

We reduce the dimensionality of the generalized sequence space by matrix factorization. Matrix factorization yields a decomposition of an $m \times n$-dimensional data matrix V, which is approximately reproduced as the product of two matrices $W * H$ of lower rank r and dimensions $n * r$ and $r * m$, respectively:

$$V \approx W * H. \quad (1)$$

The r columns of W are the basis vectors of the reduced space and H is the encoding in the new basis. In the present application, we get partitions of the protein set corresponding to a given basis of attribute vectors. The above matrix factorization can be achieved by a number of classical multivariate analysis techniques, such as vector quantization or principal components analysis (PCA). PCA imposes the constraint that the basis vectors are orthogonal. ICA (Hyvärinen , 1999) is a method of factor rotation which can be applied on top of PCA so that the transformed basis vectors are statistically independent. In PCA and ICA, the original data is reconstructed by linear combinations of the basis vectors, which involves complex cancellation of positive and negative coefficients of the attributes. NMF (Lee and Seung, 1999) imposes the constraint of non-negative coefficients instead of the orthogonality. This leads to non-subtractive combinations and a more intuitive interpretation of NMF decompositions.

In our application, the coefficients of the attributes in a basis vector reflect the frequency of particular residues in the corresponding protein set. To focus attention on a small set of attributes that are the most important determinants of the partitions, we cluster attributes in the reduced space. The metric used for clustering attributes is the covariance of attribute coefficients in the basis vectors.

The resulting clusters of attributes represent conserved sequence patterns. We validate the biological significance of the found sequence patterns by mapping the residues onto known 3D structures.

METHODS

Data set

The urease superfamily was discovered based on structure comparisons of urease, phosphotriesterase and adenosine deaminase (Holm and Sander, 1997). Based on systematic all-against-all PSI-Blast (Schaffer *et al.*, 2001) searches, we have identified 26 main branches (families) in the phylogenetic tree of the urease superfamily comprising 347 representatives at 40% identity level. A number of recent structure determinations of proteins representing different parts of the urease superfamily have verified earlier blind fold predictions.

Definition of attribute positions

In order to compare residues in different proteins with each other, we need to define equivalence classes between residues. The members of the urease superfamily are so diverse that it is difficult to generate a correct multiple alignment (Heger *et al.*, 2003). For example, PSI-Blast generates alignments between less than half of all pairs in the superfamily. However, all proteins are at least indirect PSI-Blast neighbours of each other and thus connected by transitive alignments.

The library of pairwise alignments induces an alignment trace graph where the nodes are residues and there is an edge between residues that have been aligned by PSI-Blast. We perform a greedy hierarchical clustering of the residues as described in Heger *et al.* (2003). Each cluster contains at most one residue from one protein so that there are dense connections within clusters. Residues in the same cluster are tentatively presumed to be structurally equivalent. These clusters are here called attribute positions. These positions are given arbitrary but unique labels. There were 2344 positions which were further subdivided by amino acid type (maximally 20 sub-clusters) to yield the attributes of our sequence space representation. This greedy step yields a many-to-one mapping of residues to cluster labels. One can think of these clusters as representing columns of a hypothetical multiple alignment (however, residues are not in a consistent sequential order between proteins).

Assignment of residues to attributes

To account for uncertainty (errors, in plain speak) in the above greedy clustering, we wanted to spread out the assignment of any given residue over several attribute positions (hypothetical columns). We use the idea that one can trace a transitive alignment between any two proteins in a set of connected neighbours. But there are very many alternative paths of transitive alignment in our data set. We consider all possible transitive paths to determine the statistical preferences for pairing one residue in protein A with another residue in protein B. This generates a dot matrix commonly used in sequence alignment. We know that higher dot scores correlate with more reliable alignment (Heger *et al.*, 2003), and arbitrarily excluded dots with scores lower than 40 (on a scale 0–100); this retained fewer than ten dots per row in a typical protein pair (cf. Fig. 5 in Heger *et al.*, 2003).

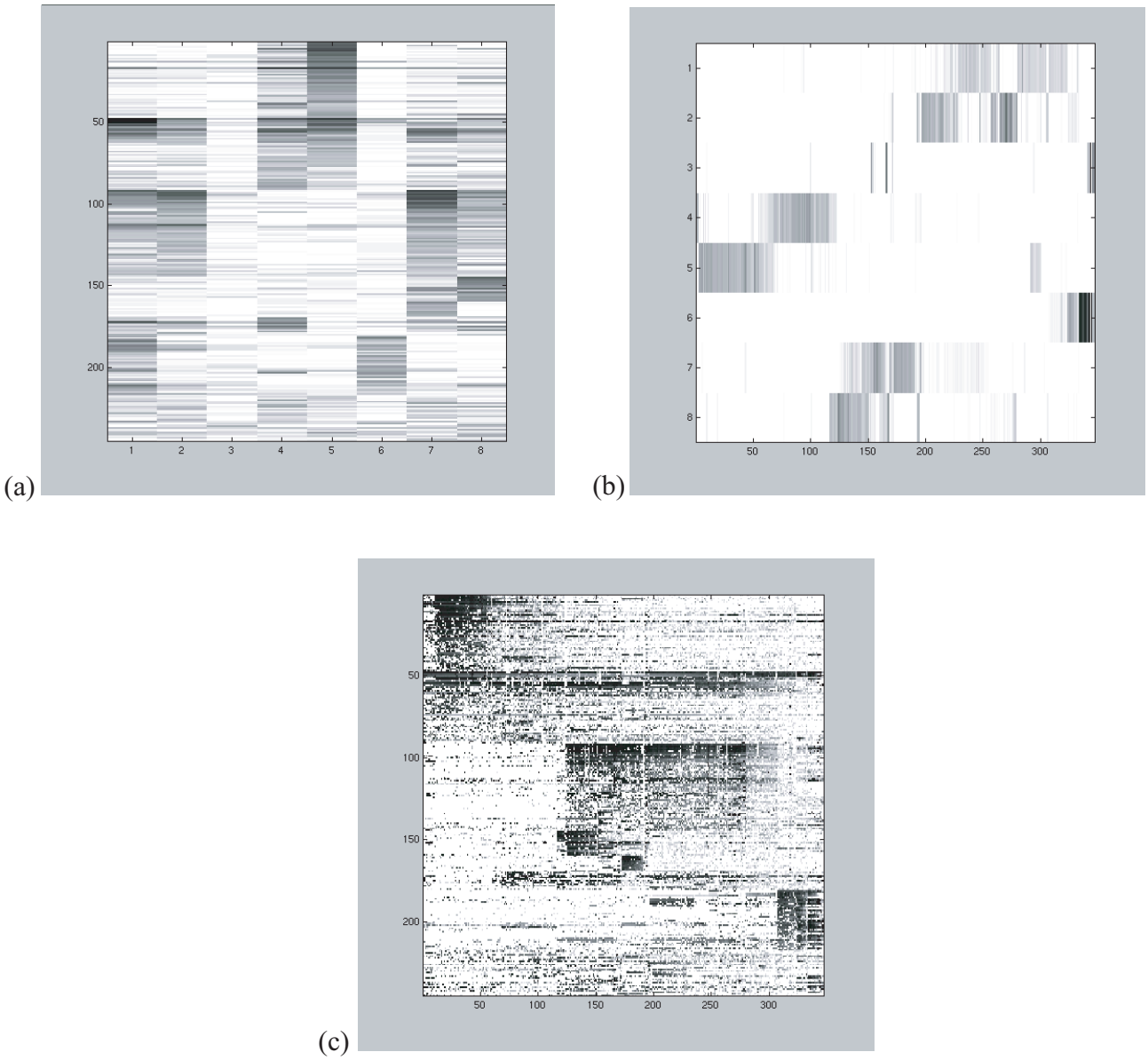

Fig. 1. Matrix factorization yields a decomposition of an $m \times n$-dimensional data matrix V as the product of two matrices $W * H$, which are of lower rank r. Shown here are submatrices that include the 254 strongest attributes selected by covariance analysis. (a) Attribute coefficients W at rank 8. Attributes are in rows and basis vectors are in columns. (b) Protein encoding matrix H at rank 8. Basis vectors are in rows and proteins are in columns. (c) Original 'attributes × proteins' matrix V. Attributes are in rows and proteins are in columns.

Let us consider the vertical protein as the query and let us attach cluster labels from the previous section to the columns. If there are several dots on a row, this means that the query residue is linked by transitive alignments to residues belonging to several different clusters. We sum the weights of dots for each row (query residue) in dot plots of a given query protein against all proteins in the superfamily. The sum of each row is then normalized to 100% to give the cluster-membership degrees of a query residue in each attribute cluster. Since weak assignments are likely to be noise, we excluded attribute assignments if the membership degree was less than 10%. When many residues from one protein mapped to the same attribute, we kept only the highest membership degree for this attribute and this protein.

The result of the above procedure is a fuzzy, probabilistic assignment of residues to cluster labels. The latter are the attributes of our sequence space (columns of the

Table 1. Mapping of the active site motif to CD

Row	Attribute	Selected (%)	3D	Others (%)
47	1375.D	D315 (73)	A	D150 (14), D316(13), D425 (10)
48	2238.H	H248 (81)	A	H237 (16)
49	2293.H	H216 (64)	A	H200 (12), H236 (18)
50	1925.A	A206 (71)	–	A184 (11)
51	1374.T	T133 (31)	D	T131 (10), T191 (10), T290 (10), T298 (14), S306 (17)
52	344.G	G271 (36)	C	N166 (15), G183 (10), N277 (16)
53	1922.H	H63 (44)	A	H120 (18)
54	1567.H	H65 (46)	A	H120 (24), H25 (13)
55	1241.D	D126 (52)	D	D35 (11), E61 (12), D129 (18)
56	362.G	G117 (54)	B	G86 (25), Q113 (13)
57	1822.G	G241 (21)	C	D180 (10), D210 (16), N256 (13)
58	363.V	I118 (22)	B	T98 (11)
59	1022.H	H216 (24)	A	H200 (14), H236 (10), H237 (10)
60	1769.E	E422 (18)	–	D350 (18), E384 (13), D388 (10)
61	1232.I	V60 (37)	B	I56 (12), V125 (19)

Col 1: row/column of attribute in Figure 1 and Figure 2. Col 2: amino acid type and arbitrary position-index of attribute. Col 3: mapping from attribute to a particular residue of CDA, membership degree in parenthesis. Col 4: spatial clusters identified from Figure 3a. Cluster A is the active site signature. Col 5: as Col 3, for all other mappings of the given attribute position. Subtract 5 to match residue numbers in PDB entry 1k6wA.

hypothetical greedy multiple alignment). Now we are ready to carry out NMF on the { attributes × proteins } matrix. When the NMF results are analysed, attributes are mapped back to the most probable single residue in a given protein. Having found an interesting attribute, we look up all residues in the target protein that match this attribute *position*. We then select that residue which has the highest membership degree with this attribute position; sometimes the amino has been mutated compared to the amino acid that defined the attribute label. Table 1 gives a worked example. For example, residue H216 of CDA is assigned with 64% confidence to attribute '2293.H'. H216 is also assigned with 24% confidence to attribute '1022.H'. H200, H236 and H237 are also assigned to the same attributes with low confidence (but their strongest assignment is to other attributes).

Dimension reduction

The dimensions of the data matrix V in the present analysis were $m = 15\,565$ attributes and $n = 347$ proteins. The matrix is sparse, with 282 800 non-zero entries. The matrix factorization was done in Matlab using a local adaptation of the NMF code by Lee & Seung (http://journalclub.mit.edu, under the Computational Neuroscience discussion category). Results were collected for ranks 2, 4, 8, 16, 32 and 64. Each rank was computed in 3–6 hours wall-clock time on a small workstation.

Covariance analysis

The NMF decomposition of proteins is quite crisp (Fig. 1b) and represents sequence-similar protein families, but post-processing is needed to focus on the most interesting attributes. A number of attributes have positive coefficients in several basis vectors (Fig. 1a). This is useful for our purposes because of 'functional hierarchy'. Functional sites are made up from a smallish set of residues. One functional site may represent a basic biochemical mechanism that is inherited unchanged by all members from the common ancestor of the superfamily, while a variety of substrate binding sites of different specificity have evolved later in different lineages. Sets of attributes that are common to two or more families are particularly interesting, because the NMF partition reflects the general trends of sequence divergence between protein families, while ancestral functional sites should be conserved across families.

To capture this 'evolutionary behaviour' of attributes, we analysed the covariance of attribute coefficients in NMF partitions at different resolutions (low to high rank). More precisely, the covariance matrix was computed for a matrix that included all basis vectors obtained at ranks 2, 4, 8, 16, 32 and 128. The first two entries in the coefficient vector of an attribute are given by rank 2 NMF, the next four by rank 4 NMF, and so forth for a total of 126 entries. Similar results were obtained using various subsets of basis vectors. We used covariance rather than correlation because the length of the coefficients vector represents the importance (conservation) of an attribute, while correlation only compares the direction of the vectors. The covariance σ_{xy} of random variables x and y is

$$\sigma_{xy} = 1/n\Sigma(x_i - \langle x\rangle)(y_i - \langle y\rangle) \qquad (2)$$

Here, the random variables are the coefficients of a given attribute in all NMF experiments. Attribute coefficients are

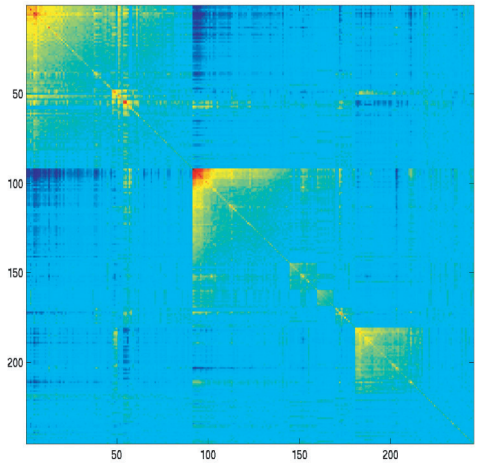

Fig. 2. Covariance matrix of attribute coefficients. Red and yellow are high positive values, green is neutral and dark blue is very negative. The ordering of attributes is the same as in Figure 1.

normalized in NMF so that the total sum per protein is one. All but ∼900 attributes had zero coefficients to three decimal places. From the ∼900 positive attributes, we selected those vectors whose length was above average. This left 254 attributes with the highest coefficients for further analysis (shown in Fig. 1 and Fig. 2). Hierarchical clusters of attributes were obtained by average linkage based on the covariance matrix of attribute coefficients (Fig. 2). The branching order is such that most attributes are added one at a time to the longest vector. This shows up in Figure 2 so that the first member of a cluster has high covariance with a large number of attributes that have lesser covariance between themselves. Motifs present in only a subset of the proteins show anti-correlation with other motifs.

RESULTS

Two-way clustering

The results represent the synthesis of 347 proteins containing about 150 000 amino acids. NMF was run at rank 2, 4, 8, 16, 32, and 64. Figure 1 shows the matrix factorization at rank 8. Attributes were ordered based on hierarchical clustering of attribute coefficients W (Fig. 2). The proteins could be ordered in the same way based on the encoding vectors H. Figure 1c indicates a block structure in the occurrence of attributes in different protein families. Similar trends were observed at all ranks, with more details emerging at higher ranks (that is, residues conserved in smaller sub-families).

Major protein families

Based on PSI-Blast searches, we have manually defined 26 main branches in the family tree of the urease superfamily. In NMF analysis, most of these families were unambiguously associated with a single basis vector at all ranks tested. Figure 1b shows the groups obtained at rank 8. Known functionalities in these groups are as follows: (1) prolidase, imidazolonepropionase, cytosine, adenine and adenosine deaminase, (2) chlorohydrolase, guanine deaminase, acylase, (3) a few outliers, (4) nodulin and many small families of exotic enzymes, (5) TatD DNase and phosphotriesterase, (6) AMP deaminases, (7) pyriminidase, N-acetylglucosamine deacetylase, (8) dihydroorotase, allantoinase, urease. Most groups include a large number of hypothetical proteins of unknown function.

Major sequence patterns

We picked the most clearly covariant clusters of attributes from Figure 2 and mapped these to 3D structures. Figure 1b shows in which proteins the pattern is present. As illustrated in Figure 3, the patterns have high contact order (i.e. long sequence separation between spatial neighbours). The spatial clustering indicates that the patterns discovered here by sequence analysis are highly non-random.

Active site

Hierarchical clustering of the covariance matrix of attribute coefficients (Fig. 2) groups together rows/columns 47–61. This pattern is present across all proteins (see Fig. 1c). The covariance analysis extracts this pattern separately from subfamily-specific motifs. Such distinction is difficult to achieve, if one analyses only one group of closely related proteins. In 3D, the pattern maps mainly to the active site. Importantly, the complete set of metal ligands and the catalytic aspartate are recovered (Fig. 3a). These four histidine and one aspartic acid residues (3D-cluster A in Table I) are located at the ends of beta-strands 1, 5, 6 and 8 of the (beta/alpha)8-barrel. Structure-based alignment shows that they are invariantly conserved, but the discovery of their co-occurrence based only on sequences is a non-trivial achievement. The motif is very sparse and invariant residues are embedded in high-entropy sequence segments. Because of this, profile-based motif discovery software (MEME, Probe, PSI-Blast) recover at most parts of the motif between distant relatives (Heger *et al.*, 2003; Holm, 1998).

Binuclear marker

This motif corresponds to rows/columns 159–168 in Figure 2. Known structures from the urease superfamily include both mononuclear (ADA, CDA) and binuclear proteins (URE, PTE, DHO). In addition to the four

histidines of the universal active site motif, the binuclear proteins have a modified lysine on strand beta-4 as a bridge between two metal ions (Fig. 3b). In some members of the PTE family, the lysine is substituted by a functionally equivalent glutamic acid (Buchbinder *et al.*, 1998). The binuclear site in the structure of DHO came as a surprise (Thoden *et al.*, 2001), as it was not predicted from sequence comparison in the original report on the urease superfamily (Holm and Sander, 1997). The lysine is difficult to detect—even by eye—in alignments among all dihydro-orotases because it is near a loop with long insertions/deletions. Here, the lysine is picked up as a conserved attribute.

Nucleotide binding

Rows/columns 170–186 in Figure 2 define a motif which is present in ADA and CDA. Despite the chemical similarity of their substrates, the unification of ADA and CDA is surprising, because their sequences are overall so diverged that they aren't confidently grouped together in phylogenetic trees based on Blast or PSI-Blast e-values. Also NMF separates ADA from the rest already in the first partition at rank 2. The motif found here contains residues which make contacts to the nucleotide base from two sides and could possibly be used as fingerprints to predict the substrate specificity of hypothetical proteins (Fig. 3c).

Small domain

Rows/columns 91–110 in Figure 2 map to the small domain and its interface with the catalytic (beta/alpha)8-barrel domain. Sequence profile methods actually find a stronger signal in the urease superfamily for the small domain than for the catalytic domain. This is because the spacing of conserved residues along the sequence is sufficiently close for local alignment profile models to detect conserved blocks (Fig. 3d). The motif has two parts, since the small domain is made up of N-terminal and C-terminal segments in the sequences. The small domain is present in CDA and URE, truncated in DHO, and absent from ADA and PTE. The function of the small domain is unknown.

Hydrophobic core

Rows/columns 1–31 in Figure 2 define a pattern which represents a conserved hydrophobic core in the large TatD and PTE families.

DISCUSSION

Summary

Non-negative matrix factorization is a productive way to identify sequence patterns that are conserved in subgroups of proteins in diverse superfamilies. We have shown that explicit multiple alignments can be bypassed in defining the space in which multivariate analysis is conducted. We applied simple post-processing to extract clusters of residues that seem to represent biologically significant sequence motifs.

Advantages

The present method is more sensitive than methods for pattern discovery in unaligned sequences (e.g. MEME Grundy *et al.*, 1997; Vilo , 1998). For example, MEME does not recognize the active site motif across all members of the urease superfamily, because only the functional residues are invariant and surrounding positions have high entropy. Thus, the statistical signal is too weak for patterns which only specify amino acid types and spacings (e.g. H.H.{110,197}H.{24,38}H.{57,88}D for the active site signature in five known structures). Our patterns use a much larger alphabet size, as a result of anchoring amino acid types to presumed structurally equivalent positions. This removes the degeneracy of, for example, the histidines in the above pattern. This lowers entropy (the number of chance matches) and increases information content (Altschul, 1991). The source of the added information is structural context from the underlying library of pairwise alignments between closer homologues, which is expanded to distant homologues via transitive alignment (Heger *et al.*, 2003).

Multiple sequence alignment is a difficult problem, especially in such large and diverse sets as the urease superfamily. We identify putative structurally equivalent positions based on transitive alignments between all pairs of proteins. An explicit multiple alignment would choose only one of many possible alternatives. Because of the uncertainty of alignment, we preferred here to use all the information and assigned relative likelihoods of cluster membership. This fuzzification allows us to start from an unaligned set of sequences. In contrast, an explicit multiple alignment is required input to the otherwise related methods Sequence Space (Casari *et al.*, 1995) and Evolutionary Trace (Madabushi *et al.*, 2002). These previous methods are based on two valuable insights. Sequence Space introduced the concept of a generalized space with amino acids types in columns of the multiple sequence alignment as attributes. We have here extended this idea to handle very distant homologs whose alignment is non-obvious. The insight behind the Evolutionary Trace method is that of using phylogenetic trees to partition proteins into informative subgroups and defining a motif as the set of all invariantly conserved residues within the group. Here, we used a very general method of partitioning proteins as well as a more general and generous measure of sequence conservation (i.e. the attribute coefficients).

We used post-processing by covariance-clustering to define motifs composed of many residues. The covariance matrices looked visually similar at ranks ranging from 2 to

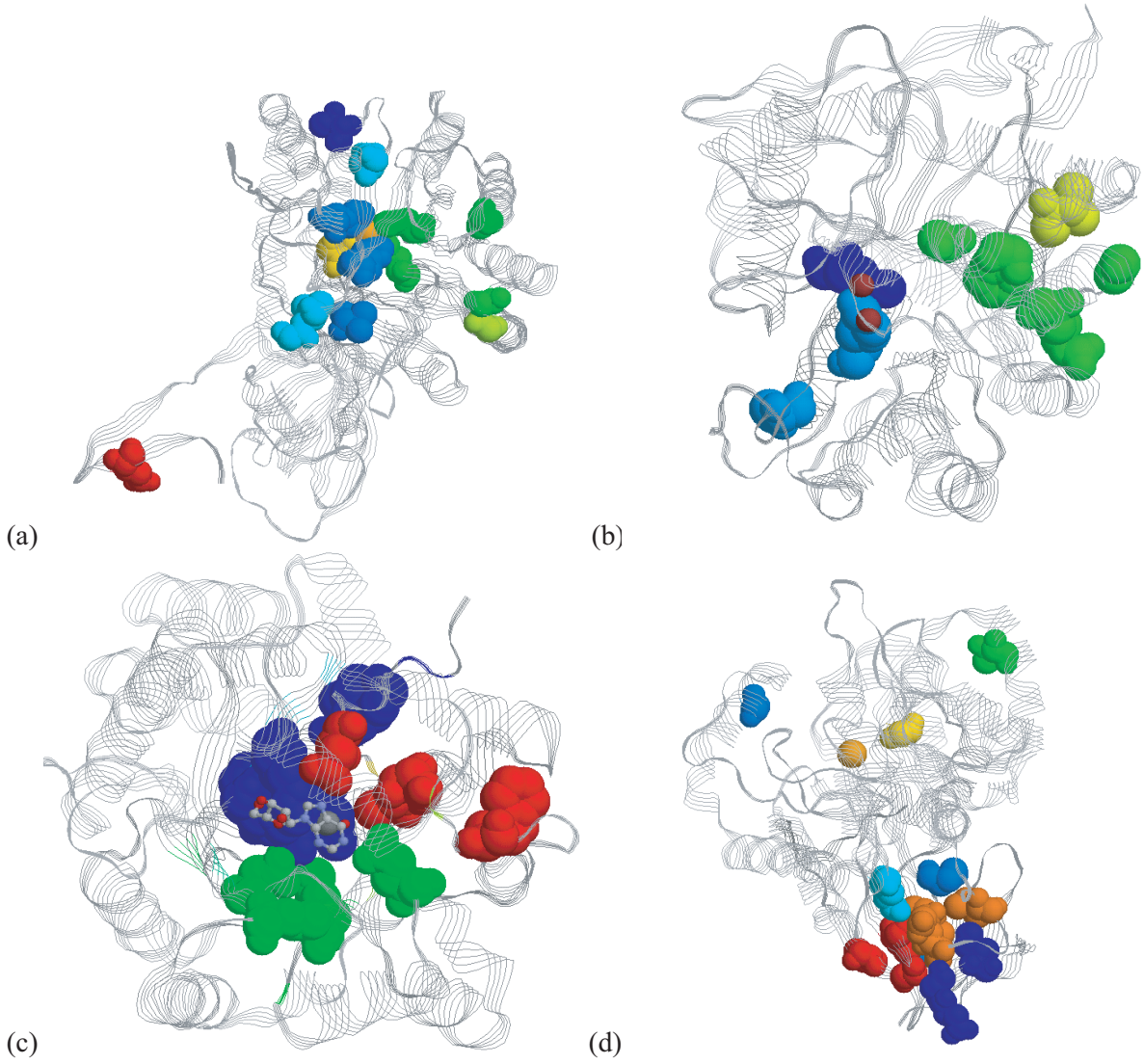

Fig. 3. Sequence-derived patterns mapped in 3D. Motif positions are space-filled and rainbow coloured blue-cyan-green-yellow-orange-red from the N-terminus to the C-terminus of the polypeptide chain. Thus, different colours in neighbouring residues indicate high contact order. The structures are oriented so that the 2D projection presents true 3D contacts. (a) Active site motif in CDA (PDB entry 1k6wA). The iron atom is orange in the middle of the big cluster and it is coordinated by four histidines. The yellow residue is the catalytic aspartic acid. (b) Binuclear marker motif in DHO (PDB entry 1j79A). The blue residue coordinating two dark metal atoms is the carbamoylated lysine on strand beta-4. The row of conserved green residues is in strand beta-7 across the (beta/alpha)8-barrel. Two further conserved residues in the upper right hand side mediate helix packing against strand beta-7. (c) Nucleotide binding motif in ADA (PDB entry 1a4mA). The metal and substrate are shown in ball-and-stick. The blue residues include strand beta-1. Two motif residues (right-most green, left-most red) recognize specificity determinants N1, N7 of the adenosine ring. The middle red residue is a cysteine packed against the aromatic ring of a histidine metal ligand. (d) Small domain motif in CDA. The catalytic domain is inserted in the middle of the small domain.

64, suggesting that there is a consistent evolutionary signal in the data and that rank is not critical to this analysis. The clustering selected a subset of attributes that are most highly conserved and correlated in a given partition of the protein set. Spatial clustering in 3D mapping suggested biological significance of the motifs. To a trained eye, the conserved sequence patterns are obvious in pileup alignments of close relatives[†], but hierarchical decomposition into universally present and more specific patterns is interesting and non-trivial. Overall, the present

[†] Structural alignments expanded by homologous sequences are available at http://www.ebi.ac.uk/dali/fssp

automatic method saves time from laborious manual examination.

Limitations

The matrix factorization analysis basically only reveals groups of proteins with some sequence similarity and gives high coefficients to those residues that are more conserved within a given group. Such sequence conservation has an evolutionary interpretation, in that selection for functional proteins leads to the conservation of functional residues and, more generally, the 3D structure. One problem is that functional sites are detected alongside structural conservation. Another problem is that while motifs hint at functional classes, we usually cannot make a guess about the common function until there is an experimentally characterized member in the class.

The fuzzification of attributes was motivated by uncertainties in the data. Further work is needed to determine how much it actually enhanced signal compared to adding noise. Small families have lesser weight in NMF. Many protocols of sequence weighting have been proposed in bioinformatic literature. Here, we used a balanced sample of representative sequences with less than 40% sequence identity.

The method is open to a number of extensions. For example, 3D structures provide very useful information for refining the motifs. 3D connectivity could be used to remove outlier residues or split motifs that describe several spatially disjoint sites.

REFERENCES

Altschul,S.F. (1991) Amino acid matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1995) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, **5**, 279–305.

Buchbinder,J.L., Stephenson,R.C., Dresser,M.J., Pitera,J.W., Scanlan,T.S. and Fletterick,R.J. (1998) Biochemical characterization and crystallographic structure of an *Escherichia coli* protein from the phosphotriesterase gene family. *Biochemistry*, **37**, 5096–5106.

Casari,G., Sander,C. and Valencia,A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Grundy,W.N., Bailey,T.L., Elkan,C.P. and Baker,M.E. (1997) Meta-MEME: motif-based hidden Markov models of protein families. *CABIOS*, **5**, 211–221.

Heger,A., Lappe,M. and Holm,L. (2003) Accurate detection of very sparse sequence motifs. *RECOMB'03*, in press.

Holm,L. (1998) Unification of protein families. *Curr. Opin. Struct. Biol.*, **8**, 372–379.

Holm,L. and Sander,C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins*, **28**, 72–82.

Hyvärinen,A. (1999) Survey of independent component analysis. *Neural Computing Surveys*, **2**, 94–128.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–191.

Madabushi,S., Yao,H., Marsh,M., Kristensen,D.M., Philippi,A., Sowa,M.E. and Lichtarge,O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Thoden,J.B., Philips,Jr,G.N., Neal,T.M., Raushel,F.M. and Holden,H.M. (2001) Molecular structure of dihydroorotase: a paradigm for catalysis through the use of a binuclear metal center. *Biochemistry*, **40**, 6989–6997.

Vilo,J. (1998) *Discovering Frequent Patterns from Strings*. Department of Computer Science, University of Helsinki.