Journal of Computer-Aided Molecular Design, **15:** 477–487, 2001.
KLUWER/ESCOM

Classification of protein disulphide-bridge topologies

J.M. Mas[a], P. Aloy[a], M.A. Martí-Renom[a], B. Oliva[a], R. de Llorens[b], F.X. Avilés[a] & E. Querol[a,*]
[a]*Institut de Biologia Fonamental i Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain;* [b]*Unitat de Bioquímica, Departament de Biologia, Facultat de Ciències, Universitat de Girona, 17071 Girona, Spain*

Received 28 August 2000; accepted 2 February 2001

Key words: disulphide bonds, protein classification

Summary

The preferential occurrence of certain disulphide-bridge topologies in proteins has prompted us to design a method and a program, KNOT-MATCH, for their classification. The program has been applied to a database of proteins with less than 65% homology and more than two disulphide bridges. We have investigated whether there are topological preferences that can be used to group proteins and if these can be applied to gain insight into the structural or functional relationships among them. The classification has been performed by Density Search and Hierarchical Clustering Techniques, yielding thirteen main protein classes from the superimposition and clustering process. It is noteworthy that besides the disulphide bridges, regular secondary structures and loops frequently become correctly aligned. Although the lack of significant sequence similarity among some clustered proteins precludes the easy establishment of evolutionary relationships, the program permits us to find out important structural or functional residues upon the superimposition of two protein structures apparently unrelated. The derived classification can be very useful for finding relationships among proteins which would escape detection by current sequence or topology-based analytical algorithms.

Abbreviations: PDB – protein Data Bank; RMSD – root mean square deviation

Introduction

Sequence database searching has been used as a powerful tool in molecular biology because some evidence of protein structure and function can be inferred from it. Searches based upon efficient alignment algorithms are applied routinely to all newly deduced protein sequences (for a review, see [1]). A similar scenario can be envisaged for three-dimensional structure comparisons. The rate at which new protein structures are being obtained exceeds one per day as a result of recent advances in crystallography and NMR spectroscopy. Given that the three-dimensional (3D) structure is highly conserved in protein evolution [2, 3], the comparison of 3D structures allows for the establishment of relationships between protein families which obviously would not arise from sequence alignment alone (for unifying two or more families into superfamilies or for discovering biologically interesting relationships [4–9]).

Protein 3D comparison requires the superimposition of the corresponding structures. There are a very large number of ways in which one could match backbone atoms, from any pair of proteins, but an extensive computational analysis is still unfeasible with today's computers. Early computer methods required manual initial alignment and were very slow or limited to close homologues [10, 11], whereas new search algorithms – generally based on simplified approaches – have been recently developed allowing fully automated and rapid similarity searches through an entire database [14]. These computer methods for structural alignment have been (and still are) very useful for the detection and classification of the building blocks of

*To whom correspondence should be addressed:
E-mail: ibfquer@blues.uab.es

protein structures [6–8, 15–20]. However, the structural alignment for the comparison and classification of proteins having low sequence similarity or lacking regular secondary structure still remains difficult, and although most modern programs do not need secondary structure to solve the problem new approaches are welcomed [21, 22].

Disulphide-containing proteins constitute a large group for which the development of tools for the 3D structural alignment and topological analysis is required, especially in the case of disulphide-rich proteins because the high content in disulphides quite probably strongly affects their fold and topological tendencies [23–25]. An important group are the small proteins with little or no regular secondary structure, in which the disulphides are indispensable for their structure and function [26, 27]. In this context, it has been previously suggested that common structural/functional features among proteins can be interrelated through the analysis of their disulphide bridges [9, 28–30]. This led to some attempts to establish a classification of disulphide-rich proteins by means of cystine geometry [31, 32] or disulphide-bridge connectivity [33–35].

This work describes a computer-based method to study disulphide bridge topologies in proteins in order to classify them and facilitates comparison of 3D structures. An initial classification of proteins containing more than two disulphides is performed. Several structural classes are defined, and a detailed comparison of one of them has been published by our group [30].

Methods and algorithms

The set of proteins

The set of proteins used for the analysis and classification here performed was extracted from the PDB [36] using the set of protein structures corresponding to the non-redundant PDB at 65% sequential homology [37, 38], obtained from the World Wide Web (http://www.embl-heidelberg.de) and dated in 1996. The final number of proteins was reduced by means of:

(a) Removal of structures of less than 30 residues and a 3D resolution lower than 3.5 Å.

(b) Removal of proteins with fewer than three disulphide bridges. Three is the most frequent number of disulphide bridges in proteins with known 3D structures [33].

(c) Proteins with more than three cystines have been classified using the best possible alignment of three of its disulphide bridges.

Topology of three-disulphide bridge knots

We wish to define the three-disulphide bridge knot of a protein independently of the sequence orientation (from N terminus to C terminus or viceversa). This implies that one disulphide bridge has to be described as a segment with no orientation (beginning and end of a vector) in the three-dimensional space. Therefore, the segment which defines one disulphide bridge is characterised by the co-ordinates in Cartesian space (R^3) of both ends. The Cα co-ordinates of each cysteine residue forming the disulphide bridge are taken as the end co-ordinates which define the segment. As a consequence, a vector in $R^3 \times R^3$ space describes one disulphide-bridge segment. The 'i' disulphide vector of protein A is defined in $\mathrm{R}^3 \times R^3$ as $r_{\mathrm{C}\alpha}^{\mathrm{ss(A,i)}} = (x, y)$, where x and y are the Cα co-ordinates of the cysteine residues which form the disulphide bridge.

The rotation of a set of disulphide-bridge segments (in $\mathrm{R}^3 \times \mathrm{R}^3$) of one protein can be obtained from the rotation in R^3 space of the whole protein. Therefore, the superimposition and RMSD between two proteins can be extended to three-disulphide bridge knots of two proteins by representing each disulphide bridge as vectors in $\mathrm{R}^3 \times \mathrm{R}^3$ space.

The next problem in the task of comparing the knot topologies of two proteins appears when one (or both) protein(s) has/have more than three disulphide bridges. For such a task it is necessary to choose the best set of disulphide bridges forming the three-disulphide bridge knot. This comparison is done by calculating the RMSD of the Cα atoms defining the disulphide bridge segments of the knot topology. Due to the number of disulphide bridges, larger than three, each combination of three-disulphide bridges for each protein has to be checked. The best set of three-disulphide bridges for two proteins is the one for which the RMSD of Cα atoms is the smallest. In conclusion, we wish to solve the problem of obtaining the best superimposition of 6 cysteines of one protein versus the 6 cysteines of the other protein, keeping the correct linkage in three disulphide bridges of each.

To solve this problem, we assign sets of disulphide bridge segments for protein A (S_A) and other for protein B (S_B) and we define the set of RMSs, $\Im^{A,B}$, as:

$$\Im^{A,B} = \{x;\, x = \mathrm{RMS}(S_A, S_B)\ \forall \text{ combinations between } S_A \text{ and } S_B\}$$

The calculation of $\Im^{A,B}$ involves all possible combinations of disulphide bridges of the two proteins to compare, taking into account that each pair of disulphide bridges presents two different orientations of its four cysteines (parallel and antiparallel). This allows the superimposition of the disulphide-bridge segments, irrespective of the sequence orientation, by means of the superimposition in R^3 of its ends.

This procedure implies that the number of calculations and the computer time increase exponentially with the number of disulphide bridges. To prevent this, a strategy was devised which allows the removal of those comparisons with high RMS. Therefore, $\Im^{A,B}$ is reduced to a new set $R\Im^{A,B}$ with less than 30 values, defined as:

$$R\Im_N^{A,B} = \{x;\, x \in \Im_N^{A,B},\ \text{where } \forall\, y \in \Im_N^{A,B} \Rightarrow x \le y \text{ and 'total of elements}' \le 30\}.$$

The $\mathrm{R}\Im^{A,B}$ set can be constructed by means of a comparative analysis of the disulphide-bridge segments, avoiding the calculations of all the combinations. The algorithm to reduce the number of calculations is based on the Basic Interatomic Distance Matching method [12], applied to the internal co-ordinates defined by a pair of disulphide bridges.

The internal co-ordinates for locating the disulphide bridges of a protein are obtained from each disulphide bridge segment being considered as a vector in R^3. This vector is taken as $v_i = x - y$ where, $(x,y) = r_{\mathrm{C}\alpha}^{ss(A,i)}$. The set of disulphide-bridge segments of a protein is grouped into pairs of disulphide bridges. After that, angles and distances (internal co-ordinates) between these cystine pairs were calculated. The results of these calculations are stored in a square matrix which contains the information of the internal co-ordinates for the comparison between pairs of disulphide bond segments. The elements of the matrix are defined according to the Basic Interatomic Distance Matching method:

$$S_{i,k} = \sum_{m=1}^{N} \frac{A\_d}{\left(\left|\frac{d_{i,i+m}^{A} - d_{i,i+m}^{B}}{\max\_d}\right| + B\_d\right)} + \frac{A\_\alpha}{\left(\left|\frac{\alpha_{k,k+m}^{A} - \alpha_{k,k+m}^{B}}{\max\_\alpha}\right| + B\_\alpha\right)},$$

where $S_{i,k}$ is the calculated value of similarity for the comparison of the cystine pair (i) in protein A and the cystine pair (k) in protein B. 'd' and 'a' are, respectively, the values of distances and angles for each disulphide-bridge pair, 'max_d' (57 Å) and 'max_α' (π radians) being the maximum expected values. 'A_d' and 'A_α' allows us to establish the ratio of distance/angle (values of 75 and 25, respectively, were used). Finally, 'B_d' and 'B_α' prevent errors of dividing by zero, and their best estimated value is 1.

Only those combinations of three-disulphide bridges of proteins A and B, using the disulphide bond segments which give the highest 30 similarity values identified in the matrix as the largest values, are taken into account to calculate $\mathrm{R}\Im^{A,B}$, therefore reducing the total number of combinations.

Clustering of disulphide-bridge knots

A double-clustering technique has been used. First, a Density Search Technique (DST) [39] was applied to group proteins by the elements of $R\Im^{A,B}$ between each pair of proteins in the group. For a clustered group of proteins the relation between two proteins (A and B) is not given by the minimum in $R\Im^{A,B}$ but by the element in $R\Im^{A,B}$ for which the superimposition of cysteines gives the smallest RMSD value when A and B are compared with themselves and with the rest of the proteins in the cluster. The average of the chosen elements in $R\Im^{A,B}$ ($\forall$ A and B in a cluster) is defined as ssRMSD. This is done by iteration upon each new member in the cluster. Applying the DST method, a new member was accepted in the cluster only if this did not shift the cluster centromer over the limit chosen. To make this classification, a 0.3 Å cut-off was used as the limit because a lower tolerance within this limit would produce many small clusters, whilst a higher limit value would produce a small number of large clusters with a large number of disulphide-bridge topologies. The smallest cluster was obliged to contain at least three proteins to be considered by the DST.

A Hierarchical Technique (HT) [39] was used to calculate the relationships between the groups defined by Density Search as a second clustering technique. The method uses a square matrix with all ssRMSD between the DST clusters and the individual, ungrouped proteins. This gives rise to a dendrogram which allows us to classify proteins using the topologic information of the disulphide-bridge knot. This dendrogram has been used for grouping the set of proteins into several classes, of clusters with ssRMSD smaller than 2.6 Å.

Program

The program KNOT-MATCH runs on Silicon Graphics computers and is available from the World Wide Web (http://luz.uab.es/biocomputing/).

Results

The use of intramolecular geometrical relationships to describe protein structures has the advantage of being independent of the co-ordinate frame [15, 40–42]. In this study, the structural comparison has been attempted using the disulphide bridges as primary units. Structural descriptors such as virtual distance and virtual angles involving C_α of the disulphide-bonded Cys residues form the basis for structure comparison. A set of proteins with less than 65% sequence identity comparison [37, 38], containing three or more disulphide bonds, has been chosen for structure analysis. The structures are compared using the program KNOT-MATCH, and they are clustered by two different techniques Density Search and Hierarchical Clustering (see Methods and algorithms). The approach is simple and the combination of various tools to detect similarities is novel. Similarities in the scaffold, in the regular secondary structures and in important structural/functional residues have been found and examined in proteins clustered in our classification. The classification shows that disulphide bridge topologies are conserved structural motifs among proteins.

Classification of disulphide-containing proteins by means of KNOT-MATCH

A classification of protein structures deposited at the Brookhaven Protein Data Bank (PDB), and containing at least three disulphide bridges, has been achieved using the two clustering techniques: first, a Density Search Technique has been used for grouping the proteins into clusters with the same pattern of disulphide bridges; second, a Hierarchical Technique was able to group these clusters into larger groups by means of the nearest neighbour. A further inspection of these groups allowed for the definition of topological classes. Grouped clusters of proteins with the same topological order of disulphide bridges showing ssRMSD shorter than 2.6 Å were defined as classes (plus an extra group of unclassified clusters). This limit corresponds to the ssRMSD of the cluster 3 that includes the major number and variation of proteins. The final illustration of classes is defined as Topological Map Classification (Figure 1), which shows a total of 13 classes.

Sixty clusters were obtained by the first approach (density search technique), fifteen of them containing more than one protein (70% of the initial set). Proteins that had the same fold type and belonged to the same functional family formed eleven out of these fifteen clusters. Classes A to M are mainly formed by proteins with larger numbers of residues than proteins included in Class M, except for class E (formed by proteins of the Insulin-like family and Kringle Modules). Also, these classes often have enzymatic activities and are rich in regular secondary structures. Finally, Class M is composed of 66 proteins where the main cluster is number 3. This cluster is formed by 64 proteins, most of them small proteins (85%) with few or no regular secondary structures, their fold being mainly organised around three or four disulphide bridges. A large number of proteins included in this cluster are growth factors, hormones, enzyme inhibitors and toxin venoms. We have also found in cluster 3 some proteins with a large number of disulphide bridges and significant regular secondary structure content. Seventeen different folds are included in the cluster and five of them (Cystine-knot Cytokines, Epidermal Growth Factor-like, Small Inhibitors, Toxins and Lectins, Snake Venom Toxin-like and Defensin-like) represent more than 50% of the total set of this cluster. This cluster is mainly formed by the known β-disulphide topology and by members of the T-knot family already studied by other authors [9, 32, 43, 44].

Analysis of the derived classification

Most of the homologous proteins were clustered together, and almost all of the proteins within the same family (as defined by SCOP) were grouped into clusters or classes with low ssRMSD values. Moreover, remote homologous, analogous and non-related proteins were frequently grouped by KNOT-MATCH within the same class. An example of this situation is shown in Cluster 3 (Class M) for the Snake Venom Toxin-like fold. This fold is represented by two different families: the Snake Venom Toxin-like, (with six representative homologous proteins), and Dendroaspin (with one representative). This last protein is analogous to the above-mentioned Snake Toxin-like according to the definitions described in the literature [20]. In addition, Cluster 3 is formed by twenty-eight families with

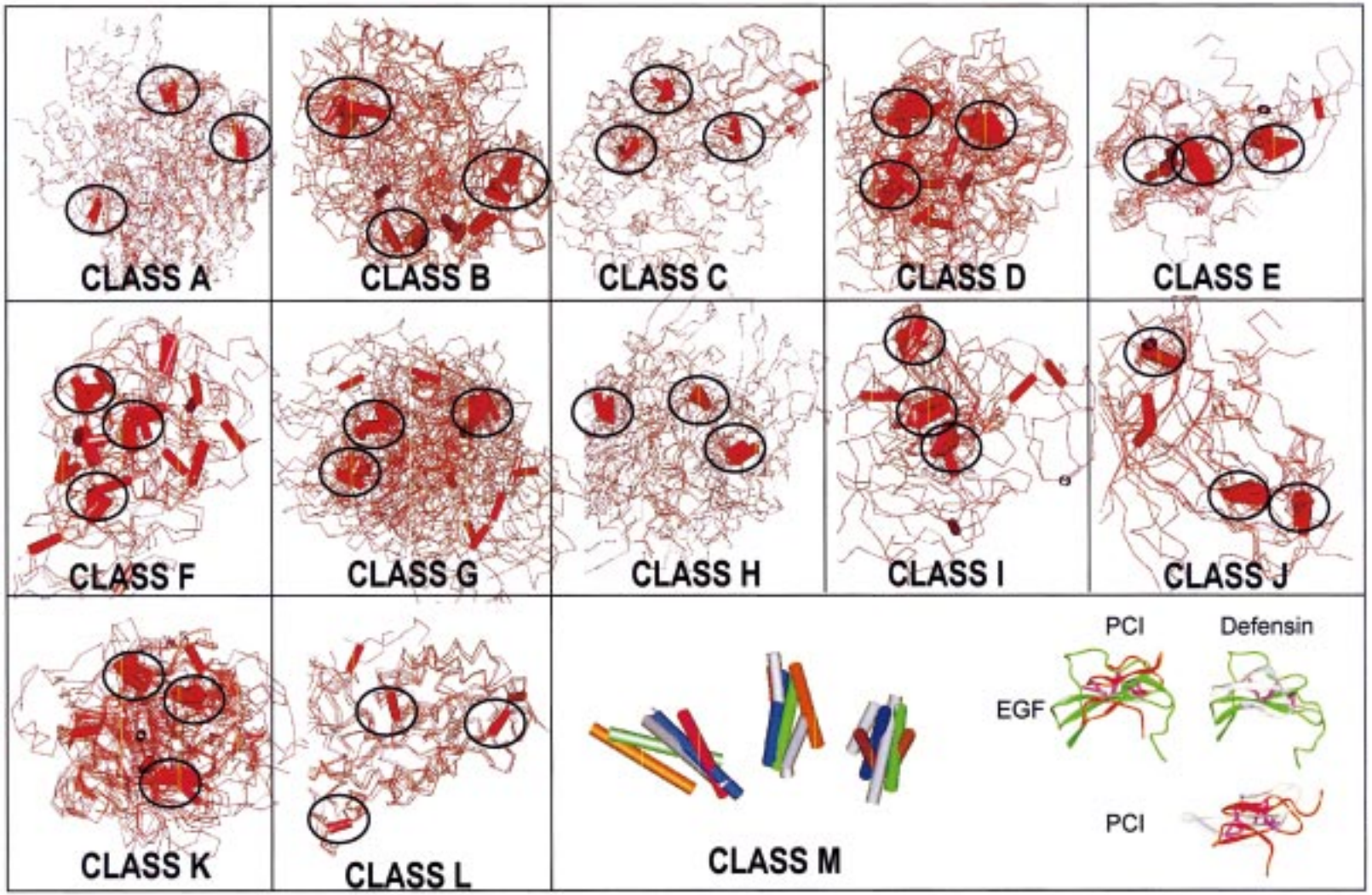

Figure 1. (a) Topological Map of the proteins classified according to their disulphide-bridge topology. Thirteen classes are obtained using topological characteristics of the disulphide bonds. ssRMSD values (in Å) and the cluster identification (bold numbers) are shown within each circle. See Table 1 for PDB codes. At the right bottom it is shown how a circle is obtained from the hierarchical clustering by using a cut-off of 2.6 Å (Class C has been used as an example).See Table 1 for PDB Codes. (b) The classes shown in the Topological Map are presented with their protein scaffolds superimposed by the disulphide-bridge topology. For Class M, having the largest number of proteins, only the disulphide bridges of a family is presented with three protein examples.

different folds and, consequently, non-related proteins under the criteria of SCOP.

Comparison between classes

Although it could be expected that homologous proteins of the same functional family or with a similar fold were joined within the same disulphide topological class, it is remarkable that there are examples of the opposite (i). Moreover, some proteins with more than three disulphide bridges could be representatives of more than one topological class (ii). The following examples describe some of these paradoxes:
(i) Classes G and H include proteins that have the same fold (Papain-like fold type). Nevertheless, it can be observed that they arise from quite distant taxonomical organisms, such as kiwi fruit or papaya for the H class and human for the G class. Therefore, it is not rare that even for the same fold type the disulphide bridge topology differs by more than 2.6 Å.
(ii) Cluster 5 (in Class D) and Cluster 1 (in Class K) are split in the topological map of Figure 1, but all of the proteins belonging to these groups are Serine Proteinases. When analysing each protein in detail, it is noticed that the members of Cluster 5 are vertebrate proteins, while those of Cluster 1 are from fungus or worms. However, chain A of γ-Chymotrypsin (3gctA) can enter in both clusters using different combinations of disulphide bridges (it has more than three cystines), a fact which allows its classification in two different classes.

How reliable are the clusters and classes?

The method uses a double clustering approach, first by density search and second by hierarchical clustering. Those proteins, that should somehow be related and escaped a cluster from the first method, are joined afterwards as a consequence of the double clustering (see Figure 1). Therefore, the method guarantees the detection of likely relations between disulphide bridge topologies, and the protein family relations within clusters and classes can substantiate this.

A statistical analysis of the proteins within the clusters and classes has been performed in order to assess the fold and family relations. The results have confirmed that the disulphide bridge topological map classification obtained by the double clustering method can show the evolutionary relationship between most of the proteins and/or their connectivity by family and/or fold, and the method is thereby validated. The first method of clustering groups 55% of proteins of the same family together and the second clustering method grouped 80% of these proteins in the same class (classes defined as in Figure 1). Moreover, only three out of a total of 60 clusters obtained by the first method included proteins from different families (clusters 3,13 and 14, this being 5% of the clusters), a fact which shows the high specificity of the first clustering method.

These results show that most of the proteins of the same family present a similar disulphide bridge topology and demonstrate the reliability of the double clustering method to classify most of these topologies. However, there is still 20% of proteins with common members of the same family that may be erroneously classified. This 20% has been thereafter studied in order to justify the method. A 75% of the proteins (within this 20%) is formed by proteins of special cases already mentioned in the text (cases (i) and (ii) above), increasing the previous 80% to a 95% of successfully explained cases. The remaining 25% is composed by the vertebrate Phospholipase A2 and by the representation of the family of Fungal Lipases. The vertebrate Phospholipase A2 appears in classes D and F as a consequence of different choices of disulphide bridge topologies. Another result characteristic of paradox (ii) is that of the family of Fungal Lipases (represented by three proteins in the current data set) which possess three different disulphide bridge topologies.

Finally, the 13 classes (A to M) defined by the disulphide bridge topology and presented on the topological map (Figure 1 and Table 1) show different percentages of structural protein classes (Table 2). It is noteworthy that 70% of the topological classes (A, B, E, F, I, J, K, L and M) are formed from a single structural type (either α, β, α/β, $\alpha+\beta$ or 'small proteins').

Structural/functional relevance of the disulphide-bridge topology

Cluster 3 (Class M) is the most populated cluster (Table 1) and was rigorously analysed in order to understand why the first clustering method included proteins from different families. Some proteins of this cluster have already been studied by Mas et al. [30] because of the various structural and functional relationships found between them. The proteins of this cluster were overlapped by means of their disulphide-bridge topologies and its side-chains and backbones were compared to obtain new insights. A number of

Table 1. Proteins in classes and clusters

Class	cluster:protein[a]
Class A	**10**:1hrnA, 3psg, 3cms; **33**:1mhlC; **39**:1tca; **46**:1ack **55**:1htrB
Class B	**9**:1lct, 1nnt, 1tfd; **28**:1hvm; **30**:1lki; **50**:1tib
Class C	**15**:1abr; **38**:1rcb; **43**:3gly;**60**:2aaiB;
Class D	**5**:1hylA, 1sgt, 3gctA,1try; **12**:3sgbI, 1pce, 2bus;**13**:1pce, 1hpt, 1kpt; **14**: 2alp, 1ncsA,1poa; **32**:1lpt; **59**:2kaiB
Class E	**6**:1pk4, 2pf1, 1pml, 1pkr; **22**: 1fbr; **40**:1thv; **51**:2hpqP; **52**:6ins; **53**:1kdu; **56**:1igl;
Class F	**2**:1poc, 1ppa, 1poa,1pp2L,1bp2,1pod; **54**:1pmc;
Class G	**8**:193l, 2eql, 1lzr, 1hml; **27**:1hucB; **31**:1lpbB; **44**:3tgl; **48**:1cpy;
Class H	**36**:1ppn; **57**:2act; **42**:3aahA; **34**:1pgs;
Class I	**7**: 1dtx, 1knt, 1dtk, 1aapA, 1shp; **18**: 1bip; **29**:1hxn; **49**:1tap;
Class J	**11**:1onc, 7rsa, 1ang; **23**:1hbq;
Class K	**1**: 1lmwB, 3gctA, 1ton, 3rp2A, 4ptp, 3est, 1ppfE, 1bit, 1hcgA;**17**:1arb;
Class L	**4**:1lgaA, 1arv, 1mnp;
Class M	**3**:1bet, 1hcnA, 1hcnB, 1pdgA, 2tgi, 1prhA,1tpg, 1hcgB, 1zaq, 4tgf, 1hrf, 3egf, 1ccf, 1erp,1erl, 1erd, 1ica, 1pnh, 2crd, 1mtx, 1nrb, 1gps, 1lpbA,1oaw, 4cpaI, 1hev, 1hyp, 1dec, 4hctI, 1pi2, 2bbi, 4sgbI, 1tgxA, 1coe, 1ntx, 1fas, 1kbaA, 2abx, 1ntn, 1cvo, 3ebx, 1tfs, 1drs, 1atx, 1dfnA, 2sh1, 1bnb, 2bds, 1ahl, 1c5a, 1cbn, 1esl, 1pp2L, 1ncsA, 2madL, 1vmoA, 1lct;
Ungrouped clusters	**47**:1ate; **19**:1bw4; **16**:1aozA; **20**:1cnsA; **21**:1esc; **24**: 1hc4; **25**:1hfh;**26**:1hgeA; **37**:1prtB; **41**:1vcaA; **45**:6taa; **58**:2fbjH;

[a]Proteins are in PDB codes

matches obtained in the 3D space were contrasted with experimental results reported in the literature.

Relationships between members of different families in a cluster

An important group of the proteins from cluster 3 have some structural similarities in backbone regions; however they have different structures and functions. These similarities have been viewed using the Turbo FRODO graphics program [30, 45]. Potato Carboxypeptidase Inhibitor (PCI), a plant protease inhibitor which is a single member of its family in the current data set, has been chosen as a reference in order to describe the likeness of the scaffolds.

The C27-C34 loop of PCI is a similar region within the members of the scorpion toxins (1agt and 2crd) and within some members of the EGF-like family (i.e., 1ccf and EGF). Also the PCI loop C18-C24 has equivalent regions with proteins of other families such as the snake venom toxins (loop K47-C54 of 1coe) and the EGF-like family (loop K47-C54 of coagulation factor $\times$ 1ccf).

The most remarkable results have been obtained for the comparison between chemically equivalent side-chain groups in the space of PCI and the EGF-like family [30]. Fifteen locations conserving the physicochemical properties of the involved residues in both families have been located; interestingly, most of these residues have been described as structurally or functionally important, either for PCI or for EGF [46–58]. These structural relationships found between PCI and EGF-like proteins could justify that PCI acts as a growth factor antagonist through its binding to EGF receptor [58].

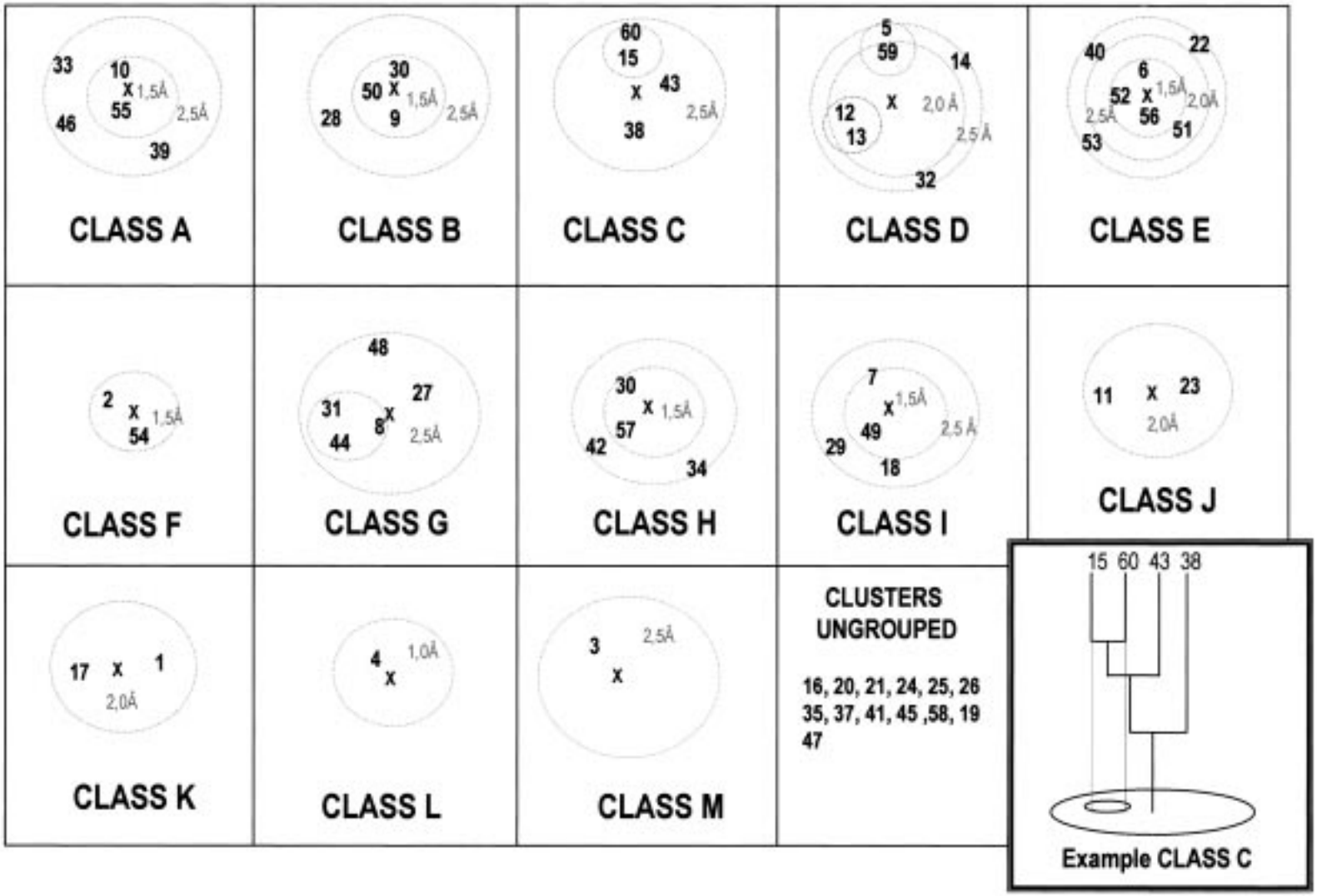

Figure 2. The EGF (3egf) and defensin (1dfn) scaffolds are shown. The members of EGF and defensin families have been independently overlapped by their disulphide-bridge topologies. A set of positions in the space with conserved physicochemical properties has been found for several members of each family. These positions are indicated in the figure with the number of the involved residue in each protein. Some of these residues, particularly for the EGF-like family, have been reported as structurally or functionally important for these proteins.

Table 2. Percentage of folds in classes

	α protein (%)	β protein (%)	α/β protein (%)	α+β protein (%)	Small protein (%)
Class A	70	–	30	–	–
Class B	–	–	100	–	–
Class C	50	50	–	–	–
Class D	20	45	–	–	35
Class E	100	–	–	–	–
Class F	85	–	–	–	15
Class G	–	–	40	60	–
Class H	–	50	–	50	–
Class I	12	12	–	–	76
Class J	–	25	–	75	–
Class K	–	100	–	–	–
Class L	8	5	2	–	85
Class M	100	–	–	–	–

Relationships between members of the same family in a cluster

Two large families of Cluster 3 have been selected for a more in-depth analysis: the Defensin-like family and the EGF-like family, previously defined by SCOP [59]. It is noteworthy that, in spite of belonging to the same disulphide topology cluster, the proteins of each of these families show low sequential homology among themselves. The analysis shows that many residues from distinct members of each family conserve related physicochemical groups in similar positions of 3D space (Figure 2), these residues being described as functionally important [30]. Thus, the superimposition of the members of the Defensin-like family [60] shows 7 amino acid locations with meaningful consensus (percentage of appearance greater than 55%). The structural or functional significance of such residue equivalence in the Defensin family is presently unknown. On the other hand, the superimposition of the members of the EGF-like family shows 15 equivalent locations, most of them reported to be unambiguously related to function [51, 52, 55, 57].

Discussion

The increasing number of known tertiary structures makes it necessary to design methods for protein structure comparison. This is because the resemblance of protein 3D structures can provide clues as to structural and functional properties or evolving trends previously hidden from current sequence-alignment algorithms. An example of using structure for obtaining evidence about the function is the product of the *Obese* gene which, upon modelling, has been predicted to be a helical cytokine, thus having important biological and clinical consequences [61]. Here, a method has been designed to classify proteins by their disulphide-bridge topology, a fact that may facilitate the above mentioned studies. Although other approaches and classifications based upon disulphide bridges have been previously reported [32–34], our procedure does not additionally require the presence of regular secondary structures in the proteins [32] or the knowledge of the sequence of linked cysteines [33, 34] to perform the analysis. Therefore, in order to compare similarities between two proteins, neither the connectivity nor the relative positions of the cystines are considered. This fact allows the versatile detection of similar spatial positioning of cysteine residues even in the absence of similar disulphide dispositions in the sequence. This feature could have particular value where the disulphide bond connectivity may be ambiguous or unusual.

The preferential occurrence of certain disulphide-bridge topologies has been observed in a database of proteins with more than two disulphides and less then 65% homology. Investigation of whether these preferences has been used to group proteins and to study the possible relationships among them. Our classification relates proteins classified into different families defined by conventional protein classifications [59, 62]. In our opinion, the latter conventional classifications

are unable to deal with disulphide-rich proteins, especially for small proteins with low content of regular secondary structure, as previously observed by Harrison and Sternberg [32]. These proteins often are excluded or specially treated in traditional approaches for protein classification. For example, proteins such as 1bnb, 1dec, 1dfnA, 1erp, 1mtx, 1pnh, 1prhA, 2crd, among others, are not catalogued by automatic methods like CATH [62] probably because they do not have a large enough number of residues. Nevertheless, it has been made possible to assign them to different groups using our approach. It has been shown that in many cases the superimposition by disulphide-bridge topology allows for a correct alignment of regular secondary structures even when proteins from different families, but with related folds, are compared (i.e., when comparing PCI and EGF or PCI and defensins). In this respect, it should also be mentioned that preferences for a certain disulphide-bridge topology may be useful in structure prediction for filtering some of the matches in fold-recognition procedures [59, 62–64].

Structural comparisons of members of the same families show that the disulphide-bridge topology can be useful in aligning important residues for proteins within these families. A good example is found for the T-Knot motif [9, 43, 44]. Several authors have described the T-Knot structural motif by its cystine pattern in sequence [65] or by the presence of a β-sheet tied by two cystines [32]. In contrast, our approach, based only on the topology of three disulphide bridges, relates a larger number of proteins as T-knot. This new definition/detection of a T-knot motif could be complementary (or more general) to the approaches described by other authors.

Moreover, do the proteins of the clusters shown in Figure 1 share functional characteristics? Most of them are eukaryotic extracellular hydrolases and inhibitors, toxins, hormones and growth factors and, in general, proteins whose functional category could be classified as cell-to-cell recognition, cell signaling, cell defense, etc., that is, functions not devoted to basic metabolism. This category of proteins is prone to shuffling and sharing entire domains (for example, there are several kringle modules on the list). In any case, a number of them share at least the disulphide-bridge core. It is probable that our clusters encode more structural and functional relationships that escape our analysis and interpretation. To decode those structural and functional relationships is a challenge to the theoretical and experimental researchers.

As stated above, one of the potential merits of the programs for protein structure superimposition is that they can find unsuspected structural relationships, which can lead to the discovery of important functional properties. This could be the case of PCI, the potato carboxypeptidase inhibitor, which shares – as shown above – a similar three-dimensional distribution in certain amino acid residues with growth factors [30].

The KNOT-MATCH program facilitates a disulphide-based 3D overlapping of proteins to visualise them with a graphical program or to perform other computer-based analyses. The KNOT-MATCH program can be downloaded from our web/ftp address (see Methods section).

Acknowledgements

The authors gratefully acknowledge Professor T. Blundell and Dr R. Sowdhamini (University of Cambridge), Dr M. J. E. Sternberg (Imperial Cancer Research Fund, London) and Dr F.X. Gomis-Rüth (Centre d'Investigacions i Desenvolpament, CSIC, Barcelona) for helpful suggestions and advice about the present work. This research has been supported by Grants BIO2000-0647, BIO98-0362, and IN94-0347 from the CICYT (Ministerio de Educación y Ciencia, Spain), by the CERBA (Centre de Referència en Biotecnología) de la Generalitat de Catalunya and by Fundació F. de Roviralta. J.M.M is a predoctoral fellowship recipient of the CIRIT-CERBA (Generalitat de Catalunya).

References

1. Bork, P., Ouzounis, C. and Sander, C., Curr. Opin. Struct. Biol., 4 (1994) 393.
2. Chothia, C. and Lesk, A.M., EMBO J., 5 (1986) 823.
3. Rost, B., Fold. Des., 2 (1997) 19.
4. Bork, P., Sander, C. and Valencia, A., Proc. Natl. Acad. Sci. USA, 89 (1992) 7290.
5. Pascarella, S. and Argos, P., J. Mol. Biol., 224 (1992) 461.
6. Pascarella, S. and Argos, P., Prot. Eng., 5 (1992) 121.
7. Holm, L., Sander, C., Schnarr, M., Rüterjans, H., Fogh, R., Boelens, R. and Kaptein, R., Prot. Eng., 7 (1993) 1449.
8. Holm, L., Murzin, A.G. and Sander, C., Nature Struct. Biol., 1 (1994) 146.
9. Sun, P.D. and Davies, D.R., Annu. Rev. Biophys. Biomol. Struct., 24 (1995) 269.
10. Rossmann, M.G. and Argos, P.J., Mol. Biol., 105 (1976) 75.
11. Matthews, B.W. and Rossmann, M.G., Meth. Enzymol., 115 (1985) 397.
12. Taylor, W.R. and Orengo, C.A., J. Mol. Biol., 208 (1989) 1.

13. Zuker, M. and Somorjai, R.L., Bull. Math. Biol., 51 (1989) 55.
14. Brenner, S.E., Trends. Genet., 11 (1995) 635.
15. Vriend, G. and Sander, C., Proteins Struct. Func. Gen., 11 (1991) 52.
16. Orengo, C.A., Brown, N.P. and Taylor, W.T., Proteins, 14 (1992) 139.
17. Alexandrov, N., Takahashi, K. and Go, N., J. Mol. Biol. 225 (1992) 5
18. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G., Protein Sci., 1 (1992) 1691.
19. Holm, L. and Sander, C., Nucleic Acid Res., 24 (1996) 206.
20. Russell, R.B., Saqi, M.A.S., Sayle, R.A., Bates, P.A. and Sternberg, M.J.E., J. Mol. Biol., 269 (1997) 423.
21. Johnson, M.S., Srinivasan, N., Sowdhamini, R. and Blundell, T.L., Biochem. Mol. Biol., 29 (1994) 1.
22. Srinivasan, N., Bax, B., Blundell, T.L. and Parker, P., J. Proteins Struct. Funct. Gen., 26 (1996) 217.
23. Chatrenet, B. and Chang, J., J. Biol. Chem., 267 (1992) 3038.
24. Chang, J., Canals, F., Shindler, P., Querol, E. and Avilés, F.X., J. Biol. Chem., 269 (1994) 33087.
25. Chang, J., Shindler, P., Ramseier, U. and Lai, P., J. Biol. Chem., 270 (1995) 9207.
26. Betz, S., Protein Sci., 2 (1993) 1551.
27. Hrabal, R., Chen, Z., James, S., Bennet, H.P.J. and Feng, N., Nature Struct. Biol., 3 (1996) 747.
28. Murray-Rust, J., McDonald, N.Q., Blundell, T.L., Hosang, M., Oefner, C., Winkler, F. and Bradshaw, R.A., Structure, 1 (1993) 153.
29. Vita, C., Toma, F. and Menez, A., Proc. Natl Acad. Sci. USA, 92 (1995) 6404.
30. Mas, J.M., Aloy, P., Martí-Renom, M.A., Oliva, B., Blanco-Aparicio, C., Molina, M.A., de Llorens, R., Querol, E. and Avilés, F.X., J. Mol. Biol., 284 (1998) 541.
31. Harrison, P.M. and Sternberg, M.J.E., J. Mol. Biol., 244 (1994) 448.
32. Harrison, P. and Sternberg, M.J.E., J. Mol. Biol., 264 (1996) 603.
33. Thornton, J.M., J. Mol. Biol., 151 (1981) 261.
34. Richardson, J.S., Adv. Prot. Chem., 34 (1981) 167.
35. Srinivisan, N., Sowdhamini, R., Ramakrishnan, C. and Balaram, P., Int. J. Pept. Prot. Res., 36 (1990) 147.
36. Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
37. Hobohm, U., Scharf, M., Schneider, R. and Sander, C., Protein Sci. 1 (1992) 409.
38. Hobohm, U. and Sander, C., Protein Sci., 3 (1994) 522.
39. Everit, B. (Heinemann Educational Books Ltd), Cluster analysis, London, U.K., 1974.
40. Sali, A. and Blundell, T.L., J. Mol. Biol., 212 (1990) 403.
41. Subbarao, N. and Haneef, I., Protein Eng., 4 (1991) 877.
42. Holm, L. and Sander, C., J. Mol. Biol., 223 (1993) 123.
43. Narasimhan, L., Singh, J., Humblet, C. and Blundell, T.L., Nature Struct. Biol., 1 (1994) 850.
44. Isaacs, N.W., Curr. Opin. Struct. Biol., 5 (1995) 391.
45. Roussel, A., Inisan, A. and Knoops-Mouthy, E., TURBO FORBO version 5.0a Manual, BIOGRAPHICS, Marseille, France: Technopole de Chateaux-Gombert, 1994.
46. Rees, D.C. and Lipscomb, W.N., J. Mol. Biol., 160 (1982) 475.
47. Brown, S.C., Mueller, L. and Jeffs, P.W., Biochemistry, 28 (1989) 593.
48. Brown, L.R. and Wüthrich, K., J. Mol. Biol., 227 (1992) 1118.
49. Montelione, G.T., Wüthrich, K., Burgess, A.W., Nice, E.C., Wagner, G., Gibson, K.D. and Scheraga, H.A., Biochemistry, 31 (1992) 236.
50. Ullner, M., Selander, M., Persson, E., Stenflo, J., Drakenberg, T. and Teleman, O., Biochemistry, 31 (1992) 5974.
51. Groenen, L.G., Nice, E.C. and Burgess, A.W., Growth Factors, 11 (1994) 235.
52. Nogata, K., Kohda, D., Hatanaka, H., Saori, I., Matsuda, S., Yamamoto, T., Suzuki, A. and Inagaki, F., EMBO J., 13 (1994) 3517.
53. Picot, D., Loll, P.J. and Garavito, R.M., Nature, 367 (1994) 243.
54. Molina, M.A., Marino, C., Oliva, B., Avilés, F.X. and Querol, E., J. Biol. Chem., 269 (1994) 21467.
55. Barbacci, E., Guarino, B., Stroh, J., Singleton, D., Rosnack., J. and Moyer, J., J. Biol. Chem., 16 (1995) 9585.
56. Jacobsen, N.E., Abadi, N., Sliwkowski, M.X., Reilly, D., Skelton, N.J. and Fairbrother, W.J., Biochemistry, 35 (1996) 3402.
57. McInnes, C., Hoyt, D.W., Harkins, R.N., Pagila, R.N., Debanne, M.T., O'Connor-McCourt, M. and Sykes, D., J. Biol. Chem., 271 (1996) 32204.
58. Blanco-Aparicio, C., Molina, M., Fernandez-Salas, E., Frazier, M., Mas, J., Querol, E., Avilés, F.X. and de Llorens, R., J. Biol. Chem., 20 (1998) 1237.
59. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., J. Mol. Biol., 247 (1995) 536.
60. Hill, C.P., Yee, J., Slested, M.E. and Eisenberg, D., Science, 251 (1991) 1481.
61. Madej, T., Boguski, M.S. and Bryant, S.H., FEBS Lett., 373 (1995) 356.
62. Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B. and Thornton, J.M., Structure, 8 (1997) 1093.
63. Cohen, F.E., Gregoret, L., Presuell, S.B. and Kuntz, I.D., J. Mol. Biol., 289 (1989) 75.
64. Rufino, S.D. and Blundell, T.L., J. Comput. Aid. Mol. Des., 8 (1994) 5.
65. Lin, S.L. and Nussinov, R., Nature Struct. Biol., 2 (1995) 835.