Towards an Intelligent System for the Automatic Assignment of Domains in Globular Proteins

Michael J E Sternberg, Hedvig Hegyi, Suhail A Islam,

Jingchu Luo and Robert B Russell

Biomolecular Modelling Laboratory
Imperial Cancer Research Fund
Lincoln's Inn Fields, London WC2A 3PX, UK
m_sternberg@icrf.icnet.uk

Abstract

The automatic identification of protein domains from coordinates is the first step in the classification of protein folds and hence is required for databases to guide structure prediction. Most algorithms encode a single concept based and sometimes do not yield assignments that are consistent with the generally accepted perception. Our development of an automatic approach to identify reliably domains from protein coordinates is described. The algorithm is benchmarked against a manual identification of the domains in 284 representative protein chains. The first step is the domain assignment by distance (DAD) algorithm that considers the density of inter-residue contacts represented in a contact matrix. The algorithm yields 85% agreement with the manual assignment. The paper then considers how the reliability of these assignments could be evaluated. Finally the use of structural comparisons using the STAMP algorithm to validate domain assignment is reported on a test case.

1 Introduction

The concept of a structural domain, first formalised by Wetlaufer (1973), is central to the description, comparison and prediction of protein structure (Orengo et al., 1994). However attempts to develop automatic algorithms to identify domains from coordinates have been limited in their capacity to reproduce the concepts used by workers employing visual inspection. In this paper we report our development of a strategy for domain assignment based on incorporating different sources of information.

The concept of a domain used by Wetlaufer in his visual inspection of 18 protein structures is that the polypeptide chain folds into distinct structural regions that can be enclosed in a compact volume. A region could be a single chain segment (i.e. a continuous domain) or formed from several segments (i.e. discontinuous). Often these domains were spatially separate parts of the protein chain. A subsequent and important review of protein structures was reported by Richardson (1981) who visually examined some 100 proteins and identified domains. Her assignment was that the domain could be independently stable and/or should be able to undergo rigid body motion. However these concepts were limited and many domains were assigned by a boot-strap procedure - i.e. whether the putative domain of one protein resembled the entire chain of a single domain protein.

These two seminal papers have provided the guidelines by which workers identify domains by inspection. Our reading of the literature suggests the following concepts are frequently used to identify domains:

i) a domain is often a spatially separate region of the chain

ii) a domain in one protein may resemble the entire chain of another protein in terms of size and packing of secondary structures

iii) each domain in a protein has a specific function

iv) a substructure in one protein resembles a domain in another protein that meets one or more of the criteria i , ii or iii.

v) the protein possess a repeating substructure and this repeat tends to meet one or more of the criteria i, ii or iii.

Clearly these are subjective criteria, but there is substantial consensus in the field as to how they are applied.

A variety of algorithms have been developed to identify automatically domains from coordinates (Rose, 1979; Wodak and Janin, 1981; Rashin, 1981; Sander, 1981; Zehfus, 1994; Zehfus and Rose, 1986; Holm and Sander, 1994; Sowdhamini and Blundell, 1995; Islam et

al., 1995; and in the QUANTA modelling package). All these approaches quantify concept (i) - the spatial separation of domains. No approach consistently yields assignments that concur with the authors' definitions. Some of the algorithm repeatedly cuts the chain leading to small segments. In others the algorithm has not been benchmarked as to its agreement with the domain assignments in the literature.

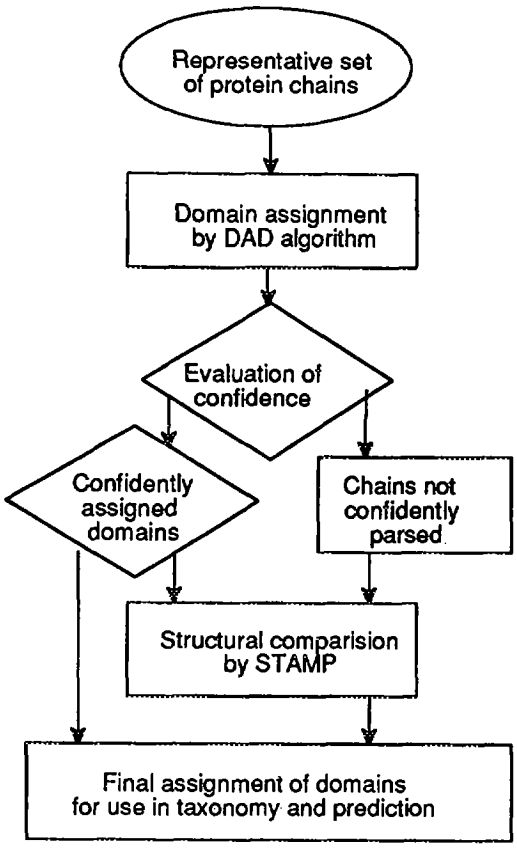


Figure 1 - Flow chart of domain assignment

Our aim is to develop an automatic algorithm for domain assignment that incorporates the concepts actually used by workers (see Figure1). We start with our computer algorithm (DAD) (Islam et al., 1995) that automatically identifies domains from coordinates based on concept (i). DAD is based in inter-residue distances and uses the contact plot method of Sander (1981). The results of this automatic approach are compared to a manual assignment of domains obtained from the literature and graphics inspection. We show that an accuracy of 85% can be achieved by this approach. The next step is to develop computational tools to identify which of the domain assignments from the DAD algorithm are reliable and which need further validation. The third step recognises that domains can also be defined by structural similarity. We propose the use of a structure comparison algorithm to aid domain assignment. The aim is that domains that are confidently assigned by the DAD algorithm can be compared to a database of structures to find similarities. We show how structure comparison can reveal domains which, although similar in three-dimensional structure, do not necessarily occur as spatially distinct units. We thus explore how concept (iv) can be included into the approach. Thus our goal is to develop tools that can reliably parse the protein coordinate sets to yield accurate domain assignments by incorporation of different sources of information.

2 Methods

2.1 Data set of domain assignments from the literature

The dataset consisted of 284 protein chains based on the 301 non redundant list of chains developed by Hobohm and Sander (1994). The list was based on coordinates available from the October 1993 release of the Brookhaven databank (Bernstein et al., 1977). (Two proteins in their list each consisted of two chains that were merged into a single set. Coordinates were unavailable for the remaining 15 chains.) We then followed Richardson's (1981) concept of domains and inspected the literature for authors assignments that, in our opinion, used similar concepts. Generally we were able to use the authors' definitions but sometimes we employed visual inspection on the graphics using our in-house program PREPI (S A Islam, unpublished but available on request). In some chains, we identified the end of one domain and the start of the next domain omitting the linking region that could not be assigned to either domain.

2.2 Domain Assignment by Distance (DAD) Algorithm

The domain assignment by a distance (DAD) algorithm cuts the chain into segment to minimise the density of inter-domain contacts in the contact plot .

Step 1 - Generation of a contact matrix

Consider a chain segment from residue a to b (for a complete chain of N residues a=1, b= N). For residues i and j , the entry in the contact matrix D_{ij} is 1 if the carbon α- carbon α distance is less than a cut off

distance d, or 0 otherwise. In a graphical representation of the matrix a point is plotted at ij if D_{ij} =1. Figure 2 shows an actual example of a domain contact map for alcohol dehydrogenase (8adh, Eklund et al., 1976).

Let the chain is divided into two domain a to k and k+1 to b, then the number of inter-domain contacts C_k, is given by:

$$C_k = \sum_{i=a}^{k} \sum_{j=k+1}^{b} D_{ij}$$

where k is an index denoting the cut position.
The maximum number of possible inter-domain contacts, M_k is given by:

$$M_k = (k - a + 1)\ (b - k)$$

The density of contact S_k is defined as C_k divided by the maximum and is:

$$S_k = C_k / M_k$$

Step 2 - Division of the chain

Calculate the value of S_k for k = a to b. A potential cut to the chain is made at the minimum value of S_k for k between a+15 and b - 15. To determine whether the chain should be subdivided, the average (A_{ab}) of S_k is then calculated by:

$$A_{a,b} = \sum_{k=a}^{b} S_k / (b - a + 1)$$

and a division made only if

$$S_k / A_{a,b} \leq F$$

where F is a cut-off value to be determined empirically.

Step 3 - Clustering into discontinuous segments

Step 2 yields a series of continuous chain segments each of which is a potential domain. Two or more of these segments might together form a discontinuous domain which is identified using the following clustering algorithm.

a) For every pair of continuous segments, calculate the value of $S_k / A_{1,N}$.

b) Store these values in a matrix X as a percentage.

c) Identify the maximum entry in X and if greater than F, then the corresponding two segments (say p and q) are assigned to the same domain, and the values in X is updated to include the assignment of segments p and q to the same domain.

d)This procedure is repeated until no further clustering of segments is allowed.

e) In addition, any segment of less than 32 residues is merged with the next section along the chain that is longer than 31 residues. A segment of less than 32 residues at the C-terminus is merged with the nearest preceding segment of more than 31 residues. As a result of this merging procedure, the algorithm will only identify domains longer than 31 residues.

Step 4 - Evaluation of Accuracy

The agreement between the program's assignments and the authors' assignments is taken as an accuracy score of the DAD algorithm Q_D:

$$Q_D = (N_{correct}/N_{total}) \times 100\%$$

where $N_{correct}$ is the number of residues identified in the authors' assignment as belonging to a particular domain that were correctly assigned to that domain by the program. A linking region in the authors' assignment is always taken as being correctly identified by the program. N_{total} is the maximum value that could be obtained for $N_{correct}$. A domain assignment to a protein is only considered correct if the number of domains found by the program is equal to the number of author-assigned domains and the accuracy $Q_D \geq 95\%$.

Runs of the program were performed for d ranging between 6.5 and 13 Å in steps of 0.5Å and for F from 40 to 65% in steps of 5% to obtain the optimum values for these two adjustable parameters assessed by the accuracy. The values selected are d = 11Å and F = 55%.

Legend to Figure 2 (next page) - Illustration of the DAD algorithm. 2a - A schematic diagram of 8adh (alcohol dehydrogenase, (Eklund et al., 1976)) showing the domain cuts according to the authors'. 2b - The distance plot with the running value of S_k and its average $A_{1,N}$. Below the plot is the secondary structure with arrows as the β-strands, rectangles as α-helices. The next line shows the cuts made. The final line gives the domain assignment after clustering. Fig 2c - Details of the cutting and clustering procedure. The table is the X matrix so the first clustering is performed between segments S1 and S3; subsequent clusterings are not shown.

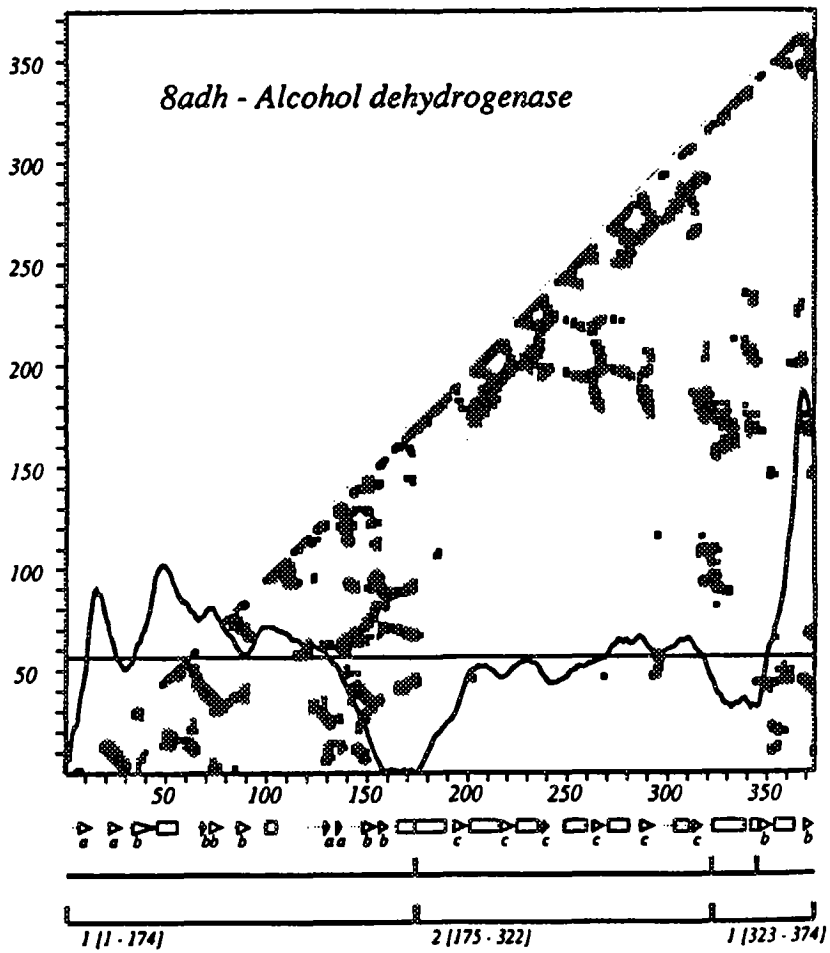


Figure 2b

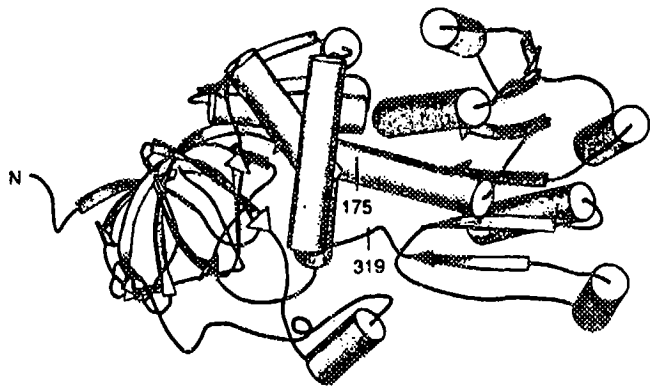
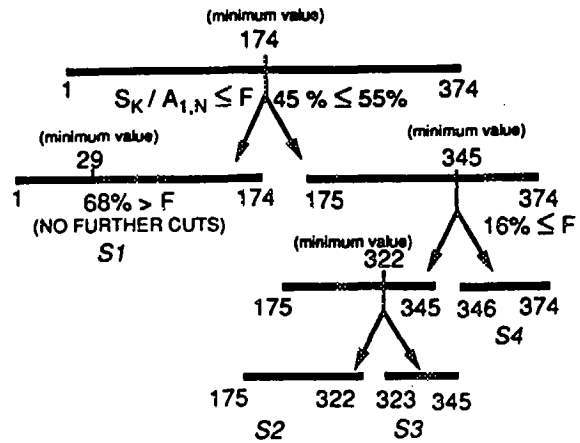


Figure 2a



	S1	S2	S3	S4
S1	0	15	138	122
S2		0	71	23
S3			0	62
S4				0

Figure 2c

2.3 Structural Comparison Algorithm - STAMP

Structure comparisons were performed using the STAMP package (Russell and Barton, 1992). Briefly, the method uses an initial alignment of query (search) and database structures to derive an initial superimposition, which is then refined to give an accurate set of structurally equivalent residues. The query sequence is overlayed on to the database sequence starting at every 5th database residue. The aligned (equivalenced) Cα atoms are used to derive an initial fit (superimposition), which is refined first by a conformation biased fit, then by a combined distance and conformation fit. Dynamic programming is used to provide a final set of structurally equivalent residues, and an overall structural similarity score (Q_S). In this study, Q_S was defined as:

$$Q_S = (T/Lq)\ (Lq - Gq)/Lq$$

Where T is the sum of residue similarities returned by dynamic programming, Lq is the length of the query structure (in residues), and Gq is the number of query residues aligned to a gap in the structural alignment. Q_S values of greater than 2.5 generally suggest structural similarity with strong similarities yielding scores above 3.5. The (Lq-Gq)/Lq term penalises those structural alignments not containing a significant portion of the query structure (i.e. short, local similarities not implying overall structural similarity). This scoring scheme was chosen because similarities involving small portions of the query domain are unlikely to aid domain assignment.

This method has been used to detect several previously un-noticed similarities between protein three-dimensional structures. Similarities have been detected between: a) the SH2 domain and domain II of *E. coli bio* operon protein(Russell and Barton, 1993); b) between HIV matrix protein p17 and interferon γ (Matthews et al., 1994); c) between domain B of disulphide bond forming protein and the C-terminal domain of thermolysin (Russell, 1994); and between the smaller domain from viral coat protein VP7 and phaseolin (Russell, 1994).

3 Results

3.1 The DAD Algorithm

Table 1 gives the agreement between the domain definition by the authors' and the identification by DAD algorithm for the representative set of 284 proteins using the optimal choice of d and F. We reported two scores - the number of times the DAD algorithm and the authors concur as to the number of domains and also imposing in addition the stricter criterion of an accuracy Q_D of ≥ 95%. With the stricter requirement, the domain assignments of 222 of the 284 chains (i.e. 78%) were correctly identified by the algorithm. 182 out of the 197 single domains (i.e. 92%) were correctly identified and 39 of the 67 two domains (i.e. 58%) were identified with $Q_D \geq 95\%$. The identification of chains split into three or more domains was poorer. Full details of the authors' and the DAD program's assignment of domains are available upon request from us.

The assignment of domains according to our interpretation of the authors' description (occasionally supplemented by inspection of the graphics) is of course subjective. We examined in detail every chain for which there was a disagreement between the number of domains assigned by the authors and the program. For 15 chains inspection suggested that there was an acceptable alternative definition as to the number of domains in the protein chain which was consistent with the results of the algorithm. For a further four chains although the number of domains was the same, there was an alternative segment definition. In addition, there was one protein (code 1aaf - HIV nucleocapsid protein) that was non globular and so the usual concept of domain does not apply. If the program is considered correct for these 20 chains, the final accuracy is 242 out of 284 chains (i.e.85%).

Inspection of the structures of the chains for which the DAD algorithm failed showed a variety of reasons for the errors. A frequent difficulty was that the domain was compact but was not spatially separate from the rest of the protein. This could be clearly seen in the identification of the dinucleotide binding domain (DNBD) (Rossmann et al., 1974). This domain consists of at least six parallel β-strands with order 321456 where most of the right-handed connections are βαβ units (Richardson, 1976; Sternberg and Thornton, 1976). In both 8atc (aspartate carbamoyltransferase chain A) and in 6ldh (lactate dehydrogenase) the algorithm did not identify this dinucleotide domain due to close packing with other parts of the chain. In contrast, the domain was identified in (1hsd) hydroxysteroid dehydrogenase, 8adh (alcohol dehydrogenase), in 4gpd (glyceraldehyde 3-phosphate dehydrogenase) and as part of a larger domain in 1gpb (glycogen phosphorylase).

We have compared (see Islam et al, 1995) our assignments with those of other workers (Rose, 1979; Rashin 1981;Wodak and Janin, 1981; Sander, 1981; Holm and Sander, 1994; Sowdhamini and Blundell, 1995). In general the locations of the cuts agree and the main difference is the decision whether to cut the chain or not. We note that Holm and Sander (1994) correctly identify the domain structure (our interpretation of the authors' assignment) of 8adh, 8atc and 4gpd but not of 6ldh, 1hsd and 1gpb.

3.2 Tools for confidence of assignment by DAD

To be useful for automatically parsing protein coordinates, one needs to know whether a particular DAD assignment is expected to be accurate. We are exploring the following approaches to estimate the expected accuracy.

i) A single inter-residue distance (d) and cut off value (F) were used by the DAD algorithm. We are evaluating whether a consistent definition resulting from runs at different cut offs provides a guide to reliability. For example the program yielded consistent results on 8adh over a range of d from 6 to 13 Å suggesting a confident assignment. In addition, often a different choice of these parameters would have yielded the authors' assignment. Thus a range of these parameters can be used to yield alternative definitions for subsequent selection.

ii) The value of the cut (S_k / $A_{a,b}$) could provide further insight into reliability. For the first putative cut, a high suggests that the chain can reliably be considered a single domain. Similarly if the DAD algorithm only implements one cut with a low value, this suggests two clearly separate domains.

iii) An alternative approach for domain assignment is to cluster secondary structures (Sowdhamini and Blundell, 1995; and QUANTA). Inspection of these results suggests that often the domain is formed from several discontinuous regions which leads to differences from most author's assignments. However if the two algorithms agree, the assignment could be considered reliable.

No of domains in chain with its frequency in authors' assignment	No of domains assigned by DAD algoritm				
	1	2	3	4	5
1 197	182(182)	15	0	0	0
2 67	20	45(39)	2	0	0
3 13	4	8	1(1)	0	0
4 6	0	5	1	0(0)	0
5 1	0	1	0	0	0(0)

Table 1 - Accuracy of domain assignment by DAD algorithm

The entries denote the number of times a chain is split into a given number of domains by the authors' and a given number by the DAD algorithm. Numbers in brackets refer to when the algorithm both agrees as to number of domains and yields an accuracy $Q_D > 95\%$

3.3 Domain assignment by structural comparison (STAMP)

The observation on the identification of the dinucleotide binding domain (DNBD) highlights that often one can readily identify a domain in one chain by virtue of its compactness and spatial separation. The presence of a similar substructure in another chain that is less spatially separate should also be identified as a domain. We explored whether recognition of structural similarity between an identified domain and substructures within chains could be used to test the internal consistency of the DAD algorithm for domain assignment.

Here we report one test case. We started with the algorithm's assignment of the dinucleotide binding domain in 8adh (residues 175 to 322). This assignment would have been found throughout the range of d values from 6 to 13Å and would thus be considered as a confident assignment. This domain was compared structurally using STAMP with every other protein chain.

For each probe protein chain:

a) The score for structural equivalence was calculated (Q_S)

b) For each domain assigned by the DAD algorithm
   i) Calculate the total number of residues in the domain (N_d)
   ii) Identify the number of residues in the domain that were structurally equivalenced to the protein chain (N_{sd})
   iii) hence evaluate the fraction of the number of residues in the domain that were equivalenced to the probe chain (N_{sd} / N_d)

c) For each domain assigned by the authors
   i) Calculate the total number of residues in the domain (N_a)
   ii) Identify the number of residues in the domain that were structurally equivalenced to the protein chain (N_{sa})
   iii) hence evaluate the fraction of the number of residues in the domain that were equivalenced to the probe chain (N_{sa} / N_a)

Table 2 gives the results for scores of $Q_S \geq 3.5$. Previous experience with structural comparisons suggested that above this score one can be confident that the structural equivalence has identified a close similarity between substantial segments of the chain. Table 2 also gives the accuracy score (Q_D) comparing the DAD algorithm with the authors' assignments. Whenever $N_{sd}/N_d < 0.67$, the DAD algorithm had previously correctly identified the domain. This is because the size of the segment(s) found by STAMP to be structurally equivalence to the DNBD of 8adh is not much less than the total number of residues in the segment identified as a domain by DAD. In contrast, whenever $N_{sd}/N_d < 0.5$, the DAD algorithm failed to split the chain into domains whereas STAMP found that only part of the chain could be equivalenced to the DNBD of 8adh. The best examples are that the six-stranded DNBD of 8atc (aspartate carbamoyltransferase) and of 6ldh (lactate dehydrogenase) are identified by STAMP but were not identified by DAD. The other examples in the Table in which STAMP finds a domain previously not identified by DAD are of folds with five β-strands with α-helical connections that have the same topology as five of the six strands of the DNBD.

Protein +chain code	STAMP score Q_S	STAMP with DAD N_{sd}	Ns_d/N_d	Acc DAD Q_D%	STAMP with author N_{sa}	Ns_a/N_a	STAMP finds
1hsdA	4.81	177	0.69	100	177	0.69	DNBD & agrees DAD
8atcA	4.71	143	0.46	-1	134	0.95	DNBD & corrects DAD
8abp	4.70	142	0.84	97	142	0.85	Related fold & agrees DAD
1dri	4.43	128	0.95	95	128	0.92	Related fold & agrees DAD
1minA	4.16	98	0.21	-1	98	0.66	Related fold & corrects DAD
1minB	4.13	104	0.21	-1	103	0.65	Related fold & corrects DAD
3gbp	4.05	143	0.87	98	143	0.88	Related fold & agrees DAD
6ldh	4.02	139	0.42	-1	139	0.85	DNBD & corrects DAD
1wsyB	3.98	105	0.27	-1	105	0.65	Related fold & corrects DAD
4gpd1	3.52	116	0.98	90	145	0.98	DNBD & roughly agrees DAD

Table 2 - Domain assignment by structural comparision

Protein chains are: 1hsd (hydroxysteroid dehydrogenase); 8atc (aspartate carbamoyltransferase); 8abp (L-arabinose binding protein); 1dri (D-ribose binding protein); 1min (nitrogenase molybdenum-iron protein); 3gpb (galactose binding protein); 6ldh (lactate dehydrogeanase); 1wsy (tryptophan synthase); 4gpd (glyceraldehyde 3-phosphate dehydrogenase).

4 Discussion and Conclusion

Given the rapid increase in the number of protein coordinates, it is becoming essential to develop tools to identify structural features. The assignment of domains is central to comparative taxonomy and to prediction of folds by threading (Orengo et al., 1994). Whilst manual assignments are of course subjective, we consider that they still reflect many concepts that are useful. The presently-available algorithm for domain identification implement a single concept and do not yield assignments that concur sufficiently with manual assignments. In this paper we show that structural comparisons can augment an algorithm based on spatial separation. The different sources of information that need to be encapsulated in manual domain identification pose a challenge for algorithm development.

References

Bernstein, F. C., Koetzle, T. F., Williams, G., Meyer, D. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. 1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535-542.

Eklund, H., Nordstrom, B., Zeppezauer, E., Soderlund, G., Ohlsson, I., Boiwe, T., Soderberg, B.-O., Tapia, O., Branden, C.-I. and Akeson, A. 1976. Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4Å resolution. *J.Mol.Biol.* 102, 27-59.

Hobohm, U. and Sander, C. 1994. Enlarged representative set of protein structures. *Prot. Sci.* 3, 522-524.

Holm, L. and Sander, C. 1994. Parser for protein folding units. *Proteins* 19, 256-268.

Islam, S. A., Luo, J. and Sternberg, M. J. E. 1995. Identification and analysis of domains in proteins. *Protein Engineering* in the press.

Matthews, S., Barlow, P., Boyd, J., Barton, R., Russell, R., Mills, H., Cunningham, M., Meyers, N., Burns, N., Clark, N., Kingsman, S., Kingsman, A. and Campbell, I. D. 1994. Three-dimensional structure of HIV-1 matrix protein p17 reveals similarity to interferon-g. *Nature* 370, 666-668.

Orengo, C. A., Jones, D. T. and Thornton, J. M. 1994. Protein superfamilies and domain superfolds. *Nature* 372, 631-634.

QUANTA. Molecular Simulations Inc, Waltham, Mass USA.

Rashin, A.A. 1981) *Nature* 291, 167-339.

Richardson, J. S. 1976. Handedness of crossover connections in beta sheets. *Proc Nat Acad Sci , USA* 73, 2619-2623.

Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 34, 167-339.

Rose, G. D. 1979. Hierarchic organization of domains in globular proteins. *J. Mol. Biol.* 134, 447-470.

Rossmann, M. G., Moras, D. and Olsen, K. W. 1994. Chemical and biological evolution of a nucleotide-binding protein. *Nature* 250, 194-199.

Russell, R. B. 1994. Domain insertion. *Protein Engineering* 7, 1407-1410.

Russell, R. B. and Barton, G. J. 1992. Multiple sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14, 309-323.

Russell, R. B. and Barton, G. J. 1993. An SH2-SH3 domain hybrid. *Nature* 364, 765.

Sander, C. 1981. Physical criteria for folding units of globular proteins. In Structural Aspects of Recognition and Assembly in Biological Macromolecules, ed. Balaban, M., Sussman, J. L., Traub, W. and Yonath, A., Balaban ISS, Rehovot and Philadelphia. pp. 183-195

Sowdhamini, R. and Blundell, T. L. 1995. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Science* in the press,

Sternberg, M. J. E. and Thornton, J. M. 1976. On the conformation of proteins: the handedness of the beta-strand / alpha-helix / beta-strand unit. *J Mol Biol* 105, 367-382.

Wetlaufer, D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA* 70, 697-701.

Wodak, S. J. and Janin, J. 1981. Location of structural domains in proteins. *Biochemistry* 20, 6544-6552.

Zehfus, M. H. 1994. Binary discontinuous compact protein domains. *Prot.ein Engineering* . 7, 335-340.

Zehfus, M. H. and Rose, G. D. 1986. Compact units in proteins. *Biochemistry* 25, 5759-5765.