Alignment of a sparse protein signature with protein sequences: application to fold prediction for three small globulins

S.C. Daniel, J.H. Parish*, J.C. Ison[1], M.J. Blades, J.B.C. Findlay

School of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT, UK

Received 23 August 1999; received in revised form 7 September 1999

Abstract A novel algorithm has been developed for scoring the match between an imprecise sparse signature and all the protein sequences in a sequence database. The method was applied to a specific problem: signatures were derived from the probable folding nucleus and positions obtained from the determined interactions that occur during the folding of three small globular proteins and points of inter-element contact and sequence comparison of the actual three-dimensional structures of the same three proteins. In the case of two of these, lysozyme and myoglobin, the residues in the folding nucleus corresponded well to the key residues spotted by examination of the structures and in the remaining case, barnase, they did not. The diagnostic performance of the two types of signatures were compared for all three proteins. The significance of this for the application of an understanding of the protein folding mechanisms for structure prediction is discussed. The algorithm is generic and could be applied to other user-defined problems of sequence analysis.

Key words: Protein signature; Protein fold; Sequence alignment; Globin; c-Type lysozyme; Barnase

1. Introduction

There are several examples in the literature of attempts to derive sequence length motifs or sparse signatures that define features that relate the protein sequence to fold or other properties. Let us imagine that there exists such a signature or sparse motif in the following hypothetical family of three proteins where 'a' represents any amino acid and the sets of residues in upper case, (M and V), (K, N and S) and (N, V and Q), are 'key' residues.

1. MaaaaaaaaaaaaaKaaaaaaaNaa
2. VaaaaaaaaaNaaaaaaaaaQaaa
3. VaaaaaaaaSaaaaaaaaaaaNaaaaaaaa

Our signature in this case might be: (M,V)(8–13 residues)(K,N,S)(7–11 residues)(N,Q). We might also wish to note whether the key residues (e.g. M, K, N in protein (1)) are in secondary structure elements (SSEs), i.e. in helical or extended regions or in coils, or in unstructured regions, as this should affect the penalty incurred by stretching/contracting the distances. We might wish further to consider a more fuzzy descriptor. The problem of using a similarity matrix (such as PAM250 [1], BLOSUM [2] or RISLER [3]) to make the amino acid scoring quantitative rather than qualitative is trivial but the problem of regarding the distances as elastic is much more difficult by conventional methods because of the combinatorial explosion that follows from the uncertainty about the entries of the form 'N_1–N_2 residues'. We note that the requirement is for an algorithm that is fast enough to scan an entire database of protein sequences and also does not have a positional bias. To illustrate the potential bias error, consider the following. If we regard the signature as beads (key residues) on an elastic string, it is not legitimate to find the first 'reasonable' fit and then look for the next one as this would bias the significance of the key residues in the N-terminal region. In this paper, we provide a solution to this problem and hence, such sparse 'protein-sized motifs' become usable and feasible.

Strategies for classification and prediction of a structure employ simplified representations of the protein structure and include residue contact profiles [4], hydrophobicity patterns [5] classes of amino acids [6] and residue environments [7,8]. Recent important developments include the discovery of local packing motifs [9] and the prediction of the protein folding nucleus from the sequence [10]. A key feature of [10] is the recognition of hydrophobic interactions that occur during the folding process. We have been exploring a complementary approach in which we seek to identify the candidacy for key residues, those residues that are in multiple contact in interactions between SSEs. The methods have been implemented to create a database of such contacts and the details of this and a survey of signatures that can be generated are in preparation (Ison et al.). As the present paper concentrates on the alignment algorithm and its application to proteins with known topohydrophobic interactions, we provide the method here in outline only. A qualitative account is already published in the non-specialist literature [11]. From a 'training set' of proteins of known structure (there can be as few as one protein in this set), we first calculate the SSEs [12], then identify the SSEs in contact (as judged by loss of solvent accessibility [13]) and then, using the same criterion, the amino acid residues in contact. Those in multiple contact are marked up as putative key residues. The resulting signature is evaluated by using the new algorithm to score the success of the signature against a database and the success is quantified by comparing 'false' and 'true' hits and is refined by sequence alignment and a view of the equivalence of SSEs (in the case where there is more than one structure in the training set).

Of the proteins in the recent paper on the detection of the folding nucleus [10], we had already derived signatures for three: myoglobin, lysozyme/α-lactalbumin and barnase. These are the examples in what follows.

*Corresponding author. Fax: (44) (0) 113 233 3148.
E-mail: howard@bmb.leeds.ac.uk

[1] Present address: HGMP, Hinxton Hall, Hinxton, Cambridge CB10 1SB, UK.

PII: S0014-5793(99)01238-7

2. Materials and methods

2.1. Sequence database

We used the Leeds OWL [14] non-redundant sequence database. For reasons of computational efficiency, the database was re-written with a simple utility program as a binary stream file of the following type: each entry consists of (i) a protein code of fixed length, (ii) an integer L (length of the sequence in amino acid residues) and (iii) the sequence itself. The entries are ordered in increasing values of L.

2.2. Pattern matching algorithm

The protein of length N residues will be $P = \{a_1, a_2....a_N\}$, where a_i is an amino acid in position i, and the signature consisting of M positions will be $S = \{s_1, s_2....s_M\}$. s_i consists of an empirical distance set to g and an empirical residue set to r. The method to find the optimal alignment of S to P is an adaptation of conventional dynamic programming [15]. We only describe the distinguishing features of our algorithm here. Residues are only sampled if they fall within a 'window' of residues of a specified size around the empirical distances specified in g. In practice, two window sizes are specified for the signature, one each for inter- and intra-SSE distances (an inter-SSE distance spans a random space between two SSEs and an intra-SSE distance spans two positions within the same SSE). A score for each s_i is derived from the chosen residue substitution matrix and a distance penalty function:

$$s_i = ((r_1 \text{ vs. a}) + (r_2 \text{ vs. a}).... + (r_x \text{ vs. a}))/x$$

'a' denotes the amino acid residue the signature position is being matched to. r_1, r_2 etc. refer to a residue identity from the empirical residue set, consisting of x residues. 'r_1 vs. a' etc. includes a distance penalty comprising an 'initialisation' penalty (applied once to all distances deviating from those in g_i) and an extension penalty (applied for each residue in that deviation). As is the case for window sizes, values for distance initialisation/extension penalties are specified for inter- and intra-SSE distances separately. In the case of the first signature position, no distance penalty is applied, which allows for the identification of sequences with long pre-, pre-pro- or other N-terminal extensions. The score for a match between S and P is equal to the mean of the scores for individual matches for every position in the signature (s_i/M).

When reading OWL, those protein sequences whose length (in residues) does not fall within a specified range are discarded. The algorithm described above has been implemented in a computer program called SIGNATURE. The parameters of SIGNATURE which are under user control are summarised in Table 1. For any particular signature, these parameters are part of the signature. The software was written in C++ and is available, together with implementation notes from the URL http://www.bioinf.leeds.ac.uk/software.html.

The software was used on a 300 MHz PC running under Linux and the timed data (Table 3) should be interpreted in light of this.

Table 1
User-definable parameters for the SIGNATURE program

Parameter	Description
Permissible sequence length	A protein within OWL must fall within this range of lengths (in residues) to be considered by SIGNATURE during a database scan.
Window size	Specified separately for inter- and intra-SSE distances. During the construction of the alignment, residues are only sampled if they fall within this window size (residues) from each empirical distance.
Distance insertion penalty	Specified separately for inter- and intra-SSE distances. This penalty is applied to the alignment scores for each distance that deviates from those within the set of empirical distances.
Distance extension penalty	Specified separately for inter- and intra-SSE distances. This penalty is applied to the alignment scores for each single residue deviation from the appropriate empirical distance.
Residue substitution matrix	Used during the scoring of a signature to a protein sequence.

3. Results and discussion

The annotated sequences of the three proteins are shown in Table 2. The rows marked PKEY show the positions highlighted by Poupon and Mornon [10] as topohydrophobic positions. The rows marked SKEY are the key residues in signatures derived from inspection of SSE interactions. These residues are all present in SSEs. We sought to establish the extent to which Poupon and Mornon's view that 'understanding the mechanisms of protein folding would allow for prediction of the three-dimensional structure' [10]. In order to do this, we constructed a 'target set' of protein sequences which achieve a statistically significant similarity score when compared against at least one of the training set proteins using the FASTA program [16]. We use the definition of 'statistical significance' given by Pearson [17], i.e. pairs of proteins with an expectation value of less than 0.05 are considered to share significant similarity.

In the implementation of our method (Ison et al., in preparation), we use as a training set as many non-redundant structures as are available. For comparison with the data deduced from the hydrophobic folding nucleus [10], we repeated the work with the globins and the lysozymes with just one such protein (there is only one non-redundant barnase structure). The results of the analysis are shown in Table 3. The finding that the SKEY signature for lysozyme is slightly less diagnostic with the larger data set reflects the fact that recruiting more structures results in more false positives being found by using a larger range of distance choices. The PKEY results for barnase are not as bad as they appear because the detailed output from SIGNATURE lists all the proteins scored as either target or non-target: the results are distorted by the fact that the barnase family is small and the highest scoring 10 sequences are non-target. In order to compare the SKEY and PKEY results, we used only one sequence in each case. The SIGNATURE program will take files containing several key residue entries (of the sort illustrated in Table 2) and, for example, there were hence 32 such entries in the case of the globin data (Table 3).

Although it was not our intention to present in this paper SKEYs as an alternative to other sequence recognition and pattern matching methods, we examined briefly the performance of the SKEY signatures using the conventional analysis of 'coverage' as a function of 'error'. As one proceeds down the list of proteins in Table 3 (SKEY entries), a cut-off point is incremented from one to the maximum number of proteins tested (1000 in this case, see Table 3). For each point, the error is the proportion of false positives above the cut-off and the coverage is the ratio of (true positives above the cut-off)/(total). The results of two values of error are given in the last two rows of Table 3. The poor coverage at a low value of error in the case of lysozyme reflects the fact that there is a non-target in the early positions (low cut-off). In this particular case, we inspected the 'rogue' protein. It turned out to be a lysozyme from *Crax fasciolata* (bird of the family *Meliphagidae*, otherwise honey eaters).

We conclude that three generalisations follow from this work.

Table 2
Key residue signatures for three proteins

```
>>> myoglobin (PDB code 1mbd)   PKEY 21; SKEY 21; common 12.

PKEY                 |    |  |             |  ||           |
SEQU  1      VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASED
SSA4         ---HHHHHHHHHHHHHHH--HHHHHHHHHHHHHHH-HHHHHH---------HHHHHH-HH
SKEY             |        |  |           |     |   |     |

PKEY       |       |   |   |          |  |          |     |  |   |  ||
       61    LKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHP
SSA4         HHHHHHHHHHHHHHHHHH---HHHHHHHHHHHHHHH----HHHHHHHHHHHHHHHHHH-
SKEY       |  ||       |                  |   |          |  |   |   |

PKEY                   |       |   |
      121   GDFGADAQGAMNKALELFRKDIAAKYKELGYQG
SSA4         ----HHHHHHHHHHHHHHHHHHHHHHHHH----
SKEY                  ||    |  |

>>> lysozyme (PDB code 2eql)   PKEY 20; SKEY 12; common 5.

PKEY        |      |   |    |      |    ||         |                    || |
SEQU  1      KVFSKCELAHKLKAQEMDGFGGYSLANWVCMAEYESNFNTRAFNGKNANGSSDYGLFQLN
SSA4   1     ----HHHHHHHHH----------HHHHHHHHHHH-----------------------
SKEY           || |   |             |  | |

PKEY                         |     |   |   |     ||     |  |  |
       61   NKWWCKDNKRSSSNACNIMCSKLLDENIDDDISCAKRVVRDPKGMSAWKAWVKHCKDKDL
SSA4         HHHH----------------------HHHHHHHHHHH---------HHHHHHH-----
SKEY                                    |   |  |              |    |

PKEY
           SEYLASCNL
      121  ---------
SKEY

>>> barnase (PDB CODE 1bnj)   PKEY 11; SKEY 11; common 3.

PKEY                                                              |
SEQU    1    AQVINTFDGVADYLQTYHKLPDNYITKSEAQALGWVASKGNLADVAPGKSIGGDIFSNRE
SSA4         ------HHHHHHHHHH---------HHHHHHH-------HHHH----EEEEEE-----
SKEY              |   |  |                |          |        |

PKEY                   |  | |           |||       ||    |  |
       61    GKLPGKSGRTWREADINYTSGFRNSDRILYSSDWLIYKTTDHYQTFTKIR
SSA4         ----------EEEEE----------EEEEE----EEEE----------
SKEY                  | |               |      | |
```

Key residues in the predicted folding nucleus are merked | in the row (PKEY) above the sequence. The row SSA4 is the secondary structure assignment. At least four residues are required for the SSE to be recognised. E, extended; H, helical; -, others. Key residues in the signature derived from SSE interactions are marked | in the row SKEY below the SSA4 row. The title row in each case includes a summary of the key residues, thus, in the case of myoglobin, there are 22 and 21 residues in the PKEY and SKEY positions, respectively, and 12 of these are in common.

(i) We agree absolutely with [10] that the folding intermediate(s) provide a route to exploring sequence-fold relationships but conclude that these are necessary but not sufficient conditions. The caveat follows from proteins, of which barnase is an interesting example, where the topohydrophobic positions are in only one part of the sequence (entirely in the C-terminal half of the sequence in this case, Table 2) and this reflects the folding pathway for what is, in effect, a two domain structure [18].

(ii) The concept of a very sparse sequence signature to define a fold is justified by this work and that of [10]. To pursue the metaphor of a protein sequence being like a language with a three-dimensional meaning [11], there is as yet no real evidence that the language shares a common syntax for all protein families.

(iii) SIGNATURE is proved to be a robust and reasonably rapid method of evaluating 'elastic signatures'. Although written for projects such as the present one, the software could be used to try out any sparse signature and for testing theories of such signatures derived by a user. The method could complement methods, such as BLAST and Ψ-BLAST [19], which rely on absolute sequence identity for very short (triplet) sequences. It could also be used to try out inventive signatures generated by those who have spotted unusual features of proteins with certain properties in common. Signatures of the types shown in Table 2 complement more conventional motif databases such as PRINTS [20] and PROSITE [21].

Acknowledgements: Section 2 of this paper describes a method developed whilst S.C.D. was supported by the Nuffield Foundation Bursary Scheme. J.C.I. was a BBSRC CASE student with DL laboratory, Daresbury. We are grateful to Dr. A J Bleasby for help and discussions and to Dr D.R. Westhead for helpful criticism of the paper.

Table 3
Summary of the results of constructing and testing the signatures of Table 2

Protein (family)	Myoglobin (globins)			Lysozyme (c-type lysozyme)			Barnase (barnase)	
	SKEY data		PKEY data	SKEY data		PKEY data	SKEY data	PKEY data
Permissible sequence length (Table 1)	130–175		130–175	100–250		100–250	80–160	80–160
Distance insertion penalty (Table 1)	Inter-SSE 11		Inter-SSE 11	Inter-SSE 15		Inter-SSE 15	Inter-SSE 11	Inter-SSE 11
Number of proteins used for signatures	32	1	1	15	1	1	1	1
True positives in top scoring family size entries	100.0%	89.7%	66.5%	96.35%	97.8%	28.72%	100.0%	0.0%
Rank of highest scoring non-target	415	602	70	133	96	5	17	1
Rank of lowest scoring target	45	7948	29713	87711	57816	97827	16	70445
Average target score	26476	2460	2431	1651	1528	2070	1540	1027
Average non-target score	5384	1931	2059	1185	948	1825	988	1057
Scan time (min:s)	1:37	1:13	0:52	1:22	575	8:49	1:47	1:03
Coverage (0.03 error)	0.97			0.20			0.89	
Coverage (0.05 error)	0.98			0.96			0.89	

The entries in the last six rows are direct output from the program SIGNATURE. The parameters other than 'permissible sequence length' and 'distance insertion penalty' (Table 1) were the same for all these proteins. The window sizes were 10 for inter-SSE and two for intra-SSE, the distance extension penalties were 10 for both inter- and intra-SSE distances. The residue substitution matrix was that of [3]. The anomalous scan time for lysozyme PKEY data was due to CPU scheduling problems on the PC which was running an unrelated program concurrently.

References

[1] Rice, D.W. and Eisenberg, D. (1997) J. Mol. Biol. 267, 1026–1038.
[2] Fiser, A., Simon, I. and Barton, G.J. (1996) FEBS Lett. 397, 225–229.
[3] Risler, J.L., Delorme, M.O. and Delacroix, A. (1988) J. Mol. Biol. 204, 1019–1029.
[4] Ouzonis, C., Sander, C., Scharf, M. and Schneider, R. (1993) J. Mol. Biol. 232, 1–19.
[5] Bowie, J.U., Clarke, N.D., Pabo, C.O. and Sauer, R.T. (1990) Proteins 7, 257–264.
[6] Taylor, W.R. (1986) J. Mol. Biol. 88, 233–258.
[7] Bowie, J.U., Luthy, R. and Eisenberg, D. (1993) Science 253, 164–169.
[8] Johnson, M.S., Overington, J.P. and Blundell, T.L. (1993) J. Mol. Biol. 231, 735–752.
[9] Jonassen, I., Eidhammer, I. and Taylor, W.R. (1999) Proteins 34, 206–219.
[10] Poupon, A. and Mornon, J.-P. (1999) FEBS Lett. 452, 283–289.
[11] Parish, J.H. (1999) in: Visual Representations and Interpretations (Paton, R. and Neilson, I., Eds.), pp. 139–145, Springer Verlag, New York.
[12] Frishman, D. and Argos, P. (1995) Proteins 23, 566–579.
[13] Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. and Scharf, M. (1995) J. Comp. Chem. 16, 273–284.
[14] Bleasby, A.J. and Wootton, J.C. (1990) Protein Eng. 3, 153–159.
[15] Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443–453.
[16] Lipman, D.J. and Pearson, W.R. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[17] Pearson, W.R. (1995) Protein Sci. 4, 1145–1160.
[18] Serrano, L., Kellis Jr., J.T., Cann, P., Matouschek, A. and Fersht, A.R. (1992) J. Mol. Biol. 224, 783–804.
[19] Altschul, S.F. and Koonin, E.V. (1998) Trends Biochem. Sci. 23, 444–447.
[20] Bairoch, A., Bucher, P. and Hofmann, K. (1997) Nucleic Acids Res. 25, 217–221.
[21] Attwood, T.K., Beck, M.E., Bleasby, A.J. and Parry-Smith, D.J. (1994) Nucleic Acids Res. 22, 3590–3596.