Vol. 139, No. 1, 1986 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
August 29, 1986 Pages 11-17

A POSSIBLE WAY FOR PREDICTION OF DOMAIN BOUNDARIES IN GLOBULAR PROTEINS FROM AMINO ACID SEQUENCE

Ferenc VONDERVISZT and István SIMON

Institute of Enzymology, Biological Research Center
Hungarian Academy of Sciences, Budapest, P.O. Box 7 H-1502

Received June 26, 1986

SUMMARY: A simple approach to domain border prediction in globular proteins is outlined relying on the amino acid sequence only. Statistically determined sequential and association preference data of amino acids were combined to generate short range preference profiles along the polypeptide chains. Domain boundaries correlate with the minima of preference profiles, but some false minima also exist. Possibilities are discussed to exclude the false minima and to further improve the efficiency of the algorithm.

Almost all but the smallest globular proteins can be subdivided into geometrically separate entities, generally referred to as domains or structural domains (1,2). The same structural domain identified by its characteristic chain fold can occur in different proteins (3). Domains are also assumed to be the folding units of proteins (1). As recently realized relative domain movements play essential role in many protein functions (4).

Several methods exist to localize structural domains in globular proteins starting from atomic coordinates (2 and references therein). However, until recently only a single attempt has been made to predict domain boundaries from the amino acid sequence only (5). In this method the folding process of a protein is simulated by an iterative algorithm starting from predicted secondary structural units and computing their optimal association pattern yielding maximal formation energy. The calculations are rather complicated and in a number of cases domain boundaries are only very crudely assigned.

0006-291X/86 $1.50

In this paper we outline a simple procedure for prediction of domain limits from the amino acid sequence based on finding minima in plots of the statistically determined short range preferences between amino acids along the polypeptide chains.

MATERIALS

Sequence data for pig glyceraldehyde 3-phosphate dehydrogenase (GAPD), papain and human hemoglobin (Hb) β-chain were taken from (6) while domain boundary data from (2). Sequence information and the values of domain limits for mouse immunoglobulin G (IgG) MOPC 21 heavy chain were obtained from (7).
All calculations were performed on a Hewlett-Packard 9825A desktop computer.

BASIC APPROACH

Recently we have studied the sequential preferences in protein sequences for pairs of amino acid residues in the n-th and n+1-th, as well as in the n-th and n+2-th, positions (8). The normalized S_{ij} and S_{i_j} frequency values of all 400 pairs of 20 amino acids revealed that general short range regularities exist in the primary structure of proteins and every amino acid has a characteristic sequential residue environment. Since the observed amino acid pairing represents mainly the preferences between amino acids inside the structural domains where the majority of the data base came from, at domain borders deviations may be expected.

There is another difference between the intra- and interdomain segments. Stability of the structural domains requests as much interaction energy as possible to overcome the structural entropy and mantain the compact structure. Short range interactions play a dominant role in the stabilization of the protein structure. However, the conformation of interdomain segments is determined mainly by domain-domain interactions. Thus the interactions between the consecutive amino acids at domain boundaries or in the hinge regions between domains are presumably less strong than inside the domains. Statistically determined association

potentials (E_{ij}) of Narayana and Argos (9) were used to describe interactions between amino acids.

For domain border prediction the available S_{ij} and S_{i_j} sequential and the E_{ij} association preference data (8, 9) were combined so that pairs preferred neither sequentially nor by their favourable interaction should stand out. We chose one of the simplest possible procedure taking[§]

$$S_{ij} \cdot E_{ji} = T_{ij} \qquad \text{and} \qquad S_{i\_j} \cdot E_{ji} = T_{i\_j} \qquad (1)$$

Values of T_{ij} and T_{i_j} for all 400 pairs of amino acids are shown in Table 1-2. T_{ij} and T_{i_j} quantitatively characterize the general preferences for first and second neighbour pairs of amino acids formed by the i-th and j-th type of amino acids in the first and second position, respectively. Let us take an ijk segment of the polypeptide chain. We calculated

$$q_1 \cdot T_{jk} + q_2 \cdot T_{i\_k} = P \qquad (2)$$

yielding what we call the general short range preference value for the k residue. q_1 and q_2 are weighting factors. It was shown that interaction between adjacent amino acids is 2.4 times as frequent as between the 2nd neighbours (5). Accordingly we used 1.0 and 0.4 for q_1 and q_2, respectively. Computing P in every position along the polypeptide chain a general short range preference profile can be obtained which is supposed to exhibit minima at domain boundaries.

The scans were smoothed convoluting linearly weighted general

---

[§] According to the original definition of association potentials the counterpart of S_{ij} is E_{ij} if we consider the preference of an amino acid residue only to its N-terminal neighbours.
In the calculations the value of 1.0 instead of 12.7 was used for association potential of Cys-Cys interaction because the latter extreme value is mainly due to disulphide bond formation not characteristic for near neighbour Cys-Cys interaction.

Table 1. General preference data of pairs of amino acids.These values were calculated from available sequential and association preference data as described in the text. The j-th figure in the i-th row of the matrix is the general preference value for the ij amino acid pair

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.92	0.58	1.13	0.76	0.84	1.27	1.07	1.00	0.59	1.26	0.68	0.75	0.54	1.05	0.49	0.78	0.78	1.02	1.44	1.00
Cys	1.06	2.21	0.42	0.35	0.94	1.09	2.92	0.98	0.22	0.65	1.00	0.65	1.32	1.67	1.50	1.07	0.74	0.60	0.51	1.92
Asp	0.79	0.23	0.93	1.13	0.64	1.00	1.02	0.65	1.69	0.48	0.25	0.84	0.44	0.48	1.95	0.78	0.84	0.42	0.40	0.79
Glu	0.45	0.20	0.89	1.21	0.61	0.74	0.76	0.43	2.28	0.55	0.50	0.94	0.50	0.46	2.34	0.54	0.44	0.37	0.64	0.51
Phe	0.68	0.73	0.71	0.64	2.81	0.86	1.07	1.22	0.82	1.63	1.17	0.40	0.95	0.75	1.28	0.68	0.76	0.99	0.88	0.99
Gly	1.12	0.98	1.02	1.05	0.87	0.93	1.01	1.04	0.97	0.83	0.72	1.02	1.42	0.95	0.92	1.14	1.29	0.91	0.54	0.74
His	0.60	1.43	0.66	1.22	2.04	1.03	1.66	0.34	0.82	1.04	0.41	0.96	1.80	0.55	0.71	0.82	0.86	0.50	3.08	0.94
Ile	0.94	0.94	0.68	0.69	1.65	0.67	0.40	1.63	0.68	1.20	1.04	0.50	0.84	0.73	0.29	0.50	0.50	0.27	0.49	0.91
Lys	0.34	0.11	0.77	1.40	0.40	0.89	0.51	0.39	0.68	0.34	0.61	0.80	0.29	0.43	0.25	0.30	0.42	0.33	0.37	1.22
Leu	0.89	0.49	0.42	0.77	1.63	0.72	1.13	1.62	0.66	1.29	1.23	0.55	0.51	0.81	0.74	0.77	0.69	2.36	0.87	0.72
Met	0.48	0.46	0.40	0.72	2.15	0.88	0.62	1.38	1.05	1.05	1.83	0.74	0.70	0.32	0.79	0.47	0.87	1.52	1.63	0.94
Asn	0.52	0.37	0.85	0.74	0.45	0.85	0.59	0.39	1.05	0.57	0.70	1.27	0.71	1.00	1.09	0.69	0.74	0.45	1.69	1.50
Pro	0.46	0.44	0.44	1.29	0.48	2.08	1.14	0.46	0.58	0.39	0.78	0.66	0.73	1.05	0.82	0.72	0.64	0.66	1.39	0.81
Gln	0.67	0.93	0.57	0.46	0.58	1.02	0.59	0.34	0.90	0.50	0.44	0.74	0.94	1.43	1.13	0.93	0.53	0.43	0.92	0.61
Arg	0.29	0.58	1.11	1.51	1.26	1.02	0.64	0.45	0.30	0.59	0.52	1.00	0.61	1.24	1.38	0.41	0.51	0.43	1.13	0.77
Ser	0.59	0.98	1.12	1.04	0.58	1.42	1.23	0.41	0.58	0.66	0.55	0.82	0.58	1.13	0.91	1.07	0.90	0.61	1.02	0.76
Thr	0.62	1.00	1.11	0.71	0.81	0.89	0.69	0.82	0.82	0.71	0.64	0.74	0.75	0.92	0.73	0.72	0.77	0.65	1.04	0.98
Val	0.94	0.80	0.65	0.73	1.38	0.83	0.75	1.26	0.50	1.75	1.71	0.60	0.56	0.74	0.55	0.65	0.70	1.41	1.70	0.95
Trp	0.60	1.09	0.35	0.47	0.77	1.31	1.40	1.63	0.85	1.09	2.25	0.90	0.50	1.71	1.39	0.55	0.63	1.73	1.09	1.85
Tyr	0.51	1.71	0.81	0.59	0.96	0.70	0.72	1.16	1.23	1.19	1.27	1.10	1.16	1.16	1.01	0.68	0.75	0.62	0.96	1.76

Table 2. General preference data of pairs of amino acids separated from each other by one residue. The j-th figure in the i-th row of the matrix represents general preference for the i_j amino acid pair

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.86	0.76	0.76	0.66	1.01	1.44	1.05	1.22	0.81	1.23	0.73	0.63	0.41	0.87	0.49	0.65	0.78	1.18	1.15	1.45
Cys	1.11	1.13	0.46	0.43	0.66	0.97	0.97	1.10	0.32	0.89	0.29	0.67	1.02	2.05	1.36	1.21	1.25	0.90	0.48	0.99
Asp	0.73	0.39	0.87	1.04	0.66	0.82	1.16	0.59	1.38	0.48	0.37	0.83	0.35	0.72	1.97	1.07	0.82	0.44	0.46	0.71
Glu	0.32	0.17	0.72	1.09	0.69	1.08	1.23	0.66	2.00	0.68	0.53	0.56	0.62	0.56	1.40	0.47	0.55	0.34	0.36	0.60
Phe	0.82	0.59	0.59	0.81	2.37	1.01	2.03	1.09	0.77	1.36	1.36	0.56	0.78	0.95	1.46	0.50	0.63	0.94	1.02	1.12
Gly	1.10	0.85	0.93	0.77	0.88	0.99	0.82	0.95	1.05	0.80	0.84	1.05	1.78	0.72	1.08	1.18	1.16	1.07	0.83	0.69
His	0.58	1.04	2.26	1.47	1.06	0.97	1.73	0.46	0.63	0.81	0.42	0.71	0.90	0.74	0.60	0.88	0.99	0.78	2.02	1.26
Ile	1.05	1.33	0.73	1.04	0.90	1.05	0.46	1.06	0.68	1.19	1.50	0.51	0.53	0.53	0.37	0.38	0.32	0.27	0.84	1.14
Lys	0.29	0.11	0.64	1.38	0.53	1.06	0.54	0.48	0.64	0.46	0.34	0.61	0.27	0.51	0.29	0.31	0.37	0.31	0.58	0.94
Leu	0.98	1.04	0.51	0.80	1.24	0.97	0.93	1.24	0.64	1.19	1.06	0.70	0.46	0.92	0.85	0.61	0.45	1.92	1.61	0.84
Met	0.61	0.91	0.28	0.75	3.57	0.78	0.78	1.17	0.86	0.76	1.35	0.87	0.85	0.52	1.08	0.44	0.63	0.93	3.20	1.21
Asn	0.49	0.51	0.97	0.75	0.49	0.80	0.78	0.36	1.44	0.54	0.86	1.69	0.40	0.97	1.21	0.58	0.59	0.50	1.22	1.31
Pro	0.41	0.75	0.53	1.02	0.68	1.48	0.99	0.47	0.57	0.57	0.66	0.74	0.73	1.18	0.94	0.62	0.89	0.66	0.99	0.81
Gln	0.53	0.58	0.40	0.49	0.77	0.75	0.95	0.50	0.77	0.64	0.39	0.88	1.37	0.91	1.04	0.71	0.71	0.55	0.87	0.73
Arg	0.27	0.60	1.62	1.86	1.38	0.67	0.64	0.48	0.25	0.49	0.50	0.79	0.63	0.85	1.30	0.68	0.78	0.35	1.13	0.77
Ser	0.60	0.88	1.22	0.88	0.67	1.07	0.97	0.42	0.54	0.76	0.68	0.95	0.64	0.99	0.89	1.36	0.95	0.56	0.80	0.84
Thr	0.78	0.73	1.10	0.91	0.64	0.87	0.68	0.53	0.72	0.54	1.24	0.81	0.84	0.84	0.88	1.05	1.17	0.48	0.40	0.86
Val	1.21	0.54	0.57	0.54	1.16	1.16	0.75	1.54	0.48	1.33	1.00	0.58	0.87	0.71	0.51	0.64	0.58	1.54	2.60	0.93
Trp	0.74	0.55	0.53	0.85	0.85	0.49	1.77	1.03	1.74	1.10	1.55	1.08	0.81	3.11	1.97	0.79	0.45	0.89	0.51	0.89
Tyr	0.75	1.68	0.81	1.05	0.66	0.69	0.73	0.59	1.37	0.82	1.71	1.46	1.36	1.37	0.85	0.56	0.59	0.72	1.29	1.67

short range preference values of neighbours up to ± 5 residues in every position applying.

$$\bar{P}_n = \frac{P_n + \sum_{m=1}^{N} \frac{N + 1 - m}{N + 1} (P_{n-m} + P_{n+m})}{N + 1} \qquad (3)$$

where $\bar{P}_n$ and P_n are general short range preference values for the residue at sequence position n after and before smoothing, respectively, m is the relative position to the central residue, while N denotes the half width of the window applied. (In our calculations N= 5 was used.)

RESULTS and DISCUSSION

To test the efficiency of the proposed algorithm calculations for several known multidomain proteins were performed. The smoothed general short range preference scans for IgG heavy chain, GAPD, papain and hemoglobin β-chain are presented in Fig.1. (Domain boundaries identified from the 3-dimensional structure are designated by dashed lines in the plots.)

It can be seen that domain boundaries appear as minima in the scans and in a few cases these are the deepest ones. However, the scans contain several other rather significant minima inside the structural domains. So far we have failed to explain in structural terms the existence of these minima. In some cases they fit well the limits of secondary structural units, but in general they are inside the α-helices or β-strands. They cannot be correlated with the chemical nature (hydrophobic, neutral, polar, charged) or the spatial arrangement (exposed, totally or partially buried) of the respective amino acids, or with the residual mobility determined by X-ray crystallography.

It is conceivable that the appearance of the undesirable minima in the plots is an inevitable consequence of the simplicity of the algorithm. Clearly, the conformation of amino acids in proteins is generally affected not only by interactions with the

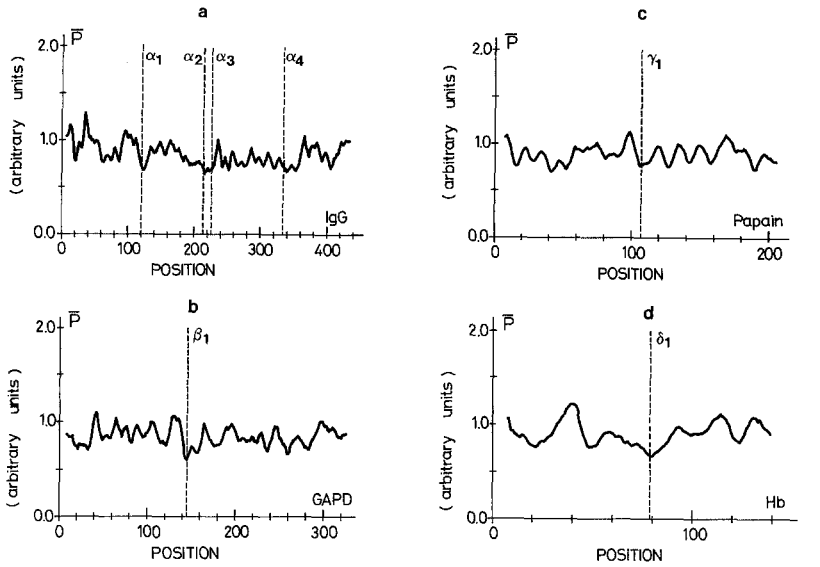

Fig.1 Smoothed general short range preference profiles for some multidomain proteins against amino acid sequence number. Domain boundaries are expected to appear as minima in these plots. Dashed lines indicate domain limits (marked with Greek letters) identified from the three-dimensional structure.

a. IgG heavy chain. α_1 is between the variable and the CH1 constant domain. The hinge region is between α_2 and α_3 .
b. Glyceraldehyde 3-phosphate dehydrogenase. The molecule is divided by β_1 into the coenzyme binding and the catalytic domains.
c. Papain
d. Hemoglobin β-chain.

1st and 2nd neighbours, but by longer range interactions as well (10). For this reason we think that the minima of our scans match those polypeptide segments which are mainly stabilized by medium and long range interactions. Consequently, the presented method could be refined by taking into account larger segments in calculating general preference scans. This would require sequential preference data for the 3rd, 4th .... etc. neighbour amino acid pairs, which are unavailable as yet. Further improvement might be achieved by using more reliable sequential and spatial amino acid preference values calculated from significantly larger data bases.

Since domain borders are between secondary structural elements, combining our approach with secondary structure prediction methods (e.g. Chou-Fasman method) could result in the exclusion of

several unwanted significant minima, thereby reducing the number of possible sites for domain limits.

We would like to emphasize that such a very simple procedure based on only 1st and 2nd neighbour amino acid preferences can already provide valuable information about the tertiary structural organization of proteins. The presented approach indicates the position of domain boundaries as minima in the scans. Although due to the false minima in this form our algorithm cannot be applied alone, it largely confines the number of potential sequence positions of domain limits and it may be useful if additional information is available. For instance, biochemical experiments such as limited proteolysis in some cases can only roughly define domain boundaries because of the severe specificity of the applied proteolytic enzymes. However, these results can be complemented with our method which can further localize domain boundaries if the rough positions are known. In addition, combination of the described approach with an independent predictive algorithm of similar effectivity, i.e. which also delimits domain boundaries to about 10-20 per cent of all positions could lead to an almost unambigous domain border assignment.

REFERENCES

1. Wetlaufer,D.B. (1973) Proc.Natl.Acad.Sci. USA 70, 697-701.
2. Janin,J. and Wodak,S.J. (1983) Prog.Biophys.molec.Biol. 42, 21-78.
3. Rao,S.T. and Rossmann,M.G. (1973) J.Mol.Biol. 76, 241-256.
4. Creighton,T.E. (1983) Proteins, W.H.Freeman and Company, New York.
5. Busetta,B. and Barrans,Y. (1984) Biochim.Biophys.Acta 790, 117-124
6. Dayhoff,M.O. (1978) Atlas of Protein Sequence and Structure vol.5, supp. 1-3, Natl.Biomed.Res.Fdn., Washington, D.C.
7. Kabat,E.A., Wu,T.T., Bilofsky,H., Reid-Miller,M. and Perry,H. (1983) Sequence of Proteins of Immunological Interest, U.S. Dept.Health Human Services, Public Health Service, N.I.H.
8. Vondervisztt,F., Mátrai,Gy. and Simon,I. (1986) Int.J.Peptide Protein Res. 27, 480-489.
9. Narayana,S.V.L. and Argos,P. (1984) Int.J.Peptide Protein Res. 24, 25-39.
10. Scheraga,H.A. (1980) Protein Folding, pp. 261-288, Elsevier North-Holland Biomedical Press, Amsterdam, New York.