JPP JOURNAL OF PROTEINS AND PROTEOMICS
3(2), July-December 2012, pp. 95-98

COMMUNICATION

DETECTING DOMAIN BOUNDARIES IN PROTEINS THROUGH PLOTTING OF THE ENERGY OF NON-BONDED INTERACTIONS (ENBI) AS A FUNCTION OF PROGRESSIVE *IN SILICO* TRUNCATION OF CHAINS IN NATIVE STRUCTURAL FORMAT

Purnananda Guptasarma*

Protein Science & Engineering Division, Institute of Microbial Technology, (IMTECH), Council of Scientific & Industrial Research (CSIR), Chandigarh 160036, India

Abstract: Several methods exist for the detection of domain boundaries in proteins. Different methods exploit different structural-biochemical characteristics distinguishing, and defining, protein domains. However, perhaps because 'domains' remain poorly defined, no single method has proved to be entirely satisfactory. Here, a new approach to defining and detecting domains is presented, along with some preliminary data from three proteins, in the form of a proof-of-concept. It is argued from first principles that protein domain boundaries may be identified through plotting of variations in the energy of non-bonded interactions of a naturally-occurring protein as a function of varying chain length (in native structural format). Such plots may be expected to show a broadly descending trend as a function of increasing chain length, marked by slope changes at domain boundaries. The approach is demonstrated with three multi-domain, single-subunit proteins, porcine pepsin (4PEP), thymidylate synthase (4TMS) and aconitase (5ACN).

Keywords: Protein domains; Protein folding; Non-bonded interactions

Introduction

A number of methods already exist for the identification of structural domains in proteins (Taylor, 1999). However, given that the structural entity commonly referred to as a domain has never actually been defined by any consensus set of structural criteria, no two methods that are currently in use for the identification of domains from structural data can yet be relied upon to identify precisely the same sets of chain segments as domains, for every protein tested. To a large extent, this is because each method utilizes a different set of criteria by which to evaluate a chain segment, or structure, for possible classification as a domain. The issue is further confounded by the fact that although the word 'domain' was originally coined to describe a minimal element of polypeptide sequence displaying the ability to fold autonomously, few structures currently referred to as domains have actually ever been tested by this criterion. On the contrary, instances are known where a protein structural motif initially widely assumed to constitute a single domain, e.g., the eight-stranded α/β-barrel, once canonized as a single domain (Levitt and Chothia, 1976; Farber and Petsko, 1990) has later turned out to contain smaller autonomously folding sub-structures (Zitzewitz *et al.*, 1999; Lang *et al.*, 2000). Clearly, until the word domain is defined by some sort of consensus criteria, different methods of structural analysis will continue to differ occasionally in regard to which chain segments of a protein constitute structural domains.

Given the above background (but without presuming to refine the meaning of the word

Corresponding Author: **Purnananda Guptasarma**
E-mail: guptasarma@iisermohali.ac.in
Present Address: Department of Biological Sciences, IISER Mohali, SAS Nagar, Punjab 140306.
Received: November 26, 2012
Accepted: November 28, 2012
Published: November 30, 2012

'domain'), a new method for identifying domains is presented here for possible further testing by other researchers. The method involves plotting of variations in the energy of non-bonded interactions as a function of varying chain length in native structural format. The basic premise, method and initial results are described below.

Amino acid residues in proteins are optimally structurally organized, particularly in respect of non-bonded interactions. Assuming that domains are entities that fold relatively autonomously of other chain segments, it can be argued that the physical folding of a domain during the uni-directional growth of a polypeptide chain (encoding multiple domains) on a ribosome must result in an overall lowering of the energy of non-bonded interactions (ENBI) of the chain, because chain folding ensures the formation of satisfactory sets of non-bonded contacts and minimizes the number of residues that remain unsatisfied in this respect. As the majority of domains comprise contiguous stretches of amino acids, and because such stretches are linked together on a single polypeptide chain, it stands to reason that when a polypeptide chain starts out on a new trajectory following completion of the synthesis and folding of a domain (on the ribosome), a change must be seen in the overall rate of change of the chain's ENBI with growth (amounting, at the very least, to a slowing-down of the rate of overall decrease accompanying growth and folding, if not actually to a small increase in ENBI).

One does not yet know for certain whether proteins fold cotranslationally, or even whether all proteins could be assumed to fold cotranslationally were some to be unequivocally established to do so. Further, even if it were discovered that proteins do fold cotranslationally, such folding would be applicable to individual domains and supersecondary structural elements capable of autonomous formation of structure. Thus, it would never be the case that individual residues would adopt a native-like structural format immediately upon synthesis. Even so, assuming that residues do indeed become structured in native format as soon as they are added to a growing chain (a false assumption, and a contrivance adopted for reasons that will become obvious later in the text), it is proposed that this becomes a basis for the framing of a novel approach for identifying domains.

Materials and Methods

Three multi-domain, single-subunit proteins were chosen for investigation, namely porcine pepsin (4PEP), thymidylate synthase (4TMS) and aconitase (5ACN). The method followed for each of these is outlined below: The PDB file containing coordinates of the protein was downloaded. The file containing the coordinates of n residues (where n is the number of residues in the chain) was processed through the following steps : (1) Lines in the PDB file corresponding to entries for atoms constituting the last residue of the chain, i.e., residue n, were deleted. This truncated file was saved with a new name as a PDB file containing the coordinates of all residues from numbers 1 to n-1. (2) The new file thus obtained was opened and lines corresponding to atoms constituting the last residue in the truncated chain, i.e., residue n-1, were deleted. Again a file was saved with a new name as a PDB file containing the coordinates of all residues from 1 to n-2……and so on, until it remained possible to generate a file containing only the coordinates of the first fifteen residues from the N-terminus. Each of these files was individually saved. The procedure for truncation was automated through the writing of a small program in BASIC that was later converted into an executable program, taking PDB files as input. No attempt was made to add carbon and oxygen atoms at the newly generated C-terminus at each stage. Each file generated above through progressive truncation of the chain was imported into the program 'SWISS PDB Viewer' [downloaded version 3.0, Nicolas Guex & Manuel Peitsch, 1996, Geneva Biomedical Research Institute, *URL http://www.expasy.ch/spdbv/mainpage.html*] and analyzed, using appropriate menu options, to obtain a measure of the total energy of non-bonded interactions (ENBI) of the piece of the protein's structure defined by the coordinates contained within that file. [H atoms and heteroatoms of unknown topology are not taken into consideration for calculations; further, calculations are not fully calibrated to the AMBER force field in this version of the program; however, this is not of consequence, since only

relative changes in ENBI are important to the method, and not absolute ENBI values]. The ENBI value for each truncated structure file was plotted as a function of the length of the truncated chain

Results and Discussion

Figure 1 shows plots obtained through the procedure described above, for three multi-domain, single-subunit proteins, namely porcine pepsin (4PEP), thymidylate synthase (4TMS) and aconitase (5ACN). Each panel plots the changes in ENBI occurring as a result of truncation of the chain in native structural format, on a residue-by-residue basis (right-to-left). The plotted trends can also be assumed to correspond to the change in ENBI that would be seen if a growing chain were to organize itself into its native structural format concurrently with its synthesis (left-to-right). In each panel, the previously known domain boundaries are shown as vertical lines intersecting the plots, with actual positions of such boundaries marked alongside.

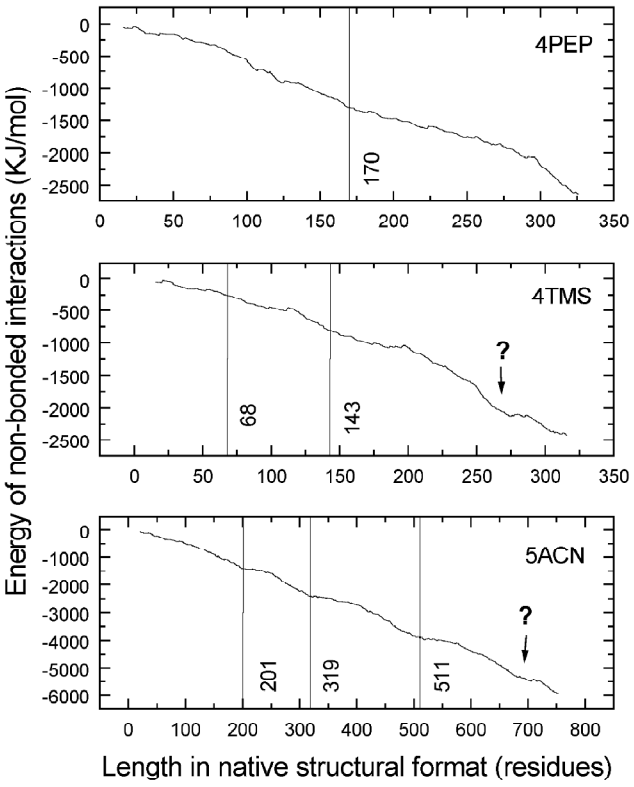

Figure 1: **Plots of variations in ENBI as functions of changing chain lengths in native structural format. Vertical lines intersecting plots are known domain boundaries (positions marked alongside). Question marks highlight putative boundaries that are not identified in the original PDB file. PDB identification numbers are marked in the top right corner of each panel. Proteins analyzed were porcine pepsin (top panel), thymidylate synthase (middle panel) and aconitase (bottom panel)**

Whereas, on the whole, ENBI can be expected to decrease with chain growth in native structural format *in silico*, the actual rate of decrease is expected to change with the nature of the contacts being made during such growth. Thus, a sharper decline in ENBI could be anticipated to accompany the completion of the structure of a domain, since such completion would result in the satisfaction of many contacts 'waiting' to be satisfied prior to the launching of the chain onto a new trajectory for the creation of a new domain. Similarly, upon completion of the structure of a domain, the launching of the chain onto a new trajectory could be expected to result in either a slowing down in the rate of decrease of ENBI, or in an actual increase in ENBI, because a chain beginning to form a new domain (with residues structured in native format) would initially place several residues in locations in which they would make no favourable non-bonded contacts whatsoever. Subsequently, of course, the chain's trajectory would return (e.g., through long-distance strand-strand, sheet-sheet, sheet-helix or helix-helix interactions) to satisfy contacts and bring down the ENBI. Together, the sharper rate of decrease resulting from completion of a domain and the slower rate of decrease resulting from the launching of a new domain, can be expected to create a discernible 'trough' in the plot, marking the domain boundary.

It is evident from two panels (4PEP and 5ACN) that plots of ENBI *vs* chain length do indeed display troughs. Remarkably, at the very point within each trough at which the change in the rate of change of ENBI becomes most discernible, one finds located the domain boundary previously identified by the depositors of the protein's coordinates. However, there is the occasional miscarriage, e.g., in the panel 4TMS, although one of the previously known domain boundaries (at residue 143) does appear to lie within a trough at approximately the position at which the change in the rate of change of ENBI is most discernible, another previously identified boundary (at residue 68) is found not to correlate with the trough at all; in fact, the trough is actually

located several tens of residues further along the chain. This discrepancy points either to a failure of the method described, or alternatively, to the previous identification of this domain boundary by entirely different criteria, e.g., through visual inspection of structure.

Notably, in panels 4TMS and 5ACN two additional troughs (marked with question marks) are seen. These do not correspond to any known domain boundary mentioned anywhere within the text of the PDB file. There may be scope, therefore, to examine once again whether domain boundaries exist at the marked locations that were somehow missed earlier, in these proteins. Alternatively, the troughs marked with question marks may reflect an, as yet unknown, drawback of the method described here.

Since the method was tested with only a few proteins, one can desist from making any excessive, or further, speculations. The reason that the method appears to work, as already explained, is that at every domain boundary the polypeptide chain strikes out in a totally different direction, away from the previous domain, within the native structure. This extended region of the chain makes non-bonded contacts only with other distant regions of the chain which are, of course, not represented in the truncated chain. Thus, the rate of drop of ENBI with increasing chain length is effectively reduced every time the chain strikes off in a different direction to create a new domain. Of course, eventually the ENBI begins to drop rapidly again when chain satisfies all requisite non-bonded interactions before starting to create yet another domain. It is these sudden changes in the reduction rate of the ENBI that the method detects, to identify domain boundaries.

Acknowledgements

I gratefully acknowledge the help of Prof. D. Guptasarma in implementing the computational aspects of the work described in this manuscript, and for discussions.

References

Farber, G.K., and Petsko, G.A. (1990). The evolution of alpha/beta barrel enzymes. Trends. Biochem. Sci. *15*, 228-234.

Lang, B., Thoma, R., Henn-Sax, M., Sterner, R., and Williams, M. (2000). Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. Science *289*, 1546-1550.

Levitt, M, and Chothia, C. (1976). Structural patterns in globular proteins. Nature *261*, 552-558.

Taylor, W.R. (1999). Prediction of structures of multidomain proteins from structures of the individual domains. Protein Eng. *12*, 203-216.

Zitzewitz., J.A., Gualfetti, P.J., Perkons, I.A., Wasta, S.A., and Matthews, C.R. (1999). Identifying the structural boundaries of independent folding domains in the alpha subunit of tryptophan synthase, a beta/alpha barrel protein. Protein Sci. *8*, 1200-1209.