BIOINFORMATICS

Vol. 16 no. 7 2000
Page 613–618

Domain size distributions can predict domain boundaries

S. J. Wheelan[1,2], A. Marchler-Bauer[1] and S. H. Bryant[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA and [2]Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

Received on July 17, 1999; revised on January 9, 2000; accepted on February 10, 2000

Abstract

Motivation: *The sizes of protein domains observed in the 3D-structure database follow a surprisingly narrow distribution. Structural domains are furthermore formed from a single-chain continuous segment in over 80% of instances. These observations imply that some choices of domain boundaries on an otherwise uncharacterized sequence are more likely than others, based solely on the size and segment number of predicted domains. This property might be used to guess the locations of protein domain boundaries.*

Results: *To test this possibility we enumerate putative domain boundaries and calculate their relative likelihood under a probability model that considers only the size and segment number of predicted domains. We ask, in a cross-validated test using sequences with known 3D structure, whether the most likely guesses agree with the observed domain structure. We find that domain boundary predictions are surprisingly successful for sequences up to 400 residues long and that guessing domain boundaries in this way can improve the sensitivity of threading analysis.*

Availability: *The DGS algorithm, for 'Domain Guess by Size', is available as a web service at http://www.ncbi.nlm.nih.gov/dgs. This site also provides the DGS source code.*

Contact: *bryant@ncbi.nlm.nih.gov*

Introduction

A number of areas of comparative sequence analysis can be aided by knowledge of domain boundaries. Fold recognition, in particular, may require that a target sequence be parsed into autonomously folding domains, likely to resemble a domain previously seen in the 3D-structure database. Comparative sequence analysis often identifies domain boundaries. However, if a sequence has no apparent similarity to other sequences, no internal repeats, and/or no regions of low complexity, one may have little clue as to the locations of domains. Predictors in the CASP3 competition, for example, faced this situation for three of the 11 fold-recognition targets where successful threading predictions were made (Marchler-Bauer and Bryant, 1999; Moult *et al.*, 1999; Murzin, 1999).

If the lengths of protein domains were fixed and domains were always formed from a single-chain continuous segment it would be a simple matter to predict domain boundaries: one need only break the sequence into pieces of the appropriate length. The size of protein domains is not fixed, of course, but neither are all domain lengths equally likely. It has already been seen that protein domain length follows a narrow distribution, and furthermore that domains identified in the 3D-structure database most often contain a single-chain continuous segment (Islam *et al.*, 1995; Sowdhamini *et al.*, 1996; Jones *et al.*, 1998). Thus, while one can not expect to perfectly predict the locations of domain boundaries based on size and segment numbers alone, one can expect that some guesses will be better than others.

Here we ask whether guessing domain boundaries based on the size and segment number of predicted domains can be accurate enough to be useful. We construct a likelihood function based on empirical distributions of domain length and segment number, as observed in the 3D-structure database. For test sequences, we then enumerate candidate domain boundaries and calculate their relative likelihood, to give a ranked list of alternative domain boundary predictions. Using a cross-validated test with sequences from the 3D-structure database we find that this method is surprisingly successful. For two-domain proteins in the test set, for example, one of DGS's top two guesses is accurate to a resolution of ± 20 residues in 57% of cases. Using threading targets from CASP3 as examples, we show that domain boundaries guessed by DGS can improve threading predictions.

*To whom correspondence should be addressed.

Methods

To determine domain-size distributions we select a training set of sequence-dissimilar chains with a known 3D structure. Structure data are taken from the Protein Data Bank (Berman *et al.*, 2000; http://www.rcsb.org/pdb/). Chains are grouped by single-linkage clustering based on a BLAST p-value (Altschul, 1997) of 10^{-7} or less and a set of 1236 group representatives selected automatically, based on completeness and resolution (Matsuo, Bryant, 1999; http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html).
This chain set contains 1882 domains, based on the domain definitions used for structure neighboring in Entrez (Wang *et al.*, 2000; http://www.ncbi.nlm.nih.gov/Entrez/). Domain definitions in Entrez are based on structural compactness. A chain is split any number of times, at points between secondary structure elements, whenever the ratio of intra- to inter-domain contacts exceeds a threshold (Madej *et al.*, 1995).

Domain guess by size calculates the likelihood of alternative partitions of a sequence into one or more domains, and ranks the alternative guesses accordingly:

$$L(n, L, S|c) = p(n|c)\,p(L, S|n, c). \quad (1)$$

Here $p(n|c)$ is the probability that a chain of length c will have n domains. We estimate $p(n|c)$ empirically, from the frequency of chains in the training set having one, two, and three (or more) domains, for discrete length intervals. The term $p(L, S|n, c)$ gives the probability that n domains will have lengths L and numbers of segments S, given that the lengths of the individual domains in vector L are constrained to sum to c. These two terms are multiplied to give the likelihood of observing a set of n domains with individual lengths L and segment numbers S, given chain length c.

In DGS we enumerate a discrete list of possible domain boundaries. Across this list we estimate $p(L, S|n, c)$ from empirical distributions for the length and segment number of individual domains, as observed in the training set:

$$p(L, S|n, c) = \prod_{L,S} p(l)\,p(s) \Big/ \sum \left[\prod_{L,S} p(l)\,p(s) \right]. \quad (2)$$

The term $p(l)$ gives the probability of observing length l for an individual domain. This is estimated as the fraction of domains in the training set whose length falls within a discrete length interval containing l. The term $p(s)$ gives the probability that an individual domain will be formed from s-chain continuous segments. This is estimated from the fraction of domains in the training set formed from a single-chain continuous segment (83.6%), two chain-continuous segments (14.7%), or three (or more) chain-continuous segments (1.7%). In DGS we ignore the chance that a domain may be formed from more than three segments. The product is taken over the n domains whose lengths and segment numbers are given by L and S. The sum in the denominator of equation (2) is taken across all the domain boundary guesses we consider. For $n = 1$ there are no alternative boundary locations, and $p(L, S|1, c)$ is equal to 1 by definition.

Figure 1 illustrates the domain-boundary enumeration algorithm implemented in DGS. Alternative partitions are explored using a 'step' size, here 20 residues. Based on the step size, all possible domain boundaries are constructed for all possible numbers of segments. For example, for a protein of length 100, with step size 20, the possible boundaries for partition into three-chain continuous segments are (20,40), (20,60), (20,80), (40,60), (40,80), (60,80). Next, the segments are labeled according to the domain to which they are assigned. For three segments the possible labels are just (1,2,1) and (1,2,3). Label (1,2,1) indicates a partition into two domains, the first split by insertion of the second, and label (1,2,3) indicates a partition into three domains. Other labels may be ignored since they place a domain segment adjacent to itself (implying a smaller number of segments) or are simply a synonymous renaming of domains. Finally, each boundary placement is labeled in each possible way and the likelihood calculated as described by equation (1).

Domain guess by size is written in C. The program uses domain number, domain size, and domain segment-number counts to dynamically calculate $p(n|c)$, $p(l)$ and $p(s)$, using the specified step size to set the width of length-interval bins. For the web service these data are updated with successive Entrez updates (Wang *et al.*, 2000). Domain guess by size returns the 25 guesses with the greatest likelihood, using a simple display to indicate domain boundaries and the assignment of chain-continuous segments to domains. The step size may be specified by the user, but we find that step sizes of less than 10 are not useful, since the top 25 guesses will in this case be very similar to one another and the run times very long. We note that for efficiency the program ignores alternative partitions with exceedingly low $\prod_{L,S} p(l)\ p(s)$ (i.e. guesses with exceedingly large or small domains), excluding them from numerator and denominator in equation (2).

Results

The sizes of the 1882 sequence-dissimilar domains in the training set are tabulated in Figure 2. This distribution peaks at around 100 residues and is relatively narrow. This size distribution has the same within-sampling error for single- and multi-domain proteins, and in DGS we therefore combine these data to estimate domain-length probabilities. These data suggest that longer proteins will

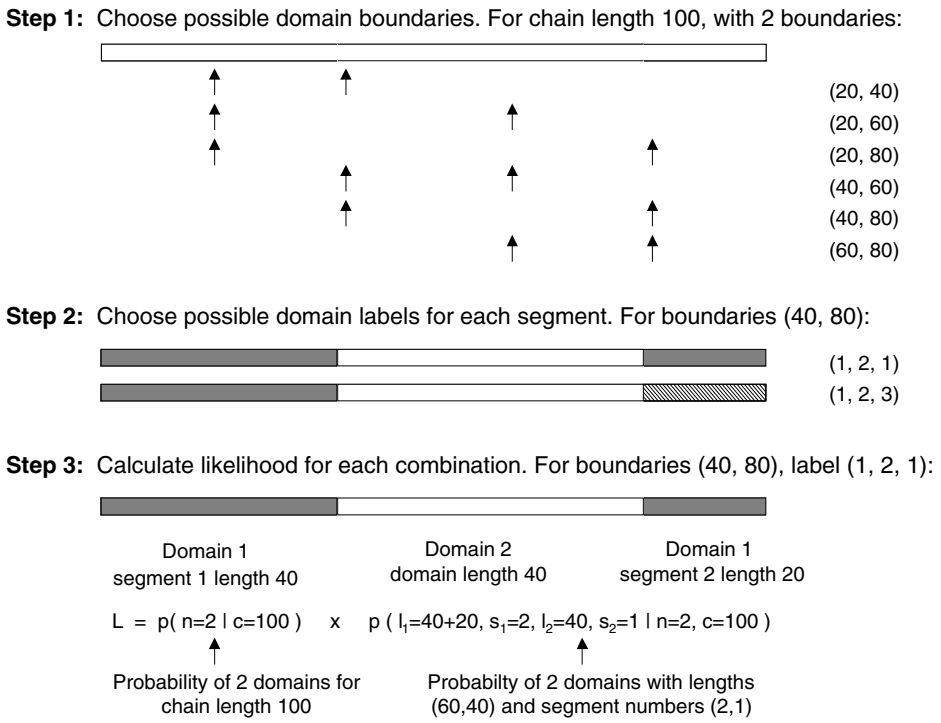

Fig. 1. The domain boundary enumeration algorithm implemented in DGS. All possible segment boundaries are first constructed, according to the specified step size (a). All possible labels for the segments are then computed (b). Lastly, the length and segment number for each domain is tabulated and the likelihood of observing a set of domains with these lengths and segment numbers is calculated as per equation (1), (c). The figure shows a hypothetical sequence of length 100 with alternative partitions using a step size of 20 residues.

tend to have more domains, and this is shown directly in Figure 3, where we tabulate domain-number frequencies for the 1236 chains in the training set.

A principal component analysis of global sequence characteristics such as average hydrophobicity, helix propensity and sheet propensity shows that only the length of the protein is correlated with the number of domains per chain and that these other factors do not improve prediction of domain boundaries (not shown). Domain definitions other than those used in Entrez (Wang *et al.*, 2000) lead to similar size distributions (Islam *et al.*, 1995; Sowdhamini *et al.*, 1996; Jones *et al.*, 1998), suggesting that the method used to identify domains in a 3D structure will not greatly affect results shown below.

To test DGS we use a cross-validation procedure, dividing the training set of 1882 domains into 10 sequence-dissimilar groups. The program was run on the chains in each group, using size and segmentation data from the other nine groups to calculate relative likelihood. Here, the results from all 10 groups are examined together. A prediction is considered a 'success' when all domain boundaries fall within the step size ($\pm$20 residues) of the domain boundaries given in Entrez. For domain predictions with more than a one-chain continuous segment, those segments must also be labeled correctly, as belonging to the same domain.

Figure 4 summarizes the overall success rate of DGS as a function of length. Figure 5 shows separately the success rate for one-domain, two-domain, and three-domain proteins. One may see that DGS does well for shorter proteins, especially single-domain proteins. Its top guess, that these are single-domain proteins, is very often correct. Domain guess by size can also predict the domain boundaries of many two-domain proteins, including a few with very complicated domain organizations (one protein had two domains, one of which was split twice and the other split once, i.e. a 12121 organization). Domain guess by size has only a small chance of correctly guessing domain boundaries for three-domain proteins, however, and we therefore make no attempt to guess partitions with more than three domains. For the test set as a whole, one of the top 10 guesses by DGS was correct for more than 50% of sequences under 400 residues in length.

To critically examine DGS's performance with two-domain proteins we may compare its success rate to that of a trivial method, guessing that the domain boundary

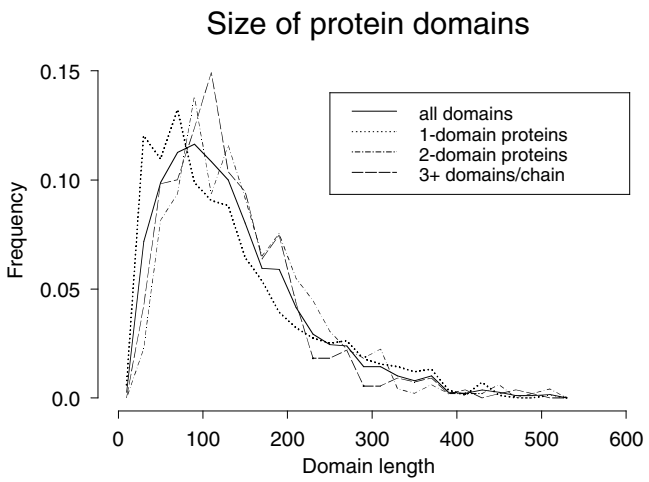

Fig. 2. Domain length distributions as observed in the 3D-structure database, according to the number of domains per chain. Data are from the non-redundant domain set described in Methods. An interval width of 20 residues is used to tabulate the number of domains of each length. The quantity $p(l)$ in equation (2) is estimated as the frequency for the length interval nearest the actual sequence length, in the distribution for all domains. The total numbers of observations are 840 1-domain chains (68%), 246 2-domain chains (20%), and 150 chains with 3-(or more) domain chains (12%).

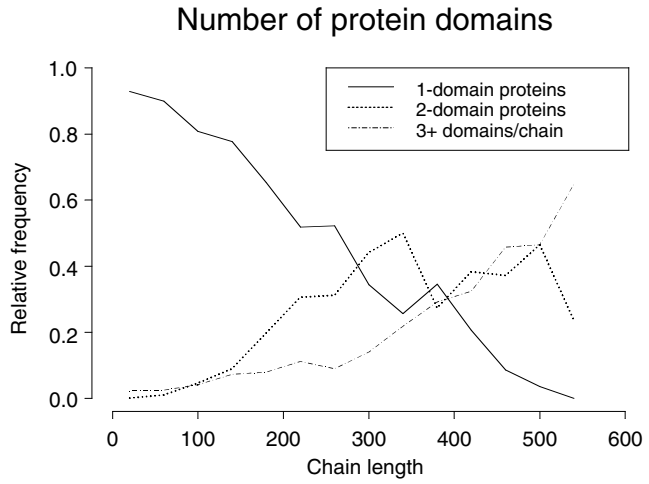

Fig. 3. The fraction of chains with 1, 2, and 3 (or more) domains per chain, as a function of chain length. Data are from the non-redundant set of chains with known 3D-structure as described in the text. A length-interval width of 40 residues is used to tabulate domain number frequency. The quantity $p(n|c)$ in equation (1) is estimated as the frequency for the length interval nearest the sequence length. We assign $p(n|c) = 0$ for domain numbers greater than 3. The total numbers of observations are as given in the caption to Figure 2.

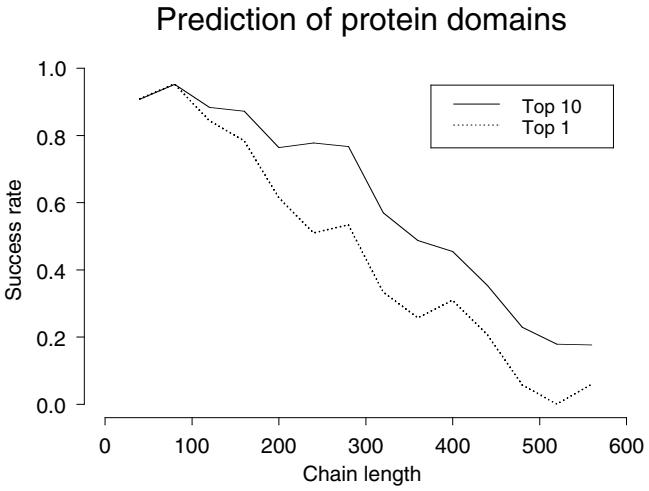

Fig. 4. Domain guess by size success rate as a function of sequence length. Each prediction is classified as a success if all predicted domain boundaries are within 20 residues of the boundaries observed in the 3D-structure. Success rates for the top prediction and the best of the top 10 predictions are shown separately. The test employs the non-redundant domain set and cross-validation procedure described in the text.

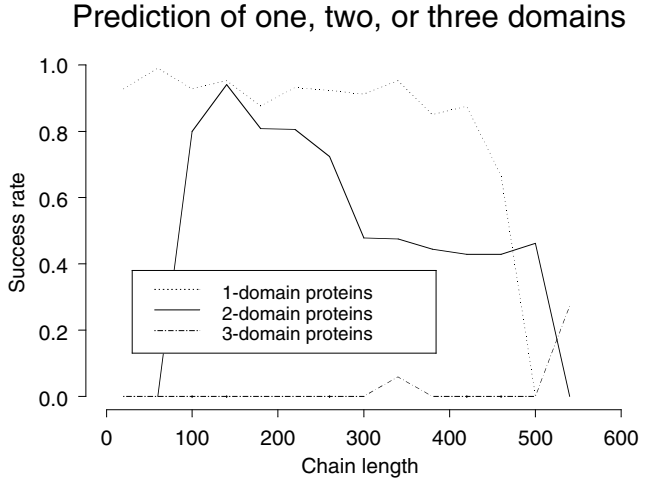

Fig. 5. Domain guess by size success rate as a function of sequence length, for 1-, 2- and 3- domain proteins. Success rate is based on the best of the top-10 predictions. Data and cross-validation procedures are as described in the text. We note that for chain lengths under roughly 500 residues a partition indicating one domain spanning the entire chain is always included in the top 10. The reported success rate for 1-domain proteins is slightly below 100%, however, since the Entrez domain definitions occasionally exclude small aglobular segments at the N- or C- termini. If this occurs, and the excluded aglobular segment is longer than 20 residues, the partition spanning the entire chain is counted as a failure.

is located at the midpoint of a sequence. Among the 246 two-domain proteins in the test set a domain boundary is located at the midpoint (±20 residues) in 20 cases (8%). This defines the success rate of the trivial method. For partition into two domains DGS usually generates two equally likely boundaries as its best guess (one offset left of the midpoint and the other offset the same distance to the right). For the 246 two-domain proteins in the test set one of these guesses is correct (±20 residues) in 140 cases (57%). Picking either guess at random one would expect to succeed in $140/2 = 70$ cases (28%), a more than 3-fold improvement over the trivial method. For two-domain proteins, partition into domains of unequal size is very frequent, and it appears that DGS predicts this effect. One of the top 10 guesses by DGS is correct in 151 cases (61%). The further improvement is due in part to guesses of discontinuous domains: 109 of these 246 proteins contain domains with more than one continuous segment, partitions the trivial method cannot predict.

Three CASP3-fold recognition targets were multi-domain proteins where domain boundaries were not identified by comparative sequence analysis, by our 'team' and most others. These are CASP3 targets 44, 63 and 83 (Moult *et al.*, 1999; Murzin, 1999). Targets 63 and 83 are two-domain proteins, and we find that one of the top two domain boundary guesses by DGS is close to that identified by structure–structure comparison. Target 44 has a complicated four-domain architecture; DGS cannot predict four domains, but its second-best guess corresponds to the first plus part of the second domain. To find out whether guessed domain boundaries can improve fold recognition we have performed threading calculations (Panchenko *et al.*, 1999) for these targets, using both the complete target sequence and domain subsequences from DGS's top-two guesses. In Table 1 we list some of the results. The table compares measures of threading model accuracy (Marchler-Bauer and Bryant, 1997) for the structurally similar template identified with lowest p-value (Bryant and Altschul, 1995), using a domain as guessed by DGS, with the corresponding values for the complete target sequence.

Examining Table 1, one may see that model accuracy generally improves when the target is partitioned into domains. For target 44 domain 1, and target 63 domain 2, contact specificity increases to 70% and 50%, respectively, from random values below 10%. For target 83 domain 1, contact specificity remains above 40%, and the other measures (such as ARms) indicate that model accuracy is good and essentially unchanged when threading the domain subsequence guessed by DGS, as compared with threading the complete target sequence. Perhaps a more interesting trend is also apparent in Table 1. For all targets the threading p-value decreases significantly when threading domain subsequences guessed by DGS. For target 44 domain 1, the p-value is perhaps low enough to make a confident prediction of a model based on this template. For target 63 domain 2, and target 83 domain 1, p-values remain in the 'twilight zone', but they are lower than for other templates, and these models would likely have been examined in analysis of threading results.

Table 1. Threading results for three CASP3 targets, with and without guesses of domain boundaries by DGS

Target	DGS	Tmplt	P-val	ARms	A%Id	Shft	CSpc
T44	None	1eps-3	0.13	6.0	19.6	1.4	9.7
T44-1	1–140	1eps-3	0.004	2.9	21.7	0.8	70.3
T63	None	1hjp-1	0.23	10.8	12.2	—	3.5
T63-2	61–138	1hjp-1	0.03	4.9	12.8	1.0	49.6
T83	None	1lmb4	0.14	2.5	26.7	0.2	48.0
T83-1	1-80	1lmb4	0.04	2.7	27.4	0.2	41.2

The target column gives the CASP3 target identification code (Moult *et al.*, 1999). The numeric domain identifier indicates the DGS domain used in the calculations, −1 for the N-terminal domain, −2 for the C-terminal domain. The DGS column indicates the domain boundary guessed by DGS, or None for the complete target sequence. Complete sequence lengths are 347, 138 and 156 residues for targets 44, 63 and 83, respectively. The Tmplt column gives the Protein Data Bank code for the structure used as a threading template. Numeric codes −1 and −3 identify a domain of that structure; definitions of compact domains for templates may be retrieved electronically using Entrez (http://www.ncbi.nlm.nih.gov/Entrez/). The p-val column gives the threading p-value, the probability that a shuffled version of the target (or target domain) sequence would score equally well (Bryant and Altschul, 1995). The ARms column gives the root mean square residual of the threading model with respect to the true structure of the target, in Angstroms. The A%Id column gives the percentage of identical residues in the threading alignment. The Shft column gives the average shift error of the threading alignment in residues, using structural superpositions by the VAST algorithm as the standard of truth (Gibrat *et al.*, 1996; Marchler-Bauer and Bryant, 1997). No value can be calculated for target 63 because the threading alignment with 1hjp-1 did not contain any of the target residues that VAST aligns with 1hjp-1. The CSpc column gives the contact specificity of the threading model, the percentage of predicted contacts present in the true structure of the target (Marchler-Bauer and Bryant, 1997).

CASP3 predictions, using the threading method we consider here were relatively successful (Panchenko *et al.*, 1999), and our 'team' was assigned 'first place' in overall fold recognition by the CASP3 assessor (Murzin, 1999). We made no predictions for targets 44 and 83, however, since threading with the complete target sequence did not indicate significant similarity to any available template structure. We made a prediction for target 63, based on threading the complete target sequence; this prediction was (perhaps fortuitously) based on a template structurally similar to the target, but model accuracy was low, as seen in Table 1 for threading with the complete target sequence. We emphasize that we present no blind prediction results using domain partitions guessed by DGS. The above retrospective analysis of threading results for

CASP3 targets suggests, however, that guessing domain boundaries can improve fold recognition, and that DGS might well have improved some predictions for CASP3.

Discussion

Domain guess by size calculates the relative likelihood of alternative domain partitions using only tables containing the numbers of domains per chain, domain lengths, and numbers of chain continuous segments per domain. This algorithm is as successful as it is only because the domain length and segment number distributions are narrow, such that a small number of guesses is likely to include the approximate locations of domain boundaries. It is far from clear, however, why protein domain sizes are constrained to a narrow distribution, or why domains are most often composed of a single-chain continuous segment. This phenomenon has been noted by others, and several possible explanations have been offered (Berman *et al.*, 1994; Trifonov, 1994).

The simplest possible explanations involve physical requirements for stable and/or rapid protein folding. Stability derives in part from burial of hydrophobic residues, and a certain minimum domain size is required. Rapid folding may at the same time favor small domains, and the balance of these effects may account for the observed narrow distribution. Other possible explanations involve the interactions with the molecular chaperones that catalyse protein folding *in vivo* (Hartl, 1996; Bukau and Horwich, 1998). Present-day proteins have co-evolved with the chaperone systems, and selection for efficient interaction with chaperones may somehow have favored certain domain sizes. An intriguing explanation involves the mechanism of genetic recombination. Domains may have evolved as recombinational units, whose size is constrained to lengths where DNA can easily form a flat circle, 275–300 bp, almost exactly corresponding to the observed peaks of protein domain size (Trifonov, 1994).

Whatever the explanation, the data would seem to speak clearly. Domain sizes and segment numbers are constrained enough so that these two factors alone are sufficient to make reasonable guesses about protein domain organization, for proteins less than about 400 residues long. In certain contexts, such as fold recognition calculations, these guesses may sometimes be accurate enough to be useful.

Acknowledgements

The authors wish to thank Anna Panchenko for useful discussions and the NIH intramural research program for support.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,A.L., Kolker,E. and Trifonov,E.N. (1994) Underlying order in protein sequence organization. *Proc. Natl. Acad. Sci. USA*, **91**, 4044–4047.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Bryant,S.H. and Altschul,S.F. (1995) Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.*, **5**, 236–244.

Bukau,B. and Horwich,A.L. (1998) The Hsp70 and Hsp60 Chaperone machines. *Cell*, **92**, 351–366.

Gibrat,J.-F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Hartl,F.U. (1996) Molecular chaperones in cellular protein folding. *Nature*, **381**, 571–580.

Islam,S.A., Luo,J. and Sternberg,M.J.E. (1995) Identification and analysis of domains in proteins. *Protein Eng.*, **8**, 513–525.

Jones,S., Stewart,M., Michie,A., Swindells,M.B., Orengo,C. and Thornton,J.M. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.

Madej,T., Gibrat,J.-F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Marchler-Bauer,A. and Bryant,S.H. (1997) Measures of threading specificity and accuracy. *Proteins*, Suppl 1, 74–82.

Marchler-Bauer,A. and Bryant,S.H. (1999) A measure of progress in fold recognition? *Proteins*, Suppl 3, 218–225.

Matsuo,Y., Bryant,S.H. (1999) Identification of homologous core structures. *Proteins*, **35**, 70–79.

Moult,J, Hubbard,T., Fidelis,K. and Pedersen,J.T. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, Suppl 3, 2–6.

Murzin,A.G. (1999) Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins*, Suppl 3, 88–103.

Panchenko,A., Marcher-Bauer,A. and Bryant,S.H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, Suppl 3, 133–140.

Sowdhamini,R., Rufino,S.D. and Blundell,T.L. (1996) A database of globular protein structural domains: clustering of representative family members into similar folds. *Fold Des.*, **1**, 209–220.

Trifonov,E.N. (1994) On the recombinational origin of protein-sequence-subunit structure. *J. Mol. Evol.*, **38**, 543–546.

Wang,Y., Addess,K.J., Geer,L., Madej,T., Marchler-Bauer,A., Zimmerman,D. and Bryant,S.H. (2000) MMDB: 3D structure data in Entrez. *Nucleic Acids Res.*, **28**, 243–245.