doi:10.1006/jmbi.2001.5387 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **316**, 839–851

JMB

SnapDRAGON: a Method to Delineate Protein Structural Domains from Sequence Data

Richard A. George and Jaap Heringa*

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

We describe a method to identify protein domain boundaries from sequence information alone based on the assumption that hydrophobic residues cluster together in space. SnapDRAGON is a suite of programs developed to predict domain boundaries based on the consistency observed in a set of alternative *ab initio* three-dimensional (3D) models generated for a given protein multiple sequence alignment. This is achieved by running a distance geometry-based folding technique in conjunction with a 3D-domain assignment algorithm. The overall accuracy of our method in predicting the number of domains for a non-redundant data set of 414 multiple alignments, representing 185 single and 231 multiple-domain proteins, is 72.4%. Using domain linker regions observed in the tertiary structures associated with each query alignment as the standard of truth, inter-domain boundary positions are delineated with an accuracy of 63.9% for proteins comprising continuous domains only, and 35.4% for proteins with discontinuous domains. Overall, domain boundaries are delineated with an accuracy of 51.8%. The prediction accuracy values are independent of the pair-wise sequence similarities within each of the alignments. These results demonstrate the capability of our method to delineate domains in protein sequences associated with a wide variety of structural domain organisation.

*Corresponding author

Keywords: protein; domain; boundaries; prediction; folding

Introduction

Understanding the domain content of a protein is a crucial step for many areas in protein science. For example, structural studies by NMR and X-ray crystallography have been greatly aided by the consideration of the modular nature of proteins.[1] The ability to build constructs based on knowledge of the domain boundaries is particularly important for structure elucidation by NMR, which requires relatively small proteins for analysis.[2] Furthermore, using sequence fragments corresponding to individual domains in a database search for related sequences is often more successful than using the whole protein sequence.[3] This is because individual domains are most likely to correspond to recurring functional and evolutionary units of a protein.[4] Nature brings many domains together with an almost infinite number of combinations. Based on this principle, structural genomics initiatives need only solve the structures for these recurring domains, and then use them as molecular templates for comparative modelling.[5]

Wetlaufer[6] first proposed the concept of the domain in 1973 after X-ray crystallographic studies of hen lysozyme,[7] papain,[8] and limited proteolysis analyses of immunoglobulins.[9,10] Wetlaufer defined domains as stable units of protein structure, which could fold autonomously. Although there is no absolute definition, domains are generally regarded as compact, semi-independent units,[11] where each domain contains an identifiable hydrophobic core.[12]

Identification of domains from protein sequence has become an intensely researched area. Most efforts of domain delineation have relied on comparative sequence searches in an attempt to infer domain boundaries from homology.[13–18] These methods have been successful at identifying modules, i.e. domains corresponding to a contiguous sequence segment, when sequence similarity is above the so-called twilight zone. At lower sequence similarity levels, however, evolutionary relationships are often discernible only at the level of tertiary structure. Furthermore, without any variation of domain connectivity within a protein

E-mail address of the corresponding author: jhering@nimr.mrc.ac.uk

0022-2836/02/030839–13 $35.00/0

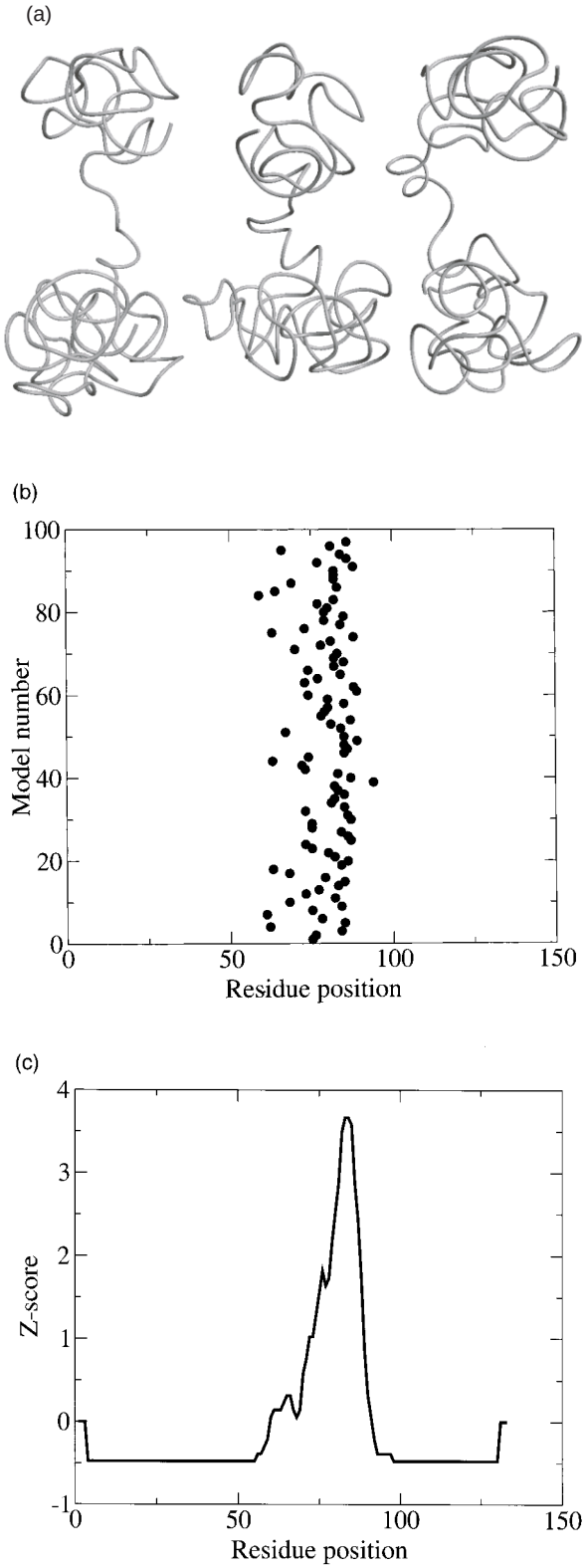

Figure 1. Overview of the SnapDRAGON method. (a) Example of three DRAGON models (visualised using the MOLMOL package[30]) generated for protein 2eifA. A total of 100 models are generated for a given multiple sequence alignment and predicted secondary structure information. (b) Plot of the boundaries predicted for each of the 100 DRAGON models *versus* the sequence position. Domain boundaries are assigned to each DRAGON model using the method by Taylor.[31] (c) Plot of the boundary *Z*-scores *versus* the sequence position. The boundaries for each DRAGON model are summed along the length of the sequence and smoothed using a biased window approach. The resulting boundary scores are then converted to *Z*-scores. Each significant peak in the graph represents a predicted boundary. For details, see the text.

family, domain positions cannot be inferred. Comparative sequence-based methods have further difficulty in boundary assignment where domains are highly associated or discontinuous, i.e. where more than one segment of a chain is required to form a complete domain. Even if the proper domain relationships are recognised in principle, the assignments by these alignment-based methods may show significant shifts relative to the exact structural boundaries between the domains.[19]

Relatively few methods have been developed for the assignment of domains based on physical principles. Busetta and Barrans[20] early on applied a protein folding method based on the interaction and accumulation of secondary structure units into domains. Regions with weak interactions between secondary structures were defined as domain boundaries. Kikuchi *et al*.[21] predicted contacts between residues based on statistical observations and then associated structural domains with areas of high contact density in a two-dimensional residue contact map. Finally, Vondervisztand Simon[22] attempted to predict domain boundaries using the concept that short-range interactions play a dominant role in domain stabilisation and that regions between domains would have a lesser preference for short-range interactions. Despite the creativity in these early approaches, none appeared successful in providing reliable domain boundary predictions.[23] Recently, Wheelan and co-workers developed a method to predict boundary locations using statistical knowledge of domain lengths.[24] However, accurate results using this method are limited to two-domain proteins with less than 300 residues.

Here, we introduce a new method for domain boundary prediction, SnapDRAGON. It incorporates an *ab initio* protein folding method, DRAGON,.[25-27] which folds a polypeptide based primarily on the notion of conserved hydrophobicity of amino acids, as well as secondary structure prediction by the PREDATOR technique.[28,29] In principle, SnapDRAGON employs the DRAGON algorithm to generate a large number of *ab initio* 3D model structures for a given multiple sequence alignment (with predicted secondary structure) and assigns automatically the domain boundaries for each of the models. Final predicted domain boundaries are derived from the consistency of the domain boundary assignments observed in the set of alternative 3D models.

Model generation by SnapDRAGON results in a set of 3D models that vary in structure with different domain contents and associated boundary positions. However, at this stage we are not interested in the details of the overall fold, but merely if we can consistently form isolated globular units given a multiple alignment and a notion of secondary structure (for a summary of the SnapDRAGON method, see Figure 1).

Results

Database

SnapDRAGON was applied to a non-redundant set of 414 multiple alignments (see Methods). Each of the alignments is associated with a known structure in the PDB depository,[32] for which the domain boundaries were assigned using a consistency criterion over three techniques (see Methods). The alignments show a wide distribution of protein lengths and domain numbers. The data set consists of 183 singular domain proteins and 231 multiple domain proteins. Of the latter, 98 structures comprise at least one discontinuous segment.

DRAGON model consistency

Of 183 single-domain proteins, an average of 66.6 % (±18.3 %) of the DRAGON models per protein show a single domain structure. Based on these models, SnapDRAGON predicts no boundaries for 86 of these single-domain proteins (47 %). Figure 2 represents the error within domain and boundary number assignment made by DRAGON for each alignment, calculated by subtracting the real number of domains from the average number assigned to the 100 DRAGON models generated for each test alignment. The average domain number assignment is slightly over-estimated in our protein set, with an average of +0.22 (±0.72) per protein. The mean number of domain boundaries assigned to each protein (Figure 2(b)) is also somewhat over-estimated with an average over all proteins of +0.66 (±1.33). This is mainly due to a tendency of the DRAGON method to form discontinuous domains, which happens particularly if modelling is carried out for longer alignments. One of the reasons for this is that DRAGON models may not always be as compact as real proteins, such that a ''single domain'' will sometimes be assigned as two domains with many inter-domain linkers.

Figure 3(a) and (b) displays, respectively, the average number of domains and boundaries for the DRAGON models generated for each alignment. Despite the above slight over-predictions, both average domain and boundary numbers correlate well with the true data, with cross-correlation coefficients of 0.56 and 0.58, respectively. The correlation coefficients for the distribution of average domain numbers, assigned to the DRAGON models as a function of chain length, and that of the real structures are 0.652 and 0.789, respectively. The corresponding regression lines both show a slope of 0.004 (Figure 3(a)), suggesting a balanced prediction of domain number over the length distribution of the test proteins. The correlation coefficients for the distribution of average boundary numbers over 100 models per protein against the chain length and the real domain numbers are 0.923 and 0.601, with slopes of 0.011 and 0.007, respectively (Figure 3(b)), confirming the

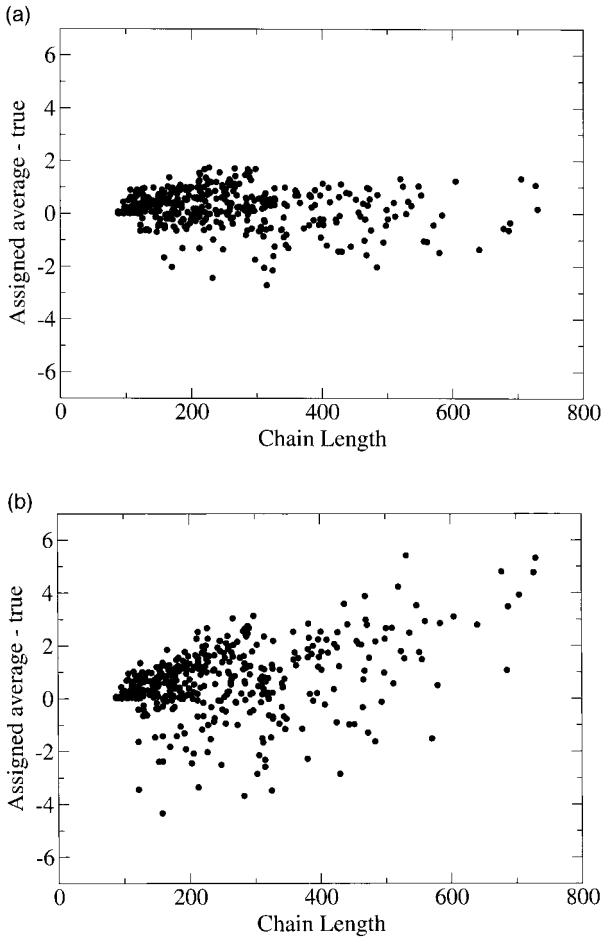

Figure 2. Error in average domain (a) and boundary (b) number over 100 DRAGON models for each protein, *versus* the sequence length of the corresponding reference sequence within each alignment. The errors were calculated by subtracting the real numbers of domains or boundaries from the corresponding average assigned numbers.

over-prediction for longer chain lengths. The bias towards more complex structures with larger alignments becomes salient when we calculate the correlation for smaller proteins (less than 400 residues), which results in regression line slopes of 0.010 for both boundary distributions. While the spread of the domain numbers across the models for each protein is almost constant over the length distribution of proteins in our reference data set (Figure 3(c)), with growing chain lengths, the standard deviation associated with the average boundary number increases (Figure 3(d)), reflecting the greater structural variability of DRAGON models for larger proteins.

Tuning the prediction protocol

Using the domain boundary assignments across 100 DRAGON models for each protein, average positional scores are calculated by sliding a window using a biased averaging protocol (see Methods) to smooth the added domain boundary assignments from each model (Figure 1). The distribution is normalised to a zero mean and unit standard deviation. A *Z*-score cut-off of 2.00 gave the best results with an average number of boundary predictions, i.e. peaks scoring beyond the cut-off, of 2.10. This number approaches the average number of 1.97 for the real domain boundaries in our data set. The biased smoothing of boundary assignments reduced the average over-prediction from +0.66, which was calculated from the average number of boundaries assigned to each DRAGON model, to +0.13 for each protein. Moreover, the distribution of the SnapDRAGON prediction errors *versus* the length of the reference proteins (Figure 4(a)) shows that the length bias has disappeared. Figure 4(b) depicts the frequencies of the types of correct and erroneous prediction of domain boundary numbers: the main diagonal corresponds to correct prediction, which corresponds to 38.1 % of proteins having their correct boundary number assigned by SnapDRAGON while 78.4 % of the reference proteins have correct boundary assignments within an error margin of ±1.

Evaluating boundary prediction

Table 1 shows the accuracy of SnapDRAGON, and that of two ''baseline'' approaches (see

Methods), in predicting domain boundary locations within the 231 multiple domain proteins. Here, a boundary prediction is regarded as successful whenever it is positioned within the stretch of a real linker within the structure associated with each of test data set entries, defined as detailed in Methods. Two measures of prediction accuracy are shown in Table 1. The first measure is the percentage of linkers in each of the test proteins accurately predicted by our method: $TP/(TP + FN)$, where TP is the number of true positive boundary predictions and FN the number of false negatives, i.e. real boundaries that have not been predicted. While in sequence database searches this number is alternatively referred to as sensitivity or coverage, it will be called linker coverage here. The second measure is the percentage of all SnapDRAGON boundary predictions that landed within observed linkers: $TP/(TP + \mathrm{FP})$. The latter measure is sometimes referred to as positive prediction value (*PPV*), but here will be named prediction success.

As a control, we compared the SnapDRAGON prediction results with those of two different baseline methods, both based on structural characteristics (see Methods). The first baseline method predicts boundary positions by separating a sequence into equally sized domains, based on the number of boundary predictions made by SnapDRAGON. The second baseline method predicts the number of domain boundaries based on the observed distribution of domain sizes, such that it can result in a different number of linker predictions than SnapDRAGON or the first baseline method.

Prediction results

The overall accuracy of SnapDRAGON domain content prediction for our data set is 72.4(±17.4)%, which was calculated using the observed number of domains in the structures associated with each database entry as a reference. Calculated over all linkers in our reference data set, the linker coverage in the continuous domain set is 55.2%, while, due to a lower coverage of 33.0% for linkers within discontinuous proteins, the overall SnapDRAGON accuracy is 42.3% (Table 1A). Overall, the first baseline method shows a linker coverage of 30.6%, while the second baseline method reaches 30.3% (Table 1A). Hence the improvement of our method is 38% and 40%, respectively, relative to the two baseline methods. Table 1B shows the linker coverage per protein. An average of 51.8(±39.1)%, 34.7(±40.8)% and 35.7(±41.3)% of true linkers are found by the SnapDRAGON and two baseline methods, respectively. The overall relative coverage increase of SnapDRAGON compared to the best baseline method is 45%. The prediction of boundaries for continuous proteins becomes relatively accurate here with a per-protein coverage of 63.9%, while the prediction for discontinuous proteins falls behind at 35.4%, constituting a relative loss in linker coverage of 45% compared to prediction for continuous proteins. Given the relatively small numbers of domain boundaries within proteins, the high error numbers of the prediction averages per protein are expected (Table 1B).

The over-prediction of SnapDRAGON is reflected in the prediction success, which shows slightly lower values. However, the success rate of the predictions (measure 2) is well balanced: Over the discontinuous proteins, the success rate of 40.7% is slightly higher than that for the continuous proteins. Nonetheless, a consequence of over-predicting is that we introduce a number of false positives. Overall, the proportion of false positive linker predictions of our method is 60%. Note, however, that other methods produce a number of top predictions for each protein in a reference set,

Table 1. Accuracy of linker prediction

		Continuous set	Discontinuous set	Full set
A. *Average prediction results per linker (%)*				
SnapDRAGON	Coverage	55.2	33.0	42.3
	Success	39.1	40.7	39.8
Baseline 1	Coverage	39.6	21.1	30.6
	Success	29.6	21.6	26.4
Baseline 2	Coverage	41.7	22.3	30.3
	Success	25.9	21.2	23.6
B. *Average prediction results per protein*				
SnapDRAGON	Coverage	63.9(±43.0)	35.4(±25.0)	51.8(±39.1)
	Success	46.8(±36.4)	44.4 (±33.9)	45.8(±35.4)
Baseline 1	Coverage	43.6(±45.3)	20.5(±27.1)	34.7(±40.8)
	Success	34.3(±39.6)	22.2(±29.5)	29.6(±36.6)
Baseline 2	Coverage	45.3(±46.9)	22.7(±27.3)	35.7(±41.3)
	Success	37.1(±42.0)	23.1(±29.6)	31.2(±37.9)

Prediction scores are given for the set of 231 multiple domain proteins. Coverage is the percentage real linkers predicted (measure 1). Success is the percentage of correct predictions made (measure 2). Baseline 1 are predictions made using the number of domains predicted by SnapDRAGON and Baseline 2 are predictions made using predicted domain number based on the observed distribution of domain sizes.

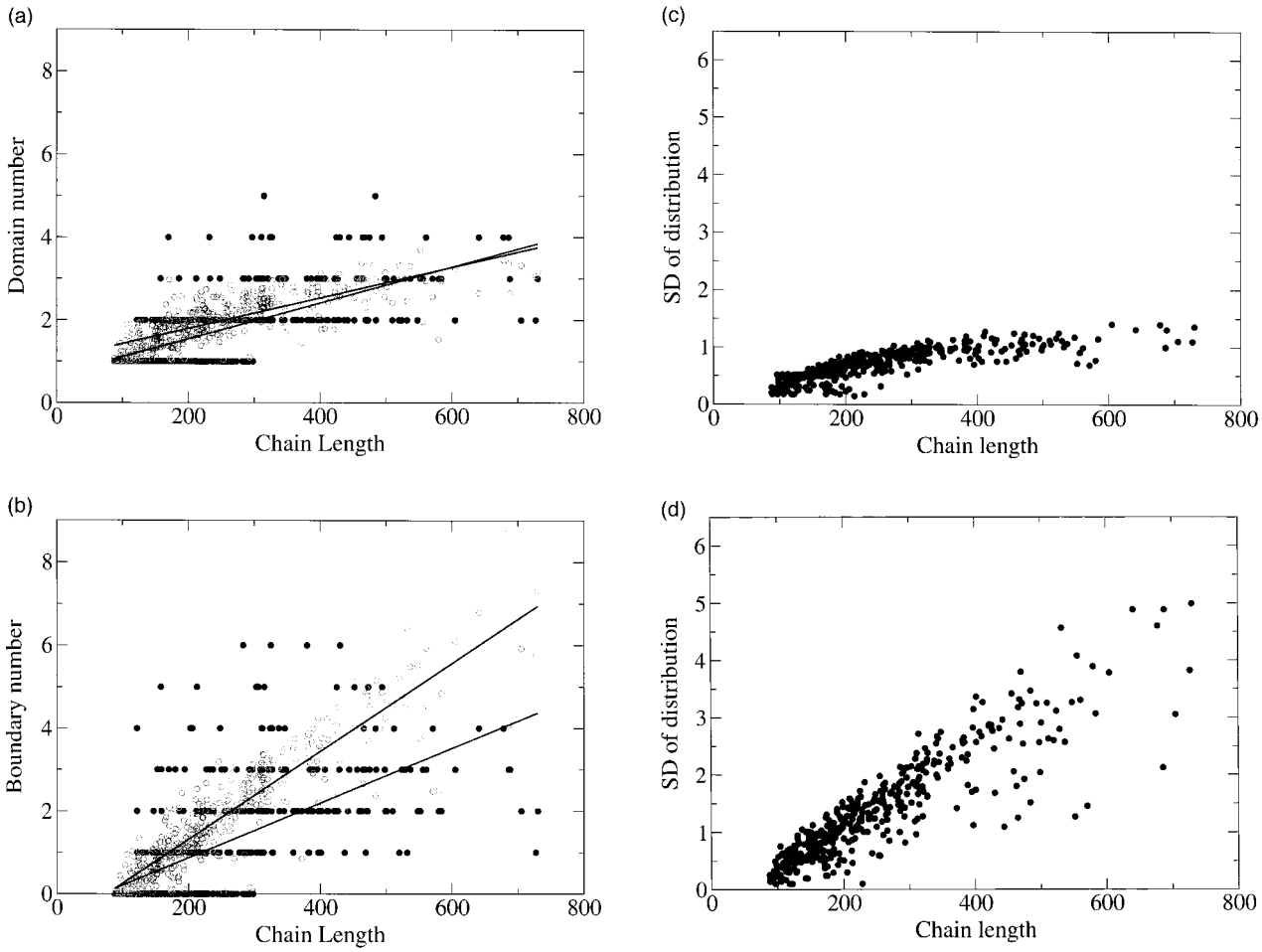

Figure 3. (a) The average number of domains (○), assigned over 100 models for each alignment, and true number (●) plotted against the reference chain length. (b) The average number of domain boundaries over 100 models for each alignment (○) and the true number (●) *versus* the reference chain length. (c) Standard deviation of the number of domains over 100 DRAGON models for each alignment against the reference chain length. (d) Standard deviation of boundary number over 100 DRAGON models for each alignment *versus* the reference chain length.

which are then evaluated. For example, the results stated by Wheelan *et al.*[24] are based on ten top predictions for each protein, based on a linker region definition of ±20 residues around true boundaries, which has been used in previous studies.[24, 33] At this resolution, SnapDRAGON attains a linker coverage of 60.7 % and a prediction success of 56.3 %, while the second baseline method scores 46.0 % coverage and 35.2 % prediction success.

While the performance of the three approaches generally falls with increasing chain lengths, it is interesting that the linker coverage for the discontinuous set does not drop with increasing sequence lengths (data not shown). The chance of successfully predicting a linker in the discontinuous set might increase with larger proteins because of the aforementioned tendency of DRAGON to generate large structures with increased complexity in domain organisation and the fact that the sequential separation of linkers is much smaller in discontinuous proteins than in the continuous set. Indeed, the average fraction of linker residues in discontinuous proteins of 20.6(±7.5) % is much larger than the fraction for continuous proteins, which averages at 14.1(±6.9) %, leading to the aforementioned overall fraction of 16.8 % over all proteins.

Yet another way of assessing our method is to compare it with random prediction, which would naively have a chance of predicting real linkers directly proportionate to the fraction of total linker regions in each of the test proteins. As the fraction of linkers per protein in our multiple domain set is 16.8(±7.8) %, random prediction, based on drawing with replacement, would lead to 16.8 % accuracy overall. Therefore, SnapDRAGON performs more than three times better than random.

Distance of predicted boundaries from linker regions

An alternative way to measure the prediction accuracy is calculating the distance between a predicted boundary and the nearest true linker region. It is calculated here by taking the number of residues separating the prediction from the nearest

end of a linker and normalising this number by the sequence length to facilitate comparison. Linkers are predicted by our method with an average distance of 7.0(±10.8)% of the sequence length. The two baseline methods attained a relative distance of 9.3(±10.9)% and 9.1(±10.6)%, respectively. The best baseline method is therefore on average more than two percentage points further away from a linker than is SnapDRAGON, which, for example, amounts to a difference of seven residues for a 350-residue protein.

Independence from sequence identity

There is no apparent correlation (-0.02) between prediction success and alignment divergence. Mean pair-wise sequence identities in each test alignment range from 28% to 71%, calculated using the Blosum62 amino acid substitution matrix and gap penalty values of 12 and 1 for gap initiation and extension, respectively. Our method appears insensitive to sequence identity in pinpointing correct domain boundaries.

Discussion

Application of the method

The identification of the exact position of N and C termini of domains within a protein is an important first step in many areas of molecular biology. Several studies have highlighted the difficulty in identifying domain boundaries, showing that incorrect assignment can lead to completely unfolded peptides.[19,34,35] We have described a method based on sampling generated 3D models, which are built using information from multiple alignments and secondary structure prediction. Our method is able to successfully predict the domain content of a protein, and can also position domain boundaries with reasonable success. This might be helpful for any structural biologist required to delineate structural domains in a protein sequence, as domain prediction will identify target proteins for structural elucidation. Alternatively, knowledge about the domain content of a query sequence might aid comparative modelling efforts. In general, the recognition of domains in protein sequences will greatly improve sequence analysis and structure prediction.[3,36]

Modelling the hydrophobic collapse

Ab initio folding of a protein sequence into its tertiary structure remains one of the greatest challenges in computational biology, a goal that has eluded researchers for the last 30 years. Although many models of protein folding have been described (for a detailed review, see Dill[37]), the main driving force of protein folding was recognised early on to be the burial of hydrophobic side-chains into the interior of the molecule, creating a hydrophobic core and hydrophilic surface.[38] Since then, many authors (e.g. Rackovsky and Scheraga[39], Dill[40]) have stressed the importance of the hydrophobic collapse for protein folding. Since hydrophobic interactions play a major role in organising and stabilising the tertiary structure of proteins,[41,42] such interactions are seen to be preserved in fold families where there is little or no sequence homology between its members.[43] Following this notion, we assume that protein folding is a result of chain collapse driven by the hydrophobic effect, effected in our tertiary model building efforts by preferentially bringing residues together in space that show conserved hydrophobicity in the associated query alignment. Model building is also based on predicted secondary structure elements, which lower the degrees of freedom of our folding method in packing hydrophobic residues. The consistency of the models to pack hydrophobic residues into multiple cores provides insight into domain formation.

Model variation

The SnapDRAGON technique requires the generation of 100 test models. Most if not all of these models are likely to contain structurally incorrect regions and will vary greatly in structure between models. For example, building 100 models for a multiple alignment associated with a protein of about 150 residues in length can lead to an average root-mean-square deviation (RMSD) of over 10 Å, when measured after all-against-all structural super-positioning.[44] It is commonly accepted that with RMSD values of more than 6 Å, any structural relationship between a pair of structures cannot be inferred anymore. However, the variation between the models gives us a means to evaluate hydrophobic packing at the course-grained level of domain folding. Although the main-chain trace through the protein core can be dramatically different in each of the models, the consistency in packing ensembles of hydrophobic residues provides a discernible signal for domain boundary placement.

Control methods

There are a number of domain prediction algorithms available, which attempt to delineate domain boundaries based on finding homologous sequence fragments in databases, for example by PSI-BLAST.[14] However, there are no prediction algorithms available that predict the boundaries relying on first principles, such as the SnapDRAGON method introduced here, with which we can readily compare our method. The approach by Wheelan *et al*.[24] is based on domain size distributions, but in practice does not give a definitive answer and often provides a list of very different solutions. We therefore designed two baseline methods to check the validity of our prediction results. Both methods naively delineate domains by cutting the sequence in equal parts, although the methods differ in determining the number of boundaries per protein:

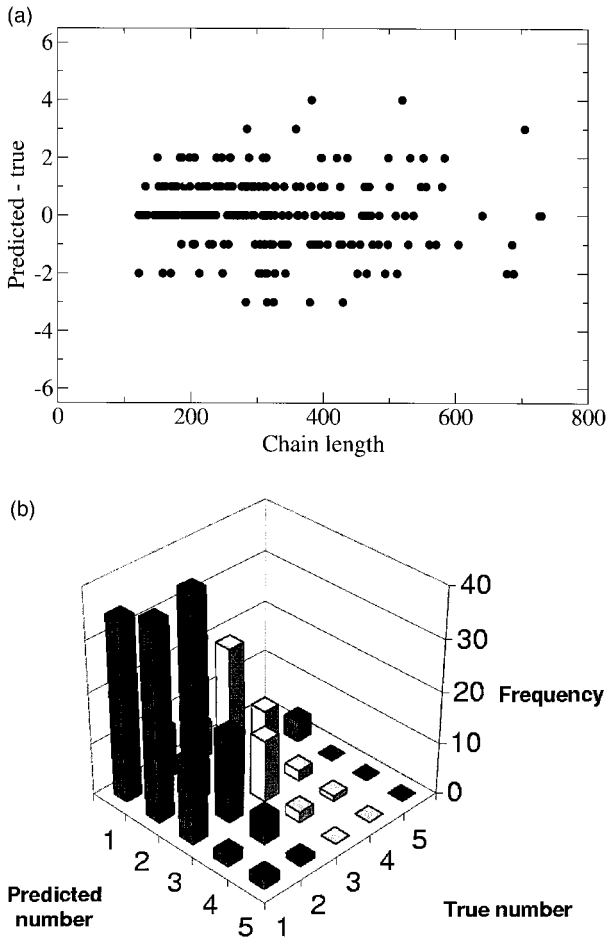

Figure 4. (a) Error in the number of boundaries assigned by SnapDRAGON to each alignment, *versus* the sequence length of the reference sequence within each alignment. The errors were calculated by subtracting the real number of domains or boundaries from their predicted number. (b) 3D histogram of predicted number of boundaries *versus* the true number for each reference sequence.

while the first baseline method uses the number predicted by SnapDRAGON, the second method contains an algorithm that optimises the number of predicted domains following the distribution of domain sizes observed in the database. This brings the second baseline close to the approach by Wheelan *et al.*[24] A further step in raising the complexity and possible predictive power of the latter baseline method might involve using the length distribution of domain types in classes such as α, β, α/β domain structures, and then use the distribution of predicted secondary structure elements in assigning the most likely types of domains with associated sizes for a given protein sequence. However, we inspected the CATH database[46] and found that the length distributions of the deposited α and β-type domains were identical with an average value of 136 residues. Although the α/β domains in CATH displayed a slightly larger average length of 180, this is probably due to an over-representation of the α/β-barrel fold in the CATH database. Given the similar size distributions of the various domain classes, and uncertainty in combining domain class-related statistics and secondary structure information, it appears that our current two baseline control methods are appropriate.

Statistical significance of DRAGON models

Here, we assess the statistical significance of boundary placements over 100 DRAGON models generated for an alignment using a biased averaging protocol and normalisation of the data to zero mean and unit standard deviation (self *Z*-scores). Peaks are then taken as significant whenever they are more than 2.0 standard deviations above the mean (Figure 1(c)). Previously we have employed a random background for statistical comparison by creating a set of five randomised alignments with associated predicted secondary structure on which DRAGON was used to generate 500 models, i.e. 100 models for each alignment. The same protocol as for the self *Z*-scores here was then used to compile an average random score value and standard deviation from the distribution of the boundaries in the 500 background models. The rationale behind scrambling the alignment columns is to assess the uniqueness of the hydro-

phobic collapse given the distribution of conserved hydrophobic positions over the alignment. As this distribution is believed to be crucial for folding in multi-domain proteins, the scrambled alignments are expected to lead to models with entirely different and spurious domain structures, thereby effectively randomising the placement of domain boundaries. However, the generation of such a random background for each protein appeared prohibitive for the number of proteins we have studied here, due to the time required for generating the vast number of models (see Methods). However, we have studied the effect of the randomisation for a smaller independent non-redundant set of 57 protein alignments, the results of which are in Table 2. The Table compares the prediction results using the random background from 500 control models, using the optimal *Z*-score cut-off of 2.0, with those resulting from a self *Z*-score threshold of 1.0 and 2.0. While caution should be taken because of the small database size, it is clear that good prediction results can be achieved when *Z*-scores from randomised control models are used; in this case leading to an average linker coverage of 54.8 %. However, the best prediction with a linker coverage 64.7 % is reached for the continuous subset using a self *Z*-score cut-off of 1.0, albeit the at the expense of over-prediction, leading to lower prediction success rates (Table 2). The evaluation of prediction results using the randomised background models is functional in SnapDRAGON and should be included in boundary predictions for small data sets.

Domain size and folding

It might be argued that the DRAGON method should ideally be used to generate structures for a single domain. However, when faced with larger protein sequences, DRAGON becomes unable to form a single hydrophobic core and hence is forced to split the protein into several domains. Such a phenomenon is also observed in nature, where domain length follows a narrow distribution with an average of about 100 residues.[45] To account for domain formation in proteins, Dill[40] developed a lattice model and predicted an optimal chain length for maximal protein stability. If a polypeptide chain is short, less than 70 residues, the protein has little interior volume to form enough favourable contacts between hydrophobic residues. On the other hand, a long polypeptide would be forced to bury many unfavourable hydrophilic residues into the interior, given a particular ratio of hydrophobic and hydrophilic residues. Long stretches of hydrophobic residues, which would enable large hydrophobic cores, are rarely observed in sequences of soluble proteins. For example, hydrophobic sequence stretches within the globular proteins in the CATH[46] domain database rarely go beyond a length of ten amino acid residues, when as hydrophobic residues the types Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp and Tyr are taken. Consequently there are no observed protein structures of more than 250 residues that contain a large single hydrophobic core.[47] Given the observed random distribution of hydrophobic residues in proteins, domain formation appears to be the optimal solution for a protein to bury its hydrophobic residues whilst keeping hydrophilic residues at the surface.

Using external information

Various sources of external information can be used to enhance the domain recognition process. For example, when dealing with very large protein sequences, it could be beneficial, given the computational expense of our method, to include information from homology searches against sequence databases such as Pfam[48] and PRODOM.[14] The consistent matching of local alignments obtained by such searches can pinpoint where a query sequence or alignment might be cut. The resulting fragments could then be given to SnapDRAGON for further refinement of the boundary predictions. Also, transmembrane and coiled coil regions are points that separate globular domains, which can be reliably identified in protein sequence using predictive tools.[49,50] Further, many domains in protein structures are repeats, which vary dramatically in divergence[51] but can be delineated using accurate prediction tools.[52] Incorporating such filter methods is likely to enhance the accuracy of domain boundary prediction from sequence. Many sources of additional information can readily be specified and used by DRAGON, such as specific distance restraints (lower and upper bounds) and predicted or experimentally determined residue solvent accessibility values.[53] This makes it relatively easy to encode domain barriers such as transmembrane segments or repeat boundaries, thus driving DRAGON to pack the segments flanking these barriers into separate domains.

Table 2. Average accuracy percentages of linker prediction over 57 proteins

		Continuous set	Discontinuous set	Full set
Randomised background *Z*-score >2	Coverage	63.3	43.6	54.8
	Success	27.2	31.1	28.9
Self-normalised *Z*-score >1	Coverage	64.7	39.5	53.5
	Success	26.6	31.7	28.9
Self-normalised *Z*-score >2	Coverage	48.7	24.3	38.7
	Success	41.3	28.3	29.9

Materials and Methods

SnapDRAGON is a suite of programs developed for the prediction of domain boundaries based on information from a multiple alignment of protein sequences and secondary structure prediction. All programs were written in ANSI C, C++, and Perl5 and run on a Linux cluster of 128 Pentium III processors. A summary of the method is presented in Figure 1.

Constructing the 3D models

We used the DRAGON method by Aszódi and Taylor[25–27] to construct the basic 3D model structures. DRAGON takes a multiple alignment and secondary structure information as input, and initially constructs a random high-dimensional C^{α} distance matrix for a target sequence. Distance geometry, incorporating a hierarchical projection method, is then used to find the 3D conformation corresponding to a pre-described target matrix. The target matrix consists of desired distances between residues, which are inferred from (i) the input multiple alignment, based on the assumption that pairs of residues that are both conserved and hydrophobic are likely to be close together in the core of the molecule,[54] and (ii) the types and interactions of secondary structure elements.

Hierarchical projection is repeated until optimal convergence between the random and the target matrix is achieved, as well as full embedding into three dimensions. The embedding into three dimensions is achieved through a ''divide and conquer'' principle; based on the classic embedding approach by Crippen and Havel,[55] which, due to its reliance on matrix diagonalisation, is a computationally expensive operation. Here, the C^{α} distance matrix is first divided into smaller clusters. Then, separately, each cluster is embedded toward a local centroid. A final structure is generated from full embedding of the multiple centroids and their local structures. Final structures generally have hydrophobic cores and hydrophilic residues on the surface. In addition to the secondary structure elements specified by the user, secondary structure formation can also spontaneously occur when chain segments approach each other below a pre-set distance threshold. In this way the hydrophobic effect influences the formation of hydrogen bonds.[56]

Using the above scenario, the DRAGON method can build a set of alternative models, where the construction of each model starts with a differently randomised C^{α} distance matrix (see above). This results in a set of final models that are likely to vary in structure with different domain boundary positions. However, at this stage we are not interested in the details of the overall fold, but merely if we can consistently form isolated units of globularity given a multiple alignment and a notion of secondary structure.

Running DRAGON

As mentioned above, DRAGON requires a multiple alignment of sequences related to a target protein and secondary structure information. Here, three-state secondary structure (α, β, coil) was predicted using PREDATOR[28,29] for each sequence in the alignment, a consensus assignment (75 %) was then used. In the case of β-strand annotation, the original DRAGON method requires a description of the strand connectivity in β-sheets. Since we assume we do not know the strand connectivity, we modified the input protocol to be able to enter strand information as distance constraints from the N to the C-terminal C^{α} atom within each strand, thus avoiding specifying their connectivity. Maximum and minimum distances for a β-strand were calculated by multiplying the number of peptide bonds in a strand by 3.2 Å (antiparallel) and 3.4 Å (parallel), respectively.[57] Helix predictions for an input alignment could be given to DRAGON directly. The DRAGON method enables the specification of confidence values for each of the specified secondary structure. We used a value of 0.4 for helices and 0.8 for strands (range 0-1), which is a stringency that biases DRAGON to follow the recommendations, while deviations can still occur. Using the above scenario, 100 model structures were generated for each test alignment (see below).

Domain boundary assignment

Once 100 DRAGON models are generated for a given alignment, a domain assignment tool is run on each of the structures. We use the technique of Taylor[31] to delineate the domain boundaries. This method initially assigns to each residue in the protein structure a numeric label (initially the sequential residue number). It then iteratively changes the labels of the residues based on their respective neighbourhoods. If a residue is surrounded by neighbours (within a given radius r) with, on average, a higher label than the central residue, its label is increased by one; otherwise it is lowered by one. This test and label reassignment is made repeatedly to each residue in the protein chain until convergence is reached. This neighbourhood-based iteration protocol results in compact regions iterating towards the same residue number. The final domain boundaries are then assigned in between such compact domain regions.

Calculating domain boundary scores

All domain boundaries obtained, if any, are recorded. The proposed boundary positions in each of the 100 DRAGON models for an alignment, are then summed along the length of the target sequence and smoothed using a biased sliding window (Figure 1). The window is slid over the sequence one residue at a time, where the values at more central positions within the window are weighted higher than values toward either end of the window:

$$S = \sum_{i=1}^{n} S_i \times W_i$$

where S_i is the score and W_i the weight at position i. The latter is defined as;

$$W_i = 1/p^2(p - |p - i|)$$

where:

$$p = 1/2(n + 1)$$

This means that:

$$\sum_{i-1}^{n} W_i = 1$$

The window score is therefore biased towards more cen-

tral positions within the window, leading to a smoother curve with defined peaks, which appeared not to be attainable with an unbiased sum-and-divide method of smoothing. The biased weighting scheme allows very close peaks to merge into one, removing trivially different boundary predictions. Throughout this work we use a window size of 5% of the sequence length, rounded to the nearest higher odd integer. For example, a sequence of 200 amino acid residues will have a window length of 11 residues. A window size of 5% was found to consistently give the best results.

The domain boundary distribution was normalised to a zero mean (m) with corresponding unit standard deviation (σ). This is used to convert the positional boundary scores for the 100 test models into Z-scores ($Z = (x - m)/\sigma$). All peaks above a Z-score of 2.00 were then proposed as domain boundaries.

Reference data set

Predictions were assessed using proteins with known 3D structure taken from a non-redundant set held at the NCBI.[58] The proteins in this set are grouped by single linkage clustering based on a BLAST P-value less than 10^{-7}, and show pair-wise sequence identity values of 25% or less.[58] As for the predicted models, each structure was assigned its domain boundaries using the Taylor algorithm[31] (see above). To ensure correct domain annotation, assignments were compared to those made in the SCOP and DALI domain databases.[59,60] As boundary placement was fairly inconsistent over the three definitions, we limited the reference set to proteins for which at least two of the above three definitions corresponded to domain boundary assignments with a separation of ten residues or less.

Since the non-redundant protein set consists largely of single-domain proteins, we randomly selected 15% of these structures for analysis. Single domain proteins that had sequence lengths greater than 300 residues were removed from this set, as it is debatable whether such larger proteins would consist of only one domain.

Defining domain linkers in reference structures

Since the Taylor algorithm assigns domain boundaries indicated by a single residue position only, we delineated the domain linker regions corresponding to these boundaries in each protein using a definition based on secondary structure and solvent accessibility. To determine the exact linker segment connecting two domains, we branch out in two directions from the boundary position. Linker assignment will end if a branch hits a secondary structure comprising three or more residues for a strand or at least four residues for a helix, where secondary structures are assigned using the program DSSP.[64] A minimum length of 20 residues is then imposed on each linker (±10 residues from the boundary position). The average sequence length of the linkers extracted following the above scenario in our set is 23.3(±4.0) residues.

Constructing multiple alignments

A multiple sequence alignment was created for each protein in our set. Homologues to a protein were first identified using the iterative database search tool PSI-BLAST[61] using an E-value cut-off of 0.001 and a maximum of four iterations. To achieve maximal information content from the PSI-BLAST results, we filtered the sequences for redundancy by using the program OBSTRUCT.[62] OBSTRUCT produces the largest possible subset of protein sequences with pairwise sequence identity scores within a particular range. Here we used a range of >20% and <60% sequence identity, which led to an average number of 11.2 sequences per alignment. For each sequence set multiple sequence alignments were generated using the method PRALINE[63] with default parameters.

Control methods

We devised two baseline prediction methods to evaluate the accuracy of SnapDRAGON. The first method to compare our results against, predicts domain boundary positions by dividing the sequence into equal parts based on the number of predicted boundaries made by SnapDRAGON. For example, a protein predicted to have two boundaries is split into three equally sized segments. This is an appropriate method of domain prediction because protein chains tend to segregate evenly between domains, rather than most of a chain being associated with only one of the domains.[65] Also, in many proteins, domain structures are repeated within a single polypeptide chain.[66]

The second control method is similar to that used by Wheelan *et al*.[24] It first predicts the number of domains based on a probability density function (PDF) of known domain sizes. The PDF is constructed from a set of 2750 domains delineated using the domain cutting method of Taylor (1999) from a non-redundant protein set at the NCBI.[58] The distribution has a mean average length of 149 residues and the most frequent domain size is 97 residues. The probability that a protein sequence with a given length has one domain, two domains, three domains and so on, is calculated by referring to the observed data. The number of domains with the highest probability is taken as the prediction and, as in the previous method, the sequence is split equally into the predicted number of domains.

Performance

One run of SnapDRAGON, which produces 100 DRAGON models, will typically take under one hour for an alignment with a length less than 400 residues, provided 100 nodes on our Linux cluster are free. However, larger sequences will often take several hours because model generation by DRAGON becomes more computationally intensive. For example, while constructing 100 models for the alignment corresponding to the shortest reference structure of 88 residues in our data set took only ten minutes on our cluster, generating this number of models for the alignment associated with the largest structure of 730 residues needed nearly ten hours to complete. Providing general run-time scaling figures is complicated by the fact that within a SnapDRAGON run to generate 100 models for a given alignment, the variance in computing time for each of the individual models can be large. Given the large number of models needed for a single query alignment, it is clear that high-performance computing is required for running SnapDRAGON.

Availability of the method

The SnapDRAGON software will be made available upon request.

Acknowledgements

We thank Drs Willie Taylor and Andras Aszódi for helpful discussions and Nigel Douglas for expert handling of our computing resources. R.A.G. is a PhD student funded by the Medical Research Council. Two anonymous referees provided helpful suggestions which improved the manuscript.

References

1. Baron, M., Norman, D. G. & Campbell, I. D. (1991). Protein modules. *Trends Biochem. Sci.* **16**, 13-17.
2. Pfuhl, M. & Pastore, A. (1995). Tertiary structure of an immunoglobulin-like domain from the giant muscle protein titin: a new member of the I set. *Structure,* **3**, 391-401.
3. Sonnhammer, E. L. & Durbin, R. (1994). A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.* **10**, 301-307.
4. Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Letters,* **286**, 47-54.
5. Elofsson, A. & Sonnhammer, E. L. (1999). A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics,* **15**, 480-500.
6. Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA,* **70**, 697-701.
7. Phillips, D. C. (1966). The three-dimensional structure of an enzyme molecule. *Sci. Am.* **215**, 78-90.
8. Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M. & Wolthers, B. G. (1968). Structure of papain. *Nature,* **218**, 929-932.
9. Edelman, G. M. (1973). Antibody structure and molecular immunology. *Science,* **180**, 830-840.
10. Porter, R. R. (1973). Structural studies of immunoglobulins. *Science,* **180**, 713-716.
11. Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167-339.
12. Swindells, M. B. (1995). A procedure for detecting structural domains in proteins. *Protein Sci.* **4**, 103-112.
13. Adams, R. M., Das, S. & Smith, T. F. (1996). Multiple domain protein diagnostic patterns. *Protein Sci.* **5**, 1240-1249.
14. Gouzy, J., Corpet, F. & Kahn, D. (1999). Whole genome protein domain analysis using a new method for domain clustering. *Comput. Chem.* **23**, 333-340.
15. Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics,* **14**, 174-187.
16. Guan, X. & Du, L. (1998). Domain identification by clustering sequence alignments. *Bioinformatics,* **14**, 783-788.
17. Park, J. & Teichmann, S. A. (1998). DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics,* **14**, 144-150.
18. Sonnhammer, E. L. L. & Kahn, D. (1994). Modular arrangement of proteins as infered from analysis of homology. *Protein Sci.* **3**, 482-492.
19. Castiglone Morelli, M. A., Stier, G., Gibson, T., Joseph, C., Musco, G., Pastore, A. & Trave, G. (1995). The KH module has an alpha beta fold. *FEBS Letters,* **358**, 193-198.
20. Busetta, B. & Barrans, Y. (1984). The prediction of protein domains. *Biochim. Biophys. Acta,* **790**, 117-124.
21. Kikuchi, T., Némethy, G. & Scheraga, H. A. (1988). Prediction of the location of structural domains in globular proteins. *J. Protein Chem.* **7**, 427-471.
22. Vonderviszt, F. & Simon, I. (1986). A possible way for prediction of domain boundaries in globular proteins from amino acid sequence. *Biochem. Biophys. Res. Commun.* **139**, 11-17.
23. Valdar, W. (1997). The prediction of domain boundaries from amino acid sequence. M.Sc, University of Manchester.
24. Wheelan, S. J., Marchler-Bauer, A. & Bryant, S. H. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics,* **16**, 613-618.
25. Aszódi, A. & Taylor, W. (1994). Folding polypeptide α-carbon backbones by distance geometry methods. *Bioploymers,* **34**, 489-505.
26. Aszódi, A., Gradwell, M. J. & Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251**, 308-326.
27. Aszódi, A. & Taylor, W. R. (1997). Hierarchic inertial projection: a fast distance matrix embedding algorithm. *Comput. Chem.* **21**, 13-23.
28. Frishman, D. & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9**, 133-142.
29. Frishman, D. & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **27**, 329-335.
30. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51-55, 29-32.
31. Taylor, W. R. (1999). Protein structural domain identification. *Protein Eng.* **12**, 203-216.
32. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.
33. Kuroda, Y., Tani, K., Matsuo, Y. & Yokoyama, S. (2000). Automated search of natively folded fragments for high-throughput structure determination in structural genomics. *Protein Sci.* **9**, 2313-2321.
34. Macias, M. J., Hyvonen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, T. M. & Oschkinat, P. H. (1996). Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature,* **382**, 646-649.
35. Schmidt, C., Henkel, B., Poschl, E., Zorbas, H., Purschke, W. G., Gloe, T. R. & Muller, P. K. (1992). Complete cDNA sequence of chicken vigilin, a novel protein with amplified and evolutionary conserved domains. *Eur. J. Biochem.* **206**, 625-634.
36. Bonneau, R., Strauss, C. E. & Baker, D. (2001). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins: Struct. Funct. Genet.* **43**, 1-11.
37. Dill, K. A. (1999). Polymer principles and protein folding. *Protein Sci.* **8**, 1166-1180.
38. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advan. Protein Chem.* **14**, 1-63.
39. Rackovsky, S. & Scheraga, H. A. (1977). Hydrophobicity, hydrophilicity, and the radial and orienta-

tional distributions of residues in native proteins. *Proc. Natl Acad. Sci. USA,* **74**, 5248-5251.
40. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry,* **24**, 1501-1509.
41. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of heamoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669-678.
42. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science,* **229**, 834-838.
43. Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225-270.
44. Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1-22.
45. Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C. & Thornton, J. M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.* **7**, 233-242.
46. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.
47. Garel, J. (1992). Folding of large proteins: multidomain and multisubunit proteins. In *Protein Folding* (Creighton, T., ed.), 1st edit., pp. 405-454, W.H. Freeman and Company, New York.
48. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
49. Rost, B., Casadio, R. & Fariselli, P. (1996). Topology prediction for helical transmembrane proteins at 86 % accuracy. *Protein Sci.* **5**, 1704-1718.
50. Lupas, A. (1997). Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.* **7**, 388-393.
51. Heringa, J. (1998). Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.* **8**, 338-345.
52. George, R. A. & Heringa, J. (2000). The REPRO server: finding protein internal sequence repeats through the Web. *Trends Biochem. Sci.* **25**, 515-517.
53. Aszódi, A., Munro, R. E. & Taylor, W. R. (1997). Protein modeling by multiple sequence threading and distance geometry. *Proteins: Struct. Funct. Genet.* **Suppl 1**, 38-42.
54. Taylor, W. R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Protein Eng.* **4**, 853-870.
55. Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation*, Research Studies Press, Taunton, England, Somerset.
56. Aszódi, A. & Taylor, W. R. (1994). Secondary structure formation in model polypeptide chains. *Protein Eng.* **7**, 633-644.
57. Creighton, T. (1993). *Proteins: Structure and Molecular Properties*, 2nd edit., W.H. Freeman, New York.
58. Matsuo, Y. & Bryant, S. H. (1999). Identification of homologous core structures. *Proteins: Struct. Funct. Genet.* **35**, 70-79.
59. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
60. Holm, L. & Sander, C. (1996). DALI/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231-234.
61. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **00**, 000-.
62. Heringa, J., Sommerfeldt, H., Higgins, D. & Argos, P. (1992). OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *Comput. Appl. Biosci.* **8**, 599-600.
63. Heringa, J. (1999). Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem.* **23**, 341-364.
64. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.
65. Islam, S. A., Luo, J. & Sternberg, M. J. (1995). Identification and analysis of domains in proteins. *Protein Eng.* **8**, 513-525.
66. Heringa, J. (1994). The evolution and recognition of protein sequence repeats. *Comput. Chem.* **18**, 233-243.

Edited by J. Thornton

(Received 2 July 2001; received in revised form 9 November 2001; accepted 18 December 2001)