Protein Domain Identification and Improved Sequence Similarity Searching Using PSI-BLAST

Richard A. George and Jaap Heringa*
Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, United Kingdom

ABSTRACT **Protein sequences containing more than one structural domain are problematic when used in homology searches where they can either stop an iterative database search prematurely or cause an explosion of a search to common domains. We describe a method, DOMAINATION, that infers domains and their boundaries in a query sequence from local gapped alignments generated using PSI-BLAST. Through a new technique to recognize domain insertions and permutations, DOMAINATION submits delineated domains as successive database queries in further iterative steps. Assessed over a set of 452 multidomain proteins, the method predicts structural domain boundaries with an overall accuracy of 50% and improves finding distant homologies by 14% compared with PSI-BLAST. DOMAINATION is available as a web based tool at http://mathbio.nimr.mrc.ac.uk, and the source code is available from the authors upon request. Proteins 2002;48:672–681.** © 2002 Wiley-Liss, Inc.

Key words: domains; modules; PSI-BLAST; sequence; genome

INTRODUCTION

Many protein families have diverged from common ancestors by evolving different combinations and associations of domains.[1–3] Domains are characterized as semi-independent three-dimensional (3D) units in proteins, often with a particular function, observed to be genetically mobile and frequently moving within and between biological systems through mechanisms of gene or exon shuffling. An understanding of the domain organization of a protein sequence is crucial for structural and functional genomics initiatives, particularly those involving structural determination of large proteins using NMR techniques due to inherent size constraints.[4,5] Other areas in protein science aided by such knowledge include comparative sequence analysis,[6] fold-recognition, and threading techniques, protein engineering, site directed mutagenesis experiments[7] and the optimization of structural prediction methods.[8]

The correct fragmentation of a protein into its putative domains is especially important in the comparative analysis of entire genome sequences. Consideration of domain architecture will shed light on the evolution, structure and function of a protein family. For example, the "Rosetta Stone" genome analysis method[9] exploits the fact that many proteins consist of multiple domains in one organism but are present as separate proteins in another organism, which strongly suggests that the corresponding separate proteins interact within the second organism. It is clear that such analysis requires accurate sequence comparison tools at the domain level.

Domain annotation of a protein sequence in the absence of structural information has proved to be a difficult problem. Early approaches explored first principles such as assembling secondary structure elements into domains,[10] predicting domains as areas with high residue contact density,[11] or as areas with a high proportion of long-range residue-residue interactions.[12] However, these early approaches were unsuccessful in providing reliable domain boundary predictions.[13] A more recent method by Wheelan et al.[14] is based on the fact that domains have a distinct size distribution, averaging at 100 residues. Accurate predictions are limited to two-domain proteins with $<$300 residues only. George and Heringa[15] recently improved the delineation of protein domain boundaries to 52% using a consistency-based protocol over sets of protein *ab initio* 3D model structures generated using distance geometry.

Currently, most annotated domain databases are based on inferring domains by searching sequence databases,[16–19] mainly based on fast alternatives of the Smith and Waterman[20] local alignment technique such as FASTA[21] and BLAST.[22] Recent improvements in database search methods have come from using a multiple sequence alignment (MSA) of a protein family to find additional family members.[23–25] MSAs contain family specific information, including structurally or functionally important positions that are not identifiable within a single sequence. To utilize information held within an MSA, a Position Specific Scoring Matrix (PSSM) is constructed,[26,27] and aligned to all sequences in a database to find new family members. Profile based methods generally improve the sensitivity of database searches compared with single sequence methods.

The most widely used database search method is Position Specific Iterative BLAST (PSI-BLAST).[25] This iterative search method creates a PSSM from stacking local-alignments of sequences found in an initial database

Grant sponsor: Medical Research Council.

*Correspondence to: Jaap Heringa, Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, NW7 1AA, UK. E-mail: jhering@nimr.mrc.ac.uk

Received 22 January 2002; Accepted 12 April 2002

© 2002 WILEY-LISS, INC.

search using gapped-BLAST.[25] The PSSM is then used to further search the database for new homologues, which are in turn added to the PSSM for additional searches. Iteration will stop when new sequences are no longer found or when the program reaches a fixed number of iterations.

Iterative sequence search methods can be a powerful way to find distant homologies but often fail when querying a multidomain protein or a protein with regions of compositional bias. For example, common conserved protein domains such as the tyrosine kinase domain can obscure weak but relevant matches to other domain types,[29] whereas sequences containing low-complexity regions, such as coiled coils and transmembrane regions, can cause an explosion of the search rather than convergence because of the absence of any strong sequence signals. Conversely, some searches may lead to premature convergence; this occurs when the PSSM is too strict only allowing matches to very similar proteins, i.e., sequences with the same domain organization as the query are detected but no homologues with different domain combinations.

An additional problem with iterative searches is "matrix migration" (also referred to as "profile wander"), which occurs when the search strategy is too permissive such that information from false-positive sequences is included into the profile, resulting in the possible loss of truly homologous sequences found in earlier rounds. A further loss of information can be incurred with PSI-BLAST, because PSI-BLAST PSSMs are trimmed to only use the highest scoring region in a search, ignoring less conserved regions. The alternative database search method QUEST[24] alleviates these problems by using an independent multiple alignment program to generate a true MSA between iterations, and not a "master-slave" alignment, thereby improving the quality of the PSSM. The QUEST method also removes any sequences that are deemed too divergent as a reliable family member, so not to "pollute" the PSSM, which leads to increased search capabilities.

A few methods exist to predict domain boundaries through postprocessing BLAST searches. The method BALLAST can be used to visualize conservation profiles for a query sequence based on sequence searching,[30] albeit the method does not delineate domain boundaries. Another technique is the method PASS (Prediction of Autonomous Folding Units based on Sequence Similarities), which uses a simple and noniterative method of domain delineation based on the stacking of sequences from a gapped-BLAST search onto the query sequence.[31] Regions along a query sequence often have a varying number of matching sequences from the BLAST data, leading to abrupt increases and decreases in sequence numbers along the query. The PASS method is based on a single BLAST run and does not use iteration to include information from distant homologues. Further, the current release of the PRODOM domain database[17] is created using the method MKDOM2,[32] which performs PSI-BLAST searches starting with the smallest sequence in the database as a query, supposed to represent a single domain. All domain sequences identified are removed from the database, after which the process is iterated with the remaining subsequences and terminated when the database becomes empty. The MKDOM2 method is an iterative protocol but does not address the aforementioned problems connected to PSSM-based iterative searches.

Here, we introduce a method, called DOMAINATION, that assigns domain boundaries by applying PSI-BLAST in a repetitive fashion. The distribution of the aligned positions of N- and C-termini from PSI-BLAST local sequence alignments is used to identify potential domain boundaries. DOMAINATION incorporates a new iterative strategy for chopping and joining domains and domain segments in an attempt to track a protein's "evolutionary pathway" from its loss and gain of domains. This allows the recognition of both continuous and discontinuous domains. For each domain inferred from the corresponding PSI-BLAST local alignments, profiles are created by filtering redundant sequences and subsequent MSAs. Each thus filtered profile is then used in further iterative database searches using PSI-BLAST. All profiles are required to contain the original query sequence at each iteration of PSI-BLAST to avoid profile wander, but parameters are set to ensure the profiles are divergent enough to capture distant sequence fragments. The whole process of iterative PSI-BLAST searches is repeated until domain assignment ends and no new homologues are found anymore. In the remainder of the article we describe the DOMAINATION method and also assess the accuracy of our protocol to assign domains by a direct comparison using known structural domain boundaries and benchmark the generally increased search performance relative to stand-alone PSI-BLAST searches.

METHOD

DOMAINATION is written in ANSI C and Perl5 and run in parallel on a 128-processor Linux cluster. Figure 1 presents an outline of the method. DOMAINATION incorporates several procedures beginning with a sequence database search using PSI-BLAST (see section 'Database search protocol'). We designed a straightforward method to cut the query sequence into domains (see section "Domain Cutting" below) while keeping track of domain deletions, permutations, and segments of discontinuous domains (see separate sections below). For each putative domain, an MSA is generated using the PSI-BLAST local alignments (see section "MSA construction" below), which are then used in further database searches. The full process is repeated until domain cutting finishes or when no more sequences are found by PSI-BLAST. Several methods of benchmarking are used to address database search performance and domain boundary prediction (see section '"Protein Test Sets and Benchmarking").

Database Search Protocol

DOMAINATION starts with an initial run of PSI-BLAST to find significant sequences in the non-redundant protein sequence database (NRDB, ftp://ncbi.nlm.nih.gov/

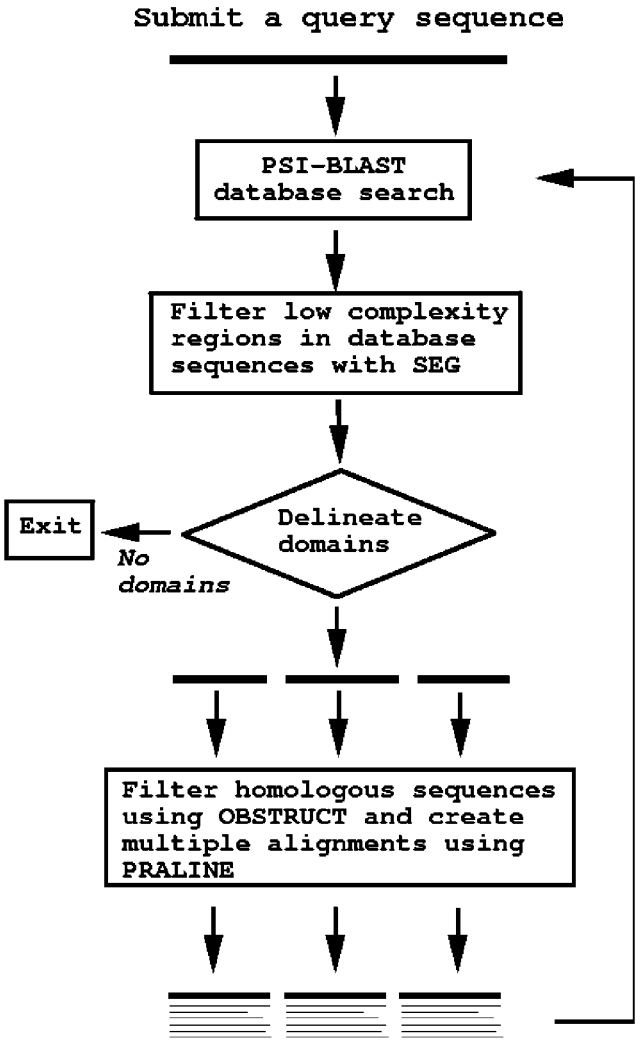

Fig. 1. Flow diagram of DOMAINATION. The method begins with an initial run of PSI-BLAST[25] to search the NRDB with a single query sequence. All significant gapped local alignments are collected from the PSI-BLAST output and filtered for sequences of low-complexity using SEG.[33] A new domain cutting protocol is then applied to the query sequence, based on the distribution of N- and C-termini of the local alignments. MSAs are generated in parallel for each domain and a nonredundant set of corresponding local alignments. Selection of sequences for the MSA is achieved using OBSTRUCT[36] to find the largest subset of sequences within a range of 20% to 60% sequence identity. Each domain sequence set must contain the original query sequence to prevent profile wander. MSAs are constructed for each domain sequence set using PRALINE.[37] The MSA are submitted simultaneously in further PSI-BLAST searches of the NRDB. The whole process iterates until no new sequences are detected or when domain cutting ends.

blast/db/). We used PSI-BLAST version 2.0.10 to identify homologues to a query sequence and used the parameter settings described by Park et al.[28]: an E-value cut-off of 0.0005 (-h0.0005 option) for selecting homologues for the PSSM, a threshold of 0.001 to reduce false positives (-e0.001 option) and an upper limit of four iterations (-j4 option). Four iterations are typically used as a maximum, as this is often sufficient for sensitive homology searching while further iterations are more likely to lead to profile wander.[28]

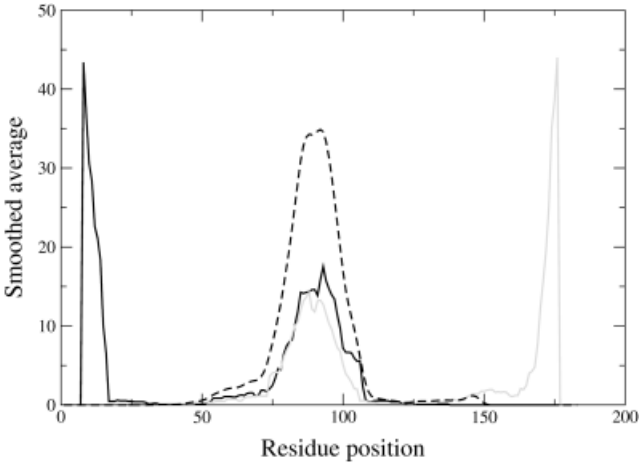

Fig. 2. Method of domain cutting. The distribution of N- and C-termini from local alignments generated from PSI-BLAST for ribonucleoprotein A1 (PDB code 2up1). The black line represents the distribution of N-termini and the grey line represents the distribution of C-termini. The dashed line represents the sum of the two distributions.

Filtering Low Complexity Regions

Low complexity regions in the query sequence are filtered using SEG[33] (default in PSI-BLAST). To complement this filtering, the sequences found in the database searches were postfiltered for compositional bias. This was done because filtering low-complexity regions in the query sequence as done by PSI-BLAST is not always successful at preventing matches to database sequences with low-complexity regions, and particularly sequences of medium complexity can lead to an explosion of false positives. Because many sequences found by PSI-BLAST can comprise low-complexity regions, postprocessing the sequences found by PSI-BLAST appears to be important for reducing false positives. Therefore, SEG is used to identify and delete regions of low complexity within sequences found after four PSI-BLAST iterations. Any local fragment containing >15% low-complexity regions are removed. A record is kept of all the sequences found in the searches.

Domain Cutting

We avoid the need for all-against-all sequence comparison and clustering by employing a simple method of domain delineation using the distribution of both the N- and C-termini for local alignments generated from a PSI-BLAST search (Fig. 2). Occurrences of N- and C-termini for each local-alignment are counted along the length of the query. In cases where the termini positions represent the real start and end positions in a database sequence (<10 residues from its termini) we are confident that they represent true domain boundaries, and they are subsequently scored twice. Two smoothing windows are then run across the summed termini, one for the N-terminal distribution and the other for the C-terminal distribution, using a window size of 15 residues. The two distributions are then combined using a biased protocol, which assigns higher weights to regions that have an abundance of both N- and C-termini, such as regions

designating the end of one domain and the start of a second domain:

if $N_i \times C_i = 0$ (either $N_i = 0$ or $C_i = 0$)

then $S_i = N_i + C_i$ (sum boundary positions)

else $S_i = N_i + C_i + (N_i \times C_i)/(N_i + C_i)$ (apply bias)

where N_i and C_i are the sum of N-termini and C-termini, respectively, at residue position i in the query sequence, and S_i is the overall sum of the termini. A window of length seven is used to smooth the sum, S, of the two distributions. The final curve is normalized by calculating self Z-scores for all positions in the summed graph, equivalent to normalizing the data to zero mean and unit standard deviation. Any peaks above a Z-score of two are then taken to represent potential domain boundaries. Because domains are rarely observed to be below 30 residues,[34] any regions between two boundaries of <30 residues are split equally between two domains or are removed if within 35 residues of the N- or C-terminal ends of the query sequence. A value of 35 was used here to avoid incorrect cutting due to the inherently large number of local alignment termini close to either end of the query sequence.

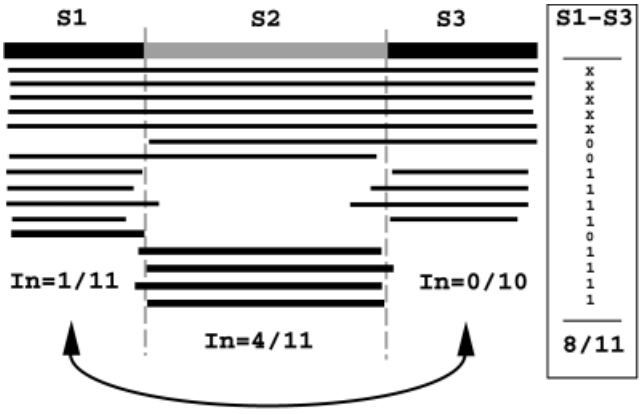

Fig. 3. Detecting discontinuous domain boundaries. An example is given of the calculation of the Independence and Association scores, given a three-segment query sequence, where segment 1 (S1) and 3 (S3) are associated with a discontinuous structural domain. "In" denotes the independence score, compiled for each query sequence segment using matching database sequence fragments. Independent local alignments are depicted in bold. The calculation of the association score between segment 1 and 3 (S1–S3), each having a required Independence score <10%, is given at the right hand side in the figure: Database sequences (designated by "x") that overlap each of the three segments with >70% are not taken into account. Since the association score is 8/11 = 73% is >50%, the two segments are considered to be parts of a discontinuous domain.

Domain Deletion

Regions that correspond to large deletions within a query sequence are likely to cause errors in domain cutting. A large deletion in the query sequence of >35 residues is likely to be the loss of a domain during the protein's evolution or the gain of a domain in a homologous protein. However, whether the missing domain would have been positioned between two domains or inserted within a single domain is not clear. Deletions will be populated by a large number of local alignment termini, which would signify a domain boundary. Therefore, we attempt to identify such sites of deletion and remove the corresponding N- and C-termini, after which smoothing and boundary assignment is repeated. A boundary is deemed a deletion site when the two adjacent segments in the query sequence, a segment being the region between two boundaries, share hits to the same sequences, whereas the local alignments show >35 intervening residues between the two segments.

Domain Shuffling

In the same way, we used local alignments to identify domains that have a different sequential order within a database sequence. A domain shuffling event is declared when two local alignments (>35 residues in length) within a single database sequence match two separate segments in the query (>70% overlap), where the sequential start and end points of alignments are reversed between the query and the database sequence.

Circular Permutation

A distinction is made here between domain shuffling and circular permutations.[35] A circular permutation is a small sequence order reversal (≤35 residues) in adjacent segments, possible corresponding to units of secondary structure. In contrast, domain shuffling must correspond to large sequence order reversals (<35 residues) that do not have to be adjacent in sequence. Because circular permutations are supposed to take place within single domains, the N- and C-termini positions of any segment involved in a circular permutation do not warrant domain delineation and therefore were removed from the distribution using the procedure applied in cases of deletion (see above).

Assigning Continuous and Discontinuous Domains

We judged if a delineated segment could exist as an isolated unit or rather be part of a discontinuous domain by calculating an "independence" score for each segment (Fig. 3). This score is based on the proportion of sequences that align with the segment and not with any other segment in the query sequence, where it is insisted that a matching sequence overlaps the segment by at least 70%. A segment is considered independent if >10% of its matching sequences do not align with any other segments. Such a region is thought to be able to exist independently.

Any segment with an independence score <10%, so that it is regarded as dependent, is then probed as a discontinuous domain segment and, if found dependent, is joined to other associated segments. To do this, an association score is assigned to all pairs of dependent segments, based on how the PSI-BLAST segments within the database sequences are matched with the query sequence: For each database sequence with homology to the query sequence, if the segments within the database sequence found by PSI-BLAST match two nonadjacent query sequence segments (>70% overlap to each) and do not match the

segment(s) in between (≤30% overlap), then this database sequence is considered to support the association; i.e., the case that the two query segments constitute a single discontinuous domain. Also a database sequence with segments matching ≥30% of either of the latter segments but <70% of the intermediate segment(s) is supposed to support the association (Fig. 3). If over all matched database sequences the majority (>50%) supports association between the two segments, where database sequences matching the considered query fragments and the intervening stretch are discarded, then the query sequence segments are declared to be associated (Fig. 3). The associated segments are joined, and the resulting reconstructed discontinuous domain sequence is then used to query the database in subsequent DOMAINATION iterations. With this approach, segments constituting discontinuous domains can be identified during iteration.

Constructing Multiple Sequence Alignments for Domain Sequence Sets

The NRDB contains non-identical sequences (<100% sequence identity), and therefore is not biologically nonredundant. To achieve maximum information content from the PSI-BLAST results, each set of sequence fragments found to pertain to a particular domain by DOMAINATION is filtered for redundancy. A non-redundant set of sequences (>20% and <60% sequence identity) is generated for all the sequence fragments matching a particular domain by the program OBSTRUCT,[36] which produces the largest possible subset of protein sequences with all pairwise sequence identity scores within a particular range.

For each thus filtered domain sequence set a MSA is created using the alignment method PRALINE[37,38] with default settings. The final MSA is converted into a format readable by PSI-BLAST. This requires that sequences extending the boundaries of the domain are trimmed and no gaps can be introduced into the query sequence so that residues in the database sequences are removed opposite positions where gaps in the query would otherwise appear. The MSA is then used to search the NRDB with PSI-BLAST options -B and -j4.

Benchmarking Domain Boundary Prediction by DOMAINATION

A set 452 multidomain protein structures ranging from two to five domains, with no more than two linkers between two domains, was used to test the performance of our method in correctly defining structural interdomain boundaries. The protein test set was derived from a set of nonhomologous proteins with known 3D structure downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html). The proteins in this set have been selected using single linkage clustering using a BLAST P-value $< 10^{-7}$, and representatives for each cluster were subsequently selected based on completeness and resolution of the structures.[39] We delineated the structural domains within each tertiary protein structure using the method described by Taylor[40] and used the resulting domain boundary definitions as a reference to evaluate the accuracy of DOMAINATION. A window of ±20 residues around each assigned reference boundary position was used, consistent with previous evaluation studies of boundary prediction methods.[14,31]

Benchmarking Sequence Searching by DOMAINATION

To evaluate the added value of DOMAINATION over stand-alone PSI-BLAST, we use the results from a PSI-BLAST database search with the sequence fragments associated with the known individual structural domains as a standard of truth, based on the assumption that searching with exactly delineated domain sequences should be optimal in finding related proteins in the database. All sequence segments corresponding to a structural domain within the non-redundant 452-protein test set were used as separate queries in PSI-BLAST searches against the NRDB (options -j4, -e0.001, and -h0.0005). Segments of discontinuous domains within the test set were joined to create a full domain sequence before searching. As before, PSI-BLAST local alignments were postfiltered for low complexity (see "Database Search Protocol") and any local fragment containing > 15% low-complexity regions was removed.

We then tested to what extent PSI-BLAST and DOMAINATION, when run on the full-length protein sequences, can capture the sequences found by the reference PSI-BLAST searches using the individual domains. Three search procedures were therefore compared: PSI-BLAST with structural domains (reference), PSI-BLAST searches with full-length sequences and DOMAINATION searches with full-length sequences.

Two separate reference sets were compiled using PSI-BLAST on the structural domains: Reference set 1 consists of database sequences for which PSI-BLAST finds all domains contained in the corresponding full-length query sequences, while reference set 2 consists of sequences found by searching with one or more of the constituent domain segments for each query sequence. Reference set 2 therefore contains many more sequences than reference set 1, as for the former only a single rather than all constituent domains of a particular query sequence need to be found in a database sequence to count that sequence. The two resulting reference sets were used to evaluate the number of sequences found by DOMAINATION and PSI-BLAST using the 452 full-length protein sequences.

Benchmarking Sequence Searching by DOMAINATION Using SSEARCH

The significance scores attached to each sequence found by PSI-BLAST do not relate to the original query sequence but rather to the PSSM created in iterative steps. This means that the aforementioned problem of profile wander (see Introduction) in PSI-BLAST searches might lead to the propagation of false positives/negatives. To reduce the number of putative false positives, we verified the statistical significance of database sequences found by relating them to the original query sequence. We used SSEARCH,[21]

an implementation of the Smith and Waterman algorithm, which calculates an E-value for each generated local alignment. We used an E-value cutoff of 0.1 to produce filtered sets of sequences found by DOMAINATION and PSI-BLAST runs using complete query sequences. The thus obtained statistical significance scores will increase confidence in the comparison of DOMAINATION with a normal running of PSI-BLAST. As before the number of significant hits found by each of the methods was compared.

RESULTS

Boundary Prediction Accuracy

The success of DOMAINATION to dissect a protein into its putative domains was measured by comparing the cutting positions to known structural domain boundaries. Of the 452 multidomain proteins in the NCBI set, 56% (254 proteins) were predicted to have more than one domain. In its first iteration, DOMAINATION made 335 boundary predictions of which 42.0% are within ±20 residues of a true boundary (see Methods). Overall we find nearly a quarter (23.3%) of all linkers in the 452-protein set. This is not a surprise because the boundaries of structural domains will not always coincide with the boundaries of a sequence domain/module. Moreover, domain boundaries cannot be recognized by sequence-based methods in the absence of a discernible signal provided by some domain shuffling among various related proteins. In such a case, only tertiary structure information can point to the actual domain boundaries.

Nonetheless, the proportion of correct predictions per protein is 49.9 ± 44.6%, reached after two iterations of DOMAINATION. This is much higher than that randomly expected per protein (19.5 ± 9.5) based on the percentage of residues constituting the linker regions of ±20 residues around the true boundaries. Prediction accuracy is not affected by using different PSI-BLAST significance thresholds, but the number of proteins with predictions increases with higher thresholds (Fig. 4).

Wheelan et al.,[14] reported prediction results based on the top two predictions for 246 two-domain proteins only, which were accurate to a resolution of ±20 residues in 57% of cases. However, their method fails when analyzing proteins with three or more domains.[14] The results of Kuroda et al.,[31] on a small test set of 52 multidomain proteins with largely solvent exposed interdomain boundaries, achieved a prediction accuracy of 52.5% within a window of ±20, but with a small coverage of only 14.4% of all linkers detected.

Joining Discontinuous Segments

Out of the 452-protein test set, 104 (23%) structures comprised one or more discontinuous domains. Of the segments corresponding to a discontinuous domain, 30% were successfully joined by DOMAINATION. Over the 104 proteins with at least one discontinuous domain, the boundary prediction accuracy was 34.2% within a resolution of ±20 residues, which is not much lower than the 49.9% reached over all proteins, given the inherent complexity of recognizing discontinuous domain fragments. This is a good result given that no structural information is used in the predictions.

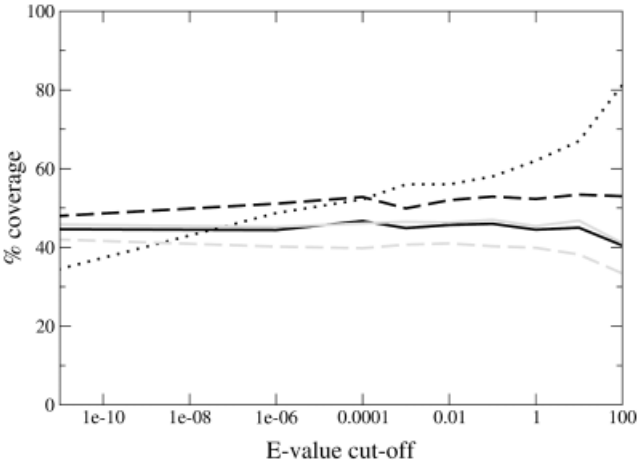

Fig. 4. Change in domain prediction accuracy when using various PSI-BLAST E-value cutoffs. PSI-BLAST is run with four iterations on each of the 452 multidomain proteins. The dotted line represents the percentage of proteins that have a boundary prediction; this is seen to increase when more sequences are found. The continuous grey line is the percentage of structural domain boundaries found per protein after the first iteration of DOMAINATION. The continuous black line is the percentage of boundary predictions made that are correct for each protein in the first iteration. The dashed lines correspond to the solid lines above, but for a second iteration of DOMAINATION. The second iteration sees more domain boundaries being found, but with a lower success rate, i.e., more false predictions. All successful predictions are within a resolution of ±20 residues from a true boundary position.

Previously we tested the ability of our method to predict discontinuous domains on a set of "pseudo-proteins" with artificial domain organizations. These multidomain proteins were constructed using a set of 15 domain families blindly selected from the Pfam (version 5) sequence-domain database[16] (data not shown). Two types of multidomain protein sequences were generated, fused, and inserted. The former simply means the concatenation of two or more domains, whereas the latter denotes the insertion of one domain into the middle of another. The majority of pseudo-proteins (67%) with a domain insertion had their discontinuous segments successfully joined by DOMAINATION and all boundaries found to within an average of ±5.7 residues. The main difference between the pseudo-proteins, and the set used here is that the boundaries in the pseudo-proteins are defined at a sequence level and can therefore more easily be identified. These data show that our method has a high potential to find discontinuous domains, but with a prerequisite that proteins with alternative intermediate domains, or those without insertions, exist in the database to provide a signal for discontinuous domain boundary detection.

Database Search Performance

Table I represents the number of database sequences found in common between the reference PSI-BLAST searches using the individual domains and those found

TABLE I. The Number of Homologous Sequences Found and Missed by the Two Methods, PSI-BLAST and DOMAINATION, for the Set of 452 Proteins with Known 3D Structure

	PSI-BLAST vs Reference set 1	DOMAINATION vs Reference set 1	PSI-BLAST vs Reference set 2	DOMAINATION vs Reference set 2
Seq's found	28,581	28,921	67,300	73,274
Seq's missed	618	278	13,542	7,568
% Missed	2.12	0.95	16.8	9.36

Reference set 1: All sequences found using searches of the full-length protein sequence that are also found by searches using PSI-BLAST with the individual domains; all sequences found contain all domains present in the full length query protein. Reference set 2: All sequences that are found that are in common with at least one of the individual PSI-BLAST domain searches; sequences found contain at least one of the domains in the full length query.

using PSI-BLAST and DOMAINATION searches using the full-length proteins (see Methods). Reference set 1 consists of 29,199 database sequences that contain all the domains as found in the original query sequences, whereas reference set 2 consists of 80,842 database sequences that have at least one of the domains within the original query sequences and is therefore a much larger set of sequences.

Table I shows that sequences found using PSI-BLAST searches with the full-length proteins include 97.9% of the sequences in reference set 1, whereas DOMAINATION finds 99.1% of these reference sequences. DOMAINATION therefore captures more than half (55%) of sequences that remain undetected by PSI-BLAST. The near 100% scores are expected because database sequences containing all the domains of the query sequences are likely to be similar and thus should relatively easily be found by both test methods.

When we increase the number of reference sequences by allowing those for which not all individual domains are detected within a sequence by the domain-PSI-BLAST runs (Reference set 2), the test becomes more difficult for the two methods (Table I). Only 83.2% of sequences found by PSI-BLAST, using the full-length proteins, are in common to those found by at least one of the individual domain searches. In contrast, DOMAINATION finds 90.6% of these sequences. Overall, DOMAINATION can detect 44% of the sequences that are missed by PSI-BLAST on the full-length query sequences.

Significance Testing

An uncertainty with the above testing scenario is that many false positives can potentially arise when performing a PSI-BLAST search for reasons mentioned earlier. In order to reduce the potential number of false positives, we used SSEARCH[21] to test the statistical significance of any found sequence with the query sequence over the NCBI test set. PSI-BLAST scores will not reflect significance to the original query sequence but to the profile(s) used during the iterative search. Therefore, significant hits found using SSEARCH are likely to reduce the number of false positives obtained in the PSI-BLAST searches and may give an improved idea of the number of true homologues found by the methods, although this filter will ignore many distant homologies that can only be identified by profile-based methods.

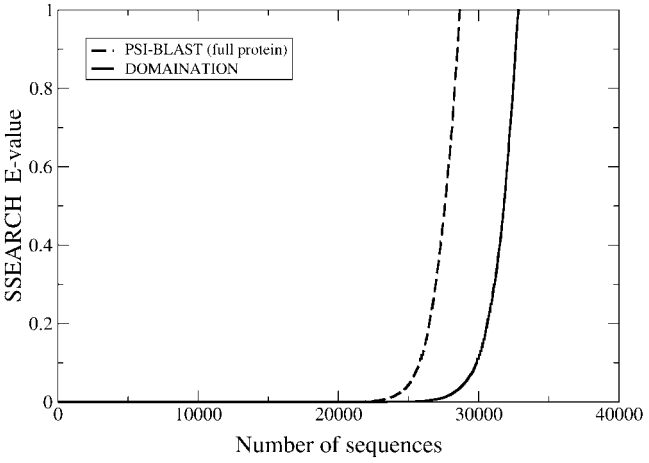

Fig. 5. Significant sequences found by PSI-BLAST. The number of significant homologues as denoted by SSEARCH comparisons of the original query sequences against the sequences found using the two search methods; DOMAINATION (continuous line) and PSI-BLAST (dashed line). Values represent the total number of sequences found from searches of proteins in the NCBI set.

Based on this scenario, searches using DOMAINATION find the majority of true homologues (Fig. 5). Using an E-value cutoff of 0.1 as a significance threshold the performance of PSI-BLAST searches using the full-length proteins is 14% below that of DOMAINATION. If we allow PSI-BLAST searches to keep iterating until convergence, we see an increase in the number of significant hits found. However, PSI-BLAST detection is still 10% below that of DOMAINATION.

The fact that DOMAINATION finds more significant hits, as assessed by SSEARCH E-values, than PSI-BLAST, suggests that searching using the complete proteins alone is not sufficient to capture all obvious homologies. PSI-BLAST seems to miss these homologous sequences as a result of profile wander and ignoring the domain content of a protein. Both methods have the same proportion (66%) of sequences found above an SSEARCH E-value of 1.0, which are removed as false positives. Although at lower E-value cutoffs more sequences are discarded, Figure 5 shows that the SSEARCH filter does not dramatically delete sequences at lower E-values down to a value of about 0.05.

TABLE II. The Number of Homologues Found and Missed by the Two Methods, PSI-BLAST and DOMAINATION, for the set of 15 Proteins with SMART Domain Annotation

	PSI-BLAST vs Reference set 1	DOMAINATION vs Reference set 1	PSI-BLAST vs Reference set 2	DOMAINATION vs Reference set 2
Seq's found	323	347	3672	5902
Seq's missed	24	0	3438	1202
% Missed	6.9	0	48.4	17.0

Reference set 1: All sequences found using searches of the full-length protein sequence that are also found by searches using PSI-BLAST with the individual domains; all sequences found contain all domains present in the full length query protein. Reference set 2: All sequences that are found that are in common with at least one of the individual PSI-BLAST domain searches; sequences found contain at least one of the domains in the full length query.

Comparing DOMAINATION and PSI-BLAST Using SMART Sequence Domains

A further comparison was made between PSI-BLAST and DOMAINATION using a set of proteins with domains assigned at the sequence level in the SMART domain database.[18] A set of fifteen multidomain proteins, ranging from two to six domains, were used (ERG_HUMAN, Q94222, Q9V9J5, A4_SAISC, AAF44820, AAG55540, AAG58344, AAH02392, MEPB_RAT, O57581, O97507, Q99PX0, Q9D398, SMZ7_BRARE, I1BC_RAT). As before, three methods were used: PSI-BLAST and DOMAINATION searches with the complete protein sequences are compared with the results of PSI-BLAST searches with the individual domains. Analogous to the data in Table I, two reference sets were created, this time by searching with the 15 multidomain query proteins. Table II shows that sequences found using PSI-BLAST searches of the full-length proteins constitute 93.1% of the sequences found in Reference 1, i.e., database sequences that consist of all the SMART domains within the query. DOMAINATION finds all (100%) of these common sequences. Again, the near 100% scores are expected because protein sequences containing all domains should easily be found by both methods. Only 51.6% of sequences found by PSI-BLAST searches using the full-length proteins were also found by at least one of the individual domain searches (Reference 2). In contrast, DOMAINATION finds 83.0% of these sequences and detects 65% of the sequences missed by PSI-BLAST searches using the full-length query sequences.

Computational Requirements

DOMAINATION takes under one hour to run a set of 452 multidomain proteins on a Linux cluster consisting of 128 nodes. The time to run an individual protein against the NRDB typically varies between 1 and 30 min. Using OBSTRUCT to filter the sequences that enter the MSA (see Methods) can not only improve the quality of the PSSM, as a more nonredundant set of sequences is used but also dramatically reduces the time taken to construct the MSAs because of the smaller number of sequences to align. Nonetheless, a few runs were seen to take over an hour. In these cases the MSA stage was found to be the limiting step when trying to align large sequence sets, so that constraining the number of sequences to be aligned will greatly reduce this time.

DISCUSSION

The concept of the domain is critical in comparative sequence analysis. Profile and iterative sequence search methods are likely to give poor results when querying a multidomain protein. It is essential that methods to delineate domains in a preprocessing step are employed. Our method takes advantage of the fact that domains are recurring evolutionary units, by collecting the N- and C-termini of local alignments to homologous sequences we can successfully isolate domains in sequence and significantly improve comparative sequence analysis by exploiting this information. In contrast, PSI-BLAST ignores the domain content of proteins and as a result misses significant domain homologies. It must be stressed that this effect also constrains the DOMAINATION searches, as DOMAINATION does not find all domain homologies because it fails to delineate over a third of the proteins in the test set, thereby effectively confining the database search to the PSI-BLAST functionality. It is particularly interesting that the results presented here demonstrate that PSI-BLAST searches using domains derived by DOMAINATION lead to enhanced sensitivity compared with searches based on structurally delineated domains. This might on the one hand just be a result of the motifs delineated and the actual proteins contained in the sequence databases or on the other hand indicate that the domains found by DOMAINATION are more associated with function than the isolated structural domains, in the sense of the Rosetta Stone approach mentioned earlier (see Introduction). The DOMAINATION protocol described here is relatively accurate and constitutes a good alternative to other domain search methods, particularly given its ability to dynamically delineate and reconstruct sequence segments associated with discontinuous domains.

Iterative Cutting and Joining of Segments

Many proteins predicted to have three or more domains often have joining events in the first iteration of DOMAINATION. This occurs when a central domain is observed to be "independent" from the other domains in a protein sequence (see Methods); such a domain will possibly have been inserted into the protein at some stage of the protein's evolution. DOMAINATION detects the insertion and joins the adjacent domains or discontinuous domain segments together. After a subsequent round of database searches the domain content of the joined segments can be

reassessed and may be delineated into domains that are used in further database queries. This iterative joining and splitting of domain segments increases the detection of homologous database sequences and thereby the chance to iteratively converge to the correct dissection of continuous and discontinuous domains. Our method differs from other domain prediction methods because it joins fragments that flank an identified domain. The iterative chopping and joining of domains by DOMAINATION not only adds to the increased performance of the database search but also allows a possible evolutionary history of the domain fusion and insertion events to be established. The method dynamically optimizes PSI-BLAST to domain searching but might also be used as a preprocessor to aid alternative database search engines such as SAM-T98,[41] HMMER2[42] or Quest.[24] The domains detected by DOMAINATION are more likely to correspond to evolutionary units and therefore turn out here to be more appropriate in iterative searches than the domains delineated using tertiary structures. The application of methods for testing comparative sequence analysis programs should therefore consider alternative benchmarking protocols in addition to just using structural domains, as we have done here by using the SMART database.

Sequence Repeat Detection

Detection of repeats in a sequence is an important preprocessing step in comparative sequence analysis; because repetitive sequences are reported to be problematic and can lead to premature convergence of PSI-BLAST.[43] Also, internal sequence repeats in a protein sequence will sometimes represent domains.[3] For example, the muscle protein titin comprises about 120 independent fibronectin-III-type and Ig-type modules connected in a linear arrangement.[44] Repeat detection techniques such as the REPRO method,[45,46] (http://mathbio.nimr.mrc.ac.uk) could be used before database searching. Any repeats found that are likely to represent a domain should be multiple aligned and used as a profile in subsequent database searches.

Delineating Domains Using Annotated Domain Databases

The most straightforward approach to infer domains is by searching the annotated domain databases.[16–19] Searching databases of domain family profiles and submitting any putative segments found to DOMAINATION might lead to improvements. For example, the Pfam[16] and SMART[18] databases can be searched easily for domains followed by DOMAINATION searches of the NRDB in subsequent iterations. It must be stressed that, although well maintained, the domain databases are incomplete, and some domain families may have incorrect boundary assignments, such that homologous of a query sequence may not be found within these databases, which stresses the importance of detecting domains "on-the-fly." Particularly, the protocols used here to detect discontinuous domain segments and domain permutations might suggest novel domain families.

ACKNOWLEDGMENTS

We are grateful to the helpful comments made by Dr. Jens Kleinjung.

REFERENCES

1. Bork P. Shuffled domains in extracellular proteins. FEBS Lett 1991;286:47–54.
2. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem 1995;64:287–314.
3. Heringa J, Taylor WR. Three-dimensional domain duplication, swapping and stealing. Curr Opin Struct Biol 1997;7:416–421.
4. Pfuhl M, Pastore A. Tertiary structure of an immunoglobulin-like domain from the giant muscle protein titin: a new member of the I set. Structure 1995;3:391–401.
5. Castiglone Morelli MA, Stier G, Gibson T, Joseph C, Musco G, Pastore A, Trave G. The KH module has an alpha beta fold. FEBS Lett 1995;358:193–198.
6. Russell RB, Ponting CP. Protein fold irregularities that hinder sequence analysis. Curr Opin Struct Biol 1998;8:364–371.
7. Nielsen PK, Yamada Y. Identification of cell-binding sites on the Laminin alpha 5 N-terminal domain by site-directed mutagenesis. J Biol Chem 2001;276:10906–10912.
8. Bonneau R, Strauss CE, Baker D. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 2001;43:1–11.
9. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402:83–86.
10. Busetta B, Barrans Y. The prediction of protein domains. Biochim Biophys Acta 1984;790:117–124.
11. Kikuchi T, Némethy, G, Scheraga HA. Prediction of the location of structural domains in globular proteins. J. Protein Chem. 1988;7: 427–471.
12. Vonderviszt F, Simon I. A possible way for prediction of domain boundaries in globular proteins from amino acid sequence. Biochem Biophys Res Commun 1986;139:11–17.
13. Valdar W. M.Sc, University of Manchester, 1997.
14. Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics 2000;16:613–618.
15. George RA, Heringa J. SnapDRAGON—a method to delineate protein structural domains from sequence data. J Mol Biol 2002; 316:839–851.
16. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2000;28:263–266.
17. Corpet F, Servant F, Gouzy J, Kahn D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res 2000;28:267–269.
18. Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res 1999;27:229–232.
19. Gracy J, Argos P. DOMO: a new database of aligned protein domains. Trends Biochem Sci 1998;23:495–497.
20. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
21. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
23. Thompson JD, Higgins DG, Gibson TJ. Improved sensitivity of profile searches through the use of sequence weights and gab excision. Comp Appl Biol Sci 1994;10:19–29.
24. Taylor WR. Dynamic sequence databank searching with templates and multiple alignment. J Mol Biol 1998;280:375–406.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–3402.
26. Gribskov M, McLachlan M, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–5358.
27. Luthy R, Xenarios I, Bucher P. Improving the sensitivity of the sequence profile method. Protein Sci 1994;3:139–146.

28. Park J, Karplus K, Barrett C, Hughey R, Haussler D. Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;184:1201–1210.
29. Sonnhammer EL, Durbin R. A workbench for large-scale sequence homology analysis. Comput Appl Biosci 1994;10:301–307.
30. Plewniak F, Thompson JD, Poch O. Ballast: blast post-processing based on locally conserved segments. Bioinformatics 2000;16:750–759.
31. Kuroda Y, Tani K, Matsuo Y, Yokoyama S. Automated search of natively folded fragments for high-throughput structure determination in structural genomics. Prot Sci 2000;9:2313–2321.
32. Gouzy J, Corpet F, Kahn D. Whole genome protein domain analysis using a new method for domain clustering. Comput Chem 1999;23:333–340.
33. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18:269–285.
34. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci 1998;7:233–242.
35. Lindqvist Y, Schneider G. Circular permutations of natural protein sequences: structural evidence. Curr Opin Struct Biol 1997;7:422–427.
36. Heringa J, Sommerfeldt H, Higgins D, Argos P. OBSTRUCT: a program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. Comput Appl Biosci 1992;8:599–600.
37. Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Computers and Chemistry 1999;23:341–364.
38. Heringa J. Local weighting schemes for protein multiple sequence alignment. Comput Chem 2002;26:459–477.
39. Matsuo Y, Bryant SH. Identification of homologous core structures. Proteins 1999;35:70–79.
40. Taylor WR. Protein structural domain identification. Protein Eng 1999;12:203–216.
41. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
42. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755–763.
43. Bateman A, Birney E. Searching databases to find protein domain organization. In: Bork P, editor. Advances in protein chemistry. 2000;54:137–157.
44. Fraternali F, Pastore A. Modularity and homology: modelling of the type II module family from titin. J Mol Biol 1999;290:581–593.
45. Heringa J, Argos P. A method to recognise distant repeats in protein sequences. Proteins Struct Funct Genet 1993;17:391–411.
46. George RA, Heringa J. The REPRO server: finding protein internal sequence repeats through the Web. Trends Biochem Sci 2000;25:515–517.