MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Protein Structural Domain Parsing by Consensus Reasoning over Multiple Knowledge Sources and Methods

Casimir A. Kulikowski [a], Ilya Muchnik [a, b], HwaSeob J. Yun [a], Aynur A. Dayanik [a], Deyou Zhang [c], Yueyue Song [a], Gaetano T. Montelione [c]

[a] *Computer Science Department, Rutgers University, Piscataway, NJ, USA*
[b] *DIMACS, Rutgers University, Piscataway, NJ, USA*
[c] *Molecular Biology and Biochemistry Department and CABM, Rutgers University, Piscataway, NJ, USA*

Abstract

Domain parsing, or the detection of signals of protein structural domains from sequence data, is a complex and difficult problem. If carried out reliably it would be a powerful interpretive and predictive tool for genomic and proteomic studies. We report on a novel approach to domain parsing using consensus techniques based on Hidden Markov Models (HMMs) and BLAST searches built from a training set of 1471 continuous structural domains from the Dali Domain Dictionary (DDD). Validation on an independent test sample of family-matched structural domain sequences from the Scop database yields a consensus prediction performance rate of 75.5%, well above the 58% obtained by simple agreement of methods.

Keywords:

Consensus Reasoning; Protein Structural Domains; Domain Parsing; Signal Detection; Sequence Database Search

Protein Structural Domains and their Parsing

Proteins are generally composed of one or more autonomously folding units known as domains [14]. Multidomain proteins in eukaryotes are often encoded by genes containing multiple exons, whose combinatorial shuffling during evolution has produced novel proteins with different domain arrangements and different associated functions. This is thought to have helped in responding to environmental challenges because, through recombinatorial events, it has allowed genomes to add, subtract, or rearrange discrete functionalities within a protein [20].

The multiple domains in proteins make them harder to express in recombinant form, and complicates their functional determination by x-ray crystallography or nuclear magnetic resonance (NMR). Expression and structure determination for single isolated domains is easier. Since isolated domains are the discrete functional units of proteins, knowing structure-function information about individual domains in a multidomain protein is usually a prerequisite for drug discovery or development for the full-length protein. Computational approaches that reliably parse protein sequence data into their constituent structural domains would therefore be most helpful in supporting the analysis of multi-domain proteins, their function, relation to disease mechanisms, and potential drug targets.

Domain parsing, or the detection of the signal of a structural domain from its sequence data is a complex problem. Definitions of protein structural domains are not standardized, and different databases will frequently report different domain boundaries for the same protein. Experimentally, proteolysis provides a practical tool to break up a large protein into domains. However, computational methods are attractive for their potentially much greater efficiency and lower cost, if they can be shown to work reliably over a broad enough range of proteins. Structural domains composed of a single, sequence-continuous module are the simplest set of targets for an initial test of consensus parsing methods, as reported here.

Protein families classified by sequence similarity can be used to derive sequence homology-based domains, and several databases, such as ProDom [4], Pfam [23], SBASE [17], and SMART [22], include such information. Protein sequences from primary sequence databases (NCBI-GenBank, SwissProt, etc.) have been typically compared to each other by all-against-all pairwise alignments, and similar sequences clustered into families. For each family, a multiple sequence alignment (MSA) is generated to represent conservative patterns, or characteristic signatures. A consensus pattern collected from a MSA, or a Hidden Markov Model (HMM) [5] of underlying transition probability (frequency) profiles can be used to identify new members of a family. Domain boundaries can also be derived from a MSA and additional information such as domain mobility and tertiary structure. Considerable expertise and careful judgement are usually necessary to validate domain assignments. And, assigned domains often tend to be smaller than the corresponding observed structural domains.

The Dali Domain Dictionary (DDD) contains a set of structural domains extracted from an all-against-all alignment of protein structures in the PDB. The resulting domains have precisely defined domain boundaries. Protein structures are also classified in the Scop [18] and CATH [19] databases by abstracting their hierarchical groupings. Both are based on detailed expert analysis, with Scop emphasizing definition of protein evolutionary relationships, while CATH emphasizes their architectures, topology and homology. These structurally based protein domain classifications provide a good point of departure for studies on the predictive power of alternative computational methods for domain parsing. Scop has already helped in the recognition of protein folds in genome data, and we chose it for our comparative studies and investigation of consensus methods.

Domain Parsing by Multicriteria Consensus Reasoning from HMM Models and BLAST Matches

Our approach applies techniques of multicriteria consensus reasoning [16][24][6] in a novel way to combine results from Hidden Markov Models [5][12][11] and BLAST [1] as tools for computational domain assignment and parsing from sequence data. To train our algorithm, we used a subset of continuous (single) domains from the DDD dataset that are also completely coincident with the corresponding domains in the Protein Data Bank (PDB). These amounted to 1,471. Expanding this dataset for each domain by addition of sequence homologs or neighbors from the nonredundant database [10], we then built a multiple sequence alignment and construct an HMM model for each domain. We selected the 10 best scoring results from BLAST search against the NR database to construct the neighborhood around the seed domain, and built our HMMs based on the multiple alignment of the DDD domain sequence and these homologous sequences with CLUSTALW [25] with the SAM package [11]. The HMMs can then predict domain boundaries in other protein sequences by aligning the protein sequence to the HMMs.

Our initial filtering criterion prior to consensus reasoning was that the two best (lowest negative log-likelihood ratios or scores) HMMs should detect some significant signal of the structural domain. For this, we align the sequence with the families of the first best and second best models, obtaining candidate domain structures for the given protein sequence with respect to the models used. Figure 1 illustrates the complexity of the comparisons needed in our study. A library of detection target sequences (for BLAST matches) and models (for HMM matches) is used to generate a set of matching scores for each against a query sequence. The detected sequence fragments derived from the first best and second best matched models against a protein sequence are shown below the reference sequences, indicating that in general, they do not need to coincide in their N or C terminal positions. However, if they detect the same underlying signal, they will typically overlap, as shown in the figure.

For HMMs, identification of a fragment is determined from the alignment of the sequence to each model by extracting the sequence starting from the first match state of the HMM to its last match state.

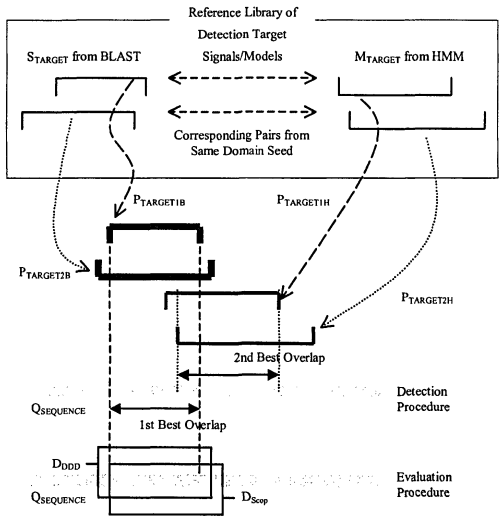

Figure 1 – Domain Detection & Prediction Evaluation

In order to assess or evaluate the effectiveness of the detected fragments we then need to compare them against the original reference sequence from the DDD database, and, for independent testing, against the corresponding Scop sequence, which may have somewhat different end-points, as indicated in the lowest line of the diagram. The above illustrates the combinatorial complexity of alignment matching between results filtered from just the two best scoring models from each method. However, there is no reason to expect that the rank order of scores produced by target model HMM and BLAST matches against query sequences needs to coincide so precisely given the many different factors that each method incorporates into their scoring function. While a very strong signal may indeed be detected among the two best ranked target models, in general one needs to extend the range of possible top ranking targets based on a comparative analysis of score values and alignment/overlap results on a well validated training set. When we did this we arrived at a conservative cutoff for preliminary screening by score values that includes the top ten candidate scores.

Table 1 – Pareto Set Extraction for a Sample Query

RutgersID BLAST:HMM	ΔN, ΔC, Lo: B ∩ H / B ∩ H ∩ S	Predicted BLAST Target ID (score)	Predicted HMM Target ID (score)
0870:0870	116, 0*, 352/352	3.1.4.1.4 (3e-71)	3.1.4.1.4 (-205.11)
1441:1441	9, 199, 49/0	4.94.1.1.1 (4e-10)	4.94.1.1.1 (-51.23)
1442:1442	9*, 101, 76/0	4.94.1.1.5 (2e-04)	4.94.1.1.5 (-35.57)
0864:0865	231, 32, 243/243	3.1.4.1.2 (3.4)	3.1.4.1.5 (-8.37)

The criteria for our Pareto Set extraction procedure must include matching scores for the two methods, and comparative alignment results. These latter involve the difference in N-terminal and C-terminal predictions, as well as the degree of alignment overlap for the predicted fragments for both methods. This is illustrated for a specific independent test query sequence extracted from the Scop database in Table 1, where ΔN and ΔC refer to the respective total number of residues that differ at the two terminals, and the overlap zone is given as the ratio of BLAST-HMM prediction overlap length to the reference domain length. Target IDs for BLAST and HMM predictions are also listed, together with their scores. This illustration shows how the top two candidates by ΔN and ΔC criteria of alignment include the best (lowest) scoring predictions by both BLAST and HMM, as listed on the first line of the table. This best prediction result also yields the maximum overlap with the reference domain (352 residues) when matched against the query, which has a length of 414 residues. Such a coincidence of best results from the multiple criteria of scoring and alignment is what we desire, but cannot be realistically expected for arbitrarily chosen query sequences. For this reason we needed to generalize the procedure to one of consensus reasoning over multiple sets of criteria for evaluating results from the multiple knowledge sources - DDD and Scop in this study.

The major steps of the general algorithm for single continuous structural domain detection by multicriteria consensus from the Pareto sets are summarized in the flow chart in Figure 2. The query sequence is scored against the 1471 target HMM models and BLASTed against the corresponding target domain reference sequences, and the top ten scoring pairs are extracted. These predictions are then assigned to their appropriate target Scop code, and aligned with the reference domain sequences from Scop corresponding to the seed domains from DDD. Parameters of the alignment are computed as described above, together with the corresponding scores, and the Pareto Set is extracted, from which the final prediction, or set of predictions are selected.

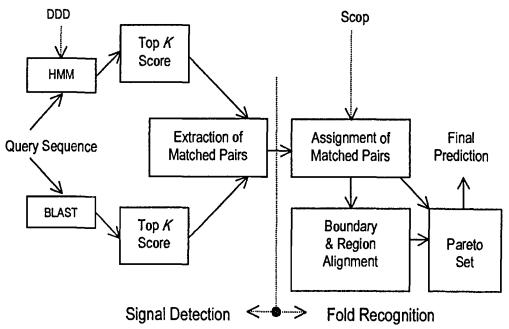

Figure 2 – Single Continuous Domain Detection Algorithm

We next describe results on the training and independent testing sets, comparing the performance of this algorithm with the simple conjunction of results from the best scoring models and searches.

Selection and Analysis of Training Set and Independent Testing Sets

The reliability of our models is evaluated by several techniques, and its domain assignment capabilities tested against the reference data from the PDB. The predictive capabilities of our HMM models have been tested against selected data from the Scop database, where the DDD domains comprise only a small fraction of the protein sequences, making it possible to divide into training samples (domains overlapping with DDD domains used in constructing the HMMs and BLAST queries), and independent testing samples for the parsing experiments.

We selected Scop target domains for realistic predictive capabilities from Scop-1.48 Astral with 40% or lower sequence identity. Of the 2683 such domains, 2490 are continuous. Requiring at least one DDD training domain and one non-DDD domain for independent testing within each Scop family reduces this to 235 families with 471 training and 611 testing domains. Further requirements of good alignment and matching between the training domains from DDD and the corresponding Scop domains reduces to 155 families with 252 training and 347 independent testing domains. For the training data, consistency of Scop and DDD domains was graded as follows: Perfect Match: No differing residues between predicted and actual test domain (150 cases); Almost Perfect Match: Maximum 5% difference at terminals & 5% overall (26 cases); N terminal within 5%; overall 10% maximum (22 cases); C terminal within 5%; overall 10% maximum (18 cases); Scop includes DDD (34 cases); DDD includes Scop (2 cases).

The framework for analysis, then, is as follows: a) Select Scop families which include families with the DDD seed structures, remove identical sequences from them. b) Divide the families into training and testing subsets. The former include proteins that contain the seed domain, and the latter do not. c) Compare against all 1471 HMMs, and carry out corresponding BLAST searches against the original seed sequence for all proteins in these families. Determine the true positive detections of the original seed domain in the independent sample of Scop proteins for each family within a superfamily. d) Compare the results from parsing with HMMs and BLAST. e) Apply the multicriteria consensus parsing algorithm and measure its improvement in parsing performance for the independent dataset in comparison with conjunctive consensus rules.

Results for simple matches on the 252 training cases show that detecting a DDD domain signal from a PDB sequence retrospectively can be as easily carried out with a BLAST search as with an HMM matching procedure. For all DDD continuous domains this is achieved at the 80% level for first best matches and 95% level for first and second best matches.

Results for matching the alignment (boundaries) of predic-

tions for the 347 independent test domains show that HMM target matching alone parses 92.2% of the 347 test domains whereas BLAST alone parses 84.7%, with agreement on 83%. However, boundary detection is very different for the HMM vs. BLAST results. Of the 347 independent test Scop domains, there is perfect alignment with the reference domain for 107 out of the 252 by HMM vs. only 3 by BLAST, or a 35:1 ratio in favor of accurate detection by HMMs. If we include the almost perfect category, this yields 200 by HMM vs. 57 by BLAST or a 3.5 to 1 ratio in favor of better HMM boundary detection.

Requiring accurate alignment with a correct assignment to the appropriate Scop domain family produces more conservative results. HMM predictions alone match 68%, BLAST alone match 65%, with agreement on a subset of 58%. This suggests that more powerful consensus reasoning methods may be needed.

Consensus Algorithm Results

When we applied our multicriterion consensus reasoning algorithm, we expanded the initial set of predicted target domains to the top 10 best scores from both methods, and then extracted matching pairs according to their Scop codes. When both seed domains from BLAST and HMM predictions share the same Scop codes at the domain level, this is the most specific candidate pair for the Pareto set. We can repeat this procedure at the higher levels of the Scop classification hierarchy, checking whether the Scop codes for family, super family, or fold level agree. For the chosen pairs at each level, N and C terminal differences between the two different method predictions are compared in terms of their differences by number of residues. Those with minimal differences between the BLAST and HMM predictions are ranked above those with larger differences, and the Pareto set is built from them. The pair with the greatest overlap length between two predictions is then chosen as the final prediction if the Pareto set has more than one element.

By this consensus reasoning method, combined BLAST and HMM predictions yields 75.5% (262/347) accuracy at the level of Scop family – a distinct improvement over the 58% obtained for the single best hit results where BLAST and HMM best scoring results coincided.

Conclusions

From the above we see that HMMs are much more reliable, or exact, in boundary detection for domain parsing than BLAST searches, but that combining HMM and BLAST results for those parses while slightly relaxing exact boundary detection yields the most reliable overall results by our consensus reasoning algorithm. The test on the 347 independently selected Scop structural domains shows the potential for applying our consensus parsing methods to the detection of structural domain signals at the genomic level. We are currently carrying out an analysis of the yeast genome with our existing system to test applicability of the consensus reasoning methods for fold recognition at the genomic level. We are also updating the software we have used for this study, and will re-compute our HMMs using the new DBCLUSTAL [26] based on the current NRDB, use PSI-BLAST with iterations for comparative scoring, and apply the IMPALA [21] method based on profiles rather than direct sequence data for our next version of the system.

Acknowledgments

We thank our colleagues in the Rutgers Bioinformatics Collaboratory for their feedback on this work. This work was supported in part by the Rutgers Bioinformatics Initiative of the Strategic Resource Opportunity Allocation (SROA), and by a New Jersey Commission on Science and R & D Excellence Program Grant.

References

[1] Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, and Lipman D. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 1997; 25: 3899-3402.

[2] Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Nucleic Acids Research 1999; 27: 260-262.

[3] Conte LL, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. Scop: a Structural Classification of Proteins database. *Nucleic Acids Research* 2000; 28: 257-259.

[4] Corpet F, Servant F, Gouzy J, and Kahn D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research* 2000; 28: 267-269.

[5] Durbin R, Eddy S, Krogh A, and G Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.

[6] Erghott, and Matthias. Multicriteria Optimization, Springer-Verlag; 2000.

[7] Gouzy J, Corpet F, and Kahn D. Whole genome protein domain analysis using a new method for domain clustering. *Computers & Chemistry* 1999; 23: 333-340.

[8] Holm L, and Sander C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Research* 1998a; 26: 316-319.

[9] Holm L, and Sander C. Dictionary of recurrent domains in protein structures. *Proteins* 1998b; 33: 88-96.

[10]Holm L, and Sander C. Removing near-neighbor redundancy from large protein sequence collections. *Bioinformatics* 1998c; 14: 423-429.

[11]Hughey R, and Krogh A. SAM: Sequence alignment and modeling software system. Technical Report (UCSC-CRL-95-7) University of California, Santa Cruz, Computer Engineering, 1995. http://www.cse.ucsc.edu/research/compbio/sam.html.

[12]Hughey R, and Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS* 1996; 12(2): 95-107.

[13]Islam SA, Luo J, and Sternberg MJ. Identification and analysis of domains in proteins. *Protein Engineering* 1995; 8: 513-525.

[14]Kim P, and Baldwin R. Intermediates in the folding reactions of small proteins. *Annual Rev Biochem* 1990; 59: 631-660.

[15]Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, and Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999; 402: 83-86.

[16]Mollaghasemi M, and Pet-Edward J. Making Multi-Objective Decisions, IEEE Computer Society Press; 1997.

[17]Murvai J, Vlahovicek K, Barta E, Cataletto B, and Pongor S. The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Research* 2000; 28: 260-262.

[18]Murzin A, Brenner SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995; 247: 536-540.

[19]Orengo C, Mitchie A, Jones S, Jones D, Swindella M, Thornton J. The CATH classification scheme of protein domain structural families. *Protein Data Bank Quarterly Newsletter* 1996; 78: 8-9.

[20]Pathy L. Protein Evolution by Exon Shuffling. R.G. Landes Co., Austin, TX; 1995.

[21]Schäffer AA, Wolf YL, Ponting CP, Koonin EV, Aravind L, and Altschul SF. Impala: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 1999; 15(12): 1000-1011

[22]Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains, *Nucleic Acids Research* 2000; 28: 231-234.

[23]Sonnhammer EL, Eddy SR, and Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997; 28: 405-420.

[24]Stewart TJ, and van der Honert. Trends in Multicriteria Decision Making, Springer-Verlag; 1999.

[25]Thompson JD, Higgins DG, and Gibson TJ. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research* 1994.

[26]Thompson JD, Plewniak F, Thierry JC, and Poch O. DbClustal: Rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acid Research. 2000; 28(15): 2919-2926.

Address for correspondence

Casimir A. Kulikowski (kulikows@cs.rutgers.edu)

Department of Computer Science

Hill Center

Busch Campus

Rutgers University

New Brunswick, NJ 08903

USA