Prediction of Protein Inter-Domain Linkers Using Compositional Index and Simulated Annealing

Maad Shatnawi and Nazar Zaki
Faculty of Information Technology
United Arab Emirates University
P.O. Box 17551, Al Ain, UAE
{shatnawi, nzaki}@uaeu.ac.ae

ABSTRACT

Protein chains are typically large and consist of multiple domains which are difficult and computationally expensive to characterize using experimental methods. Therefore, accurate and reliable prediction of protein domain boundaries is often the initial step in both experimental and computational protein research. In this paper, we propose a straightforward yet effective method to predict inter-domain linker segments by using the amino acid compositional index from the amino acid sequence information. Each amino acid in the protein sequence is represented by a compositional index which is deduced from a combination of the difference in amino acid occurrences in domains and linker segments in training protein sequences and the amino acid composition information. Further, we employ simulated annealing to improve the prediction by finding the optimal set of threshold values that separate domains from inter-domain linkers. The performance of the proposed method is compared to the current approaches on two protein sequence datasets. Experimental results show superior performance by the proposed method when compared to the state-of-the-art methods for inter-domain linker prediction.

Categories and Subject Descriptors

I.5 [PATTERN RECOGNITION]: [Structural, bioinformatics]; I.2.8 [Problem Solving, Control Methods, and Search]: [Heuristic methods]

General Terms

Algorithms, Experimentation, Performance

Keywords

Domain boundary prediction; domain linkers; protein sequences; inter-domain linker segments; simulated annealing; amino acid composition; compositional index.

GECCO'13 Companion, July 6–10, 2013, Amsterdam, The Netherlands.
Copyright © 2013 ACM 978-1-4503-1963-8/13/07...$15.00.

1. INTRODUCTION

Proteins sequences are typically long and consist of multiple domains. Domains are the basic units of protein structure which can exist, evolve, and function independently. Proteins have two types of segments: non-linker segments, which contain domains and terminal residues and linker segments which connect functional domains. Inter-domain prediction within the protein sequence is crucial for accurate determination of structural domains. Downsizing of proteins without loss of their functions is one of the major targets of protein engineering [17, 19]. This has an important effect in reducing computational cost. Many domain prediction methods first detect domain linkers, and in turn predict the location of domain segments. The knowledge of domains is used to classify proteins, predict protein-protein interaction (PPI), and understand their structures, functions and evolution. Therefore, efficient computational methods for splitting proteins into structural domains are gaining practical importance in protein engineering research [9].

Several excellent methods for predicting inter-domain linkers segments have already been developed. Suyama and Ohara [13] developed DomCut as a simple method to predict linker segments among functional domains based on differences in amino acid index between domain and linker segments. The prediction is made by calculating the linker index from the SWISS-PROT database [1] of domain/linker segments. A sequence segment is considered a linker if it satisfies three conditions; connecting two adjacent domains, in the range from 10 to 100 residues and not containing membrane spanning segments. Despite the fact that the performance of the DomCut was reasonably acceptable, the information contained in the linker index (the frequency of amino acid residue in the linker or domain segment) is simply not sufficient to accurately predict linker segments because of the lack of biological knowledge input.

Scooby-Domain (SequenCe hydrOphOBicitY predicts DOMAINs) web application was developed by George et al. [7] and extended by Pang et al. [11] to identify foldable segments in a protein sequence. Scooby-Domain uses the distribution of observed lengths and hydrophobicities in domains with known 3D structure to predict novel domains and their boundaries in a protein sequence. It utilizes a multilevel smoothing window to determine the percentage of hydrophobic amino acids within a putative domain-sized segment in the protein sequence. Each smoothing window calculates the ratio of hydrophobic residues it encapsulates along a sequence, and places the value at its central position. Using the observed distribution of domain lengths and percentage hydrophobicities, the probability that the segment can fold into a domain or be unfolded is then calculated. Scooby-Domain employs an A*-search algorithm to search through a large number of alternative domain annotations. The A*-search algorithm considers combinations of different domain sizes, using a heuristic function to conduct the search. The corresponding sequence stretch for the first predicted domain is removed from the sequence. The search process is repeated until there are less than 34 residues remaining, which is the size of the smallest domain; or until there are no probabilities greater than 0.33, which is an arbitrary cutoff, to prevent non-domain segments from being predicted as a domain. However, A* search suffers from an exponential computational time complexity which was highlighted by Russell and Norvig [12].

Yoo et al. [16] introduced DomNet (Protein Domain Boundary Prediction Using Enhanced General Regression Network and New Profiles) which was trained using a compact domain profile, secondary structure, solvent accessibility information, and inter-domain linker index to detect possible domain boundaries for a target sequence. The authors proposed a semi-parametric model that uses a nonlinear auto-associative standard regression neural network (EGRN) for filtering noise and less discriminative features.

Chatterjee et al. [4] proposed the physicochemical properties as additional features to train support vector machine (SVM) classifier to improve the prediction of multi-domains in protein chains. The extracted attribute set consists of six different features; predicted secondary structure, predicted solvent accessibility, predicted conformational flexibility profile, amino acid composition, position specific scoring matrix (PSSM), and physico-chemical properties of amino acids.

Bondugula et al. [2] introduced FIEFDom (Fuzzy Integration of Extracted Fragments for Domains) as a method to predict domain boundaries of a multi-domain protein from its amino acid sequence using a Fuzzy mean operator (FMO). Using the non-redundant sequence database together with a reference protein set (RPS) containing known domain boundaries, the operator is used to assign a likelihood value for each residue of the query sequence as belonging to a domain boundary. FMO represents a special case of the Fuzzy nearest neighbor algorithm with the number of classes set to one. The approach is a three-step procedure. First, the position specific scoring matrix (PSSM) of the query sequence is generated using a large database of known sequences. Second, the generated profile is used to search for similar fragments in the RPS. Third, the matches with the proteins in RPS are parsed, and the domain boundary propensity of the query protein is predicted using a FMO.

Ebina et al. [5] developed DROP as SVM, with a radial basis function (RBF) kernel, domain linker predictor trained by 25 optimal attributes. The optimal combination of attributes were selected from a set of 3000 features using a random forest algorithm, which calculates the average decrease Gini index (MDGI), complemented with a stepwise feature selection. The selected features were primarily related to secondary structures, PSSM elements of hydrophilic residues and prolines.

The efficiency of DROP was evaluated by DS-All dataset [6, 15], CAFASP4 (http://www.cs.bgu.ac.il/~dfischer/CAFASP4/), and CASP8 FM (http://predictioncenter.org/casp8/). DS-All contains 169 protein sequences with a maximum sequence identity of 28.6% and 201 linkers. DROP achieved a prediction recall and precision of 41.3% and 49.4%, respectively, with more than 19.9% improvement by the optimal features. DROP performances were further assessed using the Average Overlapped Score (AOS) [10] and the Normalized Domain Overlap (NDO) score [14]. The AOS is the ratio of correctly assigned residue number to the total number of residues. The NDO-Score provides a single value that evaluates (penalize/prioritize) both over- and under-predictions. DROP does not use sequence similarity to domain databases. One of the advantages of this approach is the use of random forest approach for feature selection. Instead of exhaustively searching all feature combinations, random forest was employed which provides rapid and competitive screening for the optimal features. However, random forest can possibly be trapped in local minima and suffers from over-prediction. As a result, DROP over-predicts domain linkers in single-domain targets of BDS and CAFASP4.

In general, machine-learning based approaches are computationally expensive and often suffer from low prediction accuracy and susceptible to overfitting. Therefore, a simple method for identifying domain segments is desired. In this paper, we focus on the determination of domain-linker segments using amino acid compositional (AAC) index which predicts linker segments solely from the amino acid sequence information. The compositional index is deduced from the protein sequence dataset of domain-linker segments and the amino acid composition. A preference profile is then generated by calculating the average compositional index values along the amino acid sequence using a sliding window of varying sizes. Finally, a simulated annealing (SA) algorithm was employed to enhance the prediction by finding the optimal set of threshold values that separate domains from inter-domain linker segments. The rest of this paper is organized as follows. The next section presents our proposed method. Experimental results and discussion are presented in Section 3. Section 4 provides concluding remarks and future work directions.

2. METHOD

The proposed method consists of two main steps; calculating the compositional index and then refining the prediction by detecting the optimal set of threshold values that distinguish inter-domain linkers from non-linkers. In the first step, linker and non-linker segments are extracted from the training dataset and the differences in amino acid appearances in linker segments and non-linker segments are computed. Then, the amino acid composition of the test protein sequence is computed, and finally the amino acid compositional index is calculated.

In the second step, a simulated annealing algorithm is applied to find the optimal set of threshold values that will separate linker segments from non-linker segments. In the following sections, we describe both steps. An overview of the proposed method is illustrated in Figure 1.

2.1 Compositional Index

A protein is a polypeptide which is a linear polymer of several amino acids connected by peptide bonds. The primary structure of a protein is the linear sequence of its amino acid units. There are twenty amino acids that make up polypeptides and proteins. These amino acids are represented using one-letter abbreviation as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V.

Following Zaki et al. [18], we denote by S^* the enumerated set of protein sequences in the database. From each protein sequence s_i

in S^*, we extract known linker segments and non-linker segments and store them in datasets S_1 and S_2, respectively. To represent the preference for amino acid residues in linker segments, the compositional index r is calculated. The compositional index r_i for the amino acid i will be calculated as follows:

$$r_i = -\ln\left(\frac{f_i^{linker}}{f_i^{domain}}\right) \cdot a_i \qquad (1)$$

Where f_i^{linker} and f_i^{domain} are the frequency of amino acid residue i in S_1 and S_2, respectively. This is somewhat analogous to DomCut method [13]. However, the information contained in the index values proposed by [13] has no sufficient information to accurately predict the linker segments, thus we follow the improved index proposed by Zaki et al. [18] in which amino acid compositional knowledge was incorporated. The typical AAC contains 20 components, each of which reflects the normalized occurrence frequency for one of the 20 natural amino acids in a sequence. The AAC in this case is denoted by a_i. Each residue in every testing protein sequence is represented by its corresponding compositional index r_i. Subsequently, the index values are averaged over a window that slides along the length of each protein sequence. To calculate the average compositional index values m_j^w along a protein sequence s, using a sliding window of size w, we apply the following formula:

$$m_j^w = \begin{cases} \dfrac{\sum_{i=1}^{j+((w-1)/2)} r_{s_i}}{j+((w-1)/2)}, & 1 \le j \le (w-1)/2 \\[2ex] \dfrac{\sum_{i=j-((w-1)/2)}^{j+((w-1)/2)} r_{s_i}}{w}, & (w-1)/2 \le j \le L-((w-1)/2) \\[2ex] \dfrac{\sum_{i=j}^{L} r_{s_i}}{L-j+1+((w-1)/2)}, & L-((w-1)/2) \le j \le L \end{cases} \qquad (2)$$

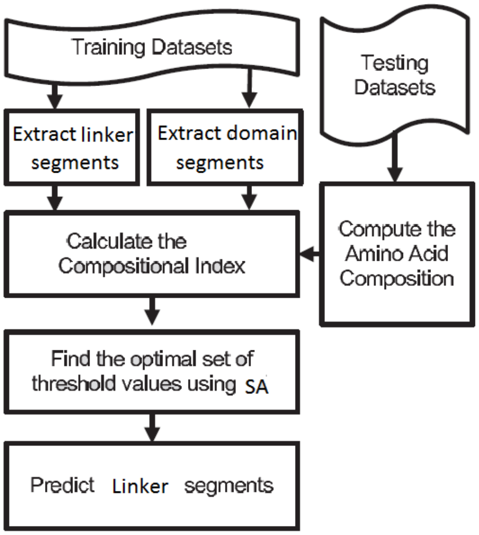
Training Datasets
Testing Datasets
Extract linker segments
Extract domain segments
Calculate the Compositional Index
Compute the Amino Acid Composition
Find the optimal set of threshold values using SA
Predict Linker segments

Figure 1: Overview of our approach.

where L is the length of the protein and s_j is the amino acid at position j in protein sequence s.

2.2 Detecting the Optimal Set of Threshold Values Using Simulated Annealing

Simulated Annealing is a probabilistic searching method for the global optimization of a given function in a large search space. It is inspired by the annealing technique which is the heating and controlled cooling of a metal to increase the size of its crystals and reduce their defects. The major advantage of SA over other optimization techniques is its ability to avoid being trapped in local optima. This is because the algorithm applies a random search which does not only accept changes that increase the objective function f (assuming a maximization problem), but also some changes that reduce it [3, 8].

In this case, the values m_j are used in conjunction with SA Algorithm to improve the prediction by detecting linkers and structural domains. This is done by first dividing each protein sequence into segments. The segment size was set to the standard linker size among the dataset. Then, starting from a random threshold value for each segment, SA is applied to predict the optimal threshold for each segment that maximizes both the recall and precision of the linker segment prediction. Recall $(\frac{\text{TP}}{\text{TP+FN}})$ is defined as the ratio of correctly predicted linkers to all of the structure-derived linkers listed in the dataset where TP is the number of amino acids within the known linker segment predicted as 'Linkers' and FN is the number of amino acid within the known linker segments predicted as 'Domains'. Precision $(\frac{\text{TP}}{\text{TP+FP}})$ is defined as the ratio of correctly predicted linkers to all of the predicted linkers where FP is the number of amino acid out of the known linker segment predicted as 'Linkers'.

Recall and precision were selected to be our measures to evaluate the performance of the proposed method due to several reasons. First, both evaluation measures were used to evaluate the performance of most of the current approaches which will allow comparing the accuracy of our method to the state-of-the-art methods. Secondly, recall and precision can handle unbalanced data situation where data points are not equally distributed among classes.

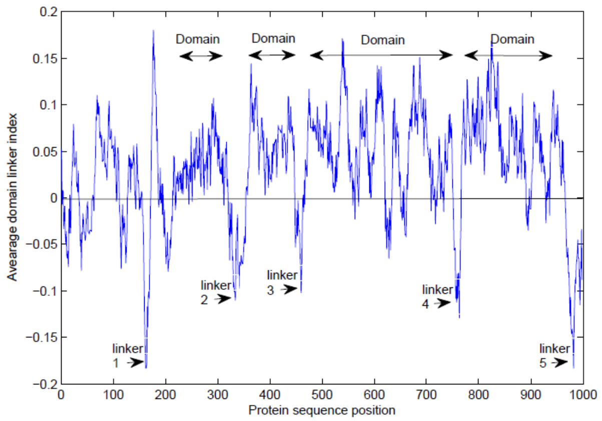

Figure 2: Linker preference profile.

```
ALGORITHM: Simulated Annealing for Linker Prediction Optimization
Set Initial state $s_0$:
    Divide the protein sequence into segments of size segmentSize
    Assign 0 as the initial threshold of each segment
Calculate the AA compositional index of the input protein sequence
Classify each AA as linker (1) or domain (0) according to its compositional index value with respect to the corresponding segment threshold.
Calculate the objective functions $f_1(s_0)$= recall and $f_2(s_0)$= precision
Maximize recall and precision:
    Select initial temperature $T_0 = 0.1$
    Select temperature reduction function $\alpha = 0.9$
    For n = 1 to Number of Segments
        Repeat
            Make a transition Tr:
                randomly increase or decrease the threshold of seg
                $s = Tr(s_0)$
            Classify each AA as linker or domain
            Calculate $f_1(s)$ and $f_2(s)$
            $\delta f_1 = f_1(s) - f_1(s_0)$, and $\delta f_2 = f_2(s) – f_2(s_0)$
            If $\delta f_1 > 0$ and $\delta f_2 > 0$
                then $s_0 = s$
            else
                generate a random number $r \epsilon R(0,1)$
                 if $r < e^{(-\delta f1/T0)}$
                   then $s_0 = s$
        Until iteration_count = 20
        Set $T = \alpha * T$
    End
Return $s_0$ as the optimal threshold values for the protein sequence segments
Return $f_1(s_0)$and $f_2(s_0)$ as the final recall and precision, respectively
```

Algorithm 1: Detecting the optimal set of threshold values using Simulated Annealing.

In this case SA will accept a transition that leads to one of the three following conditions: an increase in both recall and precision, an increase in recall if precision is not changed, or an increase in precision if recall is not changed. SA is summarized in Algorithm 1.

3. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of our proposed method was evaluated on two protein sequence datasets. The first dataset was used by Suyama and Ohara [13] to evaluate the performance of DomCut which was extracted from the Swiss-Prot data base [1]. The dataset contains non-redundant set of 273 protein sequences (486 linker and 794 domain segments). The average numbers of amino acid residues in linker and domain segments were 35.8 and 122.1, respectively. The second dataset is DS-All [6, 15] which was used to evaluate the performance of DROP [5] in predicting inter-domain linker segments.

Applying our method to DomCut/Swiss-Prot protein sequence dataset leads to an average recall of 0.452 and precision of 0.761 with the segment size of 36 AA (average linker size in the dataset). With 18 AA segment size (half of the average linker size), we achieved a recall of 0.558 and precision of 0.836. It is worth to mention that the recall and precision of DomCut were respectively 0.535 and 0.501, which was achieved at the threshold value of −0.09.

When we evaluated the performance of the proposed method on 151 protein sequences of DS-All dataset (182 linker and 332 domain segments), setting the segment size to 13 AA (average linker size in DS-All dataset), we achieved an average prediction recall of 0.592, and precision of 0.595. The comparison of the performance of our approach against the currently available domain linker prediction approaches [5] are reported in Table 1. It is clear to see that the proposed method outperformed the state-of-the-art domain-linker prediction approaches in both recall and precision.

To demonstrate the performance of the proposed method, in Figure 3, we show the optimal threshold values for an example (protein-256) in DomCut dataset as predicted by our method while in Figure 4, we show the optimal threshold values for another example (protein-151) in DS-All dataset. It is clearly shown that the compositional index threshold values at linker segments are increased by the algorithm while threshold values of domains are reduced.

Table 1: Prediction performance of publicly domain linker prediction approaches.

Approach	Recall	Precision
Proposed method	0.592	0.595
DROP	0.413	0.494
DROP-SD5.0	0.428	0.515
DROP-SD8.0	0.418	0.503
SVM-PeP	0.403	0.491
SVM-SD3.0	0.373	0.446
SVM-SD2.0	0.353	0.420
SVM-Af	0.214	0.257
Random	0.050	0.060

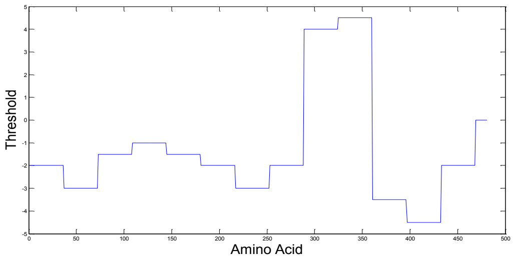

Figure 3: Optimal threshold values for protein-256 in DomCut dataset.

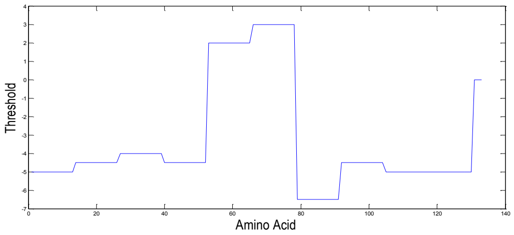

Figure 4: Optimal threshold values for protein-151 in DS-All dataset.

4. CONCLUSION AND FUTURE WORK

In this work, we examined the amino acid compositional index to predict protein inter-domain linker segments from amino acid sequence information. Then, we employed simulated annealing to improve the prediction by finding the optimal set of threshold values that separate domains from linker segments. The performance of the proposed method was compared to the state-of-the-art approaches on two well-known protein sequence datasets. Experimental results show that the proposed method outperformed the currently available approaches for inter-domain linker prediction in terms of recall and precision. We achieved a recall of 0.592 and a precision of 0.595.

This work can be extended by examining different sliding window sizes in computing the AA composition. It is expected that the combination of different window sizes will provide more flexibility in accounting for the length variation of linker segments, reduce the bias towards a fixed linker segment length, and produce a series of features for each protein sequence. Although simulated annealing has significantly improved the prediction, additional tuning and other strategy choices could accomplish more effective and flexible prediction. One of these tuning strategies is to use dynamic segment sizes which can, in turn, leads to a better optimization process. The proposed method has a potential to perform well if it is applied in all human protein sequences where novel inter-domain linkers could be identified.

5. REFERENCES

[1] Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48.

[2] Bondugula, R., Lee, M. S., and Wallqvist, A. 2009. FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator. *Nucleic Acids Research*. 37, 2, 452–462. doi:10.1093/nar/gkn944

[3] Busetti, F. 2003. *Simulated annealing overview*. Report, CiteSeer.

[4] Chatterjee, P., Basu, S, Kundu, M., Nasipuri, M., Basu, D. P. 2009. Improved prediction of Multi-domains in protein chains using a Support Vector Machine. *International Journal of Recent Trends in Engineering*. 2, 3.

[5] Ebina, T., Toh, H., and Kuroda, Y. 2011. DROP: an SVM domain linker predictor trained with optimal features selected by random forest. *Bioinformatics*. 27, 487–494. doi:10.1093/bioinformatics/btq700

[6] Ebina, T., Toh, H., and Kuroda, Y. 2009. Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics. *Biopolymers*. 92, 1–8.

[7] George, R.A., Lin, K. and Heringa, J. 2005. Scooby-domain: prediction of globular domains in protein sequence. *Nucleic Acids Res*. 33, W160–W163.

[8] Henderson, D., Jacobson, S. H., Johnson, A. W. 2003. The Theory and Practice of Simulated Annealing. Chapter 10, *Handbook of Metaheuristics. International Series in Operations Research & Management Science*. 57, 287-319.

[9] Hondoh, T., Kato, A., Yokoyama, S., and Kuroda, Y. 2006. Computer-aided NMR assay for detecting natively folded structural domains. *Protein Science*. 15, 871–883.

[10] Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C., and Thornton, J. M. 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Science*. 7: 233-242. Cambridge University Press.

[11] Pang, C. N. I. , Lin, K., Wouters, M. A., Heringa, J., and George, R. A. 2008. Identifying foldable regions in protein sequence from the hydrophobic signal. *Nucleic Acids Research*. 36, 2, 578–588. doi:10.1093/nar/gkm1070

[12] Russell, S. J. and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach.* Upper Saddle River. N.J. Prentice Hall. 97–104. ISBN 0-13-790395-2.

[13] Suyama M. and Ohara, O. 2003. Domcut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics*. 19, 5, 673–674.

[14] Tai, CH., Lee, W.J., Vincent, J.J., Lee, B. 2005. Evaluation of domain prediction in CASP6. *Proteins*. 61, Suppl. 7, 183–192.

[15] Tanaka, T., Yokoyama, S., and Kuroda, Y. 2006. Improvement of domain linker prediction by incorporating loop-length-dependent characteristics. *Biopolymers*. 84, 161–168.

[16] Yoo, P. D., Sikder, A. R., Taheri, J., Zhou, B. B., and Zomaya, A. Y. 2008. DomNet: Protein Domain Boundary Prediction Using Enhanced General Regression Network and New Profiles. *IEEE Transactions on Nanobioscience*. 7, 2, 172-181.

[17] Zaki, N. 2008. Prediction of Protein–Protein Interactions Using Pairwise Alignment and Inter-Domain Linker Region. *Engineering Letters*. 16, 4.

[18] Zaki, N., Bouktif, S., and Lazarova-Molnar, S. 2011. A Genetic Algorithm to Enhance Transmembrane Helices Topology Prediction Using Compositional Index. In *Proceedings of the ACM Genetic and Evolutionary Computation Conference (GECCO2011)*. Dublin, Ireland.

[19] Zaki, N. and Campbell, P. 2009. Domain Linker Region Knowledge Contributes to Protein-protein Interaction Prediction. In *Proceedings of the International Conference on Machine Learning and Computing*. Perth, Australia, 70-74.