Computational Biology and Chemistry 55 (2015) 23–30

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Inter-domain linker prediction using amino acid compositional index

Maad Shatnawi*, Nazar Zaki

College of Information Technology, UAEU, United Arab Emirates

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form 22 January 2015
Accepted 22 January 2015
Available online 24 January 2015

Keywords:
Domain linker prediction
Amino acid composition
Compositional index
Simulated annealing

ABSTRACT

Protein chains are generally long and consist of multiple domains. Domains are distinct structural units of a protein that can evolve and function independently. The accurate and reliable prediction of protein domain linkers and boundaries is often considered to be the initial step of protein tertiary structure and function predictions. In this paper, we introduce CISA as a method for predicting inter-domain linker regions solely from the amino acid sequence information. The method first computes the amino acid compositional index 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. Finally, the protein sequence is segmented into intervals and a simulated annealing algorithm is employed to enhance the prediction by finding the optimal threshold value for each segment that separates domains from inter-domain linkers. The method was tested on two standard protein datasets and showed considerable improvement over the state-of-the-art domain linker prediction methods.

1. Introduction

Proteins can be considered to be built up from domains, and each of which can be considered as a semi-independent structural unit of a protein capable of folding independently (Sun et al., 2013). On the other hand, several domains are often joined together in several combinations to form multi-domain protein sequences (Chothia, 1992; Yoo et al., 2008). Inter-domain linkers tie neighboring domains and support inter-domain communication in multi-domain proteins. They also provide sufficient flexibility to facilitate domain motions and regulate the inter-domain geometry (Bhaskara et al., 2012). Domain linkers play a key role in inter-domain interactions, function regulation, and protein stability (Gokhale and Khosla, 2000; Lehtinen et al., 2004; Dumontier et al., 2005). Several domain prediction methods first detect domain linkers, and, in turn, predict the location of domain regions. The knowledge of structural domains is used to classify proteins, understand their structures, functions and evolution, and predict protein–protein interactions (PPI) (Zaki, 2009). Therefore, accurate computational methods for splitting proteins into structural domains are essential in proteomics research (Hondoh et al., 2006).

Several impressive inter-domain linker prediction methods have been developed and can be generally classified into statistical-based and machine learning-based methods. One of the earlier statistical-based methods is DomCut[1] (Suyama and Ohara, 2003) which predicts inter-domain linker regions based on the differences in amino acid (AA) composition between domain and linker regions in a protein sequence. To represent the preference for AA residues in linker regions, the authors defined the linker index as: $S_i = -\ln(f_i^l / f_i^d)$, where f_i^l and f_i^d are the frequencies of AA residue i in the linker (l) and domain (d) region, respectively. A linker is predicted if there is a trough in the linker index profile and the averaged linker index value at the bottom of the trough is lower than a defined threshold value. Domcut was tested on a non-redundant 273 protein sequences and achieved 53.5% recall and 50.1% precision. Despite the fact that DomCut showed glimpse of potential success, it was reported by Dong et al. (2006) that DomCut has low sensitivity and specificity in comparison to other later published methods. Therefore, integrating more biological evidences in the linker index could enhance the prediction of the inter-domain linkers and therefore, the idea of DomCut was later expanded by several researchers such as Pang et al. (2008) to identify foldable regions and Zaki et al. (2011a,b) to identify transmembrane helical segments in a protein sequence.

Linding et al. (2003) presented another statistical-based method named GlobPlot.[2] GlobPlot allows users to plot the tendency within

* Corresponding author. Tel.: +971 506151987.
E-mail address: shatnawi@uaeu.ac.ae (M. Shatnawi).

[1] http://www.bork.embl.de/~suyama/domcut/.
[2] http://globplot.embl.de.

http://dx.doi.org/10.1016/j.compbiolchem.2015.01.006
1476-9271/

protein sequences for exploring both potential globular and disordered/flexible regions in proteins based on their AA sequence, and to identify inter-domain segments containing linear motifs. Other statistical-based methods were presented by Udwary et al. (2002) which predicts the locations of linker regions within large multifunctional proteins and Dumontier et al. (2005) which predicts domain linkers by using AA composition.

Machine learning (ML) based methods are the most commonly used approaches in protein domain linker prediction. Most of the recent approaches employ either Artificial Neural Networks (ANN) such as PRODO (Sim et al., 2005) or Support Vector Machines (SVM) such as DROP (Ebina et al., 2011) to improve the prediction. Other examples of ML-based methods include DomNet (Yoo et al., 2008), Chatterjee et al. (2009), DoBo (Eickholt et al., 2011) and ThreaDom (Xue et al., 2013). However, despite the success of the above mentioned methods they mainly suffer from the following limitations:

- Most ML-based methods such as PRODO, DROP, and DomNet are computationally expensive. They require high computational cost to generate the Position-Specific Scoring Matrix (PSSM) and/or predict secondary structure information in a single protein sequence. Finding the structural information by itself is another challenge. In contrast, predicting the domain linkers could lead to inferring the structural information.
- Some methods such as PRODO are evaluated based on the overall prediction accuracy which is not a good evaluation strategy in imbalanced data problems. Protein data are imbalanced as domain regions are much longer than linkers and, therefore, classifiers will usually be biased towards the majority class. For example, if domain regions occupy a long portion of a protein sequence, then a classifier could simply assign a domain to the whole protein sequence with high classification accuracy, however, the linkers in this case will not be predicted (all instances classified as +ve and no −ve classes are predicted).

In this work, we develop CISA, a simple but yet effective method for domain-linker prediction solely from AA information. Domain-linker regions will be determined using AA compositional index (CI), and then, a simulated annealing (SA) algorithm will be employed to enhance the prediction by finding the optimal threshold value that separates domains from linkers.

2. Method

The proposed method consists of two main steps; calculating the AA compositional index (CI) for the protein sequence of interest and then refining the prediction by detecting the optimal set of threshold values that distinguish between inter-domain linkers and non-linkers. In the first step, linker and non-linker segments are extracted from the protein sequence dataset and the frequencies of AA appearances in linker segments and non-linker segments are computed. Then, the AA composition of the query protein sequence is computed, and finally the AA compositional index is calculated. In the second step, SA algorithm is applied to find the optimal set of threshold values that separate linker segments from non-linker segments through the compositional index profile. An overview of CISA is illustrated in Fig. 1. Both steps are described in the proceeding sections.

2.1. Datasets

Two protein sequence datasets were used to evaluate the performance of CISA. The first dataset is extracted from the Swiss-Prot database (Bairoch and Apweiler, 2000) and it was previously used by Suyama and Ohara (2003) to evaluate the performance of

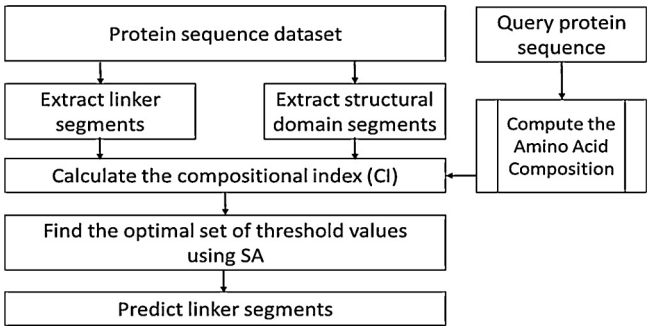

Fig. 1. Overview of CISA.

DomCut. This dataset contains 273 non-redundant protein sequences including 486 linker and 794 domain segments. The average numbers of AA residues in linker and domain segments are 35.8 and 122.1, respectively.

The second dataset is DS-All (Tanaka et al., 2006; Ebina et al., 2009) which was used to evaluate DROP (Ebina et al., 2011). This dataset was extracted from the non-redundant Protein Data Bank (nr-PDB) chain set[3] by selecting protein sequences containing two or more continuous domains as defined in version SCOP 1.69. The dataset contains 182 protein sequences including 216 linker segments. We excluded few sequences which found to be inconsistent with the PDB database,[4] ending up with 151 sequences including 334 domains and 183 linker segments. The average numbers of AA residues in linker and domain segments are 12.7 and 147.1, respectively.

2.2. Evaluation measures

The evaluation measures that we used in this work are recall (R), precision (P), and F1-measure. Recall is defined as the proportion of correctly predicted linkers to all of the structure-derived linkers listed in the dataset. Precision is defined as the proportion of correctly predicted linkers to all of the predicted linkers. Recall and precision are class-independent measures that can handle unbalanced data situation where data points are not equally distributed among classes such as domain-linker data. The F1-measure is an evaluation metric that combines precision and recall in a single value. It is defined as the harmonic mean of precision and recall (Sasaki, 2007; Powers, 2011). F1-score is used as a unified measure to compare two approaches when one approach has higher recall and lower precision than the other.

2.3. Compositional index

From each protein sequence s_i in the protein sequences database S^*, known linker segments and domain segments are extracted and saved in two datasets S_1 and S_2, respectively. The compositional index c_i of the AA is calculated to represent the existence of each AA residue in linker segments and it is defined as:

$$c_i = -\ln\left(\frac{f_i^l}{f_i^d}\right) \cdot \left(\frac{k}{a_i}\right) \quad (1)$$

This is inspired by DomCut method (Suyama and Ohara, 2003) as discussed in Section 1. However, the information encoded in the linker index (LI) is insufficient to precisely predict linker segments. Therefore, we used the compositional index proposed by

[3] http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html.
[4] http://http://www.ncbi.nlm.nih.gov/protein.