Domain-Based Approaches to Prediction and Analysis of Protein-Protein Interactions

Morihiro Hayashida, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan
Tatsuya Akutsu, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan

ABSTRACT

Protein-protein interactions play various essential roles in cellular systems. Many methods have been developed for inference of protein-protein interactions from protein sequence data. In this paper, the authors focus on methods based on domain-domain interactions, where a domain is defined as a region within a protein that either performs a specific function or constitutes a stable structural unit. In these methods, the probabilities of domain-domain interactions are inferred from known protein-protein interaction data and protein domain data, and then prediction of interactions is performed based on these probabilities and contents of domains of given proteins. This paper overviews several fundamental methods, which include association method, expectation maximization-based method, support vector machine-based method, linear programming-based method, and conditional random field-based method. This paper also reviews a simple evolutionary model of protein domains, which yields a scale-free distribution of protein domains. By combining with a domain-based protein interaction model, a scale-free distribution of protein-protein interaction networks is also derived.

Keywords: Data Mining, Domain-Domain Interactions, Prediction, Protein-Protein Interactions, Scale-Free Distribution

INTRODUCTION

Understanding of functions of genes and proteins is important in post-genomic era. Information on protein-protein interactions is useful for understanding protein functions because protein-protein interactions play a key role in many cellular processes. Since the end of the last century, some experimental techniques have been developed for comprehensive analysis of protein-protein interactions, which include two-hybrid systems and proteomics methods. Though these experimental methods revealed many unknown interactions, there were large gaps between results done by different groups (Ito et al., 2001; Uetz et al., 2000). Therefore, computational methods should be developed for inference of protein-protein interactions. For that purpose, various approaches have been proposed. In this paper, we focus on...
computational and mathematical aspects of domain-based approaches.

A protein consists of one or multiple domains, where a domain is defined as a region within a protein that either performs a specific function or constitutes a stable structural unit. Examples of structural domains are illustrated in Figure 1 though domains are sometimes defined not based on structures but based on sequence/functional similarities. In a word, domains are considered as parts of a protein. Though there is no exact or mathematical definition of protein domains, several hundreds of protein domains are currently known. In order to classify domains, several database systems have been constructed, which include Pfam (Punta et al., 2012), InterPro (Hunter et al., 2012) and ProDom (Bru et al., 2005). Furthermore, most of these databases provide facilities to identify protein domains from a given protein sequence. In Pfam, each domain is represented by an HMM (Hidden-Markov Model) and protein domains contained in a given protein sequence are identified by using these HMMs.

Utilizing information of domain organizations of proteins, several methods have been proposed for prediction of protein-protein interactions. In these methods, scores or probabilities of domain-domain interactions are first derived from known protein-protein interactions and then these are utilized for calculating the score or probability of protein-protein interaction for given protein sequences. Sprinzak and Margalit (2001) proposed the association method for computing the score of each domain pair. Kim et al. (2002) proposed similar scores and applied the scores to inference of protein-protein interactions. Deng et al. (2002) proposed an EM (Expectation-Maximization) algorithm for estimating the probability of interaction for each domain pair.

We proposed a conditional random field approach using mutual information between domain sequences (Hayashida et al., 2011). In the study, mutual information between positions of amino acid residues in domains is calculated from multiple sequence alignments because it is known that interacting residue pairs tend to have higher mutual information than non-interacting ones (Fraser et al., 2002). In addition, several methods using mutual information have been developed for inference of protein residue-residue interactions (Weight et al., 2009; Kamada et al., 2011; Jones et al., 2012). Markov and conditional random field models have been well studied in fields of natural

Figure 1. Example of protein domains. Protein \( P_1 \) consists of domains \( D_1 \) and \( D_2 \), whereas protein \( P_2 \) consists of domains \( D_3 \), \( D_4 \) and \( D_5 \). In domain-based models, it is assumed that \( P_1 \) and \( P_2 \) interact with each other if at least one domain pair interacts.
Related Content

Biological Background
(2011). Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations  (pp. 1-5). www.irma-international.org/chapter/biological-background/53892/

Evaluating a Genetics Concept Inventory
Felicia Zhang and Brett Andrew Lidbury (2013). Bioinformatics: Concepts, Methodologies, Tools, and Applications  (pp. 29-41). www.irma-international.org/chapter/evaluating-genetics-concept-inventory/76055/

Computer Aided Risk Estimation of Breast Cancer: The “Hippocrates-mst” Project
George M. Spyrou and Panos A. Ligomenides (2010). Biocomputation and Biomedical Informatics: Case Studies and Applications  (pp. 204-214). www.irma-international.org/chapter/computer-aided-risk-estimation-breast/39614/

State-of-the-Art GPGPU Applications in Bioinformatics

Integrated Features Based on Gray-Level and Hu Moment-Invariants with Ant Colony System for Retinal Blood Vessels Segmentation