Osong Public Health Res Perspect 2015 6(2), 112—120
http://dx.doi.org/10.1016/j.phrp.2015.01.006
pISSN 2210-9099 eISSN 2233-6052

ORIGINAL ARTICLE

Application of Gap-Constraints Given Sequential Frequent Pattern Mining for Protein Function Prediction

Hyeon Ah Park [a], Taewook Kim [b], Meijing Li [a], Ho Sun Shon [c], Jeong Seok Park [d], Keun Ho Ryu [a],*

[a]*Database/Bioinformatics Laboratory, College of Electrical and Computer Engineering Chungbuk National University, Cheongju, Korea.*
[b]*Syntekabio Incorporated, Korea Institute of Science and Technology, Seoul, Korea.*
[c]*Graduate School of Health Science Business Convergence, Chungbuk National University, Cheongju, Korea.*
[d]*Medical Informatics·Engineering, Korea National University of Transportation, Cheongju, Korea.*

Received: November 3, 2014
Revised: December 14, 2014
Accepted: January 16, 2015

KEYWORDS:
frequent pattern mining with gap-constraint,
graph pattern mining,
protein function prediction,
protein—protein interaction network,
sequential pattern mining

Abstract

Objectives: Predicting protein function from the protein—protein interaction network is challenging due to its complexity and huge scale of protein interaction process along with inconsistent pattern. Previously proposed methods such as neighbor counting, network analysis, and graph pattern mining has predicted functions by calculating the rules and probability of patterns inside network. Although these methods have shown good prediction, difficulty still exists in searching several functions that are exceptional from simple rules and patterns as a result of not considering the inconsistent aspect of the interaction network.

Methods: In this article, we propose a novel approach using the sequential pattern mining method with gap-constraints. To overcome the inconsistency problem, we suggest frequent functional patterns to include every possible functional sequence—including patterns for which search is limited by the structure of connection or level of neighborhood layer. We also constructed a tree-graph with the most crucial interaction information of the target protein, and generated candidate sets to assign by sequential pattern mining allowing gaps.

Results: The parameters of pattern length, maximum gaps, and minimum support were given to find the best setting for the most accurate prediction. The highest accuracy rate was 0.972, which showed better results than the simple neighbor counting approach and link-based approach.

*Corresponding author.
E-mail: khryu@dblab.chungbuk.ac.kr

Conclusion: The results comparison with other approaches has confirmed that the proposed approach could reach more function candidates that previous methods could not obtain.

1. Introduction

Defining functional characteristics of newly found protein or reassigning new functions to already-found protein has been receiving attention from scientists. Analyzing uncharacterized functions of proteins requires a sophisticated computational method, because it is impossible to manually annotate the large amount of constantly uploaded data as proteins tend to carry biological function in more than one aspect.

Although the classic way of predicting a protein function is to find the homology between the sequence of annotated protein and unannotated proteins, the question of being sensitive enough for diverse sequences still remains. Some studies have inferred the function of a protein using its three-dimensional structure [1] using the similarity of fold, but most folds are associated with only a single function whereas proteins can have multiple functions, and thus could be confusing. Later, bioinformatics techniques to analyze biological process [2,3], clustering, and classification to categorize protein function from DNA data were introduced [4—6]. After the protein—protein interaction network, which shows the functional association between proteins, was introduced, it was often used for function prediction of proteins due to its rich information [7]. The methods to exploit the network have been developed in several different ways, including majority voting method [8], global optimization method [9], labeling and weight assign method [10,11], etc. The protein network can be exploited in various ways because it is packed with a vast amount of information and be easily combined with other information in the form of annotation and weight. Interacting proteins are composed of highly complex networks referred to as protein—protein interaction networks. This successfully captures the feature of the condition of protein relationships. Interacting proteins are likely to share the same functions to serve a common purpose, but predicting protein function solely on this feature has generally demonstrated limited accuracy and efficiency for several reasons. First, protein—protein interaction networks are typically structured on very complex connectivity, therefore making the prediction procedure more challenging if proteins have too many large numbers of neighbors [19]. Second, most proteins have multiple functions under different environmental conditions, which creates more difficulty in predicting the whole, complete set of functions that a single protein may carry [7]. Finally, functional inconsistency exists between interacting proteins.

In the study of Schwikowski et al [8], which features the analysis of a large protein interaction network, function prediction relying on interacting proteins is proved to be "highly effective". Although counting the frequency of function categories among neighbor proteins works well for prediction, because of the complexity of the relationship between proteins it has encouraged applying a more sophisticated way to bring better accuracy rates in predictions. One cannot simply tell that a protein will definitely possess a function that its neighbor has—it is a matter of probability, as it is affected by a tangled relationship of proteins with some exceptions. The study by Vazquez et al [9] takes the entire picture of the network to connect all possible impact factors proteins give to each other, to decide what function each will serve. Methods considering such extra influence within networks are also well shown in the study of Chatterjee et al [12], which uses the distance between proteins, for example, and Freschi et al [13], applying rank or weight, or inserting labels as in the study by Wang et al [11]. Although adding some extra factors can better reflect the protein interaction process, all of these comprise parts of all resources gleaned from the interaction network. Combining the strong characters of each local network of strongly related proteins and the global connection flow of all local networks, and additional information tagged into the network are essential as every factor derived from this interaction network are equally effective at guessing the function of protein. Applying graph mining and involving pattern mining can provide the answer to this problem [14,15]. Even when mining a whole network, both global aspects and local aspects of connection can be easily spotted as a large or small pattern, or as in a subpattern inside a larger pattern. Because the pattern mining approach can be easily "equipped" with several constraints and weight factors [16], it makes it possible to return any good sample of how each well-known functions of proteins can indicate the unknown function of their interaction neighbors. A study by Freschi [17] suggests topology analysis that takes overlapping neighbors into account, assigning different weight to different neighboring node patterns in the end. In the study by Cho and Zhang [18], such an approach is attempted to be improved by applying a more advanced pattern mining technique. During the labeled subgraph mining for functional pattern, *a priori* pruning is applied and triangular duplicated candidate patterns are eliminated. Still, the question of inconsistency remains because no prominent, single rule of patterns exists for one particular function [21].