Methods 110 (2016) 73–80

Contents lists available at ScienceDirect

Methods

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

Conservation of hot regions in protein–protein interaction in evolution

Jing Hu [a,b], Jiarui Li [c], Nansheng Chen [c,*], Xiaolong Zhang [a,b,*]

[a] *School of Computer Science and Technology, Wuhan University of Science and Technology, China*
[b] *Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, China*
[c] *Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, Canada*

ARTICLE INFO

Article history:
Received 13 March 2016
Received in revised form 8 June 2016
Accepted 21 June 2016
Available online 21 June 2016

Keywords:
Hot region
Protein–protein interaction
Conservation
Evolution

ABSTRACT

The hot regions of protein–protein interactions refer to the active area which formed by those most important residues to protein combination process. With the research development on protein interactions, lots of predicted hot regions can be discovered efficiently by intelligent computing methods, while performing biology experiments to verify each every prediction is hardly to be done due to the time-cost and the complexity of the experiment. This study based on the research of hot spot residue conservations, the proposed method is used to verify authenticity of predicted hot regions that using machine learning algorithm combined with protein's biological features and sequence conservation, though multiple sequence alignment, module substitute matrix and sequence similarity to create conservation scoring algorithm, and then using threshold module to verify the conservation tendency of hot regions in evolution. This research work gives an effective method to verify predicted hot regions in protein–protein interactions, which also provides a useful way to deeply investigate the functional activities of protein hot regions.

1. Introduction

Though research of protein interaction interface, alanine mutation experiment discoveries that not all residues but only a few play an important role to protein binding, and those important residues are called hot spot residues [1]. Numerous studies have addressed that hot spots formed to a special conformation during process of protein interaction, instead of being distributed along protein interfaces homogeneously, hot spot residues are usually clustered within tightly packed regions, which are called hot regions [2], Fig. 1 shows the hot region in protein complex 1A22.

Hot regions of protein–protein interactions play important roles in the functions and stability of protein complexes; they are more important than hot spots in maintaining the stability of protein complexes and exerting the molecular mechanism of biological functions. The research on hot regions are very important to understanding the protein activities like disease origin, pharmaceuticals, drug effect targeting, etc.

With the development and deeply study of proteomics, intelligent computing method [3–5] more and more became supporting technology of proteomics. Discovery and discrimination of hot regions in protein-protein interaction brings lots of predicted hot regions, while there is no efficiency method to verify authenticity of predicted hot regions. As we know, biological experiment is the most direct method to verify authenticity of predicted hot regions, while because of long-time and complexity of biological experiment, it's impossible to verify every predicted result using biological experiment. The scale and number of verifying is limited by existing method mainly focused on comparing with published literatures. Thus, how to using intelligent computing method to verify hot regions has become an important and urgent topic.

In recent years, many studies have been made to hot regions. In 2005, Keskin [2] developed an algorithm to cluster hot spots into hot regions after studying the organization and contribution of structurally conserved hot spot residues. Further analysis show that hot spots in hot regions are usually more structurally conserved than other interface residues, and hot spots play roles associated with surrounding interface residues, which lead to binding free energy of hot spots are higher than other interface residues.

In 2007, Hsu [6] presented a pattern-mining approach for the identification of hot regions in protein–protein interactions, which demonstrates that the important residues associated with the interface residues may be discovered by sequential pattern-mining automatically. The proposed method aimed to locate hot region structure modules with sequence modules by multiple sequence alignment with homologous proteins and analyzing higher conserved modules with alignment results.

* Corresponding authors at: Department of Molecular Biology and Biochemistry, Simon Fraser University, Canada (N. Chen). School of Computer Science and Technology, Wuhan University of Science and Technology, China (X. Zhang).

E-mail addresses: chenn@sfu.ca (N. Chen), xiaolong.zhang@wust.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.ymeth.2016.06.020
1046-2023/© 2016 Published by Elsevier Inc.

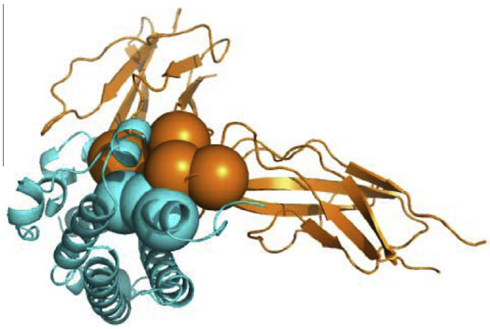

Fig. 1. Hot region in protein complex 1A22. In complex 1A22, hot spots in hot region are shown as small spheres, Spheres and their located chain are the same color, while different colors represent different chains, the cyan belt represents chain A and the orange belt represents chain B.

In 2010, Cukuroglu [7] proposed there are more hot regions are formed between hub proteins by analyzing hot regions organized structure though studying physical and chemical features of three kinds of interface, and then built a database called Hot Region [8].

In 2014, Nan [9] proposed a method to predict hot regions based on complex network and community detection. By revising false positive and false negative during the detection process, the proposed method can improve the reliability in the recognition of hot regions. In 2015, Hu [10] propose a method to predict hot regions in protein–protein interactions by combining density-based incremental clustering with feature-based classification. These method further improved precision of hot region prediction.

In the aspect of protein conservation, lots of study showed that hot spots in protein-protein interactions are usually more structurally conserved than other surface residues [1,2,11], while there is little correlation of conservation between interface residues and other surface residues, either in sequence [12] or in structure [1,13]. Hot regions are composed by hot spot residues, and have so many similar biological features with hot spots, and conservation is key and fundamental features of hot regions.

Based on above researches, we have sound reasons to use conservation features to verify hot regions in protein interactions.

2. Method

Here a sequence conservation-based method to test hot regions is proposed. For each every hot spot in a hot region, two proteins that form the complex are found in the protein database, and by isoforms we obtain the complete sequence of genes, with the complete protein sequence we get all the orthologs in different species, and then by multiple sequence alignment, we record all the sites of the hot spot residue in different species, here we apply Blocks Substitution Matrix to build conservation scoring function of hot regions for the first time, and by the scoring function we built, the conservation scores of the hot regions in different species are obtained, and at last, we calculate the hot region conservation relative to other regions on the interaction interface. We test whether hot regions are conserved in sequence in different species following four steps:

Step 1: Find isoforms of each gene in the protein complex;
Step 2: Find orthologs for each gene in different species using isoforms obtained in the first step;
Step 3: Perform multiple sequence alignments using orthologs obtained in the second step;
Step 4: Calculate a conservation score for each hot region using the scoring function and calculate the conservation probability of each hot region compared to other binding sites in different species.

For more clearly show the method, we take a protein complex 1A22 for example in every step.

2.1. Isoforms of protein

In the first step, we identified isoforms of each gene in protein complexes within hot regions. In the database of this paper, all complexes are composed of two or more proteins. We find isoforms according to each chain of every complex. Here two genes of protein complex 1A22 are GH1 (GROWTH HORMONE) and GHR (GROWTH HORMONE RECEPTOR). Through the Protein data bank [14] (PDB), we obtained the UniProt ID of each gene of every complex. In the UniProt database [15] we obtained the sequences of isoforms using UniProt ID and, if there is more than one isoform in a single gene, we use the one most closely similar by sequence alignment as this gene's isoform. There are 5 isoforms of GH1, which are list below, and though multiple sequence alignment, sp:isoform-1 is selected as isoform of GH1.

>1A22:A|PDBID|CHAIN|SEQUENCE FPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCFSESIPTPSNREETQQKSNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDSNVYDLLKDLEERIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDDALLKNYGLLYCFRKDMDKVETFLRIVQCRSVEGSCGF

>sp:isoform-1|P01241|SOMA_HUMAN Somatotropin OS=*Homo sapiens* GN=GH1 PE=1 SV=2 MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCFSESIPTPSNREETQQKSNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDSNVYDLLKDLEEGIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDDALLKNYGLLYCFRKDMDKVETFLRIVQCRSVEGSCGF

>sp:isoform-2|P01241-2|SOMA_HUMAN Isoform 2 of Somatotropin OS=*Homo sapiens* GN=GH1 MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFNPQTSLCFSESIPTPSNREETQQKSNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDSNVYDLLKDLEEGIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDDALLKNYGLLYCFRKDMDKVETFLRIVQCRSVEGSCGF

>sp:isoform-3|P01241-3|SOMA_HUMAN Isoform 3 of Somatotropin OS=*Homo sapiens* GN=GH1 MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCFSESIPTPSNREETQQKSNLELLRISLLLIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDDALLKNYGLLYCFRKDMDKVETFLRIVQCRSVEGSCGF

>sp:isoform-4|P01241-4|SOMA_HUMAN Isoform 4 of Somatotropin OS=*Homo sapiens* GN=GH1 MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCFSESIPTPSNREETQQKSNLELLRISLLLIQSWLEPVQIFKQTYSKFDTNSHNDDALLKNYGLLYCFRKDMDKVETFLRIVQCRSVEGSCGF

>sp:isoform-5|P01241-5|SOMA_HUMAN Isoform 5 of Somatotropin OS=*Homo sapiens* GN=GH1 MATGSRTSLLLAFGLLCLPWLQEGSAFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDSNVYDLLKDLEEGIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDDALLKNYGLLYCFRKDMDKVETFLRIVQCRSVEGSCGF.

2.2. Orthologs algorithm

Orthologs are homologs separated by speciation events. OrthoMCL DB [16,17] is a database of groups of orthologous protein sequences. OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. So it serves as