Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

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

Protein–protein interaction predictions using text mining methods

Nikolas Papanikolaou, Georgios A. Pavlopoulos, Theodosios Theodosiou, Ioannis Iliopoulos *

Division of Basic Sciences, School of Medicine, University of Crete, Heraklion 71003, Greece

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form 5 September 2014
Accepted 21 October 2014
Available online xxxx

Keywords:
Protein–protein interaction prediction
Text mining
Computational tools

ABSTRACT

It is beyond any doubt that proteins and their interactions play an essential role in most complex biological processes. The understanding of their function individually, but also in the form of protein complexes is of a great importance. Nowadays, despite the plethora of various high-throughput experimental approaches for detecting protein–protein interactions, many computational methods aiming to predict new interactions have appeared and gained interest. In this review, we focus on text-mining based computational methodologies, aiming to extract information for proteins and their interactions from public repositories such as literature and various biological databases. We discuss their strengths, their weaknesses and how they complement existing experimental techniques by simultaneously commenting on the biological databases which hold such information and the benchmark datasets that can be used for evaluating new tools.

1. Introduction

Proteins are the molecules that facilitate most biological processes in a cell. While most of the known proteins are characterized by a unique function, many of them act in coordination with others towards the formation of protein networks in order to deliver complex actions. Two proteins, for example, may directly interact through their physical proximity or by being members of the same protein complex [1]. At a systems biology level, the correct identification of Protein–Protein Interactions (PPIs) is of key importance for the understanding of the complex mechanisms in a cell. Such processes include cell cycle control, differentiation, protein folding, signal transduction, transcription, translation, post-translational modification and transportation.

Today, in order to better understand such systems, relatively new high-throughput methods are used to reveal protein interaction networks [2,3]. *Yeast two-hybrid system (Y2H)* or *two-hybrid screening*, for example, is being used for more than twenty years, mainly aiming to detect binary interactions [4,5] whereas other experimental methods for PPI identification are the *protein microarrays* [6] (including reverse phase protein arrays [7]), *pull down assays* [8], tandem affinity purification (TAP) [9], *immunoaffinity chromatography* (affinity-purification) in conjunction with mass spectrometry [10], *dual polarization interferometry (DPI)* [11], *microscale thermophoresis* [12], *phage display* [13,14] and *protein complex immunoprecipitation (Co-IP)* [15]. In addition, some other methods take advantage of *X-ray crystallography* [16] and *Nuclear Magnetic Resonance (NMR) spectroscopy* [17]. While most of the aforementioned high-throughput techniques have proven to be very valuable in instigating a huge growth of experimentally verified PPIs [18], they come with several shortcomings, as findings are often fractional or not conclusive, and accompanied by high false positive and false negative rates [19]. In addition, most of the experiments can often become quite costly and time consuming [20]. Therefore, algorithmic PPI predictions have become a necessity as they can provide strong indications and clues about putative PPIs and thus help steering the experimental verification to the right direction.

Non Text-mining prediction methods for PPIs can vary widely depending on the strategy they follow to infer putative interactions. Accordingly, those methods can be categorized depending on whether prediction is based on protein sequence, protein structure, genomic context, homology, experimental profiles and literature-derived associations [21]. In the case of sequences, prediction tools use artificial intelligence and machine learning approaches [22,23] to predict protein interactions through their sequence or structural characteristics [24] such as shared binding partners [25], domains [26,27] or neighboring residues [28]. Homology based prediction tools try to detect evolutionary relationships between the proteins, taking into account their structures or sequences as many known protein interactions are conserved across species [29]. The previous methods are often combined to complement each other in order to provide additional physical details about the interactions as more and more structures become available overtime [30]. According to the first subcategory of

* Corresponding author.
E-mail address: iliopj@med.uoc.gr (I. Iliopoulos).

http://dx.doi.org/10.1016/j.ymeth.2014.10.026
1046-2023/

genomic context based prediction tools, the assumption which is made, is that two proteins interact with each other according to their conservation of relative genomic locations of genes [31,32]. Alternatively, others examine gene fusion events [33] as an implication that respective fused proteins are functionally related, something that in many cases has been experimentally verified [34]. Lastly, many genomic context based prediction tools use phylogenetic profiling and base their functionality on the hypothesis that proteins involved in common pathways co-evolve in a correlated fashion across large numbers of species [35,36].

Text-mining based techniques on the other hand, try to automate the extraction of interconnected proteins through their coexistence in sentences, abstracts or paragraphs within text corpuses. This can be done by searching for statistically significant co-occurrences between gene names [37] in public repositories and online resources. Such approaches are very promising as they significantly expand the available proteome coverage, something that is currently done partially by the existing experimental approaches [38,39]. More complex Text Mining (TM) methodologies use advanced dictionaries and generate networks by *Natural Language Processing (NLP)* of text, considering gene names as nodes and verbs as edges giving a semantic notion on the graphs. Notably, even newer developments use kernel methods to predict protein interactions from literature [40,41].

While the available tools follow different concepts for predicting PPIs, a combination of the aforementioned methods along with meta-methods that combine the results of the presented tools is preferable [42–44]. This review is focused on PPI extraction through Text Mining methods as they gain importance in a large array of biological fields [45,46]. We mention the advantages and the disadvantages of the available tools of the past decade and we comment on how they perform information extraction, protein entity recognition and linking from various types of textual collections, such as Medline abstracts or other biological databases that contain textual information. We believe that this review can be a fruitful guide for researchers in the field.

2. Text mining tools

In this section, we review tools and databases according to the following criteria: first, we select tools or databases that offer, among other functionalities, PPI predictions based on Text Mining methods. This entails that publications that only describe methods or have applied an ad hoc PPI prediction approach are not included [47–50]. Furthermore, databases like *PIPS* [44] and *STITCH* [51,52], which contain PPI predictions derived from non-Text Mining methods, are also not included. Widely used systems like *DIP* [53] and *MINT* [54], which only contain manually curated data, are shortly described below. Second, we only focus on tools which come with a functional web or a standalone interface. As a result, tools like *SUISEKI* [55], which is one of the first approaches in the field or *AkaneRE* [56], are not reviewed. Similarly, databases such as *BOND* (formerly known as *BIND* [57]), are not included in the review. Lastly, the review focuses on tools that are freely available and not accompanied by a payment scheme. Therefore, systems like *MedScan* [58] are not included in the review. Following the aforementioned criteria, we have located nineteen tools. We briefly describe each approach in the following paragraphs and also present tables containing URLs, technical features, key characteristics and quality measures for each system (see Table 1, Supplementary Tables 1 and 2).

BioRAT (Biological Research Assistant for Text mining) [59] is a standalone application. Given a typical *PubMed* query by the user, *BioRAT* attempts to locate and download full papers or, if not possible, abstracts, starting from *PubMed* and following links, jumping from and to web pages. Informative terms, such as proteins and genes, are then highlighted in the collected corpus. *BioRAT* implements a general purpose Information Extraction (IE) system, the *GATE* toolbox [60]. It tries to identify bioentities such as proteins, even when their names resemble common English words using a 'part-of-speech' tagger and dictionaries (called 'gazetteers'). Following that, it extracts information using predefined or user-defined semantic templates such as *'interaction of' (PROTEIN_1) 'and' (PROTEIN_2)*. PPIs are presented in a table along with the textual information (sentences) that lead to their identification. *BioRAT* was evaluated using *DIP* [53] subsets.

eFIP (Extracting Functional Impact of Phosphorylation) [61] is a Text Mining system focused on mining protein interaction networks of phosphorylated proteins. It employs several NLP techniques in order to locate abstracts that mention protein phosphorylation alongside with indicators of PPIs and evidence of altering effects of said phosphorylation on the PPI. To that purpose, it integrates previously developed tools by the authors,

Table 1
Text mining-based PPI prediction tools.

Name	Type	Non-TM	NER	NLP	Co-occurrence	Meta	Dictionaries/ ontologies	Corpus	Results	Scoring/ ranking scheme	Benchmarking/ evaluation
BioRAT	Standalone	×	✓	✓	✓	×	✓	PubMed abstracts/full-text	XML, table	×	✓
eFIP	Online tool	×	✓	✓	✓	×	×	PubMed abstracts	Table	✓	✓
FACTA+	Online tool	×	✓	×	✓	×	✓	PubMed abstracts	Table	✓	✓
GeneWays	Online tool	×	✓	✓	×	×	✓	Full text articles	Table	✓	✓
HitPredict	Online DB	✓	×	×	×	×	✓	–	Table/graph	✓	×
hPRINT	Online DB	✓	✓	✓	✓	✓	✓	Corpus used by used DBs	Table	✓	✓
I2D	Online DB	✓	×	×	✓	✓	✓	–	Table/graph	✓	×
iHOP	Online DB	✓	✓	✓	✓	×	✓	PubMed abstracts	List of sentences	✓	✓
IMID	Online DB	×	×	✓	✓	✓	✓	Not specified	Table/graph	✓	×
Negatome	Online DB	✓	✓	✓	×	×	✓	PubMed abstracts/full-text	Downloadable list	✓	✓
openDMAP	Standalone	×	✓	✓	×	×	×	Any biomedical corpus	List	×	✓
PCorral	Online tool	×	✓	✓	✓	×	×	PubMed abstracts	Table	✓	×
PIE the search	Online tool	×	✓	✓	✓	×	×	PubMed abstracts	List	✓	✓
Polysearch	Online tool	×	✓	×	✓	×	✓	Many DBs	List	✓	✓
PPIExtractor	Standalone	✓	✓	×	✓	×	×	PubMed abstracts	Graph	✓	✓
PPI Finder	Online tool	×	✓	✓	✓	×	✓	PubMed abstracts	List	×	✓
PPInterFinder	Online tool	×	✓	✓	✓	×	✓	PubMed abstracts, integrates data from DBs	Table	×	✓
PPLook	Standalone	×	✓	✓	✓	×	✓	Any biomedical corpus	Graph	×	✓
STRING	Online DB	✓	✓	✓	✓	✓	✓	PubMed full-text	Graph/table	✓	✓

Please cite this article in press as: N. Papanikolaou et al., Methods (2014), http://dx.doi.org/10.1016/j.ymeth.2014.10.026