MAPi DOCTORAL PROGRAM IN COMPUTER SCIENCE

Thesis Proposal

Title:

Computational prediction of Inter-Species Protein-Protein Interactions

2012/2013

Candidate

Edgar Duarte Coelho (eduarte@ua.pt)

Supervisors

Joel Arrais (jpa@ua.pt)

José Luís Oliveira (jlo@ua.pt)

Research Unit

Institute of Electronics and Telematics Engineering of Aveiro

Universidade de Aveiro

Background

The assembly of all PPIs from one organism is called interactome, which can be modeled as an undirected graph, where the nodes represent proteins and the edges represent physical interactions. The representation and analysis of those graphs will allow the clarification of the pathogenesis mechanisms of various diseases and provide support in the establishment of potential means to discover new therapeutic agents, diagnosis and screening tools (1, 2).

The experimental techniques designed for PPI determination range from low throughput with high accuracy, to high-throughput with significantly lower accuracy. Although these approaches have successfully identified vast numbers of PPIs, a previous study (3) reports that approximately 70% of the identified interactions are false-positives. Moreover, the price and time required to perform experimental analysis of a species interactome makes it unfeasible.

To overcome these drawbacks, a number of computational approaches to predict PPI have been explored. One common approach consists in using text mining to extract known PPIs from the biomedical literature (4). Bock and Gough (5) applied Support Vector Machines (SVM) to predicted PPI based only on the protein sequence. Nanni *et al*. (6) added the multiple classifier system (MCS) to improve the SVM-based methods, while Guo *et al*. (7) adopted autocovariance (AC) transformation to consider the neighborhood effect. Rajasekaran *et al*. (8) employed minimotifs to improve sequence-based prediction results. To predict PPI using genomic data, Najafabadi and Salavati (9) applied a näive Bayesian network, while Lu *et al*. (10) focused on protein structures, proposing a threading algorithm.

As some PPIs occur specifically between domains of their interacting counterparts, Sprinzak and Margalit (11) applied a method to predict PPI based on domain-domain interactions (DDI), using a maximum likelihood estimator (MLE). Chen and Liu (12), improved the DDI-based PPI prediction results using a MCS. Maetschke *et al*. (13) applied Gene Ontology (GO) (14) annotation to predict PPI, combining machine learning with semantic similarity measures.

Despite giving undeniable insights about PPI networks, the aforementioned methods only predict intra-species PPI. Dyer *et al*. (15) employed a DDI-based approach, adapting the MLE algorithm (11), while the approach of Davis *et al*. (16) was based in the threading method (10), requiring resolved protein structures for comparative modeling. Lastly, a method combining multiple data sources was developed by Tastan et al. (17), which used random forest as MCS. (15-19). Nevertheless, these techniques are not successful enough in inter-species PPI prediction.

Objectives

The main goal of this project is to fully comprehend and elucidate the oral inter-species PPI network. In order to achieve this aim, the following objectives have been outlined:

1. Review of the state of the art
2. Develop a PPI prediction model
3. Design and reconstruct a complete PPI network of the oral cavity
4. Explore and analyze the PPI network using network visualization tools

5. Evaluate the performance of the prediction model under real scenarios

6. Construct a bioinformatics platform to access PPI data

Thesis planning

The biological processes occurring within a cell are performed, directly or indirectly, via a pathway of interacting proteins. For instance, PPIs play key roles in signaling and metabolic pathways, cellular structure scaffolding, and protein transport. Also, while being able to provide a comprehensive view of the interaction structure of an organism's proteome, PPIs can also offer data regarding specific interactions. This is particularly interesting, since it is well accepted that the disruption of PPI interfaces may result in the development of various disease states.

Even though there have been several efforts to predict PPI, none of them provides a solid and reliable prediction model. Such occurs due to a number of factors that influence the accuracy of the approaches. For instance, the absence of gold standard datasets to be used as training data and to evaluate system performance, the absence of literature compiling non-interacting protein pairs, the discontinuation of tools and databases, and the lack of complete genomic, proteomic and structural annotations for several organisms. Nevertheless, experimental methods cannot be considered a viable option, as they are very expensive and time-consuming, only give insight about a fraction of the total PPI network, and the most commonly used techniques retrieve a great number of false-positives, with the possibility of overlooking true-positives (3). Conversely, computational approaches grant a rapid and low-cost alternative to experimental methods.

The aim of this proposal is to overcome such obstacles in a way that allows accurate prediction of inter-species PPI, particularly within the oral cavity. The most significant methods developed in regard of this issue have been described in the state of the art. Nonetheless, it is clear that each method has its limitations. Therefore, we believe that the key to solve this issue involves the combination of multiple complementary approaches.

The approach to be explored consists in an ensemble of machine learning algorithms, joined with a multiple classifier system, as it is proven to be more accurate than an excellent single classifier. These features will then be assigned to an original prediction model, which combines DDI and protein sequence analysis, being the latter linked to an autocorrelation descriptor in order to take the amino acid neighborhood effect into account. Analysis of semantic similarities using GO will also be employed in such combination. Once the oral PPI network is established, it will be explored using network exploration tools (e.g. Cytoscape – available at http://www.cytoscape.org/). In the absence of a gold standard, we believe that the use of experimentally determined PPI data will suffice to evaluate the method. To sum up, we will develop a web tool where all the data will be readily available to the biomedical community. Below, we present an outline describing in more detail the highlighted milestones:

Prediction model of the human oral interactome

The first task consists in developing and tuning of the PPI prediction model. Our aim is to combine a few complementary methods in order to diminish limitations of the individual approaches. One of these approaches will consist in the analysis of DDIs, since there is a great

number of PPIs that occurs exclusively via specific domains pairs. The analysis of protein sequence making use of machine learning algorithms returned great results. Also, when in combination with a descriptor, such as Moran's autocorrelation descriptor, the prediction model will be able to take into account the amino acid neighboring effects and correlate two protein sequences regarding their physicochemical properties. The emerging tuned PPI prediction based on GO semantic similarity is also a very solid approach. For that matter we intend to use inducers, combined with a MCS. We believe this potential triad will yield highly significant results.

Network reconstruction, validation and analysis

The developed PPI prediction model will then be used to obtain the crude oral PPI network. This network will then be reconstructed by pruning edges based on node degree, a technique less computer intensive than testing critical edges. As for validation, we intend to use cross-validation to assess data consistency and Receiver Operating Characteristic (ROC) curves to assess performance. ROC curves are especially helpful as they are sensitive to node degree effects. Analysis of the PPI networks will be performed using Cytoscape.

Performance assessment and experimental validation

The absence of a gold standard encumbers the evaluation of the system performance. Therefore, it is crucial to build datasets consisting of high-quality curated PPI data from scientific literature. These data are found in online repositories, such as IntAct (http://www.ebi.ac.uk/intact/), MIPS (http://mips.helmholtz-muenchen.de/proj/ppi/), HPRD (http://www.hprd.org/), and STRING (http://string-db.org/). An unbiased negative dataset is required as well. The use of experimentally determined PPI data combined with the unbiased dataset will allow to promptly estimate performance measures, suchlike precision, recall, specificity, and accuracy.

Development of a web-based bioinformatics tool

Based on the results of the previous tasks, we plan to develop a web tool where the users can retrieve specific data from the whole oral PPI network. We will adopt a complex search system based in both UniProt (http://www.uniprot.org/) accession numbers and their synonyms. The end-user will be able to search by (1) single or multiple proteins, which retrieves all PPIs where the given protein is involved and the respective place inside the network, (2) multiple proteins, which retrieves the list of PPIs occurring in the given protein set, and by (3) GO annotation, yielding all the proteins in the same GO term and the respective PPIs.

Final statement

This proposal presents an innovative approach in PPI, with potential to improve the current state-of-the-art. The data resulting from this work will provide undeniable support for the scientific community, namely in the research of oral host-pathogen interactions, pathogenesis of oral ailments, drug *de novo* synthesis, and drug reuse for new targets. Also, it will enable the development of original diagnosis and screening tools.

References

1. Fechete R, Heinzel A, Perco P, Mönks K, Söllner J, Stelzer G, et al. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. PROTEOMICS – Clinical Applications. 2011;5(5-6):354-66.
2. Wang Y-C, Chen B-S. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC Medical Genomics. 2011;4(1):2.
3. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics. 2002;1(5):349-56. Epub 2002/07/16.
4. Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. Integrating protein-protein interactions and text mining for protein function prediction. BMC Bioinformatics. 2008;9 Suppl 8:S2. Epub 2008/08/05.
5. Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics. 2001;17(5):455-60.
6. Nanni L, Lumini A. An ensemble of K-local hyperplanes for predicting protein–protein interactions. Bioinformatics. 2006;22(10):1207-10.
7. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Research. 2008;36(9):3025-30.
8. Rajasekaran S, Merlin JC, Kundeti V, Mi T, Oommen A, Vyas J, et al. A computational tool for identifying minimotifs in protein-protein interactions and improving the accuracy of minimotif predictions. Proteins. 2011;79(1):153-64. Epub 2010/10/13.
9. Najafabadi HS, Salavati R. Sequence-based prediction of protein-protein interactions by means of codon usage. Genome Biol. 2008;9(5):R87. Epub 2008/05/27.
10. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins. 2002;49(3):350-64. Epub 2002/10/03.
11. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol. 2001;311(4):681-92. Epub 2001/08/24.
12. Chen X-W, Liu M. Prediction of protein–protein interactions using random decision forest framework. Bioinformatics. 2005;21(24):4394-400.
13. Maetschke SR, Simonsen M, Davis MJ, Ragan MA. Gene Ontology-driven inference of protein–protein interactions using inducers. Bioinformatics. 2012;28(1):69-75.
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9. Epub 2000/05/10.
15. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein–protein interactions. Bioinformatics. 2007;23(13):i159-i66.
16. Davis FP, Barkan DT, Eswar N, McKerrow JH, Sali A. Host pathogen protein interactions predicted by comparative modeling. Protein science : a publication of the Protein Society. 2007;16(12):2585-96. Epub 2007/10/30.
17. Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J. Prediction of interactions between HIV-1 and human proteins by information integration. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2009:516-27. Epub 2009/02/13.
18. Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, Kao CY, et al. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC Bioinformatics. 2008;9 Suppl 12:S11. Epub 2009/01/06.
19. Itzhaki Z. Domain-domain interactions underlying herpesvirus-human protein-protein interaction networks. PloS one. 2011;6(7):e21724. Epub 2011/07/16.