Computational Approach to Predict Inter-Species Oral Protein-Protein Interactions

Edgar D. Coelho[1], Joel P. Arrais[2], Sérgio Matos[1], Nuno Rosa[3], Maria José Correia[3], Marlene Barros[3], José Luís Oliveira[1]

[1] Department of Electronics, Telecommunications and Informatics (DETI), Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, Portugal
{eduarte, aleixomatos, jlo}@ua.pt
[2] Department of Informatics Engineering (DEI), Centre for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, Portugal
jpa@dei.uc.pt
[3] Department of Health Sciences, Portuguese Catholic University, Viseu, Portugal
{nrosa, mcorreia, mbarros}@crb.ucp.pt

Abstract. The majority of gene products that crowd a living cell interact, at least transiently, with other protein molecules. Concordantly, virtually all cellular events are mediated by protein-protein interactions (PPIs). The same applies to host-pathogen systems, where PPIs are essential in host colonization and infection. Some authors believe that the understanding of the human interactome will provide insight into disease development mechanisms.

Numerous experimental techniques were explored to attain the human interactome, suchlike two-hybrid screening, affinity purification mass spectrometry, DNA microarrays, protein microarrays, synthetic lethality, phage display, X-ray crystallography and nuclear magnetic resonance spectroscopy, fluorescence resonance energy transfer, surface plasmon resonance, atomic force microscopy, and electron microscopy. However, these methods possess several limitations that reduce their applicability potential in large-scale PPI prediction, as the associated time required and cost, and minimal protein interaction network coverage per run. High-throughput approaches are also often associated with low-specificity and great numbers of both false negatives and false positives. Computational approaches were the appointed alternatives for the prediction of intra-species PPIs. These methods can be categorized regarding the types of information they analyze: data mining of biomedical literature, methods based on genomic data (gene neighborhood, gene fusion, phylogenetic profiles, codon usage similarity), on protein structure (homology-based method, threading-based method), on domain information (single domain pairs, multi-domain pairs), on protein sequence, and on Gene Ontology (GO) annotation semantic similarity. In contrast, computational efforts to predict inter-species PPIs have been very limited.

We propose a computational model to predict inter-species PPIs within the oral cavity, an environment particularly prone to bacterial colonization. Rosa *et al.* suggests that the determination of the salivary interactome will clarify the role of saliva in oral biology and enable the identification of disease biomarkers. They also suggest that the presence of exfoliated epithelial cells in saliva may provide a means for diagnosis of conditions currently requiring more invasive diagnostic techniques. We defined the positive and negative datasets, and thoroughly selected the most discriminative features (concept profile similarity,

orthologous profiles, biological process, and enriched conserved domain pairs) required for the naïve Bayes classifier. Subsequently we conducted a series of tests to evaluate the performance of the proposed method and tested our approach on several oral microorganisms and human data sets. The performance of the method was validated analyzing specific network interactions in Cytoscape. We calculated the pre-test odds, likelihood ratios and respective post-test odds (PTOs) for each feature. The cumulative post-test odds (CPTOs) were also calculated to assess the discriminatory behavior of the feature through the data. When applied to the test data, our method returned 6.860.53 PPIs, of which 945.964 were considered positive. The performance of the method was evaluated by calculating the area under the receiver operating characteristic (ROC) curve (AUC), which depicts the relative tradeoffs between the true positive rate and the false positive rate. The final AUC was estimated to be 0.82.

We believe our work may be applied in several scientific areas, and even in other PPI related studies. An example is biomedical PPI screening, to assess if interactions of particular interest might occur and what is the related interaction probability. Another example is pharmacologic research, as a well-established PPI network can provide insights on potential drug targets, but also new uses for already in-market drugs. Finally, and based on the fact that the protein interaction networks are not static but dynamic, our work can support protein interaction network evolution researchers in identifying evolutionary patterns.