On the Automated Function Prediction of Big Multi-Species Networks

Matteo Re, Marco Mesiti, and Giorgio Valentini

AnacletoLab, Dipartimento di Informatica, Università degli Studi di Milano, via Comelico 39/41 - 20135 Milano, Italy

Abstract: Automated Function Prediction (AFP) of proteins in Multi-species Networks raises challenging computational problems due to the size of the generated network and the lack of scalability of traditional approaches. We present a network-based approach that allows to predict protein functions in a multi-species setting by exploiting homology relationships between species. The method adopts secondary memory-based technologies to efficiently process huge protein networks using ordinary stand-alone machines.

1. INTRODUCTION

As highlighted by a recent international challenge for the critical assessment of AFP [1], scalability and heterogeneity of the available data represent two of the main issues for the analysis of multi-species networks. Indeed on the one hand no single experimental method can fully characterize the multiplicity of the protein functions, and on the other hand the huge amount of data to be processed poses serious computational problems. The complexity of the problem is furthermore exacerbated by the different level of the functional annotation coverage in different organisms, thus making very difficult the effective transfer of the available functional knowledge from one organism to another.

2. METHODS

To face these issues, we propose a novel framework for scalable network-based learning of multi-species protein functions. In the construction of the multi-species network different types of homology between proteins, not limited to sequence similarity relationships, are exploited to transfer functional information from well annotated species to poorly annotated ones. Then, through the use of novel algorithmic approaches [2,3] and the adoption of innovative secondary memory-based technologies [4,5] we are able to provide an efficient and scalable approach for multi-species AFP. Secondary memory-based technologies allow the efficient use of the large memory available on disks, thus overcoming the main memory limitations of modern off-the-shelf computers. This approach has been applied to the analysis of a large multi-species network including more than 300 species of Bacteria and to a network with more than 200,000 proteins belonging to 13 Eukaryotic model organisms.

3. RESULTS

By exploiting orthology relationships across interspecies proteins, we constructed a multispecies network of Bacteria and we compared the AFP prediction performances for a set of GO BP terms on the integrated network with those obtained from single-species networks. Results show that the multi-species approach achieves better results in terms of both AUC, and precision at a fixed recall rate and the difference is statistically significant independently of the performance measure considered. Moreover, our approach provides a solution to the challenging main memory requirements induced by large multi-species protein networks, thus allowing the analysis of big networks using off-the-shelf machines. Indeed the AFP task for the network of 13 Eukaryotic species cannot be conducted with standard RAM-based methods, also using relatively well equipped machines, whereas our implementation requires no more than 15 seconds per GO-term on a simple notebook. Our results show that both graph DB technologies (i.e. Neo4j [5]) and secondary memory based systems for graph processing (i.e. GraphChi [4]) can be successfully applied to the analysis of large multi-species networks.

REFERENCES

1. Radivojac P, et al.. A large-scale evaluation of computational protein function prediction. *Nature Methods* . 10(3):221—227, 2013.

2. M. Re, M. Mesiti and G. Valentini, A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks, *IEEE/ACM Trans. Comput. Biology Bioinform.* 9(6) pp. 1812-1818, 2012

3. M. Mesiti, M. Re, and G. Valentini Think globally and solve locally: secondary memory-based network learning for automated multi-species function prediction, *GigaScience*, 3:5, 2014

4.Kyrola, A., Blelloch, G., Guestrin, C.: Graphchi: large-scale graph computation on just a pc. In: *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation*. OSDI'12, pp. 31–46, 2012

5.Webber, J.: A programmatic introduction to Neo4j. In: *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, pp. 217–218. ACM, Tucson, Arizona, USA, 2012