Improving Discovery in the Life Sciences using Linked Data and Semantic Web Technologies

Helena F. Deus

NSF Discovery Informatics Workshop

Abstract: Biology is becoming an information science – as such, success in the understanding and unravelling of biology will increasingly rely on the ability to use, manipulate, analyse and integrate data. This means that the next generation of biologists will be performing their experiments *in silico*, testing hypothesis through simulated studies before testing them in the lab. A key requirement for making this possible is the ability to integrate the several layers of the biological continuum.

We live embedded in a world of data. Biology itself is becoming an information science - massive amounts of experimental data are being produced at unprecedented scales, and new technologies such as next generation sequencing are creating new avenues for increasingly precise data gathering and knowledge discovery. At the same time, we are beginning to realize that biology is much more complex and redundant than initially tough: most biologists working today have been trained at looking at single systems individually. To understand, predict and eventually control living cells, it is necessary to also understand how different systems or cellular networks interact. Rather than single pathways or gene based models, it is increasingly necessary to model biology as a continuum of networks of interactions between entities such as genes, proteins and small molecules such as drugs.

Modelling and understanding the topology of these networks will require the integration of data, knowledge and expertise from multiple "omics" areas in biology, but it will also benefit stakeholders in multiple areas by creating a single framework where biology can be modelled and understood as a whole. We know, for example, that the signalling pathways controlling cell growth and differentiation are almost always altered in cancer cells [1]. The accurate prediction and modelling of growth and differentiation pathways can be used to create a baseline to aid in the prediction of what is altered in tumorous cells and how it can be targeted or reversed.

The integration of results from high-throughput experiments is particularly important as they enable probing into the cells by looking at entire genetic profiles in a single instance. However, there are two main obstacles to making full use of this data for predictive modelling: 1) experimental data typically covers only one aspect of the biological continuum - they are either genomics based (sequences or mutations of genes), transcriptomic based (expression and regulation of genes), or proteomics based (expression of proteins); and 2) The methods used to acquire high-throughput experimental evidence have still not been perfected to a point where they can be immediately trusted and useful, often requiring extensive pre-processing, background subtraction, normalization or sequence assembly.

Semantic Web and Linked Data technologies can be used to address these challenges. By relying on semantic matching, inference and federated queries, it is possible to integrate results from multiple types of "omics" experiments and annotate them to the same patient or the same position in the genome [2]. These technologies can also be used to create clusters or layers of biological entities with biological significance such as, for example, genes involved in cell differentiation (Figure 1). The time dependency between these two processes needs also to be taken into consideration: even

when both gene expression and protein expression are measured in the exact instance of cell differentiation, there may be a time delay between the observation that a gene is highly regulated and its effect at the protein level. To address challenge 2), provenance of both raw and processed data is becoming increasingly relevant. Different experimental methods can be used to collect information about, for example, protein-protein interactions. The more reliable methods provide high accuracy but low number of interactions. Other methods such as mass spectrometry can be used for screening a large number of interactions, but with lower accuracy. As such, biologists need to be able to decide how much they trust the data based on how it was acquired and what analytical process was used to obtain it [3].

Finally, it is becoming increasingly clear that the identification of these "layers" or abstractions necessary for data discovery must be identified and assembled by the domain experts themselves [4,5].

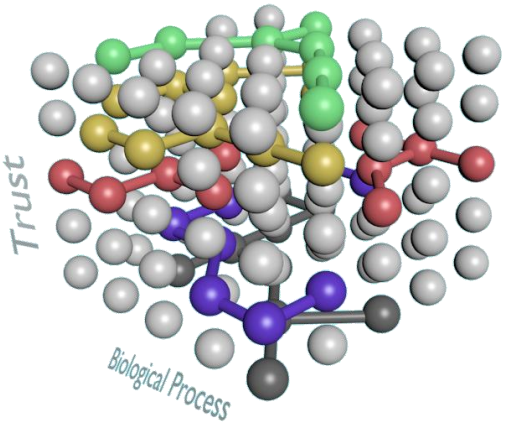

Figure 1. A layered view of biological data: data with different levels of trust can be organized as different layers which, once overlapped, provide an integrated view that can be used to discover new biological interactions. Biologists can chose to eliminate layers with less trust. A second axis can also be defined that separates biological entities such as proteins according to the biological processes where they are involved (e.g. cell differentiation).

References

[1] B.N. Kholodenko, J.F. Hancock, W. Kolch, Nature Reviews Molecular Cell Biology 11 (2010) 414-426.

[2] H.F. Deus, D.F. Veiga, P.R. Freire, J.N. Weinstein, G.B. Mills, J.S. Almeida, Journal of Biomedical Informatics 43 (2010) 998-1008.

[3] H.F. Deus, E. Prud, J. Zhao, M.S. Marshall, M. Samwald, in:, ISWC 2010 SWPM, 2010.

[4] J.S. Almeida, C. Chen, R. Gorlitsky, R. Stanislaus, M. Aires-de-Sousa, P. Eleutério, J. Carriço, A. Maretzek, A. Bohn, A. Chang, F. Zhang, R. Mitra, G.B. Mills, X. Wang, H.F. Deus, Nature Biotechnology 24 (2006) 1070-1.

[5] H.F. Deus, R. Stanislaus, D.F. Veiga, C. Behrens, I.I. Wistuba, J.D. Minna, H.R. Garner, S.G. Swisher, J.A. Roth, A.M. Correa, B. Broom, K. Coombes, A. Chang, L.H. Vogel, J.S. Almeida, PloS One 3 (2008) e2946.