Uncovering subcellular regulatory networks

Sumeet Agarwal

1 Origins

The post-genomic era has produced masses of data on subcellular biological systems: gene expression microarrays, protein-protein interactions, and metabolic pathways. A key challenge is to leverage this data to gain functional understanding of the underlying systems and mechanisms. Network biology [1, 2] seeks to do this by using the mathematical abstraction of a graph to represent systems comprising many interacting components. One question relevant to the study of these networks is whether we can identify typical structural 'signatures'; such signatures can guide the inference of networks from data. For instance, many methods exist for reconstructing Gene Regulatory Networks (GRNs) from microarray data [3–9]. However, in general it has proved difficult to sufficiently constrain such network-based models with the data available [10]. One way to address this is to consider that biological networks may have typical structural features which can be used to guide model search. I would like to investigate this possibility by examining the structure of such networks from multiple different perspectives, and attempting to detect patterns of interest in an automated, data-driven fashion.

Another question is whether we can develop richer models, by integrating interactions at multiple levels. Our picture of gene regulation has become increasingly complex, with a variety of novel data on the role of non-coding RNA and RNA interference, protein-DNA interactions, 3-D chromosome structure etc. GRNs, protein-protein interactions, metabolic pathways, protein-DNA interactions, and microRNA-target interactions all represent aspects of a single system, with myriad regulation and information flows between them [11], but there have been only a few restricted efforts to put the pieces together [12–14]. Here I propose to work towards developing a framework which allows for modelling multiple aspects of the cellular machinery at different levels, attempting to make use of a representation language based on first-order logic as described in Section 3.

2 Objectives

During this fellowship, I would like to develop theoretical approaches to address key questions in the study of biological regulatory networks. Here I summarise my objectives; in the following section I outline the methods I propose to use towards achieving these.

- Obtain 'structural signatures' of biological networks
- Develop approach to identify regulatory network models with constraints and background knowledge incorporated; first use model organism data and then attempt for human cells and other unexplored species
- Progressively extend models to include signalling, metabolic, and other sorts of relevant interaction data; focus initially on yeast as the best-studied organism, then work on others as per data availability

- Use models to predict cellular responses which can be tested in the lab; use these to refine models and also to suggest desirable interventions like drug targets

3 Methodology

3.1 Data Sources

A key task will be to obtain multiple kinds of biological data relevant to understanding subcellular networks. This will include microarray expression data [15], genetic and protein-protein interaction data [16] pathway data for metabolism and other cellular processes [17], and data on microRNA sequences [18] and targets [19]. I would also like to look at some relatively understudied data sets, such as ChIP-on-chip for protein-DNA interactions [20–22], quantitative protein expression [23], and DNA accessibility in chromatin [24] and genome proximity networks [25,26].

I intend to use progressively data from 3 kinds of organisms. Firstly, model organisms (*Saccharomyces cerevisiae* and *Caenorhabditis elegans*), for which a large quantity and variety of experimental data is available, will be used to build, verify, and refine our initial mathematical models. Secondly, humans, where substantial data sets are available for at least some cell types; these can serve as test beds on which to generate falsifiable predictions, such as transcriptional response to novel perturbations. Thirdly, I would like to look at the extent to which our approach can serve to advance understanding of a relatively unstudied organism; one possibility in this regard is the parasitic protozoan *Entamoeba histolytica*. Experimental collaborators at the JNU plan to obtain data on proteins involved in the calcium signalling network and the corresponding transcriptional response of this amoeba, which play a key role in phagocytosis and other cellular processes [27]. There will be an ensuing experimental effort to map the calcium signalling network of *E. histolytica*, along with corresponding transcriptomic profiles in the presence and absence of calcium. This data can be used to model calcium response and signalling at both the transcriptome and protein interaction levels.

3.2 High-throughput Network Analysis

During my PhD, I have developed software for computing a variety of network properties [28]. This allows us to obtain a detailed network 'signature', in the form of a vector of real numbers, or a *feature vector*: $[f_1, f_2, , ..., f_N]$. Here f_1 could be the average number of links each node has, f_2 the average path length between node pairs, and so on. We have shown how this can facilitate the identification of specific structural characteristics of a particular kind of network, such as fungal or metabolic. I will seek to obtain signatures of regulatory and physical interaction networks, which can then guide network reconstruction. The following sections outline the techniques I hope to use for this purpose.

3.3 Bayesian Inference for Networks

Bayesian inference [29] is a powerful methodology for refining knowledge on the basis of observation. For instance, suppose that we are seeking to evaluate the likelihood of a model (e.g., a GRN) given some data on the behaviour of the system (e.g., expression data). Then, using Bayes' rule, this can be expressed as

$$P(M|D) \propto P(D|M)P(M). \tag{1}$$

Here $P(M|D)$ is the *posterior* probability of model M, given that we have observed data D. Bayes' rule says that it is given by the product of $P(D|M)$, the *likelihood* of observing the data D if we assume M to be the underlying model, and $L(M)$, the *prior* probability of M being the appropriate model, before any data has been considered. The ability to make explicit one's beliefs about the system being modelled, via the prior, is a major attractive feature of Bayesian inference.

For modelling subcellular networks, I propose to use this approach to incorporate prior knowledge about what structures are likely, based on pre-deciphered networks. In the context of network evolution, I have developed an approximate Bayesian computation (ABC) algorithm for this purpose during my PhD. ABC [30–32] relies on the fact that whilst the likelihood $P(D|M)$ may be hard to evaluate for complex models, given a probabilistic model M it is usually easy to generate samples from the distribution specified by that model. By looking at what proportion of these samples are 'close' to the data, one can estimate the likelihood the of the model generating a given data set. I will attempt a similar approach for inferring regulatory network structure, using previously mapped networks as priors. However, we are interested in not just static patterns of connections but also dynamics of concentration/expression levels; for these I will use richer formal representations, as described next.

3.4 Relational Learning

We have just outlined a statistical learning approach to match models to data. A complementary framework for such machine learning tasks is provided by symbolic or relational methods, such as Inductive Logic Programming (ILP) [33]. These allow one to specify a model in terms of explicit, interpretable rules and relations (as opposed to probability distributions), expressed in a formal language such as first-order logic. For instance, if one wanted a rule disallowing open triangles, then it might read:

$$link(A, C) \wedge link(B, C) \rightarrow link(A, B). \tag{2}$$

Here A, B, C are variables denoting nodes; and $link()$ is a predicate specifying whether or not a given pair of nodes is linked. The symbol $\wedge$ denotes logical conjunction (AND), whilst $\rightarrow$ denotes implication. Thus this rule states that if two nodes have a shared interactor, they must be linked. Such approaches have also been hybridised with probabilistic models, an area broadly known as Statistical Relational Learning [34]. An example is to have rules with attached probabilities; to generate networks with 80% of triangles closed, one could assign a probability of 0.8 to Rule (2).

Recent work has sought to use such declarative approaches to reverse engineer biological networks [35–39], providing ways of directly incorporating human knowledge into the machine learning process. Advantages of ILP include the generality and interpretability of the specification language, which allows for direct incorporation of background knowledge and also uniformity of representation across multiple levels of a hierarchical system. Existing work on using ILP to construct models of biological systems has however been on a small scale and has accounted for noise or uncertainty to only a limited extent [35]. These models—such as qualitative differential equations [36] or Petri nets [37]—have also involved some coarse-graining of dynamics.

In this fellowship I will attempt to combine structural insights from our high-throughput network analysis with both relational learning (particularly ILP) and Bayesian approaches, to devise maximally effective ways of integrating data and human expertise for regulatory network inference. The major directions for development will be:

- Formulating an appropriate representation language (some subset of first-order logic) that is sufficiently rich to capture the interactions and dynamics at different levels, but also sufficiently restricts the search space over possible models to make identification feasible. Petri nets may be a powerful class of models in this respect [37, 38].
- Incorporating structural constraints as background knowledge (i.e., representing them in the given language).
- Allowing for uncertainty via probabilistic rules.
- Scaling up to using large, varied real-world data sets, via conversion to the appropriate representation language.

4 Significance

A better understanding of regulatory networks is essential for unravelling the subcellular machinery and learning how cells work. The approaches we propose to take here can help to provide a more holistic understanding of cellular circuitry than is currently available, and can contribute towards the ongoing effort of extracting meaningful information from the large quantities of experimental data being generated. Hypotheses generated using mathematical models such as those we plan to develop can serve to focus experimental efforts in fruitful directions, and contribute to the continual feedback between (and refinement of) theory and experiment, which is the hallmark of the scientific method.

Improved understanding of these systems could also have significant biomedical implications. For instance, *E. histolytica* is responsible for amoebiasis, a prevalent disease in developing countries which is estimated to cause nearly 100,000 deaths a year (making it the second deadliest protozoan parasite after *Plasmodium falciparum*) [40]. The basic biology of this organism is little-studied, and data-driven approaches can be useful in obtaining quick initial models of the underlying regulatory mechanisms, which can then generate hypotheses (e.g., regarding how the parasite causes disease and its response to drugs) that guide (and are in turn refined by) further wet lab experiments.

In summary, I believe the research proposed here will provide an opportunity to develop a more integrated understanding of how biological circuitry is organised and controlled, thus also helping us to devise more informed interventions, whether to remedy malfunctioning cells or to disable pathogenic ones.

5 Future Prospects

The lines of enquiry I suggest are part of the broader biological project of understanding life at its many different scales of organisation. One can think of organisms as comprising complex interacting systems at several levels: organ systems, organs, tissues, cells, etc. Each level builds on the one below, and cells can be seen as the most fundamental biological building blocks, the lowest level at which we see a degree of autonomous 'life'. Ultimately, this kind of effort can move us toward modelling (at a relevant level of abstraction) an entire cell, or even one of the higher-level systems, by one single formal structure, such as a network or a hypergraph [41] (providing a bird's-eye view of the relationships between different elements), with a formal mathematical model used to capture dynamics. Thus I expect there will be a lot of scope for extending the work proposed here to study additional organisms, data sets and levels of biological organisation; and this will continue to throw up novel methodological and computational challenges that will provide plenty of scientific food for thought.

References

[1] Albert-László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.

[2] Richard Bonneau. Learning biological networks: from modules to dynamics. *Nature Chemical Biology*, 4(11):658–664, October 2008.

[3] Robert J. Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS ONE*, 5(2):e9202, 02 2010.

[4] Daniel Marbach, Robert J. Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14):6286–6291, 2010.

[5] Adam A. Margolin, Kai Wang, Wei Keat K. Lim, Manjunath Kustagi, Ilya Nemenman, and Andrea Califano. Reverse engineering cellular networks. *Nature Protocols*, 1(2):662–671, June 2006.

[6] Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[7] Richard Bonneau, David J. Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S. Baliga, and Vesteinn Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, 7(5):R36, 2006.

[8] Richard Bonneau, Marc T. Facciotti, David J. Reiss, Amy K. Schmid, Min Pan, Amardeep Kaur, Vesteinn Thorsson, Paul Shannon, Michael H. Johnson, Christopher J. Bare, William Longabaugh, Madhavi Vuthoori, Kenia Whitehead, Aviv Madar, Lena Suzuki, Tetsuya Mori, Dong-Eun Chang, Jocelyne Diruggiero, Carl H. Johnson, Leroy Hood, and Nitin S. Baliga. A predictive model for transcriptional control of physiology in a free living cell. *Cell*, 131(7):1354–1365, December 2007.

[9] A Madar, A Greenfield, H Ostrer, E Vanden-Eijnden, and R Bonneau. The inferelator 2.0: A scalable framework for reconstruction of dynamic regulatory network models. In *Proceedings of the 31st Annual International Conference of the IEEE EMBS*, Minneapolis, Minnesota, 2009.

[10] Gabor Szederkenyi, Julio Banga, and Antonio Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Systems Biology*, 5(1):177, 2011.

[11] Paul Nurse. Life, logic and information. *Nature*, 454(7203):424–426, July 2008.

[12] N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symposium on Biocomputing*, pages 336–347, 2004.

[13] Jun Zhu, Bin Zhang, Erin N. Smith, Becky Drees, Rachel B. Brem, Leonid Kruglyak, Roger E. Bumgarner, and Eric E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7):854–861, July 2008.

[14] Sriram Chandrasekaran and Nathan D. Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 107(41):17845–17850, 2010.

[15] NCBI GEO (`http://www.ncbi.nlm.nih.gov/geo/`).

[16] BioGRID (`http://thebiogrid.org/`).

[17] KEGG (`http://www.genome.jp/kegg/`).

[18] miRBase (`http://www.mirbase.org/`).

[19] See `http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/`.

[20] Tong Ihn Lee, Nicola J. Rinaldi, Franois Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, Christopher T. Harbison, Craig M. Thompson, Itamar Simon, Julia Zeitlinger, Ezra G. Jennings, Heather L. Murray, D. Benjamin Gordon, Bing Ren, John J. Wyrick, Jean-Bosco Tagne, Thomas L. Volkert, Ernest Fraenkel, David K. Gifford, and Richard A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.

[21] Antti Honkela, Charles Girardot, E. Hilary Gustafson, Ya-Hsin Liu, Eileen E. M. Furlong, Neil D. Lawrence, and Magnus Rattray. Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17):7793–7798, 2010.

[22] Ahrim Youn, David J. Reiss, and Werner Stuetzle. Learning transcriptional networks from the integration of chip-chip and expression data in a non-parametric model. *Bioinformatics*, 26(15):1879–1886, 2010.

[23] See `http://www.proteinatlas.org/`.

[24] Xiao-Yong Li, Sean Thomas, Peter Sabo, Michael Eisen, John Stamatoyannopoulos, and Mark Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biology*, 12(4):R34, 2011.

[25] Bryan R. Lajoie, Nynke L. van Berkum, Amartya Sanyal, and Job Dekker. My5C: web tools for chromosome conformation capture studies. *Nature Methods*, 6(10):690–691, October 2009.

[26] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[27] Alok Bhattacharya, Narendra Padhan, Ruchi Jain, and Sudha Bhattacharya. Calcium-binding proteins of *Entamoeba histolytica*. *Archives of Medical Research*, 37(2):221–225, 2006.

[28] Sumeet Agarwal, Gabriel Villar, and Nick S. Jones. High throughput network analysis (extended abstract). In *Machine Learning in Systems Biology (MLSB), Proceedings of the Fourth International Workshop*, Edinburgh, Scotland, 2010. `http://www.physics.ox.ac.uk/systems/agarwal/mlsb10.pdf`.

[29] George E. P. Box and George C. Tiao. *Bayesian inference in statistical analysis*. Wiley, 1992.

[30] S. Tavare, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[31] J K Pritchard, M T Seielstad, A Perez-Lezaun, and M W Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798, 1999.

[32] Marin J.-M., Pierre Pudlo, Christian P. Robert, and Robin Ryder. Approximate Bayesian computational methods. January 2011.

[33] Stephen Muggleton and Luc de Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19-20(Supplement 1):629 – 679, 1994.

[34] Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[35] Jianzhong Chen, Stephen Muggleton, and José Santos. Learning probabilistic logic models from probabilistic examples. *Machine Learning*, 73:55–85, 2008.

[36] Ashwin Srinivasan and Ross D. King. Incremental identification of qualitative models of biological systems using inductive logic programming. *Journal of Machine Learning Research*, 9:1475–1533, July 2008.

[37] Ashwin Srinivasan and Michael Bain. Learning Petri net models of biological systems using ILP. In *Proceedings of the 21st International Conference on Inductive Logic Programming*, Windsor, England, 2011.

[38] Markus Durzinsky, Annegret Wagler, and Wolfgang Marwan. Reconstruction of extended petri nets from time series data and its application to signal transduction and to gene regulatory networks. *BMC Systems Biology*, 5(1):113, 2011.

[39] Max Ostrowski, Torsten Schaub, Markus Durzinsky, Wolfgang Marwan, and Annegret Wagler. Automatic network reconstruction using ASP. `arXiv:1107.5671`, July 2011.

[40] World Health Organisation. Amoebiasis. *Weekly Epidemiological Record*, 72(14):97–9, April 1997.

[41] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 05 2009.