

Topic

The proposed research will formulate novel models for evolution of regulatory networks, with emphasis on the regulation of the yeast cell cycle, and development of statistical methods for making inferences in these models.

Summary

Since Charles Darwin formulated his theory of the evolution of species, it has been of interest in biology to identify how species have diverged by evolutionary changes. Initially this was done through morphological characters—like fins changing to limbs for the first creatures crawling up from the water—where only the large scale net result of a vast number of minute changes was taken into consideration. With the increasingly easy access to biomolecular data, in particular sequential data like DNA and protein sequences, this focus has shifted to the fundamental unit of evolution, namely the instantaneous mutation in genomic material. A vast research effort has gone into defining and refining models of sequence evolution. Larger scale biochemical data is now starting to become accessible in quantities comparable to the early days of DNA sequencing, and this growth can only be expected to continue along a similar trajectory. A fundamental goal in biology is to understand the mapping from genotype, i.e. the DNA of a species, to phenotype, i.e. the actual living individuals of a species. It is thus both timely and highly relevant to start expanding our repertoire of evolutionary models from biosequences towards higher levels of organisation. We will pursue this objective for the key concept of how genes interact in regulatory networks, with initial application to the yeast cell cycle where sufficiently detailed knowledge is already available to allow sensible measures of evolutionary space and fitness to be defined.

Evolving Dynamical Systems of the Yeast Cell Cycle

The biosciences have recently seen the rise of two major trends. Firstly, the use of evolution in structural biology as seen in bioinformatics and comparative genomics. Secondly, the use of predictive mathematical modelling in systems biology and computational biology. Clearly, there is an intersection between these two major areas. Predictive mathematical models for two homologous systems should not be considered in isolation, but rather as connected through a common ancestor by evolution. Introducing a comparative angle to systems biology has several benefits, among which are

- better understanding of how biological systems change over time
- improved accuracy of model inference
- a rigorous model for knowledge transfer from well described systems to new, homologous systems

A great variety of models are used to describe the dynamics of biological systems. Such models can be discrete or continuous in time and/or space and stochastic or deterministic, or even consist of a mixture of these principles. We will initially focus on

The general setting can be formulated as determining the probability $P(N_1, N_2 | D)$ of networks N_1 and N_2 for the two homologous biological systems under consideration in the context of the data D that has been observed. Applying Bayes' Theorem, we can rewrite this as

$$\begin{aligned} P(N_1, N_2 | D) &= \frac{P(N_1, N_2)P(D | N_1, N_2)}{P(D)} \\ &= \frac{P(N_1, N_2)P(D_1 | N_1)P(D_2 | N_2)}{P(D)}. \end{aligned}$$

The second equality follows from the reasonable assumption that the data D_i obtained from system i is independent of everything, except the network N_i describing the system.

This leaves us with four entities that should be computed or estimated. The overall probability of the data $P(D)$ is difficult to define. Fortunately, it is independent of network choices and consequently can be considered a constant term in the computations. There exists sampling methods, in particular Markov chain Monte Carlo (MCMC), that allow us to sample from a distribution as long as we can compute a function proportional to the distribution. Hence, $P(D)$ can be ignored.

The two entities $P(D_1 | N_1)$ and $P(D_2 | N_2)$ express how well a network N fits observed data D . This is the focus of ongoing research in Bela Novak's group[References?], and efficient methods already present can easily be transferred from this work to our proposed application.

The last entity we need is $P(N_1, N_2)$, the probability of co-observing networks N_1 and N_2 . In the traditional approach of independently inferring networks for different systems, this has been assumed to be a product of independent probabilities of networks N_1 and N_2 , even if not explicitly stated. The novel idea in our proposal is to model the evolution of the two networks from a common ancestral network N_{anc} . Assuming independence of evolutionary paths – essentially stating that evolution has not been significantly influenced by a shared, changing environment – this allow us to write

$$P(N_1, N_2) = \sum_{N_{\text{anc}}} P(N_{\text{anc}})P(N_{\text{anc}} \text{ evolves to } N_1)P(N_{\text{anc}} \text{ evolves to } N_2),$$

or simply

$$P(N_1, N_2) = P(N_1)P(N_1 \text{ evolves to } N_2)$$

when using time reversible models of network evolution.

We propose to compute this entity by applying MCMC sampling to trajectories, $N_1 \rightarrow N^{(1)} \rightarrow N^{(2)} \rightarrow \dots \rightarrow N^{(l)} \rightarrow N_2$, connecting networks N_1 and N_2 through intermediate networks $N^{(i)}$ by simple evolutionary steps. In each MCMC iteration a small part of the current trajectory is modified, and the modification accepted or rejected according to standard MCMC methodology. We have successfully applied a similar scheme to statistical alignment [3].

Connecting two networks by a trajectory of intermediate networks requires that we address the following issues:

- What is the space the trajectories exist in, i.e. what is the universe of possible networks the $N^{(i)}$ xs are chosen from

- How is one network connected to the next network in the trajectory, i.e. what are the simple evolutionary steps changing networks
- What is the fitness of an intermediate network

An initial choice for universe will be the power set of all interactions observed in networks N_1 and N_2 . As seen in the cell cycle models of budding yeast and fission yeast, cf. Fig. 1, many interactions will be shared between homologous networks, but some will be unique to one or the other. A simple trajectory connecting N_1 and N_2 will just be a mixture of loss of interactions unique to N_1 and gain of interactions unique to N_2 , but sometimes a lost interaction may have to be regained or vice versa to ensure the fitness of the intermediate as discussed below. An obvious next step will be to include observed or hypothesised interactions not present in either N_1 or N_2 in the set of interactions intermediate networks are constructed from. Bela Novak's group has a high level of expertise in this area, and can easily supply the necessary information.

An obvious choice for simple evolutionary steps will be the change of a single interaction in the network, corresponding to a mutation affecting the efficacy of an involved gene. Such changes can naturally be divided into

- change of kinetic parameters, e.g. according to Ornstein-Uhlenbeck process
- complete loss or gain of the interaction, e.g. according to a birth-death process with birth rate λ and death rate μ

One can also view gain and loss of interactions as simply changing the corresponding kinetic parameter from or to a value of 0, and we will explore both approaches.

Perhaps the most complex of the three issues is assessing the fitness of an intermediate network. The great success of comparative genomics can be seen as circumstantial evidence that the fitness of most intermediate sequences postulated by plausible evolutionary trajectories can be ignored. However, for networks we would expect a much larger variation in the fitness of intermediate networks as the loss or gain of a single interaction can often lead to a breakdown of the intended functionality. Though there are some natural universal criteria, like overall network connectedness, this issue has to be addressed on a case-by-case basis depending on the functionality. For example for the yeast cell cycle, an obvious further criterion would be the continued presence of clearly identifiable G₁, S, G₂ and M phases. Bela Novak's group has developed computational models that can make this assessment with the computational efficiency necessary to be applied in a sampling based scheme.

There are several ways to extend on the work proposed for the initial part of this project. Probably the most straightforward extension is to include more than two systems in the analysis. Though not entirely trivial, it is possible to transfer existing methods from statistical methods. Currently the bottleneck for this extension is identification of sufficiently well studied systems with more than two homologues, a situation that is likely to change with the continued growth in the field of systems biology.

Initially, we will only be concerned about finding plausible evolutionary trajectories utilising known interactions. A more ambitious goal will be to attempt to discover interactions that may have been present in an ancestral system but not in any of the extant systems. This will require allowing evolutionary steps to insert random interactions. For this to be successful, it may be necessary to develop methods for identifying

single interactions that will significantly increase the fitness of a network. This would allow the inclusion of interactions crucial for the transition between two networks without introducing a deluge of random interactions with negligible effects.

Finally, it may be sensible to include other types of data in the process. One obvious example is sequence data for the interacting genes. A simple use could be to let the level of conservation observed for a gene affect the variation allowed for kinetic parameters of interactions involving that gene. However, more involved approaches to combine sequence and regulation evolution are certainly possible.

The methods described in [3] do allow sequences to be sampled for the evolutionary trajectory relating two species. With detailed knowledge of the genes and binding sites involved in the yeast cell cycle regulation [How much is actually known about this?], this can be combined with methods for evolution of transcription factor binding sites, as e.g. described in [1]. Whereas models for sequence evolution seem to achieve almost universal applicability without consideration of context, in particular natural restrictions of evolutionary trajectories and the fitness of intermediates, we believe it to be crucial for evolutionary models of complex systems like regulatory networks to include as much context information as possible.

References

- [1] Johannes Berg, Stana Willmann, and Michael Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4:42, 2004.
- [2] Katherine C. Chen, Attila Csikasz-Nagy, Bela Gyorffy, John Val, Bela Novak, and John J. Tyson. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Molecular Biology of the Cell*, 11(1):369–391, 2000.
- [3] István Miklós, Ádám Novák, Rahul Satija, Rune Lyngsø, and Jotun Hein. Stochastic models of sequence evolution including insertion-deletion events. *Statistical Methods in Medical Research*, 2009. Accepted.
- [4] Bela Novak, Zsuzsa Pataki, Andrea Ciliberto, and John J. Tyson. Mathematical model of the cell division cycle of fission yeast. *Chaos*, 11(1):277–286, 2001.