ANALYSIS OF BIOLOGICAL NETWORK DATA USING LIKELIHOOD-FREE INFERENCE TECHNIQUES

Carsten Wiuf[1]*, Oliver Ratmann*[2] *and Michael Knudsen*[1]

[1]Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark
[2]Department of Public Health and Epidemiology, Imperial College London, W2 1PG London, UK
wiuf@birc.au.dk, micknundsen@gmail.com, o.ratmann@imperial.ac.uk

ABSTRACT

Biological Networks have received much attention in recent years, but statistical tools for network analysis are still in their infancy. In this paper we focus on Protein Interaction Networks (PINs) that typically comprise thousands of proteins and interactions. PINs are the result of long evolutionary histories. Here we adopt simple mathematical models that capture essentials of protein evolution and develop statistical methods to estimate evolutionary PIN parameters. Our initial approach is based on a recursion for the likelihood, but it becomes computationally intractable for reasonably sized networks. Our second approach is based on summary statistics and likelihood-free inference. We discuss problems with selection of summaries, convergence, and credibility and apply the methods on *Helicobacter pylori* and *Plasmodium falciparum* data.

1. INTRODUCTION

Today it is possible to obtain massive amounts of data relating to the molecular complexity, organization and structure of a single cell or organism. These data can be obtained in a single experiment and have thus geared the biosciences towards system-level science or systems biology, where the attempt is to understand the system and its organization in broader and overall terms, rather than understanding the system's individual components one by one.

One system-level data type that is becoming available is PIN data. A PIN data set is a collection of experimentally determined interactions (physical binding between proteins). As such, a PIN data set is an incomplete observation of the interactome, the entire collection of all proteins in a cell or organism together with their interactions.

Evolution has shaped the form of an organism's interactome. In principle, we should therefore be able to learn about the processes responsible for this evolution by analyzing PIN data sets from the organism. The idea is that different evolutionary processes leave different traces in the PIN data but also that parameters describing the processes may differ between organisms. For example the authors of [1] investigate which type of model best explains a *D. melonagaster* PIN data set. However, they do not attempt to estimate the parameters in the models, but base their conclusions on how well the models (evaluated over a range of parameters) account for the motifs seen in the PIN data.

In [2] (and references therein) different distributions are fitted to the degree sequence observed in various PIN data sets. While this provides insight into the differences between organisms, it does not provide insight into the processes generating the differences – simply because the distributions are not based on evolutionary models. The approach taken in [1] has the strength that it utilizes more information in the data than just the degree sequence and thus has a higher chance of uncovering relevant features.

In this paper we present statistical analysis of PIN data sets based on mathematical model of network evolution. We focus on two data sets, a *H. pylori* data set [3] and a *P. falciparum* data set [4]. Statistical analysis of network data is far from straightforward and we discuss different approaches to inference [5, 6]. We first develop a scheme for maximum likelihood inference using a full data set (i.e. an entire network), but find that it is limited in several respects. Subsequently, we develop a likelihood-free inference (LFI) approach, based on Approximate Bayesian Computation (ABC) and summary statistics [7], and show that it is much more flexible than the likelihood approach. Importantly, we find that reliable inference requires consideration of many, carefully chosen network summaries simultaneously.

Having settled on a statistical method we apply the method to the two data sets and discuss the results in relation to biological knowledge and mathematical properties of the underlying model.

2. EVOLUTION OF THE INTERACTOME

Various processes contribute to the evolution of the interactome [8, 9]. The importance of gene duplication to biological evolution has long been recognized and substantial evidence that elucidate the importance and the mechanisms of this process in higher organisms has been collected from genomic sequence data, either in the form of whole genome duplication (WGD) or as single gene duplication (SGD) [10, 9]. In the two species we use here, *H. pylori* and *P. facliparum* there is no recorded evidence of WGD and we will simply ignore it in the following discussion, though we note that for other species such as *S. cerevisiea* WGDs have played an important role [11].

2.1. Single gene duplication

In most SGDs, a gene is tandemly duplicated. Just after a successful duplication, the child and the parental genes have exactly the same functions, but over a relatively short evolutionary time [10, 9], the two genes may diverge, resulting in different fates of the duplicates: i) one gene may be silenced (non-functionalization), ii) both genes are preserved such that one is redundant to the other, iii) one gene may acquire a new function while the function of the other is retained (neo-functionalization), and iv) both genes are changed through mutations and partly acquire new functions (sub-functionalization). The latter is very attractive [10] as it does not rely on sparse occurrences of beneficial mutations, but on loss-of-function mutations in regulatory regions. Further, sub-functionalization is a natural mechanism for specialization of gene products to different tissues and cells. In contrast, for iii) to occur the acquisition of novel interactions through benefial mutations is required.

2.2. Attachment processes

Besides SGD (and WGD) a number of other processes contribute to the evolution of the interactome, which we collectively refer to as *attachment* processes. These include various forms of horizontal transfer of genetic material between organisms (typically bacteria), integration of viral DNA into the host genome and translocation of genetic material within an organism. All of these may lead to the formation of novel genes.

2.3. The model

We adopt a model that emphasises iv) as the most important consequence of SGD and distinguish two processes, SGD and *preferential attachment* (PA). The model is a Randomly Grown Graph (RGG) and has four parameters $\theta = (\alpha, p, q, r)$. A RGG is a Markov chain in the sense that the graph (network) $\mathcal{G}_{t+1} = (\mathcal{V}_{t+1}, \mathcal{E}_{t+1})$ at step $t+1$ only depends on the graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ at step t, where t denotes the size of the network. At step $t+1$ do:

[SGD] With probability α choose a node, v_{old}, at random in $\mathcal{G}_t$ and introduce a new node v_{new}. For each neighbour, v, of v_{old}, create a link between v_{new} and v with probability p; otherwise with probability r erase the link (v_{old}, v) and create the link (v_{new}, v). Create a link between v_{old} and v_{new} with probability q.

[PA] With probability $1-\alpha$ choose a node, v_{old}, with probability proportional to its degree in $\mathcal{G}_t$ and introduce a new node v_{new}. Create a link between v_{old} and v_{new}.

The model is symmetric in r in the sense that r and $1-r$ produce statistically indistinguishable networks. In our analysis we fix r to 0.5 to reduce the number of parameters. PIN data sets are incomplete and noisy in several respects; e.g. they only contain a fraction of all proteins in the interactome (incomplete sampling) and also contain both false positive and false negative links (noise). Here we only consider incomplete sampling and assume that the sampling fraction is known (estimated from the number of open reading frames in the genome).

3. STATISTICAL METHODS

We first discuss the likelihood of an entire network, and then move on to LFI.

3.1. The likelihood of a full network

Ideally, we would compute the likelihood of an oberserved PIN data set,

$$L(\theta, \mathrm{Data}) = P(\mathrm{Data}|\theta).$$

This would allow us to perform a likelihood analysis or engage in a Bayesian analysis of the posterior distribution

$$P(\theta|\mathrm{Data}) \propto P(\mathrm{Data}|\theta)p(\theta),$$

where $p(\theta)$ is a prior on θ. However, calculating $P(\mathrm{Data}|\theta)$ is computationally very demanding even for small networks.

In [5] the likelihood is calculated recursively. Denote by $\delta(\mathcal{G}_t, v)$ the graph $\mathcal{G}_t$ with the node v removed. If it is possible to go from $\delta(\mathcal{G}_t, v)$ to $\mathcal{G}_t$ by SGD or PA then we say that v is *removable* and denote the set of removable nodes by $\mathcal{R}(\mathcal{G}_t)$. Armed with this notation, the likelihood of an entire network, $\mathcal{G}_t$, takes the form

$$L(\theta, \mathcal{G}_t) = \frac{1}{t} \sum_{v \in \mathcal{R}(\mathcal{G}_t)} \omega(\theta, \mathcal{G}_t, v) L(\theta, \delta(\mathcal{G}_t, v)), \qquad (1)$$

where

$$\omega(\theta, \mathcal{G}_t, v) = P(\mathcal{G}_t | \delta(\mathcal{G}_t, v), \theta)$$

is the conditional probability of generating $\mathcal{G}_t$ from $\delta(\mathcal{G}_t, v)$. The factor $1/t$ is the probability that v is the last added node and the quantity ω is a sum over all nodes that could have given rise to v by SGD or PA.

We note that the likelihood is written in a form that may facilitate approximate procedures such as Importance Sampling (IS) or MCMC [12, 13]. However, only in the cases $\alpha = 0$ and/or $r = 0, 1$ is the set of removable nodes fairly small; in all other cases the set consists of all nodes in the network [5] and it becomes computationally untractable.

Furthermore, as an additional complication, sampling is not taken into account in the recursion above, because sampling cannot be considered at each step in the recursion, and is best implemented after the network has achieved the desired size. Other approaches are therefore required.

3.2. Likelihood-free inference

To circumvent the problems with calculating the likelihood we turn to methods of ABC and LFI [7].

The basic idea in ABC is to combine Bayesian approaches with summary approaches. Rather than targeting the posterior distribution given the full data we aim at calculating the posterior distribution given a summary of the data. This approach in addition requires to choose a reasonable set of summaries.

For a given set of summary statistics $\mathcal{S} = (S_1, \ldots, S_k)$ we adopt a MCMC scheme to simulate the posterior distribution $P(\theta|\mathcal{S})$ – now conditional on the set of summaries and not on the full PIN data. Denote by $\mathcal{S}_0$ the set of observed summary statistics. We proceed in the following way:

[A] If now at θ, propose a move to θ' according to the proposal density $q(\theta \to \theta')$

[B] Generate a network according to θ', sample the required number of nodes and calculate the summaries $\mathcal{S}'$

[C] Define $C = \prod_{i=1}^{k} \mathbf{1}(d_i(S'_i, S_{i0}) < \epsilon_i)$ and calculate

$$h(\theta, \theta') = \min\left(1, \frac{p(\theta')q(\theta' \to \theta)}{p(\theta)q(\theta \to \theta')}C\right),$$

where $\epsilon_i > 0$ is a threshold and d_i a distance measure

[D] Accept θ' with probability $h(\theta, \theta')$ and otherwise stay at θ; go to [A].

Besides the summaries, we need to choose ϵ_i and d_i. For the thresholds we choose a tempering scheme such that the thresholds decrease during the burn-in period. The final thresholds are decided upon based on MCMC diagnostics (see e.g. [12]). The d_is are taken to be Euclidian.

4. STATISTICAL ANALYSIS OF PIN DATA

In this section we present results from the analyses of the *H. pylori* and the *P. falciparum* PIN data sets. Due to space limitations we are unable to present these results in full, but refer the reader to [6].

4.1. Summary statistics

Table 1 shows the effect of varying the summary statistics. In earlier papers only the degree sequence is used (see e.g. [14, 2]) and Table 1 clearly demonstrates that inference is unreliable when judged solely from the degree sequence. Interestingly, the estimate of p is much lower when based on the degree sequence only. However, as soon as several summary statistics are applied, the exact number and the particular choice of summaries become less important.

Choosing a distance measure and a precision threshold ϵ further influences the inference. As expected, credibility intervals become more narrow when smaller thresholds are applied; however this is at the cost a lower acceptance probability in the MCMC (h is lower) and additionally, burn-in occurs later in the MCMC.

4.2. *H. pylori* and *P. falciparum*

The *H. pylori* PIN data set comprises 675 proteins and 1,096 links [3]. The sampling fraction is estimated to 45% [6]. In contrast, the *P. falciparum* PIN data set is larger, comprising 1,271 proteins and 2,642 links [4]. The sampling fraction is 24% [6]. Table 2 shows the estimates of the three parameters.

	p	q	α
I	0.32 (0.09,0.69)	0.55 (0.19,0.87)	0.57 (0.24,0.87)
II	0.57 (0.44,0.75)	0.05 (0.01,0.10)	0.78 (0.64,0.92)
III	0.56 (0.44,0.79)	0.05 (0.00,0.09)	0.79 (0.64,0.93)

Table 1. Shown are the maximum posterior estimates of p, q and α, together with 80% credibility intervals for three sets of summary statistics. I) Degree Sequence (ND); II) Distribution of distances between nodes ('within reach', WR), Diameter (DIA), Cluster coefficient (CC), Average degree (AD), and size of largest connected component; III) WR, ND, CC and FRAG. The *H. pylori* data set is used.

	p	q	α
Hp	0.57 (0.44,0.75)	0.05 (0.01,0.10)	0.78 (0.64,0.92)
Pf	0.52 (0.46,0.59)	0.05 (0.00,0.09)	0.93 (0.87,0.98)

Table 2. Shown are the maximum posterior estimates of p, q and α, together with 80% credibility intervals for Hp) *H. pylori* and Pf) *P. falciparum*. Summary statistics: WR, DIA, CC, AD and FRAG.

The estimates are very similar for p and q. However, the 80% credibility intervals are wider for *H. pylori* than for *P. falciparum* which we attribute to the difference in network order – the *P. falciparum* PIN data set is almost twice as big. Intuitively, the difference in the estimates of α are biologically reasonable: *H. pylori* is a small bacterium, and bacteria are often subject to horizontal transfer of genetic material. In contrast, *P. falciparum* is a unicellular eukaryote, and attachment processes are believed to occur rarely in eukaryotes [9].

5. MATHEMATICAL INSIGHT

The Markov property of the model allow us to deduce a number of statements about the model. The expected number, $n_t(k)$, of nodes of degree k fulfills the relation

$$n_{t+1}(k) = \left(1 - \frac{1+kp}{t}\right) n_t(k) + \frac{1+(k-1)p}{t} n_t(k-1)$$
$$+2 \sum_{j \geq k-1} \binom{j}{k-1} \psi^k (1-\psi)^{j-k+1} \frac{n_t(j)}{t},$$

where $\psi = (1+p)/2$, $r = 1/2$ and $q = \alpha = 1$ (for convenience). A similar recursion can be obtained for an arbitrary set of parameters, but is more complicated. An argument for the correctness of the recursion can be found in [15, 16].

Here we are concerned with the existence of a limiting degree distribution as the network becomes large. We distinguish several different scenarios:

- If $\alpha p < 0.5$ then there exists an equilibrium distribution (ergodic recurrent solution)
- If $\alpha = 1$ and $p < 0.533...$ then an infinitely large network has infinitely many nodes of arbitrary degree, but

an equilibrium distribution is not guarenteed to exist (recurrent solution)

- If $\alpha = 1$ and $p > 0.562...$ then an infinitely large network has finitely many nodes of arbitrary degree, but potentially an infinite number of degree 0 (transient solution)

- If $\alpha < 1$ then an infinitely large network has infinitely many nodes of arbitrary degree, but an equilibrium distribution is not guarenteed to exist (recurrent solution).

Note that for $\alpha = 1$, there is a small window between 0.533 and 0.562 where we do not know what happens. The first bullet point is closely related to the average degree in the (infinitely large) network,

$$\frac{2 - 2(1-q)\alpha}{1 - 2\alpha p},$$

if $\alpha p < 0.5$ and otherwise infinity. Assuming the estimates in Table 2, both networks have a stable or an equilibrium distribution over time: For *H. pylori*, $\alpha p = 0.44$ and for *P. falciparum*, $\alpha p = 0.48$. However, in both cases αp is close to the point where we do not know whether the network stabilizes or not.

6. CONCLUSION

We have demonstated that using advanced statistical tools such as ABC or LFI it is possible to achieve inference on parameters describing the evolution of the interactomes of *H. pylori* and *P. falsiparum*. However, the matahematical models we apply are very basic and only mimic true evolution in an approximate sense. Nonetheless, the parameter estimates we find are in accordance with intuition and biological knowledge achieved by other means.

7. ACKNOWLEDGEMENTS

CW is supported by the Danish Cancer Society and the Danish Research Councils; OR is supported by the Welcome Trust, UK. Enette Berndt Knudsen is thanked for technical assistance.

8. REFERENCES

[1] M. Middendorf, E. Ziv, and C. H. Wiggins, “Inferring network mechanisms: The drosophila melanogaster protein interaction network,” *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 3192 – 3197, 2005.

[2] M. P. H. Stumpf, P. J. Ingram, I. Nouvel, and C. Wiuf, “Statistical model selection methods applied to biological network data,” *Lect. Notes Comput. Sc.*, vol. 3737, pp. 65 – 77, 2005.

[3] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, and et al., “The protein-protein interaction map of *helicobacter pylori*,” *Nature*, vol. 409, pp. 211 – 215, 2001.

[4] D. J. LaCount, M. Vignali, R. Chettier, A. Phansalkar, and et al., “A protein interaction network of the malaria parasite *Plasmodium falciparum*,” *Nature*, vol. 438, pp. 103 – 107, 2005.

[5] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf, “A likelihood approach to analysis of network data,” *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 7566 – 7570, 2006.

[6] O. Ratmann, O. Jørgensen, T. Hinkley, M. P. H. Stumpf, S. Richardson, and C. Wiuf, “Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*,” *PLoS Comp. Biol.*, vol. 3, pp. e320, 2007.

[7] P. Marjoram and S. Tavare, “Modern computational approaches for analysing molecular-genetic variation data,” *Nat. Rev. Genet.*, vol. 7, pp. 759 – 770, 2006.

[8] S. Ohno, *Evolution by Gene Duplication*, Springer Verlag, New York, 1970.

[9] M. Lynch, *The Origins of Genome Architecture*, Sinauer Press, New York, 2007.

[10] M. Lynch and J. S. Conery, “The evolutionary fate and consequences of duplicate genes,” *Science*, vol. 290, pp. 1151 – 1155, 2000.

[11] B. Dujon, D. Sherman, G. Fischer, P. Durrens, and et al, “Genome evolution in yeasts,” *Nature*, vol. 430, pp. 35 – 44, 2004.

[12] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman & Hall, New York, 1995.

[13] P. J. Green, N. L. Hjort, and S. Richardson, *Highly Structured Stochastic Systems*, Oxford University Press, Oxford, 2003.

[14] A. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, pp. 509 – 512, 1999.

[15] O. Hagberg and C. Wiuf, “Convergence properties of the degree distribution of some growing network models,” *Bull. Math. Biol.*, vol. 68, pp. 1275 – 1291, 2006.

[16] M. Knudsen and C. Wiuf, “A markov chain approach to randomly grown graphs,” *J. Applied Math.*, vol. 2008, pp. 190836, 2008.