Statistical Analysis of Biological Network Data

Carsten Wiuf and Oliver Ratmann

University of Aarhus and Imperial College, London

This paper discusses an area of application that seems somewhat remote from the other application areas of the symposia. The intension is not to dwell on biological and experimental issues, but briefly to introduce the biological background and setting and then move on to discuss aspects of statistical methodology that frequently are applied in the biosciences and elsewhere: Markov chain Monte Carlo (MCMC), likelihood-free inference and Bayesian Statistics. Whereas the MCMC has a long history in applied statistics, likelihood-free inference approaches are more recent, but have already found numerous applications in many areas of applied statistics; e.g. in econometrics.

Over the past decade networks have played a prominent, even central, role in many different disciplines, ranging from theoretical physics (1-4) and technology all the way to sociology (5, 6) and the humanities (7, 8). In biology they have gained particular prominence (9-11) and now descriptions in terms of networks hold a fundamental role in systems biology as well as other parts of biology. Their appeal may, at least partly, be due to the fact that in addition to being based on a rigorous mathematical base (mostly graph theory; 12, 13; and statistical physics; 14-16) they also provide a convenient graphical representation of complex processes which can – at least partially – be interpreted visually.

1. Protein Interaction Networks (PINs)

Today it is possible to obtain massive amounts of data relating to the molecular complexity, organization and structure of a single cell. These data can be obtained in a single experiment and has 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.

Organism	Nodes	Links	Fraction
S. cerevisiae	4,959	17,226	91
D. melanogaster	7,451	22,636	58
C. elegans	2,638	3,970	12
H. pylori	675	1,096	45

Table 1 Number of nodes and links in typical PIN data sets. Fraction: Number of nodes (proteins) in % of estimated number of proteins in the organism.

One type of system-level data that are becoming available is PIN data (17, 18). The cell, e.g. a human cell, contains thousand of proteins. Proteins are the products of genes. The

function(s) of a protein are (very loosely speaking) determined by the protein's interaction partners, other proteins the protein is able to bind to physically. By binding the proteins can restrain or promote molecular processes and thereby influence the functioning of the cell. A typical PIN data set consists of virtually all known proteins in an organism together with the experimentally determined physical interactions. Thus, the data can be represented by a graph or network (both terms will be used), where nodes are proteins and links are interactions. Table 1 summaries some typical PIN data sets (19-23).

PIN data sets are incomplete (19-23). First of all in the obvious sense that some proteins are unknown and thus cannot be included in the experiment; secondly the experimental techniques have limitations resulting in false and missing links, and finally the data is essentially qualitative, i.e. either there is a link or not. In reality, some links would be stronger than others, and some links will only be present in certain tissues and not in the entire organism, etc. Dealing with incompleteness is a whole issue in itself (24) and it will mainly be ignored here.

2. Mathematical Models of PIN Data

To analyse PIN data sets we need a stochastic model of a graph; stochastic in the sense that we consider each link the result of a stochastic variable, and potentially also each node or the number of nodes depending on how we construct the model and how incompleteness is taken into account.

Naturally, the model should reflect the questions, hypotheses or issues we set out to investigate. In 'early' papers (2, 6, 9) the questions was mainly about determining the shape of the degree distribution (degree = number of links of a node) and researchers would fit various distributions to the observed degree distribution; e.g. a power-law $k^{-(1+\gamma)}$, where k denotes the degree of a node and γ is a parameter to be determined. This approach could be carried out without an explicit model of the network. Also simple network models, like Erdös-Renyi graphs, were applied. An Erdös-Renyi graph has a fixed number of nodes N, and each pair of non-identical nodes is connected with probability p. If N is large and p small, then the degree of a random node is approximately Poisson with intensity $\lambda=Np$. From a biological point of view such analyses have limited value in that neither γ nor λ has a biological interpretation and the biological importance of one value of λ rather than another is difficult to assess.

More recently, a number of different graph models have been proposed (16, 25-27). They resolve around a common theme, namely creating the graph by gradually adding nodes and modifying/adding links to a small initial graph. These models have their origin in physics where they have been used to demonstrate that complex structures could emerge through application of simple rules (2, 16).

The rules for adding and modifying the graph all have resemblances to evolutionary rules or processes in biology; though the interpretation of some rules is more obvious than others, see the next section. The stochastic algorithm for generating a graph is (26, 28):

1. Start with an initial graph G_s of size s
2. At step $t+1$:

a. (Duplication) With probability α, choose a node in G_t to duplicate and modify the links of the original and the new node according to Figure 1. The modification step has two parameters p and r

b. (Preferential Attachment) With probability $1-\alpha$, choose a node in G_t proportional to its degree and create a link between the chosen node and a new node; see Figure 1

3. Continue until the graph has a predefined size given by Nodes/Fraction, where Nodes and Fraction are given in Table 1 (these are examples)
4. Sample Fraction of the nodes randomly. Keep all links between the sampled nodes

The last two steps could be made stochastic, but are kept fixed here for simplicity. Note that steps 3 and 4 are the only aspects of incompleteness that are taken into account.

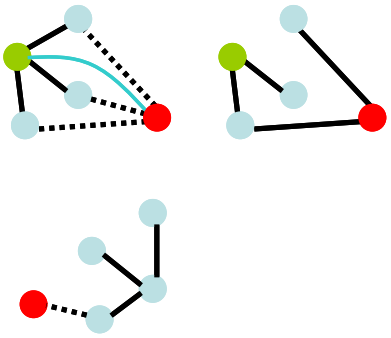

Figure 1 Top, duplication: A node is chosen at random (green node) and copied with its links (red). Afterwards an original link to a neighbour and the corresponding new link are retained with probability p. If not then one of the two links are deleted with equal probability. With probability r a link is created between the new and the original node (blue line). Bottom, preferential attachment: A node is chosen proportional to its degree and a new node is attached to it.

In total there are three parameters $\theta = (\alpha, p, r)$. Simulation of networks is straight forward and fairly efficient in terms of computational time and memory. The model is a Markov chain on graphs and at most one node in G_t and its links are used to create G_{t+1}. Typically, we consider labelled graphs; e.g. labels could be protein names. The order by which labelled nodes are introduced is called the *history* of the graph. Thus, a graph with t nodes has potentially $t!$ histories, if we start from one node. The probability of a graph and its history $P_\theta(G_t, \text{History})$ is straight forward to calculate (according to the algorithm and Figure 1), whereas the probability of $P_\theta(G_t)$ in principle requires summation over all possible histories.

Note that a simulated network of the same (large) size as an observed network will rarely be identical to an observed network. This is a common problem for complex stochastic systems: Each instance of the system has a vanishing small probability to occur.

3. Biological Relevance of the Model

A duplication event mimics the biological process of gene duplication (remember: proteins are the products of genes). After a duplication event the two genes are identical and consequently so are their products and interactions. As time go by the original gene and the duplicate might not evolve under the same selective and evolutionary constraints and some of their links might be lost. However, the organism is likely to maintain most of its

functions and require that the original and the duplicate together maintain all links to the neighbours of the original node. The probability p is then the probability that the organism maintains two proteins (the new and the original) for a given function (link).

Preferential attachment has a less clear biological meaning. However, occasionally an organism acquire a completely new gene (and hence a new protein) by horizontal transfer; i.e. import of a gene from another organism. In the mammalian world, the other organism is typical a virus and in the bacterial world, the other organism is typically also a bacteria. However, these biological processes are not understood in detail yet.

The probability α represents the balance between the two processes. It is debated to what extent both processes play a role in evolution. The relevance of the two processes and their modelling in biological terms are discussed in (29-34).

One important aspect of the model is that it is based on evolutionary ideas. In the future this should allow us to draw inference on multiple PIN data sets at the same time (e.g. related species) and thereby be able to obtain information about their joint evolution. The model as described here is very simple, though somewhat relevant, and could be improved in many ways. For the joint analysis of several PINs one important aspect in missing, namely a measure of physical time – in the model presented here time is an 'event counter'; moving one step each time a new node is included in the network.

4. Statistical Methodology

a. *The Full Likelihood*

Despite the simplicity of the model it is far from straight forward to calculate the likelihood under reasonable circumstances. It was shown in (27) that a recursive scheme could be applied to calculate the likelihood for a simpler model (and less relevant) than the one considered here. However, for networks of size >50 this approach became impractical and an Importance Sampling (IS) scheme was suggested. The idea behind the IS scheme is to sample histories of the observed network and calculate the contribution to the likelihood for each sampled history. The ideal proposal distribution is P_θ(History|Data) (35) – and not e.g. P_θ(History) – in the former case every history is supporting the data, whereas this is not the case in the latter. However, P_θ(History|Data) is difficult to characterize in the present setting.

It turns out that a reasonable proposal distribution can be chosen independently of the parameters of the model and consequently the same sampled histories can be reused to calculate the likelihood for all parameters values. This accelerates the speed of computation enormously, but introduces dependencies between likelihoods for different parameter values, because the same sampled histories are used for all parameters. However, these dependencies appear generally to be of minor importance. On the positive side counts that the sampled histories are sampled independently of each other; this guarantees that a history is chosen proportional to its proposal probability. Unfortunately, the IS scheme is likely to break down in the present case because the proposal distribution becomes intractable.

An alternative to IS is MCMC (36). The MCMC has been used in numerous contexts and has shown to be extremely useful for complex statistical systems. One approach in

continuation of the IS approach would be to devise a Markov chain on the space of histories of the network and use the histories to approximate the likelihood. However, we have not been able to construct a Markov chain that visits a large proportion of the state space in reasonable time.

Hence other approaches must be considered.

b. *Likelihood-free Inference*

Instead of considering the whole observed network one could consider a summary or a collection of summary statistics of the network. Summaries could be the degree sequence, the number of triangles etc. By carefully selecting the summary statistics one might be able to retain most of the information in the network about the parameters. However, also with this approach we run into problems, because in order to calculate the likelihood of the summaries, we need to simulate summaries according to the model, and to do so we need to simulate random graphs and histories. Further, most of the simulated histories and graphs will not match the observed summary statistics.

To circumvent these problems we loosen the criteria for matching the observed summary statistics and consider all simulated summaries to match if they are within a certain distance of the observed summaries (37-38); i.e. if $d(S_{SIM}(\theta), S_{OBS}) = |S_{SIM}(\theta) - S_{OBS}| \leq \varepsilon$, where θ indicates that simulation is done under this parameter. (Under exact simulation, $\varepsilon = 0$.)

However, in contrast to the IS scheme discussed in the previous sub-section, simulations must now be done for all parameters values (or a grid of values). To loosen the burden of computation we suggest a Bayesian scheme assuming a prior distribution on θ, and updating θ in the simulation using a MCMC approach and Metropolis-Hastings' algorithm (37-38). The goal of the inference procedure is thus to calculate (or estimate) the posterior distribution of θ given the summary statistics and not to find the maximum likelihood estimate as in Sub-section 4a. The details of the scheme are below:

1. Assign a uniform prior on θ. Calculate S_{OBS} from the data. Start at $\theta_0 = (\alpha_0, p_0, r_0)$
2. If now at θ, propose a move to θ^* according to a Gaussian proposal kernel normalized to the interval [0, 1] for each parameter. Denote the kernel by $q(\theta \rightarrow \theta^*)$
3. Generate a simulated PIN data set with parameter θ^* according to the model in Section 2. Compute the summary $S_{SIM}(\theta^*)$
4. If $|S_{SIM}(\theta^*) - S_{OBS}| \leq \varepsilon$ go to 5, otherwise go to 2 and stay at θ
5. Calculate $h = \min\{1, q(\theta \rightarrow \theta^*)/ q(\theta^* \rightarrow \theta)\}$. The term involving the prior distribution of θ disappears in h because the prior is uniform
6. Accept θ^* with probability h; otherwise stay at θ. Then return to 2

In the literature this approach is known as likelihood-free inference or Approximate Bayesian Computation (ABC) (37-38). Strictly speaking, the algorithm generates data from the posterior distribution $P(\theta | d(\cdot, S_{OBS}) \leq \varepsilon)$. In the limit as $\varepsilon \rightarrow 0$, this distribution becomes the posterior $P(\theta | S_{OBS})$.

To increase the performance of the algorithm we further adopt a simulated annealing scheme (or tempered simulation scheme), such that ε and the variance of the kernel distribution depends on how many times steps 2-6 have been performed. In the beginning, high values of ε and the variance are used to increase the Markov chain's possibilities to wander round in the parameter space. After a burn-in period the scheme is cooled down, i.e. the values are lowered. Acceptance probabilities (h) are typically in the range 15-40% for the applications we consider.

5. Statistical Analysis

a. *Simulated PIN Data*

We will validate the approach on simulated PIN data sets. We aim to show that it is possible to select reasonable summary statistics that capture most of the information in the data and to show that the posterior peaks approximately at the true parameter value.

We have chosen a number of summary statistics, some of which have frequently been used in the literature, while others are less used. Here we mention a selection of these: The degree sequence (DISTND; number of nodes with k links, $k = 0, 1,\ldots$) or just the mean degree of a node (AVGND); The within-reach distribution (WR; number of nodes that can be reaches within k links from a given node, $k = 0, 1,\ldots$) or the mean of this distribution; The diameter (DIA) of the network; The number of triangles (TRIA); The fragmentation (FRAG; a measure of the disconnectedness of the network); And the cluster coefficient (CC; a measure of groupings in the network). The frequently used summaries are DISTND (as discussed in Section 2) and CC.

One criterion for a summary statistic to be useful is that its expectation varies over the set of parameters. To illustrate this one can calculate (or approximate by simulation) the derivative of the expectation with respects to the parameter (39). In Figure 2 this is shown for DISTND and WR.

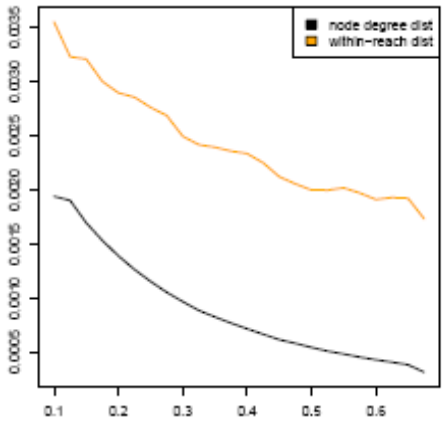

Figure 2 The figure shows the derivative of the expectation of DISTND and the derivative of the expectation of WR for α. The derivatives are normalized by α. Clearly, there appears to be more information in WR than DISTND for this choice of parameter.

In Figure 3 we show how the number of summary statistics affects the peak-ness of the posterior distribution. However, it also transpires that the commonly applied DISTND does barely capture the information in the data. PIN networks were simulated with 120 nodes and 100 retained after sampling.

Calculation of the Gelman-Rubin convergence statistic (40; see also next sub-section) indicates that the likelihood-free scheme works sufficiently well (again for small networks of size 100; results not shown here). Now, we will move on the analysis of a real data set.

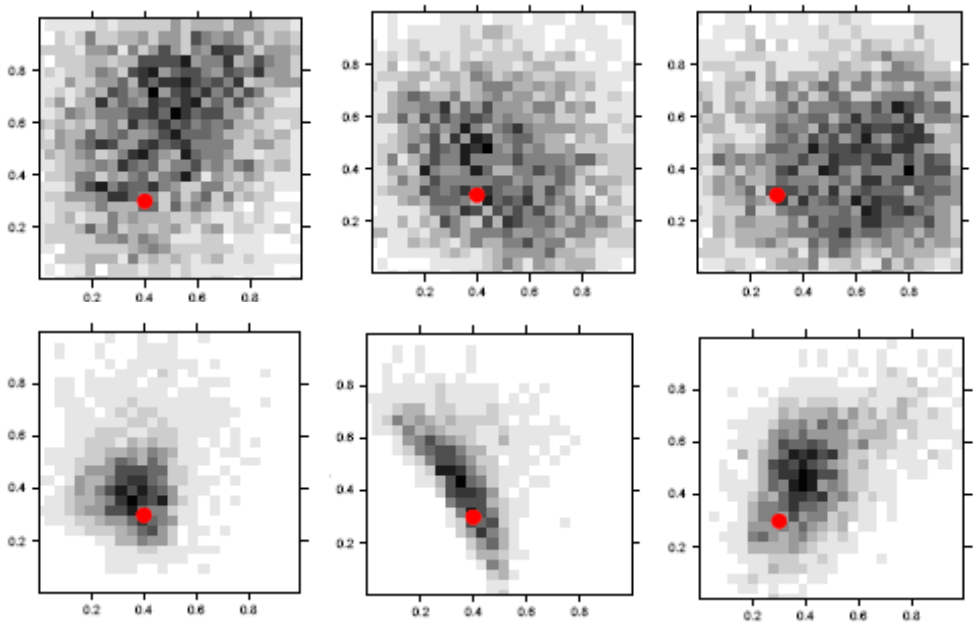

Figure 3 The top row shows heat diagrams of the posterior density for all three pairs of parameters α, p, and r when using only DISTND. The bottom row shows the same when using WR+DIA+CC+AVGND+FRAG. Clearly, the peak-ness depends strongly on the summary statistics. The red dots indicate the true values.

b. *H. pylori PIN Data*

Helicobacter pylori is a small bacteria that is associated with various forms of ulcer as well as stomach cancer. It has a relatively small genome with an approximated number of genes/proteins of 1500 (in comparison the human genome has approximately 23,000 genes). The PIN data set to be analysed here holds less than half of the estimated number of proteins; namely 675, and has 1096 links (see Table 1).

We did various runs with different choices of summary statistics, see Table 2. From the previous sub-section it was clear that to obtain a reasonable performance of the likelihood-free inference scheme more than one summary statistic should be chosen. Again it transpires that the node degree distribution (DISTND) is very unreliable and cannot alone be used for inference. Interestingly, the estimates of α indicate that both processes might play a role in the evolution of *H. pylori.*

Network Summaries	p	r	α
WR+DIA+CC+AVGND+FRAG	0.275 (0.16, 0.39)	0.034 (0.002, 0.065)	0.206 (0.02, 0.36)
WR+DISTND+CC+FRAG	0.277 (0.12, 0.40)	0.027 (0.002, 0.055)	0.154 (0.01, 0.31)
DISTND	0.518 (0.09, 0.86)	0.645 (0.12, 0.97)	0.338 (0.03, 0.74)
CC+TRIA	0.35 (0.06, 0.63)	0.236 (0.03, 0.79)	0.577 (0.12, 0.88)

Table 2 Posterior inference using different summary statistics. If one or few statistics are used then the 95% credibility intervals become very large (in parentheses) and do not always agree (or agree poorly) with inference performed using many summaries. As an example, inference on r and α is markedly affected when using only DISTND.

Figure 4 shows that the convergence properties of the likelihood-free inference scheme appear to be good for the *H. pylori* PIN network.

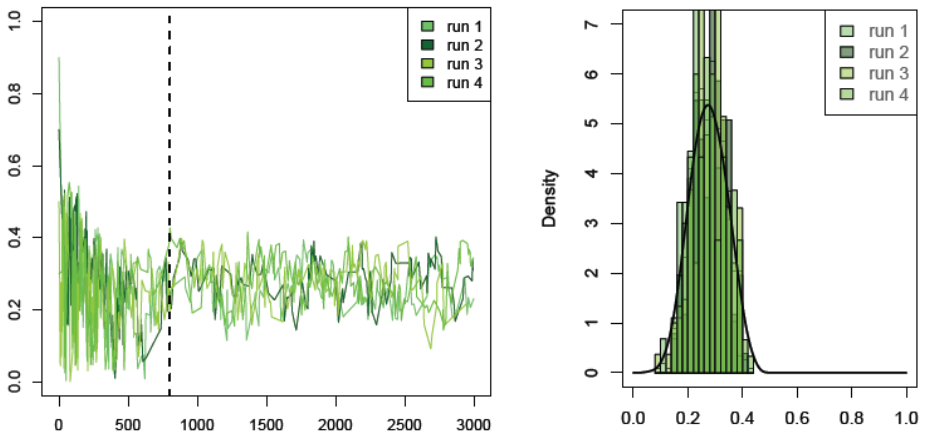

Figure 4 Four runs of the inference scheme, starting with different initial parameter values. In the left plot, the accepted moves of p are shown as function of iteration step. In the right plot the posterior distribution of p is shown for the four runs, excluding iteration steps to the left of the dashed line in the left plot. The convergence properties appear to be good.

6. Conclusions

In this paper we have discussed inference on large (biological) network data. This type of data is becoming abundant, not only in biology, but also in many other sciences including sociology, physics, and economics. We have found that the likelihood of the network is difficult to calculate even for simple models. To compensate this we suggested a likelihood-free inference scheme, based on summary statistics, MCMC and Bayesian ideas. We demonstrated that it is possible to obtain reasonable inference using this scheme, if care is taken in selecting the summary statistics. Essentially, this appears to be the main problem with the proposed approach: As shown in Table 2, poorly chosen summary statistics might lead to unreliable or even wrong statistical conclusions. We speculate that similar issues and conclusions can be made in many other sciences where complex statistical systems are being applied in the analysis of data.

Biologically, we were able to obtain estimates of biologically relevant parameters in an analysis of a *H. pylori* PIN data set. In the future, it is going to be interesting to compare the estimates obtained in this particular case with estimates obtained for other, but perhaps related, species.

6. References

1. Barabasi A, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. *Physica* A 272: 173-187
2. Barabasi A, Albert R (1999) Emergence of scaling in random networks. *Science* 286: 509-512.
3. Burda Z, Correia JD, Krzywicki A (2001) Statistical ensemble of scale-free random graphs. *Phys Rev* E 64: 046118
4. Evans T (2004) Complex networks. *Contemporary Physics* 45: 455-474.

5. Scott J (2000) *Social Network Analysis*. Sage Publications
6. Newman N, Park J (2003) Why social networks are different from other types of networks. *Phys Rev* E 68: 036122
7. Padgett J (1993) Robust action and the rise of the medici. *Am J Sociaology* 98: 1259-1319
8. Laidlaw Z (2005) *Colonial Connections 1815-1845*. Manchester University Press
9. Alm E, Arkin AP (2003) Biological networks. *Curr Opin Struct Biol* 12:193-202
10. Cork J, Purugganan M (2004) The evolution of molecular genetic pathways and networks. *Bioessay* 26: 479-484
11. de Silva E, Stumpf M (2005) Complex networks and simple models in biology. *J Roy Soc Interface* 2: 419-430
12. Bollobas B (1998) *Random Graphs*. Academic Press
13. Bollobas B, Riodan O (2003) Mathematical results on scale-free graphs. In Bornholdt S, Schuster H (eds.) *Handbook of Graphs and Networks*, pp 1-34. Wiley & Sons
14. Albert R, Barabasi A (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74: 47-97
15. Newman M (2003) The structure and function of complex networks. *SIAM Review* 45: 167-256
16. Dorogovtsev S, Mendes J (2003) *Evolution of Networks*. Oxford University Press
17. Uetz P, Finley R (2005) From protein networks to biological systems. *FEBS Lett* 579: 1821-1827
18. Vidal M (2005) Interactome modelling. *FEBS Lett* 579: 1834-1838
19. Li S *et al* (2004) A map of the interactome network of the metazoan *C. elegans. Science* 303: 540-543
20. Giot L *et al* (2003) A protein interaction map of *Drosophila melanogaster. Science* 302: 1727-1736
21. Uetz *et al* (2000) A comprehensive analysis of protein-protein interaction networks in *Saccharomyces cerevisiae*. *Nature* 403: 623-627
22. Ito T *et al* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 98: 4569-4574
23. Rain JC *et al* (2001) The protein-protein interaction map of *Helicobacter pylori. Nature* 409: 211-215
24. de Silva E, Ingram P, Agrafioti I, Swire J, Wiuf C, Stumpf MPH (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biology* 4: 39
25. Milo R *et al* (2002) Network motifs: Simple building blocks of complex networks. *Science* 298: 824-827
26. Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. *Proc Natl Acad Sci USA* 102: 3192-3197
27. Wiuf C, Brameier M, Hagberg O, Stumpf MPH (2006) A likelihood approach to analysis of network data. *Proc Natl Acad Sci USA* 103: 7566-7570
28. Ratmann O, Hinkley T, Jørgensen O, Stumpf MPH, Richardson S, Wiuf C (2007) Fitting evolution models to the protein networks of *H. pylori* and *P. falziae* with likelihood-free inference. *Submitted*

29. Ohno S (1970) *Evolution by Gene Duplication.* Springer-Verlag
30. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155
31. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302: 1401-1404
32. Force *et al* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531-1545
33. Moore RC, Purugganan MD (2005) The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol* 8: 122-128
34. Maere S *et al* (2005) Modelling gene and genome duplication in eukaryotes. *Proc Natl Acad Sci USA* 102: 5454-5459
35. Stephens M (2001) Inference under the coalescent. In Balding DJ, Bishop M, Cannings C (eds.) *Handbook of Statistical Genetics*, pp 213-238. Wiley & Sons
36. Gilks WR, Richardson S, Spiegelhalter DJ (eds.) (2002) *Markov Chain Monte Carlo in Practice.* Chapman & Hall
37. Marjoram P, Molitor J, Plagnol V, Tavare S (2003) Markov chain Monte Carlo without likelihood. *Proc Natl Acad Sci USA* 100: 15324-15328
38. Plagnol V, Tavare S (2003) Approximate Bayesian computation and MCMC. *Proceedings of MCQMC2002*
39. Heggland K, Frigessi A (2004) Estimating functions in indirect inference. *J Roy Stat Soc* B 66: 447-462
40. Gelman A (2003) *Bayesian Data Analysis.* CRC Press