Evolutionary changes of metabolic networks and their biosynthetic capacities

O. Ebenhöh, T. Handorf and D. Kahn

Abstract: The metabolic networks of different species show a large variety in their structural design. In this work, the evolution of functional properties of metabolism in relation with metabolic network structure is investigated. The metabolism of ancestral species is inferred from the metabolism of contemporary species using a Bayesian network model for metabolism evolution. Subsequently, these networks are analysed with the recently developed method of network expansion. This method allows for a structural analysis of metabolic networks as well as a quantification of network functions in terms of their synthesising capacities when they are provided with certain external resources. The evolutionary dynamics of one particular network function: the metabolic expansion of glucose is investigated.

1 Introduction

Traditionally, for the theoretical analysis of metabolic systems, models based on sets of ordinary differential equations are used for simulating their dynamic properties. For such models, detailed knowledge of the stoichiometry, regulatory interactions, and kinetic characteristics of the enzymes is required [1]. In general, these models describe systems of a relatively small size, such as single biochemical pathways or a small number of interacting pathways, for which the wiring principles of the reactions and metabolites are easily comprehensible. In recent years, a different class of models has emerged that aims at explaining the wiring principles of selected pathways [2–6]. Although there is still a lack of comprehensive knowledge of the specific kinetic characteristics of biochemical reactions, structural information about large-scale metabolic networks has become accessible with the emergence of biochemical databases such as KEGG [7, 8] or BRENDA [9]. Several approaches for the structural analysis of metabolic networks, which are feasible without considering kinetic properties of enzymes, have been introduced, such as flux balance analysis [10], the concept of elementary flux modes [11] or extreme pathways [12] as well as graph theoretical analyses [13, 14].

In this work, we are concerned with the question how structural features of metabolic networks have changed through evolution and what the consequences were for the functional properties. For our analysis, we invoke the recently developed method of network expansion [15, 16], which is based on the basic biochemical fact that only those reactions may take place, which use the available substrates, and that the products of these reactions may in turn be utilised by other reactions. With a number of given substrates (the seed), a series of metabolic networks is constructed, where in each step the network is expanded by those reactions that utilise only the seed and those metabolites that are products of reactions incorporated in previous steps. The set of metabolites within the final network is called the scope of the seed. By construction, the scope describes the synthesising capacity of a metabolic network when only the seed compounds are available as external resources.

The concept of scopes is particularly useful to systematically investigate the relations between structural and functional features of metabolic networks of a wide variety of organisms [17]. We expand this cross-species comparison by the inclusion of metabolic networks of ancestral species, which we infer from the reaction content of present day metabolic networks and available phylogenetic information. We first describe the inference of ancestral metabolic networks and then present the functional analysis of metabolic networks on the evolutionary tree.

2 Methods

2.1 Reconstruction of ancestral metabolic networks

In order to investigate evolutionary aspects of metabolism structure and function, we attempted to reconstruct most probable ancestral metabolic networks. We started from the reaction inventory of 233 species whose complete genome was analysed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology system [8]. The tree corresponding to these species was derived from the National Centre for Biotechnology Information (NCBI) taxonomy (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/) [18], using a program developed by Bru [19]. The leaves of this tree correspond to contemporary genomes, whereas 129 internal nodes correspond to predicted common ancestors. For each metabolic reaction, each node was associated with two possible states – either presence or absence of reaction in the corresponding species. Evolution of metabolism was modelled using a Bayesian network based on this tree, with conditional probabilities for reaction gain or loss associated with each edge. Non-conditional

IEE Proceedings online no. 20060014

doi:10.1049/ip-syb:20060014

Paper first received 1st March and in revised form 8th May 2006

O. Ebenhöh and T. Handorf are with the Department of Biology, Humboldt University, Berlin, Germany

D. Kahn is with the Biométrie et Biologie Evolutive, Université Lyon 1, France

E-mail: oliver.ebenhoeh@rz.hu-berlin.de

probabilities were associated to the states of the root, the last universal common ancestor (LUCA). The Bayesian tree model was implemented using the Bayesian Network Toolbox [20] for MATLAB (The MathWorks, Inc.). Conditional and non-conditional probabilities were estimated by expectation maximisation (EM) [21] so as to maximise the likelihood of the observed occurrences over all reactions in KEGG. The resulting Bayesian tree model allowed us to infer the most probable evolutionary scenario for each reaction, predicting its pattern of occurrence in ancestral species under a maximum likelihood principle. Conversely, the most probable reaction network of each ancestral species could be inferred by assembling all reactions predicted to be present from the previous analysis.

2.2 Calculation of biosynthetic capacities

As all important biochemical compounds are organic molecules, it is interesting to investigate how carbon atoms from external sources are incorporated into the cellular metabolism. In the following, we analyse as an example, which biochemical compounds organisms can produce when they are provided with glucose as the only carbon source. For this, we assume that all metabolites that do not contain a carbon atom are available. For all metabolic networks of the present species as well as for the putative networks for their common ancestors, we calculate the scopes from a seed consisting of all the non-carbon containing molecules and D-glucose. These scopes characterise the network capacities to utilise glucose as a carbon source. We will term this particular biological function of a metabolic network its biosynthetic capacity. For our calculations we assume that the most important cofactors, namely ATP/ADP, NADH/NAD^+ and Coenzyme-A, whose presence is required for many reactions, do not have to be synthesised during the expansion process. We rather assume that they act only in their function as cofactors, that is transferring phosphate groups, accepting electrons or transferring acyl groups.

3 Results

3.1 Correlation of biosynthetic capacity to network size

To assess the synthesising capacity, we plot in Fig. 1 for each organism, the number of carbon containing compounds which can be produced from glucose and inorganic substances against network size, which is defined as the numbers of reactions in the metabolic network. LUCA is marked by a black square (596 reactions, 126 compounds). All present day species are indicated by filled circles, whereas ancestral species are marked by open circles. The colour of the circles indicates to which domain the corresponding organism belongs, black symbols represent bacteria, dark-grey symbols archaea and grey symbols eukaryotes. It can be seen that as a general tendency scope size increases with network size. The organisms can be grouped into two categories, lower and upper. The lower category contains organisms with a very small synthesising capacity, and the corresponding data points are located near the horizontal axis. The metabolic networks of such organisms are unable to incorporate carbon atoms when glucose is available as the only carbon source. In fact, we found that 97 networks (70 existing and 27 ancestral) cannot produce any new carbon containing compound from glucose. The upper category includes all organisms with a large synthesising capacity and

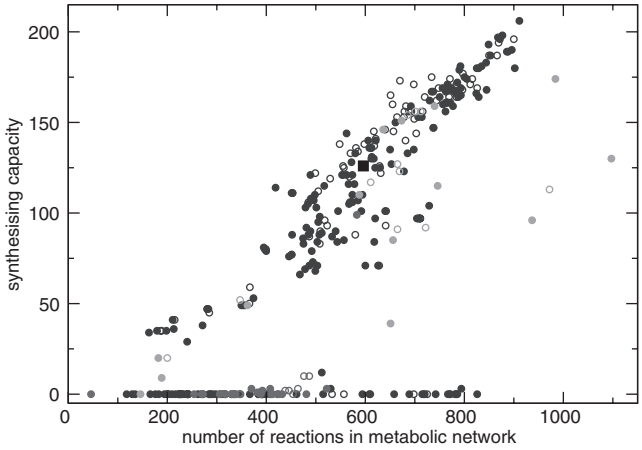

Fig. 1 *Synthesising capacity for glucose as a carbon source for 233 contemporary and 129 ancestral organisms*

Synthesising capacity is measured by counting all carbon containing compounds which can be produced from glucose and inorganic substances (y-axis)
This quantity is plotted against the network size (numbers of reactions in the metabolic network, x-axis)
Contemporary species are indicated by filled circles, ancestral species by open circles
Bacteria are denoted by black, archaea by dark-grey and eukaryotes by grey symbols
LUCA is represented by the black square

within this category there is a strong correlation between network size and synthesising capacity. It can be observed that most archaea belong to the lower category. Interestingly, while most eukaryotic networks belong to the upper category, some of them exhibit a scope of intermediary size.

3.2 Evolutionary changes of biosynthetic capacity

We next analyse how network size and synthesising capacity are related to species phylogeny by plotting the information contained in Fig. 1 on the taxonomy tree.

The evolutionary tree is depicted in Fig. 2. Each leaf node represents an existing species, whereas a non-leaf node represents a common ancestor from which the nearest outward nodes have emerged in a speciation event. In the following, we will call each such event, represented by a branch of the tree, one evolutionary step. The areas of the circles are proportional to the network size, the colour characterises the relative scope size, which we define as the number of carbon containing compounds in the scope divided by the number of all carbon containing compounds that occur in at least one reaction of the metabolic network. Light-grey nodes belong to species with a small synthesising capacity, while the dark-grey nodes can utilise glucose as the sole carbon source to synthesise ~30% of all carbon containing compounds. Both, network and scope size can change dramatically in a single evolutionary step. For a closer investigation, we systematically analyse the correlation between changes in the reaction content and changes in the synthesising capacity during an evolutionary step. In Fig. 3, the change in scope size is plotted against the change in network size. The high frequency of points on the horizontal axis near the origin of the graph indicates that most events involve moderate changes in network size and only very small changes in biosynthetic capacity. In 33 cases, the scope of glucose expansion remains exactly the same non-trivially (i.e. the scope contained other organic molecules than glucose), despite a significant change in

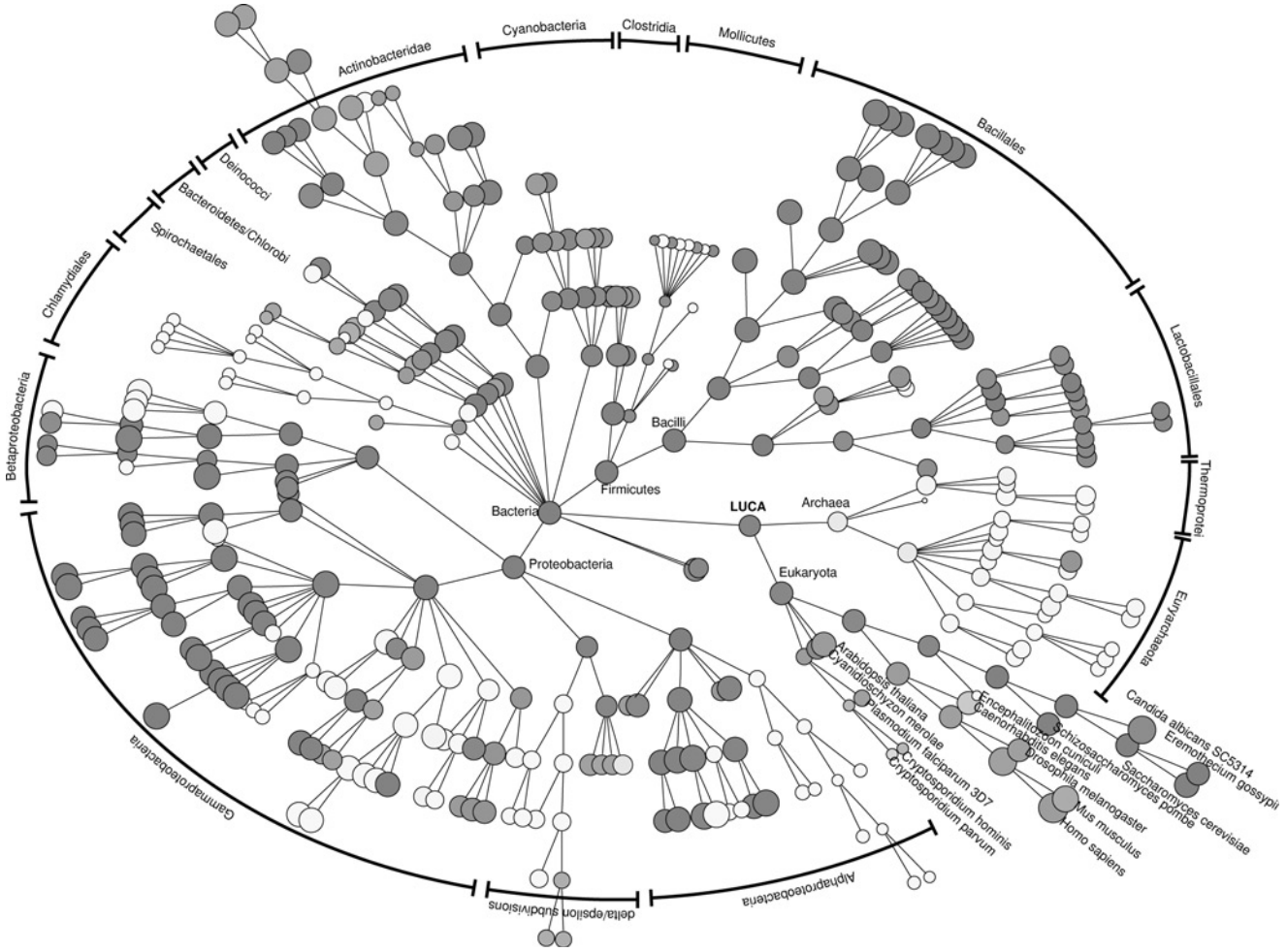

Fig. 2 *Evolution of network size and synthesising capacity*

Tree represents the evolution of the 233 considered organisms from the LUCA
Area of the circles are proportional to network size. The colour shading indicates the relative scope size, defined as the number of carbon containing compounds in the glucose scope divided by the number of all carbon containing metabolites
Continuous shading ranges from light-grey to dark-grey for 0 to 30% biosynthetic capacity, respectively

network structure. This observation is in agreement with results from earlier investigations [16, 17] that the scopes are generally robust against small alterations in the network structure. In general, there is a weak correlation between change in network size and change in scope size ($r = 0.67$). In most events, scope and network size are either both increased or both decreased, whereas there are still occasional events in which a decrease in network size leads to an increase in scope size and, conversely, an increase in network size is accompanied with a strong reduction of the scope. Interestingly, it can be observed that the synthesising capacity seemingly remains rather stable in many evolutionary branches (Fig. 2). For example, archaea tend to display a very low synthesising capacity, whereas many eukaryotes exhibit a medium capacity (refer also to Fig. 1) and most bacillales possess a high capacity. This suggests that at certain stages during evolution key events occurred that drastically changed metabolism. However, in most speciation events, predicted metabolic functions appear only minimally influenced.

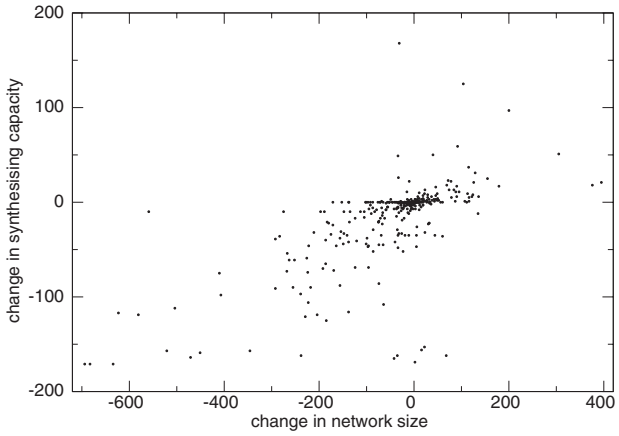

Fig. 3 *Changes in reaction content and synthesising capacity*

For each speciation event (one edge in the tree in Fig. 2), the change in synthesising capacity is plotted against the change in network size

A consequence of this hypothesis is that such organisms that are closely related in an evolutionary sense, should, as a general tendency, also display a similar functional behaviour, which in this investigation we measure in terms of the glucose scope. We verify this hypothesis by examining all pairs of organisms (past and present) and relating their distance on the evolutionary tree with the dissimilarity in their network structure as well as in their synthesising capacity. Here, the graph distance between two nodes on the species tree is used as a rough approximation of their evolutionary distance (from 1 to 16). To compare the metabolic networks and the scopes, we introduce a distance

measure quantifying the difference of two sets, which is based on the Jaccard-coefficient. For two sets M_1 and M_2, this coefficient is defined to be

$$JC(M_1, M_2) = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|} \quad (1)$$

For two identical sets, $JC = 1$, whereas for two completely disjoint sets, $JC = 0$. As a distance between two sets, we use the measure

$$d(M_1, M_2) = 1 - JC(M_1, M_2) \quad (2)$$

To assess the dissimilarity between two metabolic networks, we consider both networks as sets of reactions and calculate their distance (2). To quantify the distinctness of the synthesising capacities of two networks, we consider their scopes as sets of compounds and again apply definition (2). In Fig. 4*a*, the dependence of the differences in the network structure on the graph distance of two organisms is depicted. The bold line represents the median of network distances for all pairs of organisms with a certain distance. The thin lines indicate the 10% quantiles, one-tenth of all pairs possess a network distance smaller than the lower line and one-tenth a larger distance than the upper line. Not surprisingly, as a tendency network distance increases with increasing graph distance.

Fig. 4*b* visualises the differences in the synthesising capacities. As for those organisms with zero synthesising capacity, the scope distance of two such organisms is trivially zero while the scope distance between such an organism and any other organism amounts to one, we considered for this plot only pairs of organisms for which both exhibit a non-zero synthesising capacity. The bold line represents the median of the scope distances for all considered pairs of organisms with the same graph distance. Again, a general tendency of increasing scope distance with increasing graph distance can be observed. The 10% quantiles (thin lines) show that the variability in scope distances is larger than for network distances. This indicates that, even though biosynthetic capabilities are generally robust to small changes in the underlying network structures, the range of changes in these network functions is larger.

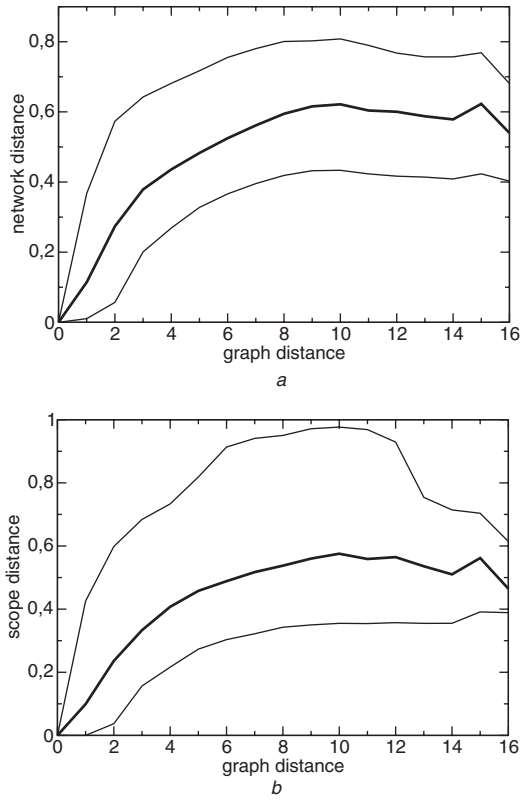

Fig. 4 *Network distance and dissimilarity in the synthesising capacities as a function of graph distance*

a Metabolic networks
b Glucose scopes
For all pairs of organisms with the same graph distance (number of edges separating them on the tree in Fig. 2, ranging from 0 to 16), the dissimilarities of their metabolic networks and their glucose scopes were determined according to (2)
Bold lines indicates the median of these values, the thin lines the 10% quantiles

4 Discussion

From the structural information on the metabolic networks of 233 organisms, which we have retrieved from the KEGG database, we have inferred scenarios for the metabolic networks of common ancestral species. Using the method of network expansion, we analysed the metabolism of present day organisms as well as the putative metabolism of ancestral species. Our investigations centred around one particular network function, biosynthetic capacity on glucose as a carbon source. Clearly, metabolic networks have to perform not only one but a multitude of biological functions. In principle, our investigations can be performed for other functions, such as the utilisation of other carbohydrates or the incorporation of phosphate or nitrogen into cellular metabolism. Glucose metabolism is closely linked to energy metabolism, one of the central cellular functions being the production of ATP. It would also be interesting to analyse how this function, measured for example in ATP production rate, changes along the evolutionary tree. However, such an analysis is not feasible with purely structural methods, as it requires detailed knowledge of kinetic parameters and regulatory mechanisms. In this work, we focused on an equally important cellular function of glucose metabolism, the preparation of precursors for other metabolic processes and biomass production from glucose as the sole carbon source.

We predicted how biosynthetic capacity on glucose as a carbon source evolved along the species tree. We found that 97 networks have zero synthesising capacity, that is the corresponding organisms, including 70 present day species, would appear not to utilise glucose at all as a carbon source. It is not yet clear whether in all cases this lack of function is an expression of the limited metabolic capabilities of the corresponding organism or whether this is a result of incomplete genome annotation. For the other organisms glucose expansion accounted for biosynthesis of up to 30% of all organic compounds that are relevant to the set of predicted reactions. This is somewhat below expectation, as several organisms included in this study are known to be prototrophic on glucose, that is they are able to grow with glucose as the sole carbon source. Even for such organisms, we cannot expect to find with our method a value of or near 100%, which can be seen as follows: The fact that these organisms can grow on glucose as the sole carbon source demonstrates that the intracellular amount of every metabolite necessary for cell growth can be increased. However, this does not mean that these metabolites can be synthesised de novo, that is when they are not present inside the cell at all. As an example, in the glycolytic pathway, two molecules of ATP are required before four molecules of ATP can be produced. Consequently, even though this pathway provides an organism with ATP, it would not be able to work if not a

single molecule of ATP had already been present. Some of the missing compounds may belong to this category while others might correspond to secondary metabolites or degradation products derived from specific substrates. Future investigation will be required to identify whether other missing compounds result from possibly incomplete annotation, faulty or ambiguous reactions or missing cofactors. Indeed, we believe that scope expansion can also be used as a consistency check of available data on metabolic systems.

In single evolutionary steps (one edge in the tree in Fig. 2), we found a weak correlation between the changes in network size and synthesising capacity. This analysis also showed a singularity in that biosynthetic capacity can remain frequently unchanged despite a change in metabolic network (Fig. 3). This singularity is consistent with earlier findings that biosynthetic capacities are essentially robust with respect to deletions of a small number of reactions [16]. For larger evolutionary distances we found, as expected, that differences in synthesising capacity and network structure tend to increase with distance. However, the range of observed changes is larger for the former than for the latter (Fig. 4). This large range of variation indicates that biosynthetic capacities can occasionally undergo drastic alterations, despite their general robustness against structural modifications of the metabolic network.

5 References

1 Heinrich, R., and Schuster, S.: 'The regulation of cellular systems' (Chapman & Hall, New York, 1996.)
2 Stephani, A., Nuño, J.C., and Heinrich, R.: 'Optimal stoichiometric design of ATP-producing systems as determined by an evolutionary algorithm', *J. Theor. Biol.*, 1999, **199**, pp. 45–61
3 Ebenhöh, O., and Heinrich, R.: 'Evolutionary optimization of metabolic pathways. Theoretical reconstruction of the stoichiometry of ATP and NADH producing systems', *Bull. Math. Biol.*, 2001, **63**, pp. 21–55
4 Ebenhöh, O., and Heinrich, R.: 'Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness', *Bull. Math. Biol.*, 2003, **65**, pp. 323–357
5 Vo, T.D., Greenberg, H.J., and Palsson, B.O.: 'Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data', *J. Biol. Chem.*, 2004, **279**, (38), pp. 39532–39540
6 Hatzimanitakis, V., Li, C., Ionita, J.A., Henry, C.S., Jankowski, M.D., and Broadbelt, L.J.: 'Exploring the diversity of complex metabolic networks', *Bioinformatics*, 2005, **21**, (8), pp. 1603–1609
7 Kanehisa, M.: 'A database for post-genome analysis', *Trends Genet.*, 1997, **13**, pp. 375–376
8 Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S. *et al.*: 'From genomics to chemical genomics: new developments in KEGG', *Nucleic Acids Res.*, 2006, **34**, D354–D357
9 Schomburg, I., Chang, A., and Schomburg, D.: 'BRENDA, enzyme data and metabolic information', *Nucleic Acids Res.*, 2002, **30**, (1), pp. 47–49
10 Kauffman, K.J., Prakash, P., and Edwards, J.S.: 'Advances in flux balance analysis', *Curr. Opin. Biotechnol.*, 2003, **14**, pp. 491–496
11 Schuster, S., Fell, D.A., and Dandekar, T.: 'A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks', *Nature Biotechnol.*, 2000, **18**, pp. 326–332
12 Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., and Palsson, B.O.: 'Metabolic pathways in the post-genome era', *TIBS*, 2003, **28**, (5), pp. 250–258
13 Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.L.: 'The large-scale organization of metabolic networks', *Nature*, 2000, **407**, pp. 651–654
14 Wagner, A., and Fell, D.A.: 'The small world inside large metabolic networks', *Proc. R. Soc. Lond. B*, 2001, **268**, pp. 1803–1810
15 Ebenhöh, O., Handorf, T., and Heinrich, R.: 'Structural analysis of expanding metabolic networks', *Genome Inform.*, 2004, **15**, (1), pp. 35–45
16 Handorf, T., Ebenhöh, O., and Heinrich, R.: 'Expanding metabolic networks: scopes of compounds, robustness and evolution', *J. Mol. Evol.*, 2005, **61**, pp. 498–512
17 Ebenhöh, O., Handorf, T., and Heinrich, R.: 'A cross species comparison of metabolic network functions', *Genome Inform.*, 2005, **16**, (1), pp. 203–213
18 Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L. *et al.*: 'Database resources of the National Center for Biotechnology Information: update', *Nucleic Acids Res.*, 2004, **32**, D35–D40
19 Bru, C.: 'Analyse évolutive des familles de domaines protéiques', Dissertation, Thesis, Université Paul Sabatier, Toulouse, France, 2005
20 Murphy, K.P.: 'The Bayes net toolbox for Matlab', *Comput. Sci. Stat.*, 2001, **33**, pp. 331–350
21 Dempster, A.P., Laid, N.M., and Rubin, D.B.: 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Stat. Soc. B*, 1977, **39**, pp. 1–38