Constructing Directed Metabolic Networks from Microarray Data

J. M. Easton, principal investigator (funded by the EPSRC), T. N. Arvanitis, co-investigator/Ph.D. supervisor (Senior Lecturer)
Biomedical Informatics, Signals & Systems Research Laboratory,
Department of Electrical, Electronic and Computer Engineering, University of Birmingham, UK
A. Peet, co-investigator/Ph.D. supervisor (Senior Research Fellow)
Institute of Child Health, University of Birmingham & Birmingham Children's Hospital NHS Trust, Birmingham, UK.
M. Viant, co-investigator/Ph.D. supervisor (NERC Advanced Fellow in Metabolomics)
School of Biosciences, University of Birmingham, UK

Abstract

Although it has been several years since metabolic networks became a commonly used analysis technique in bioinformatics, the question of how best to construct them from experimental data is still not satisfactorily resolved. Correlation-based techniques offer a quick, easy and largely unbiased way of discovering the network topology but fail to provide the directional information required for anything other than a cursory analysis of the system. Methods based on the layering of metabolite data onto the pre-existing metabolic pathway diagrams found in repositories, such as the 'Kyoto Encyclopedia of Genes and Genomes' (KEGG)[8], can force a modular structure onto the resultant network, a process which may mask other important topological features hidden within the data. Here we present a method for the construction of metabolic networks from microarray datasets using an enhanced version of the KEGG reaction database. The resulting networks are directed graphs, derived from the association between array probes and enzymes in the KEGG database. The directionality of the network is determined by the flow of catalytic reactions within the associated pathways. This method is novel in the sense that the derived directed metabolic networks are produced without the bias introduced by mapping metabolite data onto the pre-existing topology of a pathway.

Problem

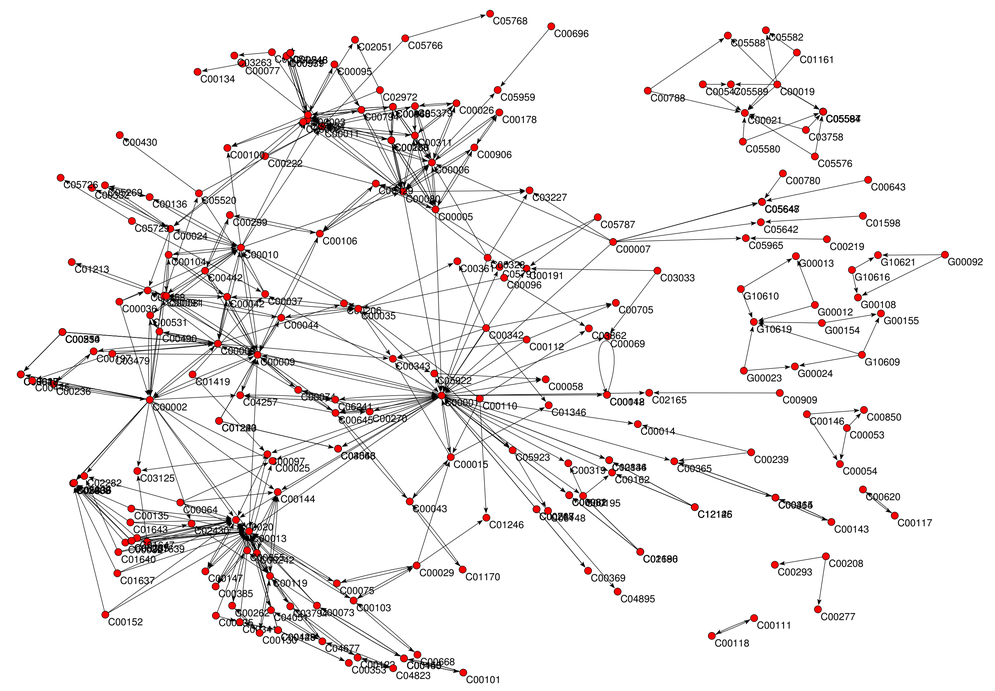
While it has become accepted in the biological community that "...a discrete biological function can only rarely be attributed to an individual molecule" [1] the question of how best to analyse the complex web of interactions now thought to be at the heart of many biological systems is still in debate. One widely used technique, that of metabolic networks, offers great promise in that it is complementary to the traditional biological approach to studying metabolism (the metabolic pathway), while also being able to draw on well-established tools & metrics from fields such as Physics and Communications where networks have been studied for many years. However, the key to a successful network analysis often lies in choosing the correct way in which to represent the field-specific data, and challengingly it is here where some of the greatest variability between research groups lies. Common approaches range from the mapping of data onto pre-defined reference pathways[3][5] found in repositories such as KEGG[8], through more traditional bioinformatics techniques such as correlation-based analysis[4] to the study of time-series data of metabolic fluxes[9][11]. Each has its advantages and drawbacks; the mapping of data onto pre-defined pathways has the advantage of being 'safe' in that the links in the network being created have already been proven experimentally, but fails to allow for alternative reaction mechanisms/pathways which may become dominant when the system is placed under stress. Correlation-based techniques allow a more 'exploratory' approach to be taken by not forcing a previously assumed architecture (which in the case of biological systems is often either modular or hierarchical) onto the system. However, the networks produced lack the directional information necessary for in-depth analysis and the results can be difficult to relate to underlying chemical processes; as Steuer, Kurths, Fiehn and Weckwerth put it "...there is no straightforward connection between the observed correlations and the underlying reaction network. We observe strong correlations between seemingly distant metabolites, whereas metabolites sharing a common reaction are not necessarily correlated."[12]. Time-series methods have produced promising results for small systems (usually no more than 10 metabolites) and *in-silico* simulations, but the number of timepoints and replicants required for the analysis of larger systems make them too impractical to be widely used. What is required therefore, is a method of building directed metabolic networks from experimental datasets which allows *a-priori* information on known metabolic reactions to be included, while at the same time displaying the dynamic architectural discovery that can be seen in correlation-based networks.

Method

The amount of information contained in a directed metabolic network means that it is hard, if not impossible, to build such a model from data on metabolite concentrations alone (at least without performing a very large number of experiments); instead it was decided to build the networks based on a microarray analysis of the sample, which metabolite concentration data could then be mapped onto at a later date. An enhanced version of the reaction component of the KEGG LIGAND[8] database was constructed containing not only definitions for each reaction but also pathway specific directional information which would allow the networks produced to be tailored to specific areas of metabolism should such information be available. The construction process itself works as follows: probes on the array are matched to their gene symbols, which are in turn linked to reactions in the KEGG database through the enzymes they produce. P-values are calculated for the up/downregulation of each of the probes on the array. Thereafter, reactions are added to either the upregulated or downregulated metabolic network if all their associated probes are significantly altered in expression. Reactions are only added to the networks if the regulation of all of their probes is of the same type, i.e. all up or all down. If there appears to be a significant amount of noise in the experimental data, which may lead to difficulties in calculating the p-values, the user also has the option to include a 'grey area' in the construction process; in this situation reactions are included if a minimum of one of the probes is significantly up/downregulated and the others fall inside the 'grey area'.

Preliminary Results

The data used in the testing phase of the work has been taken from a study of circulating tumour cells (CTCs) by Smirnov *et al.*[10] and consists of microarray data for three different types of cancers, breast, colorectal and prostate. For each cancer in their study the authors took a CTC-enriched & CTC-depleted blood sample, before running a microarray analysis. P-values for the significance of expression in the CTC-enriched fraction relative to the expression in the CTC-depleted fraction of the data were calculated using the Kruskal-Wallis test. An initial analysis of the networks produced from this dataset has shown several interesting results; in the upregulated probe network from the breast cancer sample nodal degrees have shown Isocitrate to be an important component of the system, consistent with the increased activity of Isocitrate Dehydrogenase which has been seen in some breast cancer patients[6]. The same technique has also shown Adenylosuccinate to be of interest in colorectal tumours, consistent with the increase in Adenylosuccinate lyase noted in the literature[13]. When directly comparing the networks a large number of interesting variations can be seen; for example, in the upregulated probes network for the breast cancer sample a number of reactions from the glycolysis pathway can be seen, in the colorectal cancer sample however these reactions are split across the upregulated and downregulated probes networks showing a shift from the glycolysis to the gluconeogenesis section of the pathway for that tumour type. Only around 10% of the edges are shared between networks showing that there is very little overlap between the systems. This is highly significant as it proves that the features being seen are due to the variations in the sample rather than an anomaly of the construction process. Each of the networks exhibits a power law degree distribution indicating that, as expected for biological systems, they possess a scale-free topology; an architecture which is resilient to attack against anything other than the highly connected 'hubs'.



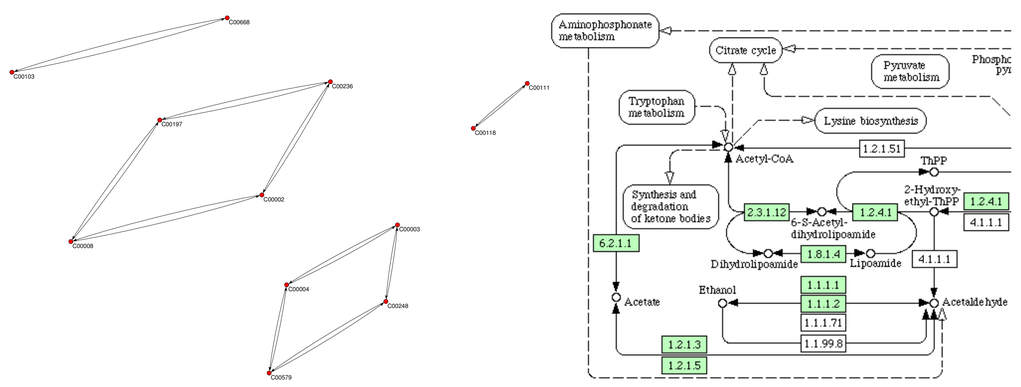


Figure 1: (Above) The directed metabolic network of reactions associated with upregulated probes for circulating tumour cells from a breast cancer, and (below) a sub-graph from the metabolic network featuring only reactions associated with the glycolysis pathway shown alongside the appropriate KEGG pathway diagram. It can bee seen that Lipoamide(C00248) shares a bidirectional link with Dihydrolipoamide(C00579) and that they are also connected to NAD+(C00003) and NADH(C00004), which also feature in the same reaction although are not shown on the KEGG diagram; however there is no link to Thiamine Diphosphate(C00068) since the P-values for probes associated with enzyme 1.2.4.1 fell outside the threshold value of P=0.05. Note that while all the links shown in the glycolysis sub-graph below are bidirectional this is not true in the full metabolic network above. Network images were produced using the Pajek graph visualisation program.[2]

Future Research

- Validation of the technique through a comparison of the networks constructed to those created by other methods (such as correlation-based analysis) and careful checking of key compounds (indicated by hubs in the network) & transformations (indicated by critical sections) against those in the literature.
- Expansion of the database which currently only includes those reactions associated with probes on the Affymetrix Human Genome Focus Array[7], a chip with approximately 8,700 probes measuring around 8,400 genes and the smallest in the Human Genome Arrays range. The choice of array was based solely on the datasets available, and for the technique to be useful to a wide range of groups the database needs to be expanded to cover all of the reactions which currently feature in KEGG LIGAND.
- Further testing of the method against other datasets and analysis of the networks produced using a range of different techniques; vulnerability metrics, clustering coefficients, betweenness centrality, nodal reciprocity, degree distributions, triadic census & motif search.
- Construction of a graphical user interface and automation of the scripts which control the construction process.

References

[1] A.-L. Barabási and Z.N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, February 2004.

[2] V. Batagelj and A. Mrvar. Pajek 1.01 - program for large network analysis. Homepage: http://vlado.fmf.uni-lj.si/pub/networks/pajek, October 2004.

[3] T. Dwyer, H. Rolletschek, and F. Schreiber. Representing experimental biological data in metabolic networks. In *Proceedings of the second conference on Asia-Pacific bioinformatics*, volume 55 of *ACM International Conference Proceeding Series*, 2004.

[4] O. Fiehn. Metabolic networks of cucurbita maxima phloem. *Phytochemistry*, 62:875–886, 2003.

[5] P. Grosu, J.P. Townsend, D.L. Hartl, and D. Cavalieri. Pathway processor: A tool for integrating whole-genome expression results into metabolic networks. *Genome Research*, 12(7):1121–1126, 2002.

[6] R. Hilf, E.D. Savlov, W.D. Rector, and J.L. Wittlief. Relationship of glycolytic enzyme activities and response of breast cancer patients to chemotherapy. *Cancer*, 38(2):695–700, 1976.

[7] Affymetrix Inc. Genechip® human genome arrays data sheet. Download from: http://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf, 2003.

[8] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[9] H. Schmidt, K. Cho, and E. Jacobsen. Identification of small-scale biochemical networks based on general type system perturbations. *The FEBS Journal*, 272:2141–2151, 2005.

[10] D.A. Smirnov, D.R. Zweitzig, B.W. Foulk, M.C. Miller, G.V. Doyle, K.J. Pienta, N.J. Meropol, L.M. Weiner, S.J. Cohen, J.G. Moreno, M.C. Connelly, L.W.M.M. Terstappen, and S.M. O'Hara. Global gene expression profiling of circulating tumor cells. *Cancer Research*, 65(12):4993–4997, June 2005.

[11] E. Sontag, A. Kiyatkin, and B.N. Kholodenko. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, 20(12):1877–1886, 2004.

[12] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth. Interpreting correlations in metabolic networks. *Biochemical Society Transactions*, 31(6):1476–1478, 2003.

[13] L. Terzuoli, F. Carlucci, A. De Martino, B. Frosi, B. Porcelli, C. Minacci, R. Vernillo, L. Baldi, E. Marinello, R. Pagani, and A. Tabucchi. Determination of p185 and adenylosuccinate lyase (asl) activity in preneoplastic colon lesions and intestinal mucosa of human subjects. *Clinical Biochemistry*, 31(7):523–528, 1998.