Modelling and simulation for metabolomics data analysis

P. Mendes[1], D. Camacho and A. de la Fuente
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington St., MC 0477, Blacksburg, VA 24061, U.S.A.

Abstract

The advent of large data sets, such as those produced in metabolomics, presents a considerable challenge in terms of their interpretation. Several mathematical and statistical methods have been proposed to analyse these data, and new ones continue to appear. However, these methods often disagree in their analyses, and their results are hard to interpret. A major contributing factor for the difficulties in interpreting these data lies in the data analysis methods themselves, which have not been thoroughly studied under controlled conditions. We have been producing synthetic data sets by simulation of realistic biochemical network models with the purpose of comparing data analysis methods. Because we have full knowledge of the underlying 'biochemistry' of these models, we are better able to judge how well the analyses reflect true knowledge about the system. Another advantage is that the level of noise in these data is under our control and this allows for studying how the inferences are degraded by noise. Using such a framework, we have studied the extent to which correlation analysis of metabolomics data sets is capable of recovering features of the biochemical system. We were able to identify four major metabolic regulatory configurations that result in strong metabolite correlations. This example demonstrates the utility of biochemical simulation in the analysis of metabolomics data.

Introduction

Powerful analytical technologies are becoming ubiquitous in biology, which are characterized by high-throughput parallel measurement of large numbers of molecular species. At the forefront are microarrays for nucleic acids and proteins and MS methods for proteins and metabolites. Even though these technologies are quite diverse in their details, they are similar in three important aspects. First, all of them are aimed at producing quantitative measurements, even if this may still not be fully realized yet. Secondly, they all enable a different way of carrying out science, in which biological systems are characterized by capturing comprehensive, and largely unbiased, snapshots of the state of the biological system. Thirdly, they have revealed many challenges in terms of how the data they generate should be best analysed. Issues about data analyses range from statistics, such as how best to design experiments [1], to higher level inferences of biological organization, for example how to uncover regulatory networks [2]. It is this third aspect of data analysis that is the focus of this paper.

Recently, researchers have been paying more attention to metabolite profiling as an extension of functional genomics. It has been postulated that metabolite profiles of the internal state of cells could aid in the identification of the function of genes, especially when mutants in that gene have no apparent phenotype [3]. The rationale is that, while the mutation would have caused an effect, the regulatory mechanisms of the cell counteract that effect, resulting in no clear macroscopic observation. However, these internal regulatory mechanisms would have changed the concentrations of metabolites in the metabolic network, and one would be able to identify the function of the mutated gene by determining the resulting metabolite profiles [3–6]. The first implication here is that, not knowing *a priori* which metabolites are expected to be altered, the profiles then need to be as comprehensible as possible; the second implication is that studying the metabolite profiles of a number of mutants might lead to the discovery of the underlying metabolic network [7–9]. This application of metabolite profiling in functional genomics is similar to transcript and protein profiling, and as with these, it would be useful to establish the complete composition of the cell in terms of metabolites – the metabolome. Metabolomics is the study of cells by measuring the profiles of all of their metabolites, or at least of a large number thereof. However, the utility of metabolomics is not restricted to functional genomics. It is useful whenever an assessment of changes in metabolite concentrations is important. Examples already exist for applications in assessing responses to environmental stress [10], toxicology [11], drug discovery [12], nutrition [13], cancer [14], diabetes [15] and natural product discovery [16]. In these applications, the focus is often on identifying metabolites that can statistically discriminate between samples – biomarkers. Metabolite profiling, whether targeted or untargeted, can also be applied as a tool in systems biology [17,18], where metabolite snapshots [19] are used to study cellular dynamics [6,20,21] through mathematical models [22,23].

Many data analysis algorithms have been proposed to analyse functional genomic data and several are in wide use. However, the performance of such algorithms has rarely

Key words: biochemical network, metabolomics, microarray, modelling, simulation, systems biology.

[1]To whom correspondence should be addressed (email mendes@vt.edu).

been demonstrated unequivocally regarding the type of information that they recover from the data. Additionally, these algorithms have rarely been compared in objective terms, in part because they may require different types of data and experimental designs. Arguments can be made to the extent that actual experimental data are of little value to assess the performance of such algorithms because the molecular networks underlying the data are not known with complete accuracy [24]. Many algorithm performance measures require such complete knowledge, for example the assessment of their accuracy in network inference.

We propose that data generated from simulations of experiments carried out from mathematical models of biochemical networks are ideally suited for the role of benchmarking data analysis algorithms [24,25]. These mathematical models have the advantage that they are completely known and so can be the basis for accurate assessments of performance. They are also well suited to compare algorithms that require different experiments because one can use the same networks to simulate the experiments required by each method; given that they would be analysing the same system, it would be easy to compare their results in terms of the characteristics of the (*in silico*) biological system. Another advantage of using biochemical models for this purpose is that their simulation provides essentially exact results (to an arbitrary precision), and one can add to them well-defined sources of noise or other experimental artifacts like restricted dynamic range. This allows for the assessment of the robustness of such methods to the level of noise in the measurements and many other similar properties.

An important consideration to be made is how noise or other experimental artifacts should be simulated. We recognize three different types of variance that can be added to the data: additive noise, intrinsic noise and biological variance. Additive noise is the type of noise introduced by the measurement apparatus and which is added to the real values. Additive noise is incorporated into simulated data by simply adding appropriate random values to the data, after simulation. Intrinsic noise is generated by time-dependent processes, an example being thermal noise that can affect the behaviour of the actual biochemical network. This type of noise must be injected into the network continuously and at all points in the network; an appropriate way to simulate these systems is through the use of stochastic differential equations (e.g. [26]). Biological variation, which is not truly a form of noise, is due to small differences in intrinsic properties between cells, cell cultures or individuals. A way of representing biological variation is to make each biological unit different from the others in the level of all proteins of the network by small amplitude random values [27]. This is easily performed before the simulation is carried out, each simulation being slightly different from the others. Obviously, real experimental data are subject to all these types of variance and possibly others, but the effect of each of these components can be studied independently or as a group.

Another artifact that is likely to be present in all analytical measurements arises from the limited dynamic range of the measuring apparatus. Dynamic range effects make all values below a certain threshold (limit of detection) zero, while all values above another threshold (saturation level) are equal to a maximal value. Dynamic range effects reduce the range of scales that the data may contain and are likely to mask some of the properties of the network. A similar problem relates to the frequency of sampling, which limits the temporal scales that are observed in an experiment. Such effects may be more limiting than noise in some circumstances and so it is important to include them in studies too.

Biochemical network models

Biochemical network models can be used to predict, to explain and to hypothesize about phenomena. When made quantitative and implemented in computer software, models can be used to carry out large numbers of simulations that are designed to answer 'what-if' questions. There are currently several software applications (e.g. [28]) that make the process of modelling biochemical networks and use them to simulate data for the purpose of assessing the efficiency of analysis algorithms. In order to fulfil the objectives delineated above in a relevant manner, it is important that the biochemical models used reflect the properties of real biological systems. Thus good candidate models are those that have been successful in representing biochemical networks and passed validation. At times, it may also be necessary to purposely create artificial models that, while not representing any real biological system, collectively contain a sufficiently wide range of properties that make them inclusive of the properties of real systems. Realistic models are important to argue about the relevance of analysis algorithms, while artificial models are important to argue about their robustness.

Examples

We have started using biochemical network models and simulation to study the utility of application of clustering algorithms to gene expression time-course data [25]. There the model was a simple branched metabolic pathway with two alternative substrate sources and including its transcriptional regulation. The combination of metabolism and gene expression was needed in this case because we were assessing how well the clustering algorithms could identify the metabolic pathway using only gene expression data. The regulation of gene expression in that network was affected indirectly by the metabolic pathway behaviour because there were feedback mechanisms of some of the metabolites on to transcription. Even though the pathway and gene networks of this model were extremely simple (six genes and six enzymes only), this exercise revealed that gene expression data analysed by clustering methods were not able to identify the metabolic pathways.

Another application has been the study of the performance of nonlinear parameter estimation algorithms [17,29]. In this case, we constructed a model that includes a linear metabolic pathway of only three enzymes, representing explicitly the enzyme synthesis and degradation and also transcript synthesis and degradation. As in the previous example,

this model was simple (eight variable concentrations only) but proved extremely hard to tackle by nonlinear least-squares methods. One of the reasons for this was the large dimensionality of the parameter space (36 parameters) and the nonlinearity of the kinetics. A conclusion of this study was that evolutionary algorithms seem to outperform all other methods tested [17,29].

In order to study several aspects of microarray statistical analysis and gene network inference algorithms, a set of artificial gene network models was created representing a wide range of topological characteristics [24]. These models illustrate a case where there is still ambiguity in our knowledge of the topology of real gene networks, and so a wide range was created. This is then being used to study several microarray analysis methods, exploring, among other characteristics, the algorithms' robustness towards the topology of the underlying networks. Robust algorithms inspire more confidence than those that only work well if the gene network follows a particular topology.

More recently, we have constructed a realistic model of yeast carbohydrate metabolism [30], by combining a model of glycolysis [31,32] with a model of glycerol synthesis [33]. This was used to study the type of knowledge that can be recovered from observation of metabolite correlations in metabolomics data. Simulations using this model associated with biological variation were instrumental in revealing a number of regulatory characteristics that originate rare but strong metabolite correlations, such as near-equilibria, moiety-conservation and asymmetric concentration-control distributions [27]. A similar approach was used in the context of intrinsic noise with similar results [26].

Conclusion

Biochemical modelling and simulation are becoming an important method to study data analysis algorithms in systems biology. This approach is expected to be particularly important in comparing competing data analysis methods that require considerably different experimental setups. In that case, modelling may be the only way to be able to study their performance in a truly comparative way.

References

1 Kerr, M.K. and Churchill, G.A. (2001) Biostatistics **2**, 183–201
2 de la Fuente, A., Brazhnik, P. and Mendes, P. (2002) Trends Genet. **18**, 395–398
3 Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998) Trends Biotechnol. **16**, 373–378
4 Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J. et al. (2001) Nat. Biotechnol. **19**, 45–50
5 Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L. and Bino, R. (2002) Plant Cell **14**, 1437–1440
6 Phelps, T.J., Palumbo, A.V. and Beliaev, A.S. (2002) Curr. Opin. Biotechnol. **13**, 20–24
7 Fiehn, O. (2001) Comp. Funct. Genom. **2**, 155–168
8 Li, X.J., Brazhnik, O., Kamal, A., Guo, D., Lee, C., Hoops, S. and Mendes, P. (2002) in Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis (Harrigan, G.G. and Goodacre, R., eds.), pp. 293–309, Kluwer Academic Publishers, Boston, Dordrecht and London
9 Fiehn, O. and Weckwerth, W. (2003) Eur. J. Biochem. **270**, 579–588
10 Viant, M.R., Rosenblum, E.S. and Tieerdema, R.S. (2003) Environ. Sci. Technol. **37**, 4982–4989
11 Nicholson, J.K., Lindon, J.C. and Holmes, E. (1999) Xenobiotica **29**, 1181–1189
12 Watkins, S.M. and German, J.B. (2002) Curr. Opin. Mol. Ther. **4**, 224–228
13 German, J.B., Roberts, M.A., Fay, L. and Watkins, S.M. (2002) J. Nutr. **132**, 2486–2487
14 Griffiths, J.R. and Stubbs, M. (2003) Adv. Enzyme Regul. **43**, 67–76
15 Watkins, S.M., Reifsnyder, P.R., Pan, H.J., German, J.B. and Leiter, E.H. (2002) J. Lipid Res. **43**, 1809–1817
16 Huhman, D.V. and Sumner, L.W. (2002) Phytochemistry **59**, 347–360
17 Mendes, P. (2001) in Foundations of Systems Biology (Kitano, H., ed.), pp. 163–186, MIT Press, Cambridge, MA
18 Weckwerth, W. (2003) Annu. Rev. Plant Biol. **54**, 669–689
19 Kell, D.B. and Mendes, P. (2000) in Technological and Medical Implications of Metabolic Control Analysis (Cornish-Bowden, A. and Cárdenas, M.L., eds.), pp. 3–25, Kluwer Academic Publishers, Dordrecht
20 Theobald, U., Mailinger, W., Baltes, M., Rizzi, M. and Reuss, M. (1997) Biotechnol. Bioeng. **55**, 305–316
21 Buchholz, A., Hurlebaus, J., Wandrey, C. and Takors, R. (2002) Biomol. Eng. **19**, 5–15
22 Rizzi, M., Baltes, M., Theobald, U. and Reuss, M. (1997) Biotechnol. Bioeng. **55**, 592–608
23 Mendes, P. (2002) Brief. Bioinformatics **3**, 1134–1405
24 Mendes, P., Sha, W. and Ye, K. (2003) Bioinformatics **19**, ii122–ii129
25 Mendes, P. (1999) in Workshop on Computation of Biochemical Pathways and Genetic Networks (Bornberg-Bauer, E., De Beuckelaer, A., Kummer, U. and Rost, U., eds.), pp. 27–34, Logos-Verlag, Heidelberg
26 Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003) Bioinformatics **19**, 1019–1026
27 Camacho, D., de la Fuente, A. and Mendes, P. (2005) Metabolomics **1**, 53–63
28 Mendes, P. (1997) Trends Biochem. Sci. **22**, 361–363
29 Moles, C.G., Mendes, P. and Banga, J.R. (2003) Genome Res. **13**, 2467–2474
30 Martins, A.M., Camacho, D., Shuman, J., Sha, W., Mendes, P. and Shulaev, V. (2004) Curr. Genomics **5**, 649–663
31 Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., van der Weijden, C.C., Schepper, M., Walsh, M.C., Bakker, B.M., van Dam, K., Westerhoff, H.V. et al. (2000) Eur. J. Biochem. **267**, 5313–5329
32 Pritchard, L. and Kell, D.B. (2002) Eur. J. Biochem. **269**, 3894–3904
33 Cronwright, G.R., Rohwer, J.M. and Prior, B.A. (2002) Appl. Environ. Microbiol. **68**, 4448–4456

Received 27 July 2005