Automated extraction of meaningful pathways from quantitative proteomics data

Josselin Noirel, Saw Yen Ow, Guido Sanguinetti, Alfonso Jaramillo and Phillip C.Wright

Advance Access publication date 7 March 2008

Abstract

Technological developments in the life sciences have resulted in an ever-accelerating pace of data production. Systems Biology tries to shed light upon these data by building complex models describing the interactions between biological components. However, extracting information from this morass of data requires the use of sophisticated computational techniques. Here, we propose a method suitable to integrate data drawn from quantitative proteomics into a metabolic scaffold and identify the metabolic pathways which are collectively up-regulated or down-regulated. The availability of such a tool is highly desirable as the extracted information could then be taken as a starting point for in-depth analyses, in particular in fields like Synthetic Biology, where datasets need be characterized routinely.

Keywords: metabolic network; quantitative proteomics; data mining

INTRODUCTION

Although science rests on human creativity and intuition [1], much effort has been and is still invested in the development and improvement of methods to ease the automatic extraction of biological information from the large and ever increasing amount of data available [2]. Systems Biology tries to describe biological systems as a whole by means of integrated approaches. It generally consists in combining the large-scale data produced during an experiment with the underlying cellular or molecular structures: the regulation networks, the protein–protein interaction network and the metabolic network. Although a simple gene list (the ‘gene list syndrome’ [3]) or a mere projection onto these different networks appear to be the simplest approaches available [4, 5], they are not the most efficient, and methods have been devised to process and further analyse the experimental data [3, 6, 7].

The emergence of microarrays has encouraged the development of many such tools [3, 6], which are much less common in quantitative proteomics because of its later rise. However, the recent technological improvements in this field, amongst which the 4- and 8-plex kits for iTRAQ-based experiments (trademarked by Applied Biosystems [8]), make this technology extremely promising,

Corresponding author. Phillip C. Wright, Biological and Environmental Systems Group, Department of Chemical and Process Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK. Tel: +44 114 2227577; Fax: +44 114 2227501; E-mail: p.c.wright@sheffield.ac.uk

Josselin Noirel is Postdoctoral Research Associate in the Department of Chemical and Process Engineering at the University of Sheffield, his research is directed towards the integration of high-throughput data into metabolic and regulatory networks, and the biological knowledge that can be drawn from this integration.

Saw Yen Ow is a PhD student in The Department of Chemical and Process Engineering at The University of Sheffield, with research interests in high throughput expression proteomics and its application to emerging cyanobacteria based systems biology.

Guido Sanguinetti is a Lecturer in Computer Science at the University of Sheffield. His interests centre on the development and application of statistical machine learning tools for modelling complex systems, in particular at the interface with the life sciences.

Alfonso Jaramillo is Assistant Professor in the Department of Biology at the École Polytechnique (France), with research interests in applying computational biology to the design of new biological systems to be used in biotechnology and biomedical science.

Phillip Wright is Professor of Systems Biology and Engineering in The Department of Chemical and Process Engineering at The University of Sheffield, with research interests in applying post-genomics tools to problems in biotechnology and biomedical science.
with applications ranging from metabolic control analysis [9] to Synthetic Biology [10]. Very little work, if any, has been devoted to the integration of the mass spectrometry data into metabolic networks. It is important to realise here that the software developed for microarray data analysis cannot be directly used. One distinctive feature of quantitative (and indeed any) proteomics is its usually fairly incomplete coverage: a few hundred proteins can be typically quantified [11, 12] and up to the few thousands of proteins can be identified [8, 13]. This introduces a great deal of uncertainty when a particular pathway is not sampled correctly, as there is much missing information. For instance, only 17% of the metabolic network is covered by the data analysed in the ‘Results’ section.

Even though substantial information may be made accessible by merely listing the enzymes that are up-regulated, several observations suggest that a method that could automatically detect up-regulated metabolic pathways would be highly beneficial to the scientific community. There are a number of reasons why this may be desirable; these are:

- Manual exploration has proven to be painful and is chiefly hypothesis driven.
- Unless very careful examination is performed, it is hard to see how different up-regulated branches of the metabolism are connected because of the intricacy of the metabolic network.
- By generating pathways where there are non-quantified proteins, such a method could suggest further experiments by pinpointing potential key enzymes.
- The routine exploration of the impact of a condition change in the cell is of utmost interest for the synthetic biologists who may produce a great number of different constructs and will have to assess the specificity of the genetic circuits they are trying to implement. In this case, research can hardly be driven by preconceived hypotheses.

The metabolic network, at least at the current level of description, appears to have the very common structure referred to as a ‘small-world scale-free network’, alongside a bow-tie and modular topology [14–16]. This small-world structure has arisen because of evolutionary requirements, because of the mechanisms involved in the creation of new functional units in the cell, or simply because our knowledge is still incomplete. Discussing these different possible causes is beyond the scope of this article; however, whatever the cause may be, the small-world structure has a great impact on the procedure that has to be devised in order to extract the potential pathways of interest.

In short, the small-world network structure, which first emerged from studies on social networks, states that, given a start point and an end point in a graph, there exists a shorter-than-expected path that connects the start and the end points. This phenomenon is also known in social networks as the ‘six degrees of separation’ principle [17]. In practice, it means that it is very easy to connect, through short paths, any pair of nodes. This observation combined with the missing data inherent in proteomics experiments implies that extra care must be taken to avoid the generation of non-meaningful pathways. A most interesting finding was made by Croes et al., [18] who showed that if the connections of the metabolic network were weighted by the number of occurrences of the metabolites, meaningful metabolic pathways could be found with a much higher degree of confidence. This choice counterbalances the ‘small-world’ effect by discouraging the passage through hub metabolites, or ‘currency metabolites’ such as water, ATP, etc., to connect the distant reactions [18, 19]. In developing our methodology, this principle has been taken up.

The cell’s metabolic reactions can be integrated into a variety of mathematical objects. (i) The most detailed abstraction is a weighted directed bipartite graph [20, 21], where some nodes are used to represent enzymes, whilst the others are used to represent metabolites. An irreversible enzymatic
reaction A $\rightarrow$ B + C catalyzed by enzyme E would appear as shown in Figure 1.

Two simpler, still very useful, representations can be used; (ii) the ‘metabolite network’ where the nodes are metabolites and where a pair of metabolites is connected whenever one metabolite is the product of the other in an enzymatic reaction [22]; (iii) the ‘enzyme network’ where nodes are enzymes a pair of which is connected whenever one enzyme can take up one of the other’s products to catalyze its further transformation (this study).

**METHODS**

**Enzyme network**

We used the KEGG database (http://www.genome.jp/kegg/) [23], though in principle other databases could be used [24, 25]. The representation adopted in this work is an ‘enzyme network’, that is to say a network of enzymes, where a pair of enzymes E1 and E2 is connected when one of the reactions catalyzed by E1 and E2 further transforms one of the products obtained from the other reaction.

An enzyme is defined as a unique set of genes. For instance, the $\alpha$, $\beta$ and $\gamma$ subunits of urease will be treated as a single entity. It is important to realize that an enzyme made up of the products of the genes $g_1$ and $g_2$ will be considered different from an enzyme composed of the expression product of gene $g_1$ only.

Because of redundancies and inconsistencies in the KEGG database, the following approach has been chosen: for each enzyme E, three sets of metabolites are attached S, P and R. The metabolite set S is the set of metabolites which are substrates in an irreversible reaction catalyzed by E, P is the set of metabolites which are products in an irreversible reaction catalyzed by E, whereas R is the set of metabolites which are substrates or products in a reversible reaction catalyzed by E (Figure 2).

The connections of the undirected enzyme network are then set up by defining a number of rules as follows: E1 and E2 are connected if and only if there exist a metabolite C such that C belongs simultaneously to S1 and P2, or S1 and R2, or P1 and S2, or P1 and R2, or R1 and S2, or R1 and P2, or R1 and R2; where $\{S_1, P_1, R_1\}$ and $\{S_2, P_2, R_2\}$ are the metabolite sets defined above for E1 and E2, respectively. In other words, a connection can exist between two enzymes only if an output metabolite of one can be the input metabolite of the other.

The network is weighted, that is, to each connection is attached a certain weight which characterizes the cost of travelling through each of them. Following Croes et al. [18] the weight of a connection E1–E2 is the number of occurrences in the overall metabolic network of the metabolite which connects E1 to E2. When several metabolites connect two enzymes, the least common is used. The metabolites having large weights have traditionally been referred to as ‘currency metabolites’ [19].

**Mapping of quantitative data**

Quantitative proteomics techniques like the iTRAQ isobaric labelling approach usually provide relative quantification (although absolute measurements are possible). A number of algorithms are available to generate the quantitative data from the tandem mass spectra [26–28]. Once this data is available, then a number of approaches have been demonstrated to understand and quantify the reliability of this data [12]. Practically, if a population of cells is grown in two conditions A and B, the iTRAQ measurement $f_E$ for a given protein will be the ratio of the absolute quantities (as indirectly measured via ion intensity-type measurements on the mass spectrometer):

$$f_E = \frac{Q_E^{(A)}}{Q_E^{(B)}}$$

where $Q_E^{(A)}$ and $Q_E^{(B)}$ are the absolute abundances of protein E in the cells grown in conditions A and B, respectively. Complications arise when an enzyme is composed of more than one protein (E1 and E2), in which case, the enzymatic activity would be a—possibly complex—function of $Q_{E1} + Q_{E2}$ or $\max(Q_{E1}, Q_{E2})$, depending on the underlying molecular mechanisms. But because the absolute
abundances $Q_{E_i}$ are not usually available (only the ratios $f_{E_i}$ are), and because molecular mechanisms are not specified in the KEGG database, the average value is used:

$$f_{E_1 E_2} = \text{mean}(f_{E_1}, f_{E_2}),$$

and so forth for more complex enzymes. Practically, this situation occurs rarely enough (about 10% of the time, data not shown) not to be regarded as a major issue. Also, for this reason, down-regulated pathways are directly not looked for as they can be extracted as up-regulated pathways if one inverts the relative measurements. If the coverage of the metabolic network were wider and if the data proved pathologically inconsistent, the inversion technique could be problematic. For instance, if the relative expression levels of two proteins are 0.4 and 1.2, the average ratio is 0.8, which would be likely to be considered as 'unchanged'. However, the values resulting from the inversion of the data are 2.5 and 0.83, the average of which is 1.67 and would likely be regarded as ‘up-regulated’.

**Extraction of up-regulated pathways**

Using the previously described method, to each node $E$ of the enzyme network is assigned a value $f_E$ representing the relative measurement of enzyme $E$ (or some reasonable value calculated from the ratios of its constituents). Two parameters $f_{\text{up}}$ and $f_{\text{mid}}$ are introduced, and we will depend on the following terminology: an enzyme $E$ is said to be ‘up-regulated’ if the value $f_E$ is greater than $f_{\text{up}}$, ‘mildly up-regulated’ if $f_E$ is comprised between $f_{\text{mid}}$ and $f_{\text{up}}$, ‘down-regulated’ if it is lower than $f_{\text{mid}}$. An enzyme to which no value $f_E$ can be attached, because none of its constituents has been identified and quantified, is said to be ‘non-quantified’.

The following heuristic method has been devised to extract up-regulated pathways:

**Local exploration**

For each up-regulated enzyme $E_1$

1. Perform a depth-first search through the enzyme network,
2. The aforementioned search generates elementary pathways $p = (E_1, E_2, \ldots, E_n)$ explored in the vicinity of enzyme $E_1$,
3. The pathways $p$’s are filtered according to the following criteria:

- $p$ must not contain any down-regulated enzyme,
- The tail of $p$ cannot be made of more than one non-quantified enzyme,
- $p$ must include more than one up-regulated enzyme (i.e. at least one up-regulated enzyme different from $E_1$),
- $p$ must not contain more than two non-quantified enzymes in a row,
- The weight of $p$ defined by the sum of the weights of the connections it is made of must not exceed the parameter $w_{\text{max}}$ (weight-limited search).

**Construction of the subgraph**

Once the elementary pathways have been generated, they are merged to reconstitute a subgraph of the metabolic network. In other words, the following network is built up: the nodes are those, which have been visited during the local exploration step, and likewise, the connections are those which have been traversed.

**Identification of the connected components**

Finally, the connected components of the subgraph are identified and listed. Before doing so, it is possible to reconsider certain connections between nodes of the subgraph that had been discarded in the first place. The weight of these connections must not exceed $w_{\text{max}}$. This step may help to connect components disconnected in the first place.

Intuitively, the weight of a connection is a measure of the non-specificity of the compound, which is used to connect two enzymatic reactions. For instance, if the parameter $w_{\text{max}}$ is equal to 2, the only connections that may be traversed are those corresponding to a compound which is specific of this connection or, in other words, which appears only twice in the metabolic network: once as the product of the first reaction and once as the reactant of the second reaction. A value $w_{\text{max}} = 8$ will allow the procedure to traverse one compound involved in eight reactions at most, or two compounds involved in four reactions, etc. As $w_{\text{max}}$ becomes lower, the commonest compounds are discarded. In terms of fluxes, the weight tells one how likely a compound is to participate in the metabolic flux identified by the software, as a high weight means that many different reactions might compete for that particular compound. This method was implemented using Perl and is available online http://www.wrightlab.group.shef.ac.uk/.
RESULTS

This heuristic method that aims at discovering the up-regulated pathways, has been run on various datasets already produced by our laboratory at the Department of Chemical and Process Engineering Department at the University of Sheffield. The data presented and analysed in [5] has been processed, and the results are discussed hereafter as an illustration.

Test case

Stensjö and collaborators studied the metabolism of Nostoc sp. PCC 7120 in N2-fixing conditions using the iTRAQ technique [5]. We focused on the metabolic pathways which were up-regulated when the cells were cultivated in BG110 medium with respect to the situation when the cells were cultivated in BG110 supplemented with ammonia. To summarize, the motivation for such a study stems from the need to understand quantitatively the metabolic changes that occur during the fixation of nitrogen in a potentially industrially valuable cyanobacterial strain. The underlying fixation process represents the fundamental process by which potential bio-H2 can be freely generated. To facilitate sufficient silencing of combined-N incorporation, depletion of combined-N, namely those of NO3 and NH4+, was achieved using minimal medium BG110. Nostoc sp. PCC 7120 grown under such conditions that promotes the activation of a chain of processes, which in essence initiates both primary and auxiliary components enabling N2 fixation. These processes have often suggested to be initiated by the NtcA global nitrogen regulator [29–32]. The systematic rationalization of the data obtained from these experiments could in turn aid the understanding of a system-wide response and metabolic clustering mechanisms that are specific towards the bio-engineering of H2 production.

The data was analysed using the method described earlier. The following parameters were used:

\[ f_{up} = 1.6, \quad f_{mid} = 1.1, \quad w_{max} = 10 \]

See the next section for more details about how to choose the parameter values. In brief, sensible values for \( f_{up} \) are around 1.5 according to studies such as the one presented in [12], owing to the technical, experimental and biological variations. The same reasoning leads us to avoid enzymes whose relative expression level is below 1, as they are presumably not significantly up-regulated. The parameter \( f_{mid} \) is set to 1.1 rather than one to be slightly more confident that the enzymes traversed are not down-regulated. Finally, by setting \( w_{max} \) to 10, one ensures that only a few, from 0 to 4, non-quantified or mildly up-regulated enzymes be traversed.

Moreover, because they cannot be assimilated to currency metabolites in this study, the weights of glutamate and pyruvate were set to four.

\[ w_{pyr} = 4, \quad w_{glu} = 4 \]

A single connected component consisting of 30 enzymes emerges from the data processed with the parameters given above. No recomputing stage was necessary to ensure that the network be fully connected. The resultant network is displayed in Figure 3, and its nodes are listed in Table 1.

This connected component consists of four mildly up-regulated enzymes, 12 up-regulated enzymes and 14 non-quantified enzymes. The key elements identified in the study [5] could be retrieved using the methodology described in this article.

In the network identified using our method, all the elements belonging to the oxidative stage of
the pentose phosphate cycle and depicted on Figure 3 in the original paper by Stensjö et al. [5], are included in the network of Figure 3 (enzymes labelled a and a'). However, the topology is slightly affected by a different definition of the metabolic network (see ‘Methods’ section), and because complementary experimental data had been used to produce the network in [5].

This example shows that it is important to allow the software to traverse non-quantified or mildly up-regulated proteins. In our case, about half of the enzymes belong to one of these categories. In particular, the full connectivity of the test network drawn from Stensjö et al.’s [5] data, depends on:

- the glutamate dehydrogenase (alr4255, non-quantified, enzyme labelled d in Figure 3) and the aspartate aminotransferase (alr1039, alr4853, alr5103, mildly up-regulated, enzyme labelled d) on one hand,
- the glutathione reductase (alr4968, mildly up-regulated), the inorganic polyphosphate/ATP-NAD kinase (all4751, alr0227, non-quantified, enzyme labelled b in Figure 3), the nicotinamide nucleotide transhydrogenase (all3408, all3410, alr0227, non-quantified, enzyme labelled b in Figure 3), the 6-phosphogluconolactonase (alr1602, non-quantified, enzyme labelled a in Figure 3), and the gluconokinase (alr3086, non-quantified, enzyme labelled b' in Figure 3), on the other hand.

These different enzymes connect the energy metabolism to the nitrogen metabolism.

Different branches of the metabolism are identified: ‘Glycolysis/gluconeogenesis’, ‘tricarboxylic acid cycle’, and ‘Nitrogen metabolism’ for instance. A linear interpretation of the up-regulated network is suggested by its topology (Figure 3). The leftmost part corresponds to the oxidative stage of the pentose phosphate cycle, which is up-regulated to allow the cell to cope with the increased energy demand. Indeed, the reduction of N₂ to ammonium necessitates 16 ATP and the oxidation of four NAD(P)H molecules. This part of the network includes, importantly, the six up-regulated enzymes (labelled a in Figure 3): the fructose-bisphosphate aldolase, the fructose-bisphosphate aldolase, the glucose-6-phosphate isomerase, the glucokinase, the glucose-6-phosphate 1-dehydrogenase and the 6-phosphogluconate dehydrogenase. Furthermore, recent evidence shows that the 6-phosphogluconolactonase, which is included in the pathway, is up-regulated [33] (enzyme labelled a in Figure 3).

Then comes the isocitrate dehydrogenase (labelled c in Figure 3), which facilitates the production of 2-oxoglutarate carbon backbone which is needed for the later fixation of nitrogen. On one hand, the connection between the pentose phosphate cycle and the isocitrate dehydrogenase is somewhat artificial, since CO₂ is the ultimate link between the 6-phosphogluconate dehydrogenase and the pentose phosphate cycle. The pentose phosphate cycle and the isocitrate dehydrogenase are included in the network of Figure 3 (enzymes labelled a and a').

### Table 1: List of up-regulated enzymes in N₂-fixing conditions (BG110) represented in the network of Figure 3

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Measurement</th>
<th>Enzymatic activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) all3735, all4563</td>
<td>2.478</td>
<td>Fructose-bisphosphate aldolase</td>
</tr>
<tr>
<td>(a) all4019</td>
<td>2.701</td>
<td>Glucose-6-phosphate 1-dehydrogenase</td>
</tr>
<tr>
<td>(a) all5002, alr2973, alr1982</td>
<td>1.475</td>
<td>Glucokinase</td>
</tr>
<tr>
<td>(a) alr1041, alr4021</td>
<td>1.644</td>
<td>Fructose-1,6-bisphosphatase</td>
</tr>
<tr>
<td>(a) alr1050</td>
<td>1.747</td>
<td>Glucose-6-phosphate isomerase</td>
</tr>
<tr>
<td>(a) alr5275</td>
<td>2.231</td>
<td>6-phosphogluconate dehydrogenase</td>
</tr>
<tr>
<td>(a') alr1602</td>
<td>NQ</td>
<td>6-phosphogluconolactonase</td>
</tr>
<tr>
<td>(b) all4968</td>
<td>1.307</td>
<td>Glutathione reductase</td>
</tr>
<tr>
<td>(b) alr0227, all4751</td>
<td>NQ</td>
<td>Inorganic polyphosphate/ATP-NAD kinase</td>
</tr>
<tr>
<td>(b) alr3409, all3410, all3408</td>
<td>NQ</td>
<td>Nicotinamide nucleotide transhydrogenase</td>
</tr>
<tr>
<td>(b) alr3086</td>
<td>NQ</td>
<td>Gluconokinase</td>
</tr>
<tr>
<td>(c) alr1827</td>
<td>1.821</td>
<td>Isocitrate dehydrogenase</td>
</tr>
<tr>
<td>(d) alr4255</td>
<td>NQ</td>
<td>Glutamate dehydrogenase</td>
</tr>
<tr>
<td>(d) alr5103, alr4853, alr1039</td>
<td>1.460</td>
<td>Aspartate aminotransferase</td>
</tr>
<tr>
<td>(e) alr2328</td>
<td>1.793</td>
<td>Glutamine synthase</td>
</tr>
<tr>
<td>(f) all440, all454, all455, alr0874</td>
<td>6.069</td>
<td>Nitrogenase</td>
</tr>
<tr>
<td>(g) alr3887</td>
<td>1.721</td>
<td>Argininosuccinate lyase</td>
</tr>
<tr>
<td>(g) all3401</td>
<td>NQ</td>
<td>Arginine decarboxylase</td>
</tr>
<tr>
<td>(h) all267</td>
<td>1.74</td>
<td>Aconitate hydratase</td>
</tr>
<tr>
<td>(i) all1092</td>
<td>1.792</td>
<td>Cysteinyl-tRNA synthetase</td>
</tr>
<tr>
<td>(j) alr0488</td>
<td>NQ</td>
<td>Pyrroline-5-carboxylate reductase</td>
</tr>
<tr>
<td>(j) alr0540</td>
<td>NQ</td>
<td>1-pyrroline-5 carboxylate dehydrogenase</td>
</tr>
<tr>
<td>(j) alr1270</td>
<td>2.120</td>
<td>Proline iminopeptidase</td>
</tr>
<tr>
<td>(j) all3731</td>
<td>NQ</td>
<td>Polyphosphate glucokinase</td>
</tr>
<tr>
<td>(j) all483</td>
<td>NQ</td>
<td>Putative phosphoketolase</td>
</tr>
<tr>
<td>(j) all3836</td>
<td>NQ</td>
<td>Glucose-1-dehydrogenase</td>
</tr>
<tr>
<td>(j) alr193, all7335</td>
<td>NQ</td>
<td>6-phosphofructokinase</td>
</tr>
<tr>
<td>(j) alr3344, all4052</td>
<td>1.127</td>
<td>Transketolase</td>
</tr>
<tr>
<td>(j) alr3528</td>
<td>NQ</td>
<td>Hypothetical protein</td>
</tr>
<tr>
<td>(j) alr4385</td>
<td>1.323</td>
<td>Triosephosphate isomerase</td>
</tr>
</tbody>
</table>

Alongside the gene names, the label used in Figure 3 is indicated within parentheses. ‘NQ’ in the measurements’ column stands for ‘non-quantified’.
and the isocitrate dehydrogenase itself. On the other hand, the pathways traversing the inorganic polyphosphate/ATP-NAD kinase, the nicotinamide nucleotide transhydrogenase, and the glutathione reductase (enzymes labelled b in Figure 3) exist due to the reduction of NADP⁺ into NADPH, which are currency metabolites.

The isocitrate dehydrogenase (labelled c in Figure 3) is connected to the glutamate dehydrogenase (non-quantified, labelled d’ in Figure 3) and the aspartate aminotransferase (mildly up-regulated, labelled d in Figure 3), two enzymes which supply the carbon skeleton in the form of glutamate molecules, to incorporate the ammonia into organic molecules thanks to the glutamine synthase (up-regulated, labelled e in Figure 3). The ammonia itself results from the chemical reaction catalyzed by the nitrogenase (labelled f in Figure 3), which is highly up-regulated.

It is interesting to notice that the network suggests other possible routes to supply the glutamate used as a substrate by the glutamine synthetase (labelled e in Figure 3). Amongst these is the arginine decarboxylase (labelled g’ in Figure 3) which catalyzes a reversible reaction converting 4-aminobutyrate into glutamate. The aconitate hydratase (labelled h in Figure 3) is only mildly up-regulated. Nonetheless, it is constantly up-regulated in two replicates out of three. This enzyme could also furnish the isocitrate dehydrogenase (labelled c in Figure 3) with isocitrate produced from citrate or aconitate. Careful scrutiny shows that the other routes (proline, argininosuccinate, cysteine degradation, enzymes labelled i, j and g in Figure 3) are very unlikely to play the same role, as the connecting metabolites are unrelated to glutamate production.

A similar analysis has been performed with inverted data to look for down-regulated networks but no down-regulated subnetwork could emerge as few enzymes in the metabolic network are markedly down-regulated. The fundamental understanding behind the fixation of N₂ have postulated strong suppression of active Photosystem II proteins and some others because of their detrimental effects on the N₂-fixing process in N₂-fixing filaments. This was supported by transcriptional analysis experiments carried out on purified cellular fractions [31]. That conclusion cannot be drawn from the iTRAQ analysis, which moreover must overcome quantitative disparities due to the seemingly poor population of heterocysts as compared with their parent cells (< 10%). However, using purified cells, we could identify down-regulated subnetworks (e.g. in the pyrimidine metabolism).

**Estimation of the parameters**

Statistical assessment is hard because of the small-world feature of the metabolic network, even when the network is weighted as indicated in the ‘Methods’ section. Nonetheless, we need a measure of plausibility for the network to estimate the best parameters. To achieve this estimation, the mass-spectrometry data was randomized 50 times; in each randomized dataset, the same proteins are identified and quantified, and the same measurements appear in order to preserve the basic statistical properties of the measurements’ distribution; however the measurements are re-assigned randomly. Therefore, in these randomized datasets, the same proteins are identified and quantified, but their value is drawn randomly from the original distribution. It is a reasonable assumption, as the intersection between two sets of quantified proteins drawn from two independent experiments represents typically from 70 to 80% of either protein list. The rationale behind this choice is that such randomized datasets, while still preserving realistic features, lose the correlation that may be explained by the regulatory consistency along the metabolic pathways. For each one of these datasets and for each parameter setting, the up-regulated network is computed. Generally, the networks which emerge from the randomized datasets look different, as depicted in Figure 4.

The parameters \(w_{\text{up}}\) and \(f_{\text{up}}\) influence dramatically the results, regarding the network’s topology and size. The influence of the parameter \(f_{\text{mid}}\) is comparatively more limited. Low values of \(w_{\text{up}}\) and high values of \(f_{\text{up}}\) imply a high stringency during the local exploration; as a consequence, only small networks constituted of tightly-clustered highly up-regulated enzymes, if any, will be found. By increasing the weight tolerance and by reducing the \(f_{\text{up}}\) parameter, many more pathways can be traversed and different parts of the metabolism can be bridged. For instance, the size of the up-regulated increases roughly linearly with \(w_{\text{up}}\) (data not shown). On the other hand, too much tolerance makes it more likely to generate implausible or biologically irrelevant pathways. In the lack of additional experimental information, the ideal parameter setting cannot be estimated rigorously.
but we can monitor when the network found with the real data departs most from the networks drawn from the randomized datasets. A very simple measure can be devised for a rough estimation of the parameter values. Figure 5 shows that a value of 10 is most likely to tell apart the network drawn from the actual data and the networks drawn from the randomized datasets. This conclusion is supported by the biological interpretation. It would be illusory however to think that there could exist an optimal value. The suitable value of $w_{\text{max}}$ is the value that produces a network, which can be supported by other observations or sensible hypotheses. Although at this stage, it seems hard to measure the significance of the inferred network, probabilistic analyses are under development.

CONCLUSIONS

The methodology detailed and developed in this article was implemented to ease the interpretation from a metabolic standpoint as well as to devise testable biological information based on high-throughput quantitative proteomic data. It allows us not only to integrate the data to comply with the paradigm of Systems Biology, but to emphasize the collective behaviour of a group of biologically related enzymes, since it is believed that a small but coherent difference in the expression of all the genes in a pathway should be more significant than a larger difference occurring in unrelated

Figure 4: Typical topologies of up-regulated networks generated using randomized datasets. The parameters used are the same as those used to generate the network in Figure 3. Most of the time, several connected compounds are found (A–D) but not always (E and F). Very often, star patterns are found (C–E). (F) shows a striking example of network drawn from randomized data whose topology is very similar to that of the network of Figure 3.

Figure 5: Number $n$ of distinct connected components as a function of the value of the parameter $w_{\text{max}}$. The dashed line is obtained from the actual data whereas the solid line is obtained by averaging the results drawn from the randomized datasets (alongside the SD at each point).
genes’ [3]. Most importantly, this tool addresses the question of the integration of large-scale but incomplete data collections, which include but are not limited to those produced by mass spectrometry. It uses the KEGG database, but could be applied to any metabolic network representation. We have provided a re-analysis of the work performed by Stensjö et al. [5] as an illustration of the advantages and the disadvantages of this method. We have shown how the key elements discussed in the original article by Stensjö et al. [5] could be identified within a single connected component: the enzymes involved in the pentose phosphate cycle, the isocitrate dehydrogenase to supply the 2-oxoglutarate further transformed in glutamate required for the fixation of the ammonia by the glutamine synthetase and the nitrogenase. This method inferred how to bridge the 2-oxoglutarate to glutamine via the reactions catalyzed by the aspartate aminotransferase and the glutamate dehydrogenase, although these enzymes are mildly up-regulated and non-quantified, respectively. The inferred network also suggests other possible routes to supply glutamate by the degradation of arginine. These hypotheses are important for a comprehensive understanding of the mechanisms activated by the N2-fixing conditions, and further investigation should focus on these enzymatic reactions. Further improvements of the methodology would try to measure the plausibility of each of these candidates in terms of a probability to be up- or down-regulated. To make the graph directed could be helpful to take account of the irreversibility of certain metabolic fluxes.

The results depend on several factors, amongst which the quality of the metabolic reconstruction, the quality [34] and the coverage of the mass spectrometry data are of the highest importance. For instance, the analysis of a Yeast dataset gives different topologies with Palsson’s reconstruction [35] and the KEGG database [23] (data not shown). Some problems originate from the inconsistencies or limitations in the metabolic network, which results in underestimated weights (hence the connections via NAD(P)H or CO2 in the network described in the ‘Results’ section). However, the underestimation of weights tends to include false positives, but not to increase the number of false negatives. This means that minute inspection of the connections should reveal the implausible pathways, which can be then discarded. In the re-analysis of Stensjö et al.’s [5] data, we have lowered the weights of glutamate, because we were aware of the important role played by this metabolite. Similarly, if former evidence suggests that a common metabolite C might be a key compound in the metabolic pathways activated in certain conditions, it should be assigned a lower value.

The eventual validation of this methodology will rest on the capability not only to ease the reconstruction of up-regulated metabolic pathways but also on the a posteriori observation that the enzymes which were not up-regulated (non-quantified or mildly up-regulated) are up-regulated or mildly up-regulated indeed. Amongst the 18 non-quantified or only mildly up-regulated enzymes identified in Nostoc’s network (Figure 3 and Table 1), 9 could be quantified in a more recent study [33]. The results are as follows:

- 1-pyrroline-5 carboxylate dehydrogenase (alr0540, originally non-quantified) quantified in a heterocyst/vegetative (114:116) experiment, 1.7, although large fluctuations seem to indicate a dependence on other factors;
- 6-phosphogluconolactonase (alr1602, originally non-quantified) quantified in three experiments: heterocyst/non-fixing filaments, heterocyst/vegetative cell (113:115), and heterocyst/vegetative cell (114:116), 1.9, 3.5 and 1.9;
- Aconitate hydratase (all1267, originally mildly up-regulated), quantified in a heterocyst/vegetative cell (113:115) experiment, 2.0;
- Arginine decarboxylase (all3401, originally non-quantified), quantified in two experiments: heterocyst/vegetative cell (114:116) and vegetative/non-fixing filaments, 1.2 and 1.2;
- Aspartate aminotransferase (alr5103, alr4853, alr1039, originally mildly up-regulated), quantified in a heterocyst/vegetative (114:116), 2.6;
- Gluconokinase (alr3086, originally non-quantified), quantified in two experiments: vegetative/non-fixing filaments, heterocyst/non-fixing filaments, 1.1 and 1.3;
- Glutathione reductase (all4968, originally mildly up-regulated), quantified in a heterocyst/vegetative cell (114:116) and vegetative/non-fixing filaments, 1.6 and 1.6;
- Polyphosphate glucokinase (all1371, originally non-quantified), quantified in three experiments: vegetative/non-fixing filaments, heterocyst/non-fixing filaments, heterocyst/vegetative (113:115), 3.0, 1.5 and 2.3;
Transketolase (alr3344, all4052, originally mildly up-regulated), quantified in three experiments: heterocyst/non-fixing filaments, heterocyst/vegetative (113:116), vegetative/non-fixing filaments, 1.3, 1.3 and 1.5.

The methodology described above depends on a very limited number of parameters. The parameters \( w_{\text{max}} \) and \( f_{\text{up}} \) are the most important in determining the inferred network. Optimal parameters may depend upon the data and the network description and, as a consequence, upon the organism considered. However, we believe that sensible values ranging from 5 to 15 would allow the treatment of most datasets and routinely identify expected and unexpected behaviours, something that would be extremely helpful in the development of synthetic organisms.

**Key Points**

- Even though proteomic data are sparse, part of the information can be retrieved by making use of the metabolic network of an organism.
- The topology of the metabolic network requires a special set-up to avoid traversing the hubs of the network.
- A heuristic method is devised to deal with sparse proteomic data, to identify up-regulated or down-regulated pathways, and to suggest possible routes of further investigation.
- This method was successfully carried out to retrieve the key actors of the processes involved in nitrogen fixation in *Nostoc* sp. PCC 7120.

**Acknowledgements**

We thank support from BiomodularH2 FP6-NEST-2005-PathSYN-043340 and from the UK’s EPSRC (EP/E036252/1 and GR/S84347/01). The authors are grateful to the anonymous reviewers for their remarks.

**References**


