Advances in network-based metabolic pathway analysis and gene expression data integration

Alberto Rezola, Jon Pey, Luis Tobalina, Ángel Rubio, John E. Beasley and Francisco J. Planes

Submitted: 17th December 2013; Received (in revised form): 23rd January 2014

Abstract

With the emergence of metabolic networks, novel mathematical pathway concepts were introduced in the past decade, aiming to go beyond canonical maps. However, the use of network-based pathways to interpret ‘omics’ data has been limited owing to the fact that their computation has, until very recently, been infeasible in large (genome-scale) metabolic networks. In this review article, we describe the progress made in the past few years in the field of network-based metabolic pathway analysis. In particular, we review in detail novel optimization techniques to compute elementary flux modes, an important pathway concept in this field. In addition, we summarize approaches for the integration of metabolic pathways with gene expression data, discussing recent advances using network-based pathway concepts.

Keywords: constraint-based modelling; elementary flux modes; genome-scale metabolic networks; gene expression data; network-based metabolic pathway analysis

INTRODUCTION

Metabolic processes are typically organized into metabolic pathways, which are commonly defined as a set of consecutive enzyme-catalyzed reactions that interconvert a set of source/target metabolites. Important individual classical (canonical) pathways can be found in biochemistry textbooks [1, 2] and several databases [3–5]. Despite their historical importance, metabolic pathways do not act as individual isolated units but act simultaneously, forming highly connected complex networks. Metabolic networks provide a more holistic view of metabolism than canonical pathways, and their systematic study, based on computational tools and high-throughput ‘omics’ experimental data, is one of the objectives underlying systems biology.

In this review, we focus on theoretical approaches to metabolic pathways from a network-based perspective, as well as their integration with ‘omics’ data, particularly gene expression data. This latter topic is especially relevant at this point in time with an explosion in the availability of ‘omics’ data [6–8] and with increasing evidence as to the limitations of canonical pathways when attempting to exploit such data [9]. The use of network-based pathway concepts opens new avenues to interpret ‘omics’ data with potential applications in different fields such as medicine and biotechnology.

Corresponding author. Francisco J. Planes, Paseo Mikeletegi, No 48, 20009, San Sebastián, SPAIN Tel: (+34) 943 212800 (ext:2802) Fax: (+34) 943 213076. E-mail: fplanes@tecnun.es

Alberto Rezola received his PhD for contextualizing metabolic processes using elementary flux modes and gene expression data. He is currently employed at the Bioinformatics department in CEIT.

Jon Pey received his PhD for the integration of ‘omics’ data and genome-scale metabolic networks using linear optimization models. He is currently employed at the Bioinformatics department in CEIT.

Luis Tobalina is a PhD student at the Bioinformatics department of CEIT. His research interests are focused on the study of metabolic and signalling networks with different mathematical and computational tools and their application in medical research.

Ángel Rubio is a researcher at CEIT since 1995. His expertise includes genomic and proteomic data analysis, statistical treatment of data and systems modelling.

John E. Beasley is a faculty member at Brunel University. His research is focused around discrete optimization problems.

Francisco J. Planes is a postdoctoral researcher in the Bioinformatics group at CEIT. His research is devoted to the development of optimization-based techniques in the area of systems biology and bioinformatics, mainly in the field of metabolic modelling.
METABOLIC NETWORK-BASED PERSPECTIVE

Recent reconstruction of genome-scale metabolic networks (GSMNs) has allowed the scientific community to study metabolism in a systematic fashion using mathematical modelling techniques. Several mathematical approaches have been developed, which can be grouped under the term constraint-based modelling (CBM) and are based on the stoichiometric matrix [10].

Assume that we have a metabolic network comprising \( R \) reactions and \( C \) metabolites. The relative number of molecules of metabolite \( \epsilon \) involved in reaction \( r \) is known as the stoichiometric coefficient \( s_{cr} \). In particular, substrates and products are represented by negative and positive stoichiometric coefficients, respectively. These coefficients are stored in the stoichiometric matrix \( S \).

The stoichiometric matrix is used to define a key constraint in CBM, the steady-state condition, which refers to the property of mass balance within the cell. At steady-state, there is no accumulation or depletion of metabolites inside the network boundaries, so that the consumption rate of a metabolite \( \epsilon \) must equal its production rate. In other words, the concentration of internal metabolites \( c \) remains constant over time.

In addition, the activity of each reaction is represented by a flux variable \( \nu_r \) \( (r = 1, \ldots, R) \). As the absolute number of molecules of a metabolite produced or consumed in a reaction is its flux \( (\nu_r) \) multiplied by its stoichiometric coefficient, the steady-state condition can be represented mathematically by the following constraint:

\[
\sum_{r=1}^{R} s_{cr} \nu_r = 0, \quad \forall c \in I
\]

A second relevant constraint in CBM refers to thermodynamic feasibility, which restricts some fluxes to being non-negative, owing to their associated Gibbs free energy. Therefore, a number of biochemical reactions are only viable in one direction and are considered irreversible \((\text{Ir})\). This constraint is:

\[
\nu_r \geq 0, \quad \forall r \in \text{Ir}
\]

The steady-state condition and thermodynamic feasibility are the two basic constraints of most CBM methods. They allow us to mathematically describe feasible functional states of reconstructed metabolic networks by a homogeneous system of linear equations. A detailed review of CBM techniques can be found in [11]. We now describe different approaches in the literature to analyse metabolic pathways from a network-based perspective [12].

NETWORK-BASED METABOLIC PATHWAY ANALYSIS

Network-based metabolic pathway analysis (NBMPA) interprets metabolic pathways as the simplest steady-state flux distributions of a metabolic network. Note that a feasible steady-state flux distribution must satisfy Equations (1) and (2), which define the solution space \( P \):

\[
P = \left\{ \nu_r \right\} \sum_{r=1}^{R} s_{cr} \nu_r = 0, \quad \forall c \in I; \quad \nu_r \geq 0, \quad \forall r \in \text{Ir}
\]

The solution space in Equation (3) is typically underdetermined, i.e. there is more than one set of fluxes \( \left\{ \nu_r \right\} \) in \( P \), and it geometrically defines a polyhedral cone in a high-dimensional space, which is termed the flux cone. The mathematical properties of the flux cone are summarized in a separate section in the Supplementary Material (available online at [http://bib.oxfordjournals.org/](http://bib.oxfordjournals.org/)) and can be analyzed using different techniques [13].

NBMPA introduces a further constraint, technically termed the non-decomposability condition, which ensures that solutions comprise a minimal number of active reactions at steady-state. This simplicity condition essentially implies that these solutions cannot be decomposed into any smaller flux distributions without violating the steady-state constraint [14, 15]. Figure 1 illustrates this non-decomposability condition in a toy example.

The non-decomposability condition can be written in mathematical notation as follows. Let \( S^* \) be a submatrix of \( S \), which exclusively involves the active reactions in a solution \( \left\{ \nu_r \right\} \) in \( P \). Then the solution \( \left\{ \nu_r \right\} \) cannot be further decomposed if the null space of \( S^* \) has dimension 1.

The simplicity condition is a biologically meaningful constraint, as it provides cohesion, as all the reactions in the solution are needed to perform a metabolic function in steady-state, and knockout of any of them will prevent such a solution from functioning.
Elementary flux modes
A flux distribution \( \{ v_r \} \) that satisfies (i) the steady-state constraint; (ii) the thermodynamic constraints; and (iii) the non-decomposability condition is termed an Elementary Flux Mode (EFM) [15]. Although there are earlier attempts to compute pathways in metabolic networks [16–21], EFMs are currently the most accepted network-based definition of metabolic pathways.

The set of EFMs is unique and can represent any possible physiological situation in a cellular system. Aside from recovering canonical pathways, the EFM approach also predicts a large number of unobserved potential pathways that emerge from a combination of active reactions in the network [22]. For this reason, the EFM approach is ideal to explore the richness inherent in a metabolic network and consequently to elucidate novel pathways [23, 24].

Convex basis-based pathway concepts
A concept closely related to EFMs is that of convex basis [25]. A convex basis is a minimal set of EFMs able to describe any \( \{ v_i \} \in P \), as we typically have more EFMs than required to construct any steady-state flux distribution. The term generating flux mode (GFM) has also been used to refer to the elements of a convex basis [26], as they constitute a set of generating vectors that describes the polyhedral cone \( P \) (Figure 2A).

In the context of NBMPA, a convex basis must contain elements that are non-decomposable, i.e. we seek a minimal generating set of EFMs. If the flux cone is pointed, it can be described by its extreme rays, which are (minimal) faces of dimension one, and therefore, the non-decomposability condition is satisfied (see Supplementary Material for a summary of mathematical definitions used for describing pointed and non-pointed polyhedral cones available online at http://bib.oxfordjournals.org/). Therefore, the set of extreme rays in a pointed flux cone is a subset of EFMs that constitute the unique convex basis. However, flux cones are typically non-pointed. In this case, their minimal (proper) faces have a dimension strictly greater than one, and therefore, the non-decomposability condition is not guaranteed. This also leads to an issue relating to the non-uniqueness of the convex basis, i.e. we can select more than one convex basis to describe a flux cone. Note that this non-uniqueness issue complicates the biological interpretation of the convex basis and is one of the reasons as to why the concept of EFMs has been preferred.

To overcome this issue, different strategies have focused on converting a non-pointed flux cone into a pointed one. In particular, two strategies exist: extreme currents (ECs) and extreme pathways (ExPas). ECs were the first strategy to be proposed [19]. This approach splits reversible reactions into two irreversible steps, which implies a move to a pointed, but higher-dimensional, flux cone. The set of ECs constitutes a unique convex basis of EFMs in the reconfigured metabolic network. However, owing to the increase in dimension, we have more ECs than EFMs. In addition, it has been shown that the set of ECs coincides with the set of EFMs when translated to the original vector space [26, 27]. This is illustrated in Figure 2, where the complete set of EFMs and ECs is determined in a toy metabolic network. For example, observe that both EC1 and EC2 map into EFM1.

With the objective of reducing the dimension of the reconfigured flux cone, ExPas were proposed [28]. For this purpose, ExPas ensure that the flux cone is pointed without splitting the entire set of reversible reactions into two irreversible steps. Instead, the ExPas approach classifies reactions into internal and exchange reactions. Exchange reactions, which represent metabolite inflows and outflows in the system, are allowed to be reversible. On the other hand, internal reactions are forced to be irreversible, and therefore, originally reversible reactions are represented by a forward and backward
reaction. Under this network configuration, the set of ExPas constitutes a unique and minimal convex basis of EFMs. This strategy also increases the dimension of the flux cone, but to a lesser extent than ECs, so that typically there are fewer ExPas than EFMs. However, the number of ExPas is still greater than the number of GFMs required to span the flux cone in the original configuration [26]. This can be observed in Figure 2, where we only have three GFMs in Figure 2A but four ExPas in Figure 2C and nine ECs in Figure 2B.

Algorithmic and computational issues

Double description method-based techniques

Classical algorithms for the computation of the complete set of EFMs and different convex basis-based pathway concepts (GFMs and ExPas) follow the same principles as procedures to enumerate the extreme rays on a pointed polyhedral cone. In particular, they are variants of the double description method introduced in [29].

In particular, this algorithm starts from an initial cone that contains the polyhedral cone under study. The positive orthant of an $n$-dimensional space is typically used, whose extreme rays are the elements of the canonical basis. Then, each constraint $j (j = 1, \ldots, m)$ is imposed one by one, and therefore, $m$ intermediate cones are studied. This method relies on the fact that any extreme ray of an intermediate cone $j + 1$ can be computed as a non-negative linear combination of two extreme rays of a previous cone $j$. When extreme rays of cone $j$ are pairwise combined, extreme rays of cone $j + 1$ arise; however, we may also have interior points, which need to be removed. After $m$ iterations, we have the extreme rays of the polyhedral cone under study. It should be emphasized that this method has to be completed so as to guarantee that the obtained solutions are EFMs, GFMs or ExPas (i.e. intermediate solutions cannot be so guaranteed).

Since the first application of the double description method in a chemical reaction system in 1988 [19], several advances have been proposed in the literature to improve its performance in metabolic networks [15, 28, 30–32]. These advances have been implemented in different bioinformatics tools, such as Metatool and its extensions, FluxAnalyzer, structural network analysis (SNA), expa, systems biology research tool (SBRT) and efmtool (Table 1).

Currently, efmtool is the most efficient algorithm to compute EFMs and GFMs [37], and it can be easily adapted to compute ExPas. efmtool has proven effective in computing the complete set of
EFMs in networks of a few hundred reactions; a convex basis of GFMs in networks near a thousand reactions [39]. However, the number of EFMs (and GFMs) explodes in a combinatorial fashion as the size of metabolic network increases [40–42], and efmtool is ineffective for networks of thousands of reactions, e.g. the human genome-scale metabolic network [43, 44]. For this reason, optimization-based techniques to compute at least a subset of the complete set of EFMs in GSMNs have emerged.

Optimization-based techniques

The use of optimization-based techniques to deal with the combinatorial explosion of EFMs in large metabolic networks has been reported in recent years. These methods enable the direct computation of a subset of EFMs of particular interest, without having to first enumerate the complete set of EFMs. Here, we briefly describe the most relevant approaches.

EFMEvolver uses a linear program (LP) to determine a single EFM [45]. In particular, it starts from the flux cone [Equation (3)]. As in ECs, reversible reactions are divided into two irreversible steps; therefore, the flux cone is pointed. This guarantees that the extreme rays of such a flux cone are EFMs. A half-space that cuts the polyhedral cone is then added to the set of constraints [Equation (5)]. This constraint forces a reaction \( u \) (or set of reactions) to be active and converts the polyhedral cone into a polytope. Thus, extreme rays involving reaction \( u \) are intersected by this constraint and they become extreme points of the solution space. Linear programming can then be used to enumerate extreme points. For this purpose, an objective function should be introduced. In particular, EFMEvolver minimizes reaction fluxes using Equation (6).

\[
\nu_n \geq 1 \tag{5}
\]

\[
\text{Min} \sum_{i=1}^{R} \nu_i \tag{6}
\]

In order to compute a subset of EFMs, this LP can be repeatedly re-solved, but preventing different reactions in a specified set, \( T \), from being active (excluded from the solution space), as observed in Equation (7):

\[
\nu_j = 0, \forall j \in T \tag{7}
\]

Linear programming is efficient for computing a single EFM. However, the computation of an EFM for each knocked-out reaction subset \( T \) is impractical. To overcome this issue, a genetic algorithm was introduced in **EFMEvolver**. This strategy efficiently explores a broad number of EFMs in large GSMNs [45].

Based on integer linear programming, a method that iteratively computes EFMs in increasing number of reactions, i.e. from shortest to longest EFMs, was presented in [46]. In [41], it was previously evidenced that finding the shortest EFMs is a nondeterministic polynomial-hard problem. However, in [46] it was shown that the computation of **K-shortest EFMs** is tractable in GSMNs.

The key concept behind this approach is that shorter metabolic pathways have better biological properties: (i) they can carry higher fluxes [47, 48]; (ii) they are better suited as a target of genetic manipulation, as it is expensive and laborious to over-express a large number of enzymes; and (iii) canonical pathways typically use a minimal number of steps to achieve their metabolic objective.

The **K-shortest EFMs** approach starts from the flux cone, as defined in Equation (3). Then, binary variables \( z_r \) are introduced, namely \( z_r = 1 \) if \( \nu_r > 0 \)

<table>
<thead>
<tr>
<th>Software</th>
<th>Algorithm</th>
<th>Pathway concept</th>
<th>Year</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>efmtool [37]</td>
<td>Recursive nullspace</td>
<td>EFMs</td>
<td>2008</td>
<td><a href="http://www.csb.ethz.ch/tools/efmtool">http://www.csb.ethz.ch/tools/efmtool</a></td>
</tr>
</tbody>
</table>
and $z_i = 0$ if $v_i = 0$. In this approach, reversible reactions are divided into two irreversible steps. The non-decomposability condition is achieved with the objective function in Equation (8). Clearly, the solution involving the minimum number of reactions (one-shortest EFM) will be non-decomposable [22]. To conserve this property through subsequent solutions ($k = 2, \ldots, K$), Equation (9) prevents $(k - 1)$-shortest EFMs from repeating, namely, each solution is forced to differ by at least one metabolic reaction with previously computed EFMs. Note that $Z^k_i$ is 0/1 parameter set to 1 if a reaction is active in the $k$-th shortest solution (zero otherwise).

$$\text{Min } \sum_{i=1}^{R} z_i$$  \hspace{1cm} (8)

$$\sum_{i=1}^{R} Z^k_i z_i \leq \left( \sum_{i=1}^{R} Z^k_i \right) - 1, \quad k = 1,\ldots,K - 1$$  \hspace{1cm} (9)

In addition, Equation (5) may be included into the model so as to compute the shortest EFMs carrying flux through a particular reaction $u$. Similarly, Equation (7) can be added to knockout particular reactions.

To capture a wider range of metabolic functions without increasing the value of $K$, the $(K,d)$-shortest EFMs approach was proposed. This technique changed the previous solution elimination constraint, guaranteeing that the obtained shortest EFMs differ by at least $d$ reactions, as shown in Figure 3. For this purpose, we only need to amend Equation (9), as shown in Equation (10). Note that in the $K$-shortest EFMs approach, $d$ was intrinsically set to 1, as Equation (9) forces solutions to differ by at least one reaction.

$$\sum_{i=1}^{R} Z^k_i z_i \leq \left( \sum_{i=1}^{R} Z^k_i \right) - d, \quad k = 1,\ldots,K - 1$$  \hspace{1cm} (10)

An example illustrating the impact of the $d$ parameter can be found in Figure 3. In Figure 3A, the $K$-shortest EFM approach is applied. The five EFMs indicated in Figure 2 are enumerated in order of increasing number of reactions, namely, one-shortest EFM involves three reactions, while the five-shortest EFM involves four reactions. Ties in the number of reactions are broken randomly. In Figure 3B, we only obtain three solutions, as we used $d = 2$ and, therefore, each solution must differ by at least two active reactions from the previous solutions. When applied to large networks, where the full set of EFMs cannot be computed, the use of the $d$ parameter increases diversity within the $K$ solutions, as discussed in [49].

Finally, the $K$-shortest EFM approach was further extended to compute the shortest GFMs in genome-scale networks [50]. GFMs can be classified into irreversible and reversible, depending on whether irreversible reactions are included. This classification is important, as the number of irreversible GFMs explodes in a combinatorial fashion, while the number of reversible GFMs is at most the maximum number of reversible reactions.

In particular, this approach computes a subset of shortest irreversible GFMs in GSMNs. To do this, the approach is based on the concept of minimal metabolic behaviours (MMBs) introduced in [51]. In this work, the authors show that, in non-pointed flux cones, the minimal proper faces are uniquely described by a minimal subset of irreversible reactions, which are precisely the MMBs. This is exploited so as to determine a particular convex basis, namely, the one with the shortest GFMs. This is related to the inner description of polyhedral cones. Details can be found in the Supplementary Material (available online at http://bib.oxfordjournals.org/).

As detailed in [50], the $K$-shortest GFMs approach is based on two main concepts: (i) MMBs can be enumerated by minimizing the number of irreversible reactions; and (ii) although a minimal (proper) face may involve different EFMs with the same MMB, we only need to select one of them to describe a convex basis.

Based on these concepts, $K$-shortest GFMs approach is presented via mixed-integer linear programming and determines the shortest irreversible GFM for each MMB. This is achieved by its objective function, which minimizes two terms: first, the number of active irreversible reactions and second, the number of active reversible reactions. Particularly, the objective function introduces a large positive constant, $W$, for irreversible reactions to guarantee that each solution involves an MMB, while the second term ensures the non-decomposability condition:

$$\text{Min } \left(W \sum_{i=1, i \in U}^{R} z_i + \sum_{i=1, i \notin U}^{R} z_i \right)$$  \hspace{1cm} (11)

In this approach, as irreversible GFMs are sought, Equation (12) ensures that at least one irreversible reaction carries flux. In addition, Equation (13)
prevents previously computed MMBs from appearing in the solution, guaranteeing one GFM per MMB. Note that, as $K$ irreversible GFMs are determined, $K$ optimization problems are sequentially solved.

$$\sum_{i=1, r \in I_{ir}}^R z_i \geq 1$$  \hspace{1cm} (12)

$$\sum_{i=1, r \in I_{ir}}^R Z_{ik}^r z_i \leq \left( \sum_{i=1, r \in I_{ir}}^R Z_{ik}^r \right) - 1, \hspace{0.2cm} k = 1, ..., K - 1$$  \hspace{1cm} (13)

Similar to $K$-shortest EFMs, if we force a particular reaction $u$ to be active by Equation (5), our approach will determine the shortest irreversible GFMs that involve such a reaction. On one hand, if $u$ is irreversible, the number of GFMs is equal to the number of MMBs containing reaction $u$, whereas on the other hand, if $u$ is reversible, the approach may not guarantee the non-decomposability condition and further considerations must be taken into account, as reflected in the seminal article [50]. Finally, note that this approach can be easily extended to ExPas, namely, by changing the network configuration.

An example of the $K$-shortest GFMs approach can be found in Figure 3C. We only have two irreversible GFMs, one for each MMB, $\{R_2, R_3\}$ and $\{R_5, R_3\}$, which are enumerated in increasing number of irreversible reactions. Ties are broken randomly. We then select GFMs with the minimum number of reversible reactions. For example, here we select the solution $\{R_1, R_2, R_3\}$ instead of the solution $\{R_6, -R_4, R_2, R_3\}$, which involves one additional reaction. Similar to the $(K,d)$-shortest procedure, the use of GFMs also increases the diversity of the solutions found, as discussed in [50].

**Figure 3**: A toy example showing the EFMs captured by the enumeration of the $K$-shortest EFMs (A), $(K,d)$-shortest EFMs (B) and $K$-shortest GFMs (C). In particular, we fixed $d = 2$ and let $K$ be a large value, e.g. $K = 10$. 

**Other network-based pathway concepts**

Apart from EFMs and convex basis-based approaches, other pathway concepts have been considered in the literature. In particular, the use of graph theory has been recurrent. Graph-based methods view metabolic pathways as a directed path from a source node to a target node. Computing paths in a graph is simpler than computing EFMs, and, for this reason, their use have been extensive for a number of applications: drug targets [52–54], disease correlation [55] and key metabolic pathways [56, 57]. However, there are issues associated with their use, as discussed in [58, 59]. The major criticism is that path-finding methods neglect reaction stoichiometry, and therefore, there is no guarantee that any path found can operate in the steady-state. This fact typically leads to solutions that are not as meaningful as other methods and brings into question the validity of this approach [58, 59]. A deeper introduction to graph-based methods can be found in [60]. Note here that in less-curated genome-scale metabolic networks, where reaction stoichiometry is not well defined, graph-based methods are the only valid approach.
for pathway analysis and this explains their early and wide application.

A recent work has directly addressed this issue with the concept of Carbon Flux Paths (CFPs) [61]. In brief, CFPs are directed paths from a given source to a target metabolite able to perform in steady-state. Instead of using graph theory, this method is built on mixed-integer linear programming (optimization), which is computationally more expensive, similar to optimization-based approaches for EFMs.

Note that CFPs are calculated on a directed graph, where nodes are metabolites and arcs represent a carbon exchange between them. The list of carbon exchanges between metabolites is currently available for *Escherichia coli* and human genome-scale metabolic networks [61, 62], and therefore, further extensions to other organisms are required. In addition, CFPs guarantee carbon exchange in each intermediate step but not between the source and target metabolite. To overcome this issue, atomic CFPs (aCFPs) have been recently introduced [63], which define a graph based on carbon atoms instead of metabolites. Although the aCFP approach improves the performance of the CFP approach, it relies on large-scale atomic metabolic reconstructions, which are currently scarce. For these reasons, the application of CFPs and aCFPs is still under development, but it is promising for a number of applications in biotechnology and health [62, 64].

Finally, the concept of Elementary Flux Patterns (EFPs) has been recently presented in [65]. Owing to the difficulties of enumerating the complete set of EFMs in large genome-scale metabolic networks, dividing the whole network into smaller sub-networks of interest and determining the EFMs for each sub-network has been used in a number of works [66, 67]. However, this procedure does not ensure the steady-state condition and/or the non-decomposability condition in the entire network; therefore, the results obtained may be biased. For this purpose, an EFP defines a minimal subset of reactions within a sub-network that is able to operate at steady-state in the whole metabolic network. EFPs consider steady-state for the whole metabolic network, while the concept of simplicity is reformulated at the sub-network level in a fashion similar to that of the non-decomposability condition in EFMs. Although their application is still under development, interesting results can be found in [68], where several novel pathways from fatty acids to glucose in humans are proposed based on EFPs.

### INTEGRATION OF GENE EXPRESSION DATA WITH NBMPA

Metabolic networks represent a blueprint of possible metabolic behaviours, but not every gene or every protein is expressed. High-throughput ‘omics’ data sets provide relevant information that should be used so as to elucidate actual metabolic behaviour. Particularly, proteomics, metabolomics and fluxomics are preferred to infer metabolic processes, as they are more closely related to the phenotype than genomics and transcriptomics that are subject to post-transcriptional regulatory circuits.

However, gene expression analyses have been the dominant approach for a wide characterization of metabolic processes for two reasons. First, the development of microarray technology has allowed the measurement of a large number of metabolic genes, while the scope of proteomics, metabolomics and fluxomics is currently limited. Second, the huge amount of publicly available microarray data sets, which can be found in public databases such as ArrayExpress [7] and GEO [6]. Note that the use of gene expression as a hint to the activity of associated reactions is certainly controversial because, as noted above, the mRNA population is subject to complex regulatory mechanisms. However, a clear correlation between gene expression and metabolic fluxes has been observed [69].

In the past few years, different efforts have been made to integrate gene expression data with network-based metabolic pathway concepts. The integration process is similar to previous approaches based on canonical pathways. First, different statistical methods are applied to process data and classify genes for a certain phenotype. Second, a representative list of metabolic pathways where gene sets can be mapped is selected. Finally, a pathway scoring technique is used to identify the most relevant pathways. These steps are schematically summarized in Figure 4 and briefly described below.

### Gene expression processing and classification

We refer to previous and more detailed works so as to explain how to measure and classify gene expression data [70, 71]. However, we would like to emphasize here two different types of gene expression
Differential gene expression analysis focuses on determining genes that have been over-expressed, unchanged or under-expressed between two conditions under study. To do this, several methods have been developed based on the fold-change and are mostly available through the limma [70], SAM [72] or EDGE [73] software packages, among others. These gene classification methods have been widely used for metabolic studies [74].

Absolute gene expression analysis determines genes that are expressed in a specific biological sample. This classification technique has not been widely explored, as an accurate measure of gene expression in a microarray is not possible because of unknown probe-specific affinities. There are several methods to establish a threshold classifying genes as present or absent, typically exploiting the introduction of control negative probes in different microarray technologies [75].

Interestingly, a recent work [76] provided absolute gene expression estimates for two human and one murine microarrays based on a probability of expression model constructed on a series of negative control experiments taken from GEO [6] and ArrayExpress [7]. Hence, each probe can be classified as present or absent in each individual sample without needing to compare with other probes in the microarray. Although specific for human and murine microarrays, this tool simplifies gene classification based on absolute expression. This is certainly of interest for metabolic pathway analysis and could complement studies based on differential expression.

Representative set of metabolic pathways
Aiming to provide a functional interpretation to gene sets and infer associated phenotypes, a list of potential metabolic pathways is selected. A number of approaches rely on canonical metabolic pathways stored in different databases, e.g. BioCyc [9] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [3]. As noted above, despite their wide application in different areas [77–79], these methods restrict novel discoveries, as pre-defined pathways do not capture the wide variety of complex metabolic states. Network-based metabolic pathway concepts, as reviewed above, aim to overcome this issue.

Path-finding techniques, from the field of graph theory, have been a recurrent approach to map gene networks.
sets. A number of tools enumerate all paths inside KEGG maps [80]. Other works compute a subset of paths in the whole network, although they require an input/output set of metabolites [81]. As noted above, pathways constructed with graph-theoretical methods face important theoretical issues [58, 59], as they do not consider reaction stoichiometry.

The use of other pathway concepts with gene expression data has been limited, mainly owing to the fact that their computation was impractical in large metabolic networks. Early work in the literature used two reductionist strategies: first, EFMs were computed for each human map from KEGG [82] and second, EFPs were determined for every subsystem within the E. coli metabolic network [83]. A more general approach was recently presented in [49], where using the (K,A)-shortest procedure discussed above, five EFMs were computed for each reaction in the human genome-scale metabolic network, overall involving ~6000 EFMs. This subset of EFMs can be used to map heterogeneous ‘omics’ data in different conditions.

Note here that, for mapping gene sets with metabolic pathways, it is essential to associate genes with reactions. This information is usually contained in metabolic network databases, e.g. [84], and it is commonly known as gene–protein reaction data. Gene–protein reaction reflects the regulation of metabolic reactions by their corresponding proteins and genes using Boolean rules.

Pathway scoring techniques
Enrichment techniques are designed to score metabolic pathways in the light of gene expression data. These techniques are highly developed [85] and are not the focus on this review article. However, we provide a brief introduction to this topic and review in detail techniques used in the context of NBMPA.

A number of techniques compare a list of up- or down-regulated genes to a gene set (reaction set) of a specific metabolic pathway so as to identify if there are equal or more matches (hits) than expected by chance [86]. These methods have been termed over-representation analysis (ORA) and typically use statistical testing to assign a P-value for each pathway. In particular, most of the ORA approaches are built on Fisher’s exact test and the hypergeometric distribution [87, 88]. However, to simplify the P-value calculation, other distributions such as chi-square [89], binomial [86] and standardized difference scores (z scores) have been used [90].

Other methods grouped into the term functional class scoring compute a statistic that summarizes the P-value of genes in the pathway [91, 92]. Using gene randomization techniques, the distribution under the null hypothesis is empirically determined, which allows the assignment of a P-value to each pathway. Note here that a previous gene classification is not required in this approach.

As in functional class scoring, gene set enrichment analysis [93] does not require establishing an arbitrary cut-off in terms of fold-change or significance. Instead, using a suitable metric, genes are ranked according to their relationship with the phenotype. The enrichment score of a particular gene set (metabolic pathway) depends on its length and position in the ranking of genes involved. Using sample randomization, the distribution under the null hypothesis is empirically determined and a P-value is then assigned. Similar methods to gene set enrichment analysis are found in the literature: gene-set analysis [94] and significance analysis of function and expression [95].

In the context of NBMPA, enrichment techniques applied to EFMs/EFPs have been mainly limited to the hypergeometric distribution [49, 82, 83], and therefore, further development is required. In this direction, it has been recently shown that the use of a multivariate hypergeometric distribution allows introducing up- and down-regulated genes (reactions) in the same statistical score, i.e. maximizing the number of up-regulated genes and minimizing the number of down-regulated genes within the gene (reaction) set [49]. Resulting EFMs are biologically more meaningful than those obtained with the standard hypergeometric distribution.

Note that, for enrichment methods, the application of multiple hypothesis testing is relevant, particularly given the high number of pathways analysed. A brief introduction to multiple hypothesis testing can be found in the Supplementary Material (available online at http://bib.oxfordjournals.org/).

Other integration methods
It should be noted that other network-based approaches have been developed to score pathways based on optimization techniques rather than using statistical analyses.

One example is YANA [96], which scores EFMs according to their contribution to reaction fluxes, which are strongly assumed to be equal to the level
of gene or protein expression. For this purpose, YANA minimizes the sum of quadratic errors between calculated reaction fluxes ($v_r$) and fluxes based on gene expression level ($v_{\text{exp}}^r$), as shown in Equation (14):

$$\min X_R r = \frac{1}{C_0} (v_r - v_{\text{exp}}^r)^2$$  \hspace{1cm} (14)$$

In addition, it exploits the property that metabolic fluxes can be written as a non-negative combination of EFMs, as shown in Equation (15):

$$v_r = \sum_{j=1}^{N} \left| \epsilon_j \right| a_j$$  \hspace{1cm} (15)$$

where $N$ and $R$ are the total number of EFMs and reactions in the metabolic network, respectively; $a_j$ is the flux activity of EFM $j$ and $\epsilon_j$ is the relative flux of reaction $r$ in EFM $j$.

Finally, YANA includes an evolutionary algorithm to find a fast and robust solution. However, YANA needs to first compute the complete set of EFMs in a metabolic network; therefore, its scope is currently limited.

An important family of approaches directly extracts pathways based on gene expression data [97–100]. An example is KeyPathway Miner [99], which integrates gene expression data to obtain most expressed graph-based paths by determining specific reaction sub-networks involving a maximum number of differentially expressed genes by means of an Ant Colony optimization strategy. A more sophisticated approach has been presented recently in [64] using the concept of CFPs, where the original formulation of CFPs was amended to incorporate expression data.

FUTURE DIRECTIONS

As detailed above, optimization-based techniques have emerged in recent years and allow us to compute EFMs and other pathway concepts in genome-scale metabolic networks. However, for their wider use, these advances require translation into user-friendly bioinformatics tools, as current software packages are not effective in GSMNs (Table 1).

On the other hand, a number of bioinformatic tools for scoring canonical metabolic pathways have been developed in the past decade [9]. However, dedicated software for undertaking an enrichment analysis of network-based pathway concepts, mainly for EFMs, is currently limited, as observed in Table 2.

Effective progress in the availability of these tools will allow us to further exploit methods from NBMPA. In particular, with the approach recently presented in [49], EFMs can be applied for different purposes. For example, using this approach and both differential and absolute gene expression data, key metabolic properties between two sub-types of non–small-cell lung cancer were distinguished and identified in [102].

In addition, as mentioned above, the integration of proteomics, metabolomics and fluxomics is currently restricted because of the lack of large and varied data sets. As they become freely and widely available, we would expect similar developments to occur.

With increasing efforts to improve the performance of algorithms to compute EFMs, the number of EFMs that can be realistically computed will be in the order of millions. In this new scenario, methods to identify the most significant EFMs based on ‘omics’ data need to be revisited and revised. A possible strategy is to directly extract the most relevant EFMs based on expression data, as done with graph-based methods and CFPs.

CONCLUSIONS

In this article, we have reviewed the field of NBMPA, with special focus on EFMs and their integration with gene expression data. The concept of EFMs is not novel; what has changed in recent years is that their computation in genome-scale metabolic networks is becoming possible for the very first time. This is owing to the emergence of optimization-based methods, reviewed in detail here. This will allow us to study cellular metabolism at an unprecedented level of complexity, possibly identifying novel metabolic pathways and key metabolites in different scenarios, as EFMs correctly capture the plasticity found in metabolic systems, going beyond canonical pathways.

We have also briefly summarized a general approach to integrate gene expression data with metabolic pathways and reviewed the latest advances for NBMPA techniques. We noticed that, although this approach is clear for canonical pathways, it has not been widely explored for network-based pathway concepts. Thus, we stressed the importance of developing better procedures and creating user-friendly
software tools so that NBMPA techniques, in the context of gene expression data, can be more widely used in different fields.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Table 2: Available software packages for the integration of gene expression data with network-based metabolic pathways

<table>
<thead>
<tr>
<th>Software</th>
<th>Method</th>
<th>Pathway concept</th>
<th>Multiple testing</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPPfinder [90]</td>
<td>z-score, hypergeometric</td>
<td>GO terms, specific pathways</td>
<td>–</td>
<td><a href="http://www.genmapp.org/">http://www.genmapp.org/</a></td>
</tr>
<tr>
<td>YANA [96]</td>
<td>% of activity for EFMs</td>
<td>EFMs from METATOOL</td>
<td>–</td>
<td>uni-wuerzburg.de</td>
</tr>
<tr>
<td>Key pathway miner</td>
<td>Graph theory and gene</td>
<td>Graph theory-based</td>
<td>–</td>
<td><a href="http://keypathwayminer.mpi-inf">http://keypathwayminer.mpi-inf</a>.</td>
</tr>
<tr>
<td></td>
<td>expression</td>
<td>pathways</td>
<td></td>
<td>mpg.de</td>
</tr>
</tbody>
</table>

**Key Points**

- Network-based metabolic pathway analysis is a promising field in systems biology that has experienced clear progress in the past few years.
- Optimization techniques have emerged as an effective strategy to determine network-based metabolic pathways.
- Thanks to these advances, integration of ‘omics’ data with network-based metabolic pathways is now possible.
- Approaches to integrate network-based metabolic pathways with gene expression data are emerging, although further development is required.
- Software packages and tools that make available the major advances in network-based metabolic pathway analysis are expected to be developed.

**FUNDING**
This work was supported by the Asociación de amigos de la Universidad de Navarra to [A.R.] and the Basque Government to [J.P.] and [L.T.].

**References**


