Methodology article

**Estimating the size of the solution space of metabolic networks**

Alfredo Braunstein¹,², Roberto Mulet*³ and Andrea Pagnani*²

Address: ¹Politecnico di Torino, Corso Duca degli Abruzzi 34, I-10129, Torino, Italy, ²ISI Foundation Viale Settimo Severo 65, Villa Gualino, I-10133 Torino, Italy and ³“Henri-Poincaré-Group” of Complex Systems and Department of Theoretical Physics, Physics Faculty, University of Havana, La Habana, CP 10400, Cuba

Email: Alfredo Braunstein - alfredo.braunstein@polito.it; Roberto Mulet* - mulet@fisica.uh.cu; Andrea Pagnani* - pagnani@isi.it

* Corresponding authors

Published: 19 May 2008

BMC Bioinformatics 2008, 9:240 doi:10.1186/1471-2105-9-240

This article is available from: http://www.biomedcentral.com/1471-2105/9/240

© 2008 Braunstein et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract**

**Background:** Cellular metabolism is one of the most investigated system of biological interactions. While the topological nature of individual reactions and pathways in the network is quite well understood there is still a lack of comprehension regarding the global functional behavior of the system. In the last few years flux-balance analysis (FBA) has been the most successful and widely used technique for studying metabolism at system level. This method strongly relies on the hypothesis that the organism maximizes an objective function. However only under very specific biological conditions (e.g. maximization of biomass for *E. coli* in reach nutrient medium) the cell seems to obey such optimization law. A more refined analysis not assuming extremization remains an elusive task for large metabolic systems due to algorithmic limitations.

**Results:** In this work we propose a novel algorithmic strategy that provides an efficient characterization of the whole set of stable fluxes compatible with the metabolic constraints. Using a technique derived from the fields of statistical physics and information theory we designed a message-passing algorithm to estimate the size of the affine space containing all possible steady-state flux distributions of metabolic networks. The algorithm, based on the well known Bethe approximation, can be used to approximately compute the volume of a non full-dimensional convex polytope in high dimensions. We first compare the accuracy of the predictions with an exact algorithm on small random metabolic networks. We also verify that the predictions of the algorithm match closely those of Monte Carlo based methods in the case of the Red Blood Cell metabolic network. Then we test the effect of gene knock-outs on the size of the solution space in the case of *E. coli* central metabolism. Finally we analyze the statistical properties of the average fluxes of the reactions in the *E. coli* metabolic network.

**Conclusion:** We propose a novel efficient distributed algorithmic strategy to estimate the size and shape of the affine space of a non full-dimensional convex polytope in high dimensions. The method is shown to obtain, quantitatively and qualitatively compatible results with the ones of standard algorithms (where this comparison is possible) being still efficient on the analysis of large biological systems, where exact deterministic methods experience an explosion in algorithmic time. The algorithm we propose can be considered as an alternative to Monte Carlo sampling methods.
Background
Cellular metabolism is a complex biological problem. It can be viewed as a chemical engine that transforms available raw materials into energy or into the building blocks needed for the biological function of the cells. In more specific terms a metabolic network is indeed a processing system transforming input metabolites (nutrients), into output metabolites (amino acids, lipids, sugars etc.) according to very strict molecular proportions, often referred as stoichiometric coefficients of the reactions.

Although the general topological properties of these networks are well characterized [1-3], and non-trivial pathways are well known for many species [4] the cooperative role of these pathways is hard to comprehend. In fact, the large sizes of these networks, usually containing hundreds of metabolites and even more reactions, makes the comprehension of the principles that govern their global function a challenging task. Therefore, a necessary step to achieve this goal is the use of mathematical models and the development of novel statistical techniques to characterize and simulate these networks.

It is well known that under evolutionary pressure, prokaryote cells like E. coli behave optimizing their growth performance [5]. Flux Balance Analysis (FBA) provides a powerful tool to predict the optimal growth and production fluxes, but is not reliable about phenotypic state the cell will acquire. This is mainly due to the fact that among the infinitely many potential network states compatible with the stoichiometric constraints, FBA chooses a single one whose biological meaning is at least questionable under generic external conditions. FBA maximizes a linear function (usually the growth rate of the cell) subject to biochemical and thermodynamic constraints [6]. On the other hand, cells with genetically engineered knockouts or bacterial strains that were not exposed to evolution pressures, need not to optimize their growth. In fact, the method of minimization metabolic adjustment [7] has shown that knockout metabolic fluxes undergo a minimal redistribution with respect to the flux configuration of the wild type. Yet, in more general situations, the results are unpredictable, therefore, a tool to characterize the shape and volume of the whole space of possible phenotypic solutions must be welcome.

So far, apart from exact algorithms evaluating the volume of the space of possible solutions, that are unsuitable for analyzing metabolic networks larger than some dozens metabolites [8,9], the best technique allowing for such a characterization is based on Monte Carlo sampling (MCS) of the steady-state flux space [10-14]. This method is known to perform very well on intermediate size metabolic networks (up to a hundred of metabolites) [10,11] where different strategies of MCS have been implemented giving comparable results. Some improved variant of MCS seems to perform well also on organism-wide where the number of variables is in the range order of a thousand [13,14]. Although MCS has in general some intrinsically associated problems, mainly due to the fact that the convergence (or mixing) time is hard to assess and often is exponential, in the case of polytope volume estimations it turns out that sampling strategies such as hit and run have mixing times that scale only polynomially with system size [15]. However in many concrete cases a practical problem is to give a precise condition for convergence therefore we believe that an alternative independent technique could be more than welcome even in the cases where MCS is applicable.

As a concrete step toward an efficient characterization of the set of fluxes compatible with the stoichiometric constraints, we propose a novel message-passing technique derived from the field of statistical physics and information theory.

Mathematical Model
As already mentioned, a metabolic network is an engine that converts metabolites into other metabolites through a series of intra-cellular intermediate steps. The fundamental equation characterizing all functional states of a reconstructed biochemical reaction network is a mass conservation law that imposes simple linear constraints between the incoming and outcoming fluxes at any chemical reaction:

\[
\frac{\partial \rho}{\partial t} = i + \hat{S} \cdot v - o
\]

where \(\rho\) is the vector of the \(M\) metabolite concentrations in the network, \(i\) (\(o\)) is the input (output) vector of fluxes, and \(v\) are the reaction fluxes governed by the \(N \times N\) stoichiometric linear operator \(\hat{S}\) (usually named stoichiometric matrix) encoding the coefficient of the \(M\) intra-cellular relations among the \(N\) fluxes.

As long as just steady-state cellular properties are concerned one can assume that a variation in the concentration of metabolites in a cell can be ignored and considered as constant. Therefore in case of fixed external conditions one can assume metabolites (quasi) stationarity and consequently the \(lhs\) of 1 can be set to zero. Under these hypotheses the problem of finding the metabolic fluxes compatible with flux-balance is mathematically described by the linear system of equations

\[
\hat{S} \cdot v = o - i \equiv b
\]
\( b \) is the net metabolite uptake by the cell. Without loss of generality we can assume that the stoichiometric matrix \( \dot{S} \) has full rows rank, \( i.e. \) that \( \text{rank}(\dot{S}) = M \), since linearly dependent equations can be easily identified and removed. Given that the number of metabolites \( M \) is lower than the number of fluxes \( N \) the subspace of solutions is a \((N-M)\)-dimensional manifold embedded in the \( N \)-dimensional space of fluxes. In addition, the positivity of fluxes, together with the experimentally accessible values for the maximal fluxes, limit further the space of feasible solutions. This fact may be expressed by the following inequalities:

\[
0 \leq \nu \leq \nu_{\text{max}} \quad (3)
\]
in such a way that together, 2 and 3, define the convex set of all the allowed time-independent phenotypic states of a given metabolic network.

**Sub-dimensional volumes**

Mathematically speaking, the space of feasible solutions consistent with the Equations 2 is an affine space \( V \subset N \) of dimension \( N - M \). The set of inequalities 3 then defines a convex polytope \( \Pi \subset V \) that, from the metabolic point of view, may be considered as the allowed configuration space for the cell states. The main goal of this work is computation of the volume of this space of solutions and of certain subspaces of it. Although conceptually simple, the notion of sub-dimensional volume like that of \( \Pi \) requires some new definitions.

Consider any linear parameterization \( \phi : N - M \rightarrow V \subset N \) (see explicative scheme in Figure 1). A popular choice for \( \phi \) is, for instance, the inverse of the so called lexicographical projection \( i.e., \) the projection over the first \( N - M \) coordinates such that its restriction to \( V \) has an inverse. Being \( \phi \) linear, the \((N - M) \times N\) Jacobian matrix \( \dot{\lambda} \) is constant and coincides with the matrix of \( \phi \) in the canonical bases.

Denoting \( \lambda = \det (\lambda^T \lambda)^{\frac{1}{2}} \), the Euclidean metric in \( N \) induces a measure on \( V \) (which does not depend on \( \phi \)):

\[
\int_V f(\nu) d\nu \equiv \lambda \int f(\phi(u)) du \quad (4)
\]

allowing to compute the volume of our polytope

\[
\text{vol}_V(\Pi) = \int_V 1_\Pi(\nu) d\nu = \lambda \int 1_{\phi^{-1}(\Pi)}(u) du \quad (5)
\]

where \( 1_\Pi(\cdot) \) is the indicator function of the set \( \Pi \). It is worth pointing out that given the linear structure of the metabolic equations, the determinant of the mapping is a (scalar) constant. Although the coefficient \( \lambda \) could be explicitly calculated, it turns out that as far as only relative volume quantities are concerned, as in the case of the *in silico* flux knock-outs introduced below, this term factors out and therefore we will drop it from the rest of the computation.

**Probabilistic framework**

The problem of describing the polytope \( \Pi \) can be cast into a probabilistic framework. We define the probability density \( \mathcal{P} \) as:

\[
\mathcal{P}(\nu) = \text{vol}_V(\Pi)^{-1} 1_\Pi(\nu) \quad (6)
\]

Marginal flux probabilities over a given set of fluxes are obtained by integrating out all remaining degrees of freedom. In particular we can define single flux marginal probability densities as integrals on the affine subspace \( W = V \cap \{ \nu_j = \vec{\nu} \} \)

\[
P_j(\nu_j) = \int_W \mathcal{P}(\nu) \prod_{j \neq i} d\nu_j = \frac{\text{vol}_W(\Pi \cap W)}{\text{vol}_V(\Pi)} \quad (7)
\]
where the normalization term $\text{vol}_W(\Pi \cap W)$ is the (subdimensional) volume of the intersection between the polytope $\Pi$ and the hyperplane $\{ \nu_i = \bar{\nu} \}$ as displayed schematically in Figure 2 where the marginal probability at point $\nu_i = \bar{\nu}$ is proportional to length of the blue segment which is the intersection between the polytope $\Pi$ (the grey triangle) and the plane $\nu_i = \bar{\nu}$.

**Approximate volume computation**

From a computational point of view, the problem of the exact computation of the volume of a polytope with current methods requires the enumeration of all its vertexes. The vertex enumeration problem is #P-hard [16,17], but even the problem of computing the volume, given the set of all vertexes is a big computational challenge. Various algorithms exist for calculating the exact volume of a polytope from its vertexes (for a review see [18]), and many software packages are available in the Internet. Computational limitations restrict however exact algorithmic strategies to cope with polytopes in relatively few dimensions (e.g. $N \cdot M$ around 10 or so). To overcome such severe limitations we will introduce a very efficient approximate computational strategy that will allow us to compute the volume and the shape of the space of solutions for real-world metabolic networks.

We will adopt the following three steps strategy:

1. We discretize the problem *a la Riemann* considering an $N$-dimensional square lattice whose elementary cell is of size $\varepsilon$. The approximated volume is then proportional to the number of cells intersecting the polytope $\Pi$. For any given $\varepsilon$ one we have then defined discrete variables $\nu_i \in \{0, 1, ..., q_i^{\text{max}}\}$, for $q_i^{\text{max}}$ equal to the integer part of $q^{\text{max}} \times \nu_i^{\text{max}}$, where the integer $q^{\text{max}}$ is the granularity of the approximation.

2. We consider an integer constraint satisfaction problem where each of the mass-balance equations set a hard constraint over the involved discretized fluxes.

3. We solve the constraint satisfaction problem using a message-passing algorithm called Belief Propagation.

**Discretization**

Consider the regular orthogonal grid $\Lambda_\varepsilon$ of side $\varepsilon$ partitioning $\mathbb{R}^N$. The number of cells $N_\varepsilon$ of $\Lambda_\varepsilon$ intersecting $\Pi$ is proportional to the numbers of cells of $\Gamma_\varepsilon$ intersecting $\varphi^{-1}(\Pi)$. Finally, the volume in Eq. 5 is proportional to $\lim_{\varepsilon \to 0} \varepsilon^N N_\varepsilon$. For any given $\varepsilon$ one we have then defined discrete variables $\nu_i \in \{0, 1, ..., q_i^{\text{max}}\}$, for $q_i^{\text{max}}$ equal to the integer part of $q^{\text{max}} \times \nu_i^{\text{max}}$, where the integer $q^{\text{max}}$ is the granularity of the approximation.
The constraint satisfaction problem

When dealing with integer coefficients \( s_{i,\nu} \) as the ones appearing in normal stoichiometric relations, the discrete version of Eqs. 2 close in the set of positive integers defining a constraint satisfaction problem. The \( \nu \)-approximation of the volume is then the number of elementary cells that are solution to the discretized mass-balance equations. It is interesting to note that in the case of fractional stoichiometric relations one can multiply all terms for the minimum common multiple of all denominators, getting an equivalent mass-balance equation with integer coefficients.

Belief Propagation

The integer constraint satisfaction is solved using Belief Propagation (BP), a local iterative algorithm that allows for the computations of marginal probability distributions. BP is exact on trees, and perform reasonably well on locally tree-like structures [19-22]. This approximation scheme allows for the computation of the logarithm of the number of solutions via the entropy that can be expressed in terms of flux marginals. We will give a detailed derivation of the equations in section Methods.

Results and Discussion

Performance on low dimensional systems

In this section we will analyze the performance of our algorithm against an exact algorithm on low dimensional polytopes. Among the different packages available in the Internet, we have chosen LRS [8], a program based on the reverse search algorithm presented by Avis and Fukuda in [9] that can compute the volume of non-full dimensional polytopes. Actually, it computes the volume of the lexicographically smallest representation of the polytope, that for the benchmark used below, coincides with the conventional volume estimated by our algorithm.

We have devised a specific benchmark generating random diluted stoichiometric matrices at a given ratio \( \alpha = M/N \) and fixed number of terms different from zero \( K \) in each of the reactions. All fluxes were constrained inside the hypercube \( 0 \leq \nu \leq 1 \). As a general strategy we have calculated several random instances of the problem and measured the volume (entropy) of the polytope using the LRS and BP algorithm. In particular, we have first generated 1000 realizations of random stoichiometric matrices with \( N = 12, M = 4 \). Note that \( N = 12 \) is around the maximum that allows simulations with LRS in reasonable time (around one hour per instance). For each polytope then we have computed the two entropies \( S_{LRS} \) and \( S_{BP} \) with both algorithms, fixing the same maximum value for the discretization \( q_{\text{max}} = 1024 \) for all fluxes.

In Figure 4 we show how the quality of the BP measure is affected by the discretization, by displaying the histogram of the relative differences \( \delta S = \frac{S_{BP} - S_{LRS}}{S_{LRS}} \) with an increasing number of bins per variable \( q_{\text{max}} = 16, 64, 256, 1024 \). One can see how a finer binning of messages improves the quality of the approximation, seemingly converging to a single distribution of errors. It is expected that for larger \( N \) the histograms would peak around the true value: upon increasing the number of fluxes, loops become larger and the overall topology of the graph becomes more locally tree-like, validating the hypothesis behind the Bethe approximation. Unfortunately, the huge increase of computer time experimented in the calculation of the volumes using LRS made impossible to test systems large enough to make any reasonable scaling analysis.

Finally we address the issue of the computational complexity of the algorithm which is a crucial one if one is interested in approaching real world metabolic networks whose size typically is at least 50 times the size of the largest network that can be analyzed with exact algorithms. In Figure 5 we display the running time of both LRS and BP as a function of the number of fluxes \( N \). Interesting, LRS outperforms BP up to sizes \( N \sim 12 \) where the running time...
of LRS explodes exponentially while BP maintains a modest almost-linear behavior.

**Distribution of fluxes in Red Blood Cell**

We used our BP algorithm and MCS [11] to obtain the marginal flux distributions for each of the reactions in the Red Blood Cell (RBC) metabolism taking the same stoichiometric matrix presented in [10,11] (see Additional file 1). The network contains 46 reactions and 34 metabolites. The first comparison of MCS with our BP algorithm is done setting all \( q \) to 1 (the resulting marginal probability distribution for the fluxes are displayed in Figure 6). The second comparison is done by setting all \( v_{\text{max}} \) of reaction \( \nu_i \) to 1 (the resulting marginal probability distribution for the fluxes are displayed in Figure 6). The second comparison is done by setting all \( v_{\text{max}} \) to the values used in [10] (results are displayed in Figure 7). In both settings we compared BP with a set of 5000 feasible solutions generated by MCS, while for the BP algorithm we used a \( q_{\text{max}} = 2048 \). As it can be seen in the two cases, the predictions of both methods compare rather well. Assuming the accuracy of MCS, differences are probably due to small loops structures in the graph. We leave for a future work the issue of a more detailed comparison of both methods. However we really think that the emerging scenario of RBC metabolism captured by BP is analogous to that obtained by MCS both quantitatively and qualitatively.

The MCS method appears to be quite efficient (the authors of [11] reported in a similar network less then 30 seconds of computer computation in a Dell Dimension 8200 to obtain their distributions) while our algorithm converged to the same result in less than 3 seconds on a similar machine (Intel CPU 6600 2.40 GHz). Of course, no stringent statement can be done at this level about the comparative performance of the two algorithms. This positive result encourages us to face the problem of metabolism at organism-wide scale.

**Analysis of gene knock-out in E. coli central metabolism**

In this section we study the influence of partial flux knock-out on the volume of the solution space. We concentrate on E. coli central metabolism [23]. The network has 62 metabolites, 104 reactions (75 internal reactions and 29 exchange fluxes). All reactions were considered irreversible following their nominal directions, and maximum flux rates were set to 1.

In this numerical experiment we first run BP on the unperturbed system measuring the volume of the solution space \( S_0 \). Then the maximum flux values are kept constant and equal to 1 for all the reactions but one (say reaction \( \nu_i \)). The partial knock-out effect on flux \( i \) is then obtained reducing repeatedly \( v_{\text{max}} \) and computing again the volume \( S_{\text{KO}}(v_{\text{max}}) \) at each time until \( v_{\text{max}} = 0 \). In principle at each reduction step of \( v_{\text{max}} \) one should converge again the BP equations. In practice for most of the fluxes convergence is very fast since at each reduction of \( v_{\text{max}} \) the new stationary point will be in general close enough to the old one. However we have experimentally noted that the larger is the impact of a given flux on the volume of the space of solutions, the longer is the convergence time at each reduction step. Let us point out that doing so we are tacitly assuming that each knock-out is independent from the others although it is known that some reactions might be associated with the same enzyme. However there is no computational restriction to analyze multiple knock-outs. An analogous technique was presented in [11], where maximal fluxes were reduced to the value of 75% of their original maximal allowed flux to mimic enzymopathies.

In Figure 8 we display the whole set of \( S_0 - S_{\text{KO}}(v_{\text{max}}) \) vs. knock-out percentage curves. We can observe how heterogeneous is the impact of the different fluxes on the volume. Moreover one can observe how different curves may cross depending on the knock-out percentage displaying thus an intriguing scenario of differential flux-reduction
impacts. Let us now concentrate on the 20 fluxes having larger impact on the space of solutions: in Figure 9 we display complete knock-outs values \( S_0 - S_{KO}(\nu_i^{max} = 0) \). Focusing only on internal fluxes (the first two fluxes are indeed exchange fluxes of water and protons), one can observe that the first half of them basically compose the backbone of glycolysis showing little pathway redundancy in the network, while reactions like FUM, ACON, SUC and SUCCD1i appear in the Krebs cycle and again show little pathway redundancy in the network (see Additional file 2).

Finally in Figure 10 we display the correlations between the changes in the entropy for different reaction knock-outs and the average flux \( \nu_i = \int \nu \mathrm{d}v \) in the unperturbed network. At 75% knock-out, two kinds of regimes are divided by a clear threshold at \( \nu \approx 0.6 \): (i) for \( \nu_i < 0.6 \), \( S_0 - S_{KO}(\nu_i^{max}) \) has a small positive correlation with \( \nu_i \), (ii) at larger average fluxes, correlations increase rapidly but with larger fluctuations. The presence of this threshold can be understood noting that reactions belonging to the linear (glycolysis) an circular (Krebs cycle) pathways are, in the wild cell, fast flux reactions, with average flux values larger than 0.5. An analogous scenario emerges in the case of a 100% knock-outs, but now fluctuation are wider, and also fluxes with intermediate average value start becoming important. This is the case for instance of the 6 large impact fluxes having average flux around 0.3. A closer inspection reveals that among them there is an exchange fluxes (glycogen) and 5 internal fluxes (G1PP, GLCP, HEX1, CS, PDH). The first three (G1PP, GLCP, HEX1) are the first steps of glycolysis, while PDH is the input of the Krebs cycle, and CS is a segment of the cycle strictly related to PDH. It is interesting that this peculiar behavior (large impact and relatively small average flux) are related to key check-points of central metabolism.

E. coli metabolism at organism-wide scale

Finally we analyze the average fluxes distribution function in the metabolic network of E. coli. The network used contains, in its original format, 1035 reactions and 626 metabolites [23]. The network has been preprocessed using Gaussian elimination procedure: when a trivial mass-balance equation of the type \( \nu_i = \nu \) is present we eliminate variable \( \nu_i \) using elementary row operations. Doing so new trivial mass balance equations appear, an then we iterate the eliminations until all trivial mass-balance equations disappear. During Gaussian elimination,
A power law distribution of these average values is displayed in Figure 11. As can be seen in [14,24,25], the probability distribution function (pdf) of the marginal flux distributions is used. It is claimed in [24] that the robustness of this value distributions have been considered using bidirectional signed fluxes. Reversible reactions have been treated as unidirectional fluxes. Using BP we are able to compute the marginal flux distributions in around 40 minutes, using $q_{\text{max}} = 64$.

We ran our algorithm on this network and computed the average fluxes in each reaction. We have performed a statistical analysis of these averages in the spirit of [14,24,25]. The probability distribution function (pdf) of these average values is displayed in Figure 11. As can be clearly seen the distribution is large and may be fitted with a power law distribution of $P(v) \propto (v + v_0)^{-\gamma}$. The fit gives an exponent $\gamma$ around 1.5, a result that compares very well with previous simulations found in: (i) [14] where FBA optimal fluxes were averaged over many different external conditions, (ii) [24] where a maximal local-output strategy is used, and (iii) [25] where a Gaussian approximation for the marginal flux distributions is used. It is claimed in [14] that the robustness of this value $\gamma$ is a signature of universality in cell’s metabolism.

A more careful analysis of the data may reveals however that the distribution of averages fluxes has a richer structure. In Figure 12 we present the cumulative distribution function $P_\gamma(v) = \int_0^v P(y)dy$ of the average fluxes of the reactions and a clear jump appears for $\nu \approx 0.5$, and smaller ones for $\nu \approx 0.4$ and $\nu \approx 0.6$. At present we do not know whether these jumps are just due to statistical fluctuations (and are correctly smeared out in the usual binning process done to plot the pdf in double logarithmic scale) or they reflect relevant biological or structural information about the network.

**Conclusion**

We proposed a novel algorithm to estimate the size and shape of the affine space of a non full-dimensional convex polytope in high dimensions. The algorithm was tested in specific benchmark, i.e. random diluted stoichiometric matrices at a given ratio $\alpha = M/N$ and fixed number of terms different from zero $K$, in each of the reactions, with results that compare very well with those of exact algorithms. Moreover, we show that while the running time of exact algorithms increases more than exponentially for already moderate sizes, our algorithm is polynomial. The program was run on the Red Blood Cell metabolism producing with shorter computational time results that are in both quantitatively and qualitatively in good agreement with those obtained by MCS presented in [10,11].
Then, our program was used to study the \textit{E. coli} central metabolism, and as expected, reactions with little redundancy turned out to be the ones with larger impact in the size of the space of the metabolic solutions. Specifically, most of the reactions associated with the transformation of glucose in pyruvate, belong to this set, as well as some reactions in the citric cycle. In addition we show strong correlations between the characteristics of the flux distributions of the wild type network and the changes in size of the space of solutions after reaction knock-outs. Finally, we calculate the distribution of the average values of the fluxes in the metabolism of the \textit{E. coli} and present results that are consistent with those of the literature. In the present approach we have mainly followed a discretization strategy that although polynomial, becomes computationally rather heavy for mass balance equation containing a large number of fluxes. We are currently investigating other representation schemes for the messages. The final hope is to obtain an algorithm that allows for on-line analysis of organism-wide metabolic networks.

Let us conclude by noting that in principle the presented approach can be extended to deal with constraints whose functional form is more general than linear, provided that the number of variables involved in each of the constraints remains small, as in the case of inequalities enforcing the second law of thermodynamics for the considered reactions \cite{26}. Work is in progress also in this direction.

**Methods**

**Belief Propagation**

In the previous section we have seen how the metabolic problem can be cast into a constraint satisfaction framework where each of $M$ mass-balance equations imposes a
constraint onto a subset of the metabolic fluxes. Let \( A \) be the set of equations and \( I \) the set of fluxes. Consider the \( a \)-th row of \( \hat{S} \), and let \( \{ i \in a \} = \{ i_1, ..., i_n \} \subset I \) be the labels of the fluxes involved in the considered equation having stoichiometric coefficients different from zero. Let also \( \{ a \in i \} = \{ a_1, ..., a_n \} \subset A \) be the labels of the equations in which flux \( i \) participates. The emerging structure is a bipartite graph, with two types of nodes: variable nodes representing the fluxes of the reactions and factor nodes imposing mass conservation. In this case marginals become \( q \)-modal probability densities that for large values of \( q \) will approximate better and better the continuous set of probabilities.

Under the hypothesis that the factor graph is a tree it can be shown [19,27] that a given flux vector \( \nu \) satisfying all flux-balance constraints can be expressed as a product of flux and reaction marginals [19-21]:

\[
P_{\nu}(\nu) = \prod_{a \in A} P_a(\{ \nu \}_{|_{\nu \neq 0}}) \prod_{i \in I} P_i(\nu_i)^{1-d_i}
\]

(8)

where \( d_i \) is the number of equations in which flux \( \nu_i \) participates (i.e. the degree of site \( i \)). The marginal probabilities are defined as:

\[
P_a(\{ \nu \}_{|_{\nu \neq 0}}) = \sum_{\{ \nu \}_{|_{\nu \neq 0}}} \mathcal{P}(\nu)
\]

\[
P_i(\nu_i) = \sum_{\nu_{\neq i}} \mathcal{P}(\nu).
\]

(9)

One can then define an entropy in terms of the marginal probability distributions which amounts to the logarithm of the number of solutions of the associated constraint satisfaction problem. From a more geometrical point of view, this entropy is a count of the (logarithm of) the number of elementary \( \epsilon \)-cells intersecting the polytope \( \Pi \) (see Figure 3).

\[
S = \sum_{\nu} \mathcal{P}(\nu) \ln P(\nu)
\]

\[
= \sum_{a \in A} \sum_{\{ \nu \}_{|_{\nu \neq 0}}} P_a(\{ \nu \}_{|_{\nu \neq 0}}) \ln P_a(\{ \nu \}_{|_{\nu \neq 0}}) - \sum_{i \in I} (d_i - 1) P_i(\nu_i) \ln P_i(\nu_i)
\]

(10)

One may wonder how such an approach could be useful in a real-world situation where the graph is not a tree. One can hope that typical loop length are large enough to ensure weak statistical dependence of neighboring sites which lay at the heart of the Bethe approximation [28,29].

It is interesting to note that the Bethe approximation is successfully used in many different problems with loopy graph topologies. This is for instance the case for LDPC error correcting codes in Information Theory [30] used in...
wireless Internet transmission technologies such as WiMAX, or in many inference problems such as graphical models [31], binary perceptron learning [32], and in constraint satisfaction problems such as Satisfiability or Coloring [33,34]. In all these cases although there is no mathematically rigorous proof about the quality of the solution, whenever the algorithm converges, the result generally provides a good estimate of marginal probability distributions.

The algorithm is based on two type of messages exchanged from variable nodes to functional nodes, and vice versa:

- $\mu_i \rightarrow a(\nu)$: the probability that flux $i$ takes value $\nu$ in the absence of reaction $a$.
- $m_i \rightarrow a(\nu)$: the non-normalized probability that the balance in reaction $a$ is fulfilled given that flux $i$ takes value $\nu$.

The two quantities satisfy the following set of functional equations:

$$m_a(\nu) = \sum_{\{\nu_i\}_{i \neq a}} \delta \left( \sum_{i \in a} s_{ai} \nu_i; b_a \right) \prod_{i \in a} u_{i \rightarrow a}(\nu_i)$$

$$\mu_{i \rightarrow a}(\nu) = C_{i \rightarrow a} \prod_{b \in i, b \neq a} m_{b \rightarrow a}(\nu_i)$$  \hspace{1cm} (11)

where $\sum_{\{\nu_i\}_{i \neq a}}$ means the sum over all values of fluxes around metabolite $a$ but $i, b \in i \setminus a$, is the set of metabolites in reaction $i$ but $a$. $C_{i \rightarrow a}$ is a constant enforcing the normalization of the probability $\mu_i \rightarrow a(\nu)$, and $\delta (\cdot; \cdot)$ is the Kronecker delta function ($\delta(a; b) = 1$ if $a = b$ and 0 otherwise). The set of Equations 11 can be solved iteratively and upon convergence of the algorithm one can compute the marginal flux distributions as:

$$P_{\nu}(\{\nu_i\}_{i \neq a}) = C_a \sum_{\{\nu_i\}_{i \neq a}} \delta \left( \sum_{i \in a} s_{ai} \nu_i; b_a \right) \prod_{i \in a} u_{i \rightarrow a}(\nu_i)$$

$$P_{\nu}(\nu) = C_i \prod_{b \in i} m_{b \rightarrow i}(\nu)$$

A brute force integration of the discrete set of equation would be much too inefficient for analyzing large networks, due to the multiple dimensional sum over $\{0, \ldots, \nu_{i \rightarrow a}^{\text{max}}\}_{i \neq a}$ in the previous equation. The first of Equations 11 includes the computation of the convolution of all $\mu_{i \rightarrow a}$ messages except one, an expression of the form $C_{i \rightarrow a} = \otimes_{i \neq a} \mu_{j \rightarrow a}$ (where denotes convolution) and requires normally $n_a - 1$ convolutions for each outputting message, i.e. $n_a (n_a - 1)$ convolutions in total for constraint $a$. In the case of the complete $E. coli$ network we have mass-balance equations with $n_a$ as large as 500, then
Authors’ contributions

Authors equally contributed to this work. All authors read and approved the final manuscript.

Additional material

Additional file 1

Human red blood cell metabolism supplementary files. The archive contains 4 text files: rbc.sto (stoichiometric matrix where each column is a reaction and each line is a metabolite), rbc.flu (one column, contains the sign of exchange fluxes: negative are incoming, positive outcoming, zero free), rbc.names (contains the name of both reactions and metabolites), and rbc.max (contains maximum flux values of the reactions). Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2105-9-240-S1.tgz]

Additional file 2

E. coli central metabolism supplementary files. The archive contains 3 text files: central.sto (stoichiometric matrix where each column is a reaction and each line is a metabolite), central.flu (one column, contains the sign of exchange fluxes: negative are incoming, positive outcoming, zero free), central.names (contains the name of both reactions and metabolites). Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2105-9-240-S2.tgz]

Acknowledgements

AB was supported by Microsoft TCI grant. RM wants to thank the Interna
tional Center for Theoretical Physics in Trieste and the Center for Molecular Immunology of La Habana, for their cordial hospitality during the completion of this work. We are also very grateful to Ginestra Bianconi, Michele Leone, Martin Weigt, and Riccardo Zecchina, for very interesting discussions, and in particular to Carlotta Martelli for sharing with us a human readable E. coli data set. We are really grateful to Jan Schellenberger for providing us the MCS data used in Figure 6 and 7.

References

1. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: The large-
scale organization of metabolic networks. Nature 2000, 407(8040):651-654.
2. Fell DA, Wagner A: The small world of metabolism. Nature Bio-
technology 2000, 18:111-1112.
3. Z D, Qin ZS: Structural comparison of metabolic networks in selected single cell organisms. BMC Bioinformatics 2005, 6(8):
4. Kanehisa M, Goto S, Hattori M, Akid-Kinosita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006, 34:D354-7.
5. Ibarra AU, Edwards J, Palsson B: Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 2002, 420:186-189.
6. Varma A, Palsson B: Metabolic Capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors. J theor Biol 1993, 165:477-502.
7. Segre D, vidup D, Church G: Analysis of optimality in natural and perturbed metabolic networks. PNAS 2002, 99:15112-15117.
8. LRS package [http://cgm.cs.mcgill.ca/~avis/C/lrs.html]
9. Avis D, Fukuda K: A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. Discrete Comput Geom 1992, 8(3):295-313.
10. Wilbeck S, Famili I, Greenberg HJ, Palsson B: Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. J Theor Biol 2004, 228:437-447.

11. Price N, Schellenberger J, Palsson B: Uniform Sampling of Steady-State Flux Spaces: Means to Design Experiments and to Interpret Enzymopathies. Biophysical Journal 2004, 87:2172-2186.

12. Thiele I, Price N, Vo T, Palsson B: Impact of diabetes, ischemia, and diet. J Biol Chem 2005, 280:11683-11695.

13. Price N, Thiele I, Palsson B: Candidate States of Helicobacter pylori’s Genome-Scale Metabolic Network upon Application of “Loop Law” Thermodynamic Constraints. Biophysical Journal 2006, 90:3919-3928.

14. Almaas E, Kovacs B, Vicsek T, Oltval Z, Barabasi AL: Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 2004, 427:839-843.

15. Simonovits M: How to compute the volume in high dimension? Math Progr 2003, 97(1-2):337-374.

16. Dyer M, Frieze A: On the complexity of computing the volume of a polyhedron. SIAM J Comput 1988, 17(5):967-97.

17. Khachiyan L: Complexity of volume computation. New trends in discrete and computational geometry, Springer-Verlag 1993:91-101.

18. Beuler B, Enge A, Fukuda K: Exact volume computation for convex polytopes: A practical study. Polytopes-combinatorics and computation, Birkhauser 2000:131-154.

19. Yedidia J, Freeman W, Weiss Y: Generalized belief propagation. Advances in Neural Information Processing Systems (NIPS) 13, Denver, CO 2001,772-778.

20. Kschischang FR, Frey BJ, Loeliger HA: Factor graphs and the sum-product algorithm. Information Theory, IEEE Transactions on 2001, 47(2):498-519.

21. Braunstein A, Mezard M, Zecchina R: Survey propagation: An algorithm for satisfiability. Random Struct Algorithms 2005, 27:201-226.

22. MacKay DJC: Information Theory, Inference, and Learning Algorithms Cambridge University Press; 2003.

23. Reed J, Vo T, Schilling C, Palsson B: An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biology 2003, 4(9):R54.

24. De Martino A, Martelli RMC, Castillo IP: Von Neumann expanding model on random graphs. Journal of Statistical Mechanics: Theory and Experiment 2007, 2007(05):P05012.

25. Bianconi G, Zecchina R: Viable flux distribution in metabolic networks. Technical report 2007 [http://aps.arxiv.org/pdf/0705.2816]. ArXiv:cond-mat/0705.2816

26. Beard D, Babson E, Curtis E, Qian H: Thermodynamic constraints for biochemical networks. J Theor Biol 2004, 228:327-333.

27. Baxter B: Exactly Solved Models in Statistical Mechanics London: Academic Press Inc; 1989.

28. Mezard M, Parisi G: The Bethe lattice spin glass revisited. The European Physical Journal B 2001, 20:217.

29. Mezard M, Parisi G: The cavity method at zero temperature. J Stat Phys 2003, 111:1.

30. Richardson T, Urbanke R: The capacity of low-density parity check codes under message passing decoding. IEEE, Trans Info Theory 2001, 47:599-618.

31. Weiss Y, Freeman WT: Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology. Neural Comp 2001, 13(10):2173-2200.

32. Baldassi C, Braunstein A, Brunel N, Zecchina R: Efficient supervised learning in networks with binary synapses. PNAS 2007, 104:11679-11684 [http://aps.arxiv.org/pdf/0707.1295]

33. Mezard M, Parisi G, Zecchina R: Analytic and Algorithmic Solution of Random Satisfiability Problems. Science 2002, 297:812 [http://aps.arxiv.org/pdf/0707.1295]

34. Mulet R, Pagnani A, Weigt M, Zecchina R: Coloring random graphs. Phys Rev Lett 2002, 89:268701 [http://aps.arxiv.org/abs/cond-mat/0208460].