POLITECNICO DI TORINO
Repository ISTITUZIONALE

A novel methodology to estimate metabolic flux distributions in constraint-based models

Original
A novel methodology to estimate metabolic flux distributions in constraint-based models / Massucci, F.A., Font-Clos, F., De Martino, A., Castillo, I.P.. - In: METABOLITES. - ISSN 2218-1989. - 3:3(2013), pp. 838-852. [10.3390/metabo3030838]

Availability:
This version is available at: 11583/2976731 since: 2023-03-10T08:50:16Z

Publisher:
MDPI

Published
DOI:10.3390/metabo3030838

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

11 November 2026

Metabolites **2013**, *3*, 838-852; doi:10.3390/metabo3030838

OPEN ACCESS

metabolites

ISSN 2218-1989

www.mdpi.com/journal/metabolites

Article

A Novel Methodology to Estimate Metabolic Flux Distributions in Constraint-Based Models

Francesco Alessandro Massucci [1,*], Francesc Font-Clos [2,3], Andrea De Martino [4,5,6] and Isaac Pérez Castillo [7]

[1] Department d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona 43007, Spain

[2] Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, Bellaterra (Barcelona) E-08193, Spain; E-Mail: fontclos@crm.cat

[3] Department de Matemàtiques, Universitat Autònoma de Barcelona, Edifici C, Bellaterra (Barcelona) E-08193, Spain

[4] Dipartimento di Fisica, Sapienza Università di Roma, p. A. Moro 2, Rome 00185, Italy; E-Mail: andrea.demartino@roma1.infn.it

[5] CNR/IPCF, UOS di Roma, Roma 00185, Italy

[6] Center for Life Nano Science@Sapienza, Istituto Italiano di Tecnologia, v. Regina Elena 291, Rome 00161, Italy

[7] Department of Mathematics, King's College London, London WC2R 2LS, UK; E-Mail: isaac.perez_castillo@kcl.ac.uk

* Author to whom correspondence should be addressed; E-Mail: francesco.massucci@urv.cat; Tel.: +34-977-558-580; Fax +34-977-559-621.

Received: 1 August 2013; in revised form: 10 September 2013 / Accepted: 10 September 2013 / Published: 20 September 2013

Abstract: Quite generally, constraint-based metabolic flux analysis describes the space of viable flux configurations for a metabolic network as a high-dimensional polytope defined by the linear constraints that enforce the balancing of production and consumption fluxes for each chemical species in the system. In some cases, the complexity of the solution space can be reduced by performing an additional optimization, while in other cases, knowing the range of variability of fluxes over the polytope provides a sufficient characterization of the allowed configurations. There are cases, however, in which the thorough information encoded in the individual distributions of viable fluxes over the polytope is required. Obtaining such distributions is known to be a highly challenging computational task when the dimensionality of the polytope is sufficiently large, and the problem of developing cost-effective *ad hoc* algorithms has recently seen a major surge of interest. Here, we propose a method that

allows us to perform the required computation heuristically in a time scaling *linearly* with the number of reactions in the network, overcoming some limitations of similar techniques employed in recent years. As a case study, we apply it to the analysis of the human red blood cell metabolic network, whose solution space can be sampled by different exact techniques, like Hit-and-Run Monte Carlo (scaling roughly like the third power of the system size). Remarkably accurate estimates for the true distributions of viable reaction fluxes are obtained, suggesting that, although further improvements are desirable, our method enhances our ability to analyze the space of allowed configurations for large biochemical reaction networks.

Keywords: metabolic networks; flux balance analysis; belief propagation algorithm

1. Introduction

The development of high throughput techniques now makes available a considerable number of high quality reconstructions of the metabolism of a variety of organisms, which include the stoichiometry of the biochemical reactions in the network and the underlying enzyme-gene associations [1–4]. Ideally, this information may be employed for kinetic modeling approaches that could shed light on issues, like the organization of a cell's metabolic phenotype and its robustness to perturbations (internal or exogenous) in a fully dynamical setting. Indeed, consider a metabolic network of N coupled chemical reactions transforming M metabolites and let $\boldsymbol{\xi} = \{\xi_i^\mu\}$ $(i = 1, \ldots, N;\ \mu = 1, \ldots, M)$ denote the stoichiometric coefficients, with the standard sign convention to distinguish substrates ($\xi_i^\mu < 0$) from products ($\xi_i^\mu > 0$) in each reaction, i. If one denotes by γ^μ the rate of change of the intracellular level of species μ, due to exchanges between the cell's interior and the environment, then, under mass action kinetics, the intracellular concentration, c^μ, of metabolite μ obeys the equation:

$$\frac{dc^\mu}{dt} = \sum_{i=1}^{N} \xi_i^\mu x^i - \gamma^\mu \tag{1}$$

where $\gamma^\mu > 0$ (resp.$\gamma^\mu < 0$) if there is a net out-take (resp. in-take) of species μ.

Unluckily, addressing the above system in full generality requires knowledge about reaction mechanisms and kinetic constants (which specify how rates depend on concentrations), which is at best only partially available. (Besides, it is not entirely clear to us that, were that information fully at our disposal, simulating (1) for a genome-scale reconstruction involving thousands of reactions and metabolites would be a sensible thing to do).

Computational studies of metabolic networks therefore generally assume that the cell operates at non-equilibrium steady-state (NESS) conditions, where the concentration of the metabolites is constant [5,6], with the rationale that as long as one is interested in the behaviour for time scales shorter than genetic ones (minutes), the much faster equilibrating biochemistry can be assumed to be "frozen" in an NESS. Under this assumption, computing NESS fluxes more modestly requires the prescription of

bounds for flux variability, *i.e.*, $x^i \in [m^i, M^i]$ (which also encode for reaction reversibility assumptions), and for exchange rates, *i.e.*, $\gamma^\mu \in [m^\mu, M^\mu]$, with which conditions (1) can be written as:

$$m^\mu \leq \sum_{i=1}^{N} \xi_i^\mu x^i \leq M^\mu \,, \qquad \mu = 1, \ldots, M \tag{2}$$

$$m^i \leq x^i \leq M^i \,, \qquad i = 1, \ldots, N \tag{3}$$

This set of inequalities is fairly general, since it may include metabolites involved only in internal reactions (for which $\gamma^\mu = 0$), as well as exchanged species (*i.e.*, with $\gamma^\mu \neq 0$). The solution space of Equation (2) is, in turn, given by:

$$S = \left\{ \boldsymbol{x} \in \mathbb{R}^N \text{ s.t. } m^\mu \leq \sum_{i=1}^{N} \xi_i^\mu x^i \leq M^\mu \;\; (\mu = 1, \ldots, M) \text{ and } m^i \leq x^i \leq M^i \;\; (i = 1, \ldots, N) \right\} \tag{4}$$

The problem we address here concerns computational methods to sample S, in order to retrieve information on quantities like the distribution of the allowed values of each flux over the entire solution space. While most studies aiming at developing predictive power on metabolic phenotypes (especially for microbes) have been thus far based on coupling (2) with a phenotypic optimization principle (the solving of which requires no sampling of S), it is being increasingly recognized that the information entailed by the structure of S might provide key insight into different aspects of metabolic network analysis, from flux-flux correlations, to robustness, to perturbations, to (more subtly) optimal experiment design [7]. The task, however, presents many challenges, especially from the viewpoint of CPU costs. The running time of the paradigmatic algorithm to sample high-dimensional polytopes, such as S, namely Hit-and-Run Monte Carlo, is known to scale, in the best case, as the third power of the number of variables in the system [8], making it impractical for large enough networks. On the other hand, it has recently been shown that the heuristics of message-passing (MP) algorithms may provide a powerful alternative [9]. In brief, MP-based protocols are designed to compute solutions to statistical inference problems, like estimating marginals of random variables, exploiting peculiar topological properties of the underlying graphs that describe the interdependence of variables (reactions in our case) and constraints (metabolites in our case). Indeed, when such a graph is locally tree-like, *i.e.*, it lacks short loops, statistical inference can be performed accurately by MP in linear time with the number of variables [10]. This property has been used in [9] to devise a highly efficient MP method to sample sets like S. Yet more recently [11], a second MP algorithm has been proposed to overcome some of the limitations of the method of [9], most notably, the inapplicability of the latter to real valued stoichiometry and, perhaps more importantly, the accuracy problems that may arise when (some) fluxes are allowed to span several orders of magnitude.

Here, we build on the work presented in [11] and apply a MP methodology, which we call *weighted Belief Propagation* (wBP), to the analysis of the solution space of metabolic reaction networks. In particular, we focus on the metabolism of the human red blood cell (hRBC), a major benchmark for sampling tools, as its size ($N = 46$, $M = 34$) makes it possible to characterize its solution space S by various methods and to compare the results. Specifically, we have evaluated results retrieved by wBP against those obtained by Hit-and-Run Monte Carlo to sample the polytope S for the hRBC (A note of caution here is needed regarding the words *Monte Carlo*, which are used throughout the

text, as well as in some of the references in a somewhat unspecific way. As is well-known, Monte Carlo techniques generically rely on repeated random sampling. Thus, Hit-and-Run is a Monte Carlo algorithm. The method used in [12] to sample a volume is also a Monte Carlo algorithm, albeit a different one, based on a rejection method. Methods to perform statistical inference by directly computing high-dimensional integrals, or sums, are often also Monte Carlo methods, as long as the integrals involved are evaluated by Monte Carlo integration. Therefore, by itself, *Monte Carlo* is perhaps too generic a term and does not reveal the specifics of an algorithm. The reader is advised to check the references carefully when comparing the various techniques). Because of the importance of Hit-and-Run as a mathematically controlled procedure with broad applicability, we have tried to devise a Monte Carlo method with computing times reduced as much as possible (the limit of standard Hit-and-Run techniques lying in their mixing time, known to scale cubically with the number of variables). The modified Hit-and-Run algorithm we have devised, based on a projection method, appears to be indeed optimized from this viewpoint with respect to, e.g., what was done in [13]. We call the resulting method the *Kernel Hit-and-Run* algorithm (KHR). As we will see, results obtained by wBP and KHR are in remarkable agreement among themselves (and with previous studies of the same problem by other, more costly, computational methods).

2. Methodology

2.1. Mathematical Statement of the Problem

Suppose that we are interested in estimating the probability distribution functions (PDFs) $P_i(x)$ for each reaction flux $i = 1, \ldots, N$. By definition, they are given by:

$$P_i(x) = \frac{\text{Vol}(S_i(x))}{\text{Vol}(S)}, \qquad S_i(x) = \{\boldsymbol{x} \in S \text{ s.t. } x^i = x\} \tag{5}$$

where $\text{Vol}(S)$ is the volume of set S. $P_i(x)$ can be mathematically written as an integral over all fluxes, but x_i, of a set of functions enforcing the constraints that define the polytope, namely (1). Denoting by F_μ these functions ($\mu = 1, \ldots, M$), we can write $P_i(x)$ for flux i as:

$$P_i(x) = \frac{1}{Z_i} \int_{D_{\setminus i}} d\boldsymbol{x}_{\setminus i} \prod_{\mu=1}^{M} F_\mu \left(\sum_{\ell=1}^{N} \xi_\ell^\mu x^\ell \right) \tag{6}$$

where $\boldsymbol{x}_{\setminus i} = (x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^N)$ denotes the collective integration variable, $D_{\setminus i} = \times_{\ell(\neq i)}^{N} [m^\ell, M^\ell]$ is the domain of integration (the set-product of the ranges of variability of all fluxes, except flux i) and $Z_i \equiv \int dx P_i(x)$ is a normalisation constant, so that each $P_i(x)$ is properly normalised to one. Each indicator function F_μ should distinguish between metabolites involved only in internal reactions ($\mu \in \mathcal{I}$ for brevity) and metabolites that are exchanged with the surrounding. A convenient parameterisation is given by:

$$F_\mu(y) = \begin{cases} \delta(y)\,, & \mu \in \mathcal{I} \\ \displaystyle\int_{m^\mu}^{M^\mu} \rho_\mu(z) \delta(y - z)\, dz\,, & \mu \notin \mathcal{I} \end{cases} \tag{7}$$

where $\delta(y)$ is a Dirac δ-distribution and ρ_μ is an *a priori* distribution for the exchange rate γ^μ. Under this choice, for intracellular metabolites, F_μ simply enforces the mass-balance constraint (1) with $\gamma^\mu = 0$ in Equation (6) through a δ-function. For exchanged chemical species, the situation is slightly more complex. If the exchange rate takes a fixed value z_0, then $\rho_\mu(z) = \delta(z - z_0)$ and:

$$F_\mu(y) = \delta(y - z_0) \tag{8}$$

corresponding, once inserted in Equation (6), to the mass-balance constraint (1) with $\gamma^\mu = z_0$. If, however, the exchanged rate is known only probabilistically, then $\rho_\mu(z)$ can be a non-trivial distribution and F_μ enforces (1) in (6) by weighing all possible values of γ^μ according to the measure ρ_μ. For instance, if there is no a *priori* information about the exchange rate, then $\rho_\mu(z)$ can be taken to be uniform. Note that when γ^μ is a random variable, one can consider it as another unknown rate, so that one could also be interested in estimating its *a posteriori* distribution $P_\mu(\gamma)$. The problem we want to face is that of computing quantities like Equation (6) for all i's.

2.2. Weighted Belief Propagation

To push mathematically forward expression (6), we need to do some type of approximation for S. Following [11], we assume the bipartite graph that describes the interdependency of reactions and metabolites to be locally tree-like. In such a case, we are supposing there are no (or only very long) cycles connecting the reactions that process a given metabolite μ. (In the following, we shall write $i \in \mu$ to indicate that reaction i processes, either as a substrate or as a product, metabolite μ.) Thus, if we imagine removing metabolite μ from the system, all reactions $i \in \mu$ become (approximately) statistically independent, as they belong to separate branches of the metabolic (tree-like) network, and their joint PDF factorizes. This is explained pictorially in Figure 1a. If we now put metabolite μ back, we see that, at a fixed value, x, of reaction i, the probability $L_{\mu\to i}(x)$ that mass balance holds for metabolite μ can be expressed, in terms of the factorized PDF computed in absence of μ, as (see Figure 1b):

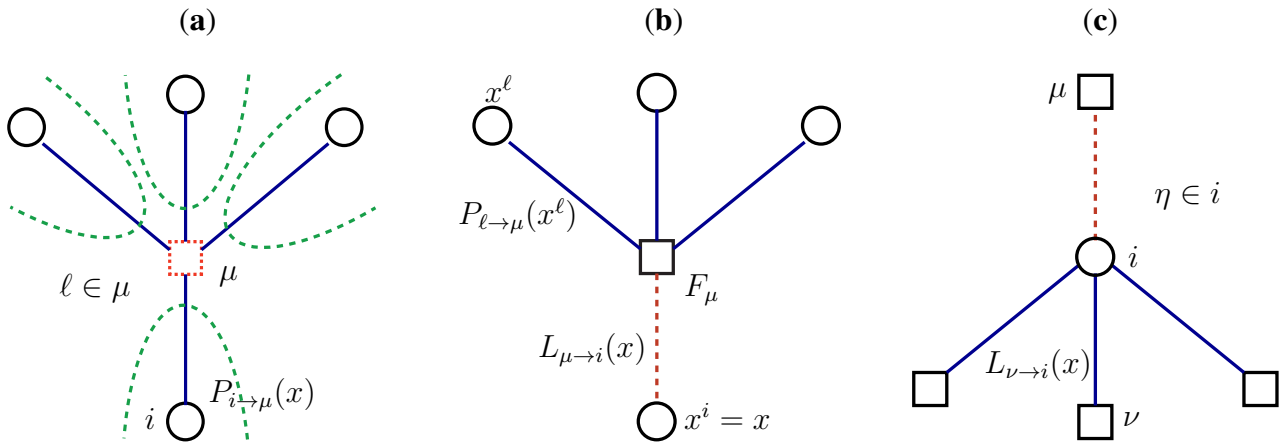
$$L_{\mu\to i}(x) = \frac{1}{L_{\mu\to i}} \int_{D_{\mu\backslash i}} dx_{\mu\backslash i} F_\mu \left(\sum_{\ell\in\mu\backslash i} \xi^\mu_\ell x^\ell + \xi^\mu_i x \right) \prod_{\ell\in\mu\backslash i} P_{\ell\to\mu}(x^\ell) \tag{9}$$

with $L_{\mu\to i}$, a normalisation constant. In this formula, we use Latin labels $(i, \ell, \ldots)$ for reactions and Greek ones $(\mu, \nu, \ldots)$ for metabolites, while the script, $\ell \in \mu\backslash i$, denotes the reactions that process μ except reaction i. Accordingly, we defined the shorthands, $dx_{\mu\backslash i} = \prod_{\ell\in\mu\backslash i} dx^\ell$ and $D_{\mu\backslash i} = \times_{\ell\in\mu\backslash i}[m^\ell, M^\ell]$. The quantity, $P_{\ell\to\mu}(x^\ell)$, is the PDF of flux ℓ taking a value x^ℓ, when metabolite μ is removed. Those PDFs are, in turn, given by the probability, for each reaction i, to satisfy the mass balance conditions for all the metabolites they process, except μ (see Figure 1c), namely:

$$P_{i\to\mu}(x) = \frac{1}{P_{i\to\mu}} \prod_{\nu\in i\backslash\mu} L_{\nu\to i}(x) \tag{10}$$

Here, the set $\nu \in i\backslash\mu$ stands for all metabolites ν, processed by reaction i, except μ, and $P_{i\to\mu}$ is a normalisation constant. Again, the above equations simply state the fact that, on locally tree-like graphs, the contributions to the PDFs coming from each node (reaction or metabolite) nicely factorize.

Figure 1. The method used to derive self-consistency Equations (9) and (10) for the conditional marginals; we only show the nearest neighbours of what one must imagine to be a large tree-like bipartite graph, where circles are reactions and squares, metabolites. (**a**) Metabolite μ and the reactions that process it ($\ell \in \mu$). If we assume removing μ from the system, all reactions connected to it belong to disjoint branches of the metabolic network, highlighted with the dashed lines. As a consequence, their joint probability distribution function (PDF) factorizes in the product of the marginals, $P_{\ell\to\mu}(x)$, of each reaction ℓ. (**b**) When metabolite μ is put back in the graph, the probability, $L_{\mu\to i}(x)$, of satisfying its mass balance condition when fixing the flux of reaction i to x depends on the marginals, $P_{\ell\to\mu}(x)$, of all neighbours, but i, and on the indicator function, F_μ. (**c**) The marginal $P_{i\to\mu}(x)$, which is computed in absence of μ, expresses the probability that i satisfies the mass balance conditions for all the metabolites it processes ($\eta \in i$), except μ. On a tree, each mass balance condition is independent, so that the probability of satisfying all of them is given by the product of the various $L_{\nu\to i}(x)$.



The conceptual step of removing metabolites from the system is the key that allows us to recast the problem in the set of self-consistency Equations (9) and (10), for the conditional probabilities (the reader should keep in mind that this is, however, just a mathematical trick with no biological interpretation whatsoever [10]). Once the fixed point of the system formed by Equations (9) and (10) is known, one can compute the actual PDFs of the fluxes in the metabolic network as:

$$P_i(x) = \frac{1}{P_i} \prod_{\nu \in i} L_{\nu\to i}(x) \tag{11}$$

where P_i is a normalisation constant. Note that Equation (11) also provides the recipe to evaluate the PDFs, $P_\mu(\gamma)$, for the exchange rates, γ^μ, once the conditional marginals, $P_{\ell\to\mu}(x^\ell)$, are known.

As discussed in [9,11], the difficulties of the problem lie not so much in the derivation of Equations (9) and (10), but in devising an efficient method to solve them. In wBP, we tackle the issue by representing the marginals (10) through a collection of $\mathcal{N}$ variables and associated weights, rather than discretizing

them as one would normally do when facing a similar problem. Let us illustrate the idea with a fairly simple example. Consider the integral:

$$\phi_x(x) = \frac{1}{C}\int_0^1 dy\phi_y(y)\int_0^1 dz\phi_z(z)\delta(x+y+z-1) \tag{12}$$

with the extra condition that $x \geq 0$, where ϕ_y, ϕ_z are known densities normalised in the interval $[0,1]$, and C is a normalisation constant. To evaluate (12), we could use Monte Carlo integration and draw $\mathcal{N}$ pairs of random variables $\{(y_i, z_i)\}_{i=1}^{\mathcal{N}}$ according to the distributions, ϕ_y and ϕ_z. Correspondingly, an estimate for $\phi_x(x)$ can be written as:

$$\phi_x(x) = \frac{\sum_{i=1}^{\mathcal{N}}\delta(x+y_i+z_i-1)\Theta(1-y_i-z_i)}{\sum_{i=1}^{\mathcal{N}}\Theta(1-y_i-z_i)} \tag{13}$$

where the term, $\Theta(1-y_i-z_i)$, accounts for draws for which the quantity $x = 1-y_i-z_i$ must be rejected due to the condition $x \geq 0$. The latter condition indeed defines a feasible triangular region in the integration plane, yz (see Figure 2), such that every extraction (y_i, z_i) falling outside this domain must be rejected. This method (basically, naive Monte Carlo integration) is, hence, poised to be rather inefficient. Fortunately, we know precisely where the rejection region is, and we can rewrite Equation (12) as follows:

$$\phi_x(x) = \frac{1}{C}\int_0^1 dy\phi_y(y)\int_0^{1-y} dz\phi_z(z)\delta(x+y+z-1) \tag{14}$$

Now, Equation (14) does not contain a rejection region, but we cannot apply Monte Carlo integration just yet, since $\phi_z(z)$ is not normalised in the interval $[0, 1-y]$. Introducing the corresponding weight:

$$w(y) \equiv \int_0^{1-y} dz\phi_z(z) \tag{15}$$

we can, however, re-cast Equation (14) in the form:

$$\phi_x(x) = \frac{1}{C}\int_0^1 dy\phi_y(y)\int_0^{1-y} dz\phi_z(z|y)w(y)\delta(x+y+z-1) \tag{16}$$

with $\phi_z(z|y) \equiv \phi(z)/w(y)$. The distributions appearing above are now properly normalized. Therefore, to evaluate Equation (16), we can simply draw $\mathcal{N}$ pairs $\{(y_i, z_i)\}_{i=1}^{\mathcal{N}}$ according to $\phi_y(y)$ and $\phi_z(z|y)$, respectively, and estimate $\phi_x(x)$ by:

$$\phi_x(x) = \sum_{i=1}^{\mathcal{N}}\alpha_i\delta(x+y_i+z_i-1)\,, \qquad \text{with } \alpha_i \equiv \frac{w(y_i)}{\sum_{j=1}^{\mathcal{N}}w(y_j)} \tag{17}$$

i.e., by $\mathcal{N}$ pairs of variables and weights $\{(x_i \equiv 1-y_i-z_i,\ \alpha_i)\}_{i=1}^{\mathcal{N}}$.

The key point of this method is that the reweighted density, $\phi_z(z|y)$, has a y-dependent support, such that rejection never occurs. Thus, at a price of computing a weight, $w(y)$, we overcome the whole rejection issue, and the method becomes much more efficient.

The great advantage of using wBP is that, at fixed $\mathcal{N}$, its running time goes as $\mathcal{O}(2Nk)$, where k is the average number of metabolites processed by each reaction. Thus, as opposed to sampling techniques

that have normally super-linear mixing times [14], wBP only scales linearly with the number of reactions (see Figure 3), making it an ideal candidate for application to genome-scale metabolic networks. In the present work, we focus, however, on the relatively small case of the hRBC, so that we are able to compare with sampling methods that yield a uniform exploration of the solution space S [8]. Due to the nature of such methods (see next section), this type of comparison is still not feasible for larger systems. This, and the fact that previous results are available [9], make the metabolic network of the hRBC the ideal testing ground for wBP.

Figure 2. Avoiding rejection. As explained in the text, from the original integration region, $(y, z) \in [0, 1] \times [0, 1]$, only the one below the line $z = 1 - y$ contributes to $\phi_x(x)$. However, in this lower triangle, the density, $\phi_y(y)\phi_z(z)$, is no longer normalised. This is easily dealt with by reweighting the integral.

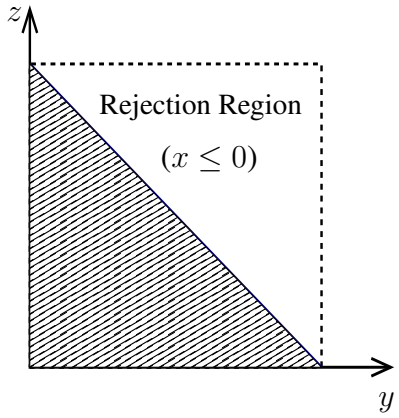
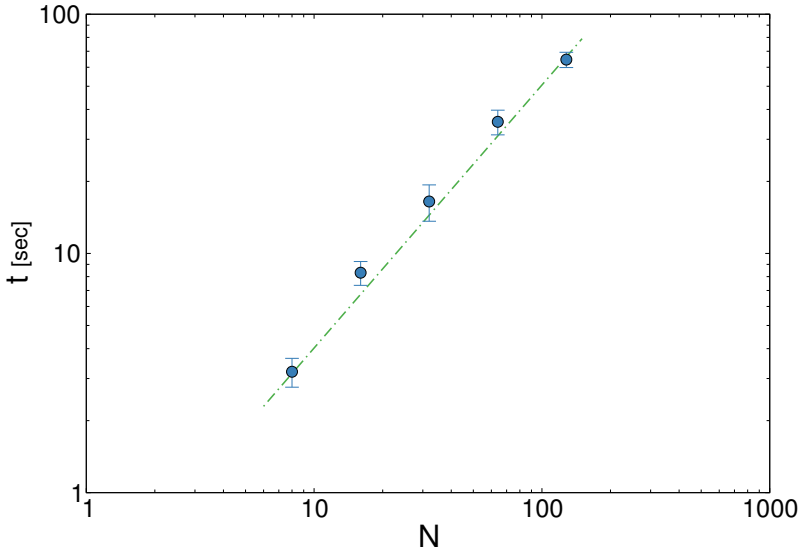


Figure 3. The running time t of the weighted Belief Propagation (wBP) algorithm *vs.* the number of reactions, N. For each value of N, we average here over 10 random synthetic metabolic networks, each having $M = N/2$ metabolites. The algorithm (blue circles) scales linearly with the system size; a linear function, $t \propto N$ (green dashed line), is plotted to guide the eye.



2.3. The Kernel Hit-and-Run (KHR) Algorithm

In order to sample the solution space S and obtain exact PDFs of individual fluxes for the hRBC by a controlled method that guarantees uniformity, we have developed an optimized version of Hit-and-Run Monte Carlo, which we call the *Kernel Hit-and-Run* method. Let us start by re-writing constraints (2) explicitly for metabolites involved in internal reactions and the rest:

$$\sum_{i\in\mu} \xi_i^\mu x^i = 0\,, \qquad \mu \in \mathcal{I} \tag{18}$$

$$m^\mu \leq \sum_{i\in\mu} \xi_i^\mu x^i \leq M^\mu\,, \qquad \mu \notin \mathcal{I} \tag{19}$$

$$m^i \leq x^i \leq M^i\,, \qquad i = 1,\ldots,N \tag{20}$$

We note that the set of $|\mathcal{I}|$ equations in (18) defines the null-space of $\boldsymbol{\xi}$, and geometrically corresponds to a family of hyperplanes passing through the origin $\boldsymbol{x} = \mathbf{0}$. Let us denote the dimension of the null space of $\boldsymbol{\xi}$ as K. Clearly, K would be at least $N - |\mathcal{I}|$ (actually $K = N - |\mathcal{I}|$ when $\boldsymbol{\xi}$ has full row rank, which can always be made to be the case and which we assume from now on). This means, obviously, that, although the number of variables in the system is N, due to the constraints in the model, the actual dimension of the solution space S is only K. As in real metabolic networks most reactions are internal, the dimension K of the null space will be significantly smaller than the original dimension of the problem N. Additionally, it turns out that the way to implement in practice such a dimensional reduction is quite straightforward: suppose that a basis of the null-space has been found, e.g., through Gaussian elimination or singular value decomposition (SVD), and let us denote as $\boldsymbol{y} = (y^1,\ldots,y^K)$ the system of coordinates with respect to such a basis, so that we can write each flux in this basis as $x^i = \sum_{j=1}^K \Phi_j^i y^j$, with $\boldsymbol{\Phi}$ an $N \times K$ matrix related to the change of basis between the original space and the null subspace. Plugging this into Equations (19) and (20) allows us to write:

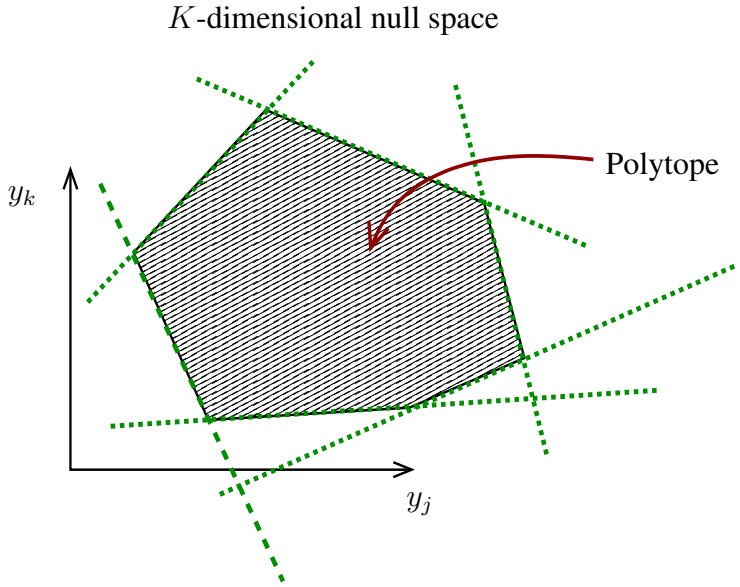
$$\begin{aligned} m^\mu &\leq \sum_{j=1}^K \Psi_j^\mu y^j \leq M^\mu\,, \qquad \mu \notin \mathcal{I} \\ m^i &\leq \sum_{j=1}^K \Phi_j^i y^j \leq M^i\,, \qquad i = 1,\ldots,N \end{aligned} \tag{21}$$

where we have defined the projected stoichiometric matrix, Ψ, with entries $\Psi_j^\mu = \sum_{i=1}^N \xi_i^\mu \Phi_j^i$. The set of Equations (21) defines a K-dimensional polytope in the null space (see Figure 4), which can be sampled *uniformly* by using the Hit-and-Run algorithm [8,15,16]. Finally, to go back to the original space, that of the reaction rates, we simply use the fact that $x^i = \sum_{j=1}^K \Phi_j^i y^j$. The sampling properties of the *Hit-and-Run algorithm* under the uniform measure were indeed mathematically proven [8], and in our case, it is very easy to see that the uniform measure in the K-dimensional null space is preserved under a linear transformation, so that the final sample in the full-dimensional space is also uniform by construction.

While the sampling measure of KHR is well controlled, a word needs to be spent on the algorithmic mixing time. For the standard *Hit-and-Run* algorithm, this scales as the square of the dimensions times the diameter of the polytope, *i.e.*, in practice, cubically with the number of dimensions [8,14]. Yet, as

mentioned before, in our approach, the dimension of the polytope is K, rather than N. This can yield a significant reduction in computation times if K is small compared to N, as will quite generally be the case. For the hRBC, for instance, we pass from $N = 46$ (which can be problematic, e.g., for Monte Carlo rejection [17]), to a much more modest $K = 12$, which is sampled quite fastly by KHR. Note also that no additional constraints need to be introduced to enclose the polytope (as opposed to [13]). This is due to the fact that the $M - |\mathcal{I}|$ metabolites that are exchanged with the environment suffice to bound the polytope in the null space.

Figure 4. A cartoonish representation of the polytope in a K-dimensional null space spanned by y-coordinates. Here, the green dashed lines represent the set of hyperplanes (21) enclosing the polytope.



Finally, it is worth mentioning that the matrix $\mathbf{\Phi}$ can be easily obtained with any standard algebra software or by standard SVD algorithms, and that no matrix inversion is required to compute the projected matrix Ψ nor to convert the obtained sample to the original, full dimensional space. Therefore, to sum up, KHR uniformly samples the solution space S, with a mixing time that scales as $\mathcal{O}(K^3)$.

3. Results and Discussion

We have applied the wBP and KHR algorithms to the study of the metabolic network of the hRBC. As mentioned in Section 2.2, such a choice is dictated by the fact that the hRBC size allows us to apply HR in a modest time and that previous results are available. We have used the same network considered in [9,12,17], which accounts for 34 metabolites, 32 internal reactions and 14 exchange reactions (see Figure 5). We are able to smoothly apply the method by using the effective reaction domains computed in [17] and used also in [9]. Such domains are derived starting from real enzymatic rates [18,19], so that they are physiologically meaningful, and span several orders of magnitude. Note that this would be a major issue if one were to discretize marginals (10), as it would require dealing with binning functions

defined on totally different scales. Thanks to our representation in terms of variables and weights, the fact that the reaction fluxes are of a different order of magnitude does not affect our method at all.

We run the wBP algorithm by representing the marginals, like Equation (10), with sets of variables/weights $\{(x_i,\ \alpha_i)\}_{i=1}^{\mathcal{N}}$ of size $\mathcal{N} = 500$. To solve the fixed point equations (9) and (10), we performed 30 iterations of our method. We started with uniform weights $\{\alpha_\ell\}_{\ell=1}^{\mathcal{N}}$ and, at each iteration $t = 1, \ldots, 30$, and for each fixed value of the variable x_i, we applied wBP $10^3 \times t$ times to evaluate the average weight α_i. Once convergence was reached, we used the variable/weight sets to compute the final 46 PDFs, $P_i(x)$ and $P_\mu(\gamma)$, according to Equation (11). In this last step, we averaged the weight values over 10^5 wBP extractions to achieve a higher accuracy. We report the results in Figures 5 and 6, where we compare our method with KHR; the agreement is excellent. The reaction PDFs obtained with both methods have indeed a very similar domain and shape in most of the cases. Notably, wBP does not perfectly capture the profile of reactions involving currency metabolites, such as ATP, ADP, NADP and NADPH. An explanation of this may lie in the fact that these compounds are highly connected in metabolic networks and likely to be involved in small loops that are not considered by the wBP method.

Figure 5. Results for human red blood cell. Here, we draw a pictorial representation of the system as a directed bipartite graph. Reaction nodes are plotted with their PDFs and metabolite nodes with green squares. Arrows entering (resp. leaving) a reaction stand for a substrate (resp. a product). We have plotted the marginals, $P_i(x)$, for the internal reactions together with the $P_\mu(\gamma)$ for the exchange rates (these are the leaves on the bipartite graph). For the densities, we have used the wBP method (red filled plots) and have compared them with the KHR algorithm (blue solid lines).

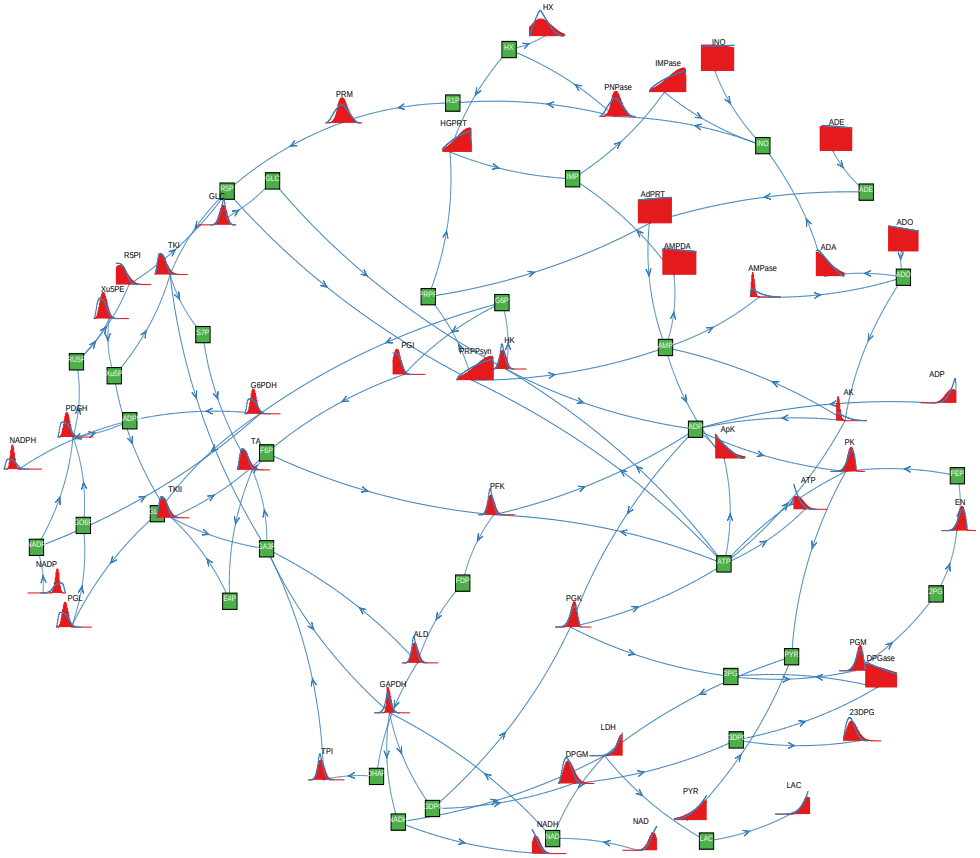
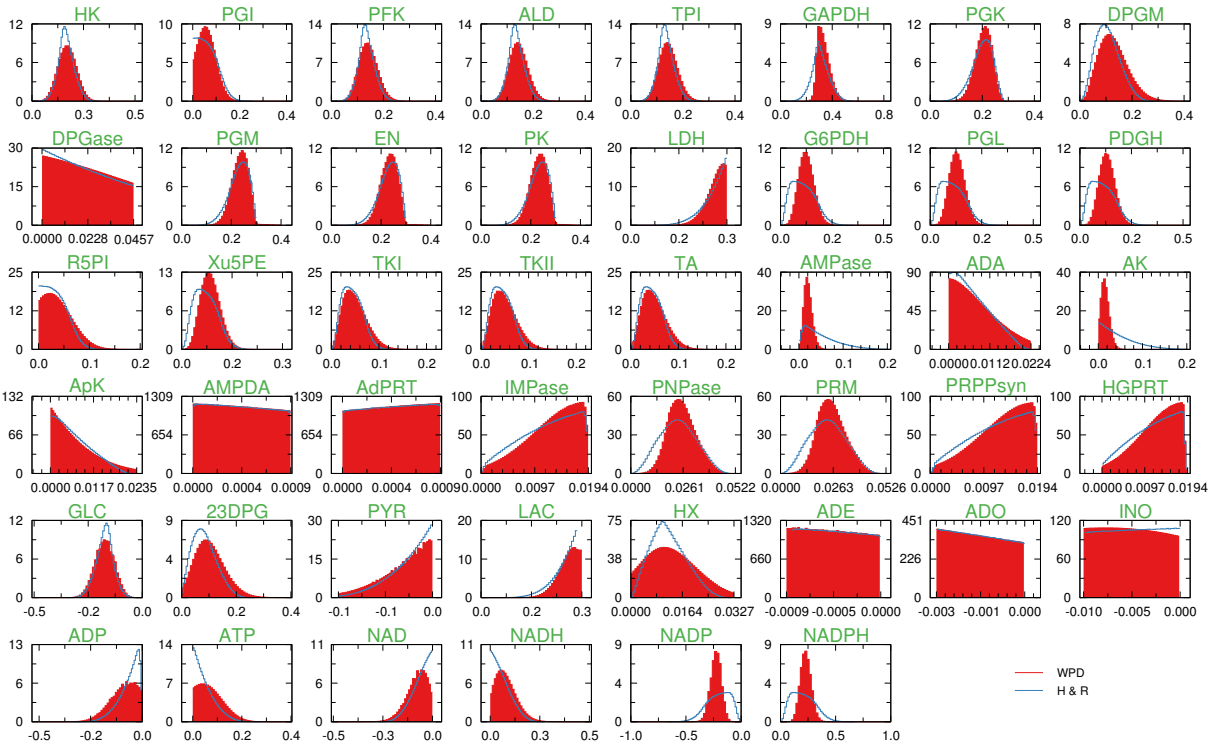


Figure 6. Results for human red blood cell. The probability density functions of the reaction rates; reaction names are the same as [9]. For the densities, we have used the wBP method (red filled plots) and have compared them with the Kernel Hit-and-Run (KHR) algorithm (blue solid lines). Note that the flux ranges span different orders of magnitude, but still, the profiles are very smooth for both weighted population dynamics and the Kernel Hit-and-Run algorithm.



Concerning the results obtained by KHR, we have been particularly careful to make sure we obtain a uniform distribution of the solution space S. As the uniform measure is guaranteed to be a limiting distribution, the difficulty lies, of course, in deciding when such a limit has practically been reached in the simulations. In this regard, we decided to apply three conservative measures to ensure this: first of all, we checked that the initial conditions did not affect the results and, after that, the simulations were averaged over the initial conditions (the initial points were generated using the MinOver algorithm [20]); secondly, the initial time-window was disregarded in the sampling; and finally, in order to avoid correlation effects in the sampling of the PDFs, we only recorded points at large spacing in time. In Figure 6, we report a panel with all the PDFs for the hRBC network to make the comparison between the methods more straightforward.

4. Conclusions

In this work, inspired by techniques employed in the statistical mechanics of disordered systems, we have presented a novel method to estimate distributions of reaction fluxes in constraint-based models of metabolic networks. The wBP methodology has, in our view, clear advantages when compared with alternative approaches. If compared to rejection-based Monte Carlo methods [12] or even to more refined sampling approaches [13], our algorithm has the significant advantage of scaling linearly with the

system size, a feature that makes it particularly suitable to study genome-scale metabolic reconstructions. Comparing wBP with similar MP-based approaches [9], our method turns out to be unaffected by the flux ranges spanning several orders of magnitude nor by the stoichiometric coefficients taking real values. Note that the former property is particularly important for the study of real metabolic systems in physiologic conditions, as enzymatic rates can vary wildly across the network [12,17].

wBP can also be integrated with optimization-based flux balance analysis (FBA), as it easily allows us to evaluate the PDFs of the enzymatic rates close to optimality (assuming a score function is known) by just injecting *a priori* to keep fluxes "close" to their optimal value.

We have also compared the performance of wBP against the KHR method. The latter is a controlled Hit-and-Run Monte Carlo taking place in the null space defined by the set of internal reactions, where a considerable effective dimensional reduction can be achieved. Indeed, starting from the original N-dimensional space of solutions, one can wind up into a space that, in the best of cases, has the same number of dimensions as the number of exchange reactions. Given the considerable gains that this observation provides, we believe that this method may be worth being explored further in its own right.

The validation of the wBP algorithm in the hRBC network, which can be considered a benchmark for the sampling problem for constrained metabolic models, opens the door to future applications of the method to more relevant organisms, such as *Mycoplasma pneumoniae* [21] or *Escherichia coli* [2]. In our opinion, the fact that its algorithmic complexity scales linearly with the number of reactions, combined with the aforementioned ability to deal with more realistic bounds, makes it a highly promising candidate for effective solutions to this challenging task.

Acknowledgments

We thank de Martino, D., Güell, O., Guimerà, R., Sales-Pardo, M., Serrano, M.A. and Sagués, F. for useful discussions and comments. F.F.C. would like to thank funding from MINECO(grant FIS2012-31324) and AGAUR(grant 2012FI_B00422). F.A.M. acknowledges financial support from European Union Grants, PIRG-GA-2010-277166 and PIRG-GA-2010-268342. A.D.M. is supported by the DREAM Seed Project of the Italian Institute of Technology (IIT). The IIT Platform Computation is gratefully acknowledged.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bowman, S.; Churcher, C.; Badcock, K.; Brown, D.; Chillingworth, T.; Connor, R.; Dedman, K.; Devlin, K.; Gentles, S.; Hamlin, N. The nucleotide sequence of Saccharomyces cerevisiae chromosome XIII. *Nature* **1997**, *387*, 90–92.
2. Feist, A.; Henry, C.; Reed, J.; Krummenacker, M.; Joyce, A.; Karp, P.D.; Broadbelt, L.J.; Hatzimanikatis, V.; Palsson, B.Ø. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **2007**, doi:10.1038/msb4100155.

3. Thiele, I.; Palsson, B.O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **2010**, *5*, 93–121.
4. Thiele, I.; Swainston, N.; Fleming, R.M.T.; Hoppe, A.; Sahoo, S.; Aurich, M.K.; Haraldsdottir, H.; Mo, M.L.; Rolfsson, O.; Stobbe, M.D.; *et al*. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **2013**, *31*, 419–425.
5. Kauffman, K.J.; Prakash, P.; Edwards, J.S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **2003**, *14*, 491–496.
6. Orth, J.; Thiele, I.; Palsson, B. What is flux balance analysis? *Nat. Biotechnol.* **2010**, *28*, 245–248.
7. Schellenberger, J.; Palsson, B.Ø. Use of randomized sampling for analysis of metabolic networks. *J. Biol. Chem.* **2009**, *284*, 5457–5461.
8. Lovász, L. Hit-and-run mixes fast. *Math. Progr.* **1999**, *86*, 443–461.
9. Braunstein, A.; Mulet, R.; Pagnani, A. Estimating the size of the solution space of metabolic networks. *BMC Bioinforma.* **2008**, doi:10.1186/1471-2105-9-240.
10. Mezard, M.; Montanari, A. *Information, Physics, and Computation*; Oxford University Press: Oxford, UK, 2009.
11. Font-Clos, F.; Massucci, F.A.; Pérez Castillo, I. A weighted belief-propagation algorithm for estimating volume-related properties of random polytopes. *J. Stat. Mech. Theory Exp.* **2012**, doi:10.1088/1742-5468/2012/11/P11003.
12. Price, N.D.; Schellenberger, J.; Palsson, B.O. Uniform sampling of steady-state flux spaces: Means to design experiments and to interpret enzymopathies. *Biophys. J.* **2004**, *87*, 2172–2186.
13. Almaas, K.; Kovacs, B.; Vicsek, T.; Oltvai, Z.M.; Barabasi, A.L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **2004**, *427*, 839–843.
14. Simonovits, M. How to compute the volume in high dimension? *Math. Progr.* **2003**, *97*, 337–374.
15. Smith, R.L. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Oper. Res.* **1984**, *32*, 1296–1308.
16. Berbee, H.; Boender, C.; Ran, A.R.; Scheffer, C.; Smith, R.; Telgen, J. Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Math. Progr.* **1987**, *37*, 184–207.
17. Wiback, S.J.; Famili, I.; Greenberg, H.J.; Palsson, B.Ø. 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.
18. Wiback, S.J.; Mahadevan, R.; Palsson, B.Ø. Reconstructing metabolic flux vectors from extreme pathways: Defining the -spectrum. *J. Theor. Biol.* **2003**, *224*, 313–324.
19. Wiback, S.J.; Palsson, B.O. Extreme pathway analysis of human red blood cell metabolism. *Biophys. J.* **2002**, *83*, 808–818.
20. Krauth, W.; Mezard, M. Learning algorithms with optimal stability in neural networks. *J. Phys. A* **1987**, doi:10.1088/0305-4470/20/11/013.

21. Wodke, J.A.H.; Puchalka, J.; Lluch-Senar, M.; Marcos, J.; Yus, E.; Godinho, M.; Gutierrez-Gallego, R.; dos Santos, V.A.P.M.; Serrano, L.; Klipp, E.; Maier, T. Dissecting the energy metabolism in Mycoplasma pneumoniae through genome-scale metabolic modeling. *Mol. Syst. Biol.* **2013**, *9*, doi:10.1038/msb.2013.6.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).