BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 15 2012, pages 2037–2044
doi:10.1093/bioinformatics/bts317

Systems biology

Advance Access publication May 29, 2012

An integrated open framework for thermodynamics of reactions that combines accuracy and coverage

Elad Noor[1,†], Arren Bar-Even[1,†], Avi Flamholz[1], Yaniv Lubling[2], Dan Davidi[1] and Ron Milo[1,*]

[1]Department of Plant Sciences and [2]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: The laws of thermodynamics describe a direct, quantitative relationship between metabolite concentrations and reaction directionality. Despite great efforts, thermodynamic data suffer from limited coverage, scattered accessibility and non-standard annotations. We present a framework for unifying thermodynamic data from multiple sources and demonstrate two new techniques for extrapolating the Gibbs energies of unmeasured reactions and conditions.

Results: Both methods account for changes in cellular conditions (pH, ionic strength, etc.) by using linear regression over the ΔG° of pseudoisomers and reactions. The Pseudoisomeric Reactant Contribution method systematically infers compound formation energies using measured K' and pK_a data. The Pseudoisomeric Group Contribution method extends the group contribution method and achieves a high coverage of unmeasured reactions. We define a continuous index that predicts the reversibility of a reaction under a given physiological concentration range. In the characteristic physiological range $3\mu M$–$3mM$, we find that roughly half of the reactions in *Escherichia coli*'s metabolism are reversible. These new tools can increase the accuracy of thermodynamic-based models, especially in non-standard pH and ionic strengths. The reversibility index can help modelers decide which reactions are reversible in physiological conditions.

Availability: Freely available on the web at: http://equilibrator.weizmann.ac.il. Website implemented in Python, MySQL, Apache and Django, with all major browsers supported. The framework is open-source (code.google.com/p/milo-lab), implemented in pure Python and tested mainly on Linux.

Contact: ron.milo@weizmann.ac.il

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 11, 2012; revised on May 21, 2012; accepted on May 23, 2012

1 INTRODUCTION

The study of metabolism has recently regained its central role in diverse areas of cell biology, physiology, medicine and systems biology. The study of metabolic pathways and networks (Haverkorn Van Rijsewijk *et al.*, 2011; Heinrich *et al.*, 1991; Ishii *et al.*, 2007; Kümmel *et al.*, 2010; Oberhardt *et al.*, 2009; Pfeiffer and Schuster, 2005) aims to better understand the complex behaviour of living organisms as well as their manipulation for human needs (Atsumi *et al.*, 2008; Bar-Even *et al.*, 2010; Sinha *et al.*, 2010; Steen *et al.*, 2010; Yim *et al.*, 2011; Zhang *et al.*, 2010). The thermodynamics of biochemical reactions (Alberty, 2003) is of special interest in the analysis and design of metabolic pathways.

The change in the Gibbs free energy (ΔG) characterizes the thermodynamic balance of biochemical reactions and dictates the direction of net flux (the difference between forward and backward fluxes) in a reaction. It is thus useful for the study of a single enzymatic reaction, for analyzing entire metabolic pathways (Vojinović and von Stockar, 2009), and for the large-scale modeling of whole-cell metabolic networks (Henry *et al.*, 2007).

A reaction at equilibrium carries no net flux. At equilibrium in a specific pH, the apparent reaction quotient Q'—the ratio of product to substrate concentrations—is termed the apparent equilibrium constant and denoted by K'. In ideal dilute solutions, the 'transformed Gibbs energy of reaction' is a function of the apparent reaction quotient: $\Delta_r G' = -\mathrm{RT}\ln\left(K'/Q'\right)$. The 'standard' transformed Gibbs energy of reaction ($\Delta_r G'^{\circ}$) is the value of $\Delta_r G'$ at standard conditions, i.e. when all compound concentrations are 1 M. Therefore, $\Delta_r G'^{\circ} = -\mathrm{RT}\ln K'$ (Fig. 1—equation I).

$\Delta_r G'$ determines the direction of net flux in a reaction. A negative $\Delta_r G'$ would correspond to a positive (forward) net flux and vice versa. Cell physiology imposes constraints on metabolite concentrations and consequently on Q'. Reactions for which $\Delta_r G'^{\circ} < 0$ for any physiological Q' can only carry a forward flux and are thus called irreversible reactions. This classification of reactions is especially important in constraint-based modeling that covers whole-cell metabolic networks and depends on the knowledge of reaction directionality for predicting flux distributions, growth rates and other large-scale metabolic phenotypes (Beg *et al.*, 2007; Burgard *et al.*, 2003; Oberhardt *et al.*, 2009). Directionality, annotations typically rely on phenomenological data and arbitrary definitions of reversibility. Recent advances in the field allow incorporating thermodynamic data directly into the model by adding explicit constraints that connect $\Delta_r G$, concentrations and reaction directionality (Fleming *et al.*, 2009; Henry *et al.*, 2007).

The NIST database for Thermodynamics of Enzyme-Catalyzed Reactions (NIST-TECR) is the most comprehensive collection of empirical thermodynamic data (Goldberg *et al.*, 2004, 2007). About 400 reactions have measured equilibrium constants and were

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

successfully mapped to KEGG identifiers (Kanehisa *et al.*, 2008). Still, this represents only about 6% of the ~5500 biochemically relevant reactions in the KEGG database. Moreover, since equilibrium constants in the NIST-TECR database are measured in a variety of pH and ionic strength levels, it is difficult to extrapolate the equilibrium constant for the conditions prevailing in an organism or system of interest. When $\Delta_r G'^{\circ}$ are not covered by the NIST-TECR database, several collections of standard Gibbs free energy of formation ($\Delta_f G'^{\circ}$) can be consulted and used as in Figure 1—equation II. The largest collection, given by Alberty (2003), contains $\Delta_f G'^{\circ}$ values for ~200 compounds.

Unfortunately, many reactions were not measured in the desired conditions (pH, ionic strength, etc.) or have never been experimentally measured at all. If the formation energy of even one of the reactants is unknown, $\Delta_r G'^{\circ}$ of a reaction cannot be derived. In order to bridge the gap between the known formation energies and the unknown ones, a method based on the group contribution assumption for biochemical compounds in aqueous solutions was described by Mavrovouniotis (1990, 1991) and later greatly improved in terms of coverage and accuracy by Jankowski *et al.* (2008). The method is based on the simplifying assumption that each functional group has a characteristic contribution to the overall formation energy of a compound and that these group contributions are independent of each other. Therefore, $\Delta_f G'^{\circ}$ is estimated by summing all contributions from the different groups composing a compound. The contribution of each such group ($\Delta_{gr} G^{\circ}$) is estimated through linear regression, which uses the known $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ and the compounds' partition into groups (Mavrovouniotis, 1990).

The group contribution method is limited in its accuracy. The group independence assumption can be overly simplistic, especially for large compounds or in conjugated systems. In addition, the definitions of the groups rely heavily on heuristics and chemical intuition. Recent improvements of the method included a better choice of groups and the introduction of group interaction corrections (Jankowski *et al.*, 2008). Recently, a promising new approach based on whole-reaction similarities (Rother *et al.*, 2010) has been shown to be more accurate, but currently does not provide as wide a coverage as group contribution methods.

Previous group contribution studies did not consider the effect of pH on the compounds' protonation levels (Alberty, 1998; Jankowski *et al.*, 2008; Mavrovouniotis, 1990, 1991) and assumed all measurements were taken at standard aqueous conditions (e.g. pH 7 and ionic strength of 0.25 M). Each compound is actually an

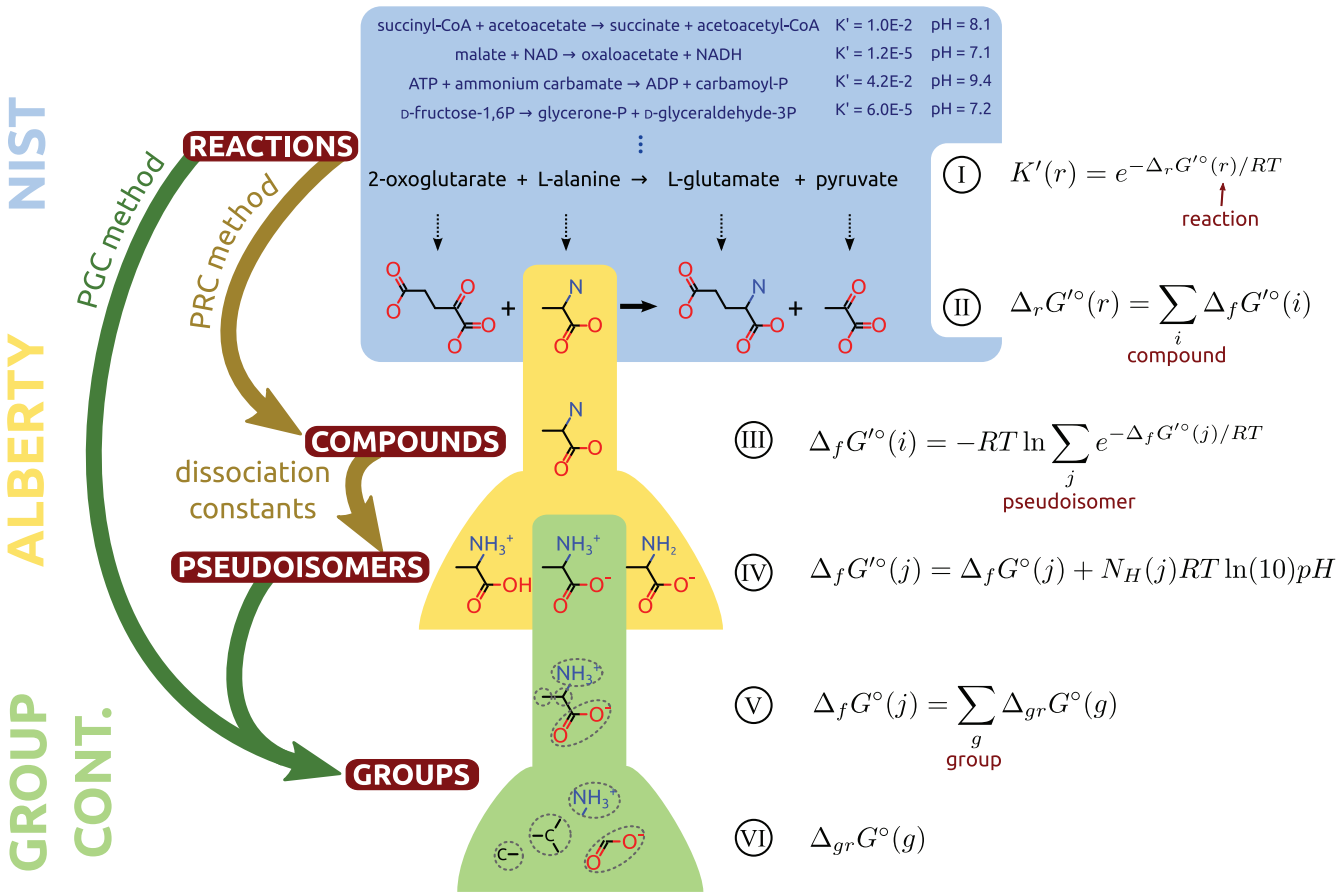

Fig. 1. An overview of the relationships between the layers of thermodynamic data. Reaction equilibrium constants can be converted to $\Delta_r G'^{\circ}$ which is calculated as the stoichiometric sum of reactant formation energies (blue shading, equations I and II). Each reactant exists as several pseudoisomers distributed according to the Boltzmann distribution, and its $\Delta_f G'^{\circ}$ can be found using the Legendre transform (yellow shading, equations III and IV). Using the group contribution assumption, each pseudoisomer can be constructed from its group components (green shading, equations V and VI)

ensemble of 'pseudoisomers' differing in their protonation levels (e.g. ATP^{4-}, $HATP^{3-}$ and H_2ATP^{2-}) and its $\Delta_f G'^{\circ}$ can be found using the Legendre transform (Fig. 1—yellow shading, equations III and IV). As pointed out by Alberty (2003), ignoring the change in the abundance of the different pseudoisomers at changing pH levels can result in errors. This is especially true for the biochemically ubiquitous phosphate groups, which have pK_a values in the range of 6–8 (Lide, 2009; Robitaille *et al.*, 1991), and hence their pseudoisomer distribution changes considerably even close to pH 7.

The concentrations of the individual pseudoisomers are determined by the Boltzmann distribution (Alberty, 2003). $\Delta_f G'^{\circ}(i)$ of a compound i is, therefore, a function of $\Delta_f G'^{\circ}(j)$ of the compound's pseudoisomers (j) (Fig. 1—equation III). Notably, $\Delta_f G'^{\circ}(j)$ of the pseudoisomers and their distribution is modulated by the pH (Fig. 1—equation IV). For example, at a high pH, the unprotonated form of an acidic compound will have lower $\Delta_f G'^{\circ}$ and thus be more abundant (Alberty, 2003). In order to use $\Delta_f G'^{\circ}$ to calculate $\Delta_r G'^{\circ}$, the different pseudoisomer forms assumed by each compound should be considered.

2 APPROACH

In this study, we present methods to accurately estimate $\Delta_r G'^{\circ}$ using two major approaches (Fig. 1). The first approach, named Pseudoisomeric Reactant Contribution (PRC), recovers pseudoisomer formation energies by applying linear regression to the entire set of reactions available in the literature (as recorded in NIST-TECR). Most measurements in the NIST-TECR database are of 'apparent' equilibrium constants, which is the equilibrium apparent reaction quotient of the total compound concentrations—i.e. the sum of all protonation states. Since the pH affects the distribution of species non-linearly, the value of $\Delta_r G'^{\circ}$ cannot be expressed as a linear sum of the reactants' formation energies. PRC applies the inverse Legendre transform to linearize the system. The inverse Legendre transform replaces each ensemble of pseudoisomers with a single representative and changes the observed value of K' accordingly (Alberty, 2002) using the known dissociation constants of the reactants. It is thus possible to find the least-squares solution for the $\Delta_f G^{\circ}$ of the pseudoisomers using linear regression (see Supplementary Material for details). Using PRC, we can retrieve $\Delta_f G'^{\circ}$ values that were previously unknown (Alberty, 2003) and enable the calculation of $\Delta_r G'^{\circ}$ for more reactions.

Compounds that were not measured previously and thus do not appear in the NIST-TECR database require a different approach. We have developed an augmented group contribution method that we call Pseudoisomeric Group Contribution (PGC). Unlike previous approaches, we decompose pseudoisomers, not compounds, into functional groups (Fig. 1—green shading). We then estimate $\Delta_f G^{\circ}$ of the pseudoisomers by summing over the contributions of their groups (equation V) and calculate the $\Delta_f G'^{\circ}$ of compounds (equations III and IV). This method provides higher accuracy and can correctly adjust the $\Delta_f G'^{\circ}$ to the different aqueous conditions described by pH and ionic strength (with pH usually being the most significant). The combination of the two approaches enables a better estimation of $\Delta_f G'^{\circ}$ for a large variety of compounds in a wide range of aqueous conditions.

3 METHODS

3.1 A PRC method systematically derives $\Delta_f G^{\circ}$

The task of estimating the formation energies of compounds given measured reaction equilibrium constants is not straightforward. The difficulty stems from the non-linearity of $\Delta_r G'^{\circ}$ as a function of $\Delta_f G^{\circ}$ and pH (Fig. 1—equations II–IV). However, if $\Delta_r G'^{\circ}$ has been measured in a specific pH and all but one of the reactants' $\Delta_f G'^{\circ}$ values is known, it is straightforward to infer the missing $\Delta_f G'^{\circ}$.

The extensive list of compound formation energies provided in (Alberty, 2003) is the product of a meticulous reconstruction such as described above and based on the data provided by many measurements of K'. As each new $\Delta_f G^{\circ}$ added to this database relies on the previously gathered values, measurement errors for a particular compound are carried on to future calculations.

Alternatively, compound acid dissociation constants (pK_as) can be used to convert the set of equations into a linear system that can be solved computationally. The idea, known as the 'Inverse Legendre Transform' (see Supplementary Material), is based on the fact that the difference between the $\Delta_r G'^{\circ}$ and $\Delta_r G^{\circ}$ of any reaction is a function of the pK_as alone and does not depend on the absolute formation energies of the reactants. The resulting linear system can then be solved using multiple linear regression. This application of the inverse Legendre transform to linearize a reaction system was introduced by Alberty (1991) and Alberty and Goldberg (1992), but to our knowledge was implemented only for a small set of reaction measurements. Here, we perform a global analysis using all the available data in NIST-TECR to achieve the best possible least-squares estimation. We refer to this method as the Pseudoisomeric Reactant Contribution (PRC) method.

Using PRC, we were able to obtain values for 407 formation energies using only the 367 reactions in the NIST-TECR dataset and the pK_a values of the participating compounds (described later in 'Acid dissociation constants'). A detailed analysis of these values in terms of prediction power and accuracy is given in Section 4. Due to linear dependencies between some of the reactions, there are 112 dimensions in the null space of the stoichiometric matrix. For instance, if two compounds always appear together (such as NAD_{ox} and NAD_{red}), the difference between the $\Delta_f G'^{\circ}$ of the pair can be inferred, but the absolute formation energy remains unknown. Similarly, element conservation rules contribute one dimension to the null space for every element which appears in the database, namely C, N, O, S and P. Commonly, as is the case in the current study, the ambiguity in the values of the formation energies is solved by defining some compounds as having $\Delta_f G'^{\circ} = 0$. Alberty's table of ~200 formation energies (Alberty, 2003) contains 18 such reference points.

3.2 A PGC method covers more reactions and conditions

In order to improve the estimations provided by the group contribution method (Mavrovouniotis, 1990) for highly pH-dependent compounds, we introduce a method that incorporates Alberty's transformed formation energies (Alberty, 2003) into the same framework used by Jankowski *et al.* (2008). Previous group contribution implementations used apparent equilibrium constants (K') and formation energies ($\Delta_f G'^{\circ}$) at pH ~7, where each compound is actually an ensemble of protonation species—or pseudoisomers (Fig. 1—equation III). For instance, the total concentration of ATP is divided between the pseudoisomers H_2ATP^{2-}, $HATP^{3-}$ and ATP^{4-} according to the Boltzmann distribution. The formation energy of the ensemble is called the standard transformed Gibbs energy as defined by the International Union of Biochemistry and Molecular Biology (IUBMB) (Alberty *et al.*, 2011). A shift in pH will change the relative abundance of the different pseudoisomers and affect $\Delta_f G'^{\circ}$ and K' non-linearly. Therefore, in order to normalize the effect of pH across different measurements, we use only formation energies of single pseudoisomers as

input for the linear regression step in our group contribution framework. For example, instead of the 'transformed' formation energy of ATP (–2292.5 kJ/mol), we use the formation energy of $HATP^{4-}$, which is –2768.1 kJ/mol (Alberty, 2003). This change requires knowing the standard formation energy of each pseudoisomer separately and the exact distribution of protons and charges in its molecular structure. In the case of reactions, we use the inverse Legendre transform—similarly to the PRC method.

In addition, groups that previously had only one form corresponding to the protonation level most abundant at pH 7 can now have multiple forms representing the different protonations at a wide range of pHs. The full list of groups is given in Supplementary Table S4. For example, the terminal phosphate group $-PO_3^{2-}$ which had only one instantiation in previous implementations (Jankowski *et al.*, 2008) ($\Delta_{gr}G'^{\circ} = -254$ kcal/mol $= -1063$ kJ/mol), has two versions in PGC, namely $-PO_3^{2-}$ and $-HPO_3^{-}$, each with its own contribution ($\Delta_{gr}G^{\circ} = -1024.2$ and -1073.9 kJ/mol, respectively). The terminal phosphate of ATP's major pseudoisomer at pH 7 is semi-protonated and corresponds to the $-HPO_3^{-}$ group. However, when this group appears in the major pseudoisomer of other compounds such as D-glucose-6-phosphate, it is fully deprotonated ($-PO_3^{2-}$). Thus, PGC adjusts the contribution of the two pseudoisomeric groups accordingly. Similarly to previous implementations (Jankowski *et al.*, 2008), the algorithm for determining the group contributions is a least-squares linear regression, except that for the same number of compounds, there are more available values of $\Delta_f G^{\circ}$ and more groups due to the use of pseudoisomers.

We use the NIST-TECR database (Supplementary Table S1), together with a list of formation energies (Alberty, 2003; Dolfing and Janssen, 1994; Thauer *et al.*, 1977) (Supplementary Table S2) and dissociation constants (Lide, 2009) (Supplementary Table S3). As part of this study, we have manually annotated each of the species in the list and determined the protonation level of each of their groups. The task was not trivial for compounds with more than one protonation site, since the order of the pK_a values of the groups determines which will deprotonate first as the pH rises, and these data are not well organized in the literature. For example, the different protonation states of nucleic acids are a particular challenge (Christensen *et al.*, 1970a, b). The details and results of this analysis and the contribution of the pseudoisomeric groups are given in Supplementary Table S4.

Using the PRC method to infer compound formation energies from NIST-TECR, it is possible to obtain predictions for a few hundreds of reactions in the KEGG database (roughly 5%)—not much more than the number of reactions in NIST-TECR itself. However, the PGC method achieves a much higher coverage of ~80% of KEGG reactions, as shown in Table 1. This coverage is similar to previous implementations (Jankowski *et al.*, 2008). Note that although there are >8000 reactions listed in KEGG, we limit our analysis to the subset of reactions whose reactants all have explicit chemical formulas (unlike wildcard formulas or generalized names such as 'donor' and 'acceptor'). Furthermore, we discard reactions that are not chemically or redox balanced.

3.3 Acid dissociation constants

Calculator Plugins were used for structure property prediction and calculation, provided by Marvin 5.5.1, 2011 from ChemAxon (http://www.chemaxon.com). The molecular description of each compound in the KEGG database was converted to a SMILES string, and given as input to the ChemAxon command-line binary which calculated the pK_a values and the charge distribution of the major pseudoisomer at pH 7. All results are listed in the Supplementary Dataset.

3.4 The NIST-TECR database benchmark

In order to evaluate the level of error for each method, we compare the measured $\Delta_r G'^{\circ}$ and the estimated one (adjusted to the same conditions as the measurement). If the chosen estimation method does not cover this reaction, its observed equilibrium constants are not included in this analysis. In order to minimize the temperature related biases, only measurements at temperatures in the range of 298–314 K (i.e. 25–40°C) were used. The root mean squared error (RMSE) of each method is calculated by giving each distinct reaction an equal weight (regardless of how many times it has been measured).

Table 1. Comparing the different methods for estimating Gibbs free energies

Feature	Alberty (2006)	PRC (current)	Jankowski (2008)	PGC (current)
Coverage of reactions (out of 5464 in KEGG)	6%	10%	77%	77%
Coverage of reactions (out of 729 in *E.coli*)[a]	18%	30%	95%	93%
Error (RMS, kJ/mol)	6.8[b]	2.4	9.9[c]	8.5
Pseudoisomers considered	Yes	Yes	No	Yes
Open source	Yes	Yes	No	Yes

[a] The relevant reactions in iAF1260.

[b] Alberty (2003) covers 2073 measurements in NIST-TECR, all other methods cover ~2950.

[c] In Jankowski *et al.* (2008), the reported result is 1.90 kcal/mol (8.0 kJ/mol). In the current article, more reactions from NIST-TECR were used for calculating RMSE since we did not filter observations according to their pH. In addition, we did not include any formation energy data in this evaluation since they are not purely empirical—many are derived from the same data which is already in NIST-TECR, and some values were derived using a group contribution approach and thus cannot be used to evaluate its precision.

The result for a leave-one-out cross-validation test for the PGC method was 11.2 kJ/mol (RMSE). All formation energy and reactions examples were included in this test. If an example could not be omitted because it was linearly independent of the other examples in the training set, its error was not included in the test. If several examples (reactions or compounds) had exactly the same group decomposition they were all treated as a single example. We used our framework to perform a cross-validation test for the standard (i.e. non-pseudoisomeric) group contribution method, which resulted in an RMSE of 16.0 kJ/mol (see Supplementary Material). It was not possible to directly compare our cross-validation results to the other methods mentioned as their implementations are not publicly available.

3.5 The reversibility index

We introduce a quantitative measure for reversibility that takes the mean concentration of metabolites and the number of substrates and products into consideration. Further details regarding the application of the index to genome-scale models and motivation for its use are given in Section 4.

We assume that all reactant concentrations lie within a range located symmetrically (in log-scale) around some characteristic physiological concentration C—here we use 100 μM (Bennett *et al.*, 2009). The variable describing this range is denoted γ: all substrates are assigned a concentration of $(1/\sqrt{\gamma})C$ and all products have a concentration of $\sqrt{\gamma}C$. Therefore, the log-scale width of the range is $\log(\sqrt{\gamma}C) - \log((1/\sqrt{\gamma})C) = \log(\gamma)$. Thus, a value of $\gamma = 100$ corresponds to the range 10 μM–1 mM and $\gamma = 1000$ corresponds to about 3 μM–3 mM.

For a reaction with N_S substrates and N_P products, the apparent reaction quotient $Q'(\gamma) = (\gamma^{1/2}C)^{N_P} \cdot (\gamma^{-1/2}C)^{-N_S} = \gamma^{N_P/2}\gamma^{N_S/2}C^{N_P-N_S} \equiv \gamma^{N/2}Q''$, where we define $N = N_P + N_S$ as the total number of reactants and $Q'' = C^{N_P-N_S}$ as the default reaction quotient. We define the reversibility index as $\hat{\gamma} = \left(K'/Q''\right)^{2/N}$, which is the required concentration range for reversing the reaction, i.e. $Q(\hat{\gamma}) = K'$. The further $\hat{\gamma}$ is from 1 the more irreversible the reaction. The reversibility index of the fructose-bisphosphate aldolase reaction, for example, is 1.04, making it clearly reversible since a change of only 4% in concentrations is required to reverse its direction.

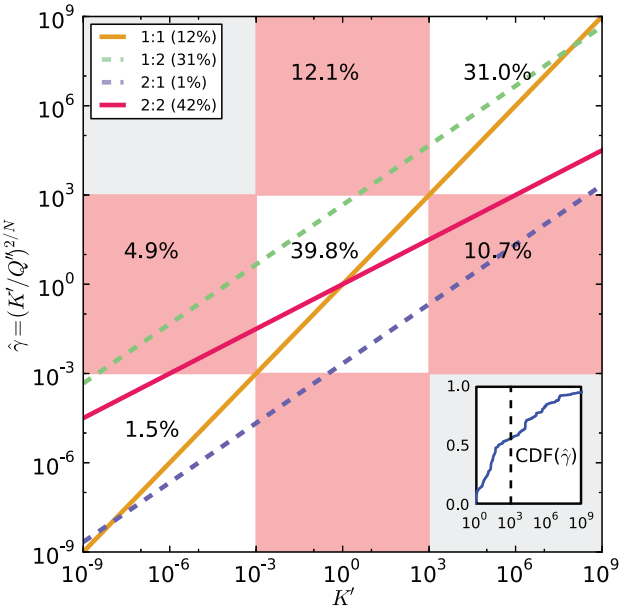

Fig. 2. A comparison between the apparent equilibrium constant K' and the reversibility index $\hat{\gamma}$. Each shaded square represents a different regime where a reaction is considered reversible or irreversible according to K' and $\hat{\gamma}$, and the number indicates the percent of reactions in the metabolism of *E.coli* (in the iAF1260 model) which occupy that regime. White areas represent regimes where both classifications agree, and pink areas are regimes where they disagree. Colored lines show the relationship between K' and $\hat{\gamma}$ for specific reactions stoichiometries and the percent of reactions in the model with that stoichiometry is given in parentheses. The inset on the bottom right shows the fraction of reversible reactions (Y-axis) as a function of the allowed concentration range—determined by the upper bound on the value of $\hat{\gamma}$ (X-axis). This is equivalent to the cumulative distribution function (CDF) of the reversibility index for all reactions in the model. Note that for the CDF, the direction of each reaction is defined such that $\hat{\gamma}$ will be larger than 1

It should be noted that for reactions with one substrate and one product, $Q''=1$ and $N=2$, and therefore $\hat{\gamma}=K'$. However, since 88% of reactions are not 1:1, $\hat{\gamma}$ is usually very different from the value of K', like in the case of fructose-bisphosphate aldolase where $\hat{\gamma}=1.04$ while $K'\sim10^{-4}$. Evaluating both K' and $\hat{\gamma}$ for all the reactions in the *E.coli* iAF1260 model (Feist *et al.*, 2007) reveals that K' plays only a partial role in determining the reversibility index (Fig. 2). In the inset, we show the fraction of reversible reactions as a function of allowed physiological concentration range reflected by $\hat{\gamma}$.

In some cases, it is useful to assign a prescribed concentration for a reactant, rather than allowing it to lie within the range defined by γ. For instance, the concentration of water is fixed and in some organisms, the concentration of P_i is kept almost constant. In such cases, the required concentration can be used to calculate Q'' instead of the default C and the fixed reactant is not included in the value of N.

3.6 Source code

The code for the implementation of the PRC and PGC methods is free and open source and can be found at:
http://milo-lab.googlecode.com/svn/branches/bioinfo-2012/
Our software is written completely in Python, depending only on free software such as Open Babel (openbabel.org) and SciPy (scipy.org).

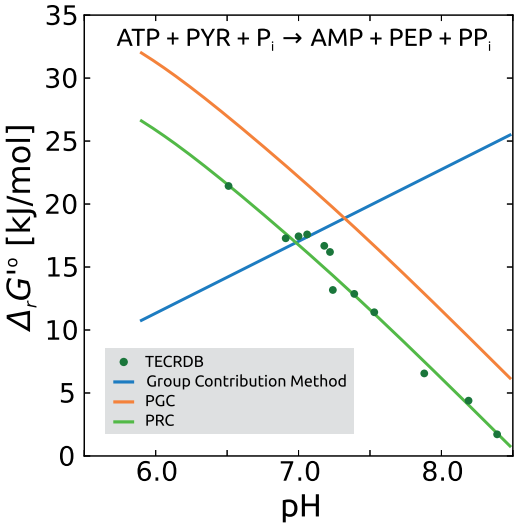

Fig. 3. The response of PPDK to pH, as recorded in the NIST-TECR database. The ionic strength for all estimation methods was set to the default 0.1 M. By taking only the most abundant psuedoisomers at pH 7 and not updating to pK_a values, the charge through this reaction is +1, i.e. one proton is added to the products. Therefore, the predicted response to pH will have a positive slope of RTln(10) (blue line). The intercept of the line was determined so that the value at pH 7 would match the predictions of Jankowski et al., 2008. PGC and PRC use pK_a values derived by ChemAxon (see Section 3) to calculate the response, which results in a better fit to the measured data (orange and green lines). Note that the lines for PGC and PRC are not exactly straight due to the combined effect of several pK_a values. The measured data (dark green) which is listed in NIST-TECR, was taken from (Reeves and Menzies, 1968)

4 RESULTS

4.1 Adjustments for pH using pK_a values can increase accuracy

Many organic compounds are weak acids and bases and, as such, assume multiple protonation levels in typical physiological conditions. Formation energies and, as a result, reaction energies are a function of the distribution of reactant psuedoisomers (Fig. 1—equations II and IV). Since the distribution of protonation levels is a function of the prevailing pH and ionic strength, $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ vary with pH and ionic strength as well. For example, the key reaction in gluconeogenesis and C4 plant photosynthesis, pyruvate + ATP + P_i ⇌ PEP + AMP + PP_i, is catalyzed by PPDK (pyruvate-phosphate dikinase, EC number 2.7.9.1) and was measured at various pH levels ranging from 6.5 to 8.4 (Fig. 3). Many of the reactants have a pK_a in this pH range, and so $\Delta_r G'^{\circ}$ of the PPDK reaction changes significantly with pH. Intracellular pH values are typically between 6 and 7.5 for most organisms (Vojinović and von Stockar, 2009), but the range can be much wider, e.g. in yeast (5.5–7.5 (Imai and Ohno, 1995; Ryan and Ryan, 1972)) or bacteria (4.8–9.0 (Breeuwer *et al.*, 1996)). As shown in Figure 3, the PGC and PRC methods can accurately predict the pH dependence of the reaction. Methods that do not account for pseudoisomers can result in a difference of >20 kJ/mol for the PPDK reaction. If each compound is assigned only a fixed protonation level, based on the most abundant pseudoisomer at pH 7, the reaction looks as follows: $PYR^{-} + ATP^{3-} + P_i^{2-} + H^{+} \rightleftharpoons PEP^{2-} + AMP^{-} + PP_i^{2-}$ which

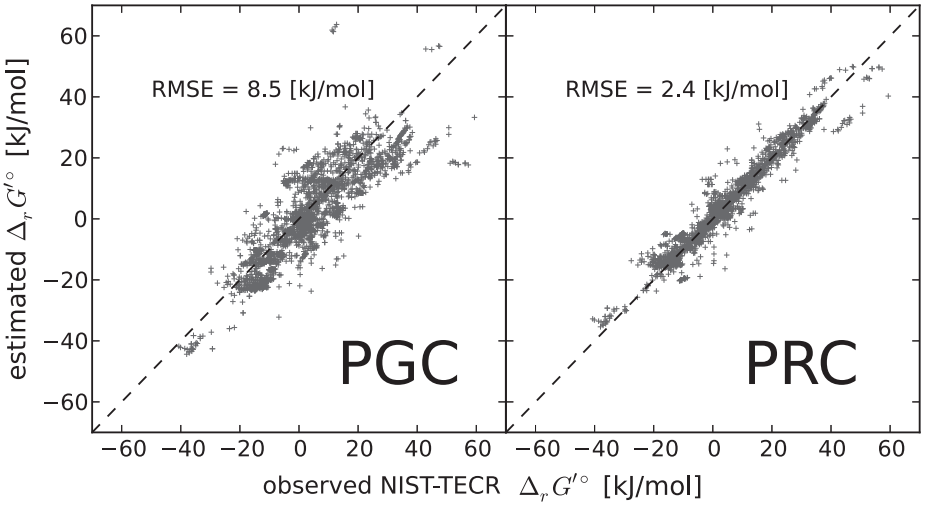

Fig. 4. Comparing the estimated reaction energies using PGC or PRC (Y-axis) to the observed data in NIST-TECR (X-axis). Each X-value corresponds to an observation of K' which is converted into $\Delta G'^\circ$, the reaction Gibbs energy in standard conditions (1 M concentrations) and in the specific pH and ionic strength of each measurement. The Y-value is calculated using PGC or PRC and adjusted to the appropriate pH and ionic strength. The dashed line marks where estimations are equal to observations. The average estimation error per reaction (RMSE) is 8.5 kJ/mol for PGC and 2.4 kJ/mol for PRC. The estimation error can be explained as follows: measurement noise/bias in the value of K', error in the values of pH and ionic strength, pseudoisomers which are unaccounted for, deviations from the theory of thermodynamics in aqueous solutions and violation of the assumption that the contribution of groups to $\Delta G'^\circ$ are completely independent (only for PGC)

implies that the reaction energy increases with the pH, since the effect of one proton is $RT\ln(10)\cdot pH$ (Fig. 3). However, the PGC and PRC methods, which take the dissociation constants of the reactants into account, show that the response has a negative slope, which corresponds well to measured data (Fig. 3).

4.2 Comparing $\Delta_r G^\circ$ estimations using the NIST-TECR database

In order to test the accuracy of the two methods described here, we use the NIST-TECR database as a benchmark. Given a measurement of K' for a given reaction, we compare the predicted reaction energy (using any of the methods described in Section 3), to the observed reaction energy (Fig. 1—equation I). The analysis of the PRC and PGC methods is given in Figure 4 and results in root mean squared error (RMSE) of 2.4 and 8.5 kJ/mol, respectively—equivalent to errors of about a factor of 3 and 30 in the estimation of K'. Note that all of the different methods, those developed in the past as well as those presented here, have used some or all of the data that appear in NIST-TECR for training the group contributions or formation energies. As a result of these dependencies and the fact that the training procedures were not made public for the previously published methods, there is no way of performing an independent cross-validation for them. We found that when using the PRC method, there is a good fit between the regression data and the observed data in NIST-TECR (RMSE of 2.4 kJ/mol) while the methods based on group contributions do not fit the NIST-TECR data as well (RMSE of 8.5 kJ/mol for the PGC method and 9.9 kJ/mol for Jankowski *et al.* (2008)). The reason for this difference in accuracy could be attributed to the fact that group contribution is based on the simplifying assumption of independence between the contributions of groups to the overall $\Delta_f G^\circ$. In addition, there are more free variables in PRC than in PGC (407 compounds versus 99 groups). A summary of the analysis for the four different estimation methods based on the data provided by NIST-TECR is given in Table 1. The goodness of fit is given by the RMSE for each of the methods. These values should be compared with the baseline method which is to use the average $\Delta_r G^\circ$ of each reaction in NIST-TECR across all its measurements. This baseline achieves the maximum accuracy but is limited in coverage to the scope of NIST-TECR. The baseline RMSE is 1.3 kJ/mol and is the average standard deviation of $\Delta_r G^\circ$ per reaction.

4.3 Determining the reversibility of reactions

As an example of the usefulness of having a framework that makes all thermodynamic data available in one location and in an open format, we discuss the issue of reaction directionality that plays a pivotal and often problematic role in many metabolic models (Feist *et al.*, 2007; Oberhardt *et al.*, 2009) and has a crucial effect on their results. A reaction is called irreversible if its net flux flows in the same direction under all allowed physiological conditions. Some reactions are indisputably irreversible, for example the reaction of ribulose-bisphosphate oxygenation (promiscuously catalyzed by the enzyme RuBisCO) which has a $\Delta_r G'^\circ$ of –530 kJ/mol and therefore $K' = 10^{93}$. It is thus tempting to use a rule-of-thumb for determining whether a reaction is irreversible (Tanaka *et al.*, 2003), by applying a threshold on its apparent equilibrium constant—e.g. $K' > 1000$ (or $K' < 0.001$ for irreversible reactions that always flow in the backward direction). This points to the fact that reactant concentrations are bound by physiological considerations and therefore a high-enough K' would require too much of an imbalance in concentrations between substrates and products for reversing a reaction.

Aside from K' itself, the reaction stoichiometry plays a crucial role in determining reversibility as well. For example, the fructose-bisphosphate aldolase reaction—fructose 1,6-*P* $\rightleftharpoons$ dihydroxyacetone-*P* + glyceraldehyde 3-*P*—has a $\Delta_r G'^{\circ}$ of 23 kJ/mol and a K' of 9.23×10^{-5} (Alberty, 2003). This might be considered irreversible (i.e. always flowing in the backward direction), but when the three reactants are at a concentration of 100 μM, a typical intracellular concentration for metabolites (Bennett *et al.*, 2009), the $\Delta_r G'$ is about 0.2 kJ/mol, very close to zero. The reason for the huge difference is due to the fact that this reaction has a different number of products and substrates, and in low-enough concentrations, the effect of the reaction quotient (Q') on $\Delta_r G'$ is significant. Therefore, any reversibility index should properly account for this stoichiometric effect. Furthermore, if a reaction involves many substrates and products, the dynamic range of $\Delta_r G'^{\circ}$ given the physiological metabolite concentrations can be much wider. That is, a reaction with one substrate and one product with $\Delta_r G'^{\circ} = 30$ kJ/mol is much more irreversible than a reaction with three substrates, three products and the same $\Delta_r G'^{\circ}$.

Previous studies (Feist *et al.*, 2007) obtained binary reversibility annotations by checking if $\Delta_r G'$ can attain positive and negative values at different physiological concentrations. Herein, we present a quantitative reversibility index, $\hat{\gamma}$, which accounts for the effects of stoichiometry and physiological concentrations and defines a convenient metric for comparing the reversibility of reactions. The reversibility index is defined using the formula $\hat{\gamma} \equiv \left(K'/Q''\right)^{2/N}$ where N is the total number of reactants (substrates plus products) and Q'' is the reaction quotient at characteristic physiological concentrations. For the purpose of calculating Q'', all metabolites are taken to have a concentration of 100 μM (see Section 3 for details). The value of $\hat{\gamma}$ signifies the fold change that all product and substrate concentrations must undergo in order to reverse a reaction. When considering a range of $\hat{\gamma} < 10^3$, which corresponds to allowing concentrations to span three orders of magnitude around 100 μM ($\sim$3 μM—3mM), about 55% of the reactions are found to be reversible (see the Supplementary Material for a detailed statistical analysis).

5 DISCUSSION

The advances in metabolic modeling (Heinrich *et al.*, 1991; Pfeiffer and Schuster, 2005; Oberhardt *et al.*, 2009; Bar-Even *et al.*, 2010), have created a need for accurate genome-wide values for reaction thermodynamic parameters. As metabolic network models for more organisms and cells emerge, it is increasingly important to have correct predictions for acidic and basic environments and for different ionic strengths. This requirement is most prominent when modeling organisms with multi-compartmented cells, and having to adjust the ΔGs to the conditions in each compartment. Most data that do exist are hard to access (e.g. in out-of-print books) and cover only a limited part of the scope of required compounds and reactions. Moreover, the attempts to use group contribution to expand this coverage are based on a closed implementation and are not extendable for outside users to add new group definitions or training examples.

In this article, we explored two approaches to estimating free energies. We show that the formation energies obtained using PRC achieve a good fit to the observed data (RMSE 2.4 kJ/mol), but do not provide genome-wide coverage of metabolic reactions ($\sim$30% of *E. coli* model). As an alternative, PGC does provide free energy estimates for the majority ($\sim$77%) of known biochemical reactions, but with larger errors (RMSE 8.5 kJ/mol). Therefore, a combination of the two methods might be beneficial, where PRC values are used whenever possible, and PGC is used to fill the gaps. There is, however, a potential problem with this approach as combinations of reactions can become inconsistent (e.g. stoichiometrically balanced cycles can have a non-zero ΔG). The challenge of combining $\Delta_r G^{\circ}$ estimations from different sources and estimation approaches in a unified and consistent manner requires an update to the methods described above (manuscript under preparation).

We hope that, with time, new measurements of reaction equilibrium constants will be published and used to improve the accuracy and coverage of both these methods (PGC and PRC). We thus join the plea of (Jankowski *et al.*, 2008), who published a table of compounds that contain groups with yet unknown contributions.

6 CONCLUSION

We believe that the tools and data that enable thermodynamic analysis of biochemical systems should be easily and freely accessible. In addition to supplying $\Delta_r G'^{\circ}$ predictions as a table for use in metabolic models, we created a website (http://equilibrator.weizmann.ac.il) with a simple user interface that enables anyone to find reactions by chemical formula or enzyme name (Flamholz *et al.*, 2011). The user can adapt the concentration of reactants and the conditions of the reaction.

The thermodynamics of biochemical reactions has a key role to play in our understanding and manipulation of metabolic pathways. An integrated and open framework that combines accuracy and coverage will facilitate the wide use of this fundamental constraint by physics on the biochemistry of life.

ACKNOWLEDGEMENTS

The authors thank the following people for critically reading this article, their helpful comments and encouraging remarks: Avi Mayo, Tomer Shlomi, Naama Tepper, Robert N. Goldberg, Wolfram Liebermeister, Igor Libourel and Hulda S. Haraldsdóttir.

Funding: The European Research Council [260392 - SYMPAC]. The Larson Charitable Foundation, Estate of David Arthur Barton, Anthony Stalbow Charitable Trust & Stella Gelerman, Canada. E.N. is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. A.B.-E. is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. Dr. Milo is the incumbent of the Anna and Maurice Boukstein Career Development Chair in Perpetuity.

Conflict of Interest: none declared.

REFERENCES

Alberty,R.A. (1991) Equilibrium compositions of solutions of biochemical species and heats of biochemical reactions Biochemistry: Alberty. *Proc. Nat. Acad. Sci. USA*, **88**, 3268–3271.

Alberty,R.A. (1998) Calculation of standard transformed formation properties of biochemical reactants and standard apparent reduction potentials of half reactions. *Arch. Biochem. Biophys.*, **358**, 25–39.

Alberty,R.A. (2002) Inverse legendre transform in biochemical thermodynamics: illustrated with the last five reactions of Glycolysis. *J. Phys. Chem. B*, **106**, 6594–6599.

Alberty,R.A. (2003) *Thermodynamics of Biochemical Reactions*. Wiley-Interscience, Hoboken. NJ.

Alberty,R.A. and Goldberg,R.N. (1992) Standard thermodynamic formation properties for the adenosine 5'-triphosphate series. *Biochemistry*, **31**, 10610–10615.

Alberty,R.A. *et al*. (2011) Recommendations for terminology and databases for biochemical thermodynamics. *Biophys. Chem.*, **155**, 89–103.

Atsumi,S. *et al*. (2008) Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, **451**, 86–89.

Bar-Even,A. *et al*. (2010) Design and analysis of synthetic carbon fixation pathways. *Proc. Nat. Acad. Sci. USA*, **107**, 8889–8894.

Beg,Q. K. *et al*. (2007) Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proc. Nat. Acad. Sci. USA*, **104**, 12663–12668.

Bennett,B.D. *et al*. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem Biol.*, **5**, 593–599.

Breeuwer,P. *et al*. (1996) A novel method for continuous determination of the intracellular pH in bacteria with the internally conjugated fluorescent Probe 5 (and 6-)-carboxyfluorescein succinimidyl ester. *Appl. Environ. Microbiol.*, **62**, 178–183.

Burgard,A.P. *et al*. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.*, **84**, 647–657.

Christensen,J.J. *et al*. (1970a) Thermodynamic pK, delta H0, delta S0, and delta Cp0 values for proton dissociation from several purines and their nucleosides in aqueous solution. *Biochemistry*, **9**, 4907–4913.

Christensen,J.J. *et al*. (1970b) Thermodynamics of proton dissociation in dilute aqueous solution. Part XIV. pK, delta H0, and delta S0 values for proton dissociation from several pyrimidines and their nucleosides at 10 and 40 degrees C. *J. Chem. Soc.*, 1643–1646.

Dolfing,J. and Janssen,D.B. (1994) Estimates of Gibbs free energies of formation of chlorinated aliphatic compounds. *Biodegradation*, **5**, 21–28.

Feist,A.M. *et al*. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mole. Syst. Biol.*, **3**, 121.

Flamholz,A. *et al*. (2011) eQuilibrator–the biochemical thermodynamics calculator. *Nucleic Acids Res.*, 1–6.

Fleming,R.M.T. *et al*. (2009) Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to *Escherichia coli*. *Biophys. Chem.*, **145**, 47–56.

Goldberg,R.N. *et al*. (2004) Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics*, **20**, 2874.

Goldberg,R.N. (2007) Thermodynamics of enzyme-catalyzed reactions: Part 7 – 2007 update. *J. Phys. Chem. Refer. Data*, **36**, 1347.

Haverkorn Van Rijsewijk,B.R.B. *et al*. (2011) Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. *Mole. Syst. Biol.*, **7**, 1–12.

Heinrich,R. *et al*. (1991) Mathematical analysis of enzymic reaction systems using optimization principles. *Federation Eur. Biochemi. Soc. J.*, **201**, 1–21.

Henry,C.S. *et al*. (2007) Thermodynamics-based metabolic flux analysis. *Biophys. J.*, **92**, 1792–1805.

Imai,T. and Ohno,T. (1995) The relationship between viability and intracellular pH in the yeast *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **61**, 3604–3608.

Ishii,N. *et al*. (2007) Multiple high-throughput analyses monitor the response of *E.coli* to perturbations. *Science*, **316**, 593–597.

Jankowski,M.D. *et al*. (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.*, **95**, 1487–1499.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**(Database issue), D480–D484.

Kümmel,A. *et al*. (2010) Differential glucose repression in common yeast strains in response to HXK2 deletion. *FEMS Yeast Res.*, **10**, 322–332.

Lide,D. (2009) *CRC Handbook of Chemistry and Physics: A Ready-reference Book of Chemical and Physical Data.* 90th ed., CRC Press.

Mavrovouniotis,M.L. (1990) Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. and Bioeng.*, **36**, 1070–1082.

Mavrovouniotis,M.L. (1991). Estimation of standard Gibbs energy changes of biotransformations. *J. Biolog. Chem.*, **266**, 14440–14445.

Oberhardt,M.A. *et al*. (2009) Applications of genome-scale metabolic reconstructions. *Mole. Syst. Biol.*, **5**, 320.

Pfeiffer,T. and Schuster,S. (2005) Game-theoretical approaches to studying the evolution of biochemical systems. *Trends Bioch. Sci.*, **30**, 20–25.

Reeves,R. and Menzies,R. (1968) The pyruvate-phosphate dikinase reaction. *J. Biol. Chem.*, **243**, 5486–5491.

Robitaille,P.-M.L. *et al*. (1991) An analysis of the pH-dependent chemical-shift phosphorus-containing metabolites behavior of phosphorus-containing metabolites. *J. Mag. Reson.*, **84**, 73–84.

Rother,K. *et al*. (2010) IGERS: inferring Gibbs energy changes of biochemical reactions from reaction similarities. *Biophysical J.*, **98**, 2478–2486.

Ryan,J.P. and Ryan,H. (1972) The role of intracellular pH in the regulation of cation exchanges in yeast. *Bioch. J.*, **128**, 139–146.

Sinha,J. *et al*. (2010) Reprogramming bacteria to seek and destroy an herbicide. *Nat. Chem. Biol.*, **6**, 464–470.

Steen,E.J. *et al*. (2010) Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature*, **463**, 559–562.

Tanaka,M. *et al*. (2003). Extraction of a thermodynamic property for biochemical reactions in the metabolic pathway. *Genome Inform.*, **371**, 370–371.

Thauer,R.K. *et al*. (1977) Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.*, **41**, 100–180.

Vojinović,V. and von Stockar,U. (2009) Influence of uncertainties in pH, pMg, activity coefficients, metabolite concentrations, and other factors on the analysis of the thermodynamic feasibility of metabolic pathways. *Biotechnol. Bioeng.*, **103**, 780–795.

Yim,H. *et al*. (2011). Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat. Chem. Biol.*, **7**, 445–452.

Zhang,K. *et al*. (2010) Expanding metabolism for total biosynthesis of the nonnatural amino acid L-homoalanine. *Proc. Nat. Acad. Sci. USA*, **107**, 6234–6239.