An algorithm for removing stoichiometric discrepancies in biochemical reaction databases†

Rahul Shaw[1], Suma Debsarma[2] and Sudip Kundu[1,*]

[1]Department of Biophysics, Molecular Biology and Bioinformatics, and [2]Department of Applied Mathematics, University of Calcutta, Kolkata 700 009, India

It is observed that a significant number of reactions are stoichiometrically unbalanced in the existing databases (e.g. MetaCyc), even in curated databases like RiceCyc. To obtain a valid model, reactions should be mass balanced. Here, we propose a systematic algorithm based on generalized inverse matrix – to search all the reactions where mass is not conserved and to calculate the appropriate balancing coefficients of reactants and products, wherever possible (feasible reactions). Using this algorithm we have identified the set of reactions which cannot be stoichiometrically balanced with the present set of substrates and products (infeasible reactions). We have also suggested names of filler metabolites to turn infeasible reactions into feasible reactions.

Keywords: Filler metabolites, metabolic model, reaction matrix, stoichiometry.

STUDIES on metabolism are a key aspect to understand the cellular physiology of a species. Researchers are trying to reconstruct and analyse small pathways[1,2] as well as genomescale metabolic models[3–8]. Metabolic reconstruction of human pathogen incorporating thermodynamic information and potential multi-drug therapy approaches was identified from a model[9]; certainly a model requires correct stoichiometry of reactions. Flux balance analysis (FBA) or a more advance enhanced flux balance analysis (eFBA) for faithful modelling of metabolic networks also requires correct stoichiometric information of reactions[10]. While the reconstruction of several small, separate pathways is possible manually using the available literature and/or experimental data, the genomescale metabolic model demands automatic procedures with manual intervention at proper stages of model building[11]. Further, the reconstructed model is analysed by different methods which can be grossly divided into two groups – kinetic modelling and structural modelling. While structural modelling requires a list of reactions (with conserved mass during interconversion and correct directionality, etc.), from which the shortest path in the reaction network, cluster and sub-cluster formation and analysis of hubs can reveal useful biological information of a species[12], and kinetic modelling needs some extra information (e.g. kinetic data). Insufficient data for kinetic modelling is one of the main reasons why most of the researchers use structural modelling for genomescale metabolic models.

Reconstructions of metabolic models start with using public databases (e.g. MetaCyc[13], KEGG[14]) and it is already reported that there exist different types of discrepancies[15]. The discrepancies include violation of mass conservation in individual reactions, missing reactions (which generate discontinuity in internal metabolic network), absence of reversible/irreversible criteria, etc. Poolman *et al.*[15] reported that 453 and 66 reactions in KEGG and BioCyc (http://biocyc.org/) respectively, violate the mass conservation principle. We have also found that there are 252 reactions in MetaCyc which are unbalanced in mass. Researchers are trying to sort or remove the inconsistencies during their model building. Efforts have also been made to address these problems and also to make a reaction repository with their substrates and products in correct form. For example, the database Rhea[16] provides manually annotated information about a large set of reactions with correct structure, formula and charge of metabolites. A large number of chemical structures available in KEGG have been corrected with respect to constitution and stoichiometry and they are available in BioMeta[17]. Also to obtain a large metabolic network from the KEGG data, one can use METANNOGEN[18]. However, neither Rhea or BioMeta contains all the available reactions in different databases nor do they provide an automated procedure based on a systematic algorithm to find mass-balanced reactions, to identify and provide correct stoichiometry of an unbalanced reaction. Félix and Valiente[19] have also identified the inconsistencies in the information about chemical compounds and biochemical reactions present in the biochemical database KEGG. They have further proposed an algorithm based on the knowledge of atomic rearrangement pattern in biochemical reactions to reduce the atom–atom mapping problem in biochemical pathways. Gevorgyan *et al.*[20] have proposed algorithms to validate a metabolic model which should satisfy two fundamental physical constraints: positivity of molecular masses of metabolites and mass conservation in all biochemical reactions. In spite of all these efforts, there is need for a systematic algorithm/software tool (publicly available which can be useful to get the biochemical reactions with correct stoichiometry of the participated metabolites) to help in the genome-scale metabolic modelling process.

There are several methods for balancing chemical equations. Some of the methods are based on trial and error, exploiting the rules of determining the change in oxidation number[21], or by balancing half reactions[22], or by balancing the gain or loss of electrons. But such a trial and error process is not always an easy task and may fail

†Supplementary data online at http://www.bioinformatics.org/mrat/sup/supplementary.zip

*For correspondence. (e-mail: skbmbg@caluniv.ac.in)

in searching all possible solutions when reactions are not unique. Moreover, they are not suitable for computer implementation and thus deterministic methods are always preferred. Das[23] used a method of partial equations. Sen *et al.*[24] described an integer programming approach to balance chemical equations. Risteski[25] provided a method based on the solution of a homogeneous matrix equation with a non-singular extended matrix. Risteski used the singular matrix method for balancing chemical equations[26] and further introduced a new complex vector method for balancing chemical equations[27]. A detailed historical background of different attempts for the problem of balancing chemical equations is given in Risteski[25–30]. Risteski[28] provided a method for balancing chemical equations using the computation of Von Neumann *g*-inverse. The method is also applicable for chemical equations involving reactants and products possessing atoms with fractional oxidation numbers. But the method is restricted to the square shaped reaction matrices (see later in the text) only. However, a large number of reactions with rectangular-shaped reaction matrices exist in reality. A rectangular-shaped reaction matrix can be converted into a square-shaped matrix by padding zero matrix of appropriate size, but then the size of the reaction matrix increases requiring greater memory and resulting in less speed. In addition, there exist a large number of reactions which cannot be mass-balanced with the present set of metabolites in those reactions. For example, the reaction 'RXNQT-4351'

$$C_{10}H_8N_2O_4S_2 \rightarrow C_{10}H_8N_2, \tag{1}$$

in MetaCyc is infeasible since one cannot balance the mass of the reaction with the present set of metabolites. We adopted the concept of Félix and Valiente[19] in our algorithm, which suggests a list of filler metabolites from known chemical space to make an infeasible reaction into a feasible one. One of the suggestions for the above infeasible reaction is

$$C_{10}H_8N_2O_4S_2 = C_{10}H_8N_2 + 2SO_2. \tag{2}$$

This suggestion is valuable for model building as it provides a testable hypothesis for the experimental biochemists.

In the present work we have developed a computer algorithm that balances chemical equations successfully and works automatically. The algorithm uses the concept of generalized inverse (*g*-inverse) matrix method. There are different types of *g*-inverse in the literature. We have chosen Moore–Penrose inverse because of its uniqueness and for non-singular matrices it becomes identical with its conventional inverse. For computing Moore–Penrose inverse there are several algorithms: iterative algorithm, Greville algorithm, Courrieu's method, etc. Singular value decomposition (SVD) algorithm[31] is efficient for computing Moore–Penrose inverse. The algorithm is applicable to any reaction matrix, whether it is square-shaped or rectangular-shaped. The algorithm works for reactants or products possessing atoms with integral or fractional oxidation numbers (see Supplementary data). Materials for this work are the two data files namely, reactions.dat (list of reactions and name of participating metabolites) and compounds.dat (empirical formula of metabolites) of MetaCyc. This is a database of non-redundant, experimentally elucidated metabolic pathways. It contains extensive information on metabolic pathways and enzymes for many organisms.

Now we present some definitions and important results of matrix algebra which are relevant to our computer code.

Let $\boldsymbol{A} \in \boldsymbol{R}^{m\times n}$, $\boldsymbol{R}^{m\times n}$ being the set of all $m \times n$ real matrices.

Definition 1. *Rank of a matrix* $\boldsymbol{A}$ *denoted by r is the number of linearly independent rows or columns of* $\boldsymbol{A}$. *Clearly*

$$r \le \min(m, n). \tag{3}$$

Theorem 1. *If rank of a matrix* $\boldsymbol{A}$ *is r, then there is at least one non-zero minor of order r and all minors of order greater than r, if any, are zeros.*

Definition 2. *For a matrix* $\boldsymbol{A}$, *its generalized inverse (g-inverse) is a matrix* $\boldsymbol{X} \in \boldsymbol{R}^{n\times m}$ *satisfying the following matrix equation:*

$$\boldsymbol{AXA} = \boldsymbol{A}. \tag{4}$$

Matrix $\boldsymbol{X}$ *satisfying eq.* (4) *will be denoted here by* $\boldsymbol{A}^{-}$.

Remark 1. Matrix $\boldsymbol{X}$ satisfying eq. (4) may not be unique. Various types of *g*-inverses are defined[32,33] satisfying in addition to eq. (4), certain other matrix equation(s).

Definition 3. *Moore–Penrose inverse of* $\boldsymbol{A}$ *is a matrix* $\boldsymbol{X} \in \boldsymbol{R}^{n\times m}$ *satisfying the following matrix equations*

$$\text{(i)} \quad \boldsymbol{AXA} = \boldsymbol{A}, \tag{5}$$

$$\text{(ii)} \quad \boldsymbol{XAX} = \boldsymbol{X}, \tag{6}$$

$$\text{(iii)} \quad (\boldsymbol{AX})^{\mathrm{T}} = \boldsymbol{AX}, \tag{7}$$

$$\text{(iv)} \quad (\boldsymbol{XA})^{\mathrm{T}} = \boldsymbol{XA}. \tag{8}$$

Here the superscript T denotes transpose matrix. Moore–Penrose inverse of $\boldsymbol{A}$, *denoted by* $\boldsymbol{A}^{+}$, *satisfying eqs* (5)–(8) *is unique. Moreover, if the matrix* $\boldsymbol{A}$ *is a non-singular*

square matrix, then $\boldsymbol{A}^+$ *becomes identical with its conventional inverse* $\boldsymbol{A}^{-1}$.

Result 1. A general solution of the homogeneous system of equations

$$\boldsymbol{A}x = 0, \tag{9}$$

is given by

$$x = (\boldsymbol{I} - \boldsymbol{A}^{-}\boldsymbol{A})c, \tag{10}$$

where c is an arbitrary vector and $\boldsymbol{A}^{-}$ is any kind of g-inverse. In particular, it can be $\boldsymbol{A}^+$ also.

Result 2. $(\boldsymbol{I} - \boldsymbol{A}^{-}\boldsymbol{A})$ *is an* $n \times n$ *matrix and its rank is* $k = n - r$.

Proposition 1. *Any chemical equation can be presented in terms of the following matrix equation*

$$\boldsymbol{A}x = 0, \tag{11}$$

where $\boldsymbol{A} = [a_{ij}]_{m\times n}$ *is known as the reaction matrix and* $x = [x_j]_{n\times 1}$ *is a column vector*.

Row index m of $\boldsymbol{A}$ counts for the number of distinct atoms or elements, while the column index n of $\boldsymbol{A}$ counts for the total number of reactants and products in a reaction. The (i, j)th entry a_{ij} of $\boldsymbol{A}$ represents the number of atoms of type i present in the jth reactant or product. a_{ij} is positive or negative according to whether it corresponds to a reactant or a product respectively. a_{ij}s are integers in most of the cases, but a_{ij}s can also be rational numbers corresponding to fractional oxidation numbers. The task of forming a reaction matrix for a given chemical equation is illustrated later in the text. In eq. (11), x is an unknown vector with integral or rational components balancing the chemical equation.

Proposition 2. *The reaction matrix is not necessarily square-shaped. Even a square-shaped reaction matrix may be non-singular. So conventional inverse of* $\boldsymbol{A}$ *does not exist. For solving the system* $\boldsymbol{A}x = 0$, *we adopt the technique of g-inverse.*

Result 3.
(i) When $r = n$, then $k = 0$ and the solution of eq. (11) becomes $x = 0$. In this case the reaction is infeasible.
(ii) When $r = n - 1$, then $k = 1$ and whatever be the value of the arbitrary vector c, $(n - 1)$ components of x are multiplies of the remaining component. In this case the reaction is unique.
(iii) When $r \le n - 2$, then $k \ge 2$ and the number of linearly independent solutions as given by eq. (11) is k. This means that the reaction is not unique within the relative proportions of the reactants and products.

If there is any chemical element E_i present only in the reactant(s) or only in the product(s), then the reaction is said to be erroneous and hence infeasible. In this case, the reaction matrix $\boldsymbol{A}$ contains a row with only non-negative entries or only non-positive entries. Thus the problem of balancing a chemical equation is reduced to the problem of solving the homogeneous system of equations, i.e. eq. (11). To solve eq. (11) we need to compute any kind of g-inverse of $\boldsymbol{A}$. Here we have used Moore–Penrose inverse $\boldsymbol{A}^+$ for solving eq. (11) so that its solution is given by

$$x = (\boldsymbol{I} - \boldsymbol{A}^{+}\boldsymbol{A})c, \tag{12}$$

where c is an arbitrary vector. In the literature there are several algorithms for computing $\boldsymbol{A}^+$. Here we have chosen the SVD algorithm.

Definition 4. *The singular values of a matrix* $\boldsymbol{A}$ *are the non-negative square roots of the eigenvalues of* $\boldsymbol{A}^T\boldsymbol{A}$.

Definition 5. *The SVD of a matrix* $\boldsymbol{A}$ *is a representation of* $\boldsymbol{A}$ *in the following form*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{13}$$

where $\boldsymbol{U}$ *is* $m \times m$ and $\boldsymbol{V}$ *is* $n \times n$ *orthogonal matrices, so that* $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$ *and* $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{I}$. $\boldsymbol{D}$ *is an* $m \times n$ *diagonal matrix having non-negative diagonal entries in descending order*.

Result 4. *It can be proved that the diagonal elements of* $\boldsymbol{D}$ *are the singular values of* $\boldsymbol{A}$, *the columns of* $\boldsymbol{U}$ *are orthonormal eigenvectors of* $\boldsymbol{A}\boldsymbol{A}^T$ *and the columns of* $\boldsymbol{V}$ *are orthonormal eigenvectors of* $\boldsymbol{A}^T\boldsymbol{A}$.

Result 5. Using SVD given in eq. (13) of the matrix $\boldsymbol{A}$, its Moore–Penrose inverse is calculated as follows

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^+U^T. \tag{14}$$

Here $\boldsymbol{D}^+$ is the Moore–Penrose inverse of $\boldsymbol{D}$. $\boldsymbol{D}^+$ is then the $n \times m$ diagonal matrix in which non-zero diagonal entries are the reciprocals of the positive diagonal entries of $\boldsymbol{D}$.

The construction of a reaction matrix is simple and is done as follows

	$R1$	$R2$	...	$R\alpha$	$P1$	$P2$	...	$P\beta$
$E1$	N_{E1R1}	N_{E1R2}	…	$N_{\mathrm{E1R}\alpha}$	$-N_{\mathrm{E1P1}}$	$-N_{\mathrm{E1P2}}$	...	$-N_{\mathrm{E1P}\beta}$
$E2$	N_{E2R1}	N_{E2R2}	…	$N_{\mathrm{E2R}\alpha}$	$-N_{\mathrm{E2P1}}$	$-N_{\mathrm{E2P2}}$	...	$-N_{\mathrm{E2P}\beta}$
$E3$	N_{E3R1}	N_{E3R2}	...	$N_{\mathrm{E3R}\alpha}$	$-N_{\mathrm{E3P1}}$	$-N_{\mathrm{E3P2}}$	...	$-N_{\mathrm{E3P}\beta}$
....								
$E\gamma$	$N_{\mathrm{E}\gamma\mathrm{R1}}$	$N_{\mathrm{E}\gamma\mathrm{R2}}$	...	$N_{\mathrm{E}\gamma\mathrm{R}\alpha}$	$-N_{\mathrm{E}\gamma\mathrm{P1}}$	$-N_{\mathrm{E}\gamma\mathrm{P2}}$	...	$-N_{\mathrm{E}\gamma\mathrm{P}\beta}$

Here Ri, $i = 1, 2,\ldots, \alpha$ denotes the reactants and Pi, $i = 1, 2,\ldots, \beta$, denotes the products. Ei, $i = 1, 2,\ldots, \gamma$ are the elements present in the reactant(s) or product(s) and N_{EiRj} (or N_{EiPk}), $i = 1, 2,\ldots, \gamma$, $j = 1, 2,\ldots, \alpha$ and $k = 1, 2,\ldots, \beta$ are the number of atoms of the ith element present in the jth reactant (or in the kth product).

Let us take an example of the chemical reaction 'UREIDOGLYCOLATE-HYDROLASE-RXN' given below

$$H_2O + C_3H_5N_2O_4 \rightarrow C_2HO_3 + NH_3 + CO_2.$$

The reaction matrix obtained by the above formulation is shown below.

	H_2O	$C_3H_5N_2O_4$	C_2HO_3	NH_3	CO_2
H	2	5	–1	–3	0
O	1	4	–3	0	–2
C	0	3	–2	0	–1
N	0	2	0	–1	0

This reaction matrix is rectangular-shaped.

The algorithm described below is for a single reaction at a time:

Step 1: Construct the reaction matrix $\boldsymbol{A}_{m,n}$.
Set $j = 0$.
Step 2: Test if any row of $\boldsymbol{A}$ contains only non-negative entries or only non-positive entries. If yes, print: 'Reaction is infeasible'
Go to Step 11.
Step 3: Determine the rank r of the matrix $\boldsymbol{A}$.
Step 4: Find $k = n - r$.
If $k = 0$, print: 'Reaction is infeasible'.
Go to Step 11.
If $k = 1$, print: 'Reaction is unique'.
Go to Step 5.
If $k \geq 2$, print: 'Reaction is not unique'.
Go to Step 5.
Step 5: Compute SVD representation of
$\boldsymbol{A}$: $\boldsymbol{A} = \boldsymbol{UDV}^T$.
Step 6: Compute Moore–Penrose inverse of
$\boldsymbol{A}$: $\boldsymbol{A}^+ = \boldsymbol{VD}^+\boldsymbol{U}^T$.
Step 7: Compute $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A})$.
Step 8: If $k = 1$,
multiply any one non-vanishing column of $\boldsymbol{B}$ by an appropriate real number so as to get its components (coefficients) in integral form.
Go to Step 10.
If $k \geq 2$,
Go to Step 9.
Step 9: Compute $S = \lambda_1 b_1 + \lambda_2 b_2 + \ldots + \lambda_n b_n$ for different integers λ_is and for non-vanishing column b_is of $\boldsymbol{B}$.
If every entry in S is non-zero and positive,
Go to Step 10.
Else
print: 'Interchange of metabolites from reactants to products or vice-versa and deletion of metabolite where coefficient is zero can result a solution'.
Step 10: Print the balancing coefficient(s).
Go to Step 12.
Step 11: Set $j = j + 1$,
if $j > N$.
Go to Step 12
(here N is the total number of filler metabolites).
Else
modify $\boldsymbol{A}$ taking into account the jth filler metabolite.
Go to Step 2 with new $\boldsymbol{A}$.
Step 12: Terminate the program.

We have used our tool for the reactions present in MetaCyc (Version: 14.6), a database of metabolic pathways containing 9299 reactions in total. The results are shown in Figure 1. Here we have observed that different types of discrepancies exist in a large number of reactions. The empirical formulas of participating metabolites of 3287 reactions are not present in the database. In 426 reactions, information about participating metabolites is absent. There are 2072 reactions which are generic in nature, i.e. general information is present, but information of a specific compound is absent (see Supplementary data). All these are mentioned as group A reactions in Figure 1. On the other hand, there are 5760 and 252 reactions which are balanced and unbalanced in mass respectively. These are presented as group B in Figure 1. Among the 252 unbalanced reactions, we have identified that 50 reactions are feasible, i.e. all these reactions can be mass-balanced with the present set of participating metabolites and 202 reactions are infeasible, which cannot be mass-balanced with the present set of metabolites. Our algorithm is able

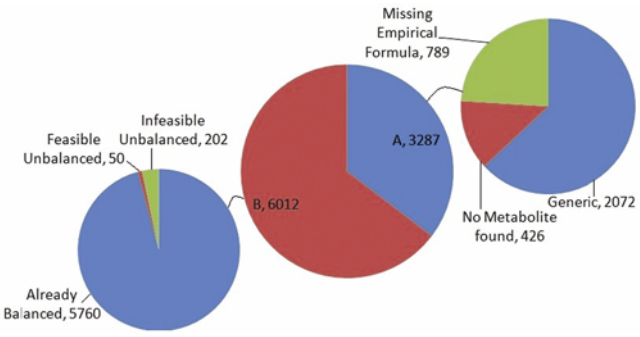

Figure 1. A represents generic reactions, those which do not have information of participating metabolites and those where formulas of participating metabolites are not present. B represents infeasible, feasible but unbalanced and already balanced reactions. Pie charts on the left and right show total number of reactions from different subgroups.

to find the correct stoichiometric coefficients for the feasible unbalanced reactions (see Supplementary data). Further, it is also able to identify the set of the possible filler metabolites for an infeasible reaction to make it feasible and provide the correct stoichiometry (see Supplementary data). Thus we have identified each category of the reactions in the database.

The reactions which violate mass conservation principle are results of different unbalanced elements in reactants and products. For example, out of 50 feasible unbalanced reactions, we find that 45 reactions which can be balanced using our method and have unique solutions, were originally stoichiometrically incorrect due to unbalance in H (27), O (6), O–H (3) and in more than two elements (9). The numbers within parenthesis give the number of unbalanced reactions of that particular type. We find five reactions which have multiple solutions, were originally stoichiometrically incorrect due to unbalance in H (4) and O (1). An example of making an infeasible reaction into a feasible one is given below.

Reaction '1.20.4.2-RXN' from MetaCyc:

$$CH_4AsO_3 + C_{10}H_{16}N_3O_6S \rightarrow CH_5AsO_2 + C_{20}H_{30}N_6O_{12}S_2.$$

Our method suggests that this reaction can be balanced using water and proton as filler metabolites; the modified reaction is given below.

$$CH_4AsO_3 + 2C_{10}H_{16}N_3O_6S + H = H_2O + CH_5AsO_2 + C_{20}H_{30}N_6O_{12}S_2.$$

The above reaction is also present in Rhea[16], a manually curated database reported recently, and the corresponding reaction ID is RHEA:15970. Our predicted result matches with the manually curated result of Rhea. Four other infeasible reactions (STIPITATONATE-DECARBOXYLASE-RXN, APIOSE-1-REDUCTASE-RXN, AQUACOBALAMIN-REDUCTASE-NADPH-RXN and AQUACOBALAMIN-REDUCTASE-RXN) of MetaCyc, which are curated and already reported in Rhea also match with our prediction (see Supplementary data). Our algorithm is able to predict the set of filler metabolites for each of the other infeasible reactions which are not reported in Rhea, but present in MetaCyc. In principle, the method can also be used for any reaction present in other databases.

Examples with performance matrices are given below.

(1) Reaction 'TRANS-HEXAPRENYLTRANST-RANSFERASE-RXN' is as follows:

$$C_{15}H_{25}O_7P_2 + C_5H_9O_7P_2 \rightarrow C_{35}H_{57}O_7P_2 + HO_7P_2.$$

The reaction matrix is shown below.

	$C_{15}H_{25}O_7P_2$	$C_5H_9O_7P_2$	$C_{35}H_{57}O_7P_2$	HO_7P_2
H	25	9	–57	–1
C	15	5	–35	0
O	7	7	–7	–7
P	2	2	–2	–2

The Moore–Penrose inverse of the above reaction matrix $\boldsymbol{A}^+$ is

$$\begin{bmatrix} -0.000296 & 0.002212 & 0.031069 & 0.004450 \\ -0.007611 & 0.000004 & 0.058372 & 0.008198 \\ -0.014336 & -0.006628 & 0.023538 & 0.003046 \\ 0.011269 & 0.001100 & -0.072024 & -0.010072 \end{bmatrix}.$$

Solution x can be obtained by the following matrix $[\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A}]$:

$$\begin{bmatrix} 0.7478260863 & -0.2347826090 & 0.2869565232 & 0.2260869566 \\ -0.2347826087 & 0.6434782609 & -0.0086956522 & 0.4173913043 \\ 0.2869565215 & -0.0086956523 & 0.1217391310 & 0.1565217392 \\ 0.2260869562 & 0.4173913042 & 0.1565217399 & 0.4869565218 \end{bmatrix}.$$

If we multiply the above matrix by $c = [1\ 1\ 1\ 0]^T$, we get $[0.8\ 0.4\ 0.4\ 0.8]^T$ and multiplication of this vector by 2.5 results in $x = [2\ 1\ 1\ 2]^T$. Thus the balanced reaction is

$$2C_{15}H_{25}O_7P_2 + C_5H_9O_7P_2 = C_{35}H_{57}O_7P_2 + 2HO_7P_2.$$

Again multiplying the above matrix by $c = [0\ 1\ 0\ 2]^T$, we get $[0.217391\ 1.478261\ 0.304348\ 1.391304]^T$ and multiplying this vector by 23 we get $x = [5\ 34\ 7\ 32]^T$. This results in the following reaction

$$5C_{15}H_{25}O_7P_2 + 34C_5H_9O_7P_2 = 7C_{35}H_{57}O_7P_2 + 32HO_7P_2.$$

Thus two different solutions (non-unique reaction) can be calculated from one solution matrix.

(2) Reaction 'RXN-11389' is as follows

$$C_{17}H_{24}N_2O_2 \rightarrow C_{16}H_{21}N_2O.$$

The reaction matrix for the above reaction is shown below

	$C_{17}H_{24}N_2O_2$	$C_{16}H_{21}N_2O$
H	24	–21
C	17	–16
O	2	–1
N	2	–2

The rank of the above matrix is equal to its column, and so we cannot balance the reaction. However, adding a filler metabolite 'GLYCOL' as product, it is possible to

balance the reaction. The modified reaction matrix is given below

	$C_{17}H_{24}N_2O_2$	$C_{16}H_{21}N_2O$	$C_2H_6O_2$
H	24	–21	–6
C	17	–16	–2
O	2	–1	–2
N	2	–2	0

The Moore–Penrose inverse of the above matrix $\boldsymbol{A}^+$ is:

$$\begin{bmatrix} 0.041702 & -0.027794 & 0.013800 \\ 0.065837 & -0.122720 & 0.036321 \\ -0.215076 & 0.301027 & -0.100243 \end{bmatrix}.$$

Solution x can be obtained by the following matrix $[\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A}]$:

$$\begin{bmatrix} 0.4444444445 & 0.4444444444 & 0.2222222222 \\ 0.4444444445 & 0.4444444444 & 0.2222222222 \\ 0.2222222222 & 0.2222222222 & 0.1111111111 \end{bmatrix}.$$

Multiplication of the above matrix by $c = [0\ \ 0\ \ 9]^T$ results in $x = [2\ \ 2\ \ 1]^T$ and the balanced reaction is

$$2C_{17}H_{24}N_2O_2 = 2C_{16}H_{21}N_2O + C_2H_6O_2.$$

Any metabolic model should obey the principle of mass conservation and conservation of atoms; otherwise the model would be erroneous and the analysis based on that model would provide incorrect predictions. Researchers start metabolic modelling by collecting and using the list of reactions from public databases, and almost all these public databases have a large number of reactions violating the mass and atom conservation principle. This algorithm would find the reactions for which one can get a mathematical solution (to balance the reaction) with the present set of metabolites, i.e. one can calculate the correct stoichiometric coefficients of metabolites. At the same time it can identify the reactions (infeasible reactions) for which mathematical solution does not exist with the present set of metabolites. The algorithm can suggest a set of metabolites from the known biochemical compounds to make an infeasible reaction feasible. In other words, when the suggested metabolite(s) are added to the existing set of metabolites of an infeasible reaction, the reaction with a new set of metabolites obeys the law of mass and atom conservation. Of course, these suggestions should be verified experimentally. Biochemists should study these infeasible reactions. Otherwise, it would create a gap in the metabolic model-building and hence in metabolic engineering. In addition, the output of the code would provide all the necessary information (correct stoichiometry of metabolites, suggested filler metabolite, etc.) about the reactions in one place, thus minimizing the labour of metabolic model builders.

We have applied our algorithm to the reactions of MetaCyc; we will include reactions of other databases like RiceCyc, AraCyc[34], KEGG, etc. in future, as the algorithm can be used for any biochemical reaction. Thus it would help in filling the gap of different scales (small pathway to genome-scale) in the metabolic modelling processes. A web-based tool using the algorithm is our next goal.

1. Fell, D. A. and Small, J. R., Fat synthesis in adipose tissue: an examination of stoichiometric constraints. *J. Biochem*., 1986, **238**, 781–786.
2. Poolman, M. G., Fell, D. A. and Raines, C. A., Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *Eur. J. Biochem*., 2003, **270**, 430–439.
3. David, H., Ozcelik, I. S., Hofmann, G. and Nielsen, J., Analysis of *Aspergillus nidulans* metabolism at the genomescale. *BMC Genomics*, 2008, **9**, 163–177.
4. Feist, A. M. *et al.*, A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol*., 2007, **3**, 121–138.
5. Poolman, M. G., Miguet, L., Sweetlove, L. J. and Fell, D. A., A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiol*., 2009, **151**, 1570–1581.
6. Raghunathan, A., Reed, J., Shin, S., Palsson, B. and Daefler, S., Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host–pathogen interaction. *BMC Syst. Biol*., 2009, **3**, 38.
7. Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S. and Palsson, B. O., Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol*., 2002, **184**, 4582–4593.
8. Raman, K. and Chandra, N., Flux balance analysis of biological systems: applications and challenges. *Brief. Bioinform.*, 2009, **10**, 435–449.
9. Thiele, I. *et al.*, A community effort towards a knowledgebase and mathematical model of the human pathogen *Salmonella typhimurium* LT2. *BMC Syst. Biol*., 2011, **5**, 8.
10. Nair, N. U., Goyal, N. and Chandra, N. R., Enhanced flux balance analysis to model metabolic networks. In BCB'10 Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, NY, USA, 2010.
11. Thiele, I. and Palsson, B. O., A protocol for generating a high-quality genomescale metabolic reconstruction. *Nature Protoc*., 2009, **5**, 93–121.
12. Verkhedkar, K. D., Raman, K., Chandra, N. R. and Vishveshwara, S., Metabolome based reaction graphs of *M. tuberculosis* and *M. leprae*: a comparative network analysis. *PLOS One*, 2007, **2**, 9, e881.
13. Caspi, R. *et al.*, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, 2008, **36**, D623–D631.
14. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. and Hirakawa, M., KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, 2010, **38**, D355–D360.
15. Poolman, M. G., Bonde, B. K., Gevorgyan, A., Patel, H. H. and Fell, D. A., Challenges to be faced in the reconstruction of metabolic networks from public databases. *Syst. Biol.* (*Stevenage*), 2006, **153**, 379–384.
16. Alcantara, R. *et al.*, Rhea – a manually curated resource of biochemical reactions. *Nucleic Acids Res.*, 2011, **40**, D754–D760.

17. Ott, M. A. and Vriend, G., Correcting ligands, metabolites, and pathways. *BMC Bioinform*., 2006, **7**, 517.
18. Gille, C., Hbner, K., Hoppe, A. and Holzhtter, H. G., Metannogen: annotation of biological reaction networks. *Bioinformatics*, 2011, **27**, 2763–2764.
19. Félix, L. and Valiente, G., Validation of metabolic pathway databases based on chemical substructure search. *Biomol. Eng*., 2007, **24**, 327–335.
20. Gevorgyan, A., Poolman, M. G. and Fell, D. A., Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, 2008, **24**, 2245–2251.
21. Woolf, A. A., Unconventional oxidation numbers. *J. Chem. Educ*., 1997, **74**, 1270.
22. Andrews, D. H. and Kokes, R. J., *Fundamental Chemistry*, Wiley, New York, 1963.
23. Das, S. C., A mathematical method of balancing a chemical equation. *Int. J. Math. Educ. Sci. Technol*., 1986, **17**, 191–200.
24. Sen, S. K., Agarwal, H. and Sen, S., Chemical equations balancing: an integer programming approach. *Math. Comput. Model*., 2006, **44**, 678–691.
25. Risteski, I. B., A new nonsingular matrix method for balancing chemical equations and their stability. *Int. J. Math. Manuscripts*, 2007, **1**, 180–205.
26. Risteski, I. B., A new singular matrix method for balancing chemical equations and their stability. *J. Chin. Chem. Soc*., 2009, **56**, 65–79.
27. Risteski, I. B., A new complex vector method for balancing chemical equations. *Mater. Technol*., 2010, **44**, 193–203.
28. Risteski, I. B., A new generalized matrix inverse method for balancing chemical equations and their stability. *Bol. Soc.*, 2008, **2**, 104–115.
29. Risteski, I. B., A new approach to balancing chemical equations. *SIAM Probl. Solutions*, 2007, 1–10.
30. Risteski, I. B., A new pseudoinverse matrix method for balancing chemical equations and their stability. *J. Korean Chem. Soc*., 2008, **52**, 223–238.
31. Cheney, W. and Kincaid, D., *Numerical Mathematics and Computing*, Brooks/Cole Publishing Company, 1999, 4th edn.
32. Ben Israel, A. and Greville, T. N. E., *Generalized Inverses: Theory and Applications*, Wiley, New York, 1974.
33. Rao, C. R. and Mitra, S. K., *Generalized Inverse of Matrices and its Applications*, Wiley, New York, 1971.
34. Jaiswal, P. *et al.*, Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res*., 2006, **34**, D717–D723.

ACKNOWLEDGEMENTS. We thank Prof. David A. Fell for valuable discussions and advice regarding filler metabolites; Dr Mark G. Poolman and Dr Md. Aftabuddin for their valuable suggestions and the Bioinformatics Centre of Calcutta University for computational facilities.

Received 2 June 2012; accepted 12 October 2012

Yttrium and rare earth element contents in seamount cobalt crusts in the Indian Ocean

V. Balaram[1,*], V. K. Banakar[2], K. S. V. Subramanyam[1], Parijat Roy[1], M. Satyanarayanan[1], M. Ram Mohan[1] and S. S. Sawant[1]

[1]National Geophysical Research Institute (CSIR), Hyderabad 500 007, India
[2]National Institute of Oceanography (CSIR), Dona Paula, Goa 403 004, India

Cobalt-rich Fe–Mn crusts occur on almost all seamounts and plateaus of the world oceans. Fe–Mn crusts are formed through layer-by-layer accretion of colloidal precipitates from cold ambient seawater onto exposed seamount rock substrates. This study reports high concentrations of rare earth elements (REE) and yttrium ranging from 1,727 to 2,511 μg/g in the crust samples collected from the Afanasy Niktin Seamount (ANS) in the Eastern Equatorial Indian Ocean. The concentrations of REE in the ANS Fe–Mn crusts are much higher than those of the mid-Pacific seamount and nodules (1,180–1,434 μg/g). Ce-enrichment up to 0.17% has been recorded in the present study as against ~ 0.1% content in global seamount Fe–Mn crusts. This enrichment is attributed to oxidative removal of Ce from seawater to the marine Fe–Mn crust. The negative Ce-anomalies obtained for seawater samples from the ANS region coupled with strong positive Ce-anomalies in Fe–Mn crusts clearly indicate that the source of Ce in ANS Fe–Mn crusts is sea-water. This investigation warrants further detailed exploration studies in order to make an estimate of these highly useful elements in the cobalt enriched Fe–Mn crusts of Indian Ocean.

Keywords: Cobalt crust, rare earth elements, seamount, yttrium.

THE future of India's and the world's enthusiastically envisaged green technologies depends on the availability of several trace and ultra-trace metals, including rare earth elements (REEs) in adequate quantity. REEs are critical constituents to many of the world's most advanced technologies such as defence metallurgy, consumer electronics, medical applications, etc. More than 97% of the world's REE ore production is from mines in China, which have restricted their exports recently for reasons unknown. Other countries with notable production are Brazil, India, Kyrgyzstan and Malaysia. Mainly, monazite from beach placers is mined in India as the principal ore mineral for REE, although xenotime holds out some

*For correspondence. (e-mail: balaram1951@yahoo.com)