Modular Decomposition of Metabolic Systems via Null Space Analysis

Mark G. Poolman[a†], Cristiana Sebu[b], Michael K. Pidcock[b] and David A. Fell[a]

a School of Life Science, Oxford Brookes University, Headington, Oxford, OX3 OBP

b Department of Mathematical Sciences, Oxford Brookes University, Wheatley, Oxfordshire, OX33 1HX

† Corresponding author.

Telephone 01865 483638

e-mail mgpoolman@brookes.ac.uk

Abstract

We describe a method by which the reactions in a metabolic system may be grouped hierarchically into sets of modules to form a metabolic reaction tree. In contrast to previous approaches, the method described here takes into account the fact that, in a viable network, reactions must be capable of sustaining a steady-state flux.

In order to achieve this decomposition we introduce a new concept - the *reaction correlation coefficient*, ϕ, and show that this is a logical extension of the concept of enzyme (or reaction) subsets. In addition to their application to modular decomposition, reaction correlation coefficients have a number of other interesting properties, including a convenient means for identifying disconnected subnetworks in a system and potential applications to metabolic engineering.

The method computes reaction correlation coefficients from an orthonormal basis of the null-space of the stoichiometry matrix. We show that reaction correlation coefficients are uniquely defined, even though the basis of the null-space is not.

Once a complete set of reaction correlation coefficients is calculated, a metabolic reaction tree can be determined through the application of standard programming techniques. Computation of the reaction correlation coefficients, and the subsequent construction of the metabolic reaction tree is readily achievable for genome-scale models using a commodity desk-top PC.

1 Introduction

The increasing online availability of annotated genomes makes routine the task of automatically† constructing genome-scale structural metabolic models (i.e. models in which only reaction stoichiometries are taken into consideration) †. Two challenges remain: ensuring the correctness of models thus generated (Poolman et al. 2006), and the subsequent analysis of such models. The latter challenge results from the fact that while there is a comprehensive set of theoretical and computational tools with which the analysis of small models ($<\approx$ 30 reactions and metabolites) may be undertaken (Schilling et al. 1999; Schilling et al. 2000; Klamt et al. 2002; Schuster et al. 2002; Lemke et al. 2004) none of these scale well when applied to large models ($>\approx$ 100 reactions and metabolites).

†1.a

†1.1

This poor scaling is not solely due to a lack of computational power, but rather to the fact that current methods tend to either produce large amounts of relatively unstructured data, describing network properties in terms of sets of individual reactions and/or metabolites, or depend upon

modelling assumptions that are not independently verifiable. Even when such calculations can be computed in a reasonable time, they may yield little new information about the overall structure of the network under consideration.

A potential solution to this problem would be the identification within a larger model of 'sub-models' within which each reaction is uniquely located. The concept of enzyme subsets (Pfeiffer et al. 1999) would appear to offer a partial solution. These are defined as sets of reactions which, at steady-state, are stoichiometrically constrained to carry flux in fixed ratio. The enzyme subsets of a model may readily be calculated from its stoichiometry matrix, as described in Pfeiffer et al. and below. It is arguable that the term *reaction subset* rather than enzyme subset is preferable, as it is reactions, not enzymes that carry flux. Unfortunately, the approach suffers from the problem that when applied to large systems, large numbers of small subsets are generated (Bonde 2006). †

†1.2

However, it is also clear that a method that separated the reactions in a system into a small number of subsystems would be equally unsatisfactory. Dividing a system of nine hundred reactions into three subsystems of three hundred reactions each is hardly preferable to dividing it into three hundred subsystems of three reactions each.

Furthermore the question as to the optimal number of reactions that a subsystem should contain (or conversely, how many subsystems a system should be divided into) is not one that can be answered *a priori*.

For these reasons a hierarchical approach is proposed whereby a metabolic system is represented as a *metabolic tree* in which the root node represents the complete system, leaf nodes represent individual reactions, and intermediate nodes represent unique subsystems of reactions.

Each node in the metabolic tree can thus be thought of as representing a *metabolic module*, capable of the net interconversion of metabolites common to reactions inside and outside the module. The hierarchical nature of the tree means that an investigator can select modules of a size convenient for a given purpose.

2 Theory and Method

Arranging a set of entities into a hierarchical tree has two prerequisites: some means by which the difference between a pair of entities can be measured, and a method for using these differences to construct the tree. These differences are conventionally represented as a square matrix of non-negative elements in which a value of zero indicates no difference between corresponding rows and columns, and increasing positive values denoting an increasing measure

of difference. We denote this the dissimilarity matrix, $\boldsymbol{\Delta}$. Once $\boldsymbol{\Delta}$ has been constructed, a number of generally applicable algorithms exist with which to construct the tree. We first describe the construction of the $\boldsymbol{\Delta}$, then that of the tree.

2.1 Construction of the dissimilarity matrix, $\boldsymbol{\Delta}$

In general, with the exceptions described below, † all reactions in a system are †2.1
capable of maintaining a steady state flux, and all steady state flux vectors must lie within the right null-space of the stoichiometry matrix, $\boldsymbol{N}$, of the system. See (Heinrich and Schuster 1996; Klamt et al. 2002; Papin et al. 2003) for recent work describing the application of linear algebra, and in particular analysis of the null-space, to metabolic systems. The null-space is spanned by the columns of the $n \times d$ null-space (or kernel) matrix $\boldsymbol{K}$, where n is the number of reactions and d the dimension of the null-space (equal to the number of reactions minus the rank of $\boldsymbol{N}$) †. Each reaction is therefore †1.b
associated with a d dimensional row vector. We may therefore construct the symmetric matrix † $\boldsymbol{\Delta}$ in terms of the angles between these vectors, such that †2.2
$\boldsymbol{\Delta}_{ij}$ denotes the angle between row vectors $\boldsymbol{K}_i$ and $\boldsymbol{K}_j$. In this paper we use the notation $\theta_{ij}^{\boldsymbol{A}}$ to represent the angle between rows i and j in matrix $\boldsymbol{A}$, so given a system of n reactions

$$\boldsymbol{\Delta}_{ij} = \theta_{ij}^{\boldsymbol{K}} : 1 \leq i \leq n, 1 \leq j \leq n.$$

Thus the minimal possible difference between reactions is $\boldsymbol{\Delta}_{ij} = 0$ (reaction vectors are parallel) and maximum absolute value $\boldsymbol{\Delta}_{ij} = \pi/2$ (reaction vectors are orthogonal).

It is possible that a number of reactions in a system will be incapable of carrying flux as a result of constraints imposed by the structure of the system. Occasionally this will not be the result of an error in the construction of the model but more often it is. Such 'dead' reactions (also called "strictly detailed balanced" reactions (Schuster and Schuster 1991)) are often (but not always, see discussion) associated with a zero row vector in $\boldsymbol{K}$, in which case they may be readily identified. Obviously, no angle can be meaningfully assigned between a zero vector and any other, and so such dead reactions must be removed before proceeding.

It is also desirable to remove isostoichiometric reactions from the model, as they add no new information to a structural model and distort the results obtained from most structural analyses, multiplying the number of elementary modes and fragmenting enzyme subsets. Of course, if the results from a structural model are to be applied in the context of a kinetic model, or

ex silico, then due consideration must be given to the enzymes that catalyse the reactions, and it is in these contexts that the presence of isoenzymes becomes relevant. It is this reason that leads us to prefer, in common with, for example, Reed and Palsson (2004), † the term "reaction subset" rather than "enzyme subset", †

†make more consistent with 1.2 also 1.c

The only apparent drawback to this approach is that $\boldsymbol{K}$ is, in most cases, non-unique and depends upon both the algorithm used for its calculation and the initial row and column order of $\boldsymbol{N}$. However, it can be shown (appendix A) that despite this, the angles between the row vectors of any $\boldsymbol{K}$ for a given $\boldsymbol{N}$ are unique, provided $\boldsymbol{K}$ is orthogonal, that is: $\boldsymbol{K}\boldsymbol{K}^T = \boldsymbol{I}$, in which case column vectors are orthonormal, and $\boldsymbol{K}$ represents an orthonormal basis of the null-space of $\boldsymbol{N}$. The use of an orthogonal basis matrix in this context is in contrast to much previous work in this area, in which $\boldsymbol{K}$ is assumed to be of the form: †

†As above

†2.21

$$\boldsymbol{K}^I = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{K'} \end{bmatrix}$$

In the rest of this paper, $\boldsymbol{K}$ is assumed to be orthogonal, unless explicitly stated otherwise.

It transpires that the elements of $\boldsymbol{\Delta}$, θ_{ij}^K, have a simple, but useful, "real-world" interpretation: $\cos(\theta_{ij}^K)$ is Pearson's (population) correlation coefficient, r_{ij}, between the fluxes carried by the pair of reactions i and j for all possible steady states of the system. We therefore denote the cosine of † this angle the *reaction correlation coefficient*, and represent it with the symbol ϕ (i.e. $\phi_{ij} = \cos(\theta_{ij}^{\boldsymbol{K}})$) and the corresponding $(n \times n)$ matrix of all ϕ, $\boldsymbol{\Phi}$. The derivation of this relationship is given in appendix B. $\boldsymbol{\Phi}$ thus provides a global, unique, and invariant set of characteristics of a metabolic system.

†2.3

For a pair of reactions, i and j, with corresponding rows in $\boldsymbol{K}$, $\boldsymbol{K}_i$, $\boldsymbol{K}_j$, ϕ_{ij} may be readily determined as: †

†2.4

$$\phi_{ij} = \frac{\boldsymbol{K}_i \boldsymbol{K}_j^T}{\sqrt{(\boldsymbol{K}_i \boldsymbol{K}_i^T)}\sqrt{(\boldsymbol{K}_j \boldsymbol{K}_j^T)}} = \cos(\theta_{ij}^{\boldsymbol{K}})$$

From the †foregoing, it is clear that ϕ must fall within the range $-1 \leq \phi \leq 1$ and that two special cases apply:

†2.5

$\phi_{ij} = \pm 1$**:** implies that row vectors $\boldsymbol{K}_i$ and $\boldsymbol{K}_j$ are parallel, and that they thus carry steady-state flux in a fixed ratio. From the original definition in Pfeiffer et al. (1999) this is equivalent to stating that the reactions are members of the same subset.

$\phi_{ij} = 0$**:** implies vectors $\boldsymbol{K}_i$ and $\boldsymbol{K}_j$ are orthogonal, and reactions i and j are in stoichiometrically disconnected subsystems, because there can be no

correlation between flux in i and j unless the system has at least one elementary mode in which both i and j have a non-zero coefficient. See appendix C for a more complete explanation. † †1.4

Thus ϕ spans the continuum of possible correlations between any pair of reaction fluxes, from being completely dependent to completely independent. Therefore a reaction correlation coefficient can be regarded as a quantitative generalisation of the qualitative concept of the reaction subset.

The sign of ϕ is dependent only upon the relative signs of non-zero elements of columns i and j in $\boldsymbol{N}$, which in turn depend upon the initial reaction specification. For example consider a reaction system containing reactions i and j capable of carrying flux at steady-state: † †2.6 start, also minor 2.7

$$\begin{aligned} i: &\quad a \longleftrightarrow b \\ j: &\quad b \longleftrightarrow c \end{aligned}$$

in which flux is defined as positive in the left $\rightarrow$ right direction, metabolites a, b and c are internal (i.e. non-boundary) metabolites, and b is involved in no other reactions. The corresponding rows of $\boldsymbol{K}$ may be written: † †2.6 cntd

$$\begin{aligned} \boldsymbol{K}_i &= [x_1 ... x_d] \\ \boldsymbol{K}_j &= [x_1 ... x_d] \end{aligned}$$

i.e. the two rows are equal vectors, hence parallel, so i and j exist in a reaction subset and carry equal flux and $\phi_{ij} = 1$. If reaction j is now defined in the opposite direction: † †2.6 cntd

$$\begin{aligned} i: &\quad a \longleftrightarrow b \\ j: &\quad c \longleftrightarrow b \end{aligned}$$

the corresponding rows in $\boldsymbol{K}$ will now be: † †2.6

$$\begin{aligned} \boldsymbol{K}_i &= [y_1 ... y_d] \\ \boldsymbol{K}_j &= [-y_1 ... - y_d] \end{aligned}$$

i and j remain in a reaction subset, but their corresponding row vectors in $\boldsymbol{K}$ are now *anti*parallel and $\phi_{ij} = -1$. Thus the sign of ϕ is determined, at least in part, by the way in which the direction of reactions are defined, and does not simply indicate competition (e.g. as would occur at a branch-point) between reactions. † †2.6

† Consequently values of ϕ used in the remainder of this paper will be taken as the absolute value. †2.6 END also minor 2.8

$\boldsymbol{\Delta}$ may be conveniently calculated by first calculating $\boldsymbol{\Phi}$:

$$\boldsymbol{\Delta}_{ij} = \cos^{-1}(\boldsymbol{\Phi}_{ij}) : 1 \leq i \leq n, 1 \leq j \leq n$$

$\boldsymbol{\Delta}$ is needed for the construction of a hierarchical tree, but $\boldsymbol{\Phi}$ contains the more readily interpretable description of relationships between reactions.

2.2 Construction of the tree.

Having obtained $\mathbf{\Delta}$, the corresponding tree is generated using the WPGMA algorithm (weighted pair group method using arithmetic averaging) (Lance and Williams 1967; Morgan and Ray 1995) as described in appendix D. A drawback to this algorithm is the potential for more than one pair of reactions to appear as nearest neighbours. Although the method of resolving ambiguities described in appendix D may appear to be rather arbitrary, to date observed ambiguities are limited to two special cases: the presence of isostoichiometric reactions or reactions that belong to a reaction subset with more than two members. Both of these cases may be readily avoided by appropriate pre-treatment of the model.

The former can be trivially solved, the latter by creating a 'condensed' model (Pfeiffer et al. 1999; Klamt and Stelling 2002) in which reactions present in each subset are replaced with a single compound reaction with a net stoichiometry corresponding to that of the subset. It is worth noting that to ensure a completely condensed model, the process of removing isostoichiometric reactions and substituting subsets should be applied iteratively; the presence of isostoichiometric reactions has the effect of breaking up reaction subsets, but it is possible for a model to contain isostoichiometric subsets.

† †2.9 Large chunk removed to appendix

3 Application to Models

Three metabolic models were analysed using the approach described: a relatively small model with a partially defined modular structure, and two genome-scale models, one based on the model of *Streptomyces coelicolor* described by Borodina et al. (2005), and the other based on the *Escherichia coli* model of Reed et al. (2003).

3.1 A model of photosynthate metabolism.

In order to determine whether the algorithm described is indeed capable of identifying modules in a metabolic system, a test model with a predefined modular structure, approximating plant photosynthate biochemistry, herafter referred to as '*photo*', was constructed with the overall structure as shown in Fig. 1. The 'source' module consists of two identical copies of a model of the Calvin cycle described previously (Poolman et al. 2001; Poolman et al. 2003). The difference between the version used here and the previously described version is that the only C_3 metabolite exported to the cytosol in exchange for inorganic phosphate is phosphoglycerate (PGA).

Dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate are not exchanged. Exported PGA is transported to 'sink' tissue, also in exchange for inorganic phosphate. The sink tissue is derived from a model describing potato tuber carbohydrate metabolism (Assmus 2005). This is divided into cytosolic and plastidic compartments and contains reactions commonly assigned to glycolysis, sucrose synthesis, and starch synthesis. Carbon entering in the form of PGA has 3 ultimate destinations: pyruvate (assumed to be the end product of glycolysis in this model), sucrose *via* sucrose synthesis, and plastidic starch from starch synthesis. Overall the model has 75 reactions, 65 internal and 8 external metabolites, and is available in the supplementary material.

3.2 The *Streptomyces coelicolor* model

A model of *Streptomyces coelicolor*, herafter referred to as '*sco*', was generated from the reaction set described by Borodina et al. (2005) and in the spreadsheet available from `http://www.genome.org/content/vol15/issue6/images/data/820/DC1/Dataset_2[1]_List_of_reactions.xls`. Reactions for macromolecule biosynthesis (with non integer stoichiometry) were removed, other non integer stoichiometries were scaled appropriately and protons were made external yielding an initial model with 954 reactions, 484 internal metabolites and 104 external metabolites. This was then subject to the following process:

1. Isostoichiometric reactions were removed.
2. Reactions identified as dead from examination of $\boldsymbol{K}$ were removed.
3. The model was then condensed, i.e. reactions forming enzyme subsets were replaced with a single reaction.

It is important that the steps are carried out in the order described, as isostoichiometric reactions form cycles, resulting in the possible failure of identification of dead reactions in step 2. Steps 1 and 3 must then be repeated until no further reduction in model size can be achieved. This is necessary as the possibility exists for the generation of isostoichiometric subsets in step 3. In practice the process appears to converge extremely rapidly, in this instance just two iterations were needed; the initial condensation produced only four isostoichiometric subsets.

This procedure resulted in a model of 363 reactions, 166 internal metabolites and 97 external metabolites. Such condensation greatly reduces the computational load of subsequent calculations and simplifies the interpretation of subsequent results. The *photo* model was not condensed as it is a

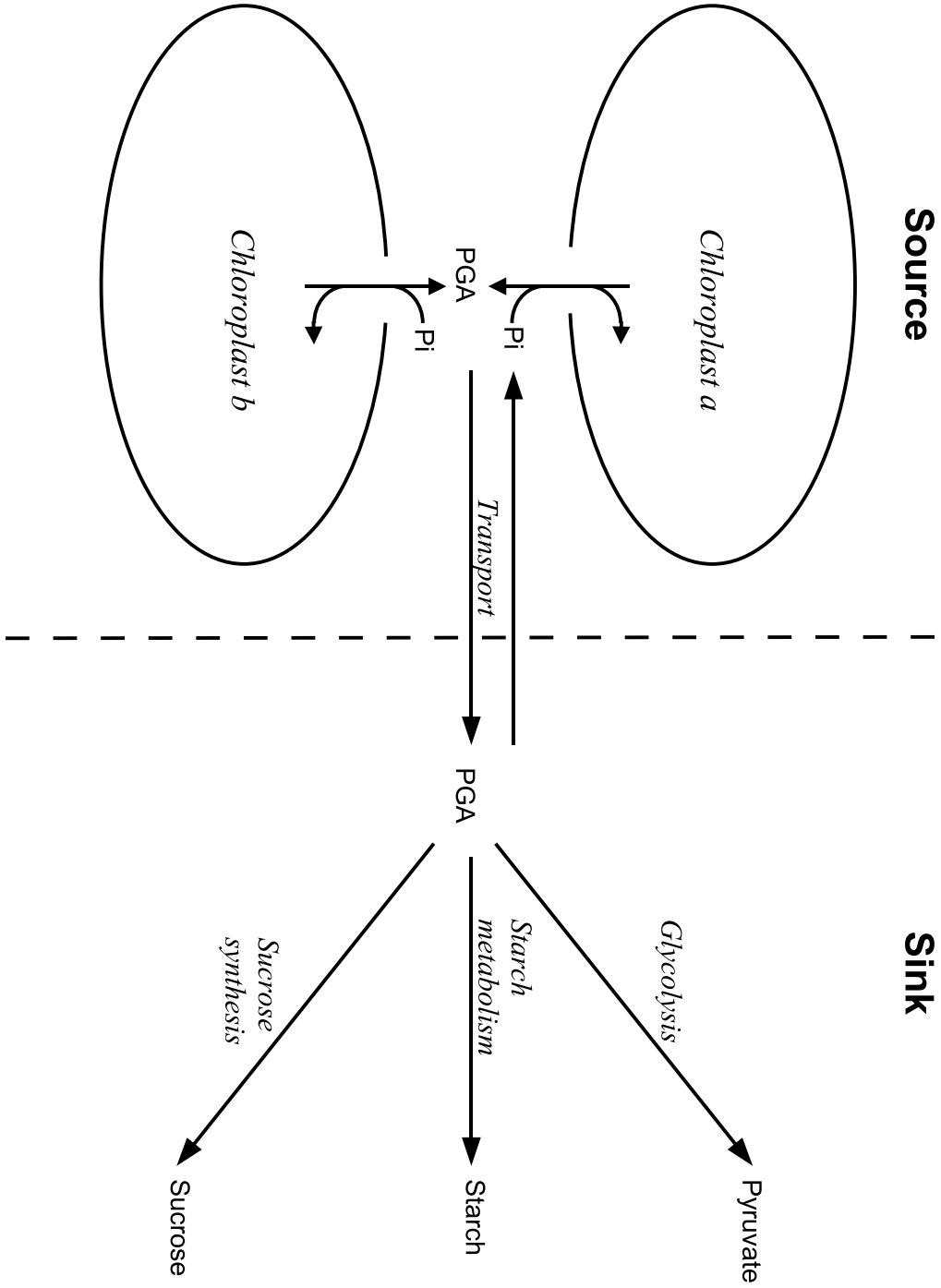


Figure 1: Modular structure of the *photo* model. This can be split into three entirely independent models (two chloroplast, and sink metabolism) by making source PGA and phosphate external.

much smaller model, and is known *a priori* not to contain isostoichiometric or dead reactions.

3.3 The *Escherichia coli* model

A model of *Escherichia coli* described by Reed et al. (2003), and hereafter described as '*eco*' was obtained from `http://systemsbiology.ucsd.edu/organisms/ecoli/ecoli_reactions.html`. This generated an initial model of 911 reactions with 1036 internal and 146 external metabolites. Processing this in the same manner as the *sco* model described above resulted in a model of 498 reactions, 202 internal and 129 external metabolites.

4 Results

4.1 Distributions of ϕ

The distribution of ϕ in all models was extremely asymmetric, and a $\log - \log$ plot of these distributions † (Fig. 2) is, at first sight, suggestive of a power †2.12 law distribution. However, plotting the normalised distribution (i.e the area under the curve is unity) † of $\log(\phi)$ (i.e. calculating $\log(\phi)$ first, and then †2.12 determining the distribution †), as shown in Figs. 3 and 4, reveals that the †2.12 *sco* and *eco* models have lognormal distributions. The distribution from the *photo* model shown in Fig. 4 (green line) is rather noisy (presumably a result of having originated from a much smaller dataset), rendering interpretation correspondingly more difficult. Repeating the plot using a number of different bin sizes appears to confirm that the distribution is bi-modal with peaks at $\log(\phi) \approx -2$ and $\log(\phi) \approx -0.5$ with a corresponding trough at $\log(\phi) \approx -1.6$.

A possible explanation for this apparently bi-modal distribution is that it is due to the highly modular nature of the network, with higher values of ϕ corresponding to correlation between pairs of reactions in the same module, and the lower valued peak resulting from correlation between pairs of reactions in different modules. This hypothesis was tested by breaking the connection between modules (by making the communicating metabolites external) resulting in three entirely independent modules, and recalculating $\mathbf{\Phi}$. The resulting histogram in Fig. 4 (red line) appears to support this hypothesis: the peak at $\log(\phi) \approx -0.5$ is amplified and that at $\log(\phi) \approx -2$ is all but absent. Furthermore, the long negative tail is notably truncated in the split model.

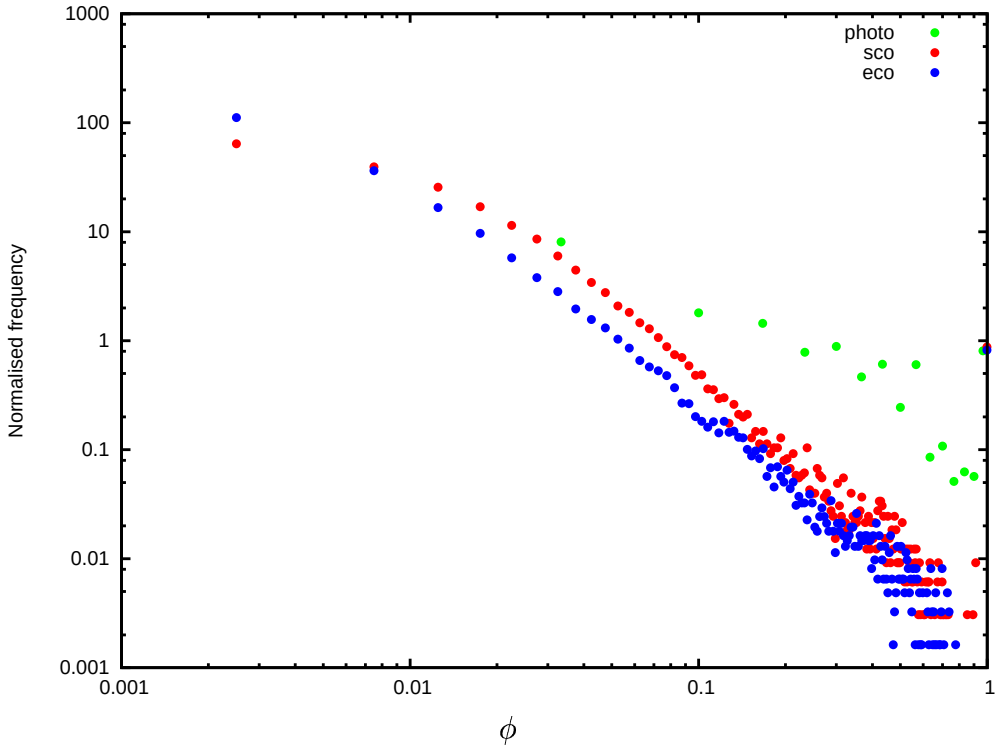


Figure 2: Normalised distributions of ϕ plotted in $\log - \log$ space from the three models described - Green circles: *photo*, Red circles: *sco*, Blue circles: *eco*.

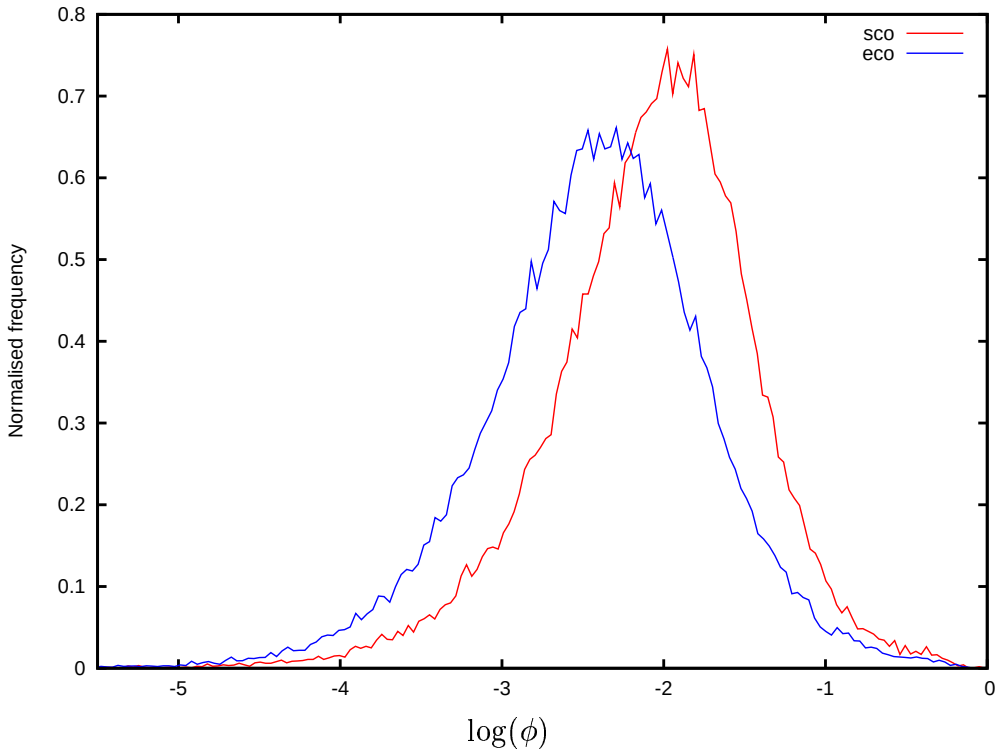


Figure 3: Normalised distributions of $\log(\phi)$ in the *sco* (red) and *eco* (blue) models.

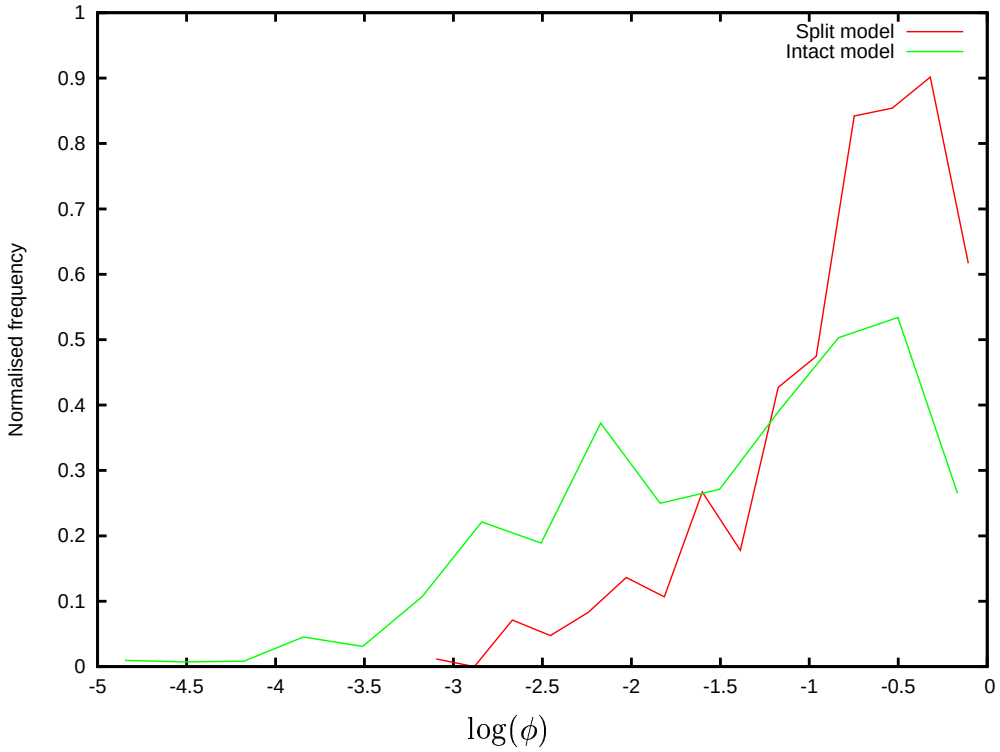


Figure 4: Normalised distribution of $\log(\phi)$ in the *photo* (green line) and split *photo* (red line) models. Connecting metabolites were removed from the *photo* model resulting in the split model which the consisted of three independent modules.

4.1.1 Metabolic trees

The reaction tree for the *photo* model is shown in Fig. 5. This quite clearly recovers the known modular structure of the model, and also reveals the enzyme subsets (which in this model were not condensed into single reactions). Another feature of interest is the close association between TPT_PGA and the PGI/PGM subset, highlighted for just one of the chloroplast modules. TPT_PGA represents the triosephosphate-phosphate translocator exporting PGA to the cytosol in exchange for cytosolic phosphate and the PGI/PGM (phosphoglucose isomerase/phosphoglucomutase) subset leads to starch. As these represent the only exit points for carbon from the chloroplast subsystems, as represented by this model, it is perhaps not surprising that they are highly correlated ($\phi = 0.991$ †), despite the fact that they are quite distant from one another on a normal pathway diagram. †2.13

Figs. 6 and 7 show the reaction trees for the *sco* and *eco* models respectively. For the sake of clarity, reaction names have been removed. Some, but not all, of the clusters in these trees have been tentatively assigned to particular biochemical roles by inspection. The clusters assigned to the TCA cycle and glycolysis are predominantly comprised of reactions traditionally associated with these classically defined pathways. The cluster most closely corresponding to a standard pathway is the oxidative pentose phosphate pathway (OPPP) shown in the *eco* reaction tree. Glycolysis is not seen as a well defined cluster in the *eco* model, but is mainly contained within the cluster identified as "misc. phospho-sugar".

Preliminary analysis of all the trees shown here, suggests the presence of more structure than has yet been characterised. However as the main objective of this paper is to present the theory and algorithm, such analyses are beyond its scope.

5 Discussion and Conclusions

5.1 General performance

The continuing improvement in computing performance means that it is now commonplace † to be able to apply algorithms to genome scale models. The only exception to this is the determination of elementary modes, which undergoes a severe combinatorial explosion in both memory and processing requirements (Klamt and Stelling 2002). †2.15

The algorithm described here depends only on the determination of the null-space matrix and subsequent operations upon it. Calculation of $\boldsymbol{K}$, $\boldsymbol{\Delta}$ and the tree itself take of the order of minutes on commodity desk-top

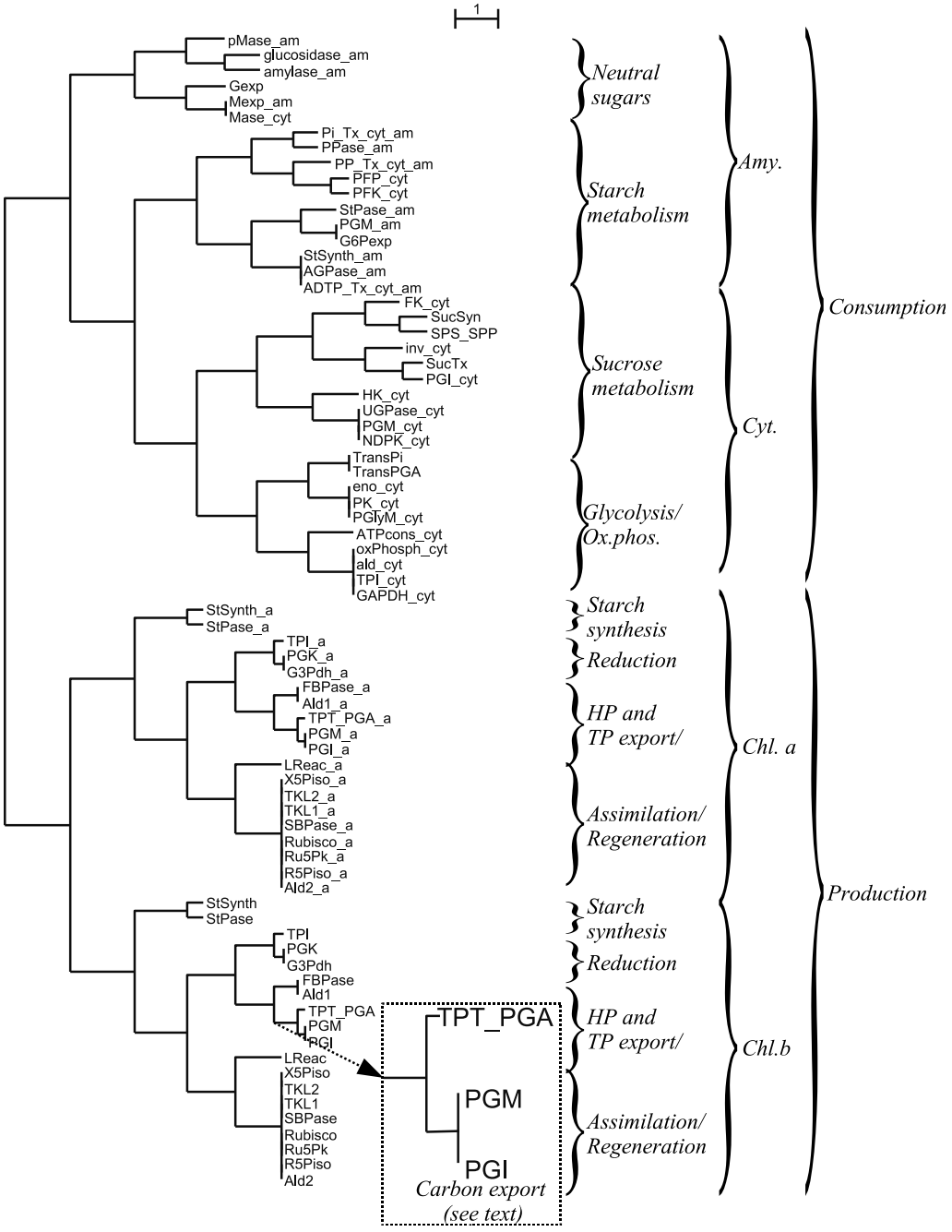


Figure 5: Reaction tree for the *photo* model, showing a number of identifiable, hierarchical and predictable modules. The scale bar indicates $\theta^{\boldsymbol{K}}_{xy} = 1$ radian. Abbreviations in module annotations : HP - hexose phosphate, TP - triose phosphate, Amy. - amyloplast, Cyt. - cytosol, Chl. - chloroplast. See supplementary material and cited references for full details of the model.

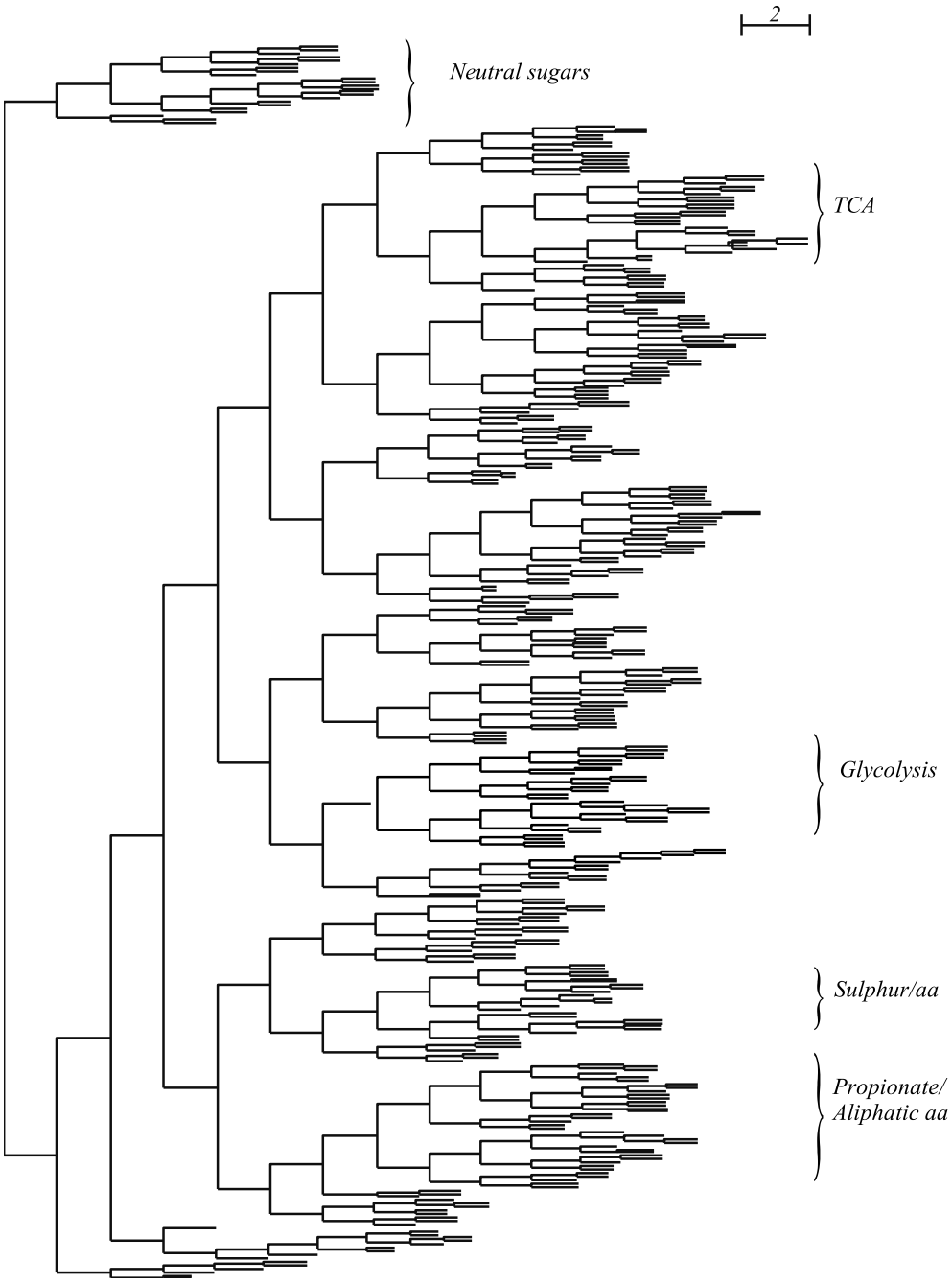


Figure 6: Reaction tree for the fully condensed *sco* model as described in section 3.2. Reaction IDs have been removed for clarity. The tree appears to show a reasonably well identified module involved in neutral sugar metabolism. Although other clusters of reactions are less clearly separated, clusters corresponding to the tricarboxylic acid cycle (TCA) and glycolysis were readily identifiable. The scale bar represents a difference of $\theta_{xy}^{\boldsymbol{K}} = 2$ radians.

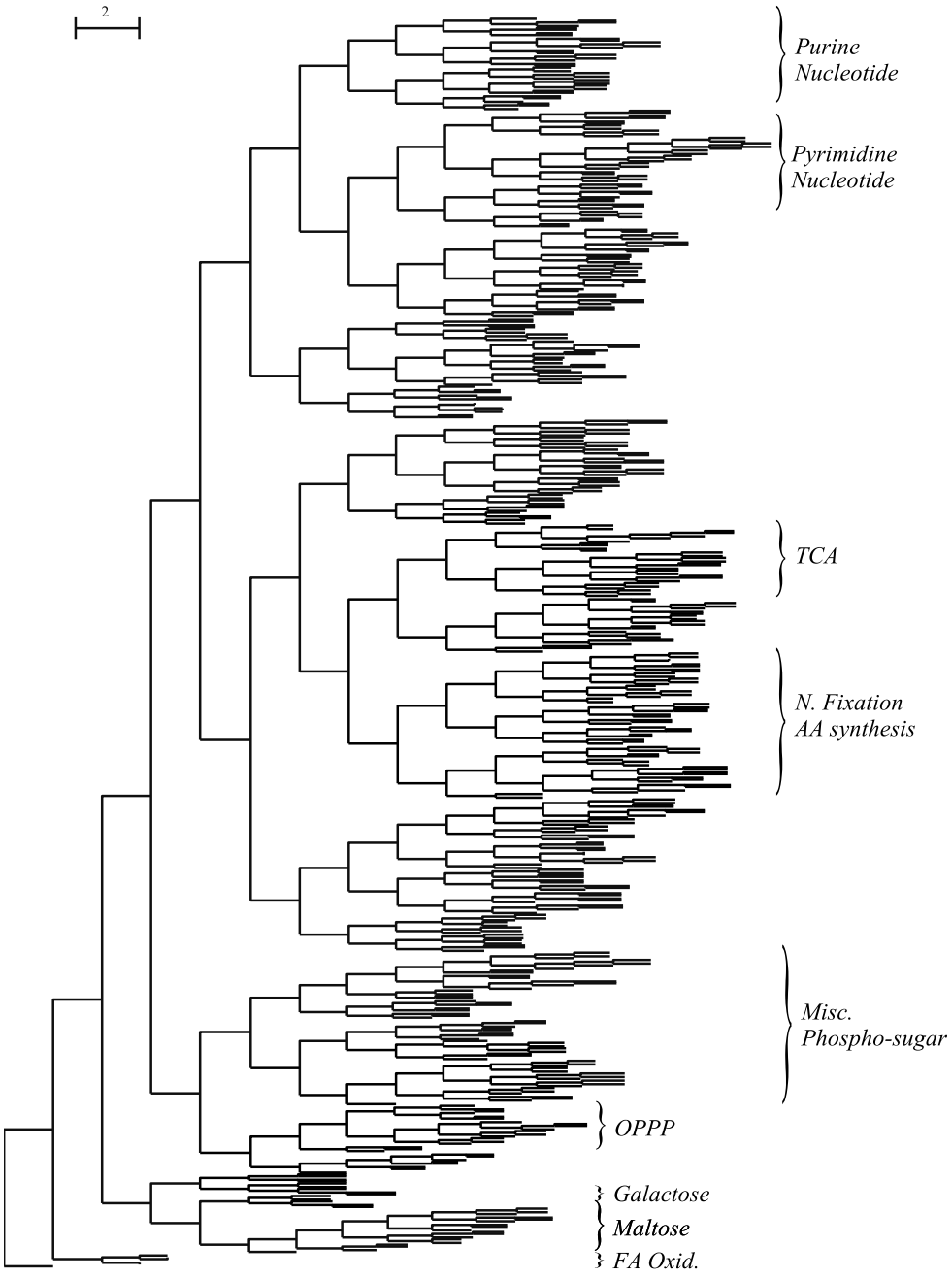


Figure 7: Reaction tree for the fully condensed *eco* model with identifiable clusters labeled. The most clearly defined were the reactions of the oxidative pentose phosphate pathway (OPPP - both the oxidative and non-oxidative parts) followed by the TCA cycle. Reactions usually associated with glycolysis did not produce a well defined cluster, but were largely contained in the cluster labeled "Misc. Phospho sugar" The scale bar represents a difference of $\theta_{xy}^{\boldsymbol{K}} = 2$ radians.

PCs using the Linux operating system and ScrumPy - metabolic modelling in Python - software (Poolman 2006). It seems unlikely that the overall size of metabolic models is going to increase greatly in the future and so computational overhead is not a problem for this form of analysis.

5.2 Comparison to related work

To the best of our knowledge, no other authors have described a null-space based approach to deriving a general, quantitative, relationship between reactions in a system, although three extreme relationships have been previously described: the dead reactions (Schuster and Schuster 1991), enzyme subsets (Pfeiffer et al. 1999) and disconnected subsystems (Schuster and Schuster 1992), and this provides a starting point for the work described here.

There have been a number of other efforts with conceptually similar goals based either on linear programming or on graph theoretic approaches. Burgard et al. (2004) generate a single level hierarchy, grouping reactions according to whether minimum and maximum steady-state flux ratios between pairs of reactions are zero, non-zero finite or infinite, based on a linear programming approach. This provides for the identification for the three extreme relationships described above; however the relationship between pairs of reactions outside of these extreme cases as described by Burgard, and the reaction correlation coefficient presented here is not certain at present. It would appear that Burgard's approach yields much more qualitative relationships between reactions than those proposed here.

Ravasz et al. (2002) used a graph theoretic approach to investigate modularity of metabolic networks. In their work, a metabolic system is represented as an (undirected) graph in which metabolites are nodes, which are connected to one another if they share a common reaction. All edges in the graph connect exactly two nodes and therefore more than one, apparently independent, edge is needed to represent reactions with more than one substrate or product. The relationship between such graph representations and metabolic networks as described above is not particularly obvious, but it is clear that the axiomatic necessity that edges connect exactly two nodes destroys much of the inherent network structure (other than networks comprised solely of isomerisations).

From such a graph-based perspective they then defined a "topological overlap" metric based on the number of common metabolites to which a pair of nodes are connected. This was then used to generate a similarity matrix to which a hierarchical cluster analysis was applied. This yielded clusters of chemically similar metabolites, but as the analysis was performed in terms of metabolites it yielded no information about the organisation of

that which, from our point of view, is rather more important: the reactions that interconvert them.

In an effort that at least partially ameliorates the problems of graph theoretic approaches, Gagneur et al. (2003) used a bipartite graph representation of a metabolic network in which nodes represent metabolites or reactions and edges connect exactly one metabolite to one reaction. Nodes in this graph are then successively amalgamated to build a (non-binary) tree without the need for an intermediate difference matrix. Nodes in the resulting tree consist of mixtures of reactions and metabolites. Although a bipartite graph representation of metabolic networks appears to be more solidly founded in reality than a unigraph, the resulting tree is rather harder to interpret then a simple binary tree. Also, when Gagneur's method was applied to the *photo* model described above, the known modular structure was not recovered.

Graph theoretic approaches are nonetheless potentially useful. As they do not depend upon the steady-state assumption there is no requirement for knowledge of external metabolites. They can therefore be usefully applied in situations were this knowledge is absent or uncertain, or in systems with no net mass conversion. Graph theoretic and null space based approaches should therefore probably be regarded as complementary, rather than exclusive, alternatives.

However, graph theoretic analyses must always proceed as a sequence of local steps: a node is chosen, certain of its properties investigated, in terms of its immediate neighbors, and then a neighbor is chosen upon which to repeat the sequence. In contrast, linear algebra approaches in general, and null-space approaches in particular, simultaneously relate every component of the system to every other component, treating the system "as a whole" rather than a collection of related components.

An approach that † has some conceptual similarity to that presented here has recently been described by Sariyara et al. (2006). These authors subjected a structural model of *E. coli* metabolism to a set of randomised input flux values and then used flux balance analysis with linear programming to assign flux values to the remaining reactions. Amongst other things, they used these values to calculate a flux correlation matrix (available as supplementary material). The distribution of the log of the absolute value of these correlation coefficients (not shown) was markedly different from those presented here, being clearly bi-modal, with peaks at ≈ 1 and ≈ 0.1. A detailed comparison of the two approaches would certainly be interesting, but is beyond the scope of the this paper. †2.19

†

†2.18 start

Another approach to understanding the modularity of metabolic networks, and, in common with the current work, also based on the steady-

state assumption rather than a graph-theoretic basis, has been from within the framework of metabolic control analysis (MCA) (e.g. Brown et al. 1990; Schuster et al. 1993; Rohwer et al. 1996). The fundamental theoretical difference is that these approaches all depend upon a knowledge (at least qualitatively) of reaction elasticities (sensitivity of reaction rates to metabolite concentrations). A major practical difference is that these efforts were directed to using an assumed modular structure of a metabolic network to investigate and describe its control properties, rather than to determine network structure in its own right. Furthermore these approaches do not readily lend themselves to the complete hierarchical decomposition of arbitrary networks that we have demonstrated here.

Nonetheless, understanding the control of flux is, of course, of great importance and the development of structural analysis of metabolic networks can, not unreasonably, be seen as a necessary step along the path to this greater ultimate goal. Furthermore, from the definition of reaction correlation coefficient given here, increasing the activity of a given reaction would be expected to have a larger impact on the flux carried by those reactions with higher flux correlation coefficients. Conceptually, this is extremely similar to the flux controll coefficient of MCA, and although identifying the mathematical relationship between the two appears potentially interesting, it must remain a goal of future endeavour.

† †2.18 end.

5.3 Applications of metabolic trees and reaction correlation coefficients

5.3.1 Identification of independent subnetworks

Although it might be intuitively predicted that any realistic metabolic network consists of a single connected system, it may well be the case that large models, especially those generated automatically from databases do not (e.g. Poolman et al. 2006). Regardless of the biological interpretation or expectation of such subnetworks, their presence, or otherwise, represents a fundamental property of a network. It has previously been proposed that the presence of such disconnected subsystems can be identified by the block diagonalisation of $\boldsymbol{K}$ (Schuster and Schuster 1992; Heinrich and Schuster 1996). Here, such a subnetwork can be readily identified in a metabolic tree: if the distance from a subtree to its parent is equal to $\pi/2$, it represents an independent subnetwork. In other words reaction correlation coefficients between a reaction internal to the subtree and any reaction external to the subtree are zero.

5.3.2 Identification of dead reactions

Dead reactions can be identified by a corresponding zero row vector in $\boldsymbol{K}$ (Schuster and Schuster 1991); however this is potentially only a subset of dead reactions as some reactions might be embedded in internal cycles (but not capable of sustaining flux at steady-state), and hence have non-zero row vectors in $\boldsymbol{K}$. Sets of such reactions can be identified in a metabolic tree if a node corresponding to an independent sub-network, as defined above, contains no leaf nodes representing transport reactions (i.e. reactions inter-converting internal and external species).

An alternative way to express this is to say that a reaction with a corresponding non-zero vector in $\boldsymbol{K}$ is nonetheless dead if the reaction correlation coefficient between it and all transport reactions is zero. Such a reaction will occur in elementary modes, but all such modes will be internal cycles and hence non-flux carrying. This is somewhat simpler than the method proposed by Burgard et al. (2004).

5.3.3 Characterisation of network modularity

The logic of the algorithm proposed here, and the results obtained when applying it to a model of known modular structure, demonstrate that it is indeed capable of recovering hierarchically nested modules, where a module is defined as a group of reactions whose steady state flux is more closely correlated to other reactions in the model than to those outside it.

However, the results obtained when applying the algorithm to the *sco* and *eco* models are somewhat equivocal. Fig. 6 shows only two easily distinguishable modules, and the *eco* metabolic tree, fig. 7, showed less obvious structure, although both trees demonstrate a leaf ordering that tends to place biochemically related reactions (and therefore metabolites), close to one another.

Nonetheless, the question as to the extent to which (or indeed whether or not) these two models are modular must be addressed. Consideration of the distributions of $\log(\phi)$ (Figs. 3 and 4) suggest that the most strongly modular model was the split *photo* model, in effect comprised of three entirely independent models. All non-zero values of ϕ in this model therefore reflect intra-modular correlations and these had a median value of $\phi \approx 10^{-0.5}$. The (intact) *photo* model in which values of ϕ were a mixture of intra- and inter-modular correlations showed a bi-modal distribution with peaks at $\phi \approx 10^{-0.5}$ and $\phi \approx 10^{-2.3}$. The conclusion to be drawn is that $\phi \leq\approx 10^{-2}$ is the representative value in a non-modular system.

The distribution of $\log(\phi)$ for the *eco* model shows a median value $\phi \approx$

$10^{-2.4}$ and for *sco* $\phi \approx 10^{-1.9}$, consistent with the conclusion suggested by inspection of the trees that while neither network can be described as highly modular, the *sco* model appears to be more modular than eco.

There would appear to be three possible, non-exclusive, explanations for this lack of observable modularity. Firstly, it is noteworthy that both of the genome scale models have a very high proportion of external metabolites, and one would expect this to have the effect of reducing possible correlations between reactions. There is evidence for this effect in the fact that the *eco* model, which appears the less modular has a higher proportion of external metabolites (61 % of total) than the *sco* model (37 %).

Secondly, there is the general problem in interpreting genome scale models in that the model contains all possible reactions that the organism is capable of catalysing. It would seem most unlikely that this represents any common biological reality, and when it becomes possible to construct models based on accurate expression or proteomic data, it is reasonable to predict that a more strongly modular pattern might be observed. This line of reasoning might, in the future be extended to a semi-quantitative position. The structural methods described here do not take actual flux observations into account. Were these available (derived independently from the techniques used here), it may transpire that a number of reactions carry a relatively insignificant flux and could therefore be discounted in the construction of the metabolic tree. However, if this does prove to be the case, modularity of a given metabolic system would then become dependent upon both the environmental/nutritional conditions to which it is subject, and to the physiological status of the organism. Under such circumstances, the utility of (metabolic) modules as a concept becomes suspect, as ideally the modularity of the system should be invariant.

The third possibility is that metabolic systems might genuinely be non (or only very weakly) modular. Modularity is a human concept originating in the engineering sciences that is an extremely powerful tool when it comes to imposing, by design, structure on a manufactured system. However, there is no law of nature which states that the structure of metabolic networks must adhere to good human engineering principles. To suggest otherwise is mere anthropomorphism.

In a well known and thought provoking paper, Lazebnik (2002) posed the question "Can a biologist fix a radio ?", and provided an entertaining and justifiable critique of an overdependence on a reductionist approach to the understanding of biological systems. Radios are, indisputably, the product of the human mind, and using nothing more than a hammer and a pair of eyes one can start to discern their modular construction. With the availability of slightly more sophisticated tools, the answer to the question is therefore

almost certainly 'Yes'. A more pertinent question might be "Can a radio technician explain metabolism ?".

5.3.4 Practical applications

Regardless of the degree of modularity of metabolic systems, the reaction correlation coefficient, ϕ, and metabolic trees derived from it, have a number of potential practical uses.

Reaction correlation coefficients describe the likely correlation between pairs of reactions, and metabolic trees have the effect of ordering all reactions accordingly. Thus, if an objective is to maximise flux in a particular reaction, reactions that are near neighbors in the reaction tree make rational candidates for manipulation.

A slightly more subtle application is the preferential optimisation of desirable pathways when alternative pathways exist for the generation of some product. Suppose an organism is capable of producing a product by utilising one of two substrates, but one of these is considered preferable to other. An optimisation can be effected by searching for knockouts that maximise the correlation of the product export step with the import step of the preferred substrate.

An exhaustive search for a single step, or two steps in combination, to be knocked out would be computationally feasible, but for greater numbers of steps it would be probably be necessary to use some form of evolutionary algorithm. If the objective function of such an algorithm was designed so as to maximise ϕ between product transport and desirable substrate transport, while simultaneously minimising the total number of reactions in the system, the end result would be to identify optimal elementary modes for the generation of product, without the need to first generate the entire set of elementary modes, the majority of which will be irrelevant to any specific task.

5.4 Conclusion

We have shown that, through analysis of the null-space of the stoichiometry matrix of a system, it is possible to identify a more complete set of relationships between steady-state reaction fluxes, in the form of reaction correlation coefficients, than has been previously described. Consideration of these allows the description of a number of system properties, and in particular allows the construction of a metabolic reaction tree.

Analysis of the metabolic reaction tree of a model with known modular structure successfully recovered those modules, but when applied to two

genome-scale models, the evidence for modularity was, at best, weak. It is possible that this apparent lack of modularity is, to some extent, an artifact of the models' definition, and that the *in vivo* systems are more modular than the models.

It may be that improved models, or refining the analysis described here to take into account expression and/or flux observations, would reveal a more strongly modular structure than observed here. However, it is also important to realise that the concept of modularity is a human construct, and that the definite of identification of modularity in a natural system is a prerequisite of its characterisation.

A Constancy of $\cos\theta$ in orthogonal kernels

Definition 1 *If we denote by* $\theta_{xy}^{\boldsymbol{A}}$ *the angle between row vectors* x *and* y *in matrix* $\boldsymbol{A}$*, then:*

$$\cos(\theta_{xy}^{\boldsymbol{A}}) = \frac{\boldsymbol{A}_x \boldsymbol{A}_y^T}{\sqrt{(\boldsymbol{A}_x \boldsymbol{A}_x^T)}\sqrt{(\boldsymbol{A}_y \boldsymbol{A}_y^T)}}$$

Thus for any pair of matrices of equal dimensions, $\boldsymbol{A}$ *and* $\boldsymbol{B}$*, if*

$$\boldsymbol{A}_x \boldsymbol{A}_y^T = \boldsymbol{B}_x \boldsymbol{B}_y^T$$

then

$$\cos(\theta_{xy}^{\boldsymbol{A}}) = \cos(\theta_{xy}^{\boldsymbol{B}}).$$

Let $\boldsymbol{N}$ be a stoichiometry matrix and let $\boldsymbol{A}$ and $\boldsymbol{B}$ † be matrices corresponding to two orthonormal bases of the null space of $\boldsymbol{N}$ (the columns are the basis elements). †2.20

Note : $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal matrices and therefore satisfy

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{B}^T \boldsymbol{B} = \boldsymbol{I}$$

Lemma 1 *Let* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be orthogonal matrices of equal dimensions, then*

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{B}\boldsymbol{B}^T \tag{1}$$

Proof : Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are both orthogonal matrices then $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}$ (i.e. their columns are orthonormal vectors):

$$\boldsymbol{A} = [a_1\, a_2\, \cdots\, a_n];\;\; \boldsymbol{B} = [b_1\, b_2\, \cdots\, b_n]$$

We can express each column vector of $\boldsymbol{B}$ as a linear combination of the column vectors of $\boldsymbol{A}$:

$$b_i = \alpha_{1i}a_1 + \alpha_{2i}a_2 + \cdots + \alpha_{ni}a_n \;\text{ for }\; i = 1,\, 2,\, \ldots\, n,$$

where $\alpha_{ki} = (a_k, b_i) = a_k^T b_i$. Thus, we can write

$$\boldsymbol{B} = \boldsymbol{A}\alpha, \quad \alpha = \boldsymbol{A}^T\boldsymbol{B} \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^T\boldsymbol{B}$$

Similarly, we can express each column vector of $\boldsymbol{A}$ as a linear combination of the column vectors of $\boldsymbol{B}$:

$$a_i = \beta_{1i}b_1 + \beta_{2i}b_2 + \cdots + \beta_{ni}b_n \;\text{ for }\; i = 1,\, 2,\, \ldots\, n,$$

where $\beta_{ki} = (b_k, a_i) = b_k^T a_i$. Thus, we can write that

$$\boldsymbol{A} = \boldsymbol{B}\beta \quad \beta = \boldsymbol{B}^T\boldsymbol{A} \quad \text{and} \quad \boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A}$$

Note that $\alpha_{ki} = (a_k, b_i) = (b_i, a_k) = \beta_{ik}$, which implies that $\alpha = \beta^T$. Both $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy

$$\boldsymbol{A}^T\boldsymbol{A} = (\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A})^T\boldsymbol{A} = (\boldsymbol{A}^T\boldsymbol{B})(\boldsymbol{B}^T\boldsymbol{A}) = \boldsymbol{I}, \tag{2}$$

$$\boldsymbol{B}^T\boldsymbol{B} = (\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{B})^T\boldsymbol{B} = (\boldsymbol{B}^T\boldsymbol{A})(\boldsymbol{A}^T\boldsymbol{B}) = \boldsymbol{I} \tag{3}$$

so that

$$(\boldsymbol{A}^T\boldsymbol{B})(\boldsymbol{B}^T\boldsymbol{A}) = (\boldsymbol{B}^T\boldsymbol{A})(\boldsymbol{A}^T\boldsymbol{B}) = \boldsymbol{I}$$

and

$$\boldsymbol{A}^T\boldsymbol{B} = (\boldsymbol{B}^T\boldsymbol{A})^{-1}.$$

Consequently,

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A}(\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{A})^T = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{A})(\boldsymbol{A}^T\boldsymbol{B})\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{I}\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{B}^T.$$

Theorem 1 *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be orthogonal matrices with equal dimensions, then*

$$\cos(\theta_{xy}^A) = \cos(\theta_{xy}^B)$$

for any x and y.

Proof: Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal matrices then by Lemma 1 $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{B}\boldsymbol{B}^T$. Consequently, $(\boldsymbol{A}\boldsymbol{A}^T)_{xy} = (\boldsymbol{B}\boldsymbol{B}^T)_{xy}$ for any x and y. However, $(\boldsymbol{A}\boldsymbol{A}^T)_{xy} = \boldsymbol{A}_x\boldsymbol{A}_y^T$ and therefore

$$\boldsymbol{A}_x\boldsymbol{A}_y^T = \boldsymbol{B}_x\boldsymbol{B}_y^T.$$

Thus

$$\cos(\theta_{xy}^A) = \cos(\theta_{xy}^B)$$

for any x and y.

B The Equivalence of r and ϕ

Assume a system with a stoichiometry matrix $\boldsymbol{N}(m \times n)$ having an orthonormal kernel, $\boldsymbol{K}(n \times d)$ and a matrix of random numbers, $\boldsymbol{R}(d \times s)$, whose elements are drawn from a distribution with mean μ and variance σ^2 where:

m	is the number of metabolites,
n	is the number of reactions,
d	is the dimension of the null space,
s	is arbitrarily large,
and	
θ_{xy}^K	is the angle between rows x and y of $\boldsymbol{K}$.

Define a matrix $\boldsymbol{V}(n \times s)$ such that

$$\boldsymbol{V} = \boldsymbol{K}\boldsymbol{R}$$

$\boldsymbol{V}$ will satisfy

$$\boldsymbol{N}\boldsymbol{V} = \boldsymbol{0}$$

i.e. any column of $\boldsymbol{V}$, $\boldsymbol{V}_j$, is a valid steady-state rate vector, in the following x and y are used exclusively to denote row indices and j is used to denote a column index. The sample correlation coefficient, $\hat{r}_{xy}$, between reactions x and y is defined

$$\hat{r}_{xy} = \frac{\sum_{j=1}^{s}(\boldsymbol{V}_{xj} - \overline{\boldsymbol{V}_x})(\boldsymbol{V}_{yj} - \overline{\boldsymbol{V}_y})}{\sqrt{\sum_{j=1}^{s}(\boldsymbol{V}_{xj} - \overline{\boldsymbol{V}_x})^2}\sqrt{\sum_{j=1}^{s}(\boldsymbol{V}_{yj} - \overline{\boldsymbol{V}_y})^2}} \tag{4}$$

Now, the j^{th} sample of reaction x is

$$\boldsymbol{V}_{xj} = \boldsymbol{K}_x\boldsymbol{R}_j$$

and thus the sample mean, of reaction x in $\boldsymbol{V}$ is

$$\overline{\boldsymbol{V}_x} = \frac{\sum_{j=1}^{s}\boldsymbol{K}_x\boldsymbol{R}_j}{s} = \boldsymbol{K}_x\hat{\boldsymbol{\mu}}$$

where $\hat{\boldsymbol{\mu}}$ is a column vector of length d of mean values, $\hat{\mu}$ of the elements of $\boldsymbol{R}$. Substituting these equivalences into equation (4), we obtain

$$\hat{r}_{xy} = \frac{\sum_{j=1}^{s}(\boldsymbol{K}_x\boldsymbol{R}_j - \boldsymbol{K}_x\hat{\boldsymbol{\mu}})(\boldsymbol{K}_y\boldsymbol{R}_j - \boldsymbol{K}_y\hat{\boldsymbol{\mu}})^T}{\sqrt{\sum_{j=1}^{s}(\boldsymbol{K}_x\boldsymbol{R}_j - \boldsymbol{K}_x\hat{\boldsymbol{\mu}})(\boldsymbol{K}_x\boldsymbol{R}_j - \boldsymbol{K}_x\hat{\boldsymbol{\mu}})^T}\sqrt{\sum_{j=1}^{s}(\boldsymbol{K}_y\boldsymbol{R}_j - \boldsymbol{K}_y\hat{\boldsymbol{\mu}})(\boldsymbol{K}_y\boldsymbol{R}_j - \boldsymbol{K}_y\hat{\boldsymbol{\mu}})^T}} \tag{5}$$

Since

$$(\boldsymbol{K}_x\boldsymbol{R}_j - \boldsymbol{K}_x\hat{\boldsymbol{\mu}})(\boldsymbol{K}_y\boldsymbol{R}_j - \boldsymbol{K}_y\hat{\boldsymbol{\mu}})^T = \boldsymbol{K}_x(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})^T\boldsymbol{K}_y^T$$

equation 5 can be rewritten as

$$\hat{r}_{xy} = \frac{\boldsymbol{K}_x \sum_{j=1}^{s}(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})^T\boldsymbol{K}_y^T}{\sqrt{(\boldsymbol{K}_x \sum_{j=1}^{s}(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})^T)\boldsymbol{K}_x^T)}\sqrt{(\boldsymbol{K}_y \sum_{j=1}^{s}(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})^T)\boldsymbol{K}_y^T)}} \tag{6}$$

If we now define the matrix $\hat{\boldsymbol{A}}(n \times n)$ as

$$\hat{\boldsymbol{A}} = \sum_{j=1}^{s}(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{R}_j - \hat{\boldsymbol{\mu}})^T$$

and substitute into equation (6) we obtain

$$\hat{r}_{xy} = \frac{\boldsymbol{K}_x\hat{\boldsymbol{A}}\boldsymbol{K}_y^T}{\sqrt{(\boldsymbol{K}_x\hat{\boldsymbol{A}}\boldsymbol{K}_x^T)}\sqrt{(\boldsymbol{K}_y\hat{\boldsymbol{A}}\boldsymbol{K}_y^T)}} \tag{7}$$

Now, as s becomes large the leading diagonal of $\hat{\boldsymbol{A}}$ approaches σ^2, and all other elements approach 0

$$\lim_{s\to\infty} \hat{\boldsymbol{A}} = \sigma^2\boldsymbol{I}$$

Therefore

$$\lim_{s\to\infty} \hat{r}_{xy} = r_{xy} = \frac{\boldsymbol{K}_x\boldsymbol{K}_y^T}{\sqrt{(\boldsymbol{K}_x\boldsymbol{K}_x^T)}\sqrt{(\boldsymbol{K}_y\boldsymbol{K}_y^T)}} = \cos(\theta_{xy}^K) \tag{8}$$

as required.

C $\phi = 0$ Implies a Stochiometrically Disconnected System

Consider a metabolic system comprised of r reactions and in which reactions i and j both carry flux. Let the elementary modes of the system be represented by the matrix $\boldsymbol{E}$ of dimension $(r \times s)$ where s is the number of elementary modes. For convenience, assume that $\boldsymbol{E}$ is arranged such that the first two rows correspond to i and j respectively, so $\boldsymbol{E}$ can be written as

$$\boldsymbol{E} = \begin{bmatrix} e_{i1} & e_{i2} & e_{i3} & \dots \\ e_{j1} & e_{j2} & e_{j3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Any valid flux vector, $\boldsymbol{v}$, can be written as:

$$\boldsymbol{v} = \boldsymbol{E}\boldsymbol{w}\,, \tag{9}$$

where $\boldsymbol{w}$ is a column vector of dimension s. The values of elements of $\boldsymbol{v}$ are thus

$$\boldsymbol{v} = \begin{bmatrix} e_{i1}w_1 + e_{i2}w_2 + e_{i3}w_3 + \dots \\ e_{j1}w_1 + e_{j2}w_2 + e_{j3}w_3 + \dots \\ \vdots \end{bmatrix}. \tag{10}$$

Lemma 2 *If there is no elementary mode which utilises both reactions i and j, then the following statements are equivalent:*

(a) There is no column, x, of the matrix $\boldsymbol{E}$ such that $e_{ix} \neq 0$ and $e_{jx} \neq 0$.

(b) The population correlation coefficient $r_{ij} = 0$.

(c) The reaction correlation coefficient $\phi_{ij} = 0$.

Proof :
$(a) \Rightarrow (b) \Rightarrow (c)$. If there is no elementary mode which utilises both reactions i and j, then there is no column, x, in $\boldsymbol{E}$ such that $e_{ix} \neq 0$ and $e_{jx} \neq 0$. Consequently, $\boldsymbol{E}$ has the following structure

$$\boldsymbol{E} = \begin{bmatrix} e_{i1} & 0 & 0 & \dots \\ 0 & e_{j2} & e_{j3} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

and likewise,

$$\boldsymbol{v} = \begin{bmatrix} e_{i1}w_1 + 0w_2 + 0w_3 + \dots \\ 0w_1 + e_{j2}w_2 + e_{j3}w_3 + \dots \\ \vdots \end{bmatrix}.$$

Thus, the summations that form the elements $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ contain no common elements from $\boldsymbol{w}$. If $\boldsymbol{v}$ are to be repeatedly sampled according to equation (9), using different instances of $\boldsymbol{w}$ with elements of random value, then the sample correlation coefficient between reactions i and j, $\hat{r}_{ij}$, will tend to zero as the sample size increases, and the population correlation coefficient, r_{ij}, will be equal to zero. As has been shown in appendix B, the reaction correlation coefficient, ϕ_{ij}, is equal to (Pearson's) population correlation coefficient for all possible instances of $\boldsymbol{v}$ in a given system, and therefore $\phi_{ij} = 0$.

$(c) \Rightarrow (b) \Rightarrow (a)$ Assume to the contrary that the reaction correlation coefficient $\phi_{ij} \neq 0$ then the population correlation coefficient $r_{ij} \neq 0$. Since $\lim_{s\to\infty} \hat{r}_{ij} = r_{ij}$ and $0 \leq \hat{r}_{ij} \leq 1$ then the sample correlation coefficient $\hat{r}_{ij} \neq 0$ provided the sample size is sufficiently large. Consequently, in the calculation of $\boldsymbol{v}$ in equation (10) the summations representing the values of $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ contain common elements from $\boldsymbol{w}$. Thus, there exists at least one column, x, of the matrix $\boldsymbol{E}$ such that $e_{ix} \neq 0$ and $e_{jx} \neq 0$. Therefore, there exists at least one elementary that utilises reactions i and j contradicting our assumption that there is no elementary mode which utilises both reactions i and j.

Theorem 2 *If $\phi_{ij} = 0$ then reactions i and j are members of stoichiometrically disconnected subsystems.*

Proof : If $\phi_{ij} = 0$ then, by Lemma 2, there is no elementary mode which utilises both reactions i and j. Hence, $\boldsymbol{E}$ is block diagonalisable with i and j in different blocks. Let $\boldsymbol{K}^I$ be a kernel of the stoichiometry matrix of the form

$$\boldsymbol{K}^I = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{K}' \end{bmatrix}.$$

The columns of $\boldsymbol{K}^I$ are a subset of the columns of $\boldsymbol{E}$, and every row in $\boldsymbol{K}^I$ corresponding to a reaction capable of carrying flux at steady-state must contain at least one non-zero element. Since there is no column, x, of the matrix $\boldsymbol{E}$ such that $e_{ix} \neq 0$ and $e_{jx} \neq 0$, $\boldsymbol{K}^I$ must be block diagonalisable with i and j in different blocks. The reactions i and j are consequently members of stoichiometrically disconnected subsystems.

D The WPGMA Algorithm

The WPGMA algorithm repeatedly removes pairs of rows and columns corresponding to nearest neighbours from a difference matrix, replacing them with a single new row and column corresponding to the subtree containing the two neighbours. The tree is built by agglomeration. Given a difference matrix containing a comparison of n items, the algorithm starts with n trees, each containing a single leaf node, at each iteration two trees are merged, when the algorithm terminates there is a single tree with n leaf nodes. In the following description, $\mathbf{\Delta}$ is assumed be a data-structure containing the entries of the difference matrix, that can be indexed by arbitrary strings. These string indices are used to represent trees.

1: **while** Dimension $\mathbf{\Delta} \neq (1,1)$ **do**
2: $i, j =$ NearestNeighbours($\mathbf{\Delta}$)
3: $\delta = \mathbf{\Delta}_{ij}$
4: Create label T, = '$(i : \delta, j : \delta)$'
5: Calculate a vector of difference values, $\boldsymbol{\delta}$, between T and other items in $\mathbf{\Delta}$
6: Remove rows and columns i and j from $\mathbf{\Delta}$
7: Generate a row and column in $\mathbf{\Delta}$ with label T and values $\boldsymbol{\delta}$

At each iteration two rows and columns are removed from $\mathbf{\Delta}$ and one is created, thus after n iterations the algorithm will terminate. The elements of $\boldsymbol{\delta}$ in step 5 are calculated as

$$\boldsymbol{\delta} = \left[\frac{\mathbf{\Delta}_{ik}+\mathbf{\Delta}_{jk}}{2}\right] \quad \text{for} \quad k = 1 \ldots n, k \neq i, k \neq j$$

Where n is the current dimension of $\mathbf{\Delta}$. The dimension of $\boldsymbol{\delta}$ is thus $n-2$, and therefore in step 7 an extra element is created such that $\mathbf{\Delta}_{TT} = 0$, where T is the row/column label created in step 4. Hence at the end of step 7, the dimension of $\mathbf{\Delta}$ is $(n-1, n-1)$.

The progress of the algorithm may be illustrated by considering an initial difference matrix:

$$\mathbf{\Delta} = \begin{array}{c|cccc} & a & b & c & d \\ a & 0.0 & 0.4 & 0.6 & 0.7 \\ b & 0.4 & 0.0 & 0.8 & 0.9 \\ c & 0.6 & 0.8 & 0.0 & 0.5 \\ d & 0.7 & 0.9 & 0.5 & 0.0 \end{array}$$

The nearest neighbours are a and b with $\delta = 0.4$, so in step 4:

$$T = \text{'}(a : 0.4, b : 0.4)\text{'}$$

denoting a tree with two leaf nodes whose distance to their common parent is 0.4. The new difference vector, $\boldsymbol{\delta}$ in step 5, representing the distances between the new tree and the items in $\boldsymbol{\Delta}$ is

$$\boldsymbol{\delta} = \left[\begin{array}{ccc} 0 & \frac{\Delta_{ac}+\Delta_{bc}}{2} & \frac{\Delta_{ad}+\Delta_{bd}}{2} \end{array}\right] = \left[\begin{array}{ccc} 0 & \frac{0.6+0.8}{2} & \frac{0.7+0.9}{2} \end{array}\right] = \left[\begin{array}{ccc} 0 & 0.7 & 0.8 \end{array}\right]$$

Removing rows a and b and columns a and b (step 6) and generating a new row and column with label '$(a : 0.4, b : 0.4)$' and values $\left[\begin{array}{ccc} 0 & 0.7 & 0.8 \end{array}\right]$ (step 7) we obtain:

$$\boldsymbol{\Delta} = \begin{array}{c|ccc} (a : 0.4, b : 0.4) & 0.0 & 0.7 & 0.8 \\ c & 0.7 & 0.0 & 0.5 \\ d & 0.8 & 0.5 & 0.0 \end{array}$$

Column labels (which are identical to row labels) are omitted for clarity.

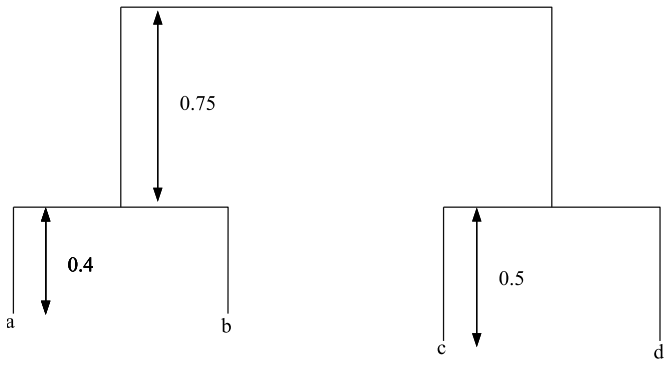
So at the beginning of the second iteration, the nearest neighbours are c and d and at the end of the second iteration:

$$\boldsymbol{\Delta} = \begin{array}{c|cc} \text{(a:0.4,b:0.4)} & 0.0 & 0.75 \\ \text{(d:0.5,c:0.5)} & 0.75 & 0.0 \end{array}$$

And at the end of the third and final iteration:

$$\boldsymbol{\Delta} = \begin{array}{c|c} ((a : 0.4, b : 0.4) : 0.75, (d : 0.5, c : 0.5) : 0.75) & 0.0 \end{array}$$

With the single row label corresponding to the tree:



A potential problem of the algorithm is that of ambiguity (or non-uniqueness (Morgan and Ray 1995)). This occurs if, in step 2 there is more than one pair of items that could be treated as nearest neighbours. In the implementation used here, ambiguities are resolved by selecting the pair that generates a new tree with the greatest number of child nodes. If this fails to resolve the ambiguity, it is resolved by lexicographic comparison of the trees involved. As long as row/column labels in $\boldsymbol{\Delta}$ are unique, this ensures a unique tree is generated.

References

Assmus, H., 2005. Modelling Carbohydrate Metabolism in Potato Tuber Cell. Ph.D. thesis, Oxford Brookes University.

Bonde, B., 2006. Metabolism and Bioinformatics: The Relationship Between Metabolism and Genome Structure. Ph.D. thesis, Oxford Brookes University.

Borodina, I., Krabben, P., Nielsen, J., 2005. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. Genome Res. 15, 820–829.

Brown, G., Hafner, R. P., Brand, D., 1990. A 'top-down' approach to the determination of control coeffecients in metabolic control theory. Eur. J. Biochem. 188, 321–325.

Burgard, A. P., Nikolaev, E. V., Schilling, C. H., Maranas, C. D., 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res. 14, 301–312.

Gagneur, J., Jackson, D. B., Casari, G., 2003. Hierarchical analysis of dependency in metabolic networks. Bioinformatics 19, 1027–1234.

Heinrich, R., Schuster, S., 1996. The Regulation of Cellular Systems, chapter 3. Chapman and Hall, London, 75–111.

Klamt, S., Schuster, S., Gilles, E. D., 2002. Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism of in purple nonsulphur metabolism. Biotechnol. Bioeng. 77, 734–750.

Klamt, S., Stelling, J., 2002. Combinatorial complexity of pathway analysis in metabolic networks. Molecular Biology Reports 29, 233–236.

Lance, G. N., Williams, W. T., 1967. A general theory of classificatory sorting strategies 1. Hierarchical systems. Comp. J. 9, 373–380.

Lazebnik, Y., 2002. Can a biologist fix a radio? - Or, what I learned while studying apoptosis. Cancer Cell 2, 179–182.

Lemke, N., Herédia, F., Barcellos, C. K., dos Reis, A. N., Mombach, J. C. M., 2004. Essentiality and damage in metabolic networks. Bioinformatics 20, 115 119.

Morgan, B. J. T., Ray, A. P. G., 1995. Non-uniqueness and Inversions in Cluster Analysis. Applied Statistics 44, 117–34

Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., Palsson, B. O., 2003. Metabolic pathways in the post-genomic era. Trends Biochem. Sci. 28, 250–258.

Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J., Montero, F., Schuster, S., 1999. Metatool: for studying metabolic networks. Bioinformatics 15, 251–257.

Poolman, M. G., 2006. ScrumPy - metabolic modelling with Python. IEE Proc. Sys. Biol. 153, 375–378.

Poolman, M. G., Bonde, B. K., Gevorgyan, A., Patel, H. H., Fell, D. A., 2006. Challenges to be faced in the reconstruction of metabolic networks from public databases. IEE Proc. Sys. Biol. 153, 379–384.

Poolman, M. G., Fell, D. A., Raines, C. A., 2003. Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. Eur. J. Biochem 270, 430–439.

Poolman, M. G., Ölcer, H., Lloyd, J. C., Raines, C. A., Fell, D. A., 2001. Computer modelling and experimental evidence for two steady states in the photosynthetic calvin cycle. Eur. J. Biochem. 268, 2810–2816.

Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabási, A.-L., 2002. Hierarchical Organization of Modularity in Metabolic Networks. Science 297, 1551–5.

Reed, J. L., Palsson, B. O., 2004. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res. 14, 1797–1805.

Reed, J. L., Vo, T. D., Schilling, C. H., Palsson, B. O., 2003. An expanded genome-scale model of *E. coli* K-12 (iJR904 GSM/GPR). Genome Biology 4, R54:1–12.

Rohwer, J., Schuster, S., Westerhoff, H., 1996. How to recognize monofunctional units in a metabolic system. J. Theor. Biol. 179, 213–28.

Sariyara, B., Perkb, S., Arkmanc, U., Hortaçsu, A., 2006. Monte Carlo sampling and principal component analysis of flux distributions yield topological and modular information on metabolic networks. J. Theor. Biol. 242, 389–400.

Schilling, C., Edwards, J., Palsson, B., 1999. Towards metabolic phenomics: analysis of genomic data using flux balances. Biotechnol. Prog. 15, 288–295

Schilling, C., Letscher, D., Palsson, B., 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway oriented perspective. J. Theor. Biol. 203, 229–248.

Schuster, S., Kahn, D., Westerhoff, H., 1993. Modular analysis of the control of complex metabolic pathways. Biophys. Chem. 48, 1–17.

Schuster, S., Klamt, S., Weckwerth, W., Moldenhauer, M., Pfeiffer, T., 2002. Use of network analysis of metabolic systems in bioengineering. Bioproc. Biosys. Eng. 24, 363–373.

Schuster, S., Schuster, R., 1991. Detecting strictly detailed balanced subnetworks in open chemical-reaction networks. J. Math. Chem. 6, 17–40.

Schuster, S., Schuster, R., 1992. Decomposition of biochemical reaction systems according to flux control insensibility. J. Chim. Phys. 89, 1887–1910.