S4 Comparison to FBA based methods

Here, we compare the results of our extension algorithm to analogous methods based on FBA[1,2]. For the system of differential equations

$$\frac{dc}{dt} = N \cdot v, \tag{1}$$

where c is the vector of metabolite concentrations, v the vector of fluxes through the reactions and N is the stoichiometric matrix of the system, FBA basically identifies feasible flux districutions v which fulfill certain constraints. In order to use FBA for the prediction of possible extensions to a metabolic draft network, a solution to (Eq. 1) has to be found minimizing the number of reactions which are carrying a flux and are not part of the draft. Such approaches have been described in[3] and[4]. Here we summarize the ideas of these works and use the following optimization problem:

$$\begin{aligned}
\text{minimize:} \quad & \sum \gamma_i && && (2)\\
\text{subject to:} \quad & N \cdot v = s && && (3)\\
& -\infty \leq s_i \leq \infty && \forall \text{ substrates } i && (4)\\
& 1 \leq s_k && \forall \text{ target metabolites } k && (5)\\
& 0 \leq s_j && \forall \text{ other metabolites } j && (6)\\
& -v_{max} \leq v_m \leq v_{max} && \forall \text{ draft reactions } m && (7)\\
& -v_{max} \cdot \gamma_n \leq v_n \leq v_{max} \cdot \gamma_n && \forall \text{ embedding reactions } n && (8)\\
& 0 \leq v_l && \forall \text{ irreversible reactions } l && (9)\\
& \gamma_n \in \{0,1\}. && && (10)
\end{aligned}$$

Equation 2 and 8 ensure a minimal utilization of embedding reactions in order to produce the target from the substrates. $\gamma_i = 1$ indicates the utilization of reaction i. The set of reactions $E = \{i|\gamma_i = 1\}$ represents a possible extension to the draft network.

We applied the above optimization procedure to the 400 randomly reduced *E.coli* networks as described in the main text. For each network one extension has been calculated using the program `lpsolve` (http://sourceforge.net/projects/lpsolve). Most apparently, the calculation of the FBA based extensions is considerably slower than our method. The FBA approach took nine hours on a normal single core desktop computer while the scope based approach took only a few seconds. In order to find solutions for all 400 networks in reasonable time, we limited the solution time of `lpsolve` to 5 minutes. With this time limit, in 375 out of the 400 cases the program could find a valid solution, even though its minimality could not be assured.

To compare the extensions obtained from the two different methods we apply the Jaccard coefficient $J = \frac{|A \cap B|}{|A \cup B|}$ which measures the similarity of two sets A and B. J is 1 if the two sets are identical and 0 for two disjoint sets. When comparing the FBA based extension with the most similar (with respect to J) scope based extension, the average Jaccard coefficient for the 400 extended networks is $\bar{J} = 0.63$.

It turns out that in the majority of cases the FBA based extensions are smaller than the extensions predicted by our approach. This is not surprising for two reasons: First, the FBA based approach is designed to find the global minimum with respect to numbers of reactions. In contrast, our approach finds minimal solutions with a preferential inclusion of those reactions for which enzymes are encoded in the genome with a high probability, but not necessarily the smallest possible extension. Second, the scope occasionally underestimates the number of producible metabolites if there exist cyclic dependencies such as in the following two-step production of metabolite B from metabolite A:

$$A + X \rightarrow Y$$
$$Y \rightarrow X + B$$

It has been argued in[5] that in this case metabolite B is in fact only producible if metabolites X or Y are replenished by other reactions. Otherwise, the strict conservation of the sum of the molecules of X and Y imply that due to the increase of cellular volume these metabolites eventually reach zero concentration. Thus, the criterion for the producibility of a metabolite based on FBA tends to overestimate the set of producible metabolites while the scope tends to underestimate this set. In fact, in the case of erroneous reaction stoichiometries, this difference is even stronger pronounced and results in the fact that our network expansion based approach is considerably more robust against such inconsistencies than the FBA based approaches (see below).
In order to check to what extent the FBA based extensions are just subsets of the scope based extensions, we defined a unsymmetrical Jaccard coefficient $J_{FBA} = \frac{|E_{FBA} \cap E_{Scope}|}{|E_{FBA}|}$ which is normalized with respect to the size of the FBA extension $|E_{FBA}|$. The average over all 400 networks is $\overline{J_{FBA}} = 0.89$. Hence, the two methods yield similar results, while the smaller FBA based extensions are to a large extent contained in the scope based extensions.

S4.1 Sensitivity of FBA based approaches against erroneous reaction stoichiometries

For any flux balance calculations, it is extremely important that all reaction stoichiometries are balanced. If there are inaccuracies as are often found in reactions retrieved from databases[6,7], it is possible that fluxes formally exist which suggest that metabolites may be created out of nothing. A simple example of an imbalanced reaction yielding absurd production fluxes is

$$\text{RNA} \rightleftharpoons \text{RNA} + \text{Nucleotide-5-phosphate}. \tag{11}$$

In contrast to FBA, for the method of network expansion, such an inconsistency is unproblematic since 'RNA' is not present. While such particularly simple examples are easy to identify, more complex cycles that have a similar overall effect but consist of a larger number of reactions, are very difficult to eradicate in an automated fashion.
As argued above, cyclic dependencies lead to an overestimation of the set of producible compounds when FBA is applied. For erroneous stoichiometries this fact is pronounced to an absurd extreme such that several metabolites are considered producible even if no nutrients are available. In contrast, such cycles lead to an underestimation of the set of producible metabolites when the method of network expansion is used. For example, an empty nutrient set will by definition always result in an empty set of producible metabolites. As a consequence, our method is considerably more robust against stoichiometric inaccuracies than FBA based approaches.

References

[1] J. S. Edwards and B. O. Palsson, *BMC Bioinformatics*, 2000, **1**, 1.

[2] K. J. Kauffman, P. Prakash and J. S. Edwards, *Curr Opin Biotechnol*, 2003, **14**, 491–496.

[3] J. L. Reed, T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong and B. O. Palsson, *Proc Natl Acad Sci U S A*, 2006, **103**, 17480–17484.

[4] V. Satish Kumar, M. Dasika and C. Maranas, *BMC Bioinformatics*, 2007, **8**, 212.

[5] K. Kruse and O. Ebenhöh, *Genome Informatics*, 2008, **20**, 91–101.

[6] M. G. Poolman, B. K. Bonde, A. Gevorgyan, H. H. Patel and D. A. Fell, *Syst Biol (Stevenage)*, 2006, **153**, 379–384.

[7] A. Gevorgyan, M. G. Poolman and D. A. Fell, *Bioinformatics*, 2008, **24**, 2245–2251.