Supplements

Inference of Cancer Mechanisms through Computational Systems Analysis

Zhen Qi and Eberhard O. Voit

Mathematical model of purine metabolism in human. A published kinetic model of human purine metabolism was used as a computational platform; it consists of 16 ordinary differential equations with 37 fluxes [1, 2]. In the literature, the fluxes were formulated either as traditional Michaelis-Menten kinetics or as power law functions under the tenets of Biochemical Systems Theory [3]. Here we chose the latter. Many parameters were obtained from experimental and clinical data in humans, and the remaining values were estimated using biological constraints, such as the ratio of adenine and guanine in nucleic acids, which is approximately 3/2, or the fact that normal subjects excrete about 420 mg per day of UA in urine. A detailed analysis of the steady-state properties of the mathematical model demonstrated that the steady state is stable and robust. Analysis also showed that the model is not sensitive to parameter changes. Simulations of normal and pathological perturbations of purine metabolism yielded consistent results with some representative biochemical and clinical observations. All these analyses are described in great detail in the literature [3].

Implementation of enzyme activities altered by cancer. Weber discovered several changes in the enzyme activities of purine metabolism in human renal carcinoma cells [4]. The affected enzymes and their fold changes compared to normal kidney cells are listed in Table S4. These alterations are expected to result in changed metabolite levels between normal human cells and human renal cell carcinoma, which were computed with the mathematical model, and the values at the corresponding steady state are shown in Table 1 in the Text.

Details regarding the inference algorithm. The first phase of the algorithm is composed of three steps, which are designed to identify primary metabolic alterations, whereas the second phase targets secondary mechanisms.

In **the first step**, the reaction rates of all enzymes, transporters, and non-enzymatic reaction steps are simultaneously, independently, and randomly varied within a range of [0.5, 1.5] times their nominal values. This range corresponds to perturbations ranging from 50% inhibition to 50% over-activation. The perturbations at each candidate site were sampled from the corresponding uniform distribution. The purpose of this first step is to reduce the number of possibly affected candidate sites based on only qualitative information, namely the direction (increase or decrease) of changes in metabolite levels observed in the cancer study. Expressed differently, we focus exclusively on the signs of changes rather than exact values. There are totally 27 candidate target sites, which determine the dimension of the original parameter space.

Each simulation of this type results in a metabolic profile, which is compared with the metabolomics data. To retain sufficiently many candidate sites for the following steps, we perform the comparison only with respect to measured metabolites that exhibit significant changes between disease and healthy cells. A threshold of ±10% appears reasonable for this step, as smaller changes are presumably not biologically significant. Of course, this threshold can be changed if it is deemed beneficial.

From millions of simulations, only those results are kept that result in metabolic profiles with the same types of increases or decreases as the experimentally or clinically measured metabolites. Collecting these results leads, for each candidate enzyme, to a distribution of admissible alterations within a range between 0.5 and 1.5 of the normal level. Within such a distribution, a value in the range [0.5, 1.0) indicates an inhibitory action, while values within the range (1.0, 1.5] show activation. A clearly skewed or a shifted distribution away from the uniform input distribution suggests that the metabolic changes are likely affected significantly by either one or the other action: inhibition or activation of the enzyme.

To determine the skewness of a distribution, we defined an index, which distinguishes whether values are distributed toward inhibition or activation and whether values on the other side are noise rather than exhibiting statistical significance. For this purpose, we compute the areas of the inhibitory part of the distribution (lower than the normal value of 1) and that of the activating part (higher than the normal value of 1). The index of skewness is the quotient between the smaller area and the whole area. Thus, each candidate enzyme has a computed index. For a high index (close to 0.5), the distribution has very low skewness, which suggests that cancer does not likely target this enzyme. We use for this first step a rather permissive value of 0.4; above this threshold, the corresponding enzyme is removed from the candidate list. Although this filter is rather coarse, experience shows that it often reduces the admissible set of solutions quite a bit, but that it does not reduce it to a degree where hardly any candidates are left.

For **the second step**, we perform simulations similar to those in the first step but now consider only those candidate sites surviving the filtering in the first step. Furthermore, we assess actual changes in the concentrations of metabolites and not just directions of change. Correspondingly, the filtering criterion is different. We run one million simulations of random perturbations simultaneously and independently sampled from uniform distributions on the list of survived candidate sites while allowing perturbations within a range of [0.2, 5.0]. This range of up to 5-fold deviations from normalcy was selected based on enzymological studies of cancer. The difference between each simulated metabolic profile and the data is again assessed using the Euclidean metric. The one thousand sets of hypothesized enzyme alterations with the smallest differences are selected for the next step. In contrast to the first step, the second step shrinks the kinetic parameter space based on quantitative comparisons of all metabolites in profiles.

In **the third step**, a genetic algorithm uses the selected one thousand sets as initial values for generating ensembles of truly "good" solutions. Specifically, for each set of enzymatic changes identified in the second step, the algorithm finds an optimized solution. However, there is no guarantee that every optimized solution is better than the best solution found in the second step, and to retain the best solutions, we select a subset of hypothesized enzyme combinations only if they are better than the best solution found in the second step. As before, we generate distributions of alterations for each candidate enzyme and analyze their skewness. At this point, we want to determine the most likely target sites, and therefore use a value 0.05 as the significance threshold. If an index is less than 0.05, cancer has a very high probability of either activating or inhibiting this enzymatic or non-enzymatic biochemical process. This conclusion implies a cancer target with statistical confidence. Thus, not only are the locations of targeted sites inferred, but the corresponding distribution also shows the intensity of a cancer-caused alteration at a site.

The above three steps yield the discovery of primary disease actions from metabolomics data, which account for the strongest effects of cancer on a metabolism. The second phase contains **steps 4 and 5**, which are designed to identify secondary mechanisms that contribute to the remaining alterations of metabolites. The two steps of the second phase are like steps 2 and 3 of the first phase, with the important exception that at this point the inferred primary actions had been fixed as median values in the model.

Again, one million random simulations are run with a perturbation range of [0.2, 5.0] during the fourth step. The top one thousand sets of hypothesized combinations of enzyme alterations are used as the initial values for a genetic algorithm in the fifth step. The optimized disease actions are again evaluated for skewness of the distributions at each candidate site. Those with significant skewness are considered as the secondary action sites of the disease.

Sampling, homogeneity, and computational load. For steps 1, 2, and 4, Monte Carlo sampling was used to obtain random parameter values in corresponding high-dimensional spaces. For each candidate site, five million (1st step), one million (2nd step), and one million (4th step) samples were randomly extracted from its uniform distribution. Therefore, the joint probability distribution of all random variables in a corresponding high-dimensional space is approximately a normal distribution following the central limit theorem. Accordingly, the samples are not homogeneous in each of the high-dimensional parameter spaces. However, sampling homogeneity is not required by our method. During each of these steps, only a subset of samples is screened out from a large number of random samples and forms an admissible subpopulation, which resides in a subspace in the corresponding high-dimensional parameter space. Provided that sufficiently many samples in the subspace are collected and entered in the admissible set, homogeneity is not crucial to the performance of our method. Our previous studies indicated that even some much smaller numbers of Monte Carlo sampling can reliably meet this criterion and produce robust conclusions about disease actions [5, 6]. Nevertheless, sampling with better homogeneity may greatly increase the chance of reaching targeted parameter values belonging to admissible subpopulations, and can thus greatly reduce the intensive computation. In this way, our method can be improved and applied to larger metabolic networks.

Computation of the Euclidean and the Jeffreys & Matusita metrics. The formulae and characteristics of the Euclidean and the Jeffreys & Matusita metrics for the distance between two metabolic profiles are shown in Table S5.

Table S1. Enzymes in purine metabolism

Enzyme or reaction	Abbreviation	EC Number
Hypoxanthine-guanine phosphoribosyltransferase	HGPRT	2.4.2.8
GMP synthetase	GMPS	6.3.5.2
Adenylosuccinate lyase	ASLI	4.3.2.2
GMP reductase	GMPR	1.7.1.7
AMP deaminase	AMPD	3.5.4.6
5’(3’) Nucleotidase	3NUC	3.1.3.31
Diribonucleotide reductase	DRNR	1.17.4.1
Adenosine deaminase	ADA	3.5.4.4
DNA polymerase	DNAP	2.7.7.7
DNases	DNAN	#
Guanine hydrolase	GUA	3.5.4.3
‘hypoxanthine excretion’	hx	$
‘xanthine excretion'	x	$
`uric acid excretion'	ua	$
Phosphoribosylpyrophosphate synthetase	PRPPS	2.7.6.1
Amidophosphoribosyltransferase	ATASE	2.4.2.14
Adenine phosphoribosyltransferase	APRT	2.4.2.7
`pyrimidine synthesis'	PYRS	#
IMP dehydrogenase	IMPD	1.1.1.205
Adenylosuccinate synthetase	ASUC	6.3.4.4
Methionine adenosyltransferase	MAT	2.5.1.6
Protein O-methyltransferase	MT	2.1.1.77, 2.1.1.80, and 2.1.1.100
S-adenosylmethionine decarboxylase	SAMD	4.1.1.50
5’-Nucleotidase	5NUC	3.1.3.5

RNA polymerase	RNAP	2.7.7.6
RNases	RNAN	#
Xanthine oxidase or xanthine dehydrogenase	XD	1.17.1.4 and 1.17.3.2

[#]: Multiple enzymes.

[$]: Non-enzymatic reaction.

Table S2. Predictions of secondary cancer mechanisms

Enzyme or reaction	Abbreviation	EC	Comment on prediction
Adenylosuccinate synthetase	ASUC	6.3.4.4	Correct prediction of site, but wrong mode of action
Amidophosphoribosyltransferase	ATASE	2.4.2.14	Correct prediction
Adenylosuccinate lyase	ASLI	4.3.2.2	Missed
adenosine monophosphate deaminase	AMPD	3.5.4.6	Missed
Uric acid excretion	VUA		Wrong prediction

Table S3. Robustness of the method for inferring primary cancer actions using randomly incomplete metabolomics data

Statistical measures	Value
Positive predictive value	92%
False discovery rate	8%
Sensitivity	19.17%

Table S4. Implementation of enzymatic alterations in human renal carcinoma cells in the mathematical model of purine metabolism

Altered enzymes	Abbreviation	Fold changes in maximum activity#
Amidophosphoribosyltransferase	ATASE	1.58
IMP dehydrogenase	IMPD	2.53
Adenylosuccinate synthetase	ASUC	1.49
Adenylosuccinate lyase	ASLI	1.76
adenosine monophosphate deaminase	AMPD	2.07
xanthine oxidase or xanthine dehydrogenase	XD	0.25

#: Fold changes were directly multiplied to maximum activity of an altered enzyme.

Table S5. Metrics for comparison of metabolic profiles and their characteristics

Metrics	*Formula*	*Characteristics*
Euclidean distance	$d(X,Y)=(\sum_{i=1}^{N}\lvert x_i-y_i\rvert^2)^{\frac{1}{2}}$	Commonly used; increases influences of errors from large components on distance to some extent
Jeffreys & Matusita distance	$d(X,Y)=(\sum_{i=1}^{N}\lvert \sqrt{x_i}-\sqrt{y_i}\rvert^2)^{\frac{1}{2}}$	Based on Euclidean distance; increases influences of errors from small components on distance to some extent

References

1. R. Curto, E. O. Voit, A. Sorribas and M. Cascante, *Math Biosci*, 1998, 151, 1-49.
2. R. Curto, E. O. Voit and M. Cascante, *The Biochemical journal*, 1998, 329 (Pt 3), 477-487.
3. E. O. Voit, *Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists*, Cambridge University Press, Cambridge, U.K., 2000.
4. G. Weber, *Clinical biochemistry*, 1983, 16, 57-63.
5. Z. Qi and E. O. Voit, *Translational Cancer Research*, 2014, 3, 233-242.
6. Z. Qi, G. W. Miller and E. O. Voit, *Toxicology*, 2014, 315, 92-101.