ModeScore: A Method to Infer Changed Activity of Metabolic Function from Transcript Profiles

Andreas Hoppe and Hermann-Georg Holzhütter

Charité University Medicine Berlin, Institute for Biochemistry, Computational Systems Biochemistry Group
hoppe@bioinformatics.org

Abstract

Genome-wide transcript profiles are often the only available quantitative data for a particular perturbation of a cellular system and their interpretation with respect to the metabolism is a major challenge in systems biology, especially beyond on/off distinction of genes.

We present a method that predicts activity changes of metabolic functions by scoring reference flux distributions based on relative transcript profiles, providing a ranked list of most regulated functions. Then, for each metabolic function, the involved genes are ranked upon how much they represent a specific regulation pattern. Compared with the naïve pathway-based approach, the reference modes can be chosen freely, and they represent full metabolic functions, thus, directly provide testable hypotheses for the metabolic study.

In conclusion, the novel method provides promising functions for subsequent experimental elucidation together with outstanding associated genes, solely based on transcript profiles.

1998 ACM Subject Classification J.3 Life and Medical Sciences

Keywords and phrases Metabolic network, expression profile, metabolic function

Digital Object Identifier 10.4230/OASIcs.GCB.2012.1

1 Background

The comprehensive study of the cell's metabolism would include measuring metabolite concentrations, reaction fluxes, and enzyme activities on a large scale. Measuring fluxes is the most difficult part in this, for a recent assessment of techniques, see [31]. Although mass spectrometry allows to assess metabolite concentrations in a more comprehensive way, the larger the set of potential metabolites, the more difficult [8]. Enzyme activities are currently measured only for selected enzymes [15, 33]. In contrast, the measurement of transcript RNA [35, 41] and protein levels [4] on a large scale is an established technique. Chip assays allow the high-throughput estimation for different conditions in a single experiment at manageable cost. Thus, we are confronted with the situation that for many cellular systems there are plenty of measured RNA profiles and a lack of accurate data for metabolite concentrations and enzyme activities. Even though the RNA will not allow accurate quantitative predictions of fluxes it is desirable to draw as many conclusions from that data as possible.

In the context of metabolic networks, transcript values are often used to select a subset of *active* reactions using a threshold [3, 38, 20, 23, 11, 22]. This has two immediate problems: (i) the selection of a proper threshold value and (ii) the subnetwork selected in that way may not be a functional network. The first problem is tackled in the GIMME algorithm [3] by gradually penalizing a reaction below a certain threshold, or in the approach by Shlomi et al. [38] by using a three-valued system — *off*, *on* and *intermediate* — with two thresholds. The second problem is solved either by the successive addition of necessary reactions considered *off* [3, 38] or by maximizing the number of concordant genes in an optimization framework

German Conference on Bioinformatics 2012 (GCB'12).
Editors: S. Böcker, F. Hufsky, K. Scheubert, J. Schleicher, S. Schuster; pp. 1–11

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

[20]. More elegantly, both problems are combined in a method which sets the threshold according to network function [22].

However, the main problem of reducing the transcript information to *on* and *off* genes is that it discards any information on gradual metabolic changes. But a complete switch-off of a metabolic enzyme is a rare event, most regulation is just the reduction or amplification of a particular function. Lee et al. predicted the metabolic state from absolute expression profiles without thresholding [28] but for expression *changes* it has not yet been done. This prompted us to develop a method focusing on quantitative changes of metabolic functionality.

To tackle this ambitious task the problem is *restated* first. While in transcript-based flux-balance analysis [3, 38, 22] the flux distribution resp. activity network is the result of the computation, here we assume that a plethora of functional flux distributions is already known. The transcript profile are used to decide which of them change their activity, in which direction, and how much.

In fact, functional reaction paths are studied throughout the last decades and are well-known for a large number of functions. On the basis of well-curated metabolic networks [10, 40] these functional flux modes can even be computed automatically [16].

In the context of the metabolism, transcript profiles are often used to find co-expressed genes, e.g. to understand the pattern of regulation [37, 21] and, thus, to predict targets of transcription factors. The Mamitsuka applied the approach to predefined metabolic reaction paths [30, 39] based on the KEGG database [25]. They used the results to rank the reaction paths [14], an idea which is also used here. However, the question of co-expression is not the main aspect here, instead it is the relative change of activity. Thus, we combine the idea of using expression profiles to score metabolic paths [14] with the idea of functional flux distributions [16]. The approach is supported by a finding of Notebaart et al. that the strongest gene co-expression occurs for coupled fluxes [32].

2 Results

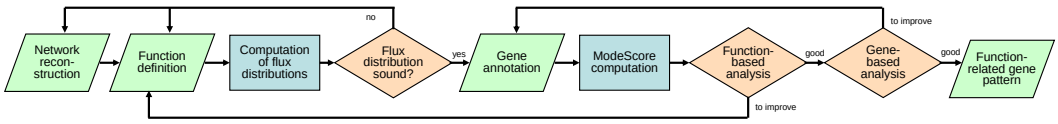

Figure 1 Flow chart of the ModeScore analysis.

The final goal of a ModeScore analysis is a set of remarkable function-related gene patterns from a series of full-genome transcript (or protein) profiles. It is a multi-step method with backwards jumps depicted in Figure 1, where the novel scoring algorithm is one of the steps. While some of the steps are computed (blue boxes in Figure 1) others involve manual curation (green rhomboids) and selective evaluation (khaki rhombi).

Preparation of flux distributions and gene annotations

For the predictivity of the ModeScore analysis sound representative flux distributions are critical. Sometimes, reference modes are already the result of the network testing process [12] but it still it may be worthwhile to tailor the flux distributions to a particular application. In a broad screening of metabolic functions, 1000s of different functions can be used with several alternate solutions computed automatically [18]. But the focus can be shifted, and also a confined analysis with just a few functions is sensible.

The gene–reaction association may be given in the reconstruction process [7] but even then it is advisable to manually refine this with the help of UniProt [1] and BRENDA [36] and the literature referenced with, e.g. if a respective isoform is expressed but at such a low level that another isoform is dominant it should be discarded, or if the enzyme is present but has a low affinity to the substrate. If several genes remain for one reaction, and if several repeat experiments have been performed their average expression is used. In the comparison between two cellular states, their expression values (as $\log_2$) are subtracted.

Scoring function

Let the k-th reference mode be denoted with $M_k = (m_i)_i$, where m_i is the flux rate of the i-th reaction. The relative expression profile is denoted by $V = (v_i)_i$, where v_i is the difference of the $\log_2$ values of the transcript abundances of one state and a reference state. Then the score of the mode is

$$\mathsf{Score}(M_k, V) = \frac{\sum_{i \in I_k} w_i \mathsf{score}_i(m_i, v_i)}{\sum_{i \in I_k} w_i}$$

where

$$I_k = \{i | \, m_i \neq 0\} \quad \text{and} \quad w_i = \sqrt{|m_i| \omega_i}$$

and

$$\mathsf{score}_i(m_i, v_i) = e^{-\frac{1}{2}\left(2\frac{\lambda v_i - m_i}{|m_i|}\right)^2}.$$

λ is chosen such that $\mathsf{Score}(M_k, V)$ is maximal. The non-negative numbers ω_i are fixed adjustments to modify the impact of a reaction on the score. If the gene expression changes are proportional to flux rates of a reference mode, the score would be equal to 1, and λ is the scaling factor between fluxes and expression changes. The rationale of this approach is that the larger the required flux increase, the larger the necessary enzyme amount and, thus, the increase in RNA transcription.

In the application below, reactions with stoichiometric factors larger than one (except protons) receive a weight ω_i which is the number of individual conversions for lumped reactions, and zero for non-enzymatic reactions, see Supplementary file 6.

The number $1/\lambda$, called amplitude in the sequel, is a measure for the function's expression difference. In fact, if all genes related to a specific function are regulated by the same amount α, then the score would be 1 and the amplitude equal to α. In this sense, the amplitudes are compatible to the expression changes.

Optimization procedure

The scoring function has been observed to have many local maxima for larger reference flux distributions, especially near zero. Therefore a specifically tuned calculation method has been recruited.

To find the global optimum of the scaling factor λ the derivation of the scoring function has been calculated algebraically. A set of probe points is calculated with the following procedure: (i) for every m_i/v_i a first level probe point is defined (ii) for every pair of first level probe points the arithmetic mean is added as another probe point (iii) for every consecutive pair (x, y) of second level probe points $n = 10 + 10(\frac{1}{1+x})$ intermediate points with equal distance are added as another probe point. This procedure is tailored to search for maxima

in superposed GAUSSian bell curves with centers at m_i/v_i. It uses the fact that a higher density of local maxima can be expected where the density of m_i/v_i is higher ($\rightarrow$ii) and the steepness of the bell curves for m_i near zero is higher, thus more probe points are generated. The huge abundance of probe points (compared to just a few maxima) to ensure a very low probability to miss solutions. The zeros of the derivation function are computed with the function `fzero` in octave [9] by supplying the probe points with different signs of the derivation function value. The zeros of the derivation function are the candidates for the global maximum which is subsequently selected.

ModeScore analysis

There are two different layers on which the ModeScore results can be analyzed, on the layer of functions and on the layer of individual genes (for one selected function).

The ranking of the amplitudes of the different functions reveals which are the most up and down-regulated functions in the comparison between two expression profiles. Additionally, three or four expression profiles can be compared in a way that the difference of amplitudes is ranked. This way, the functions with the most remarkable pattern are selected.

The next step is to check the layer of individual reactions and genes. For each function and pair of expression profiles the individual score components $\mathsf{score}_i(m_i, v_i)$ are analyzed. A high $\mathsf{score}_i(m_i, v_i)$ means that the reaction is considered belonging to the regulation pattern that led to the evaluation of the function with the amplitude $1/\lambda$. If the score is small or zero this reaction has been overruled by other reactions.

Possibly, a reaction's score $\mathsf{score}_i(m_i, v_i)$ has been the result of different genes, e.g. subunits of a protein complex. If it becomes for instance apparent that only one subunit is responsible for the regulation, other subunits can be deleted from the gene assignment to the reaction. The same should be done if several isozymes are recorded, but some of them are apparently not active in this particular experiment. It should be clear that a general network reconstruction includes all the possible reaction assignments, and only at this point of the analysis it is possible to select from the genes. After this modification, the ModeScore calculation must be repeated.

After this process has been iteratively repeated, the final reaction scores will reflect the function's regulation as shown by the expression profiles. For visualization it is now advisable to discard the genes which are not part of the regulation pattern.

In a nutshell, the following criteria are used to select the functions:

- functions with high absolute amplitudes are favored,
- central, well-known functions are favored to more peripheral functions,
- if a function can be seen as a sub-function of another function, check if the super-function is also a good candidate,
- prefer functions which are top/bottom scorer in more than one of the comparisons,
- prefer functions which show a plausible pattern of genes regulated in a consistent manner,
- prefer functions which have already been noted as specifically sensitive in the experiment,
- prefer functions which contain genes with highly remarkable regulation amplitude or pattern, and
- prefer functions whose pattern is not based on a single gene regulation.

Note that this selection process takes into account available domain knowledge.

The following criteria are used to select for each selected function the relevant genes which led to the amplitude estimation of the whole function are selected. This is also a manual process which takes into account several information sources and is guided by the following criteria:

- prefer genes with large transcript value changes,
- prefer genes with large absolute values,
- prefer genes associated to reactions which are adjacent to reactions whose genes have already been selected,
- prefer genes of highly specific enzymes or transporters,
- prefer genes of reactions which are most important for the function under regard and not for any more important function, and
- close gaps in an otherwise preferable reaction path.

Mostly, a connected reaction path (as a subset for the whole flux distribution) appears to be responsible for the regulation. Transporters are also integrated if consistent with the regulation of the enzymes. This aspect is noteworthy in comparison to the PathRanker [14] concept which ignores transporters altogether. Another important aspect in comparison to PathRanker (based on the KEGG MODULE reaction paths [26]) is that the definition of the relevant paths is not predefined (and fixed) but flexible and dependent on the actual expression values of the experiment. Often, they are in accordance with well-known reaction paths but in some cases there is a unique and noteworthy different selection of genes forming a functional unit.

Application of the method to TGFβ treatment of hepatocyte cultures

As an application of the proposed method, transcript profiles of primary mouse hepatocytes in monolayer culture (3 time points, 1h, 6h, and 24h, control versus TGFβ stimulation, 3 repeats) [6, 13, 5] are screened for remarkable functional alterations.

The functions with very high and low amplitudes of the profile comparisons 1h vs. 24h in the untreated sample and the control vs. TGFβ comparison at 24h have been analyzed for their functional relevance, see Supplementary File 4. Functions related to intermediates are replaced by their superordinate function, similar functions have been collected. For each of these functions the scores for individual reactions, see Supplementary File 5, have been analyzed for the set of genes which are responsible that the function in question has appeared with a particular regulation pattern.

The most prominent functions are the down-regulation of tyrosine degradation and ethanol degradation. The ethanol degradation pathway has already been verified to be sensitive to TGFβ [5]

However, the strong down-regulation of tyrosine degradation is a novel finding yet to be confirmed on the metabolic level, see Figure 2. The amplitudes $\frac{1}{\lambda}$ of the C1h/24h and the C/T24h comparisons are -3.5 and -0.87, respectively, belonging to the largest negative amplitudes among all functions. The scores $\mathsf{Score}(M_k, V)$ are 0.41 and 0.44 indicating a relatively strong coherence of the expression changes. The scores $\mathsf{score}_i(m_i, v_i)$ for the individual genes are especially high ($\geq$0.5) for the genes coding the reaction cascade from Phenylalanine to Acetoacetate+Fumarate, see Supplementary File 5 for details. Thus, the complete degradation pathway is down-regulated in time, amplified by TGFβ treatment, in a very consistent pattern, even suggesting a common regulation factor.

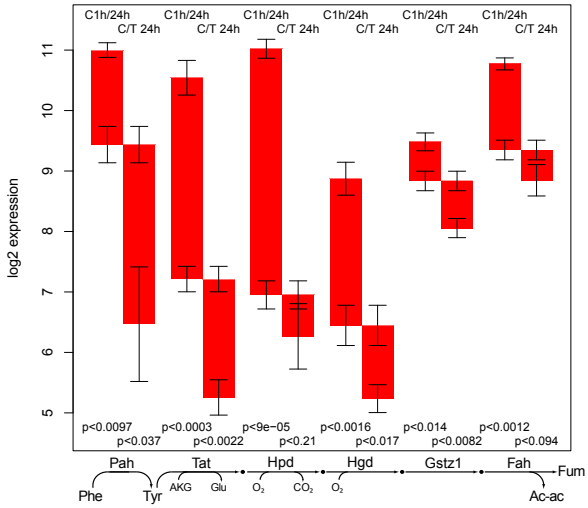

Figure 2 Regulation of the degradation cascade of phenylalanine and tyrosine. Red bars indicate down-regulation. For each gene, the respective first bar (C1h/24h) shows the comparison of the beginning of the experiment (1h) with the 24h time point of the control experiment, the second bar (C/T 24h) shows the comparison of control vs. treated sample at 24h. Error bars indicate average standard deviation of the 3 repeat experiments. P-values refer to the probability that there is an equal expression of two respective probe values, determined from the 3 repeats using Welch's t test [42], implemented in R [29]. The chart was prepared with R.

3 Discussion

Scoring algorithm

The scoring function is based on the superposition of Gaussian bell curves. The Gaussian bell curve is the typical approximation in situation where an expected behavior (the expression pattern matches the flux pattern) is blurred by a number of not quantitatively described factors which is the case for the relation of the RNA expression to the expressed enzyme and further to the reaction flux: RNA can be selectively degraded before translation, the rate of misfolded protein differs for different proteins, the life span of the enzymes differs, the reaction flux may deviate from the maximal catalytic rate, the enzyme life span is different and so is the catalytic efficiency.

The influence of the scores of individual spots to the total score of the metabolic functions is not equal but controlled by a weighting scheme. The weight mainly depends on the flux in the reference flux distribution. This is based on the assumption that the importance of the individual reactions to indicate the metabolic function as a whole can be approximated by the flux. However, that one or very few reactions dominate the scoring of the whole scoring function must be avoided. A good compromise has been found in a number of different trials as the square root of the absolute flux rate.

The luminescence signal on a gene chip is not actually measuring the RNA concentration in the sample as the affinity of the RNA to the probes is not equal and unknown for most genes. However, with a high signal the expectation value of the RNA concentration is high as well and the same is true for the resulting protein concentration, the enzyme activity, and finally the reaction flux. As the relation can be assumed to be normally distributed, the use of Gaussian bell curves is the most robust way to account for the uncertainty if the signal-flux relation.

The ModeScore concept would also be applicable if not just gene array luminescence data

but quantitatively more reliable qPCR [41], protein data, or even enzyme activity data were available. The salient point is that array data suffices to draw first conclusions before the laborious and costly other techniques are applied. On the other side, if data more closely representing metabolic function (such as enzyme activity data) was available then methods based on actual enzyme mechanisms could and should be applied such as kinetic modeling with time-resolved differential equations.

Maxima in the scoring function

In the scoring formula the global maximum of the scoring function has been used to obtain the scaling factor from the relative expression profile to the reference flux distribution. For some examples alternative local maxima have been observed with only slightly smaller scoring value. These alternative solutions can be seen as alternative ways to match the expression pattern to the flux distribution. The alternative solutions can be analyzed in the same way but generally their existence can be interpreted as an ambiguity of the expression profiles.

Reference flux distributions

The method critically depends on the quality of the predicted reference flux distributions. For this work they have been obtained in a similar way as in [12]. Different computation series have been preformed combining thermodynamic realizability [19] and flux minimization [17] in different proportions. The obtained flux distributions have individually been checked on biochemical plausibility while the one with the fewest reactions used is preferred, see Supplementary File 3.

An advantage of the ranking concept is that the flux distributions need not be representative (in the sense that overlaps are critical). Only a sufficient coverage of the whole metabolic network should be ensured (not to miss remarkable transcript changes). Not to miss a remarkable regulation abundance rather plentiful flux distributions should be used.

Comparison with other approaches

One may want to compare the ModeScore approach with a simple concept using the same reference modes: average the expression changes of all genes associated to the function. This has been tested and it showed almost no discriminative power. The problem is that even if a function contains a recognizable pattern of regulation, other genes are constant or regulated in the opposite direction. Thus the average expression changes are always close to zero. The ModeScore function however finds an amplitude that is consistent with the changes of several genes even if it is far from zero.

One may want to compare ModeScore with just filtering the largest changes of genes associated to the metabolic network. The functional context of these genes then had to be added manually. But many genes play a role in different functions, so they should be discriminated again manually. Even if this has been done, the result would not be able to detect moderately sized but very consistent regulation patterns (such as the Tyrosine degradation function above).

Often, lists of transcript changes are filtered by a certain threshold and the regulated genes (in both directions) are counted for each metabolic subsystem in either the gene ontology [2] or in the KEGG maps [24]. The main problem is that in such a collection of reactions/genes, the genes may have a completely different functional context: for instance in the amino acid metabolism maps there are reactions specific to synthesis, degradation, and

trans-amination, some of the reactions are needed to synthesize further cellular components from amino acids, for gluconeogenesis or urea synthesis. A pathway-oriented approach can not distinguish between these functions. Occasionally, even the same reaction located in different compartments participates in different functions: HMG-CoA synthase in the cytosol is involved in cholesterol synthesis while HMG-CoA synthase in the mitochondrium is only recruited for the generation of ketone bodies (in mammalian hepatocytes).

4 Methods and Materials

The stoichiometric network of the human hepatocyte's metabolism [12] has been adapted to the mouse and a plethora of about 1000 metabolic functions defined, see Supplementary File 2. The gene assignments have been obtained from the following resources: (i) Metabolic reactions with a KEGG reaction annotation in HepatoNet can be mapped to a mouse gene using KEGG [25]. (ii) Other metabolic functions with a EC number can also be mapped to a mouse gene using KEGG. (iii) Transporters with an annotation in TCDB can be mapped to a mammalian enzyme (UniProt nomenclature) using the TCDB [34]. (iv) Protein synthesis reactions have been mapped to the protein directly which makes sense for this work as the RNA is the most specific requisite for the synthesis reaction using Uniprot [1]. (v) Proteins have been mapped to their coding gene using Ensembl/BioMart [27]. (vi) Genes of the different species have been mapped to mouse genes (Ensembl nomenclature) using the computed homologies contained in BioMart database. (vii) BioMart database has also been used to assign the Affymetrix probeset identifiers to Ensembl annotations. HepatoNet was enlarged by the synthesis reactions of collagens. See Supplementary file 1 for the final network with gene annotations.

Availability

Calculation of the ModeScore scores is implemented in the freely available software FASIMU [18] in the function `modeset-score`. Supporting scripts including the score optimization process is available from the authors upon request.

Acknowledgements

The author thanks Iryna Ilkavets, Carsten Sticht, and Steven Dooley for discussions on the application example. The works was funded by the BMBF Germany as part of the VirtualLiver Network.

Supplementary files

1. Network definition in SBML (level 2 version 4) format including Ensembl annotation (xml).
2. Definition of metabolic functions (pdf).
3. Flux distributions satisfying the functional definitions (pdf).
4. Sorted tables of scores and amplitudes for functions (pdf).
5. Table of detailed reaction scores for selected functions (pdf).
6. List of weights by reaction (pdf).

The files can be downloaded from:
`http://www.bioinformatics.org/fasimu/_GCB_Modescore_Supplementary_files`

References

1 Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue):D115–D119, Jan 2004.

2 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

3 Scott A Becker and Bernhard Ø Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*, 4(5):e1000082, May 2008.

4 Grigoriy S Chaga. Antibody arrays for determination of relative protein abundances. *Methods Mol Biol*, 441:129–151, 2008.

5 Loredana Ciuclan, Sabrina Ehnert, Iryna Ilkavets, Hong-Lei Weng, Haristi Gaitantzi, Hidekazu Tsukamoto, Elke Ueberham, Nadja M Meindl-Beinker, Manfred V Singer, Katja Breitkopf, and Steven Dooley. Tgf-beta enhances alcohol dependent hepatocyte damage via down-regulation of alcohol dehydrogenase i. *J Hepatol*, 52(3):407–416, Mar 2010.

6 Steven Dooley, Jafar Hamzavi, Loredana Ciuclan, Patricio Godoy, Iryna Ilkavets, Sabrina Ehnert, Elke Ueberham, Rolf Gebhardt, Stephan Kanzler, Andreas Geier, Katja Breitkopf, Honglei Weng, and Peter R Mertens. Hepatocyte-specific smad7 expression attenuates tgf-beta-mediated fibrogenesis and protects against liver damage. *Gastroenterology*, 135(2):642–659, Aug 2008.

7 Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6):1777–1782, Feb 2007.

8 Warwick B Dunn. Mass spectrometry in systems biology an introduction. *Methods Enzymol*, 500:15–35, 2011.

9 John W. Eaton. *GNU Octave Manual*. Network Theory Limited, 2002.

10 Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, 7(2):129–143, Feb 2009.

11 Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol*, 7:501, 2011.

12 Christoph Gille, Christian Bölling, Andreas Hoppe, Sascha Bulik, Sabrina Hoffmann, Katrin Hübner, Anja Karlstädt, Ramanan Ganeshan, Matthias König, Kristian Rother, Michael Weidlich, Jörn Behre, and Herrmann-Georg Holzhütter. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol*, 6:411, Sep 2010.

13 Patricio Godoy, Jan G Hengstler, Iryna Ilkavets, Christoph Meyer, Anastasia Bachmann, Alexandra Müller, Gregor Tuschl, Stefan O Mueller, and Steven Dooley. Extracellular matrix modulates sensitivity of hepatocytes to fibroblastoid dedifferentiation and transforming growth factor beta-induced apoptosis. *Hepatology*, 49(6):2031–2043, Jun 2009.

14 Timothy Hancock, Ichigaku Takigawa, and Hiroshi Mamitsuka. Mining metabolic pathways through gene expression. *Bioinformatics*, 26(17):2128–2135, Sep 2010.

15 T. K. Harris and M. M. Keshwani. Measurement of enzyme activity. *Methods Enzymol*, 463:57–71, 2009.

16 Sabrina Hoffmann, Andreas Hoppe, and Hermann-Georg Holzhütter. Composition of metabolic flux distributions by functionally interpretable minimal flux modes (minmodes). *Genome Inform*, 17(1):195–207, 2006.

17 Hermann-Georg Holzhütter. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur J Biochem*, 271(14):2905–2922, Jul 2004.

18 Andreas Hoppe, Sabrina Hoffmann, Andreas Gerasch, Christoph Gille, and Hermann-Georg Holzhütter. FASIMU: flexible software for flux-balance computation series in large metabolic networks. *BMC Bioinformatics*, 12:28, 2011.

19 Andreas Hoppe, Sabrina Hoffmann, and Hermann-Georg Holzhütter. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol*, 1:23, 2007.

20 Carola Huthmacher, Andreas Hoppe, Sascha Bulik, and Hermann-Georg Holzhütter. Antimalarial drug targets in plasmodium falciparum predicted by stage-specific metabolic network analysis. *BMC Syst Biol*, 4(120), August 2010.

21 Jan Ihmels, Ronen Levy, and Naama Barkai. Principles of transcriptional control in the metabolic network of saccharomyces cerevisiae. *Nat Biotechnol*, 22(1):86–92, Jan 2004.

22 Paul A Jensen and Jason A Papin. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, 27(4):541–547, Feb 2011.

23 Livnat Jerby, Tomer Shlomi, and Eytan Ruppin. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol*, 6:401, Sep 2010.

24 Minoru Kanehisa. The KEGG database. *Novartis Found Symp*, 247:91–101; discussion 101–3, 119–28, 244–52, 2002.

25 Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, Jan 2008.

26 Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

27 Rhoda J Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011:bar030, 2011.

28 Dave Lee, Kieran Smallbone, Warwick B Dunn, Ettore Murabito, Catherine L Winder, Douglas B Kell, Pedro Mendes, and Neil Swainston. Improving metabolic flux predictions using absolute gene expression data. *BMC Systems Biology*, 6:73, 2012.

29 Uwe Ligges. *Programmieren mit R*. Springer-Verlag, Heidelberg, 3 edition, 2009. In German.

30 Hiroshi Mamitsuka, Yasushi Okuno, and Atsuko Yamaguchi. Mining biologically active patterns in metabolic pathways using microarray expression profiles. *SIGKDD Explorations*, 5(2):113–121, 2003.

31 Totte Niittylae, Bhavna Chaudhuri, Uwe Sauer, and Wolf B Frommer. Comparison of quantitative metabolite imaging tools and carbon-13 techniques for fluxomics. *Methods Mol Biol*, 553:355–372, 2009.

32 Richard A Notebaart, Bas Teusink, Roland J Siezen, and Balázs Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol*, 4(1):e26, Jan 2008.

33 Andrew Razgulin, Nan Ma, and Jianghong Rao. Strategies for in vivo imaging of enzyme activity: an overview and recent advances. *Chem Soc Rev*, 40(7):4186–4216, Jul 2011.

34 Milton H Saier, Can V Tran, and Ravi D Barabote. Tcdb: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res*, 34(Database issue):D181–D186, Jan 2006.

35 M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, Oct 1995.

36 Ida Schomburg, Antje Chang, Oliver Hofmann, Christian Ebeling, Frank Ehrentreich, and Dietmar Schomburg. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem Sci*, 27(1):54–56, Jan 2002.

37 Aswin S N Seshasayee, Gillian M Fraser, M. Madan Babu, and Nicholas M Luscombe. Principles of transcriptional regulation and evolution of the metabolic system in e. coli. *Genome Res*, 19(1):79–91, Jan 2009.

38 Tomer Shlomi, Moran N Cabili, Markus J Herrgård, Bernhard Ø Palsson, and Eytan Ruppin. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*, 26(9):1003–1010, Sep 2008.

39 Ichigaku Takigawa and Hiroshi Mamitsuka. Probabilistic path ranking based on adjacent pairwise coexpression for metabolic transcripts analysis. *Bioinformatics*, 24(2):250–257, Jan 2008.

40 Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121, 2010.

41 Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

42 B L Welch. The generalisation of student's problems when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.