

Predicting Physiological Concentrations of Metabolites from Their Molecular Structure

WOLFRAM LIEBERMEISTER

ABSTRACT

Physiological concentrations of metabolites can partly be explained by their molecular structure. We hypothesize that substances containing certain chemical groups show increased or decreased concentration in cells. We consider here, as chemical groups, local atomic configurations, describing an atom, its bonds, and its direct neighbor atoms. To test our hypothesis, we fitted a linear statistical model that relates experimentally determined logarithmic concentrations to feature vectors containing count numbers of the chemical groups. In order to determine chemical groups that have a clear effect on the concentration, we use a regularized (lasso) regression. In a dataset on 41 substances of central metabolism in different organisms, we found that the physical concentrations are increased by the occurrence of amino and hydroxyl groups, while aldehydes, ketones, and phosphates show decreased concentrations. The model explains about 22% of the variance of the logarithmic mean concentrations.

Key words: metabolite concentration, QSPR, molecule structure, lasso regression.

1. INTRODUCTION

RECENTLY, METABOLITE CONCENTRATIONS IN CELLS ARE GAINING new interest for the purpose of diagnostics and cell modeling. Quantitative structure-property relations (QSPR), obtained from experimental data by machine learning, have proven useful in predicting chemical and pharmacological properties of drug candidates (Clark and Pickett, 2000). We studied whether the QSPR approach could also be used to explain typical physiological concentrations of metabolites. In particular, we hypothesize that the occurrence of certain chemical groups leads to increased or decreased physiological concentrations in cells. In order to test this hypothesis, we applied a linear statistical model to feature vectors describing the molecular structure of metabolites and to logarithmic concentration data. Metabolic profiling (Goodacre *et al.*, 2004) creates large datasets on the prevalence of metabolites in different cells or tissues, and on concentration ratios between different samples. However, the sensitivity of high-throughput techniques such as mass spectrometry can depend on the molecule structure. Therefore, we tested the QSPR approach with a small but reliable dataset from a literature survey (Albe *et al.*, 1990).

2. DATA AND METHODS

2.1. Metabolite concentrations

Albe *et al.* (1990) have published a list of physiological concentrations of 41 metabolites in different cells and tissues, obtained from a literature screen. The cell types and tissues in this study comprise the bacterium *E. coli*, the yeast *S. cerevisiae*, the amoeba *D. discoideum*, red blood cells in rabbit and human, mung bean seedlings, as well as liver, muscle, and heart tissue in rat. For each concentration reported, either a single numerical value, an upper bound, or lower and upper bounds were given. To obtain a single representative value in the latter two cases, we used the upper bound or the geometric mean of upper and lower bound, respectively. All concentrations were transformed to decadic logarithms. We also studied a combined dataset. The reported overall concentration ranges depend strongly on the cell type: to eliminate this trend, we fitted a two-way analysis of variance model

$$\log c_{ik} = u_i + v_k + w + \epsilon_{ik} \quad (1)$$

where c_{ik} denotes the concentrations and the ϵ_{ik} are independent normal random variables. The subscripts i and k indicate the molecules and tissues, respectively. We regard $u_i + w$ as the typical logarithmic concentration of a molecule. As many values in the data table were missing, we estimated the model parameters by an expectation-maximization scheme (Dempster *et al.*, 1977) with the missing values as latent variables. Practically, this resulted in an iterative procedure: in each step, u_i , v_k , w , and ϵ_{ik} were determined by maximum-likelihood estimation, that is, from the row, column, and overall means of the data matrix. Then the missing values were substituted by the estimates $u_i + v_k + w$. See Tables 1, 2, and 3.

2.2. Molecule feature vectors

We downloaded the molecular structures of the metabolites under study from the LIGAND data base (Kanehisa *et al.*, 2002) at www.genome.jp/ligand/. Each molecule structure was translated into a feature (row) vector $\mathbf{x} = (x_0, \dots, x_M)$ as follows:

1. Each *local atomic configuration* (chemical group) comprising an atom, its bonds, and its neighbor atoms, is uniquely represented by a string s : for instance, C1C1N2O denotes a carbon atom forming a single bond to a carbon atom, a single bond to a nitrogen atom, and a double bond to an oxygen atom (see Fig. 1). The neighbor atoms appear in alphabetical order, and hydrogen atoms and their bonds are neglected.
2. Given the set of metabolites under study, we constructed the (alphabetically sorted, nonredundant) list of atomic configurations s_k that appear in at least two of the molecules. For each metabolite, the element x_k of the feature vector denotes how often s_k appears in this molecule. Note that the different elements of the feature vectors are statistically dependent by construction: for instance, if a carbon and an oxygen atom form a double bond, then at least two features containing this double bond must occur in the feature vector. Finally, we augmented the feature vector with an element $x_0 = 1$ and an element indicating the total number of atoms in the molecule.

2.3. Linear regression

We assume a linear model







$$y_i = \mathbf{x}_i \mathbf{a} + \sigma \eta_i. \quad (2)$$

between the feature (row) vector \mathbf{x}_i and the logarithmic concentration y_i of the i^{th} metabolite. The (column) vector \mathbf{a} contains linear weights for the chemical groups. The error term η_i is assumed to be a standard normal random variable, independent for the different metabolites, and the prefactor σ denotes the unknown standard deviation. The offset in the linear model is captured by the element x_0 in the feature vectors. A sparse weight vector \mathbf{a} is estimated as follows: we choose the vector \mathbf{a} that maximizes the score

$$J(\mathbf{a}) = \frac{1}{N} \|\mathbf{a}^T \mathbf{X} - \mathbf{y}\|^2 + \lambda \sum_k |a_k| \quad (3)$$

TABLE 1. PREDICTION OF 41 METABOLITE CONCENTRATIONS FROM THEIR MOLECULAR STRUCTURE^a

| | E. coli | S. cerev. | D. disc. | Mung bean | Rat liver | Rat heart | Rat muscle | Rabbit RBC | Human RBC | Combined data |
|------------------------------|---------|-----------|----------|--------------|--------------|--------------|---------------|---------------|--------------|------------------|
| # metabolites | 18 | 30 | 30 | 12 | 39 | 36 | 27 | 22 | 11 | 49 |
| Penalty par. λ_{opt} | 0.096 | 0.096 | 0.068 | 0.1 | 0.08 | 0.14 | 0.035 | 0.13 | 0.37 | 0.1 |
| Corr. coefficient | 0.66 | 0.24 | 0.31 | 0.64 | 0.54 | 0.6 | 0.6 | 0.55 | 0.58 | 0.52 |
| p -value | 0.0013 | 0.096 | 0.047 | 0.013 | 0.00019 | 4.8e-05 | 0.00052 | 0.0038 | 0.03 | 6.5e-05 |
| Total variance | 1.8 | 4.2 | 4.2 | 3.7 | 4.1 | 4.7 | 5.4 | 4.6 | 2.7 | 3.8 |
| Residual variance | 1.3 | 4.8 | 3.8 | 2.3 | 2.9 | 3 | 5 | 3.3 | 2 | 2.9 |
| Fraction explained | 0.28 | -0.15 | 0.082 | 0.38 | 0.29 | 0.36 | 0.071 | 0.3 | 0.27 | 0.22 |

| Feature | E. coli | S. cerev. | D. disc. | Mung bean | Rat liver | Rat heart | Rat muscle | Rabbit RBC | Human RBC | Combined data |
|---|---------|-----------|----------|--------------|--------------|--------------|---------------|---------------|--------------|------------------|
| Mean | 19.9 | 1610 | 74 | — | 455 | 107 | 371 | 20.9 | — | 108 |
| C-N NIC | 1.61 | 1.24 | — | — | 2.06 | 2.43 | — | — | — | 2.82 |
| C-O OIC | 1.85 | — | — | — | — | — | 1.23 | 1.19 | — | 1.19 |
| Size | — | 0.99 | 1.09 | 1.1 | 1.09 | 1.06 | 1.15 | 1.17 | 1.2 | 1.02 |
| C-C-C CIC1C | — | 1.06 | 1.14 | — | 1.24 | — | 1.19 | — | — | — |
| C-C CIC | — | — | — | — | — | — | 1.02 | 1.75 | — | — |
|  CIC1C1N | 1.85 | 4.96 | 3.78 | — | 1.29 | — | 3.03 | — | — | — |
|  CIC1C1O | 0.99 | — | — | — | — | — | — | — | — | — |
|  CIC1C1O1O | — | — | — | — | — | — | 1.15 | — | — | — |
|  CIC1C2O | — | — | 0.41 | — | — | — | — | — | — | — |
| C-C-N CIC1N | — | — | — | — | — | — | 0.93 | — | — | — |
|  CIC1O1O | — | — | — | 1.39 | — | — | — | — | — | — |
|  CIC1O2O | 3.74 | 0.96 | — | — | — | — | 0.996 | — | — | — |
| C-C=C CIC2C | — | — | — | — | — | — | 2.52 | — | — | — |
| C-N-C NIC1C | 0.89 | — | — | — | — | — | — | — | — | — |
| C-N-C NIC2C | — | — | 0.791 | 1.01 | — | 1.09 | — | 1.12 | — | — |
| C-O-C OIC1C | — | — | — | 3.11 | — | — | 0.37 | — | — | — |
| P-O-P O1P1P | — | — | — | — | — | — | 1.18 | — | — | — |
| P-O O1P | 1.71 | — | 1.67 | 0.59 | 0.66 | — | 1.3 | — | — | 0.87 |
| C-O O2C | — | 0.76 | 0.70 | 0.82 | 0.48 | — | 0.27 | — | — | 0.78 |
| C-O-P O1C1P | — | 0.50 | 0.13 | — | 0.28 | 0.16 | 0.043 | 0.074 | — | 0.35 |

^aFor each of the nine species and tissues as well as for the combined dataset (columns), a penalty parameter λ was chosen to yield a minimal prediction error in a 10-fold cross validation. For this optimal λ , explanative chemical groups along with the corresponding numerical weights were determined. The upper table summarizes the results from the nested-loop cross-validation to assess the generalization error. For each dataset, the table contains the number of metabolites, the optimal λ chosen, the linear correlation coefficient, and the respective p -value from the nested loop cross-validation (see text). The fraction of variance explained is the ratio (data variance-residual variance)/data variance. For the yeast *S. cerevisiae*, this ratio is negative: predictions and true values are linearly correlated, but the slope of the regression line differs from one. The lower table contains the numerical factors 10^{a_i} , where a_i is the weight associated with a feature. The features are sorted by the respective weights in the combined dataset. Features selected in neither of the datasets are not listed. RBC stands for "red blood cells."

TABLE 2. PREPROCESSED CONCENTRATION DATA^a

| | E. coli | S. cerev. | D. disc. | Mung bean | Rat liver | Rat heart | Rat muscle | Rabbit RBC | Human RBC | Combined data |
|-----------------------------|---------|-----------|----------|--------------|--------------|--------------|---------------|---------------|--------------|------------------|
| 2-Phospho-D-glycerate | — | 679.71 | — | — | 49 | 12.124 | 5 | 5 | 7 | 16.647 |
| 3-Phospho-D-glycerate | — | 161.25 | — | — | 410 | 26 | 43.818 | 46 | 48 | 60.713 |
| 6-Phospho-D-gluconate | — | 173.21 | 18 | 0.4 | 27 | — | — | 7.5 | — | 13.846 |
| ADP | 823 | 644.98 | 200 | — | 1700 | 1063.9 | 1059 | 500 | 126 | 410.2 |
| ADPglucose | — | — | — | 200 | — | — | — | — | — | 1226.6 |
| AMP | 151 | 225.83 | — | 10.9 | — | 123 | — | 60 | 50 | 80.747 |
| ATP | 2641 | 1445.7 | 700 | 10.9 | 3535 | 2366.4 | 3075 | 1700 | 1130 | 1041.3 |
| Acetyl-CoA | 350 | — | 12 | — | 39 | 9.6 | 1.3 | — | — | 11.218 |
| Citrate | 12990 | 700 | 60 | — | 375 | 164.59 | — | 138 | — | 217.07 |
| CoA | — | — | — | — | 187.35 | 58.652 | 1.7 | — | — | 19.642 |
| D-Erythrose 4-phosphate | — | — | — | 1.8 | — | — | — | — | — | 11.039 |
| D-Fructose 1,6-bisphosphate | 1900 | 2765.9 | 50 | — | 29.95 | 8.5 | 46 | 7 | 5 | 36.235 |
| D-Fructose 6-phosphate | — | 650 | 71 | 8.4 | 86.603 | 19 | 362 | 11 | 11 | 60.252 |
| D-Glucose | — | — | 500 | — | 10029 | 762 | 2377 | 6170 | — | 2229.7 |
| D-Glucose 1-phosphate | — | 100 | 20 | — | 16 | 36 | 65 | 6 | — | 19.784 |
| D-Glucose 6-phosphate | 801 | 2300 | 216 | 36 | 256.13 | 136 | 1033 | 62 | 27 | 211.94 |
| D-Ribose 5-phosphate | — | — | 26 | 4.4 | — | — | — | — | — | 41.432 |
| D-Ribulose 5-phosphate | — | — | 24 | 0.7 | — | — | — | 120 | — | 40.948 |
| D-Xylulose 5-phosphate | — | — | 14 | — | — | — | — | — | — | 14.072 |
| Dihydroxyacetone phosphate | 203 | 330 | 100 | 0.7 | 44.721 | 12 | 42.895 | 10 | 12 | 28.637 |
| Fumarate | — | — | 30 | — | 108 | 105 | — | — | — | 56.958 |
| GTP | 700 | — | — | — | — | — | — | 230 | — | 217.57 |
| Glyceraldehyde 3-phosphate | — | 692.82 | 10 | 0.6 | 16 | 3 | 21 | 3 | 4 | 11.531 |
| Glycerol-3-phosphate | 195 | — | — | — | 451.69 | 78 | 168 | — | — | 93.733 |
| Glycogen | — | — | 3420 | — | 36700 | 633 | 1500 | — | — | 2821.8 |
| Isocitrate | — | — | — | — | 29 | 34 | — | — | — | 20.363 |
| L-Alanine | — | 13229 | 970 | — | 1467.9 | 1563.1 | — | — | — | 1222.8 |
| L-Arginine | — | 18000 | — | — | — | 394 | — | — | — | 1052.4 |
| L-Asparagine | 776.79 | 6245 | 370 | — | 1354.2 | 2166.9 | — | — | — | 578.5 |
| L-Citrulline | — | 5000 | — | — | — | — | — | — | — | 1070.3 |
| L-Glutamic acid | 17363 | 22913 | 1200 | — | 3480 | 5267.6 | 2067 | — | — | 2443.9 |
| L-Glutamine | — | 22913 | — | — | 6090.9 | 4786 | — | — | — | 3418.4 |
| L-Lactate | — | — | — | — | 2340 | 4790 | 4221.6 | 3810 | — | 3411.8 |
| L-Lysine | — | — | — | — | — | 721 | — | — | — | 642.98 |
| L-Ornithine | — | 7000 | — | — | — | — | — | — | — | 1498.4 |
| L-Serine | — | — | — | — | 3708 | 344 | — | — | — | 732.39 |
| Malate | 1184.5 | — | 208 | — | 491 | 268.33 | 129 | — | — | 205.22 |
| NAD | 1615.5 | 1264.9 | 25 | — | 1097 | 426 | 310 | — | — | 210.57 |
| NADP | — | 54.772 | 25 | — | 97 | 5.8 | — | 22 | — | 18.817 |
| NADPH | — | 86.603 | 30 | — | 433 | 120 | — | — | — | 56.582 |
| NH ₃ | — | 30000 | — | — | 678 | — | 349 | — | — | 718.69 |
| Oxaloacetate | — | 50 | — | — | 10 | — | 26 | — | — | 5.4127 |
| Phosphate | — | 22000 | 9486.8 | — | 5251.5 | 4250 | 5500 | 50 | — | 2355.9 |
| Phosphoenolpyruvate | 165.23 | 30 | — | — | 143 | 4 | 8 | 10 | 12 | 14.447 |
| Pyrophosphate | — | — | — | — | 17.55 | — | — | — | — | 7.9234 |
| Pyruvate | 390 | 1600 | 60 | — | 187 | 100.7 | 56 | 90 | — | 96.251 |
| Succinate | — | — | 1670 | — | 1068 | 496 | — | — | — | 783.25 |
| Uridine diphosphate glucose | 1299 | 300 | 330 | — | 330 | — | 43 | 50 | — | 110.1 |
| alpha-Ketoglutaric acid | 476 | 1000 | 10 | — | 202 | 100.05 | 78 | — | — | 66.741 |
| Geom. mean | 871.6 | 1051 | 106.9 | 8.6 | 349 | 126.8 | 146.6 | 76.6 | 49.2 | |

^aThe data shown are based on the concentrations (μM) as published in Albe *et al.* (1990). For some metabolites, ranges for the values were given. In these cases, we used the geometric mean or the upper bound (see main text). The molecules are named according to the LIGAND database (Kanehisa *et al.*, 2002). RBC stands for “red blood cells.” The values for the combined dataset (right column) were extracted from the species-specific data as explained in the main text.

where the column vector $\mathbf{y} = (y_1, \dots, y_N)^T$ contains the concentrations, while the matrix \mathbf{X} contains the feature vectors \mathbf{x}_i as its rows. The second term stems from a lasso (L_1) prior on the weight vector and penalizes nonzero elements of \mathbf{a} (see Hastie *et al.* [2001]). By the choice of the penalty parameter λ , one can control the number of explanative features to be selected (termed here “relevant features”). To compute the weight vector, we use the algorithm from Öjelund *et al.* (2001), which can be downloaded as a matlab file at www.imm.dtu.dk/~hoe/files/lasso.m.

The predictive power of the model was measured by 10-fold cross-validation: given a fixed penalty parameter λ , the model is fitted to all data, except for a small test set containing about one tenth of the metabolites. By repeating this procedure for different test sets, we assess the mean square difference between predictions and true values for all metabolites. We can then choose an optimal penalty parameter

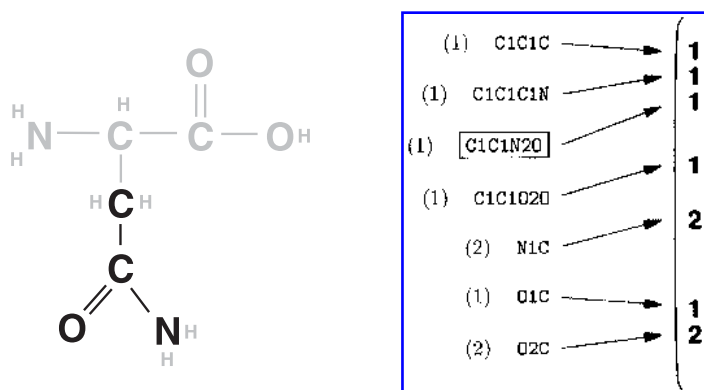


FIG. 1. Molecular structure of asparagine. **Left:** Two-dimensional structure. Hydrogen atoms, shown by small letters (H) are not considered in the analysis. **Right:** List of atomic neighborhoods in asparagine with the count numbers shown in brackets. The neighborhood C1C1N2O (box) of one carbon atom is highlighted in the scheme on the left. The feature vector (right) of asparagine contains the count numbers for all neighborhoods that appear in at least two of the metabolites under study.

to minimize this mean square prediction error. To study whether the entire procedure, including the optimization of λ , generalizes well on our data, we perform a nested-loop cross-validation: in a second level of (leave-one-out) cross-validation, the whole fitting (including optimization of λ by cross-validation) is done for the dataset without a certain metabolite i , leading to a prediction for this metabolite.

3. RESULTS

The results for the different cell types and the combined data are summarized in Table 1. For each dataset (column), the numbers denote the factors 10^{a_i} , where a_i is the weight for the i^{th} feature. In most datasets, larger molecules tend to show higher concentrations: for each atom in the molecule, the concentration is increased once by a factor larger than 1. In the combined dataset, only 5 out of the 37 atomic configurations appear as relevant: we found a positive effect for the amino group (N-C) and the hydroxyl (alcohol) group (O-C). Concentrations are decreased by (O-P) and (C-O-P) appearing in the phosphate group, and the carbonyl (aldehyde or ketone) group (O=C). In three of the individual datasets, however, the O-P group showed the opposite effect, contributing to higher concentrations.

Figure 2 shows how the model predictions generalize on the combined data: true (preprocessed) concentrations are compared to predictions from nested-loop cross-validation (see methods). The linear correlation between true and predicted values is about 0.52, with a p -value for nonzero correlation of about $6.5 \cdot 10^{-5}$ (t -test against the null hypothesis of Gaussian data without linear dependence). About 22% of the data variance (for the logarithmic concentrations) is captured by this model with relatively simple molecular descriptors.

4. DISCUSSION

Today, one of the main obstacles in mathematical modeling of cells is the lack of parameter values. For large-scale modeling, even rough estimates of typical physiological metabolite concentrations can be helpful. Here we explored the hypothesis that the chemical structure has an influence on physiological metabolite concentrations—which would QSPR turn into a tool to improve estimates of metabolite concentrations.

A crucial point in learning quantitative structure-property relations is the choice of the molecule features: our linear model, applied to logarithmic concentrations, assumes that each time a certain chemical group occurs, the physiological concentration changes by a certain factor. No interactive effects between the

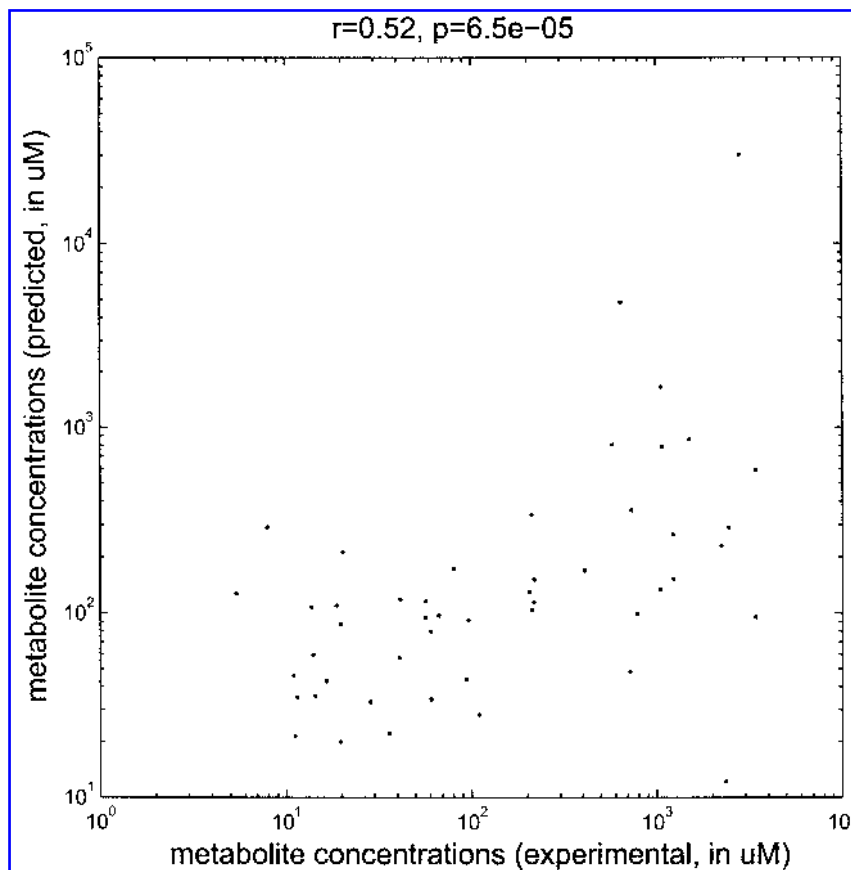


FIG. 2. Prediction of 41 metabolite concentrations in the combined dataset. Predicted concentrations from nested-loop cross-validation are plotted against the true values (in units of μM , and shown in log–log scale). The linear correlation coefficient is about 0.52, with a p -value of $6.5 \cdot 10^{-5}$ against the null hypothesis “no linear correlation.”

groups were considered. Larger predefined chemical groups (such as phenyl rings) could be incorporated into the analysis, but here we restricted ourselves to an exhaustive treatment of atomic neighborhoods, in order not to bias the study towards prior biochemical knowledge. The penalty parameter λ allows us to select the chemical groups that are most relevant for explaining the data, and it also reduces the redundancy caused by the correlations between count numbers of different features. We found that in some cases, small changes in the choice of λ can lead to a different selection of correlated features (not shown). Taking this into account, the results from the different datasets are in relatively good agreement.

Why do living cells accumulate substances with certain chemical groups? Thermodynamically, the different enthalpies of chemical groups lead to a higher or lower equilibrium concentration. However, metabolite concentrations in living cells are not in equilibrium, and moreover, they can be actively controlled by uptake and catalysis of production and degradation. Cells can, for instance, actively degrade metabolites that contain toxic chemical groups—which might also be detected by our model. For the substances from central metabolism studied here, we favor indeed a different explanation: some groups (like the amino group found in amino acids) are extensively used by the cell. They are transported through the metabolic network via different metabolites—which should therefore also show increased concentrations.

The proposed model generalizes well on the metabolites studied. We thus conclude that molecules from central metabolism show a relation between molecule structure and physiological concentrations and that our QSPR model is able to detect it. Of course, predictions based on this training set cannot be expected to remain valid for other classes of metabolites, such as large or toxic molecules. Nevertheless, we expect that training with more comprehensive datasets will permit predictions for a wider range of substances.

ACKNOWLEDGMENTS

The author would like to thank D. Kostka, F. Markowetz, and S. Schmeier for helpful discussions.

REFERENCES

- Clark, D.E., and Pickett, S.D. 2000. Computational methods for the prediction of "drug-likeness." *Drug Discovery Today* 5(2), 49–58.
- Goodacre, R., Vaidyanathan, S., Dunn, W.B., *et al.* 2004. Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology* 22(5), 245–252.
- Albe, K.R., Butler, M.H., and Wright, B.E. 1990. Cellular concentrations of enzymes and their substrates. *J. Theor. Biol.* 143, 163–195.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Royal Statist. Soc. B* 39, 1–38.
- Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at genomnet. *Nucl. Acids Res.* 30, 42–46.
- Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The Elements of Statistical Learning*, Springer-Verlag, New York.
- Öjelund, H., Madsen, H., and Thyregod, P. 2001. Calibration with absolute shrinkage. *J. Chemometr.* 15, 497–509.

Address correspondence to:

Wolfram Liebermeister
Max Planck Institute for Molecular Genetics
Kinetic Modelling Group
Innestrasse 73
14195 Berlin, Germany

E-mail: lieberme@molgen.mpg.de