ANALYSIS AND PREDICTION OF NUTRITIONAL REQUIREMENTS USING STRUCTURAL PROPERTIES OF METABOLIC NETWORKS AND SUPPORT VECTOR MACHINES

TAKEYUKI TAMURA[1] NILS CHRISTIAN[2]
tamura@kuicr.kyoto-u.ac.jp nils.christian@mpimp-golm.mpg.de
KAZUHIRO TAKEMOTO[3] OLIVER EBENHÖH[4]
takemoto@cb.k.u-tokyo.ac.jp ebenhoeh@abdn.ac.uk
TATSUYA AKUTSU[1]
takutsu@kuicr.kyoto-u.ac.jp

[1] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan*
[2] *Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany*
[3] *Graduate School of Frontier Sciences, University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, Chiba 277-8561, Japan*
[4] *Institute for Complex Systems and Mathematical Biology, University of Aberdeen, United Kingdom*

Properties of graph representation of genome scale metabolic networks have been extensively studied. However, the relationship between these structural properties and functional properties of the networks are still very unclear. In this paper, we focus on nutritional requirements of organisms as a functional property and study the relationship with structural properties of a graph representation of metabolic networks. In order to examine the relationship, we study to what extent the nutritional requirements can be predicted by using support vector machines from structural properties, which include degree exponent, edge density, clustering coefficient, degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. Furthermore, we study which properties are influential to the nutritional requirements.

Keywords: metabolic networks; support vector machines; nutritional profile; centrality.

1. Introduction

Computational analysis of biological networks is becoming important in systems biology and bioinformatics. Among these networks, detailed and large-scale studies have mostly been performed on metabolic networks.[a] It may be due to the fact that rather accurate and large-scale network data are available from such databases as KEGG [11] and EcoCyc [12], compared to protein-protein interaction networks and gene regulatory networks.

[a] Extensive studies have been done on inference of protein-protein interaction networks and gene regulatory networks, but detailed and large-scale analysis of these networks are scarce.

In order to analyze structural properties of genome scale metabolic networks, many studies have been performed. In particular, such graph features as degree exponent, clustering coefficient, edge density, frequency of network motifs and various kinds of centrality measures have been extensively studied [3, 10, 21, 23]. However, the relationships between these structural features and functional properties of metabolic networks are still very unclear.

On the other hand, for prediction of some functional properties of metabolic networks, *flux balance analysis* (FBA) has been extensively studied [18, 20]. The FBA-based approach allows to infer an optimal flux distribution when the structure of a network and the target compounds whose production should be maximized are given. This approach has been successfully applied to predict flux distributions of *E.coli* [20] and to identify knock-out targets of enzymes/genes [5, 18]. Recently, as a complementary approach, Handorf et al. (2005) proposed the concept of *scope* [9]. The scope is the set of all possible metabolites obtained from a given set of *seed* compounds and a given structure of a metabolic network. Though the scope cannot do flux optimization, it is more tolerant against errors in the network and is much faster to calculate. Handorf et al. applied the scope to infer the minimal nutritional requirements that must be met to sustain maintenance or growth of an organism [8].

In this paper, we focus on *nutritional requirements* of organisms as a functional property and study their relationship with structural features. For this purpose we examine to what extent the nutritional requirements can be predicted by using *support vector machines* (SVMs) from structural features. As for global features of metabolic networks, we use average clustering coefficient, edge density, degree exponent, cyclic coefficient, subgraph concentration, assortativity coefficient and average path length. As for local features of metabolic networks (i.e., features for each node), we use degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. As for training/test data, we use the nutritional profile generated by the scope-based method described in [8]. Furthermore, we study which structural features are influential to the nutritional requirements by examining combinations of structural features. The results show that use of either degree centrality or eigenvector centrality alone is effective for obtaining good prediction accuracy. The results also suggest that combination of centralities or combination of centrality and global features does not necessarily lead to improvement of prediction accuracy though global features are still useful for obtaining good prediction accuracy.

The organization of the paper is as follows. Section 2 and Section 3 explain the nutritional profiles and network features, respectively. Section 4 presents our prediction method using SVMs and global and local network features. Then, Section 5 provides the results of computational experiments. Finally, Section 6 discusses about the results and concludes with future work.

2. Generation of Nutritional Profiles

Each organism requires a minimal set of biochemical species, the nutrients, to allow for the production of metabolic precursors of higher molecules (e.g. proteins, RNA, DNA) and thus facilitate growth. For many organisms chemically defined nutrients are unknown. In [8] a method was introduced to computationally assess the minimal nutrients and a measure was derived that defines their essentiality, denoting how important these molecules are for the organism's growth.

Calculating nutritional profiles makes use of the method of network expansion [9], which calculates the set of producible metabolites from a given set of nutrients for a metabolic network. A biochemical reaction from the given metabolic network will operate if all its substrates are present. Subsequently, the reaction's products are added to the set of available metabolites. This procedure is iterated until no further metabolites are added, and the resulting set of metabolites is called the *scope* of the network for the given nutrients.

The method in [8] reverses this process: It calculates which nutrients are needed to produce a given set of target metabolites. Since the solution of this problem is not unique and highly combinatorial, a greedy algorithm is used to cover a large part of the solution space. Some molecules found in the solutions may be *replaceable* by others (e.g. because of nearly identical chemical composition). If a metabolite is replaceable by another, and the same holds true vice versa, these two metabolites are said to be *exchangeable*. For each organism one can therefore compile groups of so called *exchangeable resource metabolites*.

These organism specific groups are then distilled into global *resource types* by joining two metabolites if they are exchangeable in the majority of organism specific exchangeable resource metabolites. The relative occurrence of metabolites from a global resource type in the multiple nutrient sets for an organism defines the resource type's essentiality, reflected by a value in the interval $[0, 1]$. A nutrient profile for an organism is therefore formally defined as the vector of essentialities for the different resource types.

We have performed the above calculation for 447 organisms from the KEGG database [11], release 45. 45 resource types were defined. As targets we chose all metabolites that are present in at least 90% of the organisms' metabolic networks. In Fig. 1 we show the essentialities for a selected set of organisms to demonstrate

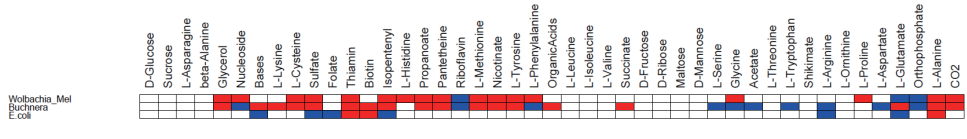

Fig. 1: Example of nutritional profiles.
White box: essentiality = 0, Blue box: 0 < essentiality < 0.9, Red box: essentiality ≥ 0.9

that the typical environment of an organism is reflected by the nutritional profiles. It can be clearly seen that the versatile *E. coli* needs a smaller set of nutrients to sustain growth compared to obligate symbionts and parasites such as *Buchnera* and *Wolbachia*, which satisfy their nutritional needs by using its host's metabolism.

The nutritional profiles thus describe a biological function, whose relationship to the structural properties of metabolic networks are studied in this work.

3. Structural Properties

To predict each nutrient profile from the network structure, we first constructed graphs from the metabolic networks, and then calculated nine global network parameters and four different centrality measures, where the centrality measures provide local features (i.e., each node's features) of a network.

3.1. *Dataset and Network Representation*

We downloaded the sets of metabolic reactions for 447 organisms from KEGG: Kyoto Encyclopedia of Genes and Genomes [11]. To emphasize the essential flow of metabolites, we neglected cofactor compounds of 71 types such as water and ATP in metabolic reactions.

The metabolic networks are represented by undirected graphs in which nodes and edges correspond to metabolites and substrate-product relationships, respectively. For example, considering a reaction S1+S2→P1+P2, metabolites S1 and S2 each connect to both products P1 and P2. That is, the edge list is as follows: (S1, P1), (S1, P2), (S2, P1), (S2, P2). In the case S1 and P1 are cofactor compounds the edge list is as follows: (S2, P2). Note that stoichiometric coefficients in the metabolic network are neglected.

To calculate network parameters (especially centrality measures), we extracted the largest connected component for each organism. In other words, small isolated clusters were removed.

3.2. *Global Network Parameters*

(i) The edge density: The edge density D is defined as the ratio of the number of edges E to the number of nodes N (i.e. $D = E/N$).
(ii) Clustering coefficient: The clustering coefficient is the average edge density at each node (i.e. at local levels), and is defined as $C = \sum_{i=1}^{N} c_i/N$. The value c_i is the local clustering coefficient defined as $2\Gamma_i/[k_i(k_i-1)]$ [1], where Γ_i and k_i are the number of edges among neighbors of node i and the number of neighbors of node i, respectively.
(iii) Degree exponent: This characterizes the heterogeneity of network connectivity. In biological networks such as metabolic networks, the frequency of nodes with k edges $P(k)$ is well known to follow power-law distributions (reviewed in [3]): $P(k) \propto k^{-\gamma}$, where γ is the degree exponent. As the degree exponent

increases, the probability that a node with large degree exists in a network decreases. That is, most nodes have similar degrees in the networks, indicating that the connectivity of the network is homogeneous. When the exponent becomes low, in contrast, the probability that a node with large degree exists in a network becomes high. That is, nodes tend to have different degrees in the networks, suggesting that the connectivity of the network is heterogeneous.

Assuming that the degree distribution of the metabolic networks follows a power law: $P(k) \propto k^{-\gamma}$, the degree exponent γ is extracted using maximum likelihood estimate given by the formula $\gamma = 1 + N\left[\sum_{i=1}^{N} \ln(k_i/k_{min})\right]^{-1}$ [16], where k_{min} is the smallest degree (the number of neighbors) in the network.

(iv) Cyclic coefficient The cyclic coefficient [13] is an extended clustering coefficient. The clustering coefficient only characterizes connections among neighbors (i.e. triangles or cycles of length 3), however, the cyclic coefficient can detect cycles of length more than 3 in addition to triangles. The cyclic coefficient is defined as $R = \sum_{i=1}^{N} r_i/N$, where r_i is $2\sum_{\langle jh\rangle}(L_{jh}^{i}-2)^{-1}/[k_i(k_i-1)]$. $\langle jh\rangle$ denotes all pairs of neighbors of node i, and L_{jh}^{i} is the length of the smallest cycle that passes through node i and its two neighbors j and h.

(v–vii) Subgraph concentrations: The (nt)-subgraph consists of a central node, $n-1$ neighbors and $n-1+t$ edges, where t denotes the number of edges among the neighbors [22]. That is, a subgraph composed of n nodes contains $(n-1)(n-2)/2+1$ different subgraphs because the maximal value of t is $\binom{n-1}{2}$.

The subgraph concentration [22] denotes a fraction of (nt)-subgraph abundance in all types of n-node subgraphs, and is defined as $S_{nt} = s_{nt}/\sum_{i=0}^{(n-1)(n-2)/2} s_{ni}$, where s_{nt} corresponds to (nt)-subgraph abundance.

In this paper, we focus on (31)-subgraphs (i.e. triangles), (42)-subgraphs (i.e. squares including two triangles), and (43)-subgraphs (i.e. 4-node complete graphs).

(viii) Assortative coefficient: The assortative coefficient [15] can be thought of as a compendium parameter of the correlation coefficient between the degree (the number of neighbors) of a node and the degrees of neighbors, and is defined as $r = (4\langle k_i k_j\rangle - \langle k_i + k_j\rangle^2)/(2\langle k_i^2 + k_j^2\rangle - \langle k_i + k_j\rangle^2)$, where k_i and k_j are the degrees of two nodes at the ends of an edge, and $\langle\cdots\rangle$ denotes the average over all edges.

(ix) Average path length: The average path length is the average length of the shortest paths between two nodes, and is defined as $L = \sum_{i,j} d_{ij}/[N(N-1)]$ [1], where d_{ij} is the shortest path length between nodes i and j.

3.3. *Centrality Measures*

(I) Degree centrality: Assuming correlation between the centrality (or importance) of a node and the degree (the number of neighbors) of the node, the degree centrality of node i is defined as $C_D(i) = k_i/(N-1)$ [7], where k_i is the degree of node i.

(II) Closeness centrality: When the average path length between a node and all other nodes is relatively short, the centrality of such a node can be considered high.

Therefore the closeness centrality of node i is defined as $C_C(i) = [\sum_{j=1,j\neq i}^{N} d_{ij}]^{-1}$ [7].
(III) Betweenness centrality: If a walker moves from one node to another via their shortest path, then well-passed nodes are defined to have a high centrality. Based on this, the betweenness centrality of node i is defined as $C_B(i) = \sum_{s\neq t\neq i} \sigma_{st}(i)/\sigma_{st}$ [7], where $\sigma_{st}(i)$ and σ_{st} are the number of shortest paths between nodes s and t on which there is node i and the number of shortest paths between nodes s and t, respectively. For normalization, the betweenness centrality is finally divided by the maximum value.
(IV) Eigenvector centrality: This is a higher version of the degree centrality. The degree centrality is only based on the number of neighbors. However, the eigenvector centrality can consider neighbors' centralities. The centrality $C_E(i)$ of node i is proportional to the average of the centralities of neighbors of node i: $C_E(i) = \lambda^{-1} \sum_{j=1}^{N} M_{ij} \cdot C_E(j)$, where λ is a constant and M_{ij} is the adjacency matrix. $M_{ij} = 1$ if node i connects to node j, and $M_{ij} = 0$ otherwise. With $\boldsymbol{x} = (C_E(1), \ldots, C_E(N))$ this equation can be rewritten as $\lambda \boldsymbol{x} = \boldsymbol{M} \cdot \boldsymbol{x}$. The eigenvector centrality is defined as the eigenvector with the largest eigenvalue [4].

4. Prediction by Support Vector Machines

SVM is a kind of statistical learning method and is basically used for binary classification. Let POS and NEG be the sets of positive examples and negative examples in a training data set, where each example is represented as a point in d-dimensional Euclidean space, and the corresponding d-dimensional vector is called a *feature vector*. Then, an SVM finds a hyperplane h such that the distance between h and the closest point is the maximum (i.e., the margin is maximized) under the condition that all points in POS lie above h, and all points in NEG lie below h. Once this h is obtained, we can infer that a new test data is positive (resp. negative) if it lies above h (resp. below h). If it is impossible to completely separate positive points from negative points, the soft margin (weighted combination of the margin and classification errors) is optimized.

SVM does not usually use feature vectors directly. It uses feature vectors in the form of kernel functions, where the kernel function is basically defined as the inner product between two feature vectors. Let $\mathbf{v}_{i_1}$ and $\mathbf{v}_{i_2}$ be the feature vectors of examples i_1 and i_2, respectively. Then, the value of the kernel function $K(\mathbf{v_{i_1}}, \mathbf{v_{i_2}})$ for this pair is calculated by $\mathbf{v_{i_1}} \cdot \mathbf{v_{i_2}}$. In some cases, more complex kernel functions are used to obtain better prediction performance. For example, $K(\mathbf{v_{i_1}}, \mathbf{v_{i_2}}) = exp(-\gamma \cdot ||\mathbf{v}_{i_1} - \mathbf{v}_{i_2}||^2)$ is frequently used. This kernel can be interpreted as the inner product between two infinite-dimensional vectors $\phi(\mathbf{v_{i_1}})$ and $\phi(\mathbf{v_{i_2}})$, where $\phi(\mathbf{v_{i_j}})$ is obtained from $\mathbf{v_{i_j}}$ (i.e., feature vector $\mathbf{v_{i_j}}$ is transformed into the infinite-dimensional feature vector $\phi(\mathbf{v_{i_j}})$). It is to be noted that we cannot show the exact form of $\phi(\mathbf{v_{i_j}})$ or calculate $\phi(\mathbf{v_{i_j}})$ explicitly, but can show the existence of $\phi(\mathbf{v_{i_j}})$ and can calculate $K(\phi(\mathbf{v_{i_1}}), \phi(\mathbf{v_{i_2}}))$ efficiently. This property (i.e., kernel functions can be efficiently computed without explicitly computing feature vectors)

is known as *kernel trick*. For details of SVMs and kernel functions, see [6, 19].

Let m and n be the numbers of compounds and organisms respectively ($m = 2140$, $n = 447$). Three types of matrices A, B, C appear in this prediction problem. As for A, $a_{i,j}$ represents the essentiality of the j-th resource type for the i-th organism. As for B, $b_{i,j}$ represents the value of the j-th global feature of the i-th organism. As for C, $c_{i,j,k}$ represents the value of the j-th local feature of the i-th compound and k-th organism. The purpose of the problem is to predict $a_{i,j}$ when $b_{i,j}$ and $c_{i,j,k}$ are given. $a_{i,j}$ is converted to a binary matrix with a threshold of 0.9. That is, if $a_{i,j} \geq 0.9$, it is treated as 1, otherwise it is treated as 0.

By shuffling organisms of A into five groups, we conducted fivefold cross validation test where information of B and C was used as the feature vector.

Each organism has at most $9+4m$ entries in the feature vector, where 9 elements are based on the information of B and $4m$ elements are based on the information from C. The kernel value for organisms i_1 and i_2 with the feature vectors $\mathbf{v}_{i_1}$ and $\mathbf{v}_{i_2}$ is calculated by $K(\mathbf{v}_{i_1}, \mathbf{v}_{i_2}) = exp(-10 \cdot ||\mathbf{v}_{i_1} - \mathbf{v}_{i_2}||^2)$.

In order to implement SVM, we used the software GIST [17]. We evaluated the prediction performance of our method by calculating sensitivity (sen), specificity (spe), negative sensitivity (senn), negative specificity (spen) and Matthew's correlation coefficient (MCC) [14] for each resource type. The definitions of these measures are as follows: Sensitivity=$\frac{TP}{TP+FN}$, Specificity=$\frac{TP}{TP+FP}$, Negative Sensitivity=$\frac{TN}{TN+FP}$, Negative Specificity=$\frac{TN}{TN+FN}$.

$$MCC(l) = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}.}$$

TP, TN, FP, FN denote true positive, true negative, false positive and false negative respectively.

5. Results

Although we conducted the computer experiments for 45 resource types, we extracted 17 resource types which satisfy $1/3 \leq (TP+FN)/(TN+FP) \leq 3$ for the following tables. This is reasonable since a fair evaluation of predictive accuracy is difficult if the number of positive examples or negative examples is too small.

Prediction results of fivefold cross validation where the feature vector consists of the 9 global parameters explained in Section 3.2 are shown in Table 1. Average of sensitivity, specificity, negative-sensitivity and negative specificity are 0.7199, 0.7344, 0.7261 and 0.7414 respectively. Since all these values are larger than 0.7, we can infer that information of global structural properties of metabolic networks is useful for predicting nutritional requirements of organisms.

However, Tables 2-4 show that some local properties of networks are more useful to predict nutritional requirements than global properties. Table 2 shows prediction results of fivefold cross validation where only degree centrality is used to calculate the feature vector, that is, the number of elements of the feature vector is m. Average

Table 1: Prediction result of fivefold cross validation where the feature vector consists of 9 global parameters explained in Section 3.1.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	63	94	96	194	0.5052	0.6037	0.7548	0.6736	0.2686
CO2	54	60	159	174	0.7260	0.7464	0.7631	0.7435	0.4896
D.Glucose	64	57	167	159	0.7455	0.7229	0.7130	0.7361	0.4587
Folate	72	36	274	65	0.8838	0.7919	0.4744	0.6435	0.3950
Glycerol	58	52	118	219	0.6941	0.6704	0.7906	0.8081	0.4816
L.Alanine	29	38	140	240	0.7865	0.8284	0.8921	0.8633	0.6851
L.Cysteine	62	87	92	206	0.5139	0.5974	0.7686	0.7030	0.2914
L.Lysine	62	45	175	165	0.7954	0.7383	0.7268	0.7857	0.5232
L.Methionine	67	48	245	87	0.8361	0.7852	0.5649	0.6444	0.4151
L.Phenylalanine	30	62	72	283	0.5373	0.7058	0.9041	0.8202	0.4819
L.Tyrosine	26	54	77	290	0.5877	0.7475	0.9177	0.8430	0.5463
Nicotinate	42	65	152	188	0.7004	0.7835	0.8173	0.7430	0.5222
Nucleoside	71	67	170	139	0.7173	0.7053	0.6619	0.6747	0.3796
OrganicAcids	38	48	77	284	0.6160	0.6695	0.8819	0.8554	0.5113
Pantetheine	54	43	253	97	0.8547	0.8241	0.6423	0.6928	0.5069
Riboflavin	77	47	236	87	0.8339	0.7539	0.5304	0.6492	0.3833
Sulfate	65	29	277	76	0.9052	0.8099	0.5390	0.7238	0.4869
Average					0.7199	0.7344	0.7261	0.7414	0.4604

of sensitivity, specificity, negative-sensitivity and negative specificity are 0.7232, 0.7922, 0.7649 and 0.7818, respectively. Since these values are larger than those of Table 1 by 0.0033, 0.0578, 0.0388 and 0.0404, it follows that degree centrality is more useful than global properties.

Similarly, Table 3 corresponds to the case where only closeness centrality is taken into consideration. However, specificity and negative sensitivity (0.5310 and 0.1816) are less than those of Table 1 by 0.2034 and 0.5445, respectively, although sensitivity and negative-specificity (0.9830 and 0.9090) are larger than those of Table 1 by 0.2631 and 0.1676. Since mcc of 0.2666 is less than mcc of Table 1 (0.4604) by 0.1938, we can conclude that closeness centrality is not useful when compared to global properties.

Finally, Table 4 corresponds to the case where only eigenvector centrality is used to calculate the feature vector. Although sensitivity (0.6907) is less than that of Table 1 by 0.0292, specificity, negative sensitivity and negative specificity (0.8431, 0.8179, 0.7774) are larger than those of Table 1 by 0.1087, 0.0918 and 0.0360, respectively. Since mcc of 0.5583 is larger than that of Table 1 (0.4604) by 0.0979 eigenvector centrality is more useful than global properties.

As said above, using only degree centrality or eigenvector centrality yields better predictive accuracy than using only global properties, but closeness centrality turns out not to be useful. The SVM calculations for betweenness centrality did not finish within three days and were aborted.

We also combined global and local properties to calculate the feature vector. Tables 5-9 show that combining global and local properties appropriately may yield

Table 2: Prediction result of fivefold cross validation where only degree centrality is used to compute the feature vector.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	54	86	104	203	0.5473	0.6582	0.7898	0.7024	0.3487
CO2	32	59	160	196	0.7305	0.8333	0.8596	0.7686	0.5960
D.Glucose	65	48	176	158	0.7857	0.7302	0.7085	0.7669	0.4957
Folate	68	17	293	69	0.9451	0.8116	0.5036	0.8023	0.5249
Glycerol	30	66	104	247	0.6117	0.7761	0.8916	0.7891	0.5334
L.Alanine	20	35	143	249	0.8033	0.8773	0.9256	0.8767	0.7414
L.Cysteine	40	66	113	228	0.6312	0.7385	0.8507	0.7755	0.4977
L.Lysine	79	36	184	148	0.8363	0.6996	0.6519	0.8043	0.4960
L.Methionine	53	33	260	101	0.8873	0.8306	0.6558	0.7537	0.5634
L.Phenylalanine	11	76	58	302	0.4328	0.8405	0.9648	0.7989	0.5043
L.Tyrosine	12	72	59	304	0.4503	0.8309	0.9620	0.8085	0.5135
Nicotinate	12	79	138	218	0.6359	0.9200	0.9478	0.7340	0.6178
Nucleoside	79	43	194	131	0.8185	0.7106	0.6238	0.7528	0.4528
OrganicAcids	15	67	58	307	0.4640	0.7945	0.9534	0.8208	0.5068
Pantetheine	58	41	255	93	0.8614	0.8146	0.6158	0.6940	0.4928
Riboflavin	64	34	249	100	0.8798	0.7955	0.6097	0.7462	0.5150
Sulfate	72	8	298	69	0.9738	0.8054	0.4893	0.8961	0.5700
Average					0.7232	0.7922	0.7649	0.7818	0.5277

Table 3: Prediction result of fivefold cross validation where only closeness centrality is used to calculate the feature vector.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	203	5	185	54	0.9736	0.4768	0.2101	0.9152	0.2684
CO2	196	0	219	32	1.0000	0.5277	0.1403	1.0000	0.2721
D.Glucose	180	3	221	43	0.9866	0.5511	0.1928	0.9347	0.2952
Folate	101	4	306	36	0.9870	0.7518	0.2627	0.9000	0.4035
Glycerol	216	2	168	61	0.9882	0.4375	0.2202	0.9682	0.2908
L.Alanine	240	1	177	29	0.9943	0.4244	0.1078	0.9666	0.1999
L.Cysteine	224	4	175	44	0.9776	0.4385	0.1641	0.9166	0.2244
L.Lysine	172	5	215	55	0.9772	0.5555	0.2422	0.9166	0.3219
L.Methionine	134	5	288	20	0.9829	0.6824	0.1298	0.8000	0.2332
L.Phenylalanine	259	2	132	54	0.9850	0.3375	0.1725	0.9642	0.2181
L.Tyrosine	259	0	131	57	1.0000	0.3358	0.1803	1.0000	0.2461
Nicotinate	206	2	215	24	0.9907	0.5106	0.1043	0.9230	0.2031
Nucleoside	165	9	228	45	0.9620	0.5801	0.2142	0.8333	0.2700
OrganicAcids	266	4	121	56	0.9680	0.3126	0.1739	0.9333	0.1868
Pantetheine	129	6	290	22	0.9797	0.6921	0.1456	0.7857	0.2448
Riboflavin	130	8	275	34	0.9717	0.6790	0.2073	0.8095	0.2957
Sulfate	110	4	302	31	0.9869	0.7330	0.2198	0.8857	0.3576
Average					0.9830	0.5310	0.1816	0.9090	0.2666

better accuracies than using only either global or local properties.

Tables 5-9 correspond to the cases where degree centrality, closeness centrality, betweenness centrality, eigenvector centrality alone and all these four properties together are used to calculate the feature vectors in addition to the 9 global features.

Table 4: Prediction result of fivefold cross validation where only eigenvector centrality is used to calculate the feature vector.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	40	77	113	217	0.5947	0.7385	0.8443	0.7380	0.4574
CO2	22	60	159	206	0.7260	0.8784	0.9035	0.7744	0.6411
D.Glucose	57	28	196	166	0.8750	0.7747	0.7443	0.8556	0.6248
Folate	69	16	294	68	0.9483	0.8099	0.4963	0.8095	0.5248
Glycerol	18	85	85	259	0.5000	0.8252	0.9350	0.7529	0.5015
L.Alanine	10	59	119	259	0.6685	0.9224	0.9628	0.8144	0.6821
L.Cysteine	17	86	93	251	0.5195	0.8454	0.9365	0.7448	0.5188
L.Lysine	47	49	171	180	0.7772	0.7844	0.7929	0.7860	0.5703
L.Methionine	37	50	243	117	0.8293	0.8678	0.7597	0.7005	0.5786
L.Phenylalanine	4	79	55	309	0.4104	0.9322	0.9872	0.7963	0.5382
L.Tyrosine	4	79	52	312	0.3969	0.9285	0.9873	0.7979	0.5283
Nicotinate	7	86	131	223	0.6036	0.9492	0.9695	0.7216	0.6201
Nucleoside	73	41	196	137	0.8270	0.7286	0.6523	0.7696	0.4887
OrganicAcids	11	72	53	311	0.4240	0.8281	0.9658	0.8120	0.4995
Pantetheine	33	56	240	118	0.8108	0.8791	0.7814	0.6781	0.5745
Riboflavin	44	43	240	120	0.8480	0.8450	0.7317	0.7361	0.5805
Sulfate	77	5	301	64	0.9836	0.7962	0.4539	0.9275	0.5627
Average					0.6907	0.8431	0.8179	0.7774	0.5583

Table 5: Prediction result of fivefold cross validation where the feature vector is calculated by 9 global parameters and degree centrality. (The total number of elements of the feature vector is $9+m$.)

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	42	90	100	215	0.5263	0.7042	0.8365	0.7049	0.3853
CO2	34	51	168	194	0.7671	0.8316	0.8508	0.7918	0.6207
D.Glucose	72	38	186	151	0.8303	0.7209	0.6771	0.7989	0.5136
Folate	77	12	298	60	0.9612	0.7946	0.4379	0.8333	0.5007
Glycerol	30	56	114	247	0.6705	0.7916	0.8916	0.8151	0.5841
L.Alanine	16	43	135	253	0.7584	0.8940	0.9405	0.8547	0.7234
L.Cysteine	32	71	108	236	0.6033	0.7714	0.8805	0.7687	0.5112
L.Lysine	81	36	184	146	0.8363	0.6943	0.6431	0.8021	0.4879
L.Methionine	59	32	261	95	0.8907	0.8156	0.6168	0.7480	0.5349
L.Phenylalanine	11	80	54	302	0.4029	0.8307	0.9648	0.7905	0.4780
L.Tyrosine	10	73	58	306	0.4427	0.8529	0.9683	0.8073	0.5210
Nicotinate	9	75	142	221	0.6543	0.9403	0.9608	0.7466	0.6501
Nucleoside	66	43	194	144	0.8185	0.7461	0.6857	0.7700	0.5102
OrganicAcids	17	64	61	305	0.4880	0.7820	0.9472	0.8265	0.5146
Pantetheine	56	35	261	95	0.8817	0.8233	0.6291	0.7307	0.5320
Riboflavin	58	29	254	106	0.8975	0.8141	0.6463	0.7851	0.5709
Sulfate	73	6	300	68	0.9803	0.8042	0.4822	0.9189	0.5784
Average					0.7300	0.8007	0.7682	0.7937	0.5422

In Table 5, sensitivity, specificity, negative sensitivity, negative specificity and mcc were 0.7300, 0.8007, 0.7682, 0.7937 and 0.5422 and they are larger than those of Table 2 by 0.0068, 0.0078, 0.0033, 0.0119 and 0.0145. Therefore, we can conclude

Table 6: Prediction result of fivefold cross validation where the feature vector is calculated by 9 global parameters and closeness centrality.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	203	5	185	54	0.9736	0.4768	0.2101	0.9152	0.2684
CO2	196	0	219	32	1.0000	0.5277	0.1403	1.0000	0.2721
D.Glucose	180	3	221	43	0.9866	0.5511	0.1928	0.9347	0.2952
Folate	101	4	306	36	0.9870	0.7518	0.2627	0.9000	0.4035
Glycerol	216	2	168	61	0.9882	0.4375	0.2202	0.9682	0.2908
L.Alanine	239	1	177	30	0.9943	0.4254	0.1115	0.9677	0.2040
L.Cysteine	224	4	175	44	0.9776	0.4385	0.1641	0.9166	0.2244
L.Lysine	172	5	215	55	0.9772	0.5555	0.2422	0.9166	0.3219
L.Methionine	134	5	288	20	0.9829	0.6824	0.1298	0.8000	0.2332
L.Phenylalanine	259	2	132	54	0.9850	0.3375	0.1725	0.9642	0.2181
L.Tyrosine	259	0	131	57	1.0000	0.3358	0.1803	1.0000	0.2461
Nicotinate	206	2	215	24	0.9907	0.5106	0.1043	0.9230	0.2031
Nucleoside	165	9	228	45	0.9620	0.5801	0.2142	0.8333	0.2700
OrganicAcids	266	4	121	56	0.9680	0.3126	0.1739	0.9333	0.1868
Pantetheine	129	6	290	22	0.9797	0.6921	0.1456	0.7857	0.2448
Riboflavin	130	8	275	34	0.9717	0.6790	0.2073	0.8095	0.2957
Sulfate	110	4	302	31	0.9869	0.7330	0.2198	0.8857	0.3576
Average					0.9830	0.5310	0.1819	0.9090	0.2668

Table 7: Prediction result of fivefold cross validation where the feature vector is calculated by 9 global parameters and betweenness centrality.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	68	87	103	189	0.5421	0.6023	0.7354	0.6847	0.2822
CO2	45	52	167	183	0.7625	0.7877	0.8026	0.7787	0.5658
D.Glucose	77	57	167	146	0.7455	0.6844	0.6547	0.7192	0.4019
Folate	76	21	289	61	0.9322	0.7917	0.4452	0.7439	0.4496
Glycerol	44	53	117	233	0.6882	0.7267	0.8411	0.8146	0.5353
L.Alanine	24	33	145	245	0.8146	0.8579	0.9107	0.8812	0.7323
L.Cysteine	48	74	105	220	0.5865	0.6862	0.8208	0.7482	0.4208
L.Lysine	67	44	176	160	0.8000	0.7242	0.7048	0.7843	0.5067
L.Methionine	64	38	255	90	0.8703	0.7993	0.5844	0.7031	0.4780
L.Phenylalanine	19	72	62	294	0.4626	0.7654	0.9392	0.8032	0.4781
L.Tyrosine	14	64	67	302	0.5114	0.8271	0.9556	0.8251	0.5520
Nicotinate	31	67	150	199	0.6912	0.8287	0.8652	0.7481	0.5665
Nucleoside	65	57	180	145	0.7594	0.7346	0.6904	0.7178	0.4512
OrganicAcids	27	53	72	295	0.5760	0.7272	0.9161	0.8477	0.5319
Pantetheine	53	39	257	98	0.8682	0.8290	0.6490	0.7153	0.5306
Riboflavin	67	34	249	97	0.8798	0.7879	0.5914	0.7404	0.4990
Sulfate	72	13	293	69	0.9575	0.8027	0.4893	0.8414	0.5365
Average					0.7322	0.7625	0.7409	0.7704	0.5011

that combining degree centrality and global properties yields better predictive accuracies than using only degree centrality.

In Table 6 mcc is 0.2668, thus taking closeness centrality into consideration still yields poor accuracies even when global properties are also taken into account.

Table 8: Prediction result of fivefold cross validation where the feature vector is calculated by 9 global parameters and eigenvector centrality.

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	51	79	111	206	0.5842	0.6851	0.8015	0.7228	0.3967
CO2	28	59	160	200	0.7305	0.8510	0.8771	0.7722	0.6154
D.Glucose	65	30	194	158	0.8660	0.7490	0.7085	0.8404	0.5819
Folate	64	22	288	73	0.9290	0.8181	0.5328	0.7684	0.5205
Glycerol	19	77	93	258	0.5470	0.8303	0.9314	0.7701	0.5360
L.Alanine	12	50	128	257	0.7191	0.9142	0.9553	0.8371	0.7119
L.Cysteine	22	82	97	246	0.5418	0.8151	0.9179	0.7500	0.5097
L.Lysine	50	49	171	177	0.7772	0.7737	0.7797	0.7831	0.5569
L.Methionine	34	50	243	120	0.8293	0.8772	0.7792	0.7058	0.5957
L.Phenylalanine	11	76	58	302	0.4328	0.8405	0.9648	0.7989	0.5043
L.Tyrosine	10	76	55	306	0.4198	0.8461	0.9683	0.8010	0.5012
Nicotinate	6	80	137	224	0.6313	0.9580	0.9739	0.7368	0.6485
Nucleoside	73	44	193	137	0.8143	0.7255	0.6523	0.7569	0.4745
OrganicAcids	17	67	58	305	0.4640	0.7733	0.9472	0.8198	0.4939
Pantetheine	36	53	243	115	0.8209	0.8709	0.7615	0.6845	0.5688
Riboflavin	42	44	239	122	0.8445	0.8505	0.7439	0.7349	0.5869
Sulfate	69	5	301	72	0.9836	0.8135	0.5106	0.9350	0.6082
Average					0.7021	0.8231	0.8121	0.7775	0.5536

Table 9: Prediction result of fivefold cross validation where the feature vector is calculated by 9 global parameters, degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. (The total number of elements of the feature vector is $9+4m$.)

Resource type	FP	FN	TP	TN	sen	spe	senn	spen	mcc
Bases	204	5	185	53	0.9736	0.4755	0.2062	0.9137	0.2646
CO2	201	0	219	27	1.0000	0.5214	0.1184	1.0000	0.2484
D.Glucose	185	3	221	38	0.9866	0.5443	0.1704	0.9268	0.2719
Folate	106	4	306	31	0.9870	0.7427	0.2262	0.8857	0.3661
Glycerol	218	1	169	59	0.9941	0.4366	0.2129	0.9833	0.2949
L.Alanine	241	1	177	28	0.9943	0.4234	0.1040	0.9655	0.1957
L.Cysteine	228	4	175	40	0.9776	0.4342	0.1492	0.9090	0.2087
L.Lysine	177	4	216	50	0.9818	0.5496	0.2202	0.9259	0.3099
L.Methionine	134	5	288	20	0.9829	0.6824	0.1298	0.8000	0.2332
L.Phenylalanine	264	2	132	49	0.9850	0.3333	0.1565	0.9607	0.2040
L.Tyrosine	264	0	131	52	1.0000	0.3316	0.1645	1.0000	0.2336
Nicotinate	206	1	216	24	0.9953	0.5118	0.1043	0.9600	0.2169
Nucleoside	167	8	229	43	0.9662	0.5782	0.2047	0.8431	0.2684
OrganicAcids	269	3	122	53	0.9760	0.3120	0.1645	0.9464	0.1906
Pantetheine	129	5	291	22	0.9831	0.6928	0.1456	0.8148	0.2557
Riboflavin	130	7	276	34	0.9752	0.6798	0.2073	0.8292	0.3048
Sulfate	116	4	302	25	0.9869	0.7224	0.1773	0.8620	0.3098
Average					0.9850	0.5278	0.1684	0.9133	0.2575

In Table 7, sensitivity, specificity, negative sensitivity and negative specificity were 0.7322, 0.7625, 0.7409 and 0.7704, respectively. Since all these values are larger than

0.7 and mcc is also larger than 0.5, we can conclude that combining betweenness centrality and global properties yields good predictive accuracies. In Table 8, although sensitivity and negative specificity (0.7021 and 0.7775) are larger than those of Table 4 by 0.0114 and 0.0001, specificity and negative-sensitivity (0.8231 and 0.8121) are smaller than those of Table 4 by 0.0200 and 0.0058. Since mcc of 0.5536 is smaller than that of Table 4 by 0.0047, it can be said that combining global properties with eigenvector centrality failed to improve the predictive accuracy. Finally, since mcc of Table 9 is 0.2575, we can conclude that combining all local and global features did not succeed in improving predictive accuracies.

6. Discussion and Conclusion

Here, we discuss the results of our computational experiments. First, we speculate about reasons for the predictability of nutrient profiles using network parameters. It is expected that metabolic pathways around compounds corresponding to nutrients are less dense because such compounds are not synthesized due to exogenous supply. In addition, the nutrient compounds are probably located at the periphery of metabolic networks for similar reasons. Thus, the global network parameters, which characterize network density, and the centralities of each node might be useful to predict nutrient profiles.

Overall, we could predict nutrient profiles from the network structure. Using the closeness centrality, however, predictions for several nutrient profiles showed relatively low accuracy. This might be caused by small-worldness of metabolic networks [23] that are represented as substrate-product relationships. As our metabolic networks are constructed by this representation, they have small-world features, indicating that all nodes are linked by short paths. Since the closeness centrality is based on the average path length as above, it is roughly homogenous among nodes due to the small-worldness. For this reason, the prediction using the closeness centrality was not effective. To predict nutrient profiles more accurately using network parameters, we might need to consider more appropriate network representations. For example, metabolic networks are not small-world when they are defined by atomic mappings [2] instead of substrate-product relationships. Accordingly, we might have good predictions because the closeness centrality would be different among nodes.

It is also seen that combination of global and local features did not necessarily lead to (considerable) improvement of the prediction accuracy. Though we have not yet identified the reason, this might be caused by overfitting. Identification of the reason is left as future work as well as introduction of some techniques to avoid overfitting.

As discussed, the results of computational experiments suggest that a combination of SVMs and structural features is useful for the prediction of nutritional requirements. Since nutritional requirements inferred by a scope-based method [8] are not necessarily perfect, it is worthy to develop alternative methods.

A combination of SVMs and structural features might be used as a complementary method to the scope-based method.

Though we have used artificially generated nutritional profiles, real nutritional requirements might be obtained by biological experiments. Therefore, use of real nutritional profiles is an important task for the future and requires a close collaboration with experimental biologists. Additional important future work is to improve the prediction accuracy. For that purpose, we need to develop novel structural features (especially local features). In particular, the introduction of local features defined for multiple nodes might be useful because there are many cases that knock-out of a single node (corresponding to a single enzyme/gene) does not affect functions of metabolic networks (because of the robustness of metabolic networks), but knock-out of multiple nodes greatly affects functions [5].

Acknowledgments

This work was partially supported by the International Research Training Group "Genomics and Systems Biology of Molecular Networks" Germany, and ITP (International Training Program) from JSPS, Japan.

References

[1] Albert, R., Barabási, A-L., Statistical mechanics of complex networks, *Rev. Mod. Phys.*, 74:47–97, 2002.

[2] Arita, M., The metabolic world of Escherichia coli is not small, *Proc. Natl. Acad. Sci. USA.*, 101:1543–1547, 2004.

[3] Barabási, A-L., Oltvai, Z. N., Network biology: understanding the cells' functional organization, *Nature Reviews Genetics*, 5:101–113, 2004.

[4] Bonacich, P., Some unique properties of eigenvector centrality, *Social Networks*, 29:555–564, 2007.

[5] Burgard, A. P., Pharkya, P., Maranas, C. D., OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnology and Bioenginerring*, 84:647–657, 2003.

[6] Cortes, C., Vapnik, V., Support-vector networks, *Machine Learning*, 20:273–297, 1995.

[7] Freeman, L. C., Centrality in social networks: Conceptual clarification, *Social Networks*, 1:215–239, 1979.

[8] Handorf, T., Christian, N., Ebenhöh, O., Kahn, D., An environmental perspective on metabolism, *Journal of Theoretical Biology*, 252:530–537, 2008.

[9] Handorf, T., Ebenhöh, O., Heinrich, R., Expanding metabolic networks: scopes of compounds, robustness, and evolution, *Journal of Molecular Evolution*, 61:498–512, 2005.

[10] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabási, A-L., Lethality and centrality in protein networks, *Nature*, 411:41–42, 2001.

[11] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y., KEGG for linking genomes to life and the environment, *Nucleic Acids Research*, 36:D480–D484, 2008.

[12] Karp, P. D., Keseler, I. M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S. M., Paulsen, I., Collado-Vides, J., Gama-Castro, S., Peralta-Gil, M.,

Santos-Zavaleta, A., Penaloza-Spinola, M. I., Bonavides-Martinez, C., Ingraham, J., Multidimensional annotation of the Escherichia coli K-12 genome, *Nucleic Acids Research*, 35:7577–7590, 2007.

[13] Kim, H-J., Kim, J. M., Cyclic topology in complex networks, *Phys. Rev. E*, 72:036109, 2005.

[14] Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochem. Biophys. Acta.*, 405:442–451, 1975.

[15] Newman, M. E. J., Assortative mixing in networks, *Phys. Rev. Lett.*, 89:208701, 2002.

[16] Newman, M. E. J., Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, 46:323–351, 2005.

[17] Pavlidis, P., Wapinski, I., Noble, W.S., Support vector machine classification on the web, *Bioinformatics*, 20(4):586–587, 2004.

[18] Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., Palsson, B. O., Comparison of network-based pathway analysis methods, *TRENDS in Biotechnology*, 22:400–405, 2004.

[19] Shawe-Taylor, J., Cristianini, N., Kernel Methods for Pattern Analysis, *Cambridge Univ. Press*, 2004.

[20] Stelling, J., Klamt, S., Bettenbrock, K, Schuster, S., Gilles, E. D., Metabolic network structure determines key aspects of functionality and regulation, *Nature* 420:190–193, 2002.

[21] Takemoto, K., Nacher, J. C., Akutsu, T., Correlation between structure and temperature in prokaryotic metabolic networks, *BMC Bioinformatics*, 8:303, 2007.

[22] Vázquez, A., Dobrin, R., Sergi, D., Eckmann, J. P., Oltvai, Z. N., Barabási, A. L,, The topological relationship between the large-scale attributes and local interaction patterns of complex networks, *Proc. Natl. Acad. Sci. USA*, 101:17940–17945, 2004.

[23] Wagner, A., Fell, D., The small world inside large metabolic networks, *Proceedings of the Royal Society of London B*, 268:1803–1810, 2001.