

系統プロファイルを利用した代謝反応ネットワーク中の 保存領域抽出手法

三宅晶子, 竹中要一, 松田秀雄

大阪大学大学院情報科学研究科バイオ情報工学専攻

概要

年々増加する代謝ネットワークの情報から生命機能情報を抽出するための方法について注目が集まりつつある。本研究では、代謝ネットワークから生物種間で保存されたサブネットワークを網羅的に探索する手法を開発した。本手法では代謝ネットワークをグラフとして扱い、グラフ中で隣接した酵素反応間の類似スコアを系統プロファイルに基づいて定義する。探索アルゴリズムは深さ優先グラフ探索に基づいている。19の生物種についてKEGGのデータを本手法に適用したところ、ほとんどの生物種に保存されたサブネットワーク、および古細菌とバクテリアで保存されたトリプトファン生合成系パスウェイが結果として得られた。

A method for extracting conserved subnetworks from metabolic network based on phylogenetic profile

Shoko Miyake, Yoichi Takenaka and Hideo Matsuda

Department of Bioinformatic Engineering,
Graduate School of Information Science and Technology, Osaka University

概要

This article presents a method to extract metabolic sub-networks conserved among organisms based on similarity between phylogenetic profiles of enzymatic reactions. We formulated a profile similarity score based on the phylogenetic profile. By using the score, we perform a connected component search algorithm. By applying our approach to the metabolic networks of 19 representative organisms selected from bacteria, archaea, and eukaryotes in the KEGG database, we detected some highly conserved sub-networks among the organisms. We obtained one of the sub-network conserved among most of the organisms and the tryptophan biosynthesis pathway conserved between bacteria and archaea.

1 はじめに

生物において、食物の消化・吸収やエネルギーの産生など、生命活動に必要な物質を合成または分解する活動のことを代謝とよぶ。代謝を構成するのは、酵素（タンパク質の一種）を触媒としてある化合物（基質）を別の化合物（生成物）に変換する酵素反応であり、ある酵素反応の生成物が別の反応の基質となることにより代謝反応ネットワークが構成される。代謝反応ネットワーク中のある基質から生成物までの反応経路は、一般に代謝反応パスウェイ（以下単にパスウェイと略記）とよばれる。その反応経路については近年よく解明されており、EcoCyc [1] や KEGG [2], WIT [3] などのデータベースに集められ、WWW 上で公開されている。

代謝反応ネットワークの構造は巨大であり、多数の分岐を伴う複雑なものとなっている。しかし、生物の進化を考える上でその構造が最初からそのような複雑なものであったとは考えにくく、初期の生命における代謝反応ネットワークはより単純な構造をしていたと考えられる。そこでの単純な機能を果たす一連の反応を代謝反応ネットワークの基本単位と考えた場合、現在のその複雑な構造は、それらの基本単位に対する環境に応じた酵素反応の追加や削除、組み合わせの結果として説明することができる。よって、これらの代謝反応ネットワークを比較解析することは、その基本的な構成単位を知り、ある生物種が進化の過程でどのように各パスウェイを獲得したか、さらにはある化合物を合成する方法についての知見を得る上で重要であるといえる。

パスウェイを計算機上で解析するこれまでの試みとしては、パスウェイ再構築 [4, 5, 6]、パスウェイのクラスタリング [7]、酵素の類似性によるパスウェイアライメント [8]、パスウェイに基づくゲノム比較 [9, 10, 11] などが挙げられる。パスウェイ再構築は、ある 2 つの化合物を指定して一方を基質、他方を生成物とするようなパスウェイを求める手法である。パスウェイのクラスタリングでは、パスウェイ中の酵素同士の配列類似性をもとにパスウェイ間の距離を計算し、パスウェイを分類する。パスウェイアライメントでは、酵素の機能階層の類似性をもとにパスウェイを分類する。パスウェイに基づくゲノム比較では、一つのパスウェイを構成する各酵素に種々のゲノム上の遺伝子を割り当てることにより比較を行っている。

本研究では、系統プロファイル [12] の類似性を導入することで、代謝反応ネットワーク中から、特定の生物種グループ間で保存されているようなサブネットワーク（以降、保存サブネットワークとよぶ）を探索するための手法を提案する。系統プロファイルとは、タンパク質の生物種間の分布パターン、すなわち、オーソログ関係にある遺伝子の産物であるタンパク質の生物種ごとの有無を 1 と 0 のビット列で表現したものである。系統プロファイルにおける生物種間の保存パターンが類似するタンパク質同士は機能的に関連があるとされ、これを利用することにより、機能的・進化的に関連のある反応からなるサブネットワークが得られることが期待される。

2 提案手法

2.1 代謝反応ネットワークのグラフ表現

本手法では、計算機による保存領域を探索するために、KEGG データベースの情報に基づき代謝反応ネットワークをグラフで表現する。代謝反応ネットワークを無向二部グラフ $G = (V_R, V_C, E)$ で記述し、これを代謝グラフとよぶ。ここで、 V_C は化合物を表す頂点（化合物頂点）の集合、 V_R は KEGG/REACTION データベースに登録されている素反応を表す頂点（酵素反応頂点）であり、 $V_R \cap V_C = \phi$ である。また、 E は $v_r \in V_R$ と $v_c \in V_C$ を結ぶ辺の集合である。酵素反応の多くが可逆反応であることが知られていることから、代謝グラフは無向グラフとして表現している。また、 V_R 中の各頂点は、EC 番号（酵素反応を分類する 4 組の数字）[13] でラベル付けされる。ラベル付けされる EC 番号は複数の場合もある。本手法では、各酵素反応に対応する酵素の情報については、KEGG データベースのアノテーション情報を参照しており、酵素反応 v_r に対する酵素集合を写像 $f_{KEGG}(v_r) \rightarrow \{e_1, e_2, \dots, e_m\}$ で表す。これに基づき、各酵素反応頂点に EC 番号をラベル付けする。代謝グラフの構成例を図 1-(A) に示す。

2.2 システムプロフィールとその類似度

システムプロフィールとは、複数の生物種に対し、あるタンパク質の各生物ゲノムでの有無を1, 0のビット列で表現したものである。あるタンパク質 x について、生物種ゲノム i でのタンパク質 x の存在の有無を、そのタンパク質をコードする遺伝子の、ゲノム i 上での有無により、次のように定義する。

$$o_i(x) = \begin{cases} 1 & x \text{ をコードする遺伝子が生物種 } i \text{ に存在} \\ 0 & \text{それ以外。} \end{cases} \quad (1)$$

あるタンパク質 x に対する、システムプロフィール $p(x)$ を N 次元のベクトル

$$p(x) = (o_1(x), o_2(x), \dots, o_N(x)) \quad (2)$$

として表現する。ここで、 N は生物種の数である。本研究では代謝グラフに含まれる全ての酵素について、システムプロフィールを構築し、そこからグラフ中の酵素反応頂点に対するシステムプロフィールを構築する。反応 $v_r \in V_R$ のシステムプロフィールは、 $p(v_r) = (o_1(e_1) \vee o_1(e_2) \vee \dots \vee o_1(e_m), \dots, o_N(e_1) \vee o_N(e_2) \vee \dots \vee o_N(e_m))$ とする。ここで、 $f_{KEGG}(v_r) \rightarrow \{e_1, e_2, \dots, e_m\}$ である。

本研究では、サブネットワークを探索する際の尺度基準として、酵素のシステムプロフィールの類似性を用いる。これはシステムプロフィールが一致または類似したタンパク質は、機能が類似する傾向にあるという仮定 [12] を利用している。これは、複合体を構成するタンパク質同士や、同じパスウェイに属するタンパク質同士は、進化の過程で同じように保存されるか、または削除されるであろうという考えに基づく。2つのシステムプロフィール $p(a) = (a_1, a_2, \dots, a_N), p(b) = (b_1, b_2, \dots, b_N)$ の間の類似度は、

$$\text{match}(p(a), p(b)) = \frac{|\{i | a_i = b_i, 1 \leq i \leq N\}|}{N} \quad (3)$$

と定め、これをシステムプロフィール間の類似スコアとよぶ。実際の類似スコアの例について、図1-(B)に示す。図中では、酵素反応 r_1 と r_2 の間の類似スコアは、システムプロフィールの4ビット中3ビットが一致していることから0.75となる。同様に r_2, r_3 間の類似スコアは0.5となる。

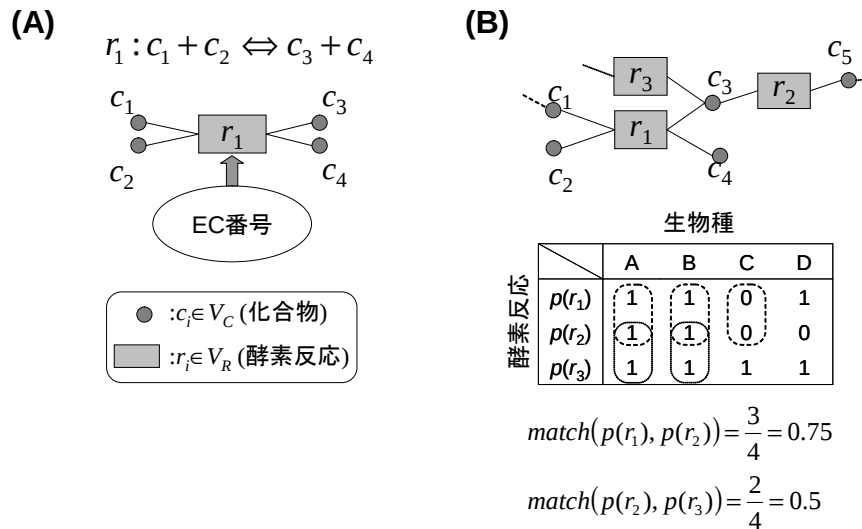


図 1: 代謝グラフの構成例-(A) と、システムプロフィールの類似スコアの例-(B)

2.3 保存サブネットワークの探索アルゴリズム

本稿では，様々な生物種のグループに対し，それぞれのグループ間で保存されている代謝反応ネットワーク中の領域を保存サブネットワークとして探索する．代謝グラフ $G(V_R, V_C, E)$ から保存サブネットワークを探索するためのアルゴリズムを次に示す．系統プロファイルの類似スコアに対する閾値 T を与え，1つの化合物頂点を介して隣接する2つの反応の類似スコアが閾値 T よりも大きいグラフの連結成分を探索する．

入力 • 代謝グラフ $G(V_R, V_C, E)$

• 反応頂点に対する系統プロファイルの集合 $P = \{p(v_r) | v_r \in V_R\}$

• 類似スコアの閾値 T

条件 • 代謝グラフ G 中で1つの化合物頂点を介して隣接する反応 $v_{r1}, v_{r2} \in V_R$ について， $match(p(v_{r1}), p(v_{r2})) > T$ ，

出力 • 保存サブネットワークの集合 $SN = \{G_i(V_{Ri}, V_{Ci}, E_i) | G_i \subseteq G\}$

アルゴリズム 深さ優先探索アルゴリズムに基づいて，条件を満たす代謝グラフ G 中の連結成分を探索する．

1. 始点となる酵素反応頂点を1つ選ぶ．
2. 1つの化合物頂点を介して隣接する酵素反応のうち，条件を満たすものを順次探索する．
3. 条件を満たす未訪問の酵素反応頂点がなくなった時点で，訪問した頂点から成る連結成分を保存サブネットワークとして出力する．

3 実験・結果

提案手法を KEGG データベース [2] から得られた代謝反応ネットワークに対して適用した．まず，KEGG に登録された生物種のうち，真核生物，バクテリア，古細菌から19種を代表生物種として選択した（表1参照）．次に，KEGG/LIGAND データベースを基に代謝グラフを構築した．頂点数は，酵素反応頂点 V_R と化合物頂点 V_C それぞれについて，4443 と 3987 となった．また，辺の数は 36998 である．系統プロファイルは KEGG/ENZYME の EC 番号のエントリ 3568 個に対して作成した．ここで，前述の 19 生物種のどのゲノムデータから見出せなかった EC 番号のエントリは除いている．作成された系統プロファイルの 1, 0 のビットパターンについては，考えられる 2^{19} 通りの組み合わせのうち，669 種類がプロファイル中に見出された．

系統プロファイルの閾値 T を 0.89 とし，提案手法を以上のデータに適用した結果，528 個の保存サブネットワークが得られた．この閾値 T の値 0.85 は，2つの系統プロファイル間のビットの相違を 19 生物種中 2 つまで許す値となっている．ここで，得られたサブネットワークに対する評価値として，サイズとスコアを定める．サブネットワーク $G_i(V_{Ri}, V_{Ci}, E_i) \subseteq G$ に対するサイズは，酵素反応頂点の数 $|V_{Ri}|$ とする．また，スコアは

$$S_{SN}(G_i) = \frac{\sum_{v_{rj} \in V_{Ri}} \sum_{k=1 \dots N} o_k(v_{rj})}{|V_{Ri}|} \quad (4)$$

表 1: 選択した 19 生物種

ドメイン	生物種名	タンパク質数	酵素数
Eucaryotes	<i>Homo sapiens</i>	19773	1900
	<i>Rattus norvegicus</i>	5487	1173
	<i>Drosophila melanogaster</i>	15925	850
	<i>Caenorhabditis elegans</i>	21521	856
	<i>Arabidopsis thaliana</i>	28159	1192
	<i>Plasmodium falciparum</i>	5314	533
	<i>Saccharomyces cerevisiae</i>	5855	1082
Bacteria	<i>Escherichia coli</i> K-12 MG1655	4289	1327
	<i>Xylella fastidiosa</i> 9a5c	2831	686
	<i>Xanthomonas campestris</i> pv. <i>campestris</i> ATCC 33913	4181	996
	<i>Bordetella bronchiseptica</i> RB50	4994	1001
	<i>Bradyrhizobium japonicum</i> USDA110	8317	1217
	<i>Bacillus subtilis</i> 168	4106	1005
	<i>Mycobacterium leprae</i> TN	1605	633
	<i>Rhodopirellula baltica</i> (<i>Pirellula</i> sp. strain 1)	7325	899
Archaea	<i>Methanococcus jannaschii</i> DSM2661	1770	536
	<i>Methanosarcina mazei</i> Goel	3371	672
	<i>Aeropyrum pernix</i> K1	2694	471
	<i>Thermoplasma acidophilum</i>	1478	476

と定める．これは，サブネットワーク中の反応を持つ生物種数の平均値である．得られたサブネットワークのスコア分布およびサイズ分布を図 2 に示す．スコアの平均は 5.19，サイズの平均は 4.57 であった．

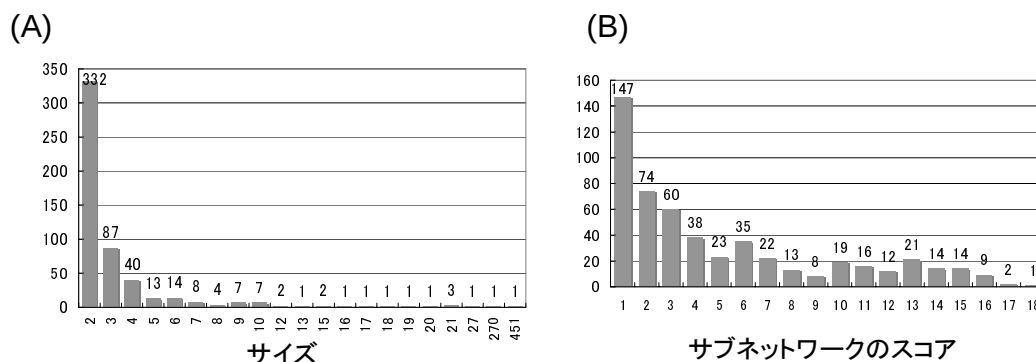


図 2: 得られたサブネットワークのサイズ分布-(A) とスコア分布-(B)

生物種間で最も保存されたサブネットワークは，生命の維持に必須のものであると考えられる．実際に得られた結果でスコアが最大のサブネットワークは，サイズが 270，スコアが 17.5 のものとなった．この巨大なサブネットワークは表 2 のような，保存されていることが既知 [14] の代謝マップを多く含んでいる．表 2 において，クラスは代謝マップの分類を，反応数は結果のサブネットワークにおいてその代謝マップに属する反応の数を示している．

図 3 に，KEGG のフェニルアラニン・チロシン・トリプトファン生合成の代謝マップ上に現れるサブネットワークを示す．図中では，色づけされた反応（四角形）および化合物（丸）が，結

果のサブネットワークに含まれているものである．このサブネットワークのサイズは 15，スコアは 13.87 であり，13 の反応が代謝マップ中のトリプトファン生合成パスウェイ上にマッピングされている．このうち，10 の反応は 19 生物種中の全てのバクテリア，古細菌と，1 つの真核生物である *Arabidopsis thaliana*（シロイヌナズナ）の間で保存されている．ただし，図 3 中の破線の円で囲まれている 3 つの反応は *A. thaliana* ではその情報が欠損している．*A. thaliana* と同様に，*Mycobacterium leprae* においても，最近まで反応 5.3.1.24 の遺伝子 (*trpF*) の情報が欠損していた．しかし，*Streptomyces coelicolor* において，*trpF* の役を *hisA* が果たしていることが近年報告され，同じことが *Mycobacterium tuberculosis* でも起こっていることが示唆されている [15]．それに伴い，*M. leprae* の反応 5.3.1.24 にも *hisA* の情報が割り当てられた．そのため，*A. thaliana* についても同様に，その遺伝子が同定される可能性があると考えられる．また，破線の四角形で囲まれた部分は，それが *E. coli* のオペロン領域によってコーディングされているパスウェイであり，進化の過程における 1 つの単位と予想されるサブネットワークが本手法によって得られたと考えられる．

表 2: 最も保存されているサブネットワークが含む主な代謝マップ

クラス	代謝マップ名	酵素反応数
Carbohydrate Metabolism	Glycolysis / Gluconeogenesis	17
	Pyruvate metabolism	9
	Citrate cycle (TCA cycle)	9
	Pentose phosphate pathway	8
Energy Metabolism	Carbon fixation	14
	One carbon pool by folate	11
Lipid Metabolism	Fatty acid biosynthesis	15
	Fatty acid metabolism	10
Nucleotide Metabolism	Purine metabolism	32
	Pyrimidine metabolism	26
Amino Acid Metabolism	Arginine and proline metabolism	12
	Glutamate metabolism	10
	Alanine and aspartate metabolism	8
Metabolism of Cofactors and Vitamins	Folate biosynthesis	8
Translation	Aminoacyl-tRNA synthetases	16

4 まとめ

系統プロファイルの類似スコアを設定することで，代謝反応ネットワークから生物種間で保存されているサブネットワークを探索する手法を提案した．選択した 19 生物種のデータに本手法を適用して実験を行い，いくつかの保存されたサブネットワークが得られた．これらの情報は，代謝ネットワークの進化的関連性の情報を示唆するものと考えられる．今後は，生物種の系統樹の情報を考慮に入れた類似スコアへの拡張を予定している．

参考文献

- [1] Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. and Karp, P.D. “EcoCyc: a comprehensive database resource for Es-

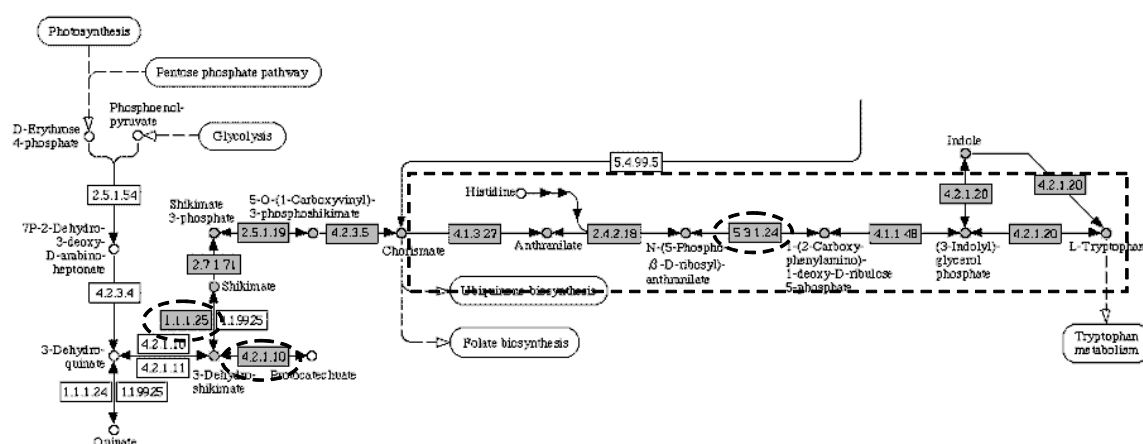


図 3: フェニルアラニン・チロシン・トリプトファン生合成系上のサブネットワーク

- cherichia coli,” *Nucleic Acids Research*, Vol. 33, pp. D334–337, 2005.
- [2] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. “The KEGG resource for deciphering the genome,” *Nucleic Acids Research*, Vol. 32, pp. D277–280, 2004.
 - [3] Overbeek, R., Larsen, N., Pusch, G.D., D’Souza, M., Selkov, E. Jr, Kyrpides, N., Fonstein, M., Maltsev, N. and Selkov, E. “WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction,” *Nucleic Acids Research*, Vol. 28, pp. 123–125, 2000.
 - [4] Gaasterland, T. and Selkov, E. “Reconstruction of metabolic networks using incomplete information,” *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pp. 127–135, 1995.
 - [5] Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K. and Kanehisa, M. “Organizing and computing metabolic pathway data in terms of binary relations,” *Pacific Symposium on Biocomputing*, pp. 175–186, 1997.
 - [6] Mavrouniotis, M.L. “AI and Molecular Biology,” *Artificial Intelligence and Molecular Biology*. AAAI Press/MIT Press, Menlo Park, pp. 325–364, 1993.
 - [7] Forst, C.V. and Schulten, K. “Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information,” *Journal of Computational Biology*, Vol. 6, pp. 343–360, 1999.
 - [8] Tohsato, Y., Matsuda, H. and Hashimoto, A. “An application of a pathway alignment method to the analysis of metabolic pathways,” *Research Communications in Biochemistry, Cell and Molecular Biology*, Vol. 5, pp. 179–191 2003.

- [9] Bono, H., Ogata, H., Goto, S. and Kanehisa, M. "Reconstruction of amino acid biosynthesis pathways from the complete genome sequence," *Genome Research*, Vol. 8, No. 3, pp. 203–210, 1998.
- [10] Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P. "Pathway alignment: application to the comparative analysis of glycolytic enzymes," *Biochemical Journal*, Vol. 343, pp. 115–124, 1999.
- [11] Ogata, H., Fujibuchi, W., Goto, S. and Kanehisa M. "A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters", *Nucleic Acids Research*, Vol. 28, No. 20, pp. 4021–4028, 2000.
- [12] Pellegrini, M., Marcotte, E., Thompson, M. J., Eisenberg, D., and Yeates, T. O. "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 96, pp. 4285–4288, 1999
- [13] Webb, E.C. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classifications of Enzymes*. Academic Press, New York, 1992.
- [14] Peregrin-Alvarez, J. M., Tsoka, S., and Ouzounis, C. A. "The Phylogenetic Extent of Metabolic Enzymes and Pathways," *Genome Research*, Vol. 13, pp. 422–427, 2003.
- [15] Barona-Gomez, F. and Hodgson, D. A. "Occurrence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis," *EMBO reports*, Vol. 4, pp. 296–300, 2003.