社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－5 (17)
2006／6／16

反応プロファイルに基づく異種生物種間の代謝ネットワークの比較

遠里由佳子
立命館大学 情報理工学部 生命情報学科

代謝ネットワークを異なる生物種間で比較・分析することは，生物の進化や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な情報である．そこで本研究では，異種生物種間での代謝反応の有無に基づいて代謝ネットワークを比較する手法を提案する．具体的には，代謝ネットワークを，酵素反応の集合として扱い，その生物が持つ代謝ネットワークを反応プロファイルと呼ぶ代謝反応の有無により１もしくは０で構成される文字列で表現する．そして，文字列間の類似度を定義し，異種生物種間の代謝ネットワークのクラスタリングを行う．実際に，MetaCyc に登録された 33 の生物種について本手法を適用し，その有効性を確かめた．

A method for Species Comparison of Metabolic Network using Reaction Profile

Yukako Tohsato
Department of Bioscience and Bioinformatics,
College of Information Science and Engineering, Ritsumeikan University

Comparative analyses of the metabolic pathways among species give important information on evolution and on pharmacological targets. In this paper, we propose a method to compare the metabolic networks among species based on enzymatic reaction. By applying our method to the metabolic networks of 33 representative organisms selected from bacteria, archaea, and eukaryotes in the MetaCyc database, we reconstructed the phylogenetic tree that represents the similarity of metabolic network based on metabolic phenotypes.

1. はじめに

生体内の細胞は，物質を食物などの形で取り入れ，膨大な化学反応によって，それらを自らの維持や成長に必要な物質に分解・合成し，活動に必要なエネルギーを得る必要がある．代謝とは，こうした生体内で起こる化学反応の総体をさす．代謝を構成するのは，酵素を触媒として，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（以下，反応と省略する）であり，ある反応の生成物が別の反応の基質となることにより，大規模で複雑な代謝ネットワーク（metabolic network）が構成される．代謝ネットワークの既存の知見は MetaCyc [1]や KEGG [2]などのデータベースに代謝マップ（metabolic map）と呼ばれる単位でまとめられ，WWW 上で公開されている．

代謝は細胞内の過程を理解する上で重要な対象である．さらに，代謝ネットワークを異なる生物種間で比較・分析することは，進化の過程で生物がどのように代謝ネットワークを獲得したかや，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な情報である．そのため近年さまざまな研究が進められている．例えば，ゲノム情報に基づくパスウェイのクラスタリング [3, 4]，酵素の類似性よるパスウェイアライメント [5] などがある．ゲノム情報に基づくパスウェイのクラスタリングでは，パスウェイを構成する各酵素にゲノム上の遺伝子を割り上げることによりクラスタリングを行っている．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．一方，従来の rRNA などのゲノム配列に基づく分子系統解析 [6]には，進化の過程で発生する遺伝子重複などの現象が十分に反映されておらず [7]，異なった立場で構築された系統樹との比較・検討が重要視されている [8]．そこで，代謝の表現型の系統とゲノム配列に基づく系統の比較も着目され，さまざまな提案が行われている [8, 9, 10]．

以上より，本研究では，代謝の表現型に基づく異種生物種間の比較手法を開発することにより，生物種間の系統における新たな知見を得ることを目的とし，代謝ネットワークを代謝反応の集合としてとらえ，代謝ネットワーク全体をそれを構成する反応の有無を表す１と０の文字列で表し，異種生物種間で比較する方法を提案する．

以下，2章では，本研究の関連研究について述べる．3章では，アルゴリズムを提案し，4章では提案手法を用いて実際に代謝ネットワークの比較した結果とその考察を報告する．そして，最後に5章でまとめと今後の課題について述べる．

2. 既存の研究と本研究の位置付け

本研究の目的と提案する手法に関連する先行研究として，Hong らにより提案されたアルゴリズムがある [8]．Hong らにより提案された手法は，代謝ネットワークを代謝マップの分類に基づき，部分ネットワークに分類し，その部分ネットワーク中の反応量 (reaction content) p_{ij}を以下のように定義している．

$$p_{ij} = 100 \times r_{ij} / R_j \qquad (1)$$

R_jは比較の対象となるすべての生物種の j 番目の代謝マップ中の反応の数，r_{ij}は，生物 i において代謝マップ R_j に存在する反応の数である．代謝マップが N種類ある場合に，生物 iの反応量 $p_{i1}, p_{i2}, \ldots, p_{iN}$ と生物 j の反応量 $p_{j1}, p_{j2}, \ldots, p_{jN}$ の類似度 D を Pearson の相関係数で求める．そして，最長距離法によりクラスタリングを行っている [8]．

しかし，提案された反応量の定義は，代謝マップに存在する反応の数が同じならば，反応の種類が異なる場合でも区別しない．例えば，生物 S1，S2，S3 において，酵素反応 r1，r2，r3，r4 の有無が，図 1 のような関係にある場合を考える．このとき，生物 S2 と S3 の反応量はともに 75 となり，存在する酵素反応の種類が異なっている場合でも，同じスコアとなる．また，Hong らによる手法のクラスタリング結果は代謝マップの分類に影響を受ける．

そこで，本論文では以上の問題点をふまえ，次章で示すようなアルゴリズムを提案する． なお，提案する手法は，Yamada らの提案する手法 [4]とよく似ているが，手法の目的や遺伝子を基準にしている点などが異なっている．

3. 提案手法

3.1. 代謝ネットワークと反応プロファイル

代謝ネットワークを構成する酵素反応の集合として扱う．生物 Sと S'の代謝ネットワーク Nと N' の比較を考える．S と S' の生物の代謝ネットワークに含まれるすべての反応の集合 R を $R = \{r1, r2, \cdots, rn\}$ とする．このとき，複数の酵素が同じ酵素反応（基質となる化合物と生成物となる化合物が同じことを意味する）に関与する場合は，同一反応として扱い，同一の酵素が複数の異なる酵素反応に関与する場合は，それぞれを別の反応として扱う．そして，集合 R において反応の種類の重複を許さない．

	反応			
	r1	r2	r3	r4
生物 S1	1	1	1	1
生物 S2	1	1	1	0
生物 S3	0	1	1	1

図 1　生物種間の反応の有無の例

このとき，R に対して，生物 X の反応プロファイルを $Px = bx1\ bx2 \cdots bxn$ とし，生物 Xに反応 ri（$1 \leq i \leq n$）が存在する場合は，それに対応するビット bxiを 1，そうでない場合は 0 とする．

3.2. 類似スコアの定義

生物 Xの反応プロファイル $Px = bx1\ bx2 \cdots bxn$ と生物 Yの反応プロファイル $Py = by1\ by2 \cdots byn$ の類似度 $T(X,Y)$ を，Tanimoto (Jaccard) 係数法に従い，以下のように定義する．

$$T(X,Y) = \frac{N_z}{N_x + N_y - N_z} \qquad (2)$$

N_xは反応プロファイル Pxで値が 1 となっているビットの数，N_yはプロファイル Pyで値が 1 となっているビットの数，N_zは Pxと Pyがともに値が 1 であるビットの数である．Tanimoto 係数法の場合，その値は常に 0 と 1 の間になり，1 に近いほど 2 つの反応プロファイル間の類似度は高く，0 に近いほど，2 つの反応プロファイル間の類似度は低い．

例えば，図 1 の生物 S1，S2，S3 の反応プロファイルはそれぞれ，1111，1110，0111 となる．このとき，T(S1, S2) = 3/4 = 0.75，T(S2, S3) = 2/4 = 0.5 となり，S1 と S2 の反応プロファイル間の類似度のほうが S2 と S3 間の類似度よりも高くなる．

類似度の定義には Pearson の相関係数をはじめとしたさまざまな数値化の方法が提案されているが，本研究では Tanimoto 係数を用いた．Tanimoto 係数は 2 つの要素の相対的な共起の強さをよくあらわす指標である．

3.3. クラスタリング

3.2 節で定義した 類似度 T を使い，反応プロファイル間の距離尺度（非類似尺度）として以下のような類似度 $D(A,B)$ を定義する．

$$D(A,B) = 1 - T(A,B) \qquad (3)$$

そして，類似度 D をもとに全生物種間の距離行列を作成したのちに，クラスタリングを行い，類似度の高い反応プロファイルから順に統合する．クラスタリングの手法には，群平均法や重心法など，さまざまな手法があるが，本研究では最長距離法 (complete linkage hierarchical clustering method) を用いた．

4. 実験と結果

4.1. 実験と結果

提案した手法の有効性を確かめるため，実際に異種生物種間の代謝ネットワークの比較を行った．実験に使用した計算機は，Pentium M 753 MHz, 1GB RAM, OS は Windows XP SP2 である．

実験データは MetaCyc [1]に収録されているゲノム解読が終了した生物の一部を用いた．実験に利用した 33 の生物種の一覧を表 1 に示す．33 の生物種のうち，古細菌は 6 種，真正細菌は 26 種，真核生物は 1 種である．表 1 には生物名と，その略記号，その生物が持っている酵素反応の数，ORF 数をあげている．図中の酵素反応の数は，各生物種で代謝ネットワークに存在する酵素反応の数となっている．ORF 数は，KEGG [2]のデータ [11]に基づいている．

本研究では MetaCyc のデータを MySQL 上に再

表 1　実験に用いた 33 種類の生物種

生物種名		略記号	反応数	ORF数
古細菌				
1	*Archaeoglobus fulgidus DSM4304*	AfD	791	2420
2	*Methanococcus jannaschii DSM2661*	MjD	693	1786
3	*Methanobacterium thermoautotrophicum delta H*	MtD	702	1873
4	Pyrococcus furiosus DSM 3638	PfD	720	2125
5	*Thermoplasma acidophilum DSM 1728*	TaD	502	1482
6	*Thermoplasma volcanium GSS1*	TvG	773	1499
真正細菌				
1	*Aquifex aeolicus VF5*	AaV	687	1560
2	*Borrelia burgdorferi B31*	BbB	473	1639
3	*Clostridium acetobutylicum ATCC824*	CaA	896	3848
4	*Caulobacter Crescentus*	Cc	812	3737
5	*Campylobacter jejuni NCTC 11168*	CjN	728	1629
6	*Campylobacter jejuni RM1221*	CjR	682	1838
7	*Escherichia coli K-12*	EcK	1041	4226
8	*Escherichia coli O157*	EcO	855	5324
9	*Enterococcus faecalis V583*	EfV	817	3265
10	*Haemophilus influenzae KW20 Rd*	HiK	836	1657
11	*Helicobacter pylori 26695*	Hp2	542	1576
12	*Helicobacter pylori J99*	HpJ	614	1491
13	*Mycobacterium leprae TN*	MlT	745	1605
14	*Neisseria meningitidis serogroup A Z2491*	NmA	790	2065
15	*Neisseria meningitidis MC58*	NmM	800	2063
16	*Pseudomonas aeruginosa PAO1*	PaP	1093	5567
17	*Porphyromonas gingivalis W83*	PgW	796	1909
18	*Streptococcus pneumoniae R6*	SpR	848	2043
19	*Streptococcus pneumoniae TIGR4*	SpT	717	2094
20	*Streptococous thermophilus LMG 18311*	StL	762	1889
21	*Streptococcus pyogenes MGAS10394*	Sy1	874	1886
22	*Streptococcus pyogenes MGAS8232*	Sy2	768	1845
23	*Streptococcus pyogenes SF370 serotype M1*	Sy3	868	1697
24	*Vibrio cholerae N16961*	VcN	848	3835
25	*Yersinia pestis C092*	YpC	1184	4067
26	*Yersinia pestis KIM*	YpK	946	4202
真核生物				
1	*Human*	Hs	1187	26657

構築した．そして，距離行列の作成までを Perl により実装し，クラスタリングおよび系統樹の図の作成には統計処理ソフト R Ver2.3.0 [12]を利用した．その結果，得られた系統樹（クラスタ樹）を図 2 に示す．図中の省略記号（例えば，MtD や MjD）は生物名を表し，表 1 に正式な生物名と対応づけられている．

4.2. 考察

図 2 の系統樹において，古細菌の 6 種類，*Methanobacterium thermoautotrophicum delta* (MtD)と *Methanococcus jannaschii DSM2661* (MjD)，*Archaeoglobus fulgidus DSM4304* (AfD)，Pyrococcus furiosus DSM 3638 (PfD)，*Thermoplasma volcanium GSS1* (TvG)，*Themoplasma acidophilum DSM 1728* (TaD) は，同じ Cluster 1 に分類されている．また，唯一の真核生物である *human* (Hu) もこれらから遠くに位置づけされており (Cluster 2)，古細菌，真正細菌，真核生物をきれいに分類することに成功している．これは，Hong らによる手法による古細菌と真正細菌のクラスタリング結果と同じ傾向となっている[8]．

また真正細菌の 26 種類は，大きく分けて，グラム陽性菌と，グラム陰性菌であるプロテオバクテリアとそれ以外に分けられる．その中で，グラム陽性菌に分類される生物は Cluster 3 に分類された *Streptococcus thermophilus LMG 18311* (StL)，*Clostridium acetobutylicum ATCC824* (CaA)，*Streptococcus pneumoniae R6* (SpR)，*Streptococcus pneumoniae TIGR4* (SpT)，*Enterococcus faecalis V583* (EfV)，*Streptococcus pyogenes MGAS10394* (Sy1)，*Streptococcus pyogenes MGAS8232X* (Sy2)，*Streptococcus pyogenes SF370 serotype M1* (Sy3)，と，*Borrelia burgdorferi B31* (BbB) の 9 種類である．グラム陽性菌の中で，BbB 以外は比較的の近隣の Cluster 3 にまとめられている．

BbB は他の真正細菌から遠くにクラスタリング

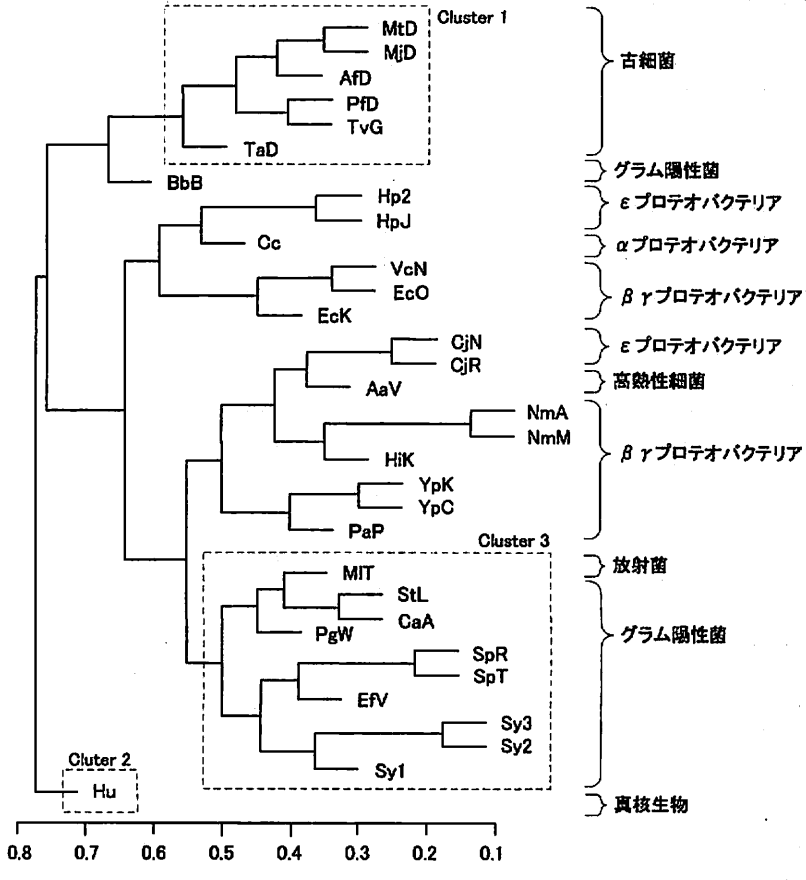

図2　33生物種の代謝ネットワークの反応プロファイルに基づくクラスタリング結果

されているが，BbBを構成する酵素反応の数は473であり，他の真正細菌でよく似たORF数を持つSy3の酵素反応の数868と比べると半分程度となっている．そのため，代謝反応のデータに欠落がある可能性が高く，その欠落がクラスタリングの結果に影響したと考えられる．

プロテオバクテリアは，α-プロテオバクテリア，ε-プロテオバクテリア，βγプロテオバクテリアに分類される．これらの生物種の代謝ネットワークは，*Neisseria meningitidis serogroup A Z2491* (NmA) と *Neisseria meningitidis MC58* (NmM)のように互いに同じ種に属する生物は近くにまとめられているが，α，ε，βγの各グループの分類は，プロテオバクテリアではない *Aquifex aeolicus VF5* (Aav) を含み，かつうまく分けられていない．この結果は，プロテオバクテリアにグラム陽性菌と似た代謝ネットワークを持つグループと，似ていないグループが存在する可能性を示している．これらの生物については，その特徴が代謝マップのどこにあるのかなど，より詳細な検証が必要だと考えられる．

5. まとめと今後の課題

代謝ネットワークを酵素反応の集合としてとらえ，反応プロファイルに基づく異種生物種間の代謝ネットワーク比較の手法を提案した．本手法は，一般的に使われている データのビット表現と，Tanimoto 係数法，クラスタリングを組み合わせており，全体として比較的簡単に実装が可能である．実際に 33 種類の生物に対して本手法を適用し，その有効性を示した．

今後，真核生物をはじめとした，さまざまな生物の代謝ネットワークのデータが充実することが見込まれる．そこで，本手法を適用することで，生物の代謝における表現型の多様化に関する知見を得ることができ，さらには，ゲノム進化と比較することで，より新たな知見を得ることが期待される．

しかしながら，ここで提案された手法に基づく代謝ネットワークの解析は，データの欠落により結果が大きく影響されるという問題がある．そのため，結果の信頼性を保障する統計的な尺度の導入が必要である．また，生物種間で同じの反応は存在しないが，類似した反応が存在する場合があり，それらの考慮も今後の課題としてあげられる．

謝辞

本研究の一部は，文部科学省ハイテク・リサーチ・センター整備事業および，2006 年度科学研究補助金（若手研究（B）課題番号 17700297）による．本研究を行うにあたり，ご意見をいただきました大阪大学大学院の松田教授，竹中助教授に深く感謝します．

参考文献

[1] Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P. and Karp, P.D., "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research*, Vol. 34, pp. D511-516, 2006.

[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] 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.

[4] Yamada, T., Goto, S. and Kanehisa M., "Extraction of Phylogenetic Network Modules from Prokayrote Metabolic Pathways," *Genome Informatics*, Vol. 15, No. 1, pp. 249-258, 2004.

[5] Tohsato, Y., Matsuda, H. and Hashimoto, A. "An application of a pathways alignment method to the analysis of metabolic pathways," *Research Communications in Biochemistry, Cell and Molecular Biology*, Vol. 5, pp. 179-191, 2003.

[6] Fitch, W.M. and Margoliash, E. "Construction of phylogenetic trees," *Science*, Vol. 155, pp. 279-284, 1967.

[7] Feng, D.F., Cho, G. and Doolittle, R.F. "Determining divergence times with a protein clock: update and reevaluation," *Proceedings of the National Academy of Science of USA*, Vol. 94, pp. 13028-13033, 1997.

[8] Hong, S.H., Kim, T.Y., and Lee, S.Y. "Phylogenetic analysis based on genome-scale metabolic pathway reaction content", *Applied Microbiology and Biotechnology*, Vol. 65, pp. 203-210, 2004.

[9] Ebenhoh, O., Handorf, T., Heinrich, R., "A Cross Species Comparison of Metabolic Network Functions," *Genome Informatics*, Vol. 16, No. 1, pp. 203-213, 2005.

[10] Clemente, J.C., Satou, K., Valiente, G. ,"Reconstruction of Phylogenetic Relationships from Metabolic Pathways Based on the Enzyme Hierarchy and the Gene Ontology," *Genome Informatics*, Vol. 16, No. 2, pp. 45-55, 2005.

[11] http://www.genome.jp/kegg/catalog/org_list.html

[12] http://www.rproject.org