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

2007－MPS－67 (47)
2007－BIO－11 (47)
2007／12／21

代謝反応パスウェイのアラインメントにおける Z値を用いたスコア補正法

西村悠†, 遠里由佳子††
†立命館大学 理工学研究科 情報理工学専攻 ††立命館大学 情報理工学部 生命情報学科

代謝反応パスウェイには多くの反応の類似が見られる．このようなパスウェイ部位の発見を目的にし，我々は化合物の構造式の類似度を用いたパスウェイのアラインメント手法を提案する．スコアの決定には Z 値を用いることにより，アラインメント長に関係なくスコアの比較を行う．この方法でパスウェイのアラインメントを行ったところ，「プリンの生合成パスウェイ」と「ピリミジンの生合成パスウェイ」との間で非常に高いパスウェイの反応の類似が見られるという結果を得ることができた．

Z-value Calculation for Metabolic Pathway Alignment Algorithm

Yu Nishimura† Yukako Tohsato††

†Information Science & Systems Engineering,
Graduate school of Science and Engineering, Ritsumeikan University
††Department of Bioscience and Bioinformatics,
College of Information Science and Engineering, Ritsumeikan University

Comparative analysis of metabolic pathways among species gives important information on evolution and on pharmacological targets. In this paper, we propose a method to align the metabolic pathways based on similarity by using the MACCS keys and Tanimoto coefficients, compute the Z-value. The effectiveness of method is demonstrated by applying the method to pathway analysis of metabolic pathways in Escherichia coli. By their results, we have found compound similarity between the purine metabolism and the pyrimidine metabolism.

1. はじめに

生体内の化学反応の多くは，酵素を触媒として，ある化合物（基質）を，別の化合物（生成物）に変換する化学反応により構成される．これらの化学反応は，ある反応の生成物が別の反応の基質となることで，複雑な反応経路のネットワークを形成する．このような一連の反応経路を一般に代謝反応パスウェイ（以下，パスウェイと省略）と呼ぶ．パスウェイを異なる生物種間や，異なる代謝反応で比較・分析することは，化合物を合成する方法についての知見を得るためや，進化の過程で生物がどのようにそのパスウェイを獲得したかを知る上で重要となる．

代謝反応パスウェイは現在の形となるまでに複雑な進化を遂げてきたと考えられており，様々な仮説が提案されている．本研究ではその仮説の１つである「パスウェイ重複」(pathway duplication) に着目する．パスウェイ重複とは，タンパク質の機能が互いに関連している一連の遺伝子がまとめてコピーされた可能性を示唆する仮説である [1]．パスウェイ重複部位の発見は，パスウェイが今日までにどのように進化を遂げてきたかを解明する重要な手がかりになると考えられる．そこで本研究では，このパスウェイ重複が起こった部位の予測手法として，パスウェイのアラインメントを提案する．

パスウェイのアラインメント手法としては，これまでに遺伝子配列 [2] や EC (Enzyme Commission [3]) 番号に基づくもの [4, 5, 6] などがさまざまな手法が提案されてきた．しかし，遺伝子配列の比較では，酵素の機能ごとにラベル付けされた EC 番号が同じでも配列がまったく異なる場合 (enzyme recruitment [7]) があることが知られており，必ずしも適切ではない．また，EC 番号を用いた比較においても，EC 番号が定められていない反応が存在することや EC 番号はあくまで人が考えて割り振ったものであるという問題がある [6]．

そこで，我々は化合物の構造式の類似性を用いたパスウェイアラインメント手法を提案した [8]．こ

れは，化合物の構造式をビット列で表現し，これを用いて反応の類似度を求め，パスウェイの類似度に拡張することによってアラインメントを行う．しかし，提案した手法ではいくつかの問題があった．それは（1）ギャップペナルティ値の設定法と（2）アラインメントのスコア補正法である．

本論文ではそれらの問題点を改善するため，適切なギャップペナルティの設定を行い，Z 値を使用することで各アラインメント長に対して個々の補正を行った．そして，実際に大腸菌の代謝反応データに適応し，その有用性を確かめた．以下，2 章では提案するパスウェイアラインメントのアルゴリズムを示し，3 章では実験結果より提案手法の有効性を検討する．最後に 4 章で結論と今後の課題を述べる．

2. 代謝反応パスウェイにおけるアラインメントアルゴリズム

2.1. 提案手法の概要

化合物の構造式の類似性に基づくパスウェイのアラインメント手法について述べる．パスウェイのアラインメントアルゴリズムは，代謝反応パスウェイを反応の系列とみなし，2 組の反応の系列が入力として与えられた場合に，動的計画法に基づくローカルアラインメントのアルゴリズムを用いて，その最適なアラインメントの組み合わせとスコアを求める．このときアラインメントは，式(1)で示す再帰式を利用する．

$$F(i,j)=\max\{F(i-1,j-1)+S(R_i,R_j),F(i-1,j)-d,F(i,j-1)-d,0\} \tag{1}$$

なお，$i=0$ のとき $F[0,j]=0$，$j=0$ のとき $F[i,0]=0$ とする．$S(R_i,R_j)$ は反応の類似度，d はギャップペナルティである．本論文で用いるアラインメントのアルゴリズムは文献 [8] に示したものと同様であり，その違いは，反応の類似度 S の算出方法とパスウェイのアラインメントスコアの補正法にある．そこで 2.2 節および 2.3 節にそれらの方法を説明する．

2.2. 反応の類似度

本研究では，反応を構成する化合物の構造式の類似度に着目し，反応を化合物の組として考え反応の類似度を定義する．化合物の構造式間の類似度を求めるために，MACCS Keys [9] という部分構造の分類を用いて，化合物の構造式をその部分構造の有無により 1 と 0 からなる 166 ビットのビット列であらわす．そして，化合物 X_i と X_j の類似度 $T(X_i,X_j)$ を，Tanimoto 係数法を用いて 0 から 1 の数で表す．このとき，1 に近づくほど 2 つのビット列間の類似度が高く，0 に近づくほど 2 つのビット列間の類似度が低いことを示す [8]．

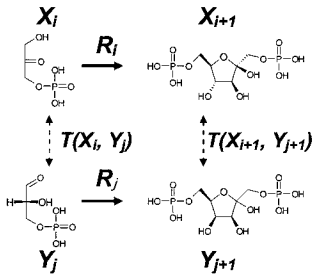

図 1　Tanimoto 係数法による反応の類似度の算出

論文 [8] では，図 1 に示す反応 R_i と R_j 間の類似スコアを求める場合，式(2)のように反応の基質と生成物におけるそれぞれの類似度を求め，その平均値を反応の類似度 $T(R_i,R_j)$ とした．

$$T(R_i,R_j)=(T(X_i,Y_i)+T(X_{i+1},Y_{j+1}))/2 \tag{2}$$

$T(R_i,R_j)$ は 0 から 1 までの数となる．しかし，この算出方法ではすべての反応類似度が正の値となるため，ギャップペナルティの値が設定できない．そこで本論文では，統計的に意味のある反応の組み合わせの場合のみその類似度がプラスになるように，すべての反応の組合せにおけるスコア $T(R_i,R_j)$ の標準偏差 T_σ と平均 T_{mean} を求め，式(3)のようにスコア $S(R_i,R_j)$ を求め，これを反応間の類似度とする．

$$S(R_i,R_j)=(T(R_i,R_j)-T_{mean})/T_\sigma \tag{3}$$

これにより，平均値よりも低い類似度には負の値を割り当て，高い類似度には正の値を割り当てることができる．また，分散を用いたことにより，類似度の偏りを考慮した上での類似度の算出及び分布を実現できる．このスコア $S(R_i,R_j)$ を反応の類似度としてアラインメント時に用いる．

2.3. Z 値を用いたスコア補正法とギャップペナルティ

ローカルアラインメントは，アラインメント長が長くなるほどスコアが大きくなりやすい．そこで，論文 [8] ではアラインメントスコアをアラインメント長で割ることにより補正を行った．しかし，アラインメント長ごとにスコアが高くなっていくが，これは必ずしも線形ではないため，アラインメント長で割る補正方法ではアラインメント結果全体を均一に補正することができない．本論文ではこれを改善するため，Z 値(標準化スコア)を用いて次の手順でアラインメントのスコアの補正を行う．

1. パスウェイのサンプリングデータを準備する．
2. サンプリングデータからアラインメント長ごとに 2000 回アラインメント結果を取り出すことを 10 回繰り返し，それらのスコアの平均と標準偏差を求めプロットする．
3. 回帰直線を用い，プロットした標準偏差を直線に近似する．
4. 各アラインメントスコアに対し，サンプリングで求めた平均値と近似直線の標準偏差を用いて，スコアと平均値の差が標準偏差の何倍かを式(4)を用いて計算する．

$$Z=\frac{X-X_m}{S_\sigma} \tag{4}$$

X_i はアラインメントのスコア，X_m はスコアのアラインメント長に対する平均値，S_σ はスコアのアラインメント長における標準偏差を表す．式(4)に示した Z 値を使用して，アラインメント結果を比較する．

この手続きの中で，サンプリングデータから作られたアラインメントされたパスウェイどうしは，進化的に関係のないパスウェイを想定したものとなっている．しかし，サンプリングデータの中には進化的に関係のあるデータが含まれる可能性があり，本来ならばそれらのデータはサンプリングデータから取り除く必要がある．しかし，配列とは違いパスウェイにおいてそれを見積もるのは困難であり，ここでは考慮していない．

3. 実験と結果

3.1. 実験データとパスウェイデータの構築

実験で使用した代謝反応パスウェイのデータは，KEGG [10] に登録された大腸菌(Escherichia coli EK-12 MG1655)のデータのうち，表 1 に示す主要な代謝マップに含まれるものを用いた (2007 年 7 月 Version 43)．表 1 には，使用した代謝マップの KEGG における ID とマップ名を示している．なお，化合物の構造式のビット列データは PubChem(http://pubchem.ncbi.nlm.nih.gov/) に記載されている各化合物の SMILES 記法(http://www.daylight.com/)のデータから MESA の Fingerprint Module [11] を用いて生成している．KEGG 中同じ代謝マップに存在する 2 つの化合物間のパスウェイは，最短経路を求める古典的なアルゴリズムの 1 種であるダイクストラ法を用いて再構築した (1 つの反応はすべて重み 1 で計算)．具体的には，reaction_main に分類される主要な反応データを用いて，その隣接関係を有向グラフとし，ダイクストラ法を用いてすべての化合物間において最短経路をパスウェイとして求めた [5, 12]．これにより，8585 通りのパスウェイを作成した．

表1　代謝マップリスト

マップID	マップ名	マップID	マップ名
map00010	Glycolysis / Gluconeogenesis	map00400	Phenylalanine, tyrosine and tryptophan biosynthes
map00020	Citrate cycle (TCA cycle)	map00450	Selenoamino acid metabolism
map00030	Pentose phosphate pathway	map00500	Starch and sucrose metabolism
map00040	Pentose and glucuronate interconversions	map00520	Nucleotide sugars metabolism
map00051	Fructose and mannose metabolism	map00530	Aminosugars metabolism
map00052	Galactose metabolism	map00561	Glycerolipid metabolism
map00130	Ubiquinone biosynthesis	map00620	Pyruvate metabolism
map00220	Urea cycle and metabolism of amino group	map00630	Glyoxylate and dicarboxylate metabolism
map00230	Purine metabolism	map00640	Propanoate metabolism
map00240	Pyrimidine metabolism	map00650	Butanoate metabolism
map00251	Glutamate metabolism	map00670	One carbon pool by folate
map00252	Alanine and aspartate metabolism	map00710	Carbon fixation
map00260	Glycine, serine and threonine metabolism	map00730	Thiamine metabolism
map00271	Methionine metabolism	map00760	Nicotinate and nicotinamide metabolism
map00280	Valine, leucine and isoleucine degradation	map00770	Pantothenate and CoA biosynthesis
map00330	Arginine and proline metabolism	map00790	Folate biosynthesis
map00340	Histidine metabolism	map00860	Porphyrin and chlorophyll metabolism
map00360	Phenylalanine metabolism	map00910	Nitrogen metabolism
map00362	Benzoate degradation via hydroxylation		

3.2. アラインメントに必要なスコア補正値の計算結果

すべての反応(2154 通り)において $T(R_i,R_j)$を求めた結果，反応の類似度 $S(R_i,R_j)$の補正に用いる平均値 T_{mean}は 0.400，標準偏差 T_σは 0.177 となった．これにより，Tanimoto 係数法でのスコアの分布が元々は 0～1 であったのに対して補正後のスコアは-2.265～3.392 となる．これを反応の類似度として用いた．

また，アラインメントスコアの補正においては，各アラインメント長において 2000 のサンプリングを取得し，これにより平均値と近似直線による標準偏差を算出してプロットした結果を図 2 に示す．この結果，傾きが 0.237，切片が 1.242 の回帰直線を得た．なお，サンプリングを 2000 ずつ取得することにおいて，アラインメント長 13 以上のアラインメント結果はアラインメントするパスウェイの総数が少なくかなり限られてしまうため，サンプリングするアラインメント長は 12 までとした．

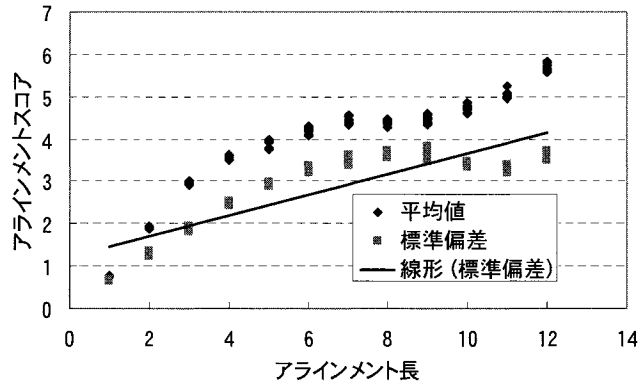

図2　サンプリングにおける平均値と標準偏差

3.3. 実験結果と考察

3.2 節で求めたパラメータに加え，アラインメントのギャップ値は 0.001 としてアラインメントを行った．その結果，最も Z 値の高いアラインメント結果は「プリンの生合成パスウェイ」と「ピリミジンの生合成パスウェイ」で検出され，Z 値は 4.793 となった．結果を以下の図 3 に示す．図 3 では，化合物に対して KEGG 上の化合物 ID を示している．また、各化合物の間に，反応に用いられる酵素名とその EC 番号，KEGG における反応 ID を示

している．各反応間の破線の矢印に対する数値は，反応の類似度を表している．また，今回の実験においては上位結果にギャップの挿入が確認できたので，ギャップが挿入された結果における最も高スコアのものを以下に図4として示す．この結果も、最高スコアのものと同じく「プリンの生合成パスウェイ」と「ピリミジンの生合成パスウェイ」である．図中の数値などは，すべて図3と同じ仕様で示している．なお，アラインメント結果の上位には同じパスウェイを含むアラインメント結果が多数見られた．そこで様々なパスウェイのアラインメント結果を確認するため，アラインメントされたパスウェイの組み合わせで同一のパスウェイを含む結果をグル

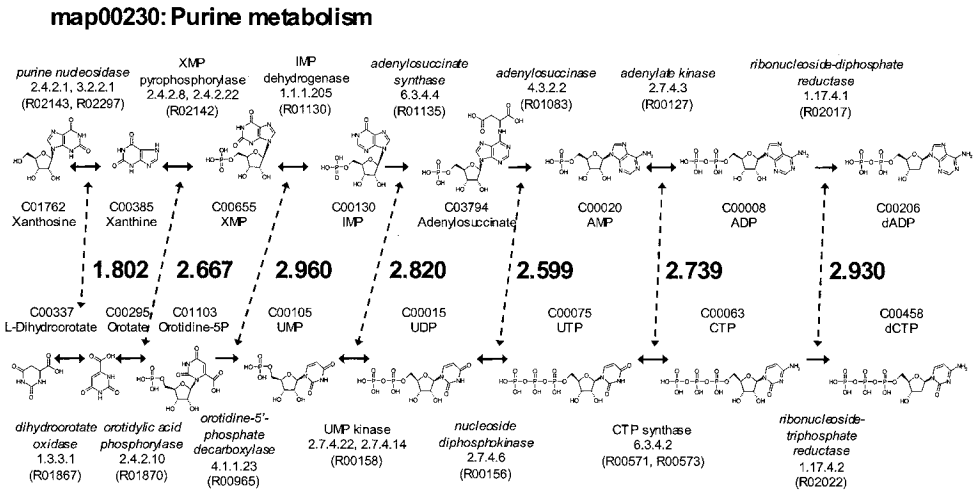

図3　最も高いZ値のアラインメント結果

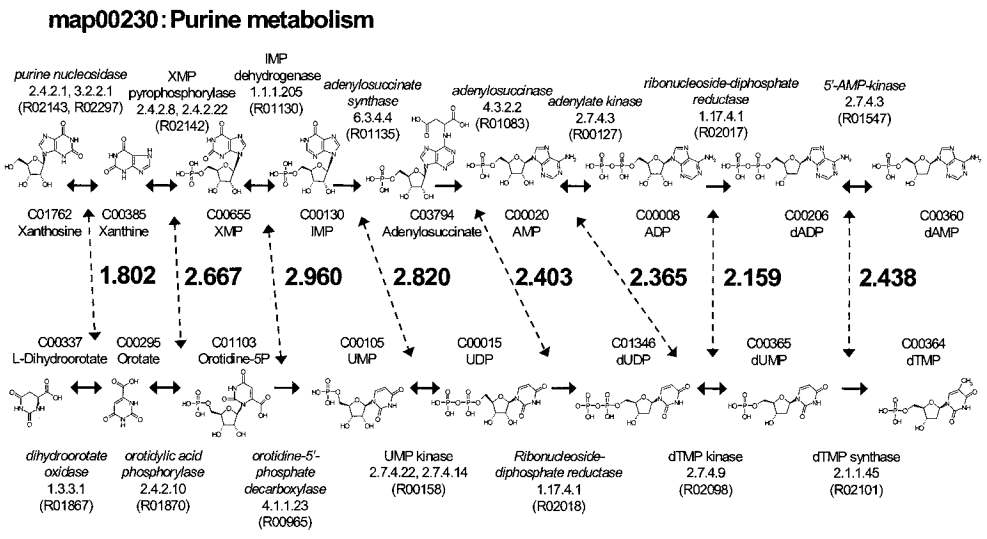

図4　ギャップが挿入されたアラインメント結果で最もZ値の高いもの

表2　アラインメント結果の上位１０ペア

Zスコア	マップID	アラインメント結果
4.793	map00230	C00206 C00008 C00020 C03794 C00130 C00655 C00385 C01762
	map00240	C00458 C00063 C00075 C00015 C00105 C01103 C00295 C00337
4.722	map00230	C00360 C00206 C00008 C00020 C03794 C00130 C00655 C00385 C01762
	map00240	C00363 C00364 C00365 C01346 C00015 C00105 C01103 C00295 C00337
4.586	map00230	C00286 C00044 C00035 C00144 C00655 C00130 C04734
	map00240	C00458 C00063 C00075 C00015 C00105 C01103 C00119
4.481	map00230	C00361 C00286 C00330 C00242 C00144 C00035 C00044 C04494
	map00240	C00239 C00881 C00526 C00106 C00105 C00015 C00075 C00063
4.471	map00051	C00118 C05378 C05345 C00644 C00392
	map00052	C00111 C03785 C01097 C06311 C01697
4.471	map00051	C00118 C05378 C05345 C01096 C00794
	map00052	C00111 C03785 C01097 C06311 C01697
4.443	map00230	C00286 C00044 C00035 C00144 C00655 C00130 C04734
	map00240	C00458 C00063 C00075 C00015 C00105 C01103 C00295
4.379	map00230	C00294 C00130 C00655 C00144 C00035 C00361
	map00240	C00299 C00105 C00015 C00075 C00063 C00458
4.377	map00230	C00286 C00044 C00035 C00144 C00655 C00130 C04734 C04677
	map00240	C00363 C00364 C00365 C01346 C00015 C00105 C01103 C00119
4.345	map00230	C00366 C00385 C00655 C00130 C04734 C04677 C04823 C04751 C03373 C04640
	map00240	C02376 C00178 C00214 C00364 C00365 C01346 C00015 C00105 C01103 C00119

ープ化し，その中で最もZ値が高いもののみを出力した．そのアラインメント結果の上位10ペアを表2に示す．この表より上位結果のほとんどは「プリンの生合成パスウェイ」と「ピリミジンの生合成パスウェイ」の組み合わせとなっていることが分かる．また，Z値で5番目となったアラインメント結果は「フルクトースとマンノース生合成パスウェイ」と「ガラクトース生合成パスウェイ」のアラインメント結果であり，これは文献[8]においてアラインメント長4で最もスコアの高いアラインメント結果と一致した．これによりギャップを含まないアラインメントに対して本論文で提案した手法は文献[8]と同様の結果が得られると考えられる．

4. おわりに

化合物の構造式に基づくパスウェイのアラインメントアルゴリズムを，Z値を用いることによってアラインメント長に関係なく比較する方法を提案した．大腸菌のパスウェイにおいて提案手法を用いたところ，「プリンの生合成パスウェイ」と「ピリミジンの生合成パスウェイ」において最もスコアの高いアラインメント結果が見られた．今後の展望としては，化合物の構造式に加えてEC番号なども視野にいれて手法の改良を行う予定である．

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，2007年度科学研究補助金(若手研究(B)課題番号17700297)による．

参考文献

[1] Schmidt, S., Sunyaev, S., Bork, P, and Dandekar, T.: Metabolites: A helping hand for pathway evolution?, *TRENDS in Biochemical Sciences* Vol.28 No.6, pp.336-341 (2003).

[2] Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P.: Pathway alignment: application to the comparative analysis of glycolytic enzymes, *Biochemical J.*, Vol.343, No.1, pp. 115-124 (1999).

[3] Webb, E. C., (Ed.): Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, Academic Press (1993).

[4] Ron Y. P., Oleg R., Esti Y. -L., and Michal Z. -U.: Alignment of metabolic pathways, *Bioinformatics*, Vol. 21, No.16, pp. 3401-3408 (2005).

[5] Clemente J.C., Satou K., and Valiente G. : Phylogenetic reconstruction from non-genomic data, *Bioinformatics*. Vol.23, No. 2, pp. e110-115 (2007).

[6] 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, Nos.3 & 4, pp. 179-191 (2003).

[7] Galperin, M. Y., Walker, D. R. and Koonin, E. V.: Analogous enzymes: independent inventions in enzyme evolution, *Genome Res.*, Vol.8, No.8, pp. 779–790 (1998).

[8] Tohsato Y., and Nishimura, Y: Metabolic pathway alignment based on similarity between chemical structures, *IPSJ Transactions on Bioinformatics*, Vol.48, No.SIG17 (TBIO3), pp.9-18(2007).

[9] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

[10] 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).

[11] MacCuish, N.E. and MacCuish, J.D.: Clustering compound data: asymmetric clustering of chemical datasets, chemometrics and cheminfomatics, *ACS Symposium Series*, Vol.894, ed. B. K. Lavine, Oxford University Press, (2005) http://www.mesaac.com/.

[12] Arita, M.: The metabolic world of Escherichia coli is not small, *Proceedings of the National Academy of Sciences USA*, Vol. 101, No.6, pp. 1543-1547 (2004).