2007－BIO－10 (2)
2007／9／13

生育温度による代謝ネットワーク構造の差異

竹本和広[1] ホセ・C・ナチェル[2] 阿久津達也[1]

1 京都大学化学研究所バイオインフォマティクスセンター
2 はこだて未来大学システム情報科学部複雑系科学科

代謝系は様々な変異によって、結果として環境に適応した表現型を獲得すると考えられている。そこで私たちは、様々な生物の代謝ネットワークの構造とその生物の成育温度の関係に注目し、成育温度とネットワーク構造の間には有意な相関があることを見出した。これは代謝ネットワークが温度によって変化することを示唆する。また、代謝ネットワークの構造を再現するモデルを提案し、そのモデルからこの構造変化の機構を考察する。結果として、この構造変化はふたつの代謝物間の反応ステップを小さくする経路の出現によることが分かった。高温においては、強い選択圧のために、この経路の出現が阻害されるため、代謝ネットワークの構造が変化すると考えられる。

Structural difference with growth temperature in metabolic networks

Kazuhiro Takemoto[1] **Jose C. Nacher**[2] **Tatsuya Akutsu**[1]

1 Bioinformatics Center, Institute for Chemical Research, Kyoto University
2 Department of Complex Systems, School of Systems Information Science, Future University-Hakodate

Most positively selected mutations cause changes in metabolism, resulting in a better-adapted phenotype. In that regard, we focus on the relationship between metabolic network structure and growth temperature, and find significant correlations in such relationship. This finding implies that metabolic networks undergo changes with temperature. In addition, we speculate on the mechanisms of the structural changes via a proposed network model. As a result, we find that the structural changes are due to the appearance of the short-cut path, which reduces the minimum distance between two nodes on a network. The metabolic networks might change because the emergence of the short-cut paths is inhibited by strong selective constraints at high temperatures.

1 研究背景

代謝系は様々な変異によって、結果として環境に適応した表現型を獲得すると考えられている。このような環境に対する適応については、温度に対する耐性がしばしば議論される。多くの生物は常温で最もよく成育するが、中には 100°C 以上の極限環境で最もよく成育する生物も存在する。このように、温度耐性は生物によって様々である。これまでに RNA の GC 含量やタンパク質のアミノ酸組成などには温度による差異が見出されている [1]。代謝はこれらの生命分子に強く依っているため、代謝系においても温度による差異が存在すると考えられる。

そこで、私たちは代謝系をネットワーク [2] として捉え、様々な原核生物について代謝ネットワークの構造特性と成育温度の相関を調査する。この調査を通してネットワーク構造の生育温度の関係を明らかにする。また、結果として、代謝ネットワークには生育温度よる差異が明らかとなり、その差異の起源についてネットワークの数理モデルを通して考察する。

2 代謝ネットワークと生育温度の関係

2.1 代謝ネットワークと生育温度

代謝ネットワークと生育温度についてはそれぞれ KEGG [3] と PGTdb [4] から得た。まず、KEGG に代謝ネットワークが登録されている生物から PGTdb によって最適生育温度が明らかとなっている 113 種の原核生物を選び出した。次に、KEGG FTP [5] から、それぞれの原核生物の代謝経路が記述されている XML ファイルをダウンロードし、代謝物をノード、それらの二項関係をエッジとして表現して代謝ネットワークを構築した。例えば、基質 S1 と S2 から生成物 P1 と P2 が生成されるような反応は図 1 のように、(S1-P1, S1-P2, S2-P1, S2-P2) というグラフで表現する。このとき化学量論係数がある場合はそれを無視する。

図 1 ある反応のネットワーク表現

また、この代謝ネットワークでは主要な代謝物（例えば糖など）の反応の流れに注目するために、反応を補助する次のような化合物を無視した：水、ATP、ADP、NAD、NADH、NADPH、二酸化炭素、アンモニア、硫酸、チオレドキシン、リン酸塩、ピロリン酸塩、そして H^+。

2.2 ネットワークの構造特性

ネットワークの構造特性については次の三つの構造特性に注目した。

■エッジ密度　これはノード数を N、エッジ数を E とすると、E/N として定義される。これは文字通りネットワークでのエッジの密度を特徴づける。

■平均クラスタ係数　これは任意のノードの最近傍ノード間の平均エッジ密度であり、グラフ理論的なモジュラリティの程度を特徴づける。ノード i のエッジ数を k_i、最近傍ノード間のエッジ数を M_i とすると、ノード i のクラスタ係数は

$$C_i = \frac{2M_i}{k_i(k_i-1)} \tag{1}$$

と定義される [2]。つまり、平均クラスタ係数は $(1/N)\sum_{i=1}^{N} C_i$ となる。

■次数指数　代謝ネットワークの次数分布 $P(k)$、エッジ数 k をもつノードの頻度、がベキ乗分布、$P(k) \propto k^{-\gamma}$、に従うことが経験的に知られている [2]。この指数 γ が次数指数であり、結合性の傾向を特徴づける。次数指数 γ が大きい場合は、ほぼ同じエッジ数を持ったノードの出現頻度が高くなるため、結合性は均一となる。逆に γ が小さい場合は、異なるエッジ数を持つノードの出現頻度が高くなるため、結合性が不均一となる。この次数指数は Newman の最尤推定法 [6] を用いて求めることができる。

2.3　結果：各構造特性と生育温度の相関

2.2 で紹介した各構造特性と生育温度の相関について調査した。相関の有意性についてはピアソンの積率相関係数 r とスピアマンの順位相関係数 r_s を用い、P 値が 0.01 より小さければ有意性ありと判定した。各相関係数と P 値については R [7] を用いて得た。

図 2 には代謝ネットワークの各構造特性と生育温度の統計的に有意な相関が示されてある。この結果は生育温度によるネットワーク構造の差異が存在することを意味する。図に示されるように、生育温度の増加に伴い、エッジ密度と平均クラスタ係数は減少し、次数指数は増加している。つまり、ネットワークは生育温度が増加することによりに次のような変化が生じると考えられる。(1) ネットワークが疎になる。(2) モジュラリティが低くなる。(3) 結合性が均一になる。

3　差異の起源に対する数理モデルを用いた考察

2 で、私たちは代謝ネットワークの構造が生育温度によって異なることを示した。次に、この生育温度による差異の起源について、数理モデルを通して仮説をたてる。

3.1　モデル

代謝ネットワークは遺伝子重複を通して進化すると考えられている [8]。遺伝子重複というイベントにより、新規のタンパク質が結果的に出現し、最終的には新規の反応が形成される。つまり、新しいエッジが出現する。この場合、次のふたつの場合が考えられる。ひとつは、新規の反応が新規代謝物-既存代謝物間に出現する (イベント I) 場合であり、もうひとつは新規の反応が既存代謝物

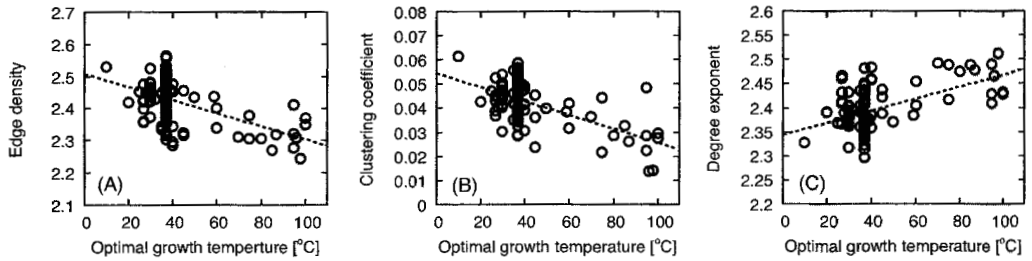

図 2　(A) エッジ密度と生育温度の相関 [$r = -0.54$ ($P < 0.01$), $r_s = -0.33$ ($P < 0.01$)]。(B) 平均クラスタ係数と生育温度の相関 [$r = -0.58$ ($P < 0.01$), $r_s = -0.30$ ($P < 0.01$)]。(C) 次数指数と生育温度の相関 [$r = 0.52$ ($P < 0.01$), $r_s = 0.27$ ($P < 0.01$)]。

間に出現する (イベント II) 場合である。これを次のようにモデル化する。確率 $1-p$ でイベント I が起こる。このとき、既存ノードをひとつランダムに選択し、新規ノードと繋げる。確率 p でイベント II が起こる。このとき、既存ノードをひとつランダムに選択し、そのノードを始点としたランダムウォークでもうひとつの既存ノードを選択する。ここでランダムウォークを採用したのは、重複ペアがネットワーク上で近接している [9] ことを考慮したためである。また、この既存-既存ノード間のエッジが出現した際に、確率 q で三角形を生成させる。

3.2　モデルの数理解析

3.2.1　次数分布と次数指数

ここで、この数理モデルの次数分布と次数指数の数理解析を示す。解析には平均場近似を利用した方法 [10] を用いる。まず、ノード i における次数 k_i の時間発展を考える。次数はイベント I が起きたとき $1/N$ の確率で 1 増加する。イベント II が起きたときは二つのノードが選択され、一方のノードはランダムに選択されるため $1/N$ の確率で次数が 1 増加する。もう一方は、ランダムに選ばれたノードを始点としたランダムウォークで選択される。ランダムウォークによって、ノード i に到達する確率は、ウォーカーのステップ数に関わらず、$k_i/\sum_j k_j$ となることが知られている [11] ので、ノード i における次数の時間発展は

$$\frac{d}{dt}k_i = (1-p)\frac{1}{N} + p\left[\frac{k_i}{\sum_j k_j} + \frac{1}{N}\right] \tag{2}$$

となる。ここで $N = (1-p)t$、$\sum_j k_j = 2t$ である。これを $k_i(t=s) = 1$ の初期条件で解く。ここで、s はノード i がネットワークに追加された時刻である。この次数の時間発展から

$$P(k) = (\gamma - 1)[A(p)+1]^{\gamma-1}[k + A(p)]^{-\gamma} \tag{3}$$

が得られる。ここで、$A(p) = 2/[p(1-p)]$ であり、次数指数は

$$\gamma = 2/p + 1 \tag{4}$$

である。

3.2.2 次数依存性クラスタ係数と平均クラスタ係数

つぎに、クラスタ係数について考える。解析には平均場近似を利用した方法 [12] を用いる。ノード i のクラスタ係数は式 (1) として定義されている。まず、ノード i の最近傍ノード間のエッジ数 M_i の時間発展について考える。M_i が増加するのはイベント II が起こり、三角形が形成された場合である。このとき M_i は三角形を構成する三つのノードで増加する。ひとつはランダム選択されたノードなので、このノードの M_i が 1 増加する確率は $1/N$ である。ほかの二つはランダムウォークによって選択されるので、それぞれのノードの M_i が 1 増加する確率は $k_i/\sum_j k_j$ である。従って、M_i の時間発展は

$$\frac{d}{dt}M_i = pq\left[\frac{1}{N}+2\frac{k_i}{\sum_j k_j}\right] \tag{5}$$

となる。3.2.1 で得た次数の時間発展を上の式に代入し、初期条件 $M_i(t=s)=0$ で解く。この M_i の時間発展を式 (1) を代入して、次数依存性クラスタ係数

$$C(k) = q\left[\frac{4}{k}-\frac{2(2-p)A(p)}{k(k-1)}\ln\frac{k+A(p)}{1+A(p)}\right] \tag{6}$$

が得られる。

平均クラスタ係数については、$P(k)$ と $C(k)$ の積の総和であるので、

$$C = \int_2^{K_m} P(k)C(k)dk \tag{7}$$

となる。ここで K_m は最大次数である。最大次数は次数の累積確率分布 $P_c(k)=\int P(k)dk$ が $1/N$ となる場合の次数、つまり $P_c(K_m)=1/N$、であるので、式 (3) より $K_m = N^{p/2}[A(p)+1]-A(p)$ となる。

3.3 パラメータの推定

ここで、私たちが提案したモデルは p と q というふたつのパラメータを持っている。これらのパラメータは次のようにして現実の代謝ネットワークから推定することができる。

まず、p である。ここで、このモデルのエッジ密度 $(=E/N)$ を考える。このモデルにおいて N と E はそれぞれ $(1-p)t$、t となる。従って、このモデルにおいて平均次数は $E/N = t/[(1-p)t] = 1/(1-p)$ となる。これを用いれば、パラメータ p は

$$p = 1-\frac{N}{E} \tag{8}$$

を用いて求めることができる。

つぎに、q である。ここではネットワークに存在する三角形の数 T を考える。このモデルにおいて三角形は近似的に確率 pq でひとつ形成されるので、$T \approx pqt$ となる。このモデルでノード数は $N=(1-p)t$ となるので、これを用いると、$T = pqN/(1-p)$ と書き直すことができる。これより、パラメータ q は

$$q = \frac{T}{N}\frac{1-p}{p} \tag{9}$$

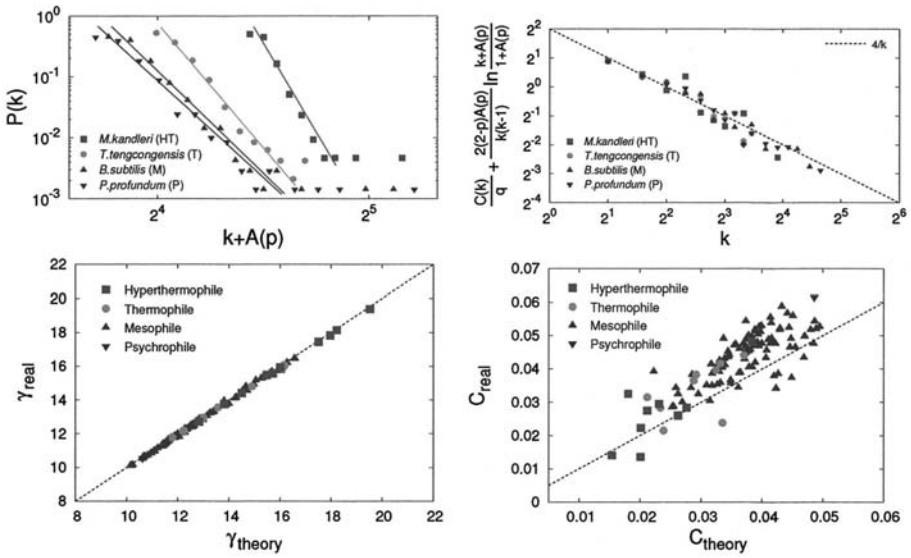

図3　モデルと現実の代謝ネットワークの比較。左上：次数分布、右上：次数依存性クラスタ係数。シンボルが現実の代謝ネットワークで、線がモデルを示している。左下：次数指数、右下：平均クラスタ係数。縦軸がモデルからの予測値、縦軸が現実の値である。破線は $y=x$ を表している。

を用いて求めることができる。

3.4　モデルと現実の代謝ネットワークの比較

ここでは 2 で調査した代謝ネットワークとモデルとの比較を行う。まず、各生物の代謝ネットワークから式 (8) と式 (9) を用いて、パラメータを得た。これを 3.2 で得られた式に代入し、モデルからの予測値を得た。なお、次数指数 γ についてはモデルとの比較がしやすいように $P(k) \propto [k+A(p)]^{-\gamma}$ を仮定し、Newman の最尤推定法 [6] を用いて得た。そのため、ここで得られた次数指数は 2.3 で示した次数指数より大きくなっている。これは、2.3 では $P(k) \propto k^{-\gamma}$ を仮定していることが原因である。また、平均クラスタ係数のモデルからの予測値 [式 (7)] は数値積分を用いて得た。

図 3 には次数分布と次数指数、次数依存性クラスタ係数と平均クラスタ係数の比較が示されている。図に示されるように、モデルと現実のデータはよく一致している。

3.5　モデルからの仮説

3.4 で、モデルは現実の代謝ネットワークの構造特性をよく再現することが示された。従って、このモデルのパラメータを用いて、生育温度による代謝ネットワーク構造の差異の起源を議論することができると考えた。

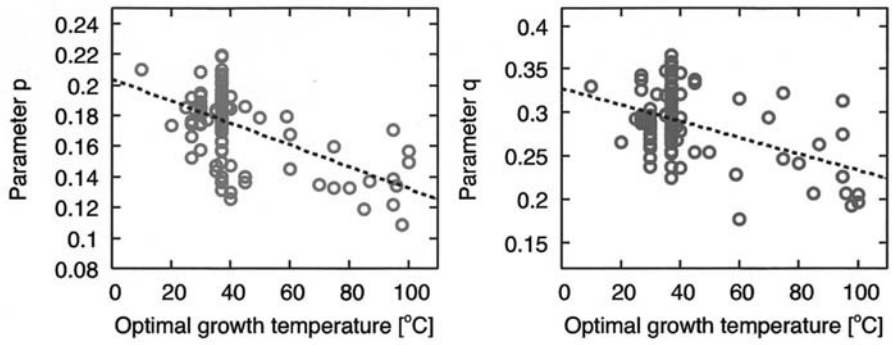

図 4　パラメータと生育温度の相関。左図：パラメータ p [$r = -0.55$ ($p < 0.01$)、$r_s = -0.33$ ($p < 0.01$)]、右図：パラメータ q [$r = -0.44$ ($p < 0.01$)、$r_s = -0.20$ ($p < 0.05$)]

図 4 には各パラメータと生育温度の相関が示されている。相関の検定については 2.3 と同様、ピアソンの積率相関係数 r とスピアマンの順位相関係数 r_s を用いている。図に示されるように、パラメータ p と q は生育温度に従って減少する。ただし、パラメータ q に関しては相関係数が大きくないことが示されている。

この結果より、ネットワーク構造の差異の起源については次のような仮説が考えられる。パラメータ p とは既存-既存ノード間にエッジが出現する頻度である。つまり生育温度が高いと、このエッジの出現が抑制される。これは、高温においてエッジの有無を決定する酵素、つまりタンパク質、に強い選択圧がかかる [13] ことが原因であると考えられる。また、パラメータ q は既存-既存ノード間にエッジが出現することにより三角形が形成される確率である。つまり、生育温度が高いと、この既存-既存ノード間にエッジは比較的長い経路のバイパスとなる。代謝ネットワークの形成において、これらの現象がネットワーク構造に差異を与えていると考えられる。

4　まとめ

私たちは、様々な温度で生育する原核生物の代謝ネットワークをグラフ理論的な方法を用いて、その構造を解析した。結果として、代謝ネットワークの構造には温度による差異が存在することを見出した。具体的には、ネットワークは温度と共に疎になり、モジュール性が低下する。また、結合性が均一になる。これらの構造変化の起源について、数理モデルを通して考えた。そのために、私たちは代謝ネットワークの構造特性をよく再現するモデルを提案した。結果として、この構造変化はふたつの代謝物間の反応ステップを小さくする経路の出現に依ることが分かった。高温における強い選択圧のために、この経路の出現が阻害されるため、代謝ネットワークの構造が変化するという仮説が得られた。

参考文献

[1] Hickey, D.A. and Singer, G.A.C. Genomic and proteomic adaptations to growth at high temperature, Genome Biology **5**, 117 (2004).

[2] Barabási, A.-L. and Oltvai, Z.N. Network biology: Understanding the cell's functional organization, Nat. Rev. Genet. **5**, 101–113 (2004).

[3] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. From genomics to chemical genomics: new developments in KEGG, Nucleic. Acids. Res. **34**, D354–357 (2006).

[4] Huang, S.L., Wu, L.C., Laing, H.K., Pan, K.T. and Horng, J.T. PGTdb: a database providing growth temperatures of prokaryotes, Bioinformatics **20**, 276–278 (2004).

[5] KEGG FTP [ftp://ftp.genome.jp/pub/kegg/xml/organisms/]

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

[7] The R Project for Statistical Computing [http://www.r-project.org/]

[8] Horowitz, N.H. On the evolution of biosynthesis, PNAS **31**, 153–157 (1945).

[9] Díaz-Mejía, J.J., Pérez-Rueda, E. and Segovia, L. A network perspective on the evolution of metabolism by gene duplication, Genome Biology **8**, R26 (2007).

[10] Barabási, A.-L., Albert, R. and Jeong, H. Mean-field theory for scale-free random networks Physica A **272**, 173-187 (1999).

[11] Saramäki, J. and Kaski, K. Scale-free networks generated by random walkers, Physica A **341**, 80–86 (2004).

[12] Takemoto, K. and Oosawa, C. Evolving networks by merging cliques, Phys. Rev. E **72**, 046116 (2005).

[13] Friedman, R., Drake, J.W. and Hughes, A.L., Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles, Genetics **167**, 1507–1512 (2004).