# 微生物工学分野へのバイオインフォマティクス

花井　泰三*・小林　元太・関口　達也・牧　幸浩・園元　謙二・岡本　正宏

近年，メタボローム，プロテオーム，トランスクリプトームなど，一度の実験で数百から数万種類のデータが得られるようになった．しかし，これらのデータを有効に利用するためには，情報科学分野で開発された技術を適応することが必要不可欠であると考えられる．このような研究分野はバイオインフォマティクスと呼ばれ，生物科学，医学および生物工学への応用が期待されている[1]．生物工学分野への応用例の一つとして，目的代謝物の生産量の向上が考えられる．目的代謝物の生産量向上のためには，代謝経路のモデル化が重要であるが，代謝経路に関わる酵素がさまざまな要素から発現制御を受けていることから，遺伝子発現制御機構の推定およびこれを組み込んだ代謝のモデル化が重要であると考えられる．このような考えに基づいて，我々はここ数年来，図1に示すように，遺伝子発現制御機構を明らかにするためのDNAマイクロアレイデータのクラスタリングおよび遺伝子相互作用ネットワーク解析，さらには遺伝子発現制御機構と代謝経路を組み合わせた動的な代謝シミュレーションの研究を行っている．

## マイクロアレイデータに対する Fuzzy *k*-means クラスタリング

遺伝子発現制御機構の解明を行うため，さまざまな実験条件下で，DNAマイクロアレイ実験などが行われる．このDNAマイクロアレイなどから得られる膨大なデータを解析する方法として，一般に解析開始時に行われるのはグループ化（クラスタリング）であり，この解析によって，パターンが類似している遺伝子などはクラスタ（グループ）に分類される．このように，数百から数万のデータを数十程度のグループに分けることで，データは理解しやすいものとなる．トランスクリプトーム解析を考えた場合，類似の発現パターンを示す遺伝子は，類似の遺伝子発現制御を受けており，機能既知遺伝子を手がかりとして，各クラスタに分類された機能未知遺伝子の機能を推定することができると考えられている．

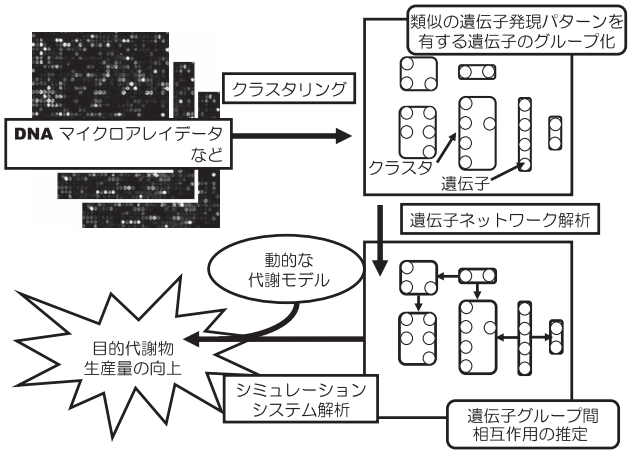


図1．微生物工学へのバイオインフォマティクスの利用

マイクロアレイデータなどをクラスタリングする際には，通常，階層型クラスタリングや*k*-means（*k*-平均）クラスタリングが用いられる[2,3]．しかし，これらの方法で，細胞内外の変化に伴う細胞の応答に関するマイクロアレイデータをクラスタリングすると，その細胞応答に関係がない遺伝子も必ずどこかのクラスタに100%の割合で属するために，クラスタ内の遺伝子を調べても，どの遺伝子がこの細胞応答に重要な遺伝子であるかがわかりにくいという問題がある（図2の上図）．これに対し，*k*-means クラスタリングにファジィ理論を組み合わせたFuzzy *k*-means クラスタリングでは，各遺伝子は各クラスタにどの程度属するのかを示す「帰属度」を持つため，細胞応答で大きく影響を受ける遺伝子はあるクラスタに高い帰属度で属し，そうでない遺伝子はさまざまなクラスタに低い帰属度で属すことができる（図2の下図）．これにより，帰属度の高い遺伝子のみに注目すれば，興味のある細胞応答に対して大きく影響を受ける，重要な遺伝子のみが抽出できるはずである．また，マイクロアレイデータには，実験誤差に起因するノイズが多く含まれているが，図2の上図で示す*k*-means クラスタリングでは，クラスタ境界付近にある遺伝子はノイズの影響によって，別のクラスタへ属してしまうが，Fuzzy *k*-means クラスタンリングの場合は，図2の下図で示すように，その遺伝子が持つ帰属度の値にわずかな変化があるものの，その他の高い帰属度を有する遺伝子は，ノイズによってあまり影響を受けないと考えられる．以上のことから，Fuzzy *k*-means クラスタリングは，マイクロアレイデータ解析のみならず，メタボローム，プロテオームデータにも広く利用可能であると考えられる．

*著者紹介　九州大学大学院農学研究院生物機能科学部門（助教授）　E-mail: taizo@brs.kyushu-u.ac.jp

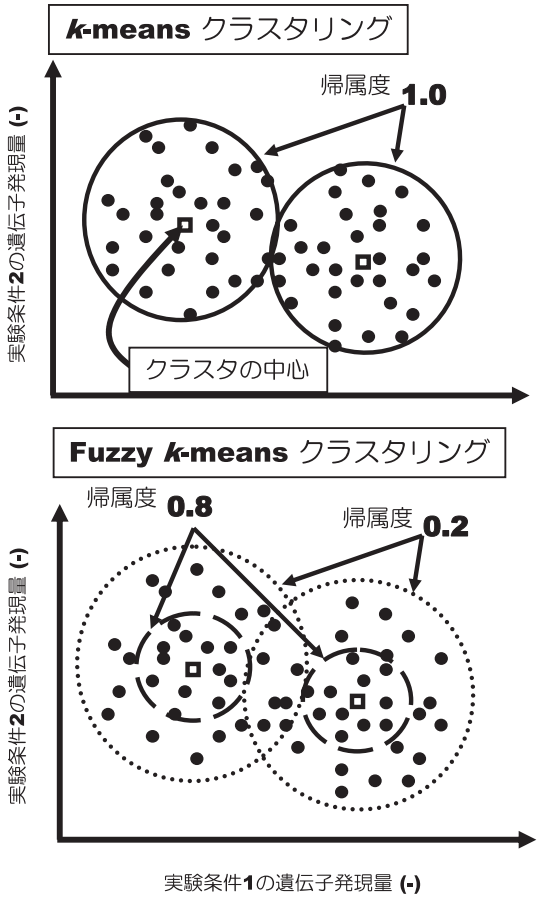


図2. *k*-meansクラスタリングとFuzzy *k*-meansクラスタリング

ここで述べたFuzzy *k*-meansクラスタリングの利点のうち，ノイズの影響に関しては現在までのところ詳しい報告がないため，我々は人工的なノイズを付加したマイクロアレイデータを用いてノイズに対する影響を調べることとした．

解析対象として，Chuら[4]のDNA遺伝子マイクロアレイによる*Saccharomyces cerevisiae*の胞子形成時の発現タイムコースデータを用いた．約6000種類の遺伝子からChuらの条件に従って，遺伝子発現レベルが著しく増加した遺伝子を抽出した．抽出した遺伝子のうち，Kupiecら[5]によって生物学的に胞子形成に関連付けられた機能を持つ45の遺伝子を選択し，解析データとした．実験で得られたこのデータをノイズのないデータ，このデータに人工的なノイズを付加したデータをノイズ付加データと仮定し，ノイズなしデータとノイズ付加データのクラスタリング結果がどの程度一致するのか（再現性）を調べることとした．

ノイズを正規分布に従って生成し，その最大値はマイクロアレイデータ値の50%および100%の値とした．Fuzzy *k*-meansクラスタリングにおいては，帰属度に閾値を設けて，閾値以上の帰属度をもつ遺伝子の再現率の計算も行った．*k*-meansとFuzzy *k*-meansクラスタリングで解析に用いたクラスタ数は，Chuらの遺伝子の分類数と同様に6とした．なお，Fuzzy *k*-meansクラスタリングで帰属度に閾値を設けない場合では，各遺伝子は最大帰属度を持つクラスタに属するとした．その結果を表1に示す．閾値を設定しない場合，*k*-meansとFuzzy *k*-meansクラスタリングの再現率は同程度であった．一方，Fuzzy *k*-meansクラスタリングでは，帰属度の閾値を0.5から0.8まで0.1刻みで上昇させたところ，帰属度の閾値の上昇に従い再現率が上昇する傾向がみられた．特に，帰属度の閾値を0.6以上とすると，ノイズが大きな場合でも99%以上の遺伝子がノイズなしの場合と同じ解析結果となることが明らかとなった．このことから，Fuzzy *k*-meansクラスタリングは，帰属度の閾値を利用することでノイズ耐性が高くなり，マイクロアレイデータ解析に有効であることが示された．

表1. ノイズを加えた際の再現性

| 方法 | 帰属度の閾値 (–) | 再現率 | |
|---|---|---|---|
| | | 最大ノイズ | |
| | | 50 （%） | 100 （%） |
| *k*-means | — | 0.942 | 0.873 |
| Fuzzy *k*-means | — | 0.953 | 0.878 |
| | 0.5 | 0.987 | 0.987 |
| | 0.6 | 0.995 | 0.993 |
| | 0.7 | 0.993 | 1.000 |
| | 0.8 | 1.000 | 1.000 |

現在は，データをいくつのクラスタに分けるべきであるのかを決定する方法[6]や遺伝子発現に関する数式モデルを利用してクラスタリングを行う方法についての研究[7]を進めている．

## 遺伝子ネットワーク解析

クラスタリング解析で得られた遺伝子のグループ間の相互作用（遺伝子発現制御機構）を明らかにするために，遺伝子（相互作用）ネットワーク解析を行う．遺伝子ネットワークの解析は，観測される遺伝子発現量のタイムコースデータなどから遺伝子（グループ）間相互作用を推定することであり，数学的には逆問題（inverse problem）[8]と考えられる．相互作用ネットワークを連立微分方程式でモデル化する方法が一般に用いられるが，現段階では遺伝子間の詳細な相互作用に関する知見が十分でなく，通常のモデル化に利用される一般質量作

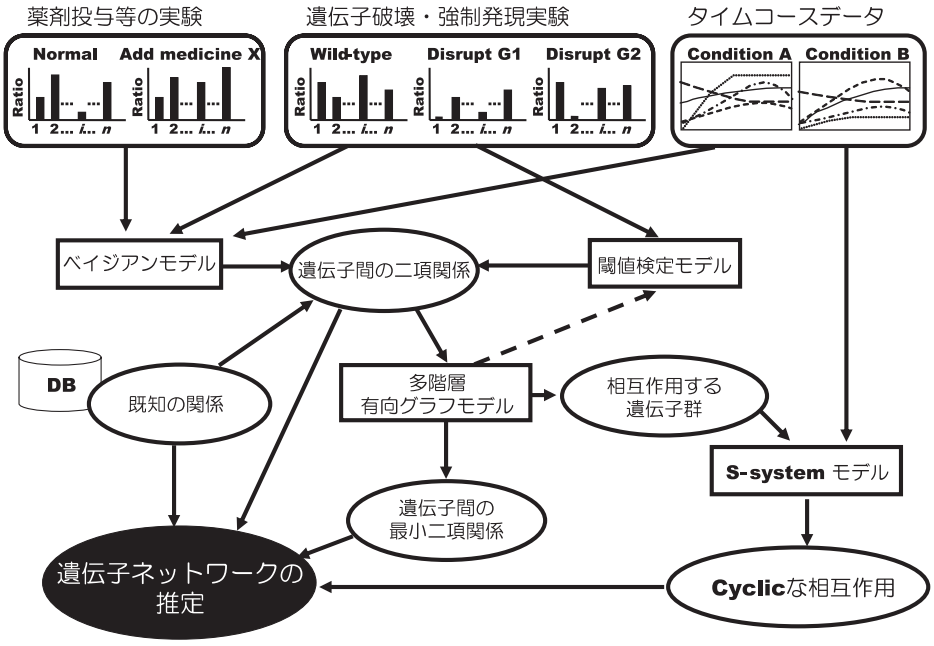


図3. モデルを組合せた遺伝子ネットワーク推定法

用則（generalized mass action law: GMA）による表記は不適当である．我々の研究グループは，これまで逆問題解決のための革新的な突破口として，微分方程式の立式に，べき乗則に基づいたS-systemモデル[9]を，観測データを再現する多数の内部パラメータの自動推定法に実数値遺伝的アルゴリズムを適用する方法を提案してきた[10]．S-systemモデルは次のようなものである．$n$個のシステム構成要素（状態変数：遺伝子ネットワークの場合は遺伝子または遺伝子グループに相当）$X_i$（$i$=1,2,…, $n$）の値（遺伝子発現量に相当）が時間的に変動し，$X_i$同士が相互作用しているネットワークシステムを考える．

$$\frac{dXi}{dt}=\alpha_i\prod_{j=1}^{n}X_j^{g_{ij}}-\beta_i\prod_{j=1}^{n}X_j^{h_{ij}}$$

（$i$=1,2,…, $n$）

この式において，$g_{ij}$は状態変数$X_i$の生成過程に関与する状態変数$X_j$の相互作用係数であり，同様に$h_{ij}$は$X_i$の分解過程（消費過程）に関与する$X_j$の相互作用係数である．たとえば，$g_{ij}$が正の値なら，$X_i$の生成過程に対し$X_j$は＋の作用を及ぼし，同様に$h_{ij}$の値が負なら，$X_i$の分解過程に対し$X_j$は－の作用を及ぼすことになる．$\alpha_i$，$\beta_i$は，それぞれ$X_i$の生成項，分解項に乗じる係数である．この式は，状態変数$X_i$の生成過程（右辺第1項）と分解過程（右辺第2項）に考えているすべての状態変数$X_j$（$j$=1,2,…, $n$）が関与していると仮定する全結線モデルである．$X_i$の生成過程（あるいは分解過程）に$X_j$が関与していない（相互作用がない）場合，$g_{ij}$（あるいは$h_{ij}$）の値はゼロということになる．しかし，生成過程，分解過程がそれぞれ1つの項で表現されているため，生成項，分解項が複数の経路で構成されている場合は，GMAを近似した表現になる．現在のところ，それぞれの遺伝子のmRNAの生成過程，分解過程の詳細な機構は明らかになっておらず，この近似表現法は有効なものと思われる．つまり，$g_{ij}$，$h_{ij}$の値を推定することで，相互作用ネットワークが推定できる．このようなS-systemモデルを用いた相互作用推定を含めて，我々は，図3で示すように，マイクロアレイデータに応じて推定モデルを組み合わせて，段階的にネットワークを推定する戦略を考案してソフトウエア化している[11]．

現在は，複数の時系列データから，影響の大きい遺伝子相互作用を推定する方法の検討[12]を行っている．

### アセトン・ブタノール発酵生産のシミュレーション

回分培養など細胞内外の条件が連続的に変化する場合は，それに応じて遺伝子発現量も連続的に変化する．そのため，明らかにした遺伝子発現制御機構を組み込んだ代謝のモデル化のためには，このような連続的な変化に対応する動的なモデルが必要となる．我々は，このような考えに基づいて，アセトン・ブタノール（ABE）発酵の動的なモデルの構築を行っている[13]．ABE発酵は，酸生成期には酢酸，酪酸を生産し，ソルベント生成期にはアセトン，ブタノール，エタノールを生産する複雑な代謝経路を持っている．そのため，発酵の制御が難しく，効率的な発酵システムは未だ構築されていない．

これまでに我々は，代謝制御経路解析用シミュレータ

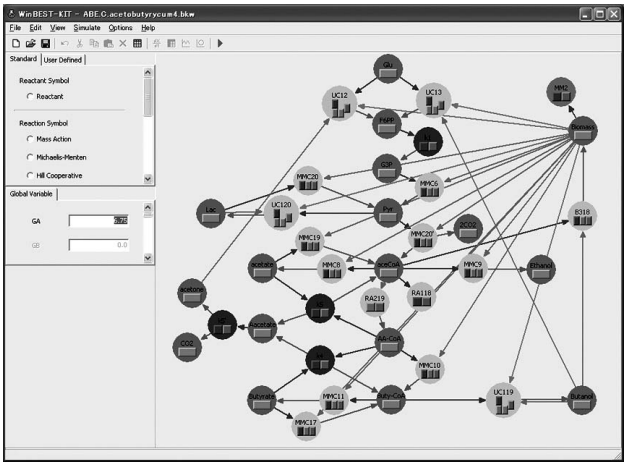

図4. WinBEST-KITによるABE発酵のモデル化

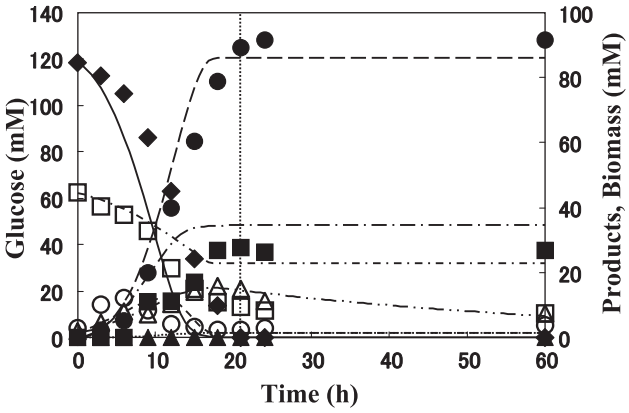


図5. 初発グルコース濃度119 mMにおけるABE発酵の経時変化（各プロット）とシミュレーション結果（実線・破線）

WinBEST-KIT を独自に開発してきた[14]．このシミュレータを使用すれば，解析対象の反応系を視覚的に構築でき，GMAに基づく立式やMichaelis-Menten式に代表されるenzyme kinetics関数近似式以外にも，ユーザが独自に関数式を定義でき，簡単にシステム解析を行うことが可能である．既知の代謝経路に基づき，このWinBEST-KITを用いてABE発酵のモデル化を行った（図4）．モデル内のパラメータは，文献値を参考として，*Clostridium saccharoperbutylacetonicum* N1-4をJar Fermenterにて，初発グルコース濃度62.0 mMで回分培養した際の実験値を再現するように決定した．パラメータ決定する際に利用していない初発グルコース濃度 35.8，119，298 mM で回分培養したときの実験データを，モデルで再現可能かを確認した．

その結果，既知の代謝経路のみを考慮した場合では，実験結果をうまく再現できなかった．そのため，酪酸の再同化がCoA Transferase（CoAT）経路および酪酸生成経路の逆経路で行われる，グルコース濃度が1 mM以下でエネルギー生産・消費が停止する，という2つの仮定を考慮してモデルを再構築した．その結果，各物質濃度の時間的挙動がモデルと定性的に一致し（図5），本モデルの有用性が検証できた．このモデルは，定常状態を仮定したいわゆる代謝流束解析とは異なり，培養状態などにより刻一刻と変化する代謝物の動的な変化を常に把握できる．このため，発酵生産過程全体のボトルネックの同定，ブタノール生産を最大化させるための培養条件の設定などに利用可能である．

現在は，発酵生産に関する主代謝経路の酵素活性を測定し，この時間的変化を表すため，遺伝子発現制御機構を組み込んだ代謝のモデル化などを進めている．

## 今後の課題

本稿では，微生物工学分野へのバイオインフォマティクスの紹介として，我々のグループが取り組んでいるクラスタリング，相互作用ネットワーク解析，動的なシミュレーションなどの現状を紹介した．現在までのところ，トランスクリプトームやメタボロームのデータは個々に解析されているため，前述の方法で明らかにした遺伝子発現制御機構を動的な代謝モデルと組み合わせることが望まれる．さらに，プロテオームやゲノムで得られたデータや知見を統合的に理解するためのアルゴリズムの開発が必要である．このような検討や開発を重ねることで，バイオインフォマティクスによる解析結果を，積極的に実験にフィードバックするIT駆動型の微生物を利用する工学分野・微生物を対象とする科学分野の研究が進んでいけば，と願っている．

## 文　献


1) Hanai, T. *et al.*: *J. Biosci. Bioeng.*, **101**, 377 (2006).
2) Eisen, M. B. *et al.*: *Proc. Natl. Acad. Sci. USA*, **95**, 14863 (1998).
3) Tavazoie, S. *et al.*: *Proc. Natl. Acad. Sci. USA*, **94**, 4262 (1997).
4) Chu, S. *et al.*: *Science*, **282**, 699 (1998).
5) Kupiec, M. *et al.*: *The Molecular and Cellular Biology of the Yeast Saccharomyces*, p.889, Cold Spring Harbor Laboratory Press (1997).
6) Arima, C. *et al.*: *Genome Informatics*, **16**, P040-1 (2005).
7) Hakamada, K. *et al.*: *Bioinformatics*, **22**, 843 (2006).
8) 富永大介，岡本正宏：化学工学論文集, **25**, 220 (1999).
9) 岡本正宏：ゲノム情報生物学, p.165, 中山書店 (2000).
10) 岡本正宏，小野　功：人工知能学会誌, **18**, 502 (2003).
11) Maki, Y. *et al.*: *J. Bioinform. Comput. Biol.*, **2**, 533 (2004).
12) Nakatsui, M. *et al.*: *Genome Informatics*, **16**, P148-1 (2005).
13) Shinto, H. *et al.*: *Genome Informatics*, **16**, P120-1 (2005).
14) Sekiguchi, T., Okamoto, M.: *J. Bioinfo. Comput. Biol.*, in press.