公募研究：2000～2003年度

微生物における転写調節系の比較ゲノミクス・プロテオミクス

●藤田 信之

国立遺伝学研究所分子遺伝研究系　（現所属：製品評価技術基盤機構バイオテクノロジー本部ゲノム解析部門）

〈研究の目的と進め方〉

微生物の転写調節系を比較ゲノミクス・プロテオミクスの立場から、また情報科学、実験科学の両面から解析することにより、ゲノムや調節系の進化、様々な外的要因に対する微生物の生存戦略、適応戦略、さらには病原性の分子基盤などを明らかにすることを目的とする。

バクテリアだけをとっても既に50種以上（2002年秋時点）について全ゲノム配列が報告されており、それぞれ独自のデータベースとして公開されている。しかし大半のデータベースは、ホモロジー検索をもとにしたオーソログの記載など、半機械的なアノテーションが行われているのみである。また、大腸菌や枯草菌など、過去に遺伝学的、生化学的な研究の蓄積がある生物種においても、それらの知識がデータベースに十分生かされているとは言い難い。

転写因子は種間の保存性が低いこと、複雑なドメイン構成を持つものが多いこと、転写因子以外の蛋白質とドメインを共有している場合が多いことなどから、ホモロジースコアのみによって機械的に同定、分類することは困難であり、また危険でもある。現実に、不適切なアノテーションが他のゲノムのアノテーションに伝播、拡大することによって、研究の現場で少なからず混乱を生じはじめている。一方、ゲノム横断的に蛋白質のモチーフやオーソロググループを抽出しデータベース化することもいくつかのグループで行われているが、用いられている検出や分類の方法は必ずしも転写因子に適したものとはなっていないために、多くの記載もれがあったり不適切な分類がなされている。そこで、転写因子に的を絞った種横断的なデータベースの構築をめざし、そのために必要なデータベース設計、コンピュータ解析、文献調査、実験データの生産を総合的に行なおうとするのが、本研究の特色である。また、得られた成果をもとに、新規のゲノム配列について転写因子の同定、分類、機能予測を高精度に行なうシステムの開発をめざす。あわせて、ゲノム間の比較により、それぞれの生物における多様な転写因子のレパートリーがどのようにして獲得されてきたかを、種分化後の遺伝子重複、ファージ等を介した水平伝達、ドメイン・スワップなどの様々な可能性を含めて考察する。様々な生物種について、大規模な遺伝子発現解析やプロテオーム解析が進められているが、それらのデータの多くは、直接もしくは間接に、特定の転写因子と関連付けて記述することができるはずである。これらのより広範なデータを集約するための核としても、転写因子を中心としたデータベースは大きな意義を持つものと思われる。

〈研究開始時の研究計画〉

1．ゲノム配列上での転写因子のサーベイと分類

全ゲノム配列が公開されている微生物（特にバクテリア）を対象として転写因子のサーベイを行なう。遺伝学的、生化学的な研究が最も進んでいる大腸菌について、文献調査を含めた予備的なサーベイを行なった結果、転写因子の数は200-250程度（全遺伝子の5%程度）と見積もられた。ほとんどの転写因子はDNA結合ドメインと調節ドメイン（リン酸化ドメイン、補助因子結合ドメイン等）を持っており、その大半がhelix-turn-helix型のモチーフを使ってDNAに結合する。すなわちほとんどの転写因子が大なり小なりよく似た構造を持っており、全体として階層的な蛋白質ファミリーを形成している。このことは、ホモロジー検索による転写因子の同定をある程度容易にしている反面、ホモロジースコアのみからオーソログを推定することを困難にしており、不適切なアノテーションの一因ともなっている。そこで転写因子の同定と分類にあたっては、上記のような転写因子の構造、機能上の特性に十分な注意を払うとともに、必要によって実験的な検証を取り入れる。

2．未知の転写因子についての機能解析

最もよく解析がすすんでいる大腸菌や枯草菌でも、転写因子と推定されている蛋白の半数近くが、標的遺伝子さえ明らかではない状況である。ゲノム配列の解析から予想された未知の転写因子について機能を推定するため、進行中のポストゲノムプロジェクトとの連携のもとに遺伝子発現解析、プロテーム解析を行ない、支配下にある遺伝子（群）の同定と調節様式の解明を行なう。

3．データベースの構築と公開

まず実験データが比較的豊富ないくつかの代表的な生物種に的をしぼって、リレーショナルデータモデルによるデータベースのプロトタイプを作成する。転写因子相互間の関係については、単純なオーソログ、パラログの記載だけではなく、ドメイン構成、高次構造、標的遺伝子、調節様式などの類似性を柔軟に記述できるように設計する。その後データベースの対象を、ゲノム配列が公開されているすべてのバクテリアに拡張する。また、可能な限り他データベースへの関連付けを行なう。データベースはWorld Wide Web上で公開する。

4．転写因子自身のプロテオーム解析

転写因子の中にはリン酸化などの翻訳後修飾やプロセシングによる調節を受けているものがあるため、転写因子自身を対象としたプロテオーム解析が今後重要になってくるものと思われる。しかしながら転写因子は一般に細胞内濃度が低く、また多くが塩基性蛋白であるため、現在一般に用いられている方法での解析は極めて困難である。有効な濃縮方法や新しい分離手法の開発を含めて、転写因子自身のプロテオームに迫る。

〈研究期間の成果〉

1．ゲノム配列上での転写因子のサーベイと分類

全ゲノム配列が明らかとなっている広範な微生物（主にバクテリア）を対象とし、ホモロジー検索、プロファイル検索、文献調査などを通して、ゲノム横断的な転写因子のサーベイと分類を行った。初年度は大腸菌、枯草菌を含む14種のバクテリアを対象とし、その後主な分類群を網羅する形で徐々に種の数を増やし、最終的には49種のバクテリアまで拡張した。また比較のために古細菌

11種と真核生物2種（出芽酵母とシロイヌナズナ）を加えた。COG、Pfamなどの既存のデータベースにも一部転写因子の記述があったが、網羅的ではなく、また転写因子と他のタンパク質ファミリーの切り分けが十分でないなど、本研究の目的には合致しなかった。そこで、実験的に機能が確かめられている転写因子をシードとし、ドメイン単位で、ホモロジー検索→アラインメント→マニュアルキュレーション（転写因子としての特徴的な構造が保存されているか等の吟味）→プロファイル作成のサイクルを繰り返すことによって、検出感度と精度の向上を図った。

上記の方法により、49種のバクテリアについて転写因子と想定されるORFを網羅的に収集した結果、8,000余りのORFが得られた。これらをドメインごとの構造およびドメインの組み合わせによって分類した結果、60以上のタンパク質ファミリーに分類できることがわかった。
各転写因子ファミリーについて、アミノ酸配列に基づいた系統解析を行なった。その結果、多くの転写因子ファミリーは相当に古い起原を持つと推定された。一方では、配列の類似性から一義的にオーソログを定義できるのは、例外的なケースを除けば、近縁のバクテリア種（プロテオバクテリアγ群、同α群、低GCグラム陽性菌など）の間に限られた。したがって進化の過程で転写因子の構造そのものは高度に保存されてきたものの、具体的な機能（標的遺伝子）は種内での遺伝子重複や水平伝達などの過程で比較的柔軟に獲得されてきたものと推測された。

個々の転写因子の内在性、外来性を評価する目的で、コドンの使用頻度に基づく主成分分析を行なった。その結果、外来性の程度すなわち水平伝達の頻度は、転写因子ファミリー間で大きく異なり、大半の転写因子ファミリーは主にゲノム内での遺伝子重複によって進化してきたものと推定された。

バクテリアにおいては、どの分類群に属するかにかかわらず、ゲノムサイズと転写因子の数との間には明らかな正の相関が見られた（図1）。ゲノムサイズの減少とともに転写因子の数は急激に減少し、ゲノムサイズが1.5M程度以下の生物種では、極端に低い値を示した。一方ゲノムサイズの大きな根粒菌、緑膿菌、放線菌などでは、TetRファミリー、LysRファミリーなどの特定の転写因子ファミリーの大規模な重複が、転写因子の増加に大きく寄与していることがわかった。またラン藻およびマイコバクテリウムは他の分類群に比較して転写因子の数が少ない傾向が見られた。

収集された配列をもとに、各転写因子ファミリーについて高品位なマルチプルアラインメントを作成し、ドメインごと（ドメイン情報がないファミリーについては保存領域ごと）のHidden Markov Model（HMM）を構築した。HMMを使ったドメインの検索結果とドメインの組み合わせ情報を使って、新規のゲノム配列から転写因子を予測し、分類するシステムの開発を行った。

２．未知の転写因子についての機能解析

転写因子の標的遺伝子の同定にはDNAマイクロアレイが有効と考えられるが、機能未知の転写因子の場合、培養条件の至適化など克服すべき問題が多い。また直接の影響か間接の影響かを見極めるのも難しい。そこでSELEX法を改良し、in vitroでの特異的なDNA結合能を利用して、ゲノムのショットガンライブラリから標的配列を選別する方法を構築し、大腸菌のLacIファミリーの転写因子2種類（うち1種は機能未知）を使ってこの方法の精度について評価を行った。いずれの転写因子もゲノム配列上の１ヶ所に極めて特異的に結合することがわかり、この方法の特異性の高さが確かめられた。またこの方法を大腸菌の転写因子CRPに応用し、ゲノム上に新規のCRP結合領域を15個所以上同定することができた。これらのDNA領域の大半は、200bp程度以上の非翻訳領域を含み、またDNaseIフットプリンティングによってCRPおよびRNAポリメラーゼの特異的な結合を確認することができた。さらに共同研究として、大腸菌の二成分制御系の転写因子や他のLacIファミリー転写因子について、この方法を使った標的遺伝子の同定を行った。今後大腸菌の機能未知の転写因子（Y遺伝子産物）についても網羅的に標的配列の同定を実施する予定である。

３．データベースの構築と公開

二成分制御系の転写因子をモデルとして、転写因子データベースのプロトタイプの開発を行なった。転写因子間の比較においては、配列の類似性だけでなく、分子系

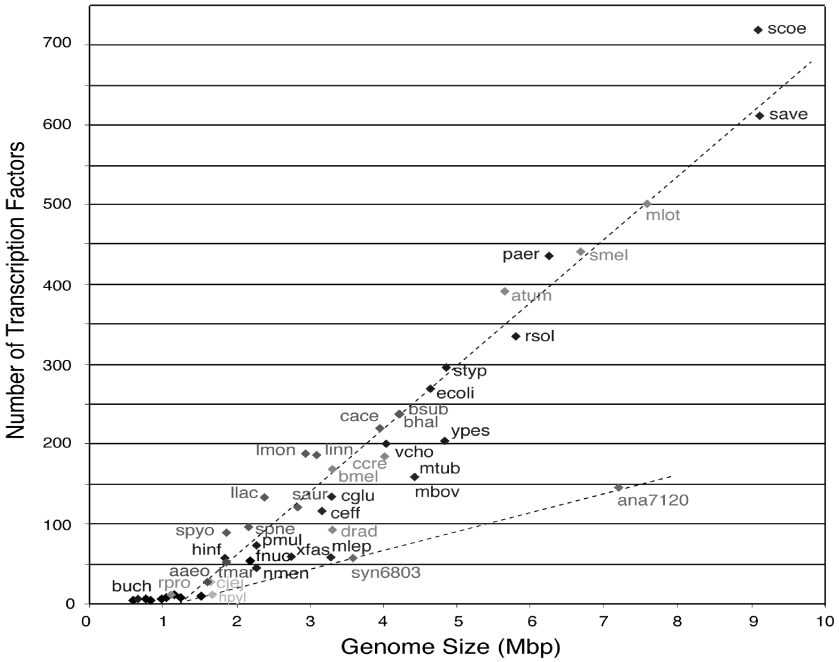

図１　ゲノムサイズと転写因子数との相関

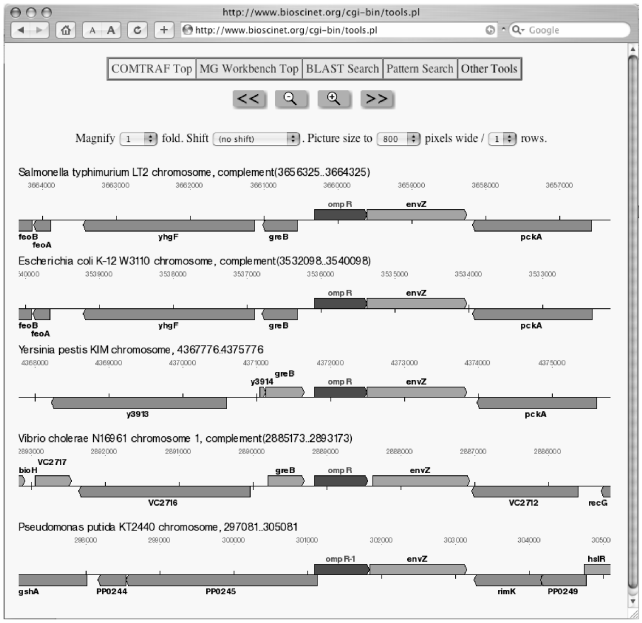

図２　Microbial Genome Workbench（オペロン構造の比較）

統樹上での位置関係、オペロン構成、ドメイン構成など様々な視点から比較を行なって結果をグラフィカルに提示できるように工夫した。また転写因子データベースのためのプラットフォームとして開発した微生物ゲノムの検索、閲覧、比較、解析のためのツールをMicrobial Genome Workbench (http://www.bioscinet.org/mgw/) として公開した（図２）。

４．転写因子自身のプロテオーム解析

通常の二次元電気泳動もしくはラジカルフリー高還元性二次元電気泳動（RFHR）とMALDI-TOF質量分析を組み合わせたペプチドマスフィンガープリント法による転写因子の分離・同定を試みたが、存在量が少ないこと、および塩基性領域での分離が不十分なことから満足のいく結果は得られなかった。

大腸菌で発現させた個々の転写因子について抗体を作成し、ウェスタンブロット法によって細胞抽出液中の量を定量する方法について検討を行った。検出のダイナミックレンジは狭いものの、検出範囲内であれば十分な定量性が得られることがわかった。また複数の抗体を組み合わせて使うことで、複数のタンパク質の同時測定も可能であった。この方法により、マイナーシグマ因子やいくつかの転写因子について、細胞内の濃度および培養条件による変動を測定した。

〈国内外での成果の位置づけ〉

タンパク質のドメインやオーソロググループを網羅的に抽出してデータベース化することはいくつかのグループで行なわれているが、用いられている検出や分類の方法は必ずしも転写因子に適したものとはなっていないために、 転写因子に関しては多くの記載もれがあったり不適切な分類がなされている。また転写因子に的をしぼった研究もあるが、PROSITE等の既知のデータベースを出発点としているために、網羅的なものとはなっていない。本研究では49種のバクテリアゲノム配列から、転写因子と見なされるORFを網羅的に抽出し、その結果をもとに、転写因子に特化したドメインごとのHidden Markov Modelのセットを作成した。helix-turn-helix型のDNA結合ドメインをより細かく分類するなどの工夫により、Pfam等の既存のドメインデータベースと比較して、より高い検出感度と精度が得られている。さらに転写因子ファミリーごとのドメインの組み合わせ情報を付加することにより、精度の高い転写因子の予測と分類が可能となった。

SELEX法はもともとは人工合成したランダムな配列からタンパク質等への結合性を持つ配列を選別・濃縮する方法である。シグマ因子やその他の転写因子への応用例も報告されているが、実在の標的配列とは異なった配列に収斂している場合が多い。これは実際の結合配列が必ずしも結合強度において最適化されたものではないことを意味している。そこでライブラリーの作成方法やPCRの条件等を工夫することにより、ゲノムのショットガンライブラリーを出発点として結合配列を選別・濃縮するオリジナルな方法を開発した。

〈達成できなかったこと、予想外の困難、その理由〉

研究代表者の所属が年度途中で変更になり、当初計画のまま研究を継続することが不可能となった。また、当初予想していたよりも遥かに速いスピードで微生物のゲノム解析が進み、研究着手時には20種程度だったバクテリアのゲノム情報が、研究終了時には120種を超えるまでに激増した（現在は300種近くまで増えている）。そのため、当初計画した全ての微生物を網羅したデータベースを構築し、維持していくことは不可能と判断せざるを得なかった。それまでの成果を最大限に活用するため、公開する情報リソースの主なターゲットを、アノテーション情報を主体としたものから、プロファイルデータベースおよびその利用技術に切りかえることにした。

In vitroでのDNA結合能を指標として転写因子の標的遺伝子を同定する方法は、LacIファミリーの転写因子やCRPについては期待以上の成果を示したが、補助因子を必要とする場合やリン酸化等の修飾を受ける場合への対応など、解決すべき問題点も多いことがわかった。

〈今後の課題〉

当初の目標のひとつであった転写因子に特化したデータベースの構築は、諸般の事情から断念せざるを得なかったが、その過程で開発したMicrobial Genome Workbenchは好評であり、現在でも固定のユーザがある。今後公的な研究費の補助は見込めないものの、できる範囲でアップデートに努めたい。

終了時までの研究の成果を最大限に生かすため、転写因子に特化したHMMのセットを作成した。所属が変更となった後も拡張と改良を続けており、現在のところ82種類の転写因子ファミリーを網羅する112個のHMMモデルおよびドメインの組み合わせ等を規定する約50のルールから構成されるまでになっている。現職場（NITEゲノム解析部門）においては新規な微生物ゲノムのアノテーションに日常的に使用しているが、より広く利用できるように、早急に公開をしたい。現状では利用可能な計算機資源が限られるため、オンラインでの検索サービスを提供することは難しいが、HMMR等の既存のソフトウェアで利用できるパッケージとして配布することを計画中である。

〈研究期間の全成果公表リスト〉

１）論文／プロシーディング

1.
Maeda, H., Jishage, M., Nomura, T., Fujita, N. and Ishihama, A.: Two extracytoplasmic function sigma subunits, E and FecI, of Escherichia coli: Promoter

selectivity and intracellular levels. J. Bacteriol., 182, 1181-1184, (2000).
2.
Katayama, A., Fujita, N. and Ishihama, A.: Mapping of subunit-subunit contact surfaces on the b' subunit of Escherichia coli RNA polymerase. J. Biol. Chem., 275, 3583-3592, (2000).
3. 0202261058
Fujita, N., Endo, S. and Ishihama, A.: Structural requirements for the interdomain linker of a subunit of Escherichia coli RNA polymerase. Biochemistry, 39, 6243-6249, (2000).
4.
Yamamoto, K., Nagura, R., Tanabe, H., Fujita, N., Ishihama, A. and Utsumi, R.: Negative regulation of the bolA1p of Escherichia coli K-12 by the transcription factor OmpR for osmolarity response genes. FEMS Microbiol. Lett. 186, 257-262, (2000).
5.
Ohnuma, M., Fujita, N., Ishihama, A., Tanaka, K. and Takahashi, H.: A carboxy-terminal 16- amino-acid region of σ38 of Escherichia coli is important for transcription under high-salt conditions and sigma activities in vivo. J. Bacteriaol., 182, 4628-4631, (2000).
6.
Wigneshweraraj, S.R., Fujita, N., Ishihama, A. and Buck, M.: Conservation of sigma-core RNA polymerase proximity relationships between the enhancer-independent and enhancer-dependent sigma classes. EMBO J., 19, 3038-3048, (2000).
7.
Maeda, H., Fujita, N. and Ishihama, A.: Competition among seven Escherichia coli sigma sbunits: relative binding affinities to the core RNA polymerase. Nucleic Acids Res., 28, 3497-3503, (2000).
8. 0202261154
Yamamoto, K, Yata, K., Fujita, N., Ishihama, A.: Novel mode of transcription regulation by SdiA, an Escherichia coli homologue of the quorum-sensing regulator. Mol. Microbiol., 41, 1187-1198, (2001).
9. 0202261236
Shin, M., Kang, S., Hyun, S.J., Fujita, N., Ishihama, A., Valentin-Hansen, P., Choy, H.E.: Repression of deoP2 in Escherichia coli by CytR: conversion of a transcription activator into a repressor. EMBO J. 20, 5392-5399, (2001).
10. 0303271829
Ozoline, O.N., Fujita, N. and Ishihama, A.: Mode of DNA-protein interaction between the C-terminal domain of Escherichia coli RNA polymerase a subunit and T7D promoter UP element. Nucleic Acids Res., 29, 4909-4919, (2001).
11. 0303271859
Colland, F., Fujita, N., Ishihama, A. and Kolb, A.: The interaction between σS, the stationary phase σ factor, and the core enzyme of Escherichia coli RNA polymerase. Genes Cells, 7, 233-247, (2002).
12. 0303271959
Yamamoto, K., Ogasawara, H., Fujita, N., Utsumi R., and Ishihama, A.: Novel mode of transcription regulation of divergently overlapping promoters by PhoP, the regulator of two-component system sensing external magnesium availability. Mol. Microbiol., 45, 423-438, (2002).
13.
Shimada, T., Fujita, N., Maeda, M. and Ishihama, A.: Systematic search for the Cra-binding promoters using genomic SELEX system. Genes Cells, 10, 907-918, (2005).

２）データベース／ソフトウェア

1. 0303272017
Microbial Genome Workbench (2003).
http://www.bioscinet.org/mgw/
(旧URL http://comtraf.lab.nig.ac.jp/mgw/)