3D-1D法を用いた全遺伝子産物同定システムの研究開発

国立遺伝学研究所　○西川 建

Constructing GTOP: Structure-Annotation Database for Genome ORFs

Ken Nishikawa;　National Institute of Genetics

Abstract:

Having a huge amount of genome sequence data, we have aimed to analyze the data and provide reliable information extracted to biological scientists. Focusing on the protein structure prediction, we have so far analyzed 40 organisms whose complete genomes were sequenced. The tools employed were PSI-BLAST, a sequence homology search method recently developed, as well as the conventional methods, FASTA/BLAST. The homology search was conducted for every ORF of a genome as a query against the sequence database, SwissProt, as well as against the structural database, PDB. A high percentage of all ORFs (40-50% for eubacteria) was hit to PDB, i.e., predictable of protein structure by homology. These results were mainly attributed to PSI-BLAST, much more powerful than FASTA/BLAST. All the data analyzed were compiled in a database called GTOP (Genome TO Protein structure and function), and available on the web: http://spock.genes.nig.ac.jp/~genome/gtop.html.

1．はじめに

大量に産出されるゲノム配列データをいち早く情報解析し、その結果をデータベースとしてまとめて公開することが、本研究課題の目的である。すでに昨年の中間報告で述べたように、ゲノム情報解析はタンパク質レベルの解析に限り、とりわけアミノ酸配列からの立体構造予測を中心とする。ただし、解析ツールとしての性能から判断して、当初計画した3D-1D法に換えて、より優れたPSI-BLAST 法を採用することにした。立体構造予測のほかに、配列ホモロジー検索、モチーフ検索、膜貫通ヘリックス　シグナル配列予測、繰り返し配列の同定などのコンピュータ解析も併せて行ない、すべての解析結果は「ＧＴＯＰデータベース」（URL は上記）としてまとめて公開している。自動解析が完了し、ＧＴＯＰに収録された生物種の数は、中間報告の時点では22種であったが、現在（2001 年9 月）では41種とほぼ倍増した。

このようにして構築されたデータベースは、個別の遺伝子　タンパク質に関する「百科辞典」として意味をもつが、さらに具体的な研究とむすびついてこそ本当の価値が生まれる。我々は、ＧＴＯＰの内容を詳細に吟味することにより、ORF の全長が既知の構造ドメインの一部分にしかヒットしない場合があること、あるいはゲノム上で隣接するORF を2つ合せるとはじめて１つの完全な構造ドメインにヒットする事例が少なからずあることを見出した。大腸菌を対象としてこれらの事例を慎重に解析した結果、これらは配列決定上の単純な実験エラーではなく、本来の遺伝子が壊れた状態、すなわち「偽遺伝子」に違いないという結論に達した。大腸菌ゲノム中にこのような偽遺伝子（ORF）が100個近く存在することが同定できたので、併せて報告したい。

2．ＧＴＯＰデータベース

ＧＴＯＰの内容に関しては、オリジナル論文にまとめ学術雑誌に投稿した（文献3)。図1は、これまでに自動解析が完了しＧＴＯＰに収録されている41種の生物種について、ゲノム中の全ORF のうち立体構造が予測されたORF の割合（%）を示したものである。種によってばらつきがあるが、全体としての平均値は44.5 % である。また、大腸菌など個別の種について、中間報告時の結果と比較してみると、約3%ほど値が上昇している。これは、構造予測のさいにサーチ対象として用いた立体構造ベータベース（PDB）の増大を反映した結果である。PDB ばかりでなく、配列データベース（SwissProt)、ファミリー分類（Pfam）などの基本データベースも随時更新されるので、それに伴いＧＴＯＰの自動解析も定期的にやり直す必要がある。ＧＴＯＰでは年間3、4回の更新を行っている。

データベースとしての管理・運用上のもう1つの大きな課題は、新規生物種のゲノム情報を取り込み、ＧＴＯＰ

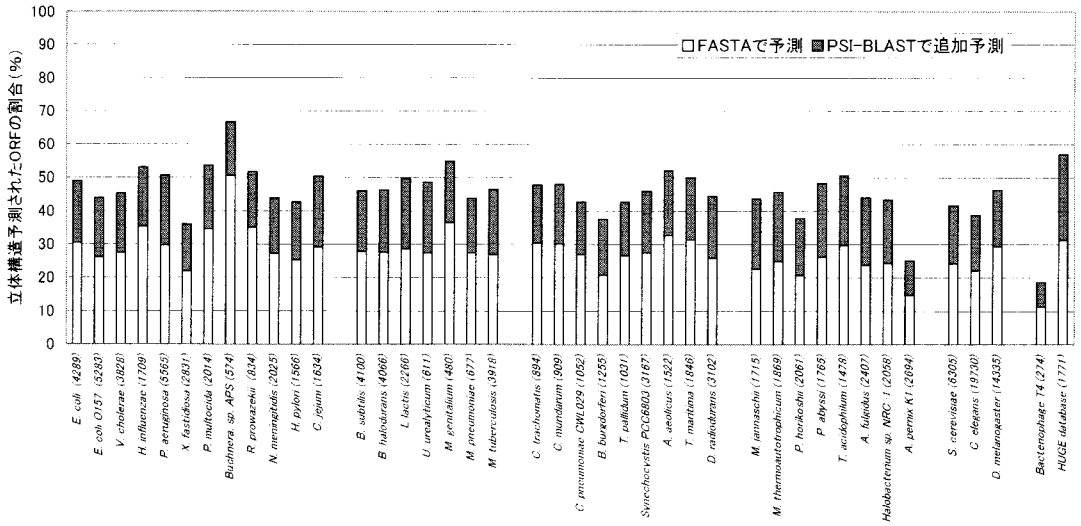

図1　各種の生物ゲノムにおいて立体構造が予測されたORFの割合

を拡大していくことである。ＧＴＯＰでは完全長ゲノムの解明されたすべての生物種の収録を目標としているが、新しくゲノムが決定され発表される生物種がますます増加している（とくに細菌は月当りの増加数が数種にのぼる）ため、その勢いに追随するのは容易ではない。完全収録の困難さは、単に自動解析に計算時間がかかるためばかりではない。ＧＴＯＰではORF ごとの個別情報だけでなく、生物種間の比較も含めたORFどうしのホモログ解析を行い、その結果はOrgPattern として提供している。ある特定のORF のパラログやホモログが他のどのような生物種に存在するか、あるいは存在しないか、という情報は利用者にとって重要である。ところが、この比較ゲノム情報のために、新規生物種の追加が難しくなるという事情が発生する。つまり、1個でも新たに生物種が加わるとすべての生物種間の比較をやり直す必要があるため、生物種を拡大するなら一度にまとめて行う方が手間がはぶける。したがって、生物種の拡大も年数回の更新時にのみ合せて行なっている。

ＧＴＯＰを利用するにあたって注意すべきは、バクテリア（真正細菌と古細菌）と酵母などを含む微生物ゲノムの情報は信頼できるが、高等生物（多細胞真核生物）のゲノム情報は一般的に信頼性に欠ける点である。その最大の原因は、塩基配列からコーディング領域を同定する段階が難航し、確定したアミノ酸配列がデータセットとして提供されないためである。ＧＴＯＰでは遺伝子として提供されたアミノ酸配列から出発して自動解析を行なっているが、一通り解析が終了しＧＴＯＰに収録された線虫やハエに対しても、元のデータセットにたびたび修正が入るので、ＧＴＯＰの解析結果も暫定的情報だと言わざるをえない。動植物ゲノムに対しては、バクテリアとは「質」の違うデータとして対応する必要があると考えている。

3．偽遺伝子の探索

ＧＴＯＰを活用して以下の研究を行った。今年初めに病原性大腸菌O157 の完全ゲノム配列が日米2チームから相次いで発表され、ＧＴＯＰには今年4月に収録した。これにより、すでにゲノムが解明されている大腸菌の標準的なK-12株と比較してみることが可能になった。K-12 株も日米の2チームが独立に配列決定しているので塩基配列データの信頼度は高い。K-12 とO157 は種は同じで菌株だけの違いなので、対応するORFを比較するとアミノ酸配列の相同性は98％以上に達する。ところが、それにもかかわらず配列の全長が大きく異なるようなペアが存在する。図2にその典型例を示す。この例では、O157 の1つのORF（Z4353）に対して、K-12 株の2つのORF（b2999, b3000）がオルソログの関係にある（前後のORF もすべて相同の関係）。したがって、K-12 側のORF は短く、予測された構造でみると、それぞれは構造ドメインの1部分にしか相当しない。このような部分的な構造にしか対応しないORF は発現したとしても正しくフォールドする（一定の形をとる）とは考えられず、むしろ本来1つの遺伝子であったものが破壊されて偽遺伝子となった姿を示していると考えるべきであろう。事実、完全な遺伝子と思われるZ4353 と同程度の長さのホモログは種内、種間の比較で見出されるが、遺伝子の断片と見なせる

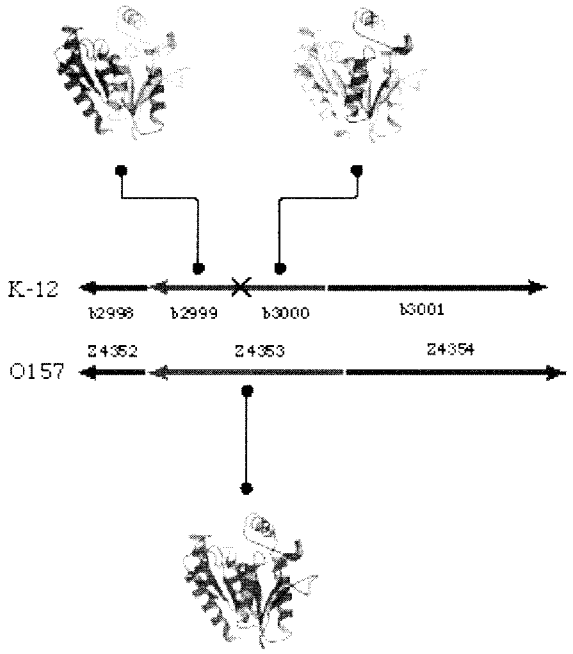

図2　大腸菌ゲノム中の2つに分断された偽遺伝子の例

b2999およびb3000と同程度の長さのホモログは皆無である。塩基配列をZ4353 と比べてみると、b3000とb2999の間には2カ所に2塩基、1塩基の欠失があり、そのために2つのORF に分断されたことが判る。遺伝子を壊すようなダメージが複数カ所あることからも単純な配列決定のエラーとは考えられず、本当に遺伝子が壊れつつある姿を示していると思われる。組織的な探索の結果、この例のような遺伝子の断片と思われるORF は大腸菌K-12株で97個、O157 株では116個が同定された（文献4）。ちなみに、これまでに大腸菌を含めて独立栄養型の細菌では偽遺伝子の存在が報告された例はない。

4．ネットワークの活用について

本研究開発では、外部のデータベースの取得・更新、研究参加機関間の計算機の相互利用・データベースの外部提供にネットワーク（SINET およびIMnet）を用いている。とくに、外部のデータベースの取得・更新は不可欠であり、できるだけ頻繁にデータベースを更新することが望ましい。本プロジェクトでは、解析対象とした生物種の全ゲノムのORF の入手やタンパク質　塩基配列データベース（約40GB に達する）等の新リリースの入手をし、日毎の更新を行っている。GTOP の公開はアクセス数が順調に伸びている。生物学研究者が自由にブラウズし、図や画像データを眺め、いろいろな角度から検索・分析できるためには、遅延が少なく一定量のデータの輸送がストレスなく行えるような、ネットワークインフラがますます重要になっている。

5．まとめ

ゲノム情報からタンパク質のアミノ酸配列を読みとり、配列ホモロジー検索などを行う解析はすでにルーチン化しているが、さらに立体構造予測にまで進めた解析は国際的にも少なくその点でＧＴＯＰは特徴あるデータベースであるといえる。とくに、国内にこのような公開データベースをもち、維持・発展させていくことは、わが国のゲノム科学の進展のために欠かすことのできない価値をもつと信じている。タンパク質の立体構造が予測できると、それを手がかりにして機能予測が可能になる場合もある。事実、そのような研究が可能であることは、我々はＴ4ファージ・ゲノムの系統的解析により示すことができた（文献1）。また、上述の偽遺伝子の探索研究のように、ＧＴＯＰを用いた比較ゲノム学的研究も大いに進めていきたいと考えている。

ＪＳＴの支援により、本プロジェクトを立ち上げることができ、データベースとして結実できたことに感謝します。データベースは継続的に更新していくことが肝要なので、ＧＴＯＰは今後とも維持・管理していく所存です。

研究実施体制

氏名	所属	役職
西川　建	国立遺伝学研究所 生命情報・DDBJ研究センター	教授
太田　元規	国立遺伝学研究所 生命情報・DDBJ研究センター	助手
川端　猛 (H11.4~H12.6)	国立遺伝学研究所 生命情報・DDBJ研究センター	JST研究員
福地　佐斗志 (H11.8~H13.9)	国立遺伝学研究所 生命情報・DDBJ研究センター	JST研究員
本間　桂一 (H12.10~H13.9)	国立遺伝学研究所 生命情報・DDBJ研究センター	JST研究員
西村　昭子	国立遺伝学研究所 系統生物研究センター	助教授
鈴木　小夜子 (H11.4~H12.3)	国立遺伝学研究所 系統生物研究センター	JST技術員
中出　晋介 (H12.4~H13.5)	国立遺伝学研究所 系統生物研究センター	JST技術員
松本　典子 (H13.6~H13.9)	国立遺伝学研究所 系統生物研究センター	JST技術員
市吉　伸行	三菱総合研究所 フロンティア科学研究部	主任研究員
伊藤　武彦	三菱総合研究所 フロンティア科学研究部	研究員
落合　孝正	三菱総合研究所 フロンティア科学研究部	主任研究員
荒木　次郎 (H12.7~H13.9)	三菱総合研究所 フロンティア科学研究部	研究員
中島　広志	金沢大学医学部	教授
山下　紗代 (H11.4~H13.3)	金沢大学医学部	JST技術員
小原　収	かずさDNA研究所 ヒト遺伝子研究部	部長

参考文献

1. Kawabata, T., Arisaka, F. and Nishikawa, K. "Structural/functional assignment of bacteriophage T4 unknown proteins by iterative database searches" ,GENE, 259, 223-233 (2000) .
2. Fukuchi, S. and Nishikawa, K., "Protein surface amino-acid composition distinctively differ between thermophilic and mesophilic bacteria" . J. Mol. Biol., 309, 835-843 (2001) .
3. Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N. and Nishikawa, K., "GTOP: A database of protein structures predicted from genome sequences" . Nuc. Acids Res., in press.
4. Homma, K., Fukuchi, S., Kawabata, T., Ota, M. and Nishikawa, K., "The Escherichia coli genome contains a significant number of pseudogenes" . (投稿中)