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

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 20 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

本研究課題を申請した２年前にくらべ、実験的にゲノム塩基配列データが解明されるテンポは一層加速され、その間にそれまで微生物に限られていた完全長ゲノムの決定が線虫をはじめとする多細胞真核生物にまで及び、ついにはヒトゲノム全体のドラフト配列が発表されるに至った。我々が目標としたのは、ゲノム上の全遺伝子（ORF）または遺伝子産物としてのタンパク質に対して立体構造予測法を適用することによって「構造アノテーション」（付加情報付け）を逸早く行い、立体構造とさらにそれから推定されるタンパク分子機能等の新規情報を広く生物学者（それぞれの生物種の専門研究者）に提供したいということであった。

本プロジェクト研究を実際に始めてみて判明したことがいくつかある。１つは、立体構造予測の解析ツールとして当初予定していた３D-１D法よりも優れた方法（PSI-BLAST 法）の存在を知り、PSI-BLAST を全面的に使うことにしたこと。その結果、立体構造が予測できるORFの割合が予想以上に多い（バクテリアでは全ORFの４～５割にのぼる）という結果が得られたこと。また、全体的なコンピュータ解析の工程はほぼ自動化できたが、予想以上に計算時間が掛ること（線虫の解析だけで約２カ月の計算時間を要した）、などである。

これまでに、２０種以上の生物種について自動解析を終了した。個別の解析結果について生物科学的な検討・吟味を行ったところ、今までまったく機能未知であったいくつかのORF について構造・機能に関する新規の知見が得られた。また、解析データを格納し検索・表示する機能をもつWebサーバ（GTOPデータベース）を立ち上げ、得られたすべての結果を公開している。

２ 解析の方法

２.1 対象とした生物種と解析の概要

現在までに真正細菌１２種、古細菌６種、真核生物２種、ウイルス１種について全ゲノムのORF の解析を行った。さらに全ゲノムではないが、ヒトの cDNA 由来のタンパク質（HUGEデータベース）に対しても同様の解析を行った。生物種名とそれぞれのORFの総数を表１に示す。

これらの生物種について配列相同性解析を主な手段として、以下のような解析を行った。

◆ 立体構造予測

立体構造既知の配列データベースに対して配列相同性解析（FASTA/PSI-BLAST）を行う。立体構造の保存性は配列に比べて極めて高いため、既知構造のタンパク質と有意な相同性があればその立体構造よく似ていると考えられる。立体構造既知の配列データベースは、Protein Data Bank (PDB)の配列相同性95% の代表リ

スト4,047配列を用いた。

表１．ゲノムサーチにより立体構造予測／ファミリー分類されたORFの割合

生物種	ＯＲＦ数	立体構造予測		ファミリー分類	
		FASTA	PSI-BLAST	BLAST	PSI-BLAST
真正細菌					
Escherichia coli（大腸菌）	4289	27.9%	46.2%	71.4%	75.5%
Vibrio cholerae（コレラ菌）	3828	24.4%	42.8%	58.2%	63.5%
Haemophilus influenzae	1709	30.8%	49.5%	72.8%	76.9%
Helicobacter pylori（ピロリ菌）	1566	22.7%	40.2%	56.8%	63.3%
Bacillus subtilis（枯草菌）	4100	25.6%	44.0%	61.7%	67.3%
Bacillus halodurans	4076	25.1%	43.7%	58.6%	65.0%
Mycoplasma genitalium	480	32.3%	49.8%	68.5%	75.2%
Mycoplasma pneumoniae	677	24.1%	39.7%	56.6%	64.1%
Chlamydia trachomatis	894	28.4%	45.1%	59.5%	66.9%
Synechocystis sp.（ラン藻）	3168	24.8%	43.2%	53.9%	62.2%
Aquifex aeolicus	1522	29.5%	49.5%	65.9%	72.7%
Thermotoga maritima	1846	28.2%	46.9%	62.6%	69.7%
古細菌					
Methanococcus jannaschii（メタン菌）	1715	19.8%	41.3%	48.5%	58.8%
Methanobacterium thermoautotrophicum	1869	21.5%	36.7%	50.5%	50.6%
Pyrococcus horikoshii	2061	17.8%	35.8%	40.7%	50.9%
Pyrococcus abyssi	1765	22.8%	45.6%	51.8%	64.2%
Archaeoglobus fulgidus	2407	21.4%	42.2%	47.9%	58.1%
Aeropyrum pernix	2694	13.2%	23.4%	26.7%	32.5%
真核生物					
Saccharomyces cerevisiae（酵母）	6307	22.1%	39.5%	61.0%	65.6%
Caenorhabditis elegans（線虫）	18576	20.6%	36.9%	42.7%	52.3%
その他					
Bacteriophage T4（T4ファージ）	275	11.3%	17.1%	48.0%	49.5%
HUGE database（ヒトcDNA）	1542	28.1%	55.3%	65.6%	72.8%

◆ ファミリー分類

機能既知の配列データベースに対して配列相同性解析（BLAST/PSI-BLAST）を行うことにより問合せタンパク質の大まかな「種類」（ファミリー）を知ることができる。タンパク質ファミリーが判れば機能が推定できる場合もあるが、立体構造と異なり進化的に相同であっても必ずしも同じ機能を持つとは限らないので、ここでは「機能予測」と区別して考える。機能既知の配列データベースはSwissProt Rel. 38の中の、最小限の機能情報が記述されている63,189配列を用いた。

◆ その他

それ以外の補助的な解析として、膜貫通ヘリックスの予測（SOSUI）、コイルドコイル領域の予測（MultiCoil）、機能モチーフの同定（ProSite）、HMM によるドメイン同定（Pfam/HMMER）や、シグナル配列予測（SignalP）などを行った。また、長い繰返し配列を検出できるプログラムを開発し解析に用いた。

2.2 PSI-BLAST について

配列相同性検索のプログラムPSI-BLASTは、近年開発されたプロフィール型配列検索プログラムで、その有効性が高く評価されるようになった（Altschul et al., 1997）。FASTA/BLASTなど従来のペアワイズ型の配列検索法に比べ、２～３倍の遠縁の相同配列まで認識できるといわれる。PSI-BLASTは、反復的にデータベースを検索してマルチプルアラインメントを作成する方法を採用しているため、自動的な検索が容易であるという長所がある。一方、計算コストはペアワイズ検索法に比べ5～10倍かかる。有意な相同性の判定条件は、E-valueが0.001以下とした（FASTA/BLASTについても同じ）。この値を用いたときの誤り率は１～５％と推定された。

3　解析結果

3.1 立体構造予測が可能なORFの割合

表１の左側に立体構造予測されたORFの比率を示した。種によって多少のばらつきがあるが、大まかな傾向として立体構造が予測されるORFの割合は、FASTAで 25％程度、PSI-BLASTを併用することで40％前後、あるいは中には50％近くにまで達する種もある。真正細菌ではすべての種で40％を超えているが、これは当初予想したよりも非常に高い値である。一方、T4ファージは他の細胞性生物とは異質であるため例外的に値が低い。また、ここでは１つのタンパク質の１ドメインでも予測されていれば１ヒットとカウントした。HUGEの値が55％と高いのはこのデータベースが500残基以上の大きなタンパク質で構成されており、部分だけの予測が多いためであろう。

表１の右側の欄に、ファミリー分類されたORFの割合を示した。ファミリー分類の割合は種によってかなり差が大きい、モデル生物としてよく研究されている大腸菌と近縁のインフルエンザ菌では70％以上（BLASTによる検索）と極めて高いが、研究が遅れている古細菌では40～50％と低い。ファミリー分類の場合、BLASTサーチにくらべPSI-BLASTサーチの適用による増分は5～10％であり、立体構造予測の場合の増分（15～20％）と比べて小さいことが注目される。このことから、PSI-BLAST法は通常のホモロジー検索（ファミリー分類）よりも立体構造予測に適した方法であり、後者においてこそ有効性が発揮されることがわかる。

3.2 GTOPデータベース

以上の結果は、ORFの総数が10万以上におよび膨大なデータ量になる。このような大量データから有効な情報を引き出すための効率的な検索エンジンの開発、解析結果のわかりやすいグラフィック表示は必須である。そのためのプログラムを開発し、すべての解析データをGTOPデータベース（Genome TO Protein structure and function）としてまとめ、次のWebサイトより公開している：http://spock.genes.nig.ac.jp/~genome/gtop.html

個別ORFに関する解析結果を示したトップページの例と、さらに予測された立体構造に関する詳細情報を示した表示画面の例を図１と図２に示す。

3.3 立体構造予測から機能予測へ

我々の解析で新規の立体構造予測といえるのは、従来法のFASTA/BLASTおよびPfam等ではヒットせず、PSI-BLASTのみでPDB配列にヒットし、かつ相同性が検出される領域が１つのドメイン以上におよぶような十分長い場合だといってよい。大腸菌を例にとれば、このような条件を満たすORFは、57個ある。さらに、タンパク質の分子機能が予測できるのは、ヒットした相手の既知タンパク質が酵素の場合か、もう少し広くリガンド結合性タンパク質の場合である。GTOPデータベースでは、X線結晶解析データからリガンド結合の情報を自動的に抽出し、その情報を利用できる形にしてある。未知タンパク質と構造既知のタンパク質の配列アラインメントを見て、リガンド結合部位のアミノ酸が保存していれば両者は同じ分子機能をもつだろうと予測できる。大腸菌では、このような機能予測が可能なORFは41個見出された。こうして機能予測されたタンパク質は、最終的には実験家の手で遺伝子を発現させて実際に機能をチェックしてもらう必要がある。

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(IMnetおよびSINET)を用いている。とくに、外部のデータベースの取得･更新は重要な部分であり、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／ゲノム配列データベース（約30GB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データベースが公開され、アクセス数が徐々に増えている。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、色々な角度から検索･分析できるためには、遅延が少なく一定量のデータの転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今後ますます求められる。

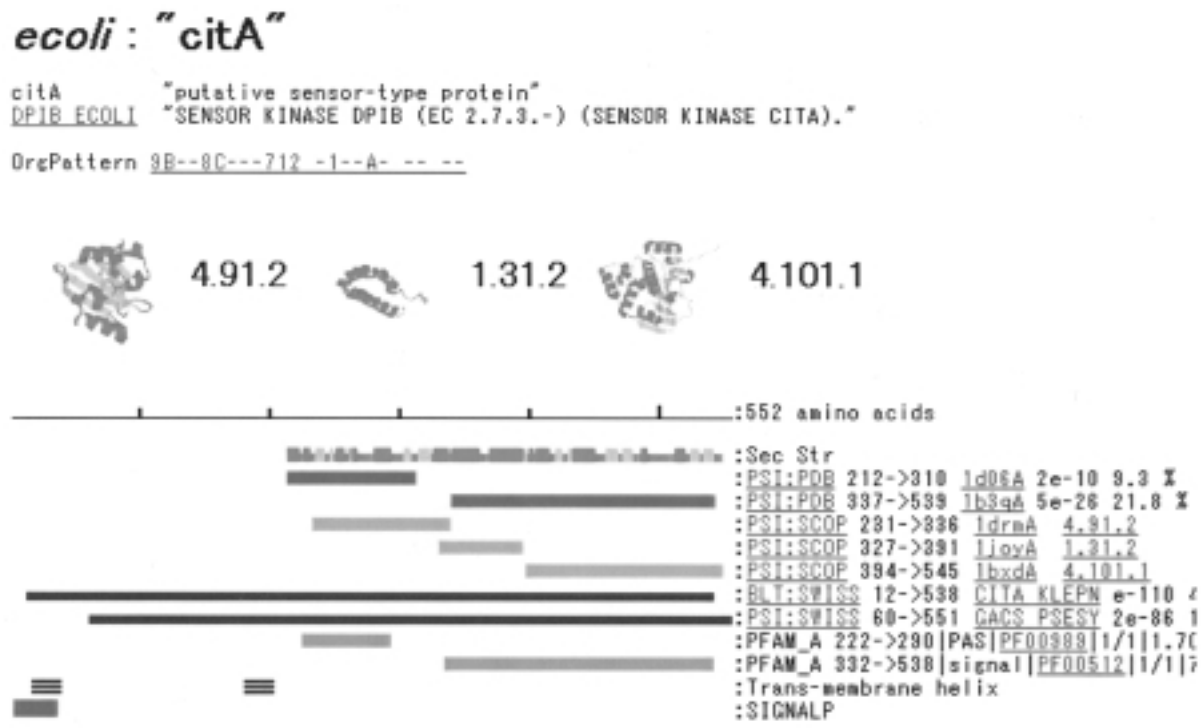

図1. GTOPの表示画面。大腸菌の個別遺伝子ファイルのトップページの例

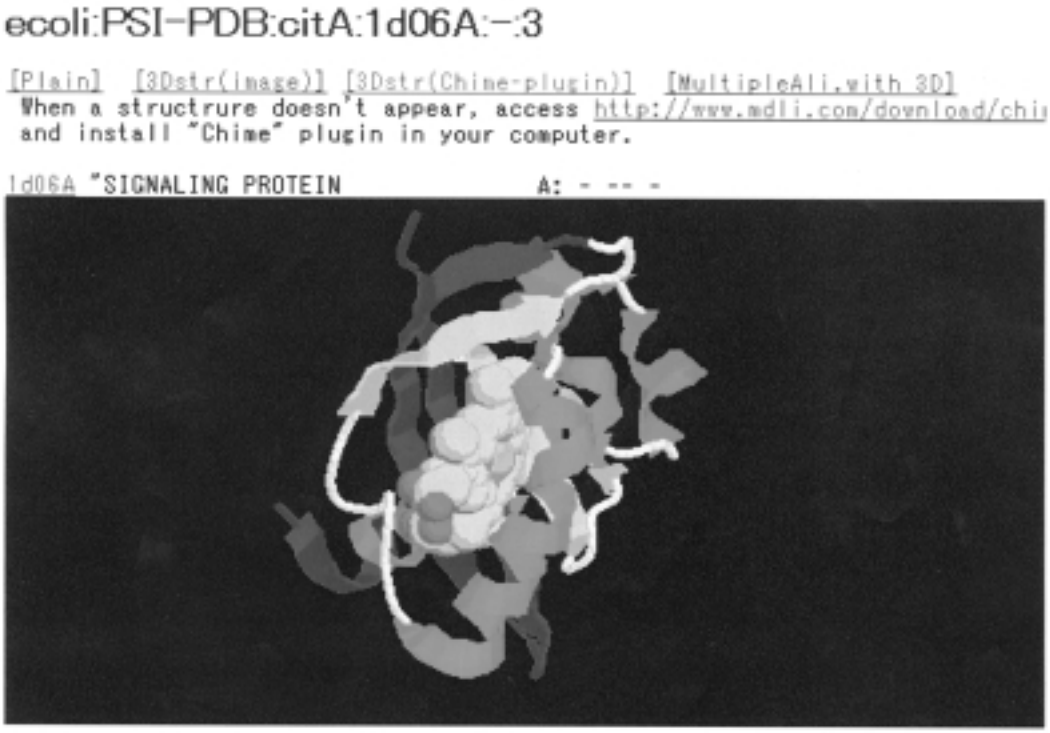

図2. 予測された立体構造。図1の画面上でクリックすると表示される。

5　まとめ及び今後の予定

ゲノムにコードされた全ORFを対象として立体構造予測を中心とする自動解析を行ない、その解析結果を公開用データベース（GTOP）に自動的にまとめ上げ、Web上に提示するというシステムを構築することができた。これにより本プロジェクトの所期の目的は、大筋としてほぼ順調に遂行されたと考える。一方、完全長ゲノムが解明されているにもかかわらず、まだ我々の解析が追いつかず処理できてない生物種もかなりある。すなわち、自動解析の「しくみ」はでき上ったが、それを完全に実行させるだけの設備（計算機パワー）が伴なってないという状態である。その点を早急に克服し、今後は解析の対象をショウジョウバエ、植物（Arabidopsis)、さらにはヒトゲノムにまで拡張する計画である。また、自動解析の結果、立体構造／機能が新規に予測されるORFが一定の割合で得られる。それらの予測可能な個別ORFについては実験データと照合して確認したり、生物学的意味・重要性について専門家の判断を仰がなければならないことが多く、生物学者との協力体制をつくる必要がある。現在はまだ不十分な状態であるが、今後はさらに広い範囲の専門家との協力関係を築くように努力したい。

6　研究実施体制

氏名	所属	役職	研究開発項目
西川　建	国立遺伝学研究所 生命情報研究センター	教授	・本研究開発の統括 ・全体システム設計･統合
太田　元規 (H10.10～H12.3)	国立遺伝学研究所 生命情報研究センター	助手	・遺伝子産物同定システム開発研究 ・全体システム設計･統合
川端　猛 (H11.4～H12.6)	国立遺伝学研究所 生命情報研究センター	JST 研究員	・遺伝子産物同定システム研究開発 ・GTOP データベースの構築
福地　佐斗志 (H11.8～)	国立遺伝学研究所 生命情報研究センター	JST 研究員	・遺伝子産物同定システム研究開発 ・GTOP データベースの構築
本間　桂一 (H12.10～)	国立遺伝学研究所 生命情報研究センター	JST 研究員	・遺伝子産物同定システム研究開発 ・GTOP データベースの構築
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	・大腸菌遺伝子のデータ解析
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	・大腸菌遺伝子のデータ解析
中出　晋介 (H12.4～)	国立遺伝学研究所 系統生物研究センター	JST 技術員	・大腸菌遺伝子のデータ解析
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	・遺伝子産物同定システム研究開発 ・自動収集システム開発
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	・遺伝子産物同定システム研究開発 ・Gene Catalog システム構築
落合　孝正 (H11.4～)	三菱総合研究所 フロンティア科学研究部	主任研究員	・遺伝子産物同定システム研究開発
荒木　次郎 (H12.7～)	三菱総合研究所 フロンティア科学研究部	研究員	・遺伝子産物同定システム研究開発 ・Web ユーザインタフェースの開発
中島　広志	金沢大学医学部	教授	・遺伝子産物同定システム研究開発 ・塩基組成によるゲノム配列解析
山下　紗代 (H11.4～)	金沢大学医学部	JST 技術員	・塩基組成によるゲノム配列解析
小原　収	かずさＤＮＡ研究所 ヒト遺伝子研究部	部長	・ヒト cDNA 配列データの解析