微生物ゲノムの共通プロトコルによる遺伝子配列情報の提供

阿部貴志
(共同研究者)
小菅武英，大城戸利久，田中直人，平畠壮規，宮崎智，菅原秀明
国立遺伝学研究所　生命情報・DDBJ センター，JST-BIRD

現在までに原核生物からヒトに至る広範な生物種のゲノム配列が決定され、ゲノム上のタンパク質コード領域 (CDS) をはじめとする遺伝子情報が国際塩基配列データベースから公開されている。公開されている微生物ゲノムの ORF の予測方法を調査したところ、各プロジェクトの間で遺伝子領域予測プログラムや ORF 判定のパラメーター設定に大きな差異が存在することが分かった。また、公開されているアノテーションの判断基準が透明でない場合や曖昧な例も多々存在した。さらに、アノテーションに使用された参照データベースのバージョンが共通でないという問題があった。したがって、公開されている遺伝子情報の内容をそのまま比較ゲノム解析に利用するのは危険であり、アノテーション情報の確認・再アノテーションの必要性が明らかになった。

ここで、GTOP[1]によるアミノ酸配列からタンパク質産物の２次構造と３次構造の予測実績を図１に示す。図１は、blast による予測に対して psi-blast による予測が勝っていることとともに、2001 年以来予測率があがっていることを示している。この予測率の上昇は実は参照データベースの一つである PDB から公開されるタンパク質の立体構造件数が増加したことによる。すなわち、図１は継続的な再解析が必要なことを明示している。そこで、我々は DDBJ から公表された微生物ゲノムデータのアノテーションの再評価を行った。以下では、本解析を Gene Trek in Prokaryote Space （GTPS）[2]と呼ぶ。

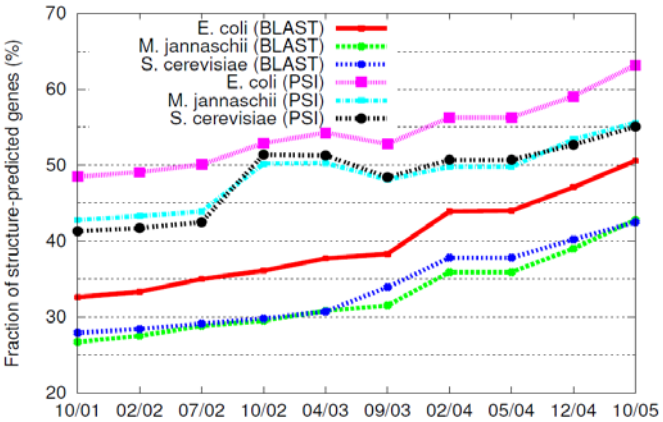

図１．GTOP によるタンパク質産物の構造予測率の変化

GTPS は平成 15 年度に、平成 15 年 7 月までに DDBJ で公開された 123 株の微生物完全長ゲノムデータを対象に実施し（2003 年版）、平成 16 年度 183 株（2004 年版）、平成 17 年度 303 株（2005 年版）と追加・更新を行っている。そのワークフローの概要を図２に示す。

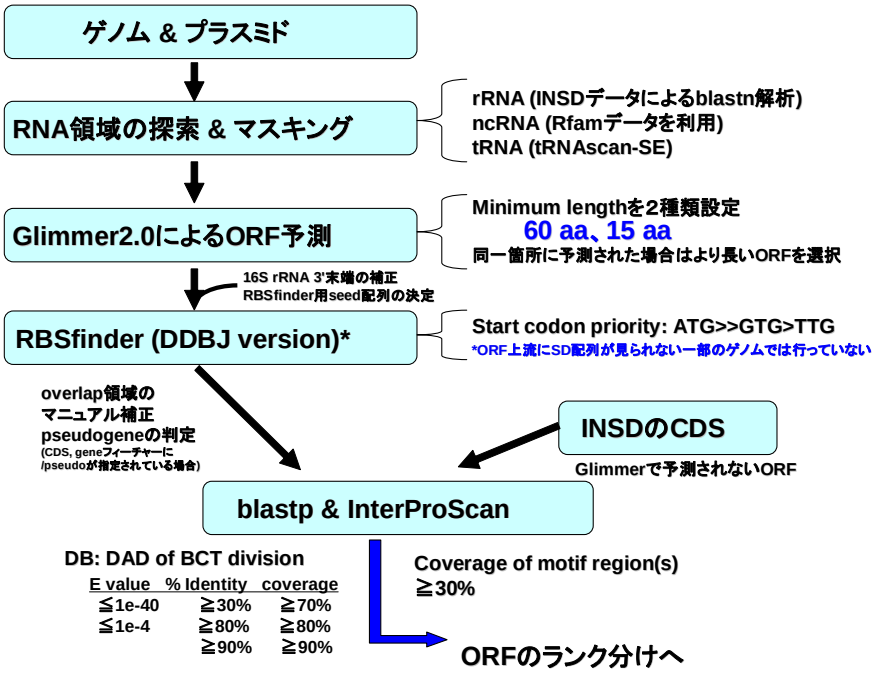

図 2. GTPS ワークフロー

GTPS のプロジェクトと平行して、DDBJ のエントリー単位ではなくゲノム単位で微生物ゲノムデータを網羅した Genome Information Broker （GIB）を維持してきたが[3]、この GIB からゲノム配列データを取得し、共通のプロトコルで網羅的に解析し、ORF 予測を行う。次に、候補 ORF をアミノ酸に翻訳後、相同性検索とモチーフ検索を行い、その結果を総合的に判断して絞り込んだORF を対象にして、予め設定したルールに従って、確実性の観点から候補 ORF を A～X のランクに分類した（表 1）。2003 年版 GTPS（表の ver.2003）では、機能が明確なランク A から機能について何らかの手がかりがあるランク D までの CDS が 374,914 件となった。これに対して、DDBJ に登録されていた ORF は 362,543 件であった。この差は GTPS によって新たに発見された ORF である可能性がある。事実、GTPS 解析で予測された CDS が実験由来のデータと一致する事例が出てきた。図 3 にその一例を示すが、図の中の赤い矢印の領域は、DDBJ から公開されているゲノム配列データには ORF としてアノテーションされていなかったが、GTPS の結果、ORF の可能性が高い領域であった。その後、Swiss-Prot でも GTPS と同じアノテーションがつけられていた。また、GTPS の成果は、大腸菌ゲノムの国際協力による再アノテーションにも採用された[4]。

表 1. ランキング結果

	ver. 2003	ver. 2004
AAAA-A	283,247	431,672
BBBB-B	7,208	10,250
C	4,680	7,511
D	79,779	107,382
E	6,788	10,225
X	466,681	687,110
	848,383	1,254,150

GTPS で対象とした延べゲノムサイズはおよそ、2003 年版 0.4 ギガ bp、2004 年版 0.6 ギガ bp そして 2005 年版で 1 ギガ bp に達し、評価対象の ORF 数も 120 万件を超えている。我々のデータ増加率の想定を越え、データが増加しており、我々の計算機資源を用いて GTPS ワークフローを実行すると 4 ヶ月を超えるようになった。特に計算時間がかかる処理は、モチーフ検索ツールである InterProScan の実行であり、他の部分と比べ、約 30 倍以上の計算時間が必要となっていた。本ワークフローでは、InterProScan を PC クラスタ上で効率よく実行するにあたり、ジョブ制御は独自に作成したスクリプトを使用していたが、現在は、理化学研究所情報基盤センターのご協力により、理化学研究所の大規模 PC クラスタ上での InterProScan の実行環境、ならびに、ジョブ制御機能を構築して頂き、大規模計算を実施することが可能となった。理化学研究所での大規模計算により、本ワークフローの迅速な更新作業を行うことが可能となった(図 4)。

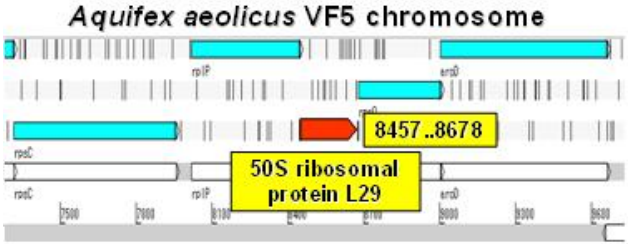

図 3. GTPS による新規 CDS（赤い矢印）がその後、Swiss-Prot により裏づけが得られた例

さて、INSD は、遺伝子配列データからゲノム配列データや cDNA 配列データへと展開してきたが、INSD の Whole Genome Shotgun（WGS）の区分に、断片配列から一定の長さまで結合した contig 配列までのデータがプロジェクト単位で大量に登録されるようになり、登録件数は微生物ゲノムに限れば、完全長ゲノム配列と逆転している（図 5）。WGS は GTPS の絶好の対象であり、今後、WGS データからの情報・知識の抽出に取り組んでいきたい。

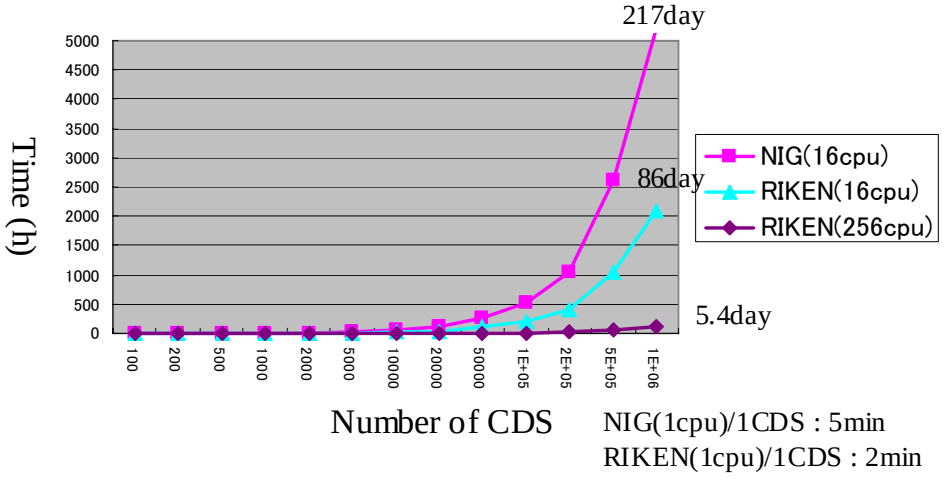

図 4. InterProScan の解析対象 CDS 数と計算時間の関係

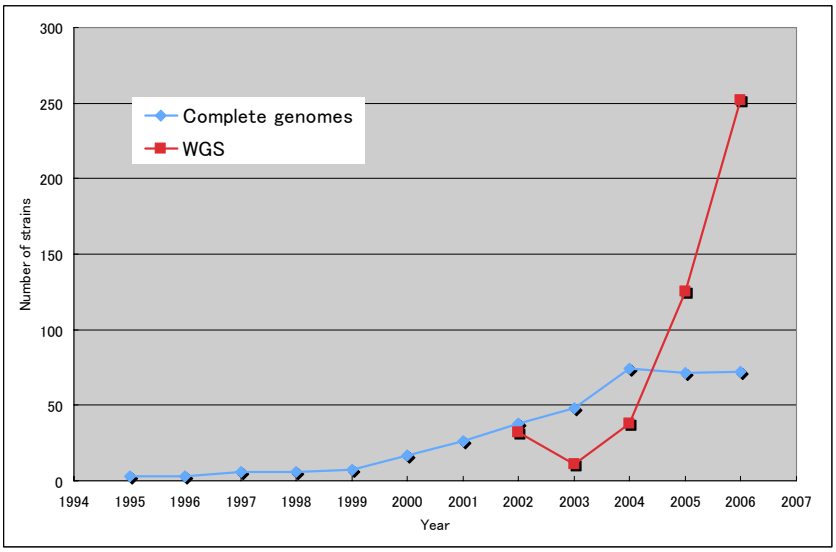

図 5. 毎年 INSD から公開される完全ゲノムの件数と WGS プロジェクト数の比較

参考文献

1. Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N., and Nishikawa, K. (2002). GTOP: a database of protein structures predicted from genome sequence. *Nucleic Acids Res.*, 30, 294-8.
2. Sugawara, H. (2004). Gene Trek in Prokaryote Space powered by a GRID environment, *Proceedings of the First International Workshop on Life Science Grid （LSGRID2004）*, May3.
3. Fumoto, M., Miyazaki, S. and Sugawara, H. Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucl. Acids Res.*, 30(1), 66-68, 2002.
4. Riley, M., Abe, T., Arnaud, B.M., Berlyn, M., Blattner, R.F., Chaudhuri, R.R., Glasner, D.J., Horiuchi, T., Keseler, M.I., Kosuge, T., Mori, H., Perna, T.N., Plunkett, G., Rudd, E.K., Serres, H.M., Thomas, H.G., Thomson, R.H., Wishart, D., and Wanner, L.B. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot—2005, *Nucleic Acids Res.*, 34, 1-9.