全ゲノムを対象とした蛋白質立体構造データベースの公開

由良 敬・山口晶大・郷 通子

ゲノムから推定される蛋白質のモデル立体構造を，筆者らはデータベースFAMSBASEとして公開してきた[1]（http://daisy.nagahama-i-bio.ac.jp/Famsbase/）．FAMSBASEは，梅山秀明らが全ゲノムを対象として全自動蛋白質立体構造モデリングソフトウェアFAMS[2]を用いて行なった，モデリング結果を集積したデータベースである．ゲノムにコードされている蛋白質は，アミノ酸配列の類似性からグループ分けすることができる．各グループの少なくとも1つの立体構造が判明すれば，グループに含まれる全蛋白質の立体構造は，ホモロジーモデリングによって推定することができる[3,4]．2005年5月末に更新した最新のFAMSBASEには，ゲノムが判明している古細菌17種，真正細菌130種，真核生物19種，ファージ111種，計277生物種由来，369,964個の蛋白質モデル立体構造が格納されている．この規模のホモロジーモデリング結果を公開しているのは，FAMSBASEだけである．277生物種には総計736,230個のORFが推定されている．よって，最新のFAMSBASEでは，約50%の立体構造が推定されている状況にある．

しかし，1つのORFの全体がホモロジーモデリングできているわけではない．モデル構造ができている部分に含まれるアミノ酸残基の割合を生物界別に調べると，古細菌と真正細菌では6割強のORFにおいてほとんど全長にわたってホモロジーモデリングができているが，真核生物では3割程度のORFでしかほぼ全長のホモロジーモデリングができていない（図1）．ほとんどの場合は，ドメイン単位のモデル構造である．ゲノムにコードされている蛋白質の全アミノ酸残基の数を基にして，モデル構造が構築できた割合の変化をみると次ようになる．

筆者らがFAMSBASEを最初に公開した2002年の段階では，古細菌では約38%，真正細菌では約40%のアミノ酸残基部分がモデル立体構造に含まれていた[1]．また2004年の段階で，真核生物ゲノムの約24%部分がモデル立体構造に含まれていた．2005年にはそれぞれ42%，46%，26%に増加している．この増加率が維持されるならば，原核生物の全水溶性蛋白質の全モデル構造を得るには，あと17年から25年かかる計算になる．ゲノムにコードされている蛋白質の25%程度は膜蛋白質と推定されており，これらは計算から除外した．このときに得られる全モデル構造は，ドメイン単位の構造であることから，各ドメインがどのような空間配置をとるのかを推定することが今後は必要になる．

文 献

1) Yamaguchi, A., Iwadate, M., Suzuki, E., Yura, K., Kawakita, S., Umeyama, H., Go, M.: *Nucl. Acids Res.*, **31**, 463-468 (2003)
2) Ogata, K., Umeyama, H.: *J. Mol. Graph. Model.*, **18**, 258-272, 305-256 (2000)
3) Vitkup, D., Melamud, E., Moult, J., Sander, C.: *Nat. Struct. Biol.*, **8**, 559-566 (2001)
4) Brenner, S. E.: *Nat. Struct. Biol.*, **7** (Suppl), 967-969 (2000)

Kei Yura[1], Akihiro Yamaguchi[2], Mitiko Go[2,3], [1]日本原子力研究所量子生命情報解析グループ，[2]長浜バイオ大学バイオサイエンス学部，[3]お茶の水女子大学

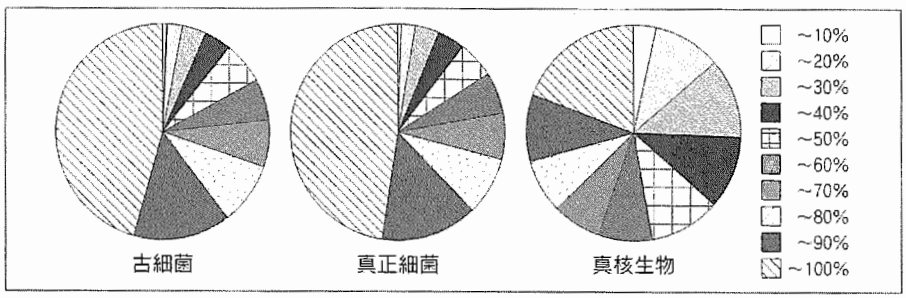

図1 1つのORFにおいてモデル構造が構築できた領域の割合を生物界ごとにまとめた

古細菌と真正細菌においては，1つのORFの90～100%の領域でモデル構造がつくれた場合が，総ORFの半分近くを占めている（円グラフの斜線部分）が，真核生物では1/4にも満たない．