CABIOS APPLICATIONS NOTE

Vol. 13 no. 4 1997
Pages 481–483

Sequence analysis of the Methanococcus jannaschii genome and the prediction of protein function

Miguel Andrade[1], Georg Casari[2], Antoine de Daruvar[1], Chris Sander[1,2] Reinhard Schneider[2], Javier Tamames[3], Alfonso Valencia[3] and Christos Ouzounis[1¶]

With the completion of the genome sequence of *Methanococcus jannaschii* (Bult *et al.*, 1996), computational analysis has revealed a number of interesting predicted functions for this organism. Although the success rate of the initial prediction was <40%, due to a conservative attitude towards possible over-interpretation (Venter, 1996), additional efforts to annotate the sequence follow, contributing a significant increase of functional assignments through a combination of different methods (Kyrpides *et al.*, 1996).

In our continuing effort to annotate the gene products for each complete genome (Casari *et al.*, 1995), we have analyzed the full genomic sequence of *M.jannaschii*, and predicted gene function by sequence similarity. We use GeneQuiz, a system for large-scale sequence analysis (Scharf *et al.*, 1994), which exploits the combination of a number of predictive methods with a rule-based engine that increases the success of predictions thanks to a collection of heuristics and a number of benchmarking cycles (Casari *et al.*, 1996). We have maintained a strong interest in the study of Archaea (Ouzounis and Sander, 1992; Ouzounis *et al.*, 1995), and the analysis of the *M.jannaschii* genome was greatly anticipated.

The scope of this communication is 3-fold: (i) to compare the performance of the GeneQuiz system against the laboriously derived (but highly accurate) manual annotation of the previous attempts; (ii) to discuss some of the cases where GeneQuiz has succeeded or failed; (iii) to assure potential users of the results that the quality of the analysis is high, despite the peculiar biochemical and phylogenetic disposition of this organism.

From a total of 1682 chromosomal open reading frames (ORFs), GeneQuiz has assigned function to 774 (46%) with high confidence ('clear' cases), and, with decreasing confidence, 118 (7%) with probable function ('tentative' cases), while 482 (29%) ORFs have a clear homolog whose function remains unknown. The remaining 308 cases (18%) did not show any detectable similarity to the database. It is interesting to mention that 140 proteins (8%) have a highly reliable homolog in the PDB database, permitting the construction of explicit three-dimensional models.

The original analysis by TIGR provided 625 function predictions (Bult *et al.*, 1996) (37% of the chromosomal genome), with the update increasing them to 809 assignments (Kyrpides *et al.*, 1996) (48%), while a manual correction of the GeneQuiz analysis has readily identified 748 functions (45%) (the 1% difference with the 774 'clear' cases represents false positives—not shown). Of these predictions, 622 assignments from the three groups agree totally, within the margins of name conventions or the identification of different members of the same family. The three assignments where the original annotation has not been confirmed refer to ORFs MJ0029 (coenzyme F420-reducing hydrogenase, alpha subunit) and MJ0941 (transcription initiation factor IIIC) (Bult *et al.*, 1996), not found by either of the other groups, and MJ1068 (O-antigen transporter), not initially identified by GeneQuiz. For another 848, all three groups agree that these proteins are either unique in the database or match a hypothetical protein, without any more information about biochemical function. The sum of the 625 known (622 plus the three conflicting cases) and 848 unknown functions, where all groups agree, amounts to 1473 ORFs, representing 87.6% of the chromosomal genome (Table I). The remaining

Table I. Comparison of the functional assignments for the 1682 chromosomal ORFs from the complete genome of *M.jannaschii*. The tick marks signify the assignment to a possible function, from each group respectively: 848 ORFs have no function, 622 ORFs had the same functional annotation, and with the exception of three cases (TIGR only, TIGR/GQ, TIGR/UIUC), there are 209 cases where there have been significant differences in assignments. Yet, with the update provided by UIUC, almost half of them have been confirmed

TIGR	UIUC	GeneQuiz	Number	Comment
			848	No function
		√	23	GQ only
	√		83	UIUC only
	√	√	103	UIUC/GQ
√			2	TIGR only
√		√	0	TIGR/GQ
√	√		1	TIGR/UIUC
√	√	√	622	All agree

[1]*The European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK,* [2]*European Molecular Biology Laboratory, Heidelberg, Germany and* [3]*National Center for Biotechnology, CSIC, Universidad Autonoma, Cantoblanco, Madrid, Spain*

E-mail: ouzounis@embl-ebi.ac.uk

Table II. The new functional annotations identified by GeneQuiz—ORF number and putative functional assignment based on the most similar protein in the database with a known function. These predictions do not necessarily imply exact function, but functional information from homologs. When sequence similarity is low or the protein family is too diverse, this is marked by 'homolog'. Results for the complete genome along with the details for functional assignment are available at: <http://www.sander.ebi.ac.uk/genequiz/genomes/mj/index.html>

MJ0103	Molybdopterin cofactor synthesis protein homolog
MJ0134	Protein-beta-aspartate methyltransferase
MJ0226	HAM1, controls HAP (6-*N*-hydroxylaminopurine) mutagenesis
MJ0252	Uridine 5′-monophosphate synthase (UMP synthase)
MJ0392	IMP dehydrogenase homolog
MJ0459	Elongation factor 1, EF-1 beta
MJ0568	Diphtheria *tox* repressor (dtxr) homolog
MJ0590	Succinyl-CoA ligase (GDP-forming), alpha chain precursor (EC 6.2.1.4) homolog
MJ0682	Replication factor C (DNA polymerase B, protein splicing)
MJ0797	Cell division protein FtsX homolog
MJ0823	CO induced hydrogenase nickel-insertion accesssory protein cooC, minD family
MJ1079	Spore germination protein B2 (putative transporter)
MJ1129	Mrp protein homolog
MJ1157	GMP synthase (glutamine-hydrolyzing) (EC 6.3.5.2) (GMP synthetase) homolog
MJ1207	Protease synthase and sporulation negative regulatory protein PAI 1
MJ1310	Na/H^+ antiporter system ORF3 homolog
MJ1318	Similar to ATP-dependent protease LA
MJ1336	ADP-heptose synthase (rfaE)
MJ1375	Putative O-antigen transporter
MJ1452	rRNA adenine *N*-6-methyltransferase (EC 2.1.1.48)
MJ1533	Mannose-sensitive hemagglutinin E (mshE) homolog
MJ1618	Polyketide synthase CurC
MJ1621	Putative O-antigen transporter (fragment), central region of MJ1068

209 ORFs (12.4%) have been checked case by case, so that different assignment sets by the three groups are identified and their performance is compared (Table I).

There were 83 (plus MJ1068, mentioned above) cases where the UIUC/TIGR groups have been more successful than the automatic sequence analysis, a fact mainly attributed to multiple sequence analysis and profile searches (Kyrpides *et al.*, 1996). Another 103 cases have been identified by the UIUC/TIGR update (Kyrpides *et al.*, 1996), also found by GeneQuiz. This update is available on the World Wide Web at the URL: http://www.tigr.org/tdb/mdb/mjdb/updates/update_090596.html.

We have been able to discover an additional 23 novel functional assignments, even after the update provided by the UIUC/TIGR groups (Table II). Among them, there are some interesting functions, such as a protein-beta-aspartate methyltransferase (MJ0134), another IMP dehydrogenase homolog (MJ0392), a succinyl-CoA ligase (GDP-forming) alpha-chain precursor (EC 6.2.1.4) homolog (MJ0590), replication factor C, a self-splicing DNA polymerase B (MJ0682), an ADP-heptose synthase (MJ1336), a rRNA adenine *N*-6-methyltransferase (MJ1452) and a mannose-sensitive hemagglutinin E homolog (MJ1533).

From a biological standpoint, it is interesting that even with the combined efforts of three groups, *M.jannaschii* still remains a unique case for genome analysis, with only 40% of its proteins having a homolog of known function, compared to 60% for *Haemophilus influenzae* (Casari *et al.*, 1995). From a methodological standpoint, it seems that despite significant experience in sequence analysis, the problem of differences in exact functional assignment with various methods and approaches remains with us. With more exchange of information during and subsequent to the original analyses, a clearer view of the genome contents will be obtained, with significant impact on the best possible dissemination of painstakingly obtained genome sequence data.

Acknowledgements

We are indebted to the UIUC group (in particular Nikos Kyrpides) for open exchange of results, and various discussions about cases of conflicting annotation. A table with the comparison of functional assignments is available at the following URL: <http://www.sander.ebi.ac.uk/genequiz/genomes/mj/mj.comparison.html>.

References

Bult,C.J. *et al.* (1996) Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.

Casari,G. *et al.* (1995) Challenging times for bioinformatics. *Nature*, **376**, 647–648.

Casari,G., Ouzounis,C., Valencia,A. and Sander,C. (1996) GeneQuiz II: automatic function assignment for genome sequence analysis. In Hunter,L. and Klein,T.E. (eds), *First Annual Pacific Symposium on Biocomputing*. World Scientific, Hawaii, pp. 707–709.

Kyrpides,N.C., Olsen,G.J., Klenk,H.-P., White,O. and Woese,C.R. (1996) *Methanococcus jannaschii* genome: revisited. *Microb. Compar. Genomics*, **1**, 329–338.

Ouzounis,C. and Sander,C. (1992) TFIIB, an evolutionary link between the transcription machineries of archaebacteria and eukaryotes. *Cell*, **71**, 189–190.

Ouzounis,C.A., Kyrpides,N.C. and Sander,C. (1995) Novel protein families in Archaean genomes. *Nucleic Acids Res.*, 1995, **23**, 565–570.

Scharf,M., Schneider,R., Casari,G., Bork,P., Valencia,A., Ouzounis,C. and Sander,C. (1994) GeneQuiz: a workbench for sequence analysis. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Intelligent Systems for Molecular Biology 1994*. AAAI Press, Stanford, CA, pp. 348–353.

Venter,J.C. (1996) More *Haemophilus* and *Mycoplasma* genes—response. *Science*, **271**, 1303–1304.

Received on February 13, 1997; revised and accepted on April 15, 1997