Article No. mb981884 J. Mol. Biol. (1998) 280, 323–326

JMB

COMMUNICATION

Homology-based Fold Predictions for *Mycoplasma genitalium* Proteins

Martijn Huynen[1], Tobias Doerks[1], Frank Eisenhaber[1] Christine Orengo[2], Shamil Sunyaev[1], Yanping Yuan[1] and Peer Bork[1]

[1]*EMBL, Meyerhoftstr.1 69012 Heidelberg and Max-Delbrück-Center for Molecular Medicine, Berlin-Buch Germany*

[2]*Dept. of Biochemistry and Molecular Biology, University College, London, UK*

**Corresponding author*

Homology search techniques based on the iterative PSI-BLAST method in combination with various filters for low sequence complexity are applied to assign folds to all *Mycoplasma genitalium* proteins. The resulting procedure (implemented as a web server) is able to predict at least one domain in 37% of these proteins automatically, with an estimated accuracy higher than 98%. Taking structural features such as coiled coil or transmembrane regions aside, folds can be assigned to more than half of the globular proteins in a bacterium just by iterative sequence comparison.

Keywords: structure prediction; genome analysis; *Mycoplasma genitalium*; homology modeling; fold assignment

In order to understand the molecular function of proteins, structural knowledge is essential, but three-dimensional structures have been determined by direct experiments only for a very small fraction of sequenced gene products. Thus, fold predictions *via* homology to a protein with known three-dimensional structure (Doolittle, 1987) or by utilizing structure-specific properties (i.e. threading methods, see Fischer *et al*., 1996), is an important step towards functional and structural characterization of any gene.

Recently, several groups have analysed complete genomes to assign structural information to sequences therein and came up with greatly varying numbers for the fraction of homologues with known three-dimensional structures that are detectable by sequence similarity searches. The respective fraction in *Mycoplasma genitalium* (frequently used as a ''benchmark'' genome because of its small size) was given as 8.5% (Genequiz consortium, 1996; http://columbra.ebi.ac.uk:8765/), 9.5% (Editorial NSB, 1997), 12.2% (Frishman & Mewes, 1997) or 16% (Fischer & Eisenberg, 1997). The latter found that an additional 9% of the *M. genitalium* proteins folds can only be predicted when incorporating structural information. It is difficult to compare all these numbers as (i) an increase of assignments in time has to be considered due to database growth (Bork & Koonin, 1998) and (ii) the authors all used different criteria to describe terms like ''significant similarity'' or ''clear homology''.

In any case, ''assignment by sequence similarity'' implied a pairwise comparison; motif and profile searches have, however, been shown to be much more sensitive (reviewed by Bork & Gibson, 1996). Thus, we have used PSI-BLAST, an iterative, profile-like approach (Altschul *et al*. 1997), to explore the fraction of *M. genitalium* proteins that have homologues with known three-dimensional structure using an expected ratio of false positives $E = 0.001$ as a threshold. Surprisingly, we found that 37% of all *M. genitalium* proteins can be automatically assigned to this fraction (for a complete list see http://www.bork.embl-heidelberg.de/3D/MG.pred). The generality of this result was obtained by repeating the procedure for *Escherichia coli* (Blattner *et al*., 1997). Here 3D assignment was possible for more than 33% of the proteins (http://www.bork.embl-heidelberg.de/3D/EC.pred).

In order to minimise the detection of false positives, we have pre-processed each *M. genitalium* sequence: Coiled coil regions were identified and masked using the Coils2 program (Lupas, 1997). SEG (Wootton & Federhen, 1996) was used to filter other compositionally biased regions. Transmembrane regions were detected and excluded using the TMpred program (Hofmann & Stoffel, 1993). Coiled coil regions were also removed from the

Abbreviations used: PDB, Protein Data Bank.

0022–2836/98/280323–04 $30.00/0

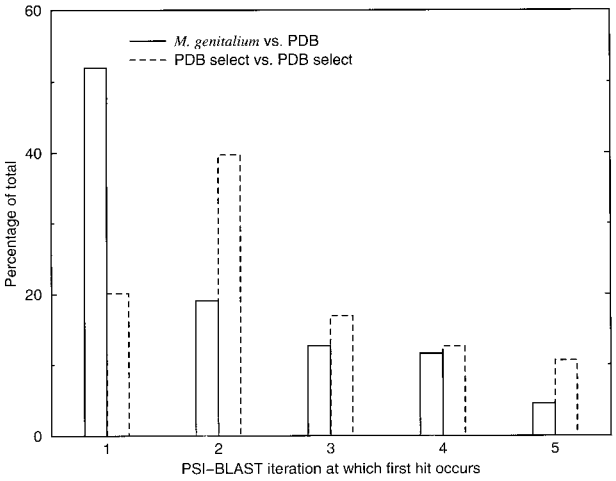

Figure 1. Comparison of the PSI-BLAST searches of *M. genitalium* sequences (left columns) with those from the selected PDB set (right columns). Due to the selection criteria for the PDB select hits, only 20% of the total hits can be found directly by gapped BLAST (first iteration). In contrast, more than half of the *M. genitalium* sequences that were identified to have a 3D homologue in the first iteration, can be related to their structural relative by a pairwise (gapped BLAST) comparison. This indicates that the accuracy of fold assignment for the *M. genitalium* proteins might be even higher than the 98% estimated for the PDB sequences. The data also allow a direct comparison of the sensitivity of PSI-BLAST *versus* gapped BLAST (only first iteration): i.e. nearly twice as many hits of *M. genitalium* to PDB proteins can be found when using family information.

PDB (Bernstein *et al*., 1977) as were all model structures and PDB entries with an insufficient primary structure (e.g. all residues being alanine or ''unknown''). Each *M. genitalium* sequence was then compared against a non-redundant (in terms of identity) protein sequence database (NRDB) at the EMBL (nrdb script provided by the NCBI) using PSI-BLAST. Hits to sequences with known three-dimensional structure below the given threshold ($E = 0.001$) were recorded.

We used this rather stringent, theoretically calculated E value to avoid false positives. In addition, we benchmarked our procedure using an approach similar to that of Brenner *et al*. (1995). A set of 685 sequences that have a pairwise level of identity of less than 25% was extracted from PDB using the PDBselect program (Hobohm & Sander, 1994). These sequences were compared with NRDB with the same procedure as described above. 602 pairs of sequences (including reciprocal hits) from the PDBselect set were identified by the procedure.

The pairs were checked for structural similarity by reference to the CATH classification of protein domain structures (see http://www.biochem.ucl.ac.uk/bsm/cath/) and by structural alignment using the SSAP program (Orengo & Taylor, 1996). The 16 pairs without clear structural similarity (SSAP scores unavailable, or smaller than 70%) were checked manually soliciting information from the literature (see http://www.bork.embl-heidelberg.de/3D/PDBsel for the complete list). This procedure resulted in only 11 false positives (mostly due to a bias in cysteine residues that had not been filtered with the programs used).

The resulting value of $1 - (11/602) = 98.2\%$ can be regarded as a lower limit for the accuracy of the fold predictions for the *M. genitalium* sequences. This is because the benchmarking was done with sequences that have a pairwise similarity that is smaller than 25%. A large fraction of all *M. genitalium* proteins (19%) have similarities to sequences in PDB that are higher than 25% and can easily be recognized in the first PSI-BLAST iteration (Figure 1). This fraction of 19% roughly corresponds to the results of previous attempts (see above) that were based on pairwise similarity searches. The gain of 18% in subsequent iterations demonstrates quantitatively the power of profile-based search techniques implicit in PSI-BLAST (Altschul *et al*., 1997).

The *M. genitalium* sequences have non-globular regions that can cause spurious hits with low E values in PSI-BLAST, while the accuracy was determined with PDB sequences which tend to be globular and might not be fully representative for bacterial sequences. However, a careful manual inspection of all the hits for *M. genitalium* sequences did not reveal any obvious hit to transmembrane or coiled coil regions, a further indication for the usefulness of the filters that were

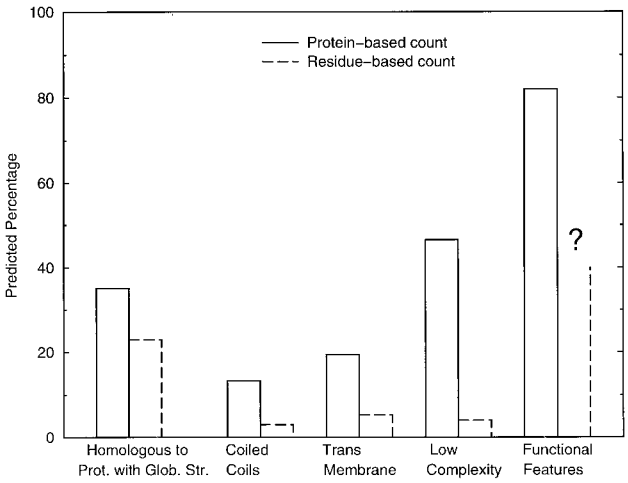

Figure 2. Prediction of functional and structural features for *M. genitalium* proteins. For each structural class both the percentage of proteins and the percentage of residues that have the various structural features are indicated. The percentage of proteins that are homologous to a protein with a determined globular structure was obtained with iterative homology searches (PSI-BLAST). The percentages of proteins with at least some functional features predicted is from Koonin *et al*. (1997). The percentages of coiled coil, trans-membrane and low complexity regions have been predicted with Coils2 (Lupas, 1997), TMpred (Hoffmann & Stoffel, 1993) and SEG (Wootton & Federhen, 1996) respectively.

Table 1. Multidomain proteins in *M. genitalium* as revealed by several disjoint PDB matches

Protein	Function	From–to	PDB-entry
MG036	Aspartyl-TRNA synthetase	3–94	1KRT
		116–362	1ADJ-A, 1ADY-A
MG069	Glucose-permease IIABC component	20–507	1OCC-A
		633–710	1IBA
		743–904	1F3G, 1GLE-F
MG089	Elongation factor G	11–170	1EFM, 821P
		226–410	1EFU-A
MG113	Asparaginyl-TRNA synthetase	10–104	1KRT
		124–435	1ADJ-A
MG136	Lysyl-TRNA synthetase	32–135	1KRS, 1KRT
		162–472	1ADJ-A
MG142	Translation initiation IF-2	120–284	1ETU, 821P
		409–619	1TUI-A, 1AIP-E
MG196	Translation initiation IF-3	2–48	1TIF
		53–139	1TIG, 1IFE
MG272	Dihydrolipoamide acetyltransferase	15–67	1BDO, 1LAC
		142–383	1EAF, 1DPB
MG305	DNAK protein	5–366	3HSC, 1ATR
		370–582	1DKZ-A
MG329	Hypothetical GTP-binding protein	4–171	1EFT
		176–352	821P, 1GNR

used (see also Bork & Koonin, 1998) and for the high accuracy of the method.

As the described fold prediction procedure is fast and highly accurate, we have opened a web-server (http://www.bork.embl-heidelberg.de/3D/) that allows reproduction of the results and that is able to check any sequence for homology with known three-dimensional structures.

Although for 35% of the *M. genitalium* proteins fold assignments are possible, for more than half of these, the predictions do not cover the entire sequence, i.e. not all putative domains were recognized. Thus, the ''coverage'' of fold assignments in *M. genitalium* should also be measured on a residue basis: about 23% of all residues in the *M. genitalium* proteins have been identified to belong to a known fold (Figure 2). Nevertheless, in some cases, several structures were matching distinct parts of the sequences and distinct domains could be assigned (Table 1). Interestingly, the shortest match covers 35 amino acids and only two more matches were smaller than 60 residues, hence we can assume that not only supersecondary structure elements but also considerable fractions of domains are aligned. In many cases, however, this may not be sufficient for homology based modeling.

Although compositionally biased regions such as coiled coil or transmembrane segments were filtered out for the homology searches, they do provide additional information about structural properties within proteins, but also about the limits of current prediction methods. At least 3% of all the residues are located in coiled coil regions (predicted using the COILS program (see Lupas, 1997) with a stringent cutoff). Another 4% of residues reside in low complexity regions (predicted using SEG (Wootton & Federhen, 1996) with stringent cutoffs) for which we do not have a good structural knowledge (Figure 2). These are conservative estimates; using standard parameters (on the cost of a few false positives) these numbers should be higher (Koonin *et al*., 1997).

Transmembrane segments are another distinct structural feature which can be predicted independently (Rost & O'Donoghue, 1997). Predictions for the fraction of membrane proteins in *M. genitalium* vary greatly. Depending on the programs, cutoffs and parameters used, the corresponding fraction (including some proteins with signal sequences) has been estimated as 18% (Fischer & Eisenberg, 1997), 24% (Koonin *et al*., 1997), 30% (Arkin *et al*., 1997) and 36% (Frischman & Mewes, 1997). In any case, transmembrane proteins might also contain globular domains, and these cannot simply be neglected for fold assignments (e.g. there are at least 16 ABC transporters in *M. genitalium* with transmembrane and ATPase domains).

In summary, we have shown that fold assignment *via* homology appears to be an extremely powerful and efficient approach. We have used stringent thresholds for automatic assignment and thus there is a great potential for further improvements of homology based fold prediction. It is clear, however, that the current level of fold prediction often does not imply any mechanistic insights with respect to binding or other functional features.

Acknowledgements

We thank an anonymous referee for his constructive comments. The work was supported by DFG Bo 1099/3-2 and BMBF MEDSEQ grants. We thank F. Milpetz for providing tools and web pages.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, G., Zhang, Z., Miller, W. & Lipman, D. J. (1997).

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
Arkin, I. T., Brunger, A. T. & Engelman, D. M. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins: Struct. Funct. Genet.* **28**, 465–466.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. *et al.* (1977). The Protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **122**, 535–542.
Blattner, F. E., Plunkett, G., III, Block, C. A. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science,* **277**, 1453–1462.
Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184.
Bork, P. & Koonin, E. V. (1998). Prediction of function from protein sequence: Where are the bottlenecks? *Nature Genet.* **18**, 313–318.
Brenner, S. E., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene duplications in *H. influenzae*. *Nature,* **378**, 140.
Doolittle, R. F. (1987). *On Urfs and Orfs*, University Science Books, Mill Valey, CA.
Editorial, (1997). Structure and the genome. *Nature Struct. Biol.* **4**, 329–330.
Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA,* **94**, 11929–11934.
Fischer, D., Rice, D., Bowie, J. U. & Eisenberg, D. (1996). Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J.* **10**, 126–136.
Frishman, D. & Mewes, H.-W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626–628.
Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.
Hofmann, K. & Stoffel, W. (1993). TMbase: a database of membrane spanning proteins. *Biol. Chem. Hoppe-Seyler,* **347**, 166.
Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of proteins sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* **25**, 619–637.
Lupas, A. (1997). Predicting coiled coil regions in proteins. *Curr. Opin. Struct. Biol.* **7**, 388–393.
Orengo, C. A. & Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617–635.
Rost, B. & O'Donoghue, S. (1997). Sisyphus and prediction of protein structure. *Comput. Appl. Biosci.* **13**, 345–356.
Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.

Edited by G. Von Heijne

(Received 19 January 1998; received in revised form 2 April 1998; accepted 3 April 1998)