doi:10.1016/j.jmb.2004.06.073 J. Mol. Biol. (2004) xx, 1–5

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

COMMUNICATION

Structural Biology Sheds Light on the Puzzle of Genomic ORFans

Naomi Siew[1,2,3] and Daniel Fischer[2,3]*

[1]*Department of Chemistry Ben Gurion University Beer-Sheva 84105, Israel*

[2]*Bioinformatics Group, Department of Computer Science, Ben Gurion University Beer-Sheva 84105, Israel*

[3]*Buffalo Centre of Excellence in Bioinformatics/Computer Science and Engineering University of Buffalo, 901 Washington St, Suite 300 Buffalo, NY 14203, USA*

*Corresponding author

Genomic ORFans are orphan open reading frames (ORFs) with no significant sequence similarity to other ORFs. ORFans comprise 20–30% of the ORFs of most completely sequenced genomes. Because nothing can be learnt about ORFans *via* sequence homology, the functions and evolutionary origins of ORFans remain a mystery. Furthermore, because relatively few ORFans have been experimentally characterized, it has been suggested that most ORFans are not likely to correspond to functional, expressed proteins, but rather to spurious ORFs, pseudo-genes or to rapidly evolving proteins with non-essential roles. As a snapshot view of current ORFan structural studies, we searched for ORFans among proteins whose three-dimensional structures have been recently determined. We find that functional and structural studies of ORFans are not as underemphasized as previously suggested. These recently determined structures correspond to ORFans from all Kingdoms of life, and include proteins that have previously been functionally characterized, as well as structural genomics targets of unknown function labeled as "hypothetical proteins". This suggests that many of the ORFans in the databases are likely to correspond to expressed, functional (and even essential) proteins. Furthermore, the recently determined structures include examples of the various types of ORFans, suggesting that the functions and evolutionary origins of ORFans are diverse. Although this survey sheds some light on the ORFan mystery, further experimental studies are required to gain a better understanding of the role and origins of the tens of thousands of ORFans awaiting characterization.

Keywords: genomic ORFans; evolution; structural biology

Genomic ORFans[1] are orphan ORFs (open reading frames) that share no significant sequence similarity with any ORFs outside the genome where they reside ("singleton" and "paralogous" ORFans), or outside a set of closely related organisms ("orthologous" ORFans).[2–4] Genome sequencing of complete organisms has demonstrated that ORFans are integral components of most newly sequenced genomes;[2,5,6] their fraction in a newly sequenced genome is usually 20–30%,[2,3] and in some cases, up to 60%.[7] Thus, ORFans are accumulating in the sequence databases at a rapid rate; over 30,000 singleton ORFans were counted in a recent census.[8] Because little can be inferred about the functions and structures of ORFans using standard bioinformatics tools, their abundance has been referred to as a "mystery".[3,9] Furthermore, because relatively few ORFans have been studied experimentally,[10–14] many speculations regarding their roles and origins have been proposed. On the one hand, if ORFans correspond to expressed, functional proteins, they may be distant members of known protein families, with similar functions and three-dimensional (3D) structures, but with sequences that have diverged beyond recognition by standard sequence comparison tools.[1,3] In this case, knowing their 3D structures becomes essential in order to assign them to their corresponding families. Alternatively, ORFans may correspond to

Present address: Both authors, Buffalo Center of Excellence in Bioinformatics/Computer Science and Engineering, University at Buffalo, 901 Washington St. Suite 300, Buffalo, NY 14203, USA.

Abbreviations used: ORF, open reading frame; ORFans, orphan ORFs; FR, fold-recognition.

E-mail address of the corresponding author: dfischer@bioinformatics.buffalo.edu

novel proteins, unique to an organism or a lineage, with possibly new functions and/or 3D structures.[1,3] In either case, the mystery of why ORFans have no homologs remains. Are ORFans the result of rapid evolution,[4,15,16] of lateral gene transfer[17] from unknown organisms or are they the result of gene-losses or of *de novo* generation?[5,6,18–20] On the other hand, it has also been suggested that most ORFans, especially the shorter ones, may correspond to non-essential, non-functional or non-expressed proteins.[15,21–24]

Here, as a snapshot view of current ORFan structural studies, we report a recent survey we conducted among newly determined 3D protein structures, and show how structural biology is already being essential in unraveling the ORFan puzzle.[1,10,25] We searched for ORFans among the PDB[26] entries released between June and December 2003 and found that out of the 172 protein chains sharing no significant sequence similarity to previously determined protein structures,[27] 17 correspond to ORFans and two correspond to "poorly conserved ORFs" or PCOs[5,14] (Table 1; for simplicity, in what follows we refer to these 19 proteins as ORFans; see below). This strongly suggests that many ORFans correspond to real, foldable proteins, and not to sequencing errors or dead proteins. The relatively large percentage of ORFans among the newly determined structures (11%) suggests that ORFans may not be as underemphasized as previously suggested[1,3] and that experimental studies of ORFans have already become routine. In what follows, we show that these 19 ORFans provide interesting examples of the various types of ORFans.

Thirteen of these 19 newly determined ORFans correspond to proteins whose function was previously characterized experimentally (at least in the broad sense), and thus, are not "orphans" with regards to their functions. This suggests that many more ORFans with still unknown 3D-structures have already been characterized functionally. These 13 ORFans cover various functional categories, with at least five involved in transcription/translation, suggesting that, because of the high sequence divergence required in these processes, many ORFans may belong to these categories. The other six ORFans correspond to proteins of unknown

Table 1. The 19 ORFans with a recently determined 3D structure

PDB code	Organism	PDB description	Length (aa)	Has homologs in
1of5B[28]	*S. cerevisiae*	Mrna export factor Mex67-Mtr2	184	–
1oq1A	*B. subtilis*	Hypothetical protein Apc1120 (protein Yesu)	223	–
1mw5A	*H. influenzae*	Hypothetical protein Hi1480	187	–
1pp8U[34]	*T. vaginalis*	Initiator binding domain (Ibp39)	132	–
1q87A[34]	*T. vaginalis*	C-domain of the Inr binding protein	138	–
1q77A	*A. aeolicus*	Putative Universal Stress Protein (Hypothetical Protein Aq_178)	221	–
1rf8B[38]	*S. cerevisiae*	Translation initiation factor Eif4E	100	–
1nycA[29]	*S. aureus*	Staphostatin B	111	*S. warneri*
1oh1[39]	*S. aureus*	Staphostatin A (hypothetical protein Sav1910)	109	*S. epidermidis*
1q8cA	*M. genitalium*	Conserved hypothetical protein (Mg027)	151	*M. pneumoniae*
1osyA[40]	*F. velupites*	Fip-Fve fungal immunomodulatory protein	115	*G. lucidum*
1r75A	*L. major*	Hypothetical protein	110	*T. brucei*
1nigA	*T. volcanium*	Hypothetical protein (Ta1238)	152	*T. acidophilum*, *F. acidaramanus*
1ofzA[41]	*A. auranta*	Fungal lectin	312	*A. oryzae*, *A. fumigatus*
1uf2P[42]	*Rice dwarf virus*	Rice dwarf virus (capsid protein)	421	Wound tumor virus, Rice gall dwarf virus
1q6aA[43]	*T. elongatus*	Circadian clock protein KaiA homolog	214	*T. erythraeum*, *Nostoc punctiforme*, *Nostoc* sp. *PCC 9709*, *PCC 7120* & *P. membranacea Synechococcus* sp. *PCC & WH*, *Synechocystis* sp. *PCC*
1qv9A[44]	*M. kandleri*	Coenzyme F420-dependent Mtd	283	*M. jannaschii*, *M. mazei*, *M. acetivorans*, *M. barkeri*, *M. thermoautotrophicus*, *M. thermoautotrophicum*, *A. fulgidus*
1q5zA[45]	*S. typhimurium*	C-terminal actin binding domain of *Salmonella* invasion protein A (Sipa)	177	*S. enteritidis*, *S. enterica Typhi*, *C. violaceum*[a]
1n93A[35]	*Borna virus*	Nucleoprotein	375	*H. sapiens*, *M. musculus*[b]

ORFans are ORFs lacking significant sequence similarity with other ORFs, except for possibly ORFs from closely related organisms. We assess significant sequence similarity using the standard PSI-BLAST[46] sequence comparison tool until convergence, and the *e*-value threshold of 0.001. References for the PDB codes are included only if they were published before March 2004.

[a] Because of the presence of a homolog in *Chromobacterium violaceum*, 1q5zA is not a proper orthologous ORFan, but rather a poorly conserved ORF or PCO.

[b] 1n93A is also a PCO, which has probably been laterally transferred to humans and mice (see the text).

function (annotated as "hypothetical proteins") whose structure was determined as part of structural genomics projects. Seven of the 19 ORFans correspond to singleton ORFans, and ten correspond to orthologous ORFans (have homologs only within closely related organisms). The 19 ORFans belong to organisms spanning all kingdoms: seven from Bacteria, three from Archea, seven from Eukarya and two viruses.

Despite the fact that ORFans show no significant sequence similarity to other proteins, the 3D structures of the majority of the ORFans clearly have previously observed folds. This suggests that they either correspond to highly divergent distant members of known protein families, with possibly similar functions or to proteins with unrelated functions whose structures have converged to a similar fold. For example, the essential messenger RNA export factor Mtr2 from yeast (1of5B), a singleton ORFan, was known to be similar in function to the metazoan p15 protein.[28] The previously determined 3D structure of the p15 protein revealed that it belongs to the NTF2-like family. The 3D structure of 1of5B revealed that Mtr2 is similar to that of p15, and thus, 1of5B represents a novel member of this family. This is a clear example of an ORFan with a highly divergent sequence, whose function and structure are similar to those of a known family. An example of an ORFan with a previously known fold, but with an unrelated function, is the bacterial virulence factor staphostatin B (1nycA), a cysteine protease inhibitor with a *Staphyloccocus*-specific function.[29] Unexpectedly, its 3D structure turned out not to be similar to other cystatins, but rather, to a variation of the lipocalin fold. This was a surprising result because staphylococci were not expected to contain lipocalin-like functions and no evidence of lipocalin-like properties has been identified in 1nycA. These examples illustrate that the 3D-structures of those ORFans whose broad function is known, can help to better understand their mechanisms of operation, and in some cases reveal their evolutionary origins.

The 3D-structures of the ORFans of unknown function, if similar to previously observed folds, can also be of help to generate verifiable hypotheses regarding their possible function and/or origin. One such example is the 3D-structure of the singleton ORFan of unknown function from *Aquifex aeolicus* (1q77A). Its structural similarity to the family of universal stress proteins lead to a putative functional assignment, which, if true, would imply that 1q77A is another highly divergent member of that family.

Could bioinformatics tools have predicted the approximate structures of those ORFans having previously observed folds? Fold-recognition (FR) methods,[30,31] applied without using the information from the recently released structures,[27] but using as templates previously determined structures, correctly predicted the structures of six of the 19 ORFans. Figure 1 shows the highly confident FR prediction of a *Mycoplasma*-specific hypothetical protein (1q8cA). This *M. genitalium* protein is an orthologous ORFan with a single homolog in *M. pneumoniae*. The FR result predicted that the 3D-structure of 1q8cA is similar to that of the transcriptor regulator NusB from *Mycobacterium*

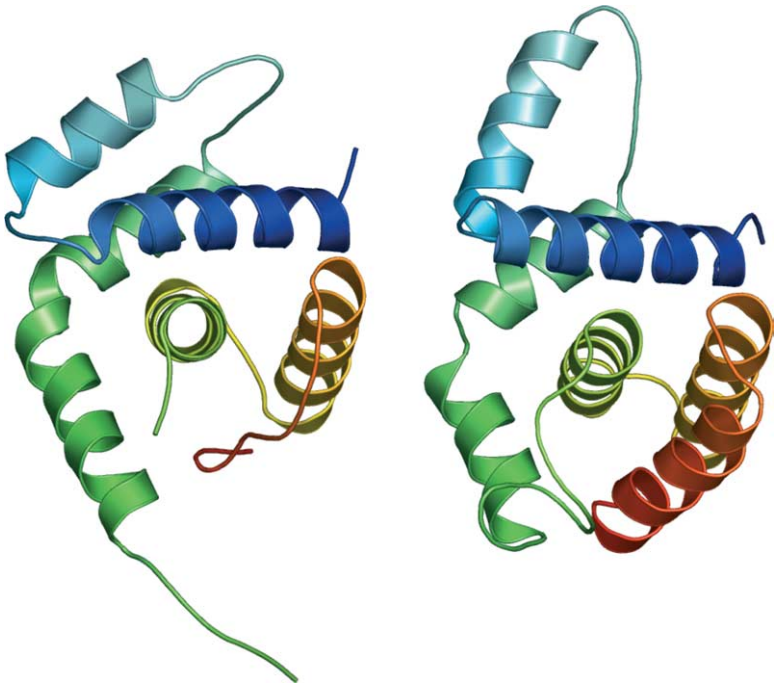

Figure 1. A relatively accurate fold recognition prediction for *M. genitalium's* conserved hypothetical protein MG027, 1q8cA. MG027 corresponds to an ORFan because with the exception of the close relative *M. pneumonia*, it shows no sequence similarity to any other protein in the databases. Nevertheless, fold recognition is able to recognize the similarity between MG027 and a protein of known structure. The fold-recognition 3D model built without using the structural information from 1q8cA is shown on the right. The model, produced by the 3D-SHOTGUN fold-recognition method,[36] is based on the predicted structural similarity of MG027 with the previously released structure of the transcriptor regulator NusB from *Mycobacterium tuberculosis* (1evyA).[32] The sequence similarity between MG027 and 1evyA is only 13%. The prediction is confirmed by the experimental structure of 1q8cA (left). Notice that no experimental data was observed for the loop region preceding the N-terminal two helices. The overall C-alpha RMSD between the predicted model and the experimental structures is 4.9 Å, with 105 residues superimposing with an RMSD of 2.5 Å.[37]

tuberculosis (1eyvA).[32] This prediction is now confirmed by the experimental structure of 1q8cA. Whether 1q8cA is also a transcriptor regulator or not, remains to be verified experimentally.

Because no clear definition of what represents a novel fold exists, it is difficult to tally the exact number of ORFans with novel folds in our survey. One accepted authority with regards to novel folds is the SCOP[33] database. Unfortunately, the latest SCOP release (1.65) only includes two of the 19 ORFans. Thus, we used structural comparison and the available published reports to determine which of the remaining ORFans correspond to novel folds. We could identify only three cases of apparent novel folds: 1q87A, 1q5zA and 1n93A. 1q87A corresponds to the structure of the C-terminal domain of the Inr binding protein from *Trichomonas vaginalis*, which is of unknown function.[34] Another interesting example of a novel fold is the nucleoprotein from Borna disease virus (1n93A).[35] This virus is a member of the *Mononegavirales* that also includes the measles and Ebola viruses. Because homologs of this protein can also be found in mammals (human and mouse), this is not a proper ORFan, but a PCO,[5,14] probably illustrating an example of lateral transfer, which could be one of the mechanisms that has generated other ORFans.

The fact that the 3D structures of the majority of ORFans in our survey correspond to previously observed folds may be satisfying to some extent because no new theories about their structural origins are required. Many more 3D structures of ORFans need to be solved in order to determine whether ORFans will turn out to be an enriched source of novel folds.

In summary, in our small survey we have found examples of ORFans that are distant members of known families, of ORFans with unknown function or with lineage-specific functions and of ORFans with novel, previously unseen folds. Although many of the mysteries concerning ORFans remain (i.e. their origins, their functions, their isolation in sequence space), it is clear from this survey that many ORFans correspond to real, functional, and in some cases, essential, proteins. Surprisingly, 11% of the newly determined structures that we considered corresponded to ORFans, suggesting that ORFans may not be as underemphasized as previously thought. Large-scale structure determinations will be required to obtain further insights about the evolution, the origin(s) and functions of the tens of thousands of ORFans awaiting functional and structural characterization.

Additional material

An expanded Table 1, containing various links to the data is available†.

† http://bioinformatics.buffalo.edu/ORFanage/3DORFans

References

1. Fischer, D. & Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.
2. Siew, N. & Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Struct. Funct. Genet.* **53**, 241–251.
3. Siew, N. & Fischer, D. (2003). Twenty thousand ORFan microbial protein families for the biologist? *Structure (Camb)*, **11**, 7–9.
4. Malpertuy, A., Tekaia, F., Casaregola, S., Aigle, M., Artiguenave, F., Blandin, G. *et al*. (2000). Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Letters*, **487**, 113–121.
5. Siew, N. & Fischer, D. (2003). Unravelling the ORFan puzzle. *Comp. Funct. Genomics* 2003, 432–441.
6. Unger, R., Uliel, S. & Havlin, S. (2003). Scaling law in sizes of protein sequence families: from superfamilies to orphan genes. *Proteins: Struct. Funct. Genet.* **51**, 569–576.
7. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W. *et al*. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.
8. Siew, N., Azaria, Y. & Fischer, D. (2004). The ORFanage: an ORFan database. *Nucl. Acids Res.* **32**, D281–D283.
9. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H. *et al*. (1996). Life with 6000 genes. *Science*, **546**, 563–567.
10. Fischer, D. (1999). Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. *Protein Eng.* **12**, 1029–1030.
11. Monchois, V., Abergel, C., Sturgis, J., Jeudy, S. & Claverie, J. M. (2001). *Escherichia coli* ykfE ORFan gene encodes a potent inhibitor of C-type lysozyme. *J. Biol. Chem.* **276**, 18437–18441.
12. Alimi, J. P., Poirot, O., Lopez, F. & Claverie, J. M. (2000). Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from *Escherichia coli* K-12 MG1655. *Genome Res.* **10**, 959–966.
13. Brenner, S. E. (2001). A tour of structural genomics. *Nature Rev. Genet.* **2**, 801–809.
14. Shmuely, H., Dinitz, E., Dahan, I., Eichler, J., Fischer, D. & Shaanan, B. (2004). Poorly conserved ORFs in the genome of the archaea *Halobacterium* sp. NRC-1 correspond to expressed proteins. *Bioinformatics*, **20**, 1248–1253.
15. Schmid, K. J. & Aquadro, C. F. (2001). The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics*, **159**, 589–598.
16. Domazet-Loso, T. & Tautz, D. (2003). An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* **13**, 2213–2219.
17. Daubin, V. & Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* **14**, 1036–1042.
18. Wolfe, K. H. & Li, W. H. (2003). Molecular evolution meets the genomics revolution. *Nature Genet.* **33**, 255–265.
19. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
20. Long, M. (2001). Evolution of novel genes. *Curr. Opin. Genet. Dev.* **11**, 673–680.
21. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. &

Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17**, 425–428.
22. Amiri, H., Davids, W. & Andersson, S. G. (2003). Birth and death of orphan genes in Rickettsia. *Mol. Biol. Evol.* **20**, 1575–1587.
23. Dujon, B. (1996). The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270.
24. Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M. R. & Cebrat, S. (1999). Origin and properties of non-coding ORFs in the yeast genome. *Nucl. Acids Res.* **27**, 3503–3509.
25. Watson, J. D., Todd, A. E., Bray, J., Laskowski, R. A., Edwards, A., Joachimiak, A. *et al*. (2003). Target selection and determination of function in structural genomics. *IUBMB Life*, **55**, 249–255.
26. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
27. Rychlewski, L., Fischer, D. & Elofsson, A. (2003). LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins: Struct. Funct. Genet.* **53**, 542–547.
28. Fribourg, S. & Conti, E. (2003). Structural similarity in the absence of sequence homology of the messenger RNA export factors Mtr2 and p15. *EMBO Rep.* **4**, 699–703.
29. Rzychon, M., Filipek, R., Sabat, A., Kosowska, K., Dubin, A., Potempa, J. & Bochtler, M. (2003). Staphostatins resemble lipocalins, not cystatins in fold. *Protein Sci.* **12**, 2252–2256.
30. Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. (2001). Structure prediction meta server. *Bioinformatics*, **17**, 750–751.
31. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
32. Gopal, B., Haire, L. F., Cox, R. A., Jo Colston, M., Major, S., Brannigan, J. A. *et al*. (2000). The crystal structure of NusB from *Mycobacterium tuberculosis*. *Nature Struct. Biol.* **7**, 475–478.
33. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
34. Schumacher, M. A., Lau, A. O. & Johnson, P. J. (2003). Structural basis of core promoter recognition in a primitive eukaryote. *Cell*, **115**, 413–424.
35. Rudolph, M. G., Kraus, I., Dickmanns, A., Eickmann, M., Garten, W. & Ficner, R. (2003). Crystal structure of the borna disease virus nucleoprotein. *Structure (Camb)*, **11**, 1219–1226.
36. Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins: Struct. Funct. Genet.* **51**, 434–441.
37. Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. (2000). MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.
38. Gross, J. D., Moerke, N. J., von der Haar, T., Lugovskoy, A. A., Sachs, A. B., McCarthy, J. E. & Wagner, G. (2003). Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E. *Cell*, **115**, 739–750.
39. Dubin, G., Krajewski, M., Popowicz, G., Stec-Niemczyk, J., Bochtler, M., Potempa, J. *et al*. (2003). A novel class of cysteine protease inhibitors: solution structure of staphostatin A from *Staphylococcus aureus*. *Biochemistry*, **42**, 13449–13456.
40. Paaventhan, P., Joseph, J. S., Seow, S. V., Vaday, S., Robinson, H., Chua, K. Y. & Kolatkar, P. R. (2003). A 1.7 Å structure of Fve, a member of the new fungal immunomodulatory protein family. *J. Mol. Biol.* **332**, 461–470.
41. Wimmerova, M., Mitchell, E., Sanchez, J. F., Gautier, C. & Imberty, A. (2003). Crystal structure of fungal lectin: six-bladed beta-propeller fold and novel fucose recognition mode for *Aleuria aurantia* lectin. *J. Biol. Chem.* **278**, 27059–27067.
42. Nakagawa, A., Miyazaki, N., Taka, J., Naitow, H., Ogawa, A., Fujimoto, Z. *et al*. (2003). The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. *Structure (Camb)*, **11**, 1227–1238.
43. Vakonakis, I., Sun, J., Wu, T., Holzenburg, A., Golden, S. S. & LiWang, A. C. (2004). NMR structure of the KaiC-interacting C-terminal domain of KaiA, a circadian clock protein: implications for KaiA-KaiC interaction. *Proc. Natl Acad. Sci. USA*, **101**, 1479–1484.
44. Hagemeier, C. H., Shima, S., Thauer, R. K., Bourenkov, G., Bartunik, H. D. & Ermler, U. (2003). Coenzyme F420-dependent methylenetetrahydromethanopterin dehydrogenase (Mtd) from *Methanopyrus kandleri*: a methanogenic enzyme with an unusual quarternary structure. *J. Mol. Biol.* **332**, 1047–1057.
45. Lilic, M., Galkin, V. E., Orlova, A., VanLoock, M. S., Egelman, E. H. & Stebbins, C. E. (2003). *Salmonella* SipA polymerizes actin by stapling filaments with nonglobular protein arms. *Science*, **301**, 1918–1921.
46. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., 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.

Edited by M. Levitt

(Received 2 March 2004; received in revised form 9 June 2004; accepted 19 June 2004)