doi:10.1016/S0022-2836(02)00467-9 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 320, 713–719

JMB

COMMUNICATION

Proteins with Class α/β Fold Have High-level Participation in Fusion Events

Sujun Hua[1], Tao Guo[1], Julian Gough[2] and Zhirong Sun[1]*

[1]*Institute of Bioinformatics Department of Biological Sciences and Biotechnology Tsinghua University, Beijing 100084, People's Republic of China*

[2]*MRC, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

**Corresponding author*

Now that complete genome sequences are available for a variety of organisms, the elucidation of potential gene products function is a central goal in the post-genome era. Domain fusion analysis has been proposed recently to infer the functional association of the component proteins. Here, we took a new approach to the analysis of the structural features of the proteins involved in fusion events. An exhaustive survey of fusion events within 30 completely sequenced genomes and subsequent structure annotations to the component proteins at a SCOP superfamily level with hidden Markov models was carried out. A domain fusion map was then constructed. The results revealed that proteins with the class α/β fold are frequently involved in fusion events, around 86% of the total 676 assigned single-domain fusion pairs including at least one component protein belonging to the class α/β fold. Moreover, the domain fusion map in our work may offer an attractive framework for designing chimeric enzymes following Nature's lead, and may give useful hints for exploring the evolutionary history of proteins.

Keywords: domain fusion; structural annotation; domain fusion map; hidden Markov models; chimeric enzymes

The major goal of genome projects is to determine the structure and function of all newly identified gene products. Domain fusion analysis was recently introduced by Marcotte *et al.*[1,2] and Enright *et al.*[3] It has been shown that the detection of fused proteins in one genome (defined as composite proteins) can be used for inferring functional associations between the homologs (defined as component proteins) that are separate in another genome.[3] The strong functional associations of the component proteins, including direct protein–protein physical interactions or similar cellular roles, have been proved further with more genomic data.[4,5]

Here, we took a new approach to the analysis of the structural features of the proteins involved in fusion events. Several important and fundamental questions can be derived, such as whether some aspects of the structure of proteins involved in fusion events predispose these proteins to form stable chimeric proteins more easily. To address these problems, an exhaustive survey of fusion events within 30 genomes and subsequent structure annotations to the component proteins at a superfamily level with hidden Markov models were carried out. A domain fusion map was then constructed and the results clearly showed that proteins with the class α/β fold are frequently involved in fusion events.

Detecting fusion events within 30 genomes

According to the criteria proposed by Yanai *et al.*,[5] fusion events were detected using the BLASTp search[6] of queried protein sequences against the non-redundant protein sequence database nrdb90.[7] Two proteins were identified as component proteins in a fusion event if each had an alignment of at least 80 residues with the same nrdb90 protein with a cut-off E-value of 1×10^{-10} and with a maximum overlap of 20 residues between these two alignments. To avoid inflation of the number of fusion events by the paralogs, the fusion links were collapsed by allowing at most a single link between any two clusters of paralogs.[5] This filtering procedure obviated the consideration of promiscuous genes or domains.[1]

Abbreviations used: HMM, hidden Markov models; SCOP, the Structural Classification of Proteins database.

E-mail address of the corresponding author: sunzhr@mail.tsinghua.edu.cn

An exhaustive survey within 30 genomes (*Pyrococcus abyssi*†, *Aquifex aeolicus*,[8] *Pseudomonas aeruginosa*,[9] *Borrelia burgdorferi*,[10] *Vibrio cholerae*,[11] *Escherichia coli*,[12] *Xylella fastidiosa*,[13] *Archaeoglobus fulgidus*,[14] *Mycoplasma genitalium*,[15] *Pyrococcus horikoshii*,[16] *Haemophilus influenzae*,[17] *Methanococcus jannaschii*,[18] *Campylobacter jejuni*,[19] *Thermotoga maritima*,[20] *Neisseria meningitidis* MC58,[21] *Mycoplasma pneumoniae*,[22] *Treponema pallidum*,[23] *Aeropyrum pernix*,[24] *Chlamydia pneumoniae* CWL029,[25] *Rickettsia prowazekii*,[26] *Helicobacter pylori* 26695,[27] *Helicobacter pylori* J99,[28] *Deinococcus radiodurans*,[29] *Bacillus subtilis*,[30] *Synechocystis* sp.,[31] *Methanobacterium thermoautotrophicum*,[32] *Chlamydia trachomatis*,[33] *Mycobacterium tuberculosis*,[34] *Ureaplasma urealyticum*,[35] and *Saccharomyces cerevisiae*[36]) yielded 10,073 fusion events.[5]

Assigning structural domains to the protein sequences

The domain definition used here is that of the Structural Classification of Protiens (SCOP) database, a hierarchical classification of all domains of known three-dimensional structure.[37] The structural and sometimes functional features of the SCOP domains in superfamilies strongly suggest a common evolutionary origin. This work used SCOP version 1.55 containing 31,474 structural domains clustered into 947 superfamilies.

The SCOP domains were assigned to protein sequences using hidden Markov models (HMMs).[38] For each non-identical SCOP domain, at the 95% sequence identity level, an HMM was generated by Gough *et al.*[39] using the iterative SAM-T99 method.[40] The protein sequences were scanned against this HMM library to find matches with a maximum expectation value of 1×10^{-5}. With this procedure, the structure assignments, or giving the SCOP superfamily identifiers, cover about 45% of prokaryote and 35% of eukaryote sequence.[41,42]

Constructing a domain fusion map

To obtain an unambiguous domain fusion relationship, we focused on the fusion events in which both component proteins were single-domain proteins. A protein sequence was considered to be a single-domain protein if the regions flanking the matched SCOP domain were less than 30 residues long. Thus, we obtained 4033 fusion pairs among the 10,073 events where at least one component was assigned as a single-domain protein, and 941 fusion pairs where both component proteins had single-domain protein assignments. Furthermore, to obtain clearly complete domain fusion relationships, we discarded 265 single-domain fusion pairs where at least a fusion segment in one of the component proteins did not cover 70% of the full sequence length. Ultimately, 676 single-domain fusion pairs were obtained where both component proteins had unique SCOP superfamily identifiers.

Structural domain assignment was used to view the 676 fusions between individual domains in terms of 129 different types of fusions between pairs of protein superfamilies. The domain fusion map in Figure 1 was used to understand how the fusions between protein superfamilies were organized on a large scale: 111 of the 947 superfamilies in SCOP were found to occur in 676 fusion links. Most of these superfamilies were observed in fusion links with less than three other superfamilies, while a few superfamilies were relatively versatile in their fusion behavior. The most versatile superfamilies included P-loop-containing nucleotide triphosphate hydrolases (SCOP identifier c.37.1), CheY-like proteins (c.23.1), FAD/NAD(P)-binding domains (c.3.1), thioredoxin-like proteins (c.47.1), PLP-dependent transferases (c.67.1), and firefly luciferase-like proteins (e.23.1) with 11, 11, seven, six, six, and six fusion partners, respectively.

As shown in Figure 1, protein superfamilies with the class α/β fold constituted about one-half (53 out of 119) of the superfamilies involved in fusion events. Furthermore, 583 fusion events, around 86% of the total of 676, included at least one component protein belonging to the class α/β fold. A total of 304 of these 583 fusion links involved component proteins that were both of the class α/β. Similarly, 100 out of 129 types of superfamily fusion pairs contained component proteins with the class α/β fold. In addition, almost all of the top 11 versatile superfamilies that have at least five fusion partners in the domain fusion map belonged to the class α/β fold, except for the firefly luciferase-like superfamily (e.23.1).

To test if the impressively large fraction of fusion events containing class α/β fold proteins was due to the structural features themselves but not other factors, for example, the overrepresentation of class α/β proteins in the Protein Data Bank (PDB) or SCOP[43] and, thus, in the annotated sequences, a rigorous statistical test was performed. Of the total annotated single-domain proteins within the 30 genomes, we randomly selected 1000 proteins with the class α/β and another 1000 proteins that were not of the class α/β. (The total annotated single-domain proteins within the 30 genomes represent 9673 unique proteins, and among these 9673 proteins, there are 5606 unique proteins with the class α/β fold.) We then calculated the number of proteins involved in detected fusion events in these two protein sets. We repeated this entire process 100 times and found that an average of 248 ± 15 components of the class α/β fold were involved in fusion events, almost twice the average of 129 ± 10 not belonging to the class α/β (Figure 2). The Wilcoxon rank sum test produced a *P*-value of less than 1×10^{-7}, which is highly significant.

† The genome sequence is available at http://www.genoscope.cns.fr

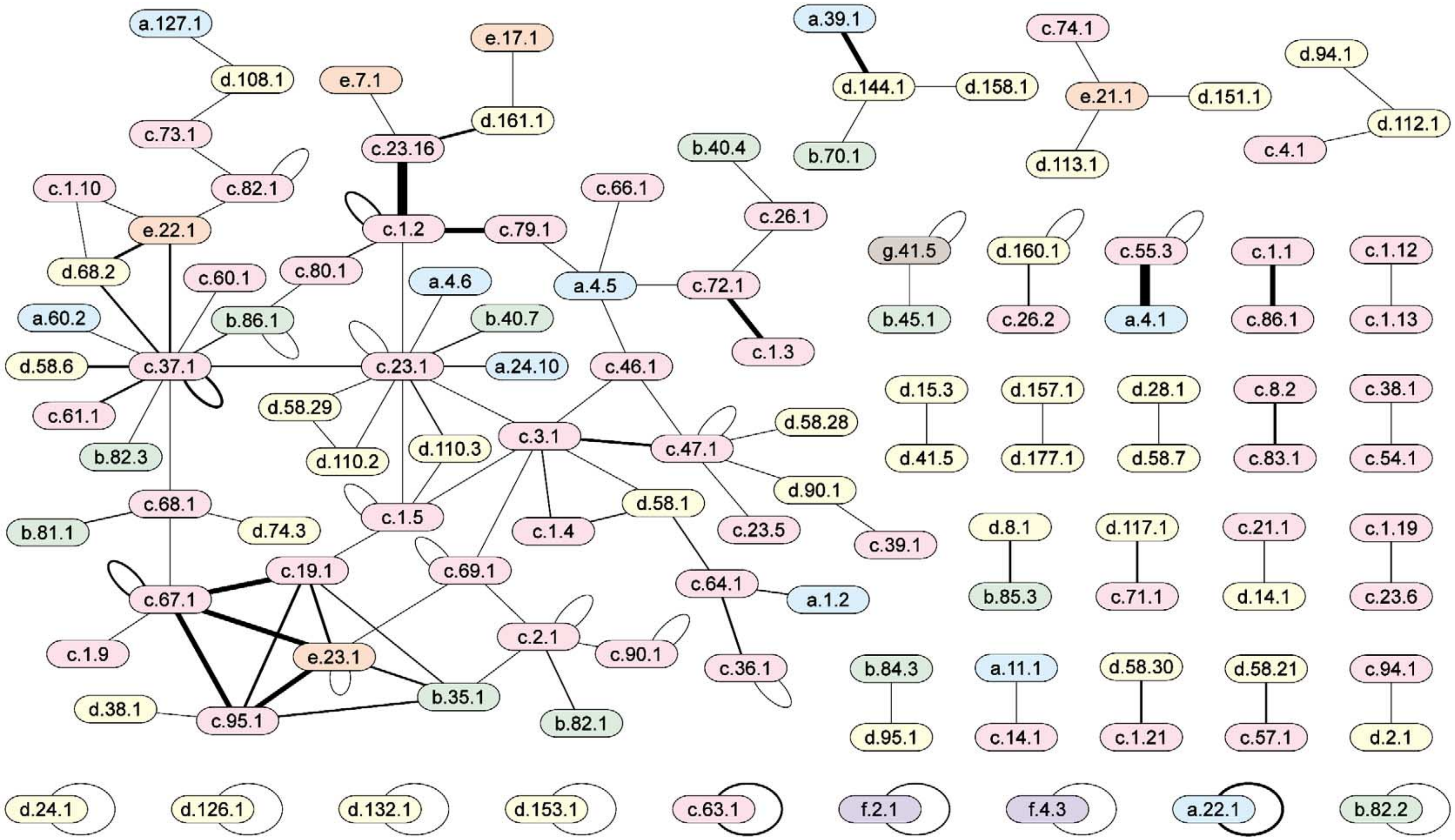

Figure 1. The domain fusion map. Each node represents a protein superfamily distinguished by the SCOP identifier. The nodes are color-coded according to the type of the fold class in SCOP (blue, all-α proteins; green, all-β proteins; pink, α/β proteins; yellow, α + β proteins; orange, multidomain proteins; purple, membrane and cell-surface proteins; gray, small proteins). A line connecting two nodes indicates that there occurs at least one fusion event with two component proteins belonging to the two respective superfamilies. If a fusion event in which two component proteins belonging to the same superfamily exists, the superfamily is labeled with a loop. The width of the line or the loop is proportional to the number of protein superfamily fusions it undergoes. In all, 61 out of the total of 119 superfamilies in the map form a big connected cluster.

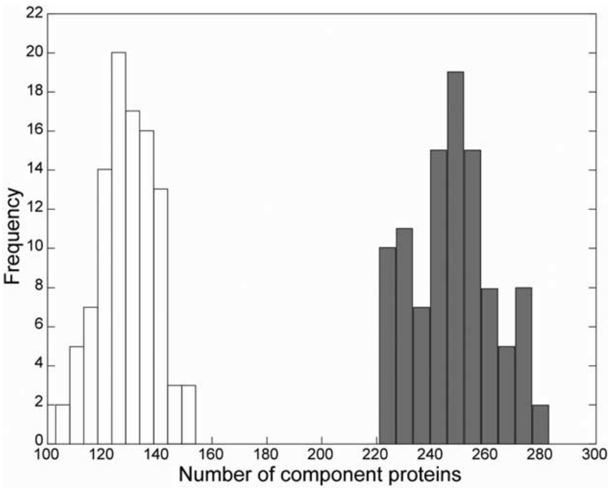

Figure 2. Proteins with the class α/β fold frequently involved in fusion events. Distribution of detected component proteins involved in fusion events from the set of 1000 randomly selected α/β fold single-domain proteins within the 30 genomes (filled bars) and from the set of 1000 randomly selected single-domain proteins not belonging to the class α/β (open bars). The data analysis showed that proteins with the class α/β fold have higher levels of participation in fusion events. The Wilcoxon rank sum test produced a P-value of less than 1×10^{-7}, which is highly significant.

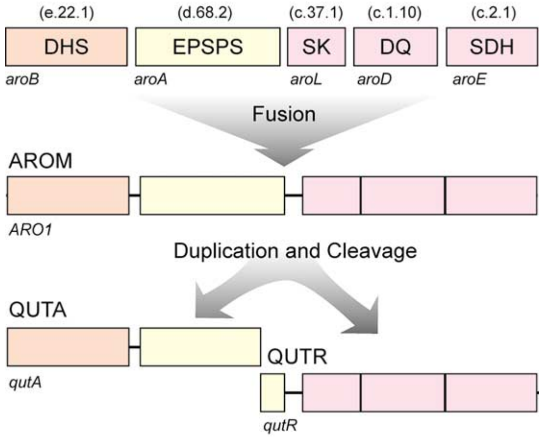

Figure 3. Modular structure of enzymes and regulatory proteins involved in the shikimate and quinate pathways. The upper five boxes represent five monofunctional enzymes catalyzing steps 2 to 6 in the prechorismate section of the shikimate pathway in *E. coli*: *aroB*, dehydroquinate synthase (DHS); *aroA*, 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS); *aroL*, shikimate kinase (SK); *aroD*, type-1 3-dehydroquinate dehydratase (DQ); and *aroE*, shikimate dehydrogenase (SDH). The structural annotation (SCOP identifier) for each enzyme is placed above the appropriate box. Among these assignments, the structures of DHS, EPSPS, SK and DQ have been proved experimentally.[60–63] The protein AROM (encoded by *ARO1* in *S. cerevisiae*), a penta-functional enzyme in the shikimate pathway, evolved from multiple gene fusions. It was also proposed by the genetic, biochemical and physiological analysis that two regulatory proteins, a transcription activator protein, QUTA, and a repressor protein, QUTR, in the quinate utilization pathway, which shares two metabolites with the shikimate pathway, have arisen *via* duplication and splitting of the gene encoding the protein AROM in *A. nidulans*.[45,46]

The data analysis has shown that proteins with class α/β fold are involved frequently in fusion events. It was noted that most pairs of component proteins with known functions appear to be metabolic enzymes.[3,44] This indicates a potential correlation between α/β folds and enzymes, especially for single-domain proteins, as was suggested by Hegyi & Gerstein.[43]

It is worthy of mention that, although most superfamilies in the map linked by fusion events have typical catalytic domains, some pairs have one catalytic component while the other component is a typical DNA-binding motif, which may help to explain the cellular complexity of higher organisms. An interesting example is shown in Figure 3.

It was revealed by our domain fusion analysis that protein AROM (encoded by *ARO1* in *S. cerevisiae*), a penta-functional enzyme in the shikimate pathway, was evolved from fusions of five separate genes (*aroB*, *aroA*, *aroL*, *aroD*, and *aroE* in *E. coli*) encoding five mono-functional enzymes in the shikimate pathway. It was suggested by the genetic, biochemical and physiological analysis that two regulatory proteins, a transcription activator protein, QUTA, and a repressor protein, QUTR, in the quinate utilization pathway, which shares two metabolites with the shikimate pathway, have arisen *via* duplication and splitting of the gene encoding the protein AROM in *Aspergillus nidulans*.[45,46] The QUTR repressor protein acts as a molecular sensor that detects the presence of quinate pathway intermediates and mediates its repressing effect by binding directly to the QUTA protein.[47,48] The QUTA activator protein contains a putative zinc binuclear cluster motif in its N-terminal domain that facilitates binding to the appropriate motif in the promoters of the quinate pathway genes.[48,49] This elaborate domain combination structure gives a structural basis for controlling the expression of genes encoding metabolic enzymes *via* the signal transduction pathway. Although the proposed evolutionary relationship between QUTA/QUTR and the protein AROM has been challenged recently by Nicholas *et al.*,[50] we can still view the QUTA or QUTR protein as an example of the composite protein that might be fused by different single-domain structures. It is anticipated that further studies on other composite proteins that are the result of fusion of a catalytic domain and a DNA-binding domain may help to elucidate the molecular mechanisms coupling the metabolic and regulatory pathways. Given that proteins at the superfamily level often show clear structural and functional homology, the domain fusion map presented here will provide useful information for studies on this aspect.

Implications for protein engineering and protein evolution

The pace at which humans modify this world will never slow. Lots of efforts have been made to create new hybrid enzymes[51,52] or modular enzymes[53–55] with multiple functions or higher catalytic efficiency by making fusions of protein modules. However, in some cases, although the individual domains or subunits can be recombined in chimeric enzymes by genetic engineering, the resulting chimeras may lose overall stability and even activity, possibly due to a structural mismatch at the fusion interface. The compatibility of the fused domains or modules will be crucial to any attempt at constructing multienzymes. The domain fusion map in our work may offer an attractive framework for designing chimeric enzymes following Nature's lead, because component proteins have co-existed stably in the natural fused proteins possibly due to some functional advantages, such as co-regulation, co-localization, substrate channelling, etc.

Understanding Nature's strategies and mechanisms for protein evolution may provide insights into the rational design of proteins with novel biological functions. Gene fusion has played an important role in the evolutionary history of contemporary proteins. The possibility of the evolution of complex folds from antecedent domain segments has been suggested.[56] In addition, structural and sequence data for some enzymes with the triosephosphate isomerase (TIM)-barrel fold or the β/α_8-barrel (SCOP identifier c.1), the most common enzyme fold,[57] strongly suggested that the β/α_8-barrel may have evolved from an ancestral β/α_4-half barrel *via* gene duplication and subsequent fusion.[58,59] As shown in Figure 1, 11 superfamilies with the β/α_8-barrel fold are involved in the fusion events, which suggests that this type of fold is still very active in the fusion behavior at a higher level that between domain and domain. The fusion map given here, which delineates evolutionary relationships among superfamilies, shows an evolutionary scenario that might have led to the evolution of some multiple-domain proteins from a series of domain fusion events.

Supplementary material

Supplementary material including the 10,073 fusion events within the 30 genomes, the structural annotation of the proteins involved in the fusion events, the 676 assigned single domain fusion pairs and a Java applet showing the domain fusion map is available at http://www.bioinfo.tsinghua.edu.cn/fusion/supplementary.htm

Acknowledgments

We thank Itai Yanai at Boston University for help with the data preparation. This work was funded by the National Natural Science Grant in China (no. 39980007) and the National Key Foundational Research Grant in China (985, 863).

References

1. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
2. Marcotte, E. M., Pellegrini, M., Thompson, M., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
3. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
4. Enright, A. J. & Ouzounis, C. A. (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.* **2**, 341–347.
5. Yanai, I., Derti, A. & DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl Acad. Sci. USA*, **98**, 7940–7945.
6. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
7. Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.
8. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E. *et al.* (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, **392**, 353–358.
9. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J. *et al.* (2000). Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.
10. Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R. *et al.* (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.
11. Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J. *et al.* (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.
12. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
13. Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R. *et al.* (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature*, **406**, 151–157.
14. Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A. *et al.* (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364–370.

15. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D. *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
16. Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S. *et al.* (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76.
17. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
18. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G. *et al.* (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.
19. Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D. *et al.* (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.
20. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H. *et al.* (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.
21. Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A. *et al.* (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**, 1809–1815.
22. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucl. Acids Res.* **24**, 4420–4449.
23. Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R. *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, **281**, 375–388.
24. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K. *et al.* (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6**, 145–152.
25. Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R. W. *et al.* (1999). Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genet.* **21**, 385–389.
26. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M. *et al.* (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.
27. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D. *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.
28. Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C. *et al.* (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.
29. White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J. *et al.* (1999). Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, **286**, 1571–1577.
30. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V. *et al.* (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
31. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y. *et al.* (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136.
32. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T. *et al.* (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155.
33. Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L. *et al.* (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, **282**, 754–759.
34. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D. *et al.* (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
35. Glass, J. I., Lefkowitz, E. J., Glass, J. S., Heiner, C. R., Chen, E. Y. & Cassell, G. H. (2000). The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature*, **407**, 757–762.
36. Cherry, J. M., Adler, C., Ball, C., Cheryitz, S. A., Dwight, S. S., Hester, E. T. *et al.* (1998). SGD: Saccharomyces Genome Database. *Nucl. Acids Res.* **26**, 73–79.
37. Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
38. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications in protein modelling. *J. Mol. Biol.* **235**, 1501–1531.
39. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919.
40. Karplus, K., Barret, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
41. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325.
42. Gough, J. & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucl. Acids Res.* **30**, 268–272.
43. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
44. Tsoka, S. & Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genet.* **26**, 141–142.
45. Hawkins, A. R., Lamb, H. K., Moore, J. D. & Roberts, C. F. (1993). Genesis of eukaryotic transcriptional activator and repressor proteins by splitting a multidomain anabolic enzyme. *Gene*, **136**, 49–54.
46. Lamb, H. K., Moore, J. D., Lakey, J. H., Levett, L. J., Wheeler, K. A., Lago, H. *et al.* (1996). Comparative analysis of the QUTR transcription repressor protein

and the three C-terminal domains of the pentafunctional AROM enzyme. *Biochem. J.* **313**, 941–950.

47. Lamb, H. K., Giles, H. N., Levett, L. J., Cairns, E., Roberts, C. F. & Hawkins, A. R. (1996). The QUTA activator and QUTR repressor proteins of *Aspergillus nidulans* interact to regulate transcription of the quinate utilization genes. *Microbiology,* **142**, 1477–1490.
48. Levett, L. J., Si-Hoe, S. M., Liddle, S., Wheeler, K., Smith, D., Lamb, H. K. *et al.* (2000). Identification of domains responsible for signal recognition and transduction within the QUTR transcription repressor protein. *Biochem. J.* **350**, 189–197.
49. Levesley, I., Newton, G. H., Lamb, H. K., van Schothorst, E., Dalgleish, R. W., Samson, A. C. *et al.* (1996). Domain structure and function within the QUTA proteins of *Aspergillus nidulans*: implications for the control of transcription. *Microbiology,* **142**, 87–98.
50. Nicholas, H. B., Arst, H. N. & Caddick, M. X. (2001). Evaluating low level sequence identities. Are *Aspergillus* QUTA and AROM homologous? *Eur. J. Biochem.* **268**, 414–419.
51. Kim, Y. G., Cha, J. & Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to *Fok* I cleavage domain. *Proc. Natl Acad. Sci. USA*, **93**, 1156–1160.
52. Nixon, A. E., Warren, M. S. & Benkovic, S. J. (1997). Assembly of an active enzyme by the linkage of two protein modules. *Proc. Natl Acad. Sci. USA*, **94**, 1069–1073.
53. Khosla, C. & Harbury, P. B. (2001). Modular enzymes. *Nature*, **409**, 247–252.
54. Mootz, H. D., Schwarzer, D. & Marahiel, M. A. (2000). Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl Acad. Sci. USA*, **97**, 5848–5853.
55. Weber, T. & Marahiel, M. A. (2001). Exploring the domain structure of modular nonribosomal peptide synthetases. *Structure*, **9**, 3–9.
56. Lupas, A. N., Ponting, C. P. & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203.
57. Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters*, **492**, 193–198.
58. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1549.
59. Hocker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. & Sterner, R. (2001). Dissection of a β/α_8-barrel enzyme into two folded halves. *Nature Struct. Biol.* **8**, 32–36.
60. Carpenter, E. P., Hawkins, A. R., Frost, J. W. & Brown, K. A. (1998). Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis. *Nature*, **394**, 299–301.
61. Stallings, W. C., Abdel-Meguid, S. S., Lim, L. W., Shieh, H. S., Dayringer, H. E., Leimgruber, N. K. *et al.* (1991). Structure and topological symmetry of the glyphosate target 5-enolpyruvylshikimate-3-phosphate synthase: a distinctive protein fold. *Proc. Natl Acad. Sci. USA*, **88**, 5046–5050.
62. Krel, T., Coggins, J. R. & Lapthorn, A. J. (1998). The three-dimensional structure of shikimate kinase. *J. Mol. Biol.* **278**, 983–997.
63. Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R. *et al.* (1999). The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction. *Nature Struct. Biol.* **6**, 521–525.

Edited by B. Holland

(Received 15 November 2001; received in revised form 10 April 2002; accepted 10 May 2002)