BIOINFORMATICS

Vol. 20 no. 8 2004, pages 1248–1253
DOI: 10.1093/bioinformatics/bth075

Poorly conserved ORFs in the genome of the archaea Halobacterium sp. NRC-1 correspond to expressed proteins

H. Shmuely[1], E. Dinitz[1], I. Dahan[1], J. Eichler[1], D. Fischer[2] and B. Shaanan[1,*]

[1]Department of Life Sciences and [2]Department of Bioinformatics and Computer Science, Ben Gurion University of the Negev, PO Box 653, Beersheva 84015, Israel

Received on July 30, 2003; revised on October 21, 2003; accepted on October 22, 2003
Advance Access publication February 10, 2004

ABSTRACT

Motivation: A large fraction of open reading frames (ORFs) identified as 'hypothetical' proteins correspond to either 'conserved hypothetical' proteins, representing sequences homologous to ORFs of unknown function from other organisms, or to hypothetical proteins lacking any significant sequence similarity to other ORFs in the databases. Elucidating the functions and three-dimensional structures of such orphan ORFs, termed ORFans or poorly conserved ORFs (PCOs), is essential for understanding biodiversity. However, it has been claimed that many ORFans may not encode for expressed proteins.

Results: A genome-wide experimental study of 'paralogous PCOs' in the halophilic archaea *Halobacterium* sp. NRC-1 was conducted. Paralogous PCOs are ORFs with at least one homolog in the same organism, but with no clear homologs in other organisms. The results reveal that mRNA is synthesized for a majority of the *Halobacterium* sp. NRC-1 paralogous PCO families, including those comprising relatively short proteins, strongly suggesting that these *Halobacterium* sp. NRC-1 paralogous PCOs correspond to true, expressed proteins. Hence, further computational and experimental studies aimed at characterizing PCOs in this and other organisms are merited. Such efforts could shed light on PCOs' functions and origins, thereby serving to elucidate the vast diversity observed in the genetic material.

Contact: bshaanan@bgumail.bgu.ac.il

INTRODUCTION

As the number of completed genome sequences grows at an increasingly rapid pace, it has become clear that one of the most pressing challenges of the post-genomic era is the interpretation of the vast amounts of data generated. One of the first steps in such endeavours is the description of the protein complement of a given organism, i.e. the proteome. Identifying the set of proteins encoded by a genome is not a trivial process. Even in prokaryotes, where the ratio of coding to non-coding DNA is relatively large and introns are few in number, cataloguing the proteome is not straightforward. For instance, the presence of an open reading frame (ORF), delineated by a start codon followed by a number of non-termination codons, does not necessary imply the existence of an encoded (let alone expressed or functional) protein. Some ORFs may correspond to pseudogenes, while others may correspond to spurious regions of DNA that do not code for proteins.

Given the lack of experimental validation for the vast majority of ORFs, prediction of a true protein can be made if significant homology to proteins identified in other genomes is found. However, most newly sequenced genomes contain a large fraction (20–30%) of ORFs for which no clear homology to protein-encoding genes identified in other sequenced genomes can be detected (Fischer and Eisenberg, 1999). We refer to these orphan ORFs as ORFans (Fischer and Eisenberg, 1999) or poorly conserved ORFs (PCOs; Siew and Fischer, 2003b).

The origins of ORFans and their biological roles remain unknown. Some ORFans may correspond to newly evolved proteins or to unique descendants of ancient proteins, with unique functions and three-dimensional (3D) structures not currently observed in other families (Fischer and Eisenberg, 1999; Vitkup *et al*., 2001; Coulson and Moult, 2002; Lee *et al*., 2003). Other ORFans may correspond to highly divergent members of known protein families, but with functions and/or 3D structures similar to proteins already known (Fischer and Eisenberg, 1999; Wood *et al*., 2001). Regardless of their origin, standard computational methods, which infer function or 3D structure of a protein on the basis of its sequence similarity to a known family, are not suitable for the study of ORFans, although recently developed non-homology-based methods (Marcotte, 2000) may eventually provide some clues about ORFans. Thus, at the present time, the function of each ORFan will need to be determined by genetic or biochemical approaches (Dujon, 1996; Oliver, 1996; Alimi *et al*., 2000).

*To whom correspondence should be addressed.

The presence of so many ORFans suggests that sequence diversity in nature may be greater than previously expected (Fraser *et al*., 2000; Boucher *et al*., 2001). However, because little can be learned about ORFans via homology, each ORFan represents a secret awaiting interpretation (Dujon, 1996; Fischer and Eisenberg, 1999; Bloom, 2000; Doolittle, 2002). If proteins in different organisms have descended from common ancestral proteins by duplication and adaptive variation, why is it that, today, so few show similarity to each other (Doolitle, 1997)? Why is it that the necessary 'intermediate sequences' that must have given rise to these ORFans are not found today? Do most ORFans correspond to rapidly diverging proteins (Schmid and Aquadro, 2001; Wood *et al*., 2001)? If so, how rapidly do they diverge, and what are the forces involved in their rapid evolution? Is their rate of change constant or did the rapid changes occur only at specific times? Do these rapidly evolving ORFans correspond to non-essential proteins or to the species-determinants (Siew and Fischer, 2003a)? In summary, ORFans are a mystery that needs solving.

Despite the apparent importance of genomic ORFans, it seems that over the last years they have been underemphasized (Fischer, 1999; Siew and Fischer, 2003a,c). The recent establishment of a number of structural genomics initiatives, designed to determine the 3D structures of a well-selected representative set of proteins, should, in theory, somewhat correct this situation. However, the vast majority of such projects have chosen to focus on determining the structures of conserved ORFs, i.e. those ORFs encoding for members of relatively large homologous protein families. Hence, focused efforts directed at specifically elucidating the structures and functions of PCOs are required.

A first step in such efforts requires the confirmation that PCOs indeed correspond to true, expressed proteins and not to errors or mis-annotated genes (e.g. Dujon, 1996; Fischer and Eisenberg, 1999; Malpertuy *et al*., 2000; Schmid and Aquadro, 2001; Skovgaard *et al*., 2001; Wood *et al*., 2001; Kunin *et al*., 2003). Especially dubious are the relatively shorter ORFs (with less than 100–150 codons), which have been referred to as smORFs (for 'small ORFs'; Basrai *et al*., 1997) or ELFs (for 'evil little fellows'; Ochman, 2002). smORFs are problematic because without evidence from homology to other ORFs, they are more likely to correspond to spurious, non-coding ORFs (Andrade *et al*., 1997; Basrai *et al*., 1997; Skovgaard *et al*., 2001; Wood *et al*., 2001; Mira *et al*., 2002; Ochman, 2002).

Although a number of studies suggest that ORFans indeed correspond to expressed proteins (e.g. Malpertuy *et al*., 2000; Alimi *et al*., 2000; Siew and Fischer, 2003b,c), to the best of our knowledge, no experimental genome-wide studies of ORFans and PCOs have been carried out. Accordingly, we now report the results of such investigation and provide the first experimental evidence that a large proportion of paralogous PCOs encoded by the halophilic archaea *Halobacterium* sp. NRC-1 do indeed code for expressed proteins.

MATERIALS AND METHODS

Target selection

For this study, we have chosen to examine the genome of the halophilic archaea *Halobacterium* sp. NRC-1 (Ng *et al*., 2000). As a first step in our genome-wide study of PCOs, we focused only on the subset of ORFs defined by us as 'paralogous PCOs'. Paralogous PCOs correspond to ORFs that have at least one homolog within *Halobacterium* sp. NRC-1, but no detectable homologs in other organisms. It should be noted that the lack of homologs in other organisms might only reflect the sensitivity of the sequence-comparison method applied to detect homologs, the threshold used and the current content of sequence databases. As such, it is plausible that with more sensitive methods, different thresholds, or with the addition of new sequences to the databases, some homologs may be found for our PCOs. Nevertheless, the fact that our PCOs do not presently have any clear homologs in other organisms means that they are poorly conserved. Thus, even if homologs are found in the future, they will remain PCOs, lacking close homologs in many organisms. Accordingly, to emphasize this fact, we use the term 'paralogous PCOs' rather than 'paralogous ORFans'.

Selection of paralogous ORFs was achieved by running BLAST for each ORF predicted in the large chromosome of *Halobacterium* sp. NRC-1 against the haloarchaeal genome itself. If a given ORF was found to have a homolog with an E-value less than 10^{-5}, then it was defined as belonging to a paralogous family. We then concentrated only on those paralogous families containing at least two members found on the large chromosome having no putative annotation (i.e. ORFs annotated as 'hypothetical' or 'conserved hypothetical'). In addition, when a sequence identical to an existing member of a paralogous group was detected, that member was eliminated from the family. To confirm that no sequences homologous to our selections exist in other organisms, each selected ORF belonging to the paralogous families was screened against the nr database at NIH (http://www.ncbi.nlm.nih.gov:80/BLAST) beginning in October 2001 and reviewed monthly until July 2003. Those ORFs for which no homolog could be detected in other organisms (at the same E-value threshold) were selected as paralogous PCOs. Other constraints employed in these searches included exclusion of any predicted protein sequences shorter than 75 amino acid residues in length, as well as proteins predicted to be membrane-associated according to the program MOMENT (Eisenberg *et al*., 1984).

Strains and growth conditions

Halobacterium sp. NRC-1 was obtained from Aharon Oren (Hebrew University of Jerusalem) and grown as previously

described (DasSarma and Fleishmann, 1995). Cells were harvested at the mid-log growth phase ($OD_{550} = 0.5–0.8$) and processed for total RNA extraction.

RNA extraction

RNA isolation was carried out according to protocols described in the HaloHandbook for Halobacterial Genetics, version 4.5 (http://www.microbiol.unimelb.edu.au/staff/mds/HaloHandbook/HaloHb.web/Halohandbook.pdf). Contaminating DNA was eliminated with a DNAFree kit (Ambion RNA). RNA concentration was quantified spectrophotometrically.

Reverse transcriptase–polymerase chain reaction (RT–PCR)

To perform RT–PCR, forward and reverse oligonucleotide primers were designed for each PCO, employing the *Halobacterium* sp. NRC-1 genomic sequence as a guide. Primers, composed to introduce appropriate restriction enzyme sites, were designed so as to overlap the start and stop regions of each ORF and the regions immediately upstream and downstream, respectively. Single-stranded cDNA was prepared for each PCO sequence from the corresponding RNA and reverse primer using the EZ-First Strand cDNA Synthesis kit for RT–PCR (Biological Industries, Israel). The single-stranded cDNA was then used as a template in a PCR reaction containing the appropriate forward and reverse primers. cDNA amplification was confirmed by electrophoresis in 1% agarose gels. The sequence of the PCR product was determined to confirm its identity. In control experiments designed to exclude any contribution from contaminating DNA, PCR amplification was performed on total RNA without prior cDNA preparation.

RESULTS

Archaea, the third and most-recently described form of life on Earth, are best known in their capacity as extremophiles, able to survive amongst the most challenging physical environmental challenges on the planet (Woese *et al*., 1990). What is presently known of archaeal biology suggests that in many systems, a mosaic of eukaryal, bacterial and archaeal traits is found. Hence, studying Archaea at the molecular level provides information not only on this unique group of microorganisms, but also on the other two Domains of Life, i.e. Eukarya and Bacteria. Moreover, the often unique stabilities of extermophilic archaeal proteins makes them well-suited for experimental studies. In the case of *Halobacterium* sp. NRC-1, the species possesses a relatively small genome with 2605 predicted ORFs distributed between a large chromosome and two smaller chromosomes, and is thereby amenable to genome-wide experimental analysis.

As the focus of our investigation was the identification of paralogous PCOs, our analysis divided the 2075 predicted ORFs on the large chromosome along the lines of paralogous groups. Thus, analysis of the annotated *Halobacterium* sp. NRC-1 genome revealed the existence of 846 paralogous ORFs that can be grouped into 284 families (see Materials and methods section). Only those paralogous families which met the following criteria were subsequently included in our study: (i) ORFs longer than 74 amino acid residues; (ii) ORFs lacking transmembrane regions and (iii) ORFs found on the large chromosome. After filtering for criteria (i)–(iii), 611 paralogues remained. Next, a BLAST filtering step was performed to exclude those paralogous groups containing at least one member possessing a homolog in other organisms with an E-value of less than 10^{-5}. Following this step, an additional 572 sequences were eliminated from our survey. In those cases when a candidate paralogous ORF did not meet our criteria, the fate of its paralogous PCO family depended on the reason that member was excluded. If the ORF possessed a homolog in another organism, the whole group was excluded from our study. If, however, the candidate ORF did not possess a homolog in another organism yet still failed to meet our other selection criteria, then the family was retained in our study, only without the ORF in question. Following this screening procedure, we were left with 14 paralogous families, containing a total of 39 paralogous PCOs. Table 1 presents the final list of 39 paralogous PCOs. The largest paralogous families contain four ORFs (family numbers 2, 4 and 14). The average length (in amino acid residues) of these 39 ORFs is 198, with 18 ORFs shorter than 150 residues and seven longer than 300.

Having chosen a set of targets, experiments were designed to confirm that mRNA derived from these genes was indeed synthesized. Transcription of mRNA encoded by a given PCO gene would offer strong support that a particular gene indeed encodes for a true protein. Thus, to detect the presence of PCO-derived mRNAs, RT–PCR was performed. The RNA content of *Halobacterium* sp. NRC-1 cells was first isolated. To digest any extraneous DNA co-captured, the samples were treated with DNase. The RNA was then used to guide the synthesis of single-stranded cDNA, using reverse transcriptase. This cDNA subsequently served as the template for PCR reactions, using primers directed to each PCO gene sequence. In control experiments, PCR was performed on RNA samples not exposed to reverse transcriptase. In this way, any resulting PCR product would appear due to the presence of residual contaminating DNA. In Figure 1, a representative example of such an experiment is shown. Finally, the obtained PCR product was subjected to sequence analysis to confirm its identity.

The results of the RT–PCR studies revealed that mRNA corresponding to 30 of the 39 PCOs was detected, corresponding to members of 13 of the 14 PCO paralogous families. No expressed mRNA was detected for PCOs of family number 10, which contains two relatively long PCOs. Expressed mRNA was found for at least one member of the other 13 PCO families, with six families expressing mRNA for all of the family members.

Table 1. Paralogous PCOs in *Halobacterium* sp. NRC-1

Paralogous PCO family	ORF name	Predicted length[a]	Archaeal expression[b]
1	Vng 0609c	215	+
	Vng 0698h	221	+
2	Vng 0053h	151	+
	Vng 1056h	321	+
	Vng 1063c	338	+
	Vng 1948h	304	−
3	Vng0258h	116	+
	Vng0751c	104	+
	Vng1426h	92	+
4	Vng1012h	90	+
	Vng1546h	79	+
	Vng2115h	81	+
	Vng2310h	96	+
5	Vng1182h	162	+
	Vng1487h	152	+
	Vng1621h	151	+
6	Vng1413h	529	+
	Vng1533h	714	+
	Vng2566h	705	−
7	Vng1096h	102	+
	Vng1490h	132	+
	Vng2414h	163	+
8	Vng1734h	111	−
	Vng1758h	85	+
9	Vng0319h	80	+
	Vng0768h	74	+
	Vng2014h	83	+
10	Vng0430h	216	−
	Vng2059h	268	−
11	Vng0725h	302	+
	Vng2230h	276	−
12	Vng0441h	211	−
	Vng2392h	210	+
13	Vng0591c	214	+
	Vng2445c	210	−
14	Vng0293h	89	+
	Vng0511h	76	+
	Vng0703h	94	−
	Vng2614h	90	+

[a]Number of amino acid residues.
[b]As revealed by RT–PCR; (+), expressed; (−), not expressed.

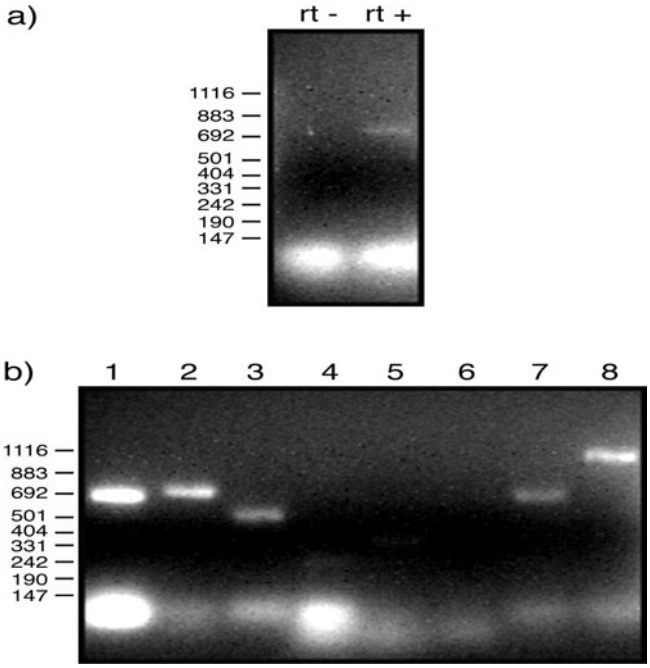

Fig. 1. Analysis of RT–PCR. To confirm the expression of paralogous *Halobacterium* sp. NRC-1 PCO genes, 0.3 μg of purified RNA was used for single-stranded cDNA synthesis in a reverse transcriptase reaction, as described in Materials and methods section. Five per cent of the reactions were then used as a template for PCR amplification, using the appropriate primers. (**a**) To verify that the purified RNA was DNA-free, PCR reactions were perfomed using total RNA as template either prior to (rt−) or following (rt+) the reverse transcriptase reaction. The results reveal that PCR amplification was only achieved in the presence of single-stranded cDNA. (**b**) Shown is a representative result obtained when the following selected PCO DNA fragments were employed as templates for PCR amplifications: Vng0609c (lane 1), Vng0698h (lane 2), Vng1182h (lane 3), Vng0430h (lane 4), Vng2059h (lane 5), Vng0441h (lane 6), Vng2392h (lane 7) and Vng1056c (lane 8). For Vng0430h, Vng2059h and Vng0441h (lanes 4–6), no expression was detected. In the other lanes, the PCR-generated fragments are of the predicted size. The lower band in each lane corresponds to the primers used in the PCR reaction. In both (a) and (b), the positions of molecular weight markers are depicted on the right.

DISCUSSION AND CONCLUSIONS

We have performed a genome-wide study of 14 paralogous PCO families of *Halobacterium* sp. NRC-1, encompassing a total of 39 ORFs. Since the mere appearance of paralogous genes within a genome as proof of the expression of protein is far from resolved (cf. Lynch and Conery, 2000; Gu *et al*., 2003), we sought to provide direct evidence that these paralogous ORFs indeed encode for true proteins. Through RT–PCR, we have identified mRNA for 30 out of the 39 paralogous PCOs. These 30 PCOs cover 13 of the 14 PCO's paralogous families identified. Hence, the presence of mRNA for these paralogous PCOs strongly suggests that they indeed direct the synthesis of true, expressed proteins.

What is the significance of the failure to detect mRNA for nine PCOs? It is possible that the conditions employed in the study are incompatible with efficient transcription of these PCOs. For instance, it is conceivable that expression of these gene products only occurs under a particular set of conditions, such as elevated temperature, reduced salt conditions or oxygen stress. However, it should be noted that in most instances where no expression was detected, mRNA encoded by other members of that family of paralogs was detected. This suggests that individual family members

may be expressed differentially. Indeed, our failure to detect expression. An alternate explanation for our failure to detect mRNA expression in these cases could have resulted from the design of primers for the detection of expression directed against a wrongly predicted start site. However, examination of the sequences of the nine non-expressed PCOs reveals that except for three cases, this is not possible. Of the remaining six sequences, two contain no additional methionine residues, while in four cases any additional methionine residues are situated too close to the C-terminus to serve as a start codon for the paralog. Another interesting observation is that only two (out of the nine) targets for which we did not find evidence of mRNA expression correspond to proteins shorter than 150 residues. For the other 16 targets shorter than 150 residues, expression evidence was found. This suggests that, in our dataset, there is no indication that shorter PCOs are less likely to be expressed.

It is important to emphasize that our definition of PCOs means only that no clear homologs were found in the sequence databases at the time of the BLAST search. It is likely that as more sequences are added to the databases, new homologs will appear for our PCOs. However, the fact that a given PCO has no homologs at present means it will remain a poorly conserved protein even if in the future a new homolog is added into the database. Indeed, by re-running BLAST near the time of submission (July 2003), we found that a homolog for the ORFs grouped into our paralogous PCO family number 3 has now been detected. This corresponds to a newly deposited 'unknown' sequence from the haloarchaeal phage PhiCh1 (Klein *et al*., 2002). This raises the hypothesis that the origin of this homology may be horizontal gene transfer (Jain *et al*., 1999). Nevertheless, the fact that homologs of unknown function are now found for these PCOs does not help in characterizing them. Thus, PCOs continue to remain poorly conserved ORFs, whose functions need to be unraveled.

In this first study, we have only focused on paralogous PCOs. Clearly, paralogs are more likely to correspond to expressed proteins than singleton PCOs. However, given their dissimilarities to any other ORF in the database, proof that PCOs, be they singleton or members of paralogous families, actually encode proteins is a necessary step before embarking on a labor-intensive structure/function project. Having now established that members of our paralogous PCOs families can direct the transcription of mRNA and hence likely correspond to expressed proteins, we have initiated further computational and experimental studies aimed at elucidating the functions and 3D structures of the encoded proteins. Preliminary results from computational studies (which will be presented elsewhere), including more sensitive sequence searches and the application of fold-recognition methods (Bujnicki *et al*., 2001), suggest that eight of our paralogous PCOs families may correspond to distant members of known families, three of them corresponding to various transcriptional regulator proteins. Experimental structural and functional determination will allow us to verify these results and to determine how many of these and other PCOs correspond to novel and unique proteins.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Miriam Cohen, former Dean of the Faculty of Natural Sciences, Ben Gurion University of the Negev, for her support and encouragement.

REFERENCES

Alimi,J.P., Poirot,O., Lopez,F. and 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.

Andrade,M.A., Daruvar,A., Casari,G., Schneider,R., Termier,M. and Sander,C. (1997) Characterization of new proteins found by analysis of short open reading frames from the full yeast genome. *Yeast*, **13**, 1363–1374.

Basrai,M.A., Hieter,P. and Boeke,J.D. (1997) Small open reading frames: beautiful needles in the hay stack. *Genome Res.*, **7**, 768–771.

Bloom,B.R. (2000) On the particularity of pathogens. *Nature*, **406**, 760–761.

Boucher,Y., Nesbo,C.L. and Doolittle,W.F. (2001) Microbial genomes: dealing with diversity. *Curr. Opin. Microbiol.*, **4**, 285–289.

Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.

Coulson,A.F. and Moult,J. (2002) A unifold, mesofold, and superfold model of protein fold use. *Proteins*, **46**, 61–71.

DasSarma,S. and Fleishmann,E.M. (ed.) (1995) Archaea: a laboratory manual—Halophiles. Cold Spring Harbor Laboratory Press, pp. 155–160.

Doolittle,R.F. (1997) A bug with excess gastric avidity. *Nature*, **388**, 515–516.

Doolittle,R.F. (2002) Biodiversity: microbial genomes multiply. *Nature*, **416**, 697–700.

Dujon,B. (1996) The yeast genome project: what did we learn? *Trends Genet.*, **12**, 263–270.

Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.

Fischer,D. (1999) Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge. *Protein Eng.*, **12**, 1029–1030.

Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.

Fraser,C.M., Eisen,J.A. and Salzberg,S.L. (2000) Microbial genome sequencing. *Nature*, **406**, 799–803.

Gu,Z., Steinmetz,L.M., Gu,X., Scharfe,C., Davis,R.W. and Li,W.H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**, 63–66.

Jain,R., Rivera,M.C. and Lake,J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci., USA*, **96**, 3801–3806.

Klein,R., Baranyi,U., Rossler,N., Greineder,B., Scholz,H. and Witte,A. (2002) *Natrialba magadii* virus phiCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. *Mol. Microbiol.*, **45**, 851–863.

Kunin,V., Cases,I., Enright,A.J., de Lorenzo,V. and Ouzounis,C.A. (2003) Myriads of protein families, and still counting. *Genome Biol.*, **4**, 401.

Lee,D., Grant,A., Buchan,D. and Orengo,C. (2003) A structural perspective on genome evolution. *Curr. Opin. Struct. Biol.*, **13**, 359–369.

Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Malpertuy,A., Tekaia,F., Casaregola,S., Aigle,M., Artiguenave,F., Blandin,G., Bolotin-Fukuhara,M., Bon,E., Brottier,P., de Montigny,J. *et al*. (2000) Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. *FEBS Lett.*, **487**, 113–121.

Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.

Mira,A., Klasson,L. and Andersson,S.G. (2002) Microbial genome evolution: sources of variability. *Curr. Opin. Microbiol.*, **5**, 506–512.

Ng,W.V., Kennedy,S.P., Mahairas,G.G., Berquistm,B., Panm,M., Shuklam,H.D., Laskym,S.R., Baliga,N.S., Thorsson,V., Sbrogna,J. *et al*. (2000) Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl Acad. Sci., USA*, **97**, 12176–12181.

Ochman,H. (2002) Distinguishing the orfs from the elfs: short bacterial genes and the annotation of genomes. *Trends Genet.*, **18**, 335–337.

Oliver,S.G. (1996) From DNA sequence to biological function. *Nature*, **379**, 597–600.

Schmid,K.J. and Aquadro,C.F. (2001) The evolutionary analysis of 'orphans' from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics*, **159**, 589–598.

Siew,N. and Fischer,D. (2003a) Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins*, **53**, 241–251.

Siew,N. and Fischer,D. (2003b) Twenty thousand ORFan microbial protein families for the Biologist? *Structure*, **11**, 7–9.

Siew,N. and Fischer,D. (2003c) Unravelling the ORFan puzzle. *Comp. Funct. Genomics*, **4**, 432–441.

Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D. and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.*, **17**, 425–428.

Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.*, **8**, 559–566.

Woese,C.R., Kandler,O. and Wheelis,M.L. (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria and eucarya. *Proc. Natl Acad. Sci., USA*, **87**, 4576–4579.

Wood,V., Rutherford,K.M., Ivens,A., Rajandream,M.A. and Barrell,B. (2001) A Re-annotation of the *Saccharomyces cerevisiae* Genome. *Comp. Funct. Genomics*, **2**, 143–154.