

MEETING REPORT

releasing Sir proteins from the Ku70p–Ku80p telomerase complex (David Shore, Univ. of Geneva, Switzerland). Cdc13p protein binds single-stranded DNA at the Ku70p–Ku80p telomerase complex (Vicki Lundblad, Baylor, USA). Nuclear organization of telomeres is important with telomeres located at the nuclear periphery (Susan Gasser, ISREC, Switzerland). Targeting DNA to the periphery using a ER–Golgi anchoring signal can produce silencing (Rolf Sternglanz, SUNY, USA). Hence, any gene brought to the nuclear periphery will

be silenced by the Sir protein complex. In summary, the importance of chromatin structure was evident in all sessions. Yeast origins, centromeres and telomeres bind elegant multiprotein complexes that act as regulatory machines to change chromatin structure and to allow important cellular processes to occur.

Further reading

- 1 Dutta, A. and Bell, S.P. (1997) *Annu. Rev. Cell Dev. Biol.* 13, 293–332
- 2 Pluta, A.F. *et al.* (1995) *Science* 270, 1591–1594

- 3 Loo, S. and Rine, J. (1995) *Annu. Rev. Cell Dev. Biol.* 11, 519–548
- 4 Smith, J.S. and Boeke, J.D. (1997) *Genes Dev.* 11, 241–254
- 5 Weaver, D.T. (1995) *Trends Genet.* 11, 388–392

Robert A. Sclafani

robert.sclafani@uchsc.edu

Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, 4200 E. Ninth Avenue, Denver, CO 80262, USA.

LETTER

Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles

Sequencing of multiple complete genomes of bacteria and archaea makes it possible to perform systematic, genome-scale comparisons that aim to delineate the genomic complement of a particular phenotype. Recently, the first genome of a hyperthermophilic bacterium, *Aquifex aeolicus*, has been sequenced¹. Previous studies based on rRNA and aminoacyl-tRNA analysis had suggested a very early divergence of *Aquifex* from the rest of the bacteria^{2,3}. *Aquifex* is

exceptional among bacteria in that it occupies the hyperthermophilic niche otherwise dominated by archaea². In the published analysis of the *Aquifex* genome, it has been concluded that the genome sequence yielded 'only a few specific indications of thermophily'¹. With three genomes of extreme thermophilic archaea (*Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus*) currently available^{4–6}, we

reasoned that a detailed comparison of the *Aquifex* and archaeal genomes could reveal genome-scale adaptations for thermophily.

The protein sequences encoded in all complete bacterial genomes were compared with the non-redundant protein sequence database using the gapped BLAST program⁷, and a phylogenetic breakdown was automatically produced using the TAX_COLLECTOR program (Ref. 8, and D.R. Walker, unpublished). The results show that the fraction of *Aquifex* gene products that have archaeal proteins as clear best hits is by far greater than for each of the other bacteria (Table 1). Taking the fraction of 'archaeal' genes in *Bacillus subtilis* (Table 1) as a conservative estimate for the random expectation in a bacterial genome and using the normal approximation of the binomial distribution, it could be estimated that the excess of 'archaeal' genes in *Aquifex* could not be explained by a random fluctuation, with $p < < 10^{-10}$. A reciprocal comparison showed that, for proteins encoded in each of the three archaeal genomes, *Aquifex* proteins are the best hits significantly more frequently than proteins from other bacteria, even those with genomes 2–3 times larger than the *Aquifex* genome, such as *Synechocystis* sp. or *B. subtilis* (Table 2). In a complementary analysis, bacterial proteins were compared with

TABLE 1. 'Archaeal' genes in bacterial genomes

Bacterial species ^a	Reliable best hits to archaeal proteins ^b
<i>Aquifex aeolicus</i>	246 (16.2%)
<i>Bacillus subtilis</i>	207 (5.0%)
<i>Synechocystis</i> sp.	126 (4.0%)
<i>Borrelia burgdorferi</i>	45 (3.6%)
<i>Escherichia coli</i>	99 (2.3%)

^aThe data on *Haemophilus influenzae*, *Helicobacter pylori* (Proteobacteria), *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (Gram-positive bacteria) are not shown because, in these species, the majority of the best hits are to homologs from larger genomes within the same phylogenetic lineages, namely *E. coli* and *B. subtilis*, respectively.

^bAll database hits with associated expectation (e) values $< 10^{-3}$ were analyzed; a 'reliable best hit' was registered when the e-value with an archaeal protein was lower than that with any bacterial or eukaryotic protein by at least a factor of 100.

LETTER

TABLE 2. 'Bacterial' proteins in archaea

Archaeal species	Reliable best hits in bacteria ^a				
	Aa	Bs	Ssp	Ec	Bb
<i>Methanococcus jannaschii</i>	193 (10.9%)	78 (4.4%)	56 (3.2%)	44 (2.5%)	16 (0.9%)
<i>Methanobacterium autotrophicum</i>	151 (8.0%)	103 (5.4%)	91 (4.8%)	41 (2.2%)	13 (0.7%)
<i>Archaeoglobus fulgidus</i>	227 (9.4%)	140 (5.8%)	80 (3.9%)	59 (2.5%)	16 (0.7%)

^aDefined as in Table 1. The bacterial species included are the same as in Table 1; abbreviations: Aa, *Aquifex aeolicus*; Bb, *Bacillus burgdorferi*; Bs, *Bacillus subtilis*; Ec, *Escherichia coli*; Ssp, *Synechocystis* sp.

protein families that are conserved in all three sequenced archaeal genomes (Ref. 9 and K. Makarova, L. Aravind, R.L. Tatusov and E.V. Koonin, unpublished). The fraction of bacterial proteins that could be included into the conserved archaeal families was essentially uniform at the level of about 20% of each of the bacterial proteomes, with a sharp deviation at 39% observed for *Aquifex* (Table 3).

Given these indications of a direct relationship between a sizeable fraction of genes in *Aquifex* and archaea, we investigated the protein families that they share in further detail using iterative database searches with the PSI-BLAST program⁷ and phylogenetic tree construction with the neighbor-joining and parsimony methods¹⁰. Of the 246 *Aquifex* proteins that are most similar to their archaeal homologs (Table 1), 26 belong to families found in archaea and *Aquifex* only. In addition, 60 of the remaining families were investigated by phylogenetic methods and, for 26, statistically significant support (>65% bootstrap replications) of the *Aquifex*/archaea grouping was observed (data not shown).

Aquifex genome contains 36 clusters of two or more adjacent 'archaeal' genes (Fig. 1); the mean length of a cluster is significantly greater ($p < 10^{-3}$) than expected on the basis of a random distribution in the genome (as calculated using a geometric distribution approximation and confirmed by computer simulation). This suggests a conserved arrangement of some genes in *Aquifex* and archaea and, indeed, three such clusters were identified, with the most prominent one including 13 *Aquifex* genes whose arrangement is partially conserved in the archaea but not in any other known bacterium (Fig. 1).

These observations suggest that there has been massive gene exchange between extreme thermophilic archaea and the lineage of bacterial hyperthermophiles represented by *Aquifex*. Convergence brought about by positive selection for thermotolerance could account for a subset of archaeal best hits among *Aquifex* proteins. Nevertheless, the highly significant differences in the level of sequence similarity between archaeal and bacterial best hits for many *Aquifex* proteins, conservation of unique domain architectures in archaea and *Aquifex*, and the phylogenetic analysis results, appear to indicate that at least 10% of the *Aquifex* have been horizontally transferred from the archaea.

The 'archaeal' genes in *Aquifex* are a functionally diverse set. Predictably, the genes that are found exclusively in archaea and *Aquifex* are functionally uncharacterized owing to the lack of experimental data on these organisms. Several of them, however, form highly conserved families that, on the basis of the observed patterns of amino acid residue conservation, could be

predicted to possess as yet unknown enzymatic activities. The remaining genes have homologs in well-characterized genomes and, accordingly, functions can be predicted in most cases. These include metabolic enzymes, transporters and proteins involved in genome replication and repair. Of particular interest are two families of ATP-dependent DNA ligases, one of which has not been described previously and is only distantly related to eukaryotic ligases, an archaeal/eukaryotic type ATPase distantly related to the bacterial RecA, and a small protein homologous to the catalytic domain of DnaG-type DNA primases. In each of these cases, *Aquifex* also encodes a typical bacterial counterpart of the 'archaeal' protein, namely the NAD-dependent DNA ligase, RecA, and a classic DnaG ortholog. Similar chimerism was observed for several enzymes, for example, tryptophan synthase β subunit, peroxidase and isopalmate dehydratase. In these cases, it seems particularly plausible that the 'archaeal' genes have been introduced into the *Aquifex* genome by horizontal transfer, on top of a

TABLE 3. Inclusion of bacterial proteins into conserved archaeal families^a

Bacterial species	Protein from the given species included in archaeal COGs
<i>Aquifex aeolicus</i>	597 (39%)
<i>Synechocystis</i> sp.	707 (22%)
<i>Bacillus subtilis</i>	910 (22%)
<i>Escherichia coli</i>	891 (21%)
<i>Borrelia burgdorferi</i>	215 (17%)

^aA total of 789 families of probable orthologs (clusters of orthologous groups, or COGs) in the three archaeal genomes were identified as previously described. Bacterial proteins were compared with these COGs using the gapped BLASTP program, and a bacterial protein was included in the given COG if its best hits to at least two archaeal genomes were among the COG's members⁹.

LETTER

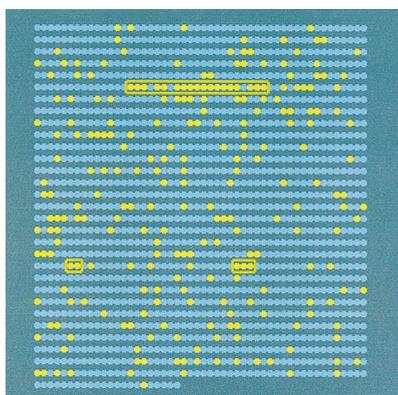


FIGURE 1. Genes of apparent archaeal origin in the genome of *Aquifex aeolicus*. Yellow circles represent genes encoding proteins with reliable best hits to archaeal homologs. Gene clusters conserved in *Aquifex* and archaea are boxed. The largest cluster contains genes for a predicted RNA helicase, a nuclease and a zinc-finger-containing nucleic acid-binding protein; the remaining genes encode uncharacterized proteins, most of which are conserved in archaea and *Aquifex* only.

typical bacterial gene repertoire, and have been retained owing to the specific selective advantage they provided by enabling the bacterium to thrive in high-temperature habitats.

The presence of the same set of genes of apparent archaeal origin in the genomes of two or more thermophilic bacteria from distant bacterial lineages would present strong evidence for the role of these apparently horizontally transferred genes in thermophily. At this time, the sequence information on bacterial thermophiles other than *Aquifex* is insufficient for generalizations. Nevertheless, several genes shared by archaeal and

bacterial thermophiles to the exclusion of mesophilic bacteria are detectable. For example, in addition to the previously described reverse gyrase found in all of the archaea, *Aquifex* and *Thermotoga maritima*^{1,11}, we detected a putative DNA methylase with a modified SAM-binding motif that is encoded not only by *Aquifex* and the archaea, but also by *Thermus aquaticus* and might be involved in additional DNA methylation contributing to thermotolerance.

We showed that the genome of *Aquifex* is a chimera that has a large component shared with the archaea, in addition to the core gene set in common with the rest of the bacteria. It seems likely that bacterial hyperthermophily has evolved secondarily within moderately thermophilic bacteria by continuous acquisition of thermotolerance genes from preadapted hyperthermophiles, namely the archaea. An alternative, in our opinion less likely, is that the preponderance of 'archaeal' genes in *Aquifex* is not the cause but just a consequence of its adaptation to the existence under extreme thermophilic environments, where archaea are dominant organisms. This dilemma is likely to be solved once genomes of other bacterial thermophiles are sequenced. If there is a causal relationship between the acquisition of archaeal genes and adaptation to extreme thermophily, the sets of genes of archaeal origin found in different thermophilic bacteria will overlap to a much greater extent than expected under a random acquisition model. Should that be

the case, theoretical and experimental analysis of these genes will be helpful for understanding the mechanisms of thermophily. A complete, annotated list of *Aquifex* genes whose products show the greatest similarity to archaeal homologs is available as supplementary information on the World Wide Web¹².

References

- 1 Deckert, G. *et al.* (1998) *Nature* 392, 353–358
- 2 Pace, N.R. (1997) *Science* 276, 734–740
- 3 Brown, J.R. and Doolittle, W.F. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 2441–2445
- 4 Bult, C.J. *et al.* (1996) *Science* 273, 1058–1073
- 5 Klenk, H.P. *et al.* (1997) *Nature* 390, 364–370
- 6 Smith, D.R. *et al.* (1997) *J. Bacteriol.* 179, 7135–7155
- 7 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
- 8 Walker, D.R. and Koonin, E.V. (1997) *ISMB* 5, 333–339
- 9 Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science* 278, 631–637
- 10 Felsenstein, J. (1996) *Methods Enzymol.* 266, 418–427
- 11 Guipaud, O. *et al.* (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 10606–10611
- 12 <http://ncbi.nlm.nih.gov/pub/koonin/aquifex/index.html>

**L. Aravind, Roman L. Tatusov,
Yuri I. Wolf, D. Roland Walker
and Eugene V. Koonin**
koonin@ncbi.nlm.nih.gov

*National Center for Biotechnology
Information, National Library of Medicine,
National Institutes of Health, Bethesda,
MD 20894, USA.*

A new section in *Trends in Genetics* for 1999 – Genome Analysis

The purpose of the Genome Analysis section is to provide a forum for original observations concerning the function, organization and evolution of genomes. With the increasing quantities of genome maps and sequence data in public databases, genome analysis and bioinformatics are providing spectacular insights into fundamental biological questions, and this trend is set to continue. In Genome Analysis, *Trends in Genetics* will publish short articles based on the analysis of publicly accessible data. Publications of outstanding quality and of interest to a broad audience of geneticists and molecular biologists will be considered, and all manuscripts will be peer reviewed by an expert panel of referees.

Manuscripts of up to 1000 words will be considered with one or two small illustrations or tables. More detailed instructions are available on request. Genome Analysis will be edited by Eugene Koonin. Potential authors are invited to contact the editor or the *Trends in Genetics* editorial office for further information.

Eugene Koonin
koonin@ncbi.nlm.nih.gov

*National Center for Biotechnology Information, National
Library of Medicine, National Institutes of Health,
Building 38A, Bethesda, MD 20894, USA.*