Title : Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model

Authors : Miklós Csűrös[1] and István Miklós[2]

Authors' affiliations: [1] Department of Computer Science and Operations Research, University of Montréal, Canada. [2] Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary.

Corresponding author: Miklós Csűrös. Département d'informatique et de recherche opérationnelle, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal, QC, H3C 3J7, Canada. Tel: +1 514 343–6111 extension 1655. Fax: +1 514 343–6111. E-mail: csuros@iro.umontreal.ca.

Abstract

Homologous genes originate from a common ancestor through vertical inheritance, duplication or horizontal gene transfer. Entire homolog families spawned by a single ancestral gene can be identified across multiple genomes based on protein sequence similarity. The sequences, however, do not always reveal conclusively the history of large families. In order to study the evolution of complete gene repertoires, we propose here a mathematical framework that does not rely on resolved gene family histories. We show that so-called phylogenetic profiles, formed by family sizes across multiple genomes, are sufficient to infer principal evolutionary trends. The main novelty in our approach is an efficient algorithm to compute the likelihood of a phylogenetic profile in a model of birth-and-death processes acting on a phylogeny.

We examine known gene families in 28 archaeal genomes using a probabilistic model that involves lineage- and family-specific components of gene acquisition, duplication, and loss. The model enables us to consider all possible histories when inferring statistics about archaeal evolution. According to our reconstruction, most lineages are characterized by a net *loss* of gene families. Major increases in gene repertoire have occurred only a few times. Our reconstruction underlines the importance of persistent streamlining processes in shaping genome composition in Archaea. It also suggests that early archaeal genomes were as complex as typical modern ones, and even show signs, in the case of the methanogenic ancestor, of an extremely large gene repertoire.

Introduction

The evolution of homologous gene families, i.e., genes of common ancestry, is enmeshed within species histories in a complex manner (Koonin, 2005). Concomitantly with the diversification of organismal lineages, gene families expand by duplications, individual genes get eliminated, and new genes arrive by lateral transfer. It is now clear that *de novo* gene formation and vertical processes (Snel *et al.*, 2002; Henikoff *et al.*, 1997), such as duplication and loss, act in concert with horizontal gene transfer (Boucher *et al.*, 2003; Gogarten and Townsend, 2005).

Gene families are identified in current practice by pairwise sequence comparisons, coupled with the clustering of postulated homolog pairs (Tatusov *et al.*, 1997; Alexeyenko *et al.*, 2006) The phylogenetic profile of a gene family comprises the family size across a set of organisms, i.e., the number of homologs within the same family in each genome. Such profiles are extremely informative even without taking the gene sequences into account: profile data sets have been used to construct organismal phylogenies (Fitz-Gibbon and House, 1999; Snel *et al.*, 1999; Tekaia *et al.*, 1999) and to infer ancestral gene content (Mirkin *et al.*, 2003; Iwasaki and Takagi, 2007); similar and complementary profiles hint at functional associations (Tatusov *et al.*, 1997; Pellegrini *et al.*, 1999). Considering various evolutionary processes in a mathematical model of gene family evolution is challenging. One main element that distinguishes the present study from past work is the elaboration of a likelihood framework for phylogenetic profiles that simultaneously accounts for gene duplication, loss, and acquisition. In particular, we describe an algorithm for the exact computation of the likelihood in a phylogenetic gain-loss-duplication model.

The present study uses a gain-loss-duplication model to address gene content evolution in Archaea. Relying on a complete set of known homolog families in 28 sequenced genomes, we inferred lineage- and family-specific statistics. In a precursory step, we constructed a plausible phylogeny using 88 universally conserved proteins, which we believe is a noteworthy result on its own, as the phylogeny resolves some problematic euryarchaeal branching orders (involving Thermoplasmatales, Methanopyrus and Methanobacteriales) confidently. Gene loss emerges in our analysis as the dominant force that has shaped archaeal genomes throughout their history. Apparently, genome streamlining has been an ongoing process in all lineages with a fairly constant intensity, apart from dramatic genome compactions in endosymbiotic Archaea. Our reconstruction suggests that early Archaea had a comparable genomic complexity to today's organisms. In particular, the euryarchaeal ancestor of two classes of methanogens had a very large genome, resulting from one of the rare upsurges in gene content, similarly to some modern lineages of Methanosarcina and Halobacteria.

Methods

Phylogenetic profiles in Archaea

Phylogenetic profiles, sequences, and functional annotations were downloaded from the arCOG database of orthologous gene clusters in Archaea (Makarova *et al.*, 2007) at `ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG`. The profiles were amended with data on lineage-specific singletons and inparalog families that have no archaeal homologs outside of one genome (Yuri Wolf, personal communication), which was produced in the process of compiling the arCOG database.

The following organisms are included in the study: *Archæoglobus fulgidus* (Arcfu), *Haloarcula marismortui* ATCC 43049 (Halma), Halobacterium sp. strain NRC-1 (Halsp), *Methanosarcina acetivorans* (Metac), *Methanococcoides burtonii* DSM 6242 (Metbu), *Methanoculleus marisnigri* JR1 (Metcu), *Methanospirillum hungatei* JF-1 (Methu), *Methanocaldococcus jannaschii* (Metja), *Methanopyrus kandleri* (Metka), *Methanosarcina mazei* (Metma), *Methanococcus maripaludis* S2 (Metmp), *Methanosphaera stadtmanæ* (Metst), *Methanothermobacter thermoautotrophicus* (Metth), *Nanoarchæum equitans* (Naneq), *Picrophilus torridus* DSM 9790 (Picto), *Pyrococcus abyssi* (Pyrab), *Pyrococcus furiosus* (Pyrfu), *Thermoplasma acidophilum* (Theac), *Thermococcus kodakaraensis* KOD1 (Theko), *Thermoplasma volcanium* (Thevo), *Æropyrum pernix* (Aerpe), *Caldivirga maquilingensis* IC-167 (Calma), *Cenarchæum symbiosum* (Censy), *Hyperthermus butylicus* (Hypbu), *Pyrobaculum ærophilum* (Pyrae), *Sulfolobus solfataricus* (Sulso); *Sulfolobus acidocaldarius* DSM 639 (Sulac), *Thermofilum pendens* Hrk 5 (Thepe) with the last eight classified as crenarchaeota. The abbreviations are those used by (Makarova *et al.*, 2007) and the arCOG database.

Reconstruction of archaeal phylogeny

The phylogeny was constructed using concatenated multiple alignments of selected orthologous protein sequences. The sequences were chosen from the arCOG database based on phylogenetic profiles: we selected all arCOG groups where every studied genome contained exactly one homolog. There are 88 such groups (see Supplemental Material for sequences), and 46 of those correspond to ribosomal proteins. Alignments were done using the program Muscle (Edgar, 2004). Phylogenies were built by likelihood maximization using PhyML (Guindon and Gascuel, 2003), with the Jones-Taylor-Thornton substitution model and eight discrete Gamma categories and invariant sites. The expected number of substitutions per amino acid site was computed on each edge for the ribosomal proteins in the JTT+I+Γ8 model by PhyML. Bootstrap support values for the branches were computed by PhyML, using 500 replicates.

Inference of gene content evolution

We maximized the likelihood (see below for the likelihood computation) of the data set using a gain-loss-duplication model with a Poisson distribution at the root and four discrete Gamma categories capturing rate variation across families, for edge length t_f and duplication λ_f each. For a given set of model parameters (three parameters — $\hat{t}_e\hat{\kappa}_e$, $\hat{t}_e\hat{\mu}_e$, $\hat{t}_e\hat{\lambda}_e$ — per edge, one for the root's Poisson parameter Γ, and two Gamma shape parameters for rate variation), the likelihood of each family was computed using (1) with the described methods of manipulating rate variation and correcting for absent profiles. The data set's likelihood (i.e., the product of family likelihoods) was then maximized numerically as a function of the model parameters, using custom-made software implementing the Broyden-Fletcher-Goldfarb-Shanno conjugate gradient method and Brent's one-dimensional optimization method (Press *et al.*, 1997). Family sizes and lineage-specific events (gains,losses,expansions,contractions) were computed using posterior probabilities in the optimized gain-loss-duplication model.

Phylogenetic birth-and-death model

A *phylogenetic birth-and-death model* formalizes the evolution of an organism-specific census variable along a rooted phylogeny T. We consider only binary phylogenies here; the full set of methods applicable to multi-furcating phylogenies is described in the Supporting Information. The model specifies edge lengths, as well as birth-and-death processes (Ross, 1996; Kendall, 1949) acting on the edges. Populations of identical individuals evolve along the tree from the root towards the leaves by Galton-Watson processes. At non-leaf nodes of the tree, populations are instantaneously copied to evolve independently along the adjoining descendant edges. Let the random variable $\xi(x) \in \{0, 1, 2, \dots\}$ denote the population count at every node $x \in \mathcal{V}(T)$. Every edge xy is characterized by a loss rate μ_{xy}, a duplication rate λ_{xy} and a gain rate κ_{xy}. If $\bigl(X(t)\colon t \geq 0\bigr)$ is a linear birth-and-death process (Kendall, 1949; Takács, 1962) with these rate parameters, then $\mathbb{P}\Bigl\{\xi(y) = m \Bigm| \xi(x) = n\Bigr\} = \mathbb{P}\Bigl\{X(t_{xy}) = m \Bigm| X(0) = n\Bigr\}$, where $t_{xy} > 0$ is the edge length, which defines the time interval during which the birth-and-death process runs. The joint distribution of $\bigl(\xi(x)\colon x \in \mathcal{V}(T)\bigr)$ is determined by the phylogeny, the edge lengths and rates, along with the distribution at the root ρ, denoted as $\gamma(n) = \mathbb{P}\{\xi(\rho) = n\}$.

It is assumed that one can observe the population counts at the terminal nodes (i.e., leaves), but not at the inner nodes of the phylogeny. Since individuals are considered identical, we are also ignorant of the ancestral relationships between individuals within and across populations. The population counts at the leaves form a *phylogenetic profile*, which is formally a function $\Phi\colon \mathcal{L}(T) \mapsto \{0, 1, 2, \dots\}$, where $\mathcal{L}(T) \subset \mathcal{V}(T)$ denote the set of leaf nodes. Our central problem is to compute the likelihood of a profile, i.e., the probability of the observed counts for fixed model parameters. Define the notation $\Phi(\mathcal{L}') = \bigl(\Phi(x)\colon x \in \mathcal{L}'\bigr)$ for the partial profile within a subset $\mathcal{L}' \subseteq \mathcal{L}(T)$. Similarly, let $\xi(\mathcal{L}') = \bigl(\xi(x)\colon x \in \mathcal{L}'\bigr)$ denote the vector-valued random variable composed of individual population counts. The *likelihood* of Φ is the probability $L = \mathbb{P}\Bigl\{\xi\bigl(\mathcal{L}(T)\bigr) = \Phi\Bigr\}$. Let T_x denote the subtree of T rooted at node x. Define the *survival count range* M_x for every node x as $M_x = \sum_{y \in \mathcal{L}(T_x)} \Phi(y)$. Clearly, the ranges can be calculated easily in a postorder traversal.

For our discussion, we borrow standard terminology applied to homologous genes (Sonnhammer and Koonin, 2002). For every edge xy, the population of node y can be split by ancestry at node x: *inparalog* groups are formed by the progenies of each individual at x and a *xenolog* group is formed by the individuals whose ancestor immigrated into the population. When $\xi(x) = n$ on the edge xy, then $\xi(y) = \eta + \sum_{i=1}^{n} \zeta_i$, where η is the xenolog group size, and ζ_i are the independent and identically distributed inparalog group sizes. The distribution of xenolog and inparalog group sizes is the well-characterized transient distribution of the appropriate linear birth-and-death processes (Karlin and McGregor, 1958; Kendall, 1949; Takács, 1962); see Supplemental Material. Namely, each ζ_i has a shifted geometric distribution, and for $\kappa > 0$, η has a negative binomial or Poisson distribution. The distributions' parameters are known functions of the edge length t_{xy}

and rates $\kappa_{xy}, \lambda_{xy}, \mu_{xy}$.

Surviving lineages

A key factor in inferring the likelihood formulas is the probability that a given individual at a tree node x has no descendants at the leaves within the subtree rooted at x. The corresponding *extinction probability* is denoted by D_x, which can be computed in a postorder traversal (Csűrös and Miklós, 2006). An individual at node x is referred to as *surviving* if it has at least one progeny at the leaves descending from x. Let $\Xi(x)$ denote the number of surviving individuals at each node x. The number of surviving xenologs and inparalogs follow the same class of distributions as the total number of xenologs and inparalogs (see Supplemental Material). Consequently, if $\xi(x) = n$ on edge xy, then $\Xi(y) = \eta + \sum_{i=1}^{n} \zeta_i$, where η is the surviving xenolog count with a Poisson or negative binomial distribution, and ζ_i are surviving paralog counts, with negative binomial distributions. The distributions' parameters can be computed explicitly using the process parameters and the extinction probabilities. In the formulas to follow, we use the probabilities $w_y^*[m|n] = \mathbb{P}\{\eta + \sum_{i=1}^{n} \zeta_i = m; \forall \zeta_i > 0\}$, which can be computed by dynamic programming for all $n, m \le M_y$ in $O(M_y^2)$ time (see Supplemental Material).

Computing the likelihood

We compute the likelihood using *conditional survival likelihoods* defined as the probability of observing the partial profile within T_x given the number of surviving individuals $\Xi(x)$: $L_x[n] = \mathbb{P}\Big\{\xi\big(\mathcal{L}(T_x)\big) = \Phi\big(\mathcal{L}(T_x)\big) \;\Big|\; \Xi(x) = n\Big\}$. For $m > M_x$, $L_x[m] = 0$. For values $m = 0, 1, \ldots, M_x$, the conditional survival likelihoods can be computed recursively as shown below.

If node x is a leaf, then

$$L_x[n] = \begin{cases} 0 & \text{if } n \neq \Phi(x); \\ 1 & \text{if } n = \Phi(x). \end{cases}$$

If x is an inner node with children x_1, x_2, then $L_x[n]$ can be expressed using $L_{x_i}[\cdot]$ and auxiliary values $B_{i;\cdot,\cdot}$ for $i = 1, 2$ in the following manner. Auxiliary values $B_{i;t,s}$ are defined for $i = 1, 2$ and $s = 0, \ldots, M_{x_i}$ as follows.

$$\begin{aligned} B_{i;0,s} &= \sum_{m=0}^{M_{x_i}} w^*_{x_i}[m|s] L_{x_i}[m] && \{0 \leq s \leq M_{x_i}\} \\ B_{2;t,M_{x_2}} &= G_{x_2}(0) B_{2;t-1,M_{x_2}} \\ B_{2;t,s} &= B_{2;t-1,s+1} + G_{x_2}(0) B_{2;t-1,s} && \{0 \leq s < M_{x_2}\} \end{aligned}$$

where $G_{x_i}(k) = \mathbb{P}\{\zeta = k\}$ for a surviving inparalog group at x_i. In the above equations, $0 < t \leq M_{x_1}$. For all $n = 0, \ldots, M_x$

$$L_x[n] = \bigl(1 - D_x\bigr)^{-n} \sum_{\substack{0 \leq t \leq M_{x_1} \\ 0 \leq s \leq M_{x_2} \\ t+s=n}} \binom{n}{s} \bigl(D_{x_1}\bigr)^s B_{1;0,t} B_{2;t,s}.$$

The complete likelihood is computed as

$$L = \sum_{m=0}^{M_\rho} L_\rho[m] \mathbb{P}\{\Xi(\rho) = m\}.$$

For some parametric distributions γ, there is a closed formula for $\mathbb{P}\{\Xi(\rho) = m\}$. In particular, if γ is the stationary distribution for a gain-loss-duplication or a gain-loss models, then $\Xi(\rho)$ has a negative binomial or Poisson distribution, respectively. The likelihood for a Poisson distribution at the root is

$$L = \sum_{m=0}^{M_\rho} L_\rho[m] \exp\Big(-\Gamma(1-D_\rho)\Big) \frac{\big(\Gamma(1-D_\rho)\big)^m}{m!} \tag{1}$$

where Γ is the mean family size at the root.

The likelihood formula (1) is corrected in order to account for the fact that the data set does not contain all-absent profiles with $\Phi(x) = 0$ for all leaves x, in a manner analogous to (Felsenstein, 1992).

Family-specific rate variation is considered by computing the likelihood values for each discrete rate category c characterized by factors $(t_c, \kappa_c, \mu_c, \lambda_c)$. The factors in our analysis are either constant 1, or correspond to the expected values within the four quartiles of a Gamma distribution with mean 1.

Results and discussion

Computational analysis of phylogenetic profiles

Birth-and-death processes are commonly used to model a population of identical individuals (Kendall, 1949; Karlin and McGregor, 1958) and waiting queues (Takács, 1962). Their use in modeling gene family evolution is justified by the fact that losses and duplications seem to occur independently between the members of multi-gene families (Nei and Rooney, 2005). The most general process we consider is a gain-loss-duplication process which is characterized by the rates of gain κ, loss μ and duplication λ: a population of size n grows by a rate of $(\lambda n + \kappa)$ and decreases by a rate of μn. In our context, the population comprises homologs of a given family in the genome. Gene acquisition occurs with a rate of κ, combining various means such as innovation and lateral transfer. We model gene family evolution in a phylogenetic setting by associating gain-loss-duplication processes with the branches of a phylogenetic tree. The corresponding phylogenetic birth-and-death model defines a probabilistic framework for the evolution of gene family size. The observed family sizes at the terminal nodes form a phylogenetic profile. In principle, a phylogenetic birth-and-death model suits likelihood-based inference since it is a probabilistic graphical model (Jordan, 2004) with a tree structure. The mathematical difficulties stem from the fact that the state space of the processes (i.e., family size) is infinitely large. Consequently, routine computational techniques used to analyze molecular sequence evolution (Felsenstein, 1981) are not applicable. Previously proposed likelihood methods (Hahn *et al.*, 2005; Spencer *et al.*, 2006; Iwasaki and Takagi, 2007) have sidestepped the infinity problem by using approximative calculations with bounds on maximal family size.

We have introduced (Csűrös and Miklós, 2006) a procedure for computing the likelihood in a restricted gain-loss-duplication model (assuming $0 < \kappa$ and $0 < \lambda < \mu$), without imposing artificial size bounds. The weakness of that procedure is potential numerical instability, due to the use of alternating sums in the formulas. We found practical cases (such as the archaeal gene content study we report below), where the numerical instability led to serious errors. The novel procedure presented here is numerically stable, as well as computationally efficient. It applies to arbitrary gain-loss-duplication models, including degenerate cases such as the one of (Hahn *et al.*, 2005) with $\lambda = \mu$ and $\kappa = 0$. The algorithm takes $O(M^2 n)$ time to complete for a phylogenetic profile over n species and M total number of genes (see Supplemental Material).

Gene content evolution in Archaea

Archaea constitute one of the three main domains of cellular life, and are notable for a spectacular diversity of adaptive strategies to extreme environments (Garrett and Klenk, 2006). We examined gene content evolution in Archaea. For the purposes of the study we have selected 28 completely sequenced genomes covering all major physiological and metabolic groups recognized in cultured Archaea: thermophiles, halophiles, acidophiles, nitrifiers and methanogens (Valentine, 2007). Homolog gene families were extracted from the arCOG (archaeal clusters of orthologous groups) database (Makarova *et al.*, 2007), and combined with groupings of genes that have no archaeal homologs outside of single genomes. The complete data set consists of 14216 families, of which 7461 are among the arCOGs.

Phylogenetic relationships

Archaeal phylogenetic relationships have been resolved to an increasing degree of confidence (Forterre *et al.*, 2006) with the aid of accumulating sequence data. Figure 1 shows our consensual phylogeny based on maximum likelihood trees for concatenated alignments of 46 ribosomal proteins (r-proteins) and 88 unique conserved proteins (uc-proteins), which are precisely those that have exactly one homolog in each sampled genome. Congruent phylogenies were proposed before (Forterre *et al.*, 2006; Gribaldo and Brochier-Armanet, 2006), based on complete phylogenomics evidence. In our study, r-proteins and uc-proteins show solid support for most recognized phylogenetic relationships, but provide contradictory signals for the placement of some euryarchaeal groups. Notably, both sequence data sets support the basal position of *N. equitans*, which was originally thought to be a specimen of a separate group from Euryarchaeota and Crenarchaeota (Waters *et al.*, 2003), but is more likely an early-branching euryarchaeal organism (Makarova and Koonin, 2005; Forterre *et al.*, 2006). The data also support the early branching position of non-thermophilic crenarchaea represented by *C. symbiosum*. In fact, non-thermophilic crenarchaea may constitute a separate phylum from Euryarchaota and Crenarchaeota, tentatively named Thaumarchaeota (Brochier-Armanet *et al.*, 2008).

[Figure 1 about here.]

The observed uncertainties about euryarchaeal groups concern the placement of Thermoplasmata, and so-called Class I methanogens (Bapteste *et al.*, 2005) comprising Methanopyrales, Methanobacteriales and Methanococcales. Thermoplasmata were originally thought to be a an early-branching lineage of Euryarchaeota

(Forterre *et al.*, 2006), but analyses of r-proteins (Matte-Tailliez *et al.*, 2002) have provided strong evidence for their late-branching position after Class I methanogens as in Figure 1. R-proteins in our study support the late-branching of Thermoplasmatales (89% bootstrap value), but a maximum-likelihood tree built from uc-proteins places Thermoplasmatales between Nanoarchaea and Thermococcales (66% BV). It has been argued that this placement is due to long-branch attraction (Matte-Tailliez *et al.*, 2002; Brochier *et al.*, 2004), a frequent systematic bias of sequence evolution models (Rodríguez-Ezpeleta *et al.*, 2007). Indeed, after we removed *N. equitans* and *C. symbiosum* from the uc-protein data set, the late-branching position of Thermoplasmatales regained solid support (100% BV).

The correct phylogenetic position of *M. kandleri* (Metka) is one of the remaining puzzles in archaeal evolution. The existence of close phylogenetic relationships between Class I methanogens is fairly certain, but different protein sets and taxonomic sampling give conflicting or weak indications (Slesarev *et al.*, 2002; Brochier *et al.*, 2004, 2005; Gao and Gupta, 2007) about the exact branching order among Methanopyrales, Methanobacteriales and Methanococcales. R-proteins in our study give a weak support for the monophyly of Methanococcales and Methanobacteriales at the exclusion of Methanopyrales (49% BV) and faintly favor the paraphyly of Class I methanogens (37% BV for the immediate split of Methanopyrales between Thermococcales and Methanobacteriales/Methanococcales; see Supplemental Material). Uc-proteins, however, solidly point to the monophyly of Class I methanogens ($> 97\%$ BV). Interestingly, the maximum-likelihood trees built from uc-proteins do not resolve well the relationships between Halobacteriales, Methanosarcinales and Methanomicrobiales (see Supplemental Material), but there is little reason to doubt that r-proteins provide a genuine phylogenetic signal about the monophyly of Class II methanogens (Bapteste *et al.*, 2005; Brochier-Armanet *et al.*, 2008), uniting Methanosarcinales and Methanomicrobiales.

We conclude that based on protein sequences, Thermoplasmatales constitute a late-branching euryarchaeal lineage, and their early-branching status is a long-branch attraction artifact. Furthermore, the sequences provide evidence of the monophyly of both Class I and Class II methanogens.

Evolutionary rates: correlations between sequence and gene content evolution

We experimented with models of increasing complexity that combine lineage- and gene-specific factors in the gain-loss-duplication processes. Specifically, we assumed that the process for family f on branch e is characterized by the rates $\kappa = \hat{\kappa}_e\kappa_f$, $\mu = \hat{\mu}_e\mu_f$, $\lambda = \hat{\lambda}_e\lambda_f$, and runs for a duration of $t = \hat{t}_e t_f$. Here, $\hat{t}_e, \hat{\kappa}_e, \hat{\mu}_e, \hat{\lambda}_e$ are branch-specific process parameters, and $t_f, \kappa_f, \mu_f, \lambda_f$ are family-specific rate variation coefficients. Starting with simple models with invariant family-specific coefficients, we introduced rate variation in a model hierarchy with increasing complexity. In more complex models, some coefficients were drawn randomly from a discretized Gamma distribution (Yang, 1994). Different family-specific coefficients do not have the same impact on the model fit. We found the largest improvement when introducing variation in edge length (t_f), followed by duplication-rate variation (λ_f). Further variation in loss and gain rates led to insignificant improvements in the model fit, and were not assumed in the analysis.

[Figure 2 about here.]

In the absence of extraneous scaling, we set $\hat{t}_e = 1$ in order to examine the total rates of gene content change on each edge e. We found a conspicuous correlation across branches between the rate of sequence evolution (expected numbers of substitutions per site for ribosomal proteins) and the component rates of gene content evolution: on this point, see Figure 2 for loss, and the Supplemental Material for duplication and gain. More precisely, the correlation holds for the lineage-specific components of loss, duplication and gain rates in a decreasing order of strength (P-values of $1.1 \cdot 10^{-11}$, $8.2 \cdot 10^{-6}$, $1.6 \cdot 10^{-4}$, respectively, by Student's t-test for Spearman rank-order correlation coefficient).

The apparent correlations between gene content and sequence evolution rates imply that a steady balance has been maintained between drift and natural selection in almost all lineages. Loss and duplication rates, in particular, have similar vagaries as amino acid substitution rates, and provide thus comparable molecular clocks. We measured each terminal node's depth by summing the rates along branches from the root to the node in question. Excluding *N. equitans* and *C. symbiosum*, the coefficient of variation of the depth is 26% for protein sequences, 23% for gene loss rates and 20% for duplication rates. Depths by gene gain rates span about a four-fold range: for substitution, loss, and duplication, the span is close to two-fold.

Genes have thus been eliminated in all archaeal lineages with a fairly universal constancy, apart from occasional accelerations. In other words, genome degradation processes seem to persist at a fairly common intensity in every lineage (Mira *et al.*, 2001). Conceivably, genome decay is counterbalanced by natural selection that eliminates deleterious mutations. The root cause of dramatically increased gene loss in obligate symbionts such as *N. equitans* (Makarova and Koonin, 2005) may be reduced selection (Hershberg *et al.*, 2007; Koonin and Wolf, 2008). Principles of population genetics imply that changes in population size alone can explain rate changes (Lynch, 2006): selection power is weaker in a smaller population, which should manifest in accelerated evolution of sequences (Ohta, 1972) and gene content.

We examined the differences between evolutionary rates in sibling terminal taxa for signs of natural selection. Figure 2 shows that gene loss and amino acid substitution rates differ in a concerted fashion for three pairs, that is, for *M. stadtmanæ*-*M. thermoautotrophicus*, Halobacterium sp.-*H. marismortuimi*, and *S. acidocaldarius*-*S. sulfolobus*. In seven other pairs, loss rates are essentially the same, even if substitution rates may differ. The agreements between substitution and gene loss rate changes attest to common selection forces and mutation processes acting on different forms of genome decay, and are predicted by population-genetic arguments (Lynch, 2006).

In the lineage leading to *M. stadtmanæ*, a human commensal (Fricke *et al.*, 2006), all rates are simultaneously larger when compared to its sibling lineage *M. thermoautotrophicus*, which may be attributed to a smaller population size for the former, which has a smaller habitat. Gene gain and duplication rates behave in general less predictably: numerical differences between loss, gain, and duplication rates on sibling lineages occur in almost all possible sign combinations. The observed fluctuations corroborate the intuition that selection pressures acting on gain and duplication are strong and variable (Wolf *et al.*, 2002). It is plausible that during episodes of massive adaptation, the selective advantages of gene acquisition may outweigh possible negative consequences of an increased genome, and thus drive elevated gene gains, especially if coupled with small population sizes. In our case, unusually large gain rates are inferred on some of the deepest branches (such as the one leading to node E1 on Figure 1 or to the halobacterial ancestor), as well as on the terminal branches leading to *M. acetivorans* (Metac), *H. marismortuimi* (Halma) and *P. ærophilum* (Pyrae).

History of archaeal gene census: streamlining and surges

We inferred a probable history of archaeal gene content using posterior probabilities for ancestral family sizes and family size changes, computed from the phylogenetic profiles in the fitted model. Figure 3 summarizes the results by lineages. (See Supplemental Material for bootstrap confidence intervals: the uncertainty in ancestral family counts is estimated to be within $\pm 19\%$ for all nodes. We note that alternate phylogenies for the Class I methanogens give similar results that fall within those confidence intervals.)

[Figure 3 about here.]

Our reconstruction suggests a recurrent theme in archaeal evolution: a major physiological or metabolic invention leads to a successful founding population in a new environment, which then further diversifies by genomic streamlining. We can see notably that Figure 3 shows only a few branches where gains prevail over losses (i.e., at least twice as many gains as losses): such is the case for some deep crenarchaeal and euryarchaeal branches, and the terminal lineages for *M. acetivorans* and *H. marismortuimi*. About half of the remaining terminal lineages and two-thirds of remaining deep lineages are dominated by loss. Moreover, there is only one ancestral node (the crenarchaeal ancestor) in the entire tree for which gain is dominant in both descendant lineages.

Why would gene loss be so prevalent? We speculate that the versatility of a large genome in such extant lineages as *M. acetivorans* (Galagan *et al.*, 2002) and *H. marismortuimi* (Baliga *et al.*, 2004) can be upheld for only relatively short time periods. Genetic drift already leads to the diversification of descendant lineages, which are frequently isolated, given the disconnectedness of the extreme environments they dwell in (Whitaker *et al.*, 2003; Escobar-Páramo *et al.*, 2005). Specialization and the loss of dispensable functions should be favorable in the descendants that are typically under significant energy stress (Valentine, 2007). Genomic streamlining should also be favored by population-size effects due to the isolation (Lynch, 2006), even in the case of slightly deleterious loss of function.

After the crenarchaeal split, the main euryarchaeal lineage has been characterized by the accumulation of new families, culminating in a large surge on the branch leading to node E1, where many new families appeared. The time interval (judging by sequence divergence in Figure 1) and the extent of gene gain is similar to what is seen with *H. marismortuimi* (Halma) and *M. acetivorans* (Metac). The inference of large gains in the E1 lineage is due to the large number of gene families shared between multiple descendant lineages, and especially between the two classes of methanogens (Slesarev *et al.*, 2002; Bapteste *et al.*, 2005; Gao and Gupta, 2007; Makarova *et al.*, 2007). In fact, this lineage may very well have been where hydrogenotrophic methanogenesis was invented, which then underwent modifications, extensions and degradations in subsequent lineages. It was noted in previous genome-scale comparisons (Bapteste *et al.*, 2005; Gao and Gupta, 2007) that it is likely that euryarchaeal lineages acquired methanogenesis predominantly by vertical inheritance, because the associated pathways are fairly complex, and neither the sequences nor the phylogenetic profiles show evidence of substantial amounts of lateral gene transfer. Figure 3 suggests that methanogenesis appeared after the split of Thermococcales in the company of more than 760 genes. Based on extant examples of archaea with such swelled genomes (Galagan *et al.*, 2002; Baliga *et al.*, 2004), it is plausible that the corresponding archaeal organisms were extremely versatile.

Our inference of ancestral gene content is quite different from previous reconstructions based on parsimony principles (Makarova *et al.*, 2007; Csűrös, 2008): at deep nodes, we postulate larger genomes. Parsimonious reconstructions (Mirkin *et al.*, 2003; Kunin *et al.*, 2005; Csűrös, 2008) aim to minimize the number of implied loss and gain events. As a consequence, parsimony inherently underestimates the age of gene families.

A major concern in ancestral gene content reconstruction is that "patchy" profiles arise from a combination of lineage-specific loss events and lateral gene transfers (LGT). Frequent lateral gene transfers imply smaller ancestral genome sizes (Dagan and Martin, 2007). Our reconstruction reveals the prevalence of differential loss, but LGT events are far from uncommon. Lineage-specific gains ("Gain column in Figure 3) account to more than 14% of families ("Families in the genome") at half of all the lineages. A probabilistic framework, such as a phylogenetic birth-and-death model, makes it feasible to take all possible gene family hostories into consideration in a mathematically sound way. A case in point is the last archaeal common ancestor (LACA), where only about 1300 families are inferred to have been present with a posterior probability of at least 90%, which is close to a parsimony-based inference of about 1000 families (Makarova *et al.*, 2007). Given the uncertainties of most family histories, the exact genome composition of LACA is hard to estimate, but the fractional probabilities point to a genome with slightly more than 2000 families, which is similar to such extant organisms as *S. sulfolobus*. Such a large genome size implies that LACA's genomic complexity was even greater than previously imagined (Makarova *et al.*, 2007), on a par with modern, moderately-sized archaeal genomes.

Acknowledgments

This work has been supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Part of the study was done while M.Cs. was a sabbatical visitor at the Rényi Institute of Mathematics, supported by a Marie-Curie Transfer-of-Knowledge fellowship. We are grateful to Yuri Wolf for providing data on lineage-specific gene families. We thank Igor Rogozin, Csaba Pál and Balázs Papp for informative discussions.

References

Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple genomes. *Bioinformatics*, **22**, e9–e15.

Baliga, N. S. *et al.* (2004). Genome sequence of Haloarcula morismurtuimi: a halophilic archaeon from the Dead Sea. *Genome Research*, **14**, 2221–2234.

Bapteste, É., Brochier, C., and Boucher, Y. (2005). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea*, **1**, 353–363.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbo, C. L., Case, R. J., and Doolittle, W. F. (2003). Lateral gene transfer and the origin of prokaryotic groups. *Annual Review of Genetics*, **37**, 283–328.

Brochier, C., Forterre, P., and Gribaldo, S. (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus paradox. *Genome Biology*, **5**, R17.

Brochier, C., Forterre, P., and Gribaldo, S. (2005). An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evolutionary Biology*, **5**, 36.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, **6**, 245–252.

Csűrös, M. (2008). Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions. *Springer Lecture Notes in Bioinformatics*, **5267**, 72–86. Proc. Sixth RECOMB Comparative Genomics Satellite Workshop.

Csűrös, M. and Miklós, I. (2006). A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Springer Lecture Notes in Bioinformatics*, **3909**, 206–220. Proc. Tenth Annual International Conference on Research in Computational Molecular Biology (RECOMB).

Dagan, T. and Martin, W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences of the USA*, **104**(3), 870–875.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Escobar-Páramo, P., Gosh, S., and DiRuggiero, J. (2005). Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophylic archaeon Pyrococcus. *Molecular Biology and Evolution*, **22**(11), 2297–2303.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

Felsenstein, J. (1992). Phylogenies from restriction sites, a maximum likelihood approach. *Evolution*, **46**, 159–173.

Fitz-Gibbon, S. T. and House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, **27**(21), 4218–4222.

Forterre, P., Gribaldo, S., and Brochier-Armanet, C. (2006). Natural history of the archaeal domain. In Garrett and Klenk (2006), chapter 2, pages 17–28.

Fricke, W. F. *et al.* (2006). The genome sequence of Methanosphaera stadtmanae reveals why this human intestinal archaeon is restricted to methanol and H_2 for methane formation and ATP synthesis. *Journal of Bacteriology*, **188**(2), 642–658.

Galagan, J. E., Nusbaum, C., Roy, A., *et al.* (2002). The genome of M. acetivorans reveals extensive metabological and physiological diversity. *Genome Research*, **12**, 532–542.

Gao, B. and Gupta, R. S. (2007). Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. *BMC Genomics*, **8**, 86.

Garrett, R. A. and Klenk, H.-P., editors (2006). *Archaea: Evolution, Physiology, and Molecular Biology*. Blackwell Publishing, Malden, Mass.

Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, **3**, 679–687.

Gribaldo, S. and Brochier-Armanet, C. (2006). The origins and evolution of Archaea: a state of the art. *Philosophical Transactions of the Royal Society of London, Series B*, **361**, 1007–1022.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and aaccurate accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., and Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, **15**, 1153–1160.

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Atwood, T. K., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.

Hershberg, R., Tang, H., and Petrov, D. A. (2007). Reduced selection leads to accelerated gene loss in Shigella. *Genome Biology*, **8**, R164.

Iwasaki, W. and Takagi, T. (2007). Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics*, **23**(13), i230–i239.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, **19**(1), 140–155.

Karlin, S. and McGregor, J. (1958). Linear growth, birth, and death processes. *Journal of Mathematics and Mechanics*, **7**(4), 643–662.

Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society Series B*, **11**(2), 230–282.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, **39**, 309–338.

Koonin, E. V. and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*. Advance access published online on October 23, 2008. DOI:10.1093/nar/gkn668.

Kunin, V., Goldovsky, L., Darzentas, N., and Ouzounis, C. A. (2005). The net of life: reconstructing the microbial phylogenetic network. *Genome Research*, **15**(7), 954–959.

Lynch, M. (2006). Streamlining and simplification of microbial genome achitecture. *Annual Review of Microbiology*, **60**, 327–349.

Makarova, K. and Koonin, E. V. (2005). Evolutionary and functional genomics of the Archaea. *Current Opinion in Microbiology*, **8**, 586–594.

Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I., and Koonin, E. V. (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biology Direct*, **2**, 33.

Matte-Tailliez, O., Brochier, C., Forterre, P., and Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Molecular Biology and Evolution*, **19**(5), 631–639.

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, **17**(10), 589–596.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y., and Koonin, E. V. (2003). Algorithms for computing evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, **3**, 2.

Nei, M. and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, **39**(1), 121–152.

Ohta, T. (1972). Population size and rate of evolution. *Journal of Molecular Evolution*, **1**, 305–314.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the USA*, **96**(8), 4285–4288.

Press, W. H., Teukolsky, S. A., Vetterling, W. V., and Flannery, B. P. (1997). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition.

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, **56**(3), 389–399.

Ross, S. M. (1996). *Stochastic Processes*. Wiley & Sons, second edition.

Slesarev, A. I. *et al.* (2002). The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences of the USA*, **99**(7), 4644–4649.

Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nature Genetics*, **21**(1), 108–110.

Snel, B., Bork, P., and Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, **12**(1), 17–25.

Sonnhammer, E. L. L. and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, **18**(12), 619–620.

Spencer, M., Susko, E., and Roger, A. J. (2006). Modelling prokaryote gene content. *Evolutionary Bioinformatics Online*, **2**, 165–186.

Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press, New York.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631–637.

Tekaia, F., Lazcano, A., and Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Research*, **9**(6), 550–557.

Valentine, D. L. (2007). Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Reviews Microbiology*, **5**, 316–323.

Waters, E. *et al.* (2003). The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences of the USA*, **100**(22), 12984–12988.

Whitaker, R. J., Grogan, D. W., and Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophylic archaea. *Science*, **301**(5635), 976–978.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. (2002). Genome trees and the Tree of Life. *Trends in Genetics*, **18**(9), 472–479.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, **39**, 306–314.

List of Figures

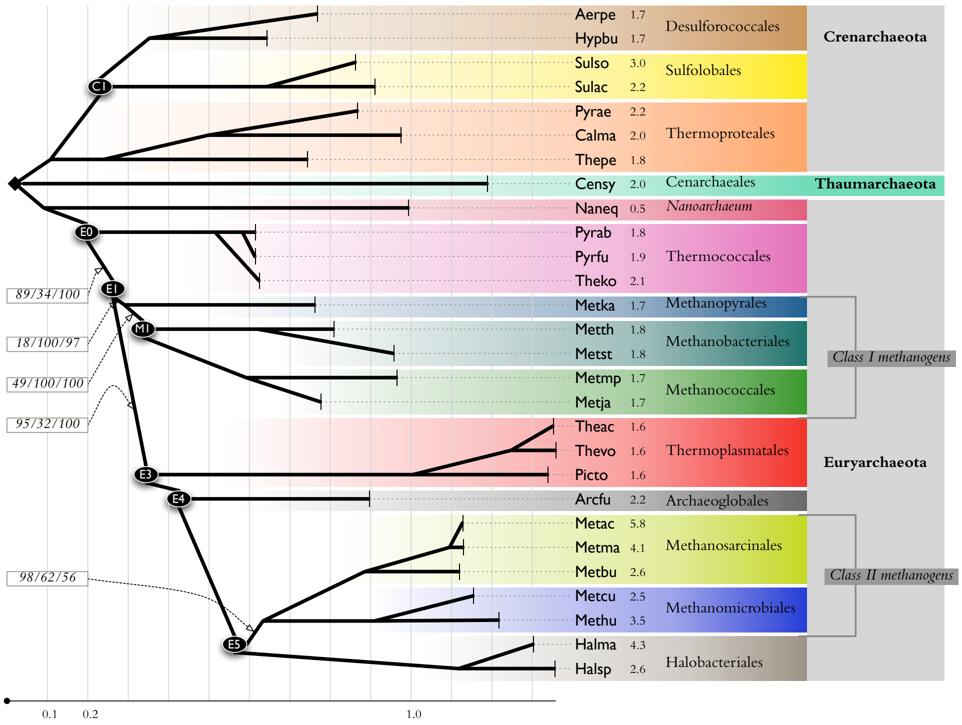
Aerpe 1.7
Hypbu 1.7
Desulforococcales
Sulso 3.0
Sulac 2.2
Sulfolobales
Pyrae 2.2
Calma 2.0
Thepe 1.8
Thermoproteales
Crenarchaeota
Censy 2.0
Cenarchaeales
Thaumarchaeota
Naneq 0.5
Nanoarchaeum
Pyrab 1.8
Pyrfu 1.9
Theko 2.1
Thermococcales
Metka 1.7
Methanopyrales
Metth 1.8
Metst 1.8
Methanobacteriales
Metmp 1.7
Metja 1.7
Methanococcales
Class I methanogens
Theac 1.6
Thevo 1.6
Picto 1.6
Thermoplasmatales
Euryarchaeota
Arcfu 2.2
Archaeoglobales
Metac 5.8
Metma 4.1
Metbu 2.6
Methanosarcinales
Class II methanogens
Metcu 2.5
Methu 3.5
Methanomicrobiales
Halma 4.3
Halsp 2.6
Halobacteriales
C1
E0
E1
M1
E3
E4
E5
89/34/100
18/100/97
49/100/100
95/32/100
98/62/56
0.1
0.2
1.0

Figure 1

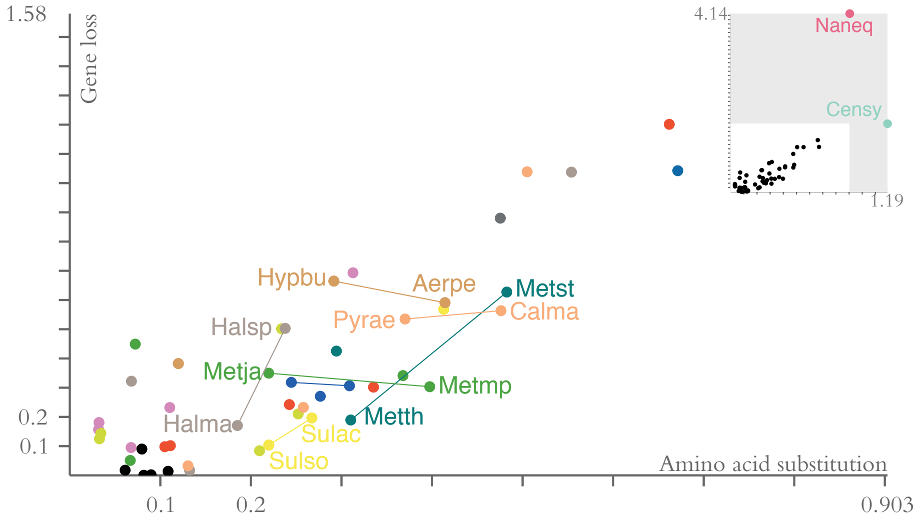
1.58
Gene loss
0.2
0.1
0.1
0.2
0.903
Amino acid substitution
Hypbu
Aerpe
Metst
Calma
Pyrae
Halsp
Metja
Metmp
Metth
Halma
Sulac
Sulso
4.14
Naneq
Censy
1.19

Figure 2

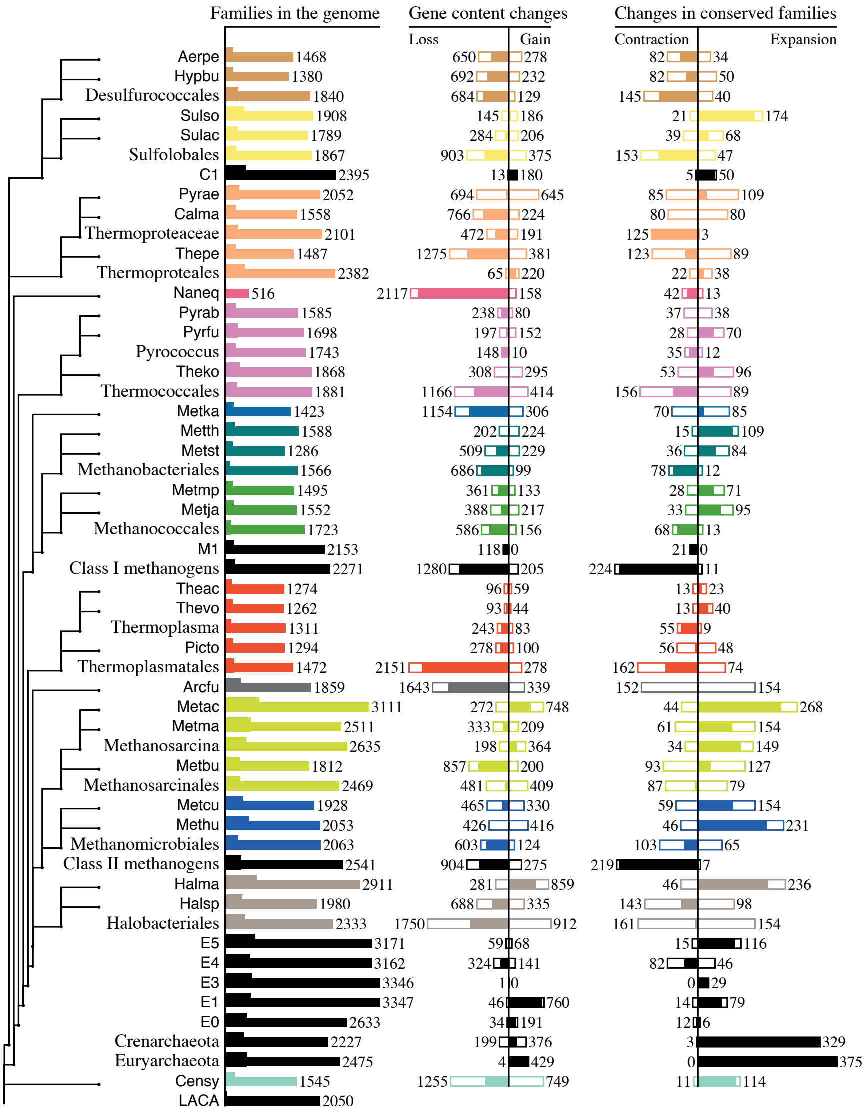
Families in the genome
Gene content changes
Loss
Gain
Changes in conserved families
Contraction
Expansion
Aerpe 1468 650 278 82 34
Hypbu 1380 692 232 82 50
Desulfurococcales 1840 684 129 145 40
Sulso 1908 145 186 21 174
Sulac 1789 284 206 39 68
Sulfolobales 1867 903 375 153 47
C1 2395 13 180 5 50
Pyrae 2052 694 645 85 109
Calma 1558 766 224 80 80
Thermoproteaceae 2101 472 191 125 3
Thepe 1487 1275 381 123 89
Thermoproteales 2382 65 220 22 38
Naneq 516 2117 158 42 13
Pyrab 1585 238 80 37 38
Pyrfu 1698 197 152 28 70
Pyrococcus 1743 148 10 35 12
Theko 1868 308 295 53 96
Thermococcales 1881 1166 414 156 89
Metka 1423 1154 306 70 85
Metth 1588 202 224 15 109
Metst 1286 509 229 36 84
Methanobacteriales 1566 686 99 78 12
Metmp 1495 361 133 28 71
Metja 1552 388 217 33 95
Methanococcales 1723 586 156 68 13
M1 2153 118 0 21 0
Class I methanogens 2271 1280 205 224 11
Theac 1274 96 59 13 23
Thevo 1262 93 44 13 40
Thermoplasma 1311 243 83 55 9
Picto 1294 278 100 56 48
Thermoplasmatales 1472 2151 278 162 74
Arcfu 1859 1643 339 152 154
Metac 3111 272 748 44 268
Metma 2511 333 209 61 154
Methanosarcina 2635 198 364 34 149
Metbu 1812 857 200 93 127
Methanosarcinales 2469 481 409 87 79
Metcu 1928 465 330 59 154
Methu 2053 426 416 46 231
Methanomicrobiales 2063 603 124 103 65
Class II methanogens 2541 904 275 219 7
Halma 2911 281 859 46 236
Halsp 1980 688 335 143 98
Halobacteriales 2333 1750 912 161 154
E5 3171 59 68 15 116
E4 3162 324 141 82 46
E3 3346 1 0 0 29
E1 3347 46 760 14 79
E0 2633 34 191 12 6
Crenarchaeota 2227 199 376 3 329
Euryarchaeota 2475 4 429 0 375
Censy 1545 1255 749 11 114
LACA 2050

Figure 3