Journal
of Molecular
Evolution

Evolution of HSP70 Gene and Its Implications Regarding Relationships Between Archaebacteria, Eubacteria, and Eukaryotes

Radhey S. Gupta,[1] G. Brian Golding[2]

[1] Department of Biochemistry, McMaster University, Hamilton, Canada L8N 3Z5
[2] Department of Biology, McMaster University, Hamilton, Canada L8N 3Z5

Received: 17 August 1992 / Accepted: 12 April 1993

Abstract. The 70-kDa heat-shock protein (HSP70) constitutes the most conserved protein present in all organisms that is known to date. Based on global alignment of HSP70 sequences from organisms representing all three domains, numerous sequence signatures that are specific for prokaryotic and eukaryotic homologs have been identified. HSP70s from the two archaebacterial species examined (viz., *Halobacterium marismortui* and *Methanosarcina mazei*) have been found to contain all eubacterial but no eukaryotic signature sequences. Based on several novel features of the HSP70 family of proteins (viz., presence of tandem repeats of a 9-amino-acid [a.a.] polypeptide sequence and structural similarity between the first and second quadrants of HSP70, homology of the N-terminal half of HSP70 to the bacterial MreB protein, presence of a conserved insert of 23–27 a.a. in all HSP70s except those from archaebacteria and gram-positive eubacteria) a model for the evolution of HSP70 gene from an early stage is proposed. The HSP70 homologs from archaebacteria and gram-positive bacteria lacking the insert in the N-terminal quadrants are indicated to be the ancestral form of the protein. Detailed phylogenetic analyses of HSP70 sequence data (viz., by bootstrap analyses, maximum parsimony, and maximum likelihood methods) provide evidence that archaebacteria are not monophyletic and show a close evolutionary linkage with the gram-positive eubacteria. These results do not support the traditional archaebacterial tree, where a close relationship between archaebacterial and eukaryotic homologs is observed. To explain the phylogenies based on HSP70 and other gene sequences, a model for the origin of eukaryotic cells involving fusion between archaebacteria and gram-negative eubacteria is proposed.

Key words: Heat shock protein HSP70—Phylogeny—Archaebacteria—Eubacteria—Eukaryotes

Introduction

The phylogenetic relationships between archaebacteria, eubacteria, and eukaryotes are of central importance to understanding the evolution of life. Based on sequence data for rRNA and a few other proteins (viz., elongation factors EF-1α, EF-2, RNA polymerase, and V- and F-type ATPases), all extant organisms have been grouped into three primary urkingdoms or domains—namely, the Archaea, the Bacteria, and the Eukarya (Woese 1987; Cedergren et al. 1988; Gogarten et al. 1989; Iwabe et al. 1989; Puhler et al. 1989; Woese et al. 1990). Each of these domains has been proposed to be monophyletic and distinct from the others (Woese et al. 1990). Although this view has gained wide acceptance, Lake and co-workers have argued that archaebacteria are not monophyletic, based on sequence characteristics of EF-1α and EF-2 as well as

Correspondence to: R.S. Gupta

other considerations (Lake 1988, 1991; Rivera and Lake 1992). Their recent results indicate that halobacteria and methanogenic archaebacteria show a closer relationship to the eubacteria, whereas extremely thermophilic archaebacteria (termed eocytes) bore sequence characteristics similar to the eukaryotic cells. The above classification has also been criticized on the grounds that "the difference in structural organization between prokaryotes and eukaryotes is an order of magnitude greater than the relatively small difference between the archaebacteria and the eubacteria" (Mayr 1990).

To obtain additional information pertinent to this question, we have analyzed sequence data for the 70-kDa heat-shock protein (HSP70) family from organisms representing all three domains. Members of the HSP70 family carry out a highly conserved molecular chaperone function in the intracellular transport of proteins and in protecting the organisms from thermal or other stress-induced damages (Lindquist and Craig 1988; Morimoto et al. 1990; Gething and Sambrook 1992). Although synthesis of some HSP70 homologs is induced by thermal or other stressors, they constitute essential and abundant proteins even in unstressed cells. In prokaryotic cells only a single HSP70 homolog (referred to as DnaK in *E. coli*) is generally found. In contrast, in eukaryotic cells, several distinct HSP70 homologs, many of which are localized in different intracellular compartments (viz., mitochondria, chloroplast, endoplasmic reticulum) have been identified (Lindquist and Craig 1988; Morimoto et al. 1990). In the past 8–10 years, due to the perceived importance of HSP70 in cell structure and function, the cDNA/genes for HSP70 have been sequenced from a large number of prokaryotic and eukaryotic species. These studies reveal that the primary structure of HSP70 is highly conserved during evolution. (See Lindquist and Craig 1988; Gupta and Singh 1992.)

In the present paper we have used HSP70 sequence data to examine the relationships between archaebacterial, eubacterial, and eukaryotic organisms. The sequence features and detailed phylogenetic analyses of HSP70 sequence data reveal novel and unexpected relationships between archaebacteria, eubacteria, and eukaryotic species, which are discussed in this paper.

Materials and Methods

Sequence Analysis. The amino acid (a.a.) sequence data for various HSP70 as well as other gene sequences were obtained from various protein and nucleic acid databases (viz., Genbank, Swiss-Prot. PIR, etc.). The sequences were initially aligned in small groups (of 10–25) employing the CLUSTAL program (Higgins and Sharp 1989; PCGene Software) or in pairs using the BESTFIT, PALIGN, and FASTA programs of the University of Wisconsin Genetics Computer Group (GCG6) program package (Pearson 1990). Based on this information, global alignment of all sequences was carried out manually, correcting for any obvious misalignments using a program written and compiled by Dr. A.L. Goldin, Dept. of Biology, California Inst. of Technology. It should be noted that about 60–100 residues in HSP70s near the C-terminal end show considerable variation among distant species. For lack of proper alignment, this region was not considered in phylogenetic analyses. The amino acid sequence identity between pairs of proteins was calculated using the PALIGN program (PCGene Software), using the structure gene matrix and unit gap and open gap costs of 1 and 7, respectively. The bootstrap neighbor-joining tree was obtained using the programs BOOT, NEIGHBOR, and CONSENSE from PHYLIP—version 3.3, program package (Felsenstein 1991). Parsimony analysis of sequences was carried out using the program PROTPARS from PHYLIP, version 3.4.

The significance of repeat polypeptide sequences to the consensus sequence was determined by comparing the alignment scores of the observed repeat with that of the randomized (1,000 times) HSP70 sequence, using PAM 250 substitution matrix (Altschul 1991). For all polypeptide repeats identified, the number of random HSP70 sequences showing higher alignment scores than observed was less than 2%.

Results

Global Sequence Alignment and Identification of Signature Sequences

A global alignment of HSP70 sequences from organisms comprising all three domains was initially carried out. The alignment consisted of a total of 66 sequences of which 57 were full length and 9 were partial. Of these sequences, 2 were from archaebacteria (a halophile and a methanogen); 11 were from the eubacterial group encompassing several gram-positive bacteria (both low G + C as well as high G + C species), chlamydiae, cyanobacteria, and purple bacteria; and 8 were from eukaryotic organelle (viz., mitochondria and chloroplast); and the remaining 45 sequences were for different eukaryotic HSP70 homologs covering all the major groups (viz., plants, animals, protists, algae, higher and lower fungi, etc.) of organisms. The observed sequence alignment was very similar to that reported previously for a limited number (12) of species. (See Fig. 4 of Gupta and Singh 1992.) Based on the alignment, numerous signature sequences were identified that provided clear distinction between prokaryotic and eukaryotic homologs. Some of these signature sequences are depicted in the partial sequence alignment shown in Fig. 1. The two archaebacterial HSP70s contain all of the eubacterial signatures but none specific for eukaryotic species. Further, no sequence features that were unique only to the archaebacteria could be inferred. (See Gupta and Singh 1992 for complete sequences.) However, several sequence features were found to be uniquely shared by archaebacteria and gram-positive eubacteria on one end, and the gram-negative eubacteria and eukaryotic HSP70 ho-

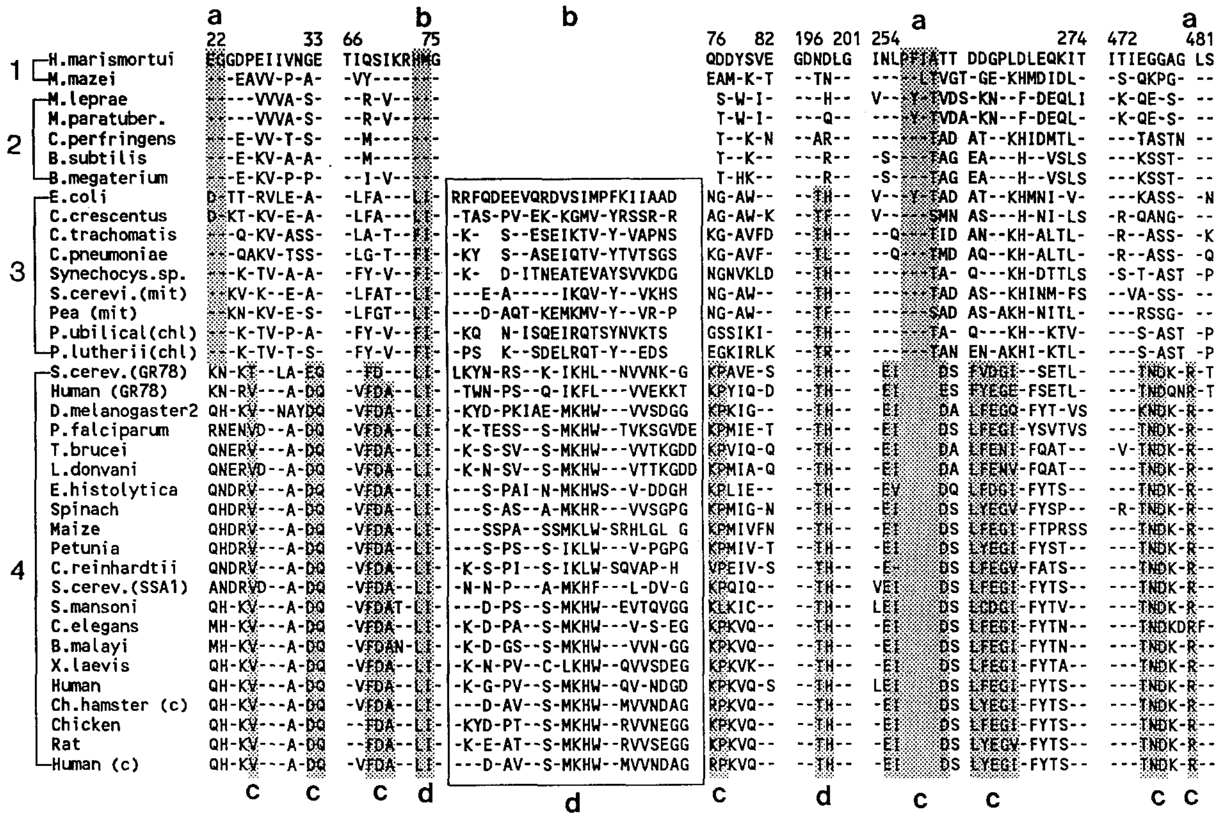


Fig. 1. Partial alignment of HSP70 sequences depicting some of the signature sequences characteristic of various prokaryotic and eukaryotic organisms. The numbers 1 to 4 on the left refer to sequences from archaebacteria (1), gram-positive bacteria (2), gram-negative bacteria and eukaryotic organeller sequences (3), and from a representative group of eukaryotic organisms (4). The numbers on the top refer to the position in *Halobacterium marismortui* sequence (Gupta and Singh 1992). The letters *a, b, c,* and *d* refer to signature sequences (*shaded regions*) that are distinctive of (a) prokaryotes, i.e., both archaebacteria and eubacteria, (b) archaebacteria and gram-positive bacteria, (c) eukaryotic species, and (d) common to gram-negative eukaryotic homologs. Not all signature sequences of the above kinds are shown. The *boxed region* shows the insert in the N-terminal quadrant. The *dashes* (-) indicate identity with a.a. in the top line. Other eukaryotic HSP70 homologs not shown also contained the indicated signature sequences. The species names are as follows: *H. marismortui, Halobacterium marismortui; M. mazei, Methanosarcina mazei; M. leprae, Mycobacterium leprae; M. tuberculosis, Mycobacterium tuberculosis; C. perfringens, Clostridium perfringens; B. subtilis, Bacillus subtilis; B. megaterium, Bacillus megaterium; E. coli, Escherichia coli; C. crescentus, Caulobacter crescentus; C. trachomatis, Chlamydiae trachomatis; C. pneumonial, Chlamydiae pneumonial; Synechocys* sp., *Synechocystis* species; *S. cerevi, Saccharomyces cerevisiae; P. ubilical, Porphyra ubilicalis; P. lutherii, Pavlova lutherii; D. melanogaster, Drosophila melanogaster; P. falciparum; Plasmodium falciparum; T. brucei, Trypanosoma brucei; L. donvani, Leischmania donvani; E. histolytica, Entamoeba histolytica; C. reinhardtii, Chlamydomonas reinhardtii; S. mansoni, Schistosoma mansoni; C. elegans, Caernorhabditis elegans; B. malayi, Brugia malayi; X. laevis, Xenopus laevis;* Ch. hamster, Chinese hamster. The notations (mit), (chl), and (c) in *parentheses* denote mitochondrial, chloroplast, or cognate forms of HSP70.

mologs on the other (signature sequences *b* and *d* in Fig. 1). The eukaryotic organeller HSP70 homologs (from mitochondria and chloroplasts) contained all of the signature sequences characteristic of gram-negative eubacteria.

Interdomain Sequence Similarity for HSP70 and Other Proteins. Global alignment of HSP70 sequences revealed that numerous long stretches of a.a. were highly conserved in this protein family (results not shown). To obtain a measure of sequence conservation in HSP70 family, as compared to other conserved proteins that are ubiquitously found (viz., elongation factor-1 and -2, glutamine synthetase, RNA polymerase, proton ATPase), we have determined interdomain a.a. identity scores for representative species from each of the three domains. These results are presented in Table 1. The minimum a.a. identity between any two HSP70 homologs from any of the three domains is about 42%. In contrast, for the other proteins which have been previously employed for deep phylogenetic analyses (Gogarten et al. 1989; Iwabe et al. 1989; Pühler et al. 1989; Pesole et al. 1991), the interdomain similarity shows greater variation between different domains and the minimum identity observed was found to be much lower. These results illustrate that HSP70 is the most conserved protein presently known that is found in all organisms.

Structural Features of HSP70 and a Model for the Evolution of HSP70 Gene/Protein. One striking

Table 1. Interdomain amino acid identity (%) for different proteins

	Eubacteria vs archaebacteria		Eubacteria vs eukaryotes		Archaebacteria vs eukaryotes	
	(a)	(b)	(c)	(d)	(e)	(f)
HSP70	50.3	57.4	46.4	47.6	42.0	47.2
EF-1 (Tu)	34.0	30.5	27.7	26.4	50.8	53.5
EF-2 (G)	26.9	26.9	22.2	23.4	33.1	30.4
ATPase	13.2	16.7	17.2	17.7	55.8	58.7
Glutamine synthetase	36.3	41.4	11.0	7.2	8.6	11.8
RNA polymerase	28.9	26.9	22.0	23.2	28.4	30.4

Amino acid identity between pairs of protein sequences was calculated by the PALIGN program of the PC GENE software. The letters *a–f* refer to the comparison between the following species representative of the domains: *(a) E. coli* vs a halobacterium; *(b) E. coli* vs a methanogen; *(c) E. coli* vs human; *(d) E. coli* vs *Saccharomyces cerevisiae; (e) Halobacteria* vs *S. cerevisiae;* and *(f)* methanogen vs *S. cevisiae.* For RNA polymerase, the identity values are taken from Pühler et al. (1989)

feature of the HSP70s from archaebacteria and the gram-positive eubacteria is the presence of a gap of 23–27 a.a. in the N-terminal quadrant (Fig. 1). This gap could result either from a deletion in HSP70 in these organisms or from an insertion in the other organisms. However, two different observations provide evidence in support of the latter view. First, we have observed that a.a. comprising the first quadrant of HSP70 from archaebacteria and gram-positive bacteria (i.e., a.a. 1–160) show significant similarity ($P < 0.001$) to those in the second quadrant (i.e., a.a. 161–320) (Gupta and Singh 1992). This similarity was not readily seen in HSP70s from other prokaryotic and eukaryotic organisms containing the additional 23– 27 a.a. in the N-terminal quadrant, indicating that these a.a. were probably inserted after the gene duplication event. The inference from sequence similarity regarding occurrence of gene duplication in the evolution of HSP70 is supported by the structural data on HSP70. The reported three-dimensional structure of the N-terminal fragment (386 a.a.) of bovine HSP70 shows the presence of two distinct lobes of approximately equal sizes and similar tertiary structures, with a deep cleft separating the two lobes (Flaherty et al. 1990). The boundaries of the two lobes, as determined from X-ray crystal structure data, show excellent correlation with the boundaries of the proposed duplicated segments (Gupta and Singh 1992). The 25 amino acids (from 82 to 106 in the bovine HSP70) corresponding to the insertion are located on the outside of lobe I and thus could have been acquired at a later time. Together, these results strongly indicate that the first two quadrants of HSP70 arose by duplication of an ancestral domain that probably lacked the insertion in the N-terminal quadrant.

A second, more compelling argument for the recent acquisition of the 23–27 a.a. in the N-terminal quadrant of HSP70s is based on the observed highly significant similarity between a protein, MreB from *E. coli* (as well as the gram-positive bacteria *Bacillus subtilis*—Doi et al. 1988; EMBL Database), and the N-terminal half of the HSP70 family of proteins (Gupta and Singh 1992). Similar to the N-terminal half of HSP70 family of proteins, MreB also contains the ancestral ATPase binding domain, and it is predicted to have similar tertiary structure to the N-terminal half of HSP70 (Bork et al. 1992). Since MreB is only about half the length of HSP70, it is very likely evolved from a predecessor of HSP70 before acquisition of the C-terminal fragment. The alignment of MreB protein with HSP70s from gram-positive bacteria (Fig. 2) or archaebacteria (Gupta and Singh 1992) required no large gaps. However, the alignment of MreB with the HSP70 homologs from either gram-negative eubacteria or eukaryotic species (Fig. 2) required introduction of a large gap in the MreB sequence, in the same position, where additional a.a. are found in these proteins. Absence of the 23–27-a.a. insert in the MreB protein strongly indicates that the ancestral HSP70 lacked this insert and that it has been introduced at a later time. The possibility that the MreB may be derived from an ancestral HSP70 by loss of the C-terminal fragment is considered unlikely because of the absence of the 23–27-a.a. insert in MreB from both gram-positive and gram-negative eubacteria.

Another novel feature of the HSP70 family of proteins that we have observed is the presence of a short polypeptide with the consensus sequence VDLGGGDFE that is repeated several times in the N-terminal half of the protein (Fig. 3). This polypeptide repeat initially came to our attention when examining degenerate oligonucleotide primers for conserved regions of HSP70 for cloning the HSP70 gene (Galley et al. 1992). It was observed that the primer sequence corresponding to one of the conserved regions in the sequence (viz., a.a. 170–178) showed significant similarity to other regions of HSP70. The visual inspection of HSP70 sequences then identified several stretches which were related to this sequence. All of the polypeptide repeats indicated in Fig. 3 have significant similarity to the

```
           1
HS70Ecol   M GKIIGIDLGTTNSCVAIMDGTTPRVLENAEGDRTTPSIIAYTQDGETLVGQPAKRQAVTNPQNTLFAIKRLIGRRFQDEEVQRDVSIMPFKIIAADNG
HS70Bsub   - S-V-------------VLE-GE-K-IA----N-----VV-FKN --RQ--EV----SI--- --IMS---HM-                        T
MreBBsub   -FARD-------A-  -L-HVKGKGI--NE       --VV-IDRNTG         KVLAVGEEARSMVG-TP-                       -I
MreBEcol   -FSNDLS-----A-  TL-YVKGQGI--NE       --VV-IR--RA     GSPKSVAAVGHDANEMLG-TP-                       -I

           DAWVEVKGQKMAPPQISAEVLLKMKKTAEDY LGEPVTEAVITVPAYFNDAQRQATKDAGRIAGLEVKRIINEPTAAALAYGLDKGTGNRTIAVYDLGGG
           -YK--IE-KDYT-QEV--II-QHL-RY--S- ---T-SK------------E--------K------E---------P-------TDEDQ--LL------
           V-IRPL-DGVI-DFE-TEAM-KYFINKLDVKSFF SKPRIL-CC-TNITSVEQK-IRE-AERS-GKTVFLEE--KV--VGA-ME IFQPSGNM-V-I---
           A-IRPM-DGVI-DFFVTEKM-QHFI-QVHSNSFMR-SPRVLVC--VGATQVE-R-IRESAQG--AREVFL-E--M---IGA--P VSEATGSM-V-I---

           TFDISIIEIDEVDGEKTFEVLATNGDTHLGGEDFDSRLINYLVEEFKKDQGIDLRNDPLAMQRLKEAAEKAKIELSSAQQTDVNLPYITADATGPKHMNI
           ---V--L-L    -DGV---RS-A--NR---D---QVI-DH--S----EN-V--SK-KM-L----D------KD--GVTS-QIS--F---GEA--L-LEV
           -T--AVLSM         GDIV-SSSIKMA-DK--MEIL--I          KRKYKL-IGE-TS-DI   --KVGTVFPGARSEELEIRGRDMVTGLPR
           -TEVAV-SL         NGVVYSSSVRI--DR--EAI---V          RRNYGS-IGEATA-RI   -H-IG--YPG-EVREIEVRGRNLAEGVPR

           KVTRAKLESLVED     LVNRSIEPLKVALQDAGLSVSDIDDVILVGGQTRMPMVQKKVAEFFGKEPRKDVNPDEAVAIGAAVQGGVLTGDVKDVLLLD 393
           SLS---FDE-SAG     --E-TMA-VRQ--K-----A-EL-K------S--I-A--DAIKKET-QD-H-G-----V--L---I------------V--- 363
           TI-VCSE-ITEALKENAAVIVQAAKGVLERTPPELSADIIDRG---T--GALLHGIDMLL--ELKVPVLIAE--MHC--V-TGIMLENIDRLP-RA-R   333
           GF-LNSN-I-EALQEPLIGIVSAVMVALEHTPPELA-DISERGMV-T--GALLRNLDRLLM-ET-IPVVVAED-LTC--R-GGKALEMIDMHGG-LFSEE 347
```

Fig. 2. Alignment of MreB protein sequences from *E. coli* and *B. subtilis* with the HSP70 sequences from the same species. MreB protein sequences are shown complete whereas for HSP70 sequences only the N-terminal portions which are homologous to MreB are shown. The *dashes* (-) denote residues identical to that shown on the top line. Gaps in sequences are shown by *blank spaces.*

9	I	D	L	G	T	T	N	S	A
81	V	E	L	D	G	E	E	Y	T
170	Y	D	L	G	G	G	T	F	D
182	L	D	L	G	G	G	V	Y	E
198	N	D	L	G	G	D	D	W	D
210	I	D	Y	L	A	D	E	F	E
271	Q	K	I	T	R	A	K	F	E
295	Q	A	L	A	D	A	D	Y	T
310	V	I	L	V	G	G	S	T	R
consensus	V I	D	L	G	G	G	D E	F	E

Fig. 3. Alignment of repeat polypeptide sequences present in the HSP70 from *E. marismortui.* The *numbers* refer to the position in *H. marismortui* sequences (Gupta and Singh 1992). All of the polypeptide repeats show significant similarity to the consensus sequence shown underneath.

consensus sequence (See Materials and Methods). It is noteworthy that many of the polypeptide repeats seen in HSP70 are located close to each other in the second quadrant. This observation suggests that an ancestral gene encoding the second quadrant of HSP70 may have evolved by tandem joining of the genes encoding this shorter polypeptide. This polypeptide sequence, because of its postulated primordial nature and high degree of conservation within HSP70 homologs, is likely to play an important role in HSP70 function.

Based upon the above observations, a model for the evolution of the HSP70 family of proteins is proposed (Fig. 4). Stages I–IV in this model are postulated to have occurred in the common ancestor to all life. In view of the known function of HSP70 in protecting the organisms from heat- or other stress (e.g., O_2 deprivation)-induced damages (see Lindquist and Craig 1988; Morimoto et al. 1990), the postulated very early evolution of this protein could have played an important role in the survival of the early life form in the primitive environment (e.g., high ambient temperature and deficient in oxygen). (See Kasting 1993.)

Phylogenetic Analyses on HSP70 Sequences. To deduce the evolutionary relationships of HSP70, all 57 sequences for which complete sequence information was available were initially analyzed using the neighbor-joining method of phylogenetic tree reconstruction (Saitou and Nei 1987). The resulting tree clearly indicated that the preferred topology branches the archaebacteria with the gram-positive bacteria (results not shown). The branching of archaebacteria within the gram-positive bacteria is an unusual result. To further investigate it, more extensive tests were performed on a subset of 18 species containing all known archaebacterial and eubacterial sequences and representative eukaryotic organisms. For these species the sequences were bootstrapped 100 times (Felsenstein 1985); for each a neighbor-joining tree was found and a consensus tree was obtained (Fig. 5a). The topology of this tree closely follows the topology of the tree found using all species. There is a clear distinction between eukaryotic and bacterial species (100 of 100) and an equally clear distinction between eukaryotic/gram-negative bacteria and archaebacteria/gram-positive bacteria (98 of 100). The placement of the archaebacteria within the gram-positive is again suggested to be polyphyletic. The species *Methanosarcina mazei* is reasonably often branched with *Bacillus subtilis* (92 of 100) but the affinities of *Halobacterium marismortui,* although within the gram-positive, are not clear. The chloroplast sequences branch with *Synechocystis* species and the mitochondrial sequences branch with *E. coli,* supporting the endosymbiotic origin of these organelles from these groups of eubacteria (Schwartz and Dayhoff 1978; Gray and Doolittle 1982; Gray 1989; Gupta et al. 1989).

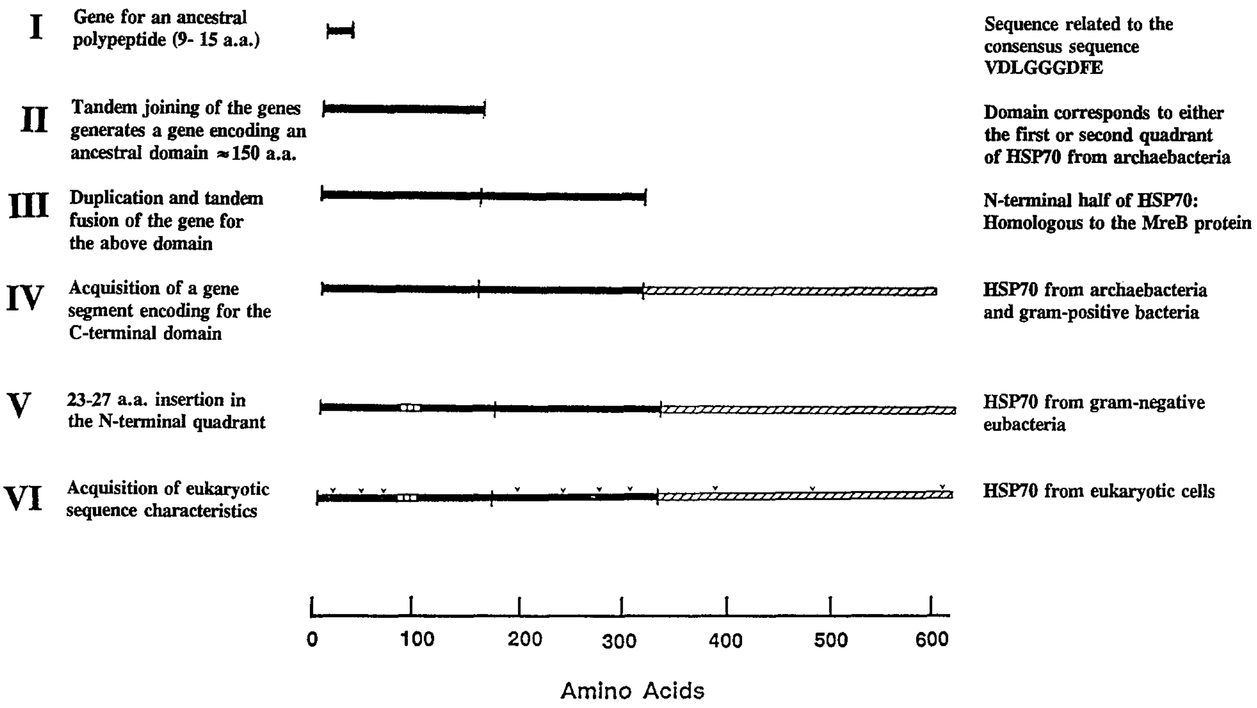

I
Gene for an ancestral polypeptide (9- 15 a.a.)
Sequence related to the consensus sequence VDLGGGDFE
II
Tandem joining of the genes generates a gene encoding an ancestral domain ≈150 a.a.
Domain corresponds to either the first or second quadrant of HSP70 from archaebacteria
III
Duplication and tandem fusion of the gene for the above domain
N-terminal half of HSP70: Homologous to the MreB protein
IV
Acquisition of a gene segment encoding for the C-terminal domain
HSP70 from archaebacteria and gram-positive bacteria
V
23-27 a.a. insertion in the N-terminal quadrant
HSP70 from gram-negative eubacteria
VI
Acquisition of eukaryotic sequence characteristics
HSP70 from eukaryotic cells
0
100
200
300
400
500
600
Amino Acids

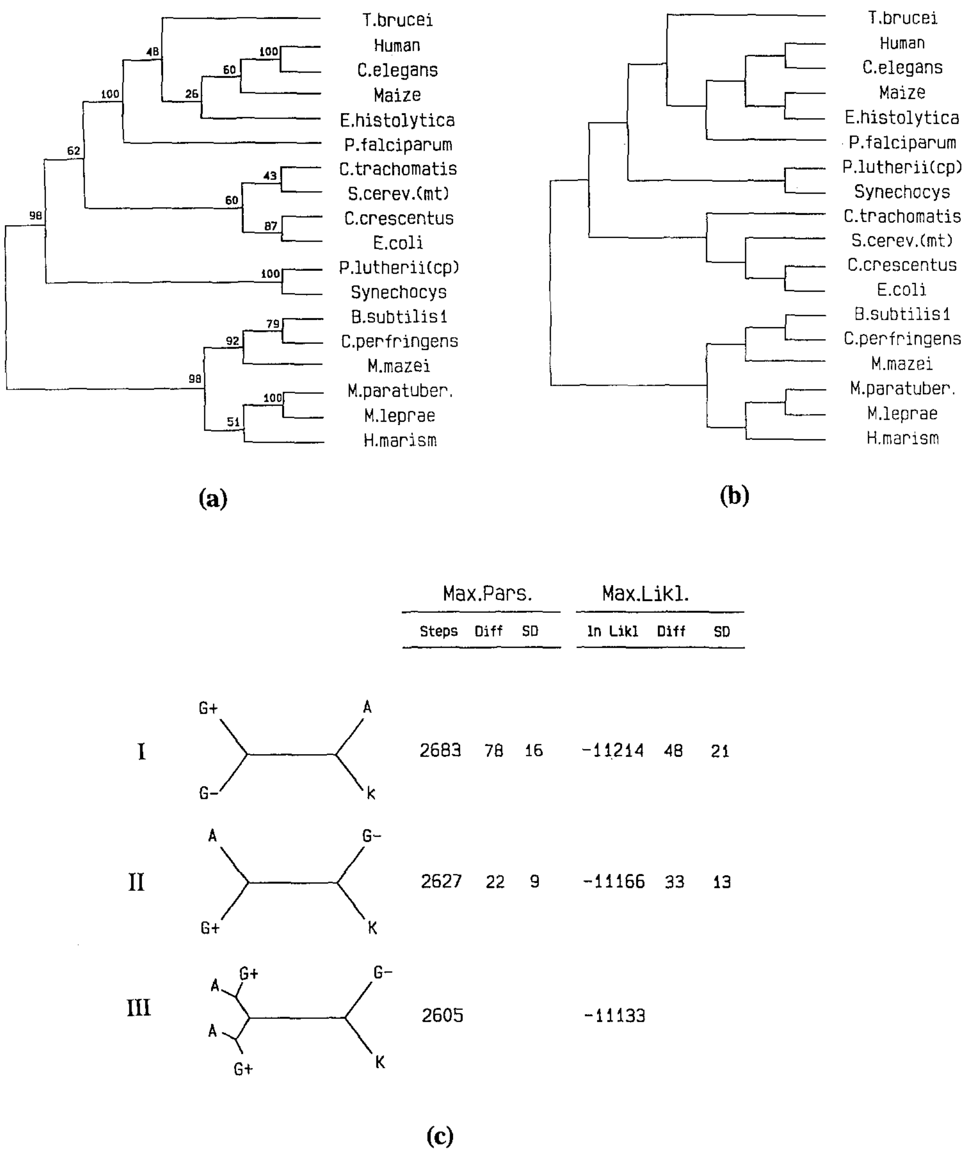
Fig. 4. A model showing various stages in the evolution of the HSP70 gene/protein. The main events characterizing different stages in the evolution are indicated on the left and some sequence features which support these events are noted on the right. A polypeptide such as that shown in stage I could also have been formed in the prebiotic world. (See Matthews 1971.)

The large insertion in the N-terminal quadrant of HSP70 is a very distinctive feature shared by the gram-negative bacteria and the eukaryotic species. To see if this insertion/deletion would affect the species placement within the phylogeny, this region was deleted from all sequences. A consensus bootstrap tree for this data is almost identical to that shown in Fig. 5a (not shown), and it provides evidence that a polyphyletic origin of the archaebacteria within the gram-positive bacteria is not dependent upon this sequence feature.

Parsimony analysis of this subset of species gives the tree shown in Fig. 5b. The parsimony tree is very similar to the bootstrap neighbor-joining tree and the placement of the archaebacteria within the gram-positive bacteria is repeated. Figure 5c diagrams three hypothetical relationships for the eukaryotes—gram-negative, gram-positive, and archaebacteria. The first shows the classical trifurcation with the archaebacteria monophyletic, and distinct from the eukaryotes and eubacteria (Woese 1987; Gogarten et al. 1989; Iwabe et al. 1989; Woese et al. 1990). A tree identical to Fig. 5b but with the archaebacteria moved to correspond with this hypothesis requires a total of 2,683 a.a. replacements. This second diagram shows the archaebacteria as monophyletic but branched within the eubacteria and with the gram-positive bacteria. This tree requires a total of 2,627 a.a. replacements and is a significant improvement over the number required for the classical trifurcation. The third diagram indicates the archaebacteria polyphyletic within the gram-positive bacteria (as per Fig. 5a,b). This topology requires a total of 2,605 a.a. replacements and again is a significant improvement over the number of changes required by either of the two other hypotheses. These trees were also evaluated using the maximum likelihood approach (Kishino et al. 1990). Again, the third tree topology yields the largest likelihood and it is a significant improvement over the other two topologies.

Discussion

The use of molecular sequence data to deduce deep phylogenetic relationships relies upon informational macromolecules which are ubiquitous and whose primary structure as well as function is highly conserved during evolution (Woese 1987, 1991). We have presented evidence that the HSP70 family of proteins constitutes the most conserved protein known, to date, that is found in all species. The high degree of sequence conservation of HSP70, in conjunction with its large size and an ancient conserved function, makes it a particularly useful system for investigating deep phylogenetic relationships. The utility of HSP70 for phylogenetic studies is further enhanced by the presence/identification of a large number of signature sequences that are either char-



	Max.Pars.			Max.Likl.		
	Steps	Diff	SD	ln Likl	Diff	SD
I	2683	78	16	-11214	48	21
II	2627	22	9	-11166	33	13
III	2605			-11133		

Fig. 5. Unrooted phylogenetic trees based on HSP70 sequences. **a** The tree shown is consensus neighbor-joining tree obtained after 100 bootstraps. **b** Parsimony tree on the same species. **c** Hypothetical trees considering different relationships between archaebacteria (A), gram-positive bacteria (G^+), gram-negative bacteria (G^-), and eukaryotes (K). The number of steps and the log likelihood are shown for each hypothetical tree. Differences between these trees and their standard deviations from the most parsimonious and likely tree.

acteristic of the main groups of organisms (viz., prokaryotes vs. eukaryotes; gram-positive bacteria vs. gram-negative bacteria) or are uniquely shared by some of them (common to gram-positive bacteria and archaebacteria; or between gram-negative bacteria and eukaryotes). Other unique structural features of HSP70 sequences (viz., presence of repeat polypeptide, internal duplication, homology to MreB protein) that have been identified provide information regarding evolution of this protein family from a very early stage. (See Fig. 4.)

Phylogenetic analyses of HSP70 sequences presented here strongly indicate a close evolutionary relationship between the gram-positive group of bacteria and the archaebacterial species. This inference is supported by sequence signatures shared uniquely by these two groups of organisms. Surprisingly, the two archaebacterial species examined do not form a coherent cluster within the gram-positive group of bacteria. These observations are inconsistent with the view that archaebacteria form a separate monophyletic domain. In addition to HSP70 sequence data, a linkage of archaebacteria with the gram-positive group of eubacteria is also indicated by a number of other gene sequences. Smith et al. (1992) have observed an association of halobacterial and methanogen species with the gram-positive eubacteria for the glutamine synthetase and superoxide dismutase gene sequences. However, these authors have attributed the observed linkage to

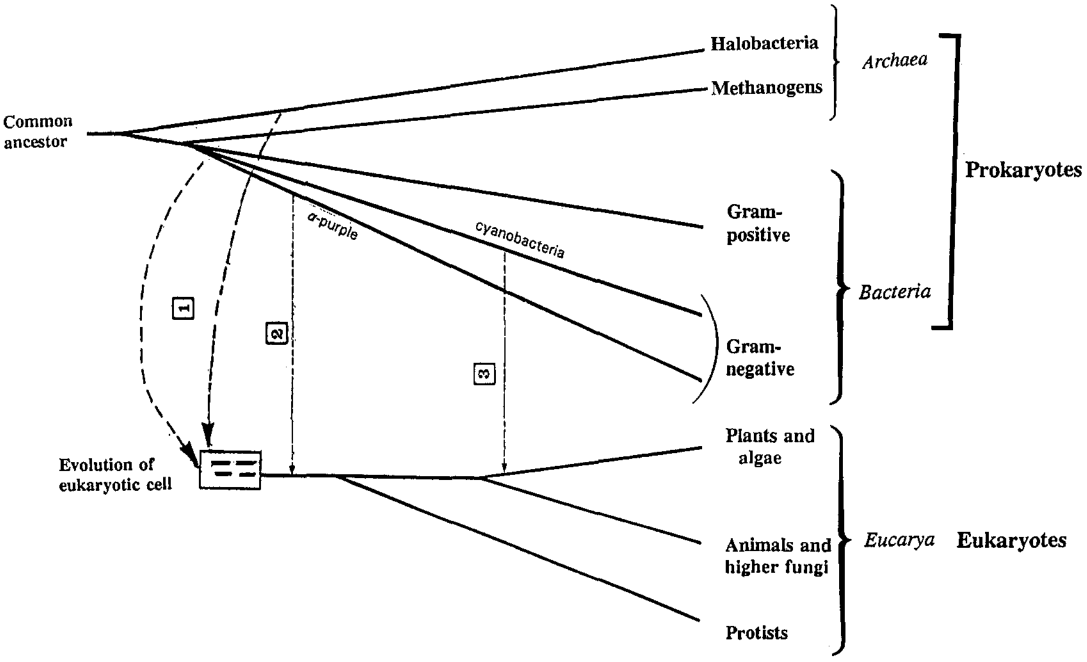


Fig. 6. Compositive hypothetical model for the origin of eukaryotic cells based on different genes. The root of the tree is placed within the archaebacteria and gram-positive eubacteria based on HSP70 sequences. The ancestral eukaryotic cell is proposed to arise (as shown by dotted line 1) by fusion of an archaebacteria and a gram-negative eubacteria. The box denotes an early stage in the evolution of the eukaryotic cell during which different genes from the two fusion partners were lost, accompanied by extensive changes in the genome. The numbers 2 and 3 refer to the endosymbiotic events leading to the genesis of mitochondria and chloroplasts, respectively.

horizontal gene transfer between species. We have independently obtained similar results for these molecules as well as glutamate dehydrogenase and aspartate aminotransferase sequences by parsimony and bootstrap neighbor-joining tree construction methods (R.S. Gupta and G.B. Golding work in progress). The gram-positive eubacteria and several of the archaebacteria also share a number of other unique characteristics including the presence of a thick homogeneous gram-positive staining cell wall. (See Cavalier-Smith 1987; Stanier et al. 1987.) The possibility that all such shared characteristics either at the gene or cellular levels are due to horizontal gene transfers between species appears overly simplistic. The possibility that a true linkage between these ancient prokaryotic species exists is currently being examined by us.

Another novel feature of HSP70 sequence data is that a relatively conserved insert is present in the same position in HSP70s from both various gram-negative eubacteria as well as *all* eukaryotic organisms. This raises the question regarding the origin of eukaryotic cells. Both these groups of HSP70s also share other sequence features that are unique to them. (See Fig. 1.) These observations and the fact that *all* eukaryotic homologs contain a large number of distinctive sequence features argue against the lateral transfer of HSP70 gene (or the insert) from gram-negative eubacteria to eukaryotic cells at a later stage in eukaryotic cell evolution. To explain the origin of eukaryotic cells, two different types of models have been previously proposed. The first type of model assumes progressive evolution of eukaryotic cell from a prokaryotic ancestor (Woese et al. 1990; Rivera and Lake 1992). Based on the sequence data for EF-1α, EF-2, F- and V-type ATPases, and RNA polymerase II and III subunits, where greater similarity between the archaebacterial and eukaryotic homologs is observed, it has been postulated that the eukaryotic cells evolved from an archaebacterial ancestor (Gogarten et al. 1989; Iwabe et al. 1989; Pühler et al. 1989). Two different phylogenetic trees have been proposed in this regard. The archaebacterial tree which postulates all archaebacteria to be monophyletic and archaebacteria to be more closely related to eukaryotic cells (Iwabe et al. 1989; Woese et al. 1990) is clearly not supported by the present data. The alternate eocyte tree implies a polyphyletic nature of the Archaea domain, and of these, it postulates that extreme thermophiles (eocytes) are the closest relatives of eukaryotes (Lake 1991; Rivera and Lake 1992). Due to the lack of HSP70 sequence data on eocytes, the validity of this tree cannot be ascertained at present. However, for the eocyte tree to explain the observed results it would require that the eocytes HSP70 sequences should be more akin to gram-negative eubacteria than to the halobacteria and methanogens, which is possible but considered less likely. The second type of model postulates eukaryotic cells to be a chimera made by fusion of two or more different cells (Sogin 1991; Zillig 1991). Since eukaryotic cells bear some characteristics of archaebacteria and others of eubacteria, Zillig (1991) has proposed that the genesis of eukaryotic cell involved fusion between an archaebacteria and eubacteria.

An extension of Zillig's model is presented in Fig. 6. The root of the tree in this case has been placed within the ancestral lineage of the archaebacteria and gram-positive group of organisms based on the observations discussed previously. (See Fig. 4.) Further, based on HSP70 sequence features, we postulate that one of the partners involved in the primary fusion event was a gram-negative eubacteria. Following the primary fusion, the early stages in the development of eukaryotic cells (denoted by the box) remain poorly understood. However, for all genes examined, eukaryotic cells contain homologs which show greater resemblance to either archaebacterial counterparts (viz., EF-1, EF-2, F- and V-type ATPase, RNA polymerase II and III) or to the eubacterial homolog (viz., HSP70). (Organeller genes which presumably were acquired by later endosymbiotic events are not considered in this regard.) To account for this, we postulate that a selection or assortment of genes from the fusion partners took place in the early eukaryotic cell(s) and the genes which were not incorporated were either gradually lost or were so drastically altered that no similarity to the original gene(s) is presently seen. The early phase in the evolution of eukaryotic cells may have had a markedly enhanced genetic activity, during which numerous changes in different genes (viz., base substitutions, additions, and deletions) which distinguish eukaryotic homologs from their prokaryotic counterparts were introduced. During this early phase other structural features characteristic of eukaryotes (e.g., nuclear membrane, cytoskeleton, etc.) also evolved by mechanisms that are not understood at present (Cavalier-Smith 1987). The model proposed here provides a conceptual framework for understanding the various contradictory phylogenetic and morphological observations. It depicts a closer relationship between the prokaryotic species (i.e., archaebacteria and eubacteria) as opposed to their relationship with the eukaryotic organisms, which both in morphological terms as well as intuitively is more appealing.

Acknowledgments. The work was supported by a research grant from the Medical Research Council of Canada to RSG. We thank Barbara Sweet for secretarial assistance.

References

Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 219:555–565

Bork P, Sander C, Valencia A (1992) An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin and hsp70 heat shock proteins. Proc Natl Acad Sci USA 89:7290–7294

Cavalier-Smith T (1987) The origin of eukaryote and archaebacterial cells. Ann NY Acad Sci 503:17–71

Cedergren R, Gray MW, Abel Y, Sankoff D (1988) The evolutionary relationships among known life forms. J Mol Evol 38:98–112

Doi M, Wachi M, Ishino F, Tomioka S, Ito M, Sakagami Y, Suzuki A, Matsuhashi M (1988) Determinations of the DNA sequence of the *mre*B gene and of the gene products of the *mre* region that function in the formation of the rodshape of *Escherichia coli* cells. J Bacteriol 170:4619–4624

Felsenstein J (1985) Confidence limits of phylogenies: An approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1991) PHYLIP manual, ver. 303. Herbarium. University of California, Berkeley

Flaherty KM, DeLuca-Flaherty C, McKay DB (1990) Three-dimensional structure of the ATPase fragment of a 79 K heat shock cognate protein. Nature 346:623–628

Galley KA, Singh B, Gupta RS (1992) Cloning of HSP70 gene from *Clostridium perfringens* using a general polymerase chain reaction based approach. Biochim Biophys Acta 1130: 203–208

Gething MJ, Sambrook JP (1992) Protein folding in the cell. Nature 355:33–45

Gogarten JP, Kibak H, Dittrich P, Taiz I, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M (1989) Evolution of the vacuolar H+-ATPases: Implications for the origin of eukaryotes. Proc Natl Acad Sci USA 86:6661–6665

Gray MW (1989) The evolutionary origin of organelles. Trends Genet 5:294–299

Gray MW, Doolittle WF (1982) Has the endosymbiont hypothesis been proven? Microbiol Rev 46:1–42

Gupta RS, Picketts DJ, Ahmad S (1989) A novel ubiquitous protein "chaperonin" supports the endosymbiotic origin of mitochondrion and plant chloroplast. Biochem Biophys Res Commun 163:780–787

Gupta RS, Singh B (1992) Cloning of HSP70 gene from *Halobacterium marismortui:* Relatedness of archaebacterial HSP70 to its eubacterial homolog and a model for the evolution of the HSP70 gene. J Bacteriol 174:4594–4605

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. Gene 73:237–244

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacterial and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Kasting JF (1993) Earth's early atmosphere. Science 259:920–926

Kishino H, Miyata T, Hasegawa H (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 30:151–160

Lake JA (1991) Tracing origins with molecular sequences: metazoan and eukaryotic beginnings. Trends Biochem Sci 16:46–50

Lake JA (1988) Origin of eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 331:184–186

Lindquist S, Craig EA (1988) The heat shock proteins. Annu Rev Genet 22:631–677

Matthews CN (1971) The origin of proteins: Heteropolypeptides from hydrogen cyanide and water. In: Buvet R, Ponnamperuma C (eds) Chemical evolution and the origin of life. North Holland, Amsterdam, pp 231–235

Mayr E (1990) A natural system of organisms. Nature 348:491

Morimoto RI, Tissières A, Georgopoulos C (1990) The stress response, function of the proteins, and perspectives. Stress Proteins Bio Med 1:1–35

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Pesole G, Bozzeti MP, Lanave C, Preparata G, Saccone C (1991) Glutamine synthetase gene evolution: A good molecular clock. Proc Natl Acad Sci USA 88:522–526

Pühler G, Leffersy H, Gropp F, Palm P, Klenk H-P, Lottspeich F, Garrett RA, Zillig W (1989) Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. Proc Natl Acad Sci USA 86: 4569–4573

Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76

Saitou N, Nei M (1987) The neighbor joining method: A new method of reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Schwartz RM, Dayhoff MO (1978) Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts. A perspective is derived from protein and nucleic acid sequence data. Science 199:395–403

Smith MW, Feng D-F, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfer. Trends Biochem Sci 17:489–493

Sogin ML (1991) Early evolution and the origin of eukaryotes. Curr Opinion Genet Dev 1:457–463

Stanier RY, Ingraham JL, Wheelis ML, Painter PR (1987) The archaebacteria. In: General microbiology, 5th ed. Macmillan, London, pp 330–343

Woese CR (1987) Bacterial evolution. Microbiol Reviews 51:221–271

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

Zillig W (1991) Comparative biochemistry of *Archae* and *Bacteria*. Curr Opinion Genet Dev 1:544–551