Research

Open Access

The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism

L Aravind, Lakshminarayan M Iyer and Vivek Anantharaman

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

Correspondence: L Aravind. E-mail: aravind@ncbi.nlm.nih.gov

Published: 8 September 2003

Genome ***Biology*** 2003, **4:**R64

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/10/R64

Received: 30 May 2003
Revised: 24 July 2003
Accepted: 31 July 2003

Abstract

Background: There is considerable heterogeneity in the phyletic patterns of major chromosomal DNA-binding proteins in archaea. Alba is a well-characterized chromosomal protein from the crenarchaeal genus *Sulfolobus*. While Alba has been detected in most archaea and some eukaryotic taxa, its exact functions in these taxa are not clear. Here we use comparative genomics and sequence profile analysis to predict potential alternative functions of the Alba proteins.

Results: Using sequence-profile searches, we were able to unify the Alba proteins with RNase P/MRP subunit Rpp20/Pop7, human RNase P subunit Rpp25, and the ciliate Mdp2 protein, which is implicated in macronuclear development. The Alba superfamily contains two eukaryote-specific families and one archaeal family. We present different lines of evidence to show that both eukaryotic families perform functions related to RNA metabolism. Several members of one of the eukaryotic families, typified by Mdp2, are combined in the same polypeptide with RNA-binding RGG repeats. We also investigated the relationships of the unified Alba superfamily within the ancient RNA-binding IF3-C fold, and show that it is most closely related to other RNA-binding members of this fold, such as the YhbY and IF3-C superfamilies. Based on phyletic patterns and the principle of phylogenetic bracketing, we predict that at least some of the archaeal members may also possess a role in RNA metabolism.

Conclusions: The Alba superfamily proteins appear to have originated as RNA-binding proteins which formed various ribonucleoprotein complexes, probably including RNase P. It was recruited as a chromosomal protein possibly only within the crenarchaeal lineage. The evolutionary connections reported here suggest how a diversity of functions based on a common biochemical basis emerged in proteins of the Alba superfamily.

Background

In all three superkingdoms of life, DNA is packaged into compact structures by proteins [1]. While there are some general similarities in the organization of the chromosomal DNA-protein complexes in all three superkingdoms, there are radical differences in the proteins that are actually involved in DNA-packaging. In bacteria, the proteins of the Hu/IHF family are the dominant chromosomal proteins and they assume

a distinctive three-dimensional fold with long β-sheets, which has thus far not been encountered in any other proteins [2]. In euryarchaea and the eukaryotes, proteins containing the α-helical histone fold are the predominant DNA-packaging components of chromatin. Four or eight copies of the histone fold domain oligomerize to form higher-order structures such as the tetrasome and the octamer, which act as the principal protein units of chromatin in eukaryotes and some archaea [3-5]. The crenarchaea lack histones, but an alternative chromosomal protein, Sul7d, which contains a fold similar to the chromodomain of eukaryotic chromosomal proteins, has been characterized in the crenarchaeal genus *Sulfolobus* [6,7]. However, homologs of this protein are not known from other crenarchaeal genera, suggesting that the dominant chromosomal protein of the crenarchaea may be vary between different genera.

More recently, a novel chromosomal protein, termed Alba, has been characterized in *Sulfolobus*, and was shown to coat DNA densely without compacting it [8-10]. It was demonstrated that Alba is acetylated on a lysine residue and deacetylated by Sir2, in a manner reminiscent of the deacetylation of the eukaryotic chromosomal proteins [8,11]. Structural studies on Alba showed that it shared a common fold with the DNase I and the carboxy-terminal domain of the translation initiation factor IF3 [12], suggesting its origin within a group of ancient nucleic acid-binding domains. Homologs of Alba were also detected in most other archaea and certain eukaryotes [10,12] suggesting that it was an ancient protein that originated prior to the divergence of the conserved genomic cores of the eukaryotes, crenarchaea and euryarchaea from their common ancestor. The Alba homolog from the euryarchaeon, *Thermococcus zilligi*, is depleted in stationary phase along with the histones, but it is not known if it physically associates with chromatin in this organism [13]. While Alba is a dominant chromosomal protein in certain crenarchaea, studies on the chromatin of eukaryotes and euryarchaea have not detected the Alba homologs as major constituents of chromosomes in these lineages [5,14]. This suggests that the highly-conserved Alba homologs, especially those from organisms where there is no evidence for a major chromosomal role, may have additional, as yet undiscovered, functions.

In this work, we investigate the evolutionary history and the functions of Alba homologs using comparative genomics and sequence profile analysis. As a result of our analysis, we were able to unify the Alba proteins with conserved protein subunits of eukaryotic RNase P and RNase MRP. We also show that the Alba proteins are derived within the ancient RNA-binding IF3-C fold and are probably closest to the YhbY superfamily of proteins, which is also involved in RNA metabolism. We present evidence that the Alba superfamily probably originated as an RNA-binding protein which formed complexes with various small RNAs, and was recruited as a chromosomal protein only, perhaps, within the crenarchaeal lineage. This evolutionary connection with proteins involved in ribonucleoprotein metabolism also throws light on the potential functions of the eukaryotic members of the Alba family.

Results and discussion

Characterization of the Alba superfamily and its relatives

To gain a better understanding of the functional diversity and provenance of the Alba protein, we initiated sequence profile searches using the PSI-BLAST program (e-value threshold for inclusion in profile was set at 0.01 for each iteration) and various archaeal homologs of the *Sulfolobus solfataricus* Alba (gi: 15897841). In the initial iterations, these searches recovered homologs from nearly all archaeal genera with completely-sequenced genomes, except *Halobacterium* and *Methanosarcina*, as well as multiple homologs from the eukaryotes such as Kinetoplastids (*Leishmania*), Apicomplexans (*Plasmodium*), plants (*Arabidopsis* and rice) and animals. The searches also recovered the conserved Rpp20p/Pop7p proteins from various eukaryotes: for example, the human Pop7 was recovered in the iteration 6 with an e-value = 10^{-5} in a search initiated with the *Archaeoglobus* protein AF1956. The Pop7/Rpp20 proteins have been demonstrated to function as subunits of two closely-related ribonucleoprotein complexes, namely RNase P and RNase MRP, which process tRNA and rRNA, respectively, in vertebrates and fungi [15-17]. Furthermore, the searches recovered a second subunit of RNase P, namely Rpp25, from vertebrates [18], as well as the RPP25-related protein Mdp2 from the ciliate *Stylonychia lemnae*, which has been shown to participate in the process of macronuclear development [19]. Reciprocal PSI-BLAST searches, seeded with various new members that were detected in these searches, detected the archaeal Alba proteins with significant e-values and recovered approximately the same set of proteins. For example, a search initiated with human Rpp25 (gi: 12803357) recovered the archaeal Alba orthologs with significant e-values (e = 10^{-4}-10^{-7}, iteration 2-3). The pairwise alignment in each of these cases spanned the entire length of the minimal archaeal Alba protein, which has been shown to comprise a single domain [12]. These observations suggested that all these proteins from diverse organisms, including the RNase P/MRP subunits Pop7/Rpp20 and Rpp25, Mdp2 and Alba define an evolutionarily-related superfamily of protein domains. Hereinafter, we refer to this superfamily of domains as the Alba superfamily. Whilst most of the eukaryotic proteins detected in these searches are larger than the archaeal Alba proteins, analysis of their composition using the SEG program suggests that the Alba domain of approximately 95 residues is the only globular domain in these larger homologs.

When the multiple alignment of the Alba superfamily is superimposed on the *Sulfolobus solfataricus* Alba crystal structure (PDB: 1h0x), a number of defining features that are

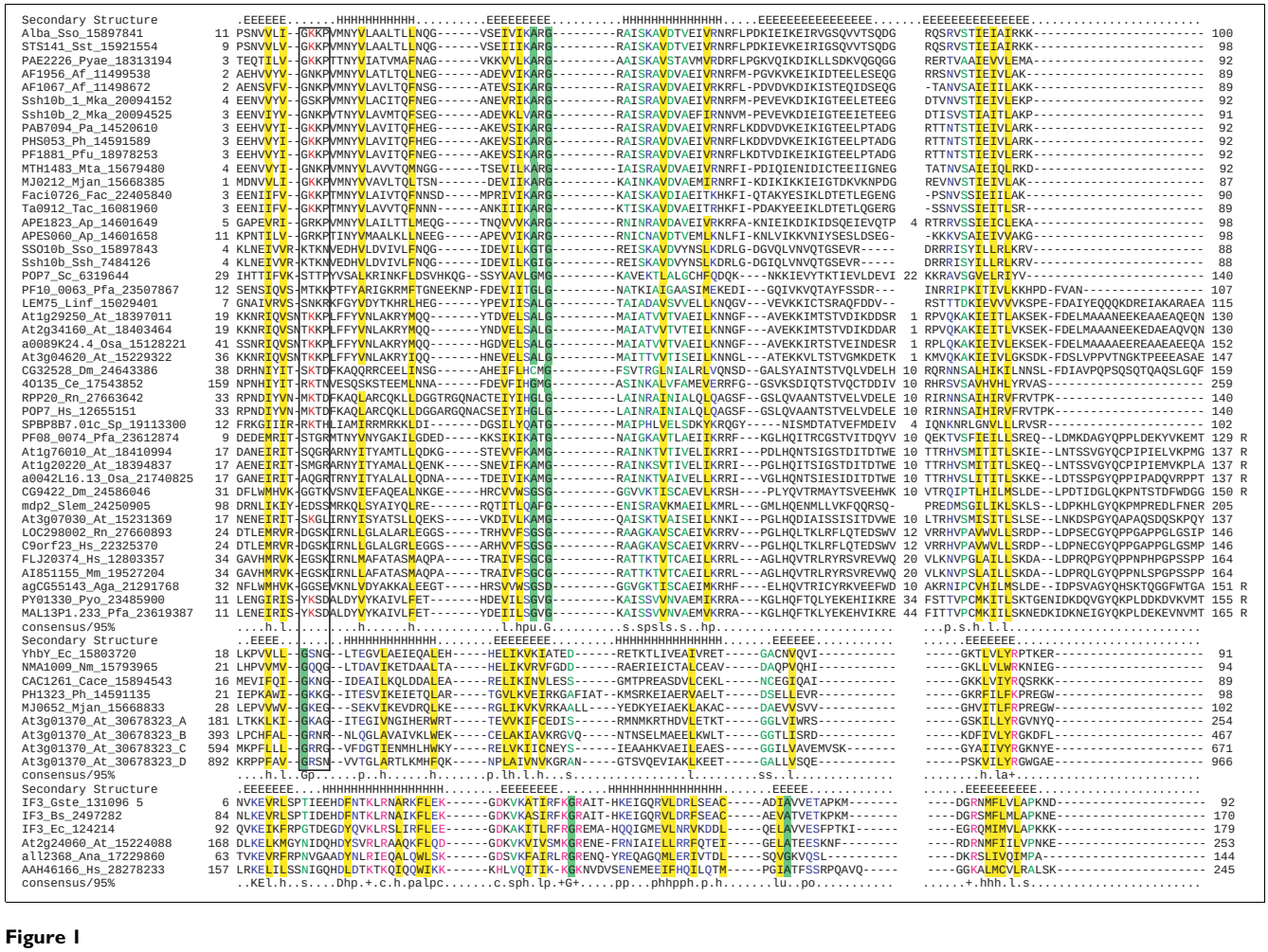

Figure 1
Multiple sequence alignment of the Alba family was constructed using the T-Coffee program after parsing high-scoring pairs from PSI-BLAST search results. The alignment of the Alba superfamily is superimposed on the structure-based alignments with members of the IF3-C and YhbY superfamilies. The secondary structure, derived from the representatives of these superfamilies with available structures, is shown above the alignment, with E representing a β strand and H an α-helix. The box shows the shared loop, which is typically bounded by small residues in Alba and YhbY. The 95% consensus shown below each family was derived using the following amino acid classes: hydrophobic (h: ALICVMYFW, yellow shading) and their aliphatic subset (l: ILV, yellow shading); aromatic (a: FHWY, yellow shading); small (s: ACDGNPSTV, green) and their tiny subset (u: GAS, green shading); charged (c: DEHKR, pink) and their basic subset (+: HKR, pink) and acidic subset (-: DE, pink); and polar (p: CDEHKNQRST, blue) and their alcoholic subset (o: ST, blue). A 'G', 'K' or 'E' shows the completely conserved amino acid in that group. The position corresponding to the lysine, which is acetylated in Alba, is shown in red. The limits of the domains are indicated by the residue positions, on each end of the sequence. The 'R' to the right of the sequence denotes those members that have the RNA-binding 'RGG' repeats at the C terminus. The numbers within the alignment are non-conserved inserts that have not been shown. The sequences are denoted by their gene name followed by the species abbreviation and GeneBank Identifier. Species abbreviations are as follows: Af, *Archaeoglobus fulgidus*; Aga, *Anopheles gambiae*; Ana, *Anabaena* sp.; Ap, *Aeropyrum pernix*; At, *Arabidopsis thaliana*; Bsub, *Bacillus subtilis*; Cace, *Clostridium acetobutylicum*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Ec, *Escherichia coli*; Fac, *Ferroplasma acidarmanus*; Gste, *Geobacillus stearothermophilus*; Hs, *Homo sapiens*; Linf, *Leishmania infantum*; Mjan, *Methanococcus jannaschii*; Mka, *Methanopyrus kandleri*; Mm, *Mus musculus*; Mta, *Methanothermobacter thermautotrophicus*; Nm, *Neisseria meningitidis*; Osa, *Oryza sativa*; Pa, *Pyrococcus abyssi*; Pfa, *Plasmodium falciparum*; Pfu, *Pyrococcus furiosus*; Ph, *Pyrococcus horikoshii*; Pyae, *Pyrobaculum aerophilum*; Pyo, *Plasmodium yoelii*; Rn, *Rattus norvegicus*; Sc, *Saccharomyces cerevisiae*; Slem, *Stylonychia lemnae*; Sp, *Schizosaccharomyces pombe*; Ssh, *Sulfolobus shibatae*; Sso, *Sulfolobus solfataricus*; Sst, *Sulfolobus tokodaii*; Tac, *Thermoplasma acidophilum*.

typical of this superfamily became apparent. These include a characteristic pattern of charged or polar residues associated with the strand-2 and the helix-2 of the domain (Figure 1). Both the eukaryotic and archaeal Alba domains possess a positive charge density of about 15-16% and share several fairly conserved charged positions (Figure 1). The region between the peculiarly elongated strands 3 and 4 in the carboxy-terminal hairpin is fairly variable: some members of the family possess short strands, while others possess low-complexity inserts of variable lengths. The majority of archaeal members of the superfamily are clearly distinguished by certain sequence motifs, such as a GXKP signature in the loop between strand-1 and helix-1 and the [KR]AVD signature in helix-2 (in the single-letter amino-acid code). Several eukaryotic and few archaeal members lack the lysine (K16), which is acetylated in the *Sulfolobus* Alba (gi: 15897841, Figure 1),

suggesting that this modification does not have a universal role in regulating the functions of this family. The residues associated with the dimer interface are also fairly well conserved throughout the superfamily implying a conserved dimeric quaternary structure. On the whole, despite differences between the various groups of the Alba family, the high overall conservation, similar charge densities and similar length of the cores of the secondary structure elements suggest that the basic biochemical properties are likely to be conserved throughout this superfamily.

To understand further the evolutionary relationships within the Alba superfamily, we constructed phylogenetic trees with the alignment of the Alba domain, using the neighbor-joining, maximum likelihood and Bayesian posterior probabilities methods. This phylogenetic analysis consistently revealed the presence of two major eukaryotic families and one archaeal family within the Alba superfamily (Figure 2a). Within the archaeal Alba family there appear to have been lineage-specific duplications in *Methanopyrus*, *Sulfolobus*, *Aeropyrum* and *Archaeoglobus*, but it is not clear whether this represents functional redundancy or specialization. One of the eukaryotic families contains Pop7/Rpp20 orthologs from yeasts, vertebrates, insects and *Caenorhabditis elegans* and related proteins from plants and protists such as *Leishmania* and *Plasmodium*. Based on the preferential grouping of the above-mentioned plant and protist proteins with the animal and fungal Pop7/Rpp20 proteins, we propose that these proteins are likely to possess the same function. Most members of this eukaryotic family are typified by a synapomorphic carboxy-terminal extension with a FDxh signature (where h is a hydrophobic residue, typically L; Figure 1).

The second eukaryotic family contains the ciliate Mdp2 protein, several animal proteins including RNase P subunit, Rpp25 and uncharacterized proteins from plants and *Plasmodium*. These proteins are unified by a synapomorphic carboxy-terminal extension, which contains a characteristic GYQXP signature. Several members of this family also have long carboxy-terminal tails with repeats of the RGG motif (see below). Members of this family have been either lost secondarily or diverged beyond recognition in the fungi and the nematode *C. elegans*. This phyletic pattern may reflect a certain degree of functional redundancy between the two eukaryotic families of the Alba superfamily. The phyletic pattern and phylogenetic tree topology suggest that the Alba protein was present in the common ancestor of the eukaryotes and archaea and subsequently duplicated in two paralogous lineages, typified by Rpp20 and Mdp2/Rpp25, very early in eukaryotic evolution. The average sequence divergence within the eukaryotic families of the Alba superfamily (mean p-distance = 0.75 for eukaryotic sequences; 0.73 within Rpp20 family and 0.65 within Mdp2 family) was much higher than that of the archaeal Alba family (mean p-distance = 0.45). While there could be a number of reasons for the accelerated divergence seen in eukaryotes, the functional diversification of the RNase P complex into the related RNase P and RNase MRP complexes, and a degree of relaxation of constraints due to functional overlap between the two eukaryotic families, could be two of the plausible explanations.

As Alba is likely to have been present in the last common ancestor of the eukaryotic and all archaeal lineages, we were interested in investigating its deeper relationships with more ancient proteins that could be traced back to the last universal common ancestor (LUCA) of the extant superkingdoms of life. Analysis of the structure of the Alba protein had shown that it possessed the IF3 carboxy-terminal domain (IF3-C) fold [12]. This fold is also present in DNase I [12], in the ribosomal protein S8, RNA 3'-terminal phosphate cyclase, R3H [20], in the carboxy-terminal domain of archaeo-eukaryotic prolyl tRNA synthetases (Pro-TRSC) [21] and in the RNA binding proteins of the YhbY family (Figures 2b and 3). In our earlier studies we had shown that the ancient RNA-binding domain, the thiouridine synthase, methylase, pseudouridine synthase (THUMP) domain, also belongs to this fold [22]. A survey of the functions of the conserved superfamilies containing the IF3-C fold indicates that the predominant theme is involvement in nucleic acid interactions, in particular RNA metabolism (Figure 2c). Earlier, based on phyletic patterns

Figure 2 *(continued on the next page)*
Phylogenetic trees of the Alba and YhbY superfamilies and an evolutionary scheme of the various superfamilies within the IF3-C fold. **(a)** Phylogenetic tree of the Alba superfamily. The tree shown here was constructed using the maximum-likelihood optimization as described in the Materials and methods section. The black circles indicate nodes with Rell-BP support of 80% or greater. The proteins are named as described in the legend to Figure 1. The divergent versions of Sso10b from the archaeal family of the Alba superfamily are indicated by a red circle. **(b)** Phylogenetic tree of the YhbY superfamily. Species abbreviations are as in the legend to Figure 1; additional species abbreviations are as follows: Blo, *Bifidobacterium longum*; Cau, *Chloroflexus aurantiacus*; Fnu, *Fusobacterium nucleatum*; Gme, *Geobacter metallireducens*; Hi, *Haemophilus influenzae*; Hsp, *Halobacterium* sp.; Lla, *Lactococcus lactis*; Pae, *Pseudomonas aeruginosa*; Spy, *Streptococcus pyogenes*; Vch, *Vibrio cholerae*; Xfa, *Xylella fastidiosa*. **(c)** An evolutionary scheme of the various superfamilies within the IF3-C fold. The horizontal lines indicate temporal epochs corresponding to certain major transitions in evolution, such as the last common ancestor of extant cellular life forms (LUCA), the divergence between the archaeo-eukaryotic lineage and the bacterial lineage and, finally, the emergence of the extant eukaryotes. The known or clearly predicted biochemical functions of the various superfamilies of this fold and their phyletic patterns have been indicated along with their names. R stands for RNA binding and D for DNA binding. The '>' as in B>E in the phyletic patterns, indicates an ancient transfer from bacteria to eukaryotes. The overall topology of the phylogram was derived using clustering based on DALI Z-scores, and specific shared derived characters. The YhhP family was first identified and predicted to function as a redox regulator in reference [52]. The red lineages are those which can be confidently traced to the LUCA, the black lineages are exclusively archaeo-eukaryotic and the green lineages are mainly bacterial with transfers to eukaryotes.

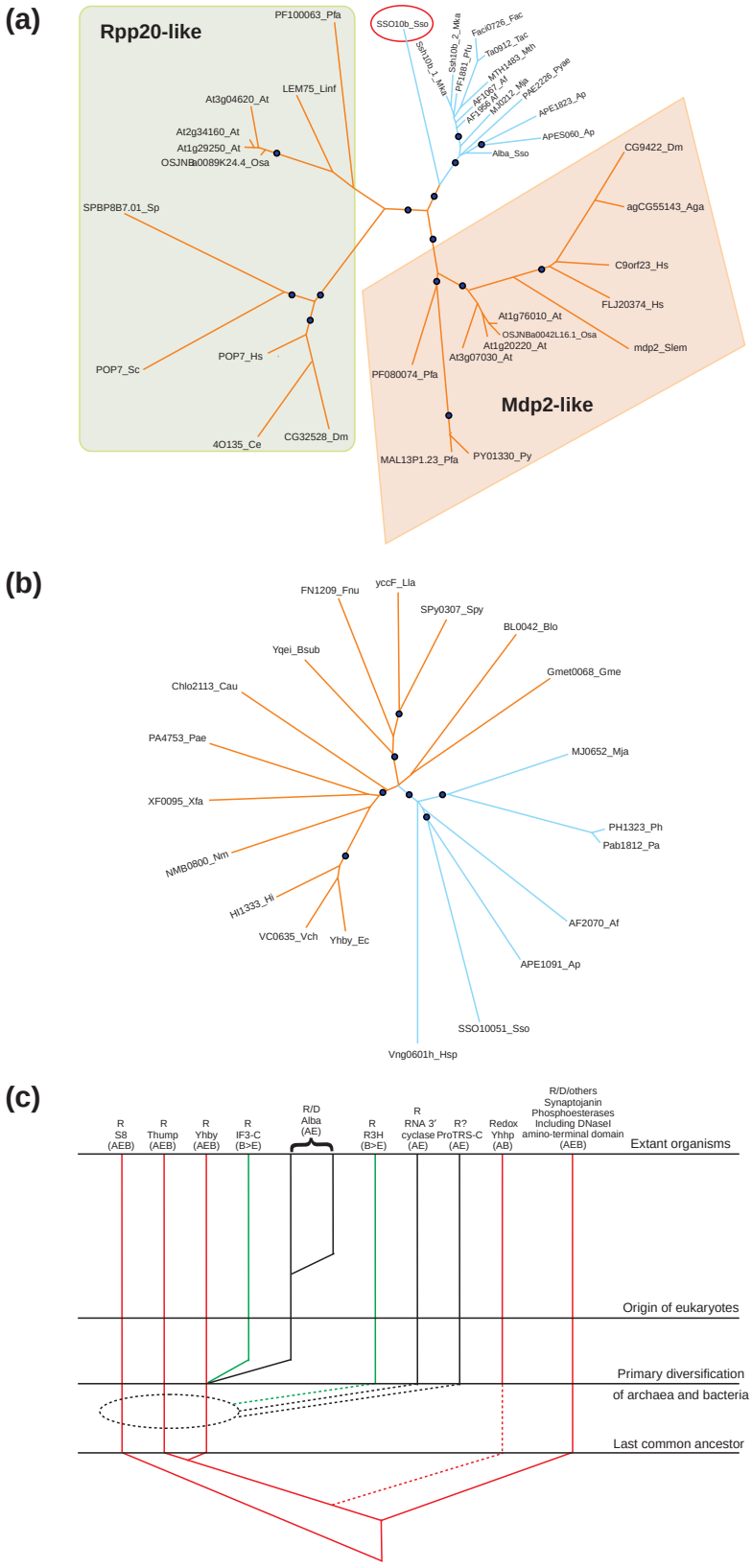

Figure 2 *(see legend on previous page)*

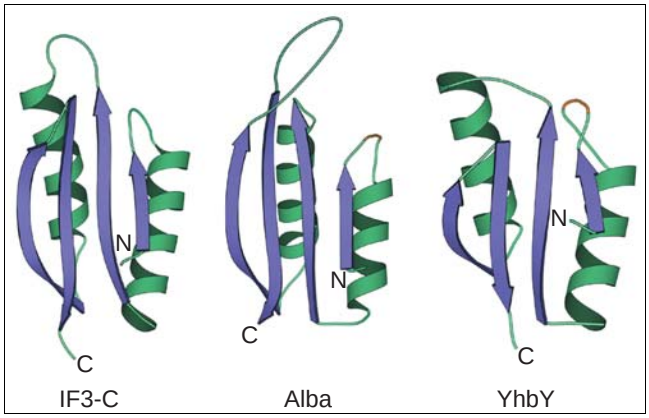

Figure 3
A ribbons representation of the IF3-C fold in the Alba, YhbY, IF3 proteins. The shared loop seen in YhbY and Alba (Figure 1) is in orange. The long carboxy-terminal β-hairpin is unique to Alba and is prone to great variability.

and phylogenetic analysis, we had traced the RNA-binding THUMP and S8 domains to LUCA [23]. The YhbY domain, which is well conserved in bacteria, archaea and plants, has been shown to be involved in intron splicing in plant chloroplasts [24-26]. It is physically associated with the ribosome in bacteria and is encoded in an operon, along with ribosomal proteins and RNase P subunits in archaea, suggesting an ancient role in translation [24]. In phylogenetic trees of the YhbY proteins, the archaeal and bacterial groups form distinct clusters which are clearly separated from each other by a distinct branch (Figure 2b). Furthermore, the archaeal and bacterial versions of YhbY are distinguished by specific sequence motifs. The tree topology, conservation pattern and function suggest that the YhbY domain was also likely to have been present in the LUCA. Hence, at least three ancient members of the IF3-C fold which possess RNA-binding properties potentially can be traced back to the point prior to the split between the bacteria and the archaeo-eukaryotic lineages, that is, the LUCA. These observations suggest that the ancestral version of the IF3-C fold is likely to have been an RNA-binding domain, which had already diversified into three distinct RNA-binding domains prior to the LUCA itself. Subsequently, Alba, the RNA 3'-terminal phosphate cyclase and Pro-TRSC appear to have been derived at the base of the archaeo-eukaryotic lineage, while IF3 and R3H were derived at the base of the bacterial lineage (Figure 2c).

Clustering, based on pairwise-structural-alignment Z-scores (obtained using the DALI program) and sequence similarity, suggests that the closest relatives of the Alba superfamily are IF3 and YhbY (Figures 2c and 3). Additionally, the YhbY and Alba domains also resemble each other in occurring typically in small (< 100 aa), single domain proteins, unlike IF3, R3H, Pro-TRSC, S8, RNA 3'-terminal cyclase, THUMP and DNase I amino-terminal domains, which occur in multidomain configurations. Furthermore, Alba and YhbY domains share a distinct basic loop, bounded by two small residues between the first strand and helix (Figures 1 and 3), which may participate in RNA binding. These observations suggest that Alba and YhbY probably diverged from each other early in the evolution of the archaeo-eukaryotic lineage.

Prediction of functional diversity within the Alba superfamily

Only three sets of proteins of the Alba superfamily are experimentally characterized: *Sulfolobus* Alba, Rpp20 proteins from yeast and vertebrates and the Rpp25 protein from vertebrates. The former has been shown to be the major chromosomal DNA-binding protein in the crenarchaeon *Sulfolobus*. However, even in *Sulfolobus*, Alba coats DNA densely but does not elicit significant compaction, suggesting that it is not sufficient for chromatin organization and requires functional partners, such as Sul7d [6]. Furthermore, at least in eukaryotes, even if Alba is a chromosomal protein it does appear to be as abundant as some of the other proteins such as HMG I/Y, HMG1 or related proteins [14]. Experimental studies on euryarchaeal chromatin have, so far, not yielded much evidence for a major chromosomal role for Alba [5,8,27]. Hence, it is conceivable that Alba was specifically recruited for a major chromosomal function only in a particular lineage of crenarchaea.

Rpp20/Pop7 from *S. cerevisiae* and humans has been demonstrated to be a shared, essential protein subunit common to both RNase P and RNase MRP [15-17]. It is likely that other members of the Rpp20 family, which are conserved widely across the eukaryotic tree, perform a similar function in RNase P. Human Rpp25, the only experimentally characterized member of the second eukaryotic family (Mdp2 family), is also a RNase P subunit [18]. Several uncharacterized members of the Mdp2 family contain tripeptide RGG repeats carboxy-terminal to their globular Alba domain (Figure 1). RGG repeats are found in a variety of eukaryotic ribonucleoproteins, such as nucleolin, FMRP and SAF-A and are known to bind a variety of RNAs, including G quartet structures [28-30]. As in the case of these eukaryotic Alba proteins, they occur typically in conjunction with some other globular RNA-binding domain. Affinity of these repeats for RNA is modulated through the methylation of arginine in many RGG-box proteins [31]. These observations suggest that, like Rpp25, other members of the Mdp2 family could also function as a RNase P/MRP subunit. The ciliate Mdp2, which is a member of the second eukaryotic family, is exclusively co-expressed along with Mdp1 and Mdp3 during macronuclear development [19]. The Mdp1 protein encodes a RNA binding protein with the Piwi and PAZ domains, while Mdp3 encodes a β-propeller protein with the kelch repeats (L.A., unpublished observations). The ciliate Piwi homologs, such as Mdp1, have been proposed to bind small RNAs that direct the process of DNA reorganization in macronucleus development, in a process related to gene silencing and repeat-induced mutagenesis in fungi [19,32]. The exclusive co-expression of Mdp2 with

Mdp1 suggests that the two may interact functionally with each other. The above observations suggest that Mdp2 and some of its relatives could potentially form other ribonucleoprotein complexes, which may perform functions in relation to the small RNAs involved in ciliate DNA reorganization, gene-silencing or allied processes. Members of the Mdp2 family are present in multiple copies in plants (Figure 2a) and may correlate with a well-developed gene-silencing system in this lineage [23,33].

Thus, an ancestral role in RNA metabolism may be interpolated for both eukaryotic families of the Alba superfamily. This is consistent with the ancient RNA-binding function of the IF3-C fold, including the closest neighbors of Alba, such as YhbY and IF3. The high degree of conservation of Alba in both euryarchaea and crenarchaea is in contrast with the heterogeneity seen in the phyletic patterns of other major DNA-packaging proteins within archaea (see below). Hence, based on its conserved archaeo-eukaryotic phyletic pattern, which is typical of several RNA-metabolism proteins [23], and by the principle of phylogenetic bracketing (Figure 2c), we would predict that most archaeal members of the Alba superfamily could have an additional or exclusive role in RNA metabolism. Of the 10 proteins of the eukaryotic RNase P and RNase MRP complexes that are conserved in eukaryotes, orthologs of Rpp30, Rpp29, Rpp21 and Pop5 (has an additional paralog Rpp14 in animals) are readily detectable in archaea [34]. Identification of the eukaryotic homologs of Alba as the subunits of RNase P/MRP implies that the archaeal Alba proteins could function possibly as the primitive equivalent of the Rpp20 and Rpp25 in the archaeal RNase P or a related ribonucleoprotein complex. This prediction is consistent with both the phyletic pattern of Alba in archaea and the requirement of Rpp20 for viability in yeast. This is also consistent with Alba retaining the basic ancestral RNA-binding property, after divergence from its functionally-related RNA-binding domains at the base of the archaeo-eukaryotic lineage. In light of these predictions, it would be of interest to investigate if the deacetylation of Alba by the Sir2 homologs might have any relevance to RNA metabolism.

The recruitment of RNA-binding domains for DNA-binding roles in chromatin appears to have occurred on different occasions in the course of eukaryotic evolution. A well-studied example of this process is provided by the HEH-fold found in the SAP and LEM domains of eukaryotic chromatin proteins. This fold is also present in proteins associated with RNA metabolism, such as a domain of the lysyl-tRNA synthetase and the bacterial transcription terminator protein Rho [35]. There is considerable diversity of chromosomal proteins across the archaea, with no major type being universally conserved [6]. All known crenarchaea lack histones but possess other proteins such as Alba or Sul7d, which is restricted to *Sulfolobus* [6]. Several euryarchaea possess histones, but others, for example, *Thermoplasma*, lack them and contain the bacterial-type HU protein instead [6]. Some euryarchaea also possess a distinct chromosomal protein of the MC1 family (incorporated into a multi-domain protein with helicase and nuclease modules in some archaea), which is also found in the PBCV virus [36,37]. Hence, it is conceivable that the archaea have explored a number of strategies for DNA organization, with different strategies being selected by different local, extreme niches. This diversity of chromosomal proteins in archaea is consistent with a model where a conserved protein such as Alba, with an original function in RNA metabolism, was recruited in some crenarchaea for chromatin-related functions. Interestingly, particular versions of Sso10b from Sulfolobales [10] are considerably divergent compared to all other archaeal Alba homologs (Figure 2a), suggesting that they may have undergone divergence to perform exclusive roles in chromatin.

Conclusions

Using sensitive sequence profile analysis, we unify the archaeal Alba proteins with the eukaryotic RNase P/MRP subunits Pop7/Rpp20 and Rpp25, and the ciliate protein Mdp2, which is involved in macronuclear development. The Alba domain superfamily comprises three well-defined families, namely the archaeal family and two eukaryotic families, typified by the Rpp20 and Mdp2 proteins, respectively. In turn, the Alba domain is closely related to the RNA-binding versions of the IF3-C fold such as YhbY and IF3-C. We present data which suggest that the ancestral function of the IF3-C fold was related to RNA interaction, and we present evidence that both the eukaryotic lineages of the Alba superfamily are principally involved in RNA metabolism. In addition to the RNase P complex, some of the Alba proteins may form complexes with other RNAs and participate in other regulatory processes, such as ciliate macronuclear development. The high degree of conservation of Alba in the archaea contrasts the poor conservation or mosaic phyletic distribution of other chromosomal proteins such as Sul7d, 7KMK, HU, MC1 and histones. These observations, along with the principle of phylogenetic bracketing, suggest that Alba in archaea may additionally or, in some lineages, exclusively, possess a role in RNA metabolism. Thus, starting as an ancestral RNA binding module, the Alba superfamily has colonized two functionally-distinct but biochemically similar niches in RNA metabolism and chromatin structure. Experimental exploration of the observations and the functional predictions reported here may help in improving our understanding of key processes in RNA metabolism.

Materials and methods

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTP program [38]. Profile searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with a default profile inclusion expectation (e)

value threshold of 0.01 (unless specified otherwise), and was iterated until convergence [38]. All large-scale sequence analysis procedures, such as determination of phyletic patterns, were carried out using the SEALS package [39,40]. Multiple alignments were constructed using the T-coffee program [41] and corrected based on PSI-BLAST pairwise alignments and structural alignments obtained with the DALI program. Similarity-based clustering of proteins was carried out using the BLASTCLUST program [42]. The MEGA program (version 2.1) was used for phylogenetic inference with the neighbor-joining algorithm [43]. The pairwise distances were estimated using the p-distance method. The Bayesian posterior probability trees were constructed using the MRBAYES program [44]. The maximum-likelihood tree was constructed using the following process: a least-squares tree was constructed using the FITCH program of the PHYLIP package or a neighbor-joining tree with the MEGA program. This was followed by local rearrangements using the ProtML program of the Molphy package to arrive at the maximum likelihood (ML) tree [45-47]. The statistical significance of various nodes of this ML tree was assessed using the relative estimate of logarithmic likelihood bootstrap (Protml RELL-BP), with 10,000 replicates.

Structural manipulations were carried out using the Swiss-PDB viewer version 3.7 program [48]. Searches of the PDB database, or custom set of target structure with a given query structure was conducted using the DALI and the VAST programs [49,50]. Interatomic distances in structure were determined using the Swiss-PDB viewer and ribbon diagrams were constructed using the MOLSCRIPT program [51].

Acknowledgements

We would like to thank an anonymous referee for useful suggestions regarding the identity of the Rpp25 subunit of RNase P.

References

1. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J: *Molecular Cell Biology.* New York: WH Freeman and Co; 1999.
2. Dorman CJ, Deighan P: **Regulation of gene expression by histone-like proteins in bacteria.** *Curr Opin Genet Dev* 2003, **13:**179-184.
3. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389:**251-260.
4. Sandman K, Bailey KA, Pereira SL, Soares D, Li WT, Reeve JN: **Archaeal histones and nucleosomes.** *Methods Enzymol* 2001, **334:**116-129.
5. Pavlov NA, Cherny DI, Nazimov IV, Slesarev AI, Subramaniam V: **Identification, cloning and characterization of a new DNA-binding protein from the hyperthermophilic methanogen *Methanopyrus kandleri*.** *Nucleic Acids Res* 2002, **30:**685-694.
6. White MF, Bell SD: **Holding it together: chromatin in the Archaea.** *Trends Genet* 2002, **18:**621-626.
7. Ball LJ, Murzina NV, Broadhurst RW, Raine AR, Archer SJ, Stott FJ, Murzin AG, Singh PB, Domaille PJ, Laue ED: **Structure of the chromatin binding (chromo) domain from mouse modifier protein 1.** *EMBO J* 1997, **16:**2473-2481.
8. Bell SD, Botting CH, Wardleworth BN, Jackson SP, White MF: **The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation.** *Science* 2002, **296:**148-151.
9. Xue H, Guo R, Wen Y, Liu D, Huang L: **An abundant DNA binding protein from the hyperthermophilic archaeon Sulfolobus shibatae affects DNA supercoiling in a temperature-dependent fashion.** *J Bacteriol* 2000, **182:**3929-3933.
10. Forterre P, Confalonieri F, Knapp S: **Identification of the gene encoding archeal-specific DNA-binding proteins of the Sac10b family.** *Mol Microbiol* 1999, **32:**669-670.
11. Zhao K, Chai X, Marmorstein R: **Structure of a Sir2 substrate, Alba, reveals a mechanism for deactylation-induced enhancement of DNA-binding.** *J Biol Chem* 2003, **278:**26071-26077.
12. Wardleworth BN, Russell RJ, Bell SD, Taylor GL, White MF: **Structure of Alba: an archaeal chromatin protein modulated by acetylation.** *EMBO J* 2002, **21:**4654-4662.
13. Dinger ME, Baillie GJ, Musgrave DR: **Growth phase-dependent expression and degradation of histones in the thermophilic archaeon *Thermococcus zilligii*.** *Mol Microbiol* 2000, **36:**876-885.
14. Wolffe AP: *Chromatin: Structure and Function. London: Academic Press* 1998.
15. Stolc V, Katz A, Altman S: **Rpp2, an essential protein subunit of nuclear RNase P, is required for processing of precursor tRNAs and 35S precursor rRNA in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 1998, **95:**6716-6721.
16. Chamberlain JR, Lee Y, Lane WS, Engelke DR: **Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP.** *Genes Dev* 1998, **12:**1678-1690.
17. Eder PS, Kekuda R, Stolc V, Altman S: **Characterization of two scleroderma autoimmune antigens that copurify with human ribonuclease P.** *Proc Natl Acad Sci USA* 1997, **94:**1101-1106.
18. Guerrier-Takada C, Eder PS, Gopalan V, Altman S: **Purification and characterization of Rpp25, an RNA-binding protein subunit of human ribonuclease P.** *RNA* 2002, **8:**290-295.
19. Fetzer CP, Hogan DJ, Lipps HJ: **A PIWI homolog is one of the proteins expressed exclusively during macronuclear development in the ciliate *Stylonychia lemnae*.** *Nucleic Acids Res* 2002, **30:**4380-4386.
20. Liepinsh E, Leonchiks A, Sharipo A, Guignard L, Otting G: **Solution structure of the R3H domain from human Smubp-2.** *J Mol Biol* 2003, **326:**217-223.
21. Yaremchuk A, Cusack S, Tukalo M: **Crystal structure of a eukaryote/archaeon-like protyl-tRNA synthetase and its complex with tRNAPro(CGG).** *EMBO J* 2000, **19:**4745-4758.
22. Aravind L, Koonin EV: **THUMP - a predicted RNA-binding domain shared by 4-thiouridine, pseudouridine synthases and RNA methylases.** *Trends Biochem Sci* 2001, **26:**215-217.
23. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, **30:**1427-1464.
24. Ostheimer GJ, Barkan A, Matthews BW: **Crystal structure of *E. coli* YhbY: a representative of a novel class of RNA binding proteins.** *Structure (Camb)* 2002, **10:**1593-1601.
25. Willis MA, Krajewski W, Chalamasetty VR, Reddy P, Howard A, Herzberg O: **Structure of HI1333 (YhbY), a putative RNA-binding protein from *Haemophilus influenzae*.** *Proteins* 2002, **49:**423-426.
26. Till B, Schmitz-Linneweber C, Williams-Carrier R, Barkan A: **CRS1 is a novel group II intron splicing factor that was derived from a domain of ancient origin.** *RNA* 2001, **7:**1227-1238.
27. Pavlov NA, Cherny DI, Jovin TM, Slesarev AI: **Nucleosome-like complex of the histone from the hyperthermophile *Methanopyrus kandleri* (MkaH) with linear DNA.** *J Biomol Struct Dyn* 2002, **20:**207-214.
28. Kiledjian M, Dreyfuss G: **Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box.** *EMBO J* 1992, **11:**2655-2664.
29. Steinert PM, Mack JW, Korge BP, Gan SQ, Haynes SR, Steven AC: **Glycine loops in proteins: their occurrence in certain intermediate filament chains, loricrins and single-stranded RNA binding proteins.** *Int J Biol Macromol* 1991, **13:**130-139.
30. Darnell JC, Jensen KB, Jin P, Brown V, Warren ST, Darnell RB: **Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function.** *Cell* 2001, **107:**489-499.
31. Liu Q, Dreyfuss G: ***In vivo* and *in vitro* arginine methylation of RNA-binding proteins.** *Mol Cell Biol* 1995, **15:**2800-2808.
32. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA: **Analysis of a piwi-related gene implicates small RNAs in genome rear-**

rangement in tetrahymena. *Cell* 2002, **110:**689-699.
33. Baulcombe D: **RNA silencing.** *Curr Biol* 2002, **12:**R82-R84.
34. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11:**240-252.
35. Aravind L, Mazumder R, Vasudevan S, Koonin EV: **Trends in protein evolution inferred from sequence and structure analysis.** *Curr Opin Struct Biol* 2002, **12:**392-399.
36. Chartier F, Laine B, Sautiere P: **Characterization of the chromosomal protein MC1 from the thermophilic archaebacterium *Methanosarcina* sp. CHTI 55 and its effect on the thermal stability of DNA.** *Biochim Biophys Acta* 1988, **951:**149-156.
37. Cam EL, Culard F, Larquet E, Delain E, Cognet JA: **DNA bending induced by the archaebacterial histone-like protein MC1.** *J Mol Biol* 1999, **285:**1011-1021.
38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
39. Walker DR, Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**333-339.
40. **SEALS** [http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html]
41. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217.
42. **BLASTCLUST - BLAST score-based single-linkage clustering** [ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt]
43. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.
44. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
45. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1:**8.
46. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266:**418-427.
47. Hasegawa M, Kishino H, Saitou N: **On the maximum likelihood method in molecular phylogenetics.** *J Mol Evol* 1991, **32:**443-445.
48. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb-Viewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
49. Holm L, Sander C: **Touring protein fold space with Dali/FSSP.** *Nucleic Acids Res* 1998, **26:**316-319.
50. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6:**377-385.
51. Kraulis PJ: **MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures.** *Journal of App Crystallography* 1991, **24:**946-950.
52. Koonin EV, Aravind L, Galperin MG: **Bacterial Stress Responses.** In *A Comparative Genomic View of the Microbial Stress Response.* Washington DC: ASM Press; 2000.