Phylogeny of related functions: the case of polyamine biosynthetic enzymes

Agnieszka Sekowska,[1,2] Antoine Danchin[1,2] and Jean-Loup Risler[3]

Author for correspondence: Antoine Danchin. Tel: +852 2816 8402. Fax: +852 2168 4427.
e-mail: adanchin@hkucc.hku.hk

[1] Regulation of Gene Expression, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

[2] Hong Kong University Pasteur Research Centre, Dexter HC Man Building, 8 Sassoon Road, Pokfulam, Hong Kong

[3] Genome and Informatics, Université de Versailles-Saint-Quentin, 45 Avenue des Etats Unis, 78035 Versailles Cedex, France

Genome annotation requires explicit identification of gene function. This task frequently uses protein sequence alignments with examples having a known function. Genetic drift, co-evolution of subunits in protein complexes and a variety of other constraints interfere with the relevance of alignments. Using a specific class of proteins, it is shown that a simple data analysis approach can help solve some of the problems posed. The origin of ureohydrolases has been explored by comparing sequence similarity trees, maximizing amino acid alignment conservation. The trees separate agmatinases from arginases but suggest the presence of unknown biases responsible for unexpected positions of some enzymes. Using factorial correspondence analysis, a distance tree between sequences was established, comparing regions with gaps in the alignments. The gap tree gives a consistent picture of functional kinship, perhaps reflecting some aspects of phylogeny, with a clear domain of enzymes encoding two types of ureohydrolases (agmatinases and arginases) and activities related to, but different from ureohydrolases. Several annotated genes appeared to correspond to a wrong assignment if the trees were significant. They were cloned and their products expressed and identified biochemically. This substantiated the validity of the gap tree. Its organization suggests a very ancient origin of ureohydrolases. Some enzymes of eukaryotic origin are spread throughout the arginase part of the trees: they might have been derived from the genes found in the early symbiotic bacteria that became the organelles. They were transferred to the nucleus when symbiotic genes had to escape Muller's ratchet. This work also shows that arginases and agmatinases share the same two manganese-ion-binding sites and exhibit only subtle differences that can be accounted for knowing the three-dimensional structure of arginases. In the absence of explicit biochemical data, extreme caution is needed when annotating genes having similarities to ureohydrolases.

Keywords: secondary metabolism, *Helicobacter pylori*, *Bacillus subtilis*, *Synechocystis* PCC6803, discriminant amino acid residue

INTRODUCTION

While the number of genome sequences increases exponentially it remains difficult to identify gene functions explicitly. Automatic annotation procedures rely mostly on sequence comparisons. They are used to build up phylogeny trees, where reference activities are assumed to spread to neighbours by contiguity. The corresponding functions are thus described tentatively as identical to that of the known reference. However, these methods do not address the central question of enzyme recruitment for new activities (Jensen, 1976; Danchin, 1989; Roy, 1999). Furthermore, genes and proteins are not simply sequences of letters, they are made from chemicals derived from cell metabolism and a single gene alteration may result in a general base or amino acid content bias (Cox & Yanofsky, 1967), changing the 'style' of an organism, possibly altering its place in calculated phylogenies, thus leading to wrong assignments of enzyme activities.

From the chemical standpoint, the origin of life is often

Abbreviation: FCA, Factorial Correspondence Analysis.

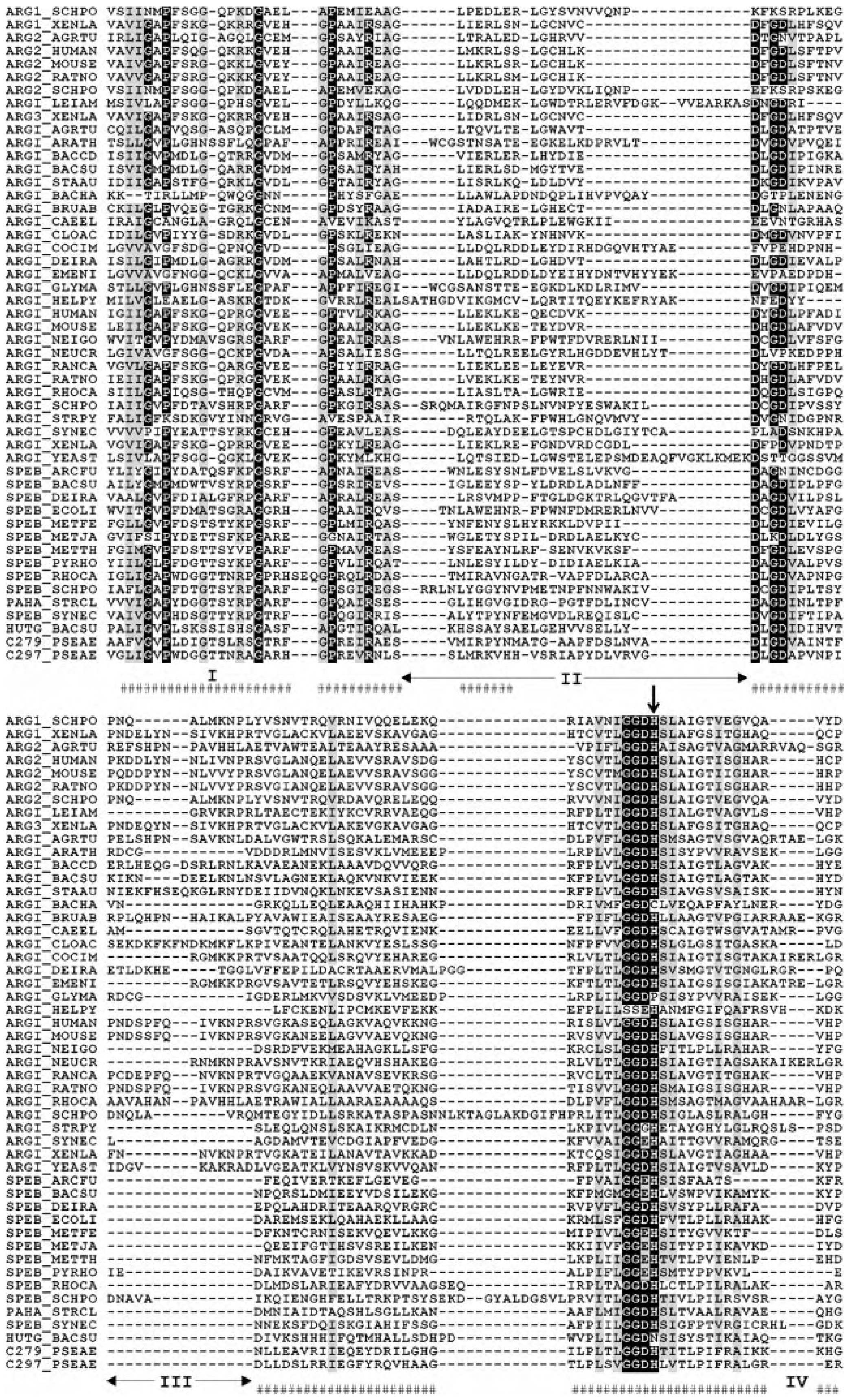


Fig. 1. For legend see page 1818.

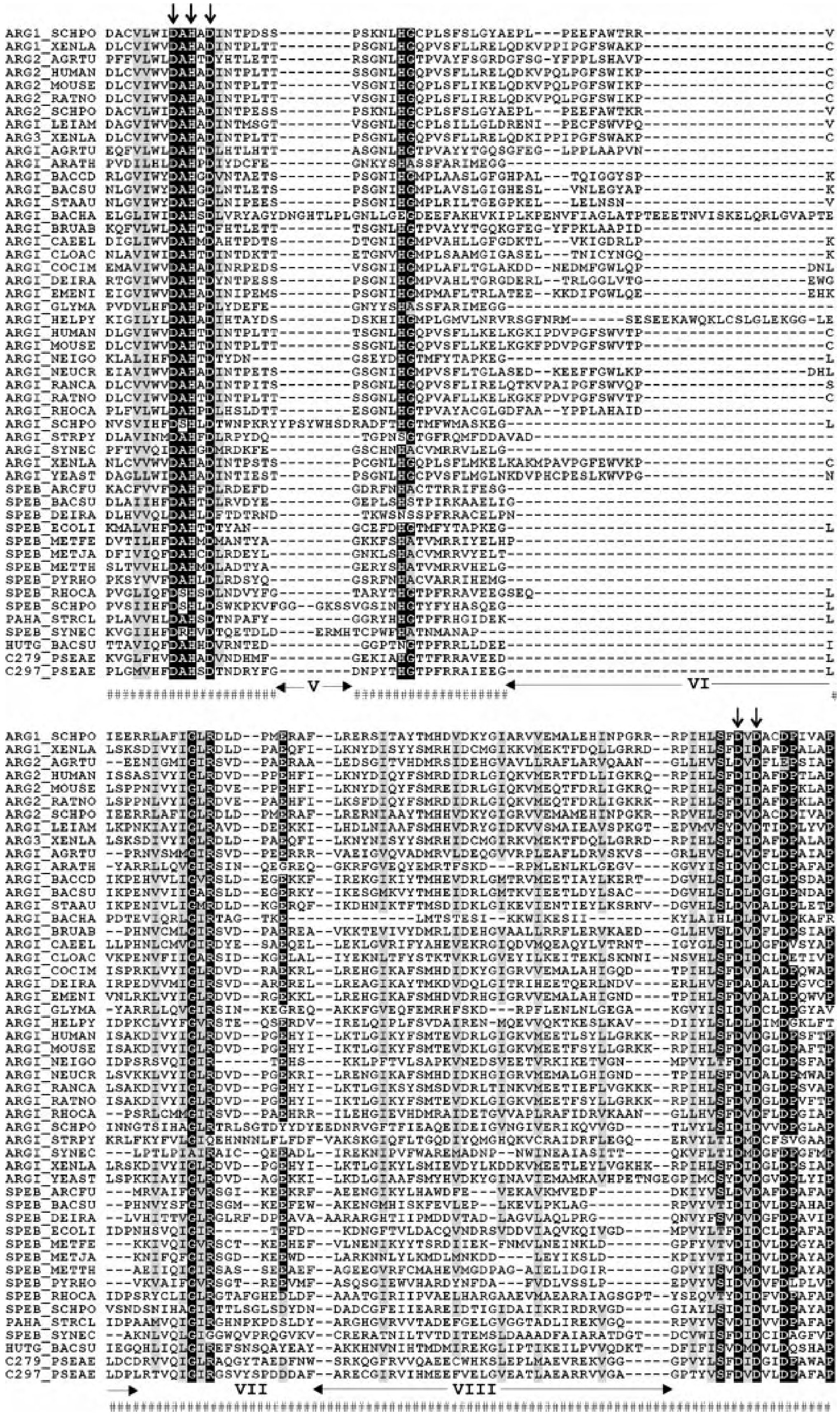
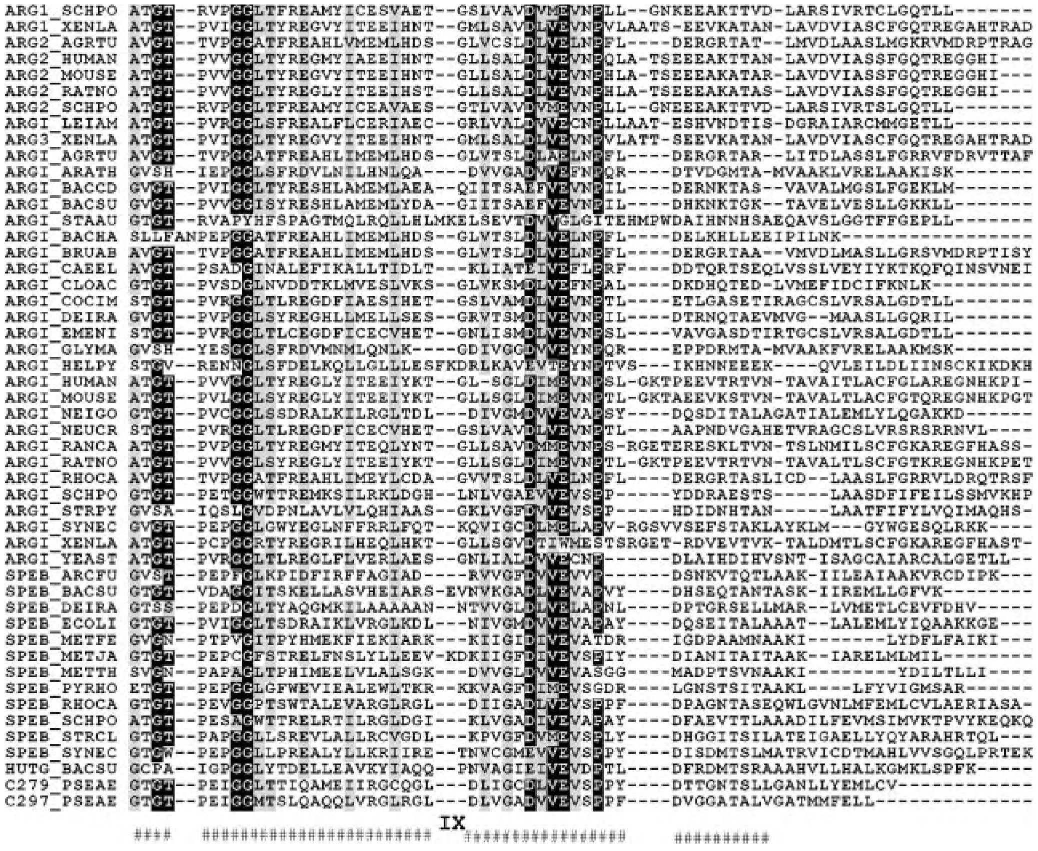


Fig. 1 (*cont.*) For legend see page 1818.



```
ARG1_SCHPO ATGT--RVPGGLTFREAMYICESVAET--GSLVAVDVMEVNPLL--GNKEEAKTTVD-LARSIVRTCLGQTLL-------
ARG1_XENLA ATGT--PVIGGLTYREGVYITEEIHNT--GMLSAVDLVEVNPVLAATS-EEVKATAN-LAVDVIASCFGQTREGAHTRAD
ARG2_AGRTU AVGT--TVPGGATFREAHLVMEMLHDS--GLVCSLDLVELNPFL----DERGRTAT--LMVDLAASLMGKRVMDRPTRAG
ARG2_HUMAN ATGT--PVVGGLTYREGMYIAEEIHNT--GLLSALDLVEVNPQLA-TSEEEAKTTAN-LAVDVIASSFGQTREGGHI---
ARG2_MOUSE ATGT--PVVGGLTYREGVYITEEIHNT--GLLSALDLVEVNPHLA-TSEEEAKATAR-LAVDVIASSFGQTREGGHI---
ARG2_RATNO ATGT--PVVGGLTYREGLYITEEIHST--GLLSALDLVEVNPHLA-TSEEEAKATAS-LAVDVIASSFGQTREGGHI---
ARG2_SCHPO ATGT--RVPGGLTFREAMYICEAVAES--GTLVAVDVMEVNPLL--GNEEEAKTTVD-LARSIVRTSLGQTLL-------
ARGI_LEIAM ATGT--PVRGGLSFREALFLCERIAEC--GRLVALDVVECNPLLAAT-ESHVNDTIS-DGRAIARCMMGETLL-------
ARG3_XENLA ATGT--PVIGGLTYREGVYITEEIHNT--GMLSALDLVEVNPVLATT-SEEVKATAN-LAVDVIASCFGQTREGAHTRAD
ARGI_AGRTU AVGT--TVPGGATFREAHLIMEMLHDS--GLVTSLDLAELNPFL----DERGRTAR--LITDLASSLFGRRVFDRVTTAF
ARGI_ARATH GVSH--IEPGGLSFRDVLNILHNLQA----DVVGADVVEFNPQR----DTVDGMTA-MVAAKLVRELAAKISK-------
ARGI_BACCD GVGT--PVIGGLTYRESHLAMEMLAEA--QIITSAEFVEVNPIL----DERNKTAS--VAVALMGSLFGEKLM-------
ARGI_BACSU GVGT--PVVGGISYRESHLAMEMLYDA--GIITSAEFVEVNPIL----DHKNKTGK--TAVELVESLLGKKLL-------
ARGI_STAAU GTGT--RVAPYHFSPAGTMQLRQLLHLMKELSEVTDVVGLGITEHMPWDAIHNNHSAEQAVSLGGTFFGEPLL-------
ARGI_BACHA SLLFANPEPGGATFREAHLIMEMLHDS--GLVTSLDLVELNPFL----DELKHLLEEIPILNK----------------
ARGI_BRUAB AVGT--TVPGGATFREAHLIMEMLHDS--GLVTSLDLVELNPFL----DERGRTAA--VMVDLMASLLGRSVMDRPTISY
ARGI_CAEEL AVGT--PSADGINALEFIKALLTIDLT---KLIATEIVEFLPRF----DDTQRTSEQLVSSLVEYIYKTKQFQINSVNEI
ARGI_CLOAC GTGT--PVSDGLNVDDTKLMVESLVKS--GLVKSMDLVEFNPAL----DKDHQTED-LVMEFIDCIFKNLK---------
ARGI_COCIM STGT--PVRGGLTLREGDFIAESIHET--GSLVAMDLVEVNPTL----ETLGASETIRAGCSLVRSALGDTLL-------
ARGI_DEIRA GVGT--PVPGGLSYREGHLLMELLSES--GRVTSMDIVEVNPIL----DTRNQTAEVMVG--MAASLLGQRIL-------
ARGI_EMENI STGT--PVRGGLTLCEGDFICECVHET--GNLISMDLVEVNPSL----VAVGASDTIRTGCSLVRSALGDTLL-------
ARGI_GLYMA GVSH--YESGGLSFRDVMNMLQNLK----GDIVGGDVVEYNPQR----EPPDRMTA-MVAAKFVRELAAKMSK-------
ARGI_HELPY STGV--RENNGLSFDELKQLLGLLLESFKDRLKAVEVTEYNPTVS---IKHNNEEEK----QVLEILDLIINSCKIKDKH
ARGI_HUMAN ATGT--PVVGGLTYREGLYITEEIYKT--GL-SGLDIMEVNPSL-GKTPEEVTRTVN-TAVAITLACFGLAREGNHKPI-
ARGI_MOUSE ATGT--PVLGGLSYREGLYITEEIYKT--GLLSGLDIMEVNPTL-GKTAEEVKSTVN-TAVALTLACFGTQREGNHKPGT
ARGI_NEIGO GTGT--PVCGGLSSDRALKILRGLTDL---DIVGMDVVEVAPSY----DQSDITALAGATIALEMLYLQGAKKD------
ARGI_NEUCR STGT--PVRGGLTLREGDFICECVHET--GSLVAVDLVEVNPTL----AAPNDVGAHETVRAGCSLVRSRSRRNVL----
ARGI_RANCA ATGT--PVPGGLTYREGMYITEQLYNT--GLLSAVDMMEVNPS-RGETERESKLTVN-TSLNMILSCFGKAREGFHASS-
ARGI_RATNO ATGT--PVVGGLSYREGLYITEEIYKT--GLLSGLDIMEVNPTL-GKTPEEVTRTVN-TAVALTLSCFGTKREGNHKPET
ARGI_RHOCA AVGT--TVPGGATFREAHLIMEYLCDA--GVVTSLDLVELNPFL----DERGRTASLICD--LAASLFGRRVLDRQTRSF
ARGI_SCHPO GTGT--PETGGWTTREMKSILRKLDGH--LNLVGAEVVEVSPP-----YDDRAESTS-----LAASDFIFEILSSMVKHP
ARGI_STRPY GVSA--IQSLGVDPNLAVLVLQHIAAS--GKLVGFDVVEVSPP-----HDIDNHTAN-----LAATFIFYLVQIMAQHS-
ARGI_SYNEC GVGT--PEPGGLGWYEGLNFFRRLFQT--KQVIGCDLMELAPV-RGSVVSEFSTAKLAYKLM---GYWGESQLRKK----
ARGI_XENLA ATGT--PCPGGRTYREGRILHEQLHKT--GLLSGVDTIWMESTSRGET-RDVEVTVK-TALDMTLSCFGKAREGFHAST-
ARGI_YEAST ATGT--PVRGGLTLREGLFLVERLAES--GNLIALDVVECNP------DLAIHDIHVSNT-ISAGCAIARCALGETLL--
SPEB_ARCFU GVST--PEPFGLKPIDFIRFFAGIAD----RVVGFDVVEVVP------DSNKVTQTLAAK-IILEAIAAKVRCDIPK---
SPEB_BACSU GTGT--VDAGGITSKELLASVHEIARS-EVNVKGADLVEVAPVY----DHSEQTANTASK-IIREMLLGFVK--------
SPEB_DEIRA GTSS--PEPDGLTYAQGMKILAAAAAN--NTVVGLDLVELAPNL----DPTGRSELLMAR--LVMETLCEVFDHV-----
SPEB_ECOLI GTGT--PVIGGLTSDRAIKLVRGLKDL---NIVGMDVVEVAPAY----DQSEITALAAAT--LALEMLYIQAAKKGE---
SPEB_METFE GVGN--PTPVGITPYHMEKFIEKIARK---KIIGIDIVEVATDR----IGDPAAMNAAKI-------LYDFLFAIKI---
SPEB_METJA GTGT--PEPCGFSTRELFNSLYLLEEV-KDKIIGFDIVEVSPIY----DIANITAITAAK--IARELMLMIL--------
SPEB_METTH SVGN--PAPAGLTPHIMEELVLALSGK---DVVGLDVVEVASGG----MADPTSVNAAKI--------YDILTLLI----
SPEB_PYRHO ETGT--PEPGGLGFWEVIEALEWLTKR--KKVAGFDIMEVSGDR----LGNSTSITAAKL----LFYVIGMSAR------
SPEB_RHOCA GTGT--PEVGGPTSWTALEVARGLRGL---DIIGADLVEVSPPF----DPAGNTASEQWLGVNLMFEMLCVLAERIASA-
SPEB_SCHPO ATGT--PESAGWTTRELRTILRGLDGI---KLVGADIVEVAPAY----DFAEVTTLAAADILFEVMSIMVKTPVYKEQKQ
SPEB_STRCL GTGT--PAPGGLLSREVLALLRCVGDL---KPVGFDVMEVSPLY----DHGGITSILATEIGAELLYQYARAHRTQL---
SPEB_SYNEC GTGW--PEPGGLLPREALYLLKRIIRE--TNVCGMEVVEVSPPY----DISDMTSLMATRVICDTMAHLVVSGQLPRTEK
HUTG_BACSU GCPA--IGPGGLYTDELLEAVKYIAQQ--PNVAGIEIVEVDPTL----DFRDMTSRAAAHVLLHALKGMKLSPFK-----
C279_PSEAE GTGT--PEIGGLTTIQAMEIIRGCQGL---DLIGCDLVEVSPPY----DTTGNTSLLGANLLYEMLCV------------
C297_PSEAE GTGT--PEIGGMTSLQAQQLVRGLRGL---DLVGADVVEVSPPF----DVGGATALVGATMMFELL--------------
                                               IX
           ####  ########################    #################    ##########
```

Fig. 1. Multiple alignment of 50 sequences of ureohydrolases and related enzymes, as obtained from CLUSTAL W and DIALIGN, manually edited in the regions containing many non-contiguous gaps. The highly divergent N and C termini have been removed. Identical and similar residues in more than 75% of the sequences are drawn on black and shaded backgrounds, respectively. Those residues that were kept for the phylogenetic reconstructions are labelled '#', the excluded parts corresponding to regions where CLUSTAL W and DIALIGN gave highly different alignments. The roman numerals point to the regions that were used for building the 'gap tree' (see Results and Fig. 3). The arrows point to the residues involved in the binding of manganese ions. In this figure and the following ones, the names of the proteins follow the SWISS-PROT convention: ARGI and SPEB indicate proteins that are noted as arginases and agmatinases, respectively. The names of the species follow the five-letter code as indicated in Table 1.

reduced to the question of incorporation of carbon into the simple building blocks needed for biopolymer synthesis. However, the universal requirement for nucleotides and amino acids demonstrates that nitrogen metabolism was extremely important in the first steps of life (Granick, 1957). It is likely that a nitrogen-fixing cycle predated life, placing nitrogen-rich molecules in the limelight (Danchin, 1989), polyamines being crucial molecules (Cohen, 1998). In this context, Ouzounis & Kyprides (1994) constructed an interesting evolutionary tree of agmatinases, the polyamine-synthesizing enzymes, with emphasis on their universal presence. Since this seminal work, many new sequences have been obtained and annotated by their similarity with the known sequences. We therefore undertook a comparative analysis of the corresponding set of sequences. Genes that were deemed important were cloned and attempts were made to identify their functions. We reconsidered the phylogeny trees of arginases/agmatinases and constructed new ones for the enzymes involved in this first step in polyamine metabolism. To incorporate evolutionary constraints of different types, we first considered the usual types of phylogeny trees constructed based on the variation of the amino acid sequence in these proteins, without taking into account the presence of gaps in the sequences. Several discrepancies with respect to the expected position of some organisms in the trees were found. In a second approach, we reconstructed trees based only on the presence and evolution of gap-containing regions in the sequences (Baldauf & Palmer, 1993; Gupta, 1998a; see Fitch & Yasunobu, 1975, for appropriate caveats), because gaps would be much less sensitive to genetic drift or amino acid metabolism. The crucial enzyme activities that presumably evolved from ancestral ureohydrolases were validated by cloning, expressing and measuring activity of the corresponding enzymes. The emerging picture is consistent with a bacterial origin of hydrolases (ureohydrolases and related activities), which later evolved to those of the *Archaea* and the *Eukarya*.

METHODS

Sequence data collection and bacterial strains. Sequence data were collected from the GenBank/EBI/DDBJ DNA sequence database. Some sequences were also obtained from ongoing genome sequencing programs.

The reference *Escherichia coli* strain for the study of polyamine biosynthesis was obtained from the *E. coli* Genetic Stock Center, through the kind help of Dr Mary Berlyn. Strain MA255 was used: K12 *thr-1 leuB6 fhuA2*? *lacY1 glnV44*(*AS*)? *gal-6* λ^- *relA1*? *can-1 speB2 speC3 rpsL133*(*strR*) *xylA7 mtlA2 thi-1*. For cloning experiments strain XL-1 Blue was used (K12 *supE44 hsdR17 recA1 endA1 gyrA46 thi relA1* Δ*lac* F′[*proAB*$^+$ *lacI*$^{\mathrm{q}}$ *lacZ*ΔM15 Tn*10*(tet$^{\mathrm{R}}$)]; laboratory collection).

Synechocystis PCC6803, *Helicobacter pylori* 26695 and *Neisseria gonorrheae* MS11-E DNAs were kind gifts from Dr N. Tandeau de Marsac (Physiologie Microbienne, Institut Pasteur, Paris), Dr H. De Reuse (Pathogénie Bactérienne des Muqueuses, Institut Pasteur, Paris) and Dr M.-K. Taha (Unité des Neisseria, Institut Pasteur, Paris), respectively.

Sequence alignment and construction of trees. The 50 ArgI and SpeB amino acid sequences and the 22 spermidine (spermine) synthase sequences currently available were aligned with the programs DIALIGN version 2 (Morgenstern *et al.*, 1998), as well as CLUSTAL W (Higgins *et al.*, 1996), using default parameters (similarity matrix BLOSUM, 30; gap open penalty, 10; gap extension penalty, 0·1). A second putative ArgI sequence (encoded by the *rocF* gene) from *H. pylori* was extracted from the Astra server for comparison to the one present in our alignments. Because the differences were very small and did not change the positions of gaps, this sequence was not incorporated further in our analysis. The resulting multiple alignment was then checked for the conservation of important residues (Perozich *et al.*, 1998) and manually edited. Regions where the two alignment programs gave widely different solutions were removed (see Fig. 1). Phylogenetic analyses were performed using the PHYLIP 3.57c suite of programs (Felsenstein, 1993). The pairwise distance matrices were calculated by PROTDIST with the Dayhoff option to estimate the expected amino acid replacements per position. The neighbour-joining (NJ) trees were obtained with NEIGHBOR. The most parsimonious trees were determined with PROTPARS. In each case, we performed 1000 bootstrap resamplings with SEQBOOT. The consensus trees were calculated by CONSENSE and drawn with the program TREEVIEW (Page, 1996).

To obtain a tree based essentially on structural features, we constructed a matrix where the presence and the length of insertions/deletions (nine values for each sequence; see Fig. 1) was given as the input. This was done without considering the amino acid variation in the sequences, only by counting the number of gaps in those regions delimited by highly conserved residues. Because evolution trends are to conserve the length of gene products, a last column with the length of the sequence as input was added in the matrix. This matrix was used to perform a multivariate analysis (Factorial Correspondence Analysis, FCA) to compute distances (χ^2) between the sequences (Hill, 1974; Lebart *et al.*, 1984). The distances between each pair of sequences were calculated from the first three co-ordinates of the FCA (73% of the total inertia). Finally, a UPGMA (Unweighted Pair Group Method using Arithmetic Averages) distance tree was calculated from the resulting distance matrix (Sneath & Sokal, 1973). Here we chose UPGMA rather than NJ because we did not intend to build a phylogeny tree, but rather to perform a simple cluster analysis which would group those species that are least different, regardless of any further refinement of the tree.

Cloning procedures and biochemical assays. Cloning was performed by PCR amplification of the DNA regions of the putative *argI* or *speB* genes under study and followed by subsequent ligation of PCR products into plasmid p*Trc*99A (Pharmacia Biotech).

To clone the putative *argI* gene from *Synechocystis* PCC6803, a DNA fragment beginning at the translational start point and ending 5 bp after the stop codon of *argI* was amplified by PCR, using primers introducing a *Bsp*HI cloning site at the 5′ end and a *Bgl*II cloning site at the 3′ end of the fragment. The PCR product was ligated and inserted into the *Nco*I and *Bam*HI sites of p*Trc*99A, creating pDIA5600. To clone the putative *speB* gene from *Synechocystis* PCC6803, a DNA fragment beginning at the translational start point and ending 52 bp after the stop codon of *speB* was amplified by PCR, using primers introducing an *Afl*III cloning site at the 5′ end and a *Bam*HI cloning site at the 3′ end of the fragment. The PCR product was ligated as described above, creating pDIA5601. To clone the *rocF* gene (encoding ArgI) from *H. pylori* 26695, a DNA fragment beginning at the translational start point and ending at the stop codon of *rocF* was amplified by PCR using primers introducing an *Nco*I cloning site at the 5′ end and a *Bam*HI cloning site at the 3′ end of the fragment. The PCR product was ligated as described above, creating pDIA5602. To clone the putative *argI* gene from *N. gonorrheae* MS11-E, a DNA fragment beginning at the translational start point and ending 4 bp after the stop codon of *argI* was amplified by PCR using primers introducing a *Bsp*HI cloning site at the 5′ end and a *Bam*HI cloning site at the 3′ end of the fragment. The PCR product was ligated as described above, creating pDIA5603.

The plasmids containing these putative *argI* or *speB* genes were expressed in an *E. coli* mutant unable to synthesize polyamines. To assay agmatinase/arginase activities, strain MA255 was used.

For determination of enzymic activities, the bacteria from 200 ml LB overnight cultures were centrifuged, washed with PBS and centrifuged again; the pellets were then weighed. Extracts were prepared by grinding the bacterial paste for several minutes in a mortar with alumina (equal to twice the pellet weight). The mixtures were resuspended in 50 mM Tris/HCl, pH 7·6, containing 1 mM EDTA and 1 mM DTT and centrifuged at 10000 r.p.m. for 30 min in the cold. Supernatants were used for determination of enzyme activities. Urea was produced as described by Hirshfield *et al.* (1970) using either agmatine or arginine as substrates (10 mM each). The assay was slightly modified by raising the pH of the Tris/HCl buffer from 7·5 to 9·0 as described by Yamamoto *et al.* (1988). Urea measurement was performed using the Blood urea nitrogen assay kit (Sigma).

RESULTS

Multialignment of arginases, agmatinases and related sequences

A collection of 50 arginase- or agmatinase-like proteins were multi-aligned using DIALIGN (Morgenstern *et al.*,

1998) and CLUSTAL W (Higgins *et al.*, 1996), and the alignments were further refined by realigning regions located between highly conserved segments (Fig. 1). Remarkably, the *H. pylori* sequence does not retain several features highly conserved in the other sequences. The sequence was checked by using the sequence published by Astra-Zeneca in addition to the sequence published at TIGR. As seen in Fig. 1, the multiple alignment is characterized by short highly conserved regions, separated by segments of much lower overall similarity. On the whole, the similarity between sequences is still large enough to identify related sequences using the BLAST programs, in particular PSI-BLAST. To be as conservative as possible, the number of gaps permitting alignment was kept to a minimum. Nine gap-containing regions were kept for further study, leaving aside gap regions that were uninformative, i.e. that either displayed too little difference between sequences or gave widely different results with the two alignment programs (see Fig. 1). Among the conserved regions are the two manganese-ion-binding sites present in these enzymes (Prosite; http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00135). The critical residues are conserved in all but a few of the sequences (particularly in the case of *H. pylori*; see Discussion). As the mean error rate in DNA sequences is of the order of 10^{-3} or higher (especially when one considers the GC swaps), the sequences in these regions should certainly be verified if one had to comment on the role of the corresponding residues (see discussion of the SpeB sequence from *Synechocystis* PCC6803 below).

Phylogeny trees

This alignment allowed us to construct phylogenies with two commonly employed approaches for phylogenetic reconstructions, i.e. distance and parsimony (maximum likelihood was also used, but did not give significantly different results; data not shown). Fig. 2 displays the trees obtained by using these methods. The trees are, on the whole, rather similar. They are firmly split into two main parts, one corresponding to a majority of identified or putative arginases (ARGI) and the other to agmatinases (SPEB). However, the bootstrap scores within the SPEB cluster are generally poor, which means that a detailed view of the evolution of agmatinases – and, in some parts, of arginases – cannot be obtained from this study. An interesting observation, however, is that the sequence of SpeB from *Bacillus subtilis* (Sekowska *et al.*, 1998) groups in both trees with the corresponding enzymes from the *Archaea*. Most of the sequences from the *Eukarya*, on the other hand, are supposed to be of the ArgI type (no agmatinases have been identified among the vertebrates) and group far from the archaeal ones. The sequences from *Schizosaccharomyces pombe* are also grouped in both trees with the sequences from Gram-negative bacteria (presumably agmatinases; see below). Enzyme activities corresponding to other functions that are related to, but distinct from urea hydrolysis (e.g. HutG encodes forminoglutamase and PahA encodes proclavaminate amidino hydrolase, which synthesizes a precursor of clavulanic acid, etc.) are generally scattered in the tree.

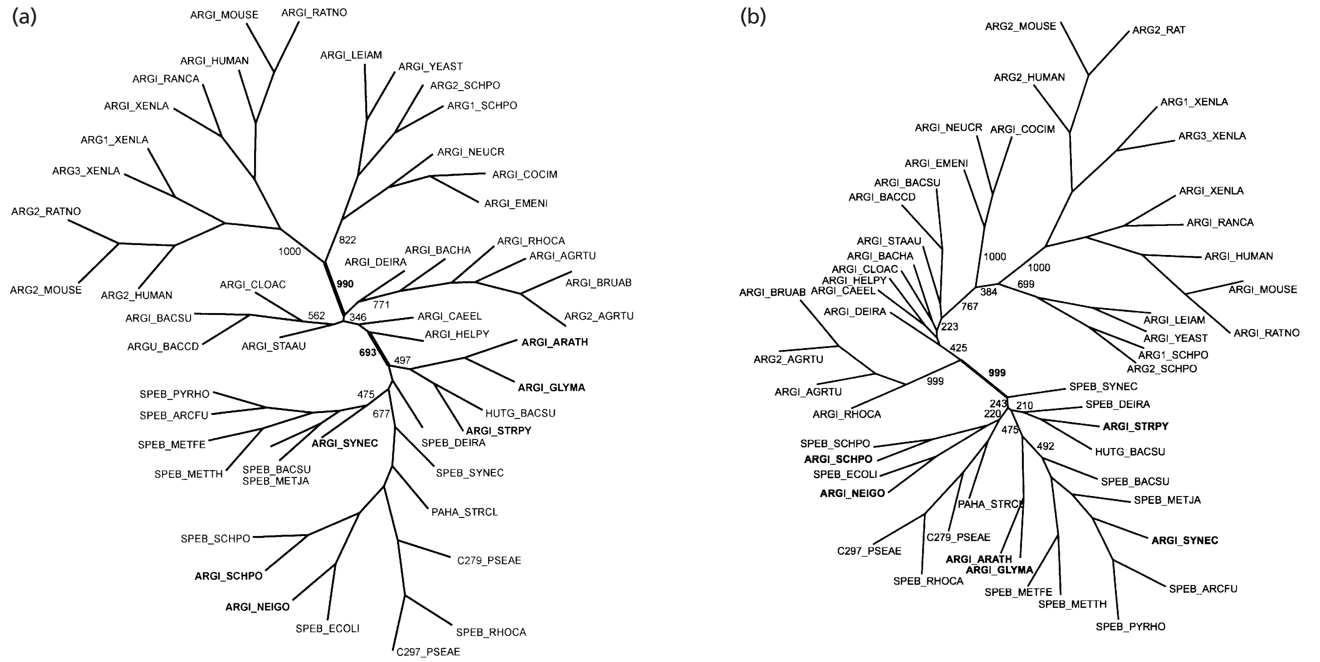


Fig. 2. Reconstructed phylogenies based on the multiple alignment shown in Fig. 1. (a) Distance tree (PROTDIST) after 1000 bootstrap resamplings; (b) most parsimonious consensus tree (PROTPARS) after 1000 bootstrap resamplings. The bootstrap scores are given for some key branches.

Functional identification of pivotal activities

An examination of the phylogeny trees revealed that some enzymes annotated as agmatinases are grouped with arginases, and *vice versa*, not in line with the internal consistency of the tree. To substantiate the validity of the trees, we cloned the presumed arginase genes from *H. pylori*, *N. gonorrheae* and *Synechocystis* PCC6803 (found in a group of archaeal enzymes containing the experimentally identified *B. subtilis* agmatinase) and the presumed agmatinase gene from *Synechocystis* and expressed them in an *E. coli* strain unable to synthesize putrescine (MA255). Urea production was subsequently measured in cell-free extracts with either agmatine or arginine as substrate. If the enzyme is an agmatinase, urea production will occur when agmatine is added; if it is an arginase, urea will be produced in the presence of arginine.

As shown in Table 1, the *H. pylori* enzyme is an arginase. In contrast, the gene product labelled ArgI in *Synechocystis* PCC6803 (Cyanobase; http://www.kazusa.or.jp/cyano/cyano.orig.html) is an agmatinase. The gene product labelled SpeB in this organism might have been a second agmatinase. However, we failed to detect this activity or to identify this gene product as an arginase under conditions where the other gene products produced urea from either arginine or agmatine. It is possible that this is an enzyme used in the degradation of histidine similar to the *hutG* gene product of *B. subtilis*. More likely, it is involved in secondary metabolism comparable to the *pahA* gene product, which is implicated in synthesis of clavulanic acid in *Streptococcus clavuligerus* (see below). The *N. gonorrheae* enzyme labelled ArgI is clearly identified as an agmatinase, not an arginase.

Discrimination between arginases and agmatinases

Arginases and agmatinases release urea and ornithine or putrescine, respectively, from substrates that differ from each other only by the presence of a carboxylate group (in arginine). Because arginases and agmatinases can be divided into two well-defined categories, we investigated whether they could be separated according to a consensus sequence. The manganese-binding sites, as shown above, make an identical consensus in both enzymes. This could be expected since these enzymes display the same catalytic activity. Ten invariant conserved residues are found in the Prosite motif (http://www.expasy.ch/cgi-bin/nicedoc.pl?PDOC00135), modified below, using the data from our study. Among these, six bind to the manganese-ion cofactors (indicated by #):

Pattern 1 [ILV]-X-[FILMV]-G-G-[ED]-H#-X-[ILMV]-[ASTV]-X-[AGP]-X(3)-[AGST];

Pattern 2 [ILMV](2)-X-[FILMVY]-D#-[AS]-H#-X-D#;

Pattern 3 [FHY]-[ILV]-[ST]-[FILMVY]-D#-[ILMV]-D#-X(3)-[APQ]-X(3)-P-[AGS]-X(7)-G.

A fourth conserved pattern also seems to be important:

Pattern 4 [AGSV]-[ACFILMV]-[DE]-[FILMTV]-[AIMTV]-E-[FILMV]-[AGHNS]-[GPS].

If only the arginase side of the tree is retained, one uncovers several further constraints that specify the arginase family. These include restriction to [ILMV] at the third residue in pattern 1 (no F), a conserved W (Y in the widely deviant *H. pylori* sequence; see Discussion) at position 3 in pattern 2 and restriction to H at position 1 in pattern 3 (neither F nor Y). In contrast, several regions generally differ between the ArgI and SpeB families precisely where the patterns are restricted in arginases. In addition, the polypeptide sequence intervening between patterns 2 and 3 is significantly shorter (20 or so residues less) in agmatinases compared to arginases. In agmatinases, position 3 of pattern 2 is more variable than in arginases, H, Q or N being present instead of W (Y). Likewise, the H at position 1 of pattern 3 of arginases is replaced by an aromatic residue in agmatinases.

Once these general features are identified, one may also remark that several sequences appear to differ significantly from the others. This is particularly true in the case of the *H. pylori* arginase. Some sequences also appear to be much longer than the mean. These include certain sequences of *Schizosaccharomyces pombe*, the sequence of *Caenorhabditis elegans* and the sequence of *Synechocystis* PCC6803 (labelled SPEB).

Alignment of gaps in the sequences

Until now, the phylogeny trees constructed were based only on the differences between amino acids at equivalent positions. However, changes yielding gaps in the alignment also correspond to one or more mutational events. In general, gaps specify the insertion or deletion of loops in the three-dimensional structure of the protein (Briozzo *et al.*, 1998). For this reason, their length is usually variable, whereas the place where they occur is fixed. This is due to structural constraints in the architecture of the protein active site(s). It is therefore interesting to consider gaps as the hallmark of some mutational events, noting that conservation of both the presence and the length of the insertions/deletions may indicate some kinship between the corresponding proteins. We therefore constructed a matrix based on the lengths of insertions occurring between well-conserved regions (see Fig. 1). As expected, the insertions/deletions often occur at places where the multiple alignment is less reliable. However, since they are anchored by highly conserved residues, their length is unequivocal. In addition, protein length was also included in the matrix because the length of the protein introduces a constraint in evolution of the same structural nature as the introduction of gaps. This matrix was used to calculate the co-ordinates of the sequences using FCA. The latter enabled us to obtain pairwise distances that were subsequently used as input in the program NEIGHBOR.

Table 1. Arginases, agmatinases and related activities extracted from public data libraries

Sequence name	Organism	Function in database	Predicted function	Identified function
ARG2_AGRTU	*Agrobacterium tumefaciens*	Arginase	Arginase	Arginase*
ARGI_ARATH	*Arabidopsis thaliana*	Arginase	Agmatinase or secondary metabolism†	Arginase*
SPEB_ARCFU	*Archaeoglobus fulgidus*	Agmatinase	Agmatinase	
ARGI_BACCD	'*Bacillus caldovelox*'	Arginase	Arginase	Arginase
ARGI_BACHA	*Bacillus halodurans*	Arginase	Arginase	
ARGI_BACSU	*Bacillus subtilis*	Arginase (RocF)	Arginase	Arginase
SPEB_BACSU	*Bacillus subtilis*	Agmatinase (YwhF)	Agmatinase	Agmatinase
HUTG_BACSU	*Bacillus subtilis*	Formiminoglutamate hydrolase	Secondary metabolism and other functions	
ARGI_BRUAB	*Brucella abortus*	Arginase	Arginase	Arginase*
ARGI_CAEEL	*Caenorhabditis elegans*	Arginase	Arginase	
ARGI_CLOAC	*Clostridium acetobutylicum*	Arginase	Arginase	
ARGI_COCIM	*Coccidioides immitis*	Arginase	Arginase	Arginase*
ARGI_DEIRA	*Deinococcus radiodurans*	Arginase	Arginase	
SPEB_DEIRA	*Deinococcus radiodurans*	Agmatinase	Agmatinase	
SPEB_ECOLI	*Escherichia coli*	Agmatinase	Agmatinase	Agmatinase
ARGI_EMENI	*Emericella nidulans*	Arginase	Arginase	Arginase
ARGI_GLYMA	*Glycine max*	Arginase	Agmatinase or secondary metabolism†	
ARGI_HELPY	*Helicobacter pylori*	Arginase (RocF)	Arginase	Arginase‡
ARGI_HUMAN	*Homo sapiens*	Arginase	Arginase	Arginase
ARG2_HUMAN	*Homo sapiens*	Arginase	Arginase	Arginase
ARGI_LEIAM	*Leishmania americana*	Arginase	Arginase	Arginase
SPEB_METFE	'*Methanococcus fervidus*'	Agmatinase	Agmatinase	
SPEB_METJA	*Methanococcus jannaschii*	Agmatinase	Agmatinase	
SPEB_METTH	*Methanobacterium thermoautotrophicum*	Agmatinase	Agmatinase	
ARGI_MOUSE	*Mus domesticus*	Arginase	Arginase	Arginase
ARG2_MOUSE	*Mus domesticus*	Arginase	Arginase	Arginase
ARGI_NEIGO	*Neisseria gonorrheae*	Arginase	Agmatinase†	Agmatinase‡
ARGI_NEUCR	*Neurospora crassa*	Arginase	Arginase	Arginase
C279_PSEAE	*Pseudomonas aeruginosa*	Unknown	Secondary metabolism and other functions	
C297_PSEAE	*Pseudomonas aeruginosa*	Unknown	Secondary metabolism and other functions	
SPEB_PYRHO	*Pyrococcus horikoshii*	Agmatinase	Agmatinase	
ARGI_RANCA	*Rana catesbeiana*	Arginase	Arginase	Arginase*
ARGI_RATNO	*Rattus norvegicus*	Arginase	Arginase	Arginase
ARG2_RATNO	*Rattus norvegicus*	Arginase	Arginase	Arginase
ARGI_RHOCA	*Rhodobacter capsulatus*	Arginase	Arginase	Arginase
SPEB_RHOCA	*Rhodobacter capsulatus*	Agmatinase	Secondary metabolism and other functions‡	
ARGI_SCHPO	*Schizosaccharomyces pombe*	Arginase	Arginase	Arginase*§
ARG1_SCHPO	*Schizosaccharomyces pombe*	Arginase	Arginase	Arginase*§
ARG2_SCHPO	*Schizosaccharomyces pombe*	Arginase	Arginase	Arginase*§
SPEB_SCHPO	*Schizosaccharomyces pombe*	Agmatinase	Arginase	Arginase*§
ARGI_STAAU	*Staphylococcus aureus*	Arginase	Arginase	
ARGI_STRPY	*Streptococcus pyogenes*	Arginase	Secondary metabolism and other functions†	
PAHA_STRCL	*Streptomyces clavuligerus*	Proclavaminate amidino hydrolase	Secondary metabolism and other functions	Proclavaminate amidino hydrolase
ARGI_SYNEC	*Synechocystis* PCC6803	Arginase	Agmatinase†	Agmatinase†

Table 1 (*cont.*)

Sequence name	Organism	Function in database	Predicted function	Identified function
SPEB-SYNEC	*Synechocystis* PCC6803	Agmatinase	Secondary metabolism or other functions†	Secondary metabolism or other functions‡
ARGI-XENLA	*Xenopus laevis*	Arginase	Arginase	Arginase*§
ARG1-XENLA	*Xenopus laevis*	Arginase	Arginase	Arginase*§
ARG3-XENLA	*Xenopus laevis*	Arginase	Arginase	Arginase*§
ARGI-YEAST	*Saccharomyces cerevisiae*	Arginase	Arginase	Arginase*

* Activity identified by complementation studies.

† Conflict.

‡ Activity identified experimentally in this study.

§ Activity identified collectively as a bulk cell arginase activity.

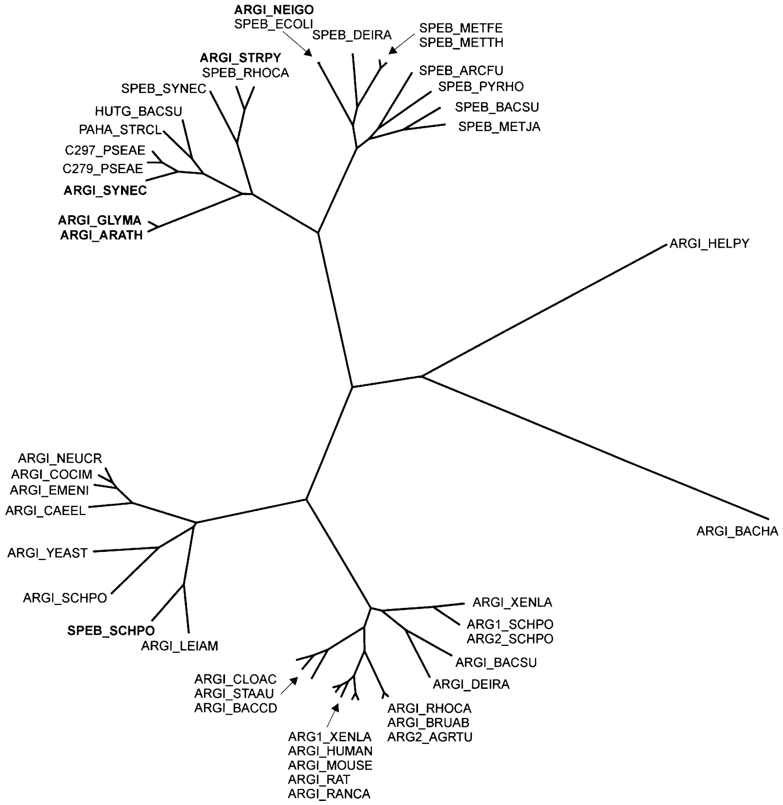


Fig. 3. Gap-tree of ureohydrolases and related enzymes, calculated by factorial analysis of the lengths of the insertions in the multiple alignments (roman numeral in Fig. 1). This is *not* a phylogenetic reconstruction, but a distance tree based on the conservation of the position and length of probable loops in the three-dimensional structure of the proteins.

The resulting distance tree is displayed in Fig. 3. One of its prominent features is that arginases and agmatinases are clearly separated. It should be noted that this separation does not merely come from the fact that the lengths of the proteins were used in the pairwise distance calculation, even if the agmatinases tend to be shorter than the arginases. For example, the protein labelled SPEB_SYNEC is one of the longest in the set (390 aa) and is nevertheless clearly not a member of the arginase family. In addition, it is probable that gap regions III and VI (Fig. 1) are the most discriminant. Interestingly, some of the sequences that were located in a continuum near the junction of the arginase and agmatinase trees when using evolution of the sequences without consideration of gaps, moved to less ambiguous positions. The ambiguous situation of the *H. pylori* enzyme has been solved by demonstrating experimentally that it is an arginase. Remarkably, the enzymes that differ both from arginases and from agmatinases also group together in this tree. The protein labelled SPEB from *Synechocystis* is located next to the agmatinase cluster in a group comprising poorly defined activities (a protein labelled ARGI from a *Streptomyces* sp. and the SPEB protein from *Rhodobacter capsulatus*). Our experiments suggest

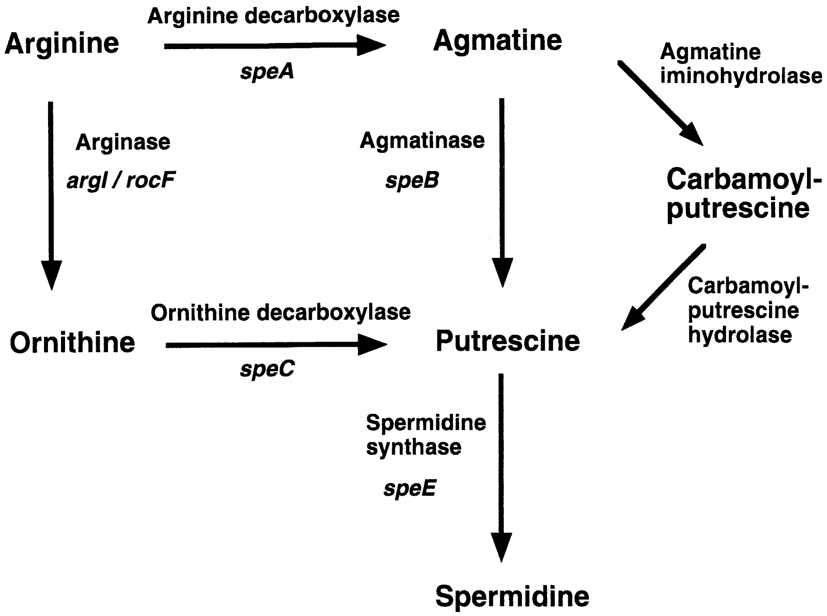


Fig. 4. A summary of polyamine metabolism. Some or all of the activities are present in the *Archaea*, *Bacteria* and *Eukarya*.

that it might not be an agmatinase but an enzyme related to secondary metabolism activities linked to polyamines (Cohen, 1998). Furthermore, the plant enzymes may be a sister group of their cyanobacterial counterpart, labelled ARGI, but that we identified as an agmatinase. Finally, we observed that the *B. subtilis* agmatinase clustered with the archaeal enzymes, as in the previous phylogenetic reconstructions, while the enzymes from *E. coli* and *N. gonorrheae* are now correctly grouped together (the *N. gonorrheae* ARGI enzyme being in fact an agmatinase).

DISCUSSION

Living organisms have been categorized into two architectural domains, according to the way they manage envelopes, membranes and skins. The protozoologist Edouard Chatton separated microbes made of a more or less complex single envelope from those that were nucleated, the prokaryotes and the eukaryotes (Chatton, 1938). Based on a phylogenetic tree of rDNA, Woese proposed in 1977 a bold hypothesis that divided the prokaryotes into two domains, the eubacteria and another one for which he coined the suggestive name 'archaea' (Woese & Fox, 1977). Since then, this paradigm has been substantiated by rDNA phylogeny trees. However, the problem was complicated by phylogenetic analysis of genes encoding the translation and the transcription machinery, when compared to those for ATP synthesis or intermediary metabolism (Ouzounis & Kyrpides, 1994, 1996; Ibba *et al.*, 1997; Koonin *et al.*, 1997; Diaz-Lazcoz *et al.*, 1998; Gupta, 1998a; Koonin & Aravind, 1998; Kyrpides & Woese, 1998; Mayr, 1998). Lack of consistency in data coming from protein sequence analysis, depending on the chosen family, suggested an important involvement of horizontal gene transfer in early evolution (Ribeiro & Golding, 1998; Woese, 1998). Therefore, identification of enzyme activities solely based on such phylogeny trees became questionable. Because nitrogen metabolism must have existed from the very beginning of life, we reinvestigated the tree of ureohydrolases, using a different approach, to evaluate the validity of phylogenies based on implicit functional homology. In this study, we found that these enzymes are split into two major families, agmatinases and arginases. This separation is observed both in standard phylogeny reconstructions and in those using gaps. It corresponds to fine alterations of the enzyme structure to accommodate the subtle differences in their substrates.

The *H. pylori* sequence is of particular interest because, although from a Gram-negative organism, it does not group with *E. coli* and *N. gonorrheae*. Furthermore, the SSEH motif differs significantly from the consensus GGD(or E)H motif present in all other arginases. Using a second type of *H. pylori* ARGI, it is evident that SSEH is specific to these bacteria and does not result from a sequencing error (data not shown). Indeed, the lifestyle of *H. pylori* is very unusual. Possibly, this is reflected in the 'style' of its genome, constraints on the availability of metabolites affecting the composition of its proteins (Danchin *et al.*, 2000). Alternatively, catalytic mechanisms might be affected in a more extreme environment. Variation in the nature of the preferred amino acids (or codons) would account for the rather odd place of the sequence in the non-gapped trees. The sequence from *Bacillus halodurans*, which is also present in bacteria living in unusual environments, is also similarly biased, a GGDC sequence replacing the (almost) universal GGD(or E)H sequence (and it groups with *H. pylori* in the gap tree). Several methods have been devised to deal with composition biases in alignments, for example by taking into account biases in the G+C content of genomes (Tourasse & Gouy, 1997; Foster & Hickey, 1999; Wilquet & Van de Casteele, 1999) or by

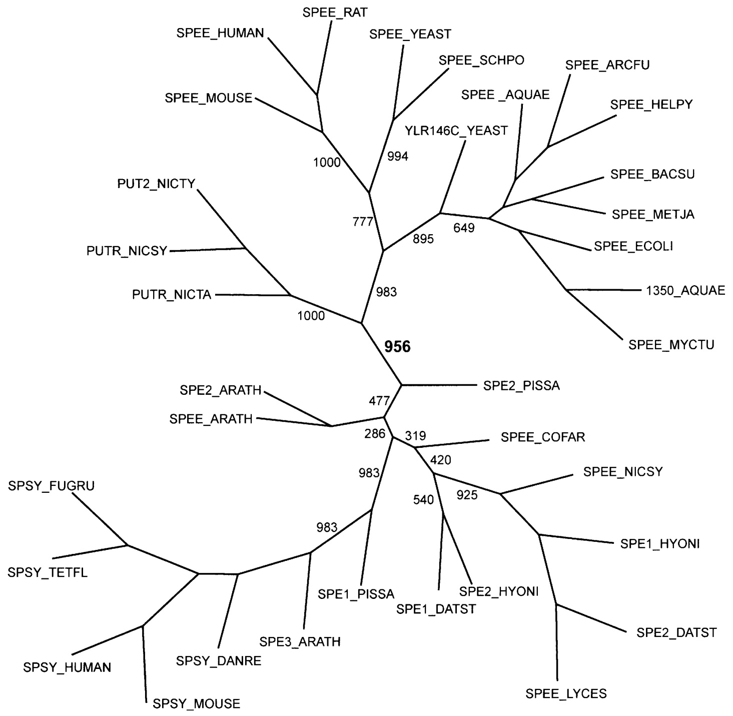


Fig. 5. Most parsimonious consensus tree (PROTPARS) after 1000 bootstrap resamplings for the family of spermidine synthase and related enzymes. Proteins labelled SPSY, SPEE and PUTR are spermine synthases, spermidine synthases and methyltransferases, respectively. In addition to the organisms listed in Table 1, several sequences are from other organisms: MYCTU, *Mycobacterium tuberculosis*; AQUAE, *Aquifex aeolicus*; COFAR, *Coffea arabica*; DANRE, *Danio rerio*; DATST, *Datura stramonium*; FUGRU, *Fugu rubripes*; HYONI, *Hyoscyamus niger*; LYCES, *Lycopersicon esculentum*; NICSY, *Nicotiana sylvestris*; NICTA, *Nicotiana tabacum*; NICTY, *Nicotiana tomentosiformis; PISSA, Pisum sativum*; TETFL, *Tetraodon fluviatilis.*

taking care of long-branch attraction artefacts in the statistical methods used (Lyons-Weiler & Hoelzer, 1997). But, as seen here, this is probably of limited interest when divergence is very ancient. As a consequence, we emphasize the utility of considering gaps when constructing trees, especially for proteins that have been evolving for a very long time (Gupta, 1998b). Gaps are less sensitive to the nature of the genetic code or amino acids, but only to the relationship between the three-dimensional structure of a protein and its function. Besides allowing better prediction of protein function, our alignment identified residues of major importance for the two binding sites of the manganese co-factor as well as residues and regions separating arginases from agmatinases.

Although not substantiated experimentally, some of the sequences which differ from arginases may have other important activities (Cohen, 1998). Secondary metabolism activities or degradation of histidine are placed on the agmatinase side in the trees. Interestingly, in the highly conserved region, [FYMLIV]-[MLIV]-[WHYNQV]-[FMLIV]-D-[AS]-H, the alanine or serine residue before the final histidine is replaced by an arginine in the *Synechocystis* SPEB sequence (clustered in the 'secondary metabolism' sequences). This alteration replaces a small non-polar or polar amino acid with a very large basic one. We did not identify the corresponding enzyme as an arginase or an agmatinase. This prompted us to check whether this was due to a sequencing error: arginine and alanine are coded by CGN or GCN codons and it is well known that C and G residues often migrate at the same rate on gels, resulting in frequent GC swaps in the final sequence. After PCR amplification of the corresponding region from the chromosome, we found that the sequence is exact (data not shown). We therefore propose that this residue is involved in defining a new specificity for the enzyme, probably for some reaction of secondary metabolism (cyanobacteria are known to produce secondary metabolites that may derive from molecules related to polyamines; Cohen, 1998). This observation illustrates the identification of an amino acid residue that may play a discriminant role in catalysis and could be a target for oriented *in vitro* evolution of enzyme activities.

The question of the origin and fate of arginine early in evolution is important because it is related both to the fixation of nitrogen and to the universality of the genetic code. Remarkably, the *Archaea* are only present in the agmatinase tree. Also, the pathways leading to putrescine (the primary polyamine, Fig. 4) involve ornithine, an amino acid not incorporated into proteins. The cell concentration of ornithine (an analogue of lysine) must therefore be finely controlled to avoid misincorporation into proteins. As a consequence, the arginase pathway is likely to have evolved later than that involving arginine decarboxylation and agmatinase. In this respect, one should note that the consensus sequence corresponding to pattern 2 is highly variable in the non-arginase sequences while it is extremely conserved in arginases, suggesting indeed that these sequences are of more recent descent. In particular, the W residue in WXDAHXD could be chosen as a discriminant residue to identify arginases (Fig. 1). One can therefore confidently assume that agmatinases predated arginases. The latter would have appeared in the *Bacteria* by recruitment of a wide specificity agmatinase and then transferred to eukaryotes (perhaps through transfer from mitochondria).

Plants have sequences labelled as arginases, but they are grouped with agmatinases in all three tree types. An arginase activity has been identified indirectly by complementation studies of a yeast mutant (Krumpelman *et al.*, 1995). Plants possess an arginine decarboxylase (Klein *et al.*, 1999), therefore they make agmatine, but the presence of agmatinase must be substantiated experimentally, because they would be the only eukaryotes with this feature. This may reflect a gene transfer from chloroplasts to the nucleus. Under the selective pressure of sexual reproduction, genes in the nucleus avert the deleterious effect of Muller's ratchet in organelles (Haigh, 1978; Bergstrom & Pritchard, 1998). This hypothesis is supported by the experiment where we showed that *Synechocystis* ARGI is in fact an agmatinase. However, the plant sequences may correspond to enzymes involved in secondary metabolism, known to be very active in plants. Noticeably, they are grouped in the gap tree near the enzymes involved in secondary metabolism. The absence of agmatinases in known sequences from the *Eukarya* is puzzling, particularly because agmatine does exist in the *Eukarya* (Sastre *et al.*, 1998; Reis & Regunathan, 1999). Recently, an agmatinase activity from the mitochondrial matrix has been identified by Sastre and co-workers (Sastre *et al.*, 1996), but we do not know whether it belongs to the same family of manganese enzymes. This is unlikely because the whole genome sequence of several eukaryotes is known. In fact, the *Eukarya* recruited another activity, agmatine iminohydrolase, to fulfil this requirement (Park & Cho, 1991; Klein *et al.*, 1999; see Fig. 4). This is not an uncommon feature among the *Eukarya* (see for example the case of aminolevulinate synthase; Duncan *et al.*, 1999). The results presented here indicate that the whole question should be reinvestigated in depth.

Remarkably, the only Gram-negative arginase is that from *H. pylori* (substantiated experimentally in this study). The *Bacteria* are split into several groups in the gap tree. In the case of agmatinases, those of *B. subtilis* and the *Archaea* go together. A group comprising the *Synechocystis* PCC6803 agmatinase (wrongly annotated as an arginase) together with enzymes involved in secondary metabolism, and in broader specificity hydrolases, is also prominent in the gap tree. If one reasons in terms of acquisitive evolution (Thompson & Krawiec, 1983), assuming that an ancient activity with broad specificity has been progressively specified during evolution (Jensen, 1976; Danchin, 1989; Roy, 1999), then the origin of the ureohydrolase tree would be near the PahA HutG group, leading to agmatinases and enzymes involved in secondary metabolism and, subsequently, to arginases. To substantiate the grouping of *B. subtilis* and archaeal polyamine enzymes, we performed a phylogenetic reconstruction of spermidine synthases. These enzymes are less universally distributed than ureohydrolases. They exert their function downstream from the synthesis of putrescine, the precursor of polyamines. Fig. 5 displays the most parsimonious tree obtained in this case. Again, *B. subtilis* groups with its archaeal counterparts, while the *Eukarya* form another well separated group. In addition, we find not only a group made of spermine synthases (SPSY and SPEE), which use spermine as the substrate instead of spermidine, but also a group of related methyltransferases, which use *S*-adenosylmethionine rather than decarboxylated *S*-adenosylmethionine as a substrate (PUTR).

Because it seems difficult to transfer systematically the corresponding genes horizontally unless at a very early stage of evolution, an examination of polyamine metabolism suggests that the empire of prokaryotes may be more a continuum than a clear-cut division between the two domains of life, perhaps reflecting an early pool of metabolic genes that would have subsequently been frozen after differentiation into specific families. The data presented here suggest that some machinery, presumably involving RNA molecules, contained polyamines at a very early time in the evolution of life. In this machinery, the ancestors of some significant part of Gram-positive bacterial and archaeal genomes appeared to cluster together. This should be taken into consideration when trying to reconstruct the tree of origin of life.

ACKNOWLEDGEMENTS

We wish to thank M. Berlyn and the *E. coli* Genetic Stock Center, N. Tandeau de Marsac, H. De Reuse and M.-K. Taha for the gift of strains or chromosomal DNA samples. We wish to express particularly our thanks to Professor Chun-Kwun Wun for critical and attentive reading of this manuscript and to Ivan Moszer for helping us in its preparation. This work was supported in part by European Union contract BIOTECH B104-CT96-0655.

REFERENCES

Baldauf, S. L. & Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci USA* **90**, 11558–11562.

Bergstrom, C. T. & Pritchard, J. (1998). Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. *Genetics* **149**, 2135–2146.

Briozzo, P., Golinelli-Pimpaneau, B., Gilles, A. M., Gaucher, J. F., Burlacu-Miron, S., Sakamoto, H., Janin, J. & Barzu, O. (1998). Structures of *Escherichia coli* CMP kinase alone and in complex with CDP: a new fold of the nucleoside monophosphate binding domain and insights into cytosine nucleotide specificity. *Structure* **6**, 1517–1527.

Chatton, E. (1938). *Titres et Travaux Scientifiques* (1906–1937). Sottano, Italy: Sete.

Cohen, S. (1998). *A Guide to the Polyamines*. Oxford: Oxford University Press.

Cox, E. C. & Yanofsky, C. (1967). Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc Natl Acad Sci USA* **58**, 1895–1902.

Danchin, A. (1989). Homeotopic transformation and the origin of translation. *Prog Biophys Mol Biol* **54**, 81–86.

Danchin, A., Guerdoux-Jamet, P., Moszer, I. & Nitschké, P. (2000). Mapping the bacterial cell architecture into the chromosome. *Philos Trans R Soc B Biol Sci* **355**, 179–190.

Diaz-Lazcoz, Y., Aude, J., Nitschké, P., Chiapello, H., Landès-Devauchelle, C. & Risler, J. (1998). Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. *Mol Biol Evol* **15**, 1548–1561.

Duncan, R., Faggart, M. A., Roger, A. J. & Cornell, N. W. (1999). Phylogenetic analysis of the 5-aminolevulinate synthase gene. *Mol Biol Evol* **16**, 383–396, erratum following 883.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.57c. Seattle: Department of Genetics, University of Washington.

Fitch, W. M. & Yasunobu, K. T. (1975). Phylogenies from amino acid sequences aligned with gaps: the problem of gap weighting. *J Mol Evol* **5**, 1–24.

Foster, P. G. & Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* **48**, 284–290.

Granick, S. (1957). Speculations on the origin and evolution of photosynthesis. *Ann NY Acad Sci* **69**, 292–308.

Gupta, R. S. (1998a). Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435–1491.

Gupta, R. S. (1998b). What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Mol Microbiol* **29**, 695–707.

Haigh, J. (1978). The accumulation of deleterious genes in a population – Muller's Ratchet. *Theor Popul Biol* **14**, 251–267.

Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**, 383–402.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Appl Statistics* **23**, 340–353.

Hirshfield, I. N., Rosenfeld, H. J., Leifer, Z. & Maas, W. K. (1970). Isolation and characterization of a mutant of *Escherichia coli* blocked in the synthesis of putrescine. *J Bacteriol* **101**, 725–730.

Ibba, M., Morgan, S., Curnow, A. W., Pridmore, D. R., Vothknecht, U. C., Gardner, W., Lin, W., Woese, C. R. & Soll, D. (1997). A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **278**, 1119–1122.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **30**, 409–425.

Klein, R. D., Geary, T. G., Gibson, A. S. & 8 other authors (1999). Reconstitution of a bacterial/plant polyamine biosynthesis pathway in *Saccharomyces cerevisiae*. *Microbiology* **145**, 301–307.

Koonin, E. V. & Aravind, L. (1998). Genomics: re-evaluation of translation machinery evolution. *Curr Biol* **8**, R266–R269.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* **25**, 619–637.

Krumpelman, P. M., Freyermuth, S. K., Cannon, J. F., Fink, G. R. & Polacco, J. C. (1995). Nucleotide sequence of *Arabidopsis thaliana* arginase expressed in yeast. *Plant Physiol* **107**, 1479–1480.

Kyrpides, N. C. & Woese, C. R. (1998). Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B alpha-beta-delta subunit families. *Proc Natl Acad Sci USA* **95**, 3726–3730.

Lebart, L., Morineau, A. & Warwick, K. A. (1984). *Multivariate Descriptive Statistical Analysis*. New York: Wiley.

Lyons-Weiler, J. & Hoelzer, G. A. (1997). Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. *Mol Phylogenet Evol* **8**, 375–384.

Mayr, E. (1998). Two empires or three? *Proc Natl Acad Sci USA* **95**, 9720–9723.

Morgenstern, B., Frech, K., Dress, A. & Werner, T. (1998). DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**, 290–294.

Ouzounis, C. A. & Kyrpides, N. C. (1994). On the evolution of arginases and related enzymes. *J Mol Evol* **39**, 101–104.

Ouzounis, C. & Kyrpides, N. (1996). The emergence of major cellular processes in evolution. *FEBS Lett* **390**, 119–123.

Page, D. M. (1996). TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.

Park, K. H. & Cho, Y. D. (1991). Purification of monomeric agmatine iminohydrolase from soybean. *Biochem Biophys Res Commun* **174**, 32–36.

Perozich, J., Hempel, J. & Morris, S. M., Jr (1998). Roles of conserved residues in the arginase family. *Biochim Biophys Acta* **1382**, 23–37.

Reis, D. J. & Regunathan, S. (1999). Agmatine: an endogenous ligand at imidazoline receptors is a novel neurotransmitter. *Ann NY Acad Sci* **881**, 65–80.

Ribeiro, S. & Golding, G. B. (1998). The mosaic nature of the eukaryotic nucleus. *Mol Biol Evol* **15**, 779–788.

Roy, S. (1999). Multifunctional enzymes and evolution of biosynthetic pathways: retro-evolution by jumps. *Proteins* **37**, 303–309.

Sastre, M., Regunathan, S., Galea, E. & Reis, D. J. (1996). Agmatinase activity in rat brain: a metabolic pathway for the degradation of agmatine. *J Neurochem* **67**, 1761–1765.

Sastre, M., Galea, E., Feinstein, D., Reis, D. J. & Regunathan, S. (1998). Metabolism of agmatine in macrophages: modulation by lipopolysaccharide and inhibitory cytokines. *Biochem J* **330**, 1405–1409.

Sekowska, A., Bertin, P. & Danchin, A. (1998). Characterization of polyamine synthesis pathway in *Bacillus subtilis* 168. *Mol Microbiol* **29**, 851–858.

Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy: the Principles and Practice of Numerical Classification*. San Francisco: Freeman.

Thompson, L. W. & Krawiec, S. (1983). Acquisitive evolution of ribitol dehydrogenase in *Klebsiella pneumoniae*. *J Bacteriol* **154**, 1027–1031.

Tourasse, N. J. & Gouy, M. (1997). Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. *Mol Biol Evol* **14**, 287–298.

Wilquet, V. & Van de Casteele, M. (1999). The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Res Microbiol* **150**, 21–32.

Woese, C. R. (1998). The universal ancestor. *Proc Natl Acad Sci USA* **95**, 6854–6859.

Woese, C. R. & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* **74**, 5088–5090.

Yamamoto, S., Nakao, H., Yamasaki, K., Takashina, K., Suemoto, Y. & Shinoda, S. (1988). Activities and properties of putrescine-biosynthetic enzymes in *Vibrio parahaemolyticus*. *Microbiol Immunol* **32**, 675–687.

Received 25 January 2000; revised 1 May 2000; accepted 5 May 2000.