GENOME LETTERS 2003, 2, 8–15
www.aspbs.com/genomelett

RESEARCH ARTICLE

Retrieving the Whole Set of Protein Modules of *Campylobacter jejuni* and *Helicobacter pylori*

Quentin Sculo, Olivier Lespinet, and Bernard Labedan*

Évolution Moléculaire et Génomique, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris-Sud, Bâtiment 409, 91405 Orsay Cedex, France

(Received: 19 May 2003; accepted: 12 June 2003)

ABSTRACT: A complete description of the whole set of proteins encoded by the complete genomes of *Campylobacter jejuni* and the two *Helicobacter pylori* strains has been deduced from an exhaustive comparison with the proteomes of the 23 other available proteobacteria. For each protein, we have determined its modular structure (identification of all structural segments of homology), its phylogenetic profile (listing of all species containing at least one ortholog), and its class (unique to its species, paralog only, ortholog only, paralog and ortholog). The exhaustive comparison allowed us to create a list of a limited set of genes that are universal to the 26 proteobacteria studied. Many of these genes encode essential functions, defining some core of the (proteo)bacterial life. Moreover, at least 64 of these universal genes in *C. jejuni* and 43 in *H. pylori* have been annotated as encoding putative or hypothetical proteins. A few cases are presented to illustrate how our homology approach may be helpful for gaining insights into protein function, especially when no experimental data are available.

Keywords: Structural Segment of Homology, Gene Duplication, Gene Fusion, Proteobacteria Epsilon, Minimal Genome, Essential Genes, Metal-dependent Hydrolase, Methylase.

The systematic sequencing of whole microbial genomes has generated a huge amount of putative, potential, or hypothetical proteins (some 25,000 orphans according to a recent review [1]). In the absence of any experimental data, assigning a function is a process currently limited to annotation by homology, a somewhat hazardous task. One of the complexities that must be unraveled to assess the validity of these homology relationships between genes and proteins is the modular structure of proteins [2, 3]. According to our working model [4, 5], we use the term *module* to mean a long (mean size 220 amino acids) structural segment of homology found in prokaryotic proteins. As already underlined [3], the dissection of proteins into their different constitutive elements may be crucial to clearing up the mechanism of combinatorial construction of a gene from ready-made basic components.

In the conceptual framework based on our module approach, we are constantly updating an exhaustive comparison of the whole set of proteins (hereafter called proteomes) encoded by completely sequenced genomes of microbial species. Accordingly, we have designed methodological approaches to identify all modules, to group them in families in order to number the putative ancestral genes that were at the origin of the present-day proteins, and to count the different events of gene duplication and gene fusion that helped to shape these present-day proteins. This paper will focus on a few case studies illustrating how these concepts may help to illuminate various aspects of the biology of poorly known pathogens such as *Campylobacter jejuni* and *Helicobacter pylori*.

To deal with the deluge of data released by the whole-genome sequencing programs, we have devised a suite of automatic programs that allows us to detect in a few steps the whole set of modules that constitute the proteome of any organism for which the complete DNA sequence is available [5]. Recently, this suite has been improved with new, more efficient programs (Sculo and Lespinet, unpublished observations) corresponding to the following three steps.

First, for each pair of species, we compare their proteomes in order to detect homologous segments, using thresholds for both evolutionary distance and alignment length. The evolu-

*Author to whom correspondence should be addressed.

1537-3053/2003/02/8/8/$17.00+.25 doi:10.1166/gl.2003.000

tionary distance separating two proteins deriving from a common ancestor and displaying significant sequence similarities is given in PAM units, a PAM unit being defined as the number of *accepted* point mutations per 100 residues separating two sequences [6]. The frequency with which any particular pair of (mutated) amino acids occurs at a given position in two properly aligned homologous proteins can be used as a PAM score to evaluate the evolutionary distance separating the two proteins. It has been shown [6] and frequently confirmed (see, for example, [7]) that for many comparisons the best scoring systems for detecting distant homologous proteins correspond to PAM 250 scores. Using a rationale based on the information theory, Altschul [8, 9] further showed that, to be statistically significant, an alignment of sequences separated by a distance of 250 PAM units must be longer than 83 residues. Therefore, to define the significance of sequence similarities in terms of putative homology between distant proteins, we adopted the following two limits: any sequence alignment must extend for at least 80 residues and have a PAM distance of less than 250 PAM units. Such an exhaustive approach allows us to collect—in one step—all paralogous (intragenomic comparison) and all orthologous (intergenomic comparison) pairs of aligned protein sequences. To retrieve the whole set of modules we adapted, as already described [4, 5], the Darwin *AllAll* program [7, 10], which uses successively the dynamic programming algorithms of Needleman and Wunsch [11] and Smith and Waterman [12], with an optimized PAM 250 matrix as a substitution score matrix [6]. Indeed, this Darwin program, based on a maximum likelihood approach, has been found to be very powerful in detecting all modules of interest, including distant homologs [2–5]. Then, the program *Module* classifies modules found in each match according to their length and location inside the aligned proteins. For example, an alignment between the first third of protein A and the last third of protein B will be interpreted as a module A_1_3 matching a module B_3_3. *In a second step*, the program *Families* gathers automatically into one family all modules that are related by a chain of similarities, collecting all relatives of both members of each pair until no further pairwise relationship is found. To meet our different experimental needs, various kinds of families were assembled, grouping either paralogs of each species, or orthologs for each pair of genomes, or all homologs for different groupings of species (see examples below). *In a third step*, the program *Protein* compiles, for each compared protein, all homology information in order to define its modular structure (identification of all events of gene duplication and gene fusion), and phylogenetic profiles (listing of all species containing at least one ortholog) for each module and for the entire protein. Moreover, the program *Protein* determines which of the four different classes previously defined [5] each gene/protein belongs to: the first two categories correspond to proteins that are found in only one species (sp) and either have a paralog (*para-sp*) or are unique to their species (*uni-sp*). The last two categories correspond to orthologous proteins that either have a paralog (*para-ortho*) or are unique to their species (*uni-ortho*). The distribution of these different classes appears to be very similar in the case of the two proteobacteria epsilon, as shown in Figure 1. In particular, the respective proportions of *para-ortho* and *uni-ortho* classes are seemingly equivalent. Similar data are obtained for other pathogens (e.g., *Neisseria meningitidinis*, *Coxiella burnetii*) whose genomes belong to the same range of size. For smaller genomes such as the obligatory intracellular pathogen *Rickettsia conorii*, the relative proportion of the *para-ortho* class is strongly lowered, whereas the *uni-sp* class is remarkably high. In contrast, nonpathogenic bacteria display a large excess of *para-ortho* proteins and a larger size for the *para-sp* class, as seen for both facultative pathogens (e.g., *Escherichia coli*, *Pseudomonas aeruginosa*) and nonpathogenic species (e.g., *Bradyrhizobium japonicum*). These contrasted distributions confirm several of our previous findings [5, 13], which suggested that many pathogens have reduced their genome size by preferentially diminishing the size of their families of paralogous genes.

To illustrate how our homology approach may help to gain new insights, we will focus on a limited set of data we obtained when comparing the three proteobacteria epsilon (*C. jejuni* and both *H. pylori* strains) with all of the other available proteobacteria (Table I): eight alpha (including three pathogens), two beta (both pathogens), and thirteen gamma (including seven pathogens and two symbionts).

Exhaustive comparison of the 89,555 proteins encoded by the 26 available proteobacterial genomes gave 303,529 modules present in 78,659 homologous proteins (87.8% of

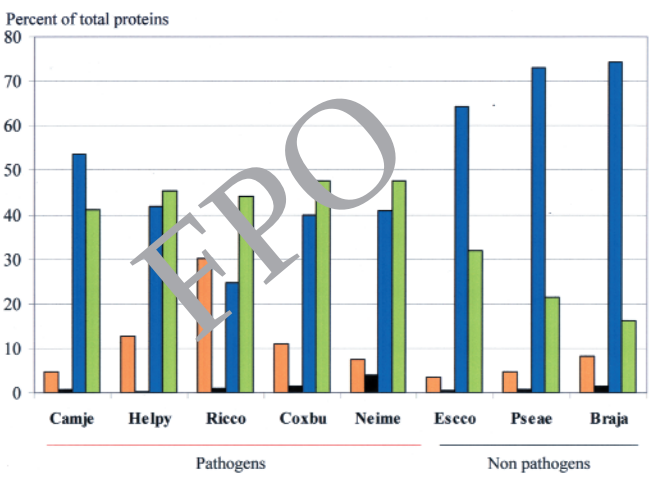

Figure 1. Distributions of the different classes of proteins encoded by various proteobacteria. Percentage of genes found in each class defined in the text (orange is for the *uni-sp* gene class, black for the *para-sp* gene class, blue for the *para-ortho* gene class, and green for the *uni-ortho* gene class) are shown for pathogenic and nonpathogenic proteobacteria. Species name abbreviations are as follows: Camje (*Campylobacter jejeuni*), Helpy (*Helicobacter pylori 26695*), Ricco (*Rickettsia conorii*), Coxbu (*Coxiella burnetii*), Neime (*Neisseria meningitidis*), Escco (*Escherichia coli*), Pseae (*Pseudomonas aeruginosa*), and Braja (*Bradyrhizobium japonicum*).

Table I. List of the 26 protebacteria used for the exhaustive comparison of their proteins.

Species name and strain	Short name	Genome accession number 00	Proteobacteria group	Pathogenic state	Interacting with
Agrobacterium tumefaciens C58	Agrtu	3062/63	Alpha	Facultative	Plants
Bradyrhizobium japonicum	Braja	4463	Alpha	No	Plants
Brucella melitensis 16M	Brume	3317/18	Alpha	Obligatory	Mammals
Caulobacter crescentus	Caucr	2696	Alpha	No	
Mesorhizobium loti	Mieslo	2678	Alpha	No	Plants
Rickettsia conorii Malish 7	Ricco	3103	Alpha	Obligatory, intracellular	Mammals
Rickettsia prowazekii Madrid E	Ricpr	0963	Alpha	Obligatory, intracellular	Mammals
Sinorhizobium meliloti 1021	Sinme	3037/47/78	Alpha	No	Plants
Neisseria meningitidis MC58	Neime	3112	Beta	Obligatory	Mammals
Ralstonia solanacearum	Ralso	3295	Beta	Obligatory	Plants
Campylobacter jejuni	Camje	2163	Epsilon	Obligatory	Mammals
Helicobacter pylori 26695	Helpy	0915	Epsilon	Obligatory	Mammals
Helicobacter pylori J99	HelpJ	0921	Epsilon	Obligatory	Mammals
Buchnera species APS	Buchn	2528	Gamma	Symbiont	Insects
Coxiella burnetii	Coxbu	2971	Gamma	Obligatory	Mammals
Escherichia coli K12 MG1655	Escco	0913	Gamma	Facultative	Mammals
Haemophilus influenzae Rd	Haein	0907	Gamma	Obligatory	Mammals
Pasteurella multocida PM70	Pasmu	2663	Gamma	Obligatory	Mammals/birds
Pseudomonas aeruginosa PAO1	Pseae	2516	Gamma	Facultative	Plants/mammals
Salmonella enterica subsp. enterica serovar Typhi	Salen	3198	Gamma	Facultative	Mammals
Shewanella oneidensis MR-1	Sheon	4347/49	Gamma	No	
Vibrio cholerae N16961	Vibch	2505/06	Gamma	Obligatory	Mammals
Wigglesworthia brevipalpis	Wigbr	4344	Gamma	Symbiont	Insects
Xanthomonas campestris pv. citri str. 306	Xanax	3919	Gamma	Obligatory	Plants
Xylella fastidiosa 9a5c	Xylfa	2488/90	Gamma	Obligatory	Plants
Yersinia pestis CO-92 Biovar Orientalis	Yerpe	3143	Gamma	Obligatory	Mammals

the total proteins). Of these modules, limited sets present in *C. jejuni* or *H. pylori* have been found to have orthologs in all of the 25 other proteobacteria. As shown in Table II, these figures are species-dependent and fall into two main categories. As many as 233 modules in *C. jejuni* and 201 in each *H. pylori* strain correspond to a unique segment of homology, the huge majority of them being entire proteins. The minor category corresponds to more complex cases where a protein appears to be universal because, besides having homologs in a large majority of the other proteobacteria, only one of its modules has orthologs in the missing species. An example of this minor category is the *C. jejuni dnaA*, which aligns along its full length with the other *dnaA* present in 24 proteobacterial species, except in the case of *W. brevipalpis*, where the homology is limited to module 1. Another, more complex case is *rplC*, with a full-length alignment limited to 11 species and a homology limited to module 2 in the 14 remaining species (Agrtu, Buchn, Caucr, Escco, Haein, Meslo, Neima, Pasmu, Pseae, Ricpr, Salen, Wibgr, Xylfa, Yerpe). Many of these universal genes also have paralogs in both *C. jejuni* (239) and *H. pylori* (165 in strain 26695, and 161 in strain J99).

Table II further shows that a significant proportion (210, 238, and 198 in *C. jejuni*, *H. pylori* 26695, and J99, respectively) of these universal genes play essential roles in basic processes such as DNA replication, transcription, and translation; cell wall biosynthesis and cell division; metabolism; and active transport. This is strikingly reminiscent of the known universal genes that have been described by different approaches as being the core of bacterial life (see, for example, [14]). Thus, our *in silico* approach may help to define the whole set of genes encoding essential functions, including those that deserve to be further ascertained by experimental studies.

Table II. Modular structure and annotation of the genes/proteins universally conserved in proteobacteria.

Module type	Camje	Helpy	HelpJ
1_1	226	197	196
1_2	3	2	2
2_2	4	2	3
Total unique	233	201	201
Complex cases	94	66	66
Total	327	267	267
Genes with an assigned function	210	238	198
Unknown modules	64	12	43
Unknown complex cases	53	17	26

Indeed, as summarized in the last two lines of Table II, a significant proportion of these omnipresent genes have no known function and have been annotated as putative or

hypothetical in *C. jejuni* and *H. pylori*. Therefore, we can infer that these universal genes for which we do not have any information are probably encoding important functions.

The 12, 43, and 64 unknown modules present in *H. pylori* 26695, *H. pylori* J99, and *C. jejuni*, respectively, are listed in Table III (after deletion of the duplicates) and arranged by their presence in one, two, or three species. Many of these genes are orthologs, and this was systematically checked by compiling for each *C. jejuni* protein its best reciprocal ortholog in each *H. pylori* strain, and conversely. However, we find 23 cases where it was difficult to assess which is the "true" ortholog when one protein has two relatives at closely similar PAM distances. Table III further shows that this simple search for the best ortholog often makes it possible to suggest by transitivity a function for the unknown proteins. For example, HP1335, which is the ortholog of a protein annotated as putative TrmU in *H. pylori* J99 and TrmU in *C. jejuni*, is most probably the tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase of *H. pylori* 26695. Such deductions have already been added in curated databases for 23 of the modules listed in Table III. Accordingly, only 87 of 119 universal modules remain annotated as putative or hypothetical, including 2, 20, and 42 genes in *H. pylori* 26695, *H. pylori* J99 and *C. jejuni*, respectively, 2 in both *H. pylori* strains, 13 in both *H. pylori* J99 and *C. jejuni*, and 8 in the three epsilon.

To go a step further and to illustrate how helpful our approach might be in improving the annotation of such essential proteins, we focused on these genes which are universal to the 26 proteobacteria and which remain annotated as *hypothetical* (= no putative function suggested) proteins in *C. jejuni*. Three of seven have recently been reannotated in well-curated databases with assigned functions (we observed that our data are in full agreement with these reannotations). For other universal *hypothetical* proteins, things are not so simple.

For example, the small protein Cj0121 (orthologs in the two *H. pylori* strains are HP1160 and JHP1087, respectively) has been reannotated as hypothetical in the current versions of SwissProt and Pfam, but it has been classified as a member of the COG0319 grouping of various predicted metal-dependent hydrolases. We further used the validation step (previously described in [5]) of our experimental approach in an attempt to further challenge this unclear case. Accordingly, we reconstructed an evolutionary tree from the multiple alignment of Cj0121 and its orthologs, using an adaptation of the PhyloTree program of the Darwin package [7, 10]. Figure 2 shows that Cj0121 and its *H. pylori* orthologs are forming a clade with three gram-positive bacteria (*Clostridium tetani*, *Bacillus cereus*, and *Thermoanaerobacter tengcongensis*). Since, in each of these recently published genomes, the respective homologous proteins have been annotated as metal-dependent hydrolase (TTE0972) or metal-binding proteins (ctc0201, bc4300), such a tree topology would suggest that Cj0121, as well as HP1160 and JHP1087, is indeed a potential metal-binding protein. Note that this tree topology further suggests that the epsilon proteobacteria may have acquired their metal-binding proteins by some event of lateral transfer from an ancestral gram-positive bacterium, inasmuch as the addition of more gram-positive bacteria gave the same topology (data not shown). Another interpretation is that all of these proteins have metal-binding properties, inasmuch as three other proteins have also been annotated accordingly (Fig. 2).

Let's turn now to a more complex case, that of the bimodular protein Cj0495. Figure 3 shows that Cj0495 affords complex homology relationships with various partners. We detected significant matches only between its module 1 and either the module 1 of its paralog Cj0722c or different modules of various proteobacterial proteins. Moreover, the Cj0722c protein, annotated as a putative DNA methylase, is aligning along its full length with the putative protoporphyrinogen oxidase HemK of many proteobacteria, in apparent agreement with a recent paper [15] which showed that there are significant similarities between protein and DNA methyltransferases. To better understand the complex relationships occurring between Cj0495 and its different orthologous and paralogous relatives, we aligned its module 1 with the homologous modules (the yellow modules in Fig. 3) to reconstruct an evolutionary tree by the same experimental approach as already described. Figure 4 shows that this tree is made of two well-separated subtrees. In the subtree located in the upper part and made essentially of the HemK family and containing the SAM-dependent methyltransferase NMB1232 of *N. meningitidis*, we found the putative DNA methyltransferase Cj0722c branching on a node common with the *H. pylori* HemK. In the subtree located in the lower part, we find Cj0495 branching on a node common with its two unknown *H. pylori* orthologs. This lower subtree contains a set of proteins that are annotated as either hypothetical or predicted methyltransferases. Note that the three ribosomal RNA small subunit methyltransferases C (rsmC) present in Xanca, Meslo, Agrtu form a clade, strongly suggesting that the *S. meliloti* SMc00325 is also a rRNA small subunit methyltransferase C. Thus, Figure 4 shows a clear evolutionary separation between the two paralogs Cj0722c and Cj0495 and their respective partners, which seems to concur with some functional separation between the different kinds of methyltransferases. This suggests that Cj0495 would be a methyltransferase but is probably not a HemK protein, inasmuch as it is found to be closer to the two *H. pylori* HP1504 and jhp1397 proteins than to the two HemK-like HP0381 and hemG.

In conclusion, we think that our modular approach may help to improve the interpretation of many unclear cases where annotation has been difficult to assess. The indications obtained remain mainly suggestive and must be interpreted with caution. However, we expect that such an approach, firmly rooted in an evolutionary basis, will increase in efficiency as the soundness of integration of the data resulting from our exhaustive comparisons progressively improves.

Table III. Modules universally conserved in proteobacteria and annotated as putative or hypothetical in epsilon strains.

Annotated as putative or hypothetical in	*H. pylori 26295* Hp	*H. pylori J99* jHp	*C. jejuni* Cj	Comments
Hp	HP0890	jhp0823	?	Undecidable "true" ortholog
Hp	HP1061	jhp0364	mttB	
jHp	HP0207	jhp0193	mrp	
jHp	HP0613	jhp0300	?	Undecidable "true" ortholog
jHp	HP0480	jhp0432	typA	
jHp	HP0600	jhp0547	?	Undecidable "true" ortholog
jHp	HP0715	jhp0653	Cj0669	
jHp	HP0743	rodA_1	mrdB	RodA (in databases)
jHp	HP1035	flhF	flhF	FlhF(in databases)
jHp	HP0853	jhp0789	Cj0426	
jHp	HP0885	jhp0817	Cj0801	
jHp	HP0955	lgt	lgt	Lgt (in databases)
jHp	HP0357	jhp1023	Cj0833c	
jHp	HP1125	jhp1054	Pal	
jHp	HP1183	jhp1109	Cj1684c	
jHp	HP1152	ffh	ffh	ffh (in databases)
jHp	HP1220	jhp1141	?	Undecidable "true" ortholog
jHp	HP1305	rpsH	rpsH	Rs8 (in databases)
jHp	HP1435	sppA	pspA	
jHp	HP1444	jhp1337	smpB	
jHp	HP1478	rep	Cj1101	
jHp	HP1560	rodA_2	Cj1038	
Cj	?	?	Cj0035c	Undecidable "true" ortholog
Cj	HP1160	jhp1087	Cj0121	
Cj	HP1360	ubiA	ubiA	UbiA (in databases)
Cj	?	?	Cj0173c	Undecidable "true" ortholog
Cj	HP0475	modC	modC	ModC (in databases)
Cj	HP0474	modB	modB	ModB (in databases)
Cj	HP0075	glmM	Cj0360	
Cj	HP1203	nusG	nusG	NusG (in databases)
Cj	?	?	Cj0481	Undecidable "true" ortholog
Cj	?	?	Cj0485	Undecidable "true" ortholog
Cj	HP1226	hemN_2	Cj0580c	
Cj	?	?	pstB	Undecidable "true" ortholog
Cj	HP1393	recN	recN	RecN (in databases)
Cj	HP1567	jhp1475	Cj0650	
Cj	HP0516	hslU	hslU	
Cj	HP1149	jhp1076	rimM	
Cj	HP0381	hemG	Cj0722c	
Cj	?	?	Cj0730	Undecidable "true" ortholog
Cj	HP0405	jhp0976	Cj0791c	
Cj	?	?	Cj0901	Undecidable "true" ortholog
Cj	?	?	glnQ	Undecidable "true" ortholog
Cj	HP1155	murG	murG	MurG (in databases)
Cj	HP0179	jhp0167	Cj1180c	
Cj	?	?	Cj1235	undecidable "true" ortholog
Cj	HP0763	ftsY	ftsY	
Cj	?	?	Cj1257c	Undecidable "true" ortholog
Cj	?	?	Cj1329	Undecidable "true" ortholog
Cj	?	?	selB	Undecidable "true" ortholog
Cj	HP0195	fabI	fabI	FabI (in databases)
Cj	HP1275	jhp1196	Cj1407c	
Cj	?	?	kpsT	Undecidable "true" ortholog
Cj	HP0734	jhp0671	Cj1454c	
Cj	HP0566	dapF	dapF	
Cj	?	?	Cj1538c	Undecidable "true" ortholog
Cj	?	?	Cj1580c	Undecidable "true" ortholog
Cj	?	?	Cj1581c	Undecidable "true" ortholog
Cj	HP0251	jhp0236	Cj1582c	
Cj	?	?	Cj1587c	Undecidable "true" ortholog
Cj	HP0936	proP	Cj1588c	

(*Continued*)

Table III. (*continued*)

Annotated as putative or hypothetical in	*H. pylori 26295* Hp	*H. pylori J99* jHp	*C. jejuni* Cj	Comments
Cj	?	?	chuC	Undecidable "true" ortholog
Cj	?	?	Cj1663	Undecidable "true" ortholog
Cj	HP1431	ksgA	ksgA	KsgA (in databases)
Hp, jHp	HP1335	trmU	trmU	trmU (in databases)
Hp, jHp	HP1343	jhp1262	Cj0186c	
jHp, Cj	HP0220	jhp0206	Cj0240c	
jHp, Cj	HP0303	jhp0288	Cj0096	
jHp, Cj	HP1024	dnaJ_1	cbpA	CbpA (in databases)
jHp, Cj	HP1014	jhp0409	Cj0807, ptmA	
jHp, Cj	HP0569	jhp0516	Cj0930	
jHp, Cj	HP0748	jhp0685	Cj1277c	
jHp, Cj	HP0834	jhp0773	Cj0386	
jHp, Cj	HP0911	jhp0847	Cj0777	
jHp, Cj	HP0939	jhp0874	glnP	
jHp, Cj	HP1181	jhp1107	Cj1375	
jHp, Cj	HP1452	thdF	thdF	TrmE (in databases)
jHp, Cj	HP1465	jhp1358	iamA	
jHp, Cj	HP1584	ydiE	Cj1344c	
Hy, jHp, Cj	HP0248	jhp0233	Cj0268c	
Hy, jHp, Cj	HP0552	jhp0499	Cj0154c	
Hy, jHp, Cj	HP0707	jhp0646	Cj0693c	
Hy, jHp, Cj	HP0787	jhp0724	Cj0941c	
Hy, jHp, Cj	HP0831	jhp0770	Cj1530	
Hy, jHp, Cj	HP1185	jhp1111	Cj1241	
Hy, jHp, Cj	HP1221	jhp1142	uppS	
Hy, jHp, Cj	HP1573	jhp1481	Cj0644	

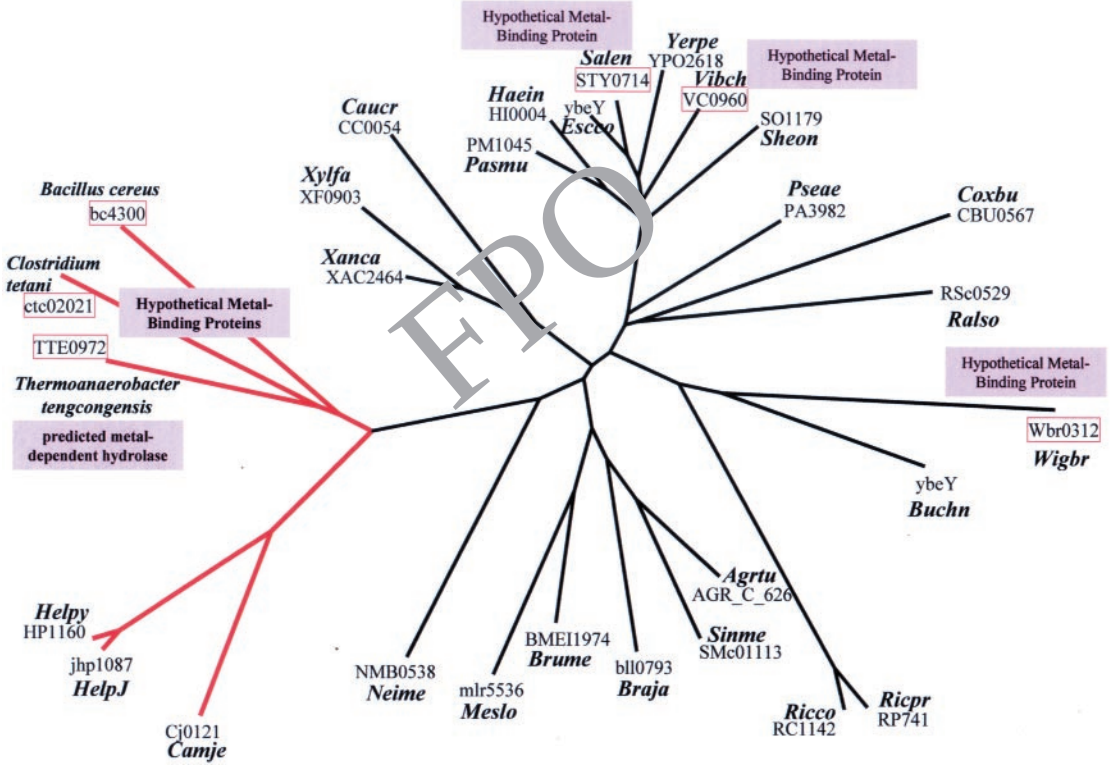

Figure 2. Annotating hypothetical universally conserved proteins: the case of the Cj0121-encoded protein. An evolutionary tree has been derived for Cj0121 and its orthologs, with the use of an adaptation of the program PhyloTree of the DARWIN package [7]. This program makes it possible to reconstruct a distance tree that is an approximation to a maximum likelihood tree [7]. Branch lengths (in PAM units) are drawn to scale. Species name abbreviations are as indicated in Table I. When known, the experimentally determined or putative function has been indicated.

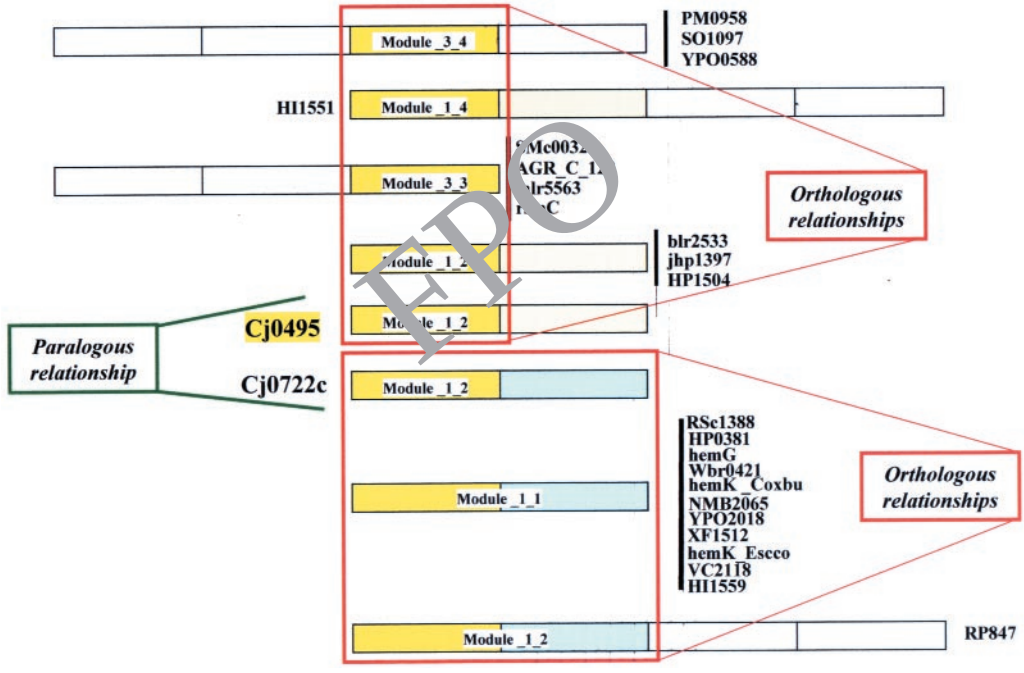

Figure 3. Modular structure of Cj0495-encoded protein and its homologs. The segments of homology found during the AllAll comparison are schematized by colored rectangles. Yellow modules are further used to make the multiple alignment and the tree shown in Figure 4.

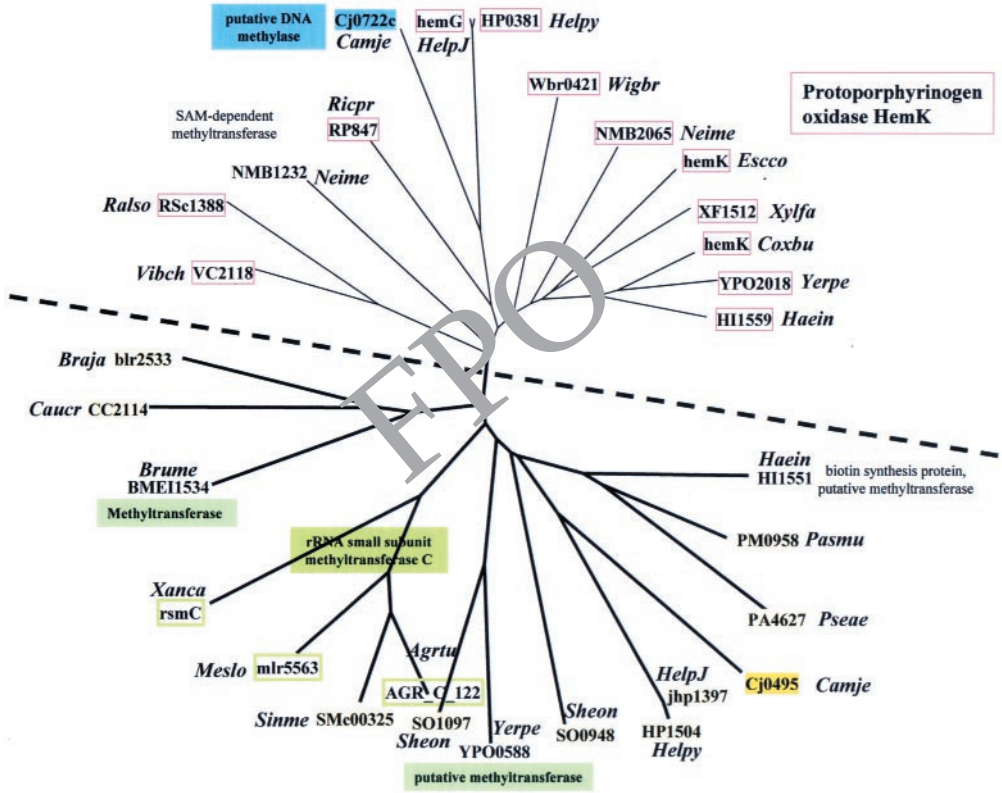

Figure 4. Annotating hypothetical universally conserved proteins: the case of the bimodular Cj0495-encoded protein and its paralog Cj0722c. An evolutionary tree has been derived for module 1 of Cj0495 and its homologous modules with the use of an adaptation of the program PhyloTree of the DARWIN package [7]. Branch lengths (in PAM units) are drawn to scale. Species name abbreviations are as indicated in Table I. A dashed line separates the two parts of this composite tree: in the upper part, a subtree made of the different protoporphyrinogen oxidases HemK and the related methylases Cj0722c and NMB1232; in the lower part, the subtree made of Cj0495 and related methylases. When known, the experimentally determined or putative function has been indicated. Hypothetical proteins are framed in light yellow.

References and Notes

1. Siew N, Fischer D. Twenty thousand ORFan microbial families for the biologist? *Structure* 11, 7 (**2003**).
2. Labedan B, Riley M. Genetic inventory: *Escherichia coli* as a window on ancestral proteins. In: *Organization of the Prokaryotic Genome*. Charlebois R, Editor. ASM Press, Washington, DC (**1999**), pp. 311–329.
3. Zouine M, Sculo Q, Labedan B. Correct assignment of homology is crucial when genomics meets molecular evolution. *Comp. Funct. Genomics* 3, 488 (**2003**).
4. Riley M, Labedan B. Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of structural segment of homology, the module. *J. Mol. Biol.* 268, 857 (**1997**).
5. Le Bouder-Langevin S, Capron-Montaland I, De Rosa R, Labedan B. A strategy to retrieve the whole set of protein modules in microbial proteomes. *Genome Res.* 12, 1961 (**2002**).
6. Schwartz RM, Dayhoff MO. Matrices for detecting distant relationships. In: *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3. Dayhoff MO, Editor. National Biomedical Research Foundation, Washington, DC (**1978**), pp. 353–358.
7. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. *Science* 256, 1443 (**1992**).
8. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *Mol. Biol.* 219, 555 (**1991**).
9. Altschul SF. A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* 36, 290 (**1993**).
10. Gonnet GH, Hallett M. *The DARWIN Manual* (**1997**).
11. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 44 (**1970**).
12. Smith TF, Waterman MS. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195 (**1981**).
13. De Rosa R, Labedan B. The evolutionary relationships between the two bacteria *Escherichia coli* and *Haemophilus influenzae* and their putative last common ancestor. *Mol. Biol. Evol.* 15, 17 (**1998**).
14. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Débarbouillé M, Dervyn E, Deuerling E, et al. Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* 100, 4678 (**2003**).
15. Nakahigashi K, Kubo N, Narita S, Shimaoka T, Goto S, Oshima T, Mori H, Maeda M, Wada C, Inokuchi H. HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and *hemK* knockout induces defects in translational termination. *Proc. Natl. Acad. Sci. USA* 99, 1473 (**2002**).