J Mol Evol (2005) 60:378–390
DOI: 10.1007/s00239-004-0198-1

JOURNAL OF MOLECULAR EVOLUTION

The Origin and Evolution of Operons: The Piecewise Building of the Proteobacterial Histidine Operon

Renato Fani,[1] Matteo Brilli,[1] Pietro Liò[2]

[1] Dipartimento di Biologia Animale e Genetica, Via Romana 17-19, I-50125, Firenze, Italy
[2] Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, CB3 0FD, Cambridge, UK

Received: 28 June 2004 / Accepted: 1 October 2004 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. The structure and organization of 470 histidine biosynthetic genes from 47 different proteobacteria were combined with phylogenetic inference to investigate the mechanisms responsible for assembly of the *his* pathway and the origin of *his* operons. Data obtained in this work showed that a wide variety of different organization strategies of *his* gene arrays exist and that some *his* genes or entire *his* operons are likely to have been horizontally transferred between bacteria of the same or different proteobacterial branches. We propose a "piecewise" model for the origin and evolution of proteobacterial *his* operons, according to which the initially scattered *his* genes of the ancestor of proteobacteria coded for monofunctional enzymes (except possibly for *hisD*) and underwent a stepwise compacting process that reached its culmination in some γ-proteobacteria. The initial step of operon buildup was the formation of the *his* "core," a cluster consisting of four genes (*hisBHAF*) whose products interconnect histidine biosynthesis to both *de novo* synthesis of purine metabolism and that occurred in the common ancestor of the α/β/γ branches, possibly after its separation from the ε one. The following step was the formation of three mini-operons (*hisGDC*, *hisBHAF*, *hisIE*) transcribed from independent promoters, that very likely occurred in the ancestor of the β/γ-branch, after its separation from the α one. Then the three mini-operons joined together to give a compact operon. In most γ-proteobacteria the two fusions involving the gene pairs *hisN–B* and *hisI–E* occurred. Finally the γ-proteobacterial *his* operon was horizontally transferred to other proteobacteria, such as *Campylobacter jejuni*. The biological significance of clustering of *his* genes is also discussed.

Key words: Operon origin — Operon evolution — Gene duplicaton — Gene fusion

Introduction

The term operon was first introduced in the early 1960s by Jacob et al. (1960) and Jacob and Monod (1961) to define a group of genes whose expression was coordinated by an operator. The same term is now used to describe any group of adjacent genes that are transcribed from a promoter into a polycistronic mRNA. The finding that genes belonging to the same metabolic pathway were organized in similar operons in microorganisms of different phylogenetic lineages, such as *Escherichia coli* and the Gram-positive *Bacillus subtilis*, led to the assumption that the clustering of genes encoding enzymes involved in the same metabolic route was a common rule in the prokaryotic world. These similarities are often considered as proof that the operon organization is an ancient character and that the assembly of gene clusters/operons might have predated the appearance of the last common ancestor (LCA). The operon organization of genes belonging to the same metabolic pathway might have been evolutionarily

Correspondence to: Renato Fani; *email:* r_fani@dbag.unifi.it

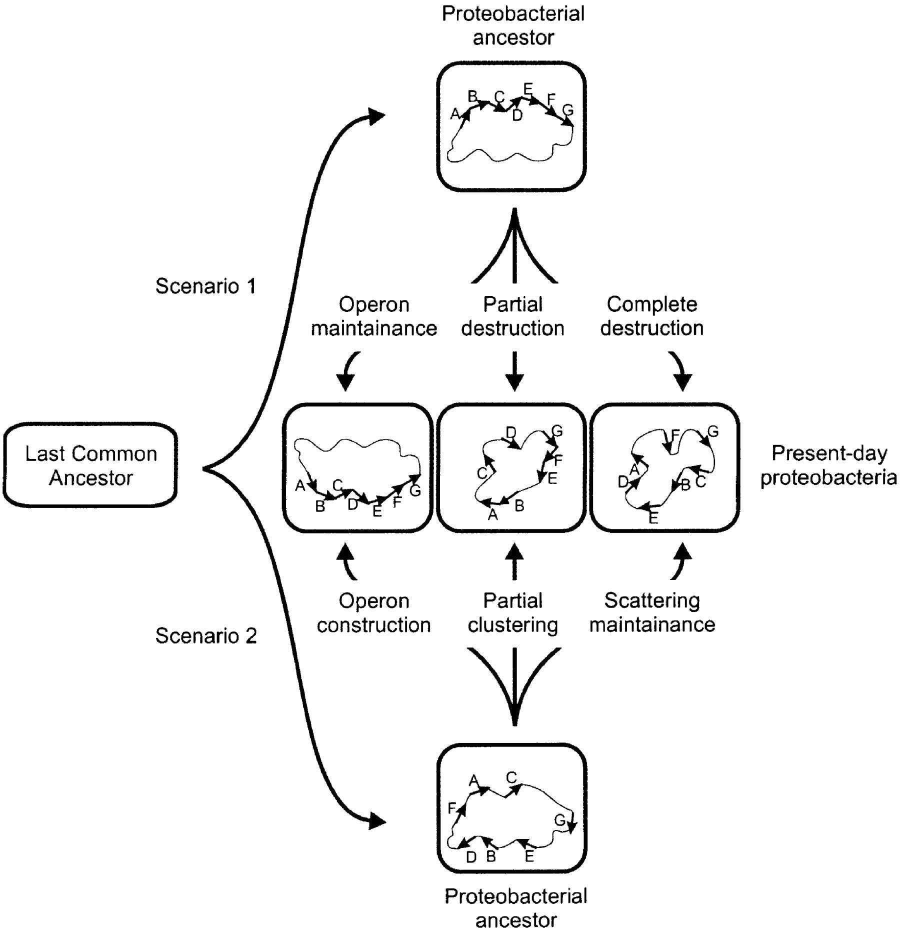

Fig. 1. Two possible alternative scenarios leading to the organization of genes involved in the same metabolic pathway in the extant proteobacteria. The stem and loop structure represents the attenuator.

advantageous in the early molecular and cellular evolution when, according to Woese (1998), the genetic temperature was high and the horizontal gene transfer should have been very frequent, allowing the exchange of entire metabolic routes.

If the idea of an ancient origin of operons is correct, this implies that whenever genes belonging to the same metabolic route are found scattered throughout the genome, the operon structure should have been somehow destroyed. The comparative analysis of several archaeal, bacterial, and eukaryal fully sequenced genomes revealed a high degree of genome instability, with drastic rearrangements of gene order occurring between both distant and close prokaryotic phylogenetic lineages (Mushegian and Koonin 1996; Watanabe et al. 1997; Kolsto 1997; Huynen and Bork 1998). In principle, the degree of gene conservation should be higher within operon structures than the outside regions, but sequence comparison of complete microbial genomes (Itoh et al. 1999) revealed that operons are unstable and that their conservation is generally low (Dandekar et al. 1998). Therefore, the conservation of operon structures is less important than expected previously, suggesting that their destruction is almost selectively neutral during long-term evolution. According to Itoh et al. (1999), functional constraints against co-expression of genes may be so weak that the organization of gene clusters in operon structures can be readily changed during evolution. A contrasting argument to the operon instability is that when an operon is destroyed and split into transcriptionally independent units, only the first one will retain the regulatory regions, and so it is quite possible that the transcription efficiency drastically decreases in the others (Itoh et al. 1999), affecting cell fitness, if the function provided by the operon is important.

However, the possibility that, at least in some cases, the operon structure is a recent invention of evolution cannot be *a priori* ruled out. As shown in Fig. 1, if a given phylogenetic lineage includes microorganisms showing a different organization of genes belonging to the same metabolic pathway, that is, complete scattering, compact operons, or partial scattering/partial clustering, at least two opposite but equally probable hypothetical scenarios can be depicted to explain such a picture.

1. The genome of the LCA contained genes organized in operons and this organization was completely or partially destroyed during evolution in some of the descendants' lineages.

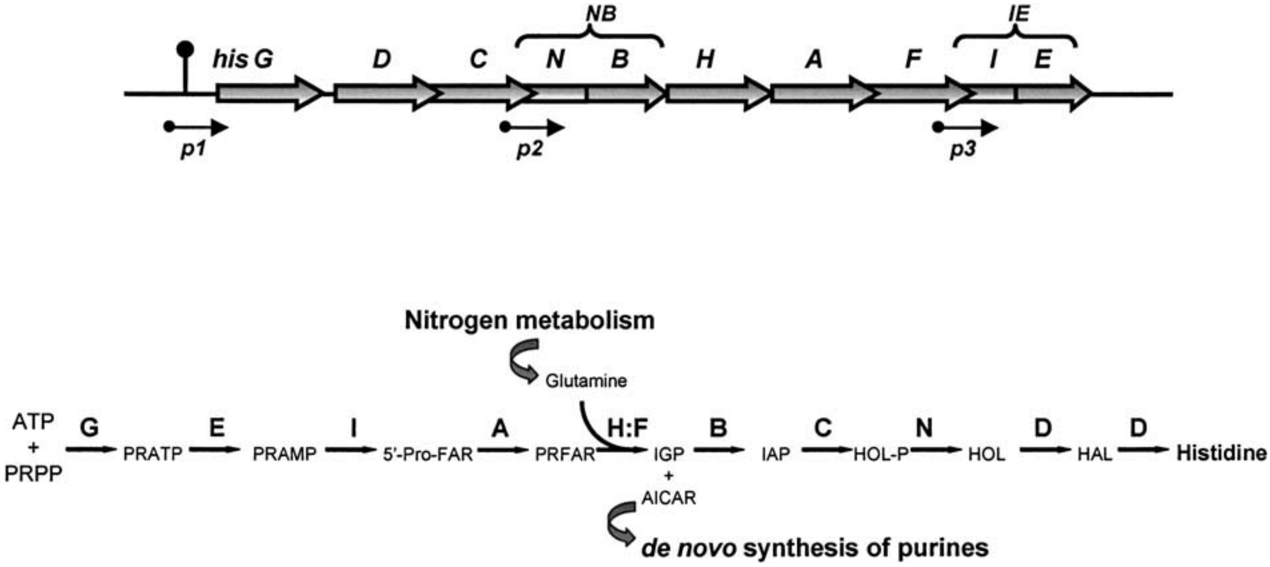

Fig. 2. Schematic representation of the *Escherichia coli* histidine biosynthetic operon (upper) and pathway (lower).

2. The LCA harbored genes (partially) scattered throughout the genome, so that in some of the descendants the construction of clusters and/or operons occurred.

Useful hints on this issue and on the molecular mechanisms/forces that might have driven the assembly/destruction of operons may be obtained by the comparative analysis of genes belonging to the same metabolic route if they are arranged in different ways in organisms belonging to the same or to different phylogenetic lineages. This comparison might permit recognition of a rule, if any, in gene organization. From this point of view the histidine biosynthetic pathway represents a very interesting case. This metabolic route has been studied for over 40 years in *E. coli* and its close relative *S. typhimurium*, leading to the accumulation of an exceptional body of biochemical, genetic, and physiological data (Alifano et al. 1996) that might be correlated with sequence data. Histidine biosynthesis is unbranched, includes a number of complex and unusual biochemical reactions, and consists of nine intermediates and of eight distinct proteins that are encoded by eight genes, arranged in *E. coli* in a very compact operon, in which the genes are located in the following order, *hisGDC(NB)HAF(IE)* (Alifano et al. 1996; Brilli and Fani 2004a and references therein). Three of the *his* genes, *hisD*, *hisNB* (formerly *hisB*), and *hisIE*, code for bifunctional enzymes (Fig. 2). As previously reported (Lazcano et al. 1992; Fani et al. 1995, 1998b) there are several independent indications for the antiquity of the histidine biosynthetic pathway, suggesting that the assembly of the entire route was completed long before the appearance of the LCA of the three extant cell domains. The available information also showed that after the building-up of the entire pathway, the *his* genes underwent major rearrangements in the three domains. In fact, a wide variety of different clustering strategies of *his* genes has been documented, suggesting that many possible histidine gene arrays exist and that there is no reason to assume the universality of the enterobacterial *his* operon (Fani et al. 1998b). Despite the large body of literature, a comparative analysis of the organization of histidine biosynthetic genes in different phylogenetic lineages has not been carried out. Therefore, the aim of this work was to use a data set of 470 histidine genes from 47 genomes and other information on His protein biochemistry to perform a detailed and comparative analysis of the structure and organization of *his* genes in proteobacteria where very different *his* gene arrays exist.

Materials and Methods

Sequence Retrieval

On October 21, 2003, a total of 54 genomes belonging to the proteobacterial lineage were completely sequenced and available in the GenBank database, with 47 of them harboring a complete set of histidine biosynthetic genes (Table 1). The 47 genomes were representatives of 40 species belonging to 27 different proteobacterial genera. Thirty-six of them were then considered for further analyses.

Sequence Alignment and Phylogenetic Tree Construction

The ClustalW program (Thompson et al. 1994) with standard parameters was used for multiple amino acid sequences alignment, followed by careful visual inspection.

Table 1. List of proteobacteria harboring a complete set of histidine biosynthetic genes and whose genome was completely sequenced on October, 21, 2003: Proteobacteria in boldface were then used for further analyses

Microorganism	Branch	Accesion No.	Length (bp)	Date of release
Agrobacterium tumefacians **str. C58 (Cereon)**	α	NC_003062	2,841,581	Oct 3 2001
		NC_003063	2,074,782	Oct 3 2001
Agrobacterium tumefaciens str. C58 (U. Washington)	α	NC_003304	2,841,490	Dec 14 2001
		NC_003305	2,075,560	Dec 14 2001
Bradyrhizobium japonicum **USDA 110**	α	NC_004463	9,105,828	Dec 27 2002
Brucella melitensis **16M**	α	NC_003317	2,117,144	Dec 27 2001
		NC_003318	1,177,787	Dec 27 2001
Brucella suis **1330**	α	NC_004310	2,107,792	Sep 30 2002
		NC_004311	1,207,381	Oct 2 2002
Caulobacter crecentus **CB15**	α	NC_002696	4,016,947	Mar 21 2001
Mesorhizobium loti	α	NC_002678	7,036,074	Sep 10 2001
Sinorhizobium meliloti	α	NC_003047	3,654,135	Oct 5 2001
Bordetella bronchiseptica	β	NC_002927	5,339,179	Aug 12 2003
Bordetella parapertussis	β	NC_002928	4,773,551	Aug 12 2003
Bordetella pertussis	β	NC_002929	4,086,189	Aug 12 2003
Chromobacterium violaceum **ATCC 12472**	β	NC_005085	4,751,080	Sep 8 2003
Neisseria meningitidis **MC58**	β	NC_003112	2,272,351	Sep 19 2001
Neisseria meningitidis Z2491	β	NC_003116	2,184,406	Sep 27 2001
Nitrosomonas europaea **ATCC 19718**	β	NC_004757	2,812,094	Apr 30 2003
Ralstonia solanacearum	β	NC_003295	3,716,413	Dec 7 2001
Buchnera aphidicola **str. APS (Acyrthosiphon pisum)**	γ	NC_002528	640,681	Sep 10 2001
Buchnera aphidicola str. Bp (Baizongia pistaciae)	γ	NC_004545	615,980	Jan 27 2003
Buchnera aphidicola str. Sg (Schizaphis graminum)	γ	NC 004061	641,454	Jun 28 2002
Candidatus ***Blochmannia floridanus***	γ	NC_005061	705,557	Aug 8 2003
Escherichia coli CFT073	γ	NC_004431	5,231,428	Jun 20 2002
Escherichia coli **K12**	γ	NC_000913	4,639,221	Oct 15 2001
Escherichia coli O157:H7	γ	NC_002695	5,498,450	Mar 7 2001
Escherichia coli O157:H7 EDL933	γ	NC_002655	5,528,445	Sep 27 2001
Haemophilua influenzae **Rd**	γ	NC_000907	1,830,138	Sep 30 1996
Pasteurella multocida	γ	NC_002663	2,257,487	Sep 10 2001
Paeudomonas aeruginosa **PA01**	γ	NC_002516	6,264,403	Sep 10 2001
Pseudomonaa putida **KT2440**	γ	NC_002947	6,181,863	Dec 16 2002
Pseudomonas syringae **pv. tomato str. DC3000**	γ	NC_004578	6,397,126	Mar 5 2003
Salmonella enterica **subsp. enterica serovar Typhi**	γ	NC_003198	4,809,037	Nov 7 2001
Salmonella enterica subsp. enterica serovar Typhi Ty2	γ	NC_004631	4,791,961	Mar 21 2003
Salmonella typhimurium **LT2**	γ	NC_003197	4,857,432	Oct 25 2001
Shewanella oneidensis **MR-1**	γ	NC_004347	4,969,803	Sep 12 2002
Shigella flexneri **2a str. 2457T**	γ	NC_004741	4,599,354	Apr 22 2003
Shigella flexneri 2a str. 301	γ	NC_004337	4,607,203	Oct 16 2002
Vibrio cholerae	γ	NC_002505	2,961,149	Sep 10 2001
		NC_002506	1,072,315	Sep 10 2001
Vibrio parahaemolyticus **RIMD 2210633**	γ	NC_004603	3,288,558	Mar 10 2003
		NC_004605	1,877,212	Mar 10 2003
Vibrio vulnificus **CMCP6**	γ	NC 004459	3,281,945	Dec 22 2002
		NC_004460	1,844,853	Dec 22 2002
Xanthomonas axonopodis **pv. citri str. 306**	γ	NC_003919	5,175,554	May 23 2002
Xanthomonas campestris **pv. campestris str. ATCC 33913**	γ	NC_003902	5,076,188	May 23 2002
Xylella fastidiosa **9a5c**	γ	NC_002488	2,679,306	Oct 2 2001
Xylella fastidiosa Temecula1	γ	NC_004556	2,519,802	Feb 3 2003
Yersinia pestis **CO92**	γ	NC_003143	4,653,728	Oct 15 2001
Yersinia pestis **KIM**	γ	NC_004088	4,600,755	Jul 26 2002
Campylobacter jejuni **subsp. jejuni NCTC 11168**	ε	NC_002163	1,641,481	Sep 27 2001
Helicobacter hepaticus **ATCC 51449**	ε	NC_004917	1,799,146	Jun 26 2003
Wolinella succinogenes	ε	NC_005090	2,110,355	Sep 4 2003

Phylogenetic analyses were performed using MEGA 2.1 (Kumar et al. 2001) for distance methods and PAML (Yang 1997) and Passml (Liò et al. 1998) for maximum likelihood (ML) methods. We use several models of evolution, implemented as substitution matrices; for amino acid sequences: Wag (Whelan and Goldman 2001), JTT, implemented using gamma "+F" parameters; for DNA sequences Jukes–Cantor and REV (see Whelan et al. [2001] for references to models of evolution).

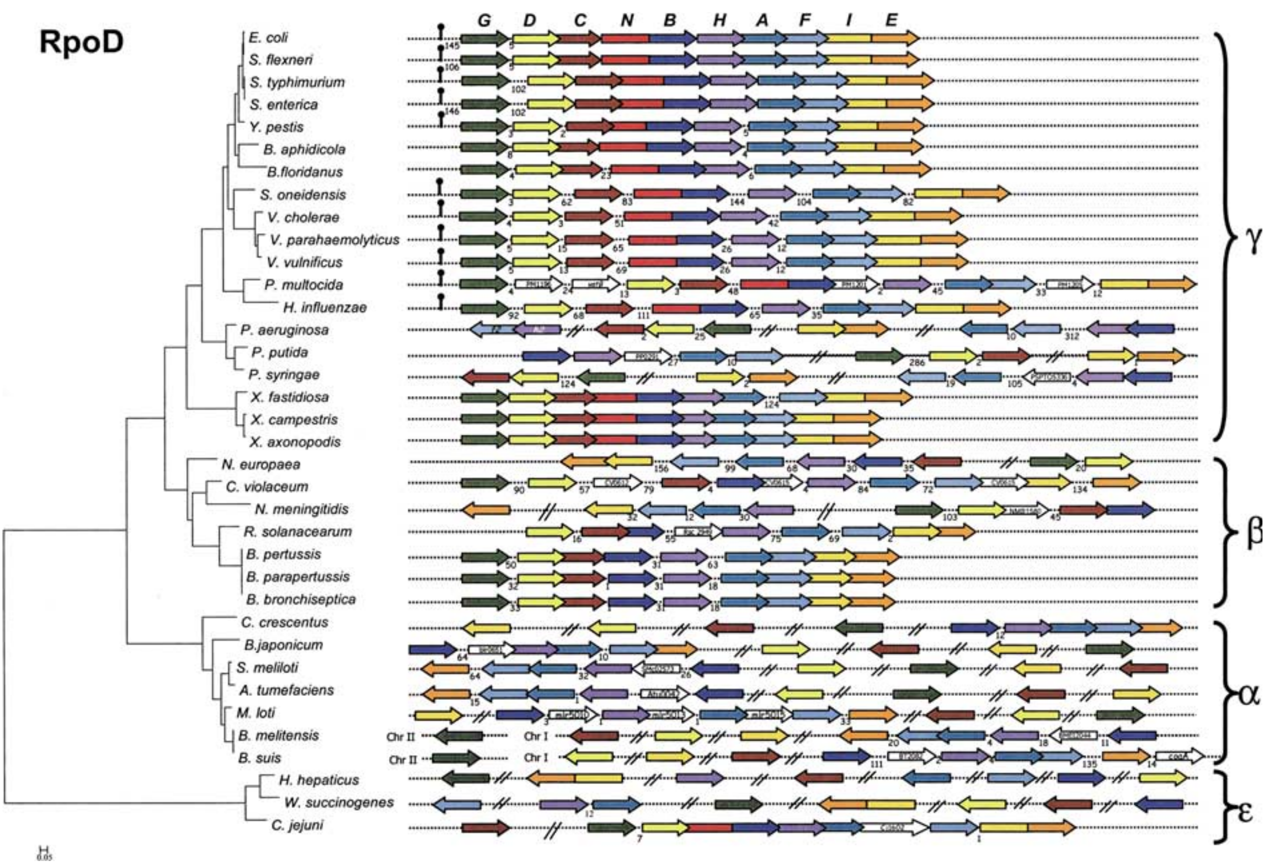

Fig. 3. Structure and organization of histidine biosynthetic genes in 36 different proteobacteria correlated with their phylogenetic position as established by RpoD analysis.

Results

The Structure of his *Biosynthetic Genes in Proteobacteria*

All of the 470 *his* gene sequences retrieved were analyzed for both their structure and their organization to investigate the mechanisms responsible for their assembly into cluster and/or operons and the extent of horizontal gene transfer (HGT) of *his* genes between organisms of the same or different phylogenetic lineages. In our opinion the organization of *his* genes in prokaryotes should not be considered disjointedly from the phylogenetic relations between the different species. For example, Fig. 3, which is also discussed later in this paper, shows the structure and organization of *his* biosynthetic genes correlated with the phylogenetic position of each bacterium as established by RpoD sequences, which protein relationships are commonly considered to give similar results to species relationships.

Our previous works have revealed that three sets of *his* genes, *hisN–hisB*, *hisA–hisF*, and *hisI–hisE*, are of particular importance from both an evolutionary and a genome organization point of view (Fani et al. 1994, 1995; Brilli and Fani 2004a).

The Structure of hisB, hisN, *and* hisNB

A detailed inspection of all the available *hisB*, *hisN*, and *hisNB* (formerly *hisB*$_{\mathrm{d}}$, *hisB*$_{\mathrm{px}}$, and *hisB*, respectively) gene products from microorganisms belonging to the three cell domains (Archaea, Bacteria, and Eukarya) was recently carried out (Brilli and Fani 2004). This analysis revealed that the bifunctional *hisNB* gene, which codes for a protein with two enzymatic abilities, an L-histidinol–phosphate phosphatase (HOL-Pase) and an imidazole–glycerol phosphate dehydratase (IGP-ase), catalyzing the sixth and the eighth steps of histidine biosynthesis (Fig. 2), is the outcome of a gene fusion event involving two cistrons (*hisN* and *hisB*) coding for HOL-Pase and IGP-ase activities, respectively. As shown in Fig. 3, a bifunctional *hisNB* gene has been detected only in some representatives of γ-proteobacteria and in the ε-proteobacterium *C. jejuni*. It has been suggested (Brilli and Fani 2004a) that *hisN* originated by duplication of a preexisting gene encoding a DDDD-type phosphatase with a broad range of specificity. The paralogous duplication gave rise to two copies: one became *hisN* and the other evolved toward *gmhB* (which is involved in the biosynthesis of a precursor of the inner core of the outer membrane lipopolysaccharides). According to the proposed model, *hisN* joined an already formed *his* operon, and its introgression was coincident with its fusion to *hisB* to give a bifunctional *hisNB*. The *his* gene organization reported in Fig. 3 shows that, whenever a bifunctional *hisNB* gene is present in the genome of a proteobacterium, it is embedded in compact operons, clearly supporting our hypothesis. The fusion event was traced in the ancestor of a γ-proteobacterial group including Enterobacteriaceae, Alteromonadaceae, Vibrionaceae, and Pasteurellaceae, after its separation from the *Pseudomonas* group. The presence of a *hisNB* gene in *C. jejuni* is very likely the result of a lateral gene transfer event; in fact the phylogenetic analysis revealed that the *C. jejuni* His proteins fell within the γ-proteobacterial ones (Brilli and Fani 2004a).

The Structure of hisA *and* hisF

We have previously shown that *hisA* and *hisF*, whose products catalyze sequential reactions (the fourth and the fifth ones) in histidine biosynthesis, are paralogs (Fani et al. 1994; see also Fani et al. 1995, 1997, 1998). Moreover, the two genes share a similar organization into two homologous modules half the size of the entire sequence (Fani et al. 1994; Fani 2004). Comparison of these two modules suggested that *hisA* and *hisF* are the results of two ancient successive in-tandem duplications. First, a *hisA1* module duplicated giving the entire *hisA*, which underwent another in-tandem duplication to give rise to the *hisF* gene. The finding that the structures of *hisA* and *hisF* are maintained through Archaea, Bacteria, and Eukarya (Fani et al. 1997, 1998) led to the assumption that the two gene duplication events occurred early in the evolution, long before the appearance of the LCA. We found that all the proteobacterial *hisA* and *hisF* genes in our dataset share the same two-module organization (not shown). Figure 3 shows that the two genes also share the same organization in almost all the proteobacteria, where they are arranged in an operon, with *hisA* located just upstream of *hisF*. Moreover, in most cases they are overlapped or very close with short intergenic space between them (Table 2).

The Structure of hisI *and* hisE

The *hisI–hisE* genes show a different structure and organization in the diverse proteobacterial lineages. In *E. coli* and its relatives, in the *Xylella*/*Xanthomonas* group, and in some ε-proteobacteria, the two genes are fused in a bifunctional one encoding a protein endowed with both phosphoribosyl-ATP-pyrophosphatase (HisE) and phosporibosyl-AMP-cyclohydrolase (Hisl) activities, which catalyze the second and third steps of histidine biosynthesis. In other proteobacteria they overlap, but are not fused, whereas in others they are separated (Fig. 3). The sequence analysis of all the available archaeal, bacterial, and eukaryal *hisI*, *hisE*, and *hisIE* genes (not shown) revealed that the two genes do not share a significant degree of sequence similarity, suggesting that they are the result of a domain-shuffling event rather than a paralogous duplication of an ancestral gene or of a gene-elongation event. It is noteworthy that when *hisI* and *hisE* are fused in a bifunctional gene, they are always arranged in the same relative order, with the *hisI* moiety located upstream of *hisE.* A gene with the two moieties arranged in the opposite order has not been found up to now, suggesting the existence of constraints in gene fusions. The existence of a *hisIE* bifunctional gene in both the ε-proteobacteria *H. hepaticus* and *W. succinogenes* is intriguing. Indeed, the other *his* genes are monofunctional and are scattered throughout the bacterial chromosome. The phylogenetic analysis revealed that these bifunctional genes very likely are native of the two bacteria (not shown). This finding suggests that the fusion of *hisI* and *hisE* may have occurred independently in different phylogenetic lineages, pointing toward a possible phenomenon of convergent evolution.

Organization of his *Genes in Proteobacteria*

The organization of histidine biosynthetic genes in the 36 proteobacteria analyzed reported in Fig. 3 revealed that consistently different *his* genes arrays exist among the proteobacterial branches and also within the same branch. In detail:

1. In the ε-proteobacteria *H. hepaticus* and *W. succinogenes* the *his* genes are scattered throughout the genome, whereas in *C. jejuni* the same genes (except for *hisC*) are organized in a compact operon harboring both *hisNB* and *hisIE* bifunctional genes. Moreover, the *C. jejuni his* gene order is identical to the enterobacterial one. This and the phylogenetic analyses based on His protein sequences suggest that the *C. jejuni his* operon originated by a lateral gene transfer (LGT) event (Brilli and Fani 2004a).
2. In bacteria belonging to the α-branch the *his* genes are partially scattered/clustered throughout the genome, differently localized on the genome and separated by several kilo base pairs. In *Brucella suis* and *B. melitensis* some genes are dislocated on different chromosomes. However, in all the representatives of this branch, five of the *his* genes (*hisBHAFE*) are clustered together in an operon-like structure that, in very few cases,

Table 2. Distance between *hisA* and *hisF* genes in 36 proteobacteria

Microorganism	Proteobacterial branch	Distance (bp)
Campylobacter jejuni subsp. jejuni NCTC 11168	ε	+941 (orf)[a]
Helicobacter hepaticus ATCC 51449	ε	+338,581
Wolinella succinogenes	ε	+538,022
Agrobacterium tumefaciens	α	−4
Bradyrhizobium japonicum USDA 110	α	+10
Brucella melitensis 16M	α	−4
Brucella suis 1330	α	−4
Caulobacter crescentus CB15	α	−1
Mesorhizobium loti	α	+832 (orf)
Sinorhizobium meliloti	α	−5
Bordetella bronchiseptica	β	−4
Bordetella parapertussis	β	−4
Bordetella pertussis	β	−4
Chromobacterium violaceum ATCC 12472	β	+72
Neisseria meningitidis MC58	β	+13
Nitrosomonas europaea ATCC 19718	β	−1
Ralstonia solanacearum	β	+69
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	γ	−19
Candidatus blochmannia floridanus	γ	−31
Escherichia coli K12	γ	−19
Haemophilus influenzae Rd	γ	−19
Pasteurella multocida	γ	−19
Pseudomonas aeruginosa PA01	γ	+10
Pseudomonas putida KT2440	γ	+11
Pseudomonas syringae pv. tomato str. DC3000	γ	+19
Salmonella enterica subsp. enterica serovar Typhi	γ	−19
Salmonella typhimurium LT2	γ	−19
Shewanella oneidensis MR-1	γ	−19
Shigella flexneri 2a str. 2457T	γ	−19
Vibrio cholarae	γ	−19
Vibrio parahaemolyticus RIMD 2210633	γ	−19
Vibrio vulnificus CMCP6	γ	−19
Xanthomonas axonopodis pv. citri str. 306	γ	−7
Xanthomonas campestris pv. campestris str. ATCC 33913	γ	−7
Xylella fastidiosa 9a5c	γ	+124
Yersinia pestis CO92	γ	−19

[a] orf, open reading frame.

includes an open reading frame(s) (ORF[s]) with unknown function.

3. A progressive clustering of *his* genes occurred in the ancestor of the β/γ proteobacteria. In fact, in all representatives of these two proteobacterial branches, the *his* genes are not scattered in the genome. In the β-branch there are two different organizations of *his* genes. *N. europaea* and *N. meningitidis* have two or three mini-operon-like structures, whereas others, with the same gene order, have increasingly compact *his* operons. In the *Bordetella* group, either the *his* genes often overlap or the intergenic regions are very short. Furthermore, no additional ORFs were found in the *Bordetella his* operon. Note that these bacteria, belonging to the α- and β-branches, lack bifunctional *hisNB* and *hisIE* genes.
4. Overall, bacteria belonging to the γ-branch showed a gene compactness higher than that found in the other branches. In most of these bacteria the *his* genes are arranged in clusters where they exhibit the same relative order. Genes belonging to these operons overlap or have null or very short intergenic space. These organisms also possess bifunctional *hisIE* and *hisNB* genes. In at least one case (*Pasteurella multocida*) the *his* operon also includes additional genes with unknown function. A different organization was found in other γ-proteobacteria, as in the different *Pseudomonas* species where the *his* genes are organized in three different clusters, *hisGDC, hisBHAF*, and *hisIE*, whose relative gene order resembles that of the enterobacterial operons, *hisGDC* (*NB*) *HAF* (*IE*).

Table 3. Comparison of maximum log-likelihood values for the three best topologies of the His, RpoD, and 16S rDNA data sets under different models of evolution

		Tree 1			Tree 2			Tree 3		
		logL	Length	α	logL	Length	α	logL	Length	α
His	Jones Standard	−80620.28	9.82		−80857.96	9.92		**-78988.58**	9.147	
	F	−80659.9	10,096		−80894.42	10.196		**−79042.38**	9.386	
	G	−76747.17	13.24	0.859	−76925.18	13.385	0.854	**−75077.08**	11.928	0.858
	Wag Standard	−79105.51	9.5297		−79326.42	9.623		**−77467.15**	8.924	
	F	−78973.16	9.657		−79187.09	9.75		**−773.49**	9.034	
	G	−75897.63	11.697	0.98698	−76070.2	11.806	0.9813	**74208.52**	10.6597	0.9843
RpoD	Jones standard	**−17737.25**	6.43319		−18594.01	7.68932		−17682.52	6.54802	
	F	**−17650.91**	6.80636		−18484.04	8.2411		−17771.32	6.92795	
	G	**−16435.49**	10.645	0.5281	−17097.86	16.959	0.46828	−16526.18	10.818	0.524
	Wag Standard	**−17600.78**	6.015		−18842.67	7.085		−17721.83	6.118	
	F	**−17570.24**	6.557		−18390.88	7.834		−17685.74	6.673	
	G	**−16433.16**	8.958	0.56978	−17102.73	−13.674	0.50815	−16522.09	9.1295	0.565
16SrDNA	JC69 Standard	**−13867.00**	1.80099		−14047.2	1,9245		−13693.7	1.80355	
	G	**−12281.80**	2.12575	0.26266	−12492.9	2,57891	0.24341	−12287.8	2.1305	0.263
	REV Standard	**−13380.50**	1.81554		−13746.6	1.94196		−13390.5	1.81841	
	G	**−11970.70**	2.29131	0.25576	–12188.2	2.84016	0.23603	−11977	2.30806	0.25542

Note. Entries are protein data sets, models of evolution, models assumptions— + F (i.e., using amino acid frequencies calculated from each data set) and Γ (i.e., considering the rate of evolution of the sites distributed as a gamma distribution); maximum log-likelihoods under different models for the three best topologies; tree lengths; and alpha values of the gamma distribution.

Phylogeny of Concatenated His Proteins

The gene organization with respect to the phylogenetic relationships in Fig. 3 suggests a progressive clustering of *his* genes, from ε ⇒ α ⇒ (β, γ) proteobacteria. Apparently, some phylogenetically distant bacteria seem to share a *his* gene organization more similar than closer ones. For instance, *C. jejuni* exhibited a *his* gene organization apparently more similar to that found in some γ-proteobacteria than in the other ε-representatives. This appeared to be true also for *Pseudomonas*. A comparative phylogenetic analysis of the His proteins cluster, 16S rDNA, and the RpoD amino acid sequences was carried out to test if these results were due to LGT events. The His phylogenetic analysis was carried out on a data set containing nine proteins (HisG, D, C, B, H, A, F, I, and E), which were concatenated into a large fusion containing 2020 positions. Phylogenetic analyses (maximum likelihood [ML] values and estimated parameters for the different topologies, models of evolution considered, and data sets) are shown in Table 3. The best topologies and branch lengths for the three data sets are shown in Figs. 3 and 4.

We analyzed topology and branch lengths of the three best ML trees. We also considered the model assumptiom + F (i.e., using amino acid frequencies calculated from each data set) and Γ (i.e., considering the rate of evolution of the sites distributed as a gamma distribution). ML branch lengths were estimated using PASSML and the following models of evolution: JTT and Wag for the amino acid sequences and JC69 and REV for the 16S rDNA sequences. We found that RpoD and 16S rDNA led to the same topology for all the models considered. The Wag matrix with the gamma distribution performed better than other models in the three data sets, the His and RpoD trees have similar lengths but different distributions of mutation rate.

The analysis of the phylogenetic tree revealed that bacteria belonging to the *Pseudomonas* group were placed within β-proteobacteria, suggesting the occurrence of a horizontal transfer of histidine biosynthetic genes. Similarly, *C. jejuni* should have acquired the *his* genes from a γ-proteobacterium; this result is in agreement with previous phylogenetic analysis carried out on single data sets of histidine biosynthetic enzymes (Brilli and Fani 2004a and references therein).

A Model for the Origin of the his *Operon in Proteobacteria*

Our findings suggest the following ‘‘piecewise’’ model for the origin of the proteobacterial *his* operon. The focal point of this model is the hypothesis that the *his*

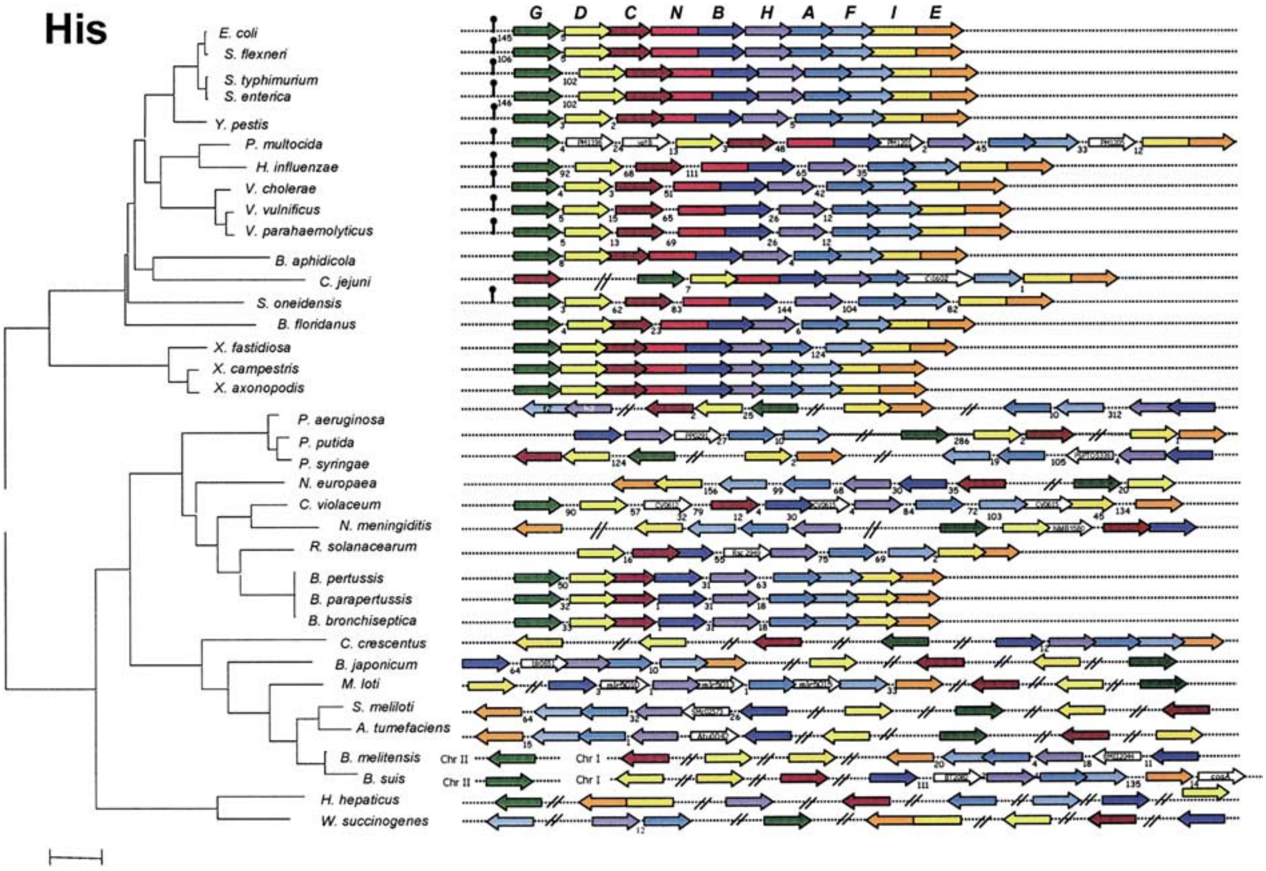

Fig. 4. Phylogenetic trees constructed using a concatenation of nine different histidine biosynthetic proteins (HisGDCBHAFIE) sequences from 36 proteobacteria.

genes were scattered on the chromosome of the ancestor of proteobacteria. All these genes (except for *hisD*) coded for monofunctional enzymes and were located in different chromosomal regions separated by DNA stretches of variable length (in some cases very long). Then these genes underwent a progressive clustering (parallel to the progressive shortening of the intergenic sequences) that culminated in some γ- and β-proteobacteria where the operons are very compact and include fused and/or overlapping genes. The model proposed, reported in Fig. 5, predicts (at least) the following possible steps:

1. The first step is the formation of the so-called "core" (Fani et al. 1995; Alifano et al. 1996; Brilli and Fani 2004a) of the *his* biosynthetic pathway, which is constituted by four genes clustered in the following order, *hisBHAF*. This cluster has been found also in other bacteria (Fani et al. 1995, 1998) and its genes encode the enzymes catalyzing the central reactions of the *his* pathway, interconnecting it to nitrogen metabolism and to the *de novo* synthesis of purines (Fig. 2) (Fani et al. 1995). This event very likely occurred in the ancestor of α/β/γ-proteobacteria. However, we cannot *a priori* rule completely out the possibility that the *his* "core" was already present in the proteobacterial ancestor and underwent destruction in *W. succinogenes* and *H. hepaticus*.
2. The following step is the clustering of *hisI* and *hisE*, whose distance on the bacterial chromosome progressively decreases until they become very close (or overlapping), giving rise to a bicistronic operon. Apparently, the clustering of *hisIE* was parallel to the formation of another gene cluster, probably *hisGDC*. In this way, an ancestral *his* regulon (Mass and Clark 1964) would have been assembled, becoming increasingly compact during the evolutionary history of the β/γ ancestor.
3. Then the histidine mini-operons joined together to form a single unit. This could have been partially destroyed in some of the modern representatives of the β-subdivision (see *N. meningitidis* and *N. europaea*). Alternatively, the compacting of the complete *his* operon happened independently in the β- and the γ-subdivisions, after their separation from their common ancestor. However, this scenario appeared less probable.
4. In some bacteria belonging to the γ-branch the recruiting of a gene coding for a HOL-P phos-

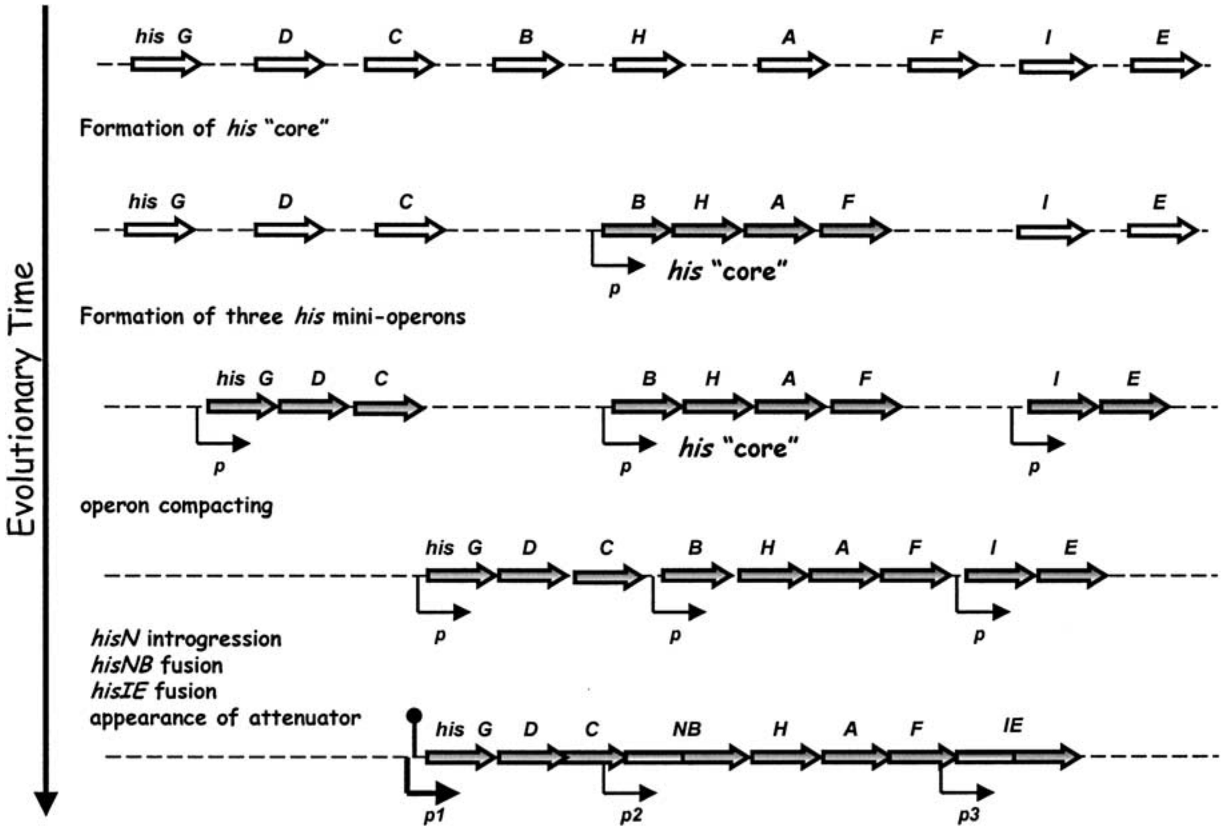

Fig. 5. An evolutionary model for the origin and evolution of histidine biosynthetic operons in the extant proteobacteria.

phatase (*hisN*) and the gene fusions involving the two pairs *hisN–hisB* and *hisl–hisE* then occurred.

5. The entire operon (or parts thereof) was then horizontally transferred to other microorganisms belonging to the same or to different branches (i.e., *Pseudomonas* and *C. jejuni*).

An alternative scenario would predict that the genome of the proteobacterial ancestor harbored a compact operon and this was somehow and differently destroyed in organisms belonging to different phylogenetic branches. If this scenario would be correct, this would mean that the destruction of operon organization would have given rise to scattered genes and/or mini-operons, with the creation of new promoters upstream of each of them (Itoh et al. 1999). In our opinion, the *E. coli*-like operon has been constructed in a relatively recent time, and its ancestrality cannot be considered plausible because it would imply that in many different phylogenetic lineages the *hisIE* gene underwent a gene fission event. It would also imply, mostly improbable, that the two moieties of the *hisNB* gene would have separated, with the concomitant loss of the *hisN* domain from the genome (note that a monofunctional *hisN* gene has not been characterized in any proteobacterium), and, therefore, the "return" of the *gmhB* gene to its ancestral state, that is, a gene encoding a DDDD phosphatase with a broad range of substrates. In our opinion, this scenario is much less probable than the first one, even though multiple gene fission events may occur during evolution as previously reported (Snel et al. 2000).

Discussion

In this paper we report the analysis of the structure and organization of histidine biosynthetic genes in proteobacteria with the aim of understanding the mechanism(s) that has (have) driven the clustering of genes and the operon construction. The analysis of the structure of the *his* genes gave strong support to the hypothesis that at least three different molecular mechanisms played an important role in shaping the pathway, that is, gene elongation, paralogous gene duplication(s), and gene fusion (Fani et al. 1995; Alifano et al. 1996; Fani 2004). The analysis of *his* gene organization in different proteobacteria revealed that several gene arrays exist within this phylogenetic lineage, with genes completely scattered throughout the bacterial chromosome, as in the case of some ε-proteobacteria, partially scattered/clustered (most representatives of the α-branch) or strictly compacted (enterobacteria, γ-branch; *Bordetella*, β-branch; and *C. jejuni*, ε-branch). Even though different scenarios can be depicted for this different organization, i.e., the presence of scattered or clustered genes in the ancestor of proteobacteria, data reported in this work support the first hypothesis. Our model suggests that the *his* gene organization in the extant proteobacteria arose from a gene scattering in the genome of the proteobacterial ancestor. The model also predicts that all or most of the *his* genes (except for *hisD*) were monofunctional and that they underwent a progressive clustering and compacting which occurred together with gene overlapping and fusion events.

This mechanism for the construction of a complete *his* operon is supported by the transcriptional analysis of the *E. coli* and *S. typhimurium* operons (Alifano et al. 1996 and references therein), which contain three different promoters (P1, P2, and P3), located upstream of *hisG, hisNB*, and *hisIE*. In this scenario, the three promoters would represent the vestiges of the ancestral promoters of the three *his* mini-operons.

It is important to notice that the order of the genes, *hisGDC*(*NB*) *HAF* (*IE*), in the enterobacterial *his* operon apparently does not match the sequence by which the enzymes they code for take part in the synthesis of histidine (HisG, E, I, A, H–F, B, C, N, D). However, if we look not at single but rather at blocks of genes/enzymatic steps, a "rule" emerges. Except for the first gene of the operon, *hisG* (which is regulated by the final product, histidine), the three gene blocks (*hisDCN, BHAF*, and *IE*) code for proteins operating in the reverse order in the metabolic pathway. In other words, the first block of genes (*hisDCN*) of the operon encodes proteins catalyzing the last block of enzymatic reactions, and the last gene block (*hisIE*) codes for proteins catalyzing the first steps (Fig. 2). It is possible that this particular gene order resulted from both regulatory and metabolic constraints (Alifano et al. 1996). Regarding the proximity between *hisG* and *hisD*, it might permit the spatial colocalization of their products and so a faster feedback inhibition of the first enzyme of the pathway, coded for by the former gene, by the end product of the pathway, histidine, released by the product of the latter.

If our model is correct, the building-up of a compact *his* operon represents a recent invention of evolution (dated in the βγ-proteobacterial ancestor) and raises the intriguing question of its biological significance. The origin and evolution of operons are still under debate, and at least six different classes of models have been proposed to explain the existence of operons: the natal model, the Fischer model, the molarity model, the coregulation model (these first four models have been reviewed extensively by Lawrence and Roth [1996], the selfish operon model (Lawrence and Roth 1996; Lawrence 1999), and the adaptation to thermophily model (Glansdorff 1999).

1. The natal model postulates that genes may be arranged in clusters if they originated by *in situ* duplication and divergence.
2. The Fischer model proposes that the physical proximity of coadapted alleles in the genome reduces the frequency of the formation of unfavorable combinations of genes by recombination events.
3. The molarity model predicts that gene clusters result in a beneficially high local concentration of proteins in the cytoplasm.
4. According to the coregulation model, genes are clustered because coregulation at a single promoter is beneficial.
5. The selfish operon hypothesis (Lawrence and Roth 1996) is based on the idea that horizontal transfer events are written in the history of all operons and that they had a key role in the origin and propagation of their organization. In other words, operons were produced in consequence of horizontal transfer events of heterogeneous sets of nonessential genes, which were later retained by the host and successively adapted to the new environment. This event would have been facilitated by the increase in fitness produced by the useful metabolic capacity guaranteed by the expression of the new set of genes.
6. Glansdorff (1999) suggested that a fundamental role in the emergence of operon structures was covered by the early adaptation to thermophily. This idea is supported by the transcription–translation coupling, which is seen as a mechanism capable of protecting the messenger RNA from the degradation caused by high temperatures.

Data reported in this work support some of the different theories described above. In detail:

1. The origin and the organization of the gene pair *hisA–hisF* support the natal model and reflect their evolutionary history (Fani et al. 1994, 1995).
2. The vicinity of HisH and HisF in the *his* operon/core is in agreement with the molarity model; indeed the two proteins must interact at a 1:1 ratio to obtain an active IGP synthase, the enzyme whose activity interconnects histidine biosynthesis to both the *de novo* synthesis of purines and nitrogen metabolism. This is also in agreement with the notion that in most cases genes coding for proteins that must interact in their functional state are very often found very close together on the genome (Dandekar et al. 1998; Tamames 2001; Itoh et al. 1999).
3. The finding that the entire histidine operon or parts thereof have been horizontally transferred between proteobacteria belonging to the same or

to different phylogenetic branch is in agreement with the proposal by Lawrence and Roth (1996), i.e., the selfish operon model. Even though no apparent attenuator sequence has been detected in *C. jejuni*, this is not in contrast with this model. In fact, it predicts that, once an entire operon has been transferred from a donor to a given host, the regulatory sequences (which, in principle, might not act in the new host) undergo mutational changes enabling the host transcriptional apparatus to recognize them and to permit expression of the introgressed operon (Lawrence and Roth 1996). This idea has recently gained experimental support by Dabizzi et al. (2001), who demonstrated that when *Azospirillum brasilense* histidine operon is transferred by plasmid-mediated conjugation to *E. coli his* mutants, the regulatory signals are not efficiently recognized by the host RNA polymerase. On the other hand, the *A. brasilense his* genes may be activated at a high frequency and over a short time scale by promoter-generating mutations occurring in *E. coli* His$^-$ populations grown in the absence of histidine (Fani et al. 1998; Dabizzi et al. 2001). Single base substitution resulted in the generation of a −10 region efficiently recognized by the *E. coli* RNA polymerase.

4. The existence of multiple and sophisticated mechanisms, such as transcription attenuation, controlling *his* expression in *E. coli*, supported the coregulation model.

Although different forces might have driven the assembly of compact *his* operons, in our opinion the major ones were those enabling the *his* genes to be coregulated finely and the protein coded for synthesized in the correct stoichometric ratio. The translational coupling due to the extensive overlapping existing between *his* genes and the presence of three genes coding for bifunctional enzymes (*hisD, hisNB*, and *hisIE*) support this idea. As discussed elsewhere (Jensen 1996; Xie et al. 2003 and references therein) gene fusions provide a mechanism for the physical association of different catalytic domains, whose fusion presumably promotes the channeling of intermediates that may be unstable and/or at low concentrations. However, gene fusion can be also viewed as a mechanism that permits obtaining a 1:1 ratio between counterparts. Another clue strongly favoring the coregulation and molarity models is the origin and evolution of bifunctional *hisNB* genes. As reported elsewhere by Brilli and Fani (2004a), the bifunctional *hisNB* gene has a narrow phylogenetic distribution. A monofunctional *hisN* gene has not been found up to now and its presence is always correlated with the presence of a bifunctional *hisNB* gene and, most important, with *his* compact operons sharing the same gene structure and relative order (Fig. 3). It has been recently proposed (Brilli and Fani 2004a) that *hisN* originated by duplication of a preexisting gene encoding a DDDD phosphatase able to catalyze the same reaction on different substrates. The paralogous duplication gave rise to two copies: one became *gmhB* (which is involved in the biosynthesis of a precursor of the inner core of the outer membrane lipopolysaccharides) and the other one evolved toward *hisN*, a process apparently coincident with its introgression into an already formed *his* operon and its fusion to *hisB*. This is also supported by the *Bordetella his* genes that are arranged in a compact operon where a *hisN* gene is not present. A possible explanation for the absence of a monofunctional *hisN* gene and of a *hisNB* gene located outside compact operons is that during the final steps of *his* operon assembly there was a need to coregulate finely all the genes involved in *his* biosynthesis. Proteobacteria lacking a bifunctional *hisNB* gene do not possess a phosphatase endowed with a narrow substrate specificity (HOL-P) and the HOL-P dephosphorylation is probably carried out by an aspecific phosphatase encoded by a *gmhB*-like gene. Therefore, this gene is not under the control of a single mechanism (i.e., histidine requirement) but is controlled by multiple mechanisms. The arrangement of *his* genes into a compact operon controlled in a similar fashion probably required that all the genes involved in histidine biosynthesis be under the same regulatory mechanism. This, in turn, might have required the presence of a HOL-Pase dedicated only to the catalysis of this sole reaction. According to this model, all those proteobacteria showing a *hisN* gene also harbor a *gmhB* gene that differs from the *gmhB* genes harbored by microorganisms laking a *hisN* gene (Brilli and Fani 2004a). Therefore, the origin and evolution of *hisN* suggest the need for both molarity and coregulation. Maybe the two forces acted simultaneously in the building-up of the *E. coli*-like *his* operon; indeed, the compactness of *his* operons is parallel to the gene fusion increase. This is in agreement with the notion that a physical interaction between different enzymes, and their stabilization by homo- or hetero-associations, may be facilitated by organizing genes into operons transcribed into a polycistronic rnRNA which is immediately translated into proteins whose closeness in the "translational environment" favors their noncovalent association or their spatial segregation in a limited volume of cell cytoplasm (Brilli and Fani 2004b). This would limit the disturbing effect of molecular crowding.

In conclusion, this work supports the idea that the β/γ-proteobacterial *his* operon is a recent invention of evolution and has been built up by piecewise gene clustering/overlapping/fusion mechanisms starting from *his* genes originally scattered in the genome of the proteobacterial ancestor. The selection pressures that have driven the operon evolution were very likely the transcript and protein molarity and coregulation. The existence of similar operons in other bacterial phylogenetic lineages, such as low-GC Gram-positive bacteria, raises the intriguing question of the organization of histidine biosynthetic genes in the LCA of three cell domains.

In this work we have discussed why we cannot *a priori* exclude that the ancestral state of *his* genes was an operon. If this is so, in some phylogenetic lineages this organization would have been destroyed, leading to a different organization of histidine biosynthetic genes, i.e., a complete or partial scattering. However, independent of the ancestral organization of the *his* genes, i.e., prior to the appearance of a proteobacterial common ancestor, the available data reported in this paper suggest the construction (or a reconstruction) of the *his* operon starting from a *his* gene scattered genome-wide scenario. However, this issue is beyond the scope of the present work, which focuses mainly on the forces and the molecular processes responsible for the assembly of genes into operons. We are fully aware that further progress on the origin and evolution of the histidine biosynthetic genes will require not only a larger number of bacterial and archaeal complete genome sequences but also information on *his* gene transcriptomics and proteomics in different microorganisms.

References

Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB (1996) Histidine biosynthetic pathway and genes: structure, regulation and evolution. Microbiol Rev 60:44–69

Brill M, Fani R (2004a) Molecular evolution of *hisB* genes. J Mol Evol 58:225–237

Brilli M, Fani R (2004b) Origin and evolution of eucaryal *HIS7* genes: from metabolons to bifunctional proteins? Gene 339:149–160

Dabizzi S, Ammannato S, Fani R (2001) Expression of horizontally transferred gene clusters: activation by promoter-generating mutations. Res Microbiol 152(6):539–549

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23:324–328

Fani R (2004) Gene duplication and gene loading. In: Miller RV, Day MJ (eds) Microbial evolution: gene establishment, survival and exchange. ASM Press, Washington, DC, pp 67–81

Fani R, Liò P, Chiarelli I, Bazzicalupo M (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the *hisA* and *hisF* genes. J Mol Evol 38:469–495

Fani R, Liò P, Lazcano A (1995) Molecular evolution of the histidine biosynthetic pathway. J Mol Evol 41:760–774

Fani R, Tamburini E, Mori E, Lazcano A, Liò P, Barberio C, Casalone E, Cavalieri D, Perito B, Polsinelli M (1997) Paralogous histidine biosynthetic genes: evolutionary analysis of the *Saccharomyces cerevisiae HIS6* and *HIS7* genes. Gene 197:9–17

Fani R, Gallo R, Fancelli S, Mori E, Tamburini E, Lazcano A (1998a) Heterologous gene expression in an *Escherichia coli* population under starvation stress conditions. J Mol Evol 47:363–368

Fani R, Mori E, Tamburini E, Lazcano A (1998b) Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. Orig Life Evol Biosph 28(4–6):555–570

Glansdorff N (1999) On the origin of operons and their possible role in evolution toward thermophily. J Mol Evol 49:432–438

Huynen MA, Bork P (1998) Measuring genome evolution. Proc Natl Acad Sci USA 95:5849–5865

Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16(3):332–346

Kolsto AB (1997) Dynamic bacterial genome organization. Mol Microbiol 24(2):241–248

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17(12):1244–1245

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356

Jacob F, Perrin D, Sanchez C, Monod J (1960) L'opèron: groupe de gènes à expression cordonnée par un opératuer. CR Acad Sci Paris 250:1727–1730

Jensen R (1996) Evolution of metabolic pathways in enteric bacteria. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella typhimurium*. ASM Press, Washington, DC, pp 2649–2662

Lazcano A, Fox GE, Oro' J (1992) Life before DNA: the origin and evolution of early Archean cells. In: Mortlock RP (ed) The evolution of metabolic function. CRC Press, Boca Raton, pp 237–339

Lawrence JG (1999) Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol 2:519–523

Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860

Liò P, Goldman N, Thorne J, Jones DT (1998) Combining protein secondary structure prediction and evolutionary inference. Bioinformatics 14:726–733

Mass WK, Clark AJ (1964) Studies on the mechanism of repression of arginine biosynthesis in *E.coli* II. Dominance of repressibility in hybrids. J Mol Biol 8:365–370

Mushegian AR, Koonin EV (1996) Gene order is not conserved in bacterial evolution. Trends Genet 12:289–290

Snel B, Bork P, Huynen M (2000) Genome evolution: Gene fusion versus gene fission. Trends Genet 16:9–11

Tamames J (2001) Evolution of gene order conservation in prokaryotes. Genome Biol 2(6):RESEARCH0020

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13(5):555–556

Xie G, Keyhani NO, Bonner CA, Jensen RA (2003) Ancient origin of the tryptophan operon and the dynamics of evolutionary change. Microbiol Mol Biol Rev 67:303–342

Watanabe H, Mori H, Itoh T, Gojobori T (1997) Genome plasticity as a paradigm for eubacteria evolution. J Mol Evol 44:557–564

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Whelan S, Liò P, Goldman N (2001) Molecular phylogenetics: state-of-art methods for looking into the past. Trends Genet 17:262–272

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