Escherichia coli Protein Sequences: Functional and Evolutionary Implications

EUGENE V. KOONIN, ROMAN L. TATUSOV, AND KENNETH E. RUDD

117

INTRODUCTION

As of August 1994, the sequence of approximately 60% of the *Escherichia coli* chromosome, which consists of about 4,720 kbp, was available; this information is being accumulated in at least three independent databases (28, 46, 56). Translation of the genes contained in the EcoSeq7 database (46) results in 2,328 protein sequences that in turn make up about 60% of all proteins encoded in the *E. coli* genome. Obviously, an enormous amount of valuable information on the physiology of the bacterial cell and its evolution is encrypted in these sequences. Extracting this information to the fullest and using it as a platform for future, rationally designed experiments is one of the principal components of the *E. coli* genome project (13). In fact, to a considerable extent, this is what makes the determination of the complete genome sequence such a worthy task; of course, this is true of any genome project, not only *E. coli* (8). The availability of the complete set of protein sequences from one organism, and especially the possibility to compare such sequence sets from different, evolutionarily distant organisms, e.g., *E. coli* and *Saccharomyces cerevisiae* or *E. coli* and *Bacillus subtilis*, will allow one to address fundamental questions that even very recently seemed to be completely out of reach. Here are a few of these basic questions. What is the extent of functional prediction based on protein sequence, and how precise can this prediction be? What is the minimal set of conserved, housekeeping genes that supports the functioning of a bacterial cell? What specific genes define the physiological differences between bacteria? How conserved or how divergent are different functional groups of proteins between bacteria and eukaryotes and between different bacteria? What is the extent of gene duplication in the bacterial genome evolution, and what is typically duplicated—a single gene, an operon, or an even larger group of genes? Which functions benefit from diversification and are performed by multiple related proteins, and which are secured by unique proteins? To what extent can we realistically hope to reconstruct the genome organization of the common ancestor of all bacteria and that of the hypothetical progenote (58), the ultimate common ancestor of bacteria, eukaryotes, and members of the *Archaea*?

Our ability to adequately address these and other important questions depends both on the completeness of the available sequence data and on the power of computer methods for sequence analysis. Obviously, the final answers still belong to the future, if only because not a single complete bacterial genome sequence is currently available, but it is not a remote future any more. The collection of 60% of the *E. coli* protein sequences seems to be an appropriate starting point for a pilot project on comprehensive analysis of the set of proteins encoded in a bacterial genome (E. V. Koonin, R. L. Tatusov, and K. E. Rudd, *Proc. Natl. Acad. Sci. USA*, in press). In this chapter, we summarize the preliminary results of such a project.

STRATEGY AND METHODOLOGY OF COMPUTER-AIDED ANALYSIS OF THE *E. COLI* PROTEIN SEQUENCE SET

The existing approaches to molecular sequence analysis can be classified into two groups—intrinsic and extrinsic methods (11). Intrinsic analysis explores statistical properties of a sequence without explicitly comparing with other sequences. With regard to protein sequences, these are amino acid composition

and charge; clustering of charged, hydrophobic, or other residues; compositional complexity, which is indicative of the globular or nonglobular structure of a protein; and others. The extrinsic analysis deals with various aspects of sequence comparison and sequence conservation between different proteins.

In this chapter, we briefly consider some intrinsic properties of *E. coli* proteins, but, by and large, we concentrate on sequence comparison. We pursue two complementary views of the bacterial genome—"from the inside" and "from the outside."

The "outside view" pertains to sequence conservation between *E. coli* proteins and proteins from other organisms. The availability of about 60% of *E. coli* protein sequences gives one the opportunity to derive reliable estimates of the number of highly conserved proteins that contain ancient regions dating back to the divergence of eubacteria, eukaryotes, and the *Archaea;* those that are conserved between *E. coli* and distantly related bacteria; and variable proteins that are conserved only in closely related bacteria or are found in *E. coli* alone. It is of obvious importance to establish correlations between the level of sequence conservation in evolution and protein functions.

The core of our analysis includes the programs of the BLAST family, which perform database searches for sequence similarity by using individual sequences as queries and producing ungapped pairwise alignments (3, 4), and the MoST program, which screens databases for conserved motifs by using ungapped multiple-alignment blocks as the input (54).

The BLAST algorithm calculates the probability (*P* value) of high-scoring sequence segment pairs to be observed by chance on the basis of the Karlin-Altschul statistics for sequence comparison (see reference 3 and references therein). The lengths of the alignments are determined automatically to obtain the highest statistical significance and therefore may vary from as few as 15 to 20 amino acid residues to as many as several hundred, depending on the level of similarity and distribution of conserved regions in related proteins. The use of the *P* value as the criterion of significance in sequence comparisons may produce false positives when a combination of several segments with low scores results in an artifactually low *P* value. Therefore, in our analysis, we defined all the cutoffs in terms of the similarity score as such rather than the *P* value. The BLASTP program is used to screen amino acid sequence databases for similarity to a protein sequence query, and the TBLASTN program is used to screen nucleotide databases translated in all six reading frames (3).

The MoST program converts an ungapped multiple-alignment block, which may be derived directly from a BLAST output by using the CAP program (54) or generated by another multiple-alignment construction method into a position-dependent weight matrix. The matrix is used for database screening, and the statistical significance of the similarity score with the matrix is determined for each segment in the database by using the recently developed theory (54). Typically, the length of alignment blocks used to generate position-dependent weight matrices varies between 12 and 40 amino acid residues.

Thus, both principal methods of sequence similarity search used in our study are oriented at detecting at least one contiguous conserved region per protein sequence. This approach may result in some marginally significant similarities being missed, but it greatly reduces the likelihood of false positives because of the existence of rigorous statistical theory.

In different types of sequence database searches, spurious "hits" with low-complexity (compositionally biased) regions present in many proteins are frequently observed (59, 60). Therefore we routinely used the SEG program (59), in conjunction with both BLAST and MoST, to filter out such regions. However, since functionally important conserved protein segments may in some cases have a compositional bias, for those proteins that failed to show significant similarity to other sequences in the initial BLAST searches, the analysis was repeated without filtering.

Recent analyses of large sets of yeast and bacterial proteins have shown that careful examination of relatively weak similarities detected in database searches performed with different, complementary computer methods is critical for increasing the rate of functional prediction and revealing evolutionary relationships (7–11, 25, 45). Accordingly, after comparing all the available *E. coli* protein sequences with the nonredundant amino acid sequences database (National Center for Biotechnology Information) by using BLASTP and listing all the alignments with similarity scores above 90 (corresponding to $P \approx 10^{-3}$) as authentic relationships, we assessed the relevance of all the alignments in the "twilight zone" (scores between 60 and 90). To this end, we used analysis of conserved motifs with the CAP and MoST

programs and multiple-alignment analysis with the MACAW program (51), as well as considerations of functional relevance. In addition, all the *E. coli* protein sequences were searched for conserved motifs typical of *E. coli* protein clusters by using CAP and MoST (see below).

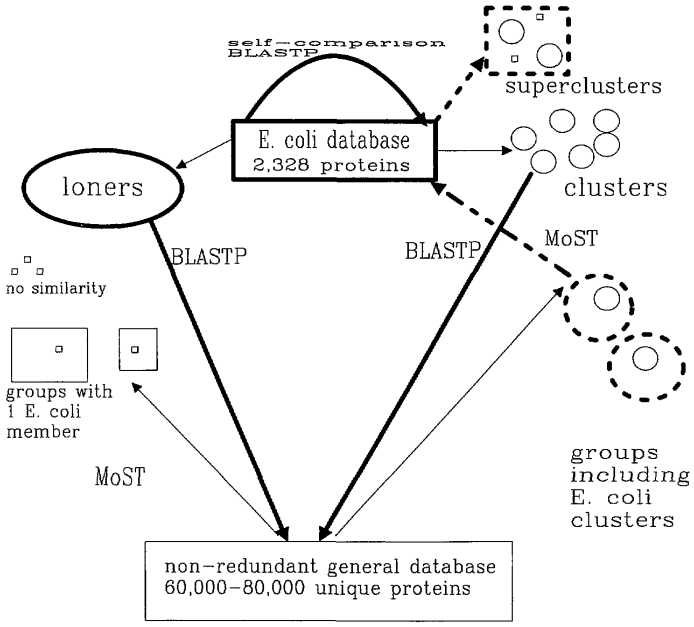


FIGURE 1 Scheme of the computer analysis of the *E. coli* protein sequence set.

Figure 1 Scheme of the computer analysis of the *E.coli* protein sequence set.

The “inside” view includes delineation of clusters of paralogous proteins within the *E. coli* protein sequence set. (Paralogs are proteins that share significant sequence similarity and, by inference, common ancestry but have distinct functions, as opposed to orthologs—related proteins that share both common ancestry and common function [14]. Accordingly, all related proteins encoded by different genes in a same organism are paralogs, whereas “the same” proteins in different organisms are orthologs. Paralogs should have evolved by intragenomic gene duplication and orthologs evolve by direct, vertical descent. The distinction between orthologs and paralogs is not universally appreciated, and they are often both simply called homologs; however, we believe that this distinction is very important for all attempts to understand the genome evolution, and we will use it throughout this text.) The existence of related genes that presumably have evolved by duplication in *E. coli* has been recognized for a long time (43, 44), but it is only now, with the sequence of a major part of the genome available, that it is becoming feasible to produce a nearly complete catalog of the clusters of paralogs. The fraction of such clusters that is represented in the current database of *E. coli* proteins probably is considerably higher than 60%, since many proteins among the 40% encoded in the not yet sequenced portions of the genome will join already known clusters. For clustering *E. coli* proteins, we used a single-linkage, “greedy” clustering algorithm. A cluster was defined as a group of protein sequences connected by similarity scores above a chosen cutoff but not requiring that each pair of sequences within a cluster had such a score. We found that a BLAST score of 70 or higher, even though not highly statistically significant, almost always corresponded to relationships between *E. coli* proteins that could be confirmed by subsequent multiple alignment analysis, search for conserved motifs, and functional assessment. Accordingly, this score was used as the cutoff for clustering. The “greedy” algorithm runs into a problem when proteins containing two or more distinct conserved domains bring together otherwise unrelated clusters. This was accounted for by deriving conserved motifs typical of each cluster and including the distinct domains of multidomain proteins in different clusters.

Motif conservation was also used as the criterion for delineating higher-rank groups of paralogous *E. coli* proteins which may include more distantly related sequences; we called such groups superclusters. A supercluster was defined as a group of proteins containing at least one unique, conserved motif. A supercluster may include several clusters of paralogs, as well as "loners," i.e., sequences not belonging to any cluster. Specifically, motifs (in the form of multiple alignment blocks) characteristic of each cluster were derived and used to search the *E. coli* protein sequence database by using MoST, in order to identify additional proteins (belonging or not belonging to other clusters) that may contain segments related to the motif.

The outside and inside views of the genome are not independent, and it is clearly of interest to explore the sequence conservation in different types of protein clusters encoded in the same genome, in our case, that of *E. coli.*

Figure 1 schematically shows the principal components of our analysis of the *E. coli* protein sequence set; the methods used for this analysis are discussed in more detail elsewhere (E. V. Koonin, R. L. Tatusov, and K. E. Rudd, *Methods Enzymol.*, in press). Another study of paralogous groups among *E. coli* proteins, based on different methods of sequence comparison, is included in this volume (see chapter 116).

SOME INTRINSIC PROPERTIES OF *E. COLI* PROTEINS

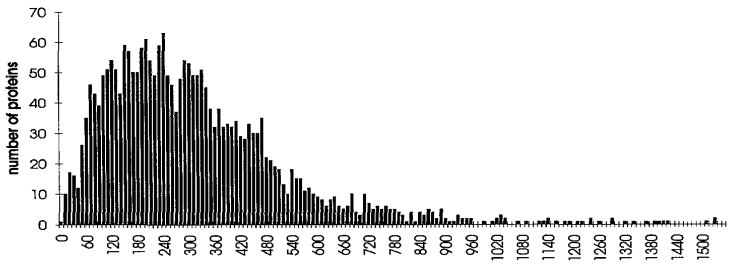


Figure 2 Length distribution of *E. coli* proteins.

Figure 2 shows the length distribution of *E. coli* proteins. The median of this distribution is 280 amino acid residues, and the mean is 321 residues, which is a typical size of an enzyme or a regulatory protein. There are examples of very large and very small proteins in *E. coli.* The three largest proteins are Lhr (large helicase-related [41a]), GltB (34), and MukB (33); in addition, a putative protein, YdbA, encoded by an open reading frame located near the replication terminus and interrupted by an insertion element, contains at least 1,540 residues (29). With the exception of GltB, these very large (putative) proteins contain extended predicted nonglobular domains (see below); Lhr and MukB also contain ATPase domains which may be involved in coupling ATP hydrolysis to mechanochemical processes. The smallest known functional gene product in *E. coli,* 29 amino acid residues in length, is the inner membrane protein KdpF (2). The smallest known enzyme is the recently identified peptidyl-prolyl isomerase, PpiC (parvulin), which contains 93 residues (40, 47).

TABLE 1 Net amino acid composition of the predicted *E. coli* proteins

Amino acid	% in *E. coli*	% in NR[a]
A	9.6	7.5
C	1.1	1.8
D	5.2	5.2
E	6.0	6.2
F	3.9	3.9
G	7.4	7.1
H	2.3	2.2
I	5.9	5.5
K	4.5	5.8
L	10.5	9.1
M	2.9	2.3
N	3.9	4.5
P	4.4	5.1
R	5.7	5.2
S	5.6	7.3
T	5.3	6.0
V	7.1	6.5
W	1.5	1.3
Y	2.8	3.2

[a] NR, nonredundant amino acid sequence database.

The net amino acid composition of *E. coli* proteins is shown in Table 1. It shows no dramatic deviations from that in the complete protein sequence database, with leucine and alanine being the most abundant and cysteine and tryptophan being the least abundant.

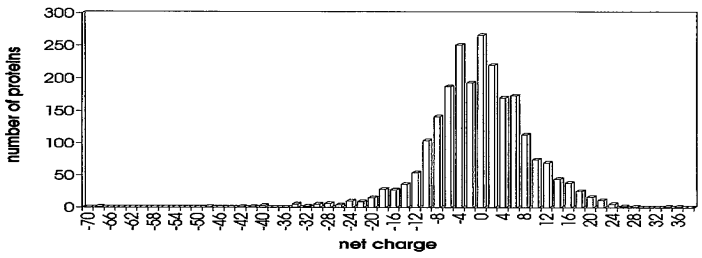


Figure 3 Net charge distribution in *E. coli* proteins.

Figure 3 shows the distribution of the net charge of *E. coli* proteins. Predictably, most of the proteins group around zero; i.e., they have only a small excess of positively charged over negatively charged residues or vice versa. However, several proteins have a conspicuously high positive or negative charge (Table 2). In most of these proteins, the charged residues are distributed diffusely and do not form prominent clusters. It is also of note that most of the species with high positive net charge are DNA- or RNA-binding proteins (Table 2).

Methods for identification of globular and nonglobular domains in proteins based on compositional complexity of their sequences have been developed recently and implemented in the SEG program (59, 60). Using this technique, we found that, predictably, the majority of *E. coli* proteins consist predominantly of globular domains. However, 707 proteins (30%) contain relatively large (>20% of the entire length) nonglobular regions and 104 proteins (4.3%) are predicted to be completely nonglobular (that is, the number of amino acid residues in the predicted globular regions is so small that they are unlikely to form a compact domain); examples of predicted nonglobular proteins are given in Table 3.

Using a modified version of the SEG program (J. C. Wootton, personal communication) to predict putative transmembrane segments in *E. coli* proteins, we observed that 394 proteins (17% of the total) contain at least one such region and 294 proteins (about 13%) contain multiple predicted transmembrane segments. Altogether, the predicted transmembrane segments make up about 3% of the total length of the known *E. coli* proteins. Notably, the fraction of (putative) membrane proteins in *E. coli* is considerably lower than the 30 to 35% predicted for *S. cerevisiae* (15, 25).

SEQUENCE CONSERVATION IN *E. COLI* PROTEINS: VIEW FROM THE OUTSIDE

Figure 4 summarizes the results of our comparison of the *E. coli* protein sequences with general sequence databases. The rate of sequence similarity detection by the approach outlined above is dramatically high—about 86% of the *E. coli* proteins were found to be related to other proteins in the database (Fig. 4A). This is a higher proportion than observed even in the most detailed recent analyses of large protein sets—about 70% for yeast proteins (25; P. Bork, personal communication) and 75% for *Mycoplasma capricolum* proteins (7). This result could be considered trivial if most of the sequence conservation was due to high similarity between *E. coli* proteins and their homologs from closely related bacteria. This is, however, not the case—about 60% of the available *E. coli* protein sequences show conservation with proteins from distantly related bacteria (defined as those outside the domain *Proteobacteria* in the overall phylogenetic tree based on rRNA sequences [35]), and about 40% contain "ancient conserved regions" (19, 20) in common with eukaryotic and/or archaeal proteins. Altogether, about two-thirds of the known *E. coli* proteins are conserved at least at the level of distantly related bacteria (Fig. 4A). It has to be kept in mind that despite the rapid accumulation of sequence information, the current databases are still incomplete, and the fraction of *E. coli* proteins for which sequence conservation is discernible will certainly further increase in the near future. These observations indicate that the great majority of the *E. coli* proteins contain sequences that are subject to a strong stabilizing selection, presumably as a result of functional constraints on the protein structure, and have been conserved through more than a billion years of evolution.

The last few years have been marked by a rapid growth of the number of resolved protein structures (22). The current version (April 1994) of the Protein Data Bank contains the structures of 56 *E. coli* proteins. A comparison of the *E. coli* protein sequences with the sequences of the proteins in the Protein Data Bank showed that an additional 169 *E. coli* proteins are significantly similar (BLAST score, >70) to proteins with known structure, thus allowing structural inferences for about 10% of the predicted *E. coli* proteins (Fig. 4B). Additional information on structural motifs could be derived from the analysis of *E. coli* protein clusters (see below).

Combining the information on sequence similarities with functional information (functional information on a somewhat smaller set of *E. coli* gene products has been summarized recently [42]), we find that for the majority of the *E. coli* proteins, there is both sequence conservation and a known (at least in general terms) biological function (Fig. 4C). The second largest fraction of the *E. coli* proteins—about one-quarter—includes proteins or predicted gene products that do not have a known function but show sequence similarity to other proteins. For about two-thirds of these proteins, the putative function could be predicted on the basis of sequence conservation; the rest are related to other uncharacterized proteins, allowing one to pinpoint regions that are important for their as yet unknown function (Fig. 4C). Clearly, prediction of the functions of uncharacterized proteins is one of the most important results of sequence analysis and of the genome project in general. Description of any significant proportion of these predictions is beyond the scope of this chapter, but below we discuss one typical example of such a novel finding. Only a small fraction of the predicted *E. coli* proteins are true unknowns, having neither an established function nor sequence conservation (Fig. 4C).

TABLE 2 Proteins and predicted gene products with high net charge in *E. coli*

Protein	No. of amino acids	Charge	Activity or function	Comment
YdbA	1,540	−68	Unknown	Predicted protein interrupted by an insertion element; rich in Asp; predicted coiled coil domains
RecE	866	−53	Exonuclease VIII; recombination	Rich in Glu; no homologs detected in *E. coli* or other organisms
FtsY	497	−48	Membrane-associated GTPase; cell division	Rich in Glu; the negatively charged N-terminal domain is distinct from the C-terminal, conserved GTPase domain
RecC	1,122	−42	Exonuclease V subunit; repair, recombination	Rich in both Asp and Glu; no homologs detected in *E. coli* or other organisms
MetH	1,227	−42	Methionine synthase	Rich in Glu; moderately conserved in distantly related bacteria
RpoD	613	−40	RNA polymerase σ^{70} factor	Rich in both Glu and Asp; contains a 24-amino-acid negative charge cluster; the negatively charged domain is located upstream of the core polymerase-binding and DNA-binding domains that are highly conserved in distantly related bacteria
RplB	273	+38	50S ribosomal protein L2	Strongly enriched in Arg-Lys doublets; highly conserved in eukaryotes and bacteria
Bax	274	+35	Unknown	Weak similarity to yeast cell cycle-regulatory protein HPC2; nucleic acid binding?
Int	387	+29	Cryptic lambdoid prophage integrase	Rich in Lys; conserved in bacteriophages
RhlE	488	+28	Putative RNA helicase	Rich in Arg; highly conserved in eukaryotes and bacteria
PriA	732	+27	DNA helicase; primosome subunit	Moderate conservation with other helicases
YfeE	289	+26	Putative amidase	Moderately conserved in distantly related bacteria
XseA	456	+26	Exonuclease VII large subunit	Rich in Arg; no conservation in *E. coli* or other organisms detected

The distribution of the *E. coli* ancient conserved proteins among functional classes showed a strong bias toward metabolic enzymes, with a particular preponderance of ATP/GTP-utilizing enzymes and oxidoreductases (Fig. 5A). Proteins of unknown function made up only a small fraction of this set of highly conserved proteins (Fig. 5A). When the *E. coli* proteins showing conservation with proteins from distantly related bacteria were similarly analyzed, the bias toward metabolic enzymes was less dramatic, with a significant proportion of the conserved regions being in regulatory proteins and in membrane proteins (Fig. 5B).

An important question with regard to the "ancient conserved regions" in *E. coli* proteins is whether most of them actually date back to the radiation of bacteria, eukaryotes, and the *Archaea* from a common ancestor or whether a significant fraction is due to the transfer of genes from the genomes of bacterial endosymbionts—chloroplasts and mitochondria—to the eukaryotic nucleus and/or actual horizontal gene transfer. Transfer of genes from the endosymbionts to the nucleus is thought to be a major evolutionary trend, whereas other forms of prokaryote-eukaryote horizontal gene transfer appear to be much less frequent (18, 36, 52). Indeed, examination of the features of eukaryotic proteins that show the highest similarity to *E. coli* proteins indicates that at least some of them function in chloroplasts or mitochondria and are likely to be encoded by relocated genes (Table 4). Some other strikingly conserved proteins, e.g., phosphoenolpyruvate (PEP) carboxylase (*E. coli* Ppc protein), represent a less trivial case of a cytoplasmic eukaryotic protein that

is much more similar to its *E. coli* ortholog than the latter is to the ortholog from a distant bacterial species (Table 4). Given the participation of PEP carboxylase in photosynthesis (50), it still appears likely that in this particular case, the unusual pattern of sequence conservation may be due to endosymbiont-nuclear gene transfer. Obviously, the endosymbiont explanation is not a reasonable one for the very high similarity between *E. coli* proteins and their archaeal homologs, e.g., formate dehydrogenase (FdhF; Table 4). In cases like this, the only possibility, besides conservation since the time of bacterial-archaeal radiation, is horizontal gene transfer.

TABLE 3 Some *E. coli* proteins with predicted large nonglobular domains

Protein	Nonglobular/ total length	% Nonglobular	Globular domain(s)	Features of the nonglobular domain(s)	Function
FliC	497/498	100	No	Enriched in threonine and serine	Flagellin
DamX	422/427	99	No	Negatively charged clusters	Unknown
FimA	173/175	99	No		Fimbrial protein
DinF	436/459	95	No	Hydrophobic, leucine rich	DNA damage inducible (repair ?)
TolA	398/421	95	No	Numerous alanine-, lysine-, and glutamic acid-rich repeats; coiled coil ?	Colicin uptake
FliD	433/468	93	No	Enriched in threonine and serine	Flagellar capping protein
SbcC	835/1,048	80	Yes	Predicted coiled-coil domains	ATPase involved in DNA recombination
LpxD	227/341	67	Yes	Heptamer repeats with isoleucine periodicity	Acyltransferase involved in lipid A biosynthesis
MukB	969/1534	63	Yes	Predicted coiled-coil domains	ATPase involved in cell partitioning; mechanochemical function ?
DnaX	394/643	61	Yes	Predicted coiled-coil domains	ATPase involved in DNA replication
Lhr	646/1538	42		Leucine-rich repeats apparently distinct from coiled coil	Putative DNA or RNA helicase

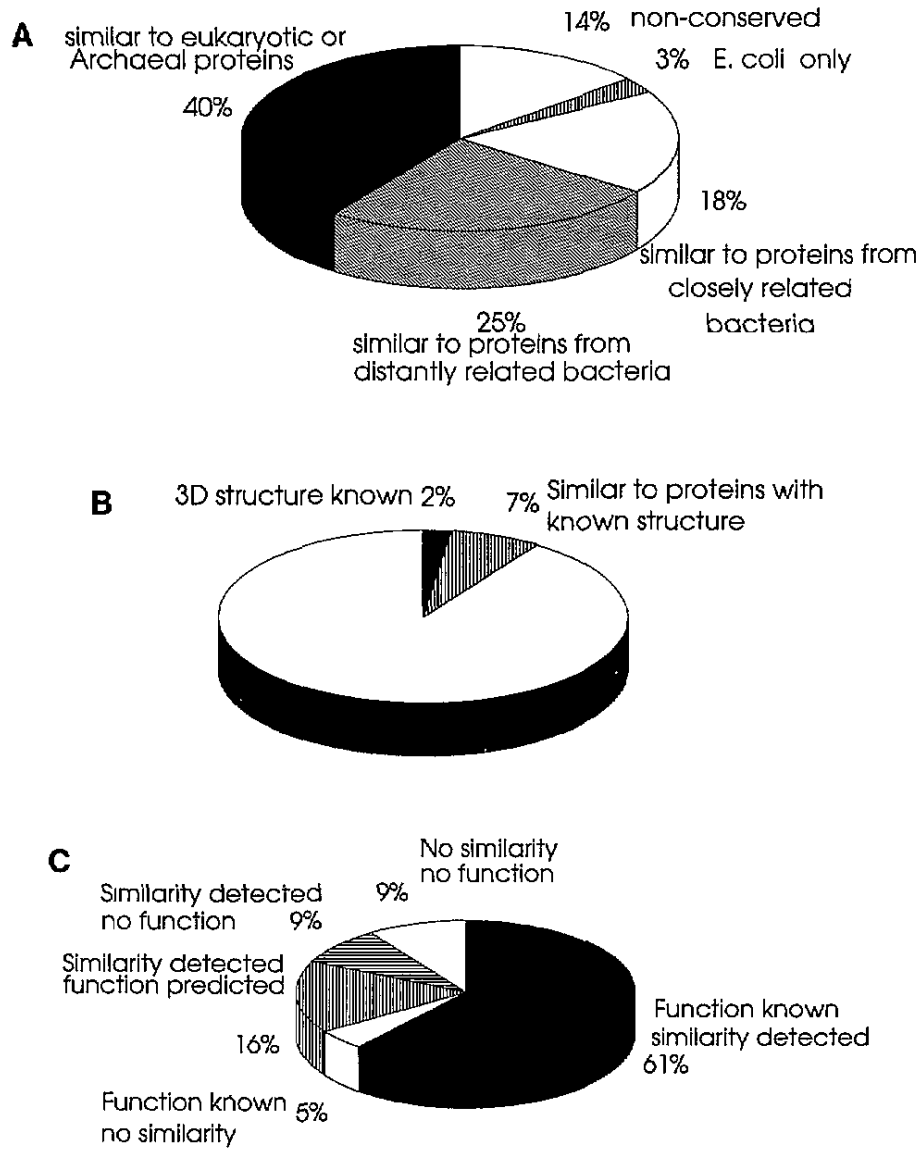


Figure 4 Sequence conservation in *E. coli* proteins. (Total, 2,419 proteins.) (A) Levels of sequence conservation. (B) Sequence conservation and information on three-dimensional structure. (C) Sequence conservation and functional information.

However, a more general and perhaps a more important observation is that the mean similarity score between *E. coli* proteins and their eukaryotic or archaeal homologs is 174 (median of the score distribution, 109), whereas a considerably higher mean score of 230 (median, 146) was observed with homologs from distantly related bacteria. A comparison of the two score distributions suggests that the majority of the "ancient conserved regions" probably date back to the original point of radiation between bacteria, eukaryotes, and the *Archaea.* Indeed, the distribution for eukaryotic (archaeal) homologs of *E. coli* proteins is shifted toward lower scores as compared with the distribution for homologs from distantly related bacteria, just as would be expected if the regions that are conserved in *E. coli* proteins and their homologs from eukaryotes and the *Archaea* have undergone longer evolution (Fig. 6). In accord with this, moderately conserved *E. coli* proteins typically have higher similarity scores with the orthologs from distantly related bacteria than with those from eukaryotes or the *Archaea* (see examples in Table 4).

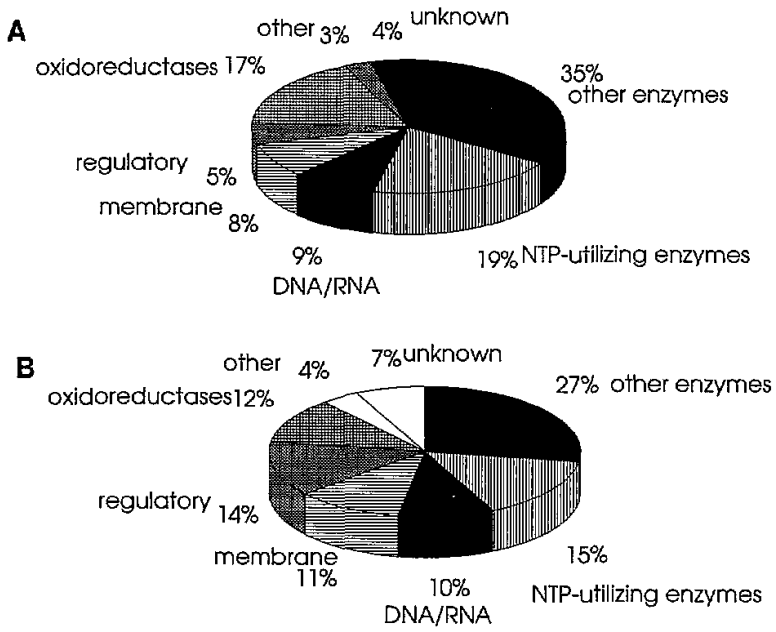


Figure 5 Functional classes of evolutionarily conserved *E. coli* proteins. (A.) Proteins with homologs in eukaryotes and/or the *Archaea*. (B) Proteins with homologs in distantly related bacteria. DNA/RNA indicated DNA-and RNA-binding proteins as well as enzymes (other than nucleoside triphosphatases) involved in gene replication, recombination, repair, and expression.

What can one surmise about the 340 *E. coli* proteins (about 14% of the total) that are not similar to any other protein in the current databases? Notably, the majority of the proteins in this set (about 70%) have not been functionally characterized. Those whose functions are known represent a spectrum of activities (Table 5), some of which simply have not yet been discovered in other organisms, whereas others (e.g., dGTP triphosphatase [39]) may be specific for *E. coli* and closely related enterobacteria.

SEQUENCE CONSERVATION IN *E. COLI* PROTEINS: VIEW FROM WITHIN

Clustering of the *E. coli* proteins sequences by using the "greedy" algorithm described above, together with motif searches designed to resolve the multidomain protein problem and to delineate superclusters, showed that about one-half of the proteins belong to 299 clusters and 70 superclusters of paralogs (Fig. 7A). Most of the clusters are small, with only 2 to 4 members, but there are several clusters with more than 10 members (Fig. 8). The observed distribution of the cluster size appears to be compatible with a stochastic duplication model; a simple simulation with uniform duplication and retention rates produced a distribution similar to the one in Fig. 8 (data not shown) but, obviously, without the larger clusters. To explain the existence of these clusters, one has to postulate that for some functions, duplication with subsequent diversification provides a significant selective advantage. Inspection of the small and large clusters from the functional point of view clearly indicates two types of functions that may be subject to such a selection: metabolite transport and regulation of gene expression. Large clusters are mostly composed of transport proteins and proteins involved in different types of regulation (Table 6), whereas metabolic enzymes typically form small clusters (Table 7). Strikingly, there seem to be essential functional classes of proteins in *E. coli* that are not prone to gene duplication, notably catalytic subunits of DNA and RNA polymerases and ribosomal proteins. It is likely that duplication is selected against in these cases, but so far we do not understand the nature of this apparent negative selection.

TABLE 4 Some highly conserved proteins in *E. coli*[a]

Protein	Length (amino acid residues)	Best hit with distant bacteria	Score, best alignment[b]	Best eukaryotic/archaeal hit	Score, best alignment	Function or activity	Paralogs in *E. coli*	Comment
Acs	652	BSSRFAD_2	111; 36%/69	PBLACOASYN_1	1,245; 57%/384	Ac-CoA[c] synthetase	Aas, EntE, EntF, FadD, YdiD′, YaaM	No bacterial ortholog
GlgP	809	PHSM_STRPN	110; 44%/47 (partial)	PHS2_HUMAN	1,227; 51%/442	Glycogen phosphorylase	MalP	No bacterial ortholog
FumC	467	FUMH_BACSU	1,000; 57%/339	FUMH_RAT	1,152; 55%/397	Fumarase	ArgH, AspA, PurB	Mitochondrial
Ppc	883	CAPP_CORGL	261; 48%/93	CAPP_MEDSA	1,016; 50%/401	PEP carboxylase	Not known	Cytoplasmic chloroplast origin ?
FdhF	715	S36605 (*Synechocou*)	374; 30%/312	FDHA_METFO	947; 48%/377	Formate dehydrogenase	Nine dehydrogenases	Archaeal ortholog
GcvP		None	NA	GCSP_PEA	928; 58%/284	Gly dehydrogenase	Not known	mitochondrial
YiaY	382	MEDH_BACMT	887; 46%/375	ADH4_YEAST	919; 61%/287	Alcohol dehydrogenase	Eight dehydrogenases	
TrpB	397	TRPB_LACLA	567; 51%/301	TRP1_ARATH	860; 58%/279	Trp synthase	Not known	Chloroplast
AcnA	891	ACON_BACSU	275; 47%/115 (partial)	IREB_RABIT	841; 42%/449	Aconitase	YacI, LeuC, LeuD	Cytoplasmic
NagB	266	MC068_1	151; 41%/78 (partial)	HUMORFKG1E_1	776; 56%/258	Glucosamine-6-phosphate isomerase	YieK	
Udk	213	None	NA	MUSURKI_1	201; 40%/90	Uridine kinase	Not known	
RpoB	1,342	BACRPLL_3	651; 68%/171	RPC2_YEAST	199; 40%/110	RNA polymerase β	Not known (none ?)	
PrfC	529	EFG_ANANI	226; 29%/154	EFGM_YEAST	179; 30%/126	Peptide release factor	14 GTPases	No ortholog
PolB	783	None	NA	DPOA_HUMAN	141; 38%/71	DNA polymerase II	Not known (none ?)	No bacterial homologs

[a]The table includes the top 10 hits between *E. coli* proteins and their homologs from eukaryotes or *Archaea* and, for comparison, four proteins with moderate similarity to eukaryotic homologs.
[b]For each protein, the highest BLASTP score, the percentage of identical amino acid residues in the best ungapped alignment, and the length of this alignment are indicated. NA, not applicable.
[c]Ac-CoA, acetyl coenzyme A.

The extent of apparent duplication in certain functional classes of *E. coli* genes becomes even more dramatic when one considers superclusters defined by the conservation of sequence motifs. The four largest superclusters account for almost one-quarter of all known *E. coli* proteins (Fig. 7B). Figure 9 shows the aligned sequences of the (predicted) ATP/GTP-binding P-loops (23, 49, 57) that form the motif defining the second largest supercluster of paralogous proteins in *E. coli.*

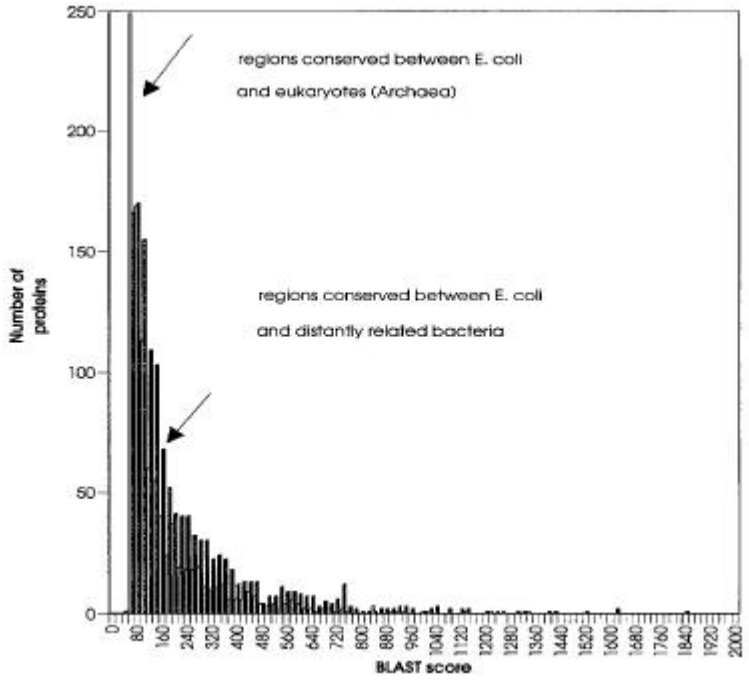


Figure 6 Distribution of similarity scores between *E. coli* proteins and their homologs from distantly related organisms. For each *E. coli* protein, the highest score with a protein from eukaryotes/*Archaea* or from distantly related bacteria (see text) produced by BLASTP is included.

The distribution of the cluster members on the *E. coli* chromosome shows a clear prominence of closely spaced, related genes (Fig. 10). About 50 pairs of such genes appear to be the result of actual tandem duplications, without intervening genes. All but three of these pairs are transcribed in the same direction and belong to the same operon (data not shown). These tandem duplications may be placed in the same category of evolutionary events with intragenic duplications, of which several cases are apparent in *E. coli*, e.g., in such members of the ATP-binding cassette (ABC) transporter ATPase cluster as AraG and RbsA (16, 21) or in ATPase subunits of ATP-dependent proteases such as ClpA and ClpB (17).

TABLE 5 Some *E. coli* proteins with known function in search of relatives

Protein	Length	Function	Proteins with analogous function in other organisms
FucI	591	L-Fucose isomerase	Unknown
RecC	1,122	Exonuclease V subunit, DNA repair and recombination	Numerous groups of exonucleases
Dgt	505	dGTPase	Unknown
PepD	485	β-Ala-His dipeptidase	Unknown but other families of dipeptidases exist
SbcB	475	Exonuclease I, DNA repair	Numerous groups of exonucleases
SelA	463	Selenocysteine synthase	Unknown
CreD	450	Inner membrane protein	Unknown
HipA	440	Resistance to inhibitors of peptidoglycan and DNA synthesis	Unknown
ProX	330	Glycine betaine-binding periplasmic protein	Several families of solute-binding proteins
HolA	343	DNA polymerase III δ subunit	Unknown
Gsk	434	IMP-GMP kinase	Numerous nucleotide kinases

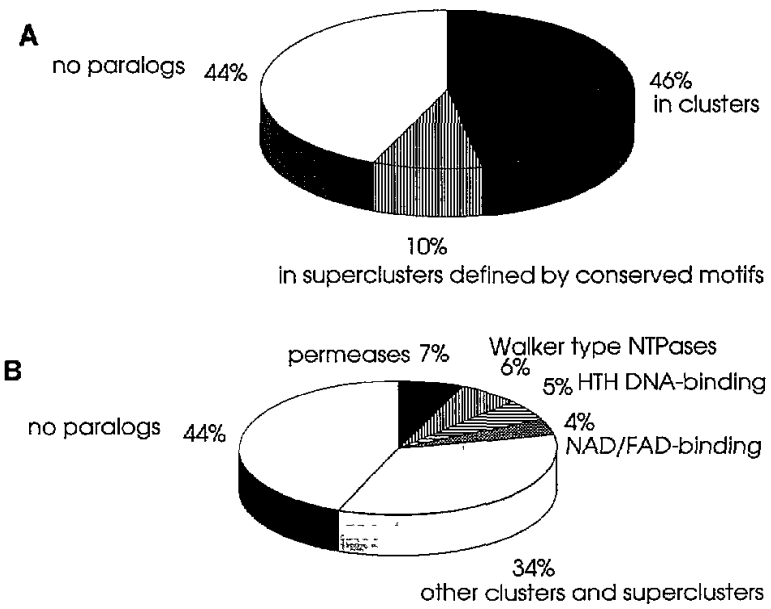


Figure 7 Clustering of paralogous *E. coli* proteins. (A) Clusters and superclusters. (B) The four largest superclusters.

Many of the duplications in the evolution of the *E. coli* chromosome apparently have involved more than one gene. The most striking example of such a cassette duplication is presented by the *dpp, opp,* and *nik* operons, which encode proteins mediating dipeptide, oligopeptide, and nickel transport, respectively, and contain five paralogous genes in a row (1, 30). There are several examples of apparent three- and two-gene cassette duplications in the *E. coli* chromosome (data not shown). Not surprisingly, most of them include genes coding for transport and regulatory proteins (see above).

A weak, large-scale periodicity in the distribution of paralogous genes along the chromosome, with the distance between paralogous genes tending to be a multiple of 6 to 7 min of the *E. coli* chromosome, was observed (Fig. 10). A similar periodicity was noticed previously in a study with a much smaller set of related *E. coli* genes (26). One possible explanation for this apparent periodicity involves duplication of large segments of the chromosome early in evolution, whereas another line of thinking may relate the periodicity to the nucleoid superstructure; additional analysis is needed to critically assess these possibilities.

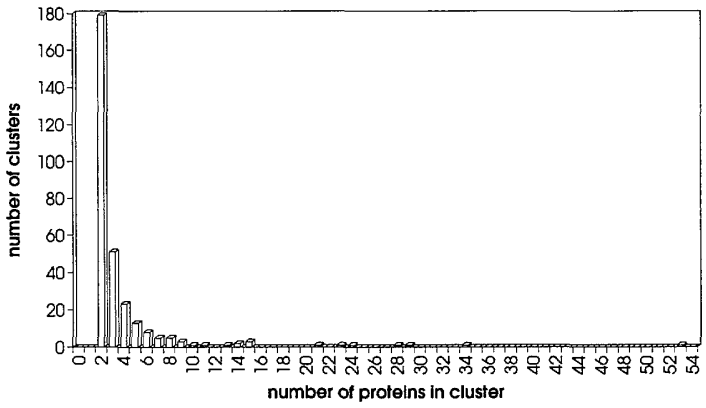


Figure 8 Distribution of the number of sequences in *E. coli* protein clusters.

Our analysis of *E. coli* protein clusters revealed 91 proteins that consist of two or more distinct, conserved domains belonging to different clusters (evidently, there are a lot more multidomain proteins altogether, but in many of them, one or both domains do not have paralogs in *E. coli*). Most of these are accounted for by the two-component, sensor-receiver signal-transducing system (37, 38), the membrane sugar phosphotransferase and transport (PTS) system (41, 48), and various proteins, in which the DNA-binding, helix-turn-helix (HTH) domain is combined with other domains, e.g., repressors containing the HTH domain and a sugar-binding domain (32, 53, 55).

Other examples of proteins with distinct conserved domains are less well known, and here we discuss one which demonstrates not only intricate multidomain organization of some proteins but also the potential of sequence analysis in predicting new protein functions. Lon is a protease involved in ATP-dependent proteolysis of abnormal and short-lived proteins in *E. coli* and other organisms (reviewed in reference 17). The ATPase domain containing the typical conserved motifs has been located in the N-terminal portion of Lon (12), whereas the catalytic serine of the protease domain has been identified in the C-terminal portion (5). In addition, a region of significant similarity has been detected between regions from the C-terminal domains of Lon and the *E. coli* protein Sms, which has been implicated in the resistance to methyl methanesulfonate but otherwise remains functionally uncharacterized (31). When the region of conservation between Sms, its *B. subtilis* homolog, and Lon proteins from different species was used to derive a position-dependent weight matrix and search the database (54), a related motif was identified in several other proteins, including *E. coli* proteins HtrA, HhoA, HhoB, and YifB. Subsequent analysis involving additional database searches and multiple alignment resulted in the delineation of a group of proteins containing ATPase and protease domains (Fig. 11). HtrA, HhoA, and HhoB are serine proteases that are related to the classical, chymotrypsin-like proteases more closely than Lon is and contain the characteristic histidine-aspartate-serine catalytic triad (6, 27; S. Bass, Q. Gu, and A. Goddard, GenBank accession number U1566 1). In Lon, HtrA, HhoA, and HhoB, the highly conserved motif is located directly after the catalytic serine, suggesting that it may be a specific form of a protein-binding site (Fig. 11A). Putative catalytic serines could also be identified in YifB and Sms (Fig. 11A). Like Lon, YifB and Sms contain clearly defined ATPase domains (Fig. 11A) (24, 31). Thus, it is possible that YifB and Sms are two new *E. coli* ATP-dependent proteases. However, the putative catalytic serine is replaced by alanine in the *B. subtilis* homolog of Sms, raising the alternative possibility that Sms has lost the protease activity. Another notable aspect of this example is the inversion of the ATPase and putative protease domains in YifB compared with Lon and Sms (Fig. 11B). Domain shuffling is also typical of some other clusters of multidomain proteins in *E. coli*, e.g. the PTS system (41).

TABLE 6 The 10 largest clusters of paralogous proteins in *E. coli*

Cluster	No. of proteins	Function	Homologs in distantly related bacteria	Homologs in eukaryotes and the *Archaea*
ABC transporter ATPases	54	ATP-dependent membrane transport; DNA repair	Yes	Yes
Permeases (AraE related)	34	Membrane transport	Yes	Yes
HTH proteins (Ada related)	29	Transcription regulation	Yes	No
Receiver domains	28	Membrane signal transduction	Yes	Yes
Sensor domains	24	Membrane signal transduction	Yes	Yes
Permeases (ArtM related)	24	Membrane transport	Yes	Yes
HTH proteins (CynR related)	21	Transcription regulation	Yes	No
Sugar-binding domains	15	Metabolite transport	Yes	
DEAD/H helicases	15	RNA/DNA duplex unwinding	Yes	Yes
GTPases	15	GTP-dependent processes	Yes	Yes

			Cluster	Function/activity
Abc	33:	IYGVIGASGA**GKSTLIRCVN**	yes	transport
UvrA	26:	LIVVTGLSGS**GKSSLAFDTL**	yes	repair
UvrA	635:	FTCITGVSGS**GKSTLINDTL**	yes	repair
MutS	609:	MLIITGPNMG**GKSTYMRQTA**	no	repair
RecN	24:	MTVITGETGA**GKSIAIDALG**	no	repair/recombination
RecF	25:	FNFLVGANGS**GKTSVLEAIY**	no	repair/recombination
SbcC	32:	LFAITGPTGA**GKTTLLDAIC**	no	repair/recombination
RecB	18:	ERLIEASAGT**GKTFTIAALY**	yes	recombination/helicase
DeaD	62:	DVLGMAQTGS**GKTAAFSLPL**	yes	helicase
MalT	34:	LALITSPAGY**GKTTLISQWA**	no	transcription regulation
ClpA	209:	NPLLVGESGV**GKTAIAEGLA**	yes	ATPase subunit of protease
ClpA	490:	SFLFAGPTGV**GKTEVTVQLS**	yes	ATPase subunit of protease
Lon	351:	ILCLVGPPGV**GKTSLGQSIA**	no	ATPase subunit of protease
DnaA	171:	PLFLYGGTGL**GKTHLLHAVG**	no	replication initiation
DnaC	101:	SFIFSGKPGT**GKNHLAAAIC**	no	replication initiation
DnaB	226:	LIIVAARPSM**GKTTFAMNLV**	no	replication/helicase
DnaX	40:	AYLFSGTRGV**GKTSIARLLA**	yes	replication
MiaA	12:	AIFLMGPTAS**GKTALAIELR**	no	tRNA modification
HydG	164:	TVLIHGDSGT**GKELVARAIH**	yes	transcription regulation
Sms	97:	AILIGGNPGA**GKSTLLLQTL**	no	ATP-dependent protease ?
RecA	62:	IVEIYGPESS**GKTTLTLQVI**	no	repair/recombination
MinD	5:	IIVVTSGKGG**GKTTSSAAIA**	yes	chromosome partitioning
HopB	217:	LVLVTGPTGS**GKTVTLYSAL**	yes	secretion
CysN	29:	RFLTCGSVDD**GKSTLIGRLL**	yes	GTPase
Ffh	102:	VVLMAGLQGA**GKTTSVGKLG**	yes	GTPase
HypB	106:	VLNLVSSPGS**GKTTLLTETL**	no	GTPase ?
YeiR	4	TNLITGFLGS**GKTTSILHLL**	yes	GTPase ?
AtpA	164:	RELIIGDRQT**GKTALAIDAI**	no	H^+ ATPase
Rho	172:	RGLIVAPPKA**GKTMLLQNIA**	no	transcription/helicase
AroK	34:	NIFLVGPMGA**GKSTIGRQLA**	yes	shikimate kinase
CoaA	90:	IISIAGSVAV**GKSTTARVLQ**	yes	panthotenate kinase
Udk	10:	IIGIAGASAS**GKSLIASTLY**	no	uridine kinase
Adk	2:	RIILLGAPGA**GKGTQAQFIM**	no	adenylate kinase
consensus		**.UUU.O....GK$.U...U.**		

Figure 9 **A**lignment of the purine nucleoside triphosphate-binding P-loops from proteins belonging to the "Walker type" ATPase/GTPase supercluster. For each sequence is indicated whether it belongs to a cluster defined on the basis of pairwise similarity or was included in the supercluster based on motif conservation. Question marks indicated proteins for which function was predicted in the course of our analysis. The consensus included amino acid residues conserved in most of the proteins containing the motif; U indicates a bulky hydrophobic residue; O indicates a small residue (G, A, or S); $ indicates serine or threonine; and a dot indicates any residue. The residues conforming with the consensus are highlighted by boldface type.

TABLE 7 Examples of small clusters of paralogous enzymes in *E. coli*

Cluster	Proteins	Enzymatic activity	Homologs in distantly related bacteria	Homologs in eukaryotes/*Archaea*	Comment
Acetyltransferase	AccB	Biotin carboxyl carrier protein	Yes	Yes	AceF and SucB are subunits of analogous enzymatic complexes and share highly significant similarity, whereas the conservation in AccB is limited to a motif around the biotin-binding site
	AceF	Dihydrolipoamide acetyltransferase	Yes	Yes	
	SucB	Dihydrolipoamide acetyltransferase	Yes	Yes	
Acetate kinase	AckA	Acetate kinase	Yes	Yes	AckA is involved in acetyl coenzyme A formation; the activity of YhaA′ can be predicted from sequence similarity
	YhaA′	??	Yes	Yes	
Acid phosphatase	Agp	Glucose-1-phosphatase	No	No	Genes located close to each other, even though not a tandem duplication; periplasmic enzymes
	AppA	pH 2.5 acid phosphatase	No	No	
Alanine racemase	Alr	Alanine racemase	Yes	No	Anabolic enzyme involved in cell wall peptidoglycan synthesis
	DadX	Alanine racemase	Yes	No	Catabolic enzyme
Aminotransferase	AspC	Aspartate aminotransferase	Yes	Yes	Biosynthetic enzymes with different substrate specificity. Both sequences are more similar to eukaryotic homologs than to those from distantly related bacteria
	TyrB	Aromatic amino acid aminotransferase	No	Yes	

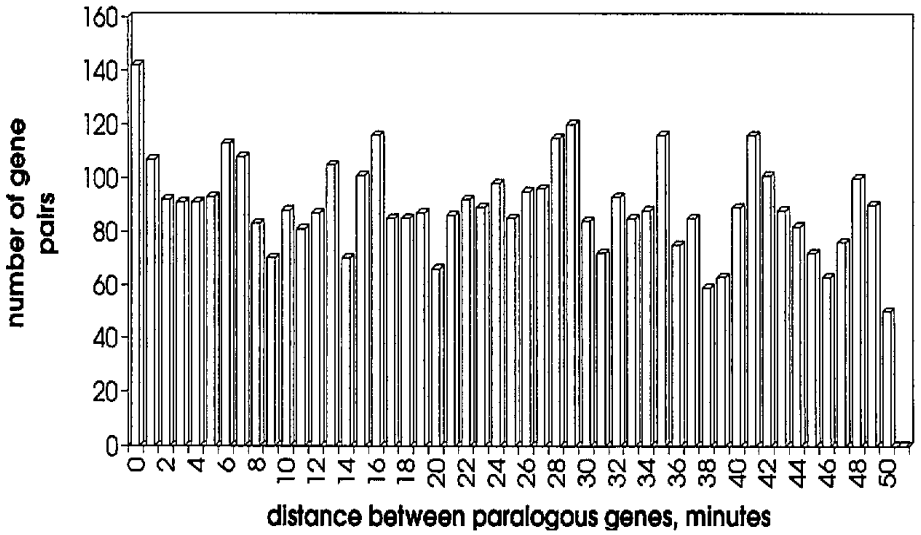
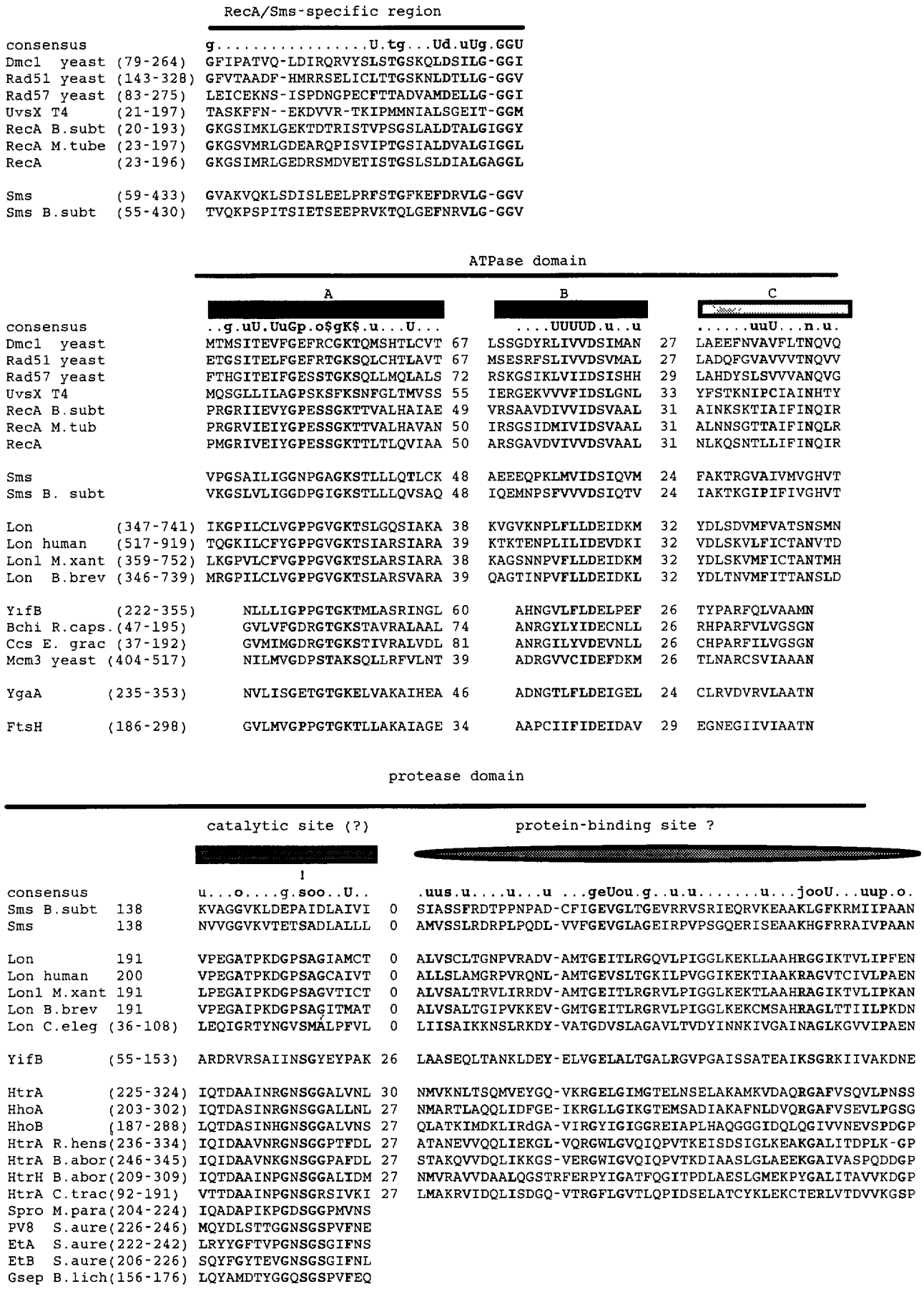


Figure 10 Distribution of genes encoding paralogous proteins along the *E. coli* chromosome.


RecA/Sms-specific region
consensus g.....U.tg...Ud.uUg.GGU
Dmc1 yeast (79-264) GFIPATVQ-LDIRQRVYSLSTGSKQLDSILG-GGI
Rad51 yeast (143-328) GFVTAADF-HMRRSELICLTTGSKNLDTLLG-GGV
Rad57 yeast (83-275) LEICEKNS-ISPDNGPECFTTADVAMDELLG-GGI
UvsX T4 (21-197) TASKFFN--EKDVVR-TKIPMMNIALSGEIT-GGM
RecA B.subt (20-193) GKGSIMKLGEKTDTRISTVPSGSLALDTALGIGGY
RecA M.tube (23-197) GKGSVMRLGDEARQPISVIPTGSIALDVALGIGGL
RecA (23-196) GKGSIMRLGEDRSMDVETISTGSLSLDIALGAGGL
Sms (59-433) GVAKVQKLSDISLEELPRFSTGFKEFDRVLG-GGV
Sms B.subt (55-430) TVQKPSPITSIETSEEPRVKTQLGEFNRVLG-GGV
ATPase domain
A
B
C
consensus ..g.uU.UuGp.ogK.u...U... ....UUUUD.u..u .....uuU...n.u.
Dmc1 yeast MTMSITEVFGEFRCGKTQMSHTLCVT 67 LSSGDYRLIVVDSIMAN 27 LAEEFNVAVFLTNQVQ
Rad51 yeast ETGSITELFGEFRTGKSQLCHTLAVT 67 MSESRFSLIVVDSVMAL 27 LADQFGVAVVVTNQVV
Rad57 yeast FTHGITEIFGESSTGKSQLLMQLALS 72 RSKGSIKLVIIDSISHH 29 LAHDYSLSVVVANQVG
UvsX T4 MQSGLLILAGPSKSFKSNFGLTMVSS 55 IERGEKVVVFIDSLGNL 33 YFSTKNIPCIAINHTY
RecA B.subt PRGRIIEVYGPESSGKTTVALHAIAE 49 VRSAAVDIVVIDSVAAL 31 AINKSKTIAIFINQIR
RecA M.tub PRGRVIEIYGPESSGKTTVALHAVAN 50 IRSGSIDMIVIDSVAAL 31 ALNNSGTTAIFINQLR
RecA PMGRIVEIYGPESSGKTTLTLQVIAA 50 ARSGAVDVIVVDSVAAL 31 NLKQSNTLLIFINQIR
Sms VPGSAILIGGNPGAGKSTLLLQTLCK 48 AEEEQPKLMVIDSIQVM 24 FAKTRGVAIVMVGHVT
Sms B. subt VKGSLVLIGGDPGIGKSTLLLQVSAQ 48 IQEMNPSFVVVDSIQTV 24 IAKTKGIPIFIVGHVT
Lon (347-741) IKGPILCLVGPPGVGKTSLGQSIAKA 38 KVGVKNPLFLLDEIDKM 32 YDLSDVMFVATSNSMN
Lon human (517-919) TQGKILCFYGPPGVGKTSIARSIARA 39 KTKTENPLILIDEVDKI 32 VDLSKVLFICTANVTD
Lon1 M.xant (359-752) LKGPVLCFVGPPGVGKTSLARSIARA 38 KAGSNNPVFLLDEIDKM 32 YDLSKVMFICTANTMH
Lon B.brev (346-739) MRGPILCLVGPPGVGKTSLARSVARA 39 QAGTINPVFLLDEIDKL 32 YDLTNVMFITTANSLD
YifB (222-355) NLLLIGPPGTGKTMLASRINGL 60 AHNGVLFLDELPEF 26 TYPARFQLVAAMN
Bchi R.caps. (47-195) GVLVFGDRGTGKSTAVRALAAL 74 ANRGYLYIDECNLL 26 RHPARFVLVGSGN
Ccs E. grac (37-192) GVMIMGDRGTGKSTIVRALVDL 81 ANRGILYVDEVNLL 26 CHPARFILVGSGN
Mcm3 yeast (404-517) NILMVGDPSTAKSQLLRFVLNT 39 ADRGVVCIDEFDKM 26 TLNARCSVIAAAN
YgaA (235-353) NVLISGETGTGKELVAKAIHEA 46 ADNGTLFLDEIGEL 24 CLRVDVRVLAATN
FtsH (186-298) GVLMVGPPGTGKTLLAKAIAGE 34 AAPCIIFIDEIDAV 29 EGNEGIIVIAATN
protease domain
catalytic site (?)
protein-binding site ?
!
consensus u...o....g.soo..U.. .uus.u....u...u ...geUou.g..u.u.......u...jooU...uup.o.
Sms B.subt 138 KVAGGVKLDEPAIDLAIVI 0 SIASSFRDTPPNPAD-CFIGEVGLTGEVRRVSRIEQRVKEAAKLGFKRMIIPAAN
Sms 138 NVVGGVKVTETSADLALLL 0 AMVSSLRDRPLPQDL-VVFGEVGLAGEIRPVPSGQERISEAAKHGFRRAIVPAAN
Lon 191 VPEGATPKDGPSAGIAMCT 0 ALVSCLTGNPVRADV-AMTGEITLRGQVLPIGGLKEKLLAAHRGGIKTVLIPFEN
Lon human 200 VPEGATPKDGPSAGCAIVT 0 ALLSLAMGRPVRQNL-AMTGEVSLTGKILPVGGIKEKTIAAKRAGVTCIVLPAEN
Lon1 M.xant 191 LPEGAIPKDGPSAGVTICT 0 ALVSALTRVLIRRDV-AMTGEITLRGRVLPIGGLKEKTLAAHRAGIKTVLIPKAN
Lon B.brev 191 VPEGAIPKDGPSAGITMAT 0 ALVSALTGIPVKKEV-GMTGEITLRGRVLPIGGLKEKCMSAHRAGLTTIILPKDN
Lon C.eleg (36-108) LEQIGRTYNGVSMALPFVL 0 LIISAIKKNSLRKDY-VATGDVSLAGAVLTVDYINNKIVGAINAGLKGVVIPAEN
YifB (55-153) ARDRVRSAIINSGYEYPAK 26 LAASEQLTANKLDEY-ELVGELALTGALRGVPGAISSATEAIKSGRKIIVAKDNE
HtrA (225-324) IQTDAAINRGNSGGALVNL 30 NMVKNLTSQMVEYGQ-VKRGELGIMGTELNSELAKAMKVDAQRGAFVSQVLPNSS
HhoA (203-302) IQTDASINRGNSGGALLNL 27 NMARTLAQQLIDFGE-IKRGLLGIKGTEMSADIAKAFNLDVQRGAFVSEVLPGSG
HhoB (187-288) LQTDASINHGNSGGALVNS 27 QLATKIMDKLIRdGA-VIRGYIGIGGREIAPLHAQGGGIDQLQGIVVNEVSPDGP
HtrA R.hens (236-334) IQIDAAVNRGNSGGPTFDL 27 ATANEVVQQLIEKGL-VQRGWLGVQIQPVTKEISDSIGLKEAKGALITDPLK-GP
HtrA B.abor (246-345) IQIDAAVNKGNSGGPAFDL 27 STAKQVVDQLIKKGS-VERGWIGVQIQPVTKDIAASLGLAEEKGAIVASPQDDGP
HtrH B.abor (209-309) IQTDAAINPGNSGGALIDM 27 NMVRAVVDAALQGSTRFERPYIGATFQGITPDLAESLGMEKPYGALITAVVKDGP
HtrA C.trac (92-191) VTTDAAINPGNSGRSIVKI 27 LMAKRVIDQLISDGQ-VTRGFLGVTLQPIDSELATCYKLEKCTERLVTDVVKGSP
Spro M.para (204-224) IQADAPIKPGDSGGPMVNS
PV8 S.aure (226-246) MQYDLSTTGGNSGSPVFNE
EtA S.aure (222-242) LRYYGFTVPGNSGSGIFNS
EtB S.aure (206-226) SQYFGYTEVGNSGSGIFNL
Gsep B.lich (156-176) LQYAMDTYGGQSGSPVFEQ

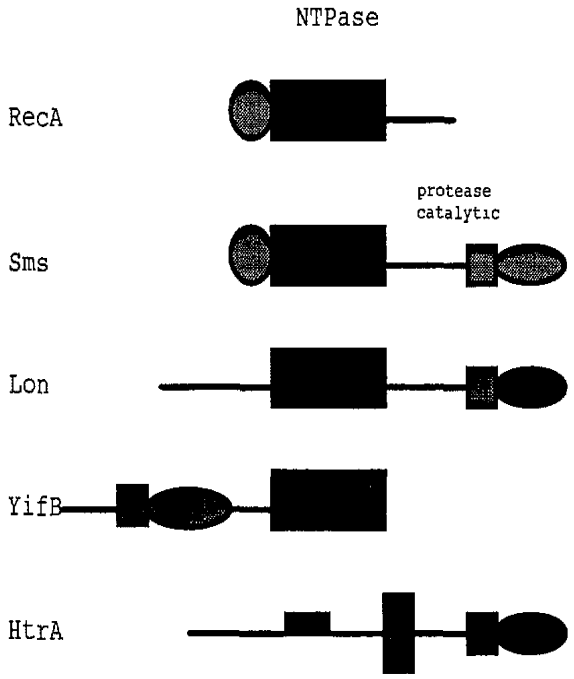


FIGURE 11 A group of multidomain proteins containing ATPase and protease domains. Two putative new ATP-dependent protesases in E. coli. (A) (Facing page) Alignment of conserved domain sequences. The C-terminal conserved motif comprising the putative protein-binding site was initially identified by using the Most program as described in the text. The multiple alignment was constructed by using the MACAW program. For each protein, the positions of the first and last amino acid residues in the sequence are indicated in parantheses. The distances between the conserved blocks are indicated by numbers. The consensus shows amino acid residues that are conserved in all of the aligned sequences (capital letters) or in most of them (lowercase letters); the designations in the consensus are as in Fig. 9. The exclamation mark shows the (predicted) catalytic serine. A group of bacterial serine proteases including the *Staphylococcus* exfoliative toxins (EtA and EtB) share similarity with HtrA in the catalytic domain but not in the putative newly identified protein-binding domain. Organism name abbreviations: B. subt, *Bacillus subtilis*; M. tube, *Mycobacterium tuberculosis*; M. xant, *Myxococcus xanthus*; R. caps., *Rhodobacter capsulata*; E. grac, *Euglena gracili*s (chloroplast); B. *brev, Bacillus brevis*; C. *eleg, Caenorhabditis elegans*; R. *hens, Rochalimea henselae*; B. *abor, Brucella abortus*; *C. trac, Chlamydia trachomatis*; M. *pseau, Mycobacterium pseudo-tubercolosis*; S. *aurem Staphylococcus aureaus*; B. *lich, Bacillus licheniformis*. The sequences of the E. coli proteins (except for HhoA and HhoB) were from the EcoSeq7 database; the other sequences were from the SWISS-PROT (Dmcl yeast, P25453; Rad51 yeast, P25451;UvsX T4, P04529; RecA B. subt, P16971; RAD57 yeast, P25301; Bchi R. caps, P26239; CCS E. grac, P31205; Mcm3 yeast, P24279; Lon human, P36776; Lonl M. xant, P36773;Lon B. brev, P36772; PV8 [V8 protease] S. aure, P04188; EtA S. aure, P09331; Lon B. brev, P36772; PV8 [V8 protease] S. aure, P04188; EtA S. aure, P09331; EtB S. aure, P09332; Gsep [glutamyl-specific endopeptidase] B. lich, P80057) or from the GenBank (Lon C. eleg, Z36719 [CEC06C3_9]; Sms B. subt, D26185 [BAC180K_149]; HhoA and HhoB, U15661; HtrA R. hens, L20127; HtrA B. abor, U07352; HtrH [htrA homolog] B. abor, U07351; Spro [serine protease] M. para, Z23092). (B) (Above) Tentative scheme of domain organization. Zn indicates the predicted N-terminal Zn finger domian in Sms (31).

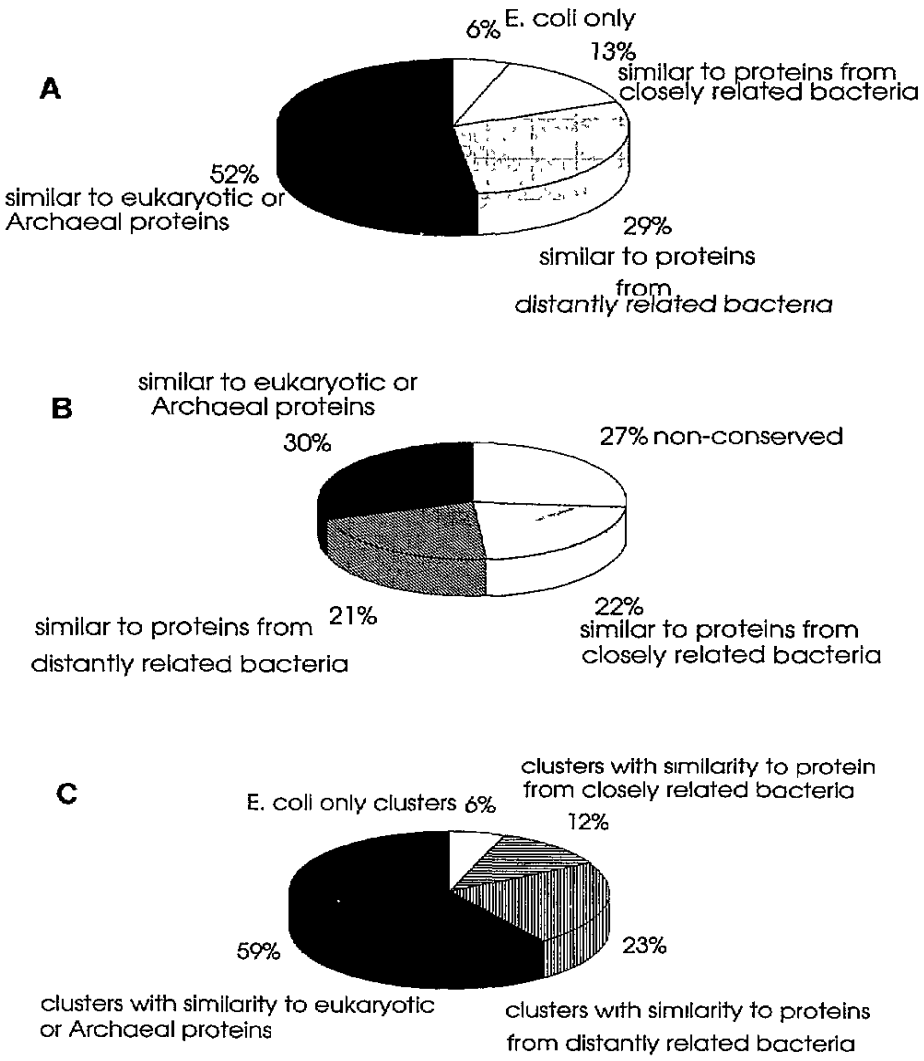


FIGURE 12 Sequence conservation and clustering of *E. coli* proteins. (A) Levels of sequence conservation in proteins belonging to clusters of paralogs. (B) Levels of sequence conservation in proteins not belonging to clusters. (C) Conserved and nonconserved clusters.

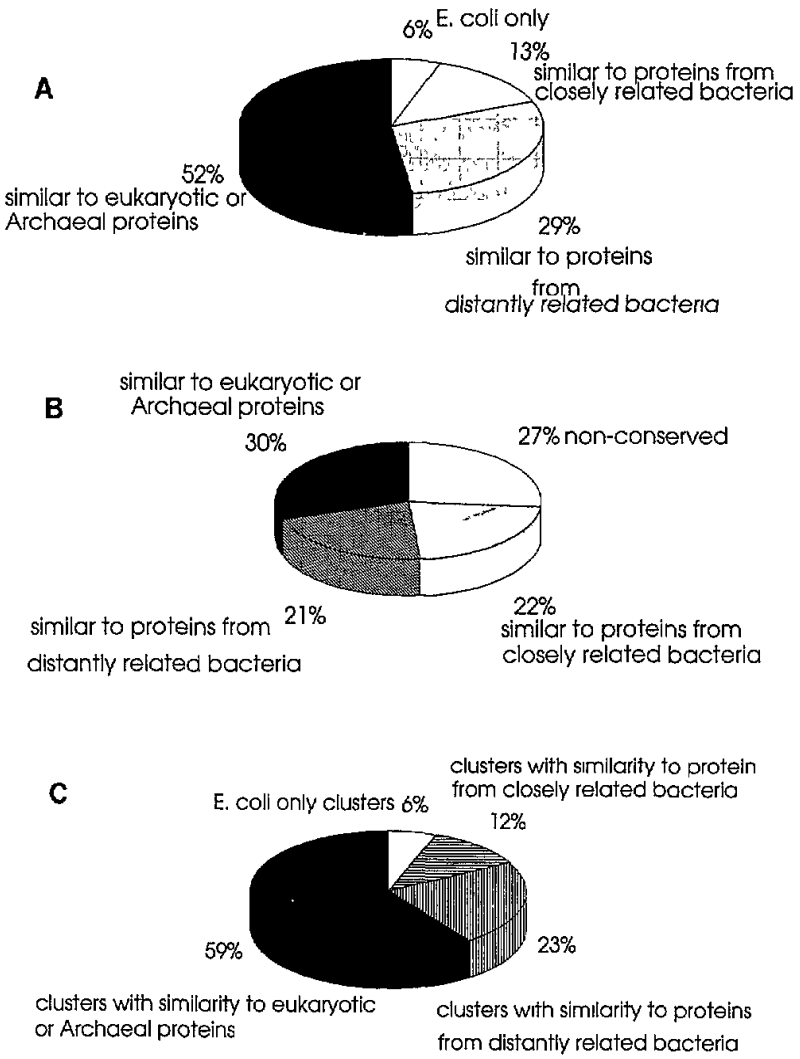


Figure 12 Sequence conservation and clustering of *E. coli* proteins. (A) Levels of sequence conservation in proteins belonging to clusters of paralogs. (B) Levels of sequence conservation in proteins not belonging to clusters. (C) Conserved and nonconserved clusters.

THE OUTSIDE AND INSIDE VIEWS OF THE *E. COLI* PROTEIN SEQUENCE SET: A JUXTAPOSITION

E. coli proteins belonging to clusters of paralogs generally show a higher evolutionary conservation than those proteins that do not have paralogs—the fraction containing ancient conserved regions is almost twice as high for the proteins in clusters (Fig. 12A and B). Predictably, the pattern of conservation is even more pronounced if analyzed for whole clusters—in most of them, at least one member is significantly similar to a eukaryotic or archaeal protein (Fig. 12C). Thus, most of the clusters correspond to critical functions that should have already been represented in the common ancestor of bacteria, eukaryotes, and the *Archaea*, even if, in some cases, by only a single member. This being so, it is particularly striking that the universal bacterial transcription regulators (HTH proteins) are not detectably similar to any proteins from eukaryotes or the *Archaea* (with the exception of the very limited similarity to homeo-domains, which in our study could be detected only when one specific family of *E. coli* HTH proteins was used to construct a motif for database search). This lack of ancient conserved regions in one of the largest *E. coli* protein superclusters may reflect the different modes of transcription regulation in bacteria, eukaryotes, and the *Archaea.* Similarly, the PTS system is composed of several clusters of highly conserved paralogous proteins (domains) without

eukaryotic homologs, suggesting that certain important pathways of metabolite transport may be restricted to bacteria (41).

CONCLUSIONS

With increasingly sensitive computer methods becoming available and sequence databases growing rapidly, even a relatively straightforward analysis of the 2,328 proteins forming about 60% of all *E. coli* gene products produced a wealth of information. For more than 90% of these proteins, either functional information or significant sequence similarity or both are available. A surprisingly high fraction—about 86%—are similar to other proteins in current databases; about two-thirds show conservation at least at the level of distantly related bacteria, and about 40% contain regions of conservation with eukaryotic or archaeal proteins. Even though gene transfer between endosymbiotic organellar genomes and the eukaryotic nuclear genome and perhaps also other forms of horizontal gene flow should be taken into account, it may be concluded that *E. coli* proteins generally have been highly conserved through very long periods of evolution. About 47% of the *E. coli* proteins belong to 286 clusters of paralogs defined on the basis of significant pairwise similarity, and an additional 10% could be included in paralogous superclusters, based on motif conservation. The majority of clusters have only two members, but there are several very large superclusters. Combined, the four largest superclusters, which include various permeases, purine NTPases with the conserved "Walker-type" motif, HTH regulatory proteins, and dinucleotide-binding proteins, account for about 25% of all known *E. coli* proteins. Genes encoding paralogous proteins appear to be nonrandomly distributed along the chromosome, with a prevalence of tandem duplications but also an apparent large-scale periodicity. Proteins belonging to paralogous clusters typically show higher conservation in evolution than do proteins that do not have paralogs within *E. coli*. Most of the clusters include proteins with ancient conserved regions and, accordingly, correspond to critical functions that should have already been encoded by the common ancestor of bacteria, eukaryotes, and the *Archaea.* Sequence similarities detected in the course of the analysis of the *E. coli* protein sequence set allow the prediction of possible functions for a number of functionally uncharacterized gene products. The complete *E. coli* chromosome sequence can be readily expected to be available within 2 years. It is our hope that the pilot project outlined in this chapter will set the stage for the complete, systematic information analysis of the whole genome.

ACKNOWLEDGMENTS

We are grateful to Amos Bairoch, Peer Bork, and John Wootton for helpful discussions and critical reading of the manuscript, and to Peer Bork and John Wootton for communicating their results prior to publication.

LITERATURE CITED

1. **Abouhamad, W. N., M. Manson, M. M. Gibson, and C. F. Higgins.** 1991. Peptide transport and chemotaxis in Escherichia coli and Salmonella typhimurium: characterization of the dipeptide permease (Dpp) and the dipeptide-binding protein. *Mol. Microbiol.* **5:**1035–1047.
2. **Altendorf, K., A. Siebers, and W. Epstein.** 1992. The KDP ATPase of *Escherichia coli. Ann. N. Y. Acad. Sci.* **671:**228–243.
3. **Altschul, S. F., M. S. Boguski, W. Gish, and J. C. Wootton.** 1994. Issues in searching molecular sequence databases. *Nat. Genet.* **6:**119–129
4. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:**403–410.
5. **Amerik, A. Y., V. K. Antonov, A. E. Gorbalenya, S. A. Kotova, T. V. Rotanova, and E. V. Shimbarevich.** 1991. Site-directed mutagenesis of La protease. A catalytically active serine residue. *FEBS Lett.* **287:**211–214.
6. **Bazan, J. F., and R. J. Fletterick.** 1990. Structural and catalytic models of trypsin-like viral proteases. *Semin. Virol.* **1:**311–322.
7. **Bork, P., C. Ouzounis, G. Casari, R. Schneider, C. Sander, M. Dolan, W. Gilbert, and P. M. Gillevet.** 1995. Exploring the *Mycoplasma capricolum* genome: a small bacterium reveals its physiology. *Mol. Microbiol.* **16:**955–967.
8. **Bork, P., C. Ouzounis, and C. Sander.** 1994. From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4:**393–403.
9. **Bork, P., C. Ouzounis, C. Sander, M. Scharf, R. Schneider, and E. Sonnhammer.** 1992. Comprehensive sequence analysis of the 182 ORFs of yeast chromosome III. *Protein Sci.* **1:**1677–1690.
10. **Borodovsky, M., E. V. Koonin, and K. E. Rudd.** 1994. New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem. Sci.* **19:**309–313.
11. **Borodovsky, M., K. E. Rudd, and E. V. Koonin.** 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.* **22:**4756–4767.
12. **Chin, D. T., S. A. Goff, T. Webster, T. Smith, and A. L. Goldberg.** 1988. Sequence of the *lon* gene in *Escherichia coli.* A heat-shock gene which encodes the ATP-dependent protease La. *J. Biol. Chem.* **263:**11718–11728.
13. **Daniels, D., G. Plunkett, V. Burland, and F. R. Blattner.** 1992. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. Science **257:**771–778.
14. **Fitch, W. M., and E. Margoliash.** 1970. The usefulness of amino acid and nucleotide sequences in evolutionary studies. *Evol. Biol.* **4:**67–77.
15. **Goffeau, A., K. Nakai, P. Slonimski, and J. L. Risler.** 1993. The membrane proteins encoded by yeast chromosome III genes. *FEBS Lett.* **325:**112–117.
16. **Gorbalenya, A. E., and E. V. Koonin.** 1990. Superfamily of UvrA- related NTP-binding proteins. Implications for rational classification of recombination/repair systems. *J. Mol. Biol.* **213:**583–591.
17. **Gottesman, S., and M. R. Maurizi.** 1992. Regulation by proteolysis: energy-dependent proteases and their targets. *Microbiol. Rev.* **56:**592–621.
18. **Gray, M. W.** 1989. The evolutionary origin of organelles. *Trends Genet.* **5:**294–299
19. **Green, P.** 1994. Ancient conserved regions in gene sequences. *Curr. Opin. Struct. Biol.* **4:**404–412.
20. **Green, P., D. J. Lipman, L. Hillier, R. Waterston, D. J. States, and J. M. Claverie.** 1993. Ancient conserved regions in new gene sequences and the protein databases. *Science* **259:**1711–1716.
21. **Higgins, C. F., I. D. Hiles, G. P. Salmond, D. R. Gill, J. A. Downie, I. J. Evans, I. B. Holland, L. Gray, S. D. Buckel, A. W. Bell, et al.** 1986. A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. *Nature* (London) **323:**448–450.
22. **Holm, L., and C. Sander.** 1994. Searching protein structure databases has come of age. *Proteins Struct. Funct. Genet.* **19:**165–173.
23. **Koonin, E. V.** 1993. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J. Mol. Biol.* **229:**1165–1174.

24. **Koonin, E. V.** 1993. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res.* **21:**2541–2547.
25. **Koonin, E. V., P. Bork, and C. Sander.** 1994. Yeast chromosome III: new gene functions. *EMBO J.* **13:**493–503.
26. **Kunisawa, T., and J. Otsuka.** 1988. Periodic distribution of homologous genes or gene segments on the *Escherichia coli* K12 genome. *Protein Sequences Data Anal.* **1:**263–267.
27. **Lipinska, B., M. Zylicz, and C. Georgopoulos.** 1990. The HtrA (DegP) protein, essential for *Escherichia coli* survival at high temperatures, is an endopeptidase. *J. Bacteriol.* **172:**1791–1797.
28. **Medigue, C., A. Viari, A. Henaut, and A. Danchin.** 1993. Colibri: a functional data base for the *Escherichia coli* genome. *Microbiol. Rev.* **57:**623–654.
29. **Moszer, I., P. Glaser, and A. Danchin.** 1991. Multiple IS insertion sequences near the replication terminus in *Escherichia coli* K-12. *Biochimie* **73:**1361–1374.
30. **Navarro, C., L. F. Wu, and M. A. Mandrand-Berthelot.** 1993. The nik operon of Escherichia coli encodes a periplasmic binding-protein-dependent transport system for nickel. *Mol. Microbiol.* **9:**1181–1191.
31. **Neuwald, A. F., D. E. Berg, and G. V. Stauffer.** 1992. Mutational analysis of the Escherichia coli serB promoter region reveals transcriptional linkage to a downstream gene. *Gene* **120:**1–9.
32. **Nichols, J. C., N. K. Vyas, F. A. Quiocho, and K. S. Matthews.** 1993. Model of lactose repressor core based on alignment with sugar-binding proteins is concordant with genetic and chemical data. *J. Biol. Chem.* **268:**17602–17612.
33. **Niki, H., A. Jaffe, R. Imamura, T. Ogura, and S. Hiraga.** 1991. The new gene *mukB* codes for a 177 kd protein with coiled-coil domains involved in chromosome partitioning of E. coli. *EMBO J.* **10:**183–193.
34. **Oliver, G., G. Gosset, R. Sanchez-Pescador, E. Lozoya, L. M. Ku, N. Flores, B. Becerill, F. Valle, and F. Bolivar.** 1987. Determination of the nucleotide sequence for the glutamate synthase structural genes of *Escherichia coli* K-12. *Gene* **60:**1–11.
35. **Olsen, G. J., C. R. Woese, and R. Overbeek.** 1994. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176:**1–6.
36. **Palmer, J. D.** 1985. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* **19:**325–354.
37. **Pao, G. M., R. Tam, L. S. Lipschitz, and M. H. Saier, Jr.** 1994. Response regulators: structure, function and evolution. *Res. Microbiol.* **145:**356–362.
38. **Parkinson, J. S., and E. C. Kofoid.** 1992. Communication modules in bacterial signaling proteins. *Annu. Rev. Genet.* **26:**71–112.
39. **Quirk, S., and M. J. Bessman.** 1991. dGTP triphosphohydrolase, a unique enzyme confined to the members of the family *Enterobacteriaceae. J. Bacteriol.* **173:**6665–6669.
40. **Rahfeld, J. U., K. P. Rucknagel, B. Schelbert, B. Ludwig, J. Hacker, K. Mann, and G. Fischer.** 1994. Confirmation of the existence of a third family among peptidyl-prolyl cis/trans isomerases. Amino acid sequence and recombinant production of parvulin. *FEBS Lett.* **352:**180–184.
41. **Reizer, A., G. M. Pao, and M. H. Saier, Jr.** 1991. Evolutionary relationships among the permease proteins of the bacterial phosphoenolpyruvate:sugar phosphotransferase system. Construction of phylogenetic trees and possible relatedness to proteins of eukaryotic mitochondria. *J. Mol. Evol.* **33:**179–193.
41a. **Reuven, N. B., E. V. Koonin, K. E. Rudd, and M. P. Deutscher.** 1995. The gene for the longest known *Escherichia coli* protein is a member of helicase superfamily II. *J. Bacteriol.* **177:**5393–5400.
42. **Riley, M.** 1993. Functions of the gene products of *Escherichia coli. Microbiol. Rev.* **57:**862–952.
43. **Riley, M., and A. Anilionis.** 1978. Evolution of the bacterial genome. *Annu. Rev. Microbiol.* **32:**519–560.
44. **Riley, M., and S. Krawiec.** 1987. Genome organization, p. 967–981. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology.* American Society for Microbiology, Washington, D.C.

45. **Robison, K., W. Gilbert, and G. M. Church.** 1994. Large scale bacterial gene discovery by similarity search. *Nat. Genet.* **7:**205–214.
46. **Rudd, K. E.** 1993. Maps, genes, sequences, and computers: an *Escherichia coli* case study. *ASM News* **59:**335–341.
47. **Rudd, K. E., H. E. Sofia, E. V. Koonin, S. Lazar, G. Plunkett III, and P. E. Rouviere.** 1995. A new family of peptidyl-prolyl isomerases. *Trends Biochem. Sci.* 20:12–14.
48. **Saier, M. H., Jr.** 1994. Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol. Rev.* **58:**71–93.
49. **Saraste, M., P. R. Sibbald, and A. Wittinghofer.** 1990. The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15:**430–434.
50. **Schaffner, A. R., and J. Sheen.** 1992. Maize C4 photosynthesis involves differential regulation of phosphoenolpyruvate carboxylase genes. *Plant J.* **2:**221–232.
51. **Schuler, G. D., S. F. Altschul, and D. J. Lipman.** 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct. Funct. Genet.* **9:**180–190.
52. **Smith, M. W., D.-F. Feng, and R. F. Doolittle.** 1992. Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem. Sci.* **17:**489–493.
53. **Tam, R., and M. H. Saier, Jr.** 1993. Structural, functional and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol. Rev.* **57:**320–346.
54. **Tatusov, R. L., S. F. Altschul, and E. V. Koonin.** 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA* **91:**12091–12095.
55. **Titgemeyer, F., J. Reizer, A. Reizer, and M. H. Saier, Jr.** 1994. Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. *Microbiology* **140:**2349–2354.
56. **Wahl, R., P. Rice, C. M. Rice, and M. Kröger.** 1994. ECD—a totally integrated database of *Escherichia coli* K12. *Nucleic Acids Res.* **22:**3450–3455.
57. **Walker, J. E., M. Saraste, M. J. Runswick, and N. J. Gay.** 1982. Distantly related sequences in the a- and b-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1:**945–951.
58. **Woese, C. R., and G. E. Fox.** 1977. Progenotes and the origin of the cytoplasm. *J. Mol. Evol.* **10:**1–6.
59. **Wootton, J. C.** 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem* **18:**269– 285.
60. **Wootton, J. C.** 1994. Sequences with ‘unusual’ amino acid composition. *Curr. Opin. Struct. Biol.* **4:**413–421.