# Physiological genomics of Escherichia coli protein families

**Ping Liang, Bernard Labedan and Monica Riley**

*Physiol. Genomics* 9:15-26, 2002. First published Mar 5, 2002; doi:10.1152/physiolgenomics.00086.2001

*Physiol Genomics* 9: 15–26, 2002.
First published March 5, 2002; 10.1152/physiolgenomics.00086.2001.

# Physiological genomics of *Escherichia coli* protein families

PING LIANG,[1] BERNARD LABEDAN,[2] AND MONICA RILEY[1]
[1]*Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts 02543; and*
[2]*Institut de Génétique et Microbiologie, Centre National de la Recherche Scientifique UMR 8621, Bâtiment 409, Université de Paris-Sud, 91405 Orsay Cedex, France*

Received 26 September 2001; accepted in final form 9 November 2001

**Liang, Ping, Bernard Labedan, and Monica Riley.** Physiological genomics of *Escherichia coli* protein families. *Physiol Genomics* 9: 15–26, 2002. First published March 5, 2002; 10.1152/physiolgenomics.00086.2001.—The well-researched *Escherichia coli* genome offers the opportunity to explore the value of using protein families within a single organism to enrich functional annotation procedures and to study mechanisms of protein evolution. Having identified multimodular proteins resulting from gene fusion, and treated each module as a separate protein, nonoverlapping sequence-similar families in *E. coli* could be assembled. Of 3,902 proteins of length 100 residues or more, 2,415 clustered into 609 protein families. The relatedness of function among members of each family was dissected in detail. Data on paralogous protein families provides valuable information in attributing putative function to unknown genes, supplementing existing function annotation. Enzymes, transporters, and regulators represent the three major types of proteins in *E. coli*. They are shown to have distinctive patterns in gene duplication and divergence and gene fusion, suggesting that details of protein evolution have been different for genes in these categories. Data for the complete list of paralogous protein families and updated functional annotation for *E. coli* K-12 are accessible in GenProtEC (**http://genprotec.mbl.edu**).

module; sequence similarity; protein family; predicting protein function; annotation; evolution

IN THIS ERA OF GENOMICS, we will be approaching an understanding of the functions of all the genes in a single cell. In the case of single-cell organisms, the full complement of gene products is the substance of the cell that enables it to be a living, self-regulating, self-reproducing entity with flexibility of response to regulatory signals. The essence of major life processes is encapsulated in single-cell organisms.

The most thoroughly studied single-cell organisms are the bacterium *Escherichia coli* and the single-cell eukaryote *Saccharomyces cerevisiae*. Both organisms use mainstream metabolic pathways that are recognizably similar to the corresponding metabolic functions in all life forms including higher eukaryotes. The entire genome sequence has been determined for both organisms. The relationships between genetics and biochemistry that constitute the fundamental processes of life in these single-cell organisms serve in a sense as a foundation for ongoing investigations on the more elaborate processes that operate in the more complex, higher forms of life.

With over 60 years of intensive study yielding a voluminous scientific literature, a great deal of the physiology and molecular biology of *E. coli* is experimentally known. We have updated recently the list of all the gene products and their immediate molecular functions (25). For *E. coli*, a high proportion, about half, of the genetically determined cell content is currently known directly by experiment. The genes are known in sequence and genetic location; the gene products (protein or RNA) have been characterized experimentally. A small fraction, 2.1%, are understood only in terms of their mutant phenotype. Function could be tentatively attributed to 29.5% of the total that are similar in sequence to genes of known functions in other organisms, but have not yet been experimentally checked in *E. coli*. Finally, 19.5% are similar to genes of unknown function in other organisms. Only 7% are currently specific to *E. coli*, and that number will decrease when comparisons with *Salmonella* species are complete. Thus we either know or have a good idea of what 81% of this organism's genes encode. To complete our understanding of the cell and its activities, we need to know not only what the gene products are and what they do, but we will also need to know how the gene products interact with one another and how their activities are regulated.

In this study, the organization of protein families of *E. coli* is addressed. Some of the proteins are multimodular, as if they arose by fusion of two or more independent genes. To identify families of proteins related by sequence, it was necessary to identify multifunctional proteins formed of two or more proteins of separate function and unrelated sequence (13, 31). These components of multifunctional proteins are what we term "modules." Different from a motif, a module represents an independent individual protein that has

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence and present address of P. Liang: Department of Cancer Genetics, Roswell Park Cancer Institute, Elm & Carlton Streets, Buffalo, NY 14263 (E-mail: Ping.Liang@RoswellPark.org).

descended in the course of evolution in many cases as a single unit in some current genomes, but is found fused to another module in some other genomes. Unless multimodular proteins are identified and the components treated as separate entities before collecting groups of proteins of similar sequence, false connections can be made, as diagrammed in Fig. 1. Figure 1 portrays a case in which at least nine groups of proteins with different functions can be incorrectly grouped together through a multimodular protein, RNE (Swiss-Prot accession no. P21513).

After separation into component parts, many *E. coli* genes can be grouped into families by their amino acid sequences. Most members in most families have related functions. Earlier studies that we completed before the entire genomic sequence was known showed that a substantial fraction of *E. coli* proteins could be clustered by sequence similarity into groups ranging in size from 2 to over 70 (14, 22). In this report we have used the completed genomic sequence (3) to assemble all the sequence-related groups of all *E. coli* genes and relate the sequence-similar families to their functions.

## METHODS

*Sequence sources.* To define the sequence-related protein families, we used the 3,902 peptide sequences of length 100 residues or greater from the Blattner genes/ORFs based on current entries in GenBank accession no. U00096 (3). We did not incorporate the conceptual frame shifts and other revisions that have been introduced in the EcoGene database (24) (**http://bmb.med.miami.edu/EcoGene/**) and Swiss-Prot (8) (**http://www.expasy.ch/**). In any case, these revisions would make little difference in the results of this type of study. Our characterization of functions of gene products is maintained in our database GenProtEC (**http://genprotec.mbl.edu**), recently updated in relation to contemporary primary literature (2, 23–25).

*Pairwise similarities between all E. coli proteins.* Exhaustive pairwise sequence comparison for each *E. coli* module against all other *E. coli* modules in the genome was performed using a locally installed Data Analysis and Retrieval With Indexed Nucleotide/Peptide Sequences package (DARWIN, version 2.0) obtained from the Computational Biochemistry Research Group at ETH, Zurich, Switzerland (10). The SearchPepAll and LocalAlignBestPAM functions were used to generate a list of all qualifying aligned pairs of peptides. These employ the dynamic programming of both Needleman-Wunsch (18) and Smith-Waterman (27) algorithms and test appropriate PAM ("accepted point mutations") score matrixes for each sequence pair. The outputs contain information related to the alignment: identifiers of the sequence pairs, the start and end positions of alignment regions for both sequences, the PAM score, variance score for a panel of substitution matrices, and the percentages of sequence identity and similarity. (PAM score of 0 signifies complete identity; higher values represent progressively less of a sequence match.) To deal with entire proteins rather than motifs and binding sites within proteins, we stipulated that all alignments have a minimal length of 100 residues. To collect evolutionarily distant relationships yet avoid artifact, we set a limit for PAM at less than 200 to reduce false matches to a minimal level. Compared with the statistical cutoffs established by Altschul (1) as a significant sequence match (a minimum length of 83 residues and a maximal PAM distance of 250), our criteria are conservative. The DARWIN algorithms and use of multiple substitution matrices have been evaluated in relation to other sequence analysis approaches and have been given high credit for sensitivity and performance (21, 28).

*Identification of modules.* Some proteins derive from compound genes such that direct translation of the genetic open reading frame (ORF) produces more than one protein functional unit. Other multifunctional proteins remain polycistronic polypeptides, expressing more than one function as a complex multisite protein. We identified such multimodular proteins in *E. coli* by the positions of regions of sequence similarity. By noting separate regions of alignment between pairs of proteins, modular composition of genes with partners in the *E. coli* genome were identified. Table 1 lists examples of multimodular proteins and functions of the modules, giving the start residue and end residue of regions of pairwise alignment with other proteins. We set arbitrary threshold values for inferring more than one module in a protein on the basis of properties of some known multimodular proteins. We defined modules to occupy more than 25% and less than 80% of an entire protein. Adjacent modules in the same protein were not allowed to overlap more than 10 residues. A heuristic method was developed to make these identifications (Le

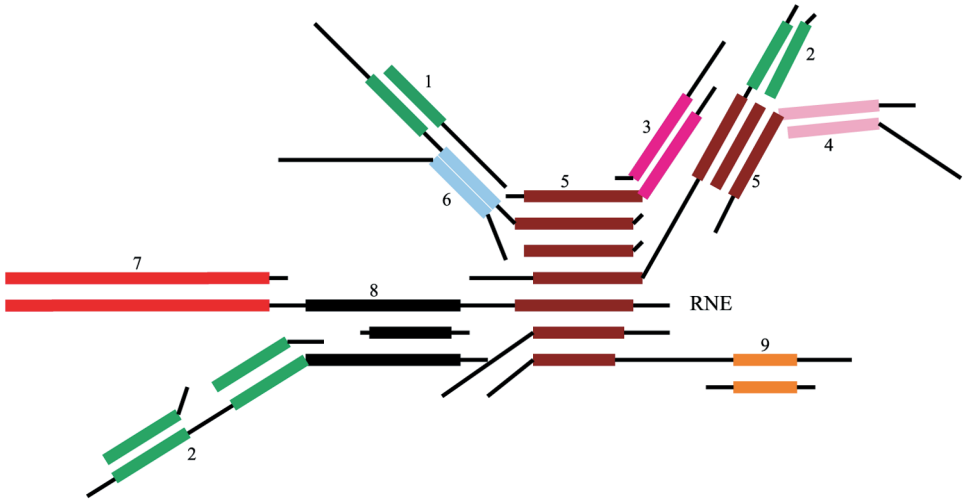


Fig. 1. Representation of network of proteins by sequence relationship showing that proteins from at least 9 different paralogous protein families can be linked together through a multimodular protein, RNE, an RNase E. All proteins are orientated from *left* ($NH_2$ terminal) to *right* (COOH terminal). Aligned regions are in thicker solid color bars, whereas unaligned regions are represented by a thin black line. Numbers label unique sequence-related alignments. Only selected matches are used to show minimal connections for any module. *1*, membrane protein subunits of NAD dehydrogenases; *2*, ATP-binding components of ABC transporters; *3*, GTPase domains; *4*, GTP-binding proteins; *5*, protein interacting sites; *6*, Fe-S subunit of oxidoreductases; *7*, part of RNase G/E; *8*, polynucleotide binding; and *9*, phage tail fiber proteins.

Table 1. *A sample list of multimodular proteins*

| SW_ID | Module | Start | End | Function |
|---|---|---|---|---|
| PTAA | b0679_1 | 1 | 311 | sugar-specific PTS: enzyme IIC, membrane component and sugar binding |
| | b0679_2 | 341 | 466 | sugar-specific PTS: enzyme IIB, phosphorylation of sugar |
| | b0679_3 | 475 | 648 | sugar-specific PTS: enzyme IIA, phosphotransferase |
| HRSA | b0731_1 | 16 | 178 | sugar-specific PTS: enzyme IIA, phosphotransferase |
| | b0731_2 | 384 | 492 | sugar-specific PTS: enzyme IIB, phosphorylation of sugar |
| | b0731_3 | 468 | 628 | sugar-specific PTS: enzyme IIC, membrane component and sugar binding |
| PTNA | b1817_1 | 2 | 128 | sugar-specific PTS: enzyme IIA, phosphotransferase |
| | b1817_2 | 163 | 317 | sugar-specific PTS: enzyme IIB, phosphorylation of sugar |
| PTGB | b1101_1 | 12 | 185 | sugar-specific PTS: enzyme IIC, membrane component and sugar binding |
| | b1101_3 | 360 | 477 | sugar-specific PTS: enzyme IIB, phosphorylation of sugar |
| RNE | b1084_1 | 4 | 427 | enolase activity |
| | b1084_2 | 501 | 789 | RNA binding |
| | b1084_3 | 885 | 1,057 | interacts with proteins in degradosome (pnpase) |
| ADHE | b1241_1 | 8 | 373 | aldehyde reductase (dehydrogenase) |
| | b1241_2 | 456 | 857 | alcohol dehydrogenase |
| | b1241 | ND | ND | deactivase of pyruvate formate-lyase* |
| FADB | b3846_1 | 8 | 206 | isomerase, hydratase, and epimerase |
| | b3846_2 | 314 | 643 | NADH-dependent dehydrogenase |
| DOGA | b3692_1 | 8 | 121 | 2-oxo-3-deoxygalactonate 6-phosphate aldolase |
| | b3692_2 | 206 | 587 | galactonate dehydratase |
| CYDC | b0886_1 | 119 | 266 | transporter, ABC superfamily (membrane component) |
| | b0886_2 | 298 | 573 | transporter, ABC superfamily (ATP-binding component) |
| ARAG | b1900_1 | 33 | 219 | transporter of arabinose, ABC superfamily (ATP-binding component) |
| | b1900_2 | 282 | 475 | transporter of arabinose, ABC superfamily (ATP-binding component) |
| DGAL | b2150_1 | 2 | 154 | transporter of galactose, ABC superfamily: periplasmic binding component |
| | b2150_2 | 157 | 264 | transporter of galactose, ABC superfamily: periplasmic binding component |

SW_ID, SwissProt names; "Start" and "End" indicate the average start and stop positions of alignments with other proteins. ND, not determined. ABC, ATP-binding cassette; PTS, phosphotransferase. *Ref. 12.

Bouder and Labedan, unpublished observations). After automatic processing, substantial hand adjustment and polishing was required to eliminate false positives and to complete missed connections.

The module names are the Blattner numbers ("b numbers," provided in GenBank, Swiss-Prot, EcoGene, and GenProtEC), plus a numerical suffix. The suffix ranges from 1 to 4, to indicate its relative position within the protein, with "1" being closest to the $NH_2$ terminus and the largest number closest to the COOH terminus. A module name without a suffix indicates the module is at least 80% of the whole protein and the unaligned regions at either end are no longer than 100 residues.

*Assembly of module-based sequence-similar groups.* The modules, instead of the whole proteins, were assembled into sequence-similar groups using a single-linkage chain clustering method to place all paired modules into groups whose memberships did not overlap. These processes were performed using the programs Module and Families, recently developed in the Labedan group (Le Bouder and Labedan, unpublished observations).

*Genealogical trees of protein groups.* To analyze the genealogical relationships among members of the larger paralogous groups, protein sequences corresponding to the involved modules were extracted and a distance matrix in PAM values for the whole group was generated using DARWIN. Trees were inferred using Fitch-Margoliash and least-squares distance methods (6) in PHYLIP (5) based on the above distance matrix.

*Statistical analysis.* The distribution patterns for PAM values are expressed in histograms. The differences of distribution patterns between groups were analyzed with the Wilcoxon rank sum test using StatView from SAS Institute (Cary, NC).

## RESULTS

*Identification of modules and assembly of sequence-related groups of proteins of E. coli.* Extending earlier studies on sequence-similar groups of proteins in *E. coli*, we have reanalyzed the whole proteome of *E. coli* using the entire sequence of the chromosome (3). The existence of multimodular proteins, composed of two or more proteins whose genes are joined together, imposes a technical challenge for analysis of sequence-related protein families. To assemble groups of sequence-related proteins, it is necessary to recognize these fusions and to deal with sequence relationships of the components independently. In earlier work this delineation was done entirely by hand, but in this analysis this problem was dealt with by using a suite of newly developed computer programs (Le Bouder and Labedan, unpublished observations) supplemented with adjustments by hand.

A simple example is the case of the ADHE protein in *E. coli*. The fact that this protein is an alcohol dehydrogenase and has aldehyde reductase activity as well could lead one to assume that the two related activities reflect merely reversibility of the reaction at a single catalytic site. But an examination of the alignment regions of ADHE produced by DARWIN, or by BLAST2, immediately tells us that these two catalytic activities are located in two sequence-unrelated regions of the protein, with the aldehyde reductase activity located in the $NH_2$-terminal region and the alcohol dehydrogenase activity in the COOH-terminal region. Clearly, ADHE is composed of two modules.

Of the 2,415 proteins with at least one sequence-related partner in *E. coli*, 287 proteins (11.8% of all paralogs) were identified as multimodular proteins, most of which contained two modules, some more. Of the 287 multimodular proteins, 229 have modules in more than one paralogous group and 58 have 2 or more modules in the same group. The former, larger class represents the occurrence of gene fusion, and the latter, smaller class represents internal duplication in the past. Both types were present in 40 of the proteins that contained 3 or more modules. Table 1 presents a sample list of multimodular proteins and the functions of the parts of the proteins, known or predicted. For example, protein ARAG has two modules belonging in the same group, both ATP binding cassettes of a transport protein, whereas protein PTAA, another type of transporter, contains three modules, three components of phosphotransferase enzyme II (A, B, and C). Combinations of the above two situations also exist. For instance, YHIH is composed of two ATP-binding components and a membrane component of an ATP-binding cassette (ABC) transporter (data not shown).

The total of 2,745 modules identified as having at least one homologous partner within *E. coli* were collected into groups having similar sequence by a clustering method by following paths of likeness among the pairs as described in METHODS. The groups were constructed transitively so that not all members of a group have detectable relatedness to all other members of the group, but no member is related in sequence to a member of any other group. There were 609 sequence-related nonoverlapping groups within the *E. coli* genome. The remaining 1,871 proteins, which we refer to as singles, do not match any of the other sequences in the same genome based on our criteria.

Homologous proteins encoded by genes in a single genome are defined as paralogs (7) and are believed to have arisen in the course of evolution by gene duplication followed by divergence in both sequence and function. Homologous sequences in different organisms are defined as orthologs.

Detailed information for the full list of the multimodular proteins and paralogous groups of *E. coli* with the position of each module within the proteins has been deposited into the GenProtEC web site (**http://genprotec.mbl.edu**).

*Anatomy of the sequence-related groups.* Not all the groups are the same size or configuration. Groups differ in number of members ranging from 2 to 94. The number of groups of a given size is inversely related to how many members are in the group. There are many more instances of groups of two members (pairs) than of any larger size. There was only one instance of each of the groups of size larger than 28 (Table 2).

The transitively assembled groups collect distantly related proteins as well as the closely related. Some groups are very tight and mutually connected among all members, such as a group of aminotransferases (Fig. 2*A*), while others have a few well-connected members plus more distantly related sequences with fewer connections in the group. This is illustrated by a group containing epimerase and dehydratase enzymes of the SDR ("short chain dehydrogenase-reductase") family (11) (Fig. 2*B*).

Table 2. *The frequency of paralogous protein group sizes in E. coli*

| | Number of Groups | | | |
|---|---|---|---|---|
| Group Size | All | E | T | R |
| 2 | 318 | 150 | 27 | 18 |
| 3 | 121 | 65 | 14 | 8 |
| 4 | 55 | 30 | 9 | 1 |
| 5 | 28 | 11 | 6 | 1 |
| 6 | 17 | 7 | 3 | |
| 7 | 17 | 7 | 3 | |
| 8 | 6 | 2 | | 2 |
| 9 | 6 | 4 | 1 | |
| 10 | 4 | 2 | 1 | |
| 11 | 8 | 4 | 2 | |
| 12 | 6 | | 4 | 1 |
| 13 | 3 | 2 | | |
| 14 | 2 | 2 | | |
| 15 | 3 | 1 | | 1 |
| 17 | 2 | 1 | | |
| 18 | 2 | 2 | | |
| 24 | 1 | | | 1 |
| 27 | 2 | | | 2 |
| 29 | 1 | 1 | | |
| 30 | 1 | | | |
| 31 | 1 | | 1 | |
| 34 | 1 | | | |
| 36 | 1 | | | 1 |
| 39 | 1 | | 1 | |
| 43 | 1 | | | 1 |
| 45 | 1 | | | 1 |
| 68 | 1 | | 1 | |
| 83 | 1 | | 1 | |
| 94 | 1 | | 1 | |

All, all *E. coli* proteins; E, enzymes; T, transporters; R, regulators.

*Functions within sequence-related protein groups.* Each protein group within a single genome is believed to have descended from a common ancestor by gene duplication. The duplication may have taken place in any one of the many ancestral progenitors in its lineage or, much more recently, in the *E. coli* genome itself. The fact that functions of group members are highly related supports identification of the groups as evolutionarily significant, not only in uniform families like the transaminases (Fig. 2*A*), but even in groups in which not all pairs of members are related by our criteria for sequence similarity. For example, the group shown in Fig. 2*B* with a limited number of detectable associations could be suspected of having some false joining artifact, especially since two kinds of enzymes are in the group, NAD-requiring dehydratases and epimerases. Yet coexistence of the two kinds of enzymes is biochemically reasonable. They are a part of a larger group of the well-recognized biochemically related SDR enzymes that use the same reaction chemistry and to drive NAD-requiring dehydrogenases and isomerases as well as dehydratases and epimerases (11). Therefore, even such a loosely related group is credible from biochemical and evolutionary perspectives.

Many pairs of paralogs are isozymes that catalyze identical or very closely related reactions. For instance, there are two paralogous alanine racemases in *E. coli* K-12, one catalytic the other biosynthetic. Some polymeric enzymes are also closely related in this way. The α-subunits of the isozymes ribonucleoside diphosphate reductase 1 and 2 (protein names PIR1 and PIR2, gene names nrdA and nrdE) are in one group, whereas their β-subunits, PIR3 and PIR4 (gene names nrdB and nrdF) make up another group. They differ only in the redox cofactor preferred. Among the 160 enzymes in *E. coli* that have at least one isozyme partner listed by Riley and Serres (23), 132 (82.5%) are found in pairs or triplets related by sequence with scores better than our threshold values.

In most of these large protein groups, there is a relationship between function and deeply branching families. Figure 3 is a distance matrix tree of a family of the 79 proteins (94 modules) of a group of ATP-binding subunits of the multimeric ABC transporters. Substrate specificities are experimentally known for 52 of the 94 group members, while 42 are annotated as putative ATP-binding components of ABC transporter with no predicted substrate information. In the tree for the protein sequences (Fig. 3), clustering is observed for different types of substrates transported: amino acids, oligopeptides, and 5- and 6-carbon sugars. The clustering represents a likely evolutionary scenario by which differentiation of replicate ATP-binding components generated families of proteins similar for a given type of substrate. Such consistency provides opportunity to supplement annotation of genes of an organism with information from internal paralogs in addition to the more customary orthologous sequence matches. Internal groupings (paralogous relationships) illuminate evolutionary relationships to bear where subsets of sequence-similar groups may contain information on type of substrate and type of reaction. In this particular case, we can predict the types of substrates used by the 11 proteins with unknown substrates in the 3 branches that show good conservation for the transported substrates. The topology of the tree agrees in principle with the one constructed with fewer members by Linton and Higgins (15).

*Enzyme, transporter, and regulator: the three major and distinctive types of proteins in E. coli.* We have classified all genes of *E. coli* K-12 by the functions of their products, both the experimentally known and predicted functions as summarized in Table 4. Genetic resources in *E. coli* are distributed into three large functional classes and several smaller ones. The largest number of genes code for enzymes (34%), followed by the genes for transport functions (13.8%), and the genes for regulatory processes (11.5%). The number of genes in the three categories accounts for more than 59% of all genes and more than 79% of genes with known or predicted functions in *E. coli* K-12. Moreover, almost all of the largest sequence-similar groups, the paralogous groups, fall into these three categories (Table 3).

The groups of proteins in each of the three categories show distinctive characteristics with respect to numbers of sequence-similar groups, sizes of groups, multimodularity, and also the ratio of unique proteins, called singles, to those with partners, the paralogs (Table 5). Enzymes and polymeric enzyme subunits tend to be clustered into smaller sequence groups as indicated by the smallest average group size and the highest percentage (32%) as singles (Table 5). Transporters and regulators occur in many fewer but much bigger groups and have a lower percentage as singles (17% and 20%, respectively) than do the enzymes. There seems to be more variety in the classes of enzymes, whereas there are not as many kinds of transporters and regulators.

The three largest paralogous groups are all transporters (Table 3), and they account for over 60% of all known and predicted transporter proteins in *E. coli*. Transporters have the highest percentage (18%) of multimodular proteins (both internal duplication and gene fusion), over double the number for enzymes (8%) and regulators (8.6%), suggesting transporters have been more often involved in gene fusion and duplication processes (Table 5).

In addition, as illustrated by Fig. 4, the levels of sequence similarity in pairwise matches within families of enzymes, transporters, and regulators have different distributions of PAM distances. The pairwise sequence similarity between related enzymes ranges

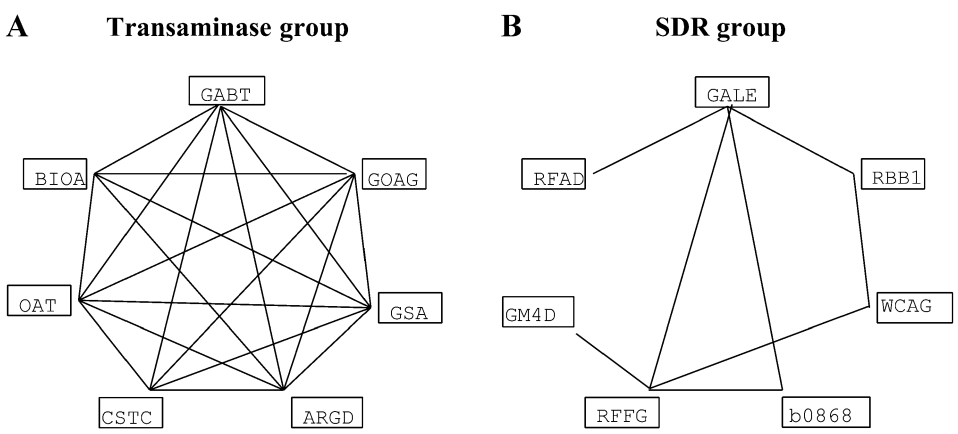


Fig. 2. The different anatomies of paralogous protein families. *A*: a group of aminotransferases. *B*: a group of short chain dehydrogenase-reductase (SDR) epimerases and dehydratases (see text). Proteins were labeled by the first section of Swiss-Prot names wherever available or by the Blattner "b number" identifier otherwise. A line between two proteins indicates that sequence similarity above the threshold (PAM ≤ 200, alignment length ≥ 100) was detected between the two proteins. PAM, accepted point mutations score.

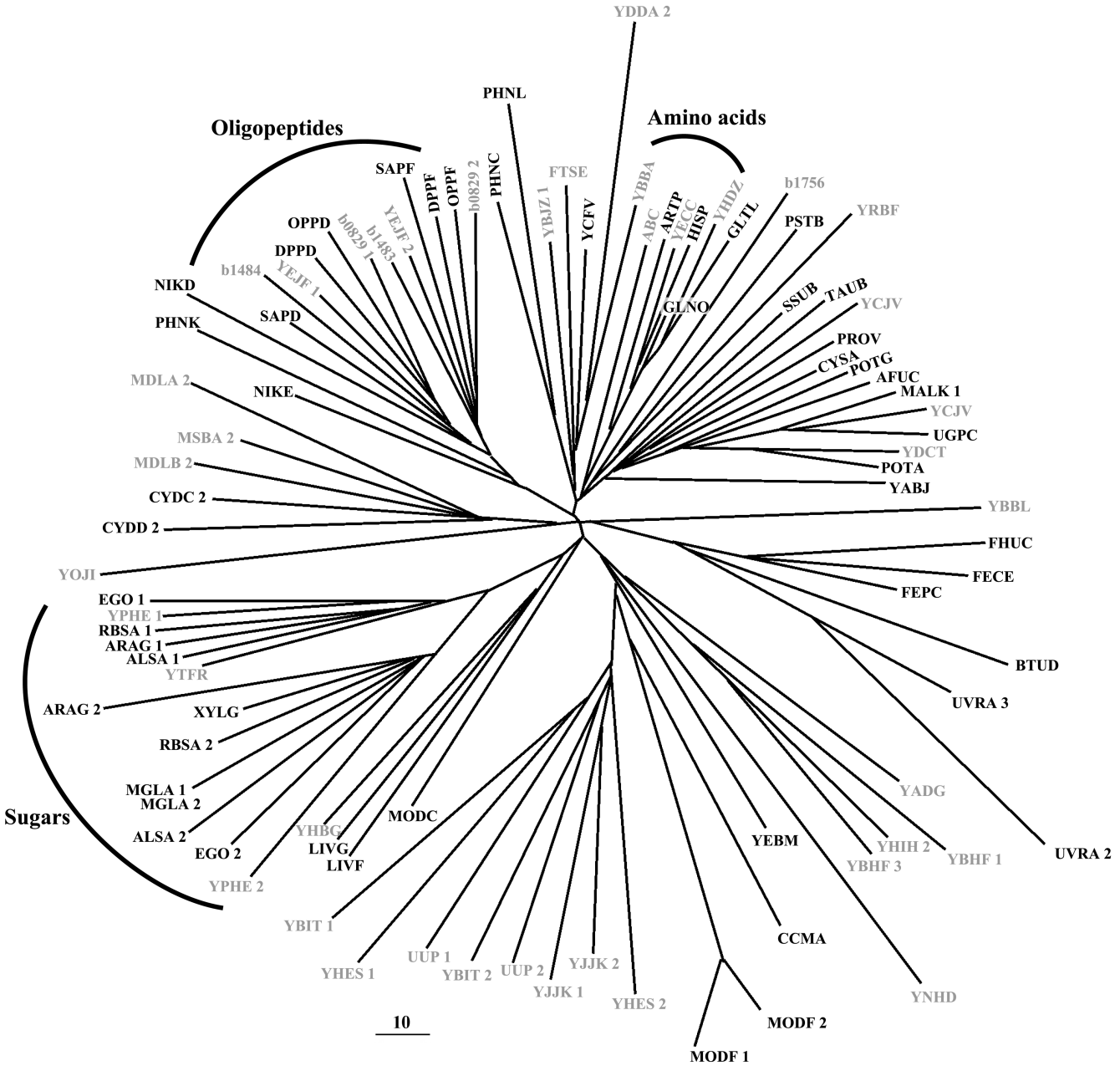


Fig. 3. An unrooted genealogy tree for a group of ATP-binding components of ATP-binding cassette (ABC) transporters. The tree was inferred as described in METHODS. Branch length is proportional to the PAM distance represented (scale bar shown). Proteins with names in black are experimentally known regarding substrates, whereas proteins with names in gray are assigned only as putative ATP components of ABC transport proteins without predicted substrate. Proteins are labeled by the first section of Swiss-Prot names wherever available or the Blattner "b number" identifier otherwise. Numerical suffixes are indications of modules. The three branches bracketed by curves have conservative types of substrates as indicated.

widely and has a median PAM value of 148 (Fig. 4*A*). In contrast, regulators are not as spread out at either the low or high ends as the enzymes are and, instead, have more instances in the midrange (Fig. 4*C*). The narrower distribution has a lower median PAM (137) than the enzymes. Transporters range as widely as enzymes, but their PAM values are more clustered at higher values and have the highest median PAM value (165) among the three types. A larger fraction of the sequence-related transporters have a PAM value over 170 compared with either the enzymes or the regulators, indicating they require the least conservation for their function as transporters compared with the other two functional categories (Fig. 4*B*). Figure 4*D* shows the differences in distribution profiles in terms of percentile as a function of PAM value.

*Distribution of protein lengths for modules and singles.* To examine the patterns for evolution by gene fusion and duplication within the *E. coli* genome, we compared the distribution of protein lengths for all *E.*

Table 3. *A list of large groups and their functions*

| Size | Function of Group |
|---|---|
| 94 | Transporters (membrane component, mainly MFS superfamily) |
| 83 | Transporters of ABC superfamily (ATP-binding component) |
| 68 | Transporters of ABC superfamily and some PTS (membrane component) |
| 45 | LysR-type transcription regulators |
| 43 | Mainly LuxR/UhpA-type 2-component regulators |
| 39 | Mainly membrane component of transporters, regulators |
| 36 | Two-component regulators (kinase component) |
| 34 | Mainly conserved proteins of unknown function |
| 31 | Transporters (mainly APC superfamily for amino acids) |
| 30 | Fimbrial proteins |
| 29 | Fe-S subunits of oxidoreductases |
| 27 | Mainly GntR-type transcriptional regulators |
| 27 | AraC/XylS-type of transcriptional regulators |
| 24 | GalR/LacI-type transcriptional regulators, and a few transporter binding proteins |
| 18 | NAD(P)-dependent oxidoreductases |
| 18 | NAD(P)-dependent alcohol dehydrogenases |
| 17 | ATP-dependent helicases |
| 17 | GTP-binding elongation factors |
| 15 | Mainly conserved proteins of unknown function, 2 flagellar-related |
| 15 | Aldehyde dehydrogenases |
| 15 | Mainly EBP transcriptional regulators |
| 14 | Methyltransferases |
| 14 | Acyl transferases |
| 13 | FAD/NAD(P) oxidoreductases |
| 13 | Formate dehydrogenases/DMSO reductases |
| 13 | Membrane proteins |
| 12 | Transport proteins (SSS, DAACS families) |
| 12 | Mainly P-type ATPases |
| 12 | Transport proteins (RND, GntP families) |
| 12 | PTS transporters, IIA component |
| 12 | DeoR type transcriptional regulators |
| 12 | Periplasmic chaperones |
| 11 | Transporters of ABC superfamily (membrane component) |
| 11 | Dehydratases/deaminases for cysteine, serine, threonine, tryptophan |
| 11 | Sugar kinases |
| 11 | Membrane components of dehydrogenases |
| 11 | Oxidoreductases, NAD(P)-binding subunits |
| 11 | Transporters of ABC superfamily (periplasmic-binding component) |
| 11 | Transporters of ABC superfamily for amino acids (periplasmic binding component) |
| 11 | Outer membrane proteins (N-terminal of following group) |
| 10 | Outer membrane proteins (mostly C-terminal of preceding group) |
| 10 | Transporters (NCS2 and GntP families) |
| 10 | Mixed functions, mostly related to DNA |
| 10 | Acetyl-CoA synthetases/ligases |

Size, number of proteins in the group.

*coli* proteins (Fig. 5, *top left*) to singles that had no sequence-related partner in *E. coli* above threshold values (Fig. 5, *top right*) and compared lengths of all paralogs (Fig. 5, *bottom left*) with the lengths of unit modules (Fig. 5, *bottom right*). The average length for all *E. coli* proteins was 340 residues. Singles, the unique proteins that do not reside in families, were smaller and averaged 268 residues in length. Paralogous proteins in sequence-related families are longer than singles and longer on average than the sum of all proteins, with their average protein length being 384.

Table 4. *Distribution of main types of functions in E. coli gene products*

| | Ratio of Experimentally Known to Putative | | |
|---|---|---|---|
| | Paralogs | Singles | Total |
| Enzyme | 616/360 | 371/180 | 987/540(34.4) |
| Transporter | 269/250 | 40/45 | 309/295(13.8) |
| Regulator | 168/118 | 49/34 | 217/292(11.5) |
| Factor | 45/26 | 63/7 | 108/33(3.2) |
| Membrane | 28/88 | 21/28 | 49/116(3.7) |
| Structure | 19/26 | 72/9 | 91/35(2.8) |
| Carrier | 20/13 | 14/13 | 34/26(1.4) |
| RNA | NA | 112/0 | 112(2.5) |
| Phage, IS | 132/33 | 149/66 | 281/99(8.6) |
| Leader | NA | 12/0 | 12(0.3) |
| Total* | 1,295/909 | 903/382 | 2,198/1,291(79.4) |
| Unknown | 215(8.9) | 698(35.2) | 913(20.6) |

Singles, proteins that have no *E. coli* homologs. Values in parentheses for "Total" indicate percentage of function class in all *E. coli* ORFs (4,438). Values in parentheses for "Unknown" indicate percentage of proteins with no assigned function. *Excluding proteins of unassigned function. IS, insertion sequence; NA, not applicable.

The bulk of the individual modules were found to have a length between 100 and 150 residues. The average value for all modules was 219. When we included shorter modules between 83 and 100 residues long, the peaks of the module length distribution did not shift, presumably because the proteins in this size range are mostly singles. With shorter lengths, evidently singles are largely composed of unimodular proteins. These data suggest that the longer proteins in sequence-related families have more often than the singles undergone gene fusion and internal gene duplication, thus becoming multimodular.

## DISCUSSION

*The value to evolution of sequence-related protein families within one species.* Since the members of sequence-related protein families in *E. coli* have proved to be related in function, the data we have collected highlight the usefulness of attributing functions to unknown proteins not only by orthologous matches as is the present custom, but also by paralogous matches. Paralogs form protein families of similar sequence within the organism whose shared functions can be used fruitfully in annotating functions of unknowns within the families. Sequence-similar protein families from a single organism are particularly useful in functional genomics and molecular evolution since, barring horizontal transfer from other sources, families of pro-

Table 5. *Characteristics of three major types of E. coli proteins*

| | Enzyme | Transporter | Regulator |
|---|---|---|---|
| Number of groups | 287 | 70 | 37 |
| Average group size | 3.6 | 8.1 | 8.3 |
| Internal duplication* | 20(2) | 26(6) | 5(1.6) |
| Gene fusion* | 63(6) | 49(12) | 22(7) |
| Singles:paralogs | 1:2.1 | 1:4.7 | 1:3.7 |

Values in parentheses are percentages. *Data limited to paralogs

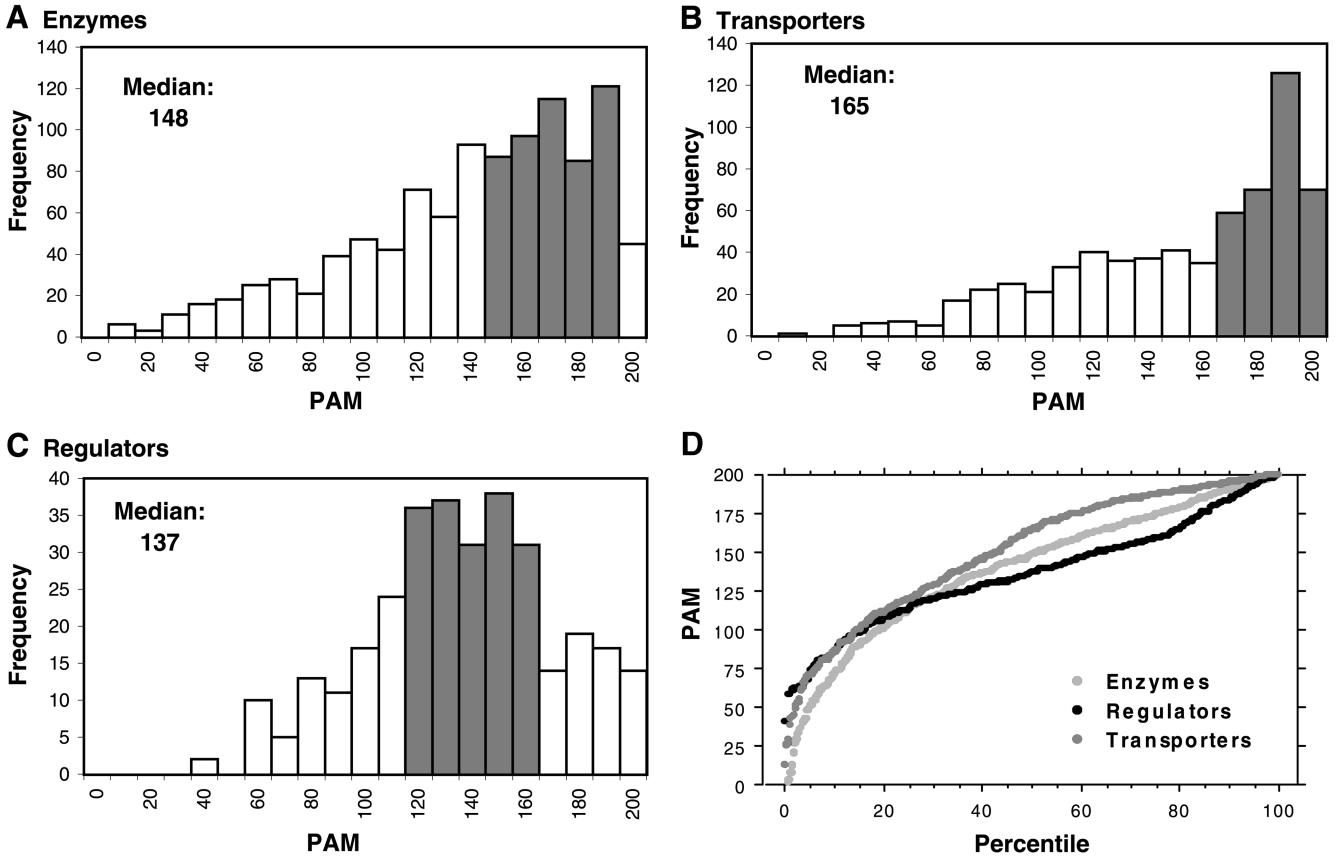


Fig. 4. PAM distribution patterns for paralogous enzymes, transporters, and regulators. For each protein, only the best match as represented by the smallest PAM value was used. The shaded bars in *A*–*C* represent the least number of consecutive histogram bars of PAM that cover 50% of proteins in the category. *D*: percentile plots of PAM values in three groups using the Wilcoxon rank sum test.

teins of similar sequence in a given organism are believed to have descended together under a common, shared set of cellular circumstances and organism demands. By contrast, when relating orthologous genes from different, often physiologically quite different, organisms, one must take into account a succession of noncomparable events over time in the two lineages that may have influenced the development of the orthologous sets of paralogous genes in ways for which we cannot correct. There are questions that are not answered by looking at the function of orthologs from other organisms; however, in some cases these questions can be answered by examining instead the functions of members of internal protein families.

Data on protein families internal to one organism also bear on evolution mechanisms, allowing us to ask whether different categories of proteins have different characteristics of molecular evolution (9, 17). The field of genomics has provided us with complete genomic sequences of over 40 organisms, most of them from unicellular organisms. In each case, composition of protein families can be windows on protein evolution and the origin of life. A generally shared view of protein evolution is that a diversity of proteins was present in the last universal common ancestor (LUCA) (4, 30), namely, the collection of cellular entities that catalyze metabolism, cell components, macromolecule synthesis, and cell division capabilities. Early genes duplicated and diverged. The descendants are present in the majority of known life forms today. Thus the sequence-related groups of proteins found in *E. coli* that are also present in nearly all organisms in the tree of life must trace back to early ancestors. Other groups present only in bacteria, for instance, or in certain kinds of bacteria, must have arisen more recently after divergence of bacteria from other domains of the tree of life (Fig. 6).

*Conservation characteristics of enzymes, regulators, and transporters.* The sizes of sequence-similar groups differ by type of function. Regulators and transporters are clustered in larger groups than are the enzymes. The sizes of the sequence-related groups presumably reflect the extent of divergence of their members from each other over time. Since enzymes spread across a broad range of family sizes, with many in small groups and pairs, it seems that many enzymes have diverged from each other more than the transporters and regulators have. Similarly, among the unpaired, single *E. coli* proteins, the numbers of enzymes is proportionally higher relative to regulators and transporters, especially the latter (Table 4). These singles may be viewed as the only surviving members of earlier paralogous

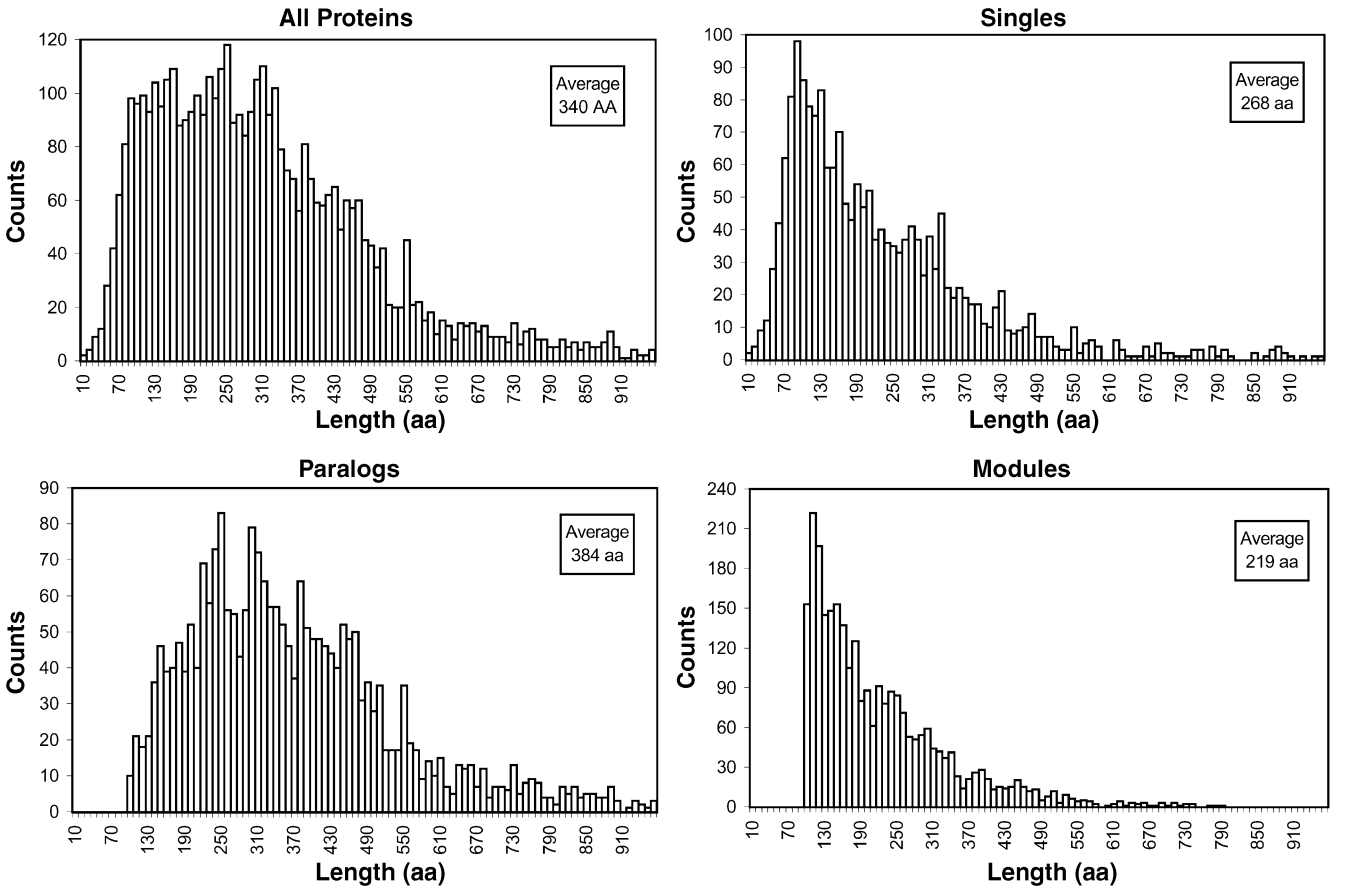


Fig. 5. Length distribution patterns for all proteins, singles, paralogs, and modules in *E. coli*. The lengths in number of amino acid residues were rounded up to the nearest 10 before plotting. Average lengths were used for modules when multiple alignments were present within the same amino acid sequence. The average length for proteins or modules having at least 100 residues in each group is indicated in the *insets*.

groups or as relatives of other *E. coli* proteins that have diverged too far from each other for the relationship to be detected (14).

One possible explanation is that transporters and regulators may have found a few "winning formulas" that were simply varied for specificity over and over, meeting the needs of the cell with this level of particularity. By contrast, enzymes may have diverged to a much greater extent to meet the very numerous and specific catalytic needs for the complex metabolic networks of the cell. A comparable analysis for functionally distinct paralogous groups for other genomes will tell us whether all transporters, regulators, and enzymes exhibit the same size distributions of internal protein families and thus that the course of evolution of types of proteins tends to be universal among all organisms.

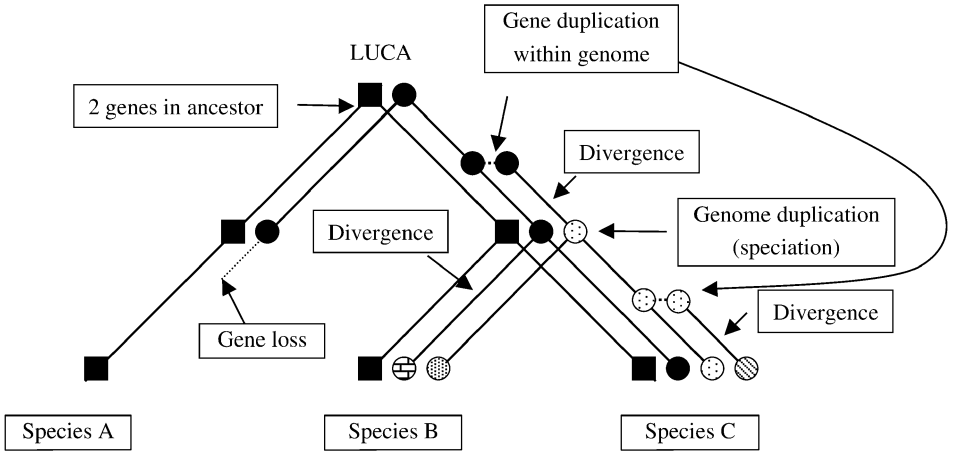


Fig. 6. A cartoon illustrating simplified patterns of protein evolution. Two ancestral proteins from the last universal common ancestor (LUCA) have been distributed to separate lines of descent by a speciation event. Within one of the species, dashed lines connect outcomes of gene duplication and divergence within one genome, events that when repeated can generate different families of related proteins.

Both amino acid sequence and function are closely related in the paralogous groups of *E. coli*. Particular types of transporters cluster together, types of transcription regulators cluster, and types of enzymes cluster (Table 3). For instance, looking at the larger families of transporters, the three types of subunits of the multimeric ABC transporters each cluster together by type (the membrane components, the ATP-binding components, the solute-binding transmembrane periplasmic subunits). Also, subsets of the major facilitator superfamily (MFS) and the amino acid/polyamine/choline (APC) superfamily of amino acid transporters each cluster by sequence similarities. In most cases the sequence-relatedness of transporters is consonant with the classification by function of types of transporters (Ref. 20; **http://www.biology.ucsd.edu/~ipaulsen/transport/**).

Types of regulators also group by function and sequence. For transcriptional repressors of the LysR type or GntR type, transcriptional activators of the AraC type, and other groups of regulators in *E. coli*, the grouping by sequence agrees with grouping by class of regulator (26) (**http://www.cifn.unam.mx/Computational_Genomics/regulondb**). Also the modules of sensor and response regulator of two-component regulators each belong to distinct sequence families, sometimes joined in multimodular proteins, sometimes in unimodular, separate proteins (29).

Types of enzymes with distinct similarities in catalytic properties also cluster by sequence. Families of enzymes with distinct similarities in properties, such as ATP-dependent helicases, GTP-binding proteins, methyltransferases, acyl transferases, transaminases, sugar kinases, dehydratases, and acetyl-CoA synthetases, each fall into a discrete sequence-similar family. Members of each family are related by chemistry of reaction but differ in substrate specificity. Other enzyme activities have more than one solution and can be achieved more than one way. Such types of enzymes split into several families. For instance, there are several sequence types of NAD(P)-requiring dehydrogenases that separate into more than one family.

*Multimodular proteins.* Multimodularity of some proteins has long been known (13). Their existence introduces complications to evolutionary reconstructions unless the individual components are identified and treated separately. Some genes are composed of multiple independently functioning modules that seem to have separate evolutionary histories but have come together by some processes of recombination leading to gene fusion.

Not all proteins in *E. coli* that are known experimentally to be multimodular were detected by our procedures, either because there are no paralogs within *E. coli* to one or more of the modules or because such relationships are not visible above the relatively conservative threshold we used to define sequence similarity. For example, with no other sequence similar to homoserine dehydrogenase detected within the *E. coli* genome, only the aspartokinase modules of AK1H (gene *thrA*) and AK2H (gene *metL*) were detected. There are, however, separate orthologs to both modules of *thrA* and *metL* in other organisms, for both the homoserine dehydrogenases and the aspartokinases. Thus additional modules can be identified by searching for orthologous matches. Therefore the actual number of multimodular proteins in *E. coli* is higher than reported here.

*Cautions for functional annotation.* It goes without saying that any data concerning sequence similarity, such as the use of sequence similarity for functional annotation purposes, must recognize and operate with the appropriate alignment region and the unit of similarity, the module, as we have done here, rather than using complete genes or proteins when they are complex. For instance, similarity to the *E. coli* AK1H only in the $NH_2$-terminal half should not be taken to indicate homoserine dehydrogenase activity in a homolog since the alignment was limited to the aspartokinase region. With respect to functional analysis, misassignment of the function of one module to any protein matching the other module instigates a chain of errors. Such misattributions contaminate databases but could be avoided by confining conclusions about functional similarity to the correct matching regions of sequence similarity between the query and the subject sequences.

Also, for different reasons caution is needed when transferring the exact function of a known protein to another of unknown function based on sequence similarity. As it happens, sequence similarity does not always spell out close similarity of catalytic function. There are functionally diverse superfamilies of proteins that spell difficulty for annotating of function by sequence similarity. There are sequence-related enzyme families that are catalytically diverse, retaining an underlying similarity of the structure of the active site, but using the same chemical mechanism for different overall reactions (9). Examples in *E. coli* sequence-similar groups, are the SDR superfamily (Kerr A and Riley M, unpublished observations) and the crotonase superfamily (McCormack T and Riley M, unpublished observations). In cases where a sequence-related group is not immediately recognized as being a superfamily of diversified proteins that catalyze related but different reactions, the attribution of an exact function of a known protein to an unknown protein can be entirely wrong.

Another kind of relationship in sequence-similar families is preservation of substrate with divergence of function. For instance, the sequence-related proteins RBSR and transporter RBSB are examples of proteins that have maintained ligand specificity (ribose) while changing action of the protein (16). RBSR is a regulator for the ribose operon, and RBSB is a transporter for ribose. The *E. coli* family to which this pair belongs contains both periplasmic binding proteins and transcriptional regulators. Although mode of evolution of function is interesting in these kinds of cases, unfortunately whenever protein families contain members of different function, in this case regulators and trans-

porters, difficulty arises for accurate function prediction of unknown proteins of similar sequence.

That being said, one should not place too much emphasis on the cases where sequence similarity indicates membership in a protein family of diverse functionality, since the great majority of the sequence-similar protein families in *E. coli* unambiguously share a primary function. Most share chemistry of reaction and differ only in specificity of ligand/substrate (17).

*Applications of paralogous protein families in genome annotation.* The results reported in this research provide useful annotation information in at least three ways. First, existing paralogous relationships of sequence similarity within a genome can be useful in attributing function to unknown proteins when there is little useful information from orthologs in current databases. The method of transitive assembly of the paralogous groups leaves some members of a group only marginally connected, yet where functions of the most distant members are known, the functions are usually clearly related to those of the group as a whole. Thus information can be derived even when the degree of sequence conservation between two paralogous proteins is sometimes below the standard threshold for detecting sequence similarity among orthologs. Internal paralog families are particularly useful in cases where the genes are unique to the studied genome or orthologs have not been found in other organisms. This is reflected in the fact that in *E. coli* paralogs as a class have a much lower percentage of unknown members compared with singles (8.9% vs. 35.2%) (Table 4). In our experience with *E. coli* K-12 and *Halobacterium* NRC-1 genomes, we were able to make putative assignments for at least 10% of genes by using this approach (19; **http://zdna.micro.umass.edu/haloweb/**).

Second, identification of multimodular proteins and location of the correct functions to the correct parts of the proteins improves the accuracy of the annotation. It is clear that multimodular proteins exist in all organisms, and in many cases only one of the functions is currently known. More complete information and location of the activities will correct misattributions and errors of omission in the annotations in current genomic databases.

Third, by examining the evolutionary relationships among the members of larger paralogous groups, additional information such as the type of enzymatic reaction or the substrate specificity may be obtained for those member proteins that are currently characterized only with a putative general function. Also, considering the finding that different degrees of sequence conservation exist among different types of proteins such as enzymes, regulators, and transporters, different thresholds may be optimal for different function types. Whereas putative transport function may be assigned with good confidence using a marginal sequence similarity, for regulators and even more so for many enzymes, level of sequence similarity may need to be more conservatively defined.

*Conclusions.* Sequence-related protein families within a single organism which are assembled with special attention paid to the existence of multimodular (composite) proteins have useful applications both in understanding elements of molecular evolution and in improving genome annotation. Paralogous protein families each presumably descended from an individual ancestral protein can be inferred from families of sequence-similar proteins encoded within an individual genome. We found that most such sequence-related families contained proteins with the same type of function. In a few cases not expanded on here, there is divergence of function among group members, showing how sequence divergence among similar family members can lead to further divergence of function. Generally speaking, paralogous group membership provides a basis for assigning putative functions to unknown members, which is particularly useful when no information is available through orthologous matches. In addition, the approach improves the ability to identify multimodular proteins, locating specific functions to different parts of a protein. Membership in well-defined clusters within large paralogous groups affords the opportunity for even more specific functional characterization. Therefore, the approaches suggested here could be useful additions to existing methods for genome annotation.

We thank Dr. S. Teichmann for providing us her latest SCOP assignments for *E. coli* proteins, and Alida Pellegrini-Toole for valuable assistance preparing the manuscript.

This work was supported by National Aeronautics and Space Administration Astrobiology Institute grant NCC2-1054 and the Merck Genome Research Institute.

## REFERENCES

1. **Altschul SF.** Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219: 555–565, 1991.
2. **Berlyn MKB.** Linkage map of *Escherichia coli* K-12, edition 10: the traditional map. *Microbiol Mol Biol Rev* 62: 814–984, 1998.
3. **Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew G, Gregor J, Davis NW, Kirkpatrick HA, Goeden M, Rose D, Mau B, and Shao Y.** The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1474, 1997.
4. **Doolittle WF.** Phylogenetic classification and the universal tree. *Science* 284: 2124–2129, 1999.
5. **Felsenstein J.** Phylogeny inference package (Version 3.2). *Cladistics* 5: 164–166, 1989.
6. **Fitch WM and Margoliash E.** Construction of phylogenetic trees. *Science* 155: 279–284, 1967.
7. **Fitch WM.** Distinguishing homologous from analogous proteins. *Syst Zool* 19: 99–113, 1970.
8. **Gasteiger E, Jung E, and Bairoch A.** SWISS-PROT: connecting biological knowledge via a protein database. *Curr Issues Mol Biol* 3: 47–55, 2001.
9. **Gerlt JA and Babbitt PC.** Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* 70: 209–246, 2001.
10. **Gonnet GH, Hallett MT, Korostensky C, and Bernardin L.** Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* 16: 101–103, 2000.
11. **Jornvall H, Hoog JO, and Persson B.** SDR and MDR: completed genome sequences show these protein families to be large, of old origin, and of complex nature. *FEBS Lett* 445: 261–264, 1999.
12. **Kessler D, Leibrecht I, and Knappe J.** Pyruvate-formate-lyase-deactivase and acetyl-CoA reductase activities of *Escherichia coli* reside on a polymeric protein particle encoded by adhE. *FEBS Lett* 281: 59–63, 1991.

13. **Kirschner K and Biswanger H.** Multifunctional proteins. *Annu Rev Biochem* 45: 143–166, 1979.
14. **Labedan B and Riley M.** Genetic inventory: *Escherichia coli* as a window on ancestral proteins. In: *Organization of the Prokaryotic Genome*, edited by Charlebois RL. Washington, DC: ASM, 1999, p. 311–329.
15. **Linton KJ and Higgins CF.** The *Escherichia coli* ATP-binding cassette (ABC) proteins. *Mol Microbiol* 28: 5–13, 1998.
16. **Mauzy CA and Hermodson MA.** Structural homology between rbs repressor and ribose binding protein implies functional similarity. *Protein Sci* 1: 843–849, 1992.
17. **Nahum LA and Riley M.** Divergence of function in sequence-related groups of *Escherichia coli* proteins. *Genome Res* 11: 1375–1381, 2001.
18. **Needleman SB and Wunsch CD.** A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48: 443–453, 1970.
19. **Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL, Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ebhardt H, Lowe TM, Liang P, Riley M, Hood L, and DasSarma S.** Genome sequence of *Halobacterium* species NRC-1. *Proc Natl Acad Sci USA* 97: 12176–12181, 2000.
20. **Paulsen IT, Sliwinski MK, and Saier MH Jr.** Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and specificities. *J Mol Biol* 277: 573–592, 1998.
21. **Pearson WR.** Comparison of methods for searching protein sequence databases. *Protein Sci* 4: 1145–1160, 1995.
22. **Riley M and Labedan B.** Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J Mol Biol* 268: 857–868, 1997.
23. **Riley M and Serres MH.** Interim report on genomics of *Escherichia coli*. *Annu Rev Microbiol* 54: 341–411, 2000.
24. **Rudd KE.** EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res* 28: 60–64, 2000.
25. **Serres MH, Gopal S, Nahum L, Liang P, Gaasterland T, and Riley M.** A functional update of the *E. coli* K-12 genome. *Genome Biol* 2: 0035.1–0035.7. (http://genomebiology.com/2001/2/9/research/0035)
26. **Salgado H, Moreno-Hagelsieb G, Smith TF, and Collado-Vides J.** Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA* 97: 6652–6657, 2000.
27. **Smith TF and Waterman MS.** Identification of common molecular subsequences. *J Mol Biol* 147: 195–197, 1981.
28. **Vogt G, Etzold T, and Argos P.** An assessment of amino acid exchange matrices in aligning protein sequences: the Twilight Zone revisited. *J Mol Biol* 249: 816–831, 1995.
29. **West AH and Stock AM.** Histidine kinases and response regulator proteins in two-component signaling systems. *Trends Biochem Sci* 26: 369–76, 2001.
30. **Woese CR, Kandler O, and Wheelis ML.** Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87: 4576–4579, 1990.
31. **Yourno J, Kohno T, and Roth JR.** Enzyme evolution: generation of a bifunctional enzyme by fusion of adjacent genes. *Nature* 228: 820–824, 1970.