J. Mol. Biol. (1997) **268**, 857–868

JMB

Protein Evolution Viewed through *Escherichia coli* Protein Sequences: Introducing the Notion of a Structural Segment of Homology, the Module

Monica Riley[1]* and Bernard Labedan[2]

[1]*Marine Biological Laboratory Woods Hole, MA 02543, USA*

[2]*Institut de Génétique et Microbiologie, CNRS URA 1354, Bâtiment 409, Université de Paris-Sud, 91405 Orsay Cedex, France*

**Corresponding author*

Paralogous genes are genes which descend from a progenitor gene which has duplicated as an ancestral gene, each copy having diverged prior to speciation. With comprehensive information available on functions of *Escherichia coli* proteins, analysis of sequence-related *E. coli* paralogous proteins can give information on the early ancestors of families of proteins now residing in many contemporary organisms, such as the enzymes of metabolism, some kinds of transport mechanisms and some kinds of regulatory mechanisms. In the first step, we have confirmed that *E. coli* contains a very high proportion of paralogous proteins. Next, we have defined two main classes of paralogous proteins. One class is formed of proteins which contain a unique structural segment homologous to a single set of related proteins. The other class corresponds to proteins which contain more than one structural segment of homology, each segment homologous to unrelated sets of proteins. We define such an independent structural segment of homology as a module. This modular structure (mean length equivalent to 209 amino acids) corresponds often to entire proteins, but there are also proteins that appear to be assembled from two or three independent modules having independent origins. Most multimodular proteins appear to have been formed early in their history, a minority appear to be relatively recent fusions of independent modules. Examining 1404 independent structural segments of homology, composed of both modules and entire proteins, we found that the segments of homology fell into 352 sequence-related groups or families. The majority of these families (ranging from 2 to 62 members) are functionally homogeneous. This strongly suggests that the 1404 present-day modules and proteins derive from a minimal set of 352 ancestral modules, each one being already of the same size and having a function similar to all members of its progeny.

Keywords: E. coli; paralogous; multimodular proteins; evolution; protein families

Introduction

To explore the evolutionary histories of proteins, we have analysed the sequence relationships among all proteins of *Escherichia coli* whose sequence is available, representing about 65% of the whole *E. coli* chromosome. Indeed, the genome of any organism may be viewed as a microcosm for the basic functions of all living beings, reflecting the entire process of evolution of specialized genes and gene products from a set of unique ancestral genes that existed at some early time in the evolution of organisms.

Many present-day genes have homologues present both in other genomes and in the same genome. These two classes of homologous genes are named orthologous and paralogous, respectively, according to Fitch (1970). Orthologous genes descended from a unique ancestral gene and their

Abbreviations used: DARWIN, Data Analysis and Retrieval With Indexed Nucleotide/Peptide Sequences; PAM, number of accepted point mutations per 100 residues separating two sequences; ORF, open reading frame; SQL, structured query language; AdhE, alcohol dehydrogenase/acetaldehyde dehydrogenase.

0022–2836/97/200857–12 $25.00/0/mb971003

divergence with comparable genes in different organisms is simply parallel to speciation. Therefore, they are good instruments for building phylogenetic trees of organisms. Paralogous genes, on the other hand, descended from copies of a gene which duplicated within a single ancestral genome. These copies have diverged prior to speciation and are good candidates for understanding protein evolution. Indeed, one can imagine that the earliest ancestral proteins had broad specificity, catalysing whole classes of reactions with one enzyme and, progressively, more specialized proteins with narrow specificity have been produced over time by duplication and divergence of the corresponding ancestral genes (Jensen, 1976). Thus, sets of paralogous genes and their gene products can inform us of the ancient evolutionary history of macromolecules. Deep evolutionary times are accessible by sequence analysis of paralogous proteins as long as the changes to sequences over time by processes of mutation, recombination and repair have not blurred the similarities so they cannot be discerned today over background noise.

In recent work, particular subsets of *E. coli* paralogous proteins have been analysed in detail: transport, receptor and transcriptional regulatory proteins (Saier, 1996) and a set of regulators and the enzymes they regulate (Otsuka *et al*., 1996). Also a set of 28 paralogous ORFs in the yeast genome have been characterized as they relate to multidrug transporters and major facilitators in yeast (Goffeau *et al*., 1996). Spreading our inquiry further and more comprehensively, we have undertaken to examine all paralogous *E. coli* proteins whose sequences are currently available. In earlier studies we used the data of the SwissProt database (Bairoch & Apweiler, 1996) releases 26 and 28 to match the sequences of all *E. coli* K12 chromosomally encoded proteins against themselves (Labedan & Riley, 1995a,b). We found that at least 52.2% of *E. coli* proteins have similarity of sequence to at least one other *E. coli* protein in SwissProt release 28 (Labedan & Riley, 1995b). This strongly suggested that early ancestry and evolutionary relatedness of proteins can be discerned from current sequence data. To take this analysis further and to learn more about the evolutionary ancestry of today's proteins, we have analysed the new, significantly increased set of *E. coli* K-12 chromosomal protein sequences present in SwissProt release 33 and we have correlated the functions of long segments present in these proteins with their sequence relatedness. Indeed, in this new study, we have chosen to examine in more detail the inheritance of entire proteins and major modules, taking into account that some proteins are multimodular, and that a significant proportion of these multimodular proteins appear to be the result of an ancient fusion of evolutionarily unrelated modules. We have evaluated how these modules affect the assembly of paralogous genes in families and we propose the module as a new unit of evolutionary descent.

Results

Assembling the pairs of similar proteins into families

We have previously shown that it is possible to identify pairs and groups of genes in the *E. coli* chromosome which are likely to have resulted from duplication of ancestral sequences (paralogous proteins) by assessing their sequence similarities and functional similarities (Labedan & Riley, 1995a,b). We have also shown that the AllAllDB program of the Darwin package (Gonnet *et al*., 1992) was very efficient in finding out, in one step, the whole set of putative paralogous proteins from all the *E. coli* protein sequences in the SwissProt database (Bairoch & Apweiler, 1996). The AllAllDB program processed 3466 sequences which were attributed to *E. coli* in SwissProt database release 33. The output contained 11,914 pairs, which we edited to remove plasmid sequences, transposon sequences, prophages, and proteins of strains of *E. coli* other than K12, leaving pairs derived from 2548 K12 chromosomal sequences. Then, as in earlier studies (Labedan & Riley, 1995a,b), we kept only sequence-related pairs having a PAM score of less than 250 and alignments of at least 100 amino acids. These two cutoffs were based on the finding that a PAM250 substitution matrix is the most efficient scoring matrix when applied to distantly related protein pairs for a minimum significant length of 83 residues (Altschul, 1991). Moreover, Figure 1 shows that many of the lengths of the alignments are greater than the arbitrarily set minimum of 100, since the distribution of alignment lengths has a median length of 209 amino acids. Therefore, in this study, we are focusing on the evolutionary history of segments of length around 200 amino acids, in strong distinction to other studies which used as a criterion of paralogy the presence of functional motifs as small as 5 to 20 amino acids (Koonin *et al*., 1995, 1996).

The cleaning of the initial AllAllDB output gave 5121 pairs of *E. coli* K12 chromosomally encoded proteins having sequence similarities along long segments extending up to the entire protein length. These 5121 pairs correspond to 1591 (including 526 open reading frames) paralogous protein sequences, many sequences belonging to families of paralogous proteins. In the next step, we tried to characterize the different families of paralogous proteins, with the objective of finding the minimal number of putative ancestral sequences, assuming that each family derives from one ancestral gene. We grouped the 5121 pairs into families of sequence-related pairs, assembled by collecting all relatives of each sequence of a pair and all relatives to those sequences and so on. Two equivalent programs (see Materials and Methods) were used to build families from all sequences in the list of 5120 pairs. Grouping the proteins by this automatic chaining method was only partially successful. Numerous small groups were found but more than

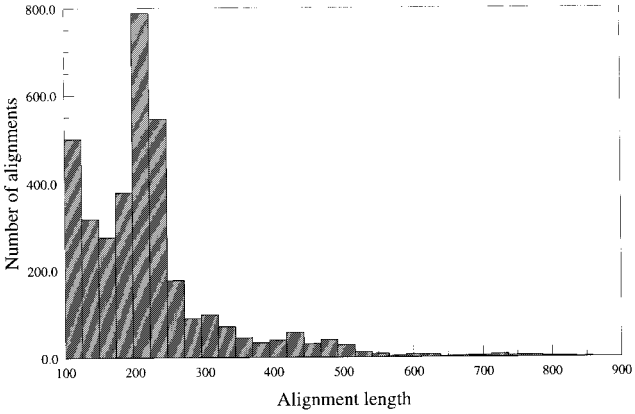

Figure 1. Distribution of lengths of amino acid sequence alignments of *E. coli* paralogous proteins of lengths of least 100 amino acids and PAM values less than 200.

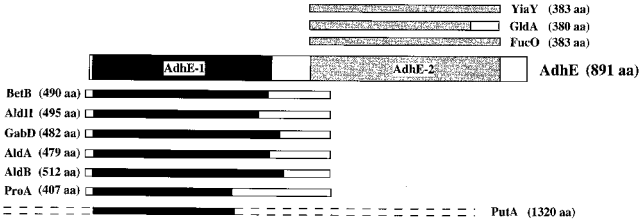

Figure 2. Example of alignments of unrelated sets of proteins with two independent modules of a multimodular protein.

67% (1066) of the sequences were in a very large group. We found that a significant fraction of all the paralogous proteins contain more than one non-overlapping alignment region in their sequence. We call these proteins multimodular proteins, a module being an independent alignment region of more than 100 amino acids. In some cases, the different modules present in one multimodular protein were found to be homologous to unrelated sets of sequences. Because of this, the automatic programs not only connected sequence-related proteins to one another but also chained together all homologues to independent modules present in the same protein, thus mixing together families of unlike sequences in a huge artificial group. This is illustrated in Figure 2 where a multimodular protein, the alcohol dehydrogenase/acetaldehyde dehydrogenase AdhE, is shown as a bimodular sequence. The N-terminal half of AdhE aligned with the major part of the sequences of several aldehyde dehydrogenases BetB, AldH (putative), GabD, AldA, AldB and ProA as well as a central module of a very large protein, the proline dehydrogenase PutA. The C-terminal half of AdhE aligned with the whole sequences of another class of dehydrogenases, belonging to the iron-containing dehydrogenase family, namely the proteins FucO, GldA and the open reading frame YiaY. Clark (1992) suggested that AdhE is the product of the fusion of two independent genes encoding alcohol dehydrogenase and acetaldehyde dehydrogenase activities. The automatic family-making programs put in the same group these two unrelated classes of dehydrogenases, which are totally dissimilar in their sequences. A modification in the approach was needed.

Defining the parameters to assemble the families

In order to avoid such artificial connections, we first required that a large percentage of a sequence be in the alignment, either 60% or 80% of the total length of one protein in each of the 5121 pairs. The number of pairs was reduced to 3320 when the 60% requirement was imposed and to 2336 when the 80% requirement was imposed. With this procedural change, the sequences were grouped by linking similar sequences processively as before. There were 1290 sequences in pairs aligned over 60% or more of their length and 1061 sequences in pairs aligned over 80% or more. Relative to the unscreened results, smaller groups were formed in the 60% aligned set but still there was one large group of 328 heterogeneous sequences. Only the 80% aligned set had groups corresponding to homogeneous families, the largest containing 55 members. This approach was helpful in identifying sequences with *bona fide* homologies and in giving us a blueprint of evolutionarily consistent families made of long evolutionarily related protein segments. However, it had the important drawback of failing to deal with a significant proportion of the data.

Therefore, we looked for a more appropriate way to identify both entire proteins and all the individual modules that belonged to separate families corresponding to separate non-overlapping alignment regions in the multimodular proteins. Independent alignments are identified in the AllAllDB output by their starting and ending residue numbers. To deal with multiple heterogeneous modules separately, we numbered modules in each multimodular protein starting from the N-terminal end, as −1, −2, or −3. For example, in the case of the AdhE protein (Figure 2), we treated the modules AdhE-1 (N-terminal half) and AdhE-2 (C-terminal half) as two independent entities.

Another change was introduced to increase the significance of the results. We had already observed (Riley & Labedan, 1996) that PAM values between 200 and 250 corresponded to the tail of the extreme value distribution and we showed that the excess of sequences it contained was due mainly to the presence of many proteins with PAM value below 250 but displaying identity values less than 20%. This is supported when we look at the shape of the distribution of PAM values after removing pairs with PAM values over 200 (Figure 3). A more normal distribution results, which has a median value of 165. For the remain-

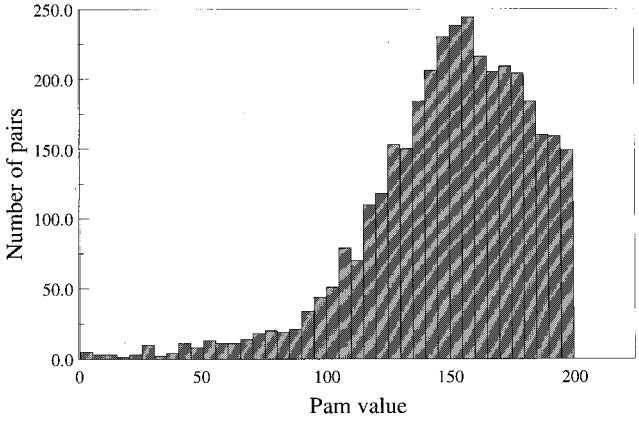

Figure 3. Distribution of PAM values of pairs of paralogous proteins from *E. coli*.

der of this work, we used the 3548 pairs having PAM values less than 200, corresponding to 1321 sequences. The 1321 sequences were composed of 1248 entire proteins and 74 multimodular proteins composed of 156 modules (Table 1). Note that many of the 1248 proteins also contain two or more modules but they did not create artificial linkings between unrelated families since they paired only with other proteins containing the same modules in the same order.

The final list of 1404 evolutionarily related segments (sequences and modules) were grouped transitively (following chains of similarities until they were exhausted), producing a total of 352 groups (Table 2). The 352 families of sequence-related proteins varied in size, ranging from two members to 62 members, as listed in Table 2.

Characterizing the functions of evolutionarily related modules

To examine the biological aspect of the groups of sequence-related proteins, members of all groups were examined for functional relatedness. To assign a cellular function to each of the gene products, we have used data previously assembled on functions of *E. coli* gene products (Riley & Labedan, 1996) as well as updates based on recent literature and information added in the current SwissProt database. Many gene products of *E. coli* have been well characterized, but some have not and are known only in terms of non-specific mutant phenotypes. We have assigned a type ''unknown'' to proteins where either the function has not been experimentally determined, or the sequence is known only as an open reading frame (ORF), or the sequence has a putative function, i.e. a function based on sequence similarities. Remarkably, 49 of the 352 groups were composed entirely of ORFs and in total 39.9% of the paralogous partners have an unknown type of function. Where function of the modules could be assessed with certainty, we found that, in most cases, all members of a group carried out the same kind of function. The preponderant functions of the modules present in the larger groups are shown in Table 3 and the functions of the smaller groups are summarized in Table 4.

The enzymes were found to group for the most part by the type of enzymatic reaction, and transport and regulation proteins grouped by mechanism (Table 3). Simple enzymes composed of one polypeptide chain, or multimers of one chain, grouped together by reaction type, as reflected in similarity of Enzyme Commission (EC) numbers (Webb, 1992). The EC numbers are informative when the enzymes in question are simple proteins. For more complex enzymes with two or more polypeptides present as subunits, the EC numbers give no information on the separate properties and roles of the subunits. However, by examining the specialized function of individual subunits of multisubunit enzymes, it was clear they fell into

Table 1. Multimodular proteins

	FTSY	NIKA	RECG	YHAU
AAS	FUMA	NLPD	RECQ	YHBX
ACRB	FUMB	NTRC	RNE*	YHCL
ACRF	GLMS	ODP2	SDHL	YHES
ADHE	GLMU	PBPA	SDHM	YHFT
ARAJ	GYRA	PBPB	SELB	YHIG
ARCB[a]	HSCA	POTE	SLT	YHIV
ATOC	HTRA	PT1A	SMS	YHJL
BARA[a]	HYDG	PTAA[a]	SYM[a]	YIBP
BCR	IF2	PTFA	YAMB[b]	YICK
CADB	IMP	PTGB	YEEF	YIDT
DNAK[a]	LIVM	PTMA	YEHU	YIHO
DP3X	MUKB[a]	PTOA	YEJH	YJCG
DPO1	MURE	RBSC	YFFG	YJGB
EVGS	NARQ	RCSC	YFHA	YJIY

[a] Identifies proteins with three modules
[b] Labels starting with letter Y are ORFs.

Table 2. Distribution of group sizes

Size	Number of groups
2	208
3	64
4	28
5	14
6	8
8	3
9	8
10	4
11	2
13	2
14	1
18	1
21	1
25	1
26	1
30	1
37	1
41	1
53	1
60	1
62	1

Table 3. Functions of the larger groups of sequence-related modules

Number of modules	(U)[a]	Predominant function for known proteins
10	(1)	Oxidoreductase, NAD(P) acceptor
10	(5)	Oxidoreductase, NAD(P) acceptor
10	(3)	Phosphotransferase component
10	(8)	Fimbriae protein
11	(1)	Oxidoreductase subunit, Mo cofactor
11	(1)	Transporters of iron complexes, vitamin B12
13	(3)	GTP binding protein
13	(3)	Oxidoreductase, iron-sulfur subunit
14	(5)	Amino acid permease
18	(5)	Transcriptional activator, ARAC family
21	(4)	Transcriptional repressor, LACI family
25	(11)	Transcriptional activator, LYSR family
26	(9)	DEAD helicase, restriction, recombination
30	(6)	Two-component sensory transduction, sensor
37	(12)	Two-component sensory transduction, regulator
41	(21)	Mixed, transmembrane protein, enzyme
53	(33)	Low affinity transporter, ARAE family
60	(19)	High affinity transporter, ARAH family
62	(26)	ABC transporter, ATP-binding component

[a] Number of unknown proteins in group

groups related by both function and sequence. Examples are iron-sulphur subunits of oxidoreductases, NAD-linked catalytic subunits of oxidoreductases, and enzymes or subunits that bind molybdenum as cofactor.

Among proteins involved in transport, there are separate groups such as the ATP-binding components of the ABC transporters (Boos & Lucht, 1996), or the integral membrane components of these ATP dependent transport systems (Boos & Lucht, 1996) or sugar-proton symport transporters (Maloney & Wilson, 1996). Similarly, two-component regulatory protein sensor domains grouped separately from response regulator domains (for a general overview, see Hoch & Silhavy, 1995). Other types of regulators, the transcriptional activators/repressors, are found in different groups such as the AraC-type, the LacI-type and the LysR-type (for a general overview, see Ninfa, 1996).

However, not all members of sequence-related groups were homogeneous in function. In the group of transcriptional regulators containing LacI and RbsR and other similar regulators, a few periplasmic binding proteins are also present: RbsB, XylF, AraF and MglB. In another family containing membrane components of transport systems, there are two integral membrane proteins with different functions in the system for transport of glycerol 3-phosphate: UhpT is the transporter, and UhpC is a highly specific receptor which is involved in the signal transduction cascade in the same system, causing expression of UhpT (Maloney & Wilson, 1996; Ninfa, 1996).

Finally, predominant group size was characteristic of type of function. The physiological functions of proteins in *E. coli* were not equally represented in large and small families of sequence-related proteins. For small groups of less than ten, enzymes are the major type of protein, many of the enzymes being in highly individual pairs or small groups. The next larger categories in small groups are transport proteins, then regulatory proteins. By contrast, for the largest groups of size ten or more, transport proteins are most common, followed by regulatory proteins, and enzymes are third (Table 4).

Table 4. Functions of modules in small (<10) or large ($\geqslant$10) groups

Small groups Number	Category	Large groups Number
399	Enzymes	85
91	Transporters	105
37	Regulators	92
17	Factors	11
15	Membrane proteins	3
8	Structural proteins	1
2	Carriers	2
360	Unknown function[a]	176
929	Total	475

[a] Described by phenotype only, by sequence similarity only, hypothetical sequence, open reading frame.

Making genealogical trees for groups of paralogous modules

The evolutionary sequence relationships between modules belonging to the same group were further examined through the Phylotree application in the DARWIN program. Many trees (examples are shown in Figures 4 to 8) illustrate the fact that our groups have sequence relationships that suggest an origin from a single ancestor for each group. This is clearly apparent, for example, in Figure 4, where all except the two ORF unknowns (YFFG module 1 and YFFE) and the putative NAPG (conceptual translation) are iron-sulphur subunits of dehydro-

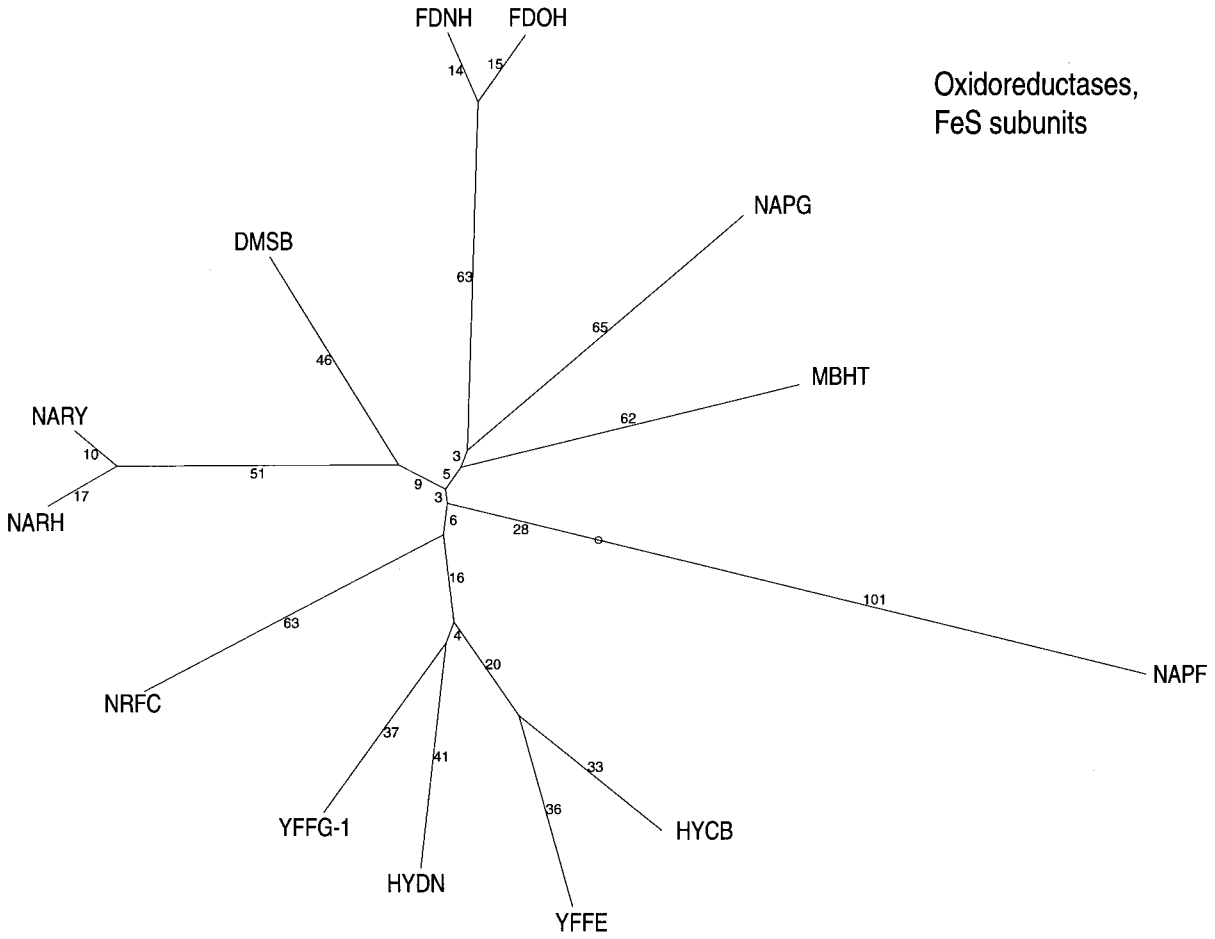

Figure 4. Geneological tree of a group of iron-sulphur subunits of oxidoreductases. Derived by DARWIN function phylotree, with distances given in PAM units. Branches are labelled with the SwissProt mnemonic of each protein. Labels starting with letter Y are ORFs

genases. Likewise, Figure 5 shows that a set of dehydrogenases binding molybdopterin cofactor belong to another independent homogeneous family. Figure 6 is composed largely of transcriptional repressors (see Discussion below) and Figure 7 is largely GTP-binding proteins. Both medium sized (Figures 4 to 7) and large (Figure 8) trees exhibit an unusual star-like topology where branches appear (1) to be of approximately equal length and (2) to emerge from an unresolved centre region.

Discussion

The bacterium *E. coli* appears to be an invaluable tool to analyse the mechanisms of protein evolution, for at least two reasons. (1) *E. coli* is one of the best known organisms where nearly all basic functions necessary to the fundamental mechanisms of life have been well studied, often in great detail (Neidhardt *et al.*, 1996). This has allowed one of us to establish a detailed organizational scheme for all the functions determined by *E. coli* genes (Riley, 1993 and more recent updates available on request or on the Web site www.mbl.edu/html/ecoli.html). These characterizations are being used systematically by people sequencing whole genomes of organisms whose biology is poorly known (e.g. Fleischman *et al.*, 1995; Bult *et al.*, 1996), organisms where ''function'' of putative open reading frames is attributed only by similarity to previously studied genes. Thus correlation of sequence similarity with similarity of physiological function is of interest to the wider microbial genomics community.

Also, (2) we have shown previously (Labedan & Riley, 1995a,b) and now confirm that many *E. coli* proteins belong to families of paralogous proteins. Such paralogous proteins can be used, at least in theory, to trace back the history of the formation of present-day families of macromolecules to the ancestral genes which first duplicated to give birth to whole families. Indeed, one expects that early ancestral proteins existed long before the emergence during evolution of any one organism in its contemporary form. Thus, studies of sets of paralogous genes and of their gene products found in present-day *E. coli* can inform us of the ancient evolutionary history of macromolecules independent of organisms.

With this objective in mind, we have tried to define all the families of paralogous proteins which

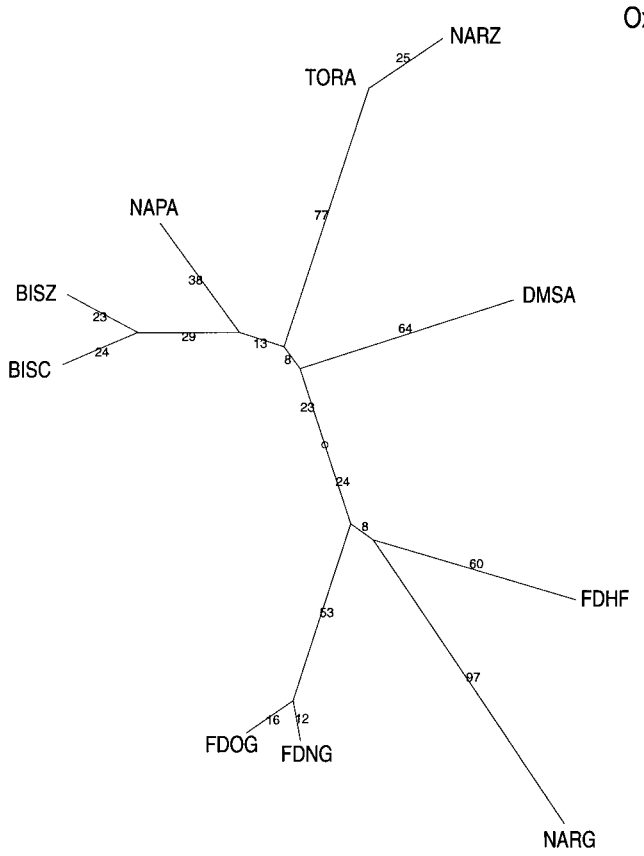

Figure 5. Geneological tree of a group of major catalytic subunits of oxidoreductases. Derived by DARWIN function phylotree, with distances given in PAM units. Branches are labelled with the SwissProt mnemonic of each protein.

could give us unambiguous information about the evolutionary relationships among *E. coli* proteins. To do that we have focused our analysis on sequence similarities extending along long segments of the matching proteins. We grouped together proteins of related sequence and systematically checked for functional similarities among the sequence-related members. This work of family assembly has led to the following conclusions.

(1) We can define two main classes of paralogous proteins. A large majority of the proteins align along the major part of their lengths. A significant minority align with two or more sets of unrelated proteins, corresponding to well-defined independent segments inside the same protein. Therefore, we have a large set of proteins with only one structural segment of homology to one set of related proteins and we have a smaller set of proteins with more than one structural segment of homology, each segment corresponding to unrelated sets of proteins (Table 1). When we replaced these 74 proteins with the 156 corresponding segments, then grouped the 1404 sequences by similarity, 352 homogeneous groups were formed with size varying from 2 to 62 members (Table 2), corresponding to families having a high degree of functional relatedness (Tables 3 and 4).

(2) We define the independent structural segment of homology as a module, corresponding to the minimal size of homology in present-day proteins. We found that this minimal size corresponds to a modular structure of around 209 amino acids (Figure 1). The module as so defined may contain more than one motif and more than one functional domain. It assorts independently in separate groups of sequence-related proteins that we believe have independent evolutionary origins.

In many cases, *E. coli* proteins are made of a single module, but there are also proteins assembled from two or three independent modules having independent origins. Among these multimodular proteins, we detect two categories according to their length of homology. The large majority of multimodular proteins form homogeneous families where all members align along one long region of homology that includes all modules. This category is most probably the offspring of successive duplications of an ancestral entity which contained more than one module as a result of a very old event of fusion. The second category, in the minority (74 instances), corresponds to proteins with a fusion of evolutionarily unrelated modules. These modules have independently duplicated some number of times as unimodular proteins before fusing to give multimodular proteins. Fusion of independent modules produced larger proteins with new properties such as regulatory proteins containing both sensor and response regions, phosphotransferases with two or three of the Enzyme IIA, IIB, and IIC elements, and complex enzymes

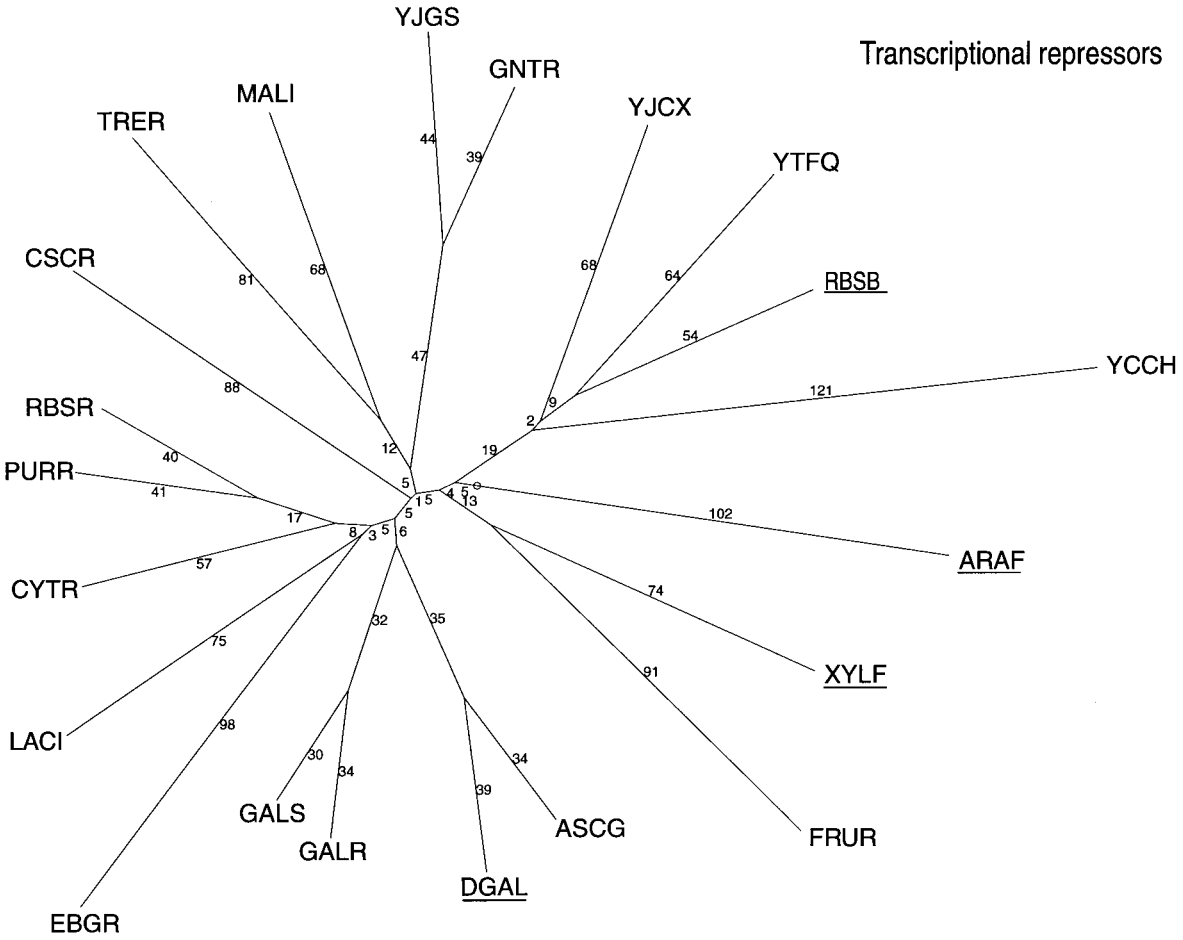

Figure 6. Geneological tree of a group in which most proteins are transcriptional repressors. Derived by DARWIN function phylotree, with distances given in PAM units. Branches are labelled with the SwissProt mnemonic of each protein. Labels starting with letter Y are ORFs.

such as homoserine dehydrogenase-aspartokinase. Fusion of enzymes during evolution seems to have been relatively common (for instance in *E. coli*, the HisIE compound enzyme phosphoribosyl-ATP-pyrophosphatase/phosphoribosyl-AMP-cyclohydrolase and HisHF, the compound two-step enzyme imidazole-glycerol-phosphate synthase). The fusions can be seen in well-studied instances to have taken different paths in different organisms, as for instance in the tryptophan biosynthetic pathway in several microorganisms (Crawford, 1989). Therefore, one can imagine that in the ancestral chromosome there were genes which had duplicated but not yet fused and others which had fused but not yet duplicated.

(3) The 1404 modules found in present-day *E. coli* K12 seem to derive from a minimal set of 352 ancestral modules. Our data suggest that each putative ancestral module was already at least as large as the alignment regions of today's modules and had a function similar to all members of its progeny. For the most part, members of a family are found to have retained a similar function and mode of action but to have specialized in substrate specificity. It seems likely that ancestral proteins were tolerant of substrate identity, then diverged to more specific derivatives.

(4) The physiological functions of proteins in *E. coli* are not equally represented in the small and large families of sequence-related proteins (Table 4). Most of the small groups of two or three are enzymes. Transport and regulator proteins follow in frequency. On the other hand, most of the large groups are composed of either transport systems or regulatory proteins, with a few groups containing enzymes and enzyme subunits. It seems that many transport and regulation proteins diversified over time by small changes while conserving sufficient primary sequence similarity as to be observable as sequence-related today. By contrast, many enzyme proteins have diverged further into small sets of highly specific catalytic agents having little or no primary sequence similarity to enzymes in other groups.

As to other types of proteins, strikingly, many proteins involved in the machinery of the information-processing systems (e.g. replication, transcription, translation) and cell structure (e.g.

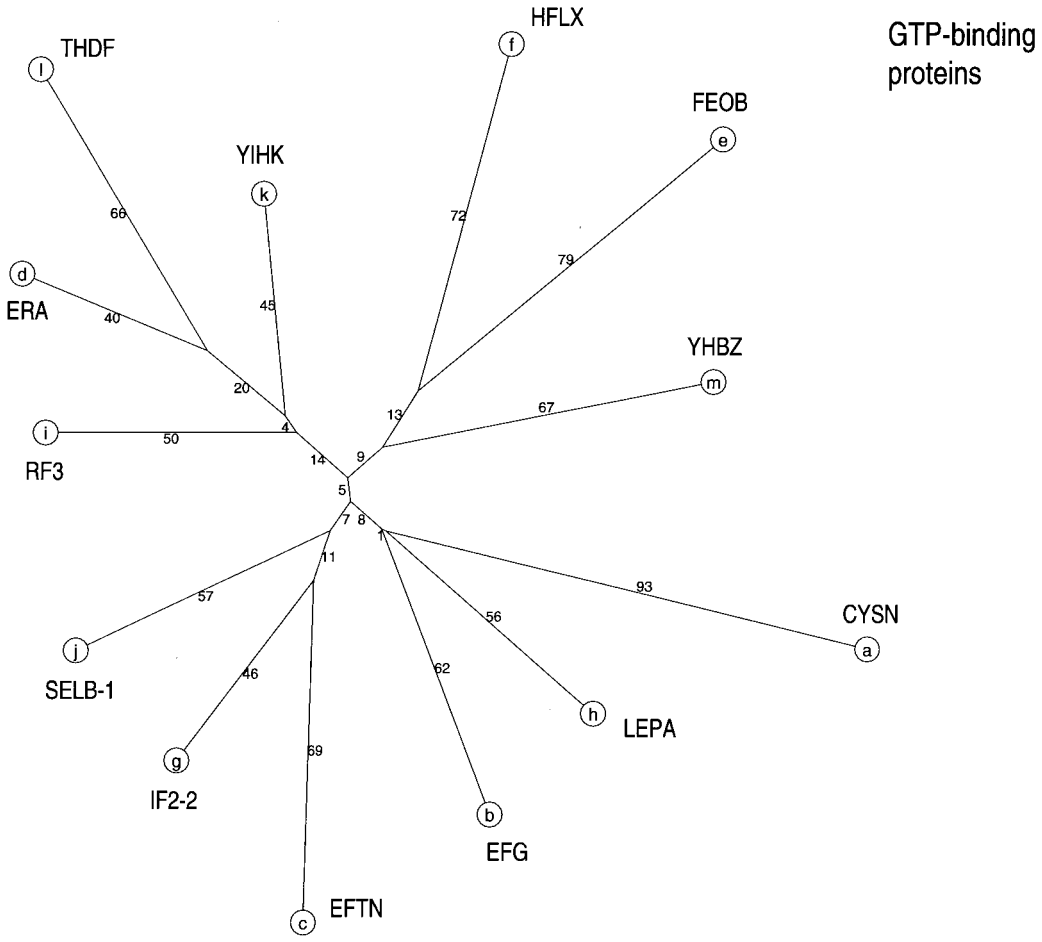

Figure 7. Geneological tree of a group in which most proteins are GTP-binding proteins. Derived by DARWIN function phylotree, with distances given in PAM units. Branches are labelled with the SwissProt mnemonic of each protein. Labels starting with letter Y are ORFs.

flagellae, ribosomal proteins) are either not part of this process of evolution by paralogy or have diverged so far that no hint of similarity of primary sequence remains. Therefore, we could not extend the conclusions based on sequence similarity to the history of these proteins.

(5) Some insight into mechanisms of evolution may be visible in the membership of the few families that contain one or a few members which are different in function from the majority. As an example, the group (Figure 6), composed largely of regulator proteins that recognize specific trigger molecules, also contains a few periplasmic transport proteins. An explanation might be that differentiation of a protein specifically recognizing ribose, for instance, gave rise to a ribose-specific repressor on one hand and a ribose-specific transport protein on the other hand (Mauzy & Hermodson, 1992).

It has not been possible to trace back the early history of each family of paralogous proteins. We have tried to reconstruct a genealogy of each module family (containing more than four members) by generating unrooted evolutionary trees. These trees show that the internal nodes of many trees (especially in the case of large families) are very poorly resolved, giving a star-like figure. Only events such as recent duplications could be easily detected in the tree topology (e.g. Figure 4). This implies that many of the primary events of the first duplications have become blurred and it will not be possible to trace back initial evolution events in the families using only this approach.

The bushyness of the protein families with most of the proteins lying about the same distance from the centre suggests that the proteins have evolved to about the same extent from their apparent origins. Although we cannot detect the origin *per se*, nevertheless it appears that the protein sequences have all arrived at a saturation point, unable to undergo more change within limits of preserving function (Meyer *et al.*, 1986). This aspect of the data will be treated more fully in a subsequent paper.

Traces of common ancestry may remain between families of proteins whose function is similar but primary structures have diverged so far that similarities in primary sequence are not detectable by the methods used in this paper. When efficient tools for comparison of proteins based on their ob-

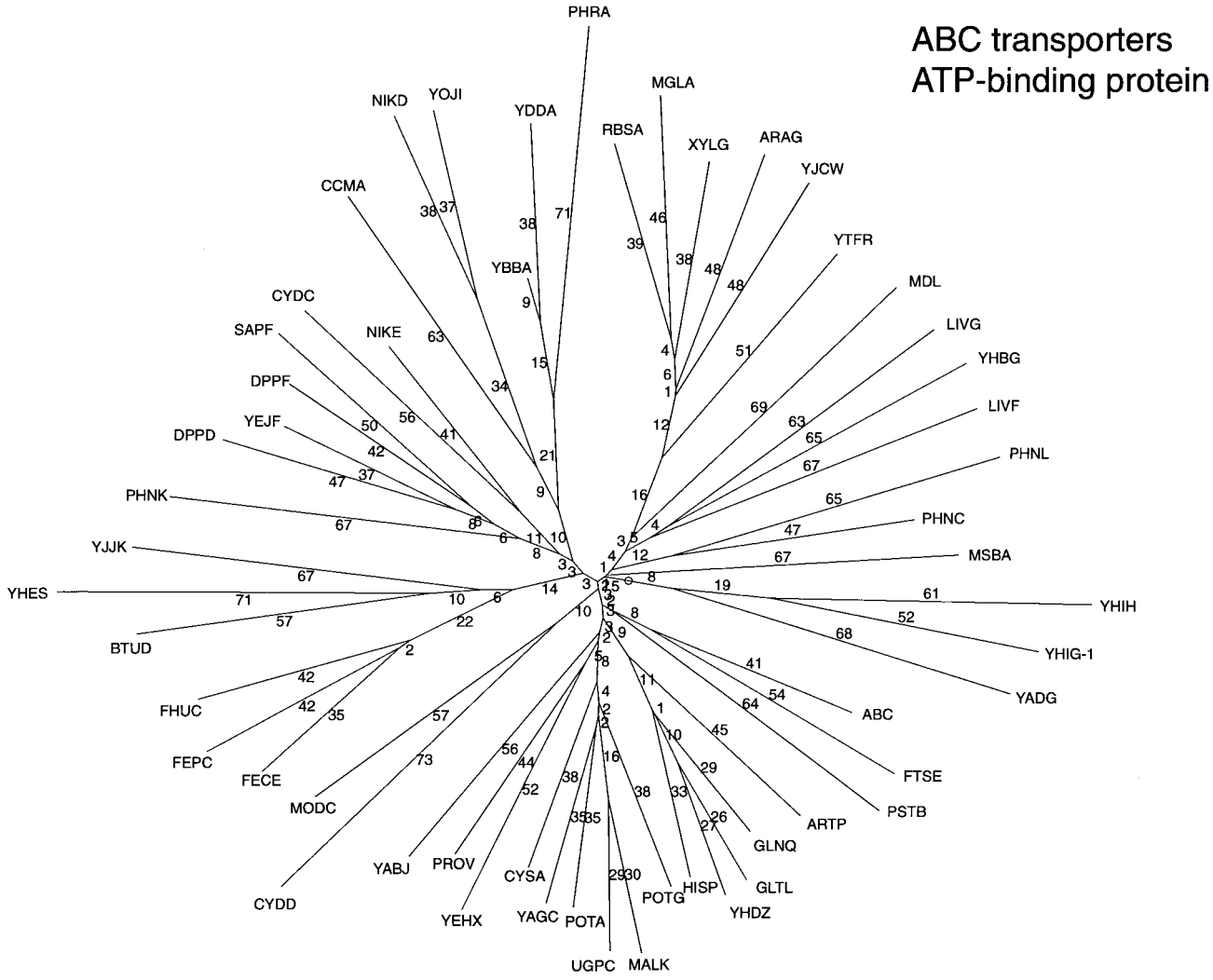

Figure 8. Geneological tree of a group in which most proteins are ATP-binding cassetes of ABC transporters. Derived by DARWIN function phylotree, with distances given in PAM units. Branches are labelled with the SwissProt mnemonic of each protein Labels starting with letter Y are ORFs.

served or predicted secondary and tertiary structures become available, we may be able to detect new levels of homology in protein structure which will help to detect more ancestral relationships. In this context, the recent work of ''mapping the protein universe'' (Holm & Sander, 1996) will be invaluable.

Materials and Methods

Finding all the significant matches among all the protein sequences of *E. coli*

All protein sequences of *E. coli* K12 were analysed for similarity of sequence with any other *E. coli* K12 protein. The sequences of the chromosomally encoded proteins of *E. coli* K12 in SwissProt release 33 (Bairoch & Apweiler, 1996) that are longer than 100 amino acid residues were assembled into a data set. This set was scanned with the AllAllDB program which allows one to search for all matches between all proteins of the SwissProt database using a complete Smith-Waterman type dynamic programming (Smith & Waterman, 1981). This AllAllDB program which was judged to be one of the best performers among sequence comparison programs (Johnson & Overington, 1993; Vogt *et al.*, 1995) is based on a landmark work that optimized the process of sequence alignments and exhaustively matched an entire protein sequence database (Gonnet *et al.*, 1992). This approach is based on the use of the unique algorithm that achieves optimal alignments of two homologous protein sequences (Needleman & Wunsch, 1970) and the development of a new mutation matrix adapted from the mutation data matrix developed initially by Dayhoff *et al.* (1978). The final outcome of such an exhaustive matching is a reorganized database which can be used as an interactive search system, the so-called DARWIN (Data Analysis and Retrieval With Indexed Nucleotide/Peptide Sequences) system available at the Computational Biochemistry Research Group server at the ETH, Zurich, Switzerland. In this system, the sequences, taken from the SwissProt database, are organized as sets of evolutionarily connected components which are characterized by an evolutionary distance measured in PAM units. The PAM distance, the number of accepted point mutations per 100 residues separating two sequences, is based on

(1) a mutation data matrix normalized to a distance of 250 PAM and recomputed for each new set of sequences, (2) a gap penalty which is itself dependent on the PAM distance intrinsic to the set of sequences studied (see Gonnet *et al*., 1992 for additional details of the method).

Forming the families of paralogous proteins and managing the data

The different matches were grouped into families of sequence-related pairs, assembled by collecting all relatives of each sequence of a pair and all relatives to those sequences and so on. Two equivalent programs, written in the Pascal computer language by Arnaud Bohelay and in SQL by David Space were used to gather automatically into one family all sequences that were related by a chain of similarities, collecting all relatives of both members of each pair until no further pairwise relationships were found. Data about pairs and families were analysed further in the context of relational database management using both Foxpro for Windows 2.6 and Claris FilemakerPro 3.03 for Macintosh. Data on function of gene products were derived from Riley & Labedan (1996) with some updates based on recent literature or information added to the SwissProt database.

Making genealogical trees for the families

For each family larger than three members, a multiple alignment and an unrooted tree were made using two functions (MulAlignment and PhyloTree) of the AllAll program, which is part of the DARWIN package. The trees obtained using this program are based on the estimated PAM distances between each pair of sequences and the deduced evolutionary distance between each node is weighted by computing the variance of the respective distance. Therefore, these distance trees are approximations to maximum likelihood trees (see Gonnet *et al*., 1992, for additional details of the method, and the booklet available at the Internet address http://cbrg.inf.ethz.ch/ServerBooklet, especially the subsection 2_3_5_1).

Acknowledgements

We thank Arnaud Bohelay, David Space, Lukas Knecht for programming assistance and Glenn Crossin for all-around assistance.

References

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol,* **219**, 555–565.

Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucl. Acids Res.* **24**, 21–25.

Boos, W. & Lucht, J. M. (1996). Periplasmic binding protein-dependent ABC transporter. In *Escherichia coli and Salmonella/Cellular and Molecular Biology* (Neidhardt, F. C. *et al*., eds), 2nd edition, pp. 1175–1209, ASM Press, Washington, D.C., USA.

Brenner, S. E., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene duplications in *H. influenzae*. *Nature,* **378**, 140.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H. P., Fraser, C. M., Smith, H. O., Woese, C. R. & Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* **273**, 1043–1045.

Clark, D. P. (1992). Evolution of bacterial alcohol metabolism. In *The Evolution of Metabolic Function* (Mortlock, R. P., ed.), pp. 105–114, CRC Press, Boca Raton, FL, USA.

Crawford, I. P. (1989). Evolution of a biosynthetic pathway: the tryptophan paradigm. *Annu. Rev. Microbiol,* **43**, 567–600.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, suppl. 3, pp. 345–352, National Biomedical Research Foundation, Washington, DC, USA.

Fitch, W. D. (1970). Distinguishing homologous from analogous proteins. *Systematic Zool.* **19**, 99–113.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, M, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrman, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496–512.

Goffeau, A., Park, J., Paulsen, J. T., Jonnlaux, J. L., Dinh, T., Mordant, P. & Saier, M. H. (1996). Yeast functional analysis reports. Multidrug-resistant transport proteins in yeast: Complete inventory and phylogenetic characterization of yeast open reading frames within the major facilitator superfamily. *Yeast,* **12**, in the press.

Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science,* **256**, 1443–1445.

Hoch, A. H. & Silhavy, T. J. (1995). *Two-component Signal Transduction,* ASM Press, Washington, DC, USA.

Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595–602.

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

Johnson, M. S. & Overington, J. P. (1993). A structural basis for sequence comparisons: An evaluation of scoring methodologies. *J. Mol. Biol.* **233**, 716–738.

Koonin, E. R., Tatsusov, T. & Rudd, K. E. (1995). Sequence similarity analysis of *Escherichia coli* proteins: Functional and evolutionary implications. *Proc. Natl Acad. Sci. USA,* **92**, 11921–11925.

Koonin, E. R., Tatsusov, T. & Rudd, K. E. (1996). *Escherichia coli* protein sequences: Functional and evol-

utionary implications. In *Escherichia coli and Salmonella/Cellular and Molecular Biology* (Neidhardt, F. C., ed.), 2nd edit., pp. 2203–2217, ASM Press, Washington, DC, USA.

Labedan, B. & Riley, M. (1995). Widespread protein sequence similarities: Origins of *E. coli* genes. *J. Bacteriol.* **177**, 1585–1588.

Labedan, B. & Riley, M. (1995b). Gene products of *Escherichia coli*: Sequence comparisons and common ancestries. *Mol. Biol. Evol.* **12**, 980–987.

Maloney, P. C. & Wilson, T. H. (1996). Ion-coupled transport and transporters. In *Escherichia coli and Salmonella/Cellular and Molecular Biology* (Neidhardt, F. C., ed.), 2nd edit., pp. 1130–1148, ASM Press, Washington, DC, USA.

Mauzy, C. A. & Hermodson, M. A. (1992). Structural homology between rbs repressor and ribose binding protein implies functional similarity. *Protein Sci.* **1**, 843–849.

Meyer, T. E., Cusanovich, M. A. & Kamen, M. D. (1986). Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proc. Natl Acad. Sci. USA,* **83**, 217–220.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Neidhardt, F. C., Curtiss III,, R., Lin, E. C. C., Ingraham, J., Low, K. B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. & Umbarger, H. E. (1996). *Escherichia coli and Salmonella/Cellular and Molecular Biology,* 2nd edit., ASM Press, Washington, DC, USA.

Ninfa, A. J. (1996). Regulation of gene transcription by extracellular stimuli. In *Escherichia coli and Salmonella/ Cellular and Molecular Biology* (Neidhardt, F. C., ed.), 2nd edit, pp. 1246–1262, ASM Press, Washington, DC, USA.

Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176**, 1–6.

Otsuka, J., Watanabe, H. & Mori, K. T. (1996). Evolution of transcriptional regulation system through promiscuous coupling of regulatory proteins with operons; suggestion from protein sequence similarities in *Escherichia coli*. *J. theor. Biol.* **178**, 183–204.

Riley, M. (1993). Functions of the gene products of *Escherichia coli*. *Microbiological Reviews,* **57**, 862–952.

Riley, M. & Labedan, B. (1996). *E. coli* gene products: Physiological functions and common ancestries. In *Escherichia coli and Salmonella/Cellular and Molecular Biology* (Neidhardt, F. C., ed.), 2nd edit., pp. 2118–2202, ASM Press, Washington, DC, USA.

Saier, M. H. (1996). Phylogenetic Approaches to the Identification and Characterization of Protein Families and Superfamilies. *Microbial. Comp. Genom.* **1**, 129–149.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the Twilight Zone revisited. *J. Mol. Biol.* **249**, 816–831.

Webb, E. C. (1992). *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes* (Webb, E. C., ed.), Academic Press, New York, USA.

Edited by J. H. Miller

(Received 9 December 1996; received in revised form 11 February 1997; accepted 12 February 1997)