Intrinsic and extrinsic approaches for detecting genes in a bacterial genome

Mark Borodovsky, Kenneth E. Rudd¹ and Eugene V. Koonin¹

School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230 and ¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received June 21, 1994; Revised and Accepted September 28, 1994

ABSTRACT

The unannotated regions of the *Escherichia coli* genome DNA sequence from the EcoSeq6 database, totaling 1,278 'intergenic' sequences of the combined length of 359,279 basepairs, were analyzed using computer-assisted methods with the aim of identifying putative unknown genes. The proposed strategy for finding new genes includes two key elements: i) prediction of expressed open reading frames (ORFs) using the GeneMark method based on Markov chain models for coding and non-coding regions of *Escherichia coli* DNA, and ii) search for protein sequence similarities using programs based on the BLAST algorithm and programs for motif identification. A total of 354 putative expressed ORFs were predicted by GeneMark. Using the BLASTX and TBLASTN programs, it was shown that 208 ORFs located in the unannotated regions of the *E. coli* chromosome are significantly similar to other protein sequences. Identification of 182 ORFs as probable genes was supported by both GeneMark and BLAST, comprising 51.4% of the GeneMark 'hits' and 87.5% of the BLAST 'hits'. 73 putative new genes, comprising 20.6% of the GeneMark predictions, belong to ancient conserved protein families that include both eubacterial and eukaryotic members. This value is close to the overall proportion of highly conserved sequences among eubacterial proteins, indicating that the majority of the putative expressed ORFs that are predicted by GeneMark, but have no significant BLAST hits, nevertheless are likely to be real genes. The majority of the putative genes identified by BLAST search have been described since the release of the EcoSeq6 database, but about 70 genes have not been detected so far. Among these new identifications are genes encoding proteins with a variety of predicted functions including dehydrogenases, kinases, several other metabolic enzymes, ATPases, rRNA methyltransferases, membrane proteins, and different types of regulatory proteins.

INTRODUCTION

The fraction of the *Escherichia coli* genome available as nucleotide sequence is now over 60% and the completion of the chromosome sequence is expected within the next two to three years (refs. 1,2 and K.E.R., unpublished). The importance of a comprehensive description of the information content of this genome is obvious. Generally, genetic information is tightly packed in bacterial chromosomes. Nonetheless, a significant proportion of the *E. coli* genome sequences is in unannotated regions containing no recognized genes. In the EcoSeq6 database (2,3) such regions comprise 359,279 basepairs (bp) of the total of $1,975 \times 10^6$ (19%). In an *E. coli* genome project, the definition of a 'gray hole', a large (at least several hundred bp) region of the chromosome devoid of known genes or long open reading frames has been introduced (4). With the goal of a complete description of the *E. coli* gene repertoire in mind, it is important to find out which of the unannotated sequences contain no genes and which actually include genes that have so far escaped detection. There may be several reasons for the failure to identify some genes in original sequencing studies, including presence of only partial open reading frames (ORFs) in the determined sequences, small size of some genes, frameshifts and other sequencing errors, and utilization of unusual initiation codons.

Computer-aided approaches to gene identification may be divided into 'intrinsic' and 'extrinsic'. The intrinsic approach includes evaluation of certain properties of the DNA sequence without explicit referral to other sequences. The relevant properties include, obviously, the length of an ORF, codon usage, the presence or absence of the ribosome-binding (Shine–Dalgarno) sequence at the appropriate distance upstream of the initiation codon, and various more subtle statistical characteristics that are believed to be typical of expressed genes as opposed to non-coding regions. Such approaches to bacterial gene prediction have been proposed already at the onset of DNA sequencing (5–13). Although there has been no optimal solution commonly recognized as sufficient for the practical purposes of gene identification, the problem of finding bacterial genes lately has been overshadowed by the problem of finding exon boundaries in eukaryotic genes interrupted by introns. To this end, several sophisticated approaches using neural networks and

dynamic programming have been suggested (14–22). Nonetheless, a reliable method for predicting protein coding capacity of DNA sequences regardless of the exon/intron organization remains a very important goal. A recent comparison of the accuracy of a number of such methods (23) has demonstrated the top performance of measures using hexamer-in-phase counts.

The GeneMark method is based on oligonucleotide-in-phase statistics with no restriction on the oligonucleotide size (24,25). Phased (non-homogeneous) Markov models are used in GeneMark for protein-coding sequences and for non-coding sequences complementary to coding regions. Ordinary (homogeneous) Markov models are used for those non-coding sequences whose complements do not code for a protein either. Parameters of the non-homogeneous Markov models are derived from oligonucleotide-in-phase counts in a training set of coding sequences, whereas the parameters of homogeneous models are defined using total oligonucleotide counts in a non-coding training set (24,25). The larger the training set, the longer the oligonucleotides for which statistics can be reliably derived and accordingly, the higher the order of the Markov model that can be used in the algorithm. The currently available *E. coli* sequence data allow one to compute the parameters for Markov models up to the fifth order, reflecting hexamer statistics for each of the six possible phases (frames). The gene prediction by GeneMark is characterized by a high level of accuracy; recently, this method has been used to improve the quality of gene prediction in one of the *E. coli* genome projects (4,26–28).

The extrinsic approach to gene prediction includes a comparison of the putative encoded amino acid sequence with protein sequence databases and a search for functional motifs. If a putative ORF product shows significant sequence similarity to one or more proteins in the database, it is almost certain that the ORF in question is a real gene.

Combining the extrinsic and intrinsic approaches has the potential of enhancing the reliability of the results obtained with each of them and is important for extracting maximum information from genomic sequences.

Here we present the results of our analysis of the complete set of unannotated sequences that are located between the annotated genes in the EcoSeq6 database using GeneMark for predicting putative coding regions and BLAST as well as motif search methods for detecting sequence similarities to proteins in the current databases. The combination of these two approaches is featured as a strategy for identifying new genes in bacterial chromosomes.

METHODS

Prediction of expressed ORFs using GeneMark

The GeneMark method (24,25) was used to predict potential expressed ORFs in unannotated DNA sequences from the EcoSeq6 database (2). Let S be a nucleotide sequence of length n ($s_1, s_2, \dots, s_n$); let n be a multiple of 3. GeneMark computes the probabilities (P_i , $i = 1, 2, \dots, 7$) of seven mutually exclusive states of S , namely:

- S is a protein-coding region, with s_1 being the first, second or third position of a codon (states 1 through 3);
- S does not code for a protein but its complement codes for a protein, with s_n being the third, second or first position of a triplet complementary to an actual codon (states 4 through 6);

— neither S nor its complement codes for a protein (state 7).

Computing of P_i includes two steps. First, auxiliary probabilities F_i ($i = 1, 2, \dots, 7$) are computed using Markov models. F_i is the probability of finding S in an infinite sequence that is in the state i . Coding regions and non-coding regions that are complementary to coding regions are described by phased Markov models; sequences that do not code for proteins and are not complementary to protein-coding sequences are described by an ordinary Markov model. A phased (non-homogeneous) Markov model incorporates the weak three-base periodicity that is observed in coding sequences or in non-coding sequence which are complementary to real genes (11,24,25,29,30). In particular, the phased Markov model of a coding sequence reflects the differences in the oligonucleotide statistics when the oligonucleotides are counted separately for each of the three possible reading frames (in-phase-oligonucleotides).

Parameters of Markov models (transition probabilities) are estimated based on oligonucleotide statistics in training sets. The currently available *E. coli* sequence data allow one to derive parameters for ordinary or phased Markov models up to the fifth order.

For the first order Markov model, we have:

$$F_1 = F_0^1(s_1) * F^1(s_2|s_1) * F^2(s_3|s_2) * F^3(s_4|s_3) * \dots * F^2(s_n|s_{n-1}), \quad (1)$$

Here, F^1 , F^2 and F^3 are the respective transition (conditional) probabilities of nucleotide s_j succeeding nucleotide s_k determined for the three codon positions in phase 1 (these probabilities reflect 'in-phase' dinucleotides frequencies) and $F_0^1(s_1)$ is the initial probability of nucleotide s_1 . The F^i values are defined from the training set of protein-coding sequences. Probabilities F_2 and F_3 are determined by a similar equation, with the superscript index permutation.

The probabilities of states 4 to 6, in which S is non-coding but is the complement of a protein-coding sequence, are computed similarly. For instance:

$$F_4 = F_0^4(s_1) * F^4(s_2|s_1) * F^5(s_3|s_2) * F^6(s_4|s_3) * \dots * F^5(s_n|s_{n-1}) \quad (2)$$

Figure 1. Distribution of GeneMark probability value for the set of known genes and for the set of ORFs in intergenic regions of EcoSeq6. The genes and intergenic regions were defined as annotated in EcoSeq6 and the scores were computed for 1,305 genes and 3,272 ORFs larger than 100 nt in the unannotated regions. The ORFs in different phases could overlap. The computation was made using the Class I gene model. The open rectangles show the prediction for known genes and the filled rectangles show the prediction for ORFs in the unannotated regions.

which includes transition probabilities determined for the three position in triplets complementary to actual codons. $F_0^N(s_1)$ is the initial probability of a nucleotide s_1 in non-coding sequence complementary to a real gene. The F^i values are defined from the training set of non-coding sequences complementary to real protein-coding regions. Probabilities F_5 and F_6 are again determined by equations similar to (2) with superscript index permutation.

Finally, the probability that S is non-coding and is not a complement of a protein-coding sequence is given by a formula using an ordinary Markov model:

$$F_7 = F_0^N(s_1) * F^N(s_2|s_1) * F^N(s_3|s_2) * F^N(s_4|s_3) * \dots * F^N(s_n|s_{n-1}) \quad (3)$$

Here, the transition probabilities F^N do not depend on a nucleotide position in a sequence and are estimated from the dinucleotide frequency statistics computed from a set of DNA sequences which do not code for a protein and are not complementary to coding sequences.

At the second step of the GeneMark calculation, F_i values are used to compute P_i ($i = 1, 2, \dots, 7$), i.e. probabilities that S actually is in one of the seven possible states. P_i are determined using the Bayes formulae:

$$P_i = F_i * R_i / \left(\sum_{j=1}^7 F_j * R_j \right) \quad (4)$$

R_i ($i = 1, 2, \dots, 7$) are *a priori* probabilities of each of the seven states. We assume $R_7 = 1/2$ (probability of an arbitrary sequence being non-coding) and $R_i = 1/12$ for $i = 1, 2, \dots, 6$.

Generalization of equations (1)–(4) to higher order Markov models is straightforward.

For each of the explored intergenic regions, the GeneMark program determines the set of all possible ORFs. For each of them, the average value of the probability function P_i (hereinafter GeneMark score) is computed using a sliding window of 96 nucleotides and a step of 12 nucleotides; the index i ($i = 1, 2, \dots, 6$) corresponds to the phase of the given ORF. If the GeneMark score is higher than the chosen cut-off, the ORF is included in the list of predicted genes. We found that the score of 0.4 is the most useful cut-off for GeneMark prediction (see below).

We assume that only one of these six possible reading frames actually encodes a protein. This assumption is important as a real

gene may 'cast a shadow' on the complementary strand, giving it some statistical characteristics of a gene. Specifically, as the reverse complement of the three stop codons are rarely used codons, a shadow phase (e.g. phase 4 if the actual gene is in phase 1) frequently may contain relatively long ORFs. In addition, the reverse complement of the common RNY codon pattern is also RNY (see ref. 8 for details). However, the shadow frame has a lower probability of being a coding region than the frame that casts the shadow and GeneMark consistently identifies the most likely frame.

GeneMark is well suited to detect putative frameshift errors in DNA sequences — such an error should be suspected each time when an ORF prediction switches from one phase to another in the same DNA strand instead of reaching a termination codon in the original phase.

The current version of GeneMark takes into account the nonhomogeneity of the *E. coli* gene set. Using Factorial Correspondence Analysis, *E. coli* genes have been partitioned into three classes differing by their codon usage (31,32). Class I includes genes that typically are expressed at a low or moderate level; class II includes genes that are constitutively expressed at a high level; and class III includes genes that apparently have been acquired by horizontal gene transfer ('foreign' genes). Generation of separate training sets for Class I and Class III genes resulted in a significant improvement in the performance of GeneMark. When the set of *E. coli* genes was divided into classes, the Class I model produced GeneMark scores greater than 0.4 for 99.9% of the Class I genes, 99.1% of the Class II genes, and only 55.4% of the Class III genes. With the Class III model, scores greater than 0.4 were obtained for 97.6% of the Class I genes, 97.3% of the Class II genes, and 88.1% of Class III genes (M. Borodovsky, J. McIninch, C. Medigue and A. Danchin, unpublished). Even though a high accuracy of Class III gene prediction could not be achieved with any of the models, presumably because of the heterogeneity of this gene set, it is obvious that for reliable analysis of an uncharacterized *E. coli* sequence, GeneMark should be run twice using the Class I and Class III gene models, respectively.

Database searches

Daily updated nucleotide and protein sequence databases were used as implemented at the National Center for Biotechnology

Table 1. Relationships between the prediction of coding regions using GeneMark and detection of sequence similarity using BLAST for the unannotated sequences in the EcoSeq6 database

	number	percent ^a
total unannotated sequence segments in EcoSeq6	1,278	NA
putative genes predicted by GeneMark including	354	100
class I+II ('native')	307	87
class III ('foreign')	51	13
corroborated by BLAST including	182	51
class I+II ('native')	155	44
class III ('foreign')	29	7
not corroborated by BLAST including	172	49
class I+II ('native')	155	43
class III ('foreign')	19	5
similarities detected by BLASTX and TBLASTN including	208	100
corroborated by GeneMark prediction	182	88
not corroborated by GeneMark prediction	26	13

^arounded to integer

Table 2. New genes in unannotated regions of the *E. coli* genome^a

REGION/ NEW SWISSPROT ENTRY ^b	UAR/ORF SIZE AND TYPE OF SEGMENT ^c	BEST HIT AND P VALUE ^d	CONSERVED MOTIF	CROSS- PHYLUM HITS ^e	Gene-Mark PREDIC- TION ^f	PREDICTED FUNCTION/ ACTIVITY	end_slt M69185 Yjkk'	N(fs) 80	TLCR_STRFR; 8.3e-11	NTP-binding A motif	ABE	F	membrane transport ATPase
aroC_end M27714 YfcB'	C 152	VIBAROC (V. anguillarum); 6.1e-74	adenine- specific DNA methylase motif	B	N	adenine- specific DNA methylase	cspA_end M30139 YiaZ	complete 50	PND2_ECOLI; 3.1e-10		none	F	membrane "killer" protein (Hok/Gef family)
end_yhaA X14430 YhaA	C (ext) 196/331	ACKA_ECOLI; 7.8e-46		B	N	acetate kinase	ald_end M64541 GspC'	C 29	GB:CPGAP_1; 9.6e-9		BE	-	dehydrogenase
phoR_end X04704 BrnQ	N 70	BRNQ_SALTY; 6.6e-44		none	N	amino acid membrane transporter	pepN_end M15273 YcbE'	C 95	GB:SPCCNRT_4 (Synechococcus); 2.2e-8	UvrA superfamily motif IV ^g	BE	N	transport ATPase
end_pspA X57560 YcjB'	N 110	NIFA_HERSE; 6.1e-41	ATP-binding "A" and "B" motifs	B	N	transcription regulator; ATPase	menD_end M21787 YfbA'	C(fs) 215	ENTC_ECOLI; 1.0e-7		AB	N	isochorismate synthase
cysP_end M32101 YfeF'	C 84	PIR:A38121 (S.t.); 5.3e-38		BE	N	dehydrogenase	end_glpR M96795 YhgB'	C(fs) 195	NIFA_KLEPN; 1.2e-7		B	N	transcription regulator; ATPase
glpC_end M20938 YfaD'	N 69	YHGA_ECOLI; 3.5e-35		none	N	unknown	ydiB_aroD X04306 YdiB	C(ext; fs) 172/223	DHQA_EMENI; 3.2e-7		E	F	dehydrogenase
sdia_end X03691 YecC'	C 75	PIR:G42600 (A. tum); 3.0e-27	UvrA superfamily motif IV ^g	BE	N	membrane transport ATPase	aroH_end M38266 YdiE'	C 63	PIR:S28041 (Y.ent.); 3.7e-7	UvrA superfamily motif IV ^g	BE	N	membrane transport ATPase
end_ung D13169 YfdD'	N 127	PIR:S22666 (S. liquefaciens); 3.2e-26		none	N	unknown	end_fesB M63308 YbaG'	C 52	GB: BACNRGABO_1 (B.s.); 1.3e-6		B	N	membrane protein
end_purK M19657 YbcI'	C 95	ARCC_PSEAE; 7.6e-25		none	N	carbamate kinase	aspS_end X53863 YecD	complete 199	ENTB_ECOLI; 2.3e-6		B	N	iso- chorismatase
araE_end J03732 YgeC'	C 38	PIR:S17771 (E. chrys); 1.6e-22		B	-	dehydrogenase	ndrB_inaA K02672 YfaE	complete(fs) 84	GB: YERFBIIAH_4 (Y.pseud); 1.0e-5		BE	N	ferredoxin
clpB_end M29364 YfhH	C (fs) 141/243 ^b	PIR:S26602 (P. aer.); 2.3e-20		B	N	unknown	end_prs X53983 M13174 YchM'	N 105	GB:SOYAKI70_1 (soybean early nodulin); 1.1e-5		E	N	membrane protein
end_yhfA X61919 YhfF'	C 62	KPPP_ALCEU; 7.2e-20		BE	N	sugar kinase	end_fimZ X51662 YbcG'	C 53	PIR:B28393 (S.t.); 3.2e-5		none	-	fimbrial protein
end_ppsA M69116 YdiD'	C (fs) 193	ENTE_ECOLI; 3.2e-16		BE	N	acyl-CoA- synthetase	end_fucO X15025 YgdG	C 59/251 ^k	DPO1_THEFL; 3.3e-5		B	N	unknown
end_maiQ M32793 YhgC'	C 57	GNTP_BACSU; 2.3e-17		B	F	permease	end_rcsB M28242 YojN'	C 129	GB:CCRDIVIX_1 (Caul. cres); 3.4e- 5		B	N	histidine kinase
gabT_end M88334 YgaE'	N 114	PIR:A41858 (P.a.); 8.2e-17	HTH motif	B	N	transcription regulator	grpE_end X07863 YijB	complete 136	U00021_19 (M. leprae); 6.3e-5		BE	N	unknown
end_lig M24278 YfeH'	C 255	PIR:S15237 (P.a.); 2.3e-16		E	N	membrane protein	ndB_ygC M16194 YgiB	complete(fs) 292	GB: AELACDEH_1 (Alc. eutr.); 7.4e-5		none	N	membrane protein
dcp_end M96235 YdeI'	N (fs) 172	PIR:S07135 (K.p.); 1.3e-15	NAD-binding motif	BE	N	dehydrogenase	yfaB_end Y00544 YfaF'	C 67	PIR:S22680 (E.c. adhesin); 9.3e-5		none	N	fimbrial protein
end_ruvB M21298 YdfG' ^h	C 182	PIR:PS0372 (C.paradoxa); 1.7e-15		B	N	unknown	end_gapA X02662 YeaA'	N 32	YCD3_YEAST; 1.4e-4		E	N	PilB-related transcription regulator
ribB_end M64472 YgiL'	N 78	GB:PMMRPA_1 (P.m.); 5.9e-14		none	F	pilin	yfbB_end X72336 YfhH'	N 91	YNGB_CLOPE; 1.8e-4		B	N	unknown
end_gltA J01619 YbgD	complete 188	FMFA_ECOLI; 8.1e-13		none	F	pilin	glgC_glgX J01616 GlgX	C (ext); 84/591	ISOA_PSEAY 1.8e-4		BE	N	amylase
end_phnQ D90227 YjcA'	C 68	PIR:P26423 (S.mut.); 2.1e-12		B	N	galactose-6- phosphate isomerase subunit	aldH_end M38433 YcjA'	N 251	GLPA_ECOLI; 2.2e-4	NAD-binding motif	BE	N	dehydrogenase
end_potE M64495 YbfI'	N(stop codon) 46/65'	KDPE_ECOLI; 7.5e-12		none	-	transcription regulator	cylB_end J03939 YbgE'	N 43	YCCB_ECOLI; 6.0e-4		none	-	putative membrane protein
end_pheP M58000 YbdE'	C(fs) 177	CZCA_ALCEU; 3.3e-11		none	N	permease	alkA_end K02498 YegD'	N 146	GR78_KLULA; 7.1e-4	HS70 phosphate- binding motif 1 ^l	BE	N	HS70 family chaperonin
yggF_end X14436 YggK'	C 40	GLPX_ECOLI; 3.6e-11		B	-	unknown	end_alkA K02498 YegE'	C 141	PIR:S18623 (A. caulinodans) 1.2e-3		none	N	unknown

Table 2. (cont.)

<i>end_pyrD</i> X02826 YcbF ^a	C	72	CS31_ECOLI; 2.6e-3	none	F	pili biogenesis factor	
<i>yjB_end</i> M26934 YdJE ^a	C	93	YABE_ECOLI; 4.0e-3	E	N	membrane protein	
<i>deaD_pmp</i> M63288 YhbM	complete	59	YC19_MARPO; 4.0e-3	TPR motif	N	unknown	
<i>end_endA</i> X65169 YggI ^a	C(fs)	96	PIR:A48830 (mouse); 5.0e-3	Zn finger motif	E	N	DNA-binding protein
<i>glgC_glgX(2)</i> J01616 YhgD	complete	114	GB:HUMBILGLY D_1; 8.6e-3		E		membrane protein
<i>trxA_end</i> X15868 YcgC ^a	C	57	PT13_RHOCA; 1.4e-2	PEP-binding motif	BE	N	PEP-utilizing enzyme
<i>end_rndA</i> M16194 YgiH	C	34/236 ^m	YAF_C_ECOLI; 1.6e-2	HTH motif	B	N	transcription regulator
<i>ung_end</i> J03725 YfiF	C	86(fs)/345 ^a	SPOU_ECOLI; 3.9e-2	SAM-binding motif	none	N	rRNA methylase ^a
<i>yggC_yggD</i> X14436 YggC	N (ext)	110/207	GB: MXABSGA_1; 6.3e-2	ATP-binding "A" motif	E	N	kinase
<i>dniR_end</i> D12650 YafG	complete	135	VP07_BPPRD; 9.3e-2	trans- glycosylase motif	none	F	lytic trans- glycosylase ^a
<i>appA_end</i> M58708 YccC ^a	C	99	GB: STRCPASBD_3; 9.4e-2	ATP-binding "B" motif	AEB	N	chromosome partitioning ATPase
<i>end_cysH</i> Y07525 YgcB ^a	N	335	GB: NGOPIFLX_1; 9.4e-2	ATP-binding "A" motif	B	F	ATPase (pilin assembly ?)

^aIn addition, 14 putative new genes that did not show significant similarity to any proteins in the current databases in the initial BLASTX searches, were shown to have homologues by screening the conceptual translation of the nucleotide sequence database using the TBLASTN program. As all of these homologues were uncharacterized and mostly incomplete putative proteins, these ORFs are not included in the table.

^bend_[genename] indicates unannotated regions between the 5'-terminus of a sequenced piece of chromosomal DNA and the beginning of the respective known gene; [genename]_end indicates regions between the end of the respective gene and the 3'-end of the sequence; and [genename]_[genename] indicates intergenic regions. The GenBank Accession Number is indicated for each entry.

^cUAR, unannotated region; the position in the putative new protein (N-terminal, C-terminal or complete) and number of residues are indicated for amino acid sequence; ext indicates preexisting ORFs that have been extended, with numerator showing the length of the extension and the denominator show the new total length; the cases of apparent frameshift (fs) are noticed.

^dThe locus ID from the SwissProt, PIR or GenBank database is indicated for each putative new gene product along with the probability of finding the observed score by chance (*P* values) calculated using the BLAST algorithm (33,35,55,56). The identified new gene products are ranked in the order of decreasing similarity (increasing *P* values). Organism name abbreviations: H.inf — *Haemophilus influenzae*; S.t. — *Salmonella typhimurium*; A. tum. — *Agrobacterium tumefaciens*; S. liquefaciens — *Serratia liquefaciens*; E.chrys. — *Erwinia chrysanthemum*; P.a. — *Pseudomonas aeruginosa*; K. p. — *Klebsiella pneumoniae*; C.par. — *Cyanophora paradoxa*; S. dys. — *Shigella dysenteriae*; P.m. — *Proteus mirabilis*; S.mut. — *Streptococcus mutans*; A. caulinodans — *Azorhizobium caulinodans*; Str. coel. — *Streptomyces coelicolor*; Y.ent. — *Yersinia enterocolitica*; B.s. — *Bacillus subtilis*; Y.pseud. — *Yersinia pseudotuberculosis*; Caul. cres. — *Caulobacter crescentus*; Alc. eutr. — *Alcaligenes eutrophus*; R.com. — *Ricinia communis*.

^eCross-phylum hits are similarities with protein sequences from phylogenetically distant organisms (42,44). A, archaea; B, distantly related bacteria; E, eukaryotes.

^fN indicates prediction of a 'native' gene (Class I or Class II), F indicates prediction of a 'foreign' gene (Class III) and dash indicates that the gene has not been predicted.

^gMotifs from ref. 57.

^hAfter this analysis has been finished, the complete coding sequence for this protein has been submitted to GenBank (ref. 58; Accession Number X57620) although the *yfiH* ORF has not been identified. The numerator shows the length of *yfiH* in our original analysis and the denominator shows the new total length.

ⁱDescribed in ref. 59, but not annotated as an ORF in current databases.

^jAfter this analysis has been finished, the complete coding sequence for this protein has been submitted to GenBank (Accession Number U08369) although the *ybfI* ORF has not been identified.

^kAfter this analysis has been finished, the complete coding sequence for this protein has been submitted to GenBank (ref. 60; Accession Number L07763) although the *ygdG* ORF has not been identified.

^lSee ref. 53.

^mAfter this analysis has been finished, the complete coding sequence for this protein has been reported; the ORF corresponding to *ygiH* has been described but has not been annotated in the database (ref. 61; GenBank Accession Number L12966).

ⁿAfter this analysis has been finished, the complete coding sequence for this protein has been reported although the *yfiF* ORF has not been identified (ref. 62; GenBank Accession Number D13169).

^oSee ref. 63.

^pSee ref. 64.

Information (NCBI, NIH). A 'non-redundant' nucleotide database merges non-identical entries from GenBank and EMBL databases; and the 'non-redundant' protein sequence database is a merger of non-identical sequences from PIR and SWISS-PROT as well as additional protein sequences annotated in the feature tables of the GenBank and EMBL databases.

All the sequences without annotated protein-coding regions were extracted from the EcoSeq6 database and compared with the non-redundant protein sequence database using the BLASTX program based on the BLAST algorithm (33–35). BLASTX compares the conceptual translation of a nucleotide sequence in all six possible reading frames with a protein database. The searches were performed using the BLOSUM62 matrix for amino acid residue comparison that recently has been shown to ensure high sensitivity in the detection of subtle similarities between amino acid sequences (36,37). The SEG program was used to filter the translated query sequences to remove low complexity (compositionally biased) segments that produce erratic results in database searches (35,38). BLASTX is instrumental in detecting putative frameshift errors; a frameshift is predicted whenever the similarity to a single protein sequence switches from one frame to another in the same strand of the query DNA sequence (34).

Two *a priori* cut-off values were chosen in order to assess the significance of the observed sequence similarities. The BLAST hits with the probability of occurrence by chance (*P*) below 10^{-4} were considered to be indicative of homology. Obviously, even with the complete set of the *E. coli* proteins (about 4,500), it is unlikely that more than one BLAST hit with such a *P* value would turn out to be a false positive. However, as the sensitivity of the BLAST search is limited, with multiple alignment methods being much more sensitive, all hits with *P* values between 10^{-1} and 10^{-4} were subjected to further scrutiny. To this end, multiple sequence alignments were generated using the MACAW program (39).

Those ORFs that have been predicted to be genes by GeneMark, but were not supported by the initial database screening using BLASTX, were subjected to an additional database search using the TBLASTN program that compares an amino acid sequence with the conceptual translation of a DNA database in all six frames (35).

Database searches with the BLAST programs were combined with search for sequence patterns by using the BLA program that detects patterns that are conserved in BLAST outputs (40). The motifs were from the PROSITE library (41), but several motifs with low information content (e.g. phosphorylation and glycosylation sites) were omitted from the search whereas several new motifs were added to the library.

Figure 2. ORF *ychM'* encodes the N-terminal portion of a putative membrane protein related to a soybean nodulin and a human tumor-specific protein. (a) GeneMark prediction for the *E. coli* DNA sequence around the *prs* gene. Each panel shows the *a posteriori* probability function for the given sequence to be a coding region under the fifth order non-homogeneous Markov model for *E. coli* genes of class I plotted separately for each of the six reading frames. The upward ticks on the midline corresponding to the 0.5 probability indicate possible initiation codons (the larger ticks show ATG and the smaller ticks show GTG) and the downward ticks show stop codons. Rectangles designate intervals between adjacent stop codons that contain predicted coding regions. The predicted coding region for YchM' is in the third panel from the bottom (frame -1) and the *prs* gene is in the lower panel (frame -2). (b) Multiple alignment of the YchM' protein sequence with plant and human membrane proteins. The distances from the protein ends are indicated. The consensus shows amino acid residues conserved in all three aligned sequences. U indicates a bulky aliphatic residue (I, L, V or M), & indicates a bulky hydrophobic residue (I, L, V, M, F, Y or W) and dot indicates any residue. The three predicted transmembrane segments are indicated. Identities between YchM' and each of the two eukaryotic proteins are shown by asterisks, and similar amino acid residues are shown by colons. GenBank Accession Numbers are indicated for the nucleotide sequences from which the aligned protein sequences have been derived.

RESULTS AND DISCUSSION

Analysis of the unannotated regions in the EcoSe6 database using GeneMark

GeneMark was run for the set of genes annotated in EcoSeq6 and for the set of ORFs longer than 100 nt found in the unannotated regions (see Methods). The resulting distribution of GeneMark scores is shown in Figure 1. These data show that with the Class I gene model, 93.7% of the known coding regions have scores greater than 0.4 whereas 88.7% of unannotated ORFs have scores below 0.4. Some of the known genes belong to Class III and some are short fragments for which a reliable prediction could not be obtained. Both of these groups contribute to the observed false negative rate of about 6.3%. Therefore the score of 0.4 appears to be a reasonable cut-off for GeneMark prediction. However, the relatively high apparent false positive rate of about 11.3% and especially the fact that a number of ORFs in the unannotated regions had a score greater than 0.8, well into the score range typical of known genes (Fig. 1), suggested that some of these ORFs are likely to be undetected genes. Characterization of these putative new genes is the goal of this study.

Matching GeneMark predictions with sequence similarities

The principal part of our analysis included matching the predictions of expressed ORFs by GeneMark with the results of database searches for amino acid sequence similarity using BLAST. The EcoSeq6 database (2) contains 1,278 unannotated regions with a total length of 359,279 basepairs located between the annotated genes. In these sequences, GeneMark identified 354 ORFs as putative genes. The comparison of GeneMark predictions with BLAST search results for this set of DNA sequences is summarized in Table 1. Almost 90 percent of the BLAST hits were supported by GeneMark prediction. Among the 26 BLAST hits that had no counterpart among the putative genes predicted by GeneMark, 17 were small portions of ORFs no longer than 30 codons. For sequences of this size, the GeneMark prediction could not be expected to be reliable (24,25).

Over one half of the GeneMark predictions were supported by sequence similarity analysis. This may be an underestimate of the actual number of gene products with similarity to proteins in the databases as many of the coding sequences found in the unannotated regions are 5'- or 3'-terminal portions of genes that typically are less conserved than internal regions.

An important question is what fraction of the predictions made by GeneMark alone correspond to real genes. The following evidence, though indirect, suggests that this fraction is high. The set of putative expressed ORFs detected by GeneMark resembles the set of known *E. coli* genes in more than one respect. First, the fraction of 'foreign' genes (about 13%) was close to the previous estimate of about 11.5% obtained with a set of 782 *E. coli* genes (Table 1; ref. 31). Second, the fraction of putative new proteins with homologues in sequence databases (which may actually reach about 60%, taking into account the lower search sensitivity with incomplete sequences) is close to the recent estimates of such genes in *E. coli* and yeast (42,43). Furthermore, 72 putative new gene products (20% of the GeneMark predictions) belong to ancient protein families that include both eubacterial and eukaryotic members. This is nearly identical to the overall proportion of such highly conserved sequences among eubacterial proteins in the SWISS-PROT database (44). Thus many of the putative expressed ORFs that are predicted by GeneMark, but show no significant sequence similarity to proteins in the current databases, nevertheless are likely to be real genes.

The number of genes in unannotated regions from EcoSeq6 may be only slightly smaller than the sum of the expressed ORFs predicted by GeneMark and those BLAST hits that have not been supported by GeneMark, totaling up to 380. EcoSeq6 contains 1,440 recognized genes (2). Thus a substantial proportion of *E. coli* genes are actually present in EcoSeq6, at least as partial sequences, but were not explicitly identified at the time when this database was being compiled.

Among the 206 putative genes for which our BLAST search revealed significant similarities, over one half have been described in the literature since the release of EcoSeq6 (mostly as 'uncharacterized ORFs'), but about 70 are, to our knowledge, unreported. Table 2 summarizes the information on these new genes in the unannotated sequences from EcoSeq6. We used the systematic nomenclature for *E. coli* ORFs (2), with each ORF name starting with the letter 'y' and the next two letters indicating the approximate position on the genome. Incomplete ORFs are indicated by a prime.

New functional identifications

The set of predicted new genes included those coding for enzymes of intermediate metabolism, different types of ATPases, dehydrogenases, membrane proteins (putative permeases), methyltransferases, and different types of regulatory proteins

Figure 3. The putative protein YgaE' is related to helix-turn-helix transcription regulators. The predicted helix-turn-helix motif is shown. Psp, *Pseudomonas* sp.; Bs, *Bacillus subtilis*. The other designations are as in Figure 2. The PIR Accession Number for ORF0, the SWISS-PROT Accession Number for GNTR and the GenBank Accession Number for the sequence encoding YgaE' are indicated.

(Table 2). The search for conserved functional motifs revealed, predictably, the wide-spread ATPase, NAD-binding, S-adenosyl-methionine (SAM)-binding, and helix–turn–helix motifs as well as some motifs that are less common in bacterial proteins, e.g. Zn finger (Table 2 and see below). Thus this set apparently was reasonably representative of functional classes of bacterial proteins and included members of a number of well known and wide-spread protein families.

The majority of the long ORFs present in the sequenced portions of the *E.coli* chromosome already have been described as either actual or at least putative genes. What is likely to be discovered in the unannotated regions falls into several categories: i) incomplete gene sequences encoding N-terminal or C-terminal regions of proteins; ii) complete small ORFs; iii) extensions of previously described ORFs that have not been identified due to sequencing errors or utilization of an unusual initiation codon;

Figure 4. ORF *yhF'* encodes the C-terminal region of a putative phosphoribulokinase. (a) GeneMark prediction for the sequence around the *yhA* ORF. The designations are as in Figure 2a. *yhF'* is in the second top panel (frame +2) and *yhA* is in the third panel from the bottom (frame -1). (b) multiple sequence alignment. The SWISS-PROT accession numbers for the bacterial phosphoribulokinases and the GenBank accession number for the sequence encoding *yhF'* are indicated. Ae, *Alcaligenes eutrophus*; Rs, *Rhodobacter sphaeroides*; Xf, *Xanthobacter flavus*.

iv) new long ORFs masked by frameshifts or other sequencing errors. In each of these situations, finding highly significant sequence similarity is less likely than it is for an average *E. coli* protein. Nevertheless, the set of putative new genes identified in this study included all these cases and while a complete discussion of the new identifications is beyond the scope of this work, we describe typical examples of each.

N-terminal fragments of new proteins. The putative protein YchM' (end__*prs*; encoded upstream of the *prs* gene in the complementary strand; Fig. 2A) was related to two membrane proteins, namely early nodulin from soybean and human DRA protein that is down-regulated in colon tumors (refs. 45, 46; Fig. 2B and Table 2). The observed similarity was by far higher than it could be expected based solely on the presence of transmembrane segments and suggested a not yet discovered common function for these proteins.

The putative protein YgaE' (*gabP*_end; encoded downstream of *gabP*) was strongly similar to an uncharacterized ORF product from *Pseudomonas* sp. and both these sequences were moderately similar to the GNTR family of helix-turn-helix motif-containing transcription regulators (Fig. 3 and Table 2). The conservation in the predicted helix-turn-helix motif was particularly striking (Fig. 3). Accordingly, it is likely that YgaE' is a new transcription regulator in *E. coli*, whereas ORF 0 product of *Pseudomonas*, which is encoded upstream of the biphenyl degradation operon, may be involved in the transcription regulation of this operon. In this case, analysis of portions of new genes in the unannotated regions of the *E. coli* genome helps in functional identification of complete, but uncharacterized genes in other bacteria.

C-terminal fragments of new proteins. The putative protein YhfF', encoded in end_yhfA, upstream of yhfA, was related to several bacterial phosphoribulokinases (Fig. 4 and Table 2). This case is illustrative of the complications arising in sequence analysis of terminal regions of proteins. While YhfF' showed significant similarity only to proteins from other Gram-negative bacteria, the latter were related also to eukaryotic phosphoribulokinases (47), indicating that YhfF' actually belongs to a conserved protein family.

A more unusual case of sequence conservation in a C-terminal fragment of a new protein was observed with YggI' which is encoded in end__*endA*, upstream of the *endA* gene. YggI' is significantly similar to the eukaryotic Zn finger domains of the

Krüppel type (Fig. 5 and Table 2). *Krüppel* proteins contain numerous repeats of a C₂H₂ Zn finger unit (48). YggI' contains a CHC₂ finger-like pattern that aligns with the cysteine pairs of two adjacent fingers in the *Krüppel* proteins; in each of the two histidine pairs that are present in this region of *Krüppel* proteins and are involved in finger formation, only one histidine is conserved in YggI' (Fig. 5). Thus the single finger domain of YggI' could have evolved from a repetitive *Krüppel*-like structure as the result of mutations eliminating part of the Zn ligands. *Krüppel* proteins are mostly developmental regulators (48) and by inference, it seems likely that YggI' may be involved in regulation of gene expression in *E. coli*.

Extension of previously described ORFs. In several cases, combined analysis by GeneMark and BLAST allowed us to extend an ORF previously described by others and to predict the function of its product. For example, GeneMark analysis indicated that the nucleotide sequence upstream of the uncharacterized ORF *yggC* (in the *yggC*—*yggD* region) was likely to be expressed (Fig. 6A). Translation of this region revealed an uninterrupted reading frame that showed highly significant similarity to *E.coli* panthothenate kinase and more distant similarities to several other kinases, strongly suggesting that the product of the extended ORF *yggC* is a new kinase (Fig. 6B and Table 2). The principal NTP-binding motif (Fig. 6B) is encoded in the upstream portion of the putative new gene that has been assumed not to be expressed (49). It is likely that this, apparently incomplete identification of the coding region was due to not taking into account of the possibility of translation initiation from a GTG codon that we now believe is used for the expression of the *yggC* gene (data not shown).

Complete small genes. The sequence downstream of the *CspA* gene (region *cspA*__end) contained, in the complementary strand, a complete small (51 codons) ORF coding for the putative protein YiaZ. YiaZ is closely related to the Gef family of cell-killing membrane proteins (Fig. 7 and Table 2). By all likelihood, the product of this gene is the third, in addition to RelF and Gef (50), member of this family encoded in the *E. coli* chromosome. Again, one of the reasons why this gene has not been described previously may be that its translation is likely to be initiated from a GTG codon. The *yiaZ* gene was predicted by GeneMark only weakly and only by the version trained on foreign (class III) genes (data not shown). The weak prediction may be attributed to the large variance of the GeneMark function value for short sequences (24,25). A 'foreign' origin for the new *gef*-related gene

			?	?		!	!		?	?		!	!		
YggI ⁺	?	RTHQFEIQS	---	VRRTFFPYRCKQCEQLT	---	VRRENVV	RGAEAVYR	CVHCGEQLVAK							X65169g
FIN12_mouse	471	RTHQKIHTG	---	EKPKY	---	CSECDCK	FIKKAANR	RIHTG	EKPKYK	SCGCK	CFIQK				197 A48830p
ZF27_MOUSE	?	RHQIKIHTG	---	EKRHG	---	CVDCGK	APTKQSTL	RHKIHTG	ERAYV	ICGGA	CFIQK				? P10077
ZN91_HUMAN	589	STHKIHTG	---	EKSYK	---	CEBGCA	FLWSL	TRRRIHTG	EKPKYK	CECG	KAFHS				549 Q05481
ZFP7_mouse	?	RHRDRIHSG	---	EKPFV	---	CKYCG	APTKQSSL	PRRIHSG	EKPKYV	CKY	CGCKFISS				? D10682g
BPZ_mouse	337	QCERIHTG	---	EKPF	---	CNCGCA	FAHSTL	QRKRIHTG	EKPKY	ENCG	CKGAFAPC				156 L20450g
Z072_XENLA	207	PEHQRTHTG	---	EKFFT	---	CTCKG	GFQTKT	LRNKHHTG	ERPYCS	CECGK	SFTEK				94 P18752
Z022_XENLA	281	INHQRTHTG	---	EKPF	---	CSECG	SFSN	QSLRV	QRTHTG	EKPKYCS	CECGK	SFVTS			101 P18745
CD2_DROME	384	AMRKIHTGEADTPKRPYT	---	CSYCGK	SFTQ	SNLT	QRTHTG	EKPKY	CFPR	CGG	CGRAFTVK				69 P20385
consensus		..H...G...G...A...T...P...K...R...Y...T...C...C...S...F...Q...S...N...L...T...Q...R...T...H...T...G...E...K...P...K...Y...C...S...C...G...K...F...P...R...C...G...G...A...F...T...V...K...													
				S			W							X	

Figure 5. ORF *YggI'* encodes a putative Zn finger protein related to *Krüppel* domains. Exclamation mark indicate the cysteine or histidine (in *YggI'*) residues in the conserved finger domain. Question marks show the histidine residues that are involved in finger formation in the *Krüppel* proteins but apparently not in *YggI'*. Each of the *Krüppel* proteins includes numerous imperfect repeat of the C₂H₂ Zn finger unit. The alignment includes the segment from each protein that showed the highest similarity to *YggI'*. Sequences of several *Krüppel* proteins are incomplete a shown by question marks. SWISS-PROT, PIR (p) or GenBank Accession numbers are indicated.

Figure 6. ORF yggC encodes a new kinase. (a) GeneMark prediction for the region around the yggC and yggD ORFs. The extended yggC coding region is in the second panel from the bottom (frame -2); note that the ORF starts with a GTG codon as indicated by the small upward tick on the midline. yggD is in the bottom panel (frame -3). The other designations are as in Figure 2a. (b) Multiple sequence alignment. Four conserved blocks delineated by the MACAW program are shown and the distances between them are indicated. The two motifs that together comprise the Walker type purine NTP-binding site (54) are denoted. The new protein sequence identified in this study using GeneMark and BLASTX is an N-terminal extension of the previously reported YggC and includes the 'A' motif of the purine NTP-binding pattern. CoaA is *E. coli* acetyl-CoA kinase; URK1 is yeast (Sc) uridylate kinase; KPPR is phosphoribulokinase from spinach (*Spinacia oleracea*; So) chloroplasts and from *Synechocystis* sp. (Ssp); and Udk is uridine kinase from *E. coli*. The SWISS-PROT or PIR (p) accession numbers for the kinases and the GenBank accession number for the sequence containing yggC and yggD are indicated.

is plausible as most of the other members of this family (Fig. 7) are plasmid proteins.

CONCLUSIONS AND PROSPECTS

In this work we have analyzed the unannotated regions of the *E. coli* genome using two complementary approaches, namely prediction of the coding potential by GeneMark program based on Markov chain models and database searches using BLAST and motif search programs. We were interested in assessing the congruency of the results obtained with each of these methods. The great majority of the ORFs that were related to known proteins were also predicted to be real genes by GeneMark. Conversely, slightly more than one half of the GeneMark predictions were supported by significant sequence similarity. With respect to the proportion of genes belonging to classes I, II and III, and the fraction of the genes that have homologues in the sequence databases, the set of putative genes predicted by GeneMark resembled the set of genuine *E. coli* genes. This leads to the conjecture that even prediction by GeneMark alone may be an indication that an ORF in *E. coli* DNA is a real gene.

It is worth mentioning that the results of this and similar studies will be useful for improving the assessment of GeneMark accuracy by providing better control sets of non-coding sequences. Obviously, if the predicted new genes that have been detected by both GeneMark and by BLAST are subtracted from the unannotated regions in EcoSeq6, this will significantly reduce the apparent false positive rate. It is expected that with the accumulation of sequence and biochemical data, an increasing number of ORFs about which we know nothing other than a high GeneMark score will prove to be real genes. Thus the false positive rate of GeneMark prediction is likely to converge to a low value which cannot be precisely predicted at this time.

This analysis shows that unannotated regions of the *E. coli* chromosome are a rich source of new genes. After this work has been completed, we became aware of two recent studies, one of which has focused on detecting sequence conservation in uncharacterized regions of bacterial DNA (51) and the other one on identifying coding regions using Hidden Markov Models (52). Many of the new genes identified here have been independently described in these studies. The more general aspects of our strategy for gene identification in a bacterial genome are described in a separate publication (53). The approach to gene finding that combines intrinsic and extrinsic analysis is likely to become increasingly important with further progress of genome sequencing projects.

NOTE

The putative new protein sequences that were identified in this study have been submitted to the SWISS-PROT database. These sequences in FASTA format, accompanied by the SWISS-PROT identifiers and Accession numbers are also available via anonymous FTP at ncbi.nlm.nih.gov in the directory repository/Eco/GeneMark. The GeneMark program can be used via E-mail server (genmark@ford.gatech.edu). The server now implements the described approach, combining coding region prediction using GeneMark and sequence similarity search, for gene identification in several bacterial and eukaryotic species.

ACKNOWLEDGEMENTS

We are grateful to J.McIninch for valuable programming assistance; to G.Bouffard for help in data handling; to Dr D.Landsman for useful discussions; and to Drs A.Krogh, I.S.Mian and D.Haussler for communicating their results prior to publication. The work of M.B. was supported by NIH grants HG00783 and GM47853.

REFERENCES

1. Kroger, M., Wahl, R. and Rice, P. (1993) *Nucleic Acids Res.* **21**, 2973–3000.
2. Rudd, K. E. (1993) *ASM News* **59**, 335–341.
3. Rudd, K. E. (1992) in *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria*, by Miller, J. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor), pp. 2.3–2.43.
4. Daniels, D., Plunkett, G., Burland, V. and Blattner, F. R. (1992) *Science* **257**, 771–778.
5. Shepherd, J.C.W. (1981) *Proc. Natl Acad. Sci. USA* **78**, 1596–1600.
6. Fickett, J. W. (1982) *Nucleic Acids Res.* **10**, 5303–5318.
7. Staden, R. and McLachlan, A. D. (1982) *Nucleic Acids Res.* **10**, 141–156.
8. Staden, R. (1984) *Nucleic Acids Res.* **12**, 505–519.
9. Gribskov, M., Devereux, J. and Burgess, R. R. (1984) *Nucleic Acids Res.* **12**, 539–549.
10. Borodovsky, M., Sprzhitsky, Yu. A., Golovanov, E.I., Alexandrov, A.A. (1986) *Molekularnaya Biologiya* **20**, 1144–1150 (English translation).
11. Borodovsky, M.Yu., Sprzhitsky, Yu. A., Golovanov, E.I., Alexandrov, A.A. (1986) *Molekularnaya Biologiya* **20**, 833–840 (English translation).
12. Claverie, J.M. and Bougueleret, L. (1986) *Nucleic Acids Res.*, **14**, 179–196.
13. Fichant, G. and Gautier, C., (1987) *Comput. Appl. Biosci.* **3**, 287–295.
14. Claverie, J., Sauvaget, I. and Bougueleret, L. (1990) *Meth. Enzymol.* **183**, 237–253.
15. Fields, C.A. and Soderlund, C.A. (1990) *Comput. Appl. Biosci.* **6**, 263–270.
16. Gelfand M. (1992) *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **8**, 19–61.
17. Lapedes, A., Barnes, C., Burks, C., Farber, R. and Sirotkin, K. (1990) In: *Computers and DNA* (Bell, G.I., Marr, T., editors) pp. 157–182, Addison-Wesley. Reading.

Figure 7. Putative protein YiaZ is a member of the GEF family of small cell-killing membrane proteins. The predicted transmembrane segment is shown. The aligned sequences were all from *E.coli* and its plasmids. The SWISS-PROT and GenBank Accession Numbers are indicated as in the other figures.

18. Konopka, A. K. and Owens, J. (1990) *Gene Anal. Techn. Appl.* **7**, 35–38.
19. Senapathy, P., Shapiro, M. B. and Harris, N. L. (1990) *Meth. Enzymol.* **183**, 252–278.
20. Uberbacher, E. C. and Mural, R. J. (1991) *Proc. Natl Acad. Sci. USA* **88**, 11261–11265.
21. Hutchinson, G. B. and Hayden, M. R. (1992) *Nucleic Acids Res.* **20**, 3453–3462.
22. Snyder, E. E and Stormo, G. D. (1993) *Nucleic Acids Res.* **21**, 607–613.
23. Fickett, J. W. and Tung, C.-S. (1992) *Nucleic Acids Res.* **20**, 6441–6450.
24. Borodovsky, M. and McIninch, J. (1993) In Proceedings of the Second International Conference on Supercomputing, Bioinformatics and Complex Genome Analysis. St. Petersburg FL, 1992, World Scientific, pp. 231–248.
25. Borodovsky, M. and McIninch, J. (1993) *Comput. Chem.* **17**, 123–133.
26. Burland, V., Plunkett, G., Daniels, D. L. and Blattner, F. R. (1993) *Genomics* **16**, 551–561.
27. Plunkett, G., Burland, V., Daniels, D. L. and Blattner, F. R. (1993) *Nucleic Acids Res.* **21**, 3391–3398.
28. Blattner, F. R., Burland, V., Plunkett, G. III., Sofia, H. J. and Daniels, D. L. (1993) *Nucleic Acids Res.* **21**, 5408–5417.
29. Tavare, S. and Song B. (1989) *Bull. Math. Biol.* **51**, 95–115.
30. Kleffe, J. and Borodovsky, M., (1992) *Comp. Appl. Biosci.* **8**, 433–441.
31. Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) *J. Mol. Biol.* **222**, 851–856.
32. Medigue, C., Viari, A., Henaut, A. and Danchin, A. (1993) *Microbiol. Rev.* **57**, 623–654.
33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
34. Gish, W. and States, D. J. (1993) *Nature Genet.* **3**, 266–272.
35. Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994) *Nature Genet.* **6**, 119–129.
36. Henikoff, S. and Henikoff, J. (1992) *Proc. Natl Acad. Sci. USA* **89**, 10915–10919.
37. Henikoff, S. and Henikoff, J. (1993) *Proteins Struct. Funct. Genet.* **17**, 49–61.
38. Wootton, J. C. and Federhen, S. (1993) *Comput. Chem.* **17**(2), 149–163.
39. Schuler, G. D., Altschul, S. F. and Lipman, D. J. (1991) *Proteins Struct. Funct. Genet.* **9**, 180–190.
40. Tatusov, R. L. and Koonin, E. V. (1994) *Comp. Appl. Biosci.* **10**, 457–459.
41. Bairoch, A. (1993) *Nucleic Acids Res.* **21**, 3097–3103.
42. Green, P., Lipman, D. J., Hillier, L., Waterston, R., States, D. and Claverie, J.-M. (1993) *Science* **259**, 1711–1716.
43. Koonin, E. V., Bork, P. and Sander, C. (1994) *EMBO J.* **13**, 493–503.
44. Claverie, J.-M. (1993) *Nature* **364**, 19–20.
45. Schweinfest, C. W., Henderson, K. W., Suster, S., Kndoh, N. and Papas, T. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 4166–4170.
46. Sandal, N. N. and Marcker, K. A. (1994) *Trends Biochem. Sci.* **19**, 19.
47. Gibson, J. L., Che, J. H., Tower, P. A. and Tabita, F. R. (1990) *Biochemistry* **29**, 8085–8093.
48. Schuh, R., Aicher, W., Gaul, U., Cote, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schroder, C., Kemler, R. and Jackle, H. (1986) *Cell* **47**, 1025–1032.
49. Alefounder, P. R., Baldwin, S. A., Perham, S. A. and Dhort, N. J. (1989) *Molec. Microbiol.* **3**, 723–732.
50. Gerdes, K., Poulsen, L. K., Thisted, T., Nielsen, A. K., Martinussen, J. and Andreasen, P. H. (1990) *New Biologist* **2**, 946–956.
51. Robison, K., Gilbert, W. and Church, G. M. (1994) *Nature Genet.* **7**, 205–214.
52. Krogh, A., Mian, I. S. and Haussler, D. *Nucleic Acids Res.* **22**, 4***–4***.
53. Borodovsky, M., Koonin, E. V. and Rudd, K. E. (1994) *Trends Biochem. Sci.* **19**, 309–313.
54. Walker, J. E., Saraste, M., Runswick, M. J. and Gay N. J. (1982) *EMBO J.* **1**, 945–951.
55. Karlin, S. and Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
56. Karlin, S. and Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
57. Gorbalenya, A. E. and Koonin, E. V. (1990) *J. Mol. Biol.* **213**, 583–591.
58. Kitagawa, M., Wada, C., Yoshioka, S. and Yura, T. (1991) *J. Bacteriol.* **173**, 4247–4253.
59. Rhiel, E., Stirewalt, V. L., Gasparich, G. E. and Bryant, D. A. (1992) *Gene* **112**, 123–128.
60. Shao, Z. and Newman, E. B. (1993) *Eur. J. Biochem.* **212**, 777–784.
61. Cain, B. D., Norton, P. J., Eubanks, W., Nick, H. S. and Allen, C. M. (1993) *J. Bacteriol.* **175**, 3784–3789.
62. Nashimoto, H. (1993) In Nierhaus, K. H. (ed) *The Translation Apparatus*, Plenum Publishing Corporation, New York, pp. 185–195.
63. Koonin, E. V. and Rudd, K. E. (1993) *Nucleic Acids Res.* **21**, 5447.
64. Koonin, E. V. and Rudd, K. E. (1994) *Trends Biochem. Sci.* **19**, 106–107.