Exact mapping of prokaryotic gene starts

Mikhail V. Baytaluk, Mikhail S. Gelfand and Andrey A. Mironov

Date received (in revised form): 30th April 2002

Mikhail Baytaluk
is a PhD student at the Institute of Molecular Biology, RAS. His research is in the area of gene recognition.

Mikhail Gelfand
is Director for Science, Integrated Genomics, Moscow. His research interests are comparative genomics, genome annotation, analysis of regulation of gene expression and gene recognition.

Andrey Mironov
is Director for Technology, Integrated Genomics, Moscow. His research interests are the creation of algorithms for sequence and structure alignments, software development and genome annotation

Keywords: *gene, genomics, gene recognition, reading frame, start of translation, computer analysis, prokaryotes*

Abstract

It is known that while the programs used to find genes in prokaryotic genomes reliably map protein-coding regions, they often fail in the exact determination of gene starts. This problem is further aggravated by sequencing errors, most notably insertions and deletions leading to frame-shifts. Therefore, the exact mapping of gene starts and identification of frame-shifts are important problems of the computer-assisted functional analysis of newly sequenced genomes. Here we review methods of gene recognition and describe a new algorithm for correction of gene starts and identification of frame-shifts in prokaryotic genomes. The algorithm is based on the comparison of nucleotide and protein sequences of homologous genes from related organisms, using the assumption that the rate of evolutionary changes in protein-coding regions is lower than that in non-coding regions. A dynamic programming algorithm is used to align protein sequences obtained by formal translation of genomic nucleotide sequences. The possibility of frame-shifts is taken into account. The algorithm was tested on several groups of related organisms: gamma-proteobacteria, the *Bacillus/Clostridium* group, and three *Pyrococcus* genomes. The testing demonstrated that, dependent on a genome, 1–10 per cent of genes have incorrect starts or contain frame-shifts. The algorithm is implemented in the program package Orthologator-GeneCorrector.

INTRODUCTION

Recent advances in sequencing of complete genomes, growth of data deposited in sequence databases, and development of computer programs for the large-scale similarity analysis make it possible to design more accurate tools for gene recognition. Systematic analysis of the performance of available software for gene recognition highlighted that while the current programs perform well at identifying genes (as opposed to random open reading frames, ORFs), gene starts are predicted with lower accuracy.

Although for many purposes the approximate location of gene starts, eg using 'the leftmost ATG' rule, is sufficient, it still creates some problems, leading to proliferation of annotation errors, complicating genomic analyses that depend on intergenic distances, eg prediction of the operon structure,[1] making it impossible to predict secreted proteins via analysis of signal peptides,[2] and obstructing analysis of translational regulation.[3]

An additional problem complicating gene recognition arises from frame-shifts that interrupt reading frames. Although there exist many biologically meaningful frame-shifts[4] or frame-shifts that indicate non-functional pseudogenes (an extreme case is the degenerating genome of *Mycobacterium leprae*[5]), at least some of them may be caused by sequencing errors. Thus it is desirable to have an algorithm for gene recognition that would not be confused by frame-shifts. Indeed, even if a frame-shift were real, it still would be useful to know the original amino acid sequence of the encoded protein.

The following features are important for gene recognition: (1) ORF length; (2) presence of a ribosome binding site (RBS) upstream of the start codon; (3) specific pattern of codon usage that is different from triplet frequencies in non-coding regions ('coding potential'), as well as

M. S. Gelfand,
Integrated Genomics – Moscow,
PO Box 348,
Moscow 117333,
Russia

Tel: +7 (095) 135 20 41
Fax: +7 (095) 132 60 80
E-mail:
gelfand@integratedgenomics.ru

other similar statistical parameters; and (4) similarity to known genes.

Intrinsic approach to gene recognition

Intrinsic, or *ab initio*, approaches use the first three types of data. Hidden Markov models (HMM) provide a convenient language for integrating these diverse parameters of candidate genes in genomic sequences. Extrinsic methods rely on the comparative analysis of genomic DNA sequences using alignment with known genes and proteins.

Ribosome binding sites

Ribosome binding sites are located in the (−20) ... (−1) region upstream of start codons and serve to direct ribosomes to the correct translation start position. A part of RBS is formed by the purine-rich Shine–Dalgarno (SD) sequence, which is complementary to the 3′ end of the 16S rRNA.[6] A number of early papers described methods for recognition of ribosome binding sites using statistical, pattern recognition or neural network modelling of experimentally mapped sites.[7–10]

There are two approaches to the recognition of ribosome binding sites in the absence of a learning sample. One possibility is to rely on the universal mechanisms of RBS recognition via base-pairing of the SD box and the 3′-terminus of the 16S rRNA.[6] It was used to predict RBSs in *Escherichia coli* by calculation of the optimal binding energy between the 16S rRNA of *E. coli* and the region upstream of a potential start codon.[11] It turned out that the reliability of this approach in *E. coli* is rather low, as the RBS pattern is weak. However, in clostridial Gram-positive bacteria, in particular *Bacillus subtilis* and *Staphylococcus aureus*, the average energy at potential RBSs tends to be much stronger. Base-pairing of the SD box and the 3′-terminus of the 16S rRNA was used to predict RBSs in *B. subtilis*, *E. coli* and *Pyrococcus furiosus*.[12] Beside the RBS binding energy, Hannenhalli *et al*.[12] took into account additional information: the distance between the RBS and the start codon, the distance from the beginning of the maximal ORF to the start codon, the start codon itself (ATG, GTG or TTG) and the coding/non-coding statistics around the start site.

The other possibility is to derive a 'pseudo-learning' sample of candidate translation initiation sites using protein-coding regions predicted by database search or statistical analysis. In the GeneMark system this sample consists of ATG codons at the 5′-ends of statistically predicted protein-coding regions.[13,14] For prediction, GeneMark uses the start codon score, the SD box score, the downstream box score, pre-start signal score and post-start signal score, all based on similarity to profiles generated from a training set. Similar ideas are implemented in RBS-Finder, a post-processing tool for GLIMMER 2.0, that finds RBSs upstream of start codons.[15]

ORPHEUS uses the similarity analysis to identify genes with only one candidate start codon, and then uses these genes to derive the recognition rule for gene starts.[16] The RBS site score is defined as the sum of the SD box score and the weight of the distance between the SD box and the start codon.

Codon usage

Gene recognition algorithms relying on the codon usage explore the idea that the codon choice is genome-specific.[17] Eighteen amino acids (not methionine and tryptophan) are encoded by two to six codons. The codon usage (the combined result of the amino acid usage and the synonymous codons usage) varies both between organisms and between different genes in the same organism. Indeed, the codon usage reflects the expression level of bacterial genes[18–21] and the history of lateral gene transfer.[22] Still, the statistical patterns in protein-coding regions (the codon usage, correlations between adjacent codons, etc.) are sufficiently strong to distinguish genes from random ORFs (for a review see Fickett and Tung[23]).

Integration: hidden Markov models

A convenient technique for integration of diverse parameters is the HMM. HMM[24] is a Markov chain of hidden states. Each state is assigned a distribution of emission probabilities (Bernoulli or Markov) that generate the observed

nucleotide sequence. The aim is, given a nucleotide sequence, to reconstruct the most probable sequence of hidden states, each corresponding to a functional state: protein-coding, non-coding intergenic, SD box, etc. This is done using a variant of the standard dynamic programming.

Extrinsic approaches

One of the most popular *ab initio* programs is GeneMark.[25] It uses non-homogeneous Markov models to describe coding sequences and ordinary Markov models for non-coding sequences. For analysis of a newly sequenced genome, parameters of the Markov models are estimated from a set of ORFs longer then 1,000 nucleotides. As an initial model for non-coding sequences, a zero-order Markov model with genome-specific nucleotide frequencies is used. The initial models are used at the first prediction step. The results of the first prediction are then used to compile a set of putative genes used at the second training step. The training and prediction steps are iterated until the set of predicted genes stabilises. Recently this algorithm was re-formulated using the language of hidden Markov models and extended to take into account information about candidate ribosome binding sites.[26]

Combined techniques

Another very popular gene recognition program, GLIMMER, relies on interpolated Markov models to take into account DNA oligomers of varying length, thus using all available data without over-training.[27] The gene start is assigned, by default, to the start codon of the longest ORF containing the predicted coding region. Then the program computes the maximum value of the hybridisation energy between the anti-SD segment in the 16S rRNA and the fragments upstream of putative start codons. If there are candidate starts where this value exceeds some threshold, the start with the highest scoring putative SD box is accepted. In a later version, GLIMMER 2.0, the sensitivity of the method was increased by resolution of overlapping genes and improvement of the probabilistic model.[28]

One more program, EcoParse, also finds the maximum likelihood parse of a DNA sequence into coding and non-coding regions using the hidden Markov model technique.[8]

Extrinsic analysis involves sequence similarity searches. Candidate gene products are searched against protein sequence databanks. BLASTX, the most popular program of this class, performs six-frame translation of the query DNA and compares the resulting amino acid sequences to known proteins.[19] In the pre-genome era, BLASTX was used to detect several hundreds of new bacterial genes missed in original publications and GenBank submissions.[29]

The simplest way to combine the extrinsic and intrinsic approaches is to apply them in parallel.[30] The complete genome sequence of *Bacillus subtilis* was screened by combination of two independent analyses by BLASTX and proFED (prokaryotic frame-shift error detection).[31] The ProFED program uses the predicted coding probabilities in the six reading frames computed by GeneMark and GLIMMER, and then attempts to reconcile overlapping or adjacent high-quality reading frames by incorporation of frame-shifts.

ORPHEUS utilises non-supervised training based on sequence similarity searches.[16] The analysis starts with database similarity search and identification of gene fragments having known reliable homologues. These fragments are used to derive the codon usage statistics and to construct the RBS scoring matrix (see above). At the prediction step, the 5′-proximal codon with sufficiently strong RBS is accepted. Unlike GeneMark and EcoParse, ORPHEUS does not rely on statistics of non-coding regions. The motivation is that only coding regions can be defined unambiguously, especially at the initial steps of the analysis.

Another gene recognition program with emphasis on accurate mapping of gene starts is CRITICA.[32] It uses BLASTN[33] at the initial stage to locate sequences in DNA database that are

highly similar to the query. If conservation of the amino acid sequence is stronger than expected given the level of conservation of the nucleotide sequence, the ORF is assumed to be coding. Similar reasoning is used to choose the correct start codon.

Even this brief review shows that the distinction between intrinsic and extrinsic methods is somewhat blurred and the most successful algorithms incorporate both approaches. In practice, putative coding regions identified by intrinsic methods are verified by similarity searches, to get support for the predicted protein. Length corrections, based on comparison with known proteins, were made in several dozens of GeneMark-predicted ORFs in the *Haemophilus influenzae*, *Methanococcus jannashii* and *Mycoplasma genitalium* genomes.[34] Tables 1 and 2 list gene recognition programs used to annotate complete prokaryotic genomes.

Benchmarking of gene recognition software is a difficult problem, since only a few genomes have been characterised experimentally to a sufficient extent. In one such study[38] it was shown that the fraction of correctly identified gene starts is highly correlated with the information content of the SD box signal.

We now turn to a description of an algorithm for correction of gene starts and frame-shifts in prokaryotic genomic sequences.

MATERIALS AND METHODS

The prediction is done in three steps:

- Building the tables of orthologues.
- Applying a dynamic programming algorithm to align pairs of orthologous genes.
- Filtering of results and identification of suspicious gene starts and possible frame-shifts.

The data flow is presented in Figure 1. The output of step 3 is evaluated before the final decision about correction of errors is made.

Data

The algorithm was tested on three groups of genomes:

- *Escherichia coli*,[39] *Vibrio cholerae*,[40] *Haemophilius influenzae*,[41] *Buchnera* sp.,[42] *Xylella fastidiosa*;[43]
- *Bacillus subtilis*,[44] *Bacillus halodurans*,[45] *Clostridium acetobutylicum*;[46]
- *Pyrococcus horikoshii*,[47] *Pyrococcus abyssi*,[48] *Pyrococcus furiosus*.[49]

Additionally, the following complete and incomplete genomes of gamma-proteobacteria were considered: *Salmonella enterica*, *S. enteritidis*, *S. paratyphi*, *S. typhi*, *S. typhimurium*, *Klebsiella oxytoca*, *K. pneumoniae*, *Yersinia enterocolitica*, *Y. pseudotuberculosis*, *Haemophilus ducreyi*, *Pasteurella haemolytica*, *P. multocida*, *Pseudomonas fluorescens*, *Ps. putida*, *Ps. stutzeri*, *Ps. syringae*, *Ps. aeruginosa*, *Vibrio anguillarum*, *V. parahaemolyticus*, *Xylella almond*, *X. oleander*, *Erwinia carotovora*, *Er. amylovora*, *Er. chrysanthemi*, *Er. herbicola*, *Buchnera aphidicola*, *Enterobacter aerogenes*, *En. cloacae*, *Shigella flexneri*, *Sh. sonnei*, *Proteus mirabilis*.

Building the orthologue tables

This is the least specific part of the algorithm: a pre-computed table produced by any external tool can be used. However we describe this step for the sake of completeness.

Although careful analysis of orthologues requires construction of a large number of phylogenetic trees, a reasonable approximation of orthology relationships in our case (closely related genomes) comes from best bidirectional hits (BETs);[50] cf. the COG (cluster of orthologous genes) system.[51]

Two genes, g from genome A and h from genome B, form a BET if the similarity between these genes $s(g,h)$

Table 1: Genomes and gene recognition programs

Organism	Gene recognition algorithm
Aquifex aeolicus	CRITICA and similarity
Archaeglobus fulgidus	GeneSmith and CRITICA
Bacillus halodurans	GeneHacker Plus
Bacillus subtilis	GeneMark
Borrelia burgdorferi	GLIMMER
Buchnera sp.	GeneHacker Plus
Campylobacter jejuni NCTC 11168	ORPHEUS and GLIMMER
Caulobacter crescentus	GLIMMER
Chlamidia pneumoniae AR39	GLIMMER
Chlamidia trachomatis MoPn Nigg	GLIMMER
Clostridium acetobutylicum	ORFs(Uniorf)
Clostridium perfringens	GeneHacker Plus
Cyanobacterium anabaena	GLIMMER
Deinococcus radiodurans R1	GLIMMER
Escherichia coli	GeneMark
Escherichia coli O157:H7	GeneHacker Plus
Halobacterium sp. NRC-1	GLIMMER
Helicobacter pylori	GeneMark and GeneSmith
Methanococcus jannaschii	GeneMark
Mycobacterium tuberculosis	TbParse
Mycoplasma genitalium	GeneMark
Mycoplasma pneumoniae	Frames
Mycoplasma pulmonis	GLIMMER
Neisseria meningitidis Z2491 (serogroup A)	GLIMMER
Pasteurella multocida Pm70	ORPHEUS and GLIMMER
Pseudomonas aeruginosa	GeneMark
Pyrococcus abyssi	ORFs and similarity
Pyrococcus horikoshii	ORFs and similarity
Ralstonia solanacearum	FrameD
Rickettsia prowazekii	ORFs(BioWish)
Salmonella enterica	GeneMark and GLIMMER
Sinorhizobium meliloti	GLIMMER and FrameD
Staphylococcus aureus	GLIMMER
Streptococcus pneumoniae TIGR4	GLIMMER
Streptococcus pyogenes	GLIMMER
Sulfolobus solfataricus P2	GLIMMER
Synechocystis sp. PCC6803	GeneMark
Termotoga maritime	GLIMMER
Thermoplasma acidophilum	ORPHEUS
Thermotoga maritima	GLIMMER
Treponema pallidum	GLIMMER
Ureaplasma urealiticum	GeneMark and GLIMMER
Vibrio cholerae	GLIMMER
Xylella fastidiosa	GLIMMER2.0 and RBSfinder
Yersinia pestis	GLIMMER

exceeds the similarity for any other choice of either member of the pair: $s(g, h) > s(x, h)$ and $s(g, h) > s(g, y)$ for every $x \neq g$ from genome A and $y \neq h$ from B.

For every gene in the basic genome (BG) to be corrected, the program identifies an orthologue in another (additional) genome (AG) from the same taxonomic group. The gene pairs are formed using BLASTP.[52] The statistical significance of similarity between a and b (E-value) must be less than 10^{-6}. Thus we obtain the table of orthologues for genomes BG and AG. The procedure is done for all additional genomes AG_i from the same taxonomic group.

The number of thus determined orthologue pairs depends on the degree of relatedness and the size of the compared genomes. For instance, the number of orthologue pairs for *E. coli* and *V. cholerae* was 2,300, whereas for *E. coli* and *H. influenzae* it was only 1,400.

The orthologue tables were constructed using the program ORTHOLOGATOR, which is a part of the created software package.[53] Processing of one pair of genomes requires approximately one hour (on a PC with Intel Pentium III, 650 MHz, RAM 128 M configuration), dependent on the size of genomes.

A dynamic programming algorithm for alignment of gene starts

Consider a pair of orthologous genes. Extend them by n_1, n_2 nucleotides at the left and m_1, m_2 nucleotides at the right

Table 2: Gene prediction methods

Gene prediction method	URL	Ref.
GeneMark	http://www.opal.biology.gatech.edu/GeneMark/	25
GLIMMER	http://www.tigr.org/softlab/glimmer/glimmer.html	27
EcoParse, TbParse	http://www.cbs.dtu.dk/krogh/EcoParse.info	35
GeneSmith	Unpublished	H. O. Smith
ORPHEUS	http://pedant.mips.biochem.mpg.de/orpheus	16
GeneHacker Plus	http://www.elmo.ims.u-tokyo.ac.jp/GH/	36
FrameD	http://www.toulouse.inra.fr/FrameD.html	37
CRITICA	ftp://rdp.life.uiuc.edu/pub/critica http://rdpwww.life.uiuc.edu	32

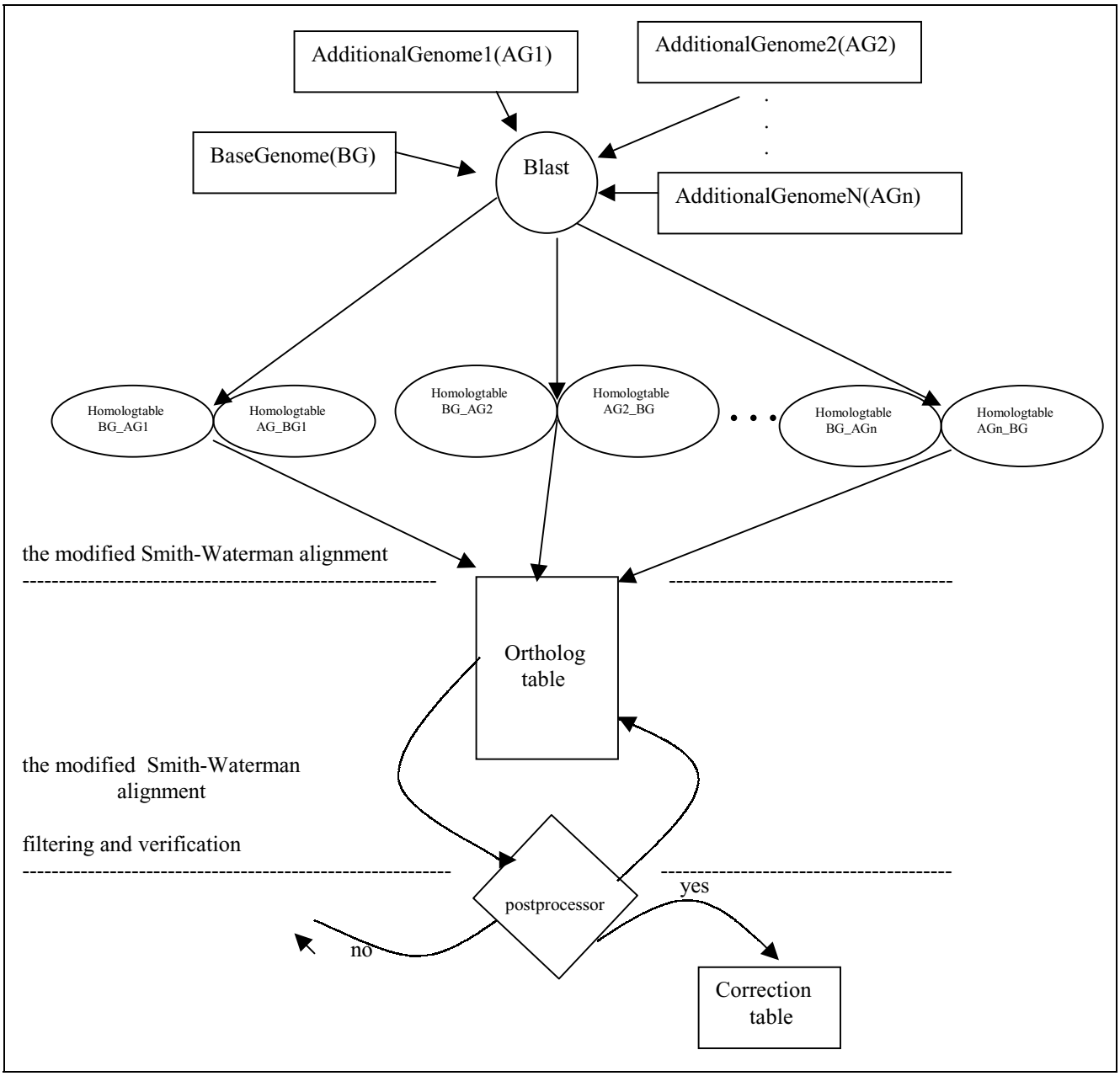

Figure 1: GeneCorrector data flow. Base genomes: *E. coli*, *V. cholerae*, *B. subtilis*, *P. horikoshii*. Additional genomes: *H. influenzae*, *Buchnera* sp., *X. fastidiosa*; *B. halodurans*, *C. acetobutylicum*; *P. abyssi*, *P. furiosus*

(the maximum extension length at each end is 200 nucleotides). Overlaps with adjacent genes translated in the same direction are excluded (Figure 2).

In a pair of extended sequences, all potential starts (ATG, TTG, GTG) and stops (TGA, TAA, TAG) around gene starts ($[s_1 - n_1, s_1 + n_1]$ for the first gene, $[s_2 - n_2, s_2 + n_2]$ for the second gene) and ends ($[e_1 - m_1, e_1 + m1]$ for the first gene, $[e_2 - m_2, e_2 + m_2]$ for the second gene) are marked. For two nucleotide sequences $[s_1 - n_1, e_1 + m_1]$ and $[s_v - n_v, e_2 + m_2]$ a variant dynamic programming algorithm similar to the Smith–Waterman local alignment procedure[54] is used to align the protein sequences generated by the formal translation of the nucleotide sequences in all three reading frames, with account to possible frame-shifts.

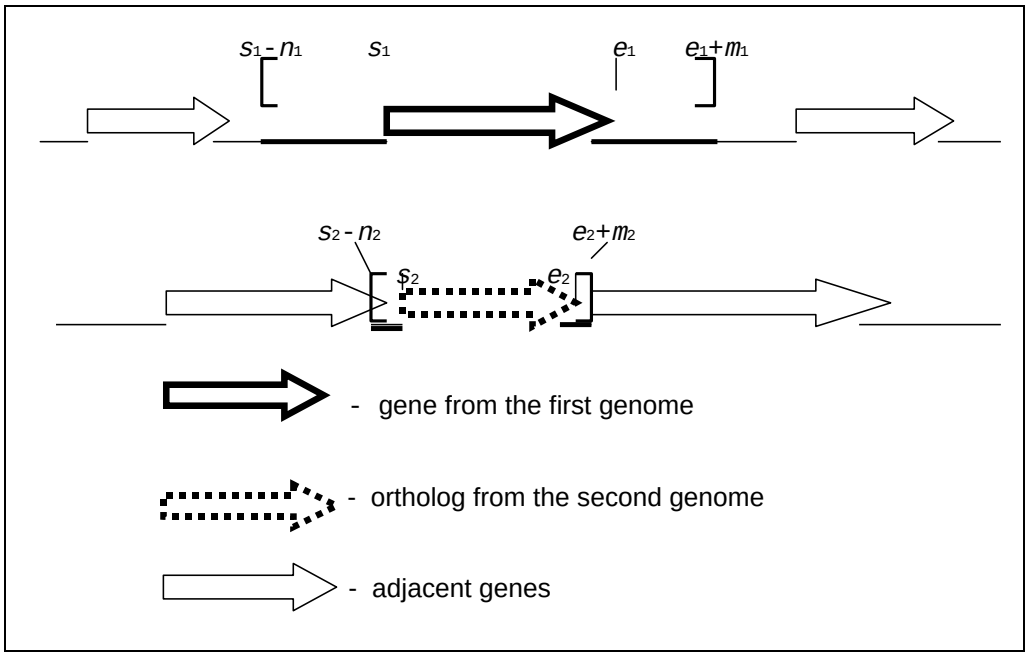

Figure 2: Expansion of gene boundaries

The recursions for the alignment are:

$$S_{i,j} = \max\begin{cases} S_{i-3,j-3} + \mathrm{d}(x_i, y_j) \\ Q_{i,j} \\ R_{i,j} \\ S_{i-2,j} + f \\ S_{i-1,j} + f \\ S_{i,j-2} + f \\ S_{i,j-1} + f \end{cases}$$

where:

$$Q_{i,j} = \max\begin{cases} S_{i,j-3} + a \\ Q_{i,j-3} + b \end{cases}$$

$$R_{i,j} = \max\begin{cases} S_{i-3,j} + a \\ R_{i-3,j} + b \end{cases}$$

Here $S_{i,j}$ is the score of the alignment at point (x_i, y_j); $\mathrm{d}(x_i, y_j)$ is the weight of matching two amino acids encoded by codons (x_{i-2}, x_{i-1}, x_i) and (y_{j-2}, y_{j-1}, y_j) codons; a is the deletion initiation penalty; b is the deletion extension penalty; f is the frame-shift penalty.

The alignment graph, whose vertices are the elements of the optimum distance matrix, is shown in Figure 3. Filling of the alignment matrix begins from the element (0, 0) and terminates on the element (l, n), where l and n are the lengths of the extended nucleotide sequences.

Algorithm attempts to begin the alignment at every pair of start codons and to terminate at every pair of stop codons in the areas around annotated start and stop respectively. The alignment does not proceed beyond points where the similarity score falls below 0; in this case the nearest already aligned pair of start or stop codons is used. Thus this procedure is similar to the Smith–Waterman local alignment, although the fact that the alignment can terminate only at selected (though multiple) points resembles the global alignment of the Needleman–Wunsch type.

The traceback in the alignment matrix initiates at the pair of stop codons having the largest cumulative alignment score among all potential stops. The traceback terminates at the pair of start codons, whose alignment score is the largest among all pairs of potential starts. If thus identified gene termini differ from the annotated ones, the gene is retained for further analysis.

The extensive testing showed that this procedure is robust as regards the choice of the amino acid substitution matrix (PAM120, PAM60, PAM30 and PAM10[55]). The deletion initiation penalty was 10, the deletion extension was 2; the frame-shift penalty was 20.

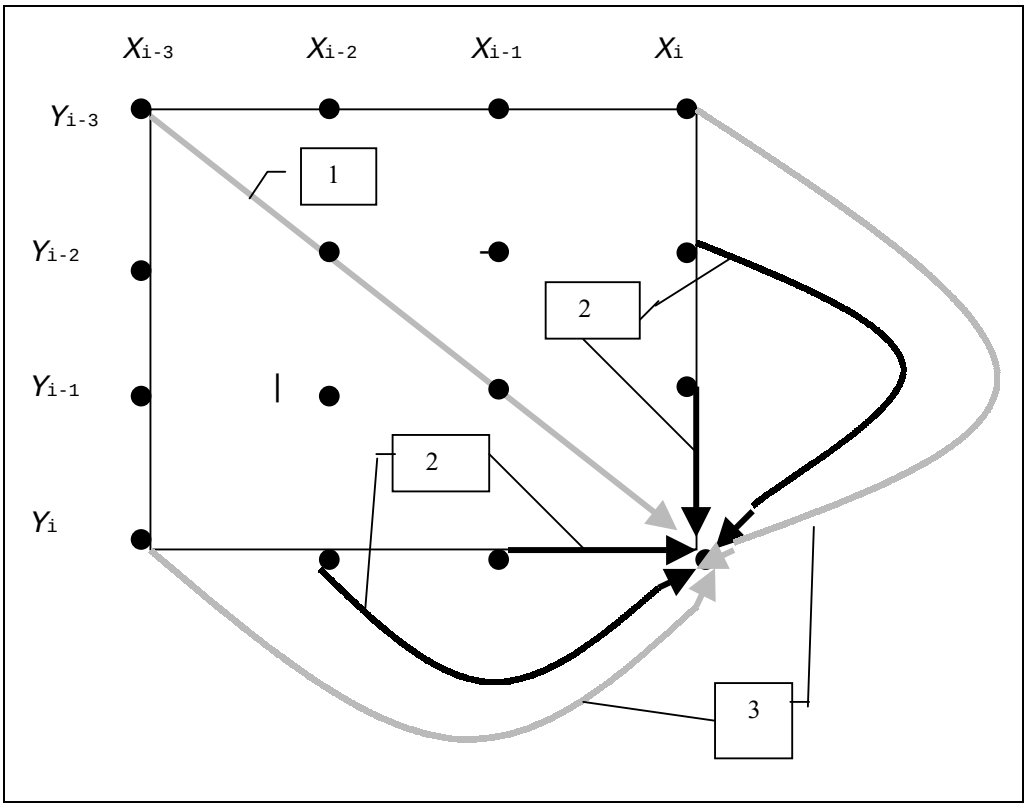

Figure 3: Fragment of the alignment graph with all types of transitions. X_i and Y_i are nucleotides: (1) from a pair of amino acids encoded by codons (X_{i-6}, X_{i-5}, X_{i-4}) and (Y_{i-6}, Y_{i-5}, Y_{i-4}) into a pair of amino acids encoded by codons (X_{i-3}, X_{i-2}, X_{i-1}) and (Y_{i-3}, Y_{i-2}, Y_{i-1}); (2) frame shift; (3) deletion of amino acids encoded by codons (X_{i-3}, X_{i-2}, X_{i-1}) or (Y_{i-3}, Y_{i-2}, Y_{i-1})

Filtering

At the filtering step possible explanations for the observed differences between the aligned and annotated gene termini are considered. Three main types of non-informative alignments have been identified during preliminary testing (Figure 4):

- Alignments that do not require correction of gene coordinates (ie confirm the existing annotation).
- Weak alignments not sufficient to suggest revision: (a) for start correction, if the relative similarity score of N-terminal region (15 per cent of the protein length) is smaller than 60 per cent, (b) for frame-shifts, if the relative similarity score of complete alignment is smaller than 40 per cent.
- Alignments with multiple transitions between the reading frames (clustered frame-shifts closer than 21 nucleotides to each other). This happens if reading frames different from the correct one encode rare amino acids with high match weight.

The alignment and filtering procedures are implemented in the program GeneCorrector.

RESULTS

Table 3 presents the number of alignments of each type for all considered pairs of genomes and the number of candidates for correction. For example, for genomes *E. coli* and *V. cholerae* with 2,300 potential pairs of orthologues, 1,254 (~55 per cent) corroborate annotation, 846 (~37 per cent) are weak alignments and 149 (~7 per cent) are alignments with multiple transitions between reading frames. The remaining 56 genes (51 alignments) in both genomes are candidates for correction of the annotation or sequencing errors.

For the retained genes, verification using a third genome was made. If correction made by the comparison of the BaseGenome (BG) with an additional genome AG*j* was confirmed in the

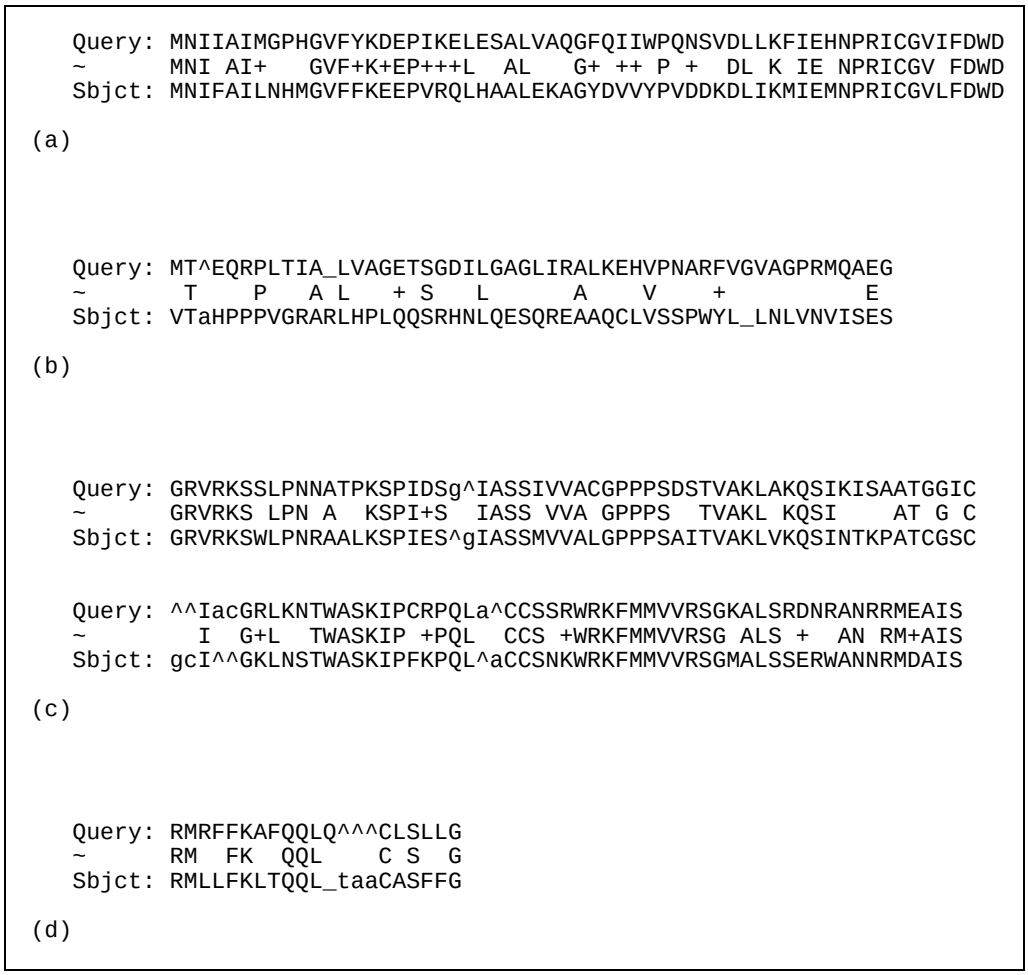

Figure 4: Alignments not leading to error correction: (a) exact coincidence with the annotation; (b) too distant genes; (c) and (d) false frame shifts

Table 3: Types of alignments obtained during testing

Pairs of organisms	Number of orthologous pairs	Coincidence with GenBank annotation	Weak alignments	Alignments with multiple transitions between reading frames	Candidates for correction (in both genomes)
E. coli–V. cholerae	2,300	1,254 (55%)	846 (37%)	149 (7%)	56
E. coli–H. influenzae	1,400	849 (61%)	452 (32%)	69 (5%)	32
E. coli–Buchnera sp.	400	235 (62%)	133 (32%)	22 (5%)	12
E. coli–X. fastidiosa	1,600	972 (61%)	563 (34%)	45 (4%)	24
V. cholerae–H. influenzae	1,050	617 (60%)	323 (33%)	75 (6%)	38
V. cholerae–Buchnera sp.	280	186 (61%)	74 (34%)	11 (4%)	18
V. cholerae–X. fastidiosa	1,200	833 (62%)	388 (33%)	54 (4%)	27
B. subtilis–B. halodurans	2,700	1,604 (60%)	946 (35%)	205 (4%)	53
B. subtilis–C. acetobutylicum	1,600	939 (58%)	577 (36%)	54 (5%)	33
P. horikoshii–P. abyssi	1,100	676 (62%)	353 (33%)	41 (4%)	38
P. horikoshii–P. furiosus	1,050	662 (63%)	333 (32%)	30 (4%)	31

comparison with one more additional genome AG_k, the new gene coordinates were accepted as final (Figure 5). For this purpose, genes from 'base' genomes (*E. coli*, *V. cholerae*, *B. subtilis* and *P. horikoshii*) were compared with orthologous genes from additional genomes of the same taxonomic groups (*H. influenzae, Buchnera* sp. and *X. fastidiosa* for *E. coli* and *V. cholerae*; *B. halodurans* and *C. acetobutylicum*

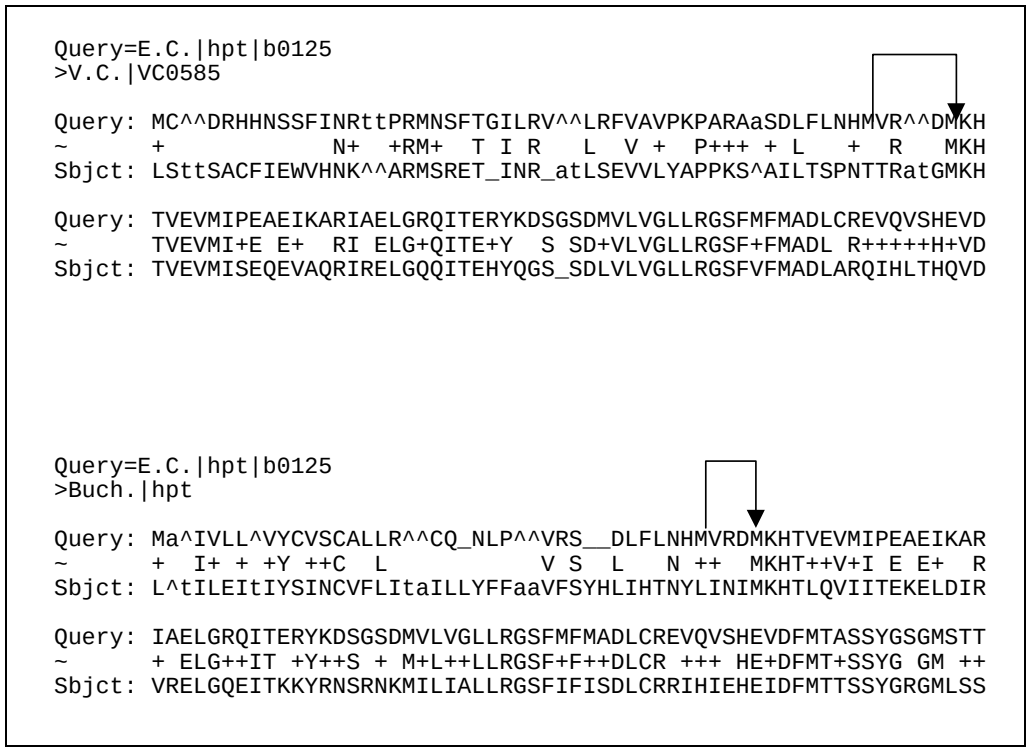

Figure 5: Correction of gene start (*hpt* from *E. coli* after comparison with *V. cholerae*) corroborated by the third genome (*Buchnera* sp.)

for *B. subtilis*; *P. abyssi* and *P. furiosus* for *P. horikoshii*).

Figure 6 illustrates the major types of errors, whereas Tables 4 and 5 presents the summary data for studied genomes. The majority of corrections in the GenBank annotation, approximately 80 per cent (dependent of the genome), were confirmed by the SWISS-PROT databank,[56] thus demonstrating high accuracy of the obtained results.

In an additional study, a strong dependency between the number of suggested corrections and the number of considered additional genomes was observed. The same procedure was applied to correction of *E. coli* genes using 32 complete or partial genomes of gamma-proteobacteria. The results were compared with 811 *E. coli* gene starts verified by N-terminal protein sequencing, extracted from the EcoGene database[57] (Table 6). This database suggests start corrections in 77 genes. Of these, 73 genes were identified by our program as well, and the remaining 4 genes were missed, because no orthologues were available for comparison. GeneCorrector suggests additional 395 start corrections, for which no experimental data are available. In no cases did GeneCorrector annotations contradict experimental data. Finally, out of 20 frame-shifts identified in the large-scale comparison, 12 are mentioned in EcoGene (only three frame-shifts were corrected when 4 genomes were used).

DISCUSSION

The above algorithms provide computational support for the gene identification experiments. We have tested the algorithm on three groups of genomes and demonstrated high reliability of predictions. Application of the comparative approach leads to a number of interesting observations. Three types of genomic sequencing and annotation errors were identified:

- Genes that had been sequenced and annotated long ago and not revised ever since. More accurate analysis showed in some cases that gene starts had been mapped incorrectly.

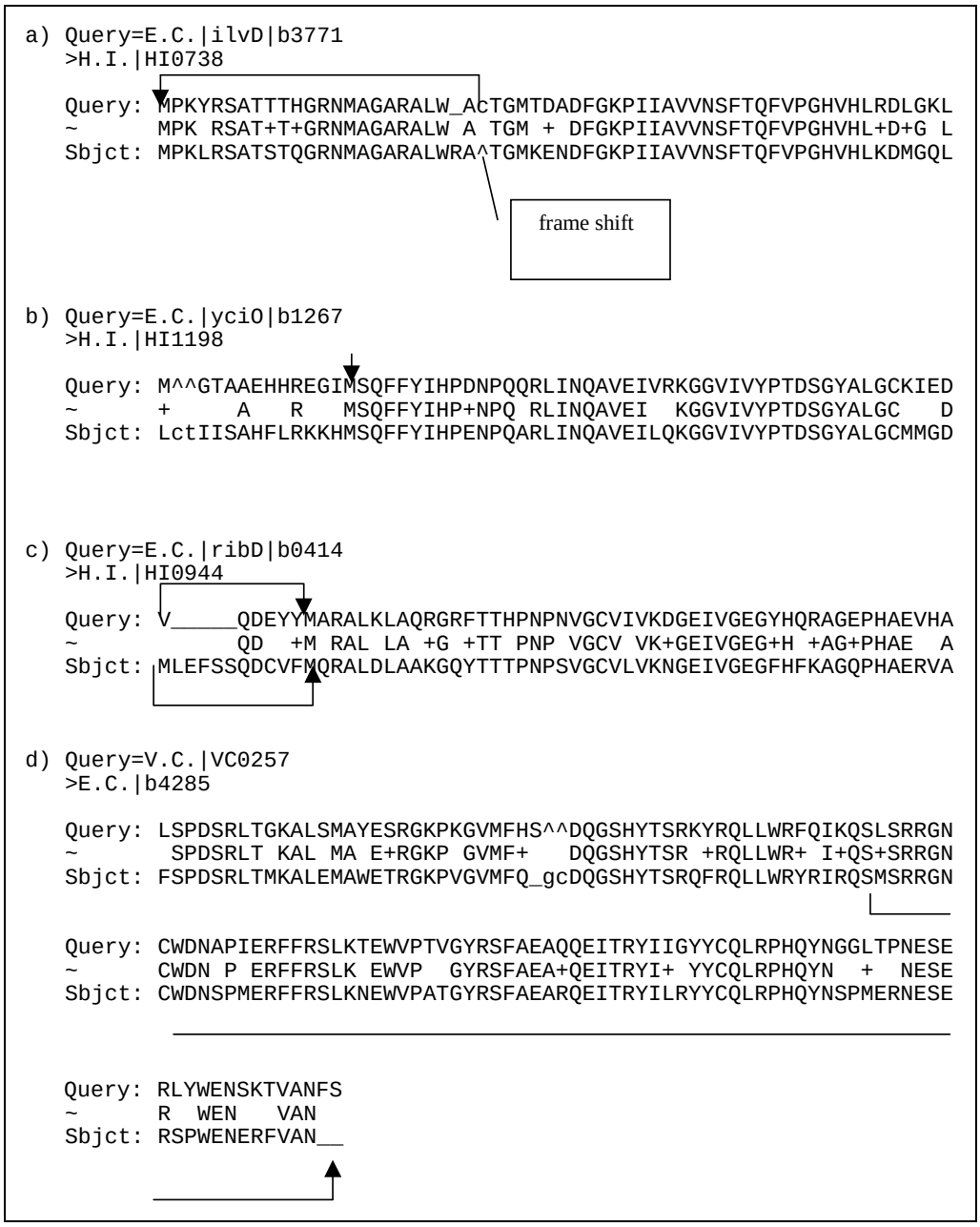

Figure 6: Examples of annotation errors: (a) gene *ilvD* from *E. coli* – frame-shift in and wrong gene start in *E. coli*; (b) gene *yciO* from *E. coli* – wrong gene start; (c) gene *ribD* from *E. coli* and *HI0944* – wrong starts of both genes; (d) hypothetical gene from *E. coli* encoding a transposase – frame-shift and wrong gene end

Table 4: Corrected gene starts and frame-shifts in 14 completely sequenced microbal genomes

Organism	Date submitted	Number of genes	Corrected starts	Frame-shifts
Escherichia coli	1997	4,288	22 (0.5%)	3
Vibrio cholerae	2000	2,736	28(1%)	5
Haemophilius influenzae	1995	1,709	15(1%)	2
Buchnera sp.	2000	564	10(2%)	5
Xylella fastidiosa	2000	2,766	5(0.01%)	2
Bacillus subtilis	1997	4,097	24(0.5%)	6
Bacillus halodurans	2000	4,066	21(0.5%)	8
Clostridium acetobutylicum	1997	3,740	22(0.5%)	5
Pyrococcus horikoshii	1998	2,058	24(1.5%)	8
Pyrococcus abyssi	2000	1,763	28(2%)	12
Pyrococcus furiosus	2000	2,208	28(1.5%)	7

Table 5: Types of errors found in the annotations of base genomes

Base genomes	Number of genes	Number of corrections	Old/obsolete annotation	Conflicting annotation (confirmed by SWISSPROT, EcoGene)	Hypothetical genes
E. coli	4,288	25	2	21	2
E. coli (large-scale comparison)	4,288	468	327	73	68
V. cholerae	2,736	33	4	26	3
B. subtilis	4,097	30	4	23	3
P. horikoshii	2,058	32	8	18	6

Table 6: Corrections of gene starts in *E. coli* using 32 additional genomes (the data for the case of only four additional genomes are in parentheses)

	GeneCorrector	No opinion
EcoGene	73 (3)	4 (74)
no opinion	395 (22)	0 (0)

- Hypothetical genes for which there is no experimental information. In such cases, the comparison corrects the results of the statistical annotation.

- Genes that have conflicting annotation in different databases.

One of the expected, but still important, results of this study was that the number of corrected starts depends on the number of considered exons. Indeed, the number of corrections suggested for *E. coli* grew from 25 to 468 when the number of additional genomes rose from 4 to 32, and the comparison with EcoGene makes it likely that most of them are valid. On the other hand, the number of missed cases decreased from 74 to only 4.

Of course, the suggested algorithm does not cover all possibilities and also does not take into account all available features. However, given its simplicity and computational flexibility, the described algorithm can be easily linked to other tools and incorporated into gene recognition software for large-scale genome annotation projects.

Acknowledgements
We are grateful to Pavel S. Novichkov and Michael Fonstein for useful discussions. This work was partially supported by grants from INTAS (99–1476), HHMI (55000309), and LICR (CRDF RBO-1268). The complete correction list and the alignment procedure are available at http://bioinform.genetika.ru/genecorrect

References

1. Overbeek, R., Fonstein, M., D'Souza, M. *et al.* (1999), 'The use of gene clusters to infer functional coupling', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 2896–2901.
2. Nielsen, H., Engelbrecht, J., Brunak, S. and von Heigne, G. (1997), 'A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites', *Int. J. Neural. Syst.*, pp. 581–599.
3. Gelfand, M. S., Mironov, A. A., Jomantas, J. *et al.* (1999), 'A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes', *Trends Genet.*, Vol. 15, pp. 439–442.
4. Neidhardt, F. (1996), '*Escherichia coli* and *Salmonella*', 2nd edn, Vol. 1, pp. 880, 916–918, 979, 982, 994–995, 2008.
5. Cole, S. T. *et al.* (2001), 'Massive gene decay in the leprosy bacillus', *Nature*, Vol. 409, pp. 1007–1011
6. Shine, J. and Dalgarno, L. (1974), 'The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites'. *Proc. Natl Acad. Sci. USA*, Vol. 71, pp. 1342–1346.
7. Stormo, G. D., Schneider, T. D., Gold, L. and Ehrenfeucht, A. (1982), 'Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*', *Nucleic Acids Res.*, Vol. 10, pp. 2997–3011.
8. Studnicka, G. M. (1986), 'Quantitative computer analysis of signal sequence

homologies in DNA', *Comput. Appl. Biosci.*, Vol. 2, pp. 269–275.

9. Barrick, D., Villanueba, K., Childs, J. *et al.* (1994), 'Quantitative analysis of ribosome binding sites in *E.coli*', *Nucleic Acids Res.*, Vol. 22, pp. 1287–1295.
10. Bisant, D. and Maizel, J. (1995), 'Identification of ribosome binding sites in *Escherichia coli* using neural network models', *Nucleic Acids Res.*, Vol. 23, pp. 1632–1639.
11. Schurr, T., Nadir, E. and Margalit, H. (1993), 'Identification and characterization of *E. coli* ribosomal binding sites by free energy computation', *Nucleic Acids Res.*, Vol. 21, pp. 4019–4023.
12. Hannenhalli, S. S. *et al.* (1999), 'Bacterial start site prediction', *Nucleic Acids Res.*, Vol. 17, pp. 3577–3582.
13. Hayes, W. S. and Borodovsky, M. (1998), in Altman, R. B. *et al.*, Eds, 'Pacific Symposium on Bioinformatics '98', World Scientific, Singapore, pp. 279–290.
14. Borodovsky, M., Lomsadze, A. and Besemer, J. (2001), 'GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions', *Nucleic Acids Res,* Vol. 29, pp. 2607–2618.
15. Suzek, B., Ermolaeva, M., Schreiber, M. and Salzberg, S. (2001), 'A probabilistic method for identifying start codons in bacterial genomes', *Bioinformatics*, Vol. 17, pp. 1123–1130.
16. Frishman, D., Mironov, A., Mewes, H.-W. and Gelfand, M. (1998), 'Combining diverse evidence for gene recognition in completely sequenced bacterial genomes', *Nucleic Acids Res.*, Vol. 26, pp. 2941–2947.
17. Grantham, R., Gautier, C., Gouy, M. *et al.* (1980), 'Codon catalog usage and the genome hypothesis', *Nucleic Acids Res.*, Vol. 8, pp. 49–62.
18. Gouy, M. and Gautier, C. (1982), 'Codon usage in bacteria: Correlation with gene expressivity', *Nucleic Acids Res.*, Vol. 10, pp. 1055–1074.
19. Ikemura, T. (1981), 'Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the expected codon in protein genes', *J. Mol. Biol.*, Vol. 146, pp. 1–21.
20. Sharp, P.M. and Li, W.-H. (1981), 'An evolutionary perspective on synonymous codon usage in unicellular organisms', *J. Mol. Evol.*, Vol. 24, pp. 28–38.
21. Shields, D. C. and Sharp, P. M., (1987), 'Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases', *Nucleic Acids Res.*, Vol. 15, pp. 8023–8040.
22. Medigue, C., Rouxel, T., Vigier, P. *et al.* (1991), 'Evidence for horizontal gene transfer in *Escherichia coli* speciation', *J. Mol. Biol.*, Vol. 222, pp. 851–856.
23. Fickett, J. W. and Tung, C. S. (1992), 'Assessment of protein coding measures', *Nucleic Acids Res.*, Vol. 20, pp. 6441–6450.
24. Eddy S. R., 'Hidden Markov models', *Curr. Opin. Struct. Biol.*, Vol. 6, pp. 361–365.
25. McIninch, J. D., Hayes, W. S. and Borodovsky, M. (1996), 'Applications of GeneMark in multispecies environments', in 'Proceedings of the 4th International Conference on Intelligent Systems in Molecular Biology', AAAI Press, Menlo Park, CA, pp. 65–75.
26. Lukashin, A. V. and Borodovsky, M. (1998), 'GeneMark.hmm: New solutions for gene finding', *Nucleic Acids Res.*, Vol. 26, pp. 1107–1115.
27. Salzberg, S. L., Delcher, A. L., Kasif, S. and White, O. (1998), 'Microbial gene identification using interpolated Markov models', *Nucleic Acids Res.*, Vol. 26, pp. 544–548.
28. Delcher, A., Harmon, D., Kasif, S. *et al.* (1999), 'Improved microbial gene identification with GLIMMER', *Nucleic Acids Res.*, Vol. 27, pp. 4636–4641.
29. Robinson, K., Gilbert, W. and Church, G. M. (1994), *Nature Genet.*, Vol. 7, pp. 205–214.
30. Borodovsky, M., Koonin, E. and Rudd, K. (1994), 'Intrinsic and extrinsic approaches for detecting genes in a bacterial genome', *Nucleic Acids Res.*, Vol. 11, pp. 4756–4767.
31. Medigue, C. *et al.* (1999), 'Detecting and analyzing DNA sequencing errors: Toward a higher quality of the *Bacillus subtilis* genome sequence', *Genome Res.*, Vol. 9, pp. 116–1127.
32. Badger, J. H. and Olsen, G. J. (1999), 'CRITICA: Coding region identification tool invoking comparative analisys', *Mol. Biol. Evol.*, Vol 16, pp. 512–524.
33. Altschul, S. F. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215, pp. 403–410.
34. Pearson, W., Wood, T., Zhang, Z. and Miller, W. (1997), 'Comparison of DNA sequences with protein sequences', *Genomics*, Vol. 46, pp. 24–36.
35. Krogh, A., Mian, I. S. and Haussler, D. (1994), 'A hidden Markov model that finds genes in *E. coli* DNA', *Nucleic Acids Res.,* Vol. 2, pp. 4768–4778.
36. Yada, T., Totoki, Y., Takagi, T. and Nakai, K. (2001), 'A novel bacterial gene-finding system with improved accuracy in locating start codons', *DNA Res.*, Vol. 8, pp. 97–106.
37. Thebault, P., Servant, F., Schiex, T. *et al.*

(2000), 'JOBIM Conference Proceedings', pp. 361–365.

38. Frishman, D., Mironov, A. and Gelfand, M. (1999), 'Starts of bacterial genes: estimating the reliability of computer predictions', *Gene*, Vol. 234, pp. 257–265.
39. Blattner, F. R. *et al.* (1997), 'The complete genome sequence of *Escherichia coli K-12*', *Science*, Vol. 277, pp. 1453–1462.
40. Heidelberg, J. F. *et al.* (2000), 'DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*', *Nature*, Vol. 406, pp. 477–483.
41. Fleischmann, R. *et al.* (1995), 'Whole-genome random sequencing and assembly of *Haemophilus influenzae* RD', *Science*, Vol. 269, pp. 496–512.
42. Shinegobu, S., Watanabe, H., Hattori, M. *et al.* (2000), 'Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp.', *Nature*, Vol. 407, pp. 81–85.
43. Simpson, A. J. G. *et al.* (2000), 'The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis', *Nature*, Vol. 406, pp. 151–156.
44. Kunst, F. *et al.* (1997), 'The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*', *Nature*, Vol. 390, pp. 249–256.
45. Hideto Takami *et al.* (2000), 'Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*', *Nucleic Acids Res.*, Vol. 28, pp. 4317–4331.
46. Nolling, J. *et al.* (2001), 'Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*', *J. Bacteriol.*, Vol. 183, pp. 4823–4838.
47. Kawarabayashi, Y. *et al.* (1998), 'Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3', *DNA Res.,* Vol. 5, pp. 55–76.
48. Helig, R. (1999) unpublished, EMBL/ GenBank/DDBJ databases.
49. Full *Pyrococcus furiosus* genome sequence – URL: www.genome.utah.edu.
50. Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000), 'The COG database: New developments in phylogenetic classification of proteins from complete genomes', *Nucleic Acids Res.,* Vol 28, pp. 6–33.
51. Tatusov, R. L, Koonin, E. V. and Lipman, D. J. (1997), 'A genomic perspective on protein families', *Science*, Vol. 278, pp. 631–637.
52. Altschul, S., Madden, T., Schaffer, A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res.,* Vol 25, pp. 3389–3402.
53. Batalyuk, M. V., Novichkov, P. S., Mironov, A. A. and Gelfand, M. D. (2000), 'Software for orthology analysis in complete bacterial genomes BGRS'2000', in 'Proceedings of the 2nd International Conference on Bioinformatics of Genome Regulation and Structure', pp. 26–27.
54. Smith, T. F. and Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J. Mol. Biol.,* Vol. 147, pp. 195–197.
55. Altschul, S. F. (1991), 'Amino acid substitution matrices from an information theoretic perspective', *J. Mol. Biol.*, Vol. 219(3), pp. 555–565.
56. URL: http://www.expasy.ch
57. Rudd, K. E. (2000), 'EcoGene: A genome sequence database for *Escherichia coli* K-12', *Nucleic Acids Res.*, Vol. 28, pp. 60–64.

Fickett, J. W. (1994), 'Inferring genes from open reading frames', *Comput. Chem.*, Vol. 18, pp. 203–205.

Huang, X. (1996), 'Fast comparison of a DNA sequence with a protein sequence database', *Microb. Compar. Genomics*, Vol. 1, pp. 281–291.

Gish, W. and States, D. J. (1993), 'Identification of protein coding regions by database similarity search', *Nature Genet.*, Vol. 3, pp. 266–272.

Gelfand, M. S., Mironov, A. A. and Pevzner, P. A. (1996), 'Gene recognition via spliced sequence alignment', *Proc. Natl Acad. Sci. USA*, Vol. 93, pp. 9061–9066.

Mural, R. J. (2000), 'ARTEMIS: a tool for displaying and annotating DNA sequence', *Brief. Bioinform.*, Vol. 1, pp. 199–200.