DNA Research **8**, 97–106 (2001)

A Novel Bacterial Gene-Finding System with Improved Accuracy in Locating Start Codons

Tetsushi Yada,[1,†] Yasushi Totoki,[1] Toshihisa Takagi,[2] and Kenta Nakai[2,*]

Genomic Sciences Center, RIKEN, Yokohama 230-0045, Japan[1] *and Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan*[2]

(Received 6 February 2001; revised 17 April 2001)

Abstract

Although a number of bacterial gene-finding programs have been developed, there is still room for improvement especially in the area of correctly detecting translation start sites. We developed a novel bacterial gene-finding program named GeneHacker Plus. Like many others, it is based on a hidden Markov model (HMM) with duration. However, it is a 'local' model in the sense that the model starts from the translation control region and ends at the stop codon of a coding region. Multiple coding regions are identified as partial paths, like local alignments in the Smith-Waterman algorithm, regardless of how they overlap. Moreover, our semiautomatic procedure for constructing the model of the translation control region allows the inclusion of an additional conserved element as well as the ribosome-binding site. We confirmed that GeneHacker Plus is one of the most accurate programs in terms of both finding potential coding regions and precisely locating translation start sites. GeneHacker Plus is also equipped with an option where the results from database homology searches are directly embedded in the HMM. Although this option does not raise the overall predictability, labeled similarity information can be of practical use. GeneHacker Plus can be accessed freely at http://elmo.ims.u-tokyo.ac.jp/GH/.

Key words: gene finding; hidden Markov model; start codon; genome analysis; ribosome binding site

1. Introduction

So far, the entire genomes of about 30 microbial species have been sequenced and more and more novel genomes are being sequenced each day. When the whole genome of an organism is sequenced, the next logical step is to locate the positions of possible coding regions. A number of such gene-finding programs have been developed,[1–10] many of which are based on probabilistic models of gene structure. For example, GeneMark.hmm[5] employs the hidden Markov model (HMM) with duration[11,12] while GLIMMER[4,9] employs an interpolated Markov model, which combines the models with various orders. Since bacterial genes can be assumed to contain no introns, relatively high prediction accuracy has been achieved in bacterial gene-finding. Indeed, both GeneMark.hmm and GLIMMER can identify most of the annotated genes and consequently have become widely used. However, some problems remain, such as the relatively low accuracy of locating the precise position of translation start sites. To further improve gene-finding accuracy, it is essential to incorporate various kinds of signal information such as the ribosome binding signal (RBS; reviewed in 13) into prediction schemes. However, this is not straightforward because most of the start sites of annotated coding regions have not been experimentally verified. In fact, Lukashin and Borodovsky[5] observed that their GeneMark.hmm, which uses the information of the RBS in its post-processing step, performs better in precisely locating start positions when using smaller data obtained from a proteome project.[14] The same group also developed a 'frame-by-frame' algorithm, which shows better accuracy in precisely predicting genes.[10] Prediction of bacterial start sites was the focus of Hannenhalli et al.,[15] who developed a specialized algorithm that detects various sequence features of start sites. They also paid great attention in assessing the accuracy of their algorithm, collecting relatively reliable data from various sources. However, since their method was not designed as a module for gene-finding programs, it cannot be directly integrated in the usual gene-finding schemes. In this paper, we report an improvement in our previous algorithm.[3] The new program, called GeneHacker Plus, in most cases shows better accuracy than existing programs. Surpris-

Communicated by Minoru Kanehisa

* To whom correspondence should be addressed. Tel. +81-3-5449-5619, Fax. +81-3-5449-5434, E-mail: knakai@ims.u-tokyo.ac.jp

† Present address: Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan

ingly, it can detect start sites with better accuracy than Hannenhalli et al.'s method in spite of the simplicity of its algorithm. We also report our attempt to integrate the results of similarity searches into the HMM prediction scheme.

2. Materials and Methods

2.1. Collection of known coding regions

We used the genome sequences and the annotations of their coding regions from the following organisms: *Archaeoglobus fulgidus*[16] (GenBank accession number: AE000782), *Bacillus subtilis*[17] (AL009126), *Escherichia coli*[18] (U00096), *Helicobacter pylori*[19] (AE000511), *Methanobacterium thermoautotrophicum*[20] (AE000666), and *Synechocystis* sp. PCC6803[21] (NC000911). For more reliable data of translation start sites, we used other sources of data, most of which are the same as that compiled by Hannenhalli et al.[15] For *E. coli*, the proteome data of Link et al.[14] were used (184 gene products with experimentally confirmed N-termini). For *B. subtilis*, two kinds of data were used: (a) 1246 'non-y' (i.e., experimentally characterized) sequences and (b) 58 sequences confirmed by comparison with homologous sequences of *B. halodurans*[22] (see below). Note that the start sites of the former sequences are not always verified experimentally. For *Pyrococcus furiosus*, 241 genes that were estimated by comparison with *P. horikoshii* were used. Since the complete annotated genome sequence of *P. furiosus* has not yet been published, we used these genes for the predictability assessment of coding sequences as well as the assessment of start site prediction. For *Synechocystis*, Sazuka et al.'s proteome data (107 genes) were used.[23]

Estimation of translation start sites based on the comparison of *B. subtilis* and *B. halodurans* was done as follows. First, using the *Entrez* system of NCBI,[24] 173 known amino acid sequences of *B. halodurans* were extracted. Second, for each of the amino acid sequences, the DNA sequence 90 residues upstream of the start site was extracted from its original nucleotide sequence and was mechanically translated to elongate the amino acid sequence on its N-terminal side by 30 codons (thus, this region can contain stop codons). Third, the TBLASTN program was used to compare these amino acid sequences in all frames with the genome sequence of *B. subtilis*.[25] Lastly, like Hannenhalli et al.,[15] 58 likely translation start sites were selected by inspecting their alignments.

2.2. Architecture of the HMM

As shown in Fig. 1, the basic architecture of GeneHacker Plus is rather simple; it consists of a model of the upstream region and two models for the coding region (CDS). As an option, the model can also include a parallel branch of the coding region for incorporating the result of a homology search (shown in dashed box). It is also noteworthy that the model itself does not cover the entire genome but only fits into one gene. Therefore, when the model is applied to an entire genome sequence, predicted genes are detected as 'partial' matches, like local alignments in the Smith-Waterman algorithm[26] as shown in Fig. 1(b) (in this sense, the standard Viterbi algorithm corresponds to 'global' alignment). A similar approach has been taken in HMMER where a variant of profile HMM was used to align a query sequence with a given profile.[12] The threshold value for gene detection is set so as to maximize the average of the sensitivity and specificity when applied to the original genome sequence containing the training data. In the case of *P. furiosus*, the lowest probability in the training set was used because the training data are not complete. Because of this local search strategy, GeneHacker Plus can detect overlapping genes without difficulty. However, from *a priori* knowledge, when there are two predicted coding regions that overlap more than 40% of either one's length in any direction, the one with the lower score is discarded.

2.3. Modeling of translation control regions

In prokaryotes, translation start sites are usually specified by the RBS[13] (also called the Shine-Dalgarno sequence). Thus, our initial model of the translation control region simply consisted of the upstream element and the subsequent spacer. The model of the element was constructed as a weight matrix of the conditional probability that a base is observed next to a given base, like Salzberg's conditional probability matrix for the prediction of splice sites.[27] The spacer region was modeled using the duration model of HMM, like GeneMark.hmm and others.[5,11] However, a proteome analysis of *Synechocystis* sp. suggests that additional conserved elements may exist in some species.[23] In addition, to deal with the problem of relatively little training data, we used the technique of pseudocounts, which enables us to avoid using exaggerated statistics.[12] Thus, we set the algorithm to construct the model of the translation control region as follows. First, the upstream 25-bp segment from the start codon of each coding region in the training set is extracted and these segments are multiply aligned,[28] excluding internal gaps. Second, a highly conserved, consecutive segment is identified using the χ^2 test at each position in the alignment. Since it is risky to set a single pre-defined cut-off value for various analyses, we determine the value manually (the only non-automatic step). The extracted segment is regarded as the RBS and the number of dinucleotides observed at each position within the segment is counted. As a pseudocount, 'one' count revised by the total base composition (i.e., four times the ratio of each nucleotide observed in the whole genome) is added for each nucleotide. Then, the conditional probability matrix of the RBS is constructed from these counts (a sample for the *E. coli* data is given in Table 1(a)). Third, the matrix

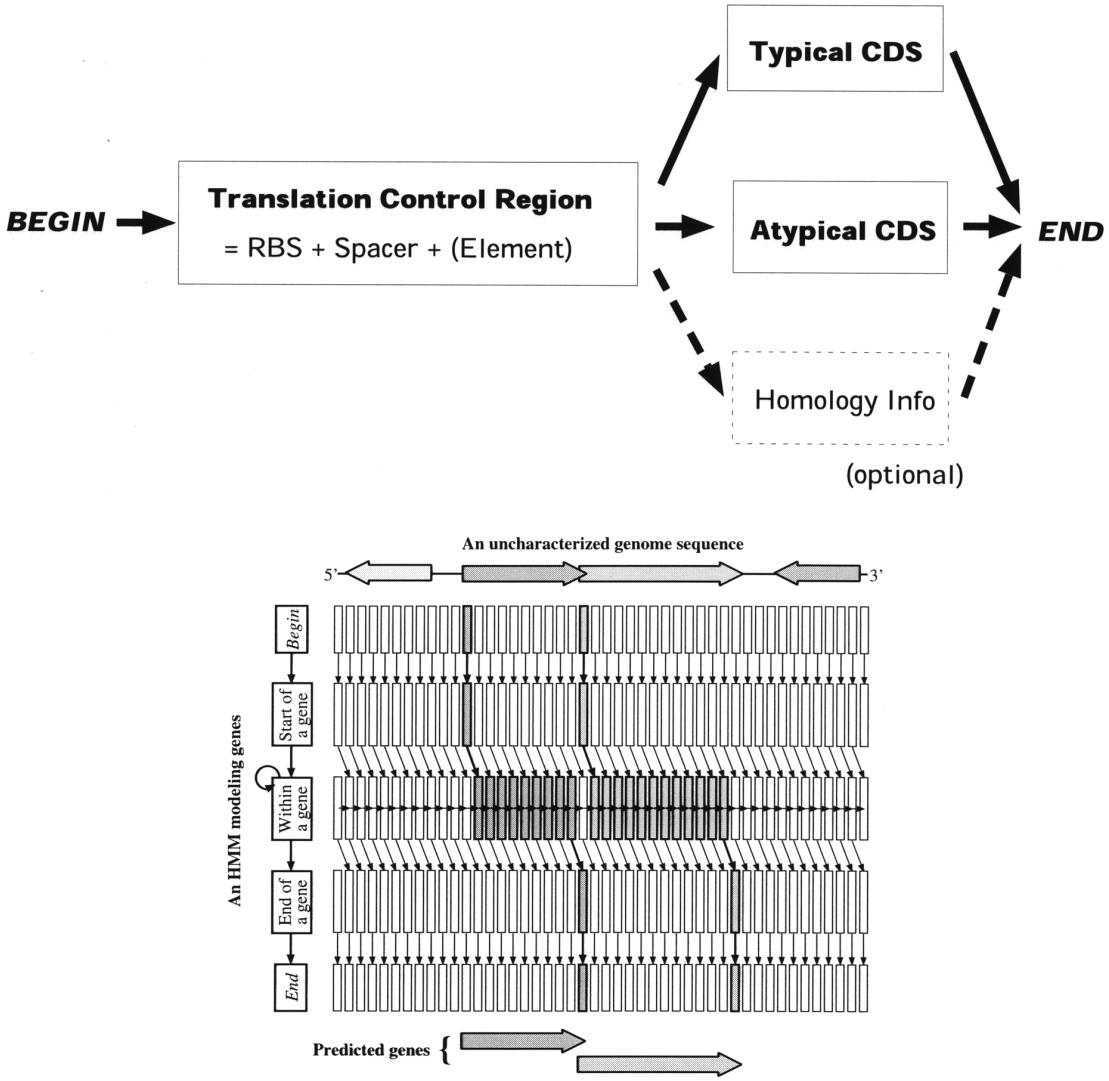

Figure 1. (a) Architecture of the HMM used. The translation control region usually consists of an RBS and a spacer but it can contain an additional element. Coding regions (CDSs) consists of two categories. The third branch for homology information is optional (shown in dashed box). (b) Schematic representation of parsing with GeneHacker Plus. The vertical axis represents a simplified gene model while the horizontal axis represents a genome sequence. Rectangles represent potential matching states between a region in the genome and a state of the corresponding HMM. Like the dynamic programming matrix, optimal matching states are selected (shown in shaded rectangles). Note that overlapping genes can be also selected. Genes in the opposite direction are detected in the second run against the complementary strand.

is applied to the original 25-bp segment and the positions of the RBS and the downstream spacer region are assigned. Fourth, the spacer sequences are simultaneously aligned without internal gaps to identify other conserved regions, if any, using the χ^2 test. If there are additional conserved regions, the above procedures are repeated to construct the conditional probability matrix and the positions of the element are re-assigned in the original sequence. Lastly, the model(s) of the remaining spacer region(s) are constructed based on their length distribution (a Gaussian curve was used for approximation) and on their dinucleotide composition considering the pseudocounts. Samples for *E. coli* data are given in Fig. 2 and Table 1(b). Note that there are two kinds of training data (i.e., either the original annotation of the whole genome or a more reliable subset as explained above) of known start codons in *E. coli*, *B. subtilis*, *P. furiosus*, and *Synechocystis* sp. (see Results).

2.4. Modeling of coding regions

The model for coding regions was constructed in a similar way to GeneMark.hmm[5] but the classification of typical and atypical coding regions was done more systematically. First, a basic model is constructed using the

Table 1. (a) Conditional probability matrix (%) of the *E. coli* RBS based on confirmed data.

P | i	1	2	3	4	5	6
$P(A_i\|A_{i-1})$	40	68	5	3	80	23
$P(C_i\|A_{i-1})$	40	12	6	7	10	10
$P(G_i\|A_{i-1})$	40	9	82	83	0	49
$P(T_i\|A_{i-1})$	40	12	7	6	10	17
$P(A_i\|C_{i-1})$	29	55	8	6	61	26
$P(C_i\|C_{i-1})$	29	13	6	13	16	17
$P(G_i\|C_{i-1})$	29	10	72	75	6	39
$P(T_i\|C_{i-1})$	29	22	13	7	16	17
$P(A_i\|G_{i-1})$	21	65	8	2	72	15
$P(C_i\|G_{i-1})$	21	13	11	5	8	17
$P(G_i\|G_{i-1})$	21	9	76	90	7	45
$P(T_i\|G_{i-1})$	21	14	6	3	13	23
$P(A_i\|T_{i-1})$	11	40	3	5	59	25
$P(C_i\|T_{i-1})$	11	25	5	16	12	12
$P(G_i\|T_{i-1})$	11	13	78	60	7	36
$P(T_i\|T_{i-1})$	11	21	14	20	22	27

(b) Conditional probability matrix (%) of the spacer between the RBS and start codon in confirmed *E. coli* data.

after | before	A	C	G	T
A	42	38	35	35
C	20	23	23	18
G	18	16	14	13
T	20	23	28	34

Nucleotide percent composition was: (A, C, G, T) = (38, 21, 16, 26).

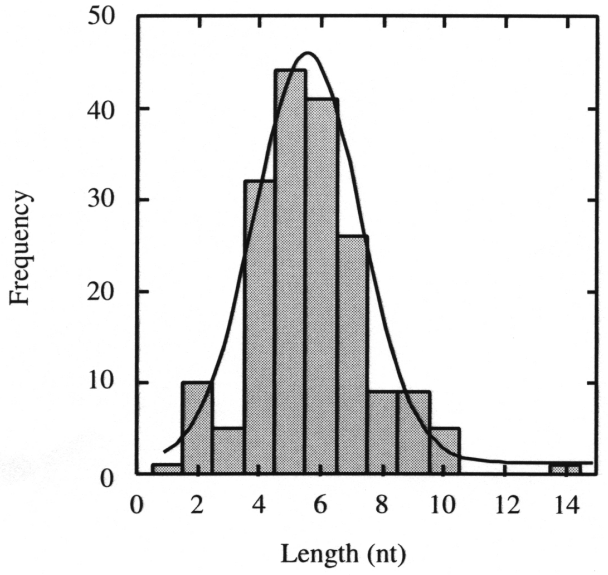

Figure 2. Length distribution of the spacer between the RBS and start codon in *E. coli* data that were confirmed experimentally. A Gaussian curve approximating the distribution is also shown.

dicodon statistics and the length distribution for the entire coding sequences in the training data (Fig. 3). Thus, we used four groups of probability values: (1) the probability of a codon used as a start codon, P(*first*); (2) the conditional probability of a second codon following a given start codon, P(*second* | *first*), because we observed that the second codons are somewhat special compared with other internal codons; (3) the conditional probability of an internal codon following a previous given internal codon, P(*internal* | *internal*); and (4) the probability of a stop codon following a given internal codon, P(*stop* | *internal*). To deal with cases when the number of known coding regions is small, we also added a pseudocount (i.e., 64 times the total codon frequency) for each codon count. Unlike the model of the translation control region, a binomial distribution was optimized to fit the observed length distribution of coding regions and this distribution was used to model all length distributions in the following steps. Second, the log-odds score for each coding sequence in the training set is calculated for the obtained model and the sequence is classified into one of two groups based on the score. The threshold value of this classification is not essential and we set the value so that 30% of the sequences are assigned to the second (atypical) group. Third, for each group of coding sequences, a new model is constructed and they are joined in parallel (Fig. 1). Fourth, the parallel model is applied to the training coding sequences and the optimal path selected is monitored for each sequence. The sequences are then reclassified based on the selection. The last two steps are repeated until the members of each group converge (we ignored any changes of less than 1% of the total number). When the constructed model is applied to genome sequences, the minimum length of the coding regions in the training data was used as the cut-off length.

By default, GeneHacker Plus does not use any homology information but it has an option where it can directly include the result of similarity searches as a third branch of the coding region (Fig. 1). In this case, TBLASTN searches are conducted using a bacterial subset of the SWISS-PROT database[29] as queries against the genomic sequence (the sequences of the same organism as the genome data are not used). When any significant matches are obtained (i.e., the length of alignment exceeds 150 nucleotides and its E-value is less than 1.0E-50), corresponding nucleotide positions in the genome are marked with the information content (bit value) of the alignment. Then, these values are used for obtaining the output symbol emission probability when parsing the genome sequence with the HMM.

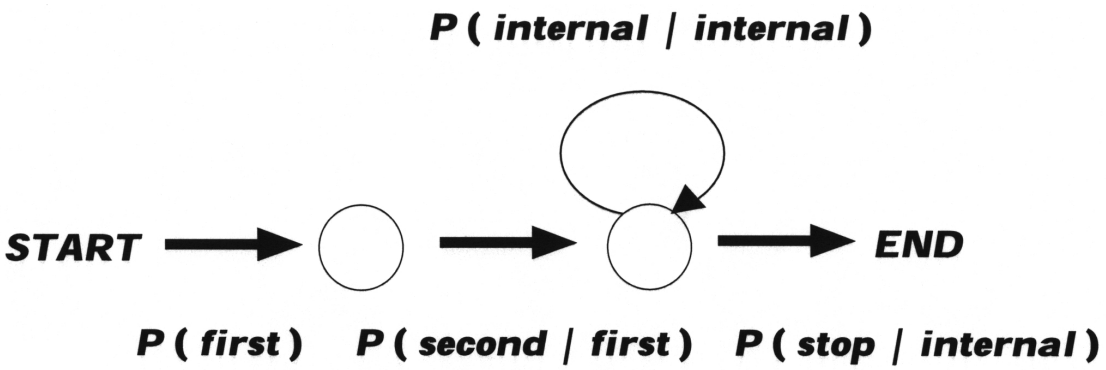

Figure 3. Basic model of coding regions. This model is used to build an initial model of the overall coding regions and is also used to build typical and atypical models, as described in the text.

3. Results

3.1. Obtained models

We constructed HMMs of the gene structure of six species. The ratio of numbers between typical genes and atypical genes are 90 : 10 for *A. fulgidus*, 84 : 16 for *B. subtilis*, 80 : 20 for *E. coli*, 90 : 10 for *H. pylori*, 90 : 10 for *M. thermoautotrophicum*, and 88 : 12 for *Synechocystis* sp. We also examined the distribution of transition probability between neighboring codons for both typical and atypical genes of each genome (data not shown). Roughly speaking, the variance of such probability distributions between genes was smaller in typical genes compared to atypical genes. This suggests that 'atypical' genes cannot be clustered into a single category (i.e., they are not of the same origin).

Another observation was made on the structure of the translation control region (Fig. 4). For all genomes except that of *Synechocystis*, the only well-conserved region was the so-called RBS. In *Synechocystis* sp., the RBS and another short element just upstream of the start codon were found to be well-conserved. It should be noted that this additional element was only significant when we used the reliable subset of start-site information and further that the element has already been proposed by the original authors.[23] We could not find a clear correlation between the quality of data (reliable data are marked with an asterisk in Fig. 4) and the conservation degree of the RBS. The RBS of *B. subtilis* appears a somewhat specific in the sense that it is longer and more strongly conserved than RBSs of other species. In addition, its spacer region is also slightly longer.

3.2. Prediction of annotated coding regions

In Table 2 (a), the accuracy of GeneHacker Plus is estimated by its ability to find the entire coding region of all regions included in the training data (self-evaluation). In this stage, we assumed that all annotations are correct and that there remain no uncharacterized coding regions to calculate specificity. Note that different data were used to construct the model of the translation control region in *B. subtilis*, *E. coli*, and *Synechocystis* sp. In column 'E', only genes for which both start and stop codons were correctly predicted were counted as positives; in column 'A', genes having the same stop codons as the annotation data were counted as positives. Both the values of sensitivity (Sn) and specificity (Sp) are given when available. For *P. furiosus*, Sp cannot be calculated because only a subset of coding regions is known. In total, GeneHacker Plus predicted 1809 genes in this genome. Corresponding results of the previous version of GeneHacker, 'frame-by-frame' algorithm, and GLIMMER (ver.2) are also shown.[3,9,10] The result of the 'frame-by-frame' algorithm was taken from the literature. Since GLIMMER was freely available, we set its sensitivity to that of GeneHacker Plus and compared their specificity. Because GLIMMER tends to report very high scores for positively-predicted genes, it was difficult to set an appropriate threshold to change its sensitivity to a specified value. As we found that longer predicted coding regions are more reliable, we first selected the coding regions with its maximum threshold value, 99, and then further selected in decreasing order of lengths until the sensitivity becaomes equal to that of GeneHacker Plus. GeneHacker Plus outperforms its previous version in most cases and the improvements are sometimes highly significant. In the 'Exact' view, the sensitivity of GeneHacker Plus is slightly inferior to that of the 'frame-by-frame' algorithm (and is much worse in *M. thermoautotrophicum*). However, in the 'Approximate' view, the sensitivity of GeneHacker Plus is at an almost comparative level with that of other programs and its specificity is significantly better.

To confirm that our program had not over-learned the training data, we also tested its accuracy by a 10-fold cross validation. Here the specificity with the exact cri-

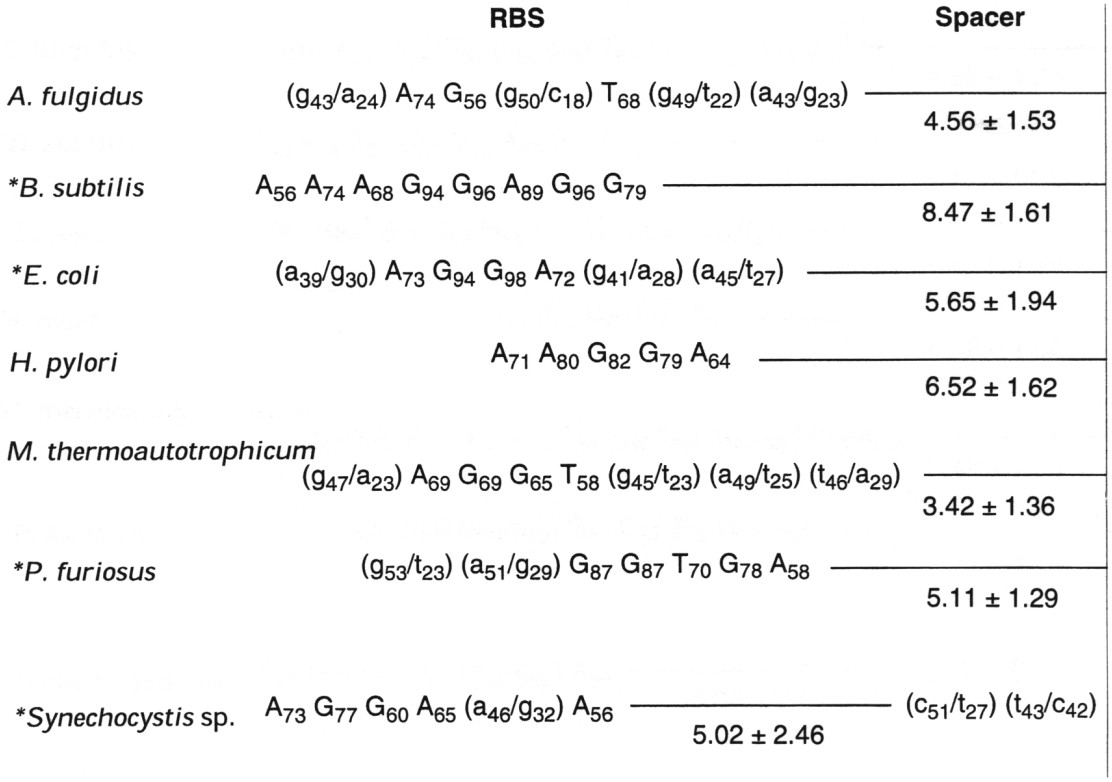

Figure 4. Model of the translation control region for each organism. Species names are shown with an asterisk if special data (other than the GenBank annotation) were used to construct the model. The consensus sequences of the RBSs are represented in conventional notation, followed by a model of the spacer regions (average length and standard deviation are shown). Note that the actual models are constructed from more complicated conditional probabilities (see Table 1). In *Synechocystis* sp., an additional element was discovered.

terion was not shown because the annotation information, especially the start site information, is not always reliable. As shown in Table 2(b), the results are rather insensitive to this procedure; only decreases of a few percent are observed in most cases. Corresponding results were also calculated using GLIMMER (its sensitivity was set to the same value with that of GeneHacker Plus). Although there are still uncharacterized genes, GeneHacker Plus seems to outperform GLIMMER in terms of specificity.

3.3. Prediction of start codons

Strictly speaking, the annotation information of coding regions is not reliable enough for the assessment of prediction accuracy. Therefore, we used a similar approach to that of Hannenhalli et al.[15] to assess the accuracy of the start-site prediction. Note that they predicted the precise position of the start site of a given open reading frame (ORF). Therefore, our prediction condition is more stringent than theirs. Nevertheless, GeneHacker Plus performs either comparable to or slightly better than Hannenhalli et al.'s method, as shown in Table 3. As for *B. subtilis*, they used the 'non-y' genes but these data are still not reliable enough. Thus, we prepared another data set based on the sequence conservation between *B. subtilis* and *B. halodurans*, similar to the way that Hannnhalli et al.[15] compared *P. furiosus* and *P. horikoshii* to obtain the *P. furiosus* data. Although the obtained data size is rather small (58 sequences), GeneHacker Plus shows better predictability on this set. In addition, we examined two kinds of additional data of *B. subtilis* for the assessment. First, the N-termini of 30 proteins were determined experimentally (T. Kodama, personal communication). GeneHacker Plus correctly predicted 24 of them; of the other 6 proteins, 5 had start positions more than 100 residues apart from the annotated sites (data not shown). It is possible that these discrepancies are due to other reasons. Second, seven additional genes were experimentally identified (reported at the home page of the Pasteur Institute; http://genolist.pasteur.fr/SubtiList/).[30] Since these genes had not been included in the original annotation of the *B. subtilis* genome, their detection has been regarded as false positives (in even our results in Table 2!). We found that GeneHacker Plus correctly predicts four of them. It is also possible that some of the remaining

Table 2. (a) Result of self-evaluation of the model.

Genomes		GH+		GH		Frame-by-frame		GLIMMER	
		E	A	E	A	E	A	E	A
A. fulgidus	Sn	75.2	97.5	74.8	96.5	78.7	98.0	75.2	97.5*
	Sp	74.4	96.5	73.8	95.2	71.7	88.3	71.8	93.1
B. subtilis	Sn	83.3	97.5	64.5	94.7	85.8	98.4	57.4	97.5*
	Sp	82.7	96.8	65.5	96.3	79.3	90.1	55.6	94.3
E. coli	Sn	73.9	96.4	69.3	93.6	75.7	93.9	74.1	96.4*
	Sp	75.2	98.1	72.8	98.4	74.2	91.9	73.2	95.3
H. pylori	Sn	84.0	96.8	78.4	93.7	86.4	96.7	74.4	96.8*
	Sp	83.6	96.4	80.3	96.0	82.0	91.3	68.8	89.5
M. thermoautotrophicum	Sn	64.6	98.3	64.2	94.5	77.6	96.5	85.9	98.3*
	Sp	65.1	99.0	67.3	99.2	73.8	91.3	75.3	86.1
P. furiosus	Sn	90.5	98.8	85.5	98.8	–	–	–	–
	Sp	–	–	–	–	–	–	–	–
Synechocystis sp.	Sn	82.1	97.9	77.5	93.4	85.9	97.3	70.7	97.9*
	Sp	82.3	98.1	82.0	98.9	80.5	90.6	68.7	95.1

All values are given in percentages (%). GH+: GeneHacker Plus (this work), GH: GeneHacker (previous version), E: Exact, A: Approximate, Sn: sensitivity, Sp: Specificity. The values of the 'frame-by-frame' algorithm were taken from Ref. 10 and the specificity of GLIMMER (ver.2) was calculated by us and was adjusted so that it had the same sensitivity as GH+.

(b) Result of 10-fold cross-validation.

Genomes	GH+			GLIMMER*	
	Exact	Approx.		Exact	Approx.
	Sn (%)	Sn (%)	Sp(%)	Sn (%)	Sp (%)
A. fulgidus	73.0	95.8	96.3	73.7	93.6
B. subtilis	81.6	96.2	97.0	56.3	91.9
E. coli	72.8	95.9	98.0	73.1	84.7
H. pylori	81.2	94.7	95.6	72.2	94.2
M. thermoautotrophicum	60.6	95.7	98.4	85.2	97.2
P. furiosus	88.0	98.3	–	–	–
Synechocystis sp.	80.2	96.5	98.0	69.0	96.0

*sensitivity was matched to that of GeneHacker Plus by the procedure described in the text.

false positives will turn out to be true positives in the future once more experiments are performed.

4. Discussion

In this work, we introduced a novel bacterial gene-finding program, GeneHacker Plus, and showed that it is one of the most accurate programs in this field both for finding coding regions and for predicting translation start sites. We used dicodon statistics rather than the more general hexamer statistics. Although the latter includes the former, the differences do not seem to be significant.[31] It is likely that our careful modeling of the first and second codons made some contribution to the higher accuracy of our method. Another reason for the high accuracy maybe the careful detection of upstream elements. Because of its features, the ability of GeneHacker Plus to detect exact positions of start sites is comparable to or better than a specifically-developed method, if we consider that the start sites of *Bacillus* 'non-y' genes are not guaranteed to be correct (Table 3). Although there maybe other reliable methods to detect RBSs (e.g., Tompa's work[32]), these methods have not fully considered the possibility that positions other than the Shine-Dalgarno sequence may be conserved. Such an example of extra conservation was observed around the confirmed start sites of *Synechocystis* sp. Although similar elements have not yet been discovered in other species, they may also exist because systematic N-terminal determination of proteins is necessary to characterize them.

Table 3. Accuracy (%) of start-site prediction.

Genomes	#genes	Self-evaluation		Cross-validation	
		GH+	Hannenhalli	GH+	Hannenhalli
B. subtilis(a)	1246	89.4	92.6	88.2	90.4
B. subtilis(b)	58	100.0	–	89.7	–
E. coli	184	96.7	94.0	95.7	84.9
P. furiosus	241	90.5	93.5	88.0	86.6
Synechocystis	107	90.7	–	91.6	–

'GH+' means GeneHacker Plus. The values in the 'Hannenhalli' columns were taken from Ref. 14. *B. subtilis* (a) are the 'non-y' genes while *B. subitlis* (b) were obtained from sequence comparison. For more details of the data, see Materials and Methods.

For a careful treatment of such uncharacterized elements GeneHacker Plus requires a manual step to set a threshold parameter, but all other parameters and models are constructed automatically. Unless sufficiently reliable start site information is available, it is reasonable to run the program fully automatically assuming only the presence of an RBS in the translation control region.

GeneHacker Plus shows relatively high prediction specificity as well as high sensitivity (Table 2). Generally speaking, it is difficult to exactly assess the sensitivity because apparent false positives may be identified later as true genes. In fact, we experienced such cases in *B. subtilis*; seven new coding regions were found after the release of the annotation. GeneHacker Plus turned out to have correctly predicted four of them. In spite of this kind of difficulty in the specificity assessment, we still think our estimation is sound because our data includes two extensively studied bacteria (*E. coli* and *B. subtilis*); it seems unlikely that these bacteria still have hundreds of false positives. For this reason, our philosophy in developing gene-finding programs is that it is preferable for bacterial gene-finding systems to minimize the number of potential false positives rather than maximize sensitivity.

Perhaps the most remarkable point of GeneHacker Plus is the simplicity of its architecture. For example, to raise the prediction accuracy of translation initiation sites, GeneMark.hmm from Borodovsky's group is equipped with a post-processing step that examines the RBS.[5] More recently, the same group proposed a 'frame-by-frame' algorithm to improve the predictability both in locating start sites and in dealing with overlapping genes.[10] In contrast, GeneHacker Plus takes a simpler approach to model the bacterial gene structure; it includes a model of translation control region within the HMM, and the HMM is applicable to the entire genome regardless of the existence of overlapping genes.

As for the computation time, GeneHacker Plus takes about 5 min to parse a 100-kb region on a 333 MHz UltraSPARC workstation. This may be slower than GeneMark.hmm (claims 1 min per 100 kb on an unknown platform),[5] but the computation time is not a serious practical consideration when analyzing a newly-determined bacterial genome. Indeed, GeneHacker Plus has been used for the analyses of several species including *Bacillus halodurans*[33] and *Buchnera* sp. APS.[34] It is also noteworthy that GeneHacker Plus outputs sufficiently reliable prediction results even though only limited numbers of known genes were used as training data in these cases.

GeneHacker Plus also has an option to include the results of a database homology search in its prediction. In gene finding, such an approach was tried by Robison et al.[35] and has been elaborated by M. S. Gelfand's group[36] as a 'spliced alignment' for eukaryotic genomes. Gelfand's group also used similarity searches to obtain a seed set of ORFs, which was used to derive various statistics for further prediction.[7] More recently, the direct integration of homology information in HMMs has been tried in large-scale gene detection from the *Drosophila* genome.[37,38] For example, Krogh reported that the use of reliable database matches with his HMM greatly improved its sensitivity.[37] We tested a similar approach to see if it could raise the predictability in bacteria. To our surprise, such homology information did not improve prediction accuracy although our results may have been due to the nature of the query sequences used. It is possible that, because the quality of bacterial gene finding has reached new heights, sequences with relatively low similarity are being picked up as background noise. Thus, we did not adopt this homology option as the default. Another problem with this option is that it drastically increases the total search time to the point where it does become a practical consideration. Nevertheless, we still realize that homology information should not be treated as a mere prediction but as a discovery. Therefore, this option does have its practical applications.

GeneHacker Plus is available for free access at our WWW site (http://elmo.ims.u-tokyo.ac.jp/GH/). Please note that the homology option is not currently supported.

Acknowledgements: We thank Andrew J. Link, Sridhar S. Hannenhalli, Hideto Takami, Takashi Sazuka and Takeko Kodama for providing us with their valuable data; Todd Taylor for critically reading the manuscript. This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Genome Informatics" of MECSST in Japan. T. T. and K. N. are also supported by the Special Coordination Funds for promoting science and technology of MECSST.

References

1. Krogh, A., Mian, I. S., and Haussler, D. 1994, A hidden Markov model that finds genes in E. coli DNA, *Nucleic Acids Res.*, **22**, 4768–4778.
2. Borodovsky, M. and McIninch, J. D. 1993, GeneMark: parallel gene recognition for both DNA strands, *Comput. Chem.*, **17**, 123–133.
3. Yada, T. and Hirosawa, M. 1996, Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model, *ISMB*, **4**, 252–260.
4. Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. 1998, Microbial gene identification using interpolated Markov models, *Nucleic Acids Res.*, **26**, 544–548.
5. Lukashin, A. V. and Borodovsky, M. 1998, GeneMark.hmm: new solutions for gene finding, *Nucleic Acids Res.*, **26**, 1107–1115.
6. Hayes, W. S. and Borodovsky, M. 1998, Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction, *Genome Res.*, **8**, 1154–1171.
7. 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.*, **26**, 2941–2947.
8. Besemer, J. and Borodovsky, M. 1999, Heuristic approach to deriving models for gene finding, *Nucleic Acids Res.*, **27**, 3911–3920.
9. Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. 1999, Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, **27**, 4636–4641.
10. Shmatkov, A. M., Melikyan, A. A., Chernousko, F. L., and Borodovsky, M. 1999, Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes, *Bioinformatics*, **15**, 874–886.
11. Rabiner, L. R. 1989, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, **77**, 257–286.
12. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998, *Biological sequence analysis: probabilisic models of proteins and nucleic acids*, Cambridge Univ. Press, Cambridge.
13. Kozak, M. 1999, Initiation of translation in prokaryotes and eukaryotes, *Gene*, **234**, 187–208.
14. Link, A. J., Robison, K., and Church, G. M. 1997, Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12, *Electrophoresis*, **18**, 1259–1313.
15. Hannenhalli, S. S., Hayes, W. S., Hatzigeorgiou, A. G., and Fickett, J. W. 1999, Bacterial start site prediction, *Nucleic Acids Res.*, **27**, 3577–3582.
16. Klenk, H. P., Clayton, R. A., Tomb, J. F. et al. 1998, The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, *Nature*, **390**, 364–370.
17. Kunst, F., Ogasawara, N., Moszer, I. et al. 1997, The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*, *Nature*, **390**, 249–256.
18. Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A. et al. 1997, The complete genome sequence of *Escherichia coli* K-12, *Science*, **277**, 1453–1474.
19. Tomb, J. F., White, O., Kerlavage, A. R. et al. 1997, The complete genome sequence of the gastric pathogen *Helicobacter pylori*, *Nature*, **388**, 539–547.
20. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C. et al. 1997, Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics, *J. Bacteriol.*, **179**, 7135–7155.
21. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, **3**, 109–136.
22. Takami, H., Nakasone, K., Ogasawara, N. et al. 1999, Sequencing of three lambda clones from the genome of alkaliphilic *Bacillus* sp. strain C-125, *Extremophiles*, **3**, 29–34.
23. Sazuka, T., Yamaguchi, M., and Ohara, O. 1999, Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing, *Electrophoresis*, **20**, 2160–2171.
24. Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. 1996, *Entrez*: molecular biology database and retrieval system, *Methods Enzymol.*, **266**, 141–162.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
26. Smith, T. F. and Waterman, M. S. 1981, Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195–197.
27. Salzberg, S. L. 1997, A method for identifying splice sites and translational start sites in eukaryotic mRNA, *Comput. Appl. Biosci.*, **13**, 365–376.
28. Hirosawa, M., Totoki, Y., Hoshida, M., and Ishikawa, M. 1995, Comprehensive study on iterative algorithms of multiple sequence alignment, *Comput. Appl. Biosci.*, **11**, 13–18.
29. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, **28**, 45–48.
30. Moszer, I., Glaser, P., and Danchin, A. 1995, SubtiList: a relational database for the *Bacillus subtilis genome*, *Microbiology*, **141**, 261–268.
31. Fickett, J. W. and Tung, C.-S. 1992, Assessment of protein coding measures, *Nucleic Acids Res.*, **20**, 6441–6450.
32. Tompa, M. 1999, An exact method for finding short motifs in sequences, with application to the ribosome bind-

ing site problem, *ISMB*, 262–271.
33. Takami, H., Nakasone, K., Takaki, Y. et al. 2000, Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*, *Nucleic Acids Res.*, **28**, 4317–4331.
34. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. 2000, Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS, *Nature*, **407**, 81–86.
35. Robison, K., Gilbert, W., and Church, G. M. 1994, Large scale bacterial gene discovery by similarity search, *Nature Genet.*, **7**, 205–214.
36. Gelfand, M. S., Mironov, A. A., and Pevzner, P. A. 1996, Gene recognition via spliced sequence alignment, *Prot. Natl. Acad. Sci. U.S.A.*, **93**, 9061–9066.
37. Krogh, A. 2000, Using database matches with for HMMGene for automated gene detection in *Drosophila*, *Genome Res.*, **10**, 523–528.
38. Reese, M. G., Kulp, D., Tammana, H., and Haussler, D. 2000, Genie—Gene Finding in *Drosophila melanogaster*, *Genome Res.*, **10**, 529–538.