ELSEVIER

Gene 234 (1999) 257–265

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Starts of bacterial genes: estimating the reliability of computer predictions

Dmitrij Frishman [a,*], Andrei Mironov [b], Mikhail Gelfand [c]

[a] *GSF-Forschungszentrum f. Umwelt und Gesundheit, Munich Information Center for Protein Sequences am Max-Planck-Institut für Biochemie, Am Klopferspitz 18, D-82152 Martinsried, Germany*

[b] *Laboratory of Mathematical Methods, National Center for Biotechnology Information NIIGENETIKA, Moscow 113545, Russia*

[c] *Institute of Protein Research, Russian Academy of Sciences, Pushchino 142292, Russia*

Received 16 December 1998; received in revised form 12 April 1999; accepted 11 May 1999; Received by B. Dujon

Abstract

Exact mapping of gene starts is an important problem in the computer-assisted functional analysis of newly sequenced prokaryotic genomes. We describe an algorithm for finding ribosomal binding sites without a learning sample. This algorithm is particularly useful for analysis of genomes with little or no experimentally mapped genes. There is a clear correlation between the ribosomal binding site (RBS) properties of a given genome and the potential gene start prediction accuracy. This correlation is of considerable predictive power and may be useful for estimating the expected success of future genome analysis efforts. We also demonstrate that the RBS properties depend on the phylogenetic position of a genome. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Complete genome; Gene recognition; Ribosomal binding site; Shine–Delgarno box; Start codon; Translation initiation

1. Introduction

A few years ago the problem of gene recognition in bacterial DNA was assumed to be essentially solved, and the main efforts were directed towards the development of algorithms for prediction of the exon–intron structure of eukaryotic genes. However, when several complete bacterial genomes had been sequenced, it turned out that the existing programs were insufficient. The problem is that the majority of completely sequenced genomes belong to little studied systematic groups and it is not possible to compile a learning sample for statistical gene recognition. A number of recently developed programs [e.g., GLIMMER (Salzberg et al., 1998)] initially define protein-coding regions as long open reading frames or segments homologous to known genes.

Another problem is the exact mapping of gene starts. A number of early papers described methods for recognition of ribosome binding sites in *Escherichia coli* using statistical, pattern recognition or neural network modelling of experimentally mapped sites (Stormo et al., 1982; Studnicka, 1986; Barrick et al., 1994; Bisant and Maizel, 1995). However, as in the case of the recognition of coding regions, these methods cannot be applied to newly sequenced genomes. The simplest solution, in some cases sufficient for subsequent protein analysis, is to use the longest open reading frame containing the identified segment. This approach was applied to the analysis of the *E. coli* genome (Blattner et al., 1997) and the *Pyrococcus horikoshii* genome (Kawarabayasi et al., 1998) and in the revised version of GenBank annotations of the *Methanococcus jannaschii* genome (Bult et al., 1996). However, more sophisticated analysis requires exact delineation of gene starts. This includes recognition of N-terminal signal peptides (von Heijne, 1984) as well as the analysis of the efficiency of translation initiation, specific signals regulating translation initiation (Vellanoweth, 1993; Draper, 1996; Keener and Nomura, 1996) and RNA secondary structure influencing it (Looman et al., 1986; de Smit and van Duin, 1990).

There are two main approaches to the recognition of ribosome binding sites in the absence of learning samples. One possibility is to rely on the universal mechanisms of RBS recognition via base-pairing of the Shine–Dalgarno box and 3′-terminus of the 16S rRNA (Shine and Dalgarno, 1974). This was used to predict

Abbreviations: RBS, ribosome binding site.

* Corresponding author. Tel:. +49-89-85782664; fax: +49-89-85782655.

E-mail address: frishman@mips.biochem.mpg.de (D. Frishman)

PII: S0378-1119(99)00200-0

RBS in *Escherichia coli* (Schurr et al., 1993) and Gram-positive bacteria, in particular *Bacillus subtilis* and *Staphilococcus aureus* (Hatzigeorgiou and Fickett, 1997). The general pattern of the Shine–Dalgarno box is also used by CRITICA (Badger and Olsen, 1997), although this program does not directly compute the base-pairing energy. The problem with this approach is that it is not a priori obvious whether the mechanisms of translation initiation are universally conserved among prokaryotes. Indeed, in many cases there seems to be no or almost no detectable base-pairing between the region upstream of start codons and 16S rRNA (Battista, 1997; Yada et al., 1997).

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′ end of statistically predicted protein-coding regions (Hayes and Borodovsky, 1998). ORPHEUS (Frishman et al., 1998) initially defines candidate coding regions as homology-derived open reading frames. The regions having only one potential start codon form the learning sample for the recognition of translation initiation sites. In a variant of this approach, the leftmost ATG, GTG and TTG codons of all long open reading frames are used (Mironov et al., 1998). A simple iterative procedure is utilized to derive the RBS recognition profile (see Methods). At that, only a fraction of candidate sites is used to update the profile at each iteration. Thus we avoid the influence of false starts (in the latter version), as well as accounting for the possibility of re-initialization of translation and overlapping genes (in the former version).

In this paper we describe the results of gene start recognition by ORPHEUS in several different prokaryotes with completely or almost completely sequenced genomes. An important issue here is the analysis of prediction reliability. We used available samples of experimentally mapped genes in order to demonstrate that the information content of the derived signal is a good estimator of the fraction of correctly identified gene starts. Another interesting observation is that the information content is approximately uniform within phylogenetic groups of genomes.

2. Methods

2.1. Datasets

One of the main problems in quantifying the success of the gene prediction accuracy algorithms is the choice of a standard of truth. This problem is usually addressed by extracting a set of trusted genes from well-annotated database entries. When techniques designed to deal with complete genomic sequences are considered, the task is further complicated since (i) for many bacterial organisms well-represented in the protein sequence databanks, no complete genomes are available yet, and (ii) for many of the newly sequenced genomes very few or no protein sequences were known in pre-genomics era.

We have implemented an automatic procedure to extract from the PIR-International database (Barker et al., 1998) protein sequences not associated with large-scale genome sequencing efforts. For every organism (e.g., *Synechocystis* sp.), the following invocation of the Sequence Retrieval System (SRS; Etzold et al., 1996) was first effected:

getz "[pir-Organism: Synechocystis sp*] !

[pir-AllText: plasmid*] ! [pir-MoleculeType: protein] !

[pir-MoleculeType: mrna] !

[pir-MoleculeType: nucleic acid] !

[pir-AllText: fragment*]" -e

to ensure that only complete genomic sequences determined by DNA-sequencing were selected. Datasets obtained in this manner were further filtered with a custom *perl* script to keep only entries with sequence modification date before 1995, the year when the first complete genome was published.

2.2. Recognition of ribosome binding sites

The procedure for prediction of ribosome binding sites is part of the ORPHEUS software for gene recognition described in Frishman et al. (1998). Coding regions are initially predicted by reliable similarity to known genes. Then the regions having only one candidate start codon are selected (Fig. 1). The regions (−20)–(−1) upstream of these codons are aligned by the following iterative procedure.

Let L be the expected length of the Shine–Dalgarno box ($L=6$). Denote by $F(b,j)$ positional nucleotide frequencies in the initial alignment ($j=(-20)\ldots(-1)$; $b=$T,C,A,G). Positional information content is:

$$H(j)=\sum_{b=\mathrm{A}}^{\mathrm{T}} F(b,j)\,\log\frac{F(b,j)}{G(b)},$$

where $G(b)$ is the genomic frequency of the base.

Initially the RBS signal was assumed to reside in positions having the maximum total information content:

$$\sum_{k=j}^{j+L-1} H(k) \;\underset{j=20\ldots-L}{\rightarrow}\; \max.$$

Then the position of the SD box in each individual sequence was determined using the following two-stage procedure.

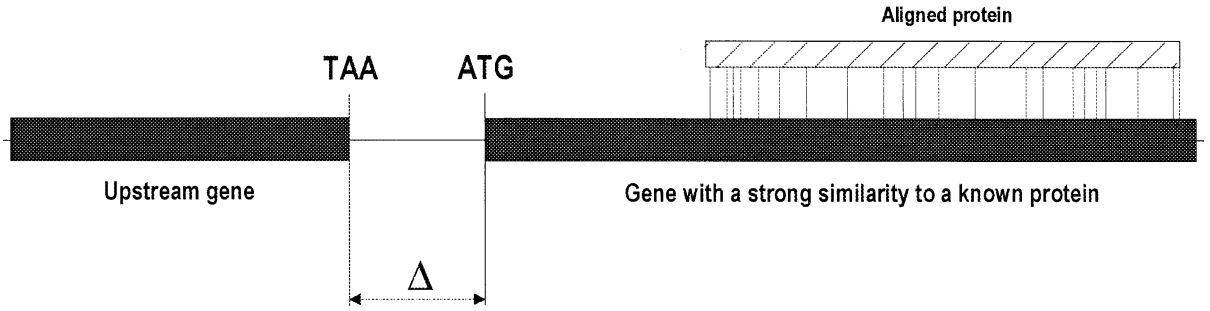

Fig. 1. Extraction of a sample of genes with only one possible start codon. A gene with a significant sequence similarity to a known protein sequence is shown, as well as an upstream gene. The DNA sequence region in which the start codon can possibly reside is thus bounded by the end of the upstream gene on one side and the end of the reliable DNA–protein alignment on the other side. Δ denotes the minimal allowed distance between the putative start codon and the stop codon of the upstream gene (typically 20 bases, although in some experiments it was set to a negative value to account for overlapping genes).

Denote by $N(b, k)$ positional nucleotide counts in the SD profile at a given iteration ($k=1\ldots L$). Following Berg and von Hippel (1988) positional nucleotide weights are:

$$W(b, k) = \log\left(\frac{N(b, k)+0.5}{\max_b N(b, k)+0.5}\right),$$

where max $N(b, k)$ is the frequency of the consensus nucleotide in the position k. The SD signal score is calculated by the formula:

$$\Delta(b_1\ldots b_L) = \sum_{k=1}^{L} W(b, k).$$

The first re-alignment stage involves the iteration until convergence of the following two steps:

- find in each sequence the segment of length L with the highest score;
- re-calculate the nucleotide weight matrix.

A distinctive feature of our algorithm is that at each optimization step only the top scoring fraction (usually top 80%) of sequences are used to produce the current weight matrix.

At the second stage the preferences for the distance between the SD box and the start codon are taken into account. Let M be a possible position of the SD box within the RBS region, and let this position occur $N(M)$ times. Denote by $N_{\max}$ the count of the most frequent position. The positional weights are calculated using the standard formula:

$$V(M) = \log\left(\frac{N(M)+0.5}{N_{\max}+0.5}\right).$$

Now the strength of the SD signal is defined as:

$$\Delta(b_1\ldots b_L) = \sum_{k=1}^{L} W(b_k, k) + V(M).$$

Thus the RBS profile is the nucleotide weight matrix and the vector of position weights. The two-step iterative procedure is used again until convergence.

2.3. Determination of the gene start prediction accuracy

The full gene prediction procedure was effected for each of the 28 genomic sequences specified in Table 1. Predicted genes were compared with the corresponding databank entries for those genomes that have sufficient number of PIR entries determined before whole genome sequencing (see above). Additionally, prediction results were compared with the gene start assignments made by the original authors using automatic methods and often supplemented by manual evaluation.

Comparison of predicted and 'known' genes, either from the PIR database or from the published complete gene sets, was based on BLAST2 (Altschul et al., 1990; Gish, unpublished) similarity searches. A gene was considered correctly predicted (true positive) if it had 97% similarity to a reference sequence. Specifically for the PIR subset of originally sequenced proteins, only for eight species was it possible to obtain 10 or more protein sequences satisfying the criteria above.

2.4. Phylogenetic analysis

The phylogenetic tree for the 28 prokaryotic organisms studied in this work was constructed by maximum likelihood analysis of small-subunit rRNA sequences. The sequences were obtained from the Ribosomal Database Project (RDP; Maidak et al., 1997). The program used to calculate and visualize the tree was PHYLIP (Felsenstein, 1993).

3. Results and discussion

The information content of the automatically derived RBS alignments in 28 completely sequenced and partial bacterial genomes is presented in Table 1. The difference in the RBS signal strength between the best and the worst scoring organism is more than twofold. In many

Table 1
Parameters of the complete and partial genomic sequences analysed in this work[a]

Organism	GC content (%)	Status[b]	Gene prediction method[c]	Information content of the RBS alignment
Helicobacter pylori	39	C	GeneMark/GeneSmith	4.35
Campylobacter jejuni	31	P	–	3.67
Enterococcus faecalis	36	P	–	3.6
Bacillus subtilis	43.5	C	GeneMark/additional searches for RBS signals	3.47
Streptococcus pyogenes	36	P	–	3.33
Streptococcus pneumoniae	40	P	–	3.24
Aquifex aeolicus	43.4	C	Similarity/Critica	3.13
Methanobacterium thermoautotrophicum	49.5	C	GenomeBrowser	3.11
Haemophilus influenzae	38	C	GeneMark	3.06
Methanococcus jannaschii	31.4	C	GeneMark	3.00
Pyrococcus horikoshii OT3	42	C	Simple ORF extraction and similarity analysis	2.93
Chlamydia trachomatis	42	P	–	2.84
Neisseria gonorrhoeae	52	P	–	2.73
Pyrococcus furiosus	41	P	–	2.73
Archaeoglobus fulgidus	48.5	C	GeneSmith/Critica	2.72
Escherichia coli	50.8	C	Manual analysis	2.55
Neisseria meningitidis	52	P	–	2.62
Actinobacillus actinomycetemcomitans	44	P	–	2.50
Treponema pallidum	52.8	C	Glimmer	2.32
Synechocystis sp.	48	C	GeneMark	2.26
Deinococcus radiodurans	66	P	–	2.2
Yersinia pestis	48	P	–	2.18
Borrelia burgdorferi	28.6	C	Glimmer	2.08
Pseudomonas aeruginosa	67	P	–	2.08
Mycobacterium leprae	58	P	–	2.03
Streptomyces coelicolor	72	P	–	1.98
Mycobacterium tuberculosis	65.6	C	TbParse	1.93

[a] The list of organisms is sorted by the information content of the RBS alignment given in the last column.
[b] Status of the genome sequencing project. P, partially studied genome; C, complete and published genome.
[c] Gene prediction methods used by the authors of the published completed genomes. References: GeneMark — Borodovsky and McIninch (1993); Glimmer — Salzberg et al. (1998); EcoParse and TbParse — Krogh et al. (1994); GeneSmith — H.O.Smith, unpublished; GenomeBrowser — Smith et al. (1997); Critica — H.H. Badger and G.J. Olsen, unpublished; Frames, GCG package — Genetics Computer Group (1994).

cases this difference can be explained by disparity between mechanisms of translation initiation. Thus, high information content is characteristic for Gram-positive bacteria from the *Clostridia*/*Bacillaceae* group. This is probably due to the absence in these species of the ribosomal protein S1 that facilitates initiation of translation by melting mRNA secondary structures (Vellanoweth, 1993). On the other hand, low information content in phylogenetically isolated Gram-positive bacterium *Deinococcus radiodurans* corresponds to the presumption of Battista (1997) that "those structural features that signal initiation of transcription in *D. radiodurans* evolved differently than they did in other prokaryotes".

In general, there is a clear correspondence between the ordering of the organisms based on the quality of their RBS and the phylogenetic ordering (see Fig. 2). The ϵ-subdivision of Gram-negative bacteria and low GC Gram-positive species tend to possess the strongest RBS, while high GC Gram-positive species bring up the rear of the list. Representatives of the γ-subdivision of Gram-negative bacteria are intermixed with the Archaea in the intermediate range of the RBS strength values. *A. aeolicus* is also located close to Archaea in Table 1, consistent with its taxonomic position at the divergence point between Archea and bacteria on the sRNA-based phylogenetic tree (Burggraf et al., 1992). Cyanobacteria and the Gram-positive radiation-resistant bacterium *D. radiodurans* seem to be phylogenetically isolated. By far the strongest RBS belong to the gastric parasite *H. pylori* (Tomb et al., 1997), followed by the closely related *C. jejuni*. We are not aware of any experimentally confirmed sequences of *H. pylori* RBS. The RBS automatically derived from the partially sequenced genome of *C. jejuni* are remarkably similar to the set of 21 putative RBS published in Wösten et al. (1998), including the distribution of the spacer length between the RBS and the start codon (Fig. 3). The window length used by our program is 6 bases, while the experimental sites vary in length from 3 to 4. The core consensus AGG or AAGG, however, is captured very well.

For some of the organisms studied we were able to associate the RBS strength with the precision of the gene start delineation. As seen in Fig. 4a, the accuracy

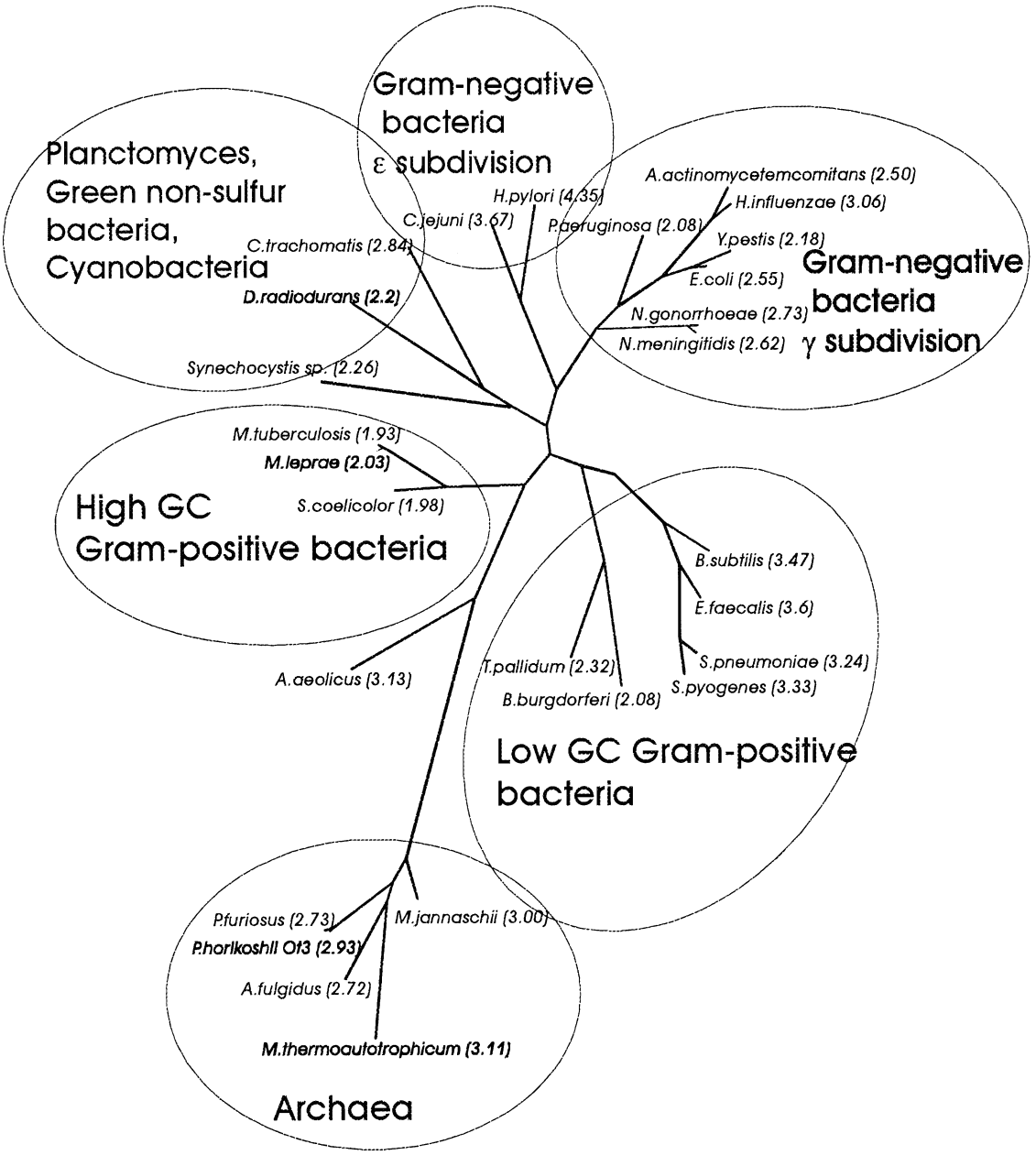

Fig. 2. Phylogenetic tree of 28 prokaryotic organisms with completely or partially sequenced genomes based on the maximum likelihood analysis of small-subunit rRNA sequences. Numbers in parentheses are the values of the RBS information content (see Table 1).

of gene start prediction as measured on the set of genes with reliably determined N-termini taken from the PIR database (see Methods) is strongly correlated with the information content of the RBS alignments and, consequently, by virtue of the phylogenetic tree in Fig. 2, with the taxonomic position of a given organism. This finding is important from the practical point of view as it allows for objective estimates, ahead of time, of the achievable gene start prediction quality in a particular genome sequencing project and thus, to some degree, of the entire annotation process.

For 13 of the completely sequenced genomes we evaluated the performance of our algorithm relative to the gene starts predicted by the genome authors using the methods given in Table 1 and often supplemented by manual analysis (Fig. 4b). Some of the data in Table 1 are consistent with earlier observations by other authors. For example, Smith et al. (1997) noted that most of the *M. thermoautotrophicum* genes are preceded with a well-defined RBS consensus. On the other hand, Sazuka and Ohara (1996), Yada et al. (1997) and Hirosawa et al. (1997) found no pronouced Shine–Dalgarno box upstream of *Synechocystis* sp. genes. ORPHEUS correctly identified 61% of 72 experimentally

A

RBS	Spacer	Start codon
ttAGGAt	5	TTG
aAAGGct	3	ATG
tAAGGAt	5	ATG
cAAGGAt	6	ATG
tAAGGAg	7	ATG
aAAGGAc	5	ATG
aAAGGAa	4	ATG
cgAGGAc	2	ATG
AAGGca	3	ATG
agAGGAg	6	GTG
aAAGGAa	6	ATG
aAAGGAg	5	ATG
aAAGGAg	5	ATG
ttAGGtg	6	ATG
tcAGGAg	7	ATG
tAAGGAa	6	ATG
cAAGGAg	5	GTG
aAAGGAt	5	ATG
ctAGcAc	5	ATG
aAAGGAg	5	ATG
aAAGGta	9	ATG

B

RBS	Spacer	Start codon
AAAGGA	6	ATG
ACATGA	13	TTG
AAAGAA	6	ATG
AAAGGA	7	ATG
AAAGGA	6	ATG
AAAGGT	5	ATG
AAAGGA	8	ATG
AAAGGA	6	ATG
AAAGGA	6	ATG
TTAGGA	7	ATG
AAAGGT	9	ATG
AAAGGA	5	ATG
AAAGGA	6	TTG
AAAGGA	6	ATG
AAAGAA	16	ATG
AAAAGA	10	GTG
AAAGAA	6	TTG
AAAGGA	7	ATG
AGAGGA	3	ATG
AAAGGA	6	ATG
CAAAGA	1	ATG
AAAGGA	6	ATG
CAAGGG	6	ATG
ATAAGA	6	TTG
AGAAGA	13	TTG
AAAGGA	6	ATG
TTAGGA	6	ATG
CAAGGA	7	ATG
GAAGGA	5	ATG
AAAGGA	2	ATG
AAGAGA	9	ATG
AAACAA	3	GTG
AAAAGA	13	ATG
GAAGGT	16	GTG

C

Position in the window	1	2	3	4	5	6
A	0	0	0	-0.626	-0-806	0
C	-0.788	-0.916	-1.079	-0.871	-1.017	-1.017
G	-0.840	-0.862	-1.022	0	0	-0.960
T	-0.840	-0.811	-1.079	-0.871	-1.017	-0.854
Consensus	A	A	A	G	G	A

Fig. 3. Ribosome binding sites in the *C. jejuni* genome. (a) A set of 21 putative RBS taken from the work of Wösten et al. (1998). (b) A set of 34 automatically derived RBS sequences (with window length 6) used by ORPHEUS to calculate the RBS weight matrix. (c) The RBS weight matrix.

derived gene starts from Sazuka and Ohara (1996) (data are not shown), in good accordance with the overall accuracy for this genome shown in Fig. 4. Blattner et al. (1997) mention difficulties in assignment of gene termini in *E. coli* and state that in the absence of homology data the leftmost possible start codon was usually selected.

When the existing GenBank annotations are consid-

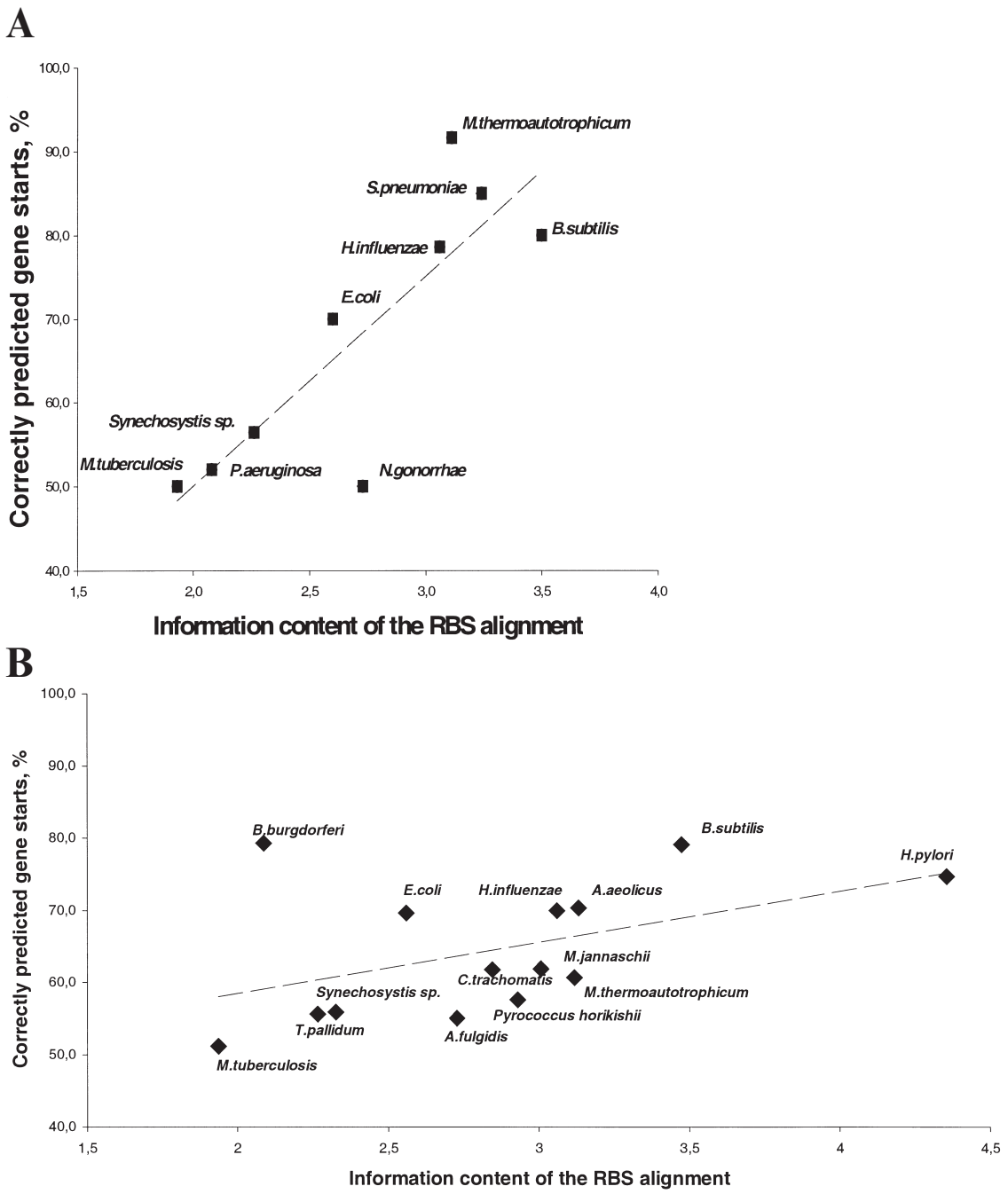

Fig. 4. Dependence of the gene start prediction accuracy on the information content of the RBS weight matrix. (a) Comparison with the selection of the proteins from the PIR databank not related to genome sequencing projects; (b) Comparison with the full gene complements of the genomes deposited in GenBank.

ered as the 'standard of truth', the correlation between the gene start prediction accuracy and the quality of the RBS derived by ORPHEUS is much less pronounced. This is at least partially due to the fact that the accuracy of gene start annotation in GenBank is rather limited, especially if the 'leftmost ATG' rule has been used (Blattner et al., 1997; Kawarabayasi et al., 1998; cf. Frishman et al., 1998).

The RBS profiles constructed by ORPHEUS are usually similar to those by GeneMark (Hayes and Borodovsky, 1998). However, our interpretation of these matrices is different. In particular, Hayes and Borodovsky ascribe the lack of the Shine–Dalgarno signal in Mycoplasmas to the possible existence of other 16S rRNAs in these genomes. They also explain the *M. jannaschii* RBS consensus GGTGA by base-pairing at

some distance from the 16S rRNA 3′-terminus (Figs. 1 and 8 of Hayes and Borodovsky, 1998). However, there is no evidence of 16S rRNAs with non-canonical 3′-termini in the *Mycoplasma* genomes. Moreover, multiple 16S rRNA genes present in some genomes (e.g., *E. coli* and *H. influenzae*) are identical in the anti-Shine–Dalgarno region involved in base-pairing with the RBS. Finally, all prokaryotic 16S rRNAs for which complete genomes are available have identical anti-Shine–Dalgarno boxes.

In the absence of other explanations, we believe that the weak signal observed in the *Synechocystis* and *Mycoplasma* genomes is in fact due to the positional nucleotide preferences within codons and is generated by incorrectly identified gene starts situated within the coding region. This conjecture is supported by the fact that there is a marked three-periodicity in the distances between the predicted RBS locations for these species and the start codon (Figs. 7 and 11 of Hayes and Borodovsky, 1998). Note that in the case of strong signals providing more reliable recognition (*E. coli*, *H. influenzae*, *M. jannaschii*) the distribution of these distances is unimodal and rather sharp, reflecting strong preference for particular distances between the RBS and the start codon (Figs. 3, 5 and 9 of Hayes and Borodovsky, 1998). Generally, we believe that multimodality, and especially periodicity, in the distribution of distances is a strong indication of low reliability of the recognition rule.

The sample of genes with a single possible start codon (Fig. 1) is arguably the best possible source of information on start codon occurrence in complete genomes. Errors are very unlikely since (i) the overall gene prediction accuracy by ORPHEUS is very high, (ii) ends of the upstream genes can be determined unequivocally, and (iii) only high-scoring DNA-protein alignments are used. The region in which the search for start codons is performed is thus delimited by reliably determined features. To take into account the possibility of ribosome re-initializing, the analysis below was done assuming non-overlapping *and* possibly overlapping genes, corresponding to positive and negative values of Δ (as defined by Fig. 1), respectively.

The results presented in this contribution have long-standing consequences for protein sequence annotation in general. Since the main bulk of the new entries in the public databases will be increasingly based on sequences determined through large-scale DNA sequencing and annotated semi-automatically, the ability to estimate the likeliness of erroneous delineation of the protein N-termini is of large practical importance.

4. Availability

All datasets and other materials mentioned in this article can be obtained from D. Frishman via e-mail. A comprehensive computational analysis of all the currently available complete bacterial genomes and many partial genomic sequences can be found at the PEDANT genome analysis server (http://pedant.mips.biochem.mpg.de). The program ORHPEUS for bacterial gene prediction is available at http://pedant.mips.biochem.mpg.de/orpheus.

Acknowledgement

We would like to thank Hans-Werner Mewes for valuable comments on the manuscript.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Badger, J.H., Olsen, G.J., 1997. CRITICA: coding region identification tool invoking comparative analysis. Manuscript available at gary@phylo.life.uiuc.edu.

Barker, W.C., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S.L., Ledley, R.S., Mewes, H.W., Pfeiffer, F., Tsugita, A., 1998. The PIR-International Protein Sequence Database. Nucleic Acids Res. 26, 27–32.

Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T.D., Lawrence, C.E., Gold, L., Stormo, G.D., 1994. Quantitative analysis of ribosome binding sites in *E. coli*. Nucleic Acids Res. 22, 1287–1295.

Battista, J.R., 1997. Against all odds: the survival strategies of *Deinococcus radiodurans*. Annu. Rev. Microbiol. 51, 203–224.

Berg, O.G., von Hippel, P.H., 1988. Selection of DNA binding sites by regulatory proteins. Trends Biochem. Sci. 13, 207–211.

Bisant, D., Maizel, J., 1995. Identification of ribosome binding sites in *Escherichia coli* using neural network models. Nucleic Acids Res. 23, 1632–1639.

Blattner, F.R., Plunkett III, G., Bloch, C.A., et al., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1462.

Borodovsky, M., McIninch, J., 1993. GeneMark: parallel gene recognition for both DNA strands. Comput. Chem. 17, 123–133.

Bult, C., et al., 1996. Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii*. Science 273, 1058–1073.

Burggraf, S., Olsen, G.J., Stetter, K.O., Woese, C.R., 1992. A phylogenetic analysis of *Aquifex pyrophilus*. Syst. Appl. Microbiol. 15, 353–356.

de Smit, M.H., van Duin, J., 1990. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc. Natl. Acad. Sci. USA 87, 7668–7672.

Draper, D.E., 1996. Translation initiation. In: Neidhardt, F.C. (Ed.), Escherichia coli and Salmonella in: Cellular and Molecular Biology vol. 1. ASM Press, Washington, DC, pp. 902–908. (Chapter 59)

Etzold, T., Ulyanov, A., Argos, P., 1996. SRS: information retrieval system for molecular biology data banks. Methods Enzymol. 266, 114–128.

Felsenstein, J., 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Frishman, D., Mironov, A., Mewes, H.W., Gelfand, M., 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res. 26, 2941–2947.

Genetics Computer Group, 1994. Program Manual for the Wiscon-

sin Package, Version 8, August 1994, 575 Science Drive, Madison, WI 53711, USA.

Hatzigeorgiou, A.G., Fickett, J.W., 1997. 1st Annual Conference on Computational Genomics, p. 8.

Hayes, W.S., Borodovsky, M., 1998. Deriving ribosomal binding sites (RBS) statistical models from unannotated DNA sequence and the use of the RBS model for N-terminal prediction. In: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E. (Eds.), Pacific Symposium on Bioinformatics '98. World Scientific, Singapore, pp. 279–290.

Hirosawa, M., Sazuka, T., Yada, T., 1997. Prediction of translation initiation sites on the genome of *Synechocystis* sp. strain PCC6803 by Hidden Markov model. DNA Res. 4, 179–184.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Kikuchi, H., 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. 5, 55–76.

Keener, J., Nomura, M., 1996. Regulation of ribosome synthesis. In: Neidhardt, F.C. (Ed.), *Escherichia coli* and *Salmonella* in: Cellular and Molecular Biology vol. 1. ASM Press, Washington, DC, pp. 1417–1431. (Chapter 90)

Krogh, A., Mian, I.S., Haussler, D., 1994. A hidden Markov model that finds genes in *E.coli* DNA. Nucleic Acids Res. 22, 4768–4778.

Looman, A.C., Bodlaender, J., de Gruyter, M., Vogelaar, A., van Knippenberg, P.H., 1986. Secondary structure as primary determinant of the efficiency of ribosomal binding sites in *Escherichia coli*. Nucleic Acids Res. 14, 5481–5497.

Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J., Woese, C.R., 1997. The RDP (Ribosomal Database Project). Nucleic Acids Res. 25, 109–111.

Mironov, A.A., Frishman, D., Gelfand, M.S., 1998. Computer analysis of regulatory patterns in complete bacterial genomes. Ribosome binding sites. Mol. Biol.. in press

Salzberg, S.L., Delcher, A.L., Kasif, S., White, O., 1998. Microbial gene identification using interpolated Markov mels. Nucleic Acids Res. 26, 544–548.

Sazuka, T., Ohara, O., 1996. Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp. strain PCC6803 by amino terminal protein sequencing. DNA Res. 3, 225–232.

Schurr, T., Nadir, E., Margalit, H., 1993. Identification and characterization of *E.coli* ribosomal binding sites by free energy computation. Nucleic Acids Res. 21, 4019–4023.

Shine, J., Dalgarno, L., 1974. Correlation between the 3′-terminal-polypirimidine sequence of 16S RNA and translational specificity of the ribosome. Proc. Natl. Acad. Sci. USA 71, 1342–1346.

Smith, D.R., et al., 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. J. Bacteriol. 179, 7135–7155.

Stormo, G.D., Schenider, T.D., Gold, L., Ehrenfeucht, A., 1982. The use of the ''Perceptron'' algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Res. 10, 2997–3011.

Studnicka, G.M., 1986. Quantitative computer analysis of signal sequence homologies. Comput. Appl. Biosci. 2, 269–275.

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.

Vellanoweth, R.L., 1993. Translation and its regulation. In: Sonenstein, A.L., Hoch, J.A., Losik, R. (Eds.), *Bacillus subtilis* and other Gram-Positive Bacteria. American Society for Microbiology, Washington, DC, (Chapter 48)

von Heijne, G., 1984. Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implication for protein export in prokaryotic and eukaryotic cells. EMBO J. 3, 2315–2318.

Wösten, M.M., Boeve, M., Koot, M.G., van Nuene, A.C., van der Zeijst, B.A., 1998. Identification of *Campylobacter jejuni* promoter sequences. J. Bacteriol. 180, 594–599.

Yada, T., Sazuka, T., Hirosawa, M., 1997. Analysis of sequence patterns surrounding the translation initiation sites in *Cyanobacterium* genome using the hidden Markov model. DNA Res. 4, 1–7.