D114–D119 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 16 October 2007*
doi:10.1093/nar/gkm799

ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes

Gang-Qing Hu[1,2], Xiaobin Zheng[1,2], Yi-Fan Yang[1,2], Philippe Ortet[3], Zhen-Su She[1,2,4] and Huaiqiu Zhu[1,2,*]

[1]State Key Lab for Turbulence and Complex System and Department of Biomedical Engineering, [2]Center for Theoretical Biology and Department of Physics, Peking University, Beijing 100871, China, [3]CEA, DSV, IBEB, LEMiRE, CNRS, Université Aix-Marseille II, CEA Cadarache, F-13108 Saint-Paul-lez-Durance, France and [4]Department of Mathematics, UCLA, Los Angeles, CA 90095, USA

Received August 15, 2007; Revised September 16, 2007; Accepted September 17, 2007

ABSTRACT

Correct annotation of translation initiation site (TIS) is essential for both experiments and bioinformatics studies of prokaryotic translation initiation mechanism as well as understanding of gene regulation and gene structure. Here we describe a comprehensive database ProTISA, which collects TIS confirmed through a variety of available evidences for prokaryotic genomes, including Swiss-Prot experiments record, literature, conserved domain hits and sequence alignment between orthologous genes. Moreover, by combining the predictions from our recently developed TIS post-processor, ProTISA provides a refined annotation for the public database RefSeq. Furthermore, the database annotates the potential regulatory signals associated with translation initiation at the TIS upstream region. As of July 2007, ProTISA includes 440 microbial genomes with more than 390 000 confirmed TISs. The database is available at http://mech.ctb.pku.edu.cn/protisa

INTRODUCTION

Over the past few years, people have witnessed an exponential growth in the number of completed microbial genomes. It is imperative to annotate a genome as precisely as possible, especially due to flourishing with genome-based experimental approaches such as DNA microarrays and protein arrays (1). Specifically, accurate translation initiation site (TIS) annotation is important for experiments such as identifying native purified proteins through N-terminal amino acid sequencing, as well as heterologous protein products (1). Meanwhile, *in silico* studies of translation initiation mechanism, gene regulation, as well as the predictions of operon, promoter and small-untranslated RNAs also rely on the correct TIS annotation (2,3).

A TIS can be reliably identified by means of the experiments such as N-terminal protein sequencing. Unfortunately, such data constitutes only a small portion of all known proteins. Even for the best-studied genome *Escherichia coli* K-12, as collected in EcoGene (4), less than a quarter of proteins have been verified in such way. Nevertheless, as the number of proteomic projects increases, the amount of TIS with experimental evidences is expected to accumulate significantly in the near future (5). On the other hand, progress has been made in reliable TIS identification through computational evidences such as sequence alignment (2,6). Frishman *et al.* (6) was the first to annotate open reading frames (ORFs) that have significant matches to known proteins with hits distributed in a way to ensure the 5′ most candidate start codon to be true. Recently, Makita *et al.* (2) introduced another method to identify high-quality TIS by sequence alignment between orthologous genes and applied the resultant dataset to evaluate the performance of TIS prediction.

The sequence patterns around TISs have been frequently used for *in silico* study of translation initiation mechanism, thus to design TIS prediction algorithm (2,7–12). In textbooks, ribosome is recruited to mRNA to initiate translation by specific signals nearby TIS such as start codon and Shine-Dalgarno (SD) signal (13). However, for genes without (or almost no) 5′UTR in the mRNA, i.e. leaderless genes, transcriptional signal such as Pribnow box (Bacteria) or TATA box (Archaea) instead of the SD signal has been found upstream of the TIS (9,14–16). Recently, a comprehensive study on hundreds of prokaryotic genomes revealed that 'non-SD-led genes are as common as SD-led genes' (17). Thus, it is reasonable to expect that the complexity of prokaryotic translation initiation mechanism will attract a closer attention as more and more genomes are being sequenced.

In addition to experimental data in the public database such as Swiss-Prot, the increasing amount of sequenced

*To whom correspondence should be addressed. Tel: 8610 6276 7261; Fax: 8610 6276 7261; Email: hqzhu@pku.edu.cn hqzhu@ctb.pku.edu.cn

genomes, which covers a wide range of prokaryotic branches, allows now a high-throughput approach to systematically collect confirmed TIS through database scanning and sequence alignment. It is also interesting to combine the state-of-the-art prediction tools to refine the current public database annotation. Moreover, annotating regulatory signals upstream of TIS will facilitate the studies of initiation mechanism. Herein, we describe a comprehensive relational database, ProTISA, which is designed to collect confirmed TIS, as well as to annotate potential transcriptional or translational signals adjacently upstream to TISs for each of the current hundreds of prokaryotic genomes. We expect that the database may serve the prokaryotic genome annotation and facilitate a wide range of studies on translation initiation.

DATA COLLECTION

Annotation on TIS location

Confirmed TISs (IPT, CDC and HSC) were collected through database scanning, literature survey, conserved domain search and sequence alignment as follows:

(i) ImPorTed (IPT): a script was written to extract high-quality manual annotation in Swiss-Prot. The feature key 'INIT_MET' is used to indicate whether the initiator methionine has been cleaved off or not. We extracted Swiss-Prot entries that are identified as being cleaved off. In addition, we collected experimentally confirmed data by literature survey. Finally, the IPT data set has been enriched by a simple N-terminal sequence comparison between closely related species.

(ii) Conserved domain confirmed (CDC): the method to identify TIS through conserved domain search is essentially similar to that in (6). We searched for each gene against the Conserved Domain Database. The TIS for a gene with only one possible start codon upstream to the 5′-most conserved domain hit was readily identified. To compensate for random matches, which would perhaps lead to an incorrect CDC-TIS annotation, we removed six amino acids from the most upstream hit before processing (18), since the frequency that a hit overlaps with the non-coding region by more than six amino acids is $<1\%$ when examined on genes with IPT TISs.

(iii) High similarity confirmed (HSC): identifying TIS through sequence alignment between orthologous genes has been described in (2). Briefly, it determines the TIS of a gene by referring to its orthologous gene with known TIS from other genera. It requires that they are aligned in the N-terminal region. In ProTISA, genes with TIS labeled as IPT or CDC constitute the references to determine HSC TISs. We have been aware that errors in the IPT or CDC TISs might propagate into the HSC TISs via sequence alignment, especially among closely related genomes. To minimize such errors, a HSC TIS is annotated only if it has support of orthologous genes from more than one different genus.

MED-Start is a TIS predictor with an iterative self-training algorithm based on a four-component statistical model to describe the TIS in prokaryotic genomes (7). The high performance of MED-Start has been demonstrated by evaluating on the *E. coli* and *Bacillus subtillis* genomes. For the present work to computationally relocate TISs for large-scale genomes of both Bacteria and Archaea, several improvements used in our another work (9) have been made to the original algorithm of MED-Start. The modification first treated with the effects of the genomic background to recover regulatory motifs as well as to characterize the sequence patterns around the motifs and the start codon. The operon structure in prokaryotic genomes was also taken into account. Further, the bias of the codon positional GC-content was applied to describe the coding potential of the context around a candidate start. Details of the improved algorithm are available from http://ctb.pku.edu.cn/main/SheGroup/MEDStartPlus.htm

The confirmed TISs via the above-mentioned evidences together with the improved predictions (MED) then served as a refined annotation resource for the public database such as RefSeq.

Annotation on regulatory signal

We implemented a MEME-like algorithm to find signals upstream of TIS (19). It combines the positional weight matrix (PWM) of the signal, the distribution of the number of nucleotides between the signal and the TIS (spacer length), and the background nucleotide frequencies into a likelihood function. An EM algorithm and a simulative annealing strategy were used to estimate the parameters.

To classify the signals, we first included two kinds of typical signals as references, i.e. the widely accepted SD consensus 'AAGGAGGTGA' (3) and the Pribnow (or TATA) box. We use the PWM of the -10 promoter in *E. coli* K-12 for the Pribnow box (20), while the PWM of the AT-rich motif found in *Archaeoglobus fulgidus* for the TATA box. Two scores were calculated for each signal, i.e. the SD score and the TA score. The former is calculated by matching the referenced consensus against the PWM of the signal, and the latter is measured by the Euclidean distance between the PWM of the signal and the referenced PWM. Interestingly, each score follows a bimodal distribution, which allows us to readily classify the signals into three categories: (i) TA-like, those resemble the Pribnow (or TATA) box; (ii) SD-like, those resemble the SD signal and (iii) atypical, those resemble neither SD signal nor Pribnow (or TATA) box.

A Bayesian methodology is employed to predict potential signal upstream of each TIS. We introduce a scoring function to measure the significance of a string as a signal comparing to a random background. A string is more likely to be a functional signal than a random sequence if the score >0. With the PWM and the spacer length distribution of the signal, we score each substring in the TIS-upstream sequence and select the one with the highest score as the potential signal.

Table 1. Statistics of confirmed TISs (as of July 2007)

Kingdom	Group	Genome No.	IPT No.	CDC No.	HSC No.	Gene No.[a]
Archaea	*Crenarchaeota*	8	207	4238	1454	4729
	Euryarchaeota	23	156	13 175	7330	15 389
	Nanoarchaeota	1	0	152	54	163
Bacteria	*Acidobacteria*	2	0	1836	932	2213
	Actinobacteria	36	286	21 137	12 561	27 297
	Aquificae	1	3	687	350	746
	Bacteroidetes/Chlorobi	11	10	6782	5431	8516
	Chlamydiae/Verrucomicrobia	11	2	3405	2212	3858
	Chloroflexi	2	0	911	701	1099
	Cyanobacteria	22	277	13 461	9 682	16 855
	Deinococcus-Thermus	4	99	2011	1276	2603
	Firmicutes	89	864	63 647	51 122	77 320
	Fusobacteria	1	1	773	419	837
	Planctomycetes	1	0	439	429	655
	Alphaproteobacteria	53	138	31 928	32 043	45 656
	Betaproteobacteria	36	52	23 893	23 601	34 828
	Gammaproteobacteria	103	6 716	93 832	93 644	127 165
	Deltaproteobacteria	14	17	8932	7416	11 996
	Epsilonaproteobacteri	11	39	6051	4143	6911
	Spirochaetes	9	23	3906	2580	4635
	Thermotogae	1	8	497	292	591
	Other Bacteria	1	0	499	359	663
Sum	–	440	8 898	302 192	258 031	394 725

[a]Number of genes with at least one confirmed TIS. A TIS might be confirmed by several evidences. About 1–2% of the genes have more than one confirmed TIS.

Details of the methods are available from http://mech.ctb.pku.edu.cn/protisa

DATA STATISTICS

As of July 2007, ProTISA provides refined annotations for 440 genomes in RefSeq, with more than 390 000 confirmed TISs: IPT (8898), CDC (302 192) and HSC (258 031) (Table 1). The percentage of confirmed TISs in a genome varies from 9% to 73% with an average of 33%. The group *Gammaproteobacteria* contributes to the majority of the data collection, which is much more evident for the IPT TISs.

Transcriptional and translational signals are classified into three classes: SD-like, TA-like and atypical signals. Of the 440 genomes, near half (212) were reported with only SD-like signals (mainly in *Firmicutes* and *Proteobacteria*), 22 genomes with only atypical signals (*Bacteroidetes/Chlorobi* and *Cyanobacteria*). The other genomes were found with dual signals: SD-like and TA-like signals were found in 76 genomes (*Actinobacteria* and *Archaea*), SD-like and atypical signals in 126 genomes (*Proteobacteria*) and TA-like and atypical signals in four genomes (Table 2).

DATA ACCESS

ProTISA was implemented under the Apache/PHP/MySQL environment on Linux platform. The basic functionalities aim to browse the stored data and to search the database with a user-specified input.

The browse page is composed of two sections. The first section shows general information for a genome such as organism name, taxonomic group and genomic GC-content. This section also displays a sequence logo (21) and a histogram of the spacer length for each signal (Figure 1). The second section contains TIS annotation with start site and initiation signal for each gene. It shows the gene coordinate, gene identity (PID and gene name), TIS evidence type (i.e. IPT or CDC or HSC or MED), and the predicted signal. It also provides links to the evidence that supports the proposed TIS to be confirmed: PMID or external database links for IPT TIS, conserved domain search results for CDC-TIS and multiple N-terminal sequence alignments among orthologous genes for HSC TIS.

The webpage provides the user with a friendly interface to search the TIS annotation by specifying a region in the genome sequence or by gene identifier such as name and PID. Users can also specify the TIS evidence types and compare the output with the RefSeq annotation.

The annotation, based on which the web server is constructed, is available for download in batch. The files can be easily imported into a database management system such as MySQL. In addition, source codes (written in C++) for the generation of CDC/HSC TIS, the new version of MED-Start, and the motif finding algorithm are freely available in our website under the GNU GPL license. Besides, referenced genes for HSC TIS creation were compiled in a FASTA file for download.

CONCLUSIONS AND FUTURE DIRECTIONS

Despite the remarkable progresses made in computational annotation, there are continuous publications concerning

Table 2. Statistics of genomes with specific signals (as of July 2007)

Kingdom	Group	SD_like only	Atypical only	SD_like and TA_like	SD_like and Atypical	TA_like and Atypical
Archaea	*Crenarchaeota*	–	–	6	2	–
	Euryarchaeota	5	–	16	–	2
	Nanoarchaeota	–	–	1	–	–
Bacteria	*Acidobacteria*	–	–	2	–	–
	Actinobacteria	1	–	33	–	2
	Aquificae	–	–	1	–	–
	Bacteroidetes/Chlorobi	–	6	–	5	–
	Chlamydiae/Verrucomicrobia	1	–	–	10	–
	Chloroflexi	1	–	1	–	–
	Cyanobacteria	–	11	–	11	–
	Deinococcus-Thermus	–	–	4	–	–
	Firmicutes	79	2	8	–	–
	Fusobacteria	1	–	–	–	–
	Thermotogae	1	–	–	–	–
	Planctomycetes	–	1	–	–	–
	Alphaproteobacteria	15	1	–	37	–
	Betaproteobacteria	2	–	–	34	–
	Gammaproteobacteria	82	1	–	20	–
	Deltaproteobacteria	9	–	2	3	–
	Epsilonaproteobacteria	11	–	–	–	–
	Spirochaetes	3	–	2	4	–
	OtherBacteria	1	–	–	–	–
Sum	–	212	22	76	126	4

TIS annotation quality in the public database such as RefSeq (1,5,9,22,23). A notable feature of ProTISA is the compilation of reliable TIS by collecting evidences from experiments, literature, conserved domain search, sequence alignment and accurate prediction. It is interesting to apply the most reliable resource, IPT TISs, to estimate the reliability of CDC TISs, HSC TISs and MED TISs. After removing redundancy from closely related genomes, we have collected a set of 3413 IPT TISs as benchmark, on which the CDC TISs report an accuracy of 99.4% and the HSC TISs report an accuracy of 99.0%. For MED TISs predicted by the modified MED-Start algorithm, the accuracy against the same benchmark achieves to 92.9%.

It is argued that the signal upstream of TIS usually implies the translation initiation mechanism (3,9,14,17,24). Another merit of ProTISA is the annotation of transcriptional and translational signals, which is visualized for each genome by a sequence logo for the signal content and a histogram for the spacer length distribution to TISs (Figure 1). This would be helpful for biologists to speculate the initiation mechanism for a specific genome (14). For example, in addition to the SD-like motif, we found in *Streptomyces coelicolor* and 'TANNNT' motif that highly resembles the Pribnow box reported previously, which generally locates at 10 bps upstream to the transcription start site (TSS) (20). Moreover, the motif has a conserved position about 10 bps upstream to the TIS, counting from TIS to the 5′ T in the 'TANNNT' motif (not shown in Figure 1). In other words, for some genes, the TSS locates just a few base pairs upstream to or even overlaps with the TIS, resulting in a leaderless gene. This would lead one to speculate the existence of initiation mechanism for leaderless gene in *S. coelicolor*, which is consistent with the results reported in (25). Interestingly, this motif was also found in several bacteria groups, for example *Actinobacteria*, *Deinococcus-Thermus* and *Firmicutes*, implying that leaderless gene may not be a marginal phenomenon as usually believed in terms of gene structure in bacteria (26).

An atypical signal is likely to be functional, especially given its conserved position to the TIS. For instance, we detected in *Synechocystis* sp. PCC 6803 a conserved dual-pyrimidine locating immediately upstream to the TISs (Figure 1D), which is consistent with the findings in (24). This phenomenon was also found in several other genera. Such signal could serve as a target for biologists to decipher its regulation role by experiments, thus leading to a better understanding of the initiation mechanism.

With the growing number of completely sequenced bacterial and archaeal genomes, the scientific value of a specific resource for TISs and the corresponding initiation signals is clear. We hope to increase the update frequency so that the database stays current as each new prokaryotic genome becomes available at NCBI. We plan to make ProTISA to be an evolving resource and add significant functionality over time. One direction of ongoing development is to explore the way of comparative genome analysis based on the divergent translation initiation mechanisms for Bacteria and Archaea. Doing this is somewhat challenging since it is difficulty to develop a quantitative model to describe these complex mechanisms. ProTISA may also add items such as genes classification based on their initiation signals. To sum up, we believe that the resource will expand to suit the needs and requests of the research community for translation initiation studies.

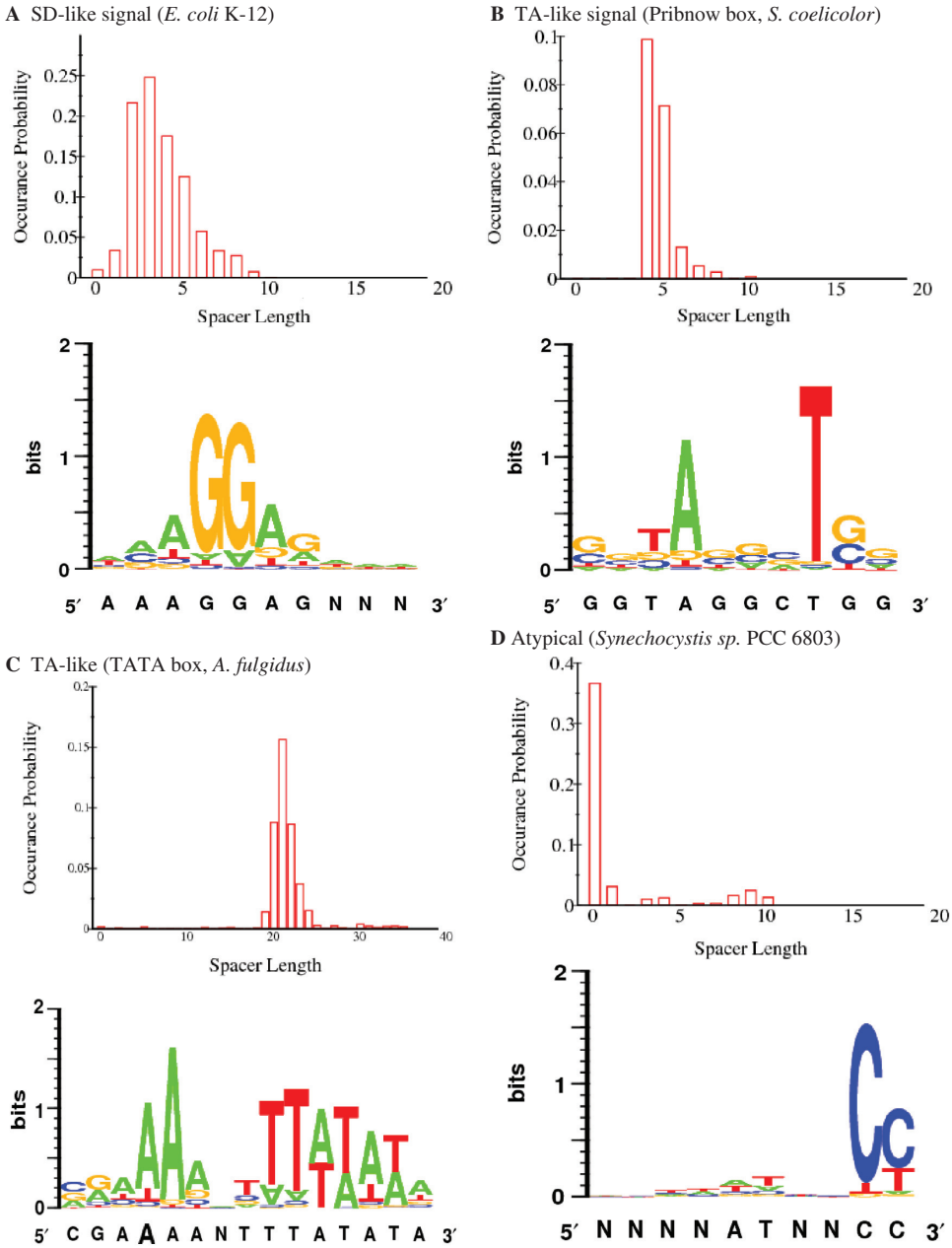

Figure 1. Sequence logo and spacer length distribution of representative signals for the genomes (**A**) *E. coli* k-12; (**B**) *S. coelicolor*; (**C**) *A. fulgidus*; and (**C**) *Synechocystis sp.* PCC 6803. The positional weight matrix of the signal is visualized by a sequence logo in which the height of a letter on a given position is proportional to its occurring frequency. A letter is bottom-up shown if the occurring frequency is lower than that from the background. The consensus is shown below the logo. The spacer length is defined as the distance (or the number of nucleotides) between the TIS and each of all annotated signals, which are calculated by the positional weight matrix visualized in sequence logo.

ACKNOWLEDGEMENTS

We thank Mr Zhou Chang-Ling and Mr Yu Da-Qi for technique support. The work received partial support by the National Natural Science Foundation (10225210 and 30770499) of China; the work was also supported by the National Basic Research Program of China (973 Program) under grant (No. 2003CB715905). Funding to pay the Open Access publication charges for this article was provided by the National Basic Research Program of China.

Conflict of interest statement. None declared.

REFERENCES

1. Poole,F.L.II, Gerwe,B.A., Hopkins,R.C., Schut,G.J., Weinberg,M.V., Jenney,F.E.Jr and Adams,M.W.W. (2005) Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. *J. Bacteriol.*, **187**, 7325–7332.
2. Makita,Y., de Hoon,M.J.L. and Danchin,A. (2007) Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes. *BMC Bioinformatics*, **8**, e47.
3. Ma,J., Campbell,A. and Karlin,S. (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted

expression levels and operon structures. *J. Bacteriol.*, **184**, 5733–5745.
4. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
5. Aivaliotis,M., Gevaert,K., Falb,M., Tebbe,A., Konstantinidis,K., Bisle,B., Klein,C., Martens,L., Staes,A. *et al.* (2007) Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. *J. Proteome Res.*, **6**, 2195–2204.
6. 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.
7. Zhu,H.-Q., Hu,G.-Q., Ouyang,Z.-Q., Wang,J. and She,Z.-S. (2004) Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics*, **20**, 3308–3317.
8. Suzek,B.E., Ermolaeva,M.D., Schreiber,M. and Salzberg,S.L. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.
9. Zhu,H., Hu,G.-Q., Yang,Y.-F., Wang,J. and She,Z.-S. (2007) MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics*, **8**, e97.
10. Tech,M. and Meinicke,P. (2006) An unsupervised classification scheme for improving predictions of prokaryotic TIS. *BMC Bioinformatics*, **7**, e121.
11. Ou,H.Y., Guo,F.B. and Zhang,C.T. (2004) GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell Biol.*, **36**, 535–544.
12. Besemer,J., Lomsadze,A. and Borodovsky,M. (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.*, **29**, 2607–2618.
13. Lewin,B. (2004) *Genes VIII*. Pearson Prentice Hall, Upper Saddle River, N.J.
14. Torarinsson,E., Klenk,H.P. and Garrett,R.A. (2005) Divergent transcriptional and translational signals in Archaea. *Environ. Microbiol.*, **7**, 47–54.
15. Londei,P. (2005) Evolution of translational initiation: new insights from the archaea. *FEMS Microbiol. Rev.*, **29**, 185–200.
16. Moll,I., Grill,S., Gualerzi,C.O. and Bläsi,U. (2002) Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol. Microbiol.*, **43**, 239–246.
17. Chang,B., Halgamuge,S. and Tang,S.L. (2006) Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene*, **373**, 90–99.
18. Larsen,T.S. and Krogh,A. (2003) EasyGene–a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, e21.
19. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
20. Hershberg,R., Bejerano,G., Santos-Zavaleta,A. and Margalit,H. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.*, **29**, 277.
21. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
22. Nielsen,P. and Krogh,A. (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, **21**, 4322–4329.
23. Starmer,J., Stomp,A., Vouk,M. and Bitzer,D. (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput. Biol.*, **2**, e57.
24. Sazuka,T. and 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.
25. Strohl,W.R. (1992) Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res.*, **20**, 961–974.
26. Wu,C.J. and Janssen,G.R. (1996) Translation of vph mRNA in *Streptomyces lividans* and *Escherichia coli* after removal of the 5′ untranslated leader. *Mol. Microbiol.*, **22**, 339–355.