D338–D343 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj060

ICDS database: interrupted CoDing sequences in prokaryotic genomes

Emmanuel Perrodou, Caroline Deshayes[1], Jean Muller, Christine Schaeffer[2], Alain Van Dorsselaer[2], Raymond Ripp, Olivier Poch, Jean-Marc Reyrat[1] and Odile Lecompte*

Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, France, [1]Inserm-UMR 570, Unité de Pathogénie des Infections Systémiques, Groupe Avenir, Paris Cedex 15, F-75730, France and [2]Laboratoire de Spectrométrie de Masse Bio-Organique (LSMBO) UMR 7512, ECPM, 25 rue Becquerel, Strasbourg F-67087 Cedex 2, France

Received August 9, 2005; Revised and Accepted October 5, 2005

ABSTRACT

Unrecognized frameshifts, in-frame stop codons and sequencing errors lead to Interrupted CoDing Sequence (ICDS) that can seriously affect all subsequent steps of functional characterization, from *in silico* analysis to high-throughput proteomic projects. Here, we describe the Interrupted CoDing Sequence database containing ICDS detected by a similarity-based approach in 80 complete prokaryotic genomes. ICDS can be retrieved by species browsing or similarity searches via a web interface (http://www-bio3d-igbmc.u-strasbg.fr/ICDS/). The definition of each interrupted gene is provided as well as the ICDS genomic localization with the surrounding sequence. Furthermore, to facilitate the experimental characterization of ICDS, we propose optimized primers for re-sequencing purposes. The database will be regularly updated with additional data from ongoing sequenced genomes. Our strategy has been validated by three independent tests: (i) ICDS prediction on a benchmark of artificially created frameshifts, (ii) comparison of predicted ICDS and results obtained from the comparison of the two genomic sequences of *Bacillus licheniformis* strain ATCC 14580 and (iii) re-sequencing of 25 predicted ICDS of the recently sequenced genome of *Mycobacterium smegmatis*. This allows us to estimate the specificity and sensitivity (95 and 82%, respectively) of our program and the efficiency of primer determination.

INTRODUCTION

The availability of numerous complete genomes and large cDNA collections provides the opportunity to investigate gene and protein function at an unprecedented scale, as demonstrated by the numerous projects in proteomics and structural genomics. These high-throughput studies are hindered by technical bottlenecks, in particular in the field of automation, but also depend on reliable sequence data for genes and proteins. Introduction of errors at the first stage of genome analysis, i.e. sequencing and gene prediction, can have a serious impact on all subsequent studies. For instance, assignment of correct gene and protein sequences is crucial for the production of a functional protein or for peptide identification in mass spectrometry. In eukaryotic genomes, the quality of gene determination is improved by the vast amount of transcriptomic data based on sequencing strategies, including EST projects, SAGE, alternative splicing determination, etc. (1). In contrast, annotation of prokaryotic genomes still largely relies on *ab initio* prediction programs. Thus, curation of predicted CoDing Sequence (CDS) in prokaryotes is a vital investment for maintaining and enhancing the use of the genomic information in the post-genomic era.

During genome annotation, the first source of errors is the sequence itself. It has been measured that error rate for finished genomes is one error in 10^3–10^5 bases (2). Most sequencing errors involve base substitutions, which have limited effects on gene prediction if they not introduce a stop codon, but some lead to insertion/deletion of bases producing artificial frameshifts in the coding region. The second source of errors is linked to gene prediction during the annotation process, errors in start codon prediction particularly and problems in detection of authentic frameshifts or in-frame stop codons in putative

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 00; Fax: +33 3 88 65 32 01; Email: lecompte@igbmc.u-strasbg.fr

genes. These generally lead to the prediction of Interrupted CoDing Sequence (ICDS) during genome annotation. Many examples of programmed translational frameshifts have been studied in different organisms, from viruses to eukaryotes (3–5). A frameshift causes the ribosome to pause at a slippery site, leading to a shift of one base, either −1 or +1. Signals involved in −1 and +1 frameshifts are distinct. In −1 frameshifts, major signalling elements consist of a slippery site where the ribosome changes reading frame and a stimulatory RNA secondary structure pseudoknots located a few nucleotides downstream. Most of the slippery site consists of a heptameric sequence (X XXY YYZ), but divergent sequences exist that often depend on the organism (6). The +1 frameshifts are less common and more difficult to detect since the slippery site is not conserved. Prediction of ICDS is also observed for in-frame insertion of stop codon that can be an indicator of a natural nonsense suppression event, i.e. the reading of stop codons as sense codons by natural suppressor tRNAs (7). This recognition requires unconventional base pairing in anticodon–codon interactions as well as a codon context effect, i.e. primary sequences and secondary structures in the vicinity of stop codons, that influence the efficiency of suppression. The codon context can simply consist of only 1–6 nt at the 3′ side of the suppressed stop codon or may involve more complex signals like stem–loop or pseudoknot structures. Whatever their origins (sequencing errors or programmed events), unrecognized frameshifts and in-frame stop codons lead to the prediction of ICDS that could code for a unique protein. Automatic detection of these mispredictions is not a trivial process since signals ruling authentic events such as programmed frameshifts are small, diverse and depend on the organism.

Several computational tools have been described to detect authentic frameshifts and/or errors during DNA sequencing. The first class of programs tries to locate authentic frameshifts using known signals of frameshifting, such as pseudoknot structures or slippery sequences independently (8–10). Moon *et al*. (11) have developed a tool to detect frameshifts, taking into account X XXY YYZ sites as well as secondary structure elements for −1 frameshifts and specific signals from +1 frameshifts that are conserved among species. As noted by the authors, these programs only permit to predict a certain type of authentic frameshift since they rely on documented signals of known frameshifts. Moreover, they are not designed to detect sequencing errors. The second class of programs has been developed to detect frameshifts of artifactual or authentic origin. Some of these programs are based on the comparison of translated DNA in all six reading frames with databases of protein sequences (12–16). Frameshift detection thus relies on the presence of a related protein in the databases. To overcome this drawback, some tools based on the intrinsic properties of coding sequences have been developed (16). Unfortunately, most of these second classes of tools are in-home programs that cannot be locally installed or are not designed for high-throughput analyses.

The spectacular increase and diversification of the protein sequence universe, covering a wide range of prokaryotic phylogenetic branches, allows now a high-throughput approach for the reliable and systematic detection of ICDS. Here, we describe a database generated using a program that detects ICDS in whole prokaryotic genomes and provides regions that require re-sequencing for the scientific community implicated in post-genomic projects (http://www-bio3d-igbmc.u-strasbg.fr/ICDS/). This database relies on a program that uses protein similarities to detect ICDS. In a first step, the program has been tested analysing three bacterial genomes in detail: two genomes of the same strain of *Bacillus licheniformis* that have permitted a cross validation of the program, and the genome of *Mycobacterium smegmatis* from whom several predicted ICDS have been re-sequenced. In a second step, the program has been run on sequenced genomes that represent a great interest for high-throughput structural projects. The number of ICDS vary from 2 to 258 ICDS/million bp. Results are accessible via a web service and will be updated for incoming prokaryotic genomes. ICDS can be retrieved by species browsing or similarity searches (blastN or blastP). Species browsing display all the ICDS for a given genome. In order to facilitate the biologist work, we provide the genome localization for each ICDS and the name of previously annotated genes, as well as optimal PCR primer sequences for each zone to be re-sequenced. This tool should help in genome sequence refinement and annotation and provide a reliable and systematic screen for potential sequencing errors, potential programmed frameshifts, internal stop and other events that can affect subsequent analysis.

MATERIALS AND METHODS

Program of interrupted CDS detection

The ICDS detection program is written in Tcl and integrated into the GScope high-throughput genomic platform that allows to handle, visualize and analyse genomic, cDNA or protein sequences in a user-friendly interface (R. Ripp, manuscript in preparation). The principle underlying our program is the detection of adjacent open reading frames (ORFs) on the same DNA strand that share common homologues. The program can scan annotated microbial genomes or raw genomic sequences. In this latter case, the genomic sequence is first analysed by an ORF prediction program, such as Glimmer (17). ORFs are translated and the protein sequences compared with a public protein database using blastP (18). The program proceeds as follows:

i. the 10 top blast hits ($E < 10^{-3}$) are extracted for each ORF;
ii. the list of homologues of an ORF is then compared with the lists obtained for adjacent ORFs. The comparison has been extended to the four neighbouring ORFs to limit effects of small overpredicted ORFs;
iii. pairs of proteins exhibiting at least one common homologue are retained. Such a pair can correspond to ICDS or to paralogous adjacent ORFs (Figure 1). If a significant similarity ($E < 10^{-3}$) is detected between the components of a pair, those ORFs are considered as paralogues and discarded from the analysis while absence of similarity define ICDS;
iv. the approximate genomic localization of the CDS rupture is calculated from the blastP HSPs.

Briefly, limits of the blastP HSPs are extracted and translated into genomic coordinates by our genomic platform that allows to physically localize each predicted genes. A region of 500 bp surrounding the CDS rupture is extracted from the genomic sequence and scanned to automatically design optimal sequencing primers.

Automatic design of primers

The sequencing primers have been designed using an optimized version of the CADO4MI program (Computed Assisted Design of Oligonucleotide for MIcroarray) (J. Muller, manuscript in preparation). The query sequence is scanned using a sliding window analysis with window length set to primer size (e.g. 21 nt) and step-size 10. The melting temperature (T_m) is calculated using the Wallace rules (19). Only 21mers with $T_m = 63 \pm 5°C$ are considered.

The 21mers are compared with the complete reference genome using the blastN program to assess specificity and to avoid hybridization with another part of the genome. Sequence selection is carried out automatically by selecting the primer pairs which have high specificity and the shortest amplification area. Primers have been searched excluding the 50 bp surrounding the ICDS and for a maximum length of 500 bp.

Database

Genome sequences were obtained from the NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) except for the genome of *M.smegmatis* from the TIGR (http://www.tigr.org/). For this study, ORF predictions were made using Glimmer (17). All the detected ICDS are organized in the Interrupted CoDing Sequence database, accessible via a web server http://www-bio3d-igbmc.u-strasbg.fr/ICDS/.

Sequencing

The chromosomal DNA of *M.smegmatis* strain mc^2155 used for PCR amplification was purified as described previously (20). Pairs of primers (Supplementary Table 1) were used for amplification using *Pfu* Turbo DNA polymerase (Stratagene) or DyNAzyme DNA polymerase (Finnzymes). PCR samples were loaded onto a 0.8% agarose gel and the fragments were cut out from the gel and purified using QIAquick PCR purification kit (Qiagen). Purified PCR fragments were used as templates in sequencing reactions with each primer used for PCR amplification.

RESULTS

General principle

The database was created using a program that relies on the analysis of physically adjacent genes to predict putative ICDS in complete genomes. Pairs of adjacent genes that exhibit at least one common homologue are defined as 'CDS containing common hits' and can correspond to pair of adjacent paralogues or to adjacent ICDS. Paralogues are excluded from the analysis by a research of homology between the two 'CDS containing common hits'. Remaining CDSs are considered as ICDS (Figure 1), indicating frameshifts or in-frame stop codon insertion, due to sequencing errors or to authentic events such as programmed frameshifts or natural nonsense suppression. The program also detects genes that exist as fusion genes in other genomes or corresponding to pseudogenes (21). Distinct neighbouring genes can match a gene fusion in the database, suggesting an apparent ICDS. In this case, genes are indicated and considered as ICDS.

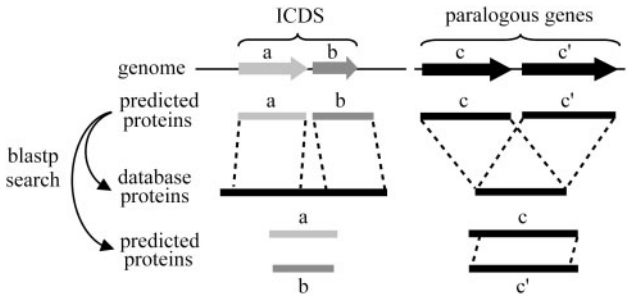

Figure 1. General principle of ICDS detection. (**a**) Pairs of proteins exhibiting at least one common homologue are retained. Such a pair can correspond to ICDS or to paralogous adjacent genes. (**b**) No significant similarity ($E < 10^{-3}$) is detected between the components of a couple, those genes are considered as ICDS. (**c**) If a significant similarity exists, genes are considered as paralogues and discarded from the analysis.

Validation using *B.licheniformis*

The program has been tested using the recently annotated genome of *B.licheniformis* (22). This genome represents a good test since the authors have manually identified some frameshifts and re-sequenced each zone they retrieved. Our program detected 144 'CDS containing common hits' pairs. Of these, 40 pairs are predicted as adjacent paralogues. A manual verification demonstrated that all were true paralogues. The remaining 104 pairs of adjacent genes were predicted as ICDS. A manual verification showed only one false positive prediction corresponding to remote paralogues. We estimated the accuracy of our program by calculating its specificity [defined as TP/(TP + FP), with TP and FP the true positive and false positive predictions, respectively] and obtained a specificity >99%. Of the 27 frameshifts identified by the authors, 23 were predicted by the program as ICDS. The others were not detected either because the genes were not predicted by Glimmer, or because of the default threshold of the blastP. Thus, our program was able to detect 89% of the manually established frameshifts and identified 81 new ICDS corresponding either to sequencing errors or to authentic events.

The genome of the same strain of *B.licheniformis* has also been sequenced by a second team (23), allowing to cross-validate our results. In this second genomic sequence, the program detected 58 ICDS (57 true positive and 1 false positive), leading to a specificity of 98%. With one exception, all of them were also identified in the first genome. ICDSs detected in only one of the two genomes (47 ICDS for the first genome and one ICDS for the second) are all due to a divergence between the two genomic sequences that create a frameshift in only one of the two genome sequences. These differences may originate from sequencing errors or from intraspecies polymorphisms (due to different isolates of *B.licheniformis* strain or to spontaneous mutations in the cloned fragment during library construction).

The sensitivity [defined as TP/(TP + FN) with FN the false negative predictions] of our program is much more complicated to estimate since we lack an appropriate benchmark: we do not know how many frameshifts, in-frame stop codons and sequencing errors a genome really contains. Thus, we created 100 artificial ICDSs by random insertion or deletion of a single base in existing CDS of the second *B.licheniformis* genome

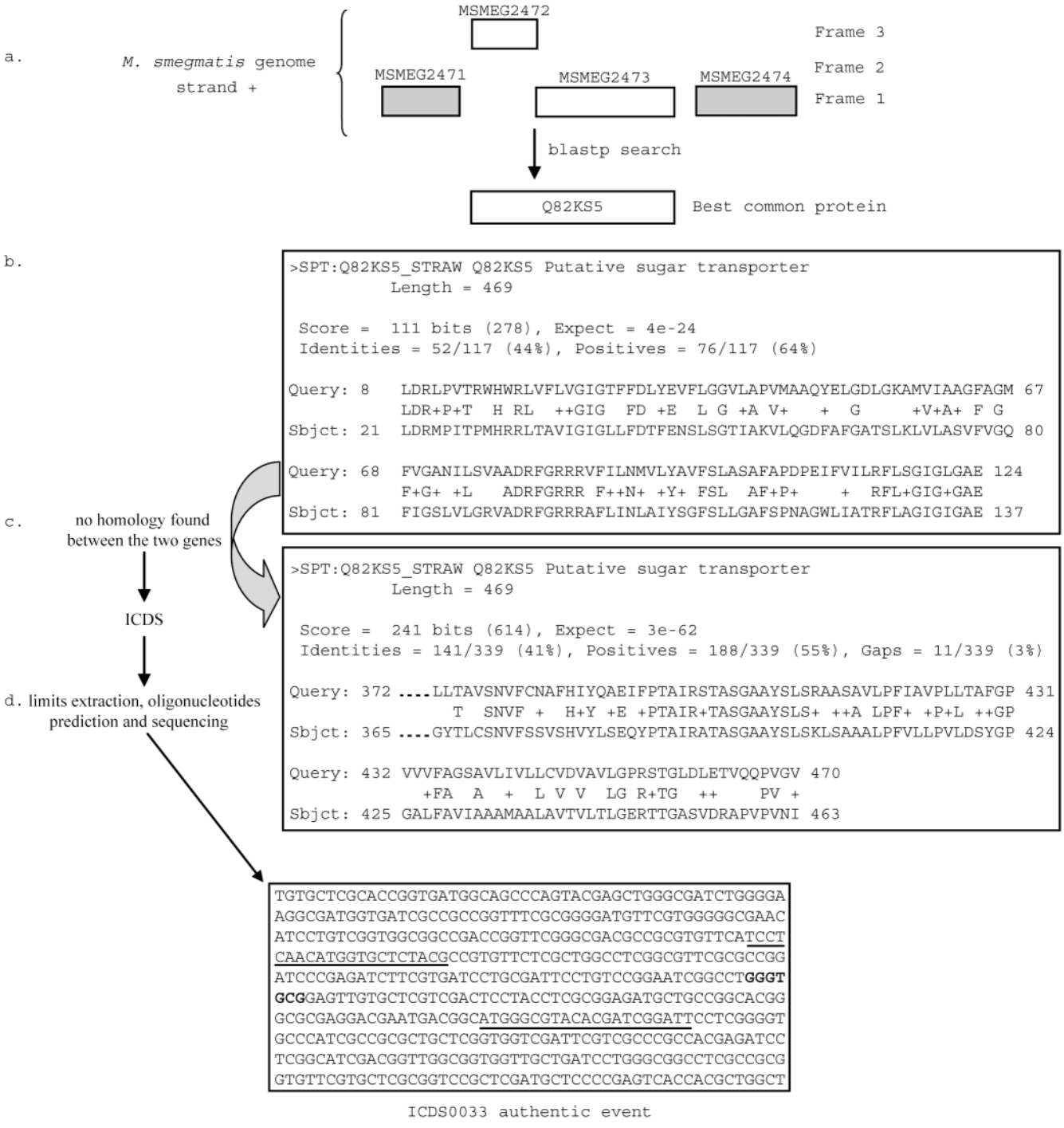

Figure 2. Example of an ICDS corresponding to an authentic event of *M.smegmatis*. (**a**) Representation of the genomic region showing the best protein. (**b**) blastP alignment of the two predicted proteins. (**c**) Research for homology between the two predicted proteins defining an ICDS. (**d**) Limits extraction and prediction of primers. The region has been re-sequenced and results are shown. The sequence in black represents the region where the frameshift occurs. The underlined sequences represent the primers.

(23). We obtained a sensitivity of 82% (82 true positives and 18 false negatives). Of the 18 artificial ICDS not detected by our program, 4 were predicted as adjacent paralogues. The remaining 14 false negatives were not detected because of absence of predicted CDS. This is a drawback of our method since only predicted genes are screened. Moreover, putative genes that do not share any similarities with sequences in the public databases are excluded.

Validation using *M.smegmatis*

We also tested our program on a recently sequenced genome: *M.smegmatis*. It belongs to the mycobacteria group that includes *Mycobacterium tuberculosis* and *Mycobacterium leprae*, the causative agents of tuberculosis and leprosy, respectively, and represents a tractable model to study some functional aspects of the pathogenic species (24). It is an ideal

test case for ICDS prediction since it has a very high G+C content that can dramatically increase the number of over-predicted genes (25).

180 'CDS containing common hits' pairs were detected by the program (86 pairs of adjacent paralogues and 94 ICDS) (see an example Figure 2). Of the 86 predicted paralogue pairs, 4 correspond to supplementary ICDS. These errors were in genes containing sequence or domain repeats. Of the 94 predicted ICDS, 3 are false positives, corresponding in fact to paralogues. The remaining 91 pairs correspond to ICDS (85 frameshifts, 6 in-frame stop codons). In particular, we detect the only putative authentic frameshift documented in *M.smegmatis* (26). This authentic frameshift occurs in a putative ABC-transport operon. These genes may cooperate to produce an integral membrane component of the transport system via a programmed translational frameshift.

In order to determine whether these events are due to sequencing errors or to programmed events, 25 regions have been re-sequenced using the predicted primers. Of the 25 re-sequenced regions, 11 events were due to sequencing errors (6 nt insertions and 5 nt deletions) (Supplementary Table 1). The remaining 14 regions do not contain sequencing errors and may correspond to non-functional genes (pseudogenes), to fission events, or to genes subject to translational frameshifts or tRNA suppression. No clear slippery site has been found within the sequence of these events, highlighting the necessity of re-sequencing to discriminate between authentic events and sequencing errors. Moreover, PCR amplification using our predicted primers permits us to validate the primer prediction program, showing the efficiency and reliability of this determination. As shown by our comparison of detected ICDS (for *B.licheniformis*) and re-sequencing studies, a great vigilance concerning sequencing errors is essential.

DISCUSSION

At the time of writing, 80 complete prokaryotic genomes are included in the ICDS Database (Supplementary Table 2). This analysis included all the prokaryotic target species chosen in worldwide initiatives of structural genomics (see http://www.rcsb.org/pdb/strucgen.html for details on ongoing projects) as well as bacteria and archaea with a variety of genome size and GC content. The user-friendly interface offers an efficient access to ICDS by species name or by similarity search. For each ICDS, we provide the definition of the reference gene, the name of the previously annotated genes and the genomic localization of the frameshift or sequencing error. We also supply two optimal PCR primers specific to the ICDS to be re-sequenced, together to the T_m and length and the sequence of the 500 bp surrounding the ICDS. Some PCR primer pairs cannot be predicted for diverse reasons:

i. existence of nearly identical paralogues in the genome (transposases for example);
ii. extended ICDS surrounding sequence (>500 bp). In this case, we do not predict optimized primers. This could be due for example to presence of a fusion protein in the databank, to false positives ICDS (paralogues) or to remote genes;
iii. absence of optimized primer in term of pre-defined T_m. This particularly occurs in G+C rich or low genomes.

We obtained an average number of 74 ICDS per tested genome and of 27.7 ICDS per million bp. The number of ICDS is variable among the analysed species: from 2 for *Nanoarchaeum equitans* to 258 ICDS/million bp for the *M.leprae*. Indeed, it has been shown that *M.leprae* undergoes reductive evolution and that less than half of the genome contains functional genes but pseudogenes (27). Surprisingly, G+C content does not seem to influence the number of ICDS per annotated gene, although sequencing errors are supposed to be more frequent in high G+C genomes.

We have shown the robustness of our program in term of specificity (>95%) and sensitivity (82%). However, some drawbacks are linked to the prediction of the genes corresponding to ICDS. We have chosen Glimmer since this program is known to overlook few genes at the expense of the number of overpredicted genes. This overprediction is important in certain genomes but should not influence the number of detected ICDS since only genes with homologues are considered. Another limit is that detection of ICDS is closely linked to the existence of at least one homologue in the sequence databases. For example, *N.equitans* is the first representative of the Nanoarchaeota, a new archaeal kingdom, and exhibits a high number of genes without homologues and the genome is supposed to contain more ICDS than detected in this study.

By providing in the database an exhaustive list of potential ICDS predicted in a genome, pre-calculated primers for sequencing projects, as well as the predicted function of the protein, we hope that some of the drawbacks observed in high-throughput projects will be removed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Caroline Deshayes is funded by a doctoral grant of Inserm–Région Ile de France. Jean-Marc Reyrat is Charge de Recherches at Inserm. This work was funded by the INSERM, the CNRS, the ULP de Strasbourg, and "Proteomique et genie des proteines" (project n° PGP 04-013) grant. This work was achieved using the RNG (Réseau National de Génopoles) Strasbourg Bioinformatics Platform infrastructures. Funding to pay the Open Access publication charges for this article was provided by the CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Bianchetti,L., Thompson,J.D., Lecompte,O., Plewniak,F. and Poch,O. (2005) vALId: validation of protein sequence quality based on multiple alignment data. *J. Bioinform. Comput. Biol.*, **3**, 1–19.
2. Weinstock,G.M. (2000) Genomics and bacterial pathogenesis. *Emerg. Infect. Dis.*, **6**, 496–504.
3. Farabaugh,P.J. (1996) Programmed translational frameshifting. *Annu. Rev. Genet*, **30**, 507–528.
4. Baranov,P.V., Gesteland,R.F. and Atkins,J.F. (2002) Recoding: translational bifurcations in gene expression. *Gene*, **286**, 187–201.
5. Baranov,P.V., Gurvich,O.L., Fayet,O., Prere,M.F., Miller,W.A., Gesteland,R.F., Atkins,J.F. and Giddings,M.C. (2001) RECODE: a

database of frameshifting, bypassing and codon redefinition utilized for gene expression. *Nucleic Acids Res.*, **29**, 264–267.
6. Farabaugh,P.J. (2000) Translational frameshifting: implications for the mechanism of translational frame maintenance. *Prog. Nucleic Acid. Res. Mol. Biol.*, **64**, 131–170.
7. Beier,H. and Grimm,M. (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res.*, **29**, 4767–4782.
8. Bekaert,M., Bidou,L., Denise,A., Duchateau-Nguyen,G., Forest,J.P., Froidevaux,C., Hatin,I., Rousset,J.P. and Termier,M. (2003) Towards a computational model for −1 eukaryotic frameshifting sites. *Bioinformatics*, **19**, 327–335.
9. Hammell,A.B., Taylor,R.C., Peltz,S.W. and Dinman,J.D. (1999) Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. *Genome Res.*, **9**, 417–427.
10. Shah,A.A., Giddings,M.C., Parvaz,J.B., Gesteland,R.F., Atkins,J.F. and Ivanov,I.P. (2002) Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18**, 1046–1053.
11. Moon,S., Byun,Y., Kim,H.J., Jeong,S. and Han,K. (2004) Predicting genes expressed via −1 and +1 frameshifts. *Nucleic Acids Res.*, **32**, 4884–4892.
12. Posfai,J. and Roberts,R.J. (1992) Finding errors in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 4698–4702.
13. Claverie,J.M. (1993) Detecting frame shifts by amino acid sequence comparison. *J. Mol. Biol.*, **234**, 1140–1157.
14. Guan,X. and Uberbacher,E.C. (1996) Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci.*, **12**, 31–40.
15. Brown,N.P., Sander,C. and Bork,P. (1998) Frame: detection of genomic sequencing errors. *Bioinformatics*, **14**, 367–371.
16. Medigue,C., Rose,M., Viari,A. and Danchin,A. (1999) Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. *Genome Res.*, **9**, 1116–1127.
17. 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.
18. 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.
19. Wallace,R.B., Shaffer,J., Murphy,R.F., Bonner,J., Hirose,T. and Itakura,K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res.*, **6**, 3543–3557.
20. Pelicic,V., Reyrat,J.M. and Gicquel,B. (1996) Generation of unmarked directed mutations in mycobacteria, using sucrose counter-selectable suicide vectors. *Mol. Microbiol.*, **20**, 919–925.
21. Lerat,E. and Ochman,H. (2004) Psi–Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res.*, **14**, 2273–2278.
22. Rey,M.W., Ramaiya,P., Nelson,B.A., Brody-Karpin,S.D., Zaretsky,E.J., Tang,M., Lopez de Leon,A., Xiang,H., Gusti,V., Clausen,I.G. *et al.* (2004) Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related Bacillus species. *Genome Biol.*, **5**, R77.
23. Veith,B., Herzberg,C., Steckel,S., Feesche,J., Maurer,K.H., Ehrenreich,P., Baumer,S., Henne,A., Liesegang,H., Merkl,R., Ehrenreich,A. and Gottschalk,G. (2004) The complete genome sequence of *Bacillus licheniformis* DSM13, an organism with great industrial potential. *J. Mol. Microbiol. Biotechnol.*, **7**, 204–211.
24. Reyrat,J.M. and Kahn,D. (2001) *Mycobacterium smegmatis*: an absurd model for tuberculosis? *Trends Microbiol.*, **9**, 472–474.
25. Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D. and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*, **17**, 425–428.
26. Barsom,E.K. and Hatfull,G.F. (1996) Characterization of *Mycobacterium smegmatis* gene that confers resistance to phages L5 and D29 when overexpressed. *Mol. Microbiol.*, **21**, 159–170.
27. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.