Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W471–W479***
doi:10.1093/nar/gki498

MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes

Stéphane Cruveiller, Jérôme Le Saux, David Vallenet, Aurélie Lajus, Stéphanie Bocs and Claudine Médigue*

Genoscope/UMR-CNRS 8030, Atelier de Génomique Comparative, 2 rue Gaston Crémieux, F-91006 Evry, France

Received February 14, 2005; Revised April 1, 2005; Accepted April 27, 2005

ABSTRACT

The annotation of newly sequenced bacterial genomes begins with running several automatic analysis methods, with major emphasis on the identification of protein-coding genes. DNA sequences are heterogeneous in local nucleotide composition and this leads sometimes to sequences being annotated as authentic genes when they are not protein-coding genes or are true but uncharacterized protein-coding genes. This first annotation step is generally followed by an expert manual annotation of the predicted genes. The genomic data (sequence and annotations) organized in an appropriate databank file format is subsequently submitted to an entry point of the International Nucleotide Sequence Database. These procedures are inevitably subject to mistakes, and this can lead to unintentional syntactic annotation errors being stored in public databanks. Here, we present a new web program, MICheck (MIcrobial genome Checker), that enables rapid verification of sets of annotated genes and frameshifts in previously published bacterial genomes. The web interface allows one easily to investigate the MICheck results, i.e. inaccurate or missed gene annotations: a graphical representation is drawn, in which the genomic context of a unique coding DNA sequence annotation or a predicted frameshift is given, using information on the coding potential (curves) and annotation of the neighbouring genes. We illustrate some capabilities of the MICheck site through the analysis of 20 bacterial genomes, 9 of which were selected for their 'Reviewed' status in the National Center for Biotechnology Information (NCBI) Reference Sequence Project (RefSeq). In the context of the numerous re-annotation projects for microbial genomes, this tool can be seen as a preliminary step before the functional re-annotation step to check quickly for missing or wrongly annotated genes. The MICheck website is accessible at the following address: http://www.genoscope.cns.fr/agc/tools/micheck.

INTRODUCTION

The wealth of sequence information produced by the numerous whole-genome sequencing projects has generated the need for rapid annotation and subsequent biological interpretation of the corresponding sequences. Despite considerable progress in the field of computational genomics, the process of annotation is still a manual, labour-intensive endeavour. Following the release of a first round of annotation, researchers have undertaken manual curation to improve the annotations, providing annotation of previously uncharacterized genes, while correcting a number of errors resulting from erroneous similarity detection. Third-party annotation of individual species has been reported in the literature by various groups, with emphasis on improvement through the re-assignment of hypothetical proteins to proteins with predicted function using the latest resources, such as improved algorithms and richer databases (1). On this occasion, it has been observed that a large discrepancy exists between functions annotated for similar proteins, partly owing to the fact that most of the functional annotations in complete genomes are based on relatively weak sequence identities (2,3).

Despite the emergence of gene finding methods based on hidden Markov models (HMMs, which enable modular modelling of DNA sequence compositional heterogeneities) (4,5), there is still a need for more accurate methods, especially for the prediction of short genes. The issue of choosing the model structure has recently been extensively discussed and has revealed that the choice of the optimal model is clearly species-specific (6). The need for gene finding methods that can overcome the problems presented by intra-genomic

*To whom correspondence should be addressed. Tel: +33 1 60 87 84 59; Fax: +33 1 60 87 25 14; Email: cmedigue@genoscope.cns.fr

variation has also previously been addressed in GeneMarkGenesis (7), which derives two models for each prokaryotic genome according to typical and atypical codon usage clusters in that genome. Our own approach consists of a systematic (semi-automatic) construction of training sets, using multivariate statistical techniques on protein-coding potential computed with the Relative Synonymous Codon Usage (RSCU) indicator. For a given genome, correspondence analysis (CA) and clustering methods are applied to the set of predicted coding DNA sequences (CDSs) or annotated genes, in order to derive gene classes used as training sets to estimate parameters for coding region composition. All these models (which take into account the compositional diversity of genes within a genome) are subsequently used together in the core of our AMIGene method (8), a gene finding program similar to GeneMark (9) in the way it parses sequences to predict CDSs. AMIGene contains an additional heuristic, making it possible to select the most likely CDS when taking into account ambiguous choices between overlapping CDSs and/or the presence of frameshifts in the DNA sequence (8). In previous studies on re-annotation of microbial genome CDSs, most of the newly found short genes revealed the presence of frameshifts that could be either artefacts or genuine frameshifts (10). We therefore subsequently combined AMIGene results with those of ProFED (Prokaryotic Frameshift Errors Detection), a method for finding potential frameshifts using only intrinsic properties of the coding sequences (11). Discrepancies between the new set of syntactic annotations thus obtained and the set of annotations stored in public databanks clearly come from 'accidentally' missing genes and/or identification of genes that other methods did not predict (or which were removed during the manual expert annotation).

In this paper we describe an integrated web-based program that enables rapid verification of CDS and frameshift annotations in a complete bacterial genome. Starting with a file in INSD format (GenBank or EMBL), MICheck first runs the AMIGene and ProFED methods (either with suitable gene models or with a new gene model computed from the user's input annotations), and then compares the set of new predicted CDSs (AMIGene CDSs) with the user annotations (user CDSs). This leads to unique CDSs from the original annotations (Unique_User) and from AMIGene predictions (Unique_AMIGene). The latter are submitted to BLAST comparisons against the UniProt databank (12). A graphical web interface has been developed to allow the annotator to investigate the MICheck results in terms of the coding potential, BLAST similarity and gene context of a unique CDS annotation.

METHODS

For each complete bacterial genome to be analysed, the set of annotated genes ('CDS' and 'gene' features) and the chromosome sequence are extracted from the input INSD file. These data are used as described in the following subsections.

Generating models for gene finding

Already, 81 organisms (a selection of the completely sequenced and annotated bacterial genomes) have been investigated in terms of codon usage differences using multivariate statistical techniques. A range from one gene model (i.e. almost all Archaea genomes) to four gene models (e.g. *Mycobacterium tuberculosis*, *Photorhabdus luminescens*) has currently been defined; these models are used simultaneously in the core of the AMIGene (8) and ProFED (11) programs. The gene models that have been computed for these genomes can be browsed or downloaded from the following URL: http://www.genoscope.cns.fr/agc/tools/micheck/html/database_status.html.

If the input genome has not yet been analysed in terms of codon usage differences (see below), or if it is distant from another genome for which specific gene models have been defined, a new gene model is computed. This step uses the sequences of the annotated genes (the coding training set) and the rest of the sequence is included in the noncoding training set [see Ref. (11) for further explanation of the way a gene model is generated].

Finding unique annotated and/or predicted CDSs

Three main types of CDSs exist following the comparison between the set of annotated databank genes (i.e. user annotations) and the set of AMIGene CDS predictions: (i) CDSs annotated both by the user (input file) and by AMIGene, (ii) CDSs annotated only by the user (this could be an annotated gene with no biological reality) and (iii) CDSs annotated only with the AMIGene method (this could be a missed CDS annotation corresponding to a putative new gene).

The MICheck method comprises three main steps (Figure 1). First, the AMIGene method is run using the input DNA sequence and the gene model(s) to determine a set of predicted genes. The model(s) used at this stage is (are) either readily available on the website or generated automatically using the annotated set of genes provided by the user as the training dataset (Figure 1, step 1). The sets of predicted genes and annotated genes are then compared primarily on the basis of their stop codon, leading to three main lists of genes (Figure 1, step 2): (i) those that are common to the two sets (status: Common), (ii) those that are unique to the AMIGene set of genes; these could be missed CDS annotations corresponding to putative new genes (status: Unique_AMIGene) and (iii) those that are unique to the annotated set of genes, which could be annotated genes with no biological reality (status: Unique_User). The final step consists of translating the unique AMIGene set of genes into proteins and performing similarity searches, using BLAST2P from the LASSAP/BIOFACET package (13), against a prokaryotic database derived from UNIPROT (12) (Figure 1, step 3).

Given the DNA sequence, the gene model(s) and the set of annotated genes (made up of the databank annotations and the additional AMIGene CDS predictions), the ProFED method automatically identifies putative frameshifts. Explanation of the method together with its default parameters and confidence levels associated with the prediction of frameshifts can be found at http://www.genoscope.cns.fr/agc/tools/micheck/html/Method.html.

The core of the MICheck program uses both the AMIGene (8) and the ProFED (11) methods implemented in C language. The binary codes for the MICheck method are available at the following address: http://www.genoscope.cns.fr/agc/ftp.

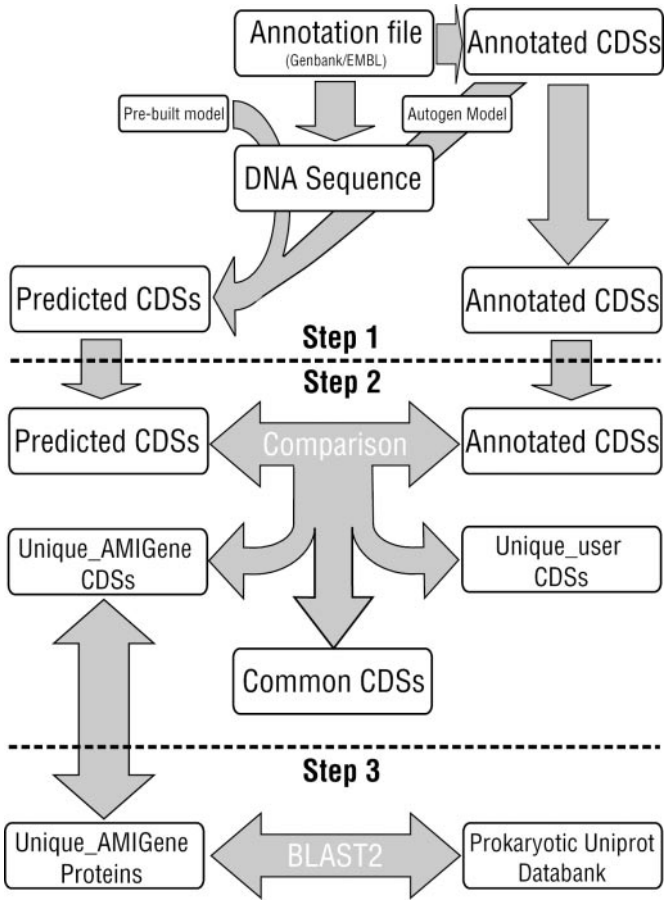

Figure 1. Schematic view of the MICheck method. Given a DNA sequence, the gene model(s) and the set of annotated genes (i.e. the user's CDSs), three main steps are executed. Step 1: the AMIGene method is run using the input DNA sequence and the gene model(s) in order to predict putative CDSs (the parameter values are set to the optimized parameters previously derived from genomes with a similar G+C content, either *Bacillus subtilis*, *Escherichia coli* or *Mycobacterium tuberculosis*) (8). Each CDS is characterized by its position in the DNA sequence and its average coding probability (the highest probability obtained with one of the input gene models). In the same way, a coding probability is computed for each annotated gene (user's CDSs). Step 2: the two sets of CDS annotations are compared for their stop codon position in the genome (some misplacement of the gene start codon may thus be revealed) and three main lists are generated: (i) the list of CDSs shared by the two compared sets of CDSs (status 'Common'); (ii) the list of additional user's CDSs having an average coding probability >0.2 (status 'Unique_User'); (iii) the list of additional AMIGene CDSs of length >300 bp and average coding probability >0.5 (status 'Unique_AMIGene'). The rest of the CDSs are labelled 'No_Status'. Using these strict parameter values, only the most obvious discrepancies between the two sets of annotations are highlighted. Step 3: the subset of unique AMIGene CDSs is translated into protein sequences and compared with a prokaryotic databank built from the UniProt databank (12), with the BLAST2 similarity search program from the LASSAP/BIOFACET package (13). The best databank hit with an *E*-value >0.001 is retained using the following criteria on alignment quality: ⩾25% similarity in amino acid sequence over >40% of the length of the smallest protein. These threshold values allow the annotator to quickly check whether the new predicted CDS is, for example, a vestigial, possibly truncated gene.

RESULTS

The MICheck web server

MICheck is available through a web interface, using an implementation of the HTML and PHP languages for its graphical interfaces. The first two steps described above take ∼5 min (for large bacterial genomes). The running time for the third step depends on the number of CDSs with the 'Unique_AMIGene' status (BLAST execution takes ∼1 min for a protein sequence 300 amino acids in length). MICheck provides its results on a web page and by sending an email to the user with an URL pointing to a web page hosting these results. Results are stored for a period of 3 weeks and can also be downloaded for further analysis (see below).

The MICheck home page. Three main sections allow for the selection of MICheck input parameters:

The first section allows the user to choose the input databank file to be analysed, in either GenBank or EMBL file format. The input can be native files from the EMBL or GenBank websites, or an EMBL file format generated by the Artemis software (14). In order to avoid incorrect results in terms of the gene model computation (choice: 'Build one new gene model'), or in terms of the number of unique AMIGene CDS annotations, we first check the number of annotated genes according to the length of the input sequence. The parsing step also takes into account heterogeneities in databank annotations, together with genes annotated as authentic frameshifts (or point mutations) in order to avoid false positives as a result of regions of the genome containing identified frameshifts. Very often, these regions are annotated using only the 'gene' feature (e.g. at the TIGR Center) or using the qualifier '/pseudo' in the 'CDS' feature (e.g. at the Sanger Center). Whereas annotations of the GenBank files are always described with both the 'gene' and 'CDS' features, in the EMBL file format, the 'gene' feature is rarely (original format) or never [Genome Reviews (GR) section] used. Consequently, frameshifted genes which have been annotated by the authors with the 'gene' feature only are missing in the annotations of the EMBL file. Using this latter file format as input to the MICheck website leads to the detection of additional new AMIGene CDS predictions corresponding to these frameshifted genes (an example of this situation is given at http://www.genoscope.cns.fr/agc/tools/micheck/html/warning.html).

The second section allows the user either to select the species for which one or several gene classes have been computed or to build a new gene model using the annotated genes from the input file. The detailed description of how these gene models are built can be found at http://www.genoscope.cns.fr/agc/tools/micheck/html/Method.html#1.

The third section allows the user to enter a prefix name for the set of AMIGene CDSs (each predicted CDS will be assigned a unique numeric identifier prefixed with 'MYSEQ' if the default value is kept). An explanation of each AMIGene parameter and the heuristic we have implemented in the core of the method can be found at http://www.genoscope.cns.fr/agc/tools/amigene/html/Method.html#2.

The MICheck output page. A typical output of a session includes the following.

(i) A summary of the number of annotated CDSs (computed from the input file), the number of AMIGene CDS predictions and the number of each type of CDS ('Common', 'Unique' and 'No status').
(ii) The lists of CDSs unique to AMIGene annotations ('Unique_AMIGene' section) and to the user's input

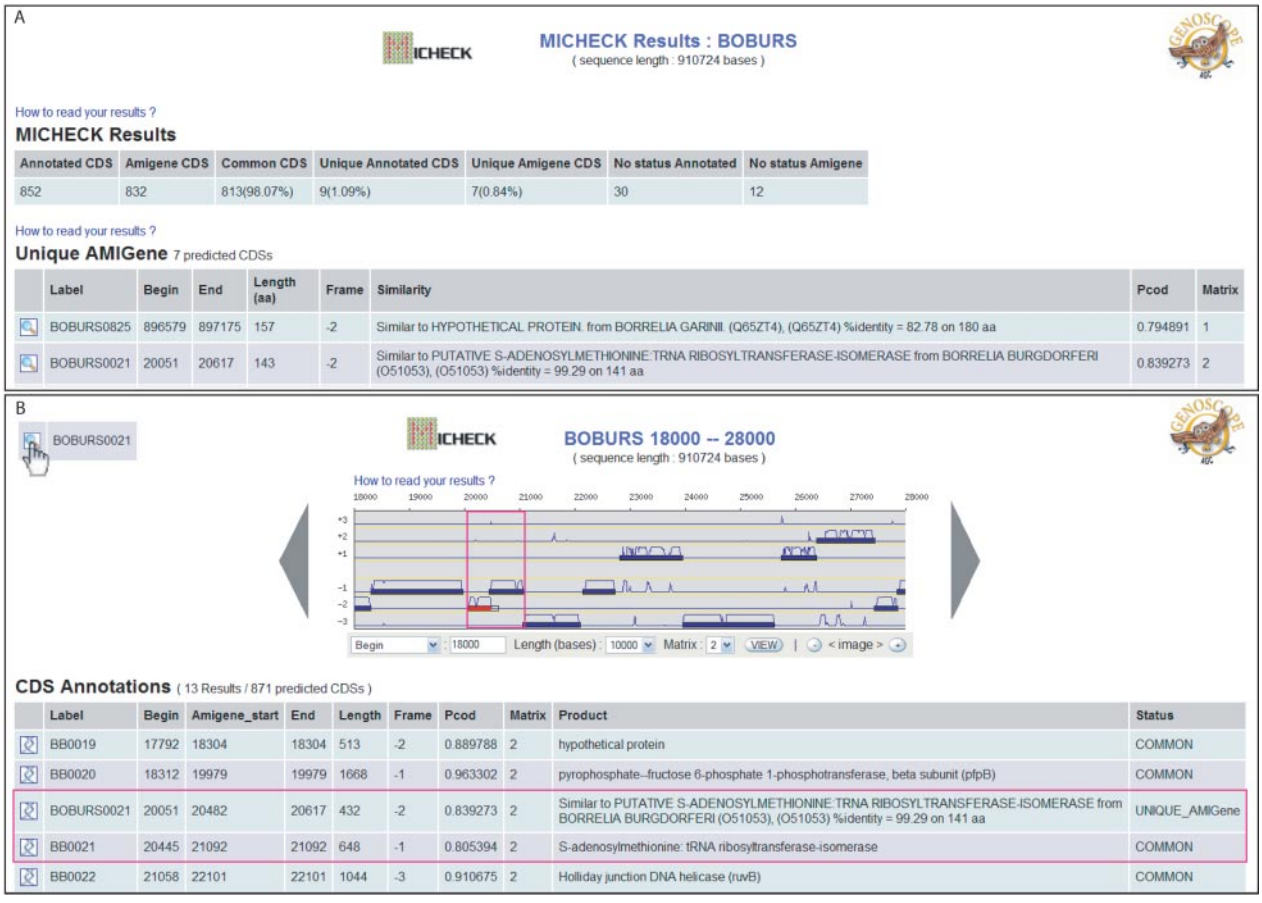

Figure 2. Sample display from an output MICheck page obtained for *Borrelia burgdorferi* genome analysis. (**A**) Partial lists of annotations unique to the AMIGene method (8) ('Unique_AMIGene') are given. Information on these unique CDSs such as their length, their coding probability ('Pcod') and the description of the best BLAST hit result ('Similarity') allows the user to quickly investigate probable missed gene annotations. (**B**) Selection of the magnifying glass icon near the 'Label' column leads to graphical MICheck output that allows the visualization of the gene context of a unique annotation (here BOBURS0021). The corresponding chromosomal segment of the *B.burgdorferi* genome, extending between positions 18 000 and 28 000 bp, is represented on this graphical map. Annotated CDSs are drawn in the six reading frames of the sequence by (i) a dark blue rectangle for 'Common' annotation and (ii) a red rectangle for 'Unique_AMIGene' annotation. For each CDS, the leftmost start position is drawn in transparency (i.e. in case of the BOBURS0021 CDS, this start position differs from the AMIGene start). Coding prediction curves, computed with the selected gene model ('Matrix' options), are superimposed on the annotated CDSs (blue curves). Information on these annotations is given in the array below, and the corresponding nucleic and proteic sequences can be retrieved independently (double helix icon near the label of a CDS). In the case of a unique AMIGene CDS, this functionality makes it possible to run a similarity search program easily, using a preferred web server.

annotation file ('Unique_User' section). If at least one significant database match has been found, the user can quickly judge (in addition to the coding probability value), whether a new AMIGene annotation has a real biological meaning. The user's final decision is guided by the examination of the corresponding gene context, in which the cartographic map shows the protein coding likeliness in terms of annotated gene positions and coding prediction curves computed, in the six reading frames, with the selected gene model ('Matrix' option, Figure 2B). This map is fully dynamic and allows the user to navigate along the genome while the corresponding list of annotated genes is updated accordingly.

(iii) The list of putative frameshifts, ordered by their level of confidence (Strong, Medium or Weak). The gene context of a single putative frameshift can be visualized in the same way as a unique CDS annotation.

(iv) Finally, this page includes several files that can also be downloaded: one file containing the MICheck results, two files containing only unique AMIGene nucleic and protein sequences, respectively, and one file containing the positions of the putative frameshifts (ProFED results).

The output results shown in Figure 2 have been obtained running MICheck on the *Borrelia burgdorferi* GenBank file, RefSeq (Reference Sequence) section (Accession no. NC_001318; annotation update: August 1, 2003). A total of seven 'Unique_AMIGene' CDSs and nine 'Unique_User' CDSs have been found (Figure 2A). One unique AMIGene CDS, BOBURS0021, is highly similar to the C-terminal part of a protein named QUEA_BORBU in the SWISSPROT databank (Accession no. O51053), annotated as a 'putative *S*-adenosylmethionine:tRNA ribosyl transferase-isomerase' (Figure 2B). The BB0021 gene, located next to BOBURS0021 and transcribed in the same strand, has been annotated with an identical biological description ('Product' column, Figure 2B). An explanation of this observation is found in the corresponding SWISSPROT entry, in which the reference number 3 refers

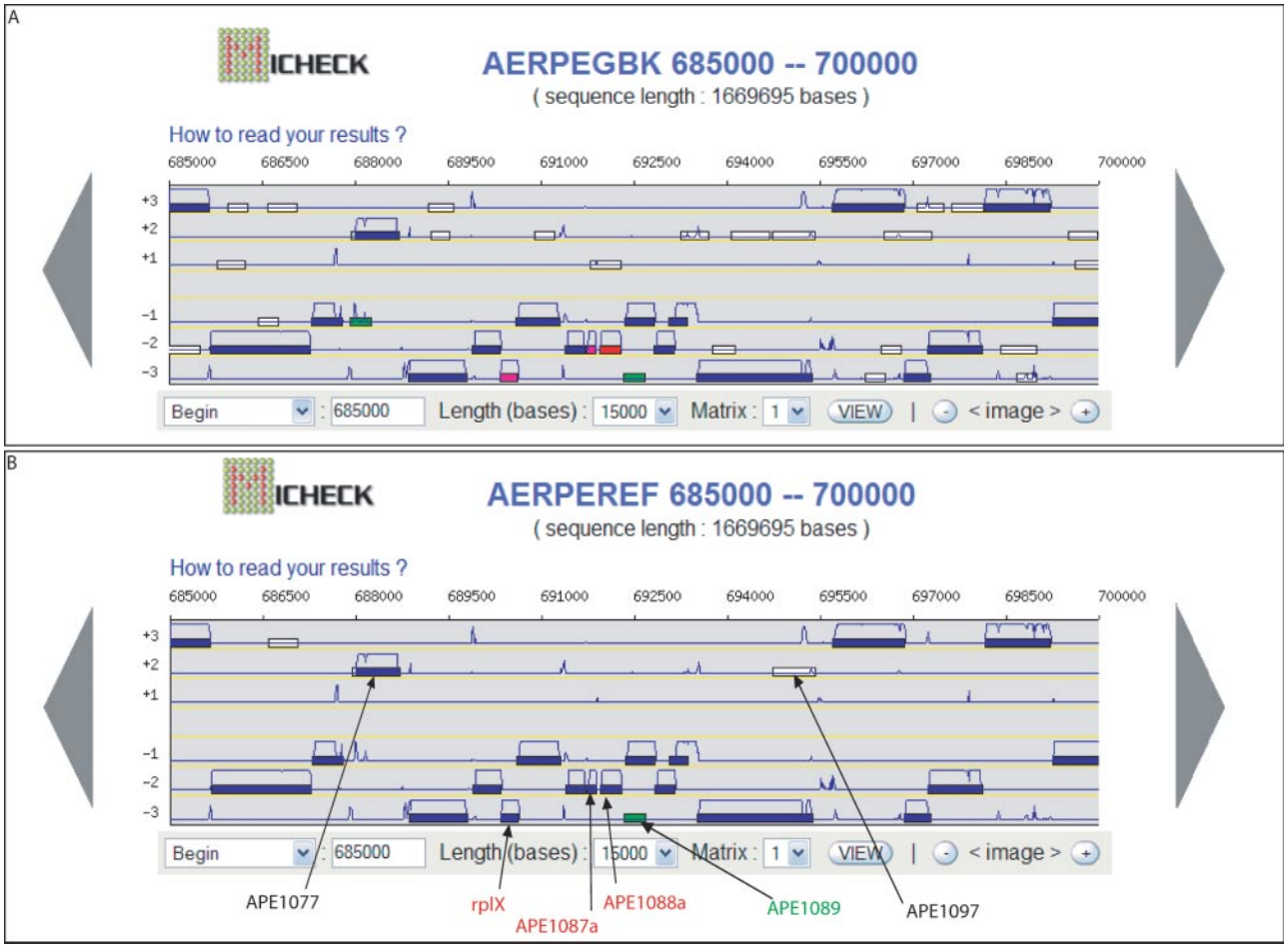

Figure 3. Comparison of MICheck annotations based on original and RefSeq *Aeropyrum pernix* annotations. The chromosomal segment of the *A.pernix* genome, extending between positions 685 000 and 700 000 bp, is represented on these two graphical maps. Annotated CDSs are drawn in the six reading frames of the sequence by (i) a dark blue rectangle for 'Common' annotation, (ii) a red rectangle for 'Unique_AMIGene' annotation, (iii) a pink rectangle for 'No_Status' AMIGene annotation, (iv) a white rectangle for 'Unique_User' annotation and (v) a green rectangle for 'No_Status' databank annotation. Annotated CDSs are drawn using the start codon position given in the input databank file; this sometimes leads to a shorter rectangle compared with the length of the corresponding prediction curve (i.e. the CDS in the middle of these maps). As shown in the 'AMIGene_Start' column (see complete *A.pernix* results available at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html), the original position is erroneous in this case. (**A**) MICheck results obtained using the original databank file as input (GenBank file format) and (**B**) MICheck results obtained using the corresponding RefSeq record as input.

to 'Identification of probable frameshift, by Zangger N.; Unpublished observations (May-2000)'. The BB0021 translation product differs from that of the SWISSPROT protein owing to a frameshift in position 205 and the unique AMIGene CDS then corresponds to the missing part of the QUEA protein.

Figure 3 illustrates re-annotation of 15 kb of the *Aeropyrum pernix* genome in which MICheck annotations have been obtained using two GenBank files as input: (i) the original one (BA000002; Figure 3A) and (ii) the Reference Sequence at the National Center for Biotechnology Information (NCBI) (NC_000854; Figure 3B). In this region, discrepancies between the original and AMIGene annotations (Figure 3A) are significant, as shown by the numerous genes that are unique to the original annotators (22 'Unique' in white rectangles and 2 'No status' in green rectangles) and the CDSs that are unique to the AMIGene predictions (1 'Unique' in the red rectangle and 2 'No status' in the pink rectangles). In the *A.pernix* RefSeq file most of these discrepancies have been removed (Figure 3B): three unique original annotations remain (all included in other genes) and three CDSs, which have been added during the reviewing process, are now 'Common' to AMIGene predicted CDSs (*rplX*, APE1087a and APE1088a).

MICheck results compared with reviewed bacterial genomes

Several years ago, the NCBI initiated the development of the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/). Derived from the primary submissions available in GenBank, RefSeq is an ongoing effort to provide a curated, non-redundant collection of sequences that includes smaller genomes such as viral, organelle and some microbial genomes (15). The RefSeq collection contains records in which experts at the NCBI have corrected or added annotations. More recently, the European Bioinformatics Institute (EBI) set up the GR project (http://www.ebi.ac.uk/GenomeReviews/), which provides curated versions of INSD (EMBL/GenBank/DDBJ) database entries, with added functional information imported from data sources such as the UniProt

knowledgebase (12), the GO annotation (16) and InterPro (17), together with many other cross-references. Similar enhancements are found in the RefSeq and GR projects in terms of sequence validation, standardized gene and product names and systematic locus tag identifiers. To date, however, the CDS re-annotation process of a complete bacterial genome can be found only in the RefSeq collection (for a limited number of organisms).

The first round of MICheck re-annotation was performed using a selection of the 19 complete bacterial genomes listed as curated in the RefSeq section of GenBank (ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/RefSeq-release9.txt). Nine of these contain the 'Reviewed RefSeq' status in the COMMENT line of the last update record (Table 1). In this list we also selected three genomes with 'Provisional RefSeq' status. Specific gene models were defined for these selected bacteria (see above), and for each genome MICheck was run three times using as input (i) the original annotations stored in the GenBank file (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/; the choice of the original GenBank file instead of the EMBL file is explained above), (ii) the set of reviewed annotation data stored in the RefSeq section (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) and (iii) the set of reviewed annotation data stored in the GR section (ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews/). Results, in terms of common and unique CDSs annotations, are given in Table 2, and complete output results are available at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html. As shown in this table, GenBank and GR input files share a similar number of annotated genes, except in the case of *Lactococcus lactis*, *Salmonella typhimurium* and *Shewanella oneidensis* (numerous original annotations have been removed in these GR records; Table 1). The number of AMIGene predicted CDSs is generally close to the number of annotated genes stored in RefSeq files (most often because of a common process of syntactic re-annotation).

MICheck results using original GenBank files. Three genomes contain no (or few) unique AMIGene annotations (*Buchnera* sp., *Haemophilus influenzae* and *Oceanobacillus iheyensis*), with generally few unique original annotations as well (Table 2). In contrast, in four other genomes (*S.oneidensis*, *A.pernix*, *Thermoplasma volcanium* and *Corynebacterium glutamicum*), the number of CDS predictions that are unique to AMIGene is high, with some interesting cases in terms of similarity results. For example, in *C.glutamicum*, the

Table 1. Sources of the reference dataset used to test MICheck

Genome	Size (Mb)	GC (%)	First release date	Accession no.			Number of annotated genes		
				GenBank	RefSeq	Genome Review	GenBank	RefSeq	Genome Review
Aeropyrum pernix	1.67	56.3	1999	BA000002	NC_000854	BA000002_GR	2695	1843	2694
Buchnera sp.	0.64	26.3	2000	BA000003	NC_002528	BA000003_GR	572	572	564
Corynebacterium glutamicum	3.31	53.8	2002	BA000036	NC_003450	BA000036_GR	3099	2993	3099
Haemophilus influenzae	1.83	38.2	1995	L42023	NC_000907	L42023_GR	1739	1716	1709
Lactococcus lactis	2.37	35.3	2001	AE005176	NC_002662	AE005176_GR	2308	2345	2266
Oceanobacillus iheyensis	3.63	37.7	2001	BA000028	NC_004193	BA000028_GR	3497	3502	3496
Pyrococcus abyssi	1.77	44.7	1996	AL096836	NC_000868	AL096836_GR	1785	1898	1785
Pyrococcus furiosus	1.91	40.7	1999	AE009950	NC_003413	AE009950_GR	2070	2130	2069
Pyrococcus horikoshi	1.74	41.9	1998	BA000001	NC_000961	BA000001_GR	2072	1959	2064
Salmonella typhimurium LT2	4.86	52.2	2001	AE006468	NC_003197	AE006468_GR	4536	4504	4453
Shewanella oneidensis	4.97	46	2002	AE014299	NC_004347	AE014299_GR	4757	4438	4630
Thermoplasma volcanium	1.58	39.9	1999	BA000011	NC_002689	BA000011_GR	1526	1506	1526

Genomes that still contain 'Provisional RefSeq' in the COMMENT line of the RefSeq record are indicated in bold.

Table 2. MICheck software results on the reference dataset

Genome	AMIGene CDSs	Common CDSs			Unique AMIgene CDSs			Unique Databank CDSs		
		GenBank	RefSeq	Genome Review	GenBank	RefSeq	Genome Review	GenBank	RefSeq	Genome Review
Aeropyrum pernix	1717	1565	1569	1565	18	35	18	941	186	941
Buchnera sp.	580	557	556	561	0	0	10	0	0	0
Corynebacterium glutamicum	2993	2906	2905	2907	15	5	15	65	14	65
Haemophilus influenzae	1770	1675	1627	1681	2	4	47	4	0	4
Lactococcus lactis	2365	2221	2260	2237	0	14	45	10	11	7
Oceanobacillus iheyensis	3458	3406	3392	3408	2	14	2	18	18	18
Pyrococcus abyssi	1892	1770	1862	1770	6	2	6	4	11	3
Pyrococcus furiosus	2090	2011	2053	2011	6	2	6	5	9	5
Pyrococcus horikoshi	1866	1677	1833	1681	7	0	7	339	91	339
Salmonella typhimurium LT2	4459	4267	4275	4314	12	7	35	9	8	1
Shewanella oneidensis	4441	4114	4127	4144	20	7	150	176	15	175
Thermoplasma volcanium	1571	1462	1462	1462	18	7	18	27	1	1

Common CDSs: the number of CDSs shared by the set of databank annotations and the set of AMIGene predictions; Unique AMIGene CDSs: the number of CDSs predicted by AMIGene strategy only; Unique Databank CDSs: the number of genes present in the annotation file only. The values are given for the original databank file (GenBank) and the reviewed annotations stored in RefSeq and Genome Review. Genomes that still contain 'Provisional RefSeq' in the COMMENT line of the RefSeq record are indicated in bold.

N-terminal part of the 2-methylcitrate dehydratase 2 (*prpD2*) gene is encoded by the AMIGene CORGLUGBK0633 CDS, while the C-terminal part is encoded by the annotated gene Cgl0657, located next to this new CDS (these two annotations are similar to the same UniProt entry, PRPD2_CORGL, a gene involved in propionate catabolism). As already noticed (10,18), the number of unique original annotations corresponding to false predictions is very high in *A.pernix*. Most surprisingly, this number remains high in the cases of *Pyrococcus horikoshii* (compared to the low number of unique AMIGene predictions) and *S.oneidensis*. These probably inaccurate original annotations are often located in front of other annotated genes and, whereas they always have a length shorter than 300 bp in *S.oneidensis*, their length can be >2000 bp in the case of *P.horikoshii* and *C.glutamicum* (Table 2 and see individual results listed at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html).

MICheck results using RefSeq reviewed files. In many cases, the NCBI review process leads to the annotation of the same missing genes as those detected by MICheck (Table 2). Moreover, the previously mentioned unique AMIGene CDS (CORGLUGBK0633) corresponds to the NCgl0627 annotation of the *C.glutamicum* RefSeq record, which has been annotated, together with NCgl0628, as 'Hypothetical protein; involved in propionate catabolism; *possible frameshift*'. This is obviously an accurate annotation for this part of the *C.glutamicum* genome, compared with the one found in the original set of annotations. Indeed, the large number of unique AMIGene annotations obtained with the *A.pernix*, *L.lactis* and *O.iheyensis* RefSeq entries is unexpected (Table 2). These cases have therefore been carefully analysed and, in addition to the unique AMIGene annotations common to the original GenBank record, we noticed that several (sometimes many) other original annotated genes have been removed. In the case of *A.pernix* these genes are annotated as 'Hypothetical protein', and half of them have a length >600 bp and a very high coding probability (0.74–0.94; see *A.pernix* results at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html). For *L.lactis* and *O.iheyensis* we noticed several (unintentional) changes in the RefSeq files which led to 'false' unique AMIGene predictions.

(i) In the case of *L.lactis* several pseudogenes are annotated with one location only; for example, the *glgB* pseudogene is described as 'gene complement (145931..147876)' in the original GenBank file and as 'gene 147277' in the RefSeq file (this is the end position of the LACLAREF0166 unique AMIGene CDS).
(ii) In the case of the *O.iheyensis* RefSeq file, insertion sequences (ISs), partial genes and pseudogenes have been annotated only with 'misc_feature', which is obviously ignored in the MICheck parsing step when it is used alone.

Finally, the number of unique RefSeq annotations is significantly lower than the number obtained with the original GenBank file (Table 2), except for *Pyrococcus abyssi*, for which eight additional RefSeq gene annotations have been added (e.g. PAB0133.1n and PAB0133.2n in a region initially annotated with two rRNA genes, 16S and 23S, and a tRNA-ala gene).

MICheck results using GR files. In most cases, the number of unique annotated GR genes is equal (or similar) to the number of unique annotations found with the original GenBank file (Table 2). When this number is lower (mainly for *S.typhimurium* and *T.volcanium*), the corresponding removed genes are always present in the list of unique AMIGene CDSs found in the original submission (see individual results given at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html). The number of unique AMIGene CDSs found with the GR record is either equal to (seven cases, Table 2) or greater than (five cases, Table 2) the number of unique CDSs detected by MICheck using the original GenBank file as input. Apart from the fact that the EBI re-annotation process does not include, to date, an automatic syntactic re-annotation of complete bacterial genomes, the main cause of mis-annotation is the initial parsing step of the original submission file: genomic features annotated by the authors with only the '/gene' or '/misc_feature' features are ignored (as previously mentioned, this is often the case when pseudogenes or partial genes are reported). This leads to extreme cases such as *L.lactis* (0 unique AMIGene CDS in the original set of annotations, compared with 45 in the GR file), *Buchnera* sp. (0 versus 10), *H.influenzae* (2 versus 47) and *S.oneidensis* (20 versus 150). All these additional unique predictions correspond to initial annotations of pseudogenes or partial genes which are missing in the corresponding GR files.

MICheck results on recently published genomes

The second round of MICheck tests have been performed using the RefSeq files containing the original annotations of eight complete genomes available since January 2005: *Xanthomonas oryzae* (February 5, 2005), *Wolbachia* sp. (February 5, 2005), *Ehrlichia ruminantium* (February 5, 2005), *Lactobacillus acidophilus* (February 2, 2005), *Thermococcus kodakaraensis* (February 2, 2005), *Staphylococcus epidermidis* RP62A (January 27, 2005), *Gluconobacter oxydans* (January 27, 2005) and *Dehalococcoides ethenogenes* (January 13, 2005). For these genomes an automatic gene model construction has been computed, using as input the set of annotated genes stored in the input file (option: 'Build one new gene model taking into account the annotations'). A summary of the corresponding re-annotation, in terms of common and unique CDS annotations, is given in Table 3 (the complete output results are available at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html). The number of unique AMIGene CDSs is very different among these genomes, with none or few unique predictions in the case of *E.ruminantium* (1.5 Mb, 27.5 GC%), *T.kodakaraensis* (2.1 Mb, 52 GC%) and *S.epidermidis* RP62A (2.6 Mb, 32.1 GC%). For this latter genome, another strain has already been annotated (ATCC12228, RefSeq file: NC_004461), and the two unique AMIGene CDSs correspond to hypothetical proteins annotated in the ATCC12228 strain, SE2189 (1182 bp in length), and SE2398 (642 bp in length). Two original RefSeq files (NC_006834 and NC_006677) contain, at least in this first release of the corresponding genomes, a large number of mis-annotations in terms of both unique AMIGene predictions and unique annotated genes. Maybe not surprisingly, the GC content of the corresponding genomes is high (61.1% for *G.oxydans* and 63.7% for *X.oryzae*) compared with the other

Table 3. Recently released genomes analysed with MICheck

Organism	Accession no.	GC (%)	AMIGene CDSs	Annotated CDSs	Common CDSs	Unique AMIGene CDSs	Unique annotated CDSs
Xanthomonas oryzae pv. *oryzae KAXCC10331*	NC_006834	63.7	4711	4637	4323	123	76
Wolbachia sp. *TRS (Brugia malayi)*	NC_006833	34.2	1071	902	785	6	0
Ehrlichia ruminantium Welgevonden	NC_006832	27.5	983	958	929	0	0
Lactobacillus acidophilus NCFM (ATCC 700396)	NC_006814	34.7	1856	1864	1725	15	6
Thermococcus kodakaraensis KOD1	NC_006624	52	2375	2307	2289	1	1
Staphylococcus epidermidis RP62A	NC_002976	32.1	2379	2553	2280	2	32
Gluconobacter oxydans 621H	NC_006677	61.1	2487	2432	2344	68	36
Dehalococcoides ethenogenes 195	NC_002936	48.8	1584	1592	1483	6	51

For details about each CDS category, see Table 2.

selected bacteria. In the case of *G.oxydans*, among the 68 CDSs that are unique to MICheck, 25 code for proteins similar to proteins with a known function (e.g. GLUOX0072 is similar to a DNA repair protein *recO*, and GLUOX1934 to a cytidine deaminase). Almost all unique original annotations (36 in total; Table 3) are included in the unique AMIGene predictions (generally located on the reverse strand). The results obtained with the *X.oryzae* original file are extreme in this set of MICheck runs given that 123 unique AMIGene CDSs, together with 76 unique annotated genes, have been found. Among these results, several interesting functions are clearly missing and many genes have been mis-annotated in the original submission (individual results for these two genomes are available at http://www.genoscope.cns.fr/agc/tools/micheck/html/samp_test.html).

CONCLUSION

We have described new web software for fast comparison of two sets of syntactic annotations of bacterial genomes. This comparison can include inaccurate or missed gene annotations that can be explored through a user-friendly web interface. Analysis of several MICheck runs on publicly available bacterial genomes illustrates the capabilities of our software. The extraction of the annotated data from the input file (i.e. the parsing step) is obviously a crucial step for the quality of the MICheck annotations. It has been shown here that, unless unexpected feature annotations are used, data from public databank files are correctly extracted. This is particularly clear in the GR file format, in which considerable effort has been devoted to standardization and homogenization of the annotations.

Although the 'Unique' status is used only in obvious cases (to highlight probable missing or mis-annotated genes), our automatic procedure will not replace validation by an expert of the results. This prompted us to develop a graphical interface to visualize the MICheck output, in which the gene context together with the coding prediction curves can help the user in making a final decision. The problem of over-annotation, and also under-annotation, cannot be ignored if one wants to compare bacterial proteomes. Apart from the set of missing or mis-annotated genes in some original submissions, common annotation rules are still missing as far as pseudogenes and partial genes are concerned. It is reasonable that the translation products of such genes are not stored in protein databanks, but the complete loss of this information in the nucleic databank files is unfortunate. Many bioinformatics groups in the world are working on gene synteny detection and comparison between bacterial genomes; they obviously first need to extract annotation data from public databanks and are faced with the problem of the heterogeneous annotations of frameshifted or partial genes (or even their total absence in several EMBL files). Even if a gene is seldom or never expressed in the cell of one particular species, it would be interesting to know whether it is involved in a synteny group conserved between several different bacteria or not.

We are not aware of any publicly available software systems that have the functionality of MICheck, and we expect that its use could contribute to a better quality of bacterial syntactic genome annotations deposited in public databanks. Indeed, in the context of the numerous microbial genome re-annotation projects, MICheck can be seen as a preliminary step before the functional re-annotation step, to quickly check for missing or wrongly annotated genes.

ACKNOWLEDGEMENTS

This work was supported by the French Centre National de la Recherche Scientifique (CNRS-UMR8030), the GENOPOLE of Evry and the French Ministry of Research (funds allocated by the ACI IMPBio). We thank Antoine Danchin, Susan Cure and Denis Bayada for their help in writing the manuscript. We thank the entire system network team of Genoscope for its essential contribution to the efficiency of the MICheck website. Funding to pay the Open Access publication charges for this article was provided by CNRG-composante Genoscope.

Conflict of interest statement. None declared.

REFERENCES

1. Ouzounis,C.A. and Karp,P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol*., **3**, 2001.1–2001.6.
2. Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet*., **17**, 429–431.
3. Iliopoulos,I., Tsoka,S., Andrade,M.A., Enright,A.J., Carroll,M., Poullet,P., Promponas,V., Liakopoulos,T., Palaios,G., Pasquier,C. *et al*. (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, **19**, 717–726.
4. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*., **26**, 1107–1115.
5. Larsen,T.S. and Krogh,A. (2003) EasyGene-a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.
6. Azard,R.K. and Borodovsky,M. (2004) Effects of choice of DNA sequence model structure on gene identification accuracy. *Bioinformatics*, **20**, 993–1005.

7. Hayes,W.S. and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res*., **8**, 1154–1171.
8. Bocs,S., Cruveiller,S., Vallenet,D., Nuel,G. and Médigue,C. (2003) AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res*., **31**, 3723–3726.
9. Borodovsky,M. and McIninch,J.D. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comp. Chem*., **17**, 123–133.
10. Bocs,S., Danchin,A. and Médigue,C. (2002) Re-annotation of genomes microbial CoDing Sequences: finding new genes and inaccurately annotated genes. *BMC BioInformatics*, **3**, 5.
11. Médigue,C., Rose,M., Viari,A. and Danchin,A. (1999) Detecting and analysing sequencing errors: toward a high quality of the *Bacillus subtilis* genome sequence. *Genome Res*., **9**, 1116–1127.
12. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
13. Glemet,E. and Codani,J.J. (1997) LASSAP, a LArge Scale Sequence compArison Package. *Comput. Appl. Biosci*., **13**, 137–143.
14. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualisation and annotation. *Bioinformatics*, **16**, 944–945.
15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
16. Camon,E., Barrell,D., Lee,V., Dimmer,E. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database—an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol*., **4**, 5–6.
17. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
18. Ussery,D.W. and Hallin,P.F. (2004) Genome Update: annotation quality in sequenced microbial genomes. *Microbiology*, **150**, 2015–2017.