BMC Bioinformatics

Research article

Re-annotation of genome microbial CoDing-Sequences: finding new genes and inaccurately annotated genes

Stéphanie Bocs*[1], Antoine Danchin[2,3] and Claudine Médigue[3]

Address: [1]Laboratoire Génome et Informatique, Université de Versailles, 91034 Evry Cedex, France, [2]HKU-Pasteur Research Center, Pokfulam, Hong-Kong and [3]Génétique des Génomes Bactériens, Institut Pasteur, 75724 Paris Cedex 15, France

E-mail: Stéphanie Bocs* - sbocs@infobiogen.fr; Antoine Danchin - adanchin@hkucc.hku.hk; Claudine Médigue - Claudine.Medigue@infobiogen.fr

*Corresponding author

Published: 5 February 2002

BMC Bioinformatics 2002, **3**:5

Received: 18 September 2001
Accepted: 5 February 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/5

Abstract

Background: Analysis of any newly sequenced bacterial genome starts with the identification of protein-coding genes. Despite the accumulation of multiple complete genome sequences, which provide useful comparisons with close relatives among other organisms during the annotation process, accurate gene prediction remains quite difficult. A major reason for this situation is that genes are tightly packed in prokaryotes, resulting in frequent overlap. Thus, detection of translation initiation sites and/or selection of the correct coding regions remain difficult unless appropriate biological knowledge (about the structure of a gene) is imbedded in the approach.

Results: We have developed a new program that automatically identifies biologically significant candidate genes in a bacterial genome. Twenty-six complete prokaryotic genomes were analyzed using this tool, and the accuracy of gene finding was assessed by comparison with existing annotations. This analysis revealed that, despite the enormous effort of genome program annotators, a small but not negligible number of genes annotated within the framework of sequencing projects are likely to be partially inaccurate or plainly wrong. Moreover, the analysis of several putative new genes shows that, as expected, many short genes have escaped annotation. In most cases, these new genes revealed frameshifts that could be either artifacts or genuine frameshifts. Some entirely unexpected new genes have also been identified. This allowed us to get a more complete picture of prokaryotic genomes. The results of this procedure are progressively integrated into the SWISS-PROT reference databank.

Conclusions: The results described in the present study show that our procedure is very satisfactory in terms of gene finding accuracy. Except in few cases, discrepancies between our results and annotations provided by individual authors can be accounted for by the nature of each annotation process or by specific characteristics of some genomes. This stresses that close cooperation between scientists, regular update and curation of the findings in databases are clearly required to reduce the level of errors in genome annotation (and hence in reducing the unfortunate spreading of errors through centralized data libraries).

Background

The main goal of large-scale genome sequencing projects is to obtain new insights into physiological and biological processes underlying the very organization of life. An essential step in this quest is gene identification, with subsequent functional annotation of the corresponding gene products. Gene recognition in bacteria is far from being always straightforward, despite the fact that bacterial genes are usually lacking introns. Extraction of all possible Open Reading Frames (ORFs) of a given length from a given DNA sequence is a trivial procedure; it is much less simple to decide which among those contain genes that are eventually expressed and code for proteins (CoDing Sequences, CDSs). The widely spread (and unfortunate) confusion between ORFs and CDSs is the sign of the lack of adequacy of many annotating systems in gene identification. Gene-finding methods are traditionally divided into two broad categories [1]. "Intrinsic" methods, which deal with DNA sequence only, use statistics or pattern recognition algorithms to find genes in DNA through detection of specific motifs or global statistical patterns. A typical example of such methods is the GeneMark software [2], a deservedly popular gene prediction program for prokaryotes, which uses periodical Markov models to find DNA regions that code for proteins. "Extrinsic" methods take into account information derived from similarity search procedures, using as queries either the genome sequence itself, or the putative proteins derived from the list of ORFs [3]. In the first case, the translation in all the six frames of the query DNA is required to compare the resulting amino acid sequences to known proteins (BLASTX program). Although this method has been shown to be relatively effective for gene finding [4], it is too time-consuming to be used as a common procedure. In addition, the prediction of such extrinsic methods entirely relies on the presence of closely related protein sequences in databanks, a dramatic limitation for gene discovery. Finally, it has been recently shown that a great many spurious short genes are generally annotated in genomes [5], and that the number of potential errors in the prediction of functional annotation is higher than is usually believed, mainly because it is based on relatively weak sequence identities and/or partial alignments [6].

Practical experience in genome analysis shows that it is necessary to incorporate as much available biologically derived evidence as possible in order to achieve reliable results [7]. An integration of several sequence analysis methods into a coherent and efficient prediction system is therefore required to obtain efficient computer-assisted annotation of DNA sequences. Several platforms integrating some of these goals have recently been developed [8–10]. Our own effort in this direction resulted in the creation of an integrated computer environment, Imagene, dedicated to genome sequence annotation and analysis [11]. In this system, both the biological data and the sequence analysis tools are uniformly represented in an object-based model, with a user interface, which allows one to display simultaneously the results produced by a variety of methods. This helps one to easily annotate interesting features of the sequence. In contrast to the approach followed by "genome crunchers" such as GeneQuiz [12] or Magpie [13], no automatic postprocessing of results has been defined in Imagene. The final synthesis and decisions are under the responsibility of the annotator and are helped by the graphic clues presented by the system. With the multiplication of genomic sequences of microorganisms, it became important to perform an efficient gene annotation using a first automatic procedure step before going to an in-depth manual annotation.

To tackle this problem, we have developed and embedded into our environment Imagene a new method (called AMIGA for Automatic MIcrobial Genome Annotation), that automatically finds out the most likely CoDing Sequences in a large contig or a complete bacterial genome (remember that a CDS is not an ORF). AMIGA relies on the combination of two gene-finding intrinsic methods and a heuristic approach allowing selection of the most likely CDSs (Bocs *et al.*, submitted; see Materials and Methods). In the present analysis, this program was run on 26 complete prokaryotic genomes. The accuracy of the method was assessed by comparing its predictions with the genome sequence annotations provided in the World-Wide DNA Data Library (GenBank/EMBL-EBI/DDBJ, WWDDL for short) [14]. We report here the results obtained by investigating discrepancies between our results and annotations present in the WWDDL. In this context we discuss both the diversity of annotation processes between authors, and the biological properties of some annotated genomes.

Results and Discussion

The AMIGA method was used to analyze 26 complete prokaryotic genomes (see Materials and Methods). The results were compared to original annotations available in the most recent release of the WWDDL. The number of CDSs in the Original Annotation (OA), in the AMIGA Prediction (AP), and Common to both OA and AP (CC) are given in Additional File 1. Subsequently, from the set of entirely missed annotated genes (*i.e.* Gene Not Found, GNF = OA-CC) and the set of newly predicted genes (*i.e.* potential New Genes, NG= AP-CC), the percentage of genes in each category is given according with reference to the value of their average coding probability (Pc). None of the AMIGA CDSs has a coding probability value below 0.2 because this value is the threshold of rejection in our method. Finally, the proportion of genes having a given status is also given in Additional File 1 (Wrong or Suspicious for the Genes Not Found having a Pc value below

0.2, and New or Ambiguous for the potential New Genes having a Pc value above 0.4). One will note that the proportion of CDSs having a Wrong or a New status (resp. New status) is generally small as compared to the total number annotated of Gene Not Found (resp. potential New Genes) (see Materials and Methods).

Accuracy of AMIGA as an automatic CDSs annotation process

A first general observation can be made on the number of CDSs identified both by the Original Annotation (OA) and by the AMIGA method Prediction (AP): the largest proportion of genes found by AMIGA demonstrates that the procedure is satisfactory in terms of gene finding accuracy. Except for *Aeropyrum pernix*, *Pyrococcus horikoshii*, and to a lesser degree, for the two strains of *Neisseria meningitidis*, more than 92% of the original annotations have been substantiated (Additional File 1). The number of annotated genes in the Gene Not Found category (*i.e.* presumed genes annotated by the authors and not retained by our method) is generally low especially in the case of the *Aquifex aeolicus*, *Haemophilus influenzae* and *Methanococcus jannaschii* genomes (about 15 genes, *i.e.* 1%, Additional File 1). Moreover, except in four cases corresponding to hypothetical genes identified either by the GeneMark method [2] or the CodonPreference method [15], no original annotated gene, having a coding probability above 0.4, has not been seen by our method. For the *Mycobacterium tuberculosis* G+C-rich genome, an important proportion of long ORFs with coding probabilities below 0.2 is observed, that probably corresponds to non-coding frames. The AMIGA method did not predict three *M. tuberculosis* annotated genes with a coding probability above 0.4 (*i.e.* 0.08%; Additional File 1). For these cases, the method did not pick out the correct translation initiation codon but chose the leftmost start. Thus, the corresponding (therefore too long) CDSs had a Pc below 0.4. In addition, they overlapped a sure CDS (Pc >= 0.4) and were consequently eliminated by AMIGA. This global analysis has revealed that the setting of AMIGA parameter values must be appropriately tuned in the case of genomes with a high G+C content.

The highest percentage of annotated GNF (*i.e.*, above 7%) is found in the *Escherichia coli*, *N. meningitidis*, *Treponoma pallidum*, *Vibrio cholerae*, *P. horikoshii* and *A. pernix* genomes. This is therefore independent of the genome size and the coding probability of most of the corresponding genes is very low (Pc < 0.2; Additional File 1). This is accounted for by features related either to the annotation process or to specific characteristics of the corresponding genomes such as portions of the sequence being horizontally transferred and/or the repetitive nature of the DNA (see below).

Evidence for a diversity of annotation processes between authors

The result of the annotation obtained with the *A. pernix* genome (42.24% of the GNF have a coding probability below 0.2, and most of them are in the Wrong category!), as well as the annotation result obtained with the *P. horikoshii* genome are somewhat surprising. Indeed, 19.86% of the annotated Genes Not Found have a Pc < 0.2 and are mainly assigned a Wrong status (14.67%; Additional File 1). The genome of both these archaebacteria has been sequenced and annotated by the same group, with the same rules. A general (trivial) criterion for assignment of potential coding regions in the genome sequence was chosen by the annotators: all the ORFs larger than 100 sense codons (*i.e.*, 300 bp in length) and starting either with ATG or GTG were retained as CDSs, whatever the results of similarity search in protein databanks, and no specific identifier of the distribution of bases in CDSs was used. Smaller ORFs (50–99 sense codons) were retained only if the sequences showed some similarity to the protein sequences (or motif sequences) in the databanks [16,17]. The simplistic nature of this procedure is also obvious when looking at the number of potential New Genes (NG) found by the AMIGA method (10.23%; Additional File 1). The large over-annotation of *A. pernix* has previously been noted [5]. This genome has also been investigated in detail using the COG system (Clusters of Orthologous Groups of proteins) [18]. Interestingly, this latter approach (which is based on phylogenetic classification of the proteins encoded in complete genomes) gave results similar to those presented here: about 32% of the originally annotated genes were not assigned to COG clusters. Not unexpectedly, this strongly suggests that these ORFs are not really genes (this could be compared to the 33.96% of original annotation having the Wrong status; Additional File 1) [19]. It is worth noting that, in term of functional annotation, looking at the genes that are members of COGs allows one to improve gene recognition in complete genomes (naturally however, new potential genes cannot be found in this way) [20].

Another example of an uncertain annotation quality is provided by the result obtained with the *Mycoplasma pneumoniae* genome, in its initial release [21]. The first analysis of the chromosome was obtained using the GCG programs package (Genetics Computer Group, Wisconsin). Recently, a new release of the genome annotation, using additional tools and methods and incorporating knowledge from the literature and new experimental data, has been published [22]. Ten new proteins were predicted in intergenic regions, sixteen protein reading frames were extended and eight shortened. In addition, the new annotation removed 23 previously annotated genes. In our approach, the number of Gene Not Found having a Wrong status is indeed very low (2 genes *i.e.* 0.29%; Addi-

tional File 1). In contrast, our proportion of predicted New Genes with a coding probability above 0.4 is quite important (95 genes, *i.e.* 11.80%; Additional File 1). Finally, it is worth noting that the genome of *Mycoplasma genitalium* also revealed an important proportion of possible New Genes (47, *i.e.* 8.55% of these genes have a coding probability above 0.4; Additional File 1). This result comes mainly from the fact that the authors intentionally did not annotate the occurrences of the three-gene operon that encodes one of the major surface proteins, the adhesin MgPa [23].

The case of the *A. aeolicus* genome illustrates the opposite situation: its annotation rests on so stringent parameters that it misses a number of genes. It is the genome for which we obtained the lowest proportion of annotated Genes Not Found by AMIGA together with an important number of potential New Genes (respectively 11 GNF, *i.e.* 0.72% and 181 NG, *i.e.* 10.57%; Additional File 1). These additional predicted genes have most often a very high coding probability (above 0.6), a length below 600 bp, and two thirds display either weak or no similarity with the non-redundant protein databank. The annotation process, briefly described by the authors [24], was based on the use of the Magpie software, which is dedicated to the complete automation of annotation [13]. It is likely that the parameter values required in the automatic assignment of CDSs from information obtained with multiple analysis tools, were chosen to be stringent in the context of the *A. aeolicus* genome annotation, in order to avoid spurious annotation.

The results obtained with the two strains of the *N. meningitidis* genome are surprising (240, *i.e.* 11.63% and 308, *i.e.* 14.47% of annotated GNF having a $Pc < 0.2$), since the two most important groups in the context of the sequencing projects (the Sanger Center and The Institute of Genome Research) have sequenced and analyzed these bacterial genomes. Generally, the predictions obtained with other genomes sequenced by these centers are indeed accurate (for example, *Campylobacter jejuni* for the Sanger Center, or *H. influenzae* for TIGR). In terms of annotation processes, the strategies used by the two groups are slightly different. Sequence analysis tools, such as Glimmer [25] for the prediction of CDSs, BlastP and FastA for the similarity searches in protein databanks or in the PFAM database of protein domains [26], are commonly used. A second annotation approach is sometimes used by the TIGR group, scanning directly the whole-genome sequence for homology searches with BlastX, BlastN and tBlastX [3], without introducing assumptions associated with defined CDSs. This analysis is particularly useful in the context of frameshift error detection. Finally, in these two sequencing centers, a graphical interface allowing the integration of the results of several analyses, and to visualize the potential CDSs, homologies, repeats, etc is used (TIGR software, unpublished; Artemis [10]). A manual annotation of the sequence and predicted proteins is subsequently performed, the results of which is a high final annotation quality. Therefore, noting that the majority of the genes missed by AMIGA in the two *N. meningitidis* strains has a coding probability below 0.2 and no status (Additional File 1), we have to conclude that the discrepancy must be accounted for by the specific properties of the bacteria themselves. In particular, *N. meningitidis* strains are naturally competent and freely take up DNA from the environment, and incorporate it into their genome. This means that large portions of the genome are horizontally transferred. In the context of this work, the CDSs having very different coding properties from the native genes are expected to be missed by the settings used in AMIGA (the minimum coding probability threshold is 0.2 and no training has been performed on horizontally transferred genes; see below, the case of *Synechocystis* sp.). Furthermore, the most striking characteristic of this genome is the presence of many repetitive elements ranging from short repeats (*i.e.*, the *Neisseria* DNA uptake sequence, which is involved in the recognition and uptake of DNA from the environment, was found in about two thousand copies in the two strains), to insertion sequences (which occasionally contain multiple frameshifts, large deletions and/or premature termination codons), and gene duplications of one kilobase or more. Many of these repeated regions seem to be involved in the genome evolution and antigenic variation in this human pathogen [27,28]. In DNA sequences harboring such repeats, the proportion of questionable coding regions detected by a method based on Markov model is generally high.

Impact of (authentic) frameshifts and horizontally transferred genes in the results

Potential New Genes that have been found by the AMIGA method are generally very short (ranging from 170 to 500 bp) and half of them code for peptide sequences showing similarities with sequences in the non-redundant protein databank ('known proteins' in the case of Bacteria and 'hypothetical proteins' in the case of Archaea). We also noticed that these small CDSs were very often located in regions where possible frameshifts have been detected by the ProFED method [29] (L. Labarre, in preparation). In addition, several of these CDSs have revealed new frameshifts, often correlated with the presence of pseudogenes. For example, in the case of the *H. influenzae* genome the 9 potential New Genes having a coding probability above 0.4 (Additional File 1) code for protein fragments (smaller than 300 bp), seven of which being identified by ProFED. Some of these new gene fragments are similar to genes of particular functions found in pathogenic organisms, such as genes involved in iron transport or in a type III secretion system.

In the case of the *T. pallidum* and *Borrelia burgdorferi* genomes, both sequenced and annotated by TIGR, about 6% of the previously annotated genes were not present in the AMIGA annotation. These genes mainly corresponded to very short CDSs (100 to 150 bp in length) having a coding probability below 0.2 (Additional File 1). A similar proportion of annotated GNF has been found in the *Synechocystis sp.* genome sequence, the majority of which where not assigned a specific status (in particular the number of Wrong CDSs is very low; Additional File 1). This indicates that the corresponding genes either showed similarity with proteins in the databank, or had a length above 900 bp (See Materials and Methods, Figure 2). This genome harbors many genes, which code for proteins containing typical Trp-Asp(WD)-repeats. These elements were originally reported as conserved repeats contained in the regulatory proteins of eukaryotes [30]. Moreover, phylogenetic relationships between cyanobacteria and plants showed that numerous proteins and tRNA genes with significant similarity to plant nuclear and plastid genes were predicted in the *Synechocystis* genome [31]. Compared to the native genes of *Synechocystis* sp., these genes may have a different codon usage bias resulting in a very low coding probability using AMIGA. Indeed, the GeneMark model used was trained on the set of all annotated genes in the WWDDL (see Materials and Methods), reflecting the codon usage bias of the majority of the genes, *i.e.* the native one. Similar observations are generally true for microbial genomes with an important proportion of horizontally transferred genes, such as *Synechocystis* PCC6803, *E. coli*, *A. aeolicus*, *Methanobacterium thermoautotrophicum*[32].

Heterogeneity in nucleotide composition and in codon usage patterns of *E. coli* genes has been analyzed and used to identify genes that could be of foreign origin [33–35]. More than 17% of the *E. coli* genes have probably been acquired from other genomes [36]. When not using the codon usage matrix associated to this particular class of genes in the AMIGA procedure [37], 7.69% of the *E. coli* annotated genes have been left out: these CDSs generally have a low coding probability and either none or a suspicious status (Additional File 1). An interesting result for this model genome is found when comparing the annotated Genes Not Found by AMIGA and having a Wrong status ([38], deposited in the WWDDL by the Blattner group in November 1998) to the curated lists of genes and proteins present both in the EcoGene and GenProtEC databases [39,40]. Among the 61 GNF having a Wrong status (1.42%; Additional File 1), 33 (54%) have disappeared from the curated databases, 22 (36%) correspond to genes whose function is either unknown or similar to that of hypothetical proteins in databanks, and 6 (10%) only are genes which have been previously identified as *bona fide* genes. These genes code for 50 amino acid long polypeptides, except for the *tnaL* gene, which codes for the tryptophanase leader peptide (25 residues). Half of them are related to insertion sequences (IS) or prophage regions. Thus, given the data of these curated databases, the number of *E. coli* Genes Not Found by AMIGA, and having the Wrong status, is in fact equal to 22 (0.6 %).

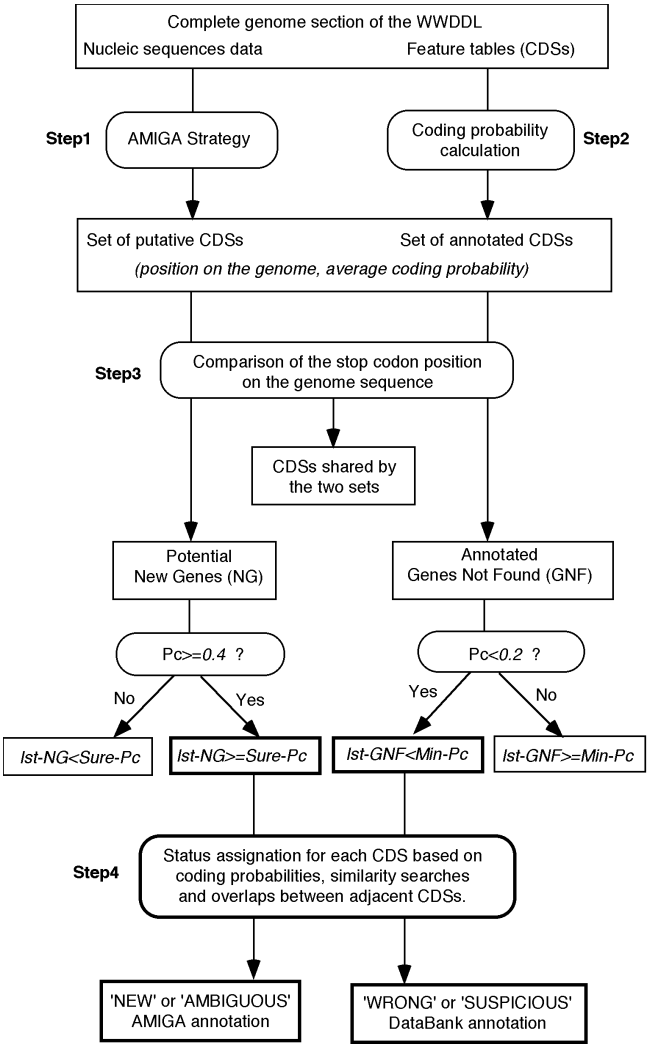

Figure 1
Overall strategy of the CDSs (re-)annotation of the bacterial genomes. The procedure involves four main steps (see text), the latter being performed on potential New Genes having a coding probability above 0.4 (list *lst-NG>=Sure-Pc*), and on annotated Genes Not Found having a coding probability below the 0.2 (list *lst-GNF<Min-Pc*). (WWDDL) World-Wide DNA Data Library (GenBank/EMBL-EBI/DDBJ); (Pc) coding probability.

Prediction of potential new genes: the cases of particular genomes

For the *E. coli* genome sequence, 0.73% of the CDSs found only by the AMIGA method have a 'New' status (Additional File 1). Using the EcoGene database [39] as a new refer-

ence (instead of the WWDDL file), we found that two thirds of these new annotations are already present in EcoGene, thus remarkably substantiating the validity of the AMIGA approach. They correspond either to previous corrections of frameshift errors in the *E. coli* DNA sequence, or to new annotations of very short CDSs. The remainder shows similarities with insertion sequences or prophage regions, and one CDS is highly similar to the C-terminus of the maltose-6'-phosphate glucosidase (MalH) protein annotated in the *Fusobacterium mortiferum* genome. This new CDS is located just behind the *glvG* gene in the *E. coli* genome, which codes for a probable 6-phospho-beta glucosidase. The corresponding SWISS-PROT databank entry (accession number P31450), contains an annotation indicating that the protein lacks the C-terminal half of the catalytic domain found in other members of this family. Finally, a new CDS is located just behind the *rpiB* gene on the *E. coli* genome, and in front of the *phnQ,* gene which has been annotated on the reverse strand. The phnQ gene (labeled 'very hypothetical protein' in the SWISS-PROT databank) has a 'Wrong' status (Additional File 1) and has been removed from the EcoGene database. It is therefore likely that it should be removed from SWISS-PROT as well. The new CDS we have found in the opposite strand is probably the correct one, but despite its high coding probability (0.63), no similarity with the protein databank has been detected.

The coordinator of the annotation consortium of the *M. tuberculosis* genome ([41], S. Cole, personal communication, 2001) has carefully analyzed results obtained in case of this bacterium. Among the potential New Genes having a New status (54 genes, *i.e.* 1.32%, Additional File 1), 61.5% have been integrated in an updated version of the annotations (to date, these corrections are only available in the Artemis system which is used to perform the *M. tuberculosis* annotation), 31.6% have been previously identified (but not integrated in the nucleic databank updates), and 7% were considered not enough convincing. One third of the previously identified new genes were in fact genes originally annotated using inappropriate databank label features (such as 'misc_feature' or 'mRNA', instead of 'CDS'). This set of newly identified genes has allowed us to confirm several original wrongly annotated CDSs (32% of the proportion of Genes Not Found having a Wrong status). More than fifty percent of our New Genes have revealed the existence of probable (authentic or not) frameshifts in the genome sequence. The products of these genes are similar to hypothetical proteins (43%), insertion sequences or prophage proteins (23%), and associated to other biological functions (34%) such as regulatory proteins, export proteins and lipoproteins. We also have found a new protein (118 aa in length), located at 1,678,550 bp between the Rv1488 (annotated as 'Hypothetical protein') and Rv1490 (annotated as 'Probable membrane protein') genes. This new gene is similar to an invasion protein of *Mycobacterium paratuberculosis* and clearly replaces the previous Rv1489c entry, which has been annotated at the same location and in the reverse strand.

Application of the method to strains 26696 and J99 of *Helicobacter pylori* produced respectively 0.7 and 0.4% of potential New Genes (Additional File 1). Several of these CDSs have been annotated in one strain but not in the other strain (2 new genes of the 26696 strain were found in the genome of the J99 strain, and 1 new gene of the J99 strain was annotated in the genome of the 26696 strain). Most of these new CDSs (65%) are located in regions where potential frameshifts have been detected by the ProFED method (L. Labarre, in preparation). One interesting new gene has been found in strain 26696. It is located at position 1,005,895 bp between two tRNA genes, coding respectively for tRNA-Gly and tRNA-Leu. This new CDS is also located in front of the gene named HP0945, which has been annotated in the reverse strand and which has a Wrong status. We found a significant similarity of its product with a TRL transfer RNA associated locus, previously described in the *H. pylori* genome. In fact, expression of this new gene has been demonstrated experimentally: this tRNA-associated locus is co-transcribed with tRNA (Gly) and reveals genetic diversity [42]. This is the reason why we did not find this CDS in the J99 strain. In this latter strain, we have found a new CDS located between the jhp0919 gene (annotated as a 'topoisomerase I, topA_2') and the jhp0920 gene (annotated as 'putative'). Its product shows similarity with a short part of the jhp0931 gene, which is found 15000 bp further in the genomic sequence. This gene has been annotated as a topoisomerase I (topA_3) and we noticed that the three adjacent genes (jhp0919+our new CDS+jhp0920) correspond to the jhp0931 gene. This gene has thus been duplicated in the genome of *H. pylori* and only one copy, jhp0931, is still functional.

Integration of the results of the present study into the SWISS-PROT databank

The present work is meant to provide efficient annotation to the High quality Automated Microbial Annotation of Proteomes (HAMAP) project. In the framework of the SWISS-PROT database [43], this project aims at automatically annotating a significant percentage of proteins originating from microbial genome sequencing projects. The annotation protocol differs from the many currently existing automatic annotation systems in that it does not try to attempt to hunt for distant similarities. The programs being developed are specifically designed to track down "eccentric" proteins, and a careful manual annotation is subsequently performed with these proteins. The results presented here are a contribution toward the achievement

Table 1: Examples of potential New Genes integrated as new SWISS-PROT entries

Entry	Accession	Description	Seq Length (aa)
RL21_AERPE	P58077	**50S RIBOSOMAL** PROTEIN L21 E	107
RL29_AERPE	P58085	**50S RIBOSOMAL** PROTEIN L29P	66
RL34_AERPE	P58026	**50S RIBOSOMAL** PROTEIN L34E	95
GYRB_ARCFU	029720	**DNA GYRASE SUBUNIT B** (EC 5.99.1.3)	632
EX7S_CHLTR	P58001	PROBABLE EXODEOXYRIBONUCLEASE VII SMALL SUBUNIT (EC 3.1.11.6) (EXONUCLEASE VII SMALL SUBUNIT)	72
YD5A_METJA	P58018	HYPOTHETICAL PROTEIN MJ135.1	364
SECG_MYCGE	P58061	PROBABLE **PROTEIN_EXPORT** MEMBRANE PROTEIN **SECG**	77
RL31_PYRHO	P58189	**50S RIBOSOMAL** PROTEIN L31 E	95
RS27_PYRHO	P58078	**30S RIBOSOMAL** PROTEIN S27E	65
SUI1_PYRHO	P58193	**PROTEIN TRANSLATION FACTOR** SUI1 HOMOLOG	99
Y56A_THEMA	P58008	HYPOTHETICAL PROTEIN TM0562.1	192
YB5A_THEMA	P58009	HYPOTHETICAL PROTEIN TM1158.1	240
YV6A_VIBCH	P58093	HYPOTHETICAL PROTEIN VCA0360.1	80

of a high level of genome annotation quality, and our identified microbial New Genes are regularly integrated into the SWISS-PROT databank. As shown in Table 1, efforts are currently focused on the completion of the catalog of essential biological functions such as those involved in the translation cellular process (*A. pernix* and *P. horikoshii*; Table 1), or in exportation systems (*M. genitalium*). New Genes showing similarities, on their full length, with other databank hypothetical proteins are also included as new SWISS-PROT entries (*M. jannaschii* and *Thermotoga maritima*; Table 1). In the case of *Archaeoglobus fulgidus*, we have found a short CDS that strongly suggested a likely frameshift error in the *gyrB* gene sequence. While the name of this gene remains identical to the one given in the nucleic databank (i.e., AF0530), the protein sequence is different since the SWISS-PROT entry has been corrected (632 aa instead of 507 aa in length).

Conclusions

We have developed a new method that automatically finds out the most likely CoDing Sequences in a large contig or a complete bacterial genome, using biological knowledge (periodical Markov chain analysis, sequence comparisons and identification of protein start sites). The method was used to analyze 26 complete prokaryotic genomes, and the accuracy of gene finding was assessed by comparison with existing annotations. Our predictions were most often in agreement with the published annotations. Discrepancies between our results and authors annotations were further investigated. We found that a sizeable amount of genes annotated within the framework of large-scale sequencing projects are likely to be partially inaccurate or plainly wrong (2%). In addition, while investigating carefully several bacterial genomes, some putative new genes have been discovered. Perhaps not unexpectedly, many short genes have been omitted in annotation, probably because of their short length. In most cases, these new genes revealed frameshifts in the DNA sequence frames that could be sequencing errors. However, because genome sequencing is now highly accurate, these frameshifts could be of a genuine type, then corresponding either to pseudogenes or to programmed translational frameshifts. Finally, in rare but very interesting cases, entirely new genes have been identified. An important conclusion of this work is that close cooperation between scientists, regular update and curation of the findings in databases are required to reduce the level of errors. We support calls for concerted efforts in re-annotation and in this context, our group actively participates to the HAMAP project (High quality Automated Microbial Annotation of Proteomes) which aims at automatically annotating a significant percentage of proteins originating from microbial genome sequencing projects [http://www.expasy.ch/prot/hamap] . The results of the present work will be available at the following Web site [http://chlora.infobiogen.fr:1234/wlag] and we encourage authors to contact us for further investigation in new potential genes which have thus been found.

Materials and Methods

Data

We have used in this study a total of 26 complete bacterial genomes available in the public databanks. The following genomes were analyzed (abbreviations in parentheses): *Aeropyrum pernix* (AERPE), *Aquifex aeolicus* (AQUAE), *Archaeoglobus fulgidus* (ARCFU), *Borrelia burgdorferi* (BOR-

BU), *Campylobacter jejuni* (CAMJE), *Chlamydia pneumoniae* (CHLPN), *Chlamydia trachomatis* (CHLTR) *Escherichia coli* (ECOLI), *Haemophilus influenzae* (HAEIN), *Helicobacter pylori* J99 (HELPJ), *Helicobacter pylori* 26695 (HELPY), *Methanococcus jannaschii* (METJA), *Methanobacterium thermoautotrophicum* (METTH), *Mycoplasma genitalium* (MYCGE), *Mycoplasma pneumoniae* (MYCPN), *Mycobacterium tuberculosis* (MYCTU), *Neisseria meningitidis* MC58 (NEIMB), *Neisseria meningitidis* Z2491 (NEIMA), *Pyrococcus abyssi* (PYRAB), *Pyrococcus horikoshii* (PYRHO), *Rickettsia prowazekii* (RICPR), *Synechocystis* sp. C125 (SYNY3), *Thermotoga maritima* (THEMA), *Treponema pallidum* (TREPA), *Ureaplasma parvum* (UREPA) and *Vibrio cholerae* (VIBCH). Sequences of all complete genomes with the accompanying information on the positions of protein-coding genes were retrieved from National Center for Biotechnology Information Entrez Genomes [http://www.ncbi.nlm.nih.gov] . Original annotation data were extracted from the feature tables of the WWDDL files (update of January 2001), and incorporated into our Prokaryotic Genome DataBase, PkGDB (L. Labarre, in preparation).

Automatic Microbial Genomes Annotation: the AMIGA strategy

The AMIGA method automatically finds out the most likely CoDing Sequences (CDSs) in a large contig or a complete bacterial genome. A precise description of this method together with a statistical evaluation of the chosen threshold values (which have currently being determined empirically, subsequent to the examination of results obtained with several AMIGA runs on various bacterial genomes) will be described elsewhere (Bocs *et al.*, submitted). The two main steps of AMIGA can be summarized as follows:

Combining two gene-finding intrinsic methods

Given the sequence of a complete genome, a CDS searching method is first executed and the positions, in the six reading frames, of the putative CDSs longer than 60 bp are kept. Then, the GeneMark method [2] is used to produce six numeric vectors corresponding to the coding probabilities along the DNA fragment for each of the six frames. Of course, preliminary calculation of the adapted GeneMark model (i.e. the matrix containing the transition probability values of the Markov model) is performed for each studied genome (MakeMat program; M. Borodovsky personal communication). The results of these two methods are merged together in order to compute for each putative CDS its coding probability, using the values of the corresponding GeneMark vector. We subsequently construct a first list containing all putative CDSs, each of them being characterized by their start and stop codon positions, their length (bp), their frame, and their average coding probability.

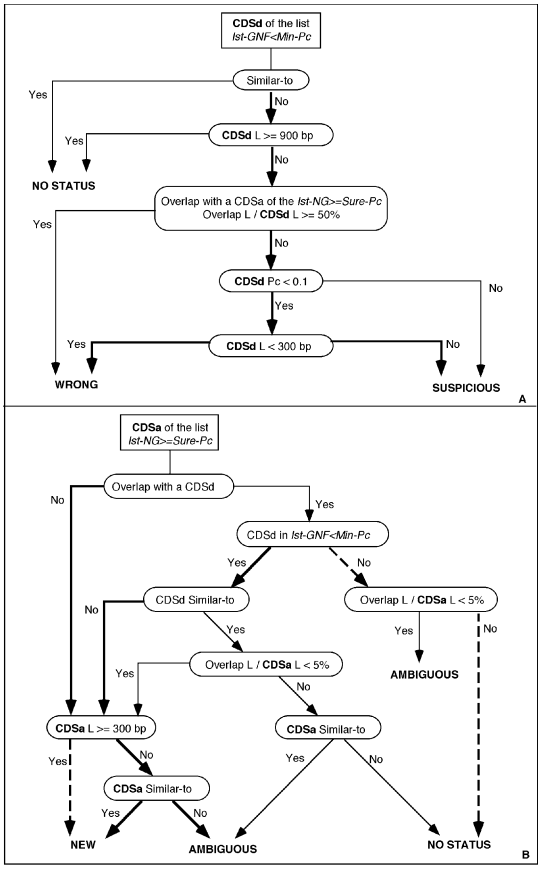

Figure 2

Assignation of a status to some additional CDSs. A. The annotated Genes Not Found by the AMIGA method (CDSd). B. The potential AMIGA New Genes (CDSa). The procedure takes into account the length of the CDS, its coding probability, results of similarity search in the non-redundant protein databank and overlaps between adjacent CDSs, these CDSs being an AMIGA CDS (CDSa) and a databank CDS (CDSd) (see text). Although all situations are investigated in the procedure, there are obviously preferred ways (thick arrows): for example a CDSa of the lst-NG>=Sure-Pc list is often found with no overlap with a CDSd. In this case, the CDSa often has a length below 300 bp and, either no similarity (AMBIGUOUS status) or similarity (NEW status) with proteins in the databank. If a CDSa does overlap a CDSd, the last one often has a weak coding probability and no similarity with proteins in the databank (in this case, the CDSa has the NEW status). Therefore it is extremely rare to found a CDSa of the lst-NG>=Sure-Pc in overlap with a CDSd having a strong coding probability, this overlap between the two CDSs being also important (broken arrows). In case of *A. pernix* and *P. horikoshii* the threshold for the CDSd length has been fixed to 600 bp instead of 300 bp. This choice is motivated by the nature of the annotation procedure of the authors of the genome sequences (see text). (L) length; (Pc) coding probability; (lst-NG>=Sure-Pc) list of CDSa having a coding probability above 0.4; (lst-GNF<Min-Pc) list of CDSd having a coding probability below 0.2.

Selecting the most likely CoDing Sequences (CDSs)

The selection of the most likely CDSs consists in the elimination of the false positives according to the coding potentials of the predicted CDSs and to overlapping criteria between adjacent CDSs, these overlaps being either total (they are called inclusion) or partial. Current experimental data do not show much evidence for existence of completely overlapping genes in prokaryotic genomes. Most of the included CDSs are then eliminated, except for inclusion being characteristic of the presence of a compensating frameshift in the sequence. In the same way, two overlapping CDSs, transcribed in the same strand, are kept in the final list of the selected CDSs [29].

Overall strategy of the CDS (re-)annotation of the bacterial genomes

For each genome of interest, the analysis consisted of the following four main steps (Figure 1). **Step1:** Starting with the AMIGA strategy, we extracted from the chromosome sequence (WWDDL files) a list of putative CDSs, which are characterized by their position in the nucleic sequence and their average coding probability. **Step2:** From the "feature" section we extracted the position of the genes as originally annotated, and a coding probability is computed for each gene (using the coding probabilities along the genome obtained in the first step). **Step3:** Subsequently, the two sets of CDSs (one from the AMIGA results and one from the authors' annotation) were compared for their stop codon position in the genome (there may be a possible misplacement of the gene start codon). Then, three main lists of CDSs are generated: (i) the list of the CDSs shared by the two compared sets of CDSs; (ii) the list of additional databank CDSs, *i.e.* annotated Genes Not Found by the AMIGA method (GNF); (iii) the list of additional AMIGA CDSs, *i.e.* putative New Genes (NG; Figure 1). **Step4:** A status being 'WRONG or SUSPICIOUS' in case of additional databank annotations, and 'NEW or AMBIGUOUS' in case of additional AMIGA predictions, was assigned according to the following procedure.

Status assignation to the two sets of additional CDSs

Two main types of CDSs annotation error can be found in databanks: an annotated CDS, which has no biological meaning, and a missed CDS annotation corresponding to a putative new gene. Therefore, from the set of annotated Genes Not Found by AMIGA, we extracted the CDSs having a coding probability below 0.2. This list of CDSs is called *lst-GNF<=Min-Pc* (Figure 2). Also, from the set of putative new genes, we selected the CDSs having a coding probability above 0.4. This list of CDSs is called *ist-NG>=Sure-Pc* (Figure 2). These two subsets were translated into protein sequences and compared to the SWISS-PROT, SPTrEMBL, and TrEMBLnew protein databanks [43], using an iterative blast2P similarity search program [3]. The statistical expect value for reporting hits was set at a threshold of 10^{-3}, the selected hits being obviously different from the original annotation. Each CDS was then characterized by its start and stop codon positions in the genome, its length (bp), its frame, its coding probability, and results of similarity search in the non-redundant protein databank if performed. In order to assign a status to these CDSs, we defined additional criteria as shown in Figure 2. A **CDSd** of the *lst-GNF<=Min-Pc* list (*i.e.*, a databank **CDS**) was assigned a 'WRONG' or a 'SUSPICIOUS' status depending on the value of: (i) its similarity with proteins in the databank; (ii) its coding probability; (iii) its length in case of very weak coding probability; (iv) its overlap with a CDSa of the *lst-NG>=Sure-Pc* list (*i.e.*, an AMIGA **CDS**) (Figure 2A). Similar criteria having different thresholds were used to assign a status to a **CDSa** of the *lst-NG>=Sure-Pc* list. In addition, for an overlapping CDSd with the **CDSa** being examined, we also took into account the value of the similarity with proteins in the databank (Figure 2B).

Additional material

Additional file

Comparison of the microbial genes annotated in GenBank files with the CDSs predicted by the AMIGA strategy. This file contains, for 26 available bacterial genomes, the number of CDSs annotated in GenBank files and the number of CDSs predicted by the AMIGA strategy. A status is assign to additional CDSs: Wrong or Suspicious for the Gene Not Found and New or Ambiguous for the potential New Genes.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-3-5-S1.xls]

Acknowledgements

This work was supported by the French Centre National de la Recherche Scientifique (CNRS), URA2171 and the Evry's GENOPOLE. We thank Anne-Lise Veuthey from the Swiss Institute of Bioinformatics (SIB) and Jean-Loup Risler from the Laboratoire Génome et Informatique for their critical comments and suggestions.

References

1. Fickett JW: **Finding genes by computer: the state of the art.** *Trends Genet* 1996, **12:**316-320
2. Borodovsky M, Mclninch JD: **GeneMark: Parallel gene recognition for both DNA strands.** *Comp* 1993, **17:**123-133
3. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402
4. Robison K, Gilbert W, Church GM: **Large scale bacterial gene discovery by similarity search.** *Nature Genetics* 1994, **7:**205-214
5. Skovgaard M, Jensen LJ, Brunak S, Ussery D, Krogh A: **On the total number of genes and their length distribution in complete microbial genomes.** *Trends Genet.* 2001, **17:**425-428
6. Devos D, Valencia A: **Intrinsic errors in genome annotation.** *Trends Genet.* 2001, **17:**429-431
7. Frishman D, Mironov A, Mewes HW, Gelfand M: **Combining diverse evidence for gene recognition in completely sequenced bacterial genomes.** *Nucleic Acids Res* 1998, **26:**2941-2947
8. Harris NL: **Genotator: A Workbench for Sequence Annotation.** *Genome Research* 1997, **7:**754-762

9. Bailey LC, Fischer S, Schug J, Crabtree J, Gibson M, Overton GC: **GAIA: Framework Annotation of Genomic Sequence. .** *Genome Research* 1998, **8:**234-250
10. Rutherford KM, Parkhill J, Crook J, Horsnell T, Rice P, Rajanfream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945
11. Médigue C, Rechenmann F, Danchin A, Viari A: **Imagene: an integrated computer environment for sequence annotation and analysis.** *Bioinformatics* 1999, **15:**2-15
12. Andrade M, Brown N, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C: **Automated genome sequence analysis and annotation.** *Bioinformatics* 1999, **15:**391-412
13. Gaasterland T, Sensen CW: **Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture.** *Biochimie* 1996, **78:**302-310
14. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29:**11-16
15. Gribskov M, Devereux J, Burgess RR: **The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression.** *Nucleic Acids Res* 1984, **12:**539-549
16. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, *et al*: **Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3.** *DNA Research* 1998, **5:**55-76
17. Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jin-no K, Takahashi M, Sekine M, Baba SI, Ankai A, *et al*: **Complete Genome Sequence of an Aerobic Hyper-thermophilic Crenarchaeon, *Aeropyrum pernix* K1.** *DNA Research* 1999, **6:**83-101
18. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV: **Towards understanding the first genome of a Crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs).** *Genome Biol* 2000, **1:**0009.1-19
19. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28
20. Natale DA, Galperin MY, Tatusov RL, Koonin EV: **Using the COG database to improve gene recognition in complete genomes.** *Genetica* 2000, **108:**9-17
21. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*.** *Nucleic Acids Res* 1996, **24:**4420-4449
22. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, Doerks T, Sanchez-Pulido L, Snel B, Suyama M, *et al*: **Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames.** *Nucleic Acids Res* 2000, **28:**3278-3288
23. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, *et al*: **The minimal gene complement of *Mycoplasma senitalium*.** *Science* 1995, **270:**397-403
24. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, *et al*: **The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*.** *Nature* 1998, **392:**353-358
25. Salzberg SL, Delcher AL, Kasif S, White O: **Microbial gene identification using interpolated Markov models.** *Nucleic Acids Res* 1998, **26:**544-548
26. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL: **Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins.** *Nucleic Acids Res* 1999, **27:**260-262
27. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, *et al*: **Complete DNA sequence of a serogroup A strain of *Neisseria meninsitidis* Z2491.** *Nature* 2000, **404:**502-506
28. Saunders NJ, Jeffries AC, Peden JF, Hood DW, Tettelin H, Rappuoli R, Moxon ER: **Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meninsitidis* stain MC58.** *Mol. Microbiol* 2000, **37:**207-215
29. Médigue C, Rose M, Viari A, Danchin A: **Detecting and Analyzing Sequencing Errors: Toward a High Quality of the *Bacillus subtilis* Genome Sequence.** *Genome Research* 1999, **9:**1116-1127
30. Neer EJ, Schmidt CJ, Nambudripad R, Smith T: **The ancient regulatory-protein family of WD-repeat proteins.** *Nature* 1994, **371:**297-300
31. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirowasa M, Sugiura M, Sasamoto S, *et al*: **Sequence analysis of the Genome of the Unicellular Cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence Determination of the Entire Genome and Assignment of Potential Protein-coding Regions.** *DNA Research* 1996, **3:**109-136
32. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304
33. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A: **Evidence for horizontal gene transfer in *Escherichia coli* speciation.** *J Mol Biol* 1991, **222:**851-856
34. Lawrence JG, Roth JR: **Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters.** *Genetics* 1996, **143:**1843-1860
35. Karlin S, Mrazek J, Campbell AM: **Codon usages in different gene classes of the *Escherichia coli* genome.** *Mol Microbiol* 1998, **29:**1341-355
36. Lawrence JG, Ochman H: **Molecular archaeology of the *Escherichia coli* genome.** *Proc Natl Acad Sci USA* 1998, **95:**9413-9417
37. Borodovsky M, McIninch J, Koonin E, Rudd K, Médigue C, Danchm A: **Detection of new genes in the bacterial genome using Markov models for three gene classes.** *Nucleic Acids Res* 1995, **23:**3554-3562
38. Blattner D, Plunkett G, Bloch C, Perna N, Burland V, Riley M, Collado-Vides J, Glasner J, Rode C, Mayhew G, *et al*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1462
39. Rudd KE: **Linkage map of Escherichia coli K-12, edition 10: the physical map.** *Microbiol Mol Biol Rev* 1998, **62:**985-1019
40. Riley M, Labedan B: **Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of structural segment of homology, the module.** *J Mol Biol* 1997, **268:**857-868
41. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, *et al*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393:**537-544
42. Dundon WG, Marshall DG, Morain CA, Smyth CJ: **A novel tRNA-assodated locus (trl) from *Helicobacter pylori* is co-transcribed with tRNA(Gly) and reveals genetic diversity.** *Microbiology* 1999, **145:**1289-1298
43. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48