BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1622–1631
doi:10.1093/bioinformatics/bth137

Development of joint application strategies for two microbial gene finders

Alice C. McHardy[1], Alexander Goesmann[1], Alfred Pühler[2] and Folker Meyer[1,]*

[1]Center for Biotechnology (CeBiTec) and [2]Lehrstuhl für Genetik, Department of Biology, Bielefeld University, 33594 Bielefeld, Germany

Received on August 18, 2003; revised on December 19, 2003; accepted on January 15, 2004
Advance Access publication February 26, 2004

ABSTRACT

Motivation: As a starting point in annotation of bacterial genomes, gene finding programs are used for the prediction of functional elements in the DNA sequence. Due to the faster pace and increasing number of genome projects currently underway, it is becoming especially important to have performant methods for this task.

Results: This study describes the development of joint application strategies that combine the strengths of two microbial gene finders to improve the overall gene finding performance. Critica is very specific in the detection of similarity-supported genes as it uses a comparative sequence analysis-based approach. Glimmer employs a very sophisticated model of genomic sequence properties and is sensitive also in the detection of organism-specific genes. Based on a data set of 113 microbial genome sequences, we optimized a combined application approach using different parameters with relevance to the gene finding problem. This results in a significant improvement in specificity while there is similarity in sensitivity to Glimmer. The improvement is especially pronounced for GC rich genomes. The method is currently being applied for the annotation of several microbial genomes.

Availability: The methods described have been implemented within the gene prediction component of the GenDB genome annotation system.

Contact: fm@CeBiTec.Uni-Bielefeld.DE

INTRODUCTION

Microbial whole genome projects have become quite common today. Following sequencing and assembly, a functional description of the sequence is produced in the annotation phase. For storage, retrieval and processing of the information involved, annotation systems such as Artemis (Rutherford *et al*., 2000), ERGO (Overbeek *et al*., 2003), GenDB (Meyer *et al*., 2003) and MAGPIE (Gaasterland and Sensen, 1996) have been developed. In the first step in annotation, gene finders are usually applied for the prediction of functional elements such as coding sequences (CDSs) in the DNA sequence.

Compared with the more complex genetic organization in higher organisms, protein coding sequences in prokaryotic genomes possess a relatively simple structure. The task in microbial CDS prediction is to separate open reading frames (ORFs) that correspond to *in vivo* transcribed and translated regions of protein-coding sequence from the non-coding ORFs, which do not constitute functional elements of an organism's chromosome. A further issue is the determination of the correct start position, which, contrary to the stop position of a coding sequence, is not uniquely defined.

Different classes of microbial gene finders exist. *Ab initio* methods rely on the evaluation of intrinsic sequence properties such as the biased distribution of DNA oligomers in coding sequences. Programs that implement this approach include Glimmer (Salzberg *et al*., 1998; Delcher *et al*., 1999), GeneMark.hmm/S (Besemer and Borodovsky, 1999; Besemer *et al*., 2001), ZCURVE (Guo *et al*., 2003) and EasyGene (Larsen and Krogh, 2003). Extrinsic gene finders additionally use pairwise sequence similarity as 'external evidence' for their predictions; examples for these are the Critica (Badger and Olsen, 1999) and Orpheus (Frishman *et al*., 1998) programs. Yet another approach uses a 'Bio-dictionary' of prokaryotic protein sequence patterns for gene identification (Shibuya and Rigoutsos, 2002). For some genomes, a performance improvement has been achieved by combining the results from two or more programs (Guo *et al*., 2003; Tech and Merkl, 2004). These methods have been named the Glimmer ∩ ZCURVE (Guo *et al*., 2003) and YACOP [Critica ∪ (Glimmer ∩ ZCURVE)] (Tech and Merkl, 2004) strategies. For start site prediction, characteristic features of gene starts and the surrounding sequence, such as preferred start codons and ribosome binding site (RBS) patterns are utilized (Besemer *et al*., 2001; Guo *et al*., 2003; Badger and Olsen, 1999; Suzek *et al*., 2001).

There is a large number of microbial genome projects either recently finished or currently under way. It is becoming increasingly important to have performant gene prediction

*To whom correspondence should be addressed.

methods. These should allow the creation of high-quality genome annotation data while reducing superfluous human validation effort. In this study, this is tackled by the development of joint gene finding strategies based on the gene finders Glimmer and Critica. Both have different strengths, are freely available and can be utilized in automated high-throughput analysis on a Unix system. Information regarding their performance is currently scarce and available only for smaller sets of 7 (Tech and Merkl, 2004) or 18 (Guo *et al*., 2003) genomes. As this may not give a representative picture for all genomes available today, initially their performance was evaluated on 113 genome sequences belonging to a wide variety of microbial organisms. Glimmer was found to be the more sensitive program but its performance decreases strongly for GC rich genomes. For example, for the genomes of *Sinorhizobium meliloti* and *Streptomyces coelicolor* there are 1507 and 5817 false positive CDS predictions, respectively. Relying on the results of the program without further modifications results in an enormous manual validation effort for human annotators in genome projects. We tackled this problem by developing joint application strategies for the two programs. Using different parameters with relevance to the gene finding problem, combined strategies with optimized performance were devised.

MATERIALS AND METHODS

Data sets

The EMBL annotations of 114 genomic sequences of eubacterial and archaeal microorganisms were used in this study. A complete list can be found at http://www.CeBiTec.Uni-Bielefeld.DE/~alice/geneprediction/Sequences. To exclude annotation ambiguities, CDSs annotated with a non-integer number of codons or ending without a stop codon were excluded. Getorf (Olson, 2002) was used for ORF determination. Critica and Glimmer-2.1 were run with the option to use RBS information to locate the correct start position. For comparison of the Glimmer performance for genomes annotated using Glimmer versus those annotated using other gene finders, the Glimmer version available at the time of obtaining the annotation data was used (Glimmer-2.10). In further analyses, the latest version (Glimmer-2.13) was used, which uses a novel method for generation of the training set of CDSs.

Data sets of genes (*known function*) with known function or other supporting evidence were prepared for all genomes based on the information given in the CDS gene product description. For this, all CDSs described without an indication of function, experimental confirmation, sequence conservation or the occurrence of functional domains were classified as uncertain. Of the total set of 305 613 CDSs annotated for the 114 genomes, this was the case for 58 889 entries. The genomic sequence data with the corresponding annotated CDSs, gene finding results and ORFs can be browsed using the GenDB web front end (http://www.CeBiTec.Uni-Bielefeld.DE/~alice/geneprediction/gendb_cds.html). The genes considered as uncertain in this study can be identified by their ‘Status function’, which was set to ‘putative’.

Measuring performance and classification accuracy

In a two-class classification problem such as discriminating between non-coding ORFs and CDSs, the classification performance of a method can be evaluated by determining the numbers of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classified items, where $\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN} = \mathrm{N}$. Positives correspond to ORFs described as CDSs in the annotation, and negatives are the remaining, non-coding ORFs. Based on the sensitivity, $x = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, and specificity, $y = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, the correlation coefficient,

$$\mathrm{CC}(P, A) = \frac{N \cdot x \cdot y - \mathrm{TP}}{[(N \cdot x - \mathrm{TP})(N \cdot y - \mathrm{TP})]^{1/2}}, \qquad (1)$$

can be determined, which represents the accuracy of the predictive classification, P, with respect to the annotation A. It is entirely symmetric in x and y and provides a summary of gene finding performance based on all four parameters (Baldi *et al*., 2000).

In gene finding, the predictive result of a gene finder is usually not identical to a classification based on a single numerical measure. Besides a numerical measure of the ‘coding potential’ of an analyzed ORF, parameters such as overlap with neighboring predictions are typically employed for the prediction. To determine the discriminatory power of an internally used scoring methodology, ROC analysis (Swets, 1988) can be used. The receiver (or relative) operating characteristic (ROC) is a plot of the sensitivity versus the FP proportion $[\mathrm{FP}/(\mathrm{FP} + \mathrm{TN})]$ of the non-coding ORFs for various settings of the decision threshold. The area under the ROC curve measures the probability of correct classification and can be used as a single-valued, general measure of classification accuracy (Swets, 1988). ROC analysis was carried out for the different scoring methodologies used by Glimmer, for which the scores assigned to the ORFs during the analysis are available from the output. As the number of non-coding ORFs in bacterial genomes largely exceeds the number of CDSs, a partial ROC was calculated, similar to that used in performance evaluation of protein database search methods (Gribskov and Robinson, 1996; Schaeffer *et al*., 2001). $\mathrm{ROC}_{0.1}$ corresponds to the area under the ROC curve up to a FP proportion of 10%. For significance estimation of the difference in overall performance, sensitivity and specificity between the two gene finding methods, two-sample t-tests were applied, with the pooled variance for similar variance samples and the Welch approximation to the degrees of freedom otherwise.

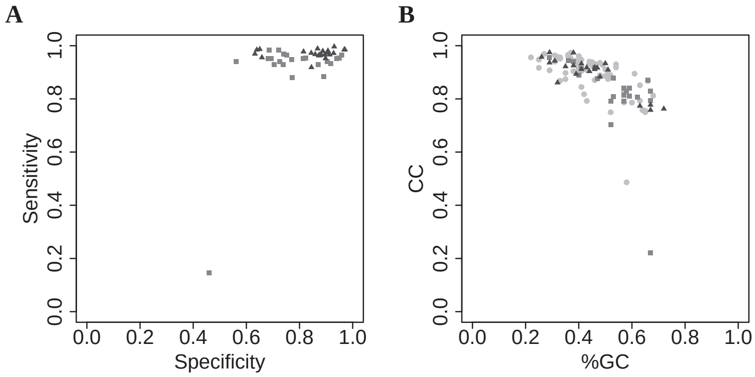

Fig. 1. Performance of Glimmer for genomes annotated using Glimmer or other gene finders in the annotation process. **(A)** Specificity versus sensitivity of Glimmer for genomes annotated using Glimmer ($\mathcal{G}$, orange squares) and genomes where other gene finders were employed ($\overline{\mathcal{G}}$, blue triangles). (**B**) Decreasing Glimmer performance with increasing GC content. Correlation of Glimmer predictions with annotation data versus genomic GC content for the $\mathcal{G}$, $\overline{\mathcal{G}}$ and remaining genomes (gray circles). This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

RESULTS

Composition of the data set

The current practice in microbial genome projects is to use one or more gene finders in combination with sequence database search methods such as BLAST (Altschul *et al.*, 1997) to locate potential coding sequences, followed by an additional manual effort of validation. It seemed necessary to first evaluate whether any of the utilized annotations mostly reflects the predictions of the employed gene finder as the CDS content, which would render it unsuitable as a standard of truth in performance evaluation. For Critica, to the best of our knowledge we do not know of any annotation in the data set where it has been applied for gene prediction. Glimmer has been frequently used. Its performance was thus compared for 22 genomes annotated using Glimmer ($\mathcal{G}$) with that for 23 genomes where other gene finders were applied ($\overline{\mathcal{G}}$; Fig. 1A). Surprisingly, the mean Glimmer performance is better for the $\overline{\mathcal{G}}$ set than for the $\mathcal{G}$ set [$\mathrm{CC}(P, A) = 0.89$ versus $\mathrm{CC}(P, A) = 0.82$]. Of the 114 sequences, for 14 Glimmer has a performance between 0.95 and 0.97. Three of the sequences belong to $\overline{\mathcal{G}}$, and only one to $\mathcal{G}$. The two for which Glimmer performs best are the genomes of *Clostridium perfringens* and *Listeria monocytogenes*, which both belong to $\overline{\mathcal{G}}$. Rather than the gene finder used, the genomic GC content has the major influence on prediction quality. Figure 1B shows decreasing Glimmer performance for genomes with higher GC content, which are more frequent in $\mathcal{G}$ than in $\overline{\mathcal{G}}$. Thus we did not exlude any genome because of the gene finder used in the annotation process.

For the genome of the archaebacterium *Aeropyrum pernix*, the sensitivity of both gene finders in reproducing the annotation data was found to be rather low (Glimmer, 0.59%; Critica, 0.56%). The *A.pernix* annotation contains all ORFs longer than 100 codons annotated as CDSs, which has been estimated to result in ~100% overannotation (Skovgard *et al.*, 2001). As this annotation thus seems not to be a good representation of CDS content, it was excluded. The remaining 113 genome sequences constitute the data set used in this study.

Gene finding performance of Glimmer and Critica

For the complete data set of 113 bacterial and archaeal genomes, the overall gene finding performance of both Glimmer and Critica is quite high. The mean correlation between predicted and annotated CDSs is 0.88 for Glimmer and 0.93 for Critica (Table 2). Glimmer has a statistically significant higher sensitivity than Critica (+5%, $p = 2.2 \times 10^{-12}$, determined with a two-sample t-test, see Methods section) but lacks in specificity (–13%, $p = 6.4 \times 10^{-22}$).

Some exceptions exist. For the *Mycobacterium leprae* genome, the specificity of Glimmer is only 22%, compared with 81% for Critica. This may be due to the unusually high content of pseudogenes among the annotated CDSs (40%). The resulting coverage of functional CDSs for this intracellular pathogen is 500 per megabase of genome sequence. This is about half the usual coverage for bacterial genomes and has been explained as an extreme case of reductive evolution (Cole *et al.*, 2001).

Also for a number of GC rich genomes, the performance of Critica is better (Figure 2A). Examples are the genomes of *Pseudomonas aeruginosa* (GC content 67%), *Ralstonia solanacearum* (67%) and, most pronounced, *S.coelicolor* (72%).

Glimmer(ct): improving gene finding performance for GC rich genomes

A problem that occurs in high GC content genomes when using Glimmer is how to obtain an adequate training set of coding sequences. This is needed for parameter estimation of the Glimmer Interpolated Context Model of CDSs. By default, Glimmer applies a script called *long-orfs* for this.

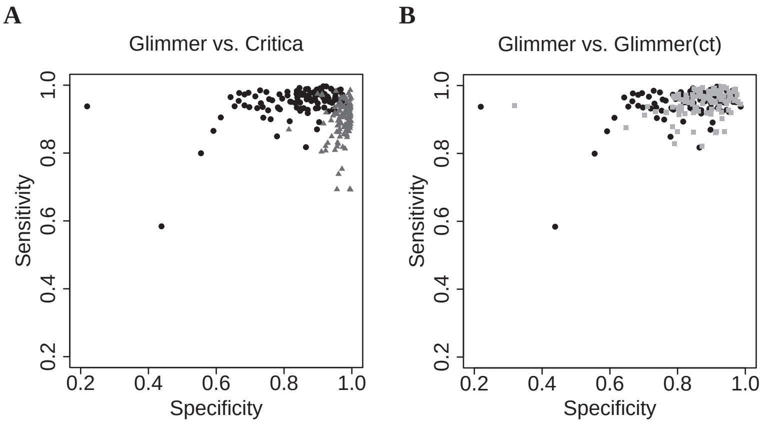

Fig. 2. Comparison of tool performance for Glimmer, Critica and the Critica-trained Glimmer(ct) on 113 prokaryotic genome sequences. **(A)** and **(B)** Sensitivity versus specificity for Glimmer (black circles) versus Critica (red triangles) and versus Glimmer(ct) (green squares). With Glimmer(ct), Critica was used to generate the training set of CDSs for parameter estimation of the Glimmer model. This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

Up to and including Glimmer, version 2.10, *long-orfs* detects all non-overlapping ORFs longer than 500 bp in a given genomic sequence. But the number of such non-overlapping, long ORFs decreases strongly with increasing GC content of a genome. At some point it is too small to be used (Guo *et al*., 2003). Recently, a novel version of *long-orfs* was released that computes an optimal minimum length of ‘long orfs’ to enlarge the training set. Also with the novel version, a difference in performance is evident for GC rich (>56%) genomes compared with sequences of lower GC content (Table 2). For the GC rich genomes, both sensitivity and specificity are reduced (−3%, −18%). Figure 3A shows decreasing performance with increasing GC content of the individual sequences.

We thus evaluated how changing the composition of the training set further can be used to improve the gene prediction performance. An iterative usage, that is using an initial set of predictions as a training set for another Glimmer run, did not lead to any improvement (data not shown). With Glimmer(ct), the more specific Critica CDS predictions were used as the training set. This results in a statistically significant 2% performance improvement compared with the standard application ($p = 0.04$, Fig. 2C). The Glimmer(ct) prediction is more specific (+3%, $p = 0.02$) without losing sensitivity. For GC rich genomes, the improvement is even more pronounced (+9% in specificity, +1% in sensitivity; Table 2).

For Critica, there is a slight loss in both sensitivity and specificity, which results in a 2%($p = 0.027$) difference in overall gene finding performance between GC rich and the remaining genomes (Fig. 3B).

Gene finding performance for different gene lengths

To examine the relation between gene length and prediction performance for Glimmer, Critica and Glimmer(ct), the sensitivity and selectivity for different settings of the minimum CDS length were compared (Fig. 3D). The values at a minimum length of 90 bp correspond to those given in Table 2. The specificity of all three gene finders decreases for shorter CDS lengths. This is more pronounced for Glimmer and Glimmer(ct) than for Critica, which is the most specific tool for all lengths. Glimmer(ct) has the highest sensitivity in detecting longer CDSs. Only when considering the complete set of CDSs longer than 90 bp, it becomes identical to that of the standard application.

Diagnostic accuracy of the Glimmer scores

Three numerical scores are available from the Glimmer output for the ORFs analyzed. These are a length-normalized raw log-score, a probability and a vote score, which is the sum of the probability scores for subregions contained within the ORF sequence analyzed in other frames. The primary decision criterion Glimmer uses is the probability score; optionally ORFs with vote scores above a certain threshold are also predicted. We were interested in determining which of these scores allows the most reliable prediction of CDSs. As a measure of predictive accuracy, $\mathrm{ROC}_{0.1}$ was determined for the different measures. Figure 4A shows a density estimate for the $\mathrm{ROC}_{0.1}$ distributions of the raw, probability and vote scores for the 113 genomes. With a mean $\overline{\mathrm{ROC}}_{0.1}$ of 0.93, the vote score allows the most accurate discrimination between CDSs and hypothetical ORFs. The raw and probability scores are less informative ($\overline{\mathrm{ROC}}_{0.1}$ of 0.81 and 0.88).

The vote score may be be used to divide Glimmer(ct) results into probably correct and less certain CDS predictions. To determine the optimal setting, this was evaluated with different threshold settings. The maximum specificity a subset of Glimmer(ct) predictions with high vote scores achieves is

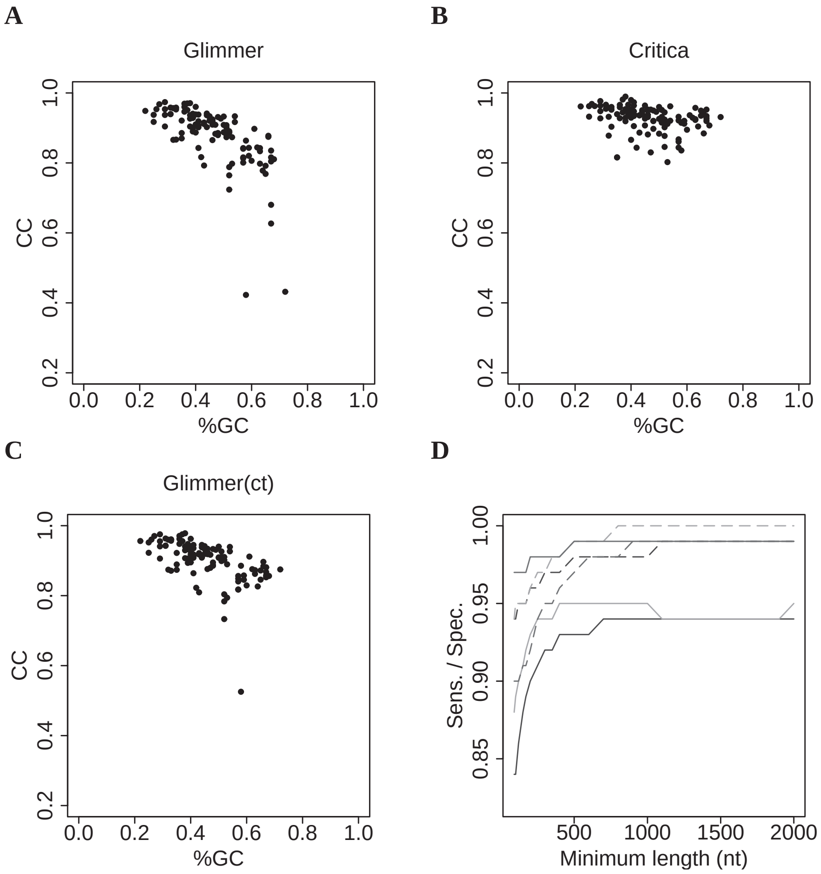

Fig. 3. Relation of gene finding performance to genomic GC content and gene length. **(A–C)** Performance of Glimmer, Critica and Glimmer(ct) versus genomic GC content for 113 microbial genomes. **(D)** Sensitivity (dashed line) and specificity (solid line) of Glimmer (blue), Glimmer(ct) (green) and Critica (red) for different minimum gene length settings.

99%. The lowest threshold setting where this specificity is reached is a vote score of 400 (Fig. 4B). About 99% of the predicted CDSs with vote scores ≥400 are thus correct predictions, which covers 56% of all annotated CDSs (Table 1). The remaining, lower scoring Glimmer(ct) predictions contain a high percentage of FPs, which makes their manual validation seem especially important.

Development of combined strategies

Typical for bacterial genome sequences, the number of non-coding ORFs largely exceeds the number of CDSs. For the genomes analyzed, the ratio of CDSs to non-coding ORFs lies between 0.03 (*M.leprae*) and 0.17 (*Sulfolobus tokodaii*) for ORFs longer than 90 bp. In manual annotation, it is therefore considerably less effort to discard FP CDS predictions rather than check for FNs among the non-coding ORFs. A gene prediction strategy based on a combination of different tool results should thus improve the specificity without significantly losing in sensitivity compared with the individual tools. To achieve this, we pursued the following idea for two parameters with relevance to the gene finding problem: given a set of very reliable, highly specific CDS predictions (Critica) and a set of additional, more uncertain ones (Glimmer(ct)), can parameter settings be determined that allow the removal of mostly the FP, additional predictions? The parameters we focused on sequentually were the allowed overlap length of additional Glimmer(ct) prediction with Critica ones and the Glimmer(ct) vote score, which was determined to be the most accurate measure for CDS prediction.

The simple union of Critica and Glimmer(ct) predictions did not result in any significant change in performance compared with Glimmer(ct), as the set of Critica predictions is almost completely contained in the Glimmer(ct) ones (Table 2).

With the overlap threshold strategy (OTS), additional Glimmer(ct) predictions are discarded if their overlap length with Critica predictions exceeds a given threshold. For parameter estimation, different settings of maximum allowed overlap length were tried, and Glimmer(ct) predictions with more overlap removed. The maximal correlation coefficient $\mathrm{CC}(P, A)$ was achieved with an allowed overlap length of 10 bp. For the individual genomes, the optimal setting was ≤50 bp for 99 genomes and between 100 and 600 bp

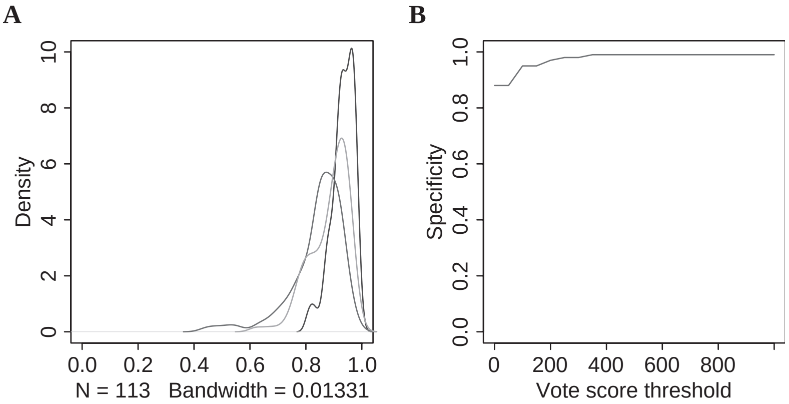

Fig. 4. Diagnostic accuracy of the different Glimmer scores. **(A)** Density estimate of the $\mathrm{ROC}_{0.1}$ distribution for the vote (blue), raw (red) and probability scores (green) for the 113 genomes. **(B)** Specificity of the remaining Glimmer(ct) predictions for different settings of the vote score threshold.

Table 1. Using the Glimmer vote score to divide predictions into probably correct ones and less certain candidates in need of manual validation

Gene finder	CC(P, A)	Sensitivity	Specificity
Glimmer(ct)	0.90	0.95	0.87
Vote score >400		0.56	0.99
OTS	0.92	0.94	0.92
Vote score >200		0.91	0.97

Given is the lowest vote score setting with which the maximal specificity could be obtained.

for another 11. Only for three genomes (*Fusobacterium nucleatum*, *Escherichia coli* CFT073 and *Leptospira interrogans*) could the performance not be increased thus. To account for genomes where the 10 bp setting is too strict, 50 bp was used as the final parameter setting with OTS. This increases specificity by 4% ($p = 6.5 \times 10^{-5}$) without significantly losing sensitivity (Table 2).

The vote score threshold strategy (VTS) uses these to further improve specificity. Additional Glimmer(ct) predictions are discarded if their vote score is lower than a given threshold setting. For determination of the optimal threshold setting, different settings of the vote score threshold between 0 and 1000 were tried. For 90 of the individual genomes, threshold settings were found that led to a performance improvement. The maximum overall performance was obtained when disregarding all predictions with vote scores <100 (Table 2). Using this parameter setting further significantly increases specificity by 4% ($p = 2.67 \times 10^{-6}$) but is also associated with some loss in sensitivity (-2%, $p = 0.004$).

As disregarding Glimmer(ct) predictions with low vote scores results in some sensitivity loss, these may instead be used to single out ‘uncertain’ candidate genes requiring human attention. Determination of the lowest vote score setting for which the set of higher scoring OTS predictions retains the maximum specificity led to a threshold of 200. In combination with the Critica predictions, the higher scoring Glimmer(ct) predictions of the OTS strategy cover 91% of the annotated CDSs, with an associated probability of 0.97 that these are correct (Table 1). The more uncertain additional Glimmer(ct) predictions with lower vote scores remaining with OTS should be given special attention in the manual validation process.

Performance evaluation

Both OTS and VTS exhibit a significant performance improvement (Figure 5; Tables 2 and 3). Compared with the Glimmer standard application, for OTS the specificity is improved by 8% ($p = 2.8 \times 10^{-8}$), without losing significantly in sensitivity (0%, $p = 0.45$). This is also true for the more reliable *known function* genes of the annotations (+4% in performance, +7% in specificity, no loss in sensitivity). VTS is even more specific (+11%, $p = 2.4 \times 10^{-17}$) but has some loss in sensitivity (-2%, $p = 3.4 \times 10^{-4}$). For the known function genes, there is no significant sensitivity loss with VTS. The performance improvement of both strategies is most pronounced for GC rich genomes (Tables 2–4). As an example, the number of FP predictions for the *S.meliloti* chromosome is reduced from 1507 for Glimmer to 100/47 with OTS and VTS (Table 4).

Of the seven genomes used for evaluation of the method, VTS has the same overall performance as YACOP (Table 5), which uses a $\left[\text{Critica} \cup (\text{Glimmer} \cap \text{ZCURVE})\right]$ combination of gene finding results (Tech and Merkl, 2004). OTS performs slightly worse. But only one of the seven genomes has a GC content >50%, and the authors state that the performance decreases for GC rich genomes. Compared with a Glimmer $\cap$ ZCURVE strategy evaluated on a four genome data set with two GC rich genomes (Guo *et al*., 2003), both OTS and VTS perform better (Table 5).

Table 2. Mean sensitivity, selectivity and overall performance of different gene finding methods on 113 bacterial and archaeal genomes

Gene finder	CC(P, A)	Sensitivity	Specificity
Glimmer[a]	0.88 ± 0.10 (0.77 ± 0.13)[b]	0.95 ± 0.08 (0.93 ± 0.16)	0.84 ± 0.12 (0.68 ± 0.11)
Glimmer[c]	0.88 ± 0.09 (0.78 ± 0.12)	0.95 ± 0.05 (0.93 ± 0.08)	0.84 ± 0.12 (0.70 ± 0.13)
Glimmer(ct)[d]	0.90 ± 0.06 (0.85 ± 0.07)	0.95 ± 0.04 (0.93 ± 0.03)	0.87 ± 0.08 (0.80 ± 0.10)
Critica	0.93 ± 0.04 (0.91 ± 0.03)	0.90 ± 0.06 (0.88 ± 0.04)	0.97 ± 0.03 (0.96 ± 0.04)
Union	0.90 ± 0.06 (0.85 ± 0.07)	0.95 ± 0.04 (0.94 ± 0.03)	0.87 ± 0.08 (0.80 ± 0.10)
OTS	0.92 ± 0.05 (0.91 ± 0.08)	0.94 ± 0.04 (0.92 ± 0.03)	0.92 ± 0.07 (0.91 ± 0.12)
VTS	0.93 ± 0.04 (0.92 ± 0.06)	0.93 ± 0.05 (0.91 ± 0.03)	0.95 ± 0.05 (0.94 ± 0.09)

[a] Version 2.10.
[b] The values in parentheses are for the 27 genomes with a genomic GC content >0.56.
[c] Version 2.13, using a new version of long-orfs for training set creation.
[d] Version 2.13, using Critica for training set creation.

Table 3. Mean sensitivity, selectivity and overall performance of different gene finding methods for genes of known function or with other confirmation

Gene finder	CC(P, A)	Sensitivity	Specificity
Glimmer	0.79 ± 0.12 (0.72 ± 0.13)[a]	0.98 ± 0.04 (0.96 ± 0.08)	0.68 ± 0.16 (0.59 ± 0.15)
Glimmer(ct)[b]	0.81 ± 0.10 (0.79 ± 0.10)	0.98 ± 0.02 (0.98 ± 0.02)	0.71 ± 0.15 (0.67 ± 0.14)
Critica	0.86 ± 0.10 (0.87 ± 0.09)	0.95 ± 0.03 (0.94 ± 0.03)	0.81 ± 0.15 (0.83 ± 0.15)
Union	0.81 ± 0.10 (0.79 ± 0.10)	0.98 ± 0.02 (0.98 ± 0.01)	0.71 ± 0.15 (0.67 ± 0.14)
OTS	0.84 ± 0.10 (0.85 ± 0.11)	0.98 ± 0.02 (0.97 ± 0.02)	0.74 ± 0.15 (0.78 ± 0.17)
VTS	0.86 ± 0.10 (0.87 ± 0.10)	0.97 ± 0.02 (0.96 ± 0.02)	0.79 ± 0.15 (0.80 ± 0.16)

[a] The values in parentheses are for the 27 genomes with a genomic GC content >0.56.
[b] Using Critica for training set creation.

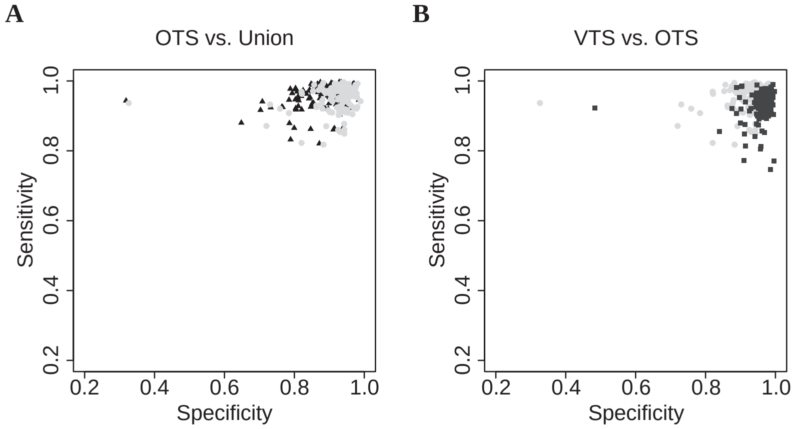

Fig. 5. Performance of the combined strategies. **(A)** and **(B)** Sensitivity versus specificity for OTS (light blue circles) versus the union set (black triangles) and versus VTS (dark blue squares). This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

DISCUSSION

This work describes the development of joint application strategies for two microbial gene finders, which combine the strengths of both tools to improve the overall gene finding performance. The comparative sequence analysis approach that Critica employs ensures its high specificity in the detection of similarity-supported genes. In the interpretation of the results of pairwise DNA sequence comparisons, Critica makes use of the degeneracy of the genetic code to discriminate conserved coding regions from conserved non-coding regions (Badger and Olsen, 1999). Similar approaches are also increasingly becoming popular in the field of eukaryotic gene prediction (Rogozin *et al.*, 1999; Moore and Lake, 2003). Compared with approaches that use similarity

Table 4. Sensitivity and FP proportion of predictions (1 − specificity) for Glimmer, Glimmer(ct), OTS and VTS for 27 genome sequences with a GC content >0.56

Organism	GenBank	Glimmer		Glimmer(ct)		OTS		VTS	
	acc. no.	sens.	1 − spec.	sens.	1 − spec.	sens.	1 − spec.	sens.	1 − spec.
Deinococcus radiodurans	AE000513	2521 (0.98)	1107 (0.31)	2483 (0.96)	517 (0.17)	2423 (0.94)	156 (0.06)	2415 (0.94)	135 (0.05)
Mycobacterium tuberculosis	AE000516	3910 (0.93)	758 (0.16)	3873 (0.93)	621 (0.14)	3780 (0.9)	300 (0.07)	3671 (0.88)	209 (0.05)
Deinococcus radiodurans	AE001825	338 (0.95)	124 (0.27)	334 (0.94)	83 (0.2)	330 (0.92)	28 (0.08)	329 (0.92)	25 (0.07)
Pseudomonas aeruginosa	AE004091	4814 (0.87)	3323 (0.41)	5375 (0.97)	1315 (0.2)	5323 (0.96)	165 (0.03)	5303 (0.95)	135 (0.02)
Caulobacter crescentus	AE005673	3584 (0.96)	1156 (0.24)	3476 (0.93)	584 (0.14)	3427 (0.92)	175 (0.05)	3404 (0.91)	155 (0.04)
Chlorobium tepidum	AE006470	2013 (0.89)	452 (0.18)	1942 (0.86)	352 (0.15)	1912 (0.85)	116 (0.06)	1829 (0.81)	80 (0.04)
Agrobacterium tumefaciens	AE008688	2579 (0.93)	846 (0.25)	2548 (0.91)	620 (0.2)	2520 (0.9)	193 (0.07)	2506 (0.9)	134 (0.05)
Agrobacterium tumefaciens	AE008689	1753 (0.93)	489 (0.22)	1721 (0.92)	373 (0.18)	1708 (0.91)	100 (0.06)	1698 (0.91)	61 (0.03)
Brucella melitensis	AE008917	1926 (0.94)	688 (0.26)	1895 (0.92)	461 (0.2)	1878 (0.91)	134 (0.07)	1858 (0.9)	58 (0.03)
Brucella melitensis	AE008918	1061 (0.93)	293 (0.22)	1055 (0.93)	245 (0.19)	1039 (0.91)	72 (0.06)	1037 (0.91)	28 (0.03)
Xanthomonas campestris Pv. *campestris*	AE008922	4083 (0.98)	2033 (0.33)	4010 (0.96)	943 (0.19)	3946 (0.94)	231 (0.06)	3933 (0.94)	122 (0.03)
Xanthomonas axonopodis Pv. *citri*	AE008923	4160 (0.96)	2320 (0.36)	4036 (0.94)	1105 (0.21)	3942 (0.91)	254 (0.06)	3931 (0.91)	113 (0.03)
Methanopyrus kandleri	AE009439	1660 (0.98)	322 (0.16)	1661 (0.98)	269 (0.14)	1639 (0.97)	180 (0.1)	1620 (0.96)	117 (0.07)
Brucella suis	AE014291	1913 (0.9)	677 (0.26)	1828 (0.86)	456 (0.2)	1819 (0.86)	147 (0.07)	1781 (0.84)	112 (0.06)
Brucella suis	AE014292	1033 (0.9)	325 (0.24)	1009 (0.88)	275 (0.21)	999 (0.87)	122 (0.11)	973 (0.85)	95 (0.09)
Bifidobacterium longum	AE014295	1612 (0.93)	625 (0.28)	1592 (0.92)	482 (0.23)	1592 (0.92)	217 (0.12)	1589 (0.92)	174 (0.1)
Pseudomonas putida	AE015451	5240 (0.98)	1768 (0.25)	5099 (0.95)	1107 (0.18)	5063 (0.95)	263 (0.05)	5006 (0.94)	213 (0.04)
Pseudomonas syringae pv. *tomato*	AE016853	5253 (0.96)	1359 (0.21)	5174 (0.95)	959 (0.16)	5152 (0.94)	346 (0.06)	5059 (0.92)	254 (0.05)
Ralstonia solanacearum	AL646052	2747 (0.8)	2204 (0.45)	3227 (0.94)	766 (0.19)	3182 (0.93)	64 (0.02)	3160 (0.92)	40 (0.01)
Mesorhizobium loti	BA000012	6649 (0.98)	2468 (0.27)	6457 (0.96)	1460 (0.18)	6348 (0.94)	320 (0.05)	6257 (0.93)	180 (0.03)
Corynebacterium efficiens	BA000035	2740 (0.93)	738 (0.21)	2713 (0.92)	488 (0.15)	2670 (0.91)	94 (0.03)	2656 (0.9)	54 (0.02)
Bradyrhizobium japonicum	BA000040	7930 (0.95)	3971 (0.33)	7665 (0.92)	2337 (0.23)	7563 (0.91)	765 (0.09)	7528 (0.91)	431 (0.05)
Halobacterium Sp. NRC-1	HSPNRC1XX	1990 (0.97)	793 (0.28)	1925 (0.94)	446 (0.19)	1871 (0.91)	89 (0.05)	1844 (0.9)	74 (0.04)
Mycobacterium leprae	MLEPRAE	1527 (0.94)	5438 (0.78)	1533 (0.94)	3285 (0.68)	1526 (0.94)	3150 (0.67)	1503 (0.92)	1603 (0.52)
Mycobacterium tuberculosis	MTBH37RV	3786 (0.97)	886 (0.19)	3776 (0.97)	692 (0.15)	3710 (0.95)	385 (0.09)	3687 (0.94)	215 (0.06)
Streptomyces coelicolor	SCO645882	4546 (0.58)	5817 (0.56)	7393 (0.95)	1473 (0.17)	7165 (0.92)	165 (0.02)	7114 (0.91)	105 (0.01)
Sinorhizobium meliloti	SME591688	3249 (0.97)	1507 (0.32)	3237 (0.97)	872 (0.21)	3227 (0.97)	100 (0.03)	3201 (0.96)	47 (0.01)

Table 5. Comparison with the YACOP and Glimmer ∩ ZCURVE combined strategies

Gene finder	CC(P, A)	Sensitivity	Specificity
I[a]			
Glimmer	0.91 ± 0.03	0.97 ± 0.01	0.87 ± 0.04
YACOP	0.96 ± 0.01	0.98 ± 0.01	0.95 ± 0.02
OTS	0.94 ± 0.02	0.97 ± 0.02	0.91 ± 0.03
VTS	0.96 ± 0.01	0.95 ± 0.01	0.97 ± 0.01
II[b]			
Glimmer	0.82 ± 0.11	0.94 ± 0.05	0.75 ± 0.14
ZCURVE ∩ Glimmer	0.94 ± 0.02	0.97 ± 0.01	0.92 ± 0.03
OTS	0.95 ± 0.02	0.96 ± 0.01	0.94 ± 0.04
VTS	0.96 ± 0.01	0.95 ± 0.00	0.97 ± 0.01

[a] For the seven-genome data set used in Tech and Merkl (2004).
[b] For the four-genome data set used in Guo *et al*. (2003).

at the amino acid level, an advantage of this is the independence from the existing accurate annotation, which is used to generate the content of protein sequence databases. If using comparisons at the amino acid level, genes may be missed whose homologs have not been annotated or annotated too short. In our analyses, we found Critica to be very robust. It performs well on sequences with a high GC-content and also on the *M.leprae* genome, which contains a large number of pseudogenes. Its strength is its high specificity, which is also evident in the detection of *known function* genes. It is also the most specific in predicting short genes.

The gene finder Glimmer relies completely on an *ab initio* approach in gene identification. It uses a very sophisticated model of sequence properties of prokaryotic CDSs (Delcher *et al*., 1999). It is also highly sensitive in the detection of genes supported by additional evidence. For GC rich genomes, it loses in prediction performance, mainly due to a specificity loss. We found that by using the very specific Critica predictions as a training set for the Glimmer CDS model, performance in terms of both sensitivity and specificity can be significantly improved.

A troublesome issue is the unknown quality of many CDS entries in the current annotation data. The annotation describes the CDS content of a genomic sequence and thus is by definition the standard of truth against which gene finding performance is evaluated. In its creation, considerable human effort is often involved to achieve high quality. Still, for no genome are all annotated CDSs supported by experimental or other convincing evidence. A comparison of the length distribution of annotated genes with genes matching a known protein led to the conclusion that many genomes might currently be over-annotated, especially concerning short genes (Skovgard *et al.*, 2001). Because of the size of the data set analyzed, the results deduced in this study are unlikely to be influenced much by the varying strategies of individual annotation projects. The performance improvement achieved by the combined strategies was also observed in evaluation with the more reliable *known function* genes of the annotations. For all methods, the sensitivity in detection of these more reliable subsets was found to be even higher than for the complete set of annotated genes.

In the development of combined gene prediction strategies, the very specific Critica predictions were initially set as fixed and combined with different subsets of additional Glimmer(ct) predictions to improve the overall performance. For specification of this additional subset, the use of two different parameters with relevance to the gene finding problem was evaluated. The first is the allowed overlap length of neighboring genes, as genes of longer overlap length are generally considered unlikely for prokaryotic organisms, although there is no systematic research on this issue. From a biological perspective, this may be explained by the extreme constraints that are placed on a sequence that is coding in two different frames. We found that by removing additional predictions with long overlaps, the specificity in gene identification can be considerably improved without a significant loss of sensitivity. The second parameter is the Glimmer(ct) vote score, which was determined to be the Glimmer scoring method that allows the most accurate discrimination between non-coding ORFs and CDSs. Discarding low vote score predictions results in a further gain in specificity but is accompanied by a slight sensitivity loss. However, for the *known function* subsets of genes, there is no significant sensitivity loss. The additional genes missed by VTS are thus both low-scoring, according to sequence composition, and without indication of function or biological activity, according to the annotation data. They are either falsely annotated or real genes that are difficult to determine, such as the genes contained in prophage DNA. Using OTS allows a considerable reduction of the necessary manual validation effort of the gene finding results for the human annotators, especially for GC rich genomes. As an example, with OTS the false positive prediction rate for the *S.meliloti* chromosome is reduced from 32% for Glimmer to 2%, without a loss of sensitivity.

The methods described have been implemented as the Reganor auto-annotation component of the GenDB genome annotation system and are currently being applied in several bacterial genome projects. We hope that the software and additional information presented in this work will be helpful to annotators in producing high-quality genome annotation.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for insightful comments leading to significant improvements. We thank Robert Giegerich, Lutz Krause, Tim Nattkemper, Marc Rehmsmeier, Ross Overbeek and Gene Myers for providing critical comments and discussion. We are also very grateful to Rainer Merkl and Maike Tech for sharing information concerning their work before publication. Furthermore, we want to thank Thierry Lombardot, Hanno Teeling and Frank-Oliver Glöckner for bringing the gene finder Critica to our attention. A.C.M. was supported by the DFG Graduiertenkolleg 635 Bioinformatik.

SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Badger,J.H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, **16**, 512–524.

Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A.F. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.

Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.

Cole,S.T., Eiglmeier,K., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D., Mungall,K. *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

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

Gaasterland,T. and Sensen,C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–78.

Gribskov,M. and Robinson,N.L. (1996) The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Guo,F.-B., Ou,H.-Y. and Zhang,C.-T. (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.*, **31**, 1780–1789.

Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.

Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. and Puhler,A. (2003) GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.*, **31**, 2187–2195.

Moore,J.E. and Lake,J.A. (Nucleic Acids Res.). Gene structure prediction in syntenic DNA segments. *Nucleic Acids Res.*, **31**, 7271–7279.

Olson,S.A. (2002) EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief. Bioinform.*, **3**, 87–91.

Overbeek,R., Larsen,N., Walunas,T., D'Souza,M., Pusch,G., Selkov,E., Liolios,K., Joukov,V., Kaznadzey,D., Anderson,I., *et al.* (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res.*, **31**, 164–171.

Rogozin,I.B., D'Angelo,D. and Milanesi,L. (1999) Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. *Gene*, **226**, 129–137.

Rutherford,K.M., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.-A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

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

Schaeffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Shibuya,T. and Rigoutsos,I. (2002) Dictionary-driven prokaryotic gene finding. *Nucleic Acids Res.*, **30**, 2710–2725.

Skovgard,M.S., Jensen,L.J., Brunak,S., Ussery,D. and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends. Genet.*, **17**, 425–427.

Suzek,B.E., Ermolaeva,M.D., Schreiber,M. and Salzberg,L.S. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.

Swets,J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Tech,M. and Merkl,R. (2004) YACOP: enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol.*, **3**, 441–451.