

## Background

The rate of prokaryotic genome sequencing is increasing rapidly and many complete genome sequences are publicly available [1]. This increased availability of genomic sequences has enabled a large-scale comparative analysis of coding sequence (CDS) sets for high-resolution comparison utilizing such information as sequence conservation, domain composition, and protein interaction. However, depending on the genome, CDS prediction is still alarmingly inconsistent. Also, it requires continuous and automatic updates as the new data from the expanding public databases accumulates [2-4] especially for large bioinformatics information providers.

There have been numerous computational methods for processing prokaryotic genomic CDSs. Most of them are *ab initio* prediction methods. They provide the genomic coordinates of candidate CDSs using their own statistical and mathematical algorithms: GeneMark [5], ECOPARSE [6], GeneHacker [7], GeneMark.hmm [8], GLIMMER [9,10], GeneMarkS [11], EasyGene [12], ZCURVE [13], and GeneLook [14].

However, the CDSs predicted by such programs usually need further time-consuming manual processes. This is mainly due to low accuracy and insufficient evidence.

Moreover, their accuracy usually depends on the quality of training sets and/or 'seed' CDSs that need manual validation for better performance. Even though GeneMarkS and GeneLook have increased the prediction accuracy, and GeneLook and the modified

EasyGene [15] have automated those manual steps, they still do not provide comprehensive information on predicted CDSs. Information such as frame-shifts, homology-based gene evidence, and best pair-wise matches against other prokaryotes is invaluable for professional curation and large-scale comparative analysis.

To complement such *ab initio* prediction methods, some CDS prediction programs add homology-based methods: ORPHEUS [16], Critica [17], Framed [18], and YACOP [19]. ORPHEUS uses the DPS (DNA-Protein Search) program [20] to compare a given genomic sequence with a non-redundant protein sequence databank. Framed can utilize a BLASTX [21] output provided by the user and provides predicted frame-shifts and conserved regions with other proteins. Critica uses the BLASTN program [21] to align a given genomic sequence with its related sequences chosen from DNA databases.

YACOP combines three gene-predicting programs, Critica, Glimmer, and ZCURVE.

However, ORPHEUS and Framed have shown relatively low CDS prediction accuracies. Although Critica and YACOP have achieved an increase in the prediction accuracy, they provide no automated prediction pipeline for the increasing number of prokaryotic genomes. Moreover, neither provides reliability information nor frame-shift or comparative genomic information for their predicted CDSs.

We have developed CONSORF, a fully automated and regularly updated prediction system for prokaryotic CDSs for large genomics information providers such as

universities and research organizations. CONSORFalso aims to improve CDS prediction sensitivity and specificity. It provides consistent and comprehensive information for predicted CDSs such as intuitive reliability scores, predicted frame-shifts, alternative start sites, conserved regions, and best pair-wise matches against other prokaryotes. The CDSs of publicly available prokaryotic genomes predicted by CONSORF are freely accessible and downloadable through our web site [22].