

Methods

Public prokaryotic sequences and information

We have downloaded the public genome and annotated proteome resources of 565 prokaryotic organisms from NCBI RefSeq [28]. We collected and utilized genome sequences from all the available prokaryotes, including both chromosome and plasmid sequences in NCBI '.fna' files and their annotated proteome sequences in '.faa' files. Other information, such as organism names, taxonomic IDs, translation tables, and CDS (coding sequence) coordinates, were parsed from GenBank ('.gbk') files. CDSs (called 'public CDSs') with stop codons ('TAA', 'TAG', or 'TGA') at the C-terminal end were collected to compare with predicted CDSs for evaluation.

Pair-wise FASTX comparisons

A pair-wise FASTX (version 3.4) comparison between the collected genome and proteome sequences was performed, avoiding self comparison, on a cluster of five Linux machines, each with an AMD Opteron™ Processor 848 and 16GByte memory. The total run time for the 384 organisms exceeds two months. However, the current update process is much faster, since only new or changed organisms are processed. Owing to the restriction on the query DNA length of the FASTX program, we have fragmented each genome into 18 kilo-base segments overlapping their neighbors on both sides over 9 kilo-bases.

Homology-based consensus CDSs

Each FASTX alignment under cutoff (e-value<0.001 and number of frame-shifts or in-frame stops < 4) was parsed and assigned to the corresponding CDS with a stop and candidate starts on both sides (Figure 2). Removing FASTX alignments significantly overlapping (more than 10% of alignment length) with an alignment with the higher bit score, non-redundant CDS sets were extracted and utilized for consensus analysis to generate the reliability information of each CDS.

We provide three different types of reliability information for all the homology-based CDSs ('homology CDSs') based on a consensus analysis: stop only (type A), stop and start only (type B), and stop, start, and frame change (type C) (Figure 2B). A frame change includes both frame-shifts and in-frame stops. Each type of reliability information for an organism's CDS includes the number of occurrences, as well as the sum of bit scores counted and added up over all the pair-wise FASTX alignments with the other organisms. Among the FASTX alignments supporting the same CDS with the greatest sum of bit scores, the alignment providing the best bit score was chosen as the representative alignment and used for the determination of the candidate starts, as well as the primary annotation (Figure 2). All the candidate 'homology CDSs' from the consensus analysis were sorted by stop (type A) reliability score, and the CDSs overlapping significantly (more than 10% of their length based on the shortest start)

with an CDS with the higher reliability score were again removed to generate the final 'homology CDSs'.

Alternative homology-based CDSs

Contrary to 'homology CDSs' that are typically based on the sum of bit scores from the consensus analysis, 'alternative CDSs' of an organism were determined from the FASTX alignments with the highest individual bit score across all the pair-wise FASTX comparisons.

Multiple *ab initio* predictions

Glimmer 2.13 was obtained from TIGR [29] and was executed in a default setting via the 'run-glimmer2' script. GeneMark 2.3.2 and prokaryotic GeneMark.hmm 2.3.2 were obtained from Gene Probe [30] and were run following the instructions. GeneMark was run with built-in matrices if available. For organisms with no available matrices, we custom built matrices by running the 'mkmat' script of the GeneMark program based on the 'homology CDSs'. Contrary to Glimmer and GeneMark, prokaryotic GeneMark.hmm was optionally run on a default setting for those organisms with built-in HMM model files.

Algorithm-based consensus CDSs

Since no frame change is considered in our *ab initio* predictions, two types of reliability information regarding stop only (type A) and both stop and start (type B) were provided, utilizing CDS nucleotide length as reliability score. Both the number of

occurrences and the sum of CDS lengths over all the available *ab initio* predictions were also denoted as in homology-based CDSs. All candidate CDSs from the algorithm-based consensus analysis were sorted by a stop (type A) reliability score and then included in the final ‘*ab initio* CDSs’.

The candidate starts of each CDS include all the probable starts between the longest start and the most probable start (called the shortest start) with the largest number of occurrences. In cases where there are multiple starts with the same largest number of occurrences, the rightmost start among them was conservatively assigned as the shortest start.

Integration of homology-based and algorithm-based CDSs

To integrate ‘homology CDSs’ and ‘*ab initio* CDSs’, the ‘homology CDSs’ with a sum of bit scores greater than a given cutoff based on the stop (type A) reliability score were chosen first. CDSs with one or more frame changes were removed for better accuracy in evaluation. Next, the ‘*ab initio* CDSs’ with a sum of CDS lengths greater than another stop (type A) score cutoff were then chosen, removing CDSs overlapping significantly with the existing ‘homology CDSs’ or with another ‘*ab initio* CDS’ with the higher consensus reliability score.

Start site refinement

To determine better start sites among candidate starts, all the pair-wise FASTX alignments containing N-terminal residue match were analyzed for each ‘integrated

CDS'. If one or more candidate start sites aligned exactly with the N-terminal end of a library protein were found, the most abundant start site among them was assigned as the new shortest start.

Evaluation

To evaluate prediction accuracy, the predicted CDSs were compared with both the annotated and the overall data sets of 'public CDSs'. The annotated and the overall 'public CDSs' exclude and include hypothetical CDSs (CDSs with no functional annotation), respectively. The sensitivity and the specificity of our predictions were calculated in the following way:

$$sensitivity = \frac{TP}{TP + FN}, \quad specificity = \frac{TP}{TP + FP}$$

Here, TP is the number of true positives (found among 'public CDSs'), FN is the number of false negatives (missed from 'public CDSs'), and FP is the number of false positives (additionally predicted CDSs). To combine sensitivity and specificity, F-measure (the harmonic average of sensitivity and specificity) was calculated in the following way:

$$F - measure = 2 \times \frac{sensitivity \times specificity}{sensitivity + specificity}$$

Four different levels of evaluation based on the strictness of the start prediction were performed for (1) stop only (level 1, 'stop only'); (2) stop, frame change, and any candidate starts (level 2, 'candidate start'); (3) stop, frame change, and the length

coverage (90%) of the shortest start (level 3, 'start coverage'); and (4) stop, frame change, and the exact position of the shortest start (level 4, 'exact start') (Table 1).