

CONSORF CDS finding

CONSORF provides five kinds of predicted CDSs in FASTA and XML formats, based on their sources of evidence and “refinement levels” in prediction (Figure 1):

- (i) homology-based *consensus* CDSs (called ‘homology CDSs’),
- (ii) alternative homology-based CDSs, predicted from the overall best match without considering the *consensus* among hits (called ‘alternative CDSs’),
- (iii) algorithm-based *ab initio* consensus CDSs (called ‘ab initio CDSs’),
- (iv) CDSs from the integration of ‘homology CDSs’ and ‘ab initio CDSs’ over certain thresholds (called ‘integrated CDSs’), and
- (v) the representative final CDSs that have undergone the refinement of start codon positions via the analysis on the N-terminal residue matches of ‘integrated CDSs’ (called ‘representative CDSs’).

The main prediction method of CONSORF is based on two complementary approaches: homology-based and algorithm-based.

In the homology-based approach, each input genome is compared to every proteome set of well-annotated prokaryotic organisms with complete genome sequences, using the FASTX program of the FASTA package [2]. The FASTX program aligns a query DNA sequence with proteins in the library after the conceptual translation that reads through potential in-frame stop codons and allows frame-shifts between codons. Every FASTX

alignment is assigned boundary stop and start codon positions enclosing the complete aligned sequence between them (Figure 2). Over the available pair-wise FASTX comparisons with each proteome set, the numbers and the alignment scores (bit scores) of those FASTX alignments enclosed by the same boundary stop and start codon positions and containing the same in-frame stop codon and frame-shift positions, if any, are summed up to generate consensus reliability scores of the supported CDSs. From the CDSs with the highest consensus reliability score, every identified CDS is checked and removed if it overlaps on the genome with an CDS with the higher consensus reliability score, generating (i) 'homology CDSs'.

In addition to 'homology CDSs', based on the consensus among the FASTX comparisons with available proteome sets, (ii) 'alternative CDSs' are also provided by CONSORF. Their start and stop codon positions and in-frame stop codon and frame-shift positions, if any, are determined from the best FASTX alignment among the comparisons with available proteome sets. Contrary to 'homology CDSs' based on the consensus reliability score, 'alternative CDSs' are more sensitive to error propagation in CDS prediction and annotation, owing to the direct dependency on the best FASTX alignment.

As well as homology-based CDS finding, algorithm-based CDS finding is applied independently. The results of multiple *ab initio* prediction programs, currently including Glimmer, GeneMark, and GeneMark.hmm, are analyzed. Over the available

ab initio prediction results, the numbers and the nucleotide lengths of the predicted CDSs with the same stop and start codon positions are summed up to generate their consensus reliability scores. In this consensus scoring system, directly dependent on the CDS nucleotide lengths, the longer CDSs are less likely to be missed. Contrary to the homology-based approach, the CDSs overlapping on the genome with another CDS are all included, generating (iii) ‘*ab initio* CDSs’.

CONSORF also integrates two complementary and independent CDSs: ‘homology CDSs’ with high specificity and ‘*ab initio* CDSs’ with high sensitivity. ‘Homology CDSs’ with consensus reliability scores over a threshold are first located on a genome, and then ‘*ab initio* CDSs’ with consensus reliability scores over another threshold are located, avoiding significant positional overlap with existing ‘homology CDSs’ or with another ‘*ab initio* CDS’ with the higher consensus reliability score, generating (iv) ‘integrated CDSs’.

To provide the final (v) ‘representative CDSs’, CONSORF inspects every N-terminal residue match of ‘integrated CDSs’ to search for more reliable start codon positions. Contrary to ‘integrated CDSs’ with consensus start codon positions that conservatively enclose conserved regions supported by FASTX alignments, ‘representative CDSs’ rely on another consensus start codon positions that align exactly with the N-terminal end of a library protein. If one or more new start codon positions are found in the inspection of

every N-terminal residue match, and they exist among probable start codon positions (called candidate starts) located between the longest start codon position (called the longest start) from the stop codon of the same CDS and the consensus start codon position (called the shortest start) of 'integrated CDSs', the most abundant start codon position is assigned as a new 'shortest start' of the 'representative CDS' (Figure 2C). The automatically predicted CDSs are freely available in FASTA and XML formats from our website [1].