Bioinformatics Advance Access published July 4, 2016

Genome Analysis

ConsPred – a rule-based (re-)annotation framework for prokaryotic genomes

Thomas Weinmaier[1], Alexander Platzer[1], Jeroen Frank[2], Hans-Jörg Hellinger[1], Patrick Tischler[3] and Thomas Rattei[1,3*]

[1]Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

[2]Department of Microbiology, Radboud University Nijmegen, Nijmegen, The Netherlands

[3]Department of Genome Oriented Bioinformatics, Technische Universität München, Freising, Germany

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: The rapidly growing number of available prokaryotic genome sequences requires fully automated and high-quality software solutions for their initial and re-annotation. Here we present ConsPred, a prokaryotic genome annotation framework that performs intrinsic gene predictions, homology searches, predictions of non-coding genes as well as CRISPR repeats and integrates all evidence into a consensus annotation. ConsPred achieves comprehensive, high-quality annotations based on rules and priorities, similar to decision-making in manual curation and avoids conflicting predictions. Parameters controlling the annotation process are configurable by the user. ConsPred has been used in the institutions of the authors for longer than 5 years and can easily be extended and adapted to specific needs.

Summary: The ConsPred algorithm for producing a consensus from the varying scores of multiple gene prediction programs approaches manual curation in accuracy. Its rule-based approach for choosing final predictions avoids overriding previous manual curations.

Implementation and availability: ConsPred is implemented in Java, Perl, and Shell and is freely available under the Creative Commons license as a stand-alone in-house pipeline or as an Amazon Machine Image for cloud computing, see https://sourceforge.net/projects/conspred/

1 INTRODUCTION

During the two decades, since the sequencing of the first bacterial genomes, the annotation of prokaryotic genome sequences has become a routine task, going along with the rapid growth of public databases (Tatusova, et al., 2015). This trend will further speed up due to recent improvements in metagenomics, which enabled the reconstruction of near-complete genome sequences of low-abundant microbial community members (Albertsen, et al., 2013; Brown, et al., 2015; Callister, et al., 2010).

High-quality annotation of genome sequences is essential in all areas of genome research and most important for comparative genomics and functional genomics. Compared to automatic tools the manual annotation achieves significantly better results (Iliopoulos, et al., 2003). However, high costs and time constraints limit this strategy to few model organisms or the annotation of very unusual or taxonomically novel genomes (e.g., Spang, et al., 2012). The vast majority of genome sequences in public databases have been annotated using automatic software without expert curation.

A variety of software tools, which implement different gene and function prediction approaches, have been developed. The annotation of prokaryotic genomes stored in primary sequence archives is thereby based on methodological choices by the authors and the particular time of submission. Re-annotation of genome sequences is therefore crucial, e.g. in comparative genomics or re-sequencing projects, to avoid artifacts due to technical inaccuracies (public pipelines e.g. discussed in Siezen and van Hijum, 2010). Standardization of annotation strategies and re-annotation has now also been addressed by the NCBI RefSeq project (Tatusova, et al., 2015).

In addition to public resources, limited in capacity and flexibility (e.g., Aziz, et al., 2008; Markowitz, et al., 2014; Vallenet, et al., 2013), also locally applicable tools for the re-annotation of prokaryotic genomes are needed. Ideally, these should incorporate all relevant types of evidence for structural and functional annotation, such as i) intrinsic prediction of coding sequences, ii) homology-based prediction of coding sequences and their function, iii) structure/sequence-based prediction of non-coding RNA genes, and iv) specific prediction of complex features, such as CRISPR repeats. Several programs have been developed for this purpose (e.g., Kang, et al., 2007; Seemann, 2014), differing in their utilization of different evidence types and their strategy for decision-making.

The authors of CONSORF (Kang, et al., 2007) have demonstrated that consensus gene prediction using hierarchical rules, allows for fully automatic, high-accuracy identification of prokaryotic genes. For large-scale re-annotation of prokaryotic genomes, this concept needed to be further extended to non-coding genes, complex features, and functional annotation. Furthermore, according to the high computational costs of sequence-similarity based approaches, re-annotation software should be able to utilize high-performance or cloud computing facilities. For these purposes, we have developed ConsPred, which facilitated numerous of our genome sequencing and comparative genomics projects during the last years (e.g., Probst, et al., 2014). ConsPred is a flexible software framework for fully automatic, integrative and comprehensive (re-)annotation of prokaryotic genomes.

2 DESCRIPTION

In ConsPred multiple *ab initio* gene prediction tools can be applied (Table S1). For homology-based prediction, all open reading frames (ORFs) are extracted and compared to the NCBI nr (Coordinators, 2015) database of protein sequences or any other user-defined protein database. In order to reduce "Shadow ORF" artifacts and to avoid alignments only resulting from synteny with closely related genomes, a specific taxonomy filter is applied. This filter excludes all proteins from closely related taxa - by default up to the own genus level - from the homology search. Conserved ORFs are trimmed to the first possible start position upstream of the alignment and also screened for neighboring ORFs, sharing similarity to adjacent regions of the same database sequence, indicative of putative pseudogenes. All *ab initio* predictions and conserved ORFs are grouped by stop coordinate and strand. The consensus genes and their start positions are determined from these groups following default rules briefly summarized here: i) *ab initio* predicted starts overrule the start of conserved ORFs, ii) *ab initio* predictions are evaluated based on configurable ranks, representing previous knowledge about the accuracy of prediction methods, e.g. for specific G+C contents, iii) overlapping *ab initio* predictions with support by conserved ORFs overrule those without (Fig. S1; Note S1).

Some non-protein-coding elements (NCEs; Table S1) are known to never overlap with CDS. ConsPred considers rRNAs, tRNAs and CRISPR repeats as blocking and others, here ncRNAs, as non-blocking. Consensus CDS overlapping blocked regions are discarded to avoid misannotations (Tripp, et al., 2011). All consensus CDS that passed the filtering are functionally annotated. The gene name, protein product and EC number are determined from sequence similarity searches first in the manually curated UniProt/SwissProt database and second, for all CDS without SwissProt hits, in the UniRef90 database (UniProt, 2015). Significant hits, whose alignment covers both query and subject by at least 70%, are used for annotation transfer. Thereby ConsPred preferably transfers annotations from the small but manually curated Swiss-Prot database, but also makes use of the much higher coverage of TrEMBL. InterProScan (Mitchell, et al., 2015) is used to predict protein domains. Additionally, all consensus CDS are compared to the KEGG (Kanehisa, et al., 2014) and eggNOG (Powell, et al., 2014) databases and assignments to KEGG KO and EC numbers, KEGG pathways and eggNOG orthologous groups and their functional categories are exported (Fig. S1). A detailed description of the entire workflow and algorithm is given in Note S1.

3 APPLICATION

ConsPred runs on a single Linux computer or in a Linux grid computing system. The database files for ConsPred are updated monthly and can be downloaded from http://fileshare.csb.univie.ac.at/conspred_data/ (~60GB in 2015). ConsPred allows independent customization for each annotation run. Depending on the genome size and the computational resources the annotation takes few hours to several days without user intervention.

Table 1. Summary of annotations for *E.coli* K-12 MG1655

	RefSeq	ConsPred	Prokka	RAST
CDS/Avg. CDS length	4,140/317	4,747/291	4,305/314	4,509/301
rRNAs/tRNAs/ncRNAs	22/89/65	22/88/208	22/89/142	22/86/0
CRISPR	0	2	2	0
Domain annotations	44,464	37,049	597	0
EC number annotations	1,117	1,540	1,611	1,283
KEGG annotation	0	3,626	0	1,285
eggNOG annotations	0	4,188	0	0

The RefSeq accession of the *E.coli* K-12 MG1655 strain used in the overview is NC_000913; both ConsPred and Prokka were executed with default settings.

4 RESULTS

We used the well-curated genome of *Escherichia coli* K-12 MG1655 to demonstrate the ConsPred annotation compared to selected other annotation platforms (Table 1). The majority of genes are consistently annotated by all of the four platforms. Most pseudogenes are annotated in the RefSeq record, resulting from human curation. RAST and ConsPred additionally predict short genes, a consequence of their lower minimal ORF length setting. Only ConsPred and RAST provide KEGG annotations and ConsPred also includes eggNOG assignments. Table 1 indicates that ConsPred provides a comprehensive genome annotation regarding types of genetic elements and functional annotations (further details and data for additional genomes with varying G+C content and assembly completeness are shown in Tables S2 and S3).

5 CONCLUSION

ConsPred is primarily useful for the annotation of finished genome sequences or high-quality genome drafts (e.g., for submission to public databases), and for genome re-annotation in comparative genomics and functional genomics projects of prokaryotes. Furthermore, the customizability of the software framework makes ConsPred a valuable toolbox for evaluating and optimizing annotation strategies.

Conflicts of Interest: None declared.

REFERENCES

Albertsen, M., *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013;31(6):533-538.

Aziz, R.K., *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;9:75.

Brown, C.T., *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 2015;523(7559):208-211.

Callister, S.J., *et al.* Analysis of biostimulated microbial communities from two field experiments reveals temporal and spatial differences in proteome profiles. *Environ Sci Technol* 2010;44(23):8897-8903.

Coordinators, N.R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2015;43(Database issue):D6-17.

Iliopoulos, I., *et al.* Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* 2003;19(6):717-726.

Kanehisa, M., *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;42(Database issue):D199-205.

Kang, S., *et al.* CONSORF: a consensus prediction system for prokaryotic coding sequences. *Bioinformatics* 2007;23(22):3088-3090.

Markowitz, V.M., *et al.* IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* 2014;42(Database issue):D560-567.

Mitchell, A., *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 2015;43(Database issue):D213-221.

Powell, S., *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* 2014;42(Database issue):D231-239.

Probst, A.J., *et al.* Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat Commun* 2014;5:5497.

Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30(14):2068-2069.

Siezen, R.J. and van Hijum, S.A. Genome (re-)annotation and open-source annotation pipelines. *Microb Biotechnol* 2010;3(4):362-369.

Spang, A., *et al.* The genome of the ammonia-oxidizing Candidatus Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. *Environ Microbiol* 2012;14(12):3122-3145.

Tatusova, T., *et al.* Update on RefSeq microbial genomes resources. *Nucleic Acids Res* 2015;43(Database issue):D599-605.

Tatusova, T., *et al.* RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* 2015;43(7):3872.

Tripp, H.J., *et al.* Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res* 2011;39(20):8792-8802.

UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;43(Database issue):D204-212.

Vallenet, D., *et al.* MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res* 2013;41(Database issue):D636-647.