Construction of PsORFs algorithm

The small size has impeded recognition of sORFs. Traditionally, minimal length cutoff of 300 nucleotides is utilized to reduce the detecting false positives in gene prediction programs [[1](#References_Orr_etal_2020)]. However, due to lack of reliable training datasets, there are few prediction methods specially designed for sORFs.

# 1. Training datasets

In general, prokaryotic genomes with similar genomic GC content share similar protein coding genes sequence properties, while the protein coding gene sequence properties differ greatly in prokaryotic genomes with different GC contents, so most traditional gene prediction algorithms are usually species specific. Our earlier work has demonstrated that that some prokaryotic genomes exhibit high universal protein coding genes properties independent of their genome sizes and genomic GC contents [[2](#References_Yu_etal_2011)]. Using the protein coding genes in these genomes as training set, we can accurately predict the protein coding genes in other prokaryotic genomes. This phenomenon provides novel ideas for gene prediction. According to the genomic GC content distribution, we downloaded 61 prokaryotic genomes (**Supplementary Table 1**) with annotated sORFs from RefSeq database. These genomes can be divided into five intervals based on their genomic GC contents, that is, <30%, 30%–39%, 40%–49%, 50%–59% and ≥60%. Then, we build the candidate datasets according to following steps: (i) Excluding the redundant sequences in each dataset using the CD-Hit program [[3](#References_Huang_2010)] with the similarity threshold of 80% at DNA level; (ii) Excluding the sORF with sequence length more than 303 nt; (iii) Excluding the sORFs whose sequence length cannot be divisible by 3; (iv) Excluding the sORFs that do not end with a stop codon; (v) Excluding the sORFs with stop codon in its sequence; (vi) Excluding the sORF start with stop codon. Next, using the filtered sORFs in each genome as training set in turn to predict the sORFs in other genomes respectively, and the genome with the best prediction performance in each genomic GC content interval was selected as the candidate genome to constitute the final training set. Finally, five genomes (NC\_009089, NC\_003103, NC\_012962, NC\_000913, NC\_008380) were determined, the filtered function known sORFs in them are used to construct the positive training set.

Lack of reliable negative samples is the key problem in sORFs prediction works. Although the intergenic region and noncoding region were usually extracted as negative samples, but there is great possibility of the existence of sORFs in these regions, which can cause the increase of prediction false negative. Moreover, for prokaryotes, there are few noncoding and intergenic regions, hence it is difficult to construct negative samples. Random sequence generation has provided useful strategy in prokaryotic gene prediction works [[2](#References_Yu_etal_2011), [4](#References_Guo_2003)]. On the same time, considering a large proportion (about 50%) of sORFs do not initiate with canonical ATG start codons [[5](#References_CaoSlavoff_2020)], we generate the negative sequence by randomly disrupting corresponding positive sORF sequence more than 50 times, and ensure that the negative sequence share the same start codon and stop codon with the positive sequence.

# 2. Feature parameters

Because of the short sequence lengths, some commonly used feature parameters are not applicable for sORFs. Researches showed that for short sequences, many traditional research methods can get better prediction results [[6](#References_Weathers_etal_2004)]. In this paper, the codon usage frequency (*cof*) is employed as numerical paramter, which is described as below:

where *ci* is the number of *i*th type codon and *N* is the sequence length.

# 3. Model training and testing

In this work, we employed Random Forest as the core algorithm to train the PsORFs model. The number of bags was set as 200 according the evaluation result during K-fold cross validation. The five-fold cross validation was used to evaluate the model performance, the accuracy and *MCC* (threshold set as 0.5) of which are 89.25% and 0.7852, respectively. During each iteration of five-fold cross validation, 4/5 of the samples were used as training set and the rest of samples were uses as testing set. This process was repeated for 100 times to avoid bias from randomness of splitting training and testing dataset. In Table 1, the feature list in accompany with feature importance score is provided.

Table 1. The feature list in accompany with feature importance score.

|  |  |
| --- | --- |
| Codons | Feature importance score |
| AAC | 1.59 |
| CTT | 1.55 |
| CGG | 1.18 |
| AAA | 1.07 |
| CAC | 1.06 |
| CTC | 0.96 |
| ATT | 0.95 |
| GAC | 0.85 |
| GTG | 0.82 |
| TTT | 0.75 |
| GCA | 0.71 |
| AAT | 0.68 |
| GAT | 0.66 |
| ATA | 0.63 |
| TGG | 0.62 |
| AAG | 0.6 |
| TCT | 0.57 |
| CCG | 0.56 |
| TGT | 0.54 |
| GGC | 0.51 |
| ATG | 0.48 |
| CCC | 0.46 |
| GTA | 0.42 |
| ACG | 0.41 |
| CGA | 0.4 |
| ACC | 0.4 |
| GAA | 0.37 |
| AGA | 0.36 |
| CCA | 0.35 |
| CAG | 0.35 |
| GGT | 0.35 |
| CAT | 0.34 |
| GGA | 0.32 |
| GTT | 0.32 |
| TCC | 0.31 |
| TTC | 0.31 |
| CTG | 0.31 |
| ATC | 0.3 |
| AGC | 0.29 |
| GCT | 0.29 |
| TGA | 0.28 |
| CCT | 0.26 |
| ACT | 0.24 |
| CTA | 0.24 |
| TCG | 0.23 |
| TAT | 0.23 |
| AGT | 0.22 |
| GCG | 0.22 |
| GCC | 0.2 |
| TAA | 0.19 |
| TAC | 0.19 |
| TAG | 0.18 |
| CGC | 0.17 |
| CAA | 0.16 |
| GTC | 0.16 |
| TCA | 0.15 |
| TTA | 0.15 |
| TGC | 0.15 |
| GGG | 0.15 |
| CGT | 0.14 |
| AGG | 0.13 |
| TTG | 0.12 |
| GAG | 0.11 |
| ACA | 0.03 |

**References**

[1] Orr MW, Mao Y, Storz G, Qian S. Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Research. 2019; 48: 1029–1042.

[2] Yu J, Xiao K, Jiang D, Guo J, Wang J, Sun X. An integrative method for identifying the over-annotated protein-coding genes in microbial genomes. DNA Research. 2011; 18: 435–449.

[3] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010; 26: 680–682.

[4] Guo FB. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Research. 2003; 31: 1780–1789.

[5] Cao X, Slavoff SA. Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. Experimental Cell Research. 2020; 391: 111973.

[6] Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Letters. 2004; 576: 348–352.