Аннотация геномов de novo – Prokka and RAST

Скачайте питоновский ноутбук – prokka_rast.ipynb

Prokka
http://www.vicbioinformatics.com/software.prokka.shtml

Workflow

Protein coding genes are annotated in two stages.

Prodigal identifies the coordinates of candidate genes, but does not describe the putative gene product. The traditional way to predict what a gene codes for is to compare it with a large database of known sequences, usually at a protein sequence level, and transfer the annotation of the best significant match. Prokka uses this method, but in a hierarchical manner, starting with a smaller trustworthy database, moving to mediumsized but domain-specific databases, and finally to curated models of protein families. By default, an e-value threshold of 106 is used with the following series of included databases:

1) An optional user-provided set of annotated proteins. These are expected to be trustworthy curated datasets and will be used as the primary source of annotation. They are searched using BLASTþ blastp (Camacho et al., 2009).

2) All bacterial proteins in UniProt (Apweiler et al., 2004) that have real protein or transcript evidence and are not a fragment. This is 16 000 proteins, and typically covers 450% of the core genes in most genomes. BLASTþ is used for the search.

3) All proteins from finished bacterial genomes in RefSeq for a specified genus. This captures domain-specific naming, and the databases vary in size and quality, depending on the popularity of the genus. BLASTþ is used for this and is optional.

4) A series of hidden Markov model profile databases, including Pfam (Punta et al., 2012) and TIGRFAMs (Haft et al., 2013). This is performed using hmmscan from the HMMER 3.1 package (Eddy, 2011).

5) If no matches can be found, label as 'hypothetical protein'

RAST Annotation Server

https://rast.nmpdr.org/rast.cgi

(Rapid Annotation using Subsystem Technology)

Что делает RAST

This is the RASTtk Default Pipeline:

1. Calls rRNAs with a custom BLAST-based tool
2. Calls tRNAs with tRNAscan
3. Calls large repeat regions
4. Calls seleno proteins
5. Calls pyrrolysyl proteins
6. Finds Streptococcus repeat regions (only if the genus is Streptococcus)
7. Calls CRISPRs
8. Calls the protein-encoding genes with Prodigal and Glimmer3
9. Annotates protein-encoding genes with k-mers (version 2),
10. Annotates remaining hypothetical proteins with k-mers (version 1),
11. Attempts to annotate remaining hypothetical proteins by blasting against close relatives (if possible)
12. Performs a basic gene overlap removal

Теперь проделаем аннотацию с помощью RAST для этого можно

1. Воспользоваться веб сервером (требуетсяс регистрация, а она может долго подтверждаться)
2. **Установить myrast (рекомендуется)**
3. Установить rasttk (будут проблемы с установкой см. далее)

Eukaryotic Genome Annotations and Tools

https://www.1010genome.com/eukaryotic-genome-annotation/