Gene Prediction in de novo Metagenomic Assemblyes

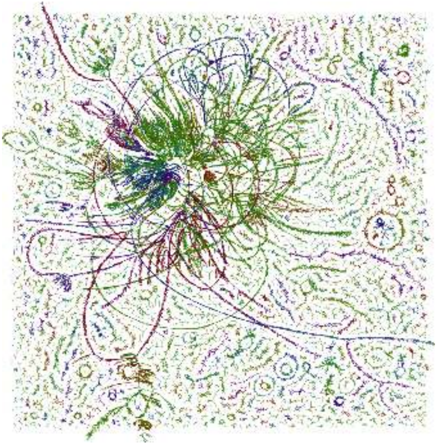
Ekaterina Sosa
Supervisor: Nikolay Vyahhi

December, 2012

Initial problem formulation

- to compare prokaryotic gene prediction tools on single-cell datasets
- to investigate metagenomics analysis methods
- to combine metagenomcs analysis and gene prediction

Introduction: about metagenomics



Meta-analysis is the study of statistically combining separate analyses.
Genomics is the study of comprehensive analysis of organisms genetic material.
Metagenomics is the study of genomic material obtained directly from the environment, instead of from culture.

Introduction: about prokaryotic gene prediction

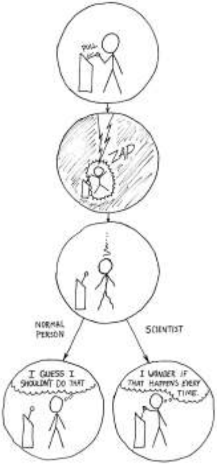
A lot of gene finders are based on Hidden Markov Model. Where transition matrix is selected by GC-content.

Tools

- GeneMark
- GeneMark-S
- GeneMark.hmm
- GLIMMER
- EasyGene

... with heuristic selection by

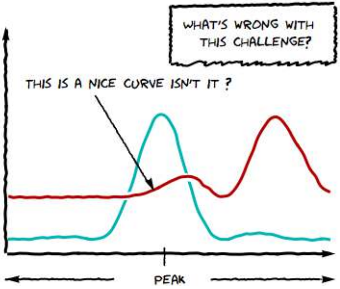
- certain GC-content for every contig
- total GC-content for assembly


ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

Comparison

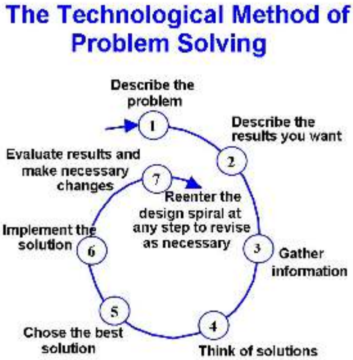
GeFComp

is a Python script for comparing performance metrics of different gene finding tools. In given a number of genome assemblies in FASTA format, GeFComp executes each tool on each of the genomes and evaluates Type I (false positives) and Type II (false negatives) errors.


WHAT'S WRONG WITH THIS CHALLENGE?
THIS IS A NICE CURVE ISN'T IT ?
PEAK

New problem formulation

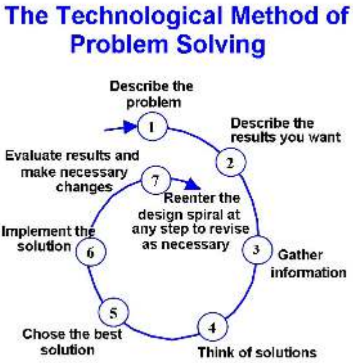
- to compare prokaryotic gene prediction tools on single-cell datasets
- **to embed the best tool into QUAST**
- to investigate metagenomics analysis methods
- to combine metagenomcs analysis and gene prediction


The Technological Method of Problem Solving
Describe the problem
1
Describe the results you want
2
Gather information
3
Think of solutions
4
Chose the best solution
5
Implement the solution
6
Evaluate results and make necessary changes
7
Reenter the design spiral at any step to revise as necessary

And the best tool was embedded into QUAST

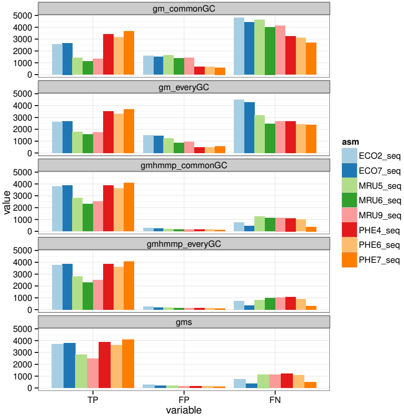
New new problem formulation

- to compare prokaryotic gene prediction tools on single-cell datasets
- to embed the best tool into QUAST
- **to embed eukariotic gene prediction into QUAST**
- to investigate metagenomics analysis methods
- to combine metagenomcs analysis and gene prediction


The Technological Method of Problem Solving
Describe the problem
1
Describe the results you want
2
Gather information
3
Think of solutions
4
Chose the best solution
5
Implement the solution
6
Evaluate results and make necessary changes
7
Reenter the design spiral at any step to revise as necessary

Current results

Hypothesis: gene was predicted
False **P**ositives is Type I error
False **N**egatives is Type II error
True **P**ositives is correct outcome
True **N**egatives is correct outcome
(TN $= \infty$)



`http://github.com/bioinf/GeneFinder`

Future directions

- to investigate metagenomics analysis methods
- to combine metagenomcs analysis and gene prediction
- to embed metagenomics gene prediction into QUAST
- to gene prediction tool