Introduction to exercise:
Genome analysis using Web Services

Tripod workshop on Perl, SQL and Web Services
November 16-20, 2009

Peter Fischer Hallin, November 19, 2009
Introduction to afternoon exercises

• We will provide all groups with different genbank accession numbers

• Each group will complete a series of genome analysis steps. Results are placed in a central MySQL table which we can monitor :-)

• You will predict protein coding genes (like you did this morning), tRNA+rRNA genes, plus the secreted proteins

• Finally, you will construct a graphical comparison of your proteome against those of all the other groups. This will allow you to identify regions where your genome is unique and where it shares homology with the other genomes.
Aim

- Genome Analysis: We have gathered a number of pre-compiled Web Services clients that you will be using.

- All scripts access CBS Web Services and will work outside of the CBS network.

- It is our vision that this will server as inspiration as to how Web Services can be embedded in workflows.

- How you can build stand-alone Web Services clients, that work exactly like local programs on your computer (getseq, prodigal etc)
Prokaryotic gene finding

- A number of prokaryotic gene finders exist, among others Glimmer3, GeneMark, EasyGene, Prodigal.

- Prodigal is a recent, and very speedy tool, which we have implemented at CBS as part of the Genome Atlas Database. Authored at Oak Ridge National Laboratory, http://compbio.ornl.gov/prodigal/.

- You will be using Prodigal to identify the protein coding genes

- Tested performance of Prodigal against public annotation

<table>
<thead>
<tr>
<th>source</th>
<th>CDS total</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>3’off</th>
<th>5’off</th>
<th>sens.</th>
<th>shared</th>
</tr>
</thead>
<tbody>
<tr>
<td>U000096 (present)</td>
<td>4,321</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>U000096 (2004)</td>
<td>4,254</td>
<td>4,172</td>
<td>82</td>
<td>109</td>
<td>1.02</td>
<td>-4.07</td>
<td>0.97</td>
<td>93%</td>
</tr>
<tr>
<td>Glimmer 3.02</td>
<td>4,476</td>
<td>4,174</td>
<td>302</td>
<td>125</td>
<td>-0.6</td>
<td>-24.09</td>
<td>0.97</td>
<td>87%</td>
</tr>
<tr>
<td>GeneMark-S 2.6</td>
<td>4,377</td>
<td>4,207</td>
<td>170</td>
<td>90</td>
<td>1.94</td>
<td>-20.17</td>
<td>0.98</td>
<td>91%</td>
</tr>
<tr>
<td>EasyGene 1.2</td>
<td>4,056</td>
<td>4,017</td>
<td>39</td>
<td>256</td>
<td>-0.28</td>
<td>-19.07</td>
<td>0.94</td>
<td>91%</td>
</tr>
<tr>
<td>Prodigal 1.1</td>
<td>4,332</td>
<td>4,200</td>
<td>132</td>
<td>97</td>
<td>0.54</td>
<td>-20.07</td>
<td>0.98</td>
<td>92%</td>
</tr>
</tbody>
</table>
rRNA genes

- The rRNA genes 5s, 16s, and 23s are central components of the ribosome.
- The ribosome is the translation machinery of the cell.
- Chromosomal position and operon copies influence the replication speed.
Prediction of rRNA genes: Alignment from secondary structure and construction of hidden markov models

Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T Ussery DW RNammer: consistent annotation of rRNA genes in genomic sequences Accepted for publication March 6 2007 in Nucleic Acids Research.

http://www.psb.ugent.be/rRNA/
tRNA genes

- tRNA are non-coding genes.
- In the translation phase, the tRNA structure directs the amino acid to match the corresponding codon during the translation.
- There is a corresponding tRNA gene to match the codons utilized by the organism.
SI DD - a way to identify promotor regions

• Stress Induced DNA duplex Destabilization - a mathematical model for estimating the free energy required to open the DNA helix. Measure in kcal/mol

• During transcription, the DNA molecule is wrapped around the RNA polymerase holoenzyme. The sigma factor recognizes more less conserved DNA motifs (e.g. the sigma-70, -35 and -10) When binding, it applies a stress (torque) to the DNA strand, causing it to pop open at what will be the transcription start site.

• To predict operons and transcript levels, this measure is relevant.

• **Visualization?**

GeneWiz / The Genome Atlas

- A way to represent numerical data and annotations pertaining to a complete chromosome

- Example, pX01 plasmid of *Bacillus anthracis*
SIDD near the rRNA operons of *E. coli* K12
The BLASTatlas method - comparing multiple genomes

The BLASTatlas method - comparing multiple genomes

Fig. 1 Mapping of protein–protein alignment to DNA. **Panel A**: mismatches and perfect matches are assigned a score of 0 and 1, respectively. Conservative mismatches are assigned a score of 0.5. In the case of DNA alignment, only scores of 0 and 1 are possible. **Panel B**: gaps in the database sequence will be rendered as being non-conserved areas (filled with zeros). **Panel C**: gaps in the reference sequence will be neglected, since they have no corresponding region in the reference genome into which they can be mapped.

downside: gaps in reference are ignored

BLASTatlas of 24 *Burkholderia* genomes

B. pseudomallei 1710b, chr II
3,181,762 bp

B. ambifaria AMMD

B. ambifaria MC40-6

B. cenocepacia AU 1054

B. cenocepacia HI2424

B. cenocepacia J2315

B. cenocepacia MC0-3

B. glumae BGR1

B. mallei ATCC 23344

B. mallei NCTC 10229

B. mallei NCTC 10247

B. mallei SAVP1

B. multivorans ATCC 17616

B. multivorans ATCC 17616

B. phytofirmans PsJN

B. pseudomallei 1106a

B. pseudomallei 668

B. pseudomallei K96243

B. sp. 383

B. thailandensis E264

B. vietnamiensis G4

B. xenovorans LB400

W) Annotations:

- CDS +
- CDS -
- rRNA
- tRNA
workflow

accession no.

getseq.pl

queryGenomes
Translation accession into organism name

getSIDD.pl
prodigal.pl
trnascan.pl
mammer.pl

your proteome

BLASTatlas
Proteome comparison

proteomes of other groups
Required resources

- Genome Atlas: http://www.cbs.dtu.dk/ws/GenomeAtlas. trnascanRun, getSeq, queryGenomes, runProdigal
- SIDDbase: http://www.cbs.dtu.dk/ws/SIDDbase. getSIDD
- RNAmmer: http://www.cbs.dtu.dk/ws/RNAmmer, runService