Gene expression & physiology

- MIAME-like standards for environmental gene expression?
  (The MIAME Checklist: Experimental Design, Samples Used, Extract Preparation & Labeling, Hybridization Procedures and Parameters, Measurement Data and Output, Array Design) (see http://www.mged.org/miame)
- Culture collections of relevant microbes
- Standard 'chips', internal controls for environmental monitoring, cross-qpcr standardization, etc.
- SgdB-like model for integrating heterogeneous datasets?
- Cross comparison/intercalibration gene x exper. (qpcr, etc.)

Proteomics & biochemistry

- Genome & peptide sequence databases & polymorphisms
- Accurate mass tag databases & experimental data
- Samples, vouchers, and antibodies
- Coordinated gene expression & proteomic studies

Dealing with 'Impedance mismatch'

- Data assimilation, analysis, archiving & integration
  (Contemporary Biological (and Oceanographic) Science is largely Information Science!)
- Field verification, process measurement & quantification
  (Beyond in silico Bioinformatics and Towards Environmetal Quantitative Biology)
- Instrumentation/methods development - benchtop/in situ
  (Make New Instruments, Measure New Things - the challenge of in situ measurement)
- Scalar and disciplinary integration (the cultural gap)
  (Earth Systems Science is Life Systems Science - better cross-talk required!)

SEQIUNCE DATA & BIOINFOMATICS

SEQUEENCE DATA TYPE
(What exactly are you looking at?)
- PCR amplicon, "Metagenome assembly", BAC sequence 454 pyrosequence reads, etc...

SAMPLE METADATA
(Context is everything!)
- Sample type, collection method, physics, chemistry, biology

DATA STANDARDS
- Quality, voucher availability, "ocean gene ontologies, MIGS/MAMS?"

DATABASE STRUCTURE/ACCESSIBILITY
- Genomic, proteomic, environmental, central vs. distributed, linkout, federated databases, etc...

AVAILABILITY/ACCESSIBILITY OF ANALYTICAL TOOLS
- Genomic/proteomic, environmental, polymorphism, metagenome analyses, data cross comparisons

How do you make sense of this ????????
Content Sensor

- **Extrinsic Content Sensors**: Local alignment, BLAST
  - Sequence from SWISSPROT, cDNA, EST
  - Intra- and inter-genomic similarity
  - Depends on quality of database

- **Intrinsic Content Sensors**: hexamer count

Gene Prediction Tools

- GENSCAN/Genome Scan
- TwinScan
- Glimmer
- GenMark
- Critica

ORF = open reading frame = start codon...[GATC]...stop codon
CDS = coding sequence (produces the actual protein/RNA species)

Softberry FGENESB annotation "pipeline":  http://softberry.com/berry.phtml

1. Finds all potential ribosomal RNA genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected RNA genes.
2. Predicts tRNA genes using tRNAscan-SE program (Washington University) and masks detected tRNA genes.
3. Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes. Generates 5th-order in-frame Markov chains for coding regions, 2nd-order Markov models for region around start codon and upstream RBS site, stop codon and probability distributions of ORF lengths.
4. Predicts operons based only on distances between predicted genes.
5. Runs BLASTP for predicted proteins against COG database, cog.pro.
6. Uses information about conservation of neighboring gene pairs in known genomes to improve operon prediction.
7. Runs BLASTP against NR for proteins having no COGs hits.
8. Predicts potential promoters (BPROM program) or terminators (BTERM) in upstream and downstream regions, correspondingly, of predicted genes.
9. Refines operon predictions using predicted promoters and terminators as additional evidences.

Typical softberry output

Expect Value (E)  (Karlin-Altschul Statistics)

\[ E = K m e^{-\lambda S} \]

or

\[ E = m n^2 S' \]

\[ S' = \text{bitscore} - (0.5 - \ln K)/\ln 2 \]

BLAST Executables & Programs

Executables:
- blastall, megablast, blastpgp, b2seq, blastclus

Blastall programs:
- blastp, blastn, blastx, tblastn, tblastx

Bare minimum for blastall:
```
./blastall -p [program] -i [fasta file] -d [database] -o [output]
```

Several different BLAST programs:

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastp</td>
<td>Compares an amino acid query sequence against a protein sequence database.</td>
</tr>
<tr>
<td>blastn</td>
<td>Compares a nucleotide query sequence against a nucleotide sequence database.</td>
</tr>
<tr>
<td>blastx</td>
<td>Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.</td>
</tr>
<tr>
<td>tblastn</td>
<td>Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.</td>
</tr>
<tr>
<td>tblastx</td>
<td>Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.</td>
</tr>
</tbody>
</table>
KEGG  
Kyoto Encyclopedia of genes and genomes  
http://www.genome.jp/kegg/

Prochlorococcus MED4  
Carbon fixation

http://www.moore.org/microgenome/

J. Craig Venter  
https://research.venterinstitute.org/moore/

http://www.megx.net/

MARINE MICRO SPECIFIC  
https://research.venterinstitute.org/moore  
http://www.moore.org/microgenome/  
http://egg.umb.es/micromar/  
http://www.megx.net/

GENERIC TOOLS AND MICROBIAL GENOME EXPLORATION  
http://genome.jgi-psf.org/mic_home.htm  
http://www.softberry.com/all.htm  
http://www.ncbi.nih.gov/  
http://img.jgi.doe.gov/cgi-bin/pub/main.cgi  
http://img.jgi.doe.gov/cgi-bin/m/main.cgi
### FGENESB Suite of Bacterial Operon and Gene Finding Programs

FGENESB automatic annotation of bacterial and archaeal genomes. The FGENESB gene algorithm is based on Markov chain models of coding regions and translation and termination sites.

**Features**

- Automatic training of gene finding parameters for new bacterial genomes using only genomic DNA as an input (optionally, pre-learned parameters from related organisms can be used)
- Mapping of rRNA and tRNA genes
- Highly accurate Markov chains-based gene prediction
- Prediction of promoters and terminators
- Operon prediction based on distances between ORFs and frequencies of different genes neighboring each other in known bacterial genomes, as well as on promoter and terminator predictions
- Automatic annotation of predicted genes by homology with COG and NR databases.
- FGENESB gene prediction algorithm is based on Markov chain models of coding regions and translation and termination sites.

### Typical softberry output

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>Stop</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>15373</td>
<td>15372</td>
<td>Promoter</td>
</tr>
<tr>
<td>2</td>
<td>15373</td>
<td>15432</td>
<td>60</td>
<td>CDS</td>
</tr>
<tr>
<td>3</td>
<td>15432</td>
<td>15457</td>
<td>24</td>
<td>Stop codon and probability distributions of ORF lengths.</td>
</tr>
<tr>
<td>4</td>
<td>15457</td>
<td>15477</td>
<td>20</td>
<td>2nd-order Markov models for region around start codons and upstream RBS site,</td>
</tr>
<tr>
<td>5</td>
<td>15477</td>
<td>15503</td>
<td>26</td>
<td>Stop codon and probability distributions of ORF lengths.</td>
</tr>
</tbody>
</table>

---

**Softberry FGENESB annotation “pipeline”**

http://softberry.com/berry.phtml

**STEP 1.** Finds all potential chromosomal genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected rRNA genes.

**STEP 2.** Predicts rRNA genes using tRNAscan-SE program (Washington University) and masks detected rRNA genes.

**STEP 3.** Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes.

**STEP 4.** Predicts operons based only on distances between predicted genes.

**STEP 5.** Refines operon predictions using predicted promoters and terminators as additional evidences.

---

**Softberry FGENESB annotation “pipeline”**

http://softberry.com/berry.phtml

**STEP 1.** Finds all potential chromosomal genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected rRNA genes.

**STEP 2.** Predicts rRNA genes using tRNAscan-SE program (Washington University) and masks detected rRNA genes.

**STEP 3.** Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes.

**STEP 4.** Predicts operons based only on distances between predicted genes.

**STEP 5.** Refines operon predictions using predicted promoters and terminators as additional evidences.

---

**Softberry FGENESB annotation “pipeline”**

http://softberry.com/berry.phtml

**STEP 1.** Finds all potential chromosomal genes using BLAST against bacterial and/or archaeal rRNA databases, and masks detected rRNA genes.

**STEP 2.** Predicts rRNA genes using tRNAscan-SE program (Washington University) and masks detected rRNA genes.

**STEP 3.** Initial predictions of long ORFs that are used as a starting point for calculating parameters for gene prediction. Iterates until stabilizes.

**STEP 4.** Predicts operons based only on distances between predicted genes.

**STEP 5.** Refines operon predictions using predicted promoters and terminators as additional evidences.