Functional Annotation & Beyond
with Integrated Microbial Genomes (IMG)

Rekha Seshadri (JGI)
ECOGEO Annotation Module
Functional Annotation: Sequence to Biology

- “Feature prediction” - location of genes and other features, such as noncoding (e.g., rRNA, tRNA), protein coding sequences (CDS), and more (e.g., regulatory sites, repeats, frameshifts)
- Next, an attempt to “name” and interpret function/role by comparing against functional “databases”
  - NCBI Guidelines on “names” (~20 rules) e.g.,
  - “concise name, not a description or phrase”
  - “The protein name should not contain specific characteristics of the protein, (e.g., subcellular location, domain structure…”)

- Archives of accumulated biological data and knowledge
  - Genome, gene sequences, mutations
  - Gene regulation, expression, splice variants
  - Protein sequence, post-translational modifications
  - Protein tertiary structure, localization, networks
  - Enzyme kinetics, metabolites, metabolic networks
  - E.g., nr, swissprot, pdb, Pfam
• Similarity is the primary predictor of homology, which is the predictor of function (*sort of*)
• How to search for sequence similarity? (*Sensitivity versus speed*)

GenBank surpasses one trillion total bases of publicly available sequence data  Thursday, January 22, 2015
Search Method: Pairwise

- Find related or similar sequences by mapping letters of two sequences, with some spacers (indels),

```
76  GGMLKPIEGGTYEVNEAMVEDLKIGVQGPHASNLGGILSNEIAKEIGKRAFIIVDPVVDE 135
     |                                                               |
61  GGMLKPIEGGTYEVNEAMVEDLKIGFEGPHAXNLGGILSNEIAKKGKRAFIIVDPVVVDX 120
```

- Parse similarities, determine “best hit”

- Examples of pairwise search tools - Basic Local Alignment Search Tool or BLAST, LAST, etc.
Sequence databases

- Non-redundant database (nr), Refseq nr (curated), Uniref
- Kyoto Encyclopedia of Genes and Genomes (KEGG) – integrates functional information into biological pathways
  - Genomic, chemical (compound, reaction, enzyme), systems (pathway), health (disease, drug)
  - 10,371 KO terms
  - Pathway database provides maps, e.g.,
Search Method: Profiles/Models

- MSA of known sequences – detect “regions of similarity” – build consensus
- Use structural and mechanistic information (catalytic sites)
- Generate profiles using Hidden Markov Models (HMMs) or Position-Specific Scoring Models (PSSMs)
- More sensitive than pairwise – detect distant relationships
- Example of profile/HMM search tools: RPS-BLAST, Hmmer
Search Method: Profile based

- Collect “seed” proteins
- Generate & Trim Alignment
- Generate Profile with HMM or PSSM
- Search New Model against all proteins

Region of good alignment and closest similarity

Compute statistical probabilities for amino acid patterns in the seed

Choose “noise” and “trusted” cutoff scores based on “known” versus “unknown” protein scores
Profile/HMM databases

- COGs – protein clusters from at least 3 complete prokaryotic genomes
  - Manually curated, full length proteins
  - 4631 COGs from 711 A & B genomes.
- Tigrfam – manually curated collection (4,488) relatively full-length multiple sequence alignments for annotation
- Pfam – large (16,230), widely used, curated collection of protein families & domains
- FOAM – HMMs trained on alignments of target KO terms

“Local” versus “Global” HMMs (e.g., Tigrfam “definitions”)
- Equivalog – full-length, all members share the same function
- Superfamily – full-length similarity, same domain architecture, but not same function
- Domain – a shorter region of homology (10-100 residues), seq similarity with or without associated function (e.g., ATP-binding site)
- Note: Pfam has different “definitions” or classifications
- Limited organization into functional hierarchies or classification systems
Homology ≠ similarity of function

- No scoring scheme provides “biological truth”
  - Any pair of sequences can be aligned
  - Finding meaning is up to you!
- Other heuristics – context-based, best reciprocal hit

![Diagram of homologs, orthologs, paralogs, and gene duplication](image-url)
Homology ≠ similarity of function

- **Paralogs problem**: Template is a paralog, more likely have diverged functionally.

- **Moonlighting problem**: Template may have more than one function.

- **Multi-domain proteins problem**: Template annotation may be based on a non-matching domain.

- **Database mis-annotations problem**: Template is mis-annotated, e.g. by homology with a multi-domain protein (see C).

Punta & Ofran. PLOS Comp Biol. 2008
Errors in databases

There are many errors in the primary sequence databases:

• In the sequences themselves:
  • sequencing errors.
  • cloning vectors sequences.
• Gene Calling errors
• In the annotations:
  • “genome rot” - inaccuracies, omissions, mistakes.
  • inconsistencies between some fields.

• List of databases:
  http://www.oxfordjournals.org/our_journals/nar/database/c
When there is “no similarity”? 

- Gene Context
- Sub-cellular localization
- Topological features
- Prediction of binding residues
HANDS-ON (25 minutes)

• **Pairwise - BLAST**

  **Objective** – determine what manner of carbon fixation pathway is favored by a “binned” genome recovered from a TARA metagenome
  
  – Generate BLAST database from multi-fasta
  
  – Use various cutoff and output options

• **HMMsearch**

  **Objective** – Identify universally conserved ribosomal protein markers* in a cyanobacterial genome recovered from a TARA metagenome
  
  • Use specific options (cutoff and display)
  
  • Review output

*16 markers used by Hug et. al. Nature Microbiology 1: 16048 (2016)
Performing Searches Locally

- **BLAST (pairwise)**
  - Objective – determine what manner of carbon fixation pathway is favored by a “binned” genome recovered from a TARA metagenome
  - Generate BLAST database from multi-fasta
  - Use various cutoff and output options

- **HMM search**
  - **Objective** – Identify universally conserved ribosomal protein markers* in a cyanobacterial genome recovered from a TARA metagenome
  - Use specific options (cutoff and display)
  - Review output

*16 markers used by Hug et. al. Nature Microbiology 1: 16048 (2016)
Performing Searches Locally

- Databases utilize both pair-wise similarity and/or HMMs to detect sequence matches
- Possible to perform these searches locally
- Why?
  - Custom made databases can undergo higher degrees of curation
  - A small search database will run quickly - no need to wait for searches against large databases
BLAST (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences (Wikipedia).

Perform Searches Locally

The following hands-on exercises utilize a genomic bin from the TARA Ocean Project data set from the Mediterranean Sea:

- Putative taxonomy → Cyanobacteria
- 35 contigs
- 1,585 putative CDS (as determined by Prodigal)
- Approx. 64.64% complete (1.29% redundancy)
Local BLAST Database – 1

Compare tara_med_examplegenhome protein sequences to a custom collection of carbon fixation related genes (downloaded from KEGG)

Collection contains marker genes for: CBB, Wood-Ljungdahl, reductive TCA, 3-hydroxypropionate, 3-hydroxypropionate/4-hydroxybutyrate

As a cyanobacteria – which carbon fixation pathway is utilized?
Local BLAST Database – 2

Create a BLAST index of ‘subject’ sequences

$ makeblastdb -in carbonfixation_markergenes.faa -dbtype prot

Creates 3 index files that end in *phr, *pin, *psq

Next, compare the tara_med_examplegenome sequences to collection
BLAST has multiple output options

-outfmt

<String>

alignment view options:

0 = pairwise,
1 = query-anchored showing identities,
2 = query-anchored no identities,
3 = flat query-anchored, show identities,
4 = flat query-anchored, no identities,
5 = XML Blast output,
6 = tabular,
7 = tabular with comment lines,
8 = Text ASN.1,
9 = Binary ASN.1,
10 = Comma-separated values,
11 = BLAST archive format (ASN.1)
Local BLAST Database – 4

BLAST – standard output format

$ blastp -query tara_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -out temp_output_file -evalue 1e-20 -num_descriptions 5 -num_alignments 5

Set minimum limit of E-value match and maximum limit for number of print matches and alignments

Query= 119286_61
Length=472

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Bits)</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>syg:sync_1967 cbbL; ribulose bisphosphate carboxylase, large su...</td>
<td>860</td>
<td>0.0</td>
</tr>
<tr>
<td>tni:TVH_2992 cbbL [H]; ribulose-1,5-bisphosphate carboxylase/...</td>
<td>777</td>
<td>0.0</td>
</tr>
<tr>
<td>tti:THIT_12370 rbcL; ribulose bisphosphate carboxylase (EC:4....</td>
<td>773</td>
<td>0.0</td>
</tr>
<tr>
<td>tvr:TVD_09485 rbcL; ribulose 1,5-bisphosphate carboxylase (EC:4...</td>
<td>755</td>
<td>0.0</td>
</tr>
<tr>
<td>tgr:Tgr7_3203 Ribulose-bisphosphate carboxylase (EC:4.1.1.39); ...</td>
<td>754</td>
<td>0.0</td>
</tr>
</tbody>
</table>

$less +11096 temp_output_file

> syg:sync_1967 cbbL; ribulose bisphosphate carboxylase, large subunit (EC:4.1.1.39); K01601 ribulose-bisphosphate carboxylase large chain [EC:4.1.1.39] (A)
Length=470

Score = 860 bits (2221), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 428/470 (91%), Positives = 434/470 (92%), Gaps = 0/470 (0%)

Query 1

MSKKYDAGKEYRDTYWTDPYVPLDSDLLACFKCKGXXGVPKEEVAAAVAESXTGTWSX 60

Sbjct 1

MSKKYDAGKEYRDTYWTDPYVPLDSDLLACFKCKGXXGVPKEEVAAAVAESXTGTWSX 60
Local BLAST Database – 5

BLAST – tabular output (fmt = 6)

Lots of custom format options for formats 6, 7, 10

qseqid means Query Seq-id
qgi means Query GI
qacc means Query accession
qaccver means Query accession.version
qlen means Query sequence length
sseqid means Subject Seq-id
sallseqid means All subject Seq-id(s), separated by a ';', sgi means Subject GI
sallgi means All subject GIs
sacc means Subject accession
saccver means Subject accession.version
sallacc means All subject accessions
slen means Subject sequence length
qstart means Start of alignment in query
qend means End of alignment in query
sstart means Start of alignment in subject
send means End of alignment in subject
$ blastp -query tara_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -out BLAST_output.tab -evalue 1e-20 -max_target_seqs 10 -outfmt '6 qseqid qstart qend sseqid slen sstart send bitscore pident evalue'

$ less BLAST_output.tab

Query ID, Query Start, Query End, Subject ID, Subject Length, Subject Start, Subject End, Bit Score, Percent Identity, E-value

Use a filter to find “real” matches
Local BLAST Database – 7

Cutoff 50% sequence identity

$ awk '{if ($9>=50) print }' BLAST_output.tab

With result sorting

$ awk '{if ($9>=50) print }' BLAST_output.tab | sort -nrk 9,9

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Query Start</th>
<th>Query End</th>
<th>Subject ID</th>
<th>Subject Length</th>
<th>Subject Start</th>
<th>Subject End</th>
<th>Bit Score</th>
<th>Percent Identity</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>119286_01</td>
<td>470</td>
<td>470</td>
<td>syg:sync_1967</td>
<td>470</td>
<td>470</td>
<td>800</td>
<td>21.00</td>
<td>77.10</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>3</td>
<td>409</td>
<td>tni:TNIR_2992</td>
<td>471</td>
<td>470</td>
<td>777</td>
<td>81.16</td>
<td>77.99</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>3</td>
<td>409</td>
<td>tli:TNIR_12370</td>
<td>471</td>
<td>470</td>
<td>773</td>
<td>81.16</td>
<td>77.99</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>2</td>
<td>409</td>
<td>tgr:Trf_1283</td>
<td>473</td>
<td>470</td>
<td>754</td>
<td>77.99</td>
<td>77.99</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>2</td>
<td>470</td>
<td>tvi:TVD_0485</td>
<td>474</td>
<td>472</td>
<td>755</td>
<td>77.99</td>
<td>77.99</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>2</td>
<td>409</td>
<td>txc:Trc_0838</td>
<td>471</td>
<td>470</td>
<td>716</td>
<td>77.99</td>
<td>77.99</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>2</td>
<td>470</td>
<td>tkn:TKI_0850</td>
<td>470</td>
<td>471</td>
<td>754</td>
<td>77.99</td>
<td>77.99</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>2</td>
<td>409</td>
<td>sdr:SCD_02031</td>
<td>473</td>
<td>470</td>
<td>724</td>
<td>76.45</td>
<td>76.45</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_01</td>
<td>3</td>
<td>409</td>
<td>afr:AFE_1691</td>
<td>473</td>
<td>470</td>
<td>724</td>
<td>76.23</td>
<td>76.23</td>
<td>0.0</td>
</tr>
<tr>
<td>119286_00</td>
<td>7</td>
<td>113</td>
<td>tni:TNIR_2993</td>
<td>114</td>
<td>113</td>
<td>166</td>
<td>69.16</td>
<td>36.55</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Query ID, Query Start, Query End, Subject ID, Subject Length, Subject Start, Subject End, Bit Score, Percent Identity, E-value

Find what we are looking for:

$ grep ‘syg:sync_1967’ carbonfixation_markergenes.faa
HMMER is used for searching sequence databases for sequence homologs and for making sequence alignments. Uses probabilistic models called profile hidden Markov models (profile HMMs)

http://pfam.xfam.org/family/PF02189/logo_image
Local HMM Database – 1

Part of the tool HMMER – perform searches

Also build new HMM models

$ hmmbuild --amino -informat afa <HMM OUTFILE NAME> <ALIGNMENT FILE>

Requires an aligned FASTA file of target sequences

Create your own HMMs: collect genes with identical function \(\rightarrow\) align (ex., using MUSCLE) \(\rightarrow\) export as an aligned fasta (afa)
Several HMM based databases – Pfam, TIGRfam, FOAM

Search tara_med_examplegenomene using an HMM database for the 16 ribosomal marker proteins used to construct Hug et al (2016) Tree of Life

HMM – hug_ribosomalmarkers.hmm

Utilizes a mixture of Pfam and TIGRfam models to identify targets in a genome
Local HMM Database – 3

$ hmmsearch --tblout HMM_output.tab --cut_tc --notextw hug_ribosomalmarkers.hmm tara_med_examplegenome.orfs.faa

--cut_tc = controls the threshold of match “trusted cutoff”

--notextw = formatting option

$ less HMM_output.tab

--- full sequence --- best 1 domain --- domain number estimation
# target name accession query name accession E-value score bias E-value score bias exp reg clu ov env dom rep inc

123536_9 - RpL5 PF00281.14 8.6e-23 76.7 0.5 1.5e-22 76.0 0.5 1.4 1 0 0 1 1 1 1
140943_11 - RpL3 PF00297.17 2.3e-36 122.3 0.8 1.9e-34 116.1 0.8 2.0 1 1 0 1 1 1 1
123536_11 - RpL6 PF00347.18 1.5e-35 118.4 2.2 1.7e-18 63.7 0.1 2.1 2 0 0 2 2 2 2
123536_6 - RpS17 PF00366.15 4.2e-29 97.2 2.8 4.9e-29 97.0 2.8 1.0 1 0 0 1 1 1 1
123536_10 - RpS8 PF00410.14 1.1e-45 151.5 0.1 1.2e-45 151.3 0.1 1.0 1 0 0 1 1 1 1
140943_12 - RpL4 PF00573.17 1.7e-55 184.2 0.3 1.9e-55 184.0 0.3 1.0 1 0 0 1 1 1 1
123536_12 - RpL18 TIGR00060 5.1e-42 139.4 1.0 5.7e-42 139.2 1.0 1.0 1 0 0 1 1 1 1
123536_3 - RpS3 TIGR01009 4.6e-77 255.0 1.7 5.4e-77 254.8 1.7 1.0 1 0 0 1 1 1 1
123536_2 - RpL22 TIGR01044 8.9e-39 128.4 0.2 1e-38 128.2 0.2 1.0 1 0 0 1 1 1 1
140943_16 - RpL22 TIGR01044 1.4e-38 127.8 0.2 1.6e-38 127.6 0.2 1.0 1 0 0 1 1 1 1
CASE STUDY: Organohalide Respirers

- Major groundwater contaminant – PCE, TCE (industrial degreaser)
- Electron acceptor for *anaerobic dehalorespiration* – incomplete however
- **Serial dechlorination by** reductive dehalogenases (RD) **with** cobalamin cofactor

![Chemical diagram showing dechlorination process]

- Known human carcinogens
  - *Dehalococcoides mccartyi* strain 195 from Ithaca Sewage Treatment Plant - **COMPLETE dechlorination** *(Science 1997: 276(5318):1568-1571)*
Integrated element

Genome Size: 1.46 Mbp
- Streamlined genome
- 13.6% integrated “islands”
- Dedicated dehalorespirer
Objective: Compare 195 vs CBDB1

<table>
<thead>
<tr>
<th>Strain 195</th>
<th>Strain CBDB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.467 Mbp</td>
<td>1.395 Mbp</td>
</tr>
<tr>
<td>Chloroethylenes, etc</td>
<td>Chlorobenzenes, etc</td>
</tr>
<tr>
<td>PCE-&gt;TCE-&gt;DCE-&gt;VC-&gt;Eth</td>
<td>PCE-&gt;TCE-&gt;DCE</td>
</tr>
<tr>
<td>17(+2) RDs</td>
<td>32 RDs</td>
</tr>
</tbody>
</table>

Use the Functional Annotation IMG Protocol to answer the following:

- Is there any synteny?
- What is unique to Strain 195?
- Can we find the RDs responsible for the terminal step(s)?
- Is de novo synthesis of cobalamin cofactor possible?
Tree of RDs from 195 and CBDB1

VC->Eth

PM
### Objective: Compare str 195 vs CBDB1

<table>
<thead>
<tr>
<th></th>
<th>Strain 195</th>
<th>Strain CBDB1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.467 Mbp</td>
<td>1.395 Mbp</td>
</tr>
<tr>
<td>chloroethylenes, etc</td>
<td>chlorobenzenes, etc</td>
<td></td>
</tr>
<tr>
<td>PCE-&gt;TCE-&gt;DCE-&gt;VC-&gt;eth</td>
<td>PCE-&gt;TCE-&gt;DCE</td>
<td></td>
</tr>
<tr>
<td>17(+2) RDs</td>
<td>32 RDs</td>
<td></td>
</tr>
</tbody>
</table>

- **Is there any synteny?**
  - Yes, extensive, some rearrangement involving RDs, several gaps
- **What proportion of genes are unique to strain 195?**
  - ~20% including RDs and Nitrogen Fixation
- **Can we find the RDs responsible for the terminal step(s)?**
  - Yes, TceA, for example
- **Is de novo biosynthesis of cobalamin cofactor likely?**
  - No, salvage of an intermediate is likely
Metagenome comparisons yield new insights

- **Carbon Fixation**
- **Aromatic HC degradation**

Not just chemoautotrophs, but potentially chemoheterotrophs!

Problems with Metagenome Comparisons

- IMG provides suite of tools for metagenome comparisons
- And ‘metadata’
- Caveats
  - Different sequencing platforms – Sanger, 454, Illumina
  - Different qualities of assembly and annotation
  - Many unpublished datasets - extensive metadata or particulars not available
  - Read data/coverage not available for non-JGI metagenomes

- JGI IMG Workshops  http://mgm.jgi.doe.gov/
Submission to IMG

- [https://img.jgi.doe.gov/cgi-bin/submit/main.cgi](https://img.jgi.doe.gov/cgi-bin/submit/main.cgi)
- ASSEMBLED DATA ONLY
- E.g., Metagenomes – pre-processing, feature prediction, functional annotation
Minimum Info about any Sequence (MIxS)

- Metadata sparsely annotated in most db
- **Standard of reporting** (GSC) - requires contextual data to be deposited at the time of sequence submission
- Set of core descriptors for genomes and metagenomes, transcriptomes (now SAGs, GFMs?)
- Genomes Online Database (GOLD) integrated within IMG adopts MIxS specifications and beyond

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property (SiGS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS-6</td>
<td>Habitat</td>
</tr>
<tr>
<td>MIGS-6.3</td>
<td>Salinity as% NaCl (w/v)</td>
</tr>
<tr>
<td>MIGS-22</td>
<td>Oxygen requirement</td>
</tr>
<tr>
<td>MIGS-15</td>
<td>Biotic relationship</td>
</tr>
<tr>
<td>MIGS-14</td>
<td>Pathogenicity</td>
</tr>
<tr>
<td>MIGS-4</td>
<td>Geographic location</td>
</tr>
<tr>
<td>MIGS-5</td>
<td>Sample collection</td>
</tr>
<tr>
<td>MIGS-4.1</td>
<td>Latitude</td>
</tr>
<tr>
<td>MIGS-4.2</td>
<td>Longitude</td>
</tr>
<tr>
<td>MIGS-4.3</td>
<td>Depth</td>
</tr>
<tr>
<td>MIGS-4.4</td>
<td>Altitude</td>
</tr>
<tr>
<td>MIGS 31</td>
<td>Finishing quality</td>
</tr>
<tr>
<td>MIGS-28</td>
<td>Libraries used</td>
</tr>
<tr>
<td>MIGS 29</td>
<td>Sequencing platforms</td>
</tr>
<tr>
<td>MIGS 31.2</td>
<td>Fold coverage</td>
</tr>
<tr>
<td>MIGS 30</td>
<td>Assemblers</td>
</tr>
<tr>
<td>MIGS 32</td>
<td>Gene calling method</td>
</tr>
<tr>
<td>MIGS 13</td>
<td>Source Material Identifier</td>
</tr>
</tbody>
</table>