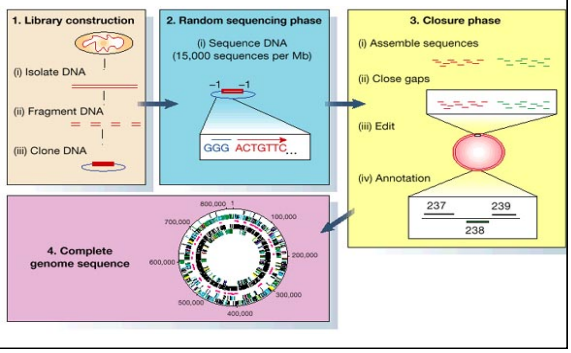
Fraser, Eisen, and Salzman **Shotgun genome sequencing**
Nature **406**, 799-803 (17 August 2000)



SEQUENCE DATA & BIOINFORMATICS

SEQUENCE 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/MIMS ?'

DATABASE STRUCTURE/ACCESSIBILITY
Genomic, Proteomic, Environmental, central vs, distributed, linkout, Federated databases, etc…

AVAILABILITY/ACCESSIBILITY of ANALYTICAL TOOLS
Genomic/Proteomic, Environmental, Polymorphs., Metagenome Analyses., Data cross comparisons

GENE EXPRESSION & PHYSIOLOGY

- MIAME-LIKE STANDARDS FOR ENVIRONMENTAL GENE EXPRESSION ?
  (*The MIAME Checklist:* Experiment Design, Samples used, extract preparation & labeling, Hybridization procedures and parameters, Measurement data and specs, Array design)
  (see http://www.mged.org/minseqe_
- CULTURE COLLECTIONS OF RELEVANT MICROBES
- STANDARD 'CHIPSETS', INTERNAL CONTROLS FOR ENVIRON. MONITORING, CROSS Q-PCR STANDARDIZATION, ETC..
- SGDB-LIKE MODEL FOR INTEGRATING HETERGENEOUS 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 ANITBODIES
- 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 Environmetnal 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 !)

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

TTCAATTGCCAAATCCATCACTAGATGAATGATTCTGAATCAGATATAAAATTACGTAATTGTAATAAGA
ATTGAGTTTTAAAACACATTATGGAAAAAAGGATTGTATTTCATGTTGATTTTGACTATTTTTATGCACA
GTGTGAAGAGATTAGAGATTCTAAAATAAAAGGGAAATGTGTTGCAGTTTGTATATTTTCTGATAGAGGA
GGAGATAGTGGAGCAATAGCTACTGCAAATTACAATGCAAGAAACTTTGGAGTAAAATCAGGAATTCCAA
TTATGCTTGCAAAAAATAAATTAAAAGAGCAAGATTCAGTATTTTTGCCAGCTGATTTTGATTATTATTC
AGAAGTATCATCAAAAGCAATGGAAATAATTGAAAAGTATGCAGATGTATTTGAATATGTTGGAAGAGAT
GAAGCATATCTTGATGTTACAAAAAAAACTGAGAGTGATTTTCATAATGCAGAACATCTAGCGCAACAGT
TGAAAAATGAAATAAGAAATAGTCTAAAACTTACATGTTCTGTAGGGATCACGCCAAATAAACTACTTTC
AAAAATTGCTTCAGATTATAAGAAACCAGATGGATTGACAACTGTAAAACCAGAACAAGTTGATGAGTTT
TTATCACCATTAAAAATAAGAGTAATTCCAGGTATAGGTAAAAAAACAGAAGATTTTTTTGTACAAATGC
ATGTGAATACAATAGAAGACCTAAGAGAAATTAATATTTTTGATTTAAACAAAATGTTTGGGAGAAAAAC
TGGAGGACATATTTTTAATTCTTCAAGAGGAATTGATGAAGAAATGGTTAAACCGAAAGCGCCTACAATT
CAGTTTAGCAAGATTATCACTTTAAAGAAAAATTCTAAAGAACTTAAATTTCTACGTGAAAACATAGAAA
AACTATGTGTGCAATTGAATGAAATTGCAATTGAAAACAACAAAATGTATCGCTCAATTGGAATCCAGTT
TGTAAACGAGGATTTATCCACAAGAACTAAATCAAAAATGATAAAAAATCCAGGAAATAATGTGATTGAA
TTAAAAAAAGTTGTAAATCAATTACTAGAAGAAGCATTGATAGAACAAGAAATGTTAGTTAGAAGAATTG
GAGTTAGAATTTCAGAGTTTTCAGATGTAGAGGGTCAGAGAGACATTACAAATTATTTTTAGATGTTTTT
TCAGATTCTATTCGCTTGAGTCGTTTTGTTTCAACTACTGTCACAATCAGTAAAAAGATGAATAACCACG
GAAGAAACATTATTTCGCCTGCTACAGAATGAAATTCTTCCCATGCTTCAGGATCAGTAGTCACTTTTAG
TGCGTAAACAGATAACGAAAAGATTCGAATCAAATTTACAAATATAGTTCCAAGAATTCCTAAAATAAAA

Content Sensor

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

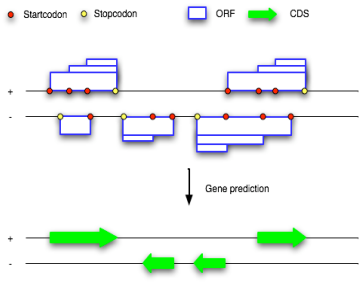
- **Intrinsic Content Sensors**: hexamer count
- HMM: Pr (Xlk previous nt), X = A, T, G, C
- GeneMark, GeneScan: 5 order
- 2 types of content sensors: 1 for coding , other for non-coding sequences.
- (some programs use both - e.g., CRITICA)

Gene Prediction Tools

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

ORF = open reading frame = start codon…[GATC]n…..stop codon

CDS = coding sequence (produces the actual protein/RNA species)



Annotation of bacterial and archaeal genomes: improving accuracy and consistency. Overbeek et al., Chem Rev. 2007. 107:3431-3447.

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

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

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

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

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

STEP 5. Runs BLASTP for predicted proteins against COG database, cog.pro.

STEP 6. Uses information about conservation of neighboring gene pairs in known genomes to improve operon prediction.

STEP 7. Runs BLASTP against NR for proteins having no COGs hits.

STEP 8. predicts potential promoters (BPROM program) or terminators (BTERM) in upstream and downstream regions, correspondingly, of predicted genes. .

STEP 9. Refines operon predictions using predicted promoters and terminators as additional evidences.

Typical softberry output

```
1     1 Op  1  21/0.000  +    CDS       407 -     1747   1311  ## COG0593 ATPase involved in DNA
                         +    Term     1786 -     1823    3.2
                         +    Prom     1847 -     1906   10.5
2     1 Op  2   3/0.019  +    CDS       1926 -    3065   1237  ## COG0592 DNA polymerase
                         +    Term     3074 -     3122    9.1
                         +    Prom     3105 -     3164    4.0
3     2 Op  1   4/0.002  +    CDS       3193 -     3405    278  ## COG2501 Uncharacterized ACR
4     2 Op  2   4/0.002  +    CDS       3418 -     4545    899  ## COG1195 Recombinational DNA
2       Op  3  16/0.000  +    CDS       4578 -     6506   2148  ## COG0187 DNA gyrase (topoisomerase II) B SU
                         +    Term     6516 -     6551    4.7
                         +    Prom     6512 -     6571    2.3
6     2 Op  4     .      +    CDS       6595 -     9066   2957  ## COG0188 DNA gyrase (topoisomerase II) A SU
                         +    Term     9067 -     9098    3.4
                         + SSU_RRNA     9308 -    10861  100.0  # AY138279 [D:1..1554] # 16S rRNA # B cereus
                         +    TRNA     10992 -    11068  101.2  # Ile GAT 0 0
                         +    TRNA     11077 -    11152   93.9  # Ala TGC 0 0
                         + LSU_RRNA    11233 -    14154   99.0  # AF267882 [D:1..2922] # 23S rRNA # Bacillus
7     3 Op  1                     .      -    CDS      14175 -     14363    158
                         + SS_RRNA     14205 -    14315   97.0  # AE017026 [D:165635..165750] # 5S rRNA # Bacil
8     3 Op  2     .      -    CDS       14353 -    15249    351  ## Similar_to_GB
9     3 Op  3     .      -    CDS       15170 -    15352     99
                         -    Prom     15373 -    15432    6.9
```

Expect Value (E) (Karlin-Altschul Statistics)

E = number of database hits you expect to find by chance, in a db of a given size

$E = Kmne^{-\lambda S}$ or $E = mn2^{-S'}$

m = query length

n = subject length (sequence or db)

K = scale for search space

λ = scale for scoring system

S'= bitscore = (λS – lnK)/ln2

More info: www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

BLAST Executables & Programs

Executables:

blastall, megablast, blastpgp, bl2seq, blastclus

Blastall programs:

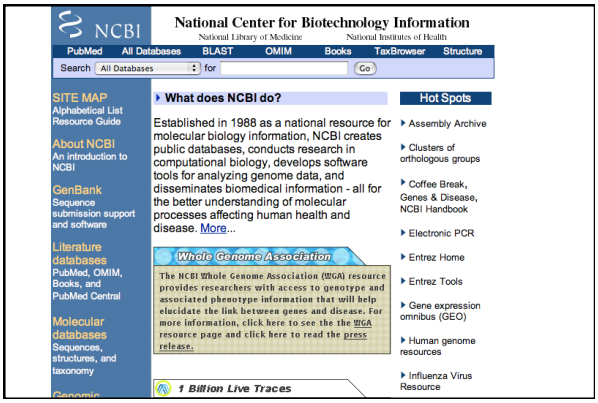
blastp, blastn, blastx, tblastn, tblastx

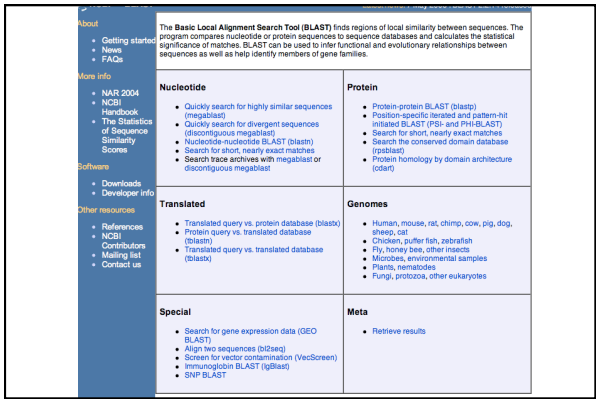
Bare minimum for blastall:

./blastall -p [program] -i [fasta file] -d [database] -o [output]

Several different BLAST programs:

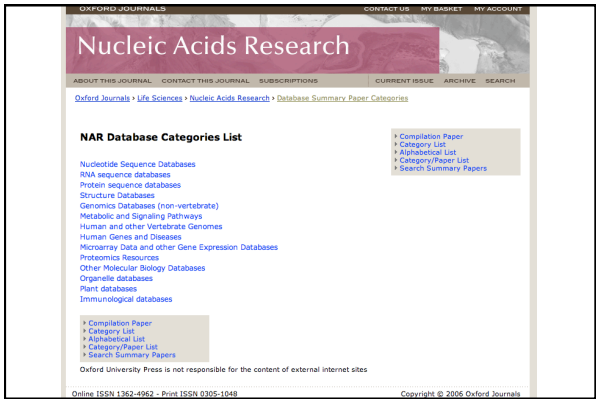
Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	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.

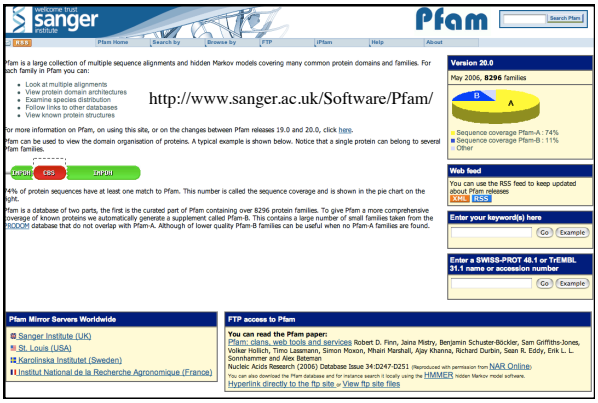


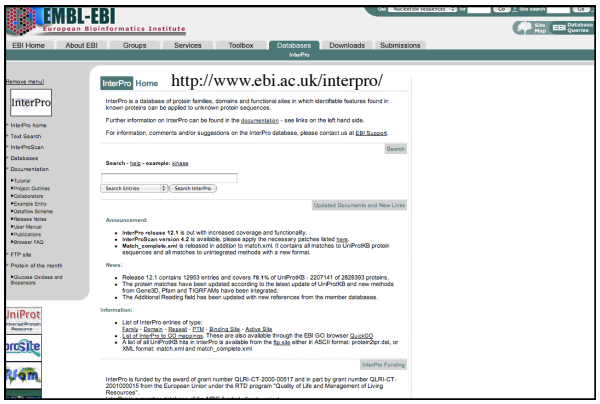


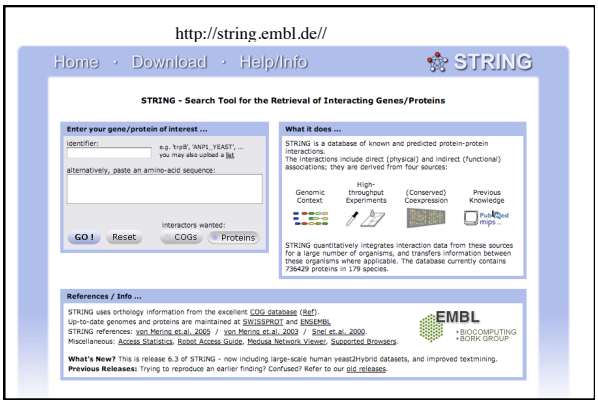

JGI
Quick Genome Search :
GO
img/m
INTEGRATED MICROBIAL GENOMES
with MICROBIOME SAMPLES
IMG/M Home
Find Genes
Find Genomes
Find Functions
Compare Genomes
MyIMG/M
Analysis Carts
About IMG/M
Using IMG/M
News
http://img.jgi.doe.gov/cgi-bin/m/main.cgi
IMG/M Genomes
finished/draft
JGI
Total
Bacteria
Archaea
Eukarya
Viruses
Microbiomes
All Genomes
Grand Total
IMG/M is an experimental system that is based on the Integrated Microbial Genomes (IMG) system. IMG/M integrates microbial community (microbiome) genome data with IMG's isolate microbial genomes and provides support for the comparative analysis of the aggregate microbiome genomes (metagenomes) in the context of isolate microbial genomes and other metagenomes.
IMG/M will be presented on Monday, August 7th, at 10:30am and demonstrated on Wednesday, August 9th, at 9:50am, at ISMB 2006.
IMG/M will be also presented on Friday, August 25, 2006 at 12:30pm at ISME 11.
Experimental Version, March 1, 2006
Questions/Comments
©2006 The Regents of the University of California
Disclaimer


OXFORD JOURNALS
CONTACT US
MY BASKET
MY ACCOUNT
Nucleic Acids Research
ABOUT THIS JOURNAL
CONTACT THIS JOURNAL
SUBSCRIPTIONS
CURRENT ISSUE
ARCHIVE
SEARCH
http://nar.oxfordjournals.org/content/vol34/suppl_1/index.dtl
Contents:
Editorial:
Articles:


OXFORD JOURNALS
CONTACT US
MY BASKET
MY ACCOUNT
Nucleic Acids Research
ABOUT THIS JOURNAL
CONTACT THIS JOURNAL
SUBSCRIPTIONS
CURRENT ISSUE
ARCHIVE
SEARCH
Oxford Journals › Life Sciences › Nucleic Acids Research › Database Summary Paper Categories
NAR Database Categories List
Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Compilation Paper
Category List
Alphabetical List
Category/Paper List
Search Summary Papers
Oxford University Press is not responsible for the content of external internet sites
Online ISSN 1362-4962 - Print ISSN 0305-1048
Copyright © 2006 Oxford Journals

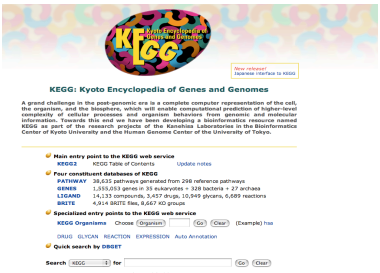

sanger institute
Pfam
Search Pfam
http://www.sanger.ac.uk/Software/Pfam/
Version 20.0
May 2006, 8296 families
Web feed
Enter your keyword(s) here
Enter a SWISS-PROT 48.1 or TrEMBL 31.1 name or accession number
Pfam Mirror Servers Worldwide
Sanger Institute (UK)
St. Louis (USA)
Karolinska Institutet (Sweden)
Institut National de la Recherche Agronomique (France)
FTP access to Pfam
You can read the Pfam paper:


EMBL-EBI
European Bioinformatics Institute
EBI Home
About EBI
Groups
Services
Toolbox
Databases
Downloads
Submissions
InterPro
InterPro Home
http://www.ebi.ac.uk/interpro/
InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.
Announcement:
News:
Information:
InterPro is funded by the award of grant number QLRI-CT-2000-00517 and in part by grant number QLRI-CT-2001000015 from the European Union under the RTD program "Quality of Life and Management of Living Resources".

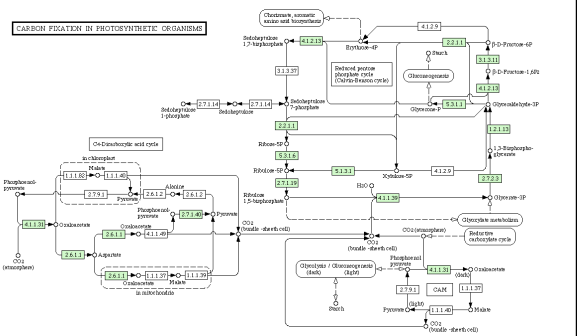

http://string.embl.de//
Home
Download
Help/Info
STRING
STRING - Search Tool for the Retrieval of Interacting Genes/Proteins
Enter your gene/protein of interest ...
Identifier:
alternatively, paste an amino-acid sequence:
GO !
Reset
interactors wanted:
COGs
Proteins
What it does ...
STRING is a database of known and predicted protein-protein interactions.
Genomic Context
High-throughput Experiments
(Conserved) Coexpression
Previous Knowledge
References / Info ...
EMBL
BIOCOMPUTING
BORK GROUP
What's New? This is release 6.3 of STRING - now including large-scale human yeast2Hybrid datasets, and improved textmining.
Previous Releases: Trying to reproduce an earlier finding? Confused? Refer to our old releases.

KEGG
Kyoto Encyclopedia of genes and genomes

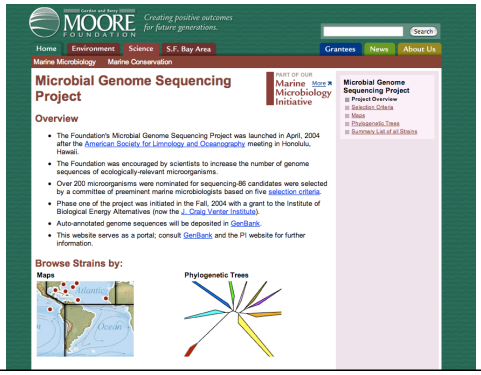
http://www.genome.jp/kegg/



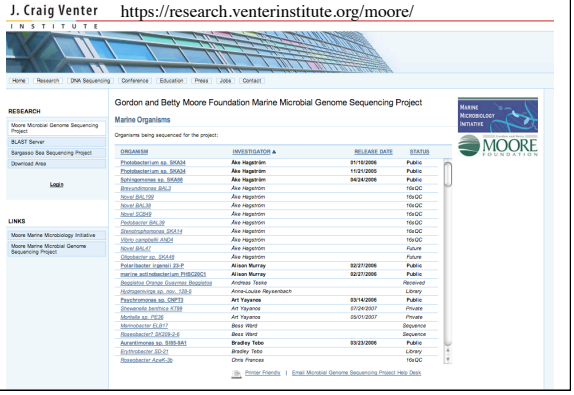
Prochlorococcus MED4 Carbon fixation



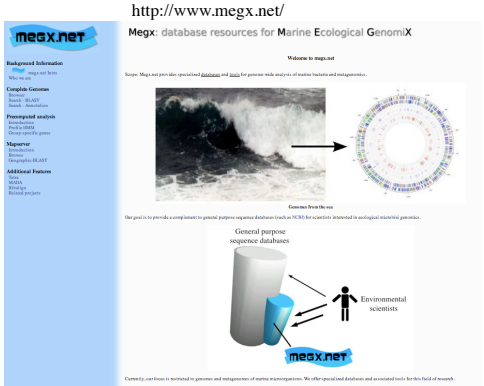
http://www.moore.org/microgenome/



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



http://www.megx.net/



MARINE MICRO SPECIFIC
https://research.venterinstitute.org/moore

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

http://egg.umh.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

RIBOSOMAL RNA DATABASE AND PROBE RESOURCES AND TOOLS
http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

http://www.microbial-ecology.net/probebase/

http://www.arb-home.de/

http://rdp.cme.msu.edu/

TARGETED PROTEIN DATABASES AND SEARCH TOOLS ONLINE
http://nar.oxfordjournals.org/content/vol34/suppl_1/index.dtl

http://www.genome.jp/kegg/

http://www.ncbi.nlm.nih.gov/COG/

http://www.sanger.ac.uk/Software/Pfam/

http://www.ebi.ac.uk/interpro/

http://string.embl.de

Gene Ontology (GO; not a database - rather, sequence semantics !)
http://www.geneontology.org/GO.doc.shtml

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 bac terial genomes using only genomic DNA as an i nput (optionally, pre-learned parameters from related organism can be used)
- Mapping of tRNA and rRNA genes
- Highly accurate Markov chains-based gene prediction_Prediction of promoters and terminators
- Operon prediction based on distances between ORFs and f requencies of different genes neighboring each other in known bacterial genomes, as we ll 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.

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

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

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

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

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

STEP 5. Runs BLASTP for predicted proteins against COG database, cog.pro.

STEP 6. Uses information about conservation of neighboring gene pairs in known genomes to improve operon prediction.

STEP 7. Runs BLASTP against NR for proteins having no COGs hits.

STEP 8. predicts potential promoters (BPROM program) or terminators (BTERM) in upstream and downstream regions, correspondingly, of predicted genes. .

STEP 9. Refines operon predictions using predicted promoters and terminators as additional evidences.

Typical softberry output

```
1     1 Op  1  21/0.000   +    CDS       407 -     1747   1311  ## COG0593 ATPase involved in DNA
                          +    Term     1786 -     1823    3.2
                          +    Prom     1847 -     1906   10.5
2     1 Op  2   3/0.019   +    CDS      1926 -     3065   1237  ## COG0592 DNA polymerase
                          +    Term     3074 -     3122    9.1
                          +    Prom     3105 -     3164    4.0
3     2 Op  1   4/0.002   +    CDS      3193 -     3405    278  ## COG2501 Uncharacterized ACR
4     2 Op  2   4/0.002   +    CDS      3418 -     4545    899  ## COG1195 Recombinational DNA
2       Op  3  16/0.000   +    CDS      4578 -     6506   2148  ## COG0187 DNA gyrase (topoisomerase II) B SU
                          +    Term     6516 -     6551    4.7
                          +    Prom     6512 -     6571    2.3
6     2 Op  4     .       +    CDS      6595 -     9066   2957  ## COG0188 DNA gyrase (topoisomerase II) A SU
                          +    Term     9067 -     9098    3.4
                          + SSU_RRNA     9308 -    10861  100.0  # AY138279 [D:1..1554] # 16S rRNA # B cereus
                          +    TRNA    10992 -    11068  101.2  # Ile GAT 0 0
                          +    TRNA    11077 -    11152   93.9  # Ala TGC 0 0
                          + LSU_RRNA   11233 -    14154   99.0  # AF267882 [D:1..2922] # 23S rRNA # Bacillus
7     3 Op  1                       .       -    CDS      14175 -    14363    158
                          + 5S_RRNA    14205 -    14315   97.0  # AE017026 [D:165635..165750] # 5S rRNA # Bacil
8     3 Op  2     .       -    CDS      14353 -    15249    351  ## Similar_to_GB
9     3 Op  3     .       -    CDS      15170 -    15352     99
                          -    Prom     15373 -    15432    6.9
```