

Tools for Bioinformatics and Genome Analysis

Gen 240B

Thomas Girke

April 5, 2007

Revolution in Biological and Medical Sciences

- ▶ Modern genome sciences has transformed biology into an information science discipline.
- ▶ Do living systems simply consist of information (genome) plus chemical components?
 - ▶ We can create viruses just by knowing their sequence.
 - ▶ Pretty soon the same will be possible with simple bacteria.
- ▶ Still, our understanding of living systems is very superficial.
- ▶ Organismal systems biology is the future and it will be dominated by computational approaches.

Part I: Tools and Databases

ROADMAP SESSION

- A. Software environments for bioinformatics
- B. Biological databases
- C. Sequence analysis
- D. Structural bioinformatics
- E. Cheminformatics in drug discovery
- F. Homework assignment (sequence analysis)

Part II: Microarray Informatics

- G. Analysis of microarray experiments (homework)

Download Slide Show

http://bioinfo.ucr.edu/~tgirke/HTML_Presentations/Gen240B/2007/Gen240B_Apr05.pdf

Standard Bioinformatics Areas

- ▶ Biostatistics & computational biology
- ▶ Algorithms
- ▶ Sequence analysis
- ▶ Phylogenetics
- ▶ Gene expression analysis
- ▶ Proteomics & metabolic profiling
- ▶ Network & systems analysis

Structural Bioinformatics

- ▶ Cheminformatics
- ▶ Drug design
- ▶ Molecular modeling

Biological Databases

- ▶ Data management/curation, web programming, etc.

The Genome Analysis Perspective of Bioinformatics

- ▶ Fragment Assembly: ESTs and genomic fragments
- ▶ Mapping
- ▶ Annotation
 - ▶ Gene predictions
 - ▶ ORFs, UTRs, introns, exons, promoters
 - ▶ Many errors in eukaryote genomes!!
 - ▶ Similarity searches
 - ▶ BLAST, FASTA, Smith-Waterman
 - ▶ Pathway and gene ontology annotations
- ▶ Gene/protein families
 - ▶ Domain databases
 - ▶ Multiple alignments
- ▶ Structure/Function
 - ▶ 2D, 3D structure (availability?)

Software Environments for Bioinformatics

General Areas

1. Algorithms
2. Statistic tools
3. Programming languages
4. Bio* projects
5. UNIX/LINUX
6. Database engines
7. Open-source projects

- ▶ Bioalgorithms.info: <http://www.bioalgorithms.info>
- ▶ Algorithms Archive: <http://www.aic.nrl.navy.mil/galist/>

Why LINUX in Bioinformatics?

- ▶ Free access to hundreds of bioinformatics tools (more up-to-date)
- ▶ Multiuser, multitasking & remote access
- ▶ Better performance, advanced parallel computing on LINUX clusters
- ▶ Access to shell, open-source projects, programming languages, database engines
- ▶ Many more reasons

Selection of important languages for bioinformatics

- ▶ C/C++: books & many URLs
- ▶ Perl: <http://www.perl.org>
- ▶ Python: <http://www.python.org>
- ▶ Ruby: <http://www.ruby-lang.org>
- ▶ JAVA: <http://java.sun.com>
- ▶ R: <http://cran.at.r-project.org>
- ▶ Others ...

Bio* modules for processing data from databases and software applications

- ▶ BioPerl: <http://bio.perl.org>
- ▶ BioPython: <http://biopython.org>
- ▶ BioJava: <http://www.biojava.org>
- ▶ BioRuby: <http://bioruby.org>

Selection of biostatistical tools

- ▶ S-Plus: <http://www.insightful.com/products/s/default.asp>
- ▶ SAS: <http://www.sas.com>
- ▶ MATLAB: <http://www.mathworks.com>
- ▶ R: <http://cran.at.r-project.org>
- ▶ BioConductor: <http://www.bioconductor.org>
- ▶ TIGR MeV: <http://www.tm4.org>
- ▶ Others ...

Database Engines

- ▶ Open-source
 - ▶ MySQL: <http://www.mysql.com>
 - ▶ PostgreSQL: <http://www.postgresql.org>
- ▶ Commercial
 - ▶ Oracle: <http://www.oracle.com>

Database Curation

- ▶ biocurator.org: curation of biological data
- ▶ Literature
- ▶ Gene annotations
- ▶ Many more

Selection of important software collections

- ▶ Sequence and phylogenetic analysis
 - ▶ NCBI Toolkit: <ftp://ftp.ncbi.nih.gov/blast/executables>
 - ▶ EMBOSS: <http://emboss.sourceforge.net/apps/cvs/index.html>
 - ▶ GCG: <http://www.accelrys.com/about/gcg.html> (now commercial!)
 - ▶ PHYLIP: <http://evolution.genetics.washington.edu/phylip.html>
- ▶ Molecular modeling and cheminformatics
 - ▶ Swiss-Model & Swiss-PDB Viewer (Deep View):
<http://swissmodel.expasy.org>
 - ▶ CHARMM: <http://yuri.harvard.edu>
 - ▶ Open Babel: <http://openbabel.sourceforge.net>

Biological Databases

Selection of important databases

- ▶ DNA sequence depositories: [GenBank](#), [EMBL](#) and [DDBJ](#)
- ▶ NCBI: <http://www.ncbi.nlm.nih.gov>
- ▶ EBI: <http://www.ebi.ac.uk>
- ▶ Ensembl: <http://www.ensembl.org>
- ▶ BioMart: <http://www.biomart.org>
- ▶ Swiss-Prot: <http://us.expasy.org/sprot>
- ▶ UniProt: <http://www.pir.uniprot.org>
- ▶ TIGR: <http://www.tigr.org>
- ▶ JGI: <http://www.jgi.doe.gov>

Selection of important databases

- ▶ PFAM: <http://www.sanger.ac.uk/Software/Pfam>
- ▶ PROSITE: <http://us.expasy.org/prosite>
- ▶ ProDom: <http://prodes.toulouse.inra.fr>
- ▶ TIGRFAMs: <http://www.tigr.org/TIGRFAMs>
- ▶ InterPro: <http://www.ebi.ac.uk/interpro>
- ▶ Others

Continued

- ▶ Promoter Databases
 - ▶ EPD: <http://www.epd.isb-sib.ch>
 - ▶ TESS: <http://www.cbil.upenn.edu/tess>
 - ▶ Organism specific collections:
<http://faculty.ucr.edu/~tgirke/Links.htm>
- ▶ Transcriptional Profiling
 - ▶ NCBI GEO: <http://www.ncbi.nlm.nih.gov/geo>
 - ▶ Microarray @ EBI: <http://www.ebi.ac.uk/microarray>
 - ▶ SMD: <http://genome-www5.stanford.edu>
- ▶ Proteomics
 - ▶ SWISS-2DPAGE: <http://us.expasy.org/ch2d>

Opportunities: Expression Data in Public Domain

Example: [Gene Expression Omnibus \(GEO\)](#)

	Public	Unreleased	Total
Platforms	2,968	312	3,280
Samples	113,749	26,988	140,737
Series	4,861	1,036	5,897

Reference: NAR, 2007, Vol. 35, Database issue D760-D765

Process and Organism Databases

- ▶ Pathways and process classification
 - ▶ KEGG: <http://www.genome.jp/kegg>
 - ▶ Gene Ontology: <http://geneontology.org>
- ▶ Organism specific databases
 - ▶ Human, crops, model organisms, etc.
- ▶ Process oriented databases
 - ▶ Specialized resources for pathways, gene families, feature predictions, etc.

- ▶ Protein Structure
 - ▶ Protein Data Bank (PDB): <http://www.rcsb.org/pdb>
 - ▶ Structural Classification of Proteins (SCOP):
<http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html>
- ▶ Bioactive Compounds (drugs)
 - ▶ ChemBank: <http://chembank.med.harvard.edu>
 - ▶ PubChem: <http://pubchem.ncbi.nlm.nih.gov>
 - ▶ NCI: <http://cactus.nci.nih.gov>
 - ▶ ChemMine: <http://bioweb.ucr.edu/ChemMineV2>

Example: NCBI

- ▶ Main Page: <http://www.ncbi.nlm.nih.gov/>
- ▶ GenBank
- ▶ Literature: PubMed & HubMed
- ▶ Alphabetical Link Table:
<http://www.ncbi.nlm.nih.gov/Sitemap/AlphaList.html>
- ▶ FTP Downloads: <ftp://ftp.ncbi.nih.gov/blast/>
- ▶ Useful protein structure viewing tool:
<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>

Sequence Analysis (homework)

Similarity Searching



Feature Predictions



Structure Homologs

Only a brief overview can be provided in this section. More details in module III.

1. EMBOSS: a multipurpose toolbox
2. Alignments & sequence searching
3. BLAST and HMMER
4. Consensus & feature predictions

A Multipurpose Toolbox

A collection of over 120 useful sequence analysis tools is available in the EMBOSS package:

- ▶ Command-line:
<http://emboss.sourceforge.net/apps/cvs/index.html>
- ▶ Several online implementations: [EBI](#) and [UCR](#)

Sequence Searching

- ▶ Sequence similarity searching
 - ▶ BLAST ('local alignment search')
 - ▶ FASTA ('best alignment search')
 - ▶ Pattern matching
 - ▶ Regular Expressions
 - ▶ PSCAN
 - ▶ FUZZPRO
 - ▶ PatScan
 - ▶ ...
- ▶ Remote Homology Detection
 - ▶ Psi-BLAST/RPS-BLAST
 - ▶ Smith-Waterman (SSEARCH, dynamic programming)
 - ▶ HMMs: HMMER, SAM
 - ▶ Domain and motif databases
 - ▶ Fold recognition approaches (Meta Servers)

Alignments

- ▶ Pairwise alignments
 - ▶ Fast local alignment: [BLAST2](#)
 - ▶ Smith-Waterman local alignment: [WATER](#)
 - ▶ Needleman-Wunsch global alignment: [NEEDLE](#)
 - ▶ Long sequences: [SUPERMATCHER](#)
 - ▶ Genome alignments: [MUMMER](#)
- ▶ Multiple alignments
 - ▶ ClustalW multiple alignment: [EMMA](#)
 - ▶ Hierarchical clustering: [MultAlign](#)
 - ▶ For diverse sequences: [T-Coffee](#)
 - ▶ For diverse sequences: [MUSCLE](#)
 - ▶ For local similarities (long gaps): [DIALIGN](#)
 - ▶ DNA alignment guided by protein alignment: [TRANALIGN](#)
 - ▶ Align cDNAs to genome: [EST2GENOME](#)

Comparisons of Diverse Sequences

- ▶ Multiple Alignments
 - ▶ Challenging with multiple domain proteins or sequence identities below 25%
- ▶ Pairwise all-against-all comparisons
 - ▶ Build phylogenetic tree from pairwise scores
- ▶ Descriptor-based approach
 - ▶ Use numeric property values as similarity measure: AA composition, physical properties, etc.

BLAST and HMMER

- ▶ BLAST Flavors: <http://www.ncbi.nlm.nih.gov/BLAST>
 - ▶ blastall: command-line collection of BLAST tools
 - ▶ BLAST: BLASN, BLASTP, TBLASTN, TBLASTX
 - ▶ Psi-BLAST: Position-Specific Iterated BLAST
 - ▶ RPS-BLAST: Reverse Position-Specific BLAST
 - ▶ Phi-BLAST: Pattern Hit Initiated BLAST
 - ▶ Mega-BLAST: 10 faster than BLASTN
 - ▶ BLAST2: pairwise comparisons
 - ▶ WU-BLAST: Washington University BLAST
- ▶ HMMER: <http://hmmer.wustl.edu/>
 - ▶ hmmbuild: hmm from alignment
 - ▶ hmsearch: searches sequence database with hmm
 - ▶ hmmpfam: searches hmm database with sequence
 - ▶ hmalign: aligns multiple sequences to hmm

Pattern Search



Pattern Discovery

Motif Databases

- ▶ The Eukaryotic Promoter Database (EPD)
- ▶ TRANSFAC
- ▶ ...

Pattern matching tools

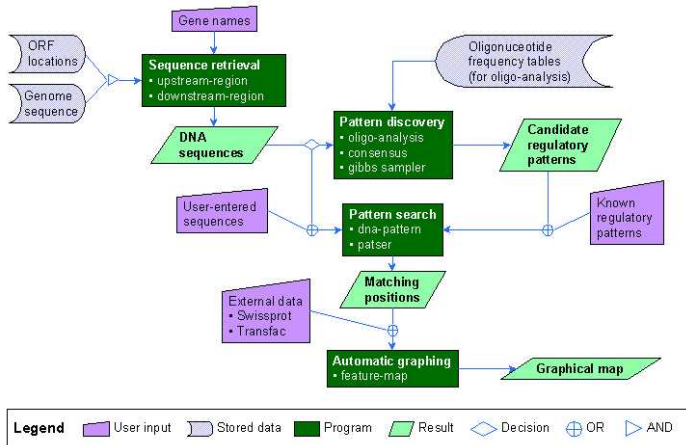
- ▶ Regular Expressions
- ▶ FUZZNUC
- ▶ PatScan
- ▶ ...

Pattern Discovery

- ▶ MotifCut: <http://motifcut.stanford.edu>
- ▶ AlignACE & ScanACE:
<http://arep.med.harvard.edu/mrnadata/mrnasoft.html>
- ▶ MEME and META-MEME, San Diego Super Computer Center: <http://www.sdsc.edu/Research/biology>
- ▶ Regulatory Sequence Analysis Tools (RSA):
<http://rsat.ulb.ac.be/rsat>
- ▶ Gibbs Motif Sampler, Coldspring Harbor:
<http://argon.cshl.org/ioschikz/gibbsDNA/mgibbsDNA-form.html>
- ▶ Motif Sampler, searches for over-represented motifs:
<http://www.esat.kuleuven.ac.be/thijs/Work/MotifSampler.html>
- ▶ Stanford, motif finding in upstream sequences:
<http://genome-www4.stanford.edu/cgi-bin/ewing/oligoAnalysis.pl>

Example: RSA

<http://rsat.ulb.ac.be/rsat>



- ▶ Miscellaneous prediction tools
 - ▶ Gene predictions (e.g.: [GENSCAN](#))
 - ▶ Membrane domains, targeting signals, etc. ([EMBOSS](#))
 - ▶ Secondary structure prediction (e.g.: [garnier](#), [einverted](#), [mfold](#))
 - ▶ Prediction of disordered regions: [DisEMBL](#) & [GlobPlot](#)
 - ▶ Many additional tools

Structural Bioinformatics

- ▶ Protein structure access
 - ▶ PDB General: <http://www.rcsb.org/pdb>
 - ▶ PDB Downloads: <ftp://ftp.rcsb.org/pub/pdb/data>
- ▶ Homology modeling (threading-based structure prediction)
 - ▶ Swiss-Model & Swiss-PDB Viewer (Deep View):
<http://swissmodel.expasy.org>
 - ▶ Prospect:
<http://compbio.ornl.gov/structure/prospect2/index.html>
- ▶ Molecular simulations
 - ▶ CHARMM: <http://yuri.harvard.edu>
- ▶ Tool collection: <http://faculty.ucr.edu/~tgirke/Links.htm>

Cheminformatics in Drug Discovery

Databases

- ▶ PubChem: <http://pubchem.ncbi.nlm.nih.gov>
- ▶ ChemBank: <http://chembank.broad.harvard.edu>
- ▶ NCI: <http://cactus.nci.nih.gov>
- ▶ ChemMine: <http://bioweb.ucr.edu/ChemMineV2>

Tools

- ▶ Open Babel - structure formats:
<http://openbabel.sourceforge.net>
- ▶ JOELib - descriptors: http://www-ra.informatik.uni-tuebingen.de/software/welcome_e.html
- ▶ PerIMol - Perl modules for molecular chemistry:
<http://perlmol.org>
- ▶ ChemPython - Python modules for molecular Chemistry:
<http://chempython.org>
- ▶ CACTVS - collection of computational chemistry programs:
<http://www2.ccc.uni-erlangen.de/software/cactvs>
- ▶ Corina - 3D structure prediction: http://www2.chemie.uni-erlangen.de/software/corina/free_struct.htm
- ▶ PASS - prediction of activity spectra for compounds:
<http://www.ibmh.msk.su/PASS/default.htm>

Homework Assignment

Homework Assignment

1. Go to <http://www.ncbi.nlm.nih.gov>, select protein DB, run query: P450 & hydroxylase & human [organism], select under Limits SwissProt
 - (a) Report final query syntax from Details page.
2. Save GIs from this final query to file (select GI List format under display)
 - (a) Report the number of retrieved GIs.
3. Retrieve the corresponding sequences through Batch-Entrez (<http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi>) using GI list file as query input -> save sequences in FASTA format
4. Generate multiple alignment and tree of these sequences using Multalign (<http://prodes.toulouse.inra.fr/multalin/multalin.html>)
 - (a) Save multiple alignment and tree to file
 - (b) Identify putative heme binding cysteine
5. Open corresponding SwissProt page (<http://us.expasy.org/sprot>) for first P450 sequence in your list
 - (a) Compare putative heme binding cysteine and compare with consensus pattern from Prosite database ([Syntax](#))
 - (b) Report corresponding Pfam ID
 - (c) How many mouse (*Mus musculus*) sequences are in this family (use species tree on Pfam db)
6. BLASTP against nr database (use again first P450 in your list), select on See Conserved Domains from CDD (this runs RPS-BLAST), click on red P450 domain.
 - (a) Compare resulting alignment with result from MultAlin
 - (b) View 3D structure in Cn3D, save structure (screen shot) and highlight heme binding cysteine