Biotools at UMassMed

David Lapointe, Ph.D.
Director Scientific Computing
University of Massachusetts Medical School
biotools.umassmed.edu
Bioinformatics has a large toolset
Bioinformatics?

• Bioinformatics covers a large territory
  – Sequence and Genome Analysis
  – Computational Biology
  – Databases
  – Visualization
  – Programming

• Informatics applied to Biology
Motivation

Biotools started 2001

First as a resource for the Bioinformatics Course

Later expanded access to UMass system, Worcester Colleges, and global.

Last month, 12000 visits from 73 countries.
Welcome to the BioTools site at the University of Massachusetts Medical School.

Restriction and Pattern Analysis
- Restriction mapping tool
- Rebase Query Tool
- Transcription Factor Site Scan

DNA Sequence Analysis
- ORF Plotting Tool
- Primer Selection Tool

Protein Sequence Analysis
- Signal Sequence Cleavage Tool
- Peptide/Protein Statistics
- Garnier Secondary Structure
- PRINTS protein motifs scan
- MHC Motif Predictor

Utilities
- PrettyPlot MSF files
- Translate Nucleic Acids

Links
- BioNetbook
- CoBRA (Biostatistics)
- HSLS Online
- BioGRID
- OregAnno
- PrimerBank
- Signaling Gateway
- BioInformatics.Org
- GeneCards

Last updated: Friday February 27, 2009
Comments and Suggestions welcome.

Research Computing/Information Services @ UMass Medical School
55 Lake Avenue N, Worcester MA 01655
BioTools

The BioTools site is a resource of tools for bioinformatics for the five schools of the UMass System. Some resources are licensed to the UMass Medical School and only usable at that site. Comments and Suggestions are welcome.

EMBOSS-WWW
  EmbossGUI
  Restriction and Pattern Analysis
    Restriction mapping tool
    Rebase Query Tool
    Transcription Factor Site Scan
  DNA Sequence Analysis
    ORF Plotting Tool
    Primer Selection Tool
  Protein Sequence Analysis
    Signal Sequence Cleavage Tool
    Peptide/Protein Statistics
    Garnier Secondary Structure
    PRINTS protein motifs scan
    MHC Motif Predictor
  Utilities
    PrettyPlot MSF files
    Translate Nucleic Acids
    ClustalW Multiple Alignment
    Matrix2PNG
    WebLogo
    Guide to Unix

Links
  BioNetbook
  CoBRA (Biostatistics)
  HSLS Online
  BioGRID
  OregAnno
  PrimerBank
  Signaling Gateway
  Bioinformatics.Org
  GeneCards

Last updated: Friday February 27, 2009
Comments and Suggestions welcome.
BioTools

The BioTools site is a resource of tools for bioinformatics for the five schools of the UMass System. Some resources are licensed to the UMass Medical School and only usable at that site. Comments and Suggestions are welcome.

- EMBOSS-WWW
  - EmbossGUI
- Restriction and Pattern Analysis
  - Restriction mapping tool
  - Rebase Query Tool
  - Transcription Factor Site Scan
- DNA Sequence Analysis
  - ORF Plotting Tool
  - Primer Selection Tool
- Protein Sequence Analysis
  - Signal Sequence Cleavage Tool
  - Peptide/Protein Statistics
  - Garnier Secondary Structure
  - PRINTS protein motifs scan
  - MHC Motif Predictor New!
- Utilities
  - PrettyPlot MSF files
  - Translate Nucleic Acids
  - ClustalW Multiple Alignment
  - Matrix2PNG
  - WebLogo New!
  - Guide to Unix

- High Performance Computing
  - UmassMed HPC Wiki
- UMMS only
  - BioTools Message Forum
  - Statistics Message Forum
  - TRANSFac [info]
  - TRANSCompel [info]
- Links
  - BioNetbook
  - CoBRA (Biostatistics)
  - HSLS Online
  - BioGRID New!
  - OregAnno New!
  - PrimerBank
  - Signaling Gateway
  - Bioinformatics.Org
  - GeneCards

Last updated: Friday February 27, 2009
Comments and Suggestions welcome.
BioTools

The BioTools site is a resource of tools for bioinformatics for the five schools of the UMass System. Some resources are licensed to the UMass Medical School and only usable at that site. Comments and Suggestions are welcome.

EMBOSS WWW

EmbossGUI

Restriction and Pattern Analysis
- Restriction mapping tool
- Rebase Query Tool
- Transcription Factor Site Scan

DNA Sequence Analysis
- ORF Plotting Tool
- Primer Selection Tool

Protein Sequence Analysis
- Signal Sequence Cleavage Tool
- Peptide/Protein Statistics
- Garnier Secondary Structure
- PRINTS protein motifs scan
- MHC Motif Predictor

Utilities
- PrettyPlot MSF files
- Translate Nucleic Acids
- ClustalW Multiple Alignment
- Matrix2PNG
- WebLogo
- Guide to Unix

High Performance Computing
- UmassMed HPC Wiki

UMMS only
- BioTools Message Forum
- Statistics Message Forum
- TRANSFac [info]
- TRANSCompel [info]

Links
- BioNetbook
- CoBRA (Biostatistics)
- HSLS Online
- BioGRID [New!]
- OregAnno [New!]
- PrimerBank
- Signaling Gateway
- BioInformatics.Org
- GeneCards

Last updated: Friday February 27, 2009
Comments and Suggestions welcome.
Select a category to locate programs by function

<table>
<thead>
<tr>
<th>Alignment Consensus</th>
<th>Nucleic Motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment Differences</td>
<td>Nucleic Mutation</td>
</tr>
<tr>
<td>Alignment Dot Plots</td>
<td>Nucleic Primers</td>
</tr>
<tr>
<td>Alignment Global</td>
<td>Nucleic Profiles</td>
</tr>
<tr>
<td>Alignment Local</td>
<td>Nucleic Repeats</td>
</tr>
<tr>
<td>Alignment Multiple</td>
<td>Nucleic Restriction</td>
</tr>
<tr>
<td>Display</td>
<td>Nucleic Translation</td>
</tr>
<tr>
<td>Edit</td>
<td>Phylogeny</td>
</tr>
<tr>
<td>Enzyme Kinetics</td>
<td>Protein 2D Structure</td>
</tr>
<tr>
<td>Feature Tables</td>
<td>Protein Composition</td>
</tr>
<tr>
<td>Nucleic 2D Structure</td>
<td>Protein Motifs</td>
</tr>
<tr>
<td>Nucleic Codon Usage</td>
<td>Protein Mutation</td>
</tr>
<tr>
<td>Nucleic Composition</td>
<td>Protein Profiles</td>
</tr>
<tr>
<td>Nucleic CpG Islands</td>
<td></td>
</tr>
<tr>
<td>Nucleic Gene Finding</td>
<td></td>
</tr>
</tbody>
</table>

ALIGNMENT CONSENSUS

<table>
<thead>
<tr>
<th>Program name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cons</td>
<td>Creates a consensus from multiple alignments</td>
</tr>
<tr>
<td>megamerger</td>
<td>Merge two large overlapping nucleic acid sequences</td>
</tr>
<tr>
<td>merger</td>
<td>Merge two overlapping nucleic acid sequences</td>
</tr>
</tbody>
</table>
BioTools

The BioTools site is a resource of tools for bioinformatics for the five schools of the UMass System. Some resources are licensed to the UMass Medical School and only usable at that site. Comments and Suggestions are welcome.

EMBOSS-WWW
- EmbossGUI
- Restriction and Pattern Analysis
  - Restriction mapping tool
  - Rebase Query Tool
  - Transcription Factor Site Scan
- DNA Sequence Analysis
  - ORF Plotting Tool
- Protein Sequence Analysis
  - Signal Sequence Cleavage Tool
  - Peptide/Protein Statistics
  - Garnier Secondary Structure
  - PRINTS protein motifs scan
  - MHC Motif Predictor
- Utilities
  - PrettyPlot MSF files
  - Translate Nucleic Acids
  - ClustalW Multiple Alignment
  - Matrix2PNG
  - WebLogo
- Guide to Unix

High Performance Computing
- UmassMed HPC Wiki
- UMMS only
  - BioTools Message Forum
  - Statistics Message Forum
  - TRANSFac
  - TRANSCompel
- Links
  - BioNetbook
  - CoBRA (Biostatistics)
  - HSLS Online
  - BioGRID
  - OregAnno
  - PrimerBank
  - Signaling Gateway
  - BioInformatics.Org
  - GeneCards

Last updated: Friday February 27, 2009
Comments and Suggestions welcome.
Gateways to local resources

Lists

Wikis

Tools
<table>
<thead>
<tr>
<th>Date</th>
<th>Subject</th>
<th>Replies</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-03-06 12:54:59</td>
<td>FW: Funding opportunities for construction/renovation and instrumentation</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2009-03-05 12:40:33</td>
<td>FW: News on the ARRA Stimulus</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2009-02-25 09:43:43</td>
<td>Desktop Sequence Software</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2009-02-19 13:18:39</td>
<td>FW: Leveraging Data Pipelining for On-the-Fly Data Integration of Multiple Experiments</td>
<td>2</td>
<td>Kittler, Ellen Ph. D.</td>
</tr>
<tr>
<td>2009-02-15 09:53:30</td>
<td>Re: Ingenuity Software</td>
<td>3</td>
<td>Elizabeth Luna</td>
</tr>
<tr>
<td>2009-02-05 08:57:57</td>
<td>FW: [BioC] [JOB] National Cancer Institute, Bethesda, MD</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-11-24 17:14:33</td>
<td>Mathworks Seminar Dec 3</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-09-08 10:48:26</td>
<td>FW: [blast-announce] New BLAST 2 Sequences Interface</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-08-27 09:55:46</td>
<td>Ingenuity Software</td>
<td>1</td>
<td>Elizabeth Luna</td>
</tr>
<tr>
<td>2008-08-12 16:16:00</td>
<td>Mathworks Seminar</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-07-28 09:54:14</td>
<td>Simulation Tools/Software from Simbios</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-07-22 09:23:50</td>
<td>FW: [blast-announce] Primer-BLAST now available - addendum</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-07-21 11:36:31</td>
<td>FW: [Gene-announce] Enhancements to Entrez Gene</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-06-16 22:01:10</td>
<td>FW: RECOMB Regulatory Genomics, Systems Biology, and DREAM3 2008 announcement</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-06-10 16:27:57</td>
<td>FW: NIH Notices and Funding Opportunities</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-06-06 14:35:20</td>
<td>FW: [Gene-announce] Entrez Gene removes links to GDB</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-06-05 12:16:30</td>
<td>FW: Reminder: Webinar Invitation - Advances in GPCR Research, June 17</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
<tr>
<td>2008-06-02 13:32:18</td>
<td>FW: [Bioclusters] NRCR Summer Institute 2008 -- Cyberinfrastructure for Biomedicine</td>
<td>0</td>
<td>Lapointe, David</td>
</tr>
</tbody>
</table>
Main Page

High Performance Computing at UMassMed

The High Performance Wiki is maintained by Academic and Research Computing Services at the University of Massachusetts Medical School. If you would like to participate in this wiki contact us.

What is HPC?

High Performance Computing refers to the use of supercomputers and computer clusters to solve computationally intensive problems applied generally to scientific research. In the biomedical arena, high performance computing is used to solve problems of

- Medical Physics
- Protein Structure
- Molecular Dynamics
- Comparative Genomics
- Computational Biology
to name a few.

Topics

Overview

- Binar
- HPCC
- Getting Access
- Resources
- UNIX Related QA

Using Clusters

- Queues
- Software
PMID Lookup to HTML

Here's where to locate the pmid for a publication on PubMed. Just use the number on the form or create a text file of pmids to upload.

Enter PMIDs here (one pmid/line):

12036939
1769064
2501299

or upload file with PMIDS: 1 per line

Get Refs!  Reset
References

Download zipped references


Last updated: Thursday April 02, 2009

Comments and Suggestions welcome.
Links to External resources
BioTools

The BioTools site is a resource of tools for bioinformatics for the five schools of the UMass System. Some resources are licensed to the UMass Medical School and only usable at that site. Comments and Suggestions are welcome.

EMBOSS-WWW
  EmbossGUI
Restriction and Pattern Analysis
  Restriction mapping tool
  Rebase Query Tool
  Transcription Factor Site Scan
DNA Sequence Analysis
  ORF Plotting Tool
  Primer Selection Tool
Protein Sequence Analysis
  Signal Sequence Cleavage Tool
  Peptide/Protein Statistics
  Garnier Secondary Structure
  PRINTS protein motifs scan
  MHC Motif Predictor [New]
Utilities
  PrettyPlot MSF files
  Translate Nucleic Acids
  ClustalW Multiple Alignment
  Matrix2PNG
  WebLogo [New]
  Guide to Unix

High Performance Computing
  UmassMed HPC Wiki
UMMS only
  BioTools Message Forum
  Statistics Message Forum
  TRANSFac [info]
  TRANSCompel [info]
Links
  BiNetbook
  CoBRA (Biostatistics) [New]
  RCSB Protein Data Bank
  BioGRID [New]
  OregAnno [New]
  PrimerBank
  Signaling Gateway
  Bioinformatics.Org
  GeneCards

Last updated: Friday February 27, 2009
Comments and Suggestions welcome.
### Browse COBRA by Subject

- Categorical Data Analysis (71)
- Clinical Epidemiology (55)
- Clinical Trials (101)
- Computation (120)
- Computational Biology/Bioinformatics (218)
- Design of Experiments and Sample Surveys (49)
- Disease Modeling (89)
- Epidemiology (187)
- General Biostatistics (272)
- Genetics (104)
- Health Services Research (33)
- Institutional and Historical (3)
- Laboratory and Basic Science Research (19)
- Longitudinal Data Analysis and Time Series (140)
- Medical Specialties (27)
- Microarrays (198)
- Multivariate Analysis (132)
- Statistical Models (372)
- Statistical Theory and Methods (591)
- Survival Analysis (269)
- Vital and Health Statistics (11)

---

### Peer-Reviewed Journals

- Statistical Applications in Genetics and Molecular Biology*
- The International Journal of Biostatistics*

**Propose a journal**

---

* guest registration

### Browse COBRA by Institution

- Bioconductor Project
- Collection of Biostatistics Research Archive
- Columbia University
- Duke University
- Harvard University
- Johns Hopkins University
- Memorial Sloan-Kettering Cancer Center
- University of California, San Francisco
- University of California, Berkeley
- University of Michigan School of Public Health
- University of North Carolina at Chapel Hill
- University of Pennsylvania
- University of Texas, MD Anderson Cancer Center
- University of Washington

**Create a new series**
GeneCards® is a searchable, integrated database of human genes that provides concise genomic, proteomic, transcriptomic, genetic and functional information on all known and predicted human genes. Information featured in GeneCards includes orthologies, disease relationships, mutations and SNPs, gene expression, gene function, pathways, protein-protein interactions, related drugs & compounds and direct links to cutting edge research reagents and tools such as antibodies, recombinant proteins, clones, expression assays and RNAi reagents.

SAMPLE GENE: CASP3

GeneCards Guide
Getting Started
Statistics

Search the GeneCards Human Gene Database

Search by:
- Keywords
- Gene Symbol only
- Symbol/alias
- GC id
- Symbol/External id

Examples:
- tay sachs
dimerization AND diabetes
- neurodagenorative OR smile
- ESR1
- wnt*
- FRAXA
- QC17M03*
- 3395
- 7431
- P12004
- ENSG00000185515

Options:
- Show microcards only
- Show minicards also (slower)
- Sort microcards alphabetically (faster)
- Sort microcards by relevance

More search examples

The GeneCards search is case insensitive

GeneCards Gene Database statistics
Next Up

Provide links to local resources
Faculty developed applications
Portal to data storage

Develop applications for Web 3.0
Web 3.0?
Where is this going?

The web is a great vehicle for distributing information, creating resources.

The web, however, is human readable. It is difficult to harvest information from websites.

Resource creators are moving to semantic web along with web service models to allow machine harvest of information.
So a researcher might come to me or you and ask

I have 200 Entrez gene Ids. What pathways do these genes belong to?

What resource would you direct them toward?
There is always Google!

Find workflows and Web services for bioinformatics

Pathways

Search

Creator
Name: Antoon Goderis
Member since: Nov 3, 2006

Contributors [Volunteer to contribute]
José M.... Anonymous  Franck  Duncan  Paul Fish...
Stian Sd... Alan R Wi...

Search engine details
Find Web services and workflows for bioinformatics applications, compatible with the open source Taverna workflow editor;
http://www.mygrid.org.uk/taverna

searches 136 sites, including:
http://www.mygrid.org.uk/wiki/Mygrid/BiologicalWebServices,
workflows.mygrid.org.uk, www.biojava.org/docs/
taverna.sourceforge.net/index.php?doc=services.html
bioweb.pasteur.fr/docs/EMBOSS/

Keywords: biology, bioinformatics, web services, workflows, Taverna, myGrid, registry, myExperiment

Last updated: Aug 14, 2007
Add this search engine to your Google homepage
Add this search engine to your blog or webpage »
Create your own Custom Search Engine »
Hmm, mostly journal articles. Let see Workflows only

Refine results for **Pathways:**

<table>
<thead>
<tr>
<th>No workflows</th>
<th>No local applications</th>
<th>Workflows only</th>
<th>WSDL only</th>
</tr>
</thead>
</table>

**BioMed Central | Full text | Systems biology for identifying liver ...**
Mar 10, 2009 ... The objective of this paper is to describe systems biology methods for identifying pathways involved in liver toxicity induced by free fatty ...
www.biomedcentral.com/1753-6561/3/1/S2/S2
by Z Li - 2009
Labeled No workflows WSDL only

**Genome Biology | Full text | Computational prediction of human ...**
PathoLogic predicts the metabolic pathways of the organism, providing new .... Current knowledge of human nutrition based on metabolic pathways is derived ...
genomebiology.com/2004/6/1/R2
by P Romero - 2004 - Cited by 80 - Related articles - All 13 versions
Labeled No workflows WSDL only

**Bioinformatics—from genes to pathways - Nature Methods**
Combined with the right computational tools, genomic data can uncover unknown pathways to cellular processes. Because few researchers have the resources to ... www.nature.com/nmeth/journal/v1/n2/full/nmeth1104-159.html
by L Bonetta - 2004 - Cited by 12 - Related articles
Labeled WSDL only No workflows

**Genome Biology | Full text | Reactome: a knowledgebase of ...**
Reactome: a knowledge base of biologic pathways and processes .... Pathways can be part of larger pathways, Reactome represents glycolysis and .... genomebiology.com/2007/8/3/r39
by I Vastrik - 2007 - Cited by 101 - Related articles - All 7 versions
Labeled No workflows WSDL only

**PATIKAweb: a Web interface for analyzing biological pathways ...**
Summary: PATIKAweb provides a Web interface for retrieving and analyzing biological pathways in the PATIKA database, which contains data integrated from ... bioinformatics.oxfordjournals.org/cgi/content/abstract/22/3/374
by U Dogrusoz - 2006 - Cited by 19 - Related articles - All 8 versions
Labeled No workflows WSDL only
Ok This looks better

Refine results for **Pathways**:

- myExperiment.org - Workflows - omim and **pathways** (Katy ...)
  Mar 3, 2009 ... Title: omim and pathways. Type: Taverna 1 ... pathways, KeggGenestoPathways, pathway_by_genes, pathwayDescriptions, KeggGenestoPathways: ...
  www.myexperiment.org/workflows/688
  Labeled Workflows only

- myExperiment - Workflows - Entrez Gene to KEGG **Pathway** (Paul ...)
  These gene ids are then cross-referenced to KEGG gene ids. Each KEGG gene id is then sent to the KEGG **pathway** database and its relevant **pathways** returned: ...
  www.myexperiment.org/workflows/15
  Labeled Workflows only

- myExperiment.org - Workflows - **Pathways** and Gene annotations for ...
  Workflow Entry: **Pathways** and Gene annotations for Arabidopsis affy data: ... The KEGG gene identifiers are then used to search for **pathways** in the KEGG: ...
  www.myexperiment.org/workflows/725
  Labeled Workflows only

- myExperiment.org - Workflows - Mouse **Pathways** and Gene annotations ...
  The KEGG gene identifiers are then used to search for **pathways** in the KEGG **pathway** database: ... **Pathways** and Gene annotations for Arabidopsis affy data: ...
  www.myexperiment.org/workflows/16
  Labeled Workflows only

- myExperiment.org - Workflows - Mapping microarray data onto ...
  Nov 22, 2007 ... This workflow maps microarray data onto metabolic **pathway** diagrams represented as SBML models drawn using Cell Designer: ...
  www.myexperiment.org/workflows/79
  Labeled Workflows only

- myExperiment - Workflows - KEGG **pathways** common to both QTL and ...
  This workflow takes in two lists of KEGG **pathway** ids. These are designed to come from **pathways** found from genes in a QTL (Quantitative Trait Loci) region: ...
  www.myexperiment.org/workflows/13
  Labeled Workflows only
This workflow takes in Entrez gene ids then adds the string "ncbi-geneid:" to the start of each gene id. These gene ids are then cross-referenced to KEGG gene ids. Each KEGG gene id is then sent to the KEGG pathway database and its relevant pathways returned.
www.myexperiment.org

myExperiment makes it really easy to find, use and share scientific workflows and other files, and to build communities.

Use myExperiment to...
- Find Workflows
- Find Files
- Share Your Workflows and Files
- Create and Find Packs of Items
- Create and Join Groups
- Find People and Make Friends
- Send Messages
- Get Feedback
- Tag and Rate things
- Write Reviews and Comments
- Build your Profile and Reputation

Explore

Register

or Login:

Username or Email:
Password:
Remember me:

Or use OpenID:

Forgot Password?

About myExperiment
Join the Mailing List
Give us Feedback
For Developers
The myGrid Project
Taverna Workflow Workbench
The BioCatalogue Project
myExperiment Publications

myExperiment has over 1500 users, 130 groups, 560 workflows, 150 files and 40 packs

BioCatalogue will provide a single registration point for Web Service providers and a single search site for scientists and developers.

BioCatalogue will also act as a place where the community can find contacts and meet the experts and maintainers of these services.

The BioCatalogue team is currently working with the Embrace team to merge their registries. In the meantime, if you are keen to register your web services, please use the Embrace Registry, the contents of which will be merged with BioCatalogue in due course.

The BioCatalogue team is currently working on its first release, the pilot BioCatalogue, which has been released for testing to our biocatalogue-friends mailing list.

More information about the pilot and current BioCatalogue activities can be found on the BioCatalogue public wiki.
The end result is that we are approaching eScience (EU)/cyberinfrastructure(USA)

Genomics (NCBI, Ensembl, UCSC,GMOD)

meets System Biology (KEGG,BIND,GO)