Web-based Bioinformatics (Proteomics) Applications

Chiquito Crasto
Department of Genetics, UAB
chiquito@uab.edu
January 30, 2013

Philosophical underpinnings …

• Bioinformatics is here to stay—simply because computers are part of everyday life. This is not going to change in the near or distant future
• Students, researchers, etc., will be better served embracing bioinformatics ideas even if they do not necessarily want to pursue bioinformatics-driven careers, and opt to be “bench” scientists
  – By bioinformatics-driven, one means developmental aspects, e.g., developing software to do sequence-similarity searches
• There is significant tool development that will allow scientists to access these to enhance their research (data-analysis, information dissemination, etc.) without having to recourse to collaborations with bioinformatics specialists—unless if specific tools have to be developed
• One should not ignore the intellectualism that goes into conceptualizing and developing tools
• It makes sense then to be able to access and understand how to use these tools
Interoperability & Database Accessibility

- Interoperability: the ability of systems to interoperate, that is exchange information in meaningful ways without having to reproduce information
- Integration: accessing and presenting information that is stored in different resources
  - This precludes the need to store the same information in different resources
  - Examples, how information is stored in the NCBI databases

Theme of the today’s class—web-based proteomics applications

- Isocitrate dehydrogenase (EC 1.1.1.42) and (EC 1.1.1.41), also known as IDH, is an enzyme that participates in the citric acid cycle. It catalyzes the third step of the cycle: the oxidative decarboxylation of isocitrate, producing alpha-ketoglutarate (α-ketoglutarate) and CO₂ while converting NAD⁺ to NADH.

http://en.wikipedia.org/wiki/File:Citric_acid_cycle_with_aconitate_2.svg
NCBI (National Center for Biotechnology Information)

Selected Applications through NCBI

- GenBank—resource for genes
- BioSystems
- BLAST
- Pubmed
- Computational Resources from NCBI's Structure Group
- Conserved Domain Database (CDD)
- Peptidome
- Protein Clusters
- Protein Database
- Structure (Molecular Modeling Database)

Genbank (Search Nucleotide)

GenBank Overview
What is GenBank?
GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequence (Nucleic Acids Research 2007 Jan 18;39(Database issue):D27-7). There are approximately 192,037,651 nucleotide sequences in 172,466,934 sequence records in the GenBank division as of April 2013.
The complete GenBank can be searched using BLAST (Basic Local Alignment Search Tool), which recognizes the DNA Database of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Access to GenBank

There are several ways to search and retrieve data from GenBank:
- Search GenBank for sequence identifiers and annotations with GenBank Overview, which is divided into three distinct sections: GenBank Overview (the main collection), BLAST (Basic Local Alignment Search Tool), and 3DSSRO (Drosophila Survey Sections).
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches GenBank, dbEST, and dbGSS independently, using BLASTn for nucleotide searches and BLASTp for protein searches.
- Search, view, and download sequences programmatically using LDBL utilities.

GenBank Data Usage
The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide written or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.
Protein Sequence in Genbank (isocitrate dehydrogenase)

Note that the protein sequence and the rest of the entries are formatted similar to that of the nucleotide sequences in Genbank.

BioSystems
Pubmed—repository of biomedical abstracts

Information in Pubmed is available in several formats. Abstracts can be downloaded 500 at a time. Abstracts can be specified in terms of date of publication, author lists, etc. If subscriptions are available, a user can access the full text of articles. NCBI has made several utility tools available to automatically download abstracts.
A single Abstract in Pubmed

Ataxia telangiectasia mutated influences cytochrome c oxidase activity.

Pate AY, MacDonald TN, Specks UD, Cheng JY, Fisher JE.
Department of Biology, Saint Louis University, St. Louis, MO 63105, USA.

Abstract

Cells lacking ataxia telangiectasia mutated (ATM) have impaired mitochondrial function. Furthermore, mammalian cells lacking ATM have increased levels of reactive oxygen species (ROS) as well as mitochondrial DNA (mtDNA) deletions in the region encoding the cytochrome c oxidase (COX). We hypothesized that ATM specifically influences COX activity in skeletal muscle. COX activity was ∼45% lower in diabetic animals from ATM-deficient mice than in wild-type mice (P<0.01, n=14/group). However, there were no ATM-related differences in activity of succinate dehydrogenase, isocitrate dehydrogenase, aconitase, citrate synthase, and succinate thiocarboxylate isomerase. Incubation of wild-type or ATM-deficient skeletal muscle homogenates with the ATM inhibitor K-55233 caused a >50% reduction (P<0.05, n=4/group). In COX activity compared to muscle incubated with vehicle alone. Among the control muscles, no significant trend with the ATM inhibitor. COX activity was (normalized to ECL, P<0.05) with activity of glutamate 5-phosphate dehydrogenase, a key determinant of antioxidant defense through production of NADPH. Overall, the findings suggest that ATM has a protective role for COX activity.

PHEM: 23561616 (Published by Elsvier Inc.)

Computational Resources from NCBI's Structure Group

### Three-dimensional structure views in Genbank—STRUCTURE

<table>
<thead>
<tr>
<th>#</th>
<th>Protein Name</th>
<th>Description</th>
<th>Taxonomy</th>
<th>Proteins</th>
<th>Chemicals</th>
<th>Date Published</th>
<th>PDB ID</th>
<th>MDD ID</th>
<th>View in Cn3D</th>
<th>Published</th>
<th>Protein</th>
<th>Similar Structures</th>
<th>Conserved Domains</th>
<th>PubChem Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Isocitrate Dehydrogenase Y160F Mutant Steady-State Intermediate Complex (Laue</td>
<td></td>
<td>Escherichia coli</td>
<td>2</td>
<td>3</td>
<td>2012/11/01 00:00</td>
<td>1hJ3</td>
<td>1B6F</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Crystal Structure Of Isocitrate Dehydrogenase Mutant K230m Bound To Isocitrate And Mn2+</td>
<td></td>
<td>Escherichia coli</td>
<td>2</td>
<td>3</td>
<td>2012/10/13 00:00</td>
<td>11Z7</td>
<td>1J9P</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Low Temperature Structure Of Wls-Type Iib Complexed With Mg-Isocitrate</td>
<td></td>
<td>Escherichia coli</td>
<td>2</td>
<td>3</td>
<td>2012/11/13 00:00</td>
<td>11Z7</td>
<td>1J9P</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Regulation Of An Enzyme By Phosphorylation At The Active Site Residue</td>
<td></td>
<td>Escherichia coli</td>
<td>2</td>
<td>3</td>
<td>2013/01/15 00:00</td>
<td>11Z7</td>
<td>1J9P</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

---

### Structure of Actin—Genbank Structure View

- **Link to Protein Databank**
- **Visualization software**
Structure of Domains in Genbank

List of domains related to or associated with Isocitrate Dehydrogenase

Conserved domain database (CDD) in Genbank

This page is used for searching by text term (other search methods allow queries by protein sequence).
- Enter one or more search terms (e.g., chloride channel).
- Use search fields to narrow your search (e.g., family).
- Advanced search options are available in the Limits, Previewsheets, and History folder tabs.
- Use quotes to force a phrase search (e.g., "voltage gate")
- Use a wildcard (e.g., *glut* or *glut*) to search for a word stem.
- Search results and conserved domain records are described in the help document.

About the Database

Conserved domains are functional units within a protein that have been used as building blocks in molecular evolution and recombined in various arrangements to make proteins with different functions. The Conserved Domain Database (CDD) brings together several collections of multiple sequence alignments representing conserved domains, including NCBI-curated domains, which use 3D structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases (e.g., SMART, CDD, PFAM, TIGRFAM).

The data are then used for pairwise functional annotation of protein query sequences based on matches to specific CDD (illustrated example is superfamily). Identification of proteins with similar domain architectures, and protein classification. The Search conserved domains and Protein Classification overview pages provide more information about the resources available and how they can be used.
Clustering Proteins in terms of Sequence Similarities--Genbank

[Image of the Protein Clusters website interface]

Clustering Proteins in terms of Sequence Similarities--Genbank

[Image of the top patterns section of the Protein Clusters website]

12
ENSEMBL—European version of Genbank—now focused exclusively on genome wide applications

Sample Ensembl Result—Chromosomal location and other features for downloading information

Chromosome 15: 90,288,632-90,483,231

Region in detail

Location: 15:90289632-90483231  Go
Gene:  Go
ENSEMBL—Gene Summary

ENSEMBL—Protein
UniProt combines SwissProt and TrEMBL

“UniProtKB/TrEMBL (unreviewed) contains protein sequences associated with computationally generated annotation and large-scale functional characterization. UniProtKB/Swiss-Prot (reviewed) is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions” --http://www.uniprot.org/help/uniprotkb

UniProt has replaced SwissProt

**Mirror Sites**
SwissProt—search for Proteins

Results

- SwissProt/Swiss-Prot
- UniprotKB/Swiss-Prot
- Trembl
- InterPro
- IPA
- MEROPS
- TIGRFAMs
- PIR
- SWISS-2.8
- OMA
- DDJ
- GO
- CDCAST
- IndexDB
- PDB
- CD-PDB
- BRENDA
- REACTOME
- Sequence Retrieval System (SCOP)
- PRO公开数据库

EXPASY-Databases and Features

Translate
Swiss 2D-PAGE

Swiss 2D-PAGE – Isocitrate dehydrogenase

Query Results:

<table>
<thead>
<tr>
<th>Accession</th>
<th>ID</th>
<th>Description</th>
<th>Gene</th>
<th>Keywords</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>O00200</td>
<td></td>
<td>Isocitrate dehydrogenase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O00205</td>
<td></td>
<td>Isocitrate dehydrogenase</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Swiss 2D-PAGE – Isocitrate dehydrogenase

Search by description (DE), entry name (ID), gene name (GN) or UniProt/Swiss-Prot keywords (KW)

Enter search keywords: (isocitrate dehydrogenase)

Limit by: (Options: DE, GN, KW)

Include external UniProt/Swiss-Prot data in search: (Options: O, GN, DE)

Set by: (Options: Accession number, Protein ID, Gene name)

Please enter a keyword. This may be any word or partial word appearing in the entry identifier (ID), the description (DE), the gene names (GN) or a UniProt/Swiss-Prot keyword (KW). For example, you may type specie, human, or just apo, or ATP or AP01, ROBAM.

If you enter more than one keyword, entries having any keyword will be listed. Please do NOT use any boolean operators (and, or, etc.).
Welcome to the SWISS-MODEL Repository

The SWISS-MODEL Repository is a database of annotated three-dimensional comparative protein structure models generated by the fully automated homology-modelling pipeline SWISS-MODEL.

Example Queries:
- PDB2RMS [1S1F_A5002] [1PFL_A2181] [1PFL_A3750] [1PFL_A4200] [1PFL_A4900] [1PFL_A5400]
- PDB2RMS -- iopritrats dehydrogenase Accession Number

SEARCH

The current release of the SWISS-MODEL-Repository (10.2.2) consists of 3,021,165 model entries for 224,452 unique sequences in the UniProt database.

NOTE: The SWISS-MODEL repository contains theoretically calculated models, which may contain significant errors.

SwissModel Repository ...
Uniref—Clustering of Proteins

Cluster: lactic dehydrogenase (NADP) (56%)

Published January 11, 2011
Built on seed sequence APN2111 Unit component clusters with 100% or 99% identity

Filter
I - 25 of 1,137 members here 163 organisms
Dataset
[Taxonomy (11,213)]
[Filter | Reset]

Members

<table>
<thead>
<tr>
<th>Member</th>
<th>Entry name</th>
<th>Status</th>
<th>Protein name</th>
<th>Organism</th>
<th>Component cluster</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>O08522</td>
<td>B11611_HELPF</td>
<td>Isozyme</td>
<td>lactic dehydrogenase (NADP)</td>
<td>Helicobacter pylori (strain P10)</td>
<td>U000011_HELPF</td>
<td>425</td>
</tr>
<tr>
<td>O09883</td>
<td>B11611_HELPF</td>
<td>Isozyme</td>
<td>lactic dehydrogenase (NADP)</td>
<td>Helicobacter pylori (strain P10)</td>
<td>U000011_HELPF</td>
<td>425</td>
</tr>
<tr>
<td>O11584</td>
<td>B11611_HELPF</td>
<td>Isozyme</td>
<td>lactic dehydrogenase (NADP)</td>
<td>Helicobacter pylori (strain P10)</td>
<td>U000011_HELPF</td>
<td>425</td>
</tr>
<tr>
<td>O12006</td>
<td>B11611_HELPF</td>
<td>Isozyme</td>
<td>lactic dehydrogenase (NADP)</td>
<td>Helicobacter pylori (strain P10)</td>
<td>U000011_HELPF</td>
<td>425</td>
</tr>
<tr>
<td>S01106</td>
<td>B11611_HELPF</td>
<td>Isozyme</td>
<td>lactic dehydrogenase (NADP)</td>
<td>Helicobacter pylori (strain P10)</td>
<td>U000011_HELPF</td>
<td>425</td>
</tr>
</tbody>
</table>

KEGG (Kyoto Encyclopedia of Genes and Genomes)

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

KEGG Pathway

CITRATE CYCLE (TCA CYCLE)

Isocitrate dehydrogenase
Isocitrate Dehydrogenase in KEGG

MASCOT—Protein Identification from Mass Spectroscopy Data

- Peptide Mass Fingerprinting
- Sequence Query
- MS/MS Ion Search
MRM-Path

MRMPath ...
MRMPath results for isocitrate dehydrogenase

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Start</th>
<th>Length (bp)</th>
<th>Parent Ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHP1</td>
<td>1237</td>
<td>1258</td>
<td></td>
</tr>
<tr>
<td>SHP2</td>
<td>1320</td>
<td>1329</td>
<td></td>
</tr>
<tr>
<td>MRP1</td>
<td>1028</td>
<td>1044</td>
<td></td>
</tr>
<tr>
<td>MRP2</td>
<td>1157</td>
<td>1162</td>
<td></td>
</tr>
<tr>
<td>MRP2</td>
<td>1202</td>
<td>1207</td>
<td></td>
</tr>
<tr>
<td>MRMP</td>
<td>1204</td>
<td>1207</td>
<td></td>
</tr>
<tr>
<td>MRA</td>
<td>1210</td>
<td>1217</td>
<td></td>
</tr>
<tr>
<td>MRMP</td>
<td>1213</td>
<td>1217</td>
<td></td>
</tr>
<tr>
<td>MRA</td>
<td>1215</td>
<td>1217</td>
<td></td>
</tr>
<tr>
<td>MRMP</td>
<td>1217</td>
<td>1217</td>
<td></td>
</tr>
</tbody>
</table>

MRM-Mutation

- Analysis of Protein Mutations

MRMPath is a methodology that allows the user to select individual proteins and determine whether they have known mutations. This is determined by examining the MRMP path database. Each of the protein sequences is subjected to multiple sequence alignment which is then used to select the best model for each protein. Once the model is selected, it is used to identify the best representation of the mutation. The user can then select the mutation to be explored and the computer software will generate a result. The result is a list of all mutations that are present in the protein.
Mass Spectrometry Tools—EXPASY

http://www.expasy.org/resources/search/keywords:mass%20spectrometry

Interesting Papers—Mass Spectrometry and Bioinformatics

- http://www.ingentaconnect.com/content/ben/cbio/2012/00000007/0000001/art00010
Protein Data Bank-PDB

- [http://www.rcsb.org/pdb/home/home.do](http://www.rcsb.org/pdb/home/home.do)
- “A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the [wwPDB](http://www.rcsb.org/pdb/home/home.do), the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.”

Problems during Protein Identification

- No sequence in database --- nothing to correlate with
- Problems with entries in database: human errors in entering information (typographical errors and curation); sequencing errors; errors during transcription
- Modifications in large proteins: degradation, oxidation of methionine, deamidation of N and Q, remember glycosylations, phosphorylations, and acetylations ….

[http://www.unimod.org/](http://www.unimod.org/) lists the possible modifications that can occur