Web-based Bioinformatics (Proteomics) Applications

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

Philosophical underpinnings …

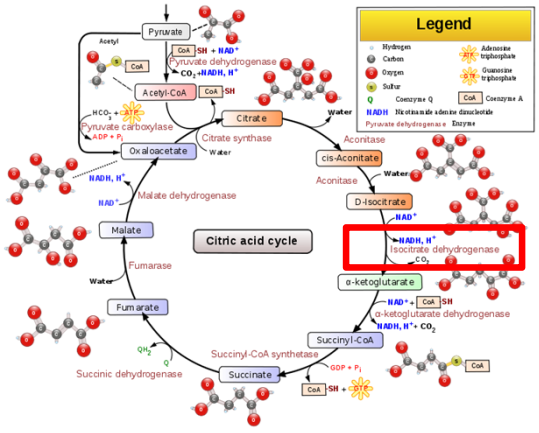
- Bioinformatics is here to stay—simply because computers are part of everyday life. This is not going to change in the near or distant future
- Students, researchers, etc., will be better served embracing bioinformatics ideas even if they do not necessarily want to pursue bioinformatics-driven careers, 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

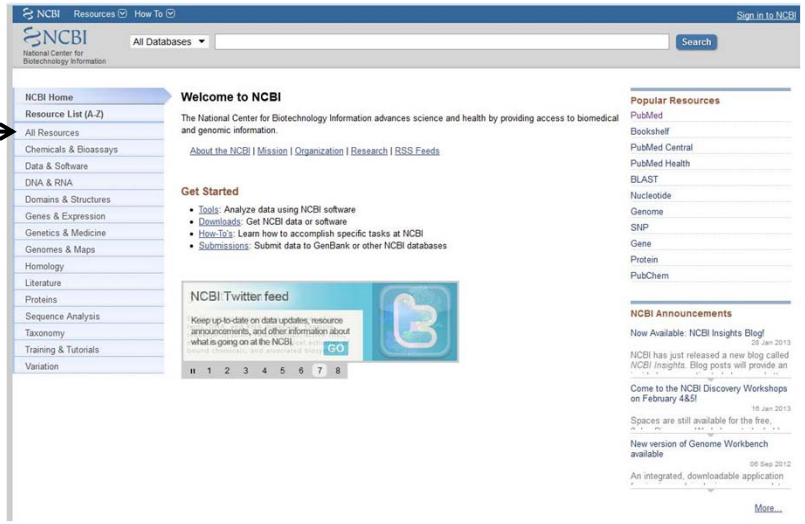
- **Isocitrate dehydrogenase** (EC 1.1.1.42) and (EC 1.1.1.41), also known as **IDH**, is an enzyme that participates in the citric acid cycle. It catalyzes the third step of the cycle: the oxidative decarboxylation of isocitrate, producing alpha-ketoglutarate (α-ketoglutarate) and CO_2 while converting NAD^+ to NADH.



http://en.wikipedia.org/wiki/File:Citric_acid_cycle_with_aconitate_2.svg

NCBI (National Center for Biotechnology Information)

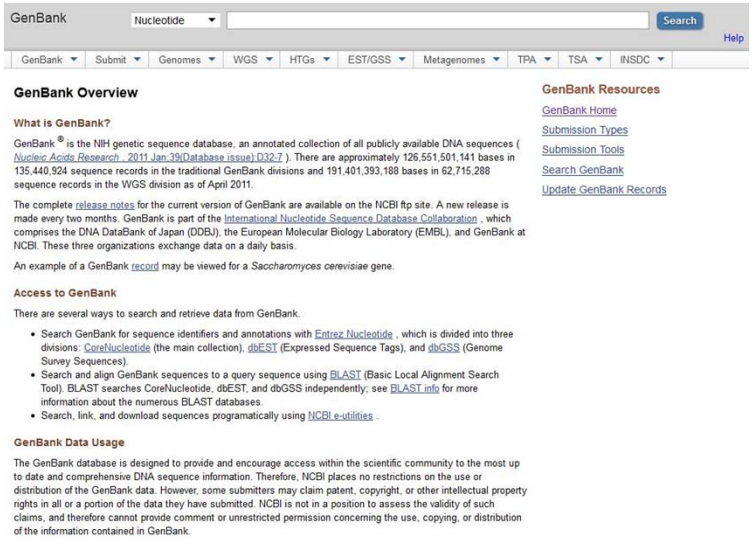
http://www.ncbi.nlm.nih.gov/



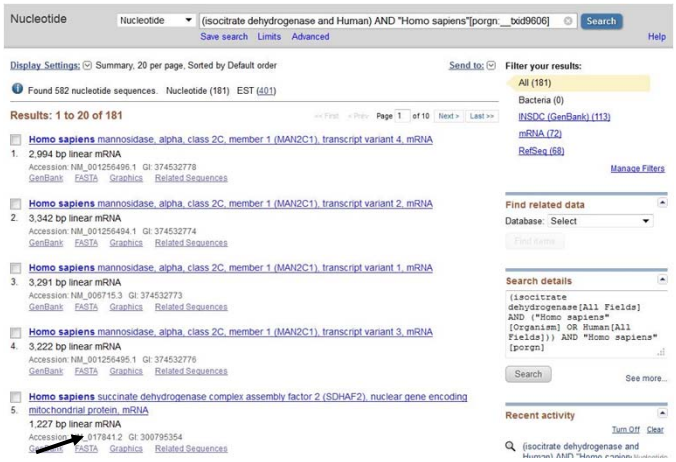
Selected Applications through NCBI

- 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)

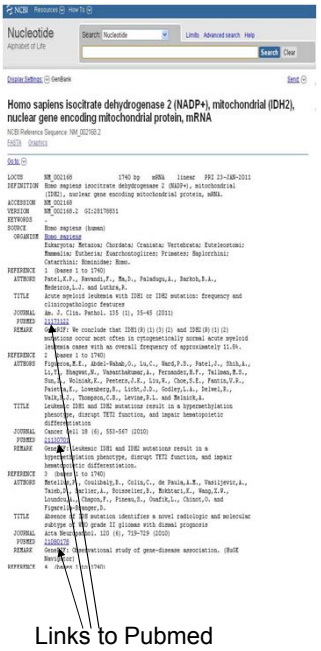


Nucleotide-Genbank's gene repository



Accession Number

A Nucleotide Entry in Genbank

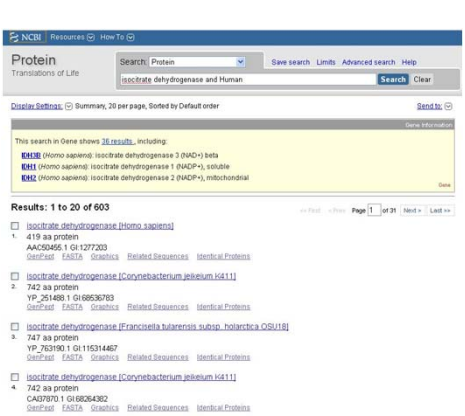


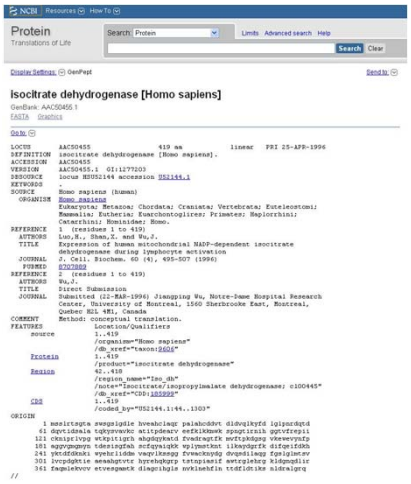
Protein Sequence

Links to Pubmed

Gene Sequence

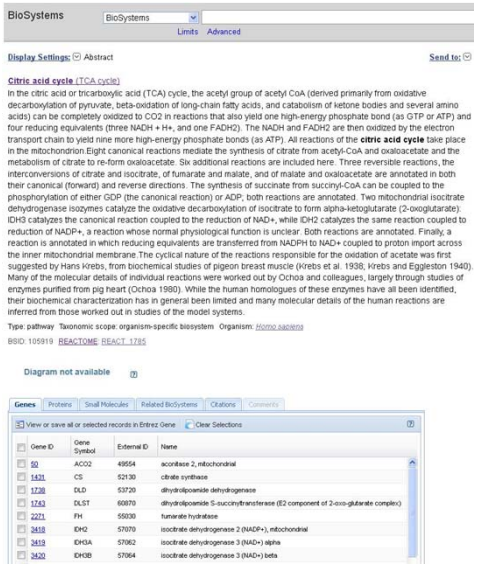
Protein Sequence in Genbank (isocitrate dehydrogenase)



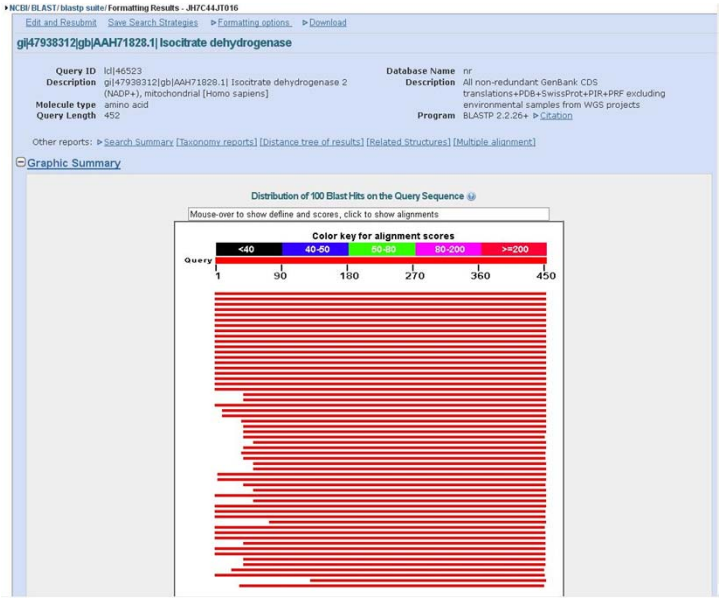


Note that the protein sequence and the rest of the entries are formatted similar to that of the nucleotide sequences in Genbank.

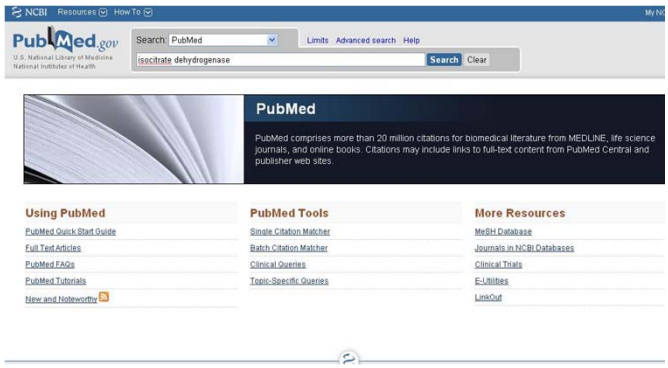
BioSystems



BLAST Results

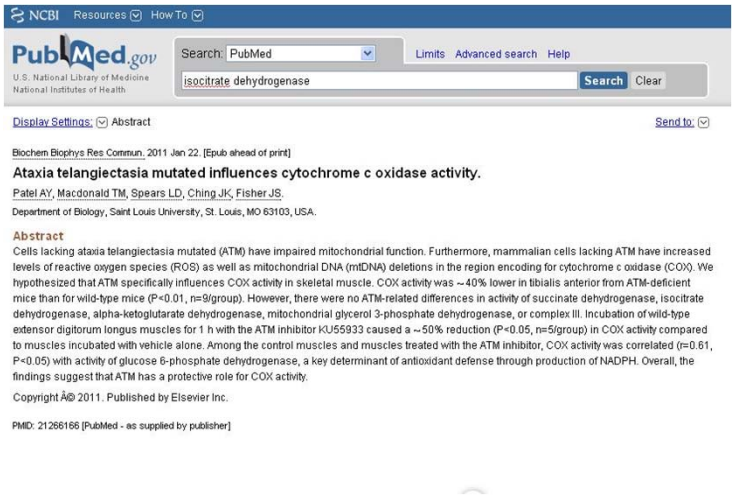


Pubmed—repository of biomedical abstracts



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

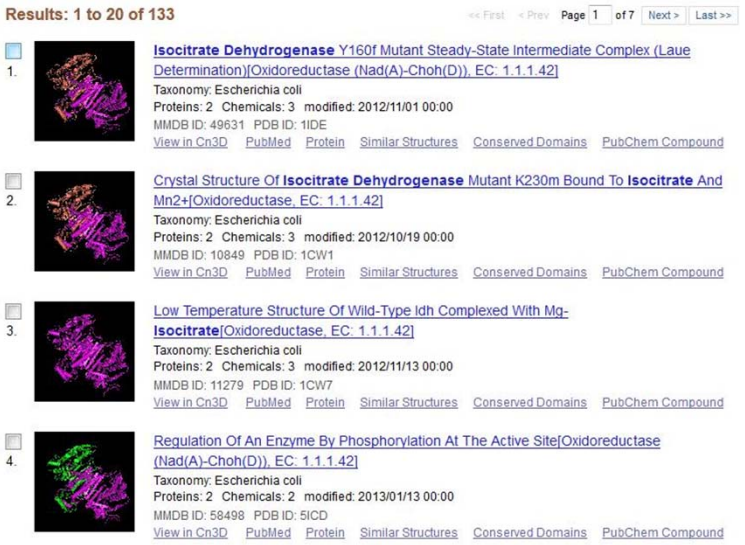
A single Abstract in Pubmed



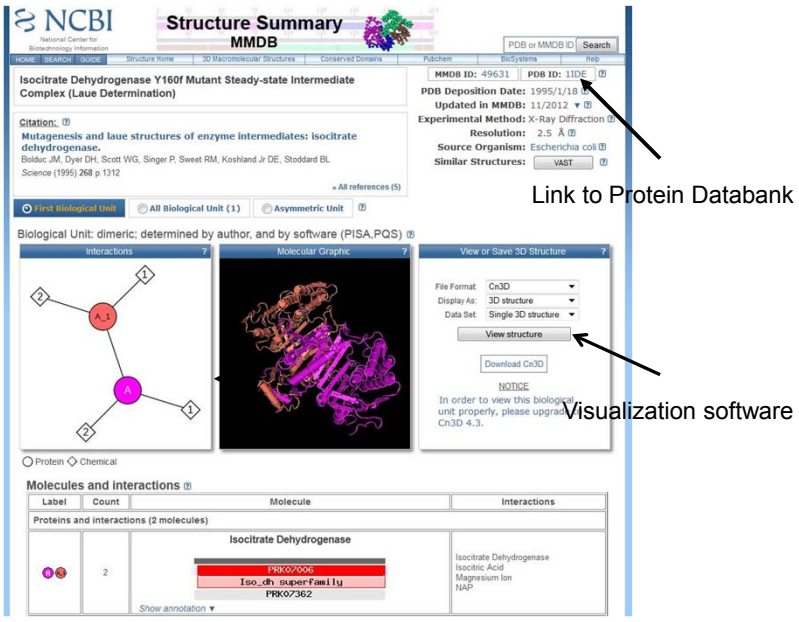
Computational Resources from NCBI's Structure Group

http://www.ncbi.nlm.nih.gov/Structure/index.shtml

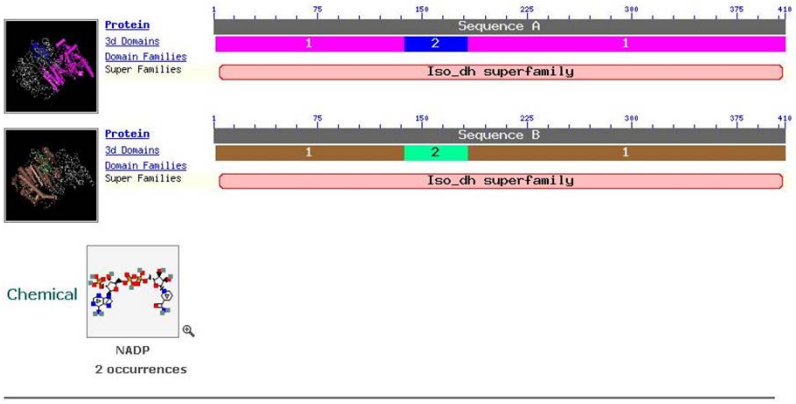
Three-dimensional structure views in Genbank--STRUCTURE



Structure of Actin—Genbank Structure View



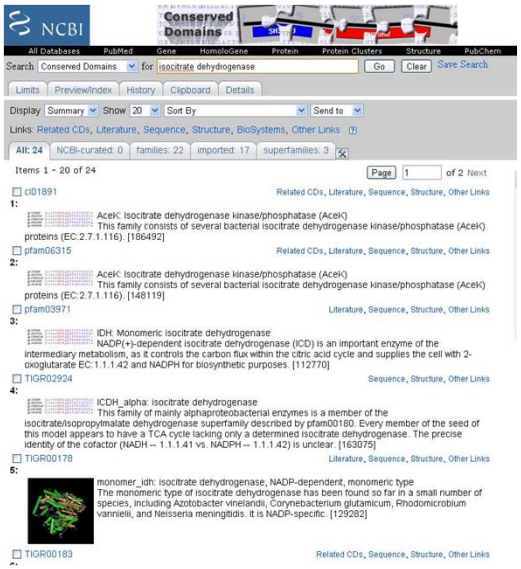
Structure of Domains in Genbank



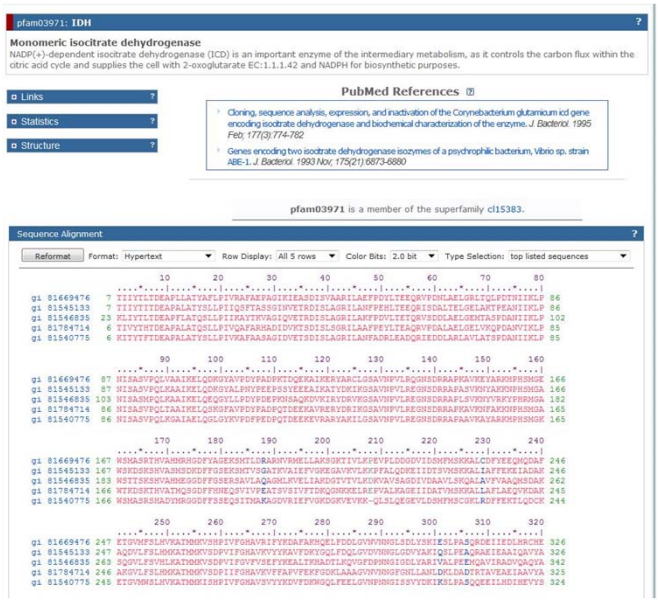
List of domains related to or associated with Isocitrate Dehydrogenase

Conserved domain database (CDD) in Genbank

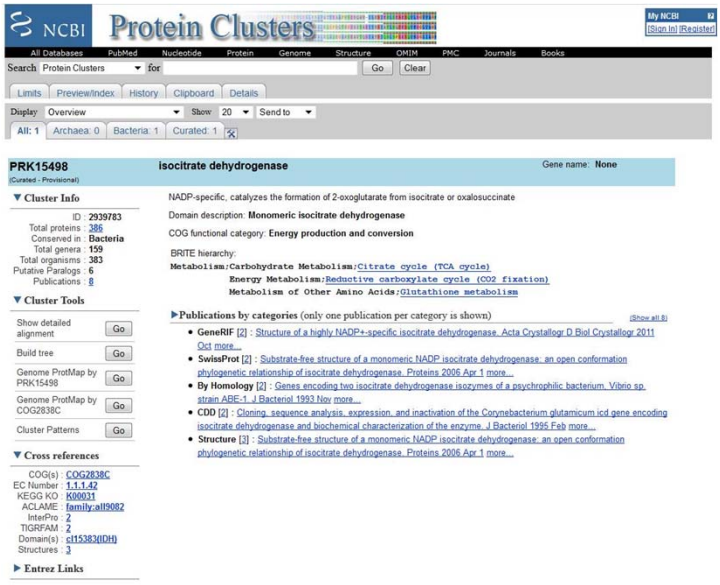
CDD …


NCBI
Conserved Domains
All Databases PubMed Gene HomoloGene Protein Protein Clusters Structure PubChem
Search Conserved Domains for isocitrate dehydrogenase Go Clear Save Search
Limits Preview/Index History Clipboard Details
Display Summary Show 20 Sort By Send to
Links: Related CDs, Literature, Sequence, Structure, BioSystems, Other Links
All: 24 NCBI-curated: 0 families: 22 imported: 17 superfamilies: 3
Items 1 - 20 of 24
Page 1 of 2 Next
cl01891 Related CDs, Literature, Sequence, Structure, Other Links
1:
AceK: Isocitrate dehydrogenase kinase/phosphatase (AceK)
This family consists of several bacterial isocitrate dehydrogenase kinase/phosphatase (AceK) proteins (EC:2.7.1.116). [186492]
pfam06315 Related CDs, Literature, Sequence, Structure, Other Links
2:
AceK: Isocitrate dehydrogenase kinase/phosphatase (AceK)
This family consists of several bacterial isocitrate dehydrogenase kinase/phosphatase (AceK) proteins (EC:2.7.1.116). [148119]
pfam03971 Literature, Sequence, Structure, Other Links
3:
IDH: Monomeric isocitrate dehydrogenase
NADP(+)-dependent isocitrate dehydrogenase (ICD) is an important enzyme of the intermediary metabolism, as it controls the carbon flux within the citric acid cycle and supplies the cell with 2-oxoglutarate EC:1.1.1.42 and NADPH for biosynthetic purposes. [112770]
TIGR02924 Sequence, Structure, Other Links
4:
ICDH_alpha: isocitrate dehydrogenase
This family of mainly alphaproteobacterial enzymes is a member of the isocitrate/isopropylmalate dehydrogenase superfamily described by pfam00180. Every member of the seed of this model appears to have a TCA cycle lacking only a determined isocitrate dehydrogenase. The precise identity of the cofactor (NADH -- 1.1.1.41 vs. NADPH -- 1.1.1.42) is unclear. [163075]
TIGR00178 Literature, Sequence, Structure, Other Links
5:
monomer_idh: isocitrate dehydrogenase, NADP-dependent, monomeric type
The monomeric type of isocitrate dehydrogenase has been found so far in a small number of species, including Azotobacter vinelandii, Corynebacterium glutamicum, Rhodomicrobium vannielii, and Neisseria meningitidis. It is NADP-specific. [129282]
TIGR00183 Related CDs, Sequence, Structure, Other Links

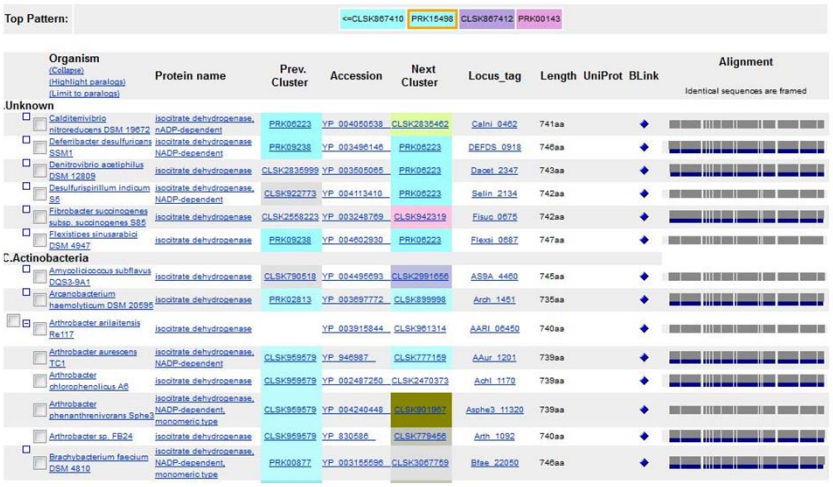
CDD…


pfam03971: IDH
Monomeric isocitrate dehydrogenase
NADP(+)-dependent isocitrate dehydrogenase (ICD) is an important enzyme of the intermediary metabolism, as it controls the carbon flux within the citric acid cycle and supplies the cell with 2-oxoglutarate EC:1.1.1.42 and NADPH for biosynthetic purposes.
Links
Statistics
Structure
PubMed References
Cloning, sequence analysis, expression, and inactivation of the Corynebacterium glutamicum icd gene encoding isocitrate dehydrogenase and biochemical characterization of the enzyme. J. Bacteriol. 1995 Feb; 177(3):774-782
Genes encoding two isocitrate dehydrogenase isozymes of a psychrophilic bacterium, Vibrio sp. strain ABE-1. J. Bacteriol. 1993 Nov; 175(21):6873-6880
pfam03971 is a member of the superfamily cl15383.
Sequence Alignment
Reformat Format: Hypertext Row Display: All 5 rows Color Bits: 2.0 bit Type Selection: top listed sequences

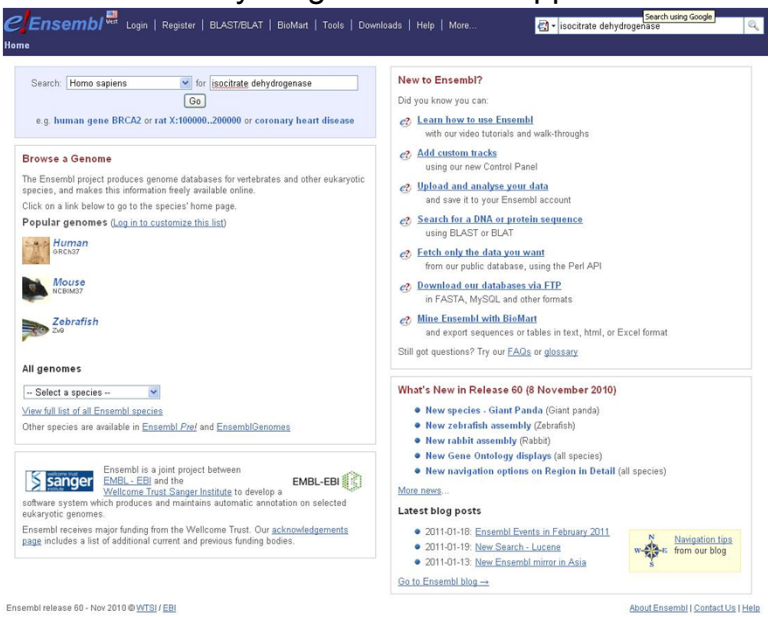
Clustering Proteins in terms of Sequence Similarities--Genbank



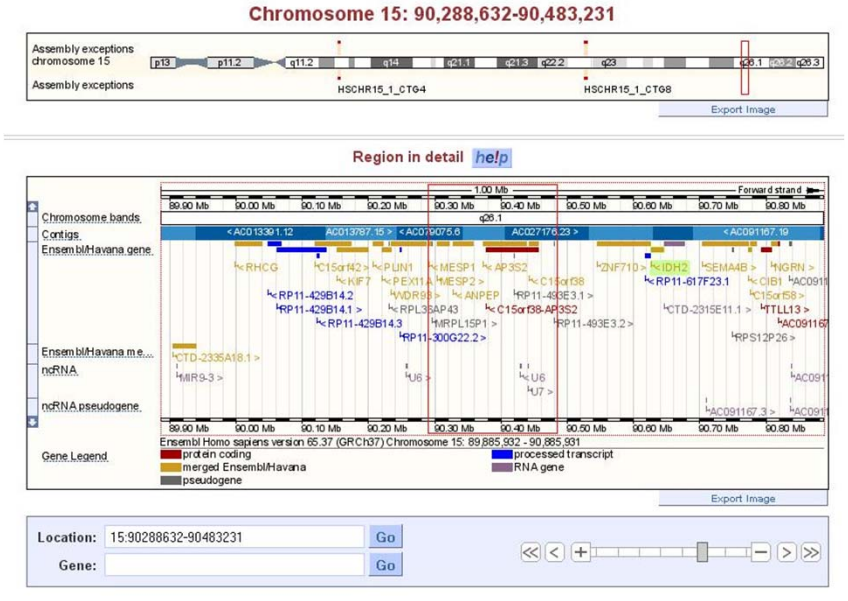
Clustering Proteins in terms of Sequence Similarities--Genbank



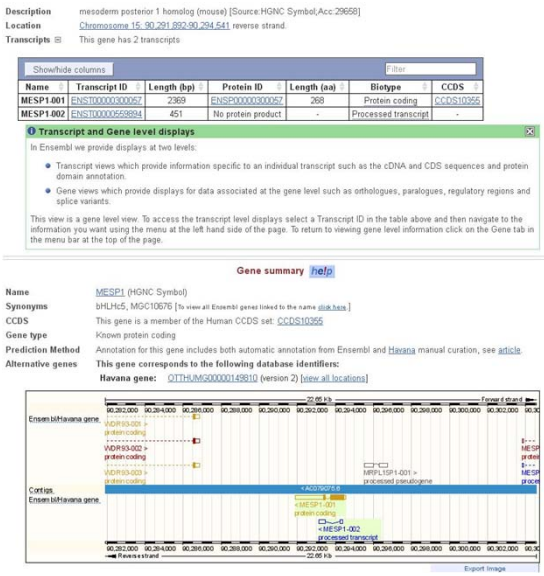
ENSEMBL—European version of Genbank—now focused exclusively on genome wide applications



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



ENSEMBL—Gene Summary



Description mesoderm posterior 1 homolog (mouse) [Source:HGNC Symbol;Acc:29658]
Location Chromosome 15: 90,291,892-90,294,541 reverse strand.
Transcripts This gene has 2 transcripts

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
MESP1-001	ENST00000300057	2369	ENSP00000300057	268	Protein coding	CCDS10355
MESP1-002	ENST00000559894	451	No protein product	-	Processed transcript	-

Transcript and Gene level displays

In Ensembl we provide displays at two levels:

- Transcript views which provide information specific to an individual transcript such as the cDNA and CDS sequences and protein domain annotation.
- Gene views which provide displays for data associated at the gene level such as orthologues, paralogues, regulatory regions and splice variants.

This view is a gene level view. To access the transcript level displays select a Transcript ID in the table above and then navigate to the information you want using the menu at the left hand side of the page. To return to viewing gene level information click on the Gene tab in the menu bar at the top of the page.

Gene summary *help*

Name MESP1 (HGNC Symbol)
Synonyms bHLHc5, MGC10676 [To view all Ensembl genes linked to the name click here.]
CCDS This gene is a member of the Human CCDS set: CCDS10355
Gene type Known protein coding
Prediction Method Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.
Alternative genes **This gene corresponds to the following database identifiers:**
Havana gene: OTTHUMG00000149810 (version 2) [view all locations]

ENSEMBL—Protein

Transcript: MESP1-001 ENST00000300057

Description mesoderm posterior 1 homolog (mouse) [Source:HGNC Symbol;Acc:29658]
Location Chromosome 15: 90,291,892-90,294,541 reverse strand.
Gene This transcript is a product of gene ENSG00000166823 - This gene has 2 transcripts

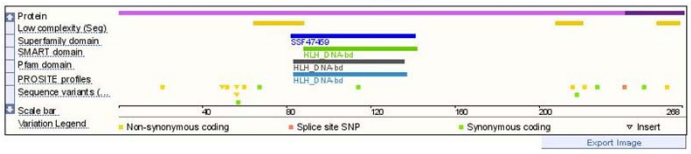
Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
MESP1-001	ENST00000300057	2369	ENSP00000300057	268	Protein coding	CCDS10355
MESP1-002	ENST00000559894	451	No protein product	-	Processed transcript	-

Transcript and Gene level displays

Views in Ensembl are separated into gene based views and transcript based views according to which level the information is more appropriately associated with. This view is a transcript level view. To flip between the two sets of views you can click on the Gene and Transcript tabs in the menu bar at the top of the page.

Protein summary *help*



Statistics
Ave. residue weight: 106.348 g/mol
Charge: 8.0
Isoelectric point: 9.0165
Molecular weight: 28,501.38 g/mol
Number of residues: 268 aa

SWISSPROT--http://www.expasy.ch/

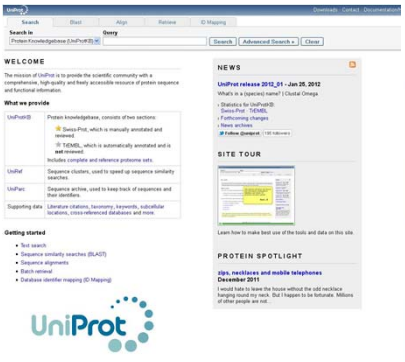
– UniProt combines SwissProt and TrEMBl

"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

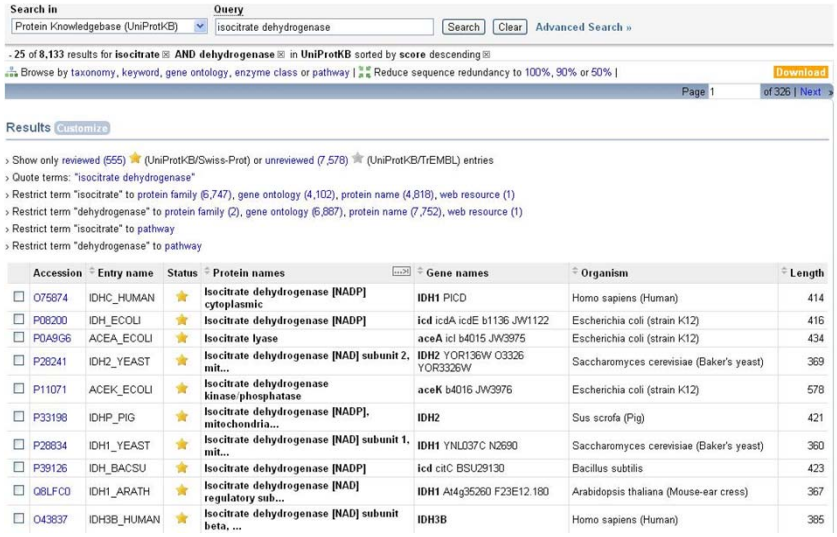
UniProt has replaced SwissProt

Mirro Sites
Switzerland: **http://www.expasy.org/** at Swiss Institute of Bioinformatics, Geneva
Australia: **http://au.expasy.org/** at Australian Proteome Analysis Facility, Sydney
Brazil: **http://br.expasy.org/** at Laboratório Nacional de Computação Científica, Petrópolis
Canada: **http://ca.expasy.org/** at Canadian Bioinformatics Resource, Halifax
China: **http://cn.expasy.org/** at Peking University
Korea: **http://kr.expasy.org/** at Yonsei Proteome Research Center, Seoul

UNIPROT SWISSPROT



SwissProt—search for Proteins



Search in: Protein Knowledgebase (UniProtKB) | Query: isocitrate dehydrogenase | Search | Clear | Advanced Search »

25 of 8,133 results for **isocitrate** AND **dehydrogenase** in **UniProtKB** sorted by **score** descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Download

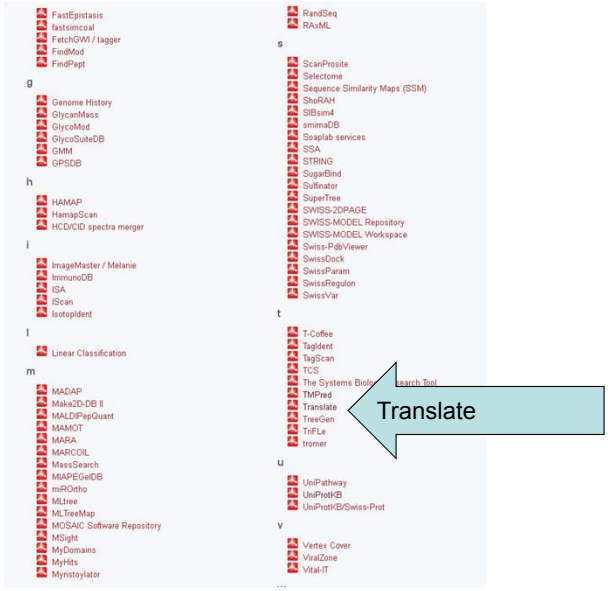
Page 1 of 326 | Next »

Results Customize

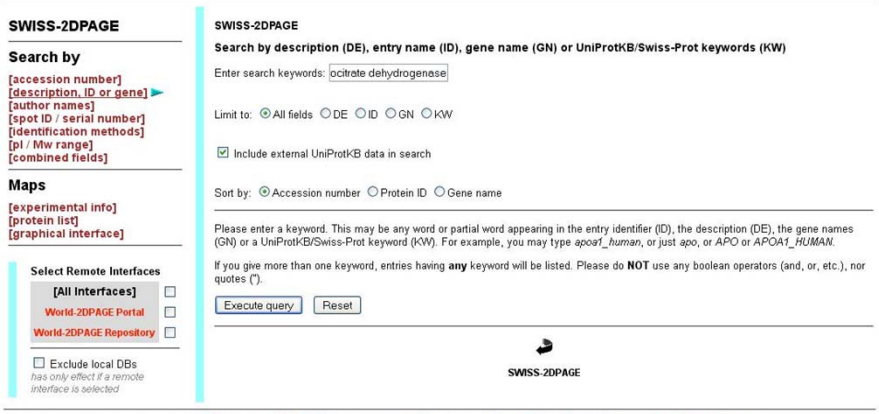
› Show only reviewed (555) (UniProtKB/Swiss-Prot) or unreviewed (7,578) (UniProtKB/TrEMBL) entries
› Quote terms: "isocitrate dehydrogenase"
› Restrict term "isocitrate" to protein family (6,747), gene ontology (4,102), protein name (4,818), web resource (1)
› Restrict term "dehydrogenase" to protein family (2), gene ontology (6,887), protein name (7,752), web resource (1)
› Restrict term "isocitrate" to pathway
› Restrict term "dehydrogenase" to pathway

Accession	Entry name	Status	Protein names	Gene names	Organism	Length
O75874	IDHC_HUMAN	★	**Isocitrate dehydrogenase [NADP] cytoplasmic**	**IDH1** PICD	Homo sapiens (Human)	414
P08200	IDH_ECOLI	★	**Isocitrate dehydrogenase [NADP]**	**icd** icdA icdE b1136 JW1122	Escherichia coli (strain K12)	416
P0A9G6	ACEA_ECOLI	★	**Isocitrate lyase**	**aceA** icl b4015 JW3975	Escherichia coli (strain K12)	434
P28241	IDH2_YEAST	★	**Isocitrate dehydrogenase [NAD] subunit 2, mit...**	**IDH2** YOR136W O3326 YOR3326W	Saccharomyces cerevisiae (Baker's yeast)	369
P11071	ACEK_ECOLI	★	**Isocitrate dehydrogenase kinase/phosphatase**	**aceK** b4016 JW3976	Escherichia coli (strain K12)	578
P33198	IDHP_PIG	★	**Isocitrate dehydrogenase [NADP], mitochondria...**	**IDH2**	Sus scrofa (Pig)	421
P28834	IDH1_YEAST	★	**Isocitrate dehydrogenase [NAD] subunit 1, mit...**	**IDH1** YNL037C N2690	Saccharomyces cerevisiae (Baker's yeast)	360
P39126	IDH_BACSU	★	**Isocitrate dehydrogenase [NADP]**	**icd** citC BSU29130	Bacillus subtilis	423
Q8LFC0	IDH1_ARATH	★	**Isocitrate dehydrogenase [NAD] regulatory sub...**	**IDH1** At4g35260 F23E12.180	Arabidopsis thaliana (Mouse-ear cress)	367
O43837	IDH3B_HUMAN	★	**Isocitrate dehydrogenase [NAD] subunit beta, ...**	**IDH3B**	Homo sapiens (Human)	385

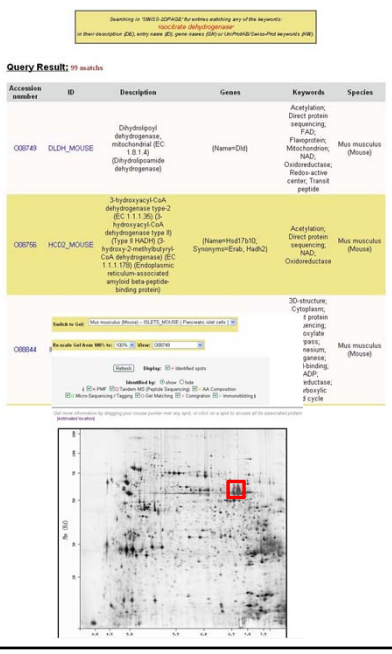
EXPASY-Databases and Features



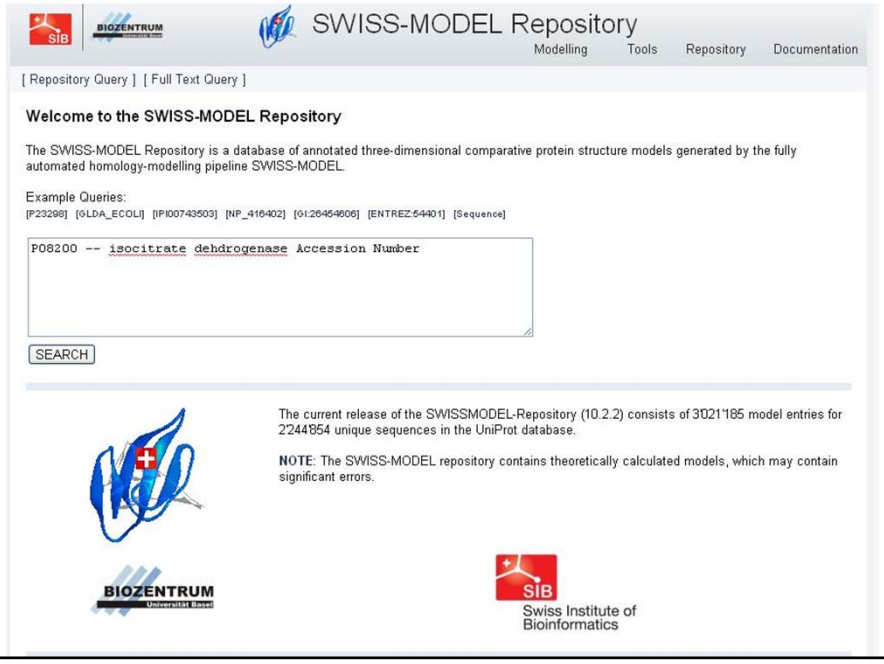
Swiss 2D-PAGE



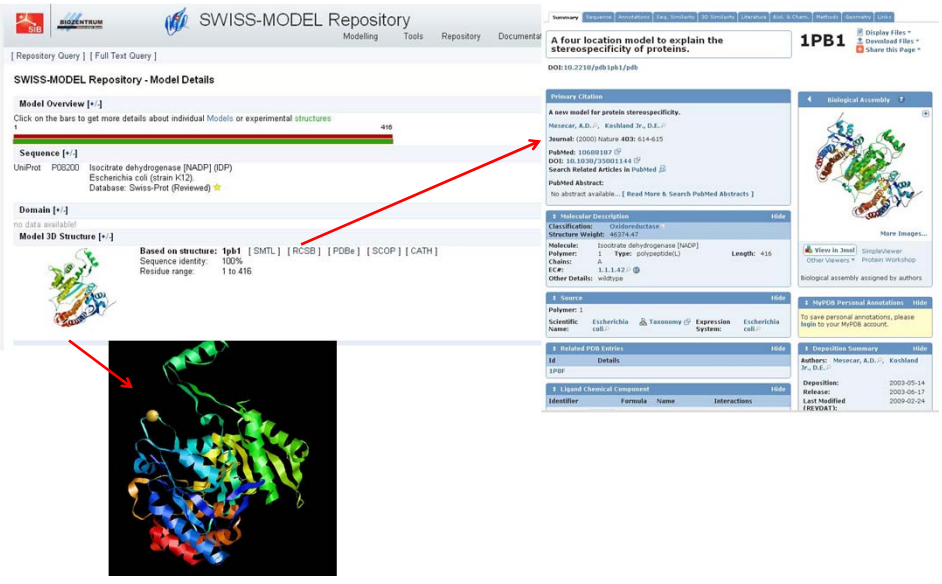
Swiss 2DPAGE –Isocitrate dehydrogenase



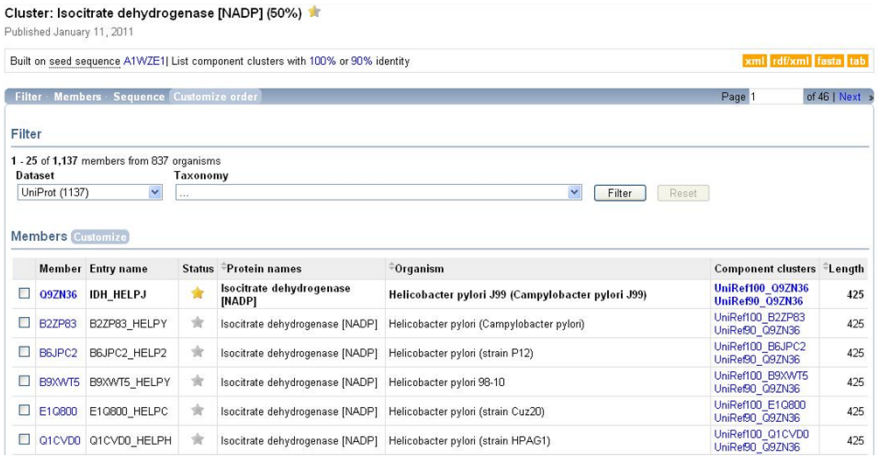
Swiss PDB Model Respository



SwissModel Repository …



Uniref—Clustering of Proteins



Cluster: Isocitrate dehydrogenase [NADP] (50%)

Published January 11, 2011

Built on seed sequence A1WZE1| List component clusters with 100% or 90% identity

xml rdf/xml fasta tab

Filter · Members · Sequence · Customize order

Page 1 of 46 | Next »

Filter

1 - 25 of 1,137 members from 837 organisms

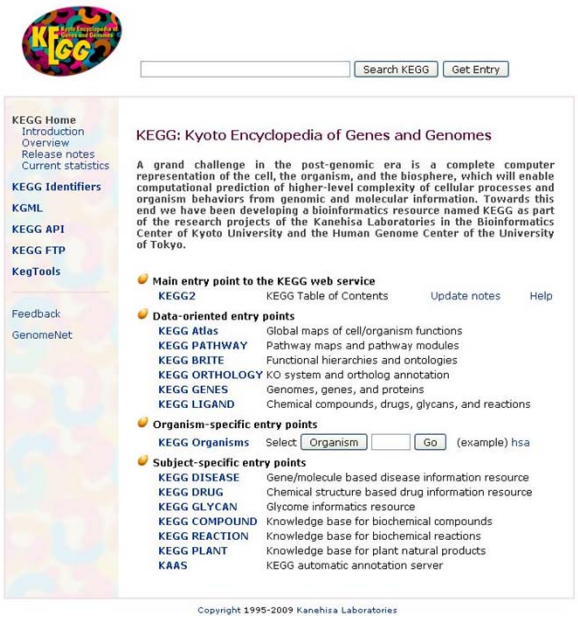
Dataset: UniProt (1137) Taxonomy: ... Filter Reset

Members Customize

Member	Entry name	Status	Protein names	Organism	Component clusters	Length
Q9ZN36	IDH_HELPJ	★	**Isocitrate dehydrogenase [NADP]**	**Helicobacter pylori J99 (Campylobacter pylori J99)**	**UniRef100_Q9ZN36 UniRef90_Q9ZN36**	425
B2ZP83	B2ZP83_HELPY	☆	Isocitrate dehydrogenase [NADP]	Helicobacter pylori (Campylobacter pylori)	UniRef100_B2ZP83 UniRef90_Q9ZN36	425
B6JPC2	B6JPC2_HELP2	☆	Isocitrate dehydrogenase [NADP]	Helicobacter pylori (strain P12)	UniRef100_B6JPC2 UniRef90_Q9ZN36	425
B9XWT5	B9XWT5_HELPY	☆	Isocitrate dehydrogenase [NADP]	Helicobacter pylori 98-10	UniRef100_B9XWT5 UniRef90_Q9ZN36	425
E1Q800	E1Q800_HELPC	☆	Isocitrate dehydrogenase [NADP]	Helicobacter pylori (strain Cuz20)	UniRef100_E1Q800 UniRef90_Q9ZN36	425
Q1CVD0	Q1CVD0_HELPH	☆	Isocitrate dehydrogenase [NADP]	Helicobacter pylori (strain HPAG1)	UniRef100_Q1CVD0 UniRef90_Q9ZN36	425

KEGG (Kyoto Encyclopedia of Genes and Genomes)

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



Search KEGG Get Entry

KEGG Home
Introduction
Overview
Release notes
Current statistics
KEGG Identifiers
KGML
KEGG API
KEGG FTP
KegTools
Feedback
GenomeNet

KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

Main entry point to the KEGG web service
KEGG2 KEGG Table of Contents Update notes Help

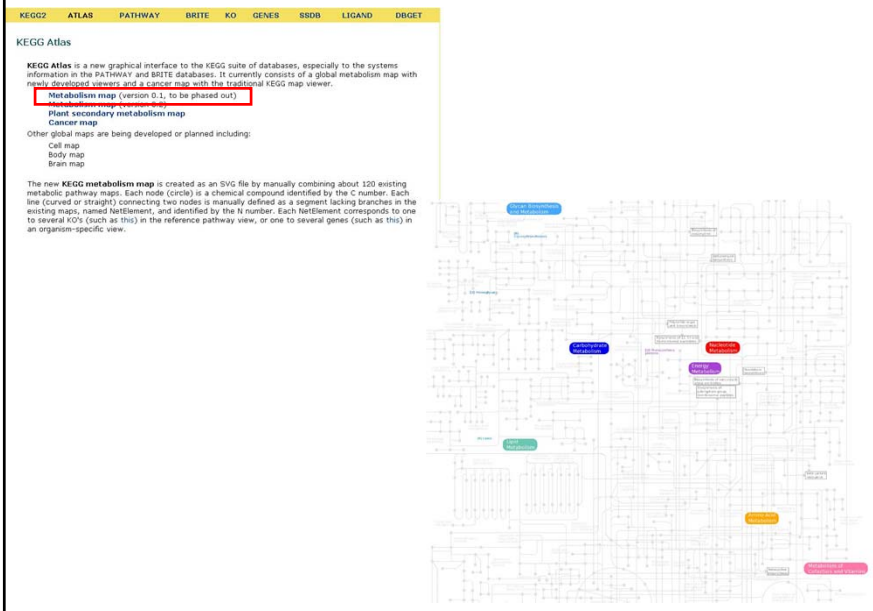
Data-oriented entry points
KEGG Atlas Global maps of cell/organism functions
KEGG PATHWAY Pathway maps and pathway modules
KEGG BRITE Functional hierarchies and ontologies
KEGG ORTHOLOGY KO system and ortholog annotation
KEGG GENES Genomes, genes, and proteins
KEGG LIGAND Chemical compounds, drugs, glycans, and reactions

Organism-specific entry points
KEGG Organisms Select Organism Go (example) hsa

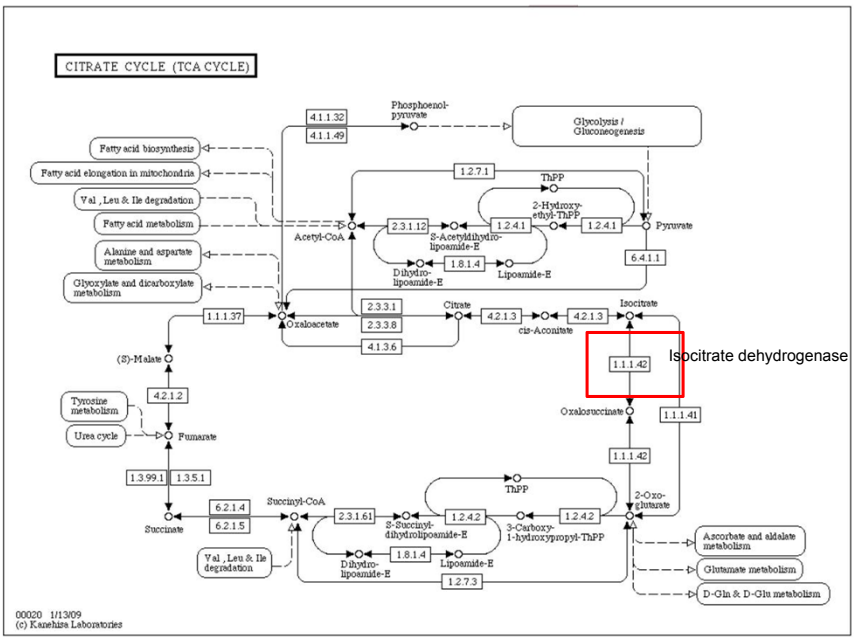
Subject-specific entry points
KEGG DISEASE Gene/molecule based disease information resource
KEGG DRUG Chemical structure based drug information resource
KEGG GLYCAN Glycome informatics resource
KEGG COMPOUND Knowledge base for biochemical compounds
KEGG REACTION Knowledge base for biochemical reactions
KEGG PLANT Knowledge base for plant natural products
KAAS KEGG automatic annotation server

Copyright 1995-2009 Kanehisa Laboratories

Kegg Atlas



KEGG Pathway



Isocitrate Dehydrogenase in KEGG

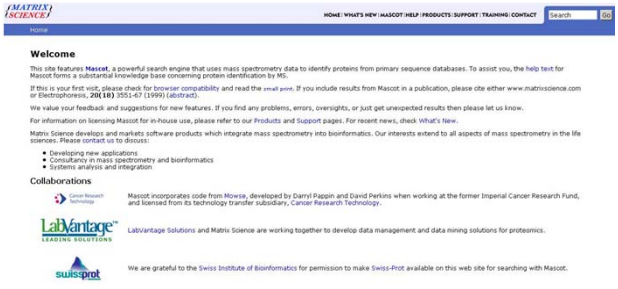
KEGG Homo sapiens (human): 3417 Help

Entry	3417 CDS H.sapiens
Gene name	IDH1, IDCD, IDH, IDP, IDPC, PICD
Definition	isocitrate dehydrogenase 1 (NADP+), soluble (EC:1.1.1.42)
Orthology	K00031 isocitrate dehydrogenase [EC:1.1.1.42]
Pathway	hsa00020 Citrate cycle (TCA cycle) hsa00480 Glutathione metabolism hsa01100 Metabolic pathways hsa04146 Peroxisome
Class	Metabolism; Carbohydrate Metabolism; Citrate cycle (TCA cycle) [PATH:hsa00020] Metabolism; Metabolism of Other Amino Acids; Glutathione metabolism [PATH:hsa00480] Cellular Processes; Transport and Catabolism; Peroxisome [PATH:hsa04146] BRITE hierarchy
SSDB	Ortholog Paralog GFIT
Motif	Pfam: Iso_dh PGA2 DUF505 PROSITE: IDH_IMDH Motif
Other DBs	NCBI-GI: 28178825 NCBI-GeneID: 3417 OMIM: 147700 HGNC: 5382 HPRD: 00984 Ensembl: ENSG00000138413 UniProt: O75874 Q6FHQ6
Structure	PDB: 3INM 1TOL 1T09 3MAP 3MAS 3MAR Thumbnails Jmol
Position	2q33.3

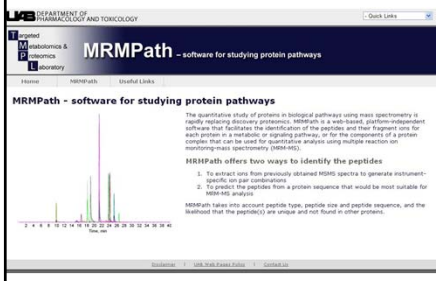
…

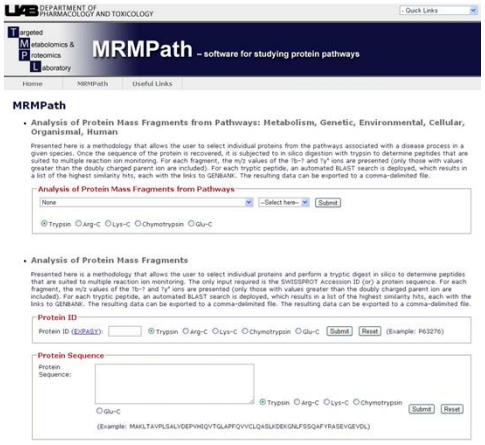
MASCOT—Protein Identification from Mass Spectroscopy Data

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

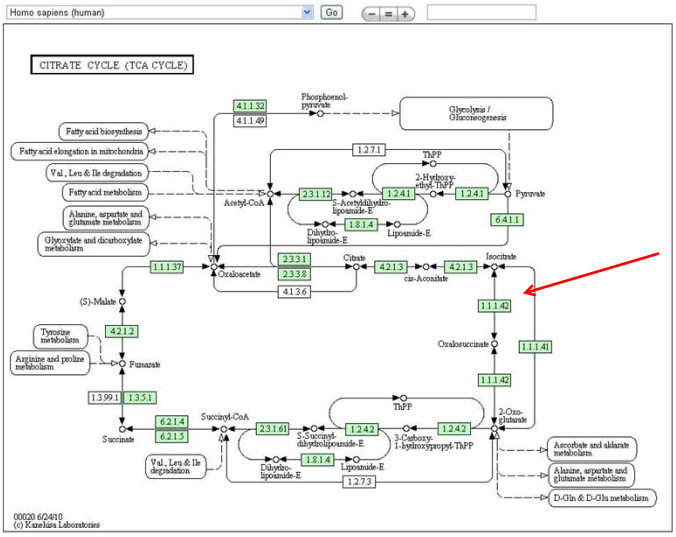


MRM-Path





MRMPath …



Isocitrate dehydrogenase

MRMPath results for isocitrate dehydrogenase

Click here to download this into an Excel sheet
NOTE: Please click on the 'YES' button if a warning appears when you try to open the excel sheet

hsa:3417 IDH1, IDCD, IDH, IDP, IDPC, PICD; isocitrate dehydrogenase 1 (NADP+), soluble (EC:1.1.1.42); K00031 isocitrate dehydrogenase [EC:1.1.1.42] (A)

BLAST ALL FRAGMENTS

Sequence	m/z Parent Ion	B Ion Mass	Y > Parent Ions
BLAST IIWELIK	457.792	542.2979	801.4921
		655.3819	688.4080
		768.4660	502.3287
		896.5609	
BLAST LIFPYVELDLHSYDLGIENR	1203.6235	1203.6666	2293.1551
		1340.7256	2180.0710
		1427.7576	2033.0026
		1590.8209	1935.9499
		1705.8479	1772.8865
		1818.9320	1673.8181
		1875.9535	1544.7755
		1989.0376	1431.6915
		2118.0801	1316.6645
		2232.1231	
		2388.2241	
BLAST DATNDQVTK	496.2411	517.1894	876.4474
		645.2480	805.4103
		744.3164	704.3626
		845.3641	590.3197
		973.4591	
BLAST DAAEAIK	359.1954	387.1516	602.3560
		458.1887	531.3189
		571.2728	460.2818
		699.3678	
BLAST SPNGTIR	372.7065	457.2047	657.3731
		570.2887	560.3204
		726.3898	446.2774
			389.2560

....

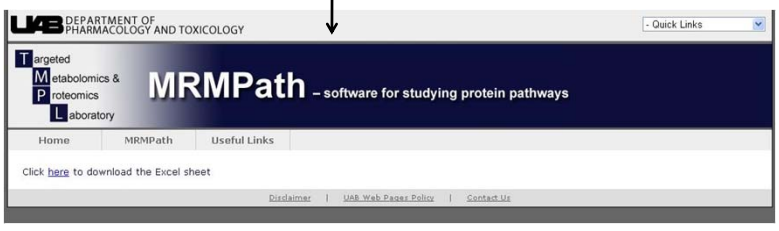
MRM-Mutation

- **Analysis of Protein Mutations**

MRMutation is a methodology that allows the user to select individual proteins and determine whether they have known mutations. This is determined by examining the EXPaSY.org database. Each of the protein sequence is subjected to trypsin digestion in silico to determine whether these peptides with mutations are suited to multiple reaction ion monitoring. The input required is the UNIPROT Accession ID. The output spreadsheet contains the m/z values of the first three 'b' and 'y' ions (only those with values greater than the doubly charged parent ion are included), the start and end residues of the peptide with respect to the parent protein and the mutation.

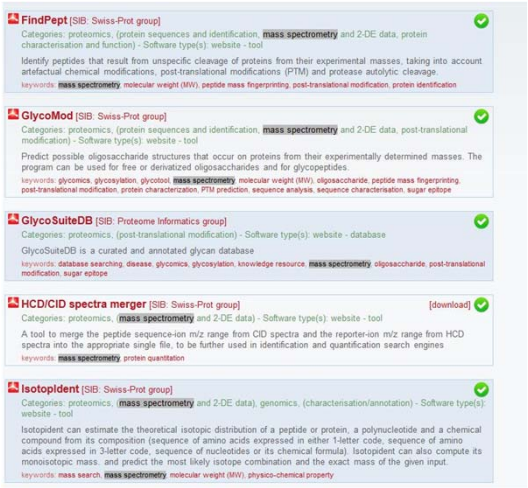
Protein ID

Protein ID (EXPASY): Submit Reset (Example: P04632)



Mass Spectrometry Tools–EXPASY

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



Interesting Papers—Mass Spectrometry and Bioinformatics

- http://masspec.scripps.edu/publications/public_pdf/72_art.pdf
- http://www.sciencedirect.com/science/article/pii/S0014579309002208
- http://www.ingentaconnect.com/content/ben/cbio/2012/00000007/00000001/art00010

Protein Data Bank-PDB

- 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, 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/ lists the possible modifications that can occur