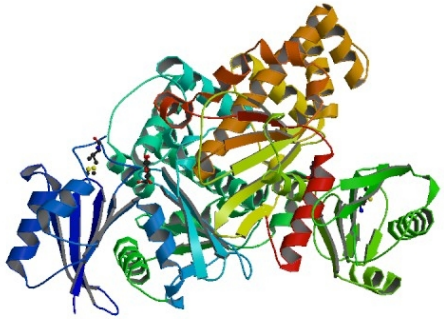
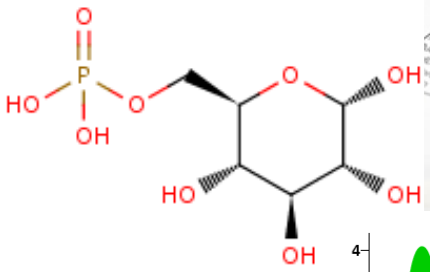
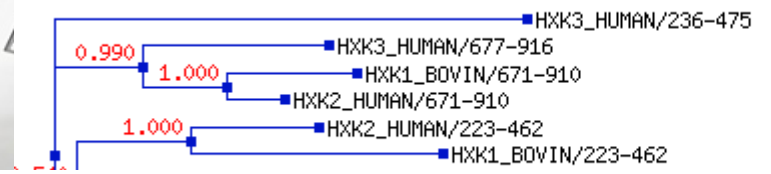
Biological Databases

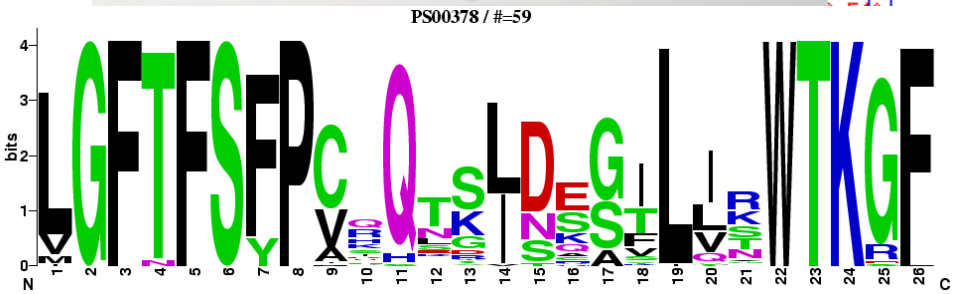
Matthias König
[05|05|10]

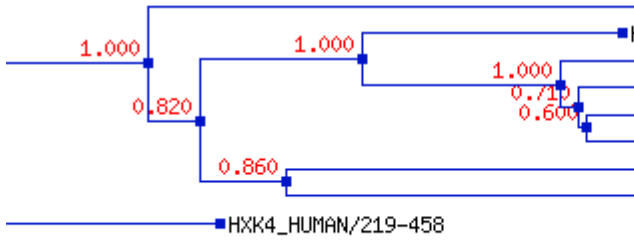


```
   1 ggacactaag ccccacagct caacacaacc aggagagaa
  61 gcgcccagca atggccctgc ctggagaaca tccaggctc
 121 gaatgcttgc cgactcgttg gagaacaatg aaaaggagg
 181 ccccaaacca gcccgaggag aaccacattc tcccaggga
 241 tgcggcggag aagccttgga tatttccact tcagaagcc
```









Outline biological databases

I. Introduction & Overview

II. Examples

III. Sequence alignment & fragment search

IV. Database tools and implementation

I Introduction and overview

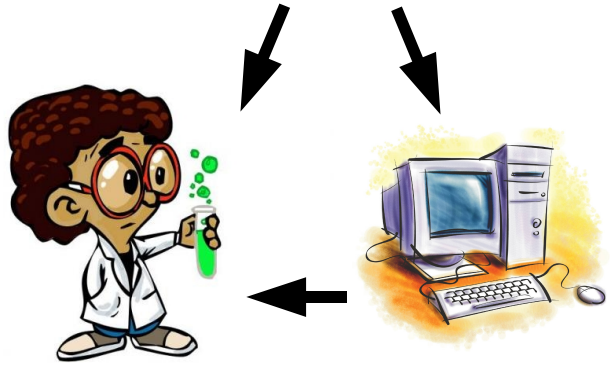
Why databases ?

- biology has turned into **data-rich science**
  - High-throughput genomics, proteomics, metabolomics, ...
  - Vast amount of data generated in experiments (like MS peptide fragments)
- need for storing and communicating large datasets has grown tremendously
  - archiving, curation, analysis and interpretation of all of these datasets are a challenge
  - convenient methods for proper storing, searching & retrieving necessary
- **Databases are the means** to handle this data overload

What can databases do ?

- **Make biological data available** ...

  1. … **to scientists**.
  2. … **in computer-readable form**.
     - Analysis (computer based)
     - Handle and share large volumes of data
     - Interface for computer based systems (Algorithms, Web interfaces)



- Store data
  - Defined formats
  - Automated storage and retrieval of experimental data
- Link knowledge with external resources

Database classification I

- **Type of data**
  - Nucleotide or protein sequences
  - Protein sequence patterns and motifs
  - Macromolecular 3D structures
  - Gene expression data
  - Metabolic pathways
  - ...
- **Data entry and quality control**
  - Scientists deposit data directly
  - Appointed curators add and update
  - Type and degree of error checking
  - Consistency, redundancy, conflicts, updates

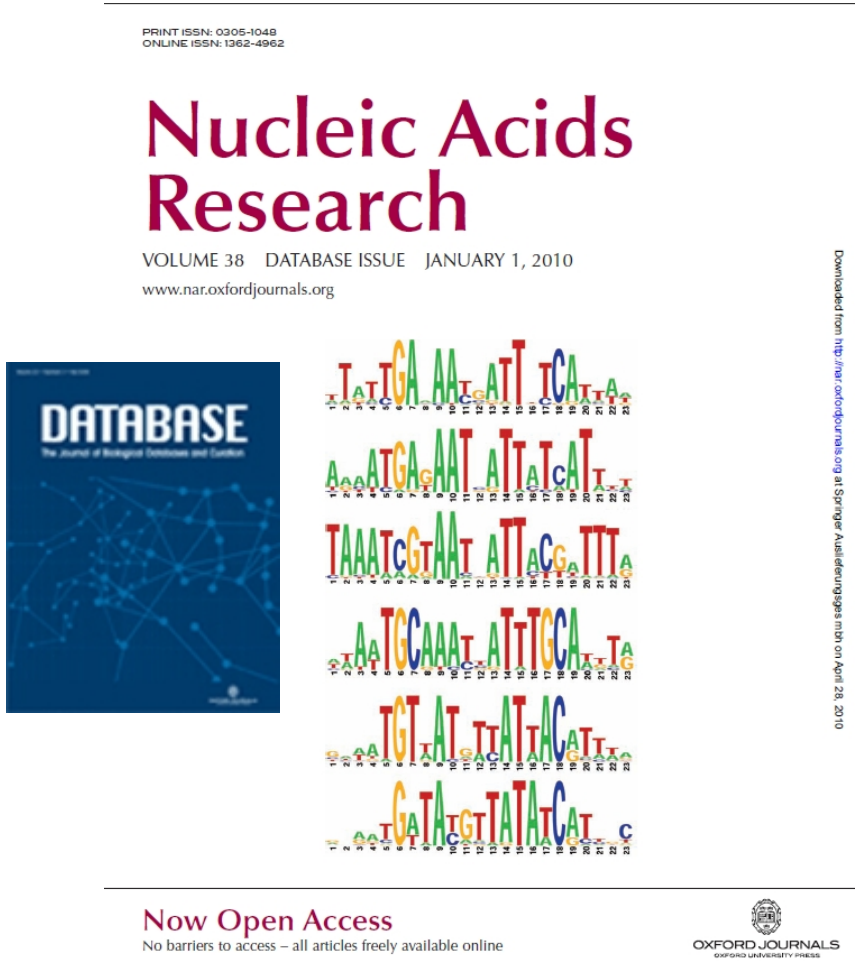
Database classification II

- **Primary or derived data**
  - Primary: experimental results directly into database
  - Secondary: results of analysis of primary databases
- **Technical design**
  - Flat-files
  - Relational database (SQL)
  - Object-oriented database
  - Exchange/publication technologies (FTP, HTML, COBRA, XML, SOAP)

- **Maintainer status**
  - Large, public institution funded by government (EMBL, NCBI)
  - Academic group or scientist
  - Commercial company

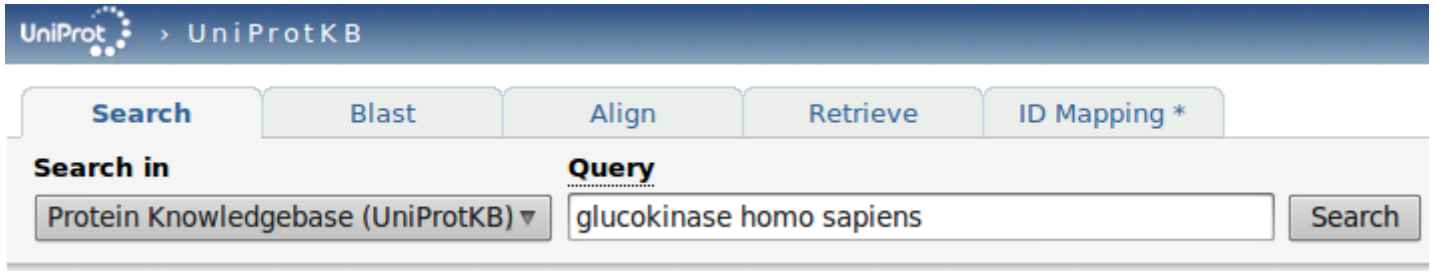
How to find my database ?

- Nucleic Acid Research offers database issue every year
- Database Journals
  - Database: The Journal of Biological Databases and Curation
- Database portals
  - DBD (database of biological database)
  - Pathguide
- Websearch
  - http://lmgtfy.com/



How to access the data ?

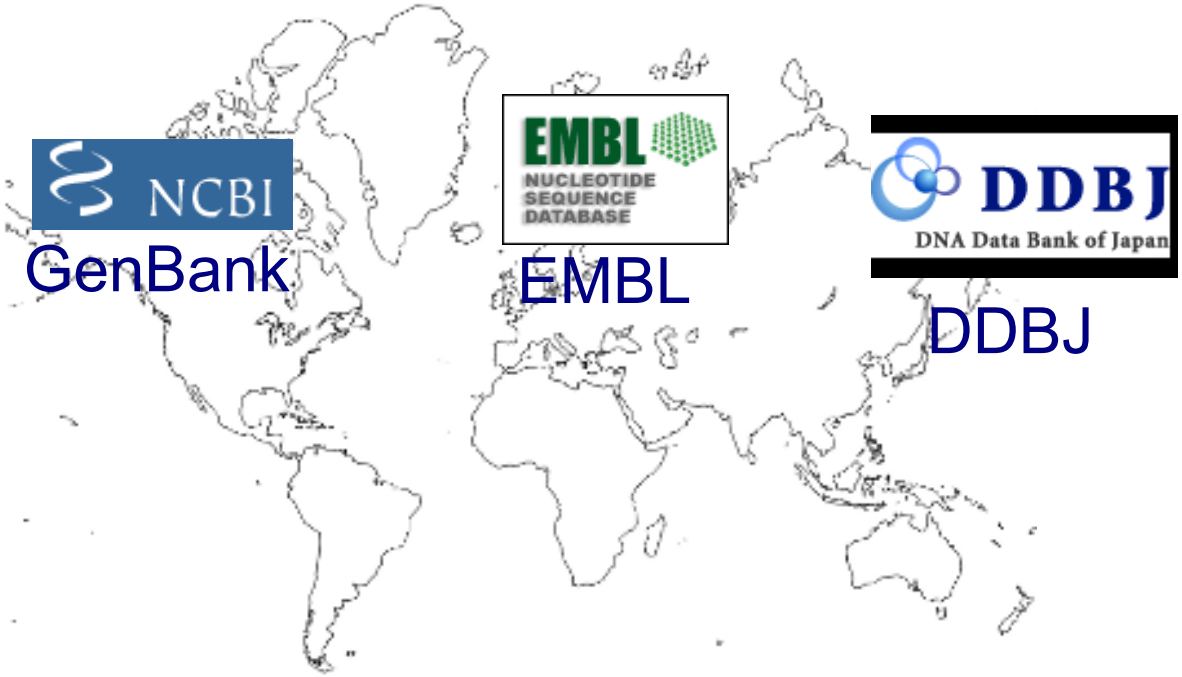
- Human **Web interface** (web based, small scale)
    - Common mode of search are keywords with modifiers or identifiers
    - Cross-references link the information of different databases



- **Web service** (SOAP, CORBA)
- **Flat files** (script based, large scale)
- **Database dump** (script based, large scale)

II Examples of biological databases

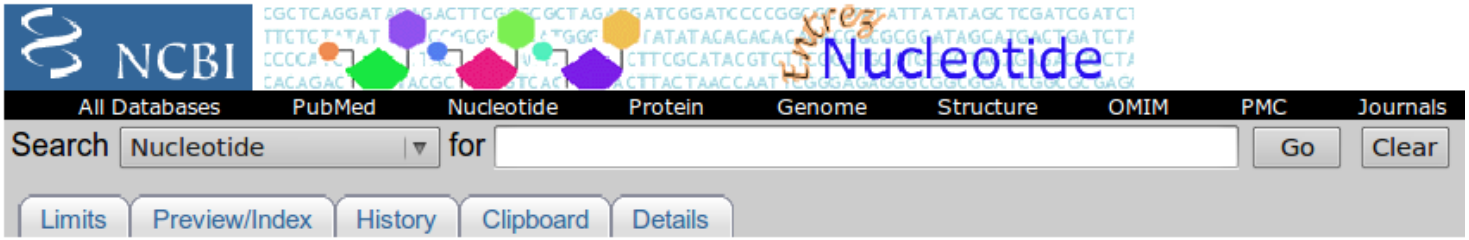
Nucleotide sequence databases



- sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents.
- entries in the EMBL, GenBank and DDBJ databases are **synchronized** on a daily basis
- **accession numbers** are managed in a consistent manner
- comparatively little error checking and fair amount of redundancy.

Nucleotide sequence example

- Glucokinase (hexokinase 4), mRNA [GenBank]



NCBI Reference Sequence: NM_000162.3

Homo sapiens glucokinase (hexokinase 4) (GCK), transcript variant 1, mRNA

Comment Features Sequence

```
LOCUS       NM_000162               2741 bp    mRNA    linear   PRI 25-APR-2010
DEFINITION  Homo sapiens glucokinase (hexokinase 4) (GCK), transcript variant
            1, mRNA.
ACCESSION   NM_000162
VERSION     NM_000162.3  GI:167621407
KEYWORDS    .
```

```
   1 gagcaggaaa tgccgagcgg cgcctgagcc ccagggaagc aggctaggat gtgagagaca
  61 cagtcacctg cagcctaatt actcaaaagc tgtccccagg tcacagaagg gagaggacat
 121 ttcccactga atctgtctga aggacactaa gccccacagc tcaacacaac caggagagaa
 181 agcgctgagg acgccaccca agcgcccagc aatggccctg cctggagaac atccaggctc
 241 agtgaggaag ggtccagaag ggaatgcttg ccgactcgtt ggagaacaat gaaaaggagg
 301 aaactgtgac tgaacctcaa accccaaacc agcccgagga gaaccacatt ctcccaggga
 361 cccagggcgg gccgtgaccc ctgcggcgga gaagccttgg atatttccac ttcagaagcc
 421 tactggggaa ggctgagggg tcccagctcc ccacgctggc tgctgtgcag atgctggacg
 481 acagagccag gatggaggcc gccaagaagg agaaggtaga gcagatcctg gcagagttcc
 541 agctgcagga ggaggacctg aagaaggtga tgagacggat gcagaaggag atggaccgcg
 601 gcctgaggct ggagacccat gaagaggcca gtgtgaagat gctgcccacc tacgtgcgct
 661 ccaccccaga aggctcagaa gtcggggact tcctctccct ggacctgggt ggcactaact
```

Protein sequence databases

- UniProt KB
  - mission to provide a comprehensive, high-quality and freely accessible resource of protein sequence and functional information
  - **SWISS-PROT** is a protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
  - **TrEMBL** is a **computer-annotated supplement** of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.
- PIR
- SWISS-PROT and PIR are different from the nucleotide databases in that they are both **curated**

Protein sequence example

- Glucokinase homo sapiens [P35557 (HXK4_HUMAN)]

Reviewed, UniProtKB/Swiss-Prot **P35557** (HXK4_HUMAN)

Last modified March 2, 2010. Version 115. History...

Clusters with 100%, 90%, 50% identity | Documents (6) | Third-party data | Customize display

Names and origin · Protein attributes · General annotation (Comments) · Ontologies · Bina
Relevant documents

Names and origin Hide

Protein names	*Recommended name:* **Glucokinase** EC=2.7.1.2 *Alternative name(s):* Hexokinase type IV Short name=HK IV Hexokinase-4 Short name=HK4 Hexokinase-D
Gene names	Name:**GCK**
Organism	**Homo sapiens (Human)** [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata ›

Protein attributes Hide

Sequence length	465 AA.
Sequence status	Complete.
Protein existence	Evidence at protein level.

```
        10         20         30         40
MLDDRARMEA AKKEKVEQIL AEFQLQEEDL KKVMRRMQKE
        70         80         90        100
YVRSTPEGSE VGDFLSLDLG GTNFRVMLVK VGEGEEGQWS
       130        140        150        160
MLFDYISECI SDFLDKHQMK HKKLPLGFTF SFPVRHEDID
       190        200        210        220
VVGLLRDAIK RRGDFEMDVV AMVNDTVATM ISCYYEDHQC
       250        260        270        280
VELVEGDEGR MCVNTEWGAF GDSGELDEFL LEYDRLVDES
       310        320        330        340
LVRLVLLRLV DENLLFHGEA SEQLRTRGAF ETRFVSQVES
       370        380        390        400
TTDCDIVRRA CESVSTRAAH MCSAGLAGVI NRMRESRSED
       430        440        450        460
ERFHASVRRL TPSCEITFIE SEEGSGRGAA LVSAVACKKA
```

Peptide related information

- MEROPS - Peptidase Database
- Peptide Database (Cancer) [example]
- PeptideMass
  - cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL) or a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides.
- SYSFPEITHI
  - SYFPEITHI is a database comprising more than 7000 peptide sequences known to bind class I and class II MHC molecules. The entries are compiled from published reports only.

- PeptideAtlas
  - multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments.

Peptide	Welcome to the Antimicrobial Peptide Database	http://aps.unmc.edu/AP/main.php
	Bioactive Polypeptide Database	http://biopd.bjmu.edu.cn/help.asp
	CoPS - Comprehensive Peptide Signature Database	http://203.90.127.70/copsv2/index.asp
	Peptide Database (Cancer)	http://www.cancerimmunity.org/peptidedatabase/Tcelle
	SPdb : A Signal Peptide Database	http://proline.bic.nus.edu.sg/spdb/
	Welcome to the JenPep Database. JenPep is a database of quantitative binding data for immunological protein-peptide interactions.	http://www.jenner.ac.uk/JenPep/
	Database of Nonribosomal Peptide Synthetases	http://linux1.nii.res.in/~zeeshan/webpages/home.html
	Peptide Antigen Database	http://www.proteinlounge.com/peptide_home.asp
	NRPSDB: A Database of NONRIBOSOMAL PEPTIDE SYNTHETASES	http://203.90.127.50/~zeeshan/webpages/nrpsall.html
	The MHC-Peptide Interaction Database version T (MPID-T) is a new generation database for sequence-structure-function information on T cell receptor/peptide/MHC interactions. It contains all structures of TcR/pMHC and pMHC complexes, with emphasis on the structural characterization of these complexes.	http://surya.bic.nus.edu.sg/mpidt/
	Bioactive Peptide Database - Details	http://www.molecularstation.com/bioinformatics/link/de
	AntiJen Database:Peptide Library Search (A kinetic, thermodynamic and cellular Database V2)	http://www.jenner.ac.uk/AntiJen/aj_peplib.htm
	PepSeeker: a database of proteome peptide identifications for investigating fragmentation patterns.	http://nwsr.bms.umist.ac.uk/pepseeker
	A Database Of Mhc Ligands And Peptide Motifs (Ver. 1.0)	http://www.syfpeithi.de/
	System Biology Database	http://www.proteinlounge.com/
	GenomeNet icon (DBGET Search – LITDB) LITDB is a database of key words and key phrases for compounds, fact data, and subjects from the protein and peptide literature.	http://www.genome.ad.jp/htbin/www%5Fbfind?litdb
	Arabidopsis Unannotated Secreted Peptide Database	http://peptidome.missouri.edu/
	Comprehensive Database of Mechanisms of Peptide Fragmentation	http://www.highchem.com/publications/comprehensive
	Antimicrobial Peptide Database	http://aps.unmc.edu/AP/main.html
	PeptideMass cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL) or a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides.	http://www.expasy.ch/tools/peptide-mass.html
	Peptide Database	http://www.peptideatlas.org/
	S A P D - Synthetic Antibiotic Peptide Database	http://www.molecularstation.com/bioinformatics/link/de
	The MHC-Peptide Interaction Database (MHCP) is a curated MySQL database for sequence-structure-function information on MHC-Peptide interactions.	http://sege.ntu.edu.sg/wester/mhcp/intro.htm
	Peptide Antigen Database	http://www.peptides.net/index.aspx?ID=76269
	BciPep a database of B cell epitopes (**Bcipep** is collection of the peptides having the role in Humoral immunity.)	http://www.imtech.res.in/raghava/bcipep/info.html
	PEPSEEKER (Peptide Identification Database)	http://www.nwsr.manchester.ac.uk/cgi-bin/pepseeker/p
	ProteinProspector	http://prospector.ucsf.edu/
	Peptide Conformation Database	http://www.peptidome.org/products/conformation.htm
	PenBase is a curated database devoted to Penaeidins	http://www.penbase.immunaqua.com/
	DOMINO: a database of domain-peptide interactions.	http://mint.bio.uniroma2.it/domino/search/searchWelco
	Peptide Database List	http://www.peptidestation.com/peptide-links/index.php
	Bioinformatics Centre , IISc / Research (Peptide Database)	http://www.physics.iisc.ernet.in/~dichome/rh.htm
	PepBank - a database of peptides based on sequence text mining and public peptide data sources	http://pepbank.mgh.harvard.edu/
	ANTIMIC - Database of natural antimicrobial peptides	http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/

Enzymes

- BRENDA [glucokinase 2.7.1.2]

  - Comprehensive enzyme information system

SYSTEMATIC NAME	IUBMB Comments
ATP:D-glucose 6-phosphotransferase	A group of enzymes found in invertebrates and microorganisms that are highly

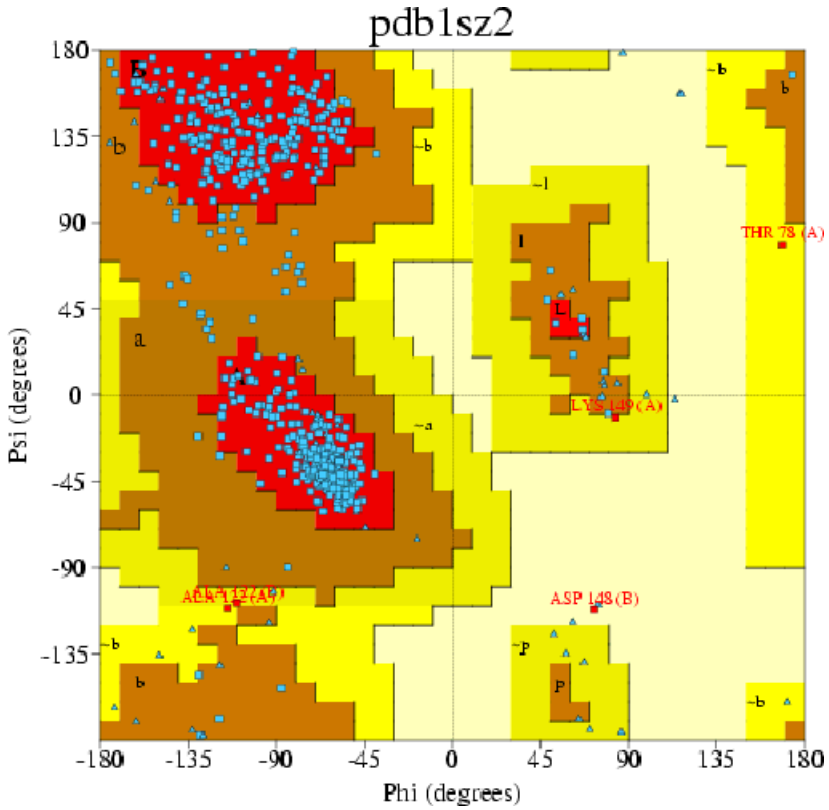
SYNONYMS	ORGANISM	COMMENTARY	LITERATURE
GCK	Homo sapiens	-	661567
GK	Homo sapiens	-	663393
GlkB	Homo sapiens	-	688144
glucokinase (phosphorylating)	-	-	-
glucokinase B	Homo sapiens	-	688144
hexokinase IV	Homo sapiens	-	660963, 688144
kinase, gluco- (phosphorylating)	-	-	-

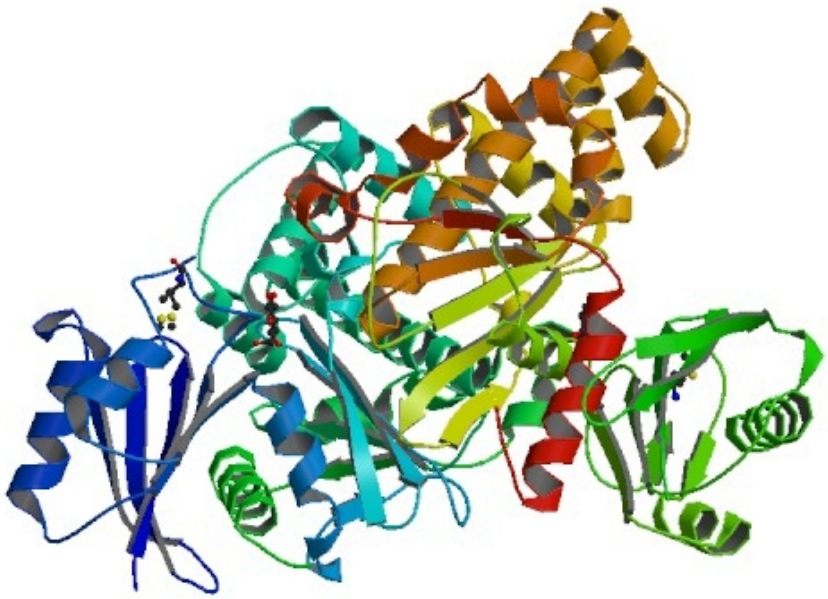
- KEGG Enzymes [glucokinase 2.7.1.2]
- Ensemble [GCK ENSG00000106633]

  - The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Structure databases

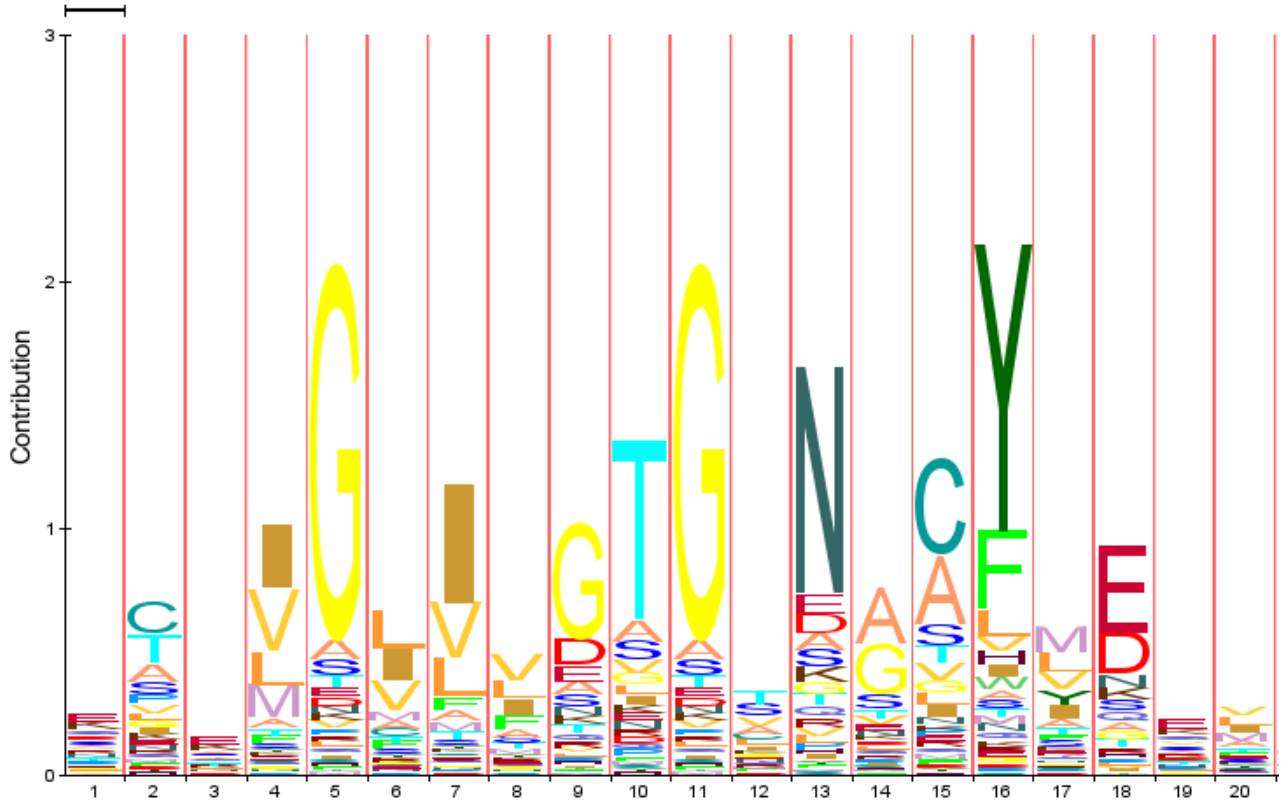
- PDB [glucokinase 1SZ2]
- PDBsum





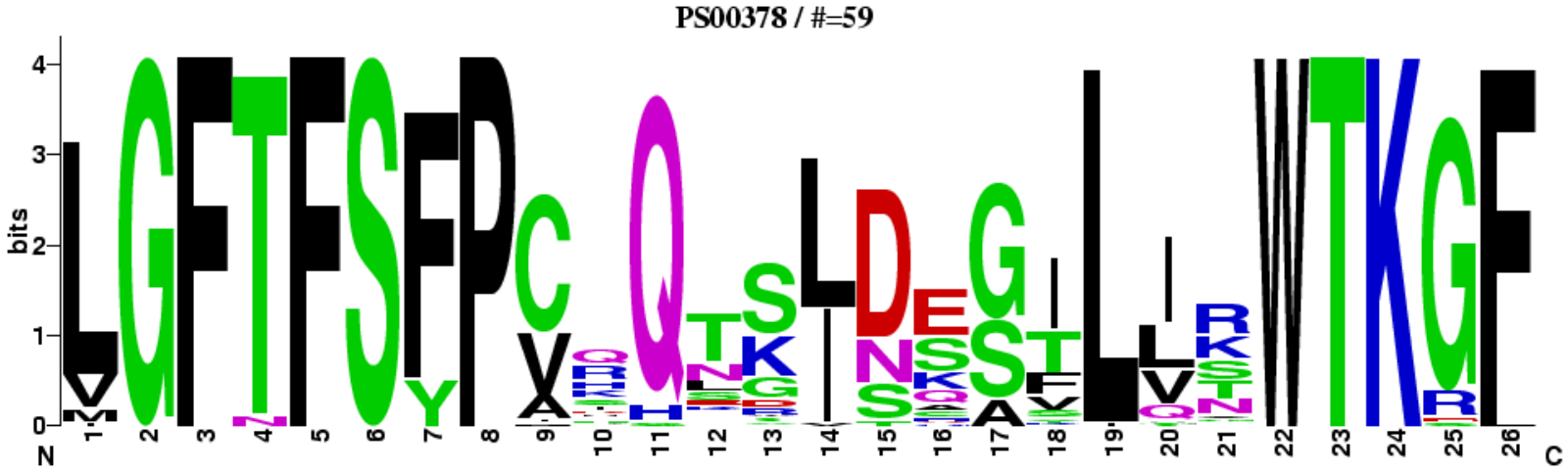
Secondary Databases

- Sometimes known as **pattern databases**
- Contain results from the **analysis of the sequences** in the primary databases
- Examples
  - PROSITE
  - Pfam
  - PRINTS



Motifs and secondary structure

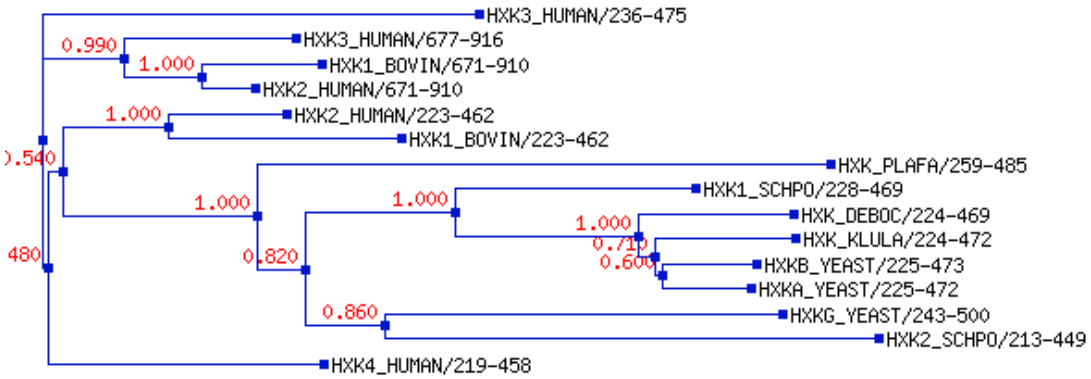
- PROSITE [HEXOKINASES PS00378]
  - Database of protein domains, families and functional sites
  - Hexokinases signature: Pattern [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-x(3)-[LIVM]-x(2)-W-T-K-x- [LF].



Motifs and secondary structure

- Pfam [Hexokinase_2 PF03727]

  - The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)



HXK2_HUMAN

This is the summary of UniProt entry HXK2_HUMAN (P52789).

Description:	Hexokinase-2 EC=2.7.1.1
Source organism:	Homo sapiens (Human) (NCBI taxonomy View Pfam proteome data.
Length:	917 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniPro
removed *after* a Pfam release, these entries will not be removed from Pfam until the *ne*

Pfam domains

This image shows the arrangement of the Pfam domains that we found
boundaries for each of the domains.

Source	Domain	Start	End
Pfam A	Hexokinase_1	16	221
Pfam A	Hexokinase_2	223	462
Pfam A	Hexokinase_1	464	669
Pfam A	Hexokinase_2	671	910

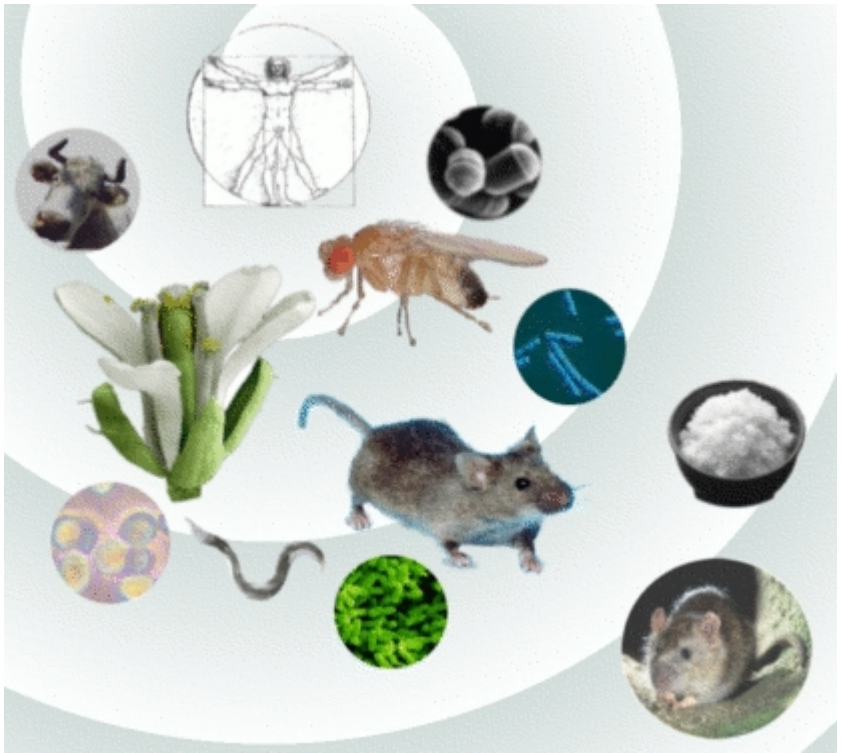
Literature Databases

- PubMed / MEDLINE
  - Database of citations and abstracts for biomedical literature
- OMIM (Online Mendelian Inheritance in Man) [Glucokinase]
  - Catalog of human genes and genetic disorders with textual information and copious links to scientific literature
- Google Scholar
- CiteXplore
  - combines literature search with text mining tools for biology.
- Arxiv
  - Open access to 601,910 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

Taxonomy

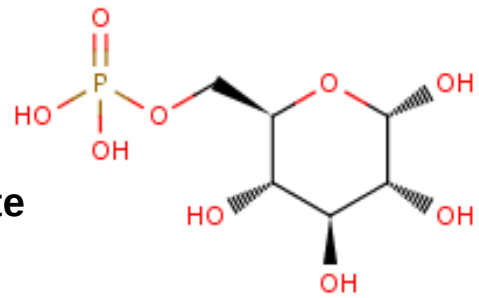
- UniProt taxonomy [homo sapiens]
  - Organisms are classified in a hierarchical tree structure.
  - next to manually verified organism names, external links, organism strains and viral host information is provided.
- NCBI taxonomy [homo sapiens]

Mnemonic	HUMAN
Taxon identifier	9606
Scientific name	Homo sapiens
Common name	Human
Synonym	-
Other names	› man
Rank	Species
Lineage	› cellular organisms › Eukaryota › Fungi/Metazoa group › Metazoa › Eumetazoa › Bilateria › Coelomata › Deuterostomia › Chordata › Craniata › Vertebrata



Chemical entities

- ChEBI (Chemical Entities of Biological Interest EBI)
  - freely available dictionary of molecular entities focused on ‘small’ chemical compounds
- Kegg Compounds
  - KEGG COMPOUND is a chemical structure database for metabolic compounds and other chemical substances that are relevant to biological systems.
  - Peptide entries in KEGG COMPOUND are designated with "Peptide" in the first Entry line
- PubChem



α-D-glucose 6-phosphate

CHEBI:17665

KEGG:C00668

PubChem:5958

Properties Computed from Structure:

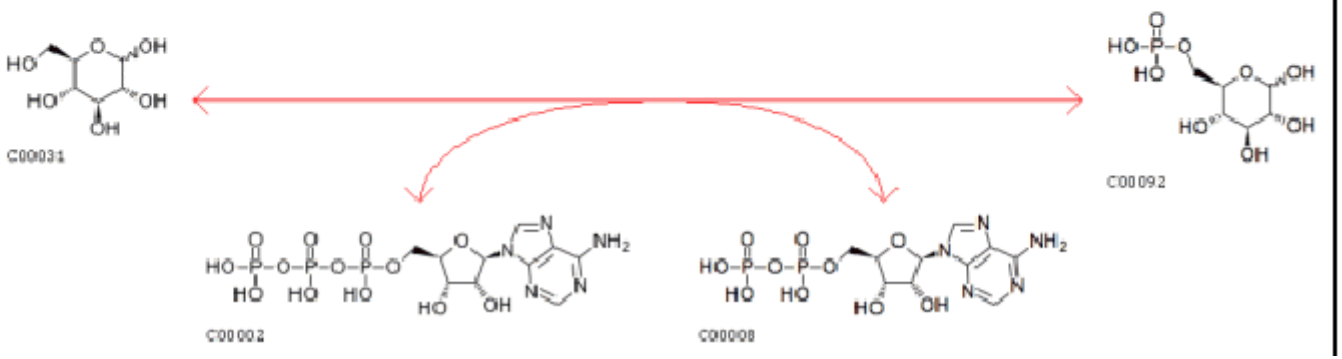
Molecular Weight	260.135781 [g/mol]
Molecular Formula	$C_6H_{13}O_9P$
XLogP3-AA	-4.2
H-Bond Donor	6
H-Bond Acceptor	9
Rotatable Bond Count	3
Exact Mass	260.029719
MonoIsotopic Mass	260.029719
Topological Polar Surface Area	157
Heavy Atom Count	16
Formal Charge	0

Reactions

- Kegg Reactions [R00299]

KEGG **REACTION: R00299** Help

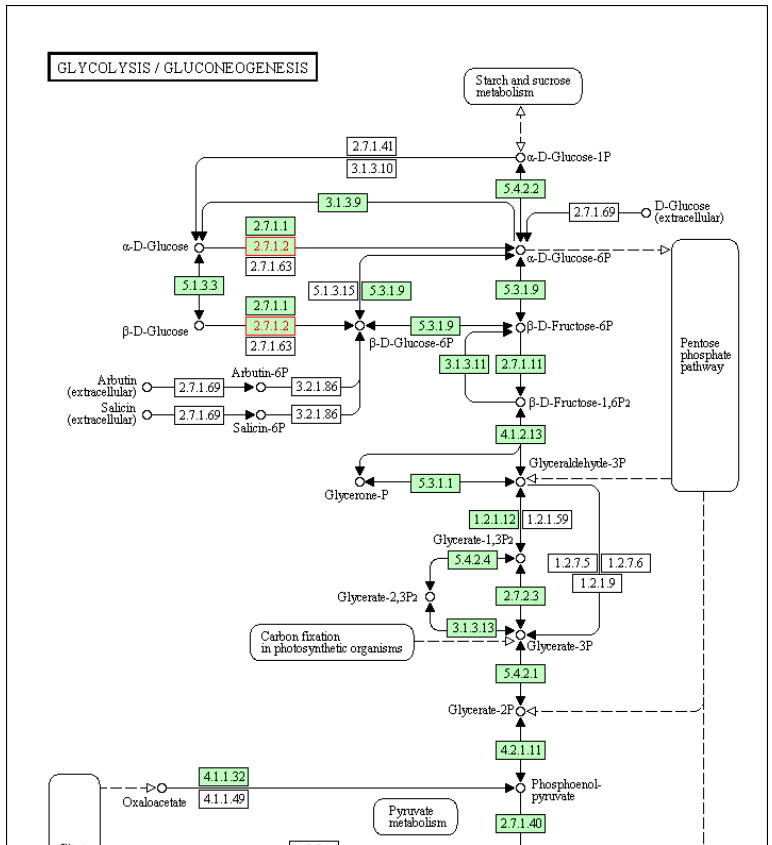
Entry	R00299 Reaction
Name	ATP:D-glucose 6-phosphotransferase
Definition	ATP + D-Glucose <=> ADP + D-Glucose 6-phosphate
Equation	C00002 + C00031 <=> C00008 + C00092

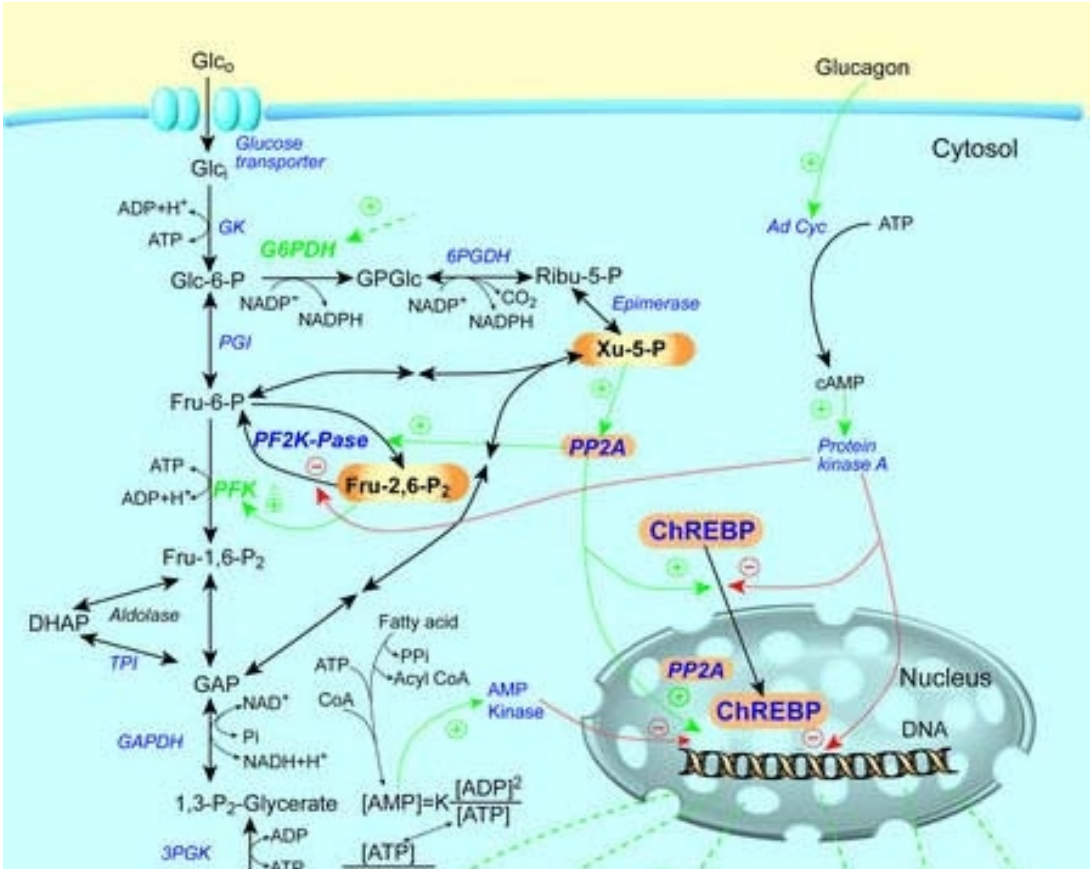


- Rhea [17828]
  - Rhea is a freely available, manually annotated database of chemical reactions

Metabolic networks - pathways

- Kegg Pathways [glycolysis / gluconeogenesis hsa]
- MetaCyc (HumanCyc)
- Reactome - a curated knowledgebase of biological pathways





III Sequence Alignment – Fragment search with BLAST

Sequence Alignment - BLAST

- BLAST, is an algorithm for comparing primary biological sequence information (amino-acid or nucleotide sequences)
    - Enables to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.
- BLAST is one of the most widely used bioinformatics programs
    - it addresses a fundamental problem
    - the algorithm emphasizes speed over sensitivity (practical on the huge genome databases currently available
- Variants
    - Nucleotide-nucleotide BLAST (blastn)
    - Protein-protein BLAST (blastp)
    - Nucleotide 6-frame translation-protein (blastx)

BLAST

- To run, BLAST requires a **query sequence** to search for, and a sequence to search against (also called the target sequence) or a sequence database containing multiple such sequences.
- Input: **sequences** in FASTA or Genbank format.
- Output: graphical format showing the **hits** found, a table showing sequence identifiers for the hits with **scoring data**, as well as **alignments** for the sequence of interest and the hits received with corresponding BLAST scores for these.
- NCBI - http://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST Results

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage
XP_002713907.1	PREDICTED: glucokinase [Oryctolagus cuniculus]	101	101	100%
EFB16270.1	hypothetical protein PANDA_013319 [Ailuropoda melanoleuca]	101	101	100%
3IMX_A	Chain A, Crystal Structure Of Human Glucokinase In Complex With A Synthetic Activator	101	101	100%
ACD40312.1	glucokinase [Rhabdosargus sarba]	101	101	100%
3FR0_A	Chain A, Human Glucokinase In Complex With 2-Amino Benzamide Activator >pdb|3A0I|X Chain X, Human	101	101	100%
3F9M_A	Chain A, Human Pancreatic Glucokinase In Complex With Glucose And Activator Showing A Mobile Flap >pd	101	101	100%
XP_001495888.2	PREDICTED: glucokinase (hexokinase 4) [Equus caballus]	101	101	100%
ABS31137.1	glucokinase [Homo sapiens]	101	101	100%
NP_001095772.1	glucokinase [Bos taurus] >gb|AAI50098.1| GCK protein [Bos taurus]	101	101	100%
XP_001520120.1	PREDICTED: similar to glucokinase [Ornithorhynchus anatinus]	101	101	100%
NP_001096321.1	glucokinase (hexokinase 4) [Xenopus (Silurana) tropicalis] >gb|AAI35717.1| LOC100124905 protein [Xeno	101	101	100%

GENE ID: 3101 HK3 | hexokinase 3 (white cell) [Homo sapiens]
(Over 10 PubMed links)

Sort alignments

for this subject sequence by:

E value Score

Percent identity

Query start position Subject start position

```
 Score = 62.6 bits (140),  Expect = 5e-09
 Identities = 21/27 (77%), Positives = 23/27 (85%), Gaps = 0/27 (0%)
Query  4    LPLGFTFSFPVRHEDIDKGILLNWTKG  30
            LPLGFTFSFP R   +D+GILLNWTKG
Sbjct  602  LPLGFTFSFPCRQLGLDQGILLNWTKG  628
```

IV database design and implementation

Database Tools

- Database design (Model building)
  - Determine the relationships between the different data elements.
  - Superimpose a logical structure upon the data on the basis of these relationships.
  - Scheme development (paper & pencil)
  - Scheme implementation and refinement (database designer like MicroOLAP DB Designer)
- Relational database (Storage)
  - MySQL, PostgreSQL, SQLite
- Interfaces (Access)
  - SQL queries
  - Administration tools (phpMySQL, phpPgAdmin)
  - Frameworks & Webinterfaces (Django (Python), Hypernate (Java))

Thanks

- Computational Systems Biochemistry group
- Prof. Holzhütter & Michael Weidlich

Genomics and Systems Biology of Molecular Networks
International Research Training Group

Network Systems Biology
HepatoSys

CHARITÉ
COMPUTATIONAL SYSTEMS BIOCHEMISTRY

Presentation available at

http://www.charite.de/sysbio/people/koenig/

Sources

Biological databases

- Nucleic Acid Research
- 2001 Per Kraulis – Databases in bioinformatics - Stockholm Bioinformatics Center, SBC, Lecture notes, http://www.avatar.se/molbioinfo2001/databases.html
- Lim Yun Ping – Biological databases - National University of Singapore - www.s-star.org/downloads/tutorial/t1b.pdf
- Klipp & Liebermeister – Systems Biology (Databases)
- Wikipedia http://en.wikipedia.org/wiki/Biological_database
-

Sequence Alignment & BLAST

- Wikipedia - http://en.wikipedia.org/wiki/BLAST
- 2001 Per Kraulis – Sequence alignments - Stockholm Bioinformatics Center, SBC, Lecture notes http://www.avatar.se/molbioinfo2001/multali.html
-

Sources

Database design

- Wikipedia http://en.wikipedia.org/wiki/Database_design
- Database Design and Modeling Fundamentals http://www.sqlteam.com/article/database-design-and-modeling-fundamentals
-

Database and Database Management

- Wikipedia - http://en.wikipedia.org/wiki/BLAST
- 2001 Per Kraulis – Sequence alignments - Stockholm Bioinformatics Center, SBC, Lecture notes http://www.avatar.se/molbioinfo2001/multali.html