Databases for Life Science Research

Ulf Leser
This Lecture

• What this lecture is not
• RDBMS in ten slides
• Classification & Properties
• Some Examples
This is not a Course in Databases

- A first course on technical foundation of databases requires a semester
- Databases are among the most important types of applications existing (from a commercial point of view)
- “True” databases are almost always relational
  - Open source: MySQL, Postgres
  - Commercial: Oracle, Ibm/DB2, MS-SQLServer
  - Exotic: XML, NoSQL, DMS, embedded DBs, …
- People call many things a database
Properties of a Database Management System

- Universal **data model** to store all types of data
  - Usually: Relational (think of tables)

- **Query language** to retrieve „arbitrary“ pieces of the data
  - Usually: SQL, the Structured Query Language

- Way of storing the data is **independent of future usages**

- Ability to serve multiple queries at the same time while multiple clients are manipulating data
  - „Serve“: Provide a **consistent view on the data**
  - „Manipulate“: Update, insert, delete

- Ability to survive almost any **type of crash** without damaging the data

- Multimedia, performance, scalability, …
Architecture: Client-Server

- JAVA (JDBC)
- Native (SQL*Plus, OCI)
- Other databases

- DB1-Server
- DB2-Server

- Disk

- Listener

- Consistency
- Parallelism
- Recovery
- Load Balancing
- Authentication
- Authorization

....
Example

- RDBMS
- Data Import
- Export
- (Web-Based) User Interface
- Pathways (Reactome, Transfac, etc.)
- Function (Enzyme, UniProt, etc.)
- Links & Integration
Direct Access to a Database
Accessing a Typical Biological Database

• **Web interface** – search and read
  - Good for exploratory modes and single-object searches

• **Automatic control of the web interface**
  - Write a program that *mimics a surfer*
  - Searching: GET/POST; reading: Parse HTML
  - This is simple!

• **Download**
  - Usually in special flatfile format or as XML
  - Requires a parser – extract the pieces you are looking for
  - Some database can be downloaded directly as MySQL/Oracle files

• **Web Services**
  - Link controlling the web interface but simpler
Libraries

• Don’t forget: You are not the first to try this
• BioPerl, BioPython, BioJava, …
  – Large libraries containing everything you dream of
  – Also includes parsers for the most important databases
• BioSQL
  – Readily build parsers to transform many BDB into a uniform relational schema
• …
• Search before you start developing something on your own
• But: Writing something on your own may be faster than performing extensive modifications to somebody else’s code
Modeling a Microarray Experiment

- BioSource
- RNA
- Sample preparation
- labeled RNA:Sample
- Experiment
- Image acquisition
- Array:Probe
- ArrayDesign
- Array preparation
- Normalisation
- Hybridisation
- Protocol
A Simple Model

- **Sample** has a source and was extracted under some conditions.
- **Array** has many probes arranged in a grid; has a manufacturer.
- **Experiment** was performed by someone and at some point in time.
- **Results** are generated through experiments plus some data processing.
This Lecture

• What this lecture is not
• Architecture
• BDB: Classification & Properties
• Some Examples
In many, many different species
There are 100reds of Them (Published)

Number of existing (circles) and new databases (triangles) are plotted from 1996 to 2011. New databases are difference between the number of existing databases for each year. DBcat (red) is shown with NAR (blue) counts.

Copyright Geospiza 2011
### Global identifier
<table>
<thead>
<tr>
<th>ID</th>
<th>HSIGHAF</th>
<th>standard; RNA; HUM; 1089 BP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>J00231</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NI</td>
<td>g185041</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>17-DEC-1994 (Rel. 42, Last updated, Version 6)</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>Human Ig gamma3 heavy chain disease OMM protein mRNA.</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KW</td>
<td>C-region; gamma heavy chain disease protein;</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC</td>
<td>Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RN</td>
<td>[1]</td>
<td></td>
</tr>
<tr>
<td>RP</td>
<td>1-1089</td>
<td></td>
</tr>
<tr>
<td>RX</td>
<td>MEDLINE; 82247835.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RC</td>
<td>GDB; 119339; IGHG3.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>GDB; G00-119-339.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>The protein isolated from patient OMM is a gamma heavy chain</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PH</td>
<td>23.964</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>/codon_start=1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>567112&quot;</td>
<td></td>
</tr>
<tr>
<td>XX</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| SQ      | Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other; |}

### Description
- **ID**: HSIGHAF
- **Global identifier**: standard; RNA; HUM; 1089 BP.
- **Date**: 17-DEC-1994 (Rel. 42, Last updated, Version 6)
- **Description**: Human Ig gamma3 heavy chain disease OMM protein mRNA.
- **Taxonomy**: Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates.
- **References**: GDB; 119339; IGHG3.
- **Cross-Links**: The protein isolated from patient OMM is a gamma heavy chain.
- **Feature**: CDS
- **Sequence**: 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
- **Raw data**: CCTGGAACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TCTCTCTTCC TGGTGCCACGC 60
- **TTCCAGATGG GTCTGTCCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC 120"
Properties

ID   HSIGHAF    standard; RNA; HUM; 1089 BP.
XX  AC   J00231;
XX  NI   g185041
XX  DT   17-DEC-1994 (Rel. 42, Last updated, Version 6)
XX  DE   Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX  KW   C-region; gamma heavy chain disease protein;
XX  OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
XX  RN   [1]
XX  RP   1-1089
XX  RX   MEDLINE; 82247835.
XX  DR   GDB; 119339; IGHG3.
XX  DR   GDB; G00-119-339.
XX  DR   The protein isolated from patient OMM is a gamma heavy chain
XX  PH   23. .964
XX  FT   /codon_start=1
XX  FT   567112"
XX  SQ   Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
       CCTGGAACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TTCCTTCTCC TGGGGAAGCC
       TCCCAGATGG GTCCTGTCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC
       ...
Biological Databases Today

• Flat files or XML for export / exchange
• Almost all maintained in relational systems
• “Read-only”, no transactions
• Web-based user interfaces
  – Very very few direct SQL accesses (but dump files for own use)
  – Simplicity is king: “Google style”
• Most are available entirely for download
• New releases every X months
• Many are manually curated
• Some trends towards “open” databases
  – ProteinWiki, Oreganno, …
Data Model

- Many started as books
  - One object - one page
    - Sequence, gene, protein, diseases, ...
    - Object may be an abstract concept (gene) or a measurement (sequence)
  - Entry-based: A primary object and its (nested) annotations
  - Perfectly suited for XML and flatfiles
- More recent databases often are much more “multi-object”
  - Multiple primary objects
  - Multiple links between objects
  - More “normal” database
Classes of Databases

• **Primary – secondary**
  - Taking experimental data (sequences) are conclusions drawn from experiments (genes)
  - Relatively few primary (20?), many secondary (100reds)

• **Species-specific – type-specific**
  - All stuff on one species (MGD)
  - All on one topic across species (GenBank)

• **Curated or not**
  - Most secondary databases are created manually

• **Quality issues: Correctness & completeness**

• **Some primary databases are de-facto standards**
  - Sequences: Genbank, proteins: UniProt, structures: PDB, …
Links

- BDB maintain links to many other BDBs
  - Instance level
  - Browsing support
- Stored as external ID
- Presented as hyperlink

- No central authority for ID or links
- No consistency – "link hell" is a truth
Different Cultures

- BDB developer often are more similar to BDB users (LS) than to database researchers (DR)
- DR publishes methods, LS publishes results
  - 1,300 databases = 1,300+ LS papers
  - 1,300 databases = ~10 DR papers
- DR: Often little willingness to get domain-specific
  - Building a BDB usually is not considered CS research (no papers, no PhD)
- DR: Often little willingness to consider CS as science
  - Too abstract, no concrete results on physical objects
- A VLDB paper on BDB is by no means certainly a contribution to LS
- A NAR paper on a BDB is by no means certainly interesting for a DR
Types of “Data”

- Knowledge
  - Confirmed, abstract, condensed
  - Text, graphics
  - Publications

- Information
  - Interpreted, filtered
  - Objects, annotations
  - BDB – secondary databases

- Data
  - Measured - raw, noisy, context-free
  - Numbers, sequences, metadata
  - BDB – primary databases
Data to Information

Scanning → Image → Spot detection & assignment

- Background correction
- spot-to-gene, …

Raw data
Data to Information

1. Raw data
2. Scanning
3. Image
4. Spot detection & assignment
5. Background correction
   - spot-to-gene, ...
6. Experiment-level normalization
7. Differentially expressed genes
8. Signatures
9. Clustering
10. Functional analysis
11. Classification
Data? Information? Knowledge?

- Scanning
- Image
- Spot detection & assignment

- Background correction
  - spot-to-gene, …

- Experiment-level normalization

- Differentially expressed genes

- Signatures
- Clustering
- Functional analysis
- Classification
Steps with a wide Choice of Methods

- Scanning
- Image
- Spot detection & assignment

- Background correction
- spot-to-gene, ...

- Experiment-level normalization

- Differentially expressed genes

- Signatures
- Clustering
- Functional analysis
- Classification
Data to Knowledge 2

Statistical Data Integration

- Scanning → Image → Spot detection & assignment
- Background correction → Spot-to-gene, ...
- Raw data

- Statistical Data Integration
  - Signatures
  - Clustering
  - Functional analysis
  - Classification

- Raw data → Scanning → Image → Spot detection & assignment
- Background correction → Spot-to-gene, ...
- Raw data
Data and Analysis Workflow

• High-throughput experiments require a multi-step analysis pipeline
• Many different suggestions for each step and for their composition into a process
• User only interested in result: Which genes are over-expressed in acute lymphoma?
• Data (information) may be integrated at various levels
  – Which may result in dramatically different final results
• Rule-of-thumb: The later, the less comparable numerically
  – You may write a survey after mentally aggregating the results, but you cannot compute further with them
Semantic Heterogeneity – What is a Gene?

- A stretch of DNA (with holes) on a chromosome that at some stage gets translated into a protein
What is a Gene?

- A re-assembly of stretches of DNA that are transcribed together plus some further editing on the mRNA level
What is a Gene?

- Like Def.2, plus **parts of the sequence downstream** that is necessary to regulate transcription of the gene
What is a Gene? [GBR+07]

• The same gene?
  - Genes may generate different assemblies (differential splicing)
  - Genes may have interspersed genes
  - Gene duplications in a genome
  - The „same“ gene in another organism
  - Mutation of a gene
  - Genes with a different start site
  - Post-translational modifications

• A gene?
  - Pseudo genes (never transcribed, yet highly similar)
  - Non-coding genes
  - miRNA

• Gene definitions change(d) over centuries, decades, and last years
Does it Matter?

• Sometimes yes
  – E.g. to study differential splicing
  – E.g. to study regulatory relationships

• Sometimes no
  – E.g. to study gene function (without too many details)
  – E.g. to study gene interactions (without too many details)

• Most studies today are carried our “without too much detail”

• E.g., detailed knowledge on splice variants and their functional differences is still almost non-existing

• Researcher know they are doing wrong, but it is the best they can do now
Is this a Problem?

Yes, if you plan to create a **stable, precise, comprehensive integrated gene database**

No, if you are pursuing a **specific study** taking into account your selection of genes
This Lecture

• What this lecture is not
• RDBMS in ten slides
• Classification & Properties
• Some Examples
  - Tour
  - Important biological databases
  - Redundancy
A quick tour of some databases

• Suppose ...
  - Child: severe mental retardation, seizures, intensive irritability. Furthermore, doctor notices abnormal pigmentation and a strange, mouse-like odor. Similar symptoms reported in close relatives
  - Let’s see if this is a hereditary disease!

  - Disease known with these symptoms?
  - Which gene is involved?
  - On which chromosome is this gene localized?
  - What is the EC number of the coded enzyme?
DESCRIPTION

Phenylketonuria is an inborn error of metabolism resulting from a deficiency of phenylalanine hydroxylase (EC 1.14.16.1) and characterized by mental retardation. There are other causes of hyperphenylalaninemia. Sirver et al. (1994) reviewed the hyperphenylalaninemas of man and mouse.

CLINICAL FEATURES

Early diagnosis of phenylketonuria (PKU), a cause of mental retardation, is important because it is treatable by dietary means. Features other than mental retardation in untreated patients include "mousy" odor; light pigmentation; peculiarities of gait, stance, and sitting postures; and epilepsy (Pance, 1957). Kawashima et al. (1988) suggested that cataracts and brain calcification may be frequently overlooked manifestations of classic untreated PKU. Brain calcification has been reported in dihydropteridine reductase (DHPR) deficiency (261660). Pitt and O'Day (1971) found only 3 persons with cataracts among 46 adults, aged 28 to 71 years, with untreated PKU. They concluded that PKU is not a cause of cataracts. Levy et al. (1970) screened the serum of 280,919 'normal' teenagers and adults whose blood had been submitted for syphilis testing. Only 3 adults with the biochemical findings of PKU were found. Each was mentally subnormal. Normal mentality is very rare among patients with phenylketonuria who have not received dietary therapy.

The basic defect in PKU is phenylalanine hydroxylase deficiency. Evidence of heterogeneity in phenylketonuria was presented by Auerbach et al. (1967) and by Auerbach et al. (1968).

Cockburn et al. (1990) observed scleroderma in 2 infants with PKU. Improvement in the skin lesions after commencement of a low phenylalanine diet supported the possibility of a causal relationship.

Widespread screening of neonates for phenylketonuria brought to light a class of patients with a disorder of phenylalanine metabolism.
DIAGNOSIS

Normal blood phenylalanine levels are 58 +/- 15 micromoles/liter in adults, 60 +/- 13 micromoles/liter in teenagers, and 62 +/- 18 micromoles/liter (mean +/- SD) in childhood. In the newborn, the upper limit of normal is 120 micromoles/liter (2 mg/dl) (Scrivner et al 1985, Gregory et al., 1986). In untreated classical PKU, blood levels as high as 24 mM/liter can be found.

By the use of RFLPs related to the phenylalanine hydroxylase gene, Erskie et al. (1985) achieved prenatal diagnosis of a PKU homozygote and a PKU heterozygote. In 1988, I was described experience with prenatal diagnosis of PKU by RFLP analysis. It pointed out that in those cases in which the affected child had died but a phenotypically normal brother or sister is available for investigation, full genetic predictability could be obtained only if this child proved to be homozygously healthy in the phenylalanine-loaded heterozygote test.

Taking advantage of the 'illegitimate' transcription of the PAH gene in circulating lymphocytes, Abadie et al. (1993) succeeded in make the DNA diagnosis of phenylketonuria. Furthermore, they identified 3 novel mutations in 2 patients.

Kalaydjieva et al. (1991) identified 3 silent mutations in the PAH gene, in codons 232, 245, and 385, linked to specific RFLP haplotypes in several Caucasian populations. All 3 mutations created a new restriction site and were easily detected on PCR-amplified DNA. The combined analysis of these markers and 1 or 2 PKU mutations formed a simple panel of diagnostic tests with full informativeness in a proportion of PKU families. Forrest et al. (1991) used a modification of the chemical cleavage of mismatch (CCM) method to identify mutations in PAH in PKU. They stated that 'judicious choice of probes gives the CCM method the potential to detect close to 100% single-base mutations.'
**Database Links**

| Gene Map | locus: 12q24.1 |

Note: pressing the 'Gene' button will find the citations in MEDLINE whose text most closely matches the text of the preceding OMIM paragraph, using the Entrez MEDLINE neighboring function.

**TEXT**

**DESCRIPTION**

Phenylketonuria is an inborn error of metabolism resulting from a deficiency of phenylalanine hydroxylase (EC 1.14.16.1) and characterized by mental retardation. There are other causes of hyperphenylalaninemia, *Scriner et al.* (1994) reviewed the hyperphenylalaninemas of man and mouse.

**CLINICAL FEATURES**

Early diagnosis of phenylketonuria (PKU), a cause of mental retardation, is important because it is treatable by dietary means. Features other than mental retardation in untreated patients include a 'moosy' odor, light pigmentation; peculiarities of gait, stance, and sitting posture; seizures, and epilepsy (Papke, 1957). Kawashima et al. (1988) suggested that cataracts and brain calcification may be frequently overlooked manifestations of classic untreated PKU. Brain calcification has been reported in dihydropteridine reductase (DHPR) deficiency (261630). *Tin and O'Day* (1991) found only 5 persons with cataracts among 46 adults, aged 28 to 71 years, with untreated
Action

• OK. What else do we know about the gene?
  - What’s the sequence?
  - What’s its length (in human, complete mRNA)?
  - What is the protein sequence?
  - What’s the function of the enzyme? In which biological processes is it involved?

  - EMBL & SWISS-PROT: (via SRS): http://srs.ebi.ac.uk
<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE638857</td>
<td>Homo sapiens phenylalanine hydroxylase gene, 5' flanking region. (908 nt)</td>
</tr>
<tr>
<td>L47725</td>
<td>Homo sapiens phenylalanine hydroxylase (PAH) mutant Q20stop mRNA (183 nt)</td>
</tr>
<tr>
<td>L11952</td>
<td>Homo sapiens phenylalanine hydroxylase (PAH) mRNA, complete cds. (706 nt)</td>
</tr>
<tr>
<td>A38020</td>
<td>Human phenylalanine hydroxylase mRNA, complete cds. (2429 nt)</td>
</tr>
<tr>
<td>AJ599152</td>
<td>human, Genomic Mutant, 26 nt</td>
</tr>
<tr>
<td>S70378</td>
<td>phenylalanine hydroxylase (exon 2) human, mRNA Mutant, 26 nt</td>
</tr>
<tr>
<td>AK004677</td>
<td>Branchiostoma floridae mRNA for phenylalanine hydroxylase, partial (2070 nt)</td>
</tr>
<tr>
<td>M28032</td>
<td>D. melanogaster phenylalanine hydroxylase (pah) mRNA, complete cds. (1529 nt)</td>
</tr>
<tr>
<td>A0001733</td>
<td>Drosophila melanogaster mRNA for phenylalanine hydroxylase (henna gene)</td>
</tr>
<tr>
<td>A0001721</td>
<td>Drosophila melanogaster mRNA for phenylalanine hydroxylase, Henna recessive 3 allele, Type I (Henna gene)</td>
</tr>
<tr>
<td>A0001722</td>
<td>Drosophila melanogaster mRNA for phenylalanine hydroxylase, Henna recessive 3 allele, Type II (Henna gene)</td>
</tr>
<tr>
<td>A0001717</td>
<td>Drosophila melanogaster Henna gene, Henna recessive 3 allele, 5' flanking region and partial exon 1 (717 nt)</td>
</tr>
<tr>
<td>A0001716</td>
<td>Drosophila melanogaster Henna gene, Henna recessive 3 allele, partial exon 2 (404 nt)</td>
</tr>
<tr>
<td>A0001715</td>
<td>Drosophila melanogaster Henna gene, Henna recessive 3 allele, partial exons 2 and 3 and joined CDS (771 nt)</td>
</tr>
<tr>
<td>A0001758</td>
<td>Drosophila melanogaster Henna gene, Henna recessive 3 allele, from exon 3 (partial) to exon 5 (partial) (759 nt)</td>
</tr>
</tbody>
</table>
Ulf Leser: Bioinformatics, Winter Semester 2010/2011

RX PUBLMED: 22884570.
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX
XX

At positions 105-114 a sequence 'ctctacccg' has partial homology to the 3' end of the 18S ribosomal RNA and may function as a recognition site for binding the 18S ribosomal RNA during translation. Two inverted repeats (positions 41-56 and 108-120) could be involved in a stable stem-loop structure. Poly-A signals are located at nucleotides 2195, 2200, 2344-2349 and 2411-2416. The AA consensus phosphorylation site, retaining the determinants common to substrates of the cAMP-dependent protein kinase, is preserved in the human AA sequence as compared to that of the cat. The mutation causing phenylketonuria replaces an arginine with a tryptophan at residue 408 of phenylalanine hydroxylase [2].
A phenylalanine hydroxylase amino acid polymorphism with implications for molecular diagnostics.


Position
VARIANTS PKU TRP-252 AND THR-318, AND VARIANT GLU-274.

Medline 21134732
PubMed 1151129

<table>
<thead>
<tr>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATALYTIC ACTIVITY</td>
</tr>
<tr>
<td>L-PHENYLALANINE + TETRAHYDROBIOPTERIN + O(2) = L-TYROSINE + DIHYDROBIOPTERIN + H(2)O.</td>
</tr>
<tr>
<td>COFACTOR</td>
</tr>
<tr>
<td>FERROUS ION.</td>
</tr>
<tr>
<td>ENZYME REGULATION</td>
</tr>
<tr>
<td>N-TERMINAL REGION OF PAH IS THOUGHT TO CONTAIN ALLOSTERIC BINDING SITES FOR PHENYLALANINE AND TO CONSTITUTE AN &quot;INHIBITORY&quot; DOMAIN THAT REGULATES THE ACTIVITY OF A CATALYTIC DOMAIN IN THE C-TERMINAL PORTION OF THE MOLECULE.</td>
</tr>
<tr>
<td>PATHWAY</td>
</tr>
<tr>
<td>CATABOLISM OF PHENYLALANINE; FIRST (RATE LIMITING) STEP.</td>
</tr>
<tr>
<td>SUBUNIT</td>
</tr>
<tr>
<td>HOMODIMMER.</td>
</tr>
<tr>
<td>POLYMORPHISM</td>
</tr>
<tr>
<td>THE GLU-274 VARIANT OCCURS ON APPROXIMATELY 4% OF AFRICAN-AMERICAN PAH ALLELES. THE ENZYME ACTIVITY OF THE VARIANT PROTEIN IS INDISTINGUISHABLE FROM THAT OF THE WILD-TYPE FORM.</td>
</tr>
<tr>
<td>DISEASE</td>
</tr>
<tr>
<td>DEFECTS IN PAH ARE THE CAUSE OF PHENYLKETONURIA (PKU) [MIM:261600]. PKU IS AN AUTOSOMAL RECESSIVE INBORN ERROR OF PHENYLALANINE METABOLISM, DUE TO SEVERE PHENYLALANINE HYDROXYLASE DEFICIENCY. IT IS CHARACTERIZED BY BLOOD CONCENTRATIONS OF PHENYLALANINE PERSISTENTLY ABOVE 1200 MUMOL (NORMAL CONCENTRATION 150 MUMOL) WHICH USUALLY CAUSES MENTAL RETARDATION (UNLESS LOW PHENYLALANINE DIET IS INTRODUCED EARLY IN LIFE). THEY TEND TO HAVE LIGHT PIGMENTATION, RASHES SIMILAR TO</td>
</tr>
</tbody>
</table>

MIM: GO:0004505: Phenylalanine 4-monooxygenase activity; TAS.
GO:0006652: Pantothenate and ubiquinone biosynthesis; TAS.
InterPro: IPQ01273: Aae_hydroxylase.
| InterPro |
| IPQ02911: ACT. |
| IPQ005961: Phenylalanine4hydroxylase. |
| Pfam |
| PF01642: ACT.1 |
| PF03551: biotemn_H.1 |
| PRINTS |
| PFQ0252: FWHydroxLASE. |
Action

- OK, we know about the gene
  - In which pathway is this enzyme involved? Which other proteins are involved in this pathway, and how are they related?
  - KEGG: http://www.kegg.com/
Action

- How does the (human) protein look like?
  - PDB: http://www.rcsb.org/pdb/
Your query found 19 structures in the current PDB release and you have selected 0 structures so far. (There are currently 2 structures being processed or "on hold" matching your query.) You can select specific structures by clicking on the checkbox next to their ID. If you do not select any structures, certain options will default to all structures. To examine an individual structure select the Explore link.

Pull down to select option: [New Search] Go

KEY: = Download compressed (GNU zipped) PDB file = View PDB file = Structure viewing options
PDB
Action

- Anything else out there?

  - www.google.com, „phenylalanine hydroxylase“
A proper PAH database

Welcome to the PAHdb World Wide Web site. This site is run with the following aim: To provide users with access to up-to-date information about mutations at the phenylalanine hydroxylase locus. Here you will have access to the content of the database in the form of electronic reports. The database is updated manually off-line by the curators to ensure that no erroneous information is appended. The curators now also accept data electronically via the submission form. For a better synopsis of the site's function, please read the PAHdb Knowledgebase Abstract.

What's New:
- [05.03.03] BH$_4$-responsive hyperphenylalaninemia (BH$_4$ Alleles Part IV)
- [15.01.03] Tetrahydrobioprotein as an Alternative Treatment for Mild Phenylketonuria
- [31.10.02] Entire web user interface change
- [29.08.02] A Phenylnalanine Hydroxylase (PAH) Acceptor Amino Acid Polymorphism in the Curators' page

Information about PAH mutations
Mutations in the phenylalanine hydroxylase (PAH) gene, the majority of which result in deficient enzyme activity and cause hyperphenylalaninemia, occur in all 13 exons of the gene and flanking sequence. Some cause phenylketonuria, others cause non-PKU hyperphenylalaninemia, while still others are silent.

Who is navigating this site?

Country: 
Choose your country
This Lecture

• What this lecture is not
• RDBMS in ten slides
• Classification & Properties
• Some Examples
  - Tour
    - Important biological databases
  - Redundancy
The Most Important Ones

• Of course: Importance depends on your question

• (DNA) Sequences
  – Genbank / EMBL / DDBJ
  – Ensembl
    – UCSC Genome Browser

• Proteins
  – UniProt (SwissProt + PIR + TrEMBL)
  – PDB (3D structure)
  – BRENDA (enzymes)

• Pathways
  – KEGG
  – Reactome
The Most Important Ones 2

- Chips
  - GEO
  - ArrayExpress

- Genes
  - Ensembl
  - VEGA
  - GeneCards

- Mutations and diseases
  - OMIM
  - dnSNP, HapMap
  - Several x-thousand genomes projects running now
The Most Important Ones 3

- Enzymes
  - UniProt
  - BRENDA
- Species
  - MGD, MIPS, RGD, TAIR, SGD, EcoCyc, …
- BioModels
- Protein-Protein-Interactions
- …
Extreme Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually created "source" DBs
Extreme Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually created “source” DBs
- DBs integrated other and HT data sets
Extreme Example: Protein-Protein-Interactions

• There are >300 BDBs related to PPI and pathways
  – See http://www.pathguide.org

• Manually created “source” DBs
• DBs integrated other and HT data sets
• Predicted interactions
Extreme Example: Protein-Protein-Interactions

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually created “source” DBs
- DBs integrated other and HT data sets
- Predicted interactions
- Pathway DBs – consisting of PPI

- [KP10]
A Mess [KP10]

• **Inconsistent understanding** of what a PPI actually is
  - Binary, physical interaction
  - Complexes
  - Transient, functional association

• Some integrated DBs have imported more data than there is in the sources

• Source databases **overlap to varying degrees**
  - Effort to sort things out in IMex consortium

• Largely **different reliability** of content
  - Literature-curated, high-throughput experiments (false positive rate), results transferred from orthologs, ...

• **Literature-curated DBs** do not exhibit higher quality than HT [CYS08]
  - Re-annotation reveals inconsistencies, subjective judgments, errors in gene assignment, ...
Take-Home Messages

• Whatever you search, it is not unlikely that someone already has build a database of it
  - Starting point for searching: *Nucleic Acid Research*
  - For trendy stuff, very likely there are many databases

• Database usage is Zipf-distributed

• Before using data from external DB
  - Carefully check how a DB is maintained (last update?)
  - Criteria for including data: automatic, predicted, curated, …

• Do not be afraid: It is not difficult to access a DB
  - But mind semantic clashes and different IDs

• Before you start building a database, ask someone who knows how to do it