Biological Databases Review

B-IT LSI/CEMBIO Winter Semester 2015-2016

Charlie Hoyt[1], Alex Masny[1]

[1]LSI, B-IT, Dahlmannstraße 2, 53113 Bonn, Germany

Table of Contents

Basic Knowledge

AKA high school biology. If my 17 year old sisters in AP biology know this, so should we.

Fundamental Principles of Databases

- Annotation
  - the process of adding information to a particular entity
  - a combination of comments, notations, references and citations, that describe all the experimental and inferred information about a gene or protein
- Provenance = source of info
  - ability to trace back the original source of information
  - e.g. evidence
- Curation
  - why?
    - eliminating redundancies and removing outliers
    - unifying semantics
    - check quality control, reliability of information
- Versioning
  - accession numbers
    - generated automatically
  - submission process
    - why you want an accession number
      - unique identifier = readable by human
    - you have submitted; for publication
      - the mechanism to force people to share info;

- Data integration
  - done by shared key = standards
    - the same table level, which identical in 2 databases
  - # Uniprot = hub: have links, pointers to other DBs
- Usage of biodb (2 types)
  - humans – interface
  - used by machines (machine-machine)
    - web services
- Ownership of info: distinguish db with
  - primary data
  - derived data
    - derived: submitter = owner
- Staging of db / updating mechanism
  - archive
  - gene db: annual update and packaged together with previous years
  - builds of db (context of human genome); build = version
    - complex object like genome;
    - you assemble it and refined
      - many different sequences, context = areas, continuous DNA stretch;
      - problem with repetitive sequence
        - clarify: one, two or more repeats;
    - consensus about info
- Information fusion and aggregation
  - # ensembl
    - documentation! http://www.ensembl.org/info/index.html
  - strategies to aggregate different types of information
  - need sort of shared keys -– smth you have to map on
- Distribution looks awfully exponential, or cut gamma

Omics

- Genomics
- Epigenomics
- Transcriptomics
- Proteomics
- Metabolomics

- the scientific study of chemical processes involving metabolites. Specifically, metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles. The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes.
- Metabonomics
  - the study in the change of metabolites. Be careful when talking to a researcher on metabolites. Whether they identify with the L or the N crew is incredibly important to their ego.
- Lipidomics
  - could be included in the metabolomics (most definitely, people who study lipids just want their own name)
- glycomics = study of sugars
  - Actually pretty interesting to Cembio peeps because of the way that each organism has its own glycosylation pattern on proteins. Important when making antibodies in a non-human cell line, since they have to be fixed so human immune systems don't go after them

General Questions

1. Object orientation: What is an object in a Biology and what "methods" can biologiocal objects in execute? Provide at least three examples for "methods" executed by bioobjects.
2. What are typical attributes of Nucleic Acids? Name at least three of them!
3. Which categories of biomolecules do you know?
4. Which categories of bioDATABASES correspond to these categories of bioMOLECULES?
5. Please give at least one example for each category of biodatabase as we see them categorized at the EBI SRS interface!
6. Which major portals to biodata do you know?
7. What sort of discrepancy exists between biodatabases that represent information on genes & genomes as opposed to biodatabases that store information on gene expression?
8. What is a "flat file database"?
9. What features would you assign to "Bioinformatics" and how does it differ from "Systems Biology"?
10. Annotate a sketch of biomolecules with entity-types from biodatabases you know.
11. ~~How are BioDatabases integrated in SRS? What does the SRS documentation say about the mechanisms used for linking between biodatabases?~~
12. What does "computer readable knowledge" mean?
13. The EBI call itself "the portal to knowledge". How is biomedical knowledge represented in biodatabases?
14. In one of the links to "relevant background information", a primer on molecular biology mentions "linkage disequilibrium". What does this term mean?
15. In the 2012 database issue of the journal Nucleic Acid Research the category "protein sequence databases" is subdivided into 6 sub-categories: list at least three of them.
16. The online version of the 2012 biodatabase issue of NAR comprises how many entries in its biodatabase list?
17. The ENTREZ documentation mentions "E-utilities". A link on the ENTREZ side leads to the documentation of E-utilities .... Please explain, what E-utilities are and what they can be used for.
18. ~~What categories of biodatabases are integrated under SRS and which ones are not?~~
19. ~~How do you link query results in SRS and how do you perform "facetted searches " (which is the usage of search results from one query as the starting group for the next query) using SRS?~~

Literature Databases, Controlled Vocabularies and Gene Catalogs

1. In the description file for the TAXONOMY database, the usage of taxonomy entries in other databases is mentioned. Which types of other databases refer to TAXONOMY entries?
2. What is a catalogue?

3. What is an index? How does SRS “index” over several databases?
4. What is a hierarchy? Which relationship-type is used in hierarchies?
5. What is a taxonomy? Give a brief definition of a taxonomy!
6. What is an ontology? What are the essential features of an ontology that distinguishes it from a taxonomy?
7. Which of the above mentioned controlled vocabularies has a tree structure?
8. What is a directed acyclic graph (DAG) and which type of knowledge representation is based on such a DAG?
9. Please explain / characterize the content of PubMed: how does a typical minimum data set look like in PubMed? I refer to the „anatomy of search results age“ mentioned in the PubMed documentation.
10. What is the difference between PubMed and MEDLINE? Explain in brief!
11. What is PubMedCentral? What does it contain and how does it differ from PubMed?
12. Name at least three searchable types (categories) of information that are contained in MEDLINE abstracts
13. What are MeSH terms and what is their purpose?
14. Give a short summary of the structure of MeSH and
15. How can a search result be “expanded” (I refer to the PubMed help, where “expansion of search results” is a separate point)
16. Explain “information retrieval” with an example involving Medline and MeSH terms
17. When do we speak of synonyms and when do we speak of homonyms?
18. Sketch the major concepts and the conceptual schema of MedLine
19. Explain the differences between OMIM and MedLine
20. What are the three root concepts of GO?
21. What means “annotation”?
22. Which controlled vocabularies do you know besides GO?
23. What is a database?5
24. What is a data service?
25. What is an API?
26. What is a web service?
27. How is data stored?
28. What is the difference between knowledgebase vs database?
29. What are the advantages of each pathway database?

Nucleotide-related Databases:

1. What is a 3 ÚTR ?
2. What is a transcription factor site and why would you collect information on these sites in a database?
3. Explain the fundamentals of gene regulation (activation of transcription; features of DNA that mediate and control transcription)
4. Explain how in silico prediction of transcription factor binding sites can be validated through molecular biology experiments.
5. What is a 5 ÚTR?
6. How does the transcriptional machinery know about the beginning and the end of a “gene” (a transcript)?
7. Which properties (features) define a class of transcription factors in TFFACTOR?
8. Name at least three different classes (types) of transcription factors
9. How are evidences for the presence of a certain domain or motif represented in TFFACTOR? (I refer to the FT line in TFFACTOR entries).

Gene Expression Databases

Gene Expression databases hold information about the levels of RNA expression. Data is acquired through microarray, RNA-seq, and other platforms. EMBL-EBI has a good primer on functional genomics to read first.

Data Platforms

Expressed sequence tags (EST) is a short subsequence of a cDNA sequence that's useful to each of these techniques in identifying transcripts.

Hybridization Methods - Microarray

RNA is expressed in a cell. The cell is popped, the RNA is isolated with biology magic, and reverse transcribed with fluorescent dNTP's into short DNA strips. The microarray experiment consists of pouring the mess of cDNA's onto a chip that has lots of short complementary DNA sequences. Light is shined and a picture is taken that can be interpreted by a computer, and reduced into information about how much binding happens at each short sequence. In theory, multiple short subsequences are selected from each gene (or expressed DNA region) to provide robustness - and identify false positive and promiscuous results.

Some microarray experiments are done as a comparison of two types of cells - think of a sample of cells from a cancer patient versus a sample of the same patient's normal cells within the same tissue. Each cell is popped and the same process is done, but different color fluorescent dNTP's are used for each sample. The fluorescent cDNA's from both samples are put together on the same chip, so the chip can measure a proxy for the relative expression of each gene. Keep in mind that this requires proper normalization of data and lots of care, but the idea is simple enough.

This method can only detect elements that are pre-planned. Dynamic range of 100 fold or 200 fold.

What to say about its limits of detection? Speed? Higher limit of detection (bad) and slower. Sucks!

Sequencing Methods - RNA-Seq ("Next-Gen Sequencing")

RNA-Seq is the shinier, hunkier grandson of Sanger sequencing. Once again, RNA is converted to cDNA, then amplified, then shotgun sequenced massively and parallely. Shotgun sequencing means the cDNA's are broken randomly into two chunks each, which are smaller and more reasonably sequenced. A computer can reassemble the sequence in the end because there is a good coverage of many possible breaks in the same sequence, especially after amplification. To do this, each sequence also has to have an extra element called an "adaptor" added, but the specifics of this aren't so important for this class.

No upper limit of quantification - no detector to saturate. Large dynamic range - up to 9000 fold differences detectable.

Useful Sources:

- http://rnaseq.uoregon.edu/
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/

Chromatin Immunoprecipitation (ChIP)

Identifies protein-binding sites on DNA. Very useful for histone protein patterns, DNAase identification of open chromatin, etc.

Data Standardization

The Functional Genomics Data Society (FGED) published these standards to improve interpretability and reproducibility of functional genomics and expression experiments. They've also published some standards on data formats, but that wasn't really mentioned in BioDB class; I just found it on their site's projects directory.

Minimum Information About a Microarray Experiment (MIAME)

This standard was established to ensure microarray experimental results are less ambiguous, easier to interpret, and more reproducible. It consists of the following points:

1. Raw Data
2. Normalized Data
3. Sample Annotation - compounds used, concentrations, etc.
4. Experimental Design and Sample/Data Relationships
5. Data Annotation (gene identifiers, genomic coordinates, probe sequence, reference array catalog number)
6. Laboratory and data processing protocols

MIAME Notation in Markup Language (MINiML)

MINiML (MIAME Notation in Markup Language, pronounced 'minimal') is a data exchange format optimized for microarray gene expression data, as well as many other types of high-throughput molecular abundance data. MINiML assumes only very basic relations between objects: Platform (e.g., array), Sample (e.g., hybridization), and Series (experiment). MINiML captures all components of the MIAME checklist, as well as any additional information that the submitter wants to provide. MINiML uses XML Schema as syntax.

Sources

- http://dx.doi.org/10.1038/ng1201-365
- http://fged.org/projects/miame/
- MINiML: http://www.ncbi.nlm.nih.gov/geo/info/MINiML.html

Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE)

This standard was established to ensure RNA-Seq experimental results are less ambiguous and easy to reproduce. Mufasra specifically said this would be on the test during her lecture on it, so yeah. It consists of the following points:

1. Description of biological system, samples, and experimental variables
2. Sequence read data for each assay (FASTQ)
3. Final data for each assay
4. Data-sample relationships, associated publication, summary of experiment
5. Experimental and data processing protocols

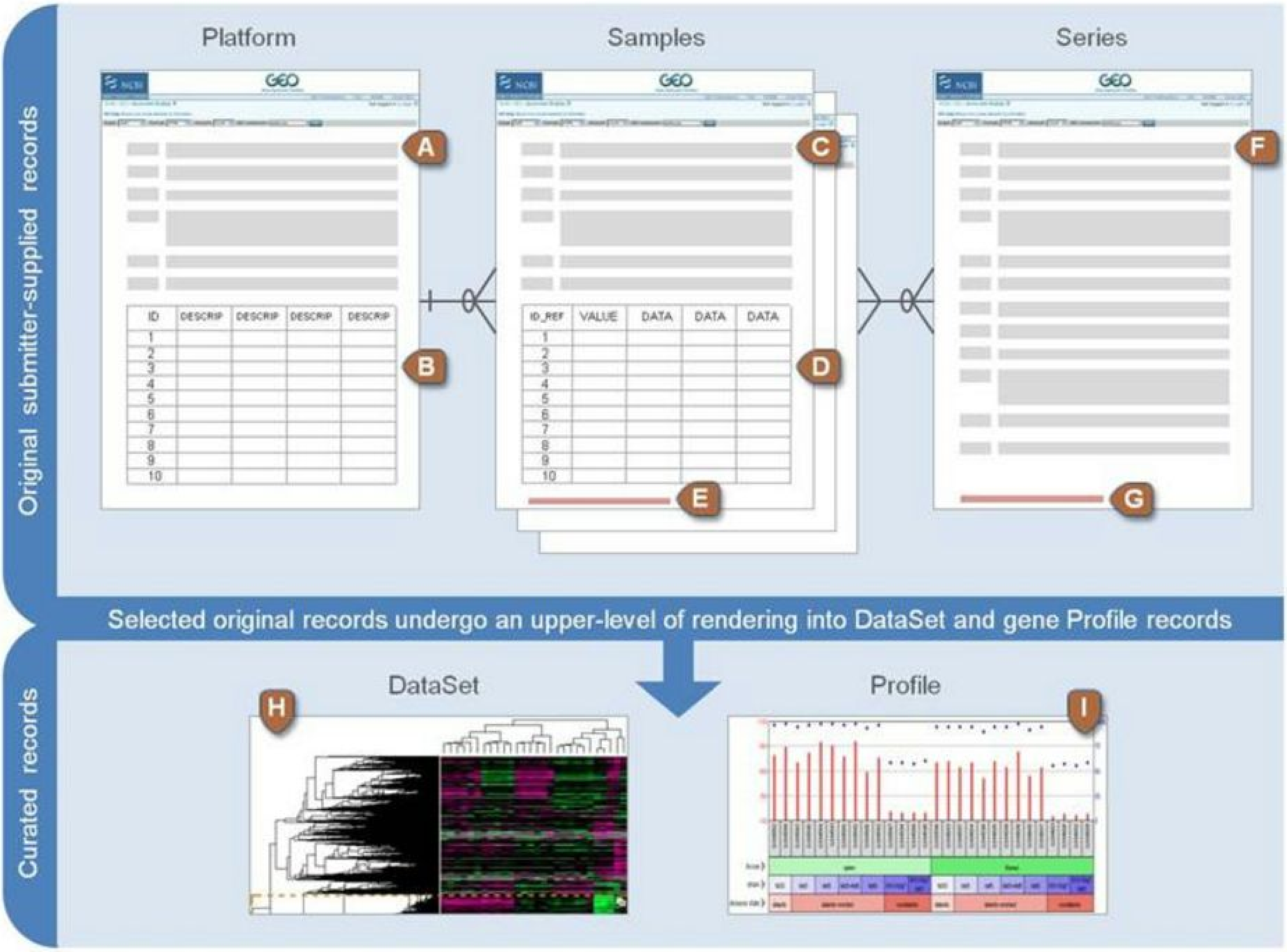
Links

- http://fged.org/projects/minseqe/

Gene Expression Omnibus (GEO)

A NCBI public repository of functional genomics data supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. GEO stores lots of different types of expression data, while ArrayExpress just has microarray data and is better for access.

Organization schema of GEO records


Original submitter-supplied records
Platform
Samples
Series
A
B
C
D
E
F
G
ID
DESCRIP
ID_REF
VALUE
DATA
Selected original records undergo an upper-level of rendering into DataSet and gene Profile records
Curated records
DataSet
Profile
H
I

A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template.

A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it.

A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses.

To quickly locate data relevant to your interests, search GEO DataSets and GEO Profiles:

- GEO DataSets is a *study-level* database which users can search for studies relevant to their interests. The database stores descriptions of all original submitter-supplied records, as well as curated DataSets. DataSet records are assembled by GEO curators.
- GEO Profiles is a *gene-level* database which users can search for gene expression profiles relevant to their interests. A Profile consists of the expression measurements for an individual gene across all Samples in a DataSet.

Links

- http://www.ncbi.nlm.nih.gov/geo/
- Overview: http://www.ncbi.nlm.nih.gov/geo/info/overview.html

ArrayExpress

A database of functional genomics data, part of EMBL-EBI. Data in ArrayExpress is gathered from researchers (using Annotare tool) and imported from GEO. It contains information on experiments, assays, raw data files of microarray and high-throughput sequencing (HTS) analysis that are described and archived according to the community guidelines for microarray (MIAME) and HTS (MINSEQE).

One can use ArrayExpress to:

- Search by keywords or experiment's properties (e.g. citation, transcriptomics platform, species or sample annotation) and identify experiments of interest;
- Download data associated with experiment(s), alongside its annotation, for your own analysis;
- Submit microarray or HTS data that you want to publish.

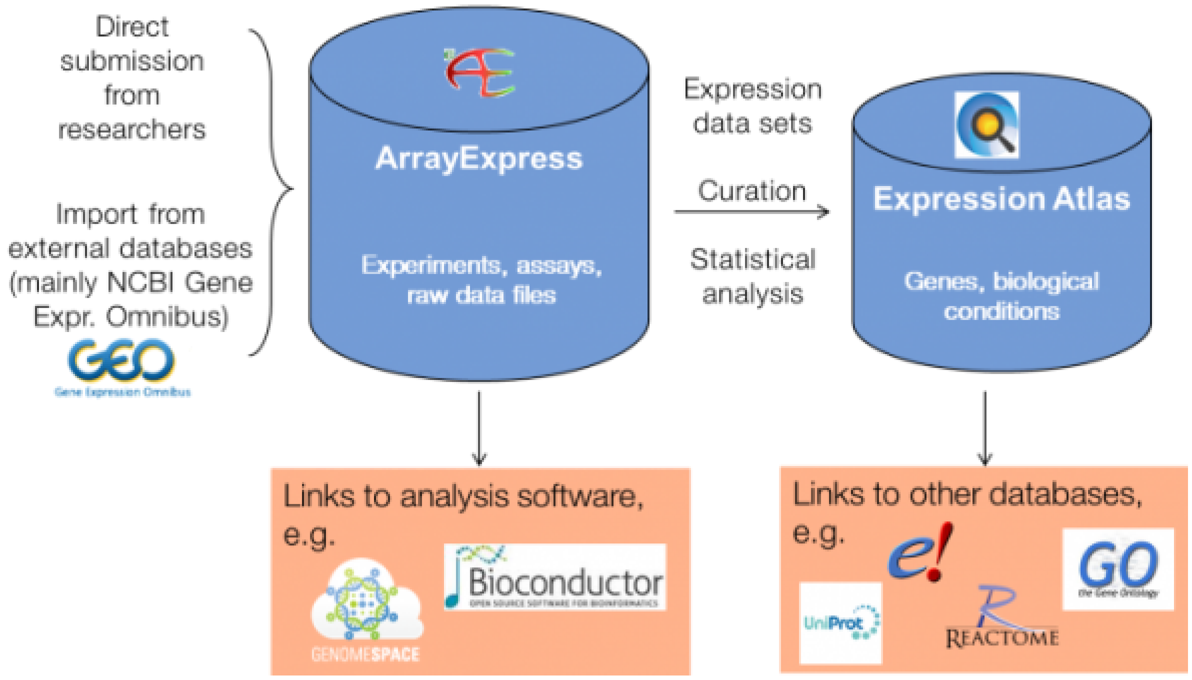
Curated data is stored into **Expression Atlas**. A baseline is calculated for normal/untreated tissues/common cell lines. Data is then compared to the baseline and annotations for genes being up-regulated, down-regulated, or not regulated are added for differential experimental conditions (like a cancer cell vs normal cell).

Contains links to analysis software, like **Bioconductor** – a project which also provides tools for comprehension of high-throughput genomic data. It is based primarily on the R programming language.

Links

- https://www.ebi.ac.uk/arrayexpress/
- Training course: http://www.ebi.ac.uk/training/online/course/arrayexpress-quick-tour-1

Schema of relationship between ArrayExpress and Expression Atlas


Direct submission from researchers
Import from external databases (mainly NCBI Gene Expr. Omnibus)
GEO
Gene Expression Omnibus
ArrayExpress
Experiments, assays, raw data files
Expression data sets
Curation
Statistical analysis
Expression Atlas
Genes, biological conditions
Links to analysis software, e.g.
GENOMESPACE
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Links to other databases, e.g.
e!
UniProt
REACTOME
GO
the Gene Ontology

Expression Atlas (the 'Atlas')

A database containing analysed gene expression data derived from sets (gene expression patterns under different biological conditions, like different cell types, organisms parts, diseases, compound treatments and genotypes) stored in ArrayExpress. Contains two components:

- Baseline Atlas - providing gene expression data for normal, untreated tissues or commonly used cell lines;
- Differential Atlas - allowing queries on genes that are up- or down- regulated in different experimental conditions.

Unlike ArrayExpress which focuses on experiments, the Atlas focuses on genes and biological conditions, allowing you to ask biological questions such as:

- What genes are expressed in normal human liver?

- What genes are expressed across a panel of ENCODE cell lines?
- What genes are up- or down-regulated in drought and salt tolerance (DST) mutant Japanese rice plants vs wild type controls?

How the Atlas is produced

The Atlas is composed of a sub-set of datasets from ArrayExpress, namely those on expression profiling, which are manually curated and then analysed in-house by a standard statistical pipeline. The manual curation step ensures only well-annotated data sets from well-designed experiments are included in the Atlas. For example, for an experiment to be considered for the differential atlas, it must have at least three biological replicates for each condition for proper downstream statistical analysis, and the intent of the experiment must be clear. Various quality-control metrics are also implemented during statistical analysis to discard sub-standard data, e.g. microarray data with lots of background noise.

The difference between the two databases is that ArrayExpress is built around experiments (containing information on data files, sample annotation and others), whereas the Atlas is built around genes and biological conditions and is used to visualise changes in gene expression associated with different biological conditions.

Links:

- https://www.ebi.ac.uk/gxa/home
- Quick start: https://www.ebi.ac.uk/gxa/help/index.html

The Reference Sequence (RefSeq)

The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses.

Links:

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

UniGene

UniGene computationally identifies transcripts from the same locus; analyzes expression by tissue, age, and health status; and reports related proteins (protEST) and clone resources.

UniGene is an NCBI database of the transcriptome and thus, despite the name, not primarily a database for genes. Each entry is a set of transcripts that appear to stem from the same transcription locus (i.e. gene or expressed pseudogene). Information on protein similarities, gene expression, cDNA clones, and genomic location is included with each entry. (https://en.wikipedia.org/wiki/UniGene)

Links:

- http://www.ncbi.nlm.nih.gov/unigene
- Typical Entry http://www.ncbi.nlm.nih.gov/nucest/BF732171.1

Gene Expression Questions

1. What are microarrays?

2. What are alternative gene expression determination technologies?
3. Explain the microarray workflow in the laboratory and map the major MAGE-OM classes the workflow
4. What is the difference between one-colour (one channel) and two-colour (twchannel) microarray assays?
5. What are the consequences of one-channel versus two-channel hybridization for normalization and comparison between chip experiments?
6. Give an example for a one-colour microarray platform
7. Give an example for a two-colour microarray platform
8. Explain the following MAGE-OM classes:
    a. Array
    b. Biomaterial
    c. bioSource
    d. Hybridization
    e. Feature
    f. Feature extraction Compound
    g. Ontology-entry
9. What are the key concepts used in the conceptual model of ArrayExpress?
10. Why is it essential to capture all data that describe the origin of the biosample?
11. Outline the major differences between the conceptual design of GEO and ArrayExpress
12. What are "abundantly expressed" genes?
13. What is the typical distribution of all mRNA species expressed in a cell?
    a. OCCUPY DISTRIBUTION!!!
14. Is this distribution cell-type specific?
15. Name at least two foreign keys that could link a microarray database to a nucleotide sequence database or UniProt KB?
16. Is there an accession number for microarray data?
17. Are images taken from microarray scanners part of the database schema of ArrayExpress?
18. How are experimental series represented in GEO?
19. If you ever visited ArrayExpress, you should have read about "Gene Expression Atlas". What does the "Gene Expression Atlas" comprise?
20. In ArrayExpress, you will find data sets with the designator "tiling array" or "genome tiling experiment". What is the difference between a "classical" microarray experiment and a genome tiling experiment? Explain!
21. What Boolean operators are allowed for querying ArrayExpress (advanced search)?
22. What fields can be searched in GEO and what fields can be browsed?
23. What are GEO profiles?

Nucleotide Sequence Databases

Basically all of these databases are the same but because of ego and a huge pissing contest, they're all still existing. I vote we knock out the americans and the japanese and keep europe. Sorry, NSA luv u tho xoxo. The INSDC [http://www.insdc.org/] basically gives these guys the rights to hand out accession numbers for each sample, without which Gondor would indeed fall.

European Nucleotide Archive (ENA)

The ENA is part of EMBL-EBI. It contains sequences in the FASTA, FASTQ, CRAM, and other formats from different periods in their analyzed lifetimes. It contains short sequences and a section for whole genomes.

Links:

- http://www.ebi.ac.uk/ena
- Introduction: https://www.youtube.com/watch?v=ZeDB3X4G1gU

GenBank

- http://www.ncbi.nlm.nih.gov/genbank/
- Tutorial: https://www.youtube.com/watch?v=j7hV10gYz1Q

DNA Databank of Japan (DDBJ)

- http://www.ddbj.nig.ac.jp/
- Tutorial: https://www.youtube.com/watch?v=UF640aBB_c0

Nucleotide Database Questions

1. Name three of the most important / most informative entity-types that can be found in EMBL or Entrez Gene?
2. Which objects in biology correspond to these entity-types?
3. Define a gene and name three attributes
4. Who is the owner of an entry in EMBL?
5. Name the most important data types / entity types that you have to provide with an entry in EMBL
6. What means “vector clipping”?
7. How come that so many genes have more than one name?
8. Explain the difference between a data repository and a curated database
9. What is a knowledge base?
10. What are mRNA, hnRNA, cDNA, rRNA and tRNA and how are they represented in EMBL?
11. What means “coding sequence” and what is a “non-coding sequence”?
12. how is information on the exon-intron-structure of a gene represented in an EMBL-entry?
13. Give examples for “values” of the entity type “molecule” in EMBL
14. Who issues an accession number?
15. What is the difference between an accession number and a database identifier?
16. Why is EMBL synchronized with DDBJ and NCBI/GenBank?
17. Why is EMBL split in EMBL, EMBL updates and EMBL (whole genome shotgun) at the SRS interface?
18. What is a contig?
19. What means “whole genome assembly”?
20. What differentiates EMBL from ENSEMBL?
21. Sketch the process of sequencing and identify possible sources for errors
22. Name at least three attributes listed under “Features” in a standard EMBL entry
23. Sketch the major parts of a typical EMBL entry: what categories does a “normal” EMBL entry have?
24. What are “EST sequences” ? and which database comprises information on EST sequences?

Genome Databases

Genome databases contain the entire nucleotide sequences for an organism, and information about the assembly used.

Builds are basically the name for an entirely assembled chromosome, or genome. They have versions because they’re not perfect and are always being improved with more reliable sequencing and analysis techniques.

BLAST!

File Formats

- FASTA: Annotations plus {ACTG}* with 60? characters per line
- FASTQ: extra information on top of FASTA plus quality scores for each line

EBI Genomes within the European Nucleotide Archive (ENA)

Submission requires having annotations about samples, experiments, and runs. Like, what's in your cow? And I did a triplicate sequence on it. Nice. Includes viruses, bacteria, and eukaryotes.

- http://www.ebi.ac.uk/genomes/
- There will probably be a question about how to submit data to this database (https://www.youtube.com/watch?v=atkePLnse7A)

NCBI Genomes

- http://www.ncbi.nlm.nih.gov/genome/
- Help Page: http://www.ncbi.nlm.nih.gov/books/NBK3837/

Ensembl

Contains full genomes with alignments to entries in NCBI Reference Sequence and UniProt. Ensembl produces automated gene annotations by finding clusters of these entries.

Its genome viewer also contains other powerful annotation overlays to contigs, sequence variations (SNPs) from NCBI dbSNP, comparative genomics, and functional genomics from ENCODE. One of these annotations is Human and Vertebrate Analysis aNd Annotation (HAVANA), a set of manually curated gene annotations for whole genomes. Manually curated sets are a good benchmark to include against an automated process.

The automated annotations Include annotations of introns, exons, and noncoding regions. All annotations are colored by source.

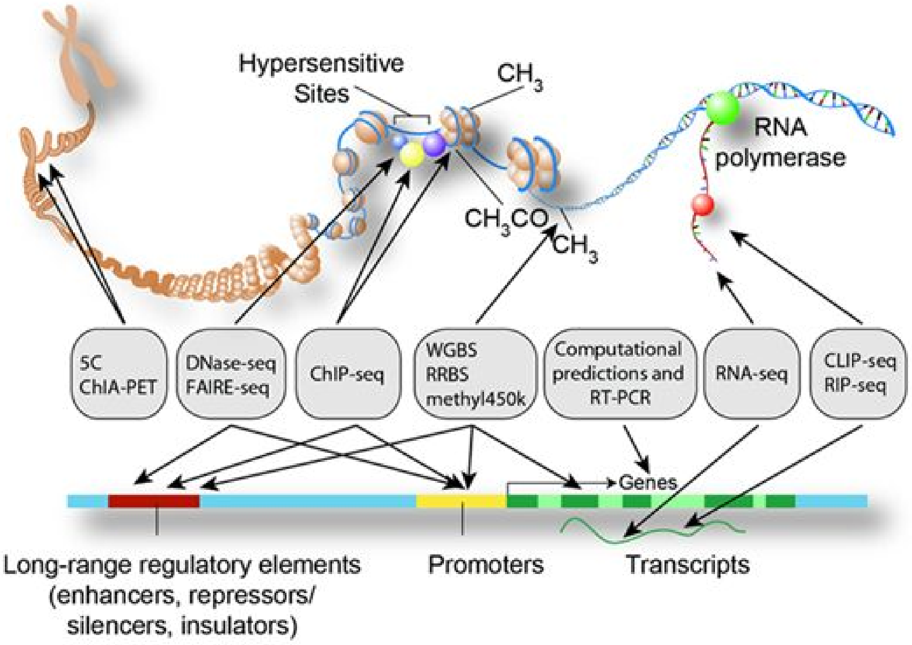
Links:

- http://www.ensembl.org/
- Overview: https://youtu.be/ZpnQOOxXufM
- Ensembl contains an interface called BioMart (http://www.ensembl.org/biomart/) that allows for bulk download of data

Encyclopedia of DNA Elements (ENCODE)

Encode is a a set of hella experiments on regulatory data- available information about DNA Methylation, Histone Protein locations and transcription binding with chromatin immunoprecipitation (ChIP), open chromatin tests with DNAase1, methylation with H3L4Me, etc. Really impressive.

From their site "*Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing.*"



- https://www.encodeproject.org/

Other Organism Specific Genome Databases

- Saccharomyces Genome Database (SGD) http://www.yeastgenome.org/
  - a comprehensive integrated biological information for the budding yeast *S. cerevisiae* along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and gene products in fungi and higher organisms.
- Rat Gene Database http://rgd.mcw.edu/
  - the premier site for genetic, genomic, phenotype, and disease data generated from rat research. In addition, it provides easy access to corresponding human and mouse data for cross-species comparisons.
- Zebrafish Model Organism Database http://zfin.org
- Plant Genome Database (PlantGDB) http://plantgdb.org/
  - PlantGDB provides sequence data for >70,000 plant species, custom EST assemblies (PUT) for over 150 species, web tools and plant genome browsers, as well as an outreach portal for plant genomics

Gene Databases

Hugo Gene Nomenclature Committee (HGNC)

HGNC provides unique and stable identifiers for human loci, which includes protein coding genes, ncRNA genes, pseudogenes, and other stuff.

Links:

- http://www.genenames.org/

Entrez Gene

Pronounced ahn-tray. This is the NCBI's internal gene accession number. It's not so stable, or sensical, for that matter. Smear.

Links:

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

SNP Databases

Refresher on the difference between SNP versus Mutation: There needs to be a certain threshold frequency of a change in the population for a mutation to be considered a SNP, while a mutation is any change from one base to another. Mutations have a strong biological impact and are easier to speculate on than SNPs. SNPs are often found in the intergenic regions.

A haplotype is a genetic pattern that is inherited together.

In the future, whole genome sequences will help to identify rare SNPs, ones that don't even make the cut of normal SNP chips.

Laboratory Methods and Background

Genome-wide Association Study (GWAS)

A Genome-wide Association Study (GWAS) tries to build a statistical correlation between the occurrence of a certain SNP and a phenotype. It uses very simple statistics, like the Fisher's Exact Test and Bonferroni Correction for Multiple-Hypothesis Testing. This means that since many different conditions/phenotypes are being analyzed by scientists, we need to set the bar for significance very high so we don't get results by accident.

Interestingly, GWAS studies often find disease associated to intergenic SNPs.

Knockdown and Knockout Experiments

GWAS experiments are often followed up by a knockdown experiment in mice to help get a mechanistic understand of how a SNP contributes to the measured phenotype. This is pretty simple, and commonly done with RNAi, shRNA, and now CRISPR-Cas9. Further characterization is usually needed.

Expression quantitative trait loci (eQTL)

Expression quantitative trait loci (eQTLs) are genomic loci that contribute to variation in expression levels of mRNAs.

While a GWAS can associate a phenotype to certain SNPs, eQTL studies attempt to link SNPs to gene expression (in a specific cell type).

https://en.wikipedia.org/wiki/Expression_quantitative_trait_loci

Linkage Disequilibrium

Linkage disequilibrium looks at 2 SNPs together and asks how well they correlate. In the HapMap, cliques of SNPs who all have a correlation of > 0.8 are called LD-bins. These bins reflect areas with little or no recombination (if the SNPs occur close in physical location, though that's often not the case). Since these bins are experimentally determined though, we get results of physically distant SNPs correlating highly.

For a real experiment, it's hard to measure all of the SNPs, so a couple representative SNPs are picked from each LD-bin, called Tag-SNPs.

NCBI dbSNP

dbSNP and Ensembl are both genetic variation databases. In addition to its genome data, Ensembl also contains information about structural genetic variation, such as copy number, inversions, and translocations. Like most NCBI websites, dbSNP looks terrible. This website lets you look up information about a SNP based on its RS number, like genomic position, chromosome number, related genes, etc.

Links:

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

Genome.gov and GWAS Catalog

NCBI's Genome.gov and EBI's each hold catalogs for published GWASs.

Links:

- http://www.genome.gov/
- http://ebi.ac.uk/gwas

Regulatory Elements Database

Regulatory regions can be very far away from a gene, even 1 megabase. One element can even regulate multiple genes.

The Regulatory Elements Database is part of the ENCODE project. It's useful to start with a given genomic coordinate and find regulatory regions, or start with a gene and get regulatory elements.

Unmet need: what cell type is your gene mutation affecting most, causing a disease?

Links:

- http://dnase.genome.duke.edu/

RegulomeDB

RegulomeDB contains information about what regulatory elements in the intergenic space that a given SNP has been measured/predicted to interact with/be a part of. These regions shot DNAase hypersensitivity, are the binding sites of transcription factors, and/or are promoter regions.

Links:

- http://regulomedb.org/

HapMap

There are 8 million common SNPs that are matched with obvious traits (eyes, hair, ethnicity, etc.). Rare allele frequency is <10%, and if you have a <1% occurrence, it's not really an allele anymore.

The goal of HapMap is to identify the different haplotypes in different ancestries (African, European, Asian) and produce tags to go with each haplotype (see LD-Bins above)

Links:

- http://hapmap.ncbi.nlm.nih.gov/

- Finding Genes for Human Disease by Lynn Marquis https://www.youtube.com/watch?v=w1IJO_0t7Jg

Functional SNP Database

This database/service relates data out of lots of other sources including Ensembl, dbSNP, HapMap, RegulomeDb to predict the functional effect of SNPs on transcription, translation, PTMs, etc.

Links:

- http://compbio.cs.queensu.ca/F-SNP/

Disease Specific Databases

Disease specific databases curate data, and add annotations from the data's sources that aren't included in the general databases. Useful tool is the MISO Sequence Ontology Browser.

- Catalog of Somatic Mutations in Cancer (COSMIC) [http://cancer.sanger.ac.uk/cosmic]

Proteins

Protein Data Bank (PDB)

The PDB contains 3D X-ray crystal structures and some NMR structures. It has its own implementation of the BLAST algorithm for sequence alignments, and it also has 3D similarity searches as well.

Links:

- http://www.rcsb.org/

UniProt Knowledgebase

The UniProt Knowledgebase is the central resource for data on proteins. It contains:

- GO Annotations for function and biological processes
- Annotations to pathology

It also links to information about:

- Protein-protein interactions (Reactome)
- Inhibitors (ChEMBL, DrugBank, BindingDB)
- 3D Structure (PDB)
- Families and Domains (Pfam, InterPro, etc.)
- Amino Acid Sequence (UniParc)
- Genetic Variants (dbSNP)
- Tons of cross-references

Each annotation in UniProtKB also contains a link to its evidence, whether it be a database, manual curation, assertion by similarity, a publication, or automatic curation. UniProtKb consists of TrEMBL, which automatically annotates proteins, and SwissProt, which consists of manually curated annotations. The Protein Information Resource also is part of the UniProt Consortium.

TrEMBL

Trembl will look at cDNA sequences in both the 3’->5’ and 5’->3’ direction and look for start codons in all 3 possible frames. It translates in silico, and usually keep the longest as a putative protein.

Links:

- http://www.uniprot.org/uniprot
- http://pir.georgetown.edu/

UniParc and UniProt Reference Clusters (UniRef)

UniParc contains protein sequences. They are imported from different databases and have their own stable accession numbers.

UniRef is a collection of three neural networks (look up “radial basis function network”!!!) that cluster protein sequences from UniParc by three different conditions. Users can query it with a seed sequence to assign it to a cluster. Each entry in UniRef represents one cluster.

1. UniRef100 clusters identical sequences over 11 residues from any organism into a single UniRef entry
2. UniRef90 analyzes the longest sequence from each cluster in UniRef100, and builds new clusters of groups of sequences that all have at least 90% identity with every other element of the cluster. Each cluster is given an entry. Each element in the cluster must have at least 80% overlap with a seed sequence to be given as a result.
3. UniRef50 does the same as UniRef90 by taking the longest sequence from each cluster in UniRef90 and making new clusters sharing 50% identity. Each element in the cluster must have at least 80% overlap with a seed sequence to be given as a result

Links:

- http://www.uniprot.org/uniparc/
- http://www.uniprot.org/uniref/

Structural Classification of Proteins (SCOP)

SCOP provides a hierarchical classification for proteins with PDB entries based on their 2D and 3D structural features.

The tree is divided in the following order:

1. Class
2. Fold
3. Superfamily
4. Family
5. Protein
6. Species

Links:

- http://scop.mrc-lmb.cam.ac.uk/scop/
- http://supfam.org/SUPERFAMILY/

Protein-Protein Interactions (PPIs)

PPI’s are the final frontier of drug discovery. It’s still incredibly novel to find small molecules that can disrupt the binding between two proteins. Maybe allostery can muck it up, but it would be really exciting to find molecules that get in the way of these interactions. On a biological level, these databases help make assertions about which proteins are doing what, and later help give evidence when building pathways.

Laboratory Techniques

- Forster Resonance Electron Transfer (FRET)
- Co-immunoprecipitation (Co-IP)
- Yeast 2 Hybrid
- Synthetic Gene Array

BioGrid

BioGrid is a curated database of protein-protein and protein-genetic interactions. This was the one presented in class. Does network analysis of connectivity coefficient of connected neighbors to a given gene.

Links:

- http://thebiogrid.org/

Other PPI Databases

- Biological Interaction Network Database (BIND) [https://www.bindingdb.org/]
- Database of Interacting Proteins (DIP) [http://dip.doe-mbi.ucla.edu/]
- IntAct [http://www.ebi.ac.uk/intact]
- Molecular Interactions Database (MINT) [http://mint.bio.uniroma2.it/mint]
- Human Protein Reference Dataset (HPRD) [http://www.hprd.org/]

Agile Protein Interaction DataAnalyzer (APID)

APID provides access to experimentally validated PPIs in BIND, BioGRID, DIP, HPRD, IntAct, and MINT. Allows for exploration of the “interactome” network. It’s notable because it aggregates the results of all of these databases, which aren’t synchronized. This was the one presented in class.

Links:

- http://bioinfow.dep.usal.es/apid/

Other PPI Data Aggregation Services

- Microbial Protein Interaction Database (MPIDB) [http://jcvi.org/mpidb/]
- Protein Interaction Network Analysis (PINA) [http://cbg.garvan.unsw.edu.au/pina/]

PPI Predictions

These are more web services than databases. They use information from PPI databases and machine learning to make predictions about binding surfaces and interactions. STRING uses network analysis, PIP uses naïve bayes (lol), and MiMI is discontinued, because Michigan isn't even a real state.

- STRING [http://string-db.org/]
- PIP [http://www.compbio.dundee.ac.uk/www-pips/]
- MiMI [http://mimi.ncibi.org/]

Pathway Databases

Pathway Databases store informations about biochemical, signaling, metabolic, and other pathways. Though pathways are all part of larger and more complex networks, these databases allow for reasoning over smaller units.

Kyoto Encyclopedia of Genes and Genomics (KEGG)

KEGG is a huge knowledge base, consisting of a database of pathways, enzymes, transformations/biochemical reactions, and tons of other stuff. It's like the Japanese EBI. It has lots of cross-links to other databases (within KEGG, albeit) but they provide lots of context that other databases can't. Go Japan. And GIF image maps. Hell yeah.

- KEGG Enzymes contains data about enzymes
- KEGG Reactions contains all data about reactions
- KEGG RPAIR contains compound pairs from enzymatic reactions and their enzyme

Broad Definitions of Reactions contains:

- Transportation
- Classical Biochemical Transformation
- Binding
- Dissociation
- Degradation
- Phosphorylating
- Dephosphorylation

Links:

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

WikiPathways

WikiPathways is built on the community editing, open sourced software behind MediaWiki. It has a custom editor that allows for community curation of pathways. They're stored in GPML, an extension of XML compatible with visualization software such as Cytoscape. WikiPathways also contains data from KEGG

Links:

- http://www.wikipathways.org/

Reactome

Reactome is an open-source, curated pathway database, with lots of cross-references to other databases. It has an incredibly powerful browser as well.

- http://www.reactome.org/
- http://www.reactome.org/PathwayBrowser/

PubMed and Medline

MEDLINE is the database of documents (medical literature, books, publications) that PubMed uses as a service to link to those documents. This is the first stop in all searches for informations about biological stuff. Faceted search can be performed and MeSH terms can be used to make even more powerful queries. Hofmann loves to ask questions about the publication on certain subject vs. time histogram.

Links:

- http://www.ncbi.nlm.nih.gov/pubmed
- https://health.ebsco.com/products/medline

Ontologies

Medical Subject Headings (MeSH)

MeSH is a hierarchical standardization of controlled vocabulary and synonyms used by the National Library of Medicine. It isn't exactly an ontology, but more of a thesaurus (hierarchical, but with multiple inheritance). As papers are published, they get the appropriate MeSH terms annotated to them. This makes MeSH an invaluable tool for literature search on repositories like Medline and PubMed. The synonyms dictionary is also useful for text mining.

Links:

- http://www.ncbi.nlm.nih.gov/mesh
- https://www.nlm.nih.gov/mesh/MBrowser.html

Gene Ontology (GO)

GO consists of annotations for proteins functions, their involvement in processes, and their cellular location. These annotations have been widely adopted across platforms describing proteins and gene products in many organisms. Each ontology is hierarchical to allow for powerful searches of variable granularity. You'll notice GO Terms in UniProtKB.

Three GO's:

- GO Process
- GO Component
- GO Function

Links:

- http://geneontology.org/

The Sequence Ontology (SO)

The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. Examples are *binding_site* and *exon*. Biomaterial features are those which are intended for use in an experiment such as aptamer and *PCR_product*. There are also experimental features which are the result of an experiment.

The Sequence Ontologies are provided as a resource to the biological community. They have the following obvious uses:

- To provide for a structured controlled vocabulary for the description of primary annotations of nucleic acid sequence, e.g. the annotations shared by a DAS server (BioDAS, Biosapiens DAS), or annotations encoded by GFF3.
- To provide for a structured representation of these annotations within databases. Were genes within model organism databases to be annotated with these terms then it would be possible to query all these databases for, for example, all genes whose transcripts are edited, or trans-spliced, or are bound by a particular protein. One such genomic database is Chado.
- To provide a structured controlled vocabulary for the description of mutations at both the sequence and more gross level in the context of genomic databases.

Links:

- http://www.sequenceontology.org/
- wiki: http://www.sequenceontology.org/so_wiki/index.php/Main_Page

MISO: the Sequence Ontology Browser

Useful for the following:

- Search for a SO term by entering a SO term name or synonym in the query box above;
- Explore the structure of SO and browse for SO terms using the expandable, cascading tree on the left;
- Go to the detail page for a term where you can:
  - Get details about a term, it's definition and relationships;
  - See graphical views of a term's place in the ontology and link to it's neighbors;
  - Export details about a term in a variety of formats;
  - And access and contribute detailed documentation about a term and it's history by linking through to the SO wiki.

Links:

- http://www.sequenceontology.org/browser/obob.cgi

The Ontology for Biomedical Investigations (OBI)

The Ontology for Biomedical Investigations (OBI) addresses the need for controlled vocabularies to support integration and joint ("cross-omics") analysis of experimental data, a need originally identified in the transcriptomics domain by the FGED Society, which developed the MGED Ontology as an annotation resource for microarray data. OBI uses the Basic Formal Ontology upper level ontology as a means of describing general entities that do not belong to a specific problem domain. As such, all OBI classes are a subclass of some BFO class.

The ontology has the scope of modeling all biomedical investigations and as such contains ontology terms for aspects such as:

- biological material - for example blood plasma
- instrument (and parts of an instrument therein) - for example DNA microarray, centrifuge
- information content - such as an image or a digital information entity such as an electronic medical record
- design and execution of an investigation (and individual experiments therein) - for example study design, electrophoresis material separaition
- data transformation (incorporating aspects such as data normalization and data analysis) - for example principal components analysis dimensionality reduction, mean calculation

Less 'concrete' aspects such as the role a given entity may play in a particular scenario (for example the role of a chemical compound in an experiment) and the function of an entity (for example the digestive function of the stomach to nutriate the body) are also covered in the ontology.

Links:

- http://obi-ontology.org/page/Main_Page

International Classification of Diseases (ICD)

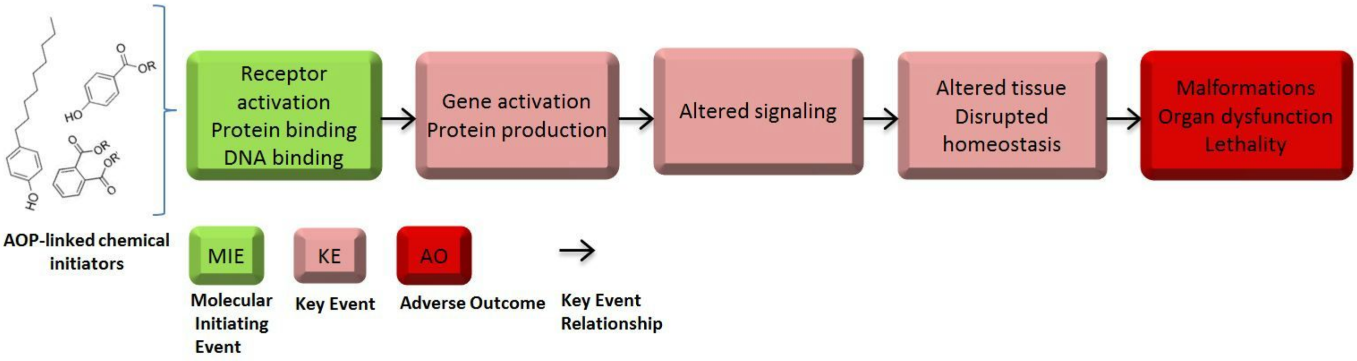
The ICD is the World Health Organization's ontology for diseases. It is not the same as MeSH, though there is a bit of shared stuff between them.

Links:

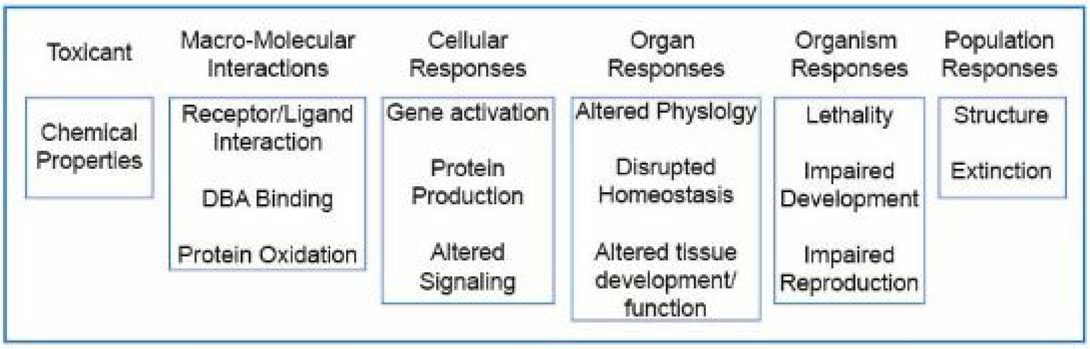
- https://bioportal.bioontology.org/ontologies/ICD10

Challenges

Adverse Outcomes Pathology



Source: http://aopkb.org/background.html



Toxicant	Macro-Molecular Interactions	Cellular Responses	Organ Responses	Organism Responses	Population Responses
Chemical Properties	Receptor/Ligand Interaction	Gene activation	Altered Physlogy	Lethality	Structure
	DBA Binding	Protein Production	Disrupted Homeostasis	Impaired Development	Extinction
	Protein Oxidation	Altered Signaling	Altered tissue development/ function	Impaired Reproduction	

Source: http://www.oecd.org/chemicalsafety/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm

Links:

- Adverse Outcome Pathway KB (AOP-KB) [http://aopkb.org/]
- http://www.oecd.org/chemicalsafety/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm
- http://aopkb.org/background.html - read this. it basically has the answer to hofmann's question

Systems Toxicology

- Toxicology - how chemicals mess stuff up. See ADME**T**
- How do chemicals modulate biological pathways
- Make models of this = systems toxicology
- Omis, Transcript, protein, and metabolite profiling

Useful Databases:

- BRENDA
- KEGG
- Reactome

Links:

- http://www.ncbi.nlm.nih.gov/pubmed/22562485
- http://www.nature.com/nrg/journal/v5/n12/execsumm/nrg1493.html
- http://pubs.rsc.org/en/content/articlelanding/2015/tx/c4tx00058g#!divAbstract