BioDatabases January 2016

Questions for the exam

In this file all questions are collected that have ever been seen in the questions lists of Prof. Hofmann and also in the exams of the senjor students.

All questions are divided into 9 groups:

1. Basics

2. Literature databases, controlled vocabularies and gene catalogues

3. Nucleotide Databases

4. Nucleotide related databases

5. Microarray databases

6. Protein databases

7. Interaction databases

8. Enzyme and metabolic pathway databases

9. Chemical databases

Red questions show that they have been found in the real exam!

Sometimes we have found more answers to the same question (from different students), so there after the question you can see more then one possible solution (showed with "OR").

Sorry for the huge amount of misprints and just grammar mistakes!) Sometimes we could not see the full answer or just could not understand the handwriting))

Possible Questions on Bio-Databases

1. Basics

1.1. Object orientation: What is an object in Biology and what "methods" can biological objects execute? Provide at least three examples for "methods" executed by bio-objects.

Similarly to OOP (Object Oriented Programming), biological entities of real world and their interactions can be represented as "bio-objects". A bio-object is a real world abstraction where a defined entity can be classified and assigned to a class model. This model tries to reflect real world properties and interactions. Thanks to that we can store data related to these interactions.

A class can also be hierarchical represented, can be son of other class and inheriting properties ("is-a" relationship), or be part of other class ("has-a" relationship).

Object: instance of class that contains unique properties. The actions that objects can play are the methods.

DNA methods:

- Strand separation
- Replication
- Packaging in/with chromatin
- Re-association/hybridization

RNA methods:

- Interaction with ribosomes
- Enzymatic functions
- Structure functions
- Transport functions

Protein methods:

- Interaction with other biomolecules
- Enzymatic functions
- Structure functions
- Transport functions

Lipid methods:

- Interaction with other biomolecules
- Association to form membranes
- Binding to proteins

Carbohydrate methods:

- Interaction with other biomolecules
- Polymerization
- Conjugation

1.2. What are typical attributes of Nucleic Acids? Name at least three of them.

Attributes of NucleicAcid Objects

The attributes of a NucleicAcid object are:

Name	Type	Description
adjacent_ligands	list of Ligands	List of ligands whose BindingSites include at least one residue from this nucleic acid. The list is ordered by the size of the nucleic acid/ligand interaction, i.e. when there are two or more ligands in the list, the first will have more of this nucleic acid's residues involved in its binding site than will the second. If the nucleic acid has been subjected to a geometrical transformation then all ligands in the list will be transformed in the same way.
atoms	list of Atoms	The Atom objects in the nucleic acid.
bonds	list of Bonds	The Bond objects in the nucleic acid.
chain_id	string	Nucleic acid chain identifier, e.g. *B* (returns the single-character string "-" if no identifier available).
n_atom	integer	Number of atoms in the nucleic acid.
n_unit	integer	Number of units (i.e. residues) in the nucleic acid chain; equivalent to len(nucleic_acids).
pdb	PDB	The PDB object that contains the nucleic acid.
residues	list of Residues	The Residue objects in the nucleic acid.
sequence_3d	string	String containing the nucleic acid sequence as one-letter codes, e.g. *ATTAGTA*. The sequence returned is that determined from the residues in the PDB ATOM records, not the sequence defined in PDB SEQRES records. These may differ, e.g. the experimental sequence would not include residues whose 3D atomic positions were not determined because of crystallographic disorder.
type	string	String that identifies this object type. For a NucleicAcid object, this will be the string *nucleic_acid*

So I suggest the following attributes:

- Ligand binding or interaction
- Bases: either ATGC (DNA) or AUGC (RNA)
- Nucleotides: Desoxyribonucleotide, Ribonucleotide
- Genetic coding

Or:

- Sequence
- Secondary structure
- Chromatin packaging
- Patterns and motifs as binding sites for proteins

1.3. Which categories of biomolecules do you know?

. Nucleic acids (DNA, RNA)
. Proteins and enzymes
. Lipids
. Sugars
. Small molecules (<600 daltons, substrates for enzymatic reactions, catabolism products, signaling molecules)
. Conjugates and polymers

1.3.* List at least three different classes of biomolecules and their corresponding databases. EXAM!

Molecule	DB
DNA	EMBL, GeneBank, DDBJ
RNA	Sequence: EMBL, GeneBank, DDBJ Expression: ArrayExpress, GEO BMD
Protein	Domain: ---- 3D: PDB, MMDB Sequence: UniProt/IntAct, ---?

1.4. Which categories of Bio-Databases correspond to these categories of biomolecules?

☐ Nucleotide databases or genome databases correspond to DNA and RNA (EMBL GenBank)

☐ Protein databases correspond to proteins (SWISSPROT)

☐ Small molecule databases correspond to lipids sugars and other small molecules (LIPIDBANK)

- ☐ Macromolecular structure databases correspond to conjugates and polymers (MSD by EBI)

1.5. Please give at least one example for each category of Bio-Databases as we see them categorized at the EBI SRS interface.

Literature, bibliography and reference databases: MEDLINE, OMIM

Gene Dictionaries and Ontologies: GO, HGNC, UniGene, ENTREZ

Nucleotide databases: EMBL, ENSEMBL

Protein databases: UniProt, IPI, InterPro

Structure databases: MSD

Nucleotide related databases: TF (SITE, FACTOR, CELL, GENE)

Protein Function, structure and interaction DB: InterPro, PFam, PDB, IntAct

Enzymes, reactions and metabolic pathway: ??

Small molecule databases: ChEBI

Mutation and SNP

ArrayExpress

1.6. Which major portals to bio-data do you know? EXAM

- SRS at EBI
- ENTREZ at NCBI
- Academic info/biodata .net
- Science Magazines websites like sciencemag.org
- Annual update published by NAR (Journal of Nucleic Acid Research)
- Expasy.org (not updated anymore)

1.7. What sort of discrepancy exists between Bio-Databases that represent information on genes and genomes as opposed to Bio-Databases that store information on gene expression and what are the consequences for the database design? EXAM

Sequences that are displayed by genes and genomes databases have a consistent representation

inter-species wise. Gene expression data is more confined to the special sample, experiment reflecting the status of our bio-sample at a specific time i.e. the results cannot be generalized over the whole species.

BioDBs – gene and genome	**BioDBs - gene expression**
➢ static information, does not change with time	➢ dynamic information, changes with time, conditions and org
➢ contains data regarding gene sequence, location on the chromosome etc.	➢ contains data about experiment conditions, cell types, time etc on which gene expression
➢ comparatively less volume of data	➢ huge volume of data, increases with time (splicing, posttranscriptional modifications)
➢ ensemble	➢ array express, GEO

OR

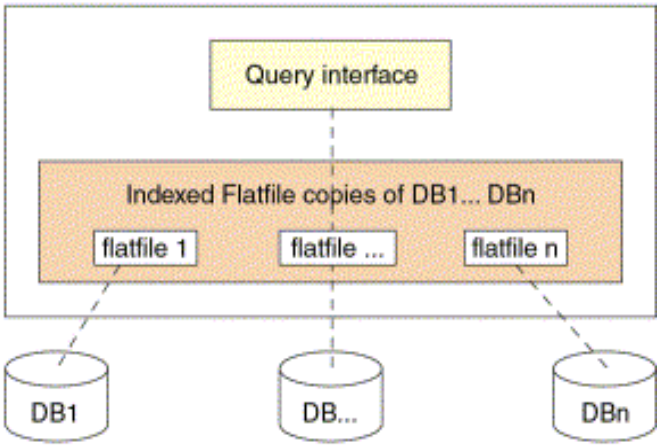
Gene/Genome information is represented as a sequence, a string of characters. It also has annotation attached with information about the sequence. On the other hand we want to quantify the amount of mRNA of a sample with some treatment. Here the treatment of the sample as of maximal ------ and the quantity in normally expressed as a difference with a controlled sample.

All of them should be captured when strong gene information experimental design, or raw data, processed data, treatment information.

1.8. What is a "flat file database"?

A flat file database is described by a very simple database model, where all the information is stored in a plain text file, which is ordinary unformatted file that can be only be read or written sequentially, e.g. XML, PDB format. One database record per line. Example: GenBank

Indexing flat-files: Allows to integrate high number of heterogeneous databases Principle: databases to be integrated are provided as flat-files. Indexing flat-files system can handle many simultaneous users.



1.9. What features would you assign to "Bioinformatics" and how does it differ from "Systems Biology"?

Bioinformatics:

Combination of Biology and informatics. It's the application of computer science and information technology to the field of biology and medicine. It deals with algorithms, databases, and information systems and soft computing data mining etc.

Systems Biology:

Aims to explain how higher level properties of complex biological systems arise from the interaction among their parts.

Bioinformatics can help in Systems Biology, but their field covers more topics. Genome, transcriptome and proteome are the features that can be assigned to Bioinformatics. Since Bioinformatics gives information about them.

For System Biology, in additional to genome, transcriptome and proteome, metabolome can also be assigned as a feature. System Biology is a computational and experimental research group while Bioinformatics is advanced genome information technology research group.

1.9* Define systems biology EXAM

Study of an organism viewed as an integrated and interacting network of genes, proteins and biochemical reactions which give use to life. Networks come from graph theory.
A model - Representing of a real world system
Alternative types of models in system biology?

- Systems biology applications
- Pharmaceuticals – drug discovery, biomarker discovery
- Biodefense
- Biomedication

1.10. Annotate a sketch of biomolecules with entity-types from Bio-Databases you know

("Funktionseinheit" such as structure of protein, interaction, functions).

I guess here we should match biomolecules with entity-types in databases.

. Entity type	. BioDB
. DNA and RNA	. EMBL, Genbank
. Protein	. Uniprot
. Lipid	. Lipidbank
. Carbohydrates	. EurocarbDB

1.11. How are Bio-Databases integrated in SRS? What does the SRS documentation say about the mechanisms used for linking between Bio-Databases?

SRS was developed at the European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI). It allows linear databases to be indexed to other linear databases. In the data warehouse approach to data integration, data from individual databases is usually transformed into a common format and then stored in a large single database called a data warehouse. The advantages of physical integration are that queries can be executed rapidly because all the data are located in one place, and the end user sees a homogeneous, integrated data source. SRS allows virtually any linear database to be included in the data warehouse, but linking requires explicit cross-referencing. Hyperlinking is extensive in SRS.

SRS is listing all the databases contains a link to a description page about the database including the date on which it was last updated. You select one or more of the databases to search before entering your query. A databank entry may contain references to other databanks, and vice versa. In SRS these relationships are known as links and can be used to extend a query across multiple databanks. Thus you can obtain all the entries in one databank that are linked to an entry (or entries) in another databank. From a user

perspective there are two types of link: hypertext links and index links (query links). Hypertext links are links between entries which are displayed as hypertext. These are hardcoded into SRS and you can use them whenever you wish. They are useful for examining entries that are referenced directly from entries. Index links are built into the SRS indices at the same time as databanks are added. They allow you to construct queries using relationships between databanks. They require SRS to search through entries or indices in other databanks, looking for matches. It is assumed that you are already familiar with hypertext links, so only a limited demonstration of them is given in this chapter (see section "Hypertext Links"). The remainder of this chapter is devoted to explaining index links.

SRS is service which provide access to the databanks and tools which have been available. Public SRS service contains details of the databanks.

SRS leaves the underlying data sources in their original formats (e.g. GenBank in flat file format, GO in XML or relational tables [MySQL]). The databases are integrated through meta-data.

Three ways of cross-reference:

. Hypertext links:
- Links between entries which are displayed as hypertext. These are hardcoded into SRS and you can use them whenever you wish. They are useful for examining entries that are referenced directly from entries.

. Indexed links:

- Built into the SRS indices at the same time as databanks are added. They allow you to construct queries using relationships between databanks. They require SRS to search through entries or indices in other databanks, looking for matches.
- Example: “give me all entries in Swiss-Prot that are linked to EMBL”

. Composite structures

1.12. What does “computer readable knowledge” mean?

Or machine readable knowledge. It means data can be stored data in computer readable format.

We store knowledge in a computer-readable form, usually for the purpose of having automated deductive reasoning applied to them. They contain a set of data, often in the form of rules that describe the knowledge in a logically consistent manner. An ontology can define the structure of stored data - what types of entities are recorded and what their relationships are. Logical operators, such as *And* (conjunction), *Or* (disjunction),

implication and *negation* may be used to build it up from simpler pieces of information. Consequently, classical deduction can be used to reason about the knowledge in the knowledge base.

- Storing data in computer readable format

- Having automated deductive reasoning

biomolecules	**Bio-Database**	**Entity-Type**
DNA and RNA	EMBL, GenBank	Molecule-value
Protein	Uniprot	ProteinID
Lipid	Lipidbank	lipidname
Carbohydrates	EurocarbDB	

1.13. The EBI call itself "the portal to knowledge". How is biomedical knowledge represented in Bio-Databases?

The main missions of the European Bioinformatics Institute (EBI) centre on building, maintaining and providing biological databases and information services to support data deposition and exploitation. Integrating biological and molecular annotations are based on the semantic knowledge ***represented*** in cross- references between databases. EBI is like a gateway to BioDatabases. Biomedical knowledge is classified by specific area and stored in databases. EBI is the integration of these databases. SRS can be used to search relevant biological related information on it.

EBI is a kind of gateway to different freely available Bio-Databases and provides advanced bioinformatics training = EBI combines several Bio-databases. One can access the different Bio-Databases via the SRS Library Page. Different Databases use different formats to present their data.

1.14. In one of the links to "relevant background information", a primer on molecular biology mentions "linkage disequilibrium". What does this term mean?

-population genetics

> "Linkage Disequilibrium" is the presence of the statistical associations between alleles at different loci. Factors are different including selection, rate of recombination, mutation, and genetic drift. It is used when the occurrence of some combinations of alleles or genetic markers in a population is more often or less often than it would be expected from random formation of haplotypes from alleles based on their frequencies. In other words, it is the non-random association of alleles at 2 or more loci, not necessarily on the same chromosome.

In population genetics, **linkage disequilibrium** is the non-random association of alleles at two or more loci, not necessarily on the same chromosome. . In other words, linkage disequilibrium is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes (is a combination of alleles (DNA sequences) at adjacent locations (loci) on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci) from alleles based on their frequencies. It is not the same as linkage, which is the association of two or more loci on a chromosome with limited recombination between them. The amount of linkage disequilibrium depends on the difference between observed and expected (assuming random distributions) allelic frequencies. Populations where combinations of alleles or genotypes can be found in the expected proportions are said to be in **linkage equilibrium**.

This is related to the **International HapMap Project** is an organization that aims to develop a **hap**lotype **map** (**HapMap**) of the human genome, which will describe the common patterns of human genetic variation. HapMap is a key resource for researchers to find genetic variants affecting health, disease and responses to drugs and environmental factors.

1.15. In the 2012 database issue of the journal NAR (Nucleic Acid Research) the category "protein sequence databases" is subdivided into 6 sub-categories: list at least three of them.

- General sequence databases
- Protein properties
- Protein localization and targeting
- Protein sequence motifs and active sites
- Protein domain databases; protein classification

- Databases of individual protein families

1.16. The online version of the 2012 Bio-Database issue of NAR comprises how many entries in its Bio-Database list?

1380 entries of Databases sorted into 14 categories and 41 subcategories.

1.17. 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.

E-utilities = Entrez Programming Utilities

These are tools that provide access to Entrez data outside of the regular web query interface and may be helpful for retrieving search results for future use in another environment.

Entrez is NCBI's primary text search and retrieval system that integrates the PubMed database of biomedical literature with 39 other literature and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression.

The Entrez Programming Utilities (E-utilities) are a set of eight server-side programs that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI). The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. The E-utilities are therefore the structured interface to the Entrez system, which currently includes 38 databases covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature.

Summary on its USAGE:

A piece of software posts an E-utility URL to NCBI, then retrieves the result and processes them as required. Combining E-utilities components to form customized data pipelines within these applications, is a powerful approach to data manipulation.

1.18. What categories of Bio-Databases are integrated under SRS and which ones are not?

. **Library Group**	. **Databank name**	. **Site**
. Active protein	. REFSEQP,	.

sequence		
. Enzymes, reaction and metabolic pathways	. ENZYME,	.
. Gene dictionaries	. UniGene	.
. Literature, bibliography and reference DB	. Medline, OMIM, Taxonomy	.
. Mutations	.	.
. Nucleotide related DB	.	.
. Nucleotide sequence	. EMBL,	.
. Protein function DB	.	.
. Protein interactions DB	. IntAct	.
. Protein structure	. PDB	.

Analysis tools: CLUSTAL, FASTA, BLAST

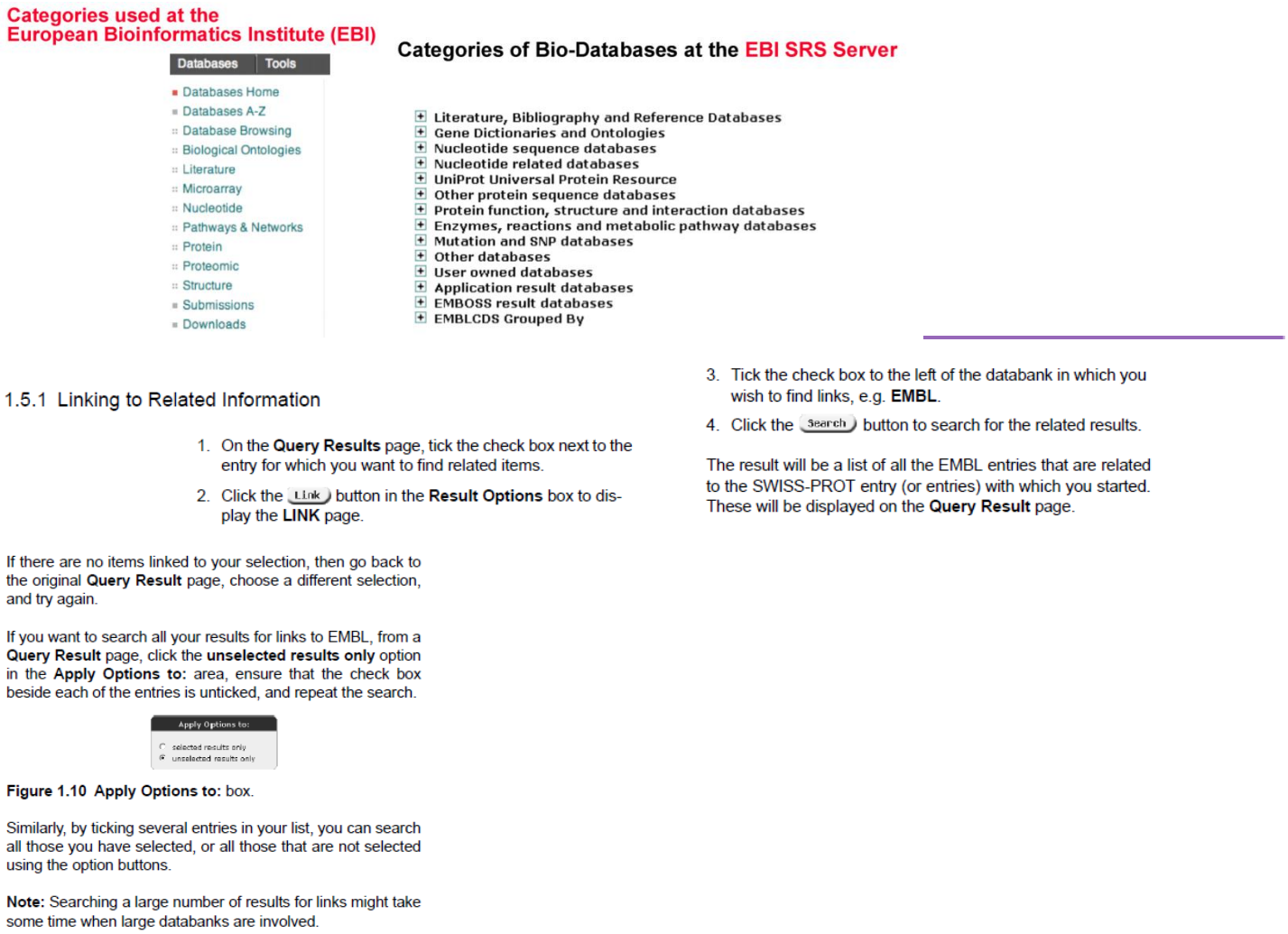
Not integrated: Metadatabase, Mathematical model databases, PCR/Real time PCR database. Genome DB, Gene expression DB, Pathway and Networks, Proteomics are not there.

1.19. 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? ???????

From: http://srs.ebi.ac.uk/srs/doc/srsuser.pdf

First select the library page and execute a query search. After the search, we obtain a subset. Now, click on the right panel - *Link to related information:* , Again, select the new library and click on search. Subset of the first query which has any link with the database you have chosen now will be shown.

Faceted search, also called **faceted navigation** or **faceted browsing**, is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters.



Categories used at the European Bioinformatics Institute (EBI)

Databases | Tools

- Databases Home
- Databases A-Z
- Database Browsing
- Biological Ontologies
- Literature
- Microarray
- Nucleotide
- Pathways & Networks
- Protein
- Proteomic
- Structure
- Submissions
- Downloads

Categories of Bio-Databases at the EBI SRS Server

- Literature, Bibliography and Reference Databases
- Gene Dictionaries and Ontologies
- Nucleotide sequence databases
- Nucleotide related databases
- UniProt Universal Protein Resource
- Other protein sequence databases
- Protein function, structure and interaction databases
- Enzymes, reactions and metabolic pathway databases
- Mutation and SNP databases
- Other databases
- User owned databases
- Application result databases
- EMBOSS result databases
- EMBLCDS Grouped By

1.5.1 Linking to Related Information

1. On the **Query Results** page, tick the check box next to the entry for which you want to find related items.
2. Click the Link button in the **Result Options** box to display the **LINK** page.

If there are no items linked to your selection, then go back to the original **Query Result** page, choose a different selection, and try again.

If you want to search all your results for links to EMBL, from a **Query Result** page, click the **unselected results only** option in the **Apply Options to:** area, ensure that the check box beside each of the entries is unticked, and repeat the search.

Apply Options to:
selected results only
unselected results only

Figure 1.10 Apply Options to: box.

Similarly, by ticking several entries in your list, you can search all those you have selected, or all those that are not selected using the option buttons.

Note: Searching a large number of results for links might take some time when large databanks are involved.

3. Tick the check box to the left of the databank in which you wish to find links, e.g. **EMBL**.
4. Click the Search button to search for the related results.

The result will be a list of all the EMBL entries that are related to the SWISS-PROT entry (or entries) with which you started. These will be displayed on the **Query Result** page.

1.20. Which bio databases that you know are based on the relational schema?

Relational DBs organize data into one or more tables (relations) of columns and rows, with a unique key identifying each row.
Majority of the BioDBs are not relational DBs. But Microarray Databases (GEO and ArrayExpress) are based on relational BioDBs. PDB, BRENDA, IntAct, REACTOME are also such BioDBs.

1.21. Describe the main feature of a relational database and the advantages over unstructured information?

A relational database is a collection of data items organized as a set of formally described tables from which data can be accessed easily. A relational database is created using the relational model. The software used in a relational database is called a relational database management system (RDBMS). RDBMS is the software program facilitating creation and updation of database. It is based on relational model.

Advantages:

- Avoids data duplication
- Avoids inconsistent data
- Easier to change data
- Easier to change data format
- Data can be added and removed easily
- Easier to maintain security

OR

Relational databases are composed of tables of data (each similar to a flat database). Each column in a table describes some attribute of the rows of the table. Each row is an actual object (usually called a record) in the database. Records in relational databases are unique (you cannot have two copies of the same data in a table). Each type of database has its use. The type of application that needs to access the data and the amount of data stored will determine the best type of database to use.

RDBMS main features:

1) Based on relational model (ER diagram)
2) Easy data retrieval
3) Easy to update data and data type
4) Easy to add or delete records from the data
5) Less redundancy
6) Consistent
7) Stable enables multiple users to work at the same time
8) Secure

1.22. Develop a simple database schema for a database that captures knowledge on the "basic dogma of molecular biology"

DNA --- RNA---- Protein

1.22* Develop a simple database – schema for a database that capture knowledge about "systems biology". The database schema should comprise the most relevant entities and relationships to represent information about the basic dogma of molecular biology. EXAM

Dogma: DNA --transcription-- RNA –translation -- Protein

There would be three kinds of entities, each with its own code and with links to the other entities: a gene would link to all known mRNA species and they could link to there protein sequence and its information.
??

1.23 Develop a simple database schema for a database that captures knowledge on "conjugates"

e.g. phopho-lipids; proteo-glycans; glyco-lipids

1.24 Name at least four different bio-objects and list at least three different attributes for each bio-object

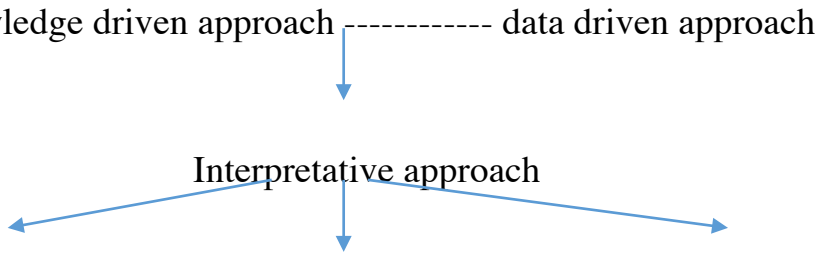
1.25 What are "high throughput data"? give examples for high throughput data generation technologies!

1.26 What is data, what is information and what is knowledge?

1.27 What is the difference between a database and a knowledge base? EXAM

Database is in essence, only a storage of data it may, not have context, meaning on relation to other data. In a knowledge base we store data such a way, that it provides information about connection to other data and has a context and a meaning, to answer the questions.

OR



Implication:

- Can be used for sparse data
- When there is lack of data
- Different biomaterial
- Technology constraint

1.28 What is the first normalization form? Give an example.

A relation is in first normal form if the domain of each attribute contains only atomic values, and the value of each attribute contains only a single value from that domain

1.29 Describe the structure of FASTA format.

- text-based format for representing either nucleotide sequences or peptide sequences

- begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional)

1.30 What is XML and XSD? Which steps are needed to transfer the information from a XML file to a database?

XML: Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

XSD: An XML schema is a road map for the XML document similar to a Document Type

Definition (DTD). Schemas describe the elements and map out the presentation and nesting of XML documents. Essentially, the schema enables all applications to understand the flow of the page and validate the elements.

How to transfer information from XML to database: Use script to translate hierarchical information in XML file into SQL statements.

1.31 How would you represent spatial and temporal changes in biological systems in a conceptual model of a biodatabase? Exam 2008

1.32 Give two examples for relationship-types used in a relational biodatabase!

1.33 Explain the difference between a knowledge base and a data repository and provide one example for each one - exam 2008

1.34 What is meant by curation? Provide at least two examples for "curated" Databases. Exam 2014

Curation means standardization and annotation of data so that everyone can understand. To create the database; curators need to acstract and organize data from literature. They also need to describe the data in terms of standarts, protocols, vocabulary.

Example: PDB, SwissProt, GEO, etc.

1.35 Which advantages have database management systems in comparison to a) text and b) XML files? Exam 2014

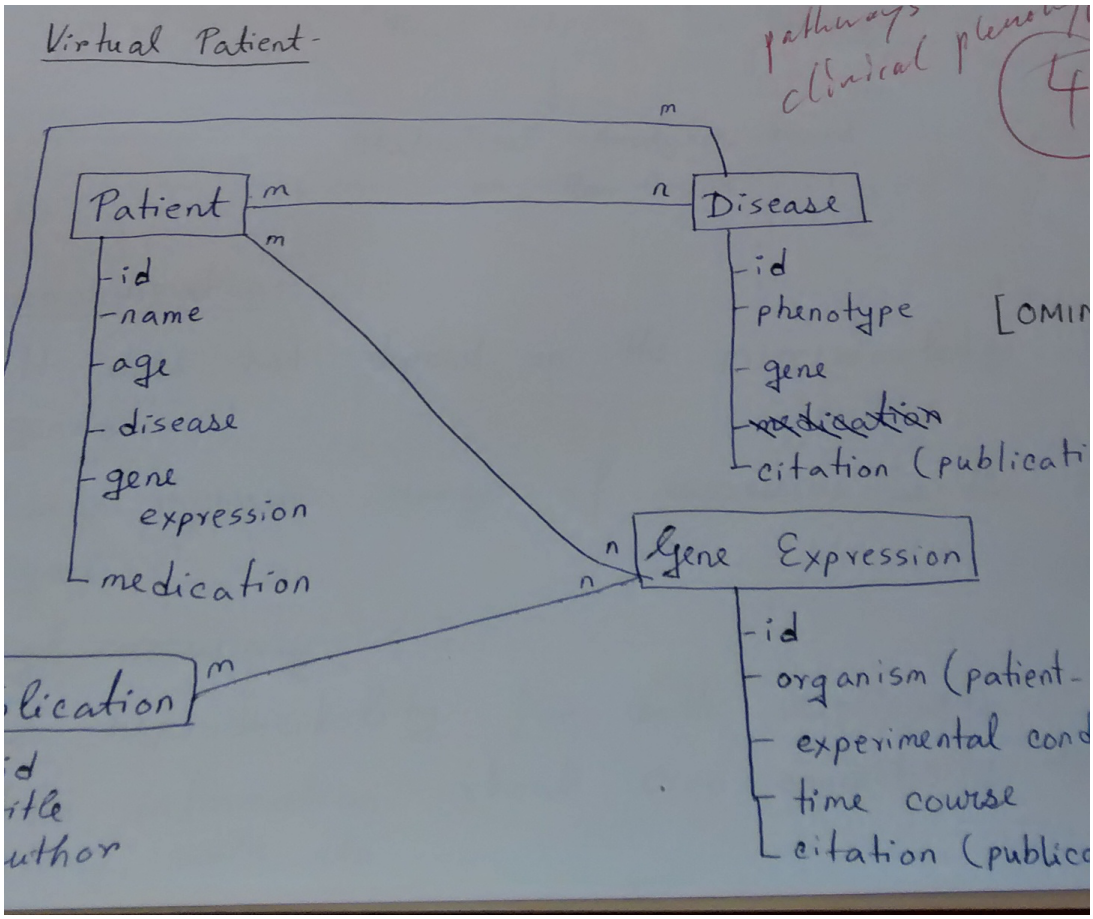
a)

- less data redundancy
- less data inconsistency
- security
- atomicity

b)

- user friendly
- easy retrieval
- easy to understand
- concurrent operation by multiple users

1.36 Develop a conceptual model of a database that stores information about a "virtual patient". Think first! What elemnts does a "virtual patient" have? What sort of information do you need to represent in the conceptual model to be able to use that model for the purpose of "personalized medicine"? exam 2014 not full amount of points!



1.37 You are doing your Master Thesis with some clinical researcher at Venusberg. The medical researchers expect you to help them with in silico methods to identify moleculat determinants of Alzheimer disease. What databases will you access and mine to support them and what sort of information do you get from these databases?

Exam 2014 not full amount of points

1) OMIM – phenotype and responsible gene information for Alzheimer.

2) Nucleotide database (DDBJ, EMBL) – gene related information using the gene from OMIM

3) ENSEMBL – genome information including variants of the gene from OMIM.

4) UniProt KB – protein information endcoded by the using OMIM gene.

5) PDB – 3D structure of the protein using UniProt KB accession number

6) InterPro – protein family, domain, classification information using UniProt KB or PDB Protein.

7) GEO – gene expression data using EMBL gene ID

8) GO – description of the gene and protein in terms of molecular function, biological process and cellular component

9) MEDLINE/ PubMed – related citation and reference using the PubMed ID present in all the databases.

10) SNPs??????

1.38 Regular expression for MIA (or MTA???) EXAM

- MIA molecular information agent?

1.39 Form tables with the given names and attributes EXAM

1.40 Difference among the metabolomics, genomics, proteomics, bioinformatics and systems biology EXAM

1.40 Triples??? EXAM

Triples are subject predicate object expressions. Predicate help in establishing the relationships and thus in extraction of knowledge. In complex knowledge, triples are dependent on each other which makes the relationships more complex because they form a network like structure. In this case we can travel through the graph to extract knowledge.

1.41 How is database integrated with other databases? EXAM

1.42 Build a model database for a given disease from scratch EXAM

2. Literature databases, controlled vocabularies and gene catalogues

2.1. In the description file for the TAXONOMY database, the usage of taxonomy entries in other data bases is mentioned. Which types of other databases refer to TAXONOMY entries?

- The taxonomy database of the International Sequence Database Collaboration contains the names of all organisms that are represented in the sequence databases with at least one nucleotide or protein sequence.
- Many of the Entrez databases (Nucleotide, Protein, Genome, etc.) include an Organism field, [orgn], that indexes entries in that database by taxonomic group. All of the names associated with a taxon (scientific name, synonyms, common names, and so on) are indexed in the Organism field and will retrieve the same set of entries. The Organism field will retrieve all of the entries below the term and any of their children. (http://www.ncbi.nlm.nih.gov/books/NBK21100/ ; Taxonomy Fields in Other Entrez Databases)

  Many of the Entrez databases (Nucleotide, Protein, Genome, etc.) include an Organism field [orgn] that indexes entries in that database by taxonomic group. All of the names associated with a taxon (scientific name, synonyms, common names, and so on) are indexed in the Organism field and will retrieve the same set of entries. The Organism field will retrieve all of the entries below the term and any of their children.

2.2. What is a catalogue?

A catalogue is a systematically arranged list of entities.

Catalogue is an organized, detailed, descriptive list of items arranged systematically.

A catalogue is a complete enumeration of items arranged systematically with descriptive details (or) A catalogue is a collection of entities. A catalogue describes data set attributes and indicates the volumes on which a data set is located.

The database catalog of a database instance consists of metadata in which definitions of database objects such as base tables, views (virtual tables), synonyms, value ranges, indexes, users, and user groups are stored. In computing, a catalog is a directory of information about data sets, files, or a database. A catalog usually describes where a data set, file or database entity is located and may also include other information, such as the type of device on which each data set or file is stored.

2.3. What is an index? How does SRS "index" over several databases?

An index is a set of systematically organized pointers that are used to accelerate access to a set of data.

(OR) A listing of the number, type, label and sequence of all the genes identified within the genome of a given organism. Gene indices are usually created by assembling overlapping EST sequences into clusters, and then determining if each cluster corresponds to a unique gene

(OR) Interface for indexing (multiple) EMBL/Swissprot.dat files (i.e. flat file EMBL/Swissprot format).

(OR) An index is a collection of entities with a full set of entrance points to the entities.

(OR) An index is a set of systematically organized pointers that are used to accelerate access to a set of data. Indexing flat-files (SRS is based on):

- Allows to integrate high number of heterogeneous databases

- Integration system uses a script (specific to each database) to index flat-files from a database. Script is also responsible for discrimination data types and generating links to other relevant databases

- Users search several databases via their indexes

- Maintenance of integrated database schema is not required

- Adding/removal of any number of databases is easy

SRS indexing process: SRS is updated daily; it uses an update mechanism whereby external and local ftp sites are checked for new data files on a daily basis. In this way the system always provides the most up to date data that is available. The system can index plain text, html and xml formatted data files. These data files are broken down by a parser into entries and subsequently into fields. These field indices can then be used for data retrieval or for generating searchable links between different database entries. SRS indexes database records using a word by word approach. Queries can be broadened or refined by using any of the logical operators – 'and', 'or' and 'but not'.

2.4. What is a hierarchy? Which relationship-type is used in hierarchies?

Hierarchy is an organization of entities, where each element (except the top one) has one parent. Every child element has the features of the parent element.

(OR) It's an organizational structural representation of a domain where each level represents a certain shared feature which descends from a parent (node/level).

(OR) A hierarchy is a categorization of a controlled vocabulary where each term of the vocabulary is related to one parent term.

(OR) In hierarchy, the lower level object "is-a" member of higher class, like: human is a primate (Is-a).

2.5. What is a taxonomy? Give a brief definition of a taxonomy

A taxonomy is a collection of controlled vocabulary terms organized into a hierarchical structure (or) taxonomy is a nomenclature of a certain field of knowledge. Each term in a taxonomy is in one or more parent-child relationships to other terms in the taxonomy.

There may be different types of parent-child relationships in a taxonomy (e.g., whole- part, genus-species, type instance), but good practice limits all parent-child relationships to a single parent to be of the same type. Some taxonomies allow poly- hierarchy, which means that a term can have multiple parents. This means that if a term appears in multiple places in a taxonomy, then it is the same term. Specifically, if a term has children in one place in a taxonomy, then it has the same children in every other place where it appears.

2.6. What is an ontology? What are the essential features of an ontology that distinguishes it from a taxonomy? EXAM

Ontology is a human and machine readable formalized representation of a certain domain that mostly has a DAG structure. A formal ontology is a controlled vocabulary expressed in an ontology representation language which allows for modeling objects creation and hence domain models realization by employing a set of grammar rules (constraints) which could be strict or lax.

A taxonomy only relationships of is-a are allowed in a hierarchical way. Ontologies are more general and also includes the relationship part of, and interactions not necessarily in a hierarchical way. Unlike Ontologies, Taxonomies do not contain explicit grammar rules to constrain how to use controlled vocabulary term to express something meaningful within a domain of interest (formal semantics). In comparison to taxonomies, ontologies imply a broader scope of information.

OR

Both terms refer to the organisation of a controlled vocabulary. In a taxonomy the term related through sample “PS a” relationship with. Ontology has a grammar that allows for higher ----- between terms. It cant be modelled as a tree (normally) as a directed acylic graph (DAG) and has multiple ----

2.7. Which of the above mentioned controlled vocabularies has a tree structure?

Taxonomy which is a collection of controlled vocabulary terms organized into a hierarchical structure.

According to the lecture Ontology and consequently a DAG do not exhibit a tree structure!

2.8. What is a directed acyclic graph (DAG) and which type of knowledge representation is based on such a DAG?

It’s a directed graph where one cannot traverse back to a node V having already visited it (due to the lack of cycles). Ontologies. Taxonomies are directed trees, which are also a special case of DAG. Directed acyclic graph: It is the directed version of a cycle graph, with all edges oriented in the same direction, which means there is a route from node A to node B and no way back.

The gene ontology comprises of 3 separate ontologies to describe the attributes of gene products. They are molecular function, biological process and the cellular component. These ontologies are structure vocabularies, which are in the form of direct acyclic graph that represent a network in which each term may be a child of one or more than one parent.

2.9. 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 page” mentioned in the PubMed documentation.

PubMed provides access to bibliographic information that includes Medline or otherwise it is the search engine of Medline. It provides access to citations from biomedical literature, other entrez molecular biology resources. Out- of –scope (eg articles on plate tectonics or astrophysics) from certain Medline journals, primarily general science and chemical journals for which life sciences articles are indexed for Medline.

Citations that precede the date that a journal was selected for Medline indexing. Some additional life science journals that submit full text to PubMed central and receives a quantitative review by NLM.

PubMed search results are displayed in a summary format, see the anatomy of search results page below. Citations are initially displayed 20 items per page with the most recently entered citations displayed first. You can mouse over a journal's title abbreviation to display the full journal name.

Anatomy of the Summary Results:

2.10. What is the difference between PubMed and MEDLINE? Explain in brief!

MEDLINE is a bibliographic database containing citations and abstracts of bioscience articles. PubMed is a service under NCBI Entrez search and retrieval system. PubMed provides access to bibliographic information that includes MEDLINE and some other resources (PubMedCentral and articles from journals before MEDLINE-inclusion and out-of-scope articles). It also provides links to free full-text articiles (if available).

MEDLINE is the largest component of PubMed. A distinctive feature of MEDLINE is that the records are indexed with NLM's (National Library of Medicine) controlled vocabulary, the Medical Subject Headings (MeSH).

PubMed includes:

- In-process citations (before indexed in MEDLINE)
- Citations that precede the date a journal was selected for MEDLINE indexing
- Old MEDLINE entries
- Citations to out-of-scope articles
- Some life science and physics journals

PubMed comprises over 21 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of biomedicine and health, covering portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering. PubMed also provides access to additional relevant web sites and links to the other NCBI molecular biology resources. PubMed is a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH).

PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Compiled by the United States National Library of Medicine (NLM), MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system.

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

2.11. What is PubMed Central? What does it contain and how does it differ from PubMed? EXAM

PubMed Central® (PMC) is a free archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). PMC serves as a digital counterpart to NLM’s extensive print journal collection.

- provide permanent access to all of its content
- the value of PMC lies in its capacity to store and cross-reference data from diverse sources using a common format within a single repository

PubMed Central is a free digital database of full-text scientific literature in biomedical and life sciences. PubMed Central provides openly available peer-reviewed scientific research. PubMed Central does not include any unreviewed research articles. The full text of all PubMed Central articles is available free, but use of the material still is subject to the copyright and/or related license terms of the respective authors or publishers. PubMed provides access to citations from biomedical literature. PubMed does not display the full text of articles.

OR

PubMed is a database of abstract of scientific journals. Its content cames from MEDLINE but also from PubMed central. It is organized through MeSH terms and related articles. PubMed Central specialises in full text. PubMed with open access licences.

2.12. Name at least three searchable types (categories) of information that are contained in MEDLINE abstracts.

- PMID: PubMed Unique Identifier
- AB: Abstract
- AU: Author
- Title
- Journal Issue
- Text words

2.13. What are MeSH terms and what is their purpose? EXAM

Medical Subject Headings - The National Library of Medicine's (NLM's) controlled vocabulary thesaurus which has a set of naming descriptors in a hierarchical structure that enables annotation and further retrieval of information with various specificity.

MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Engines designed to search MEDLINE (such as Entrez and PubMed) generally use a Boolean expression combining MeSH terms, words in abstract and title of the article, author names, date of publication, etc. Entrez and PubMed can also find articles similar to a given one based on a mathematical scoring system that takes into account the similarity of word content of the abstracts and titles of two articles.

(OR) MeSH is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE / PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts.

2.14. Give a short summary of the structure of MeSH and how can a search result be "expanded" (I refer to the PubMed help, where "expansion of search results" is a separate point).

One possibility: When PubMed searches a MeSH term, it will automatically include narrower terms in the search, if applicable. This is also called "automatic explosion." Some terms occur in more than one place in the hierarchy. For example, "Eye" appears under the Anatomy branch, but also under the Sense Organs branch. Automatic explosion will include narrower terms from

all instances of the term in the hierarchy. automatic explosion (explode) - In PubMed, MeSH (Medical Subject Headings) terms (as well as any subheading that is the top of a "subheading tree") are "exploded" automatically to retrieve citations that carry the specified MeSH heading (or subheading) and also retrieve citations that carry any of the more specific MeSH headings (or subheadings) indented beneath it in the Tree structure.

One other possibility: Based on the fact the MeSH implys a hierarchical tree structure following a naive headline search one can further browse the specific desired sub nodes utilizing the MeSH browser which has tree expansion functionality.

2.15. Explain "information retrieval" with an example involving Medline and MeSH terms.

Information retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web.

Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within databases. We search for Alzheimer at Medline; we get many entries about the citations and abstracts from journals according to this disease. In Mesh we get several terms relating this disease so Mesh represents a glossary with terms and definitions with respect to the certain disease. Mesh also provides links to Medline. Therefore, Mesh terms are used for indexing of Medline entries. The indexing procedure is base on the identification of identical terms in abstracts and assigning matches of concepts from Mesh to an abstract.

MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Engines designed to search MEDLINE (such as Entrez and PubMed) generally use a Boolean expression combining MeSH terms, words in abstract and title of the article, author names, date of publication, etc. Entrez and PubMed can also find articles similar to a given one based on a mathematical scoring system that takes into account the similarity of word content of the abstracts and titles of two articles.

MeSH terms are used for indexing Medline entries. The more MeSH terms are there in an abstract, the higher is the relevance of the term to the entry in Medline. Thus, if we perform a query in the MEDLINE database using a MeSH term, the results will involve all the entries (abstracts) that are associated with this MeSH term.

EXAMPLE: Heart Attack Aspirin Prevention is translated as ("myocardial infarction"[MeSH Terms] OR ("myocardial"[All Fields] AND "infarction"[All Fields]) OR

"myocardial infarction"[All Fields] OR ("heart"[All Fields] AND "attack"[All Fields]) OR "heart attack"[All Fields]) **AND** ("aspirin"[MeSH Terms] OR "aspirin"[All Fields]) **AND** ("prevention and control"[Subheading] OR ("prevention"[All Fields] AND "control"[All Fields]) OR "prevention and control"[All Fields] OR "prevention"[All Fields])

2.16. When do we speak of synonyms and when do we speak of homonyms?

Synonyms

Two words that can be interchanged in a context as they have the same semantics are said to be synonymous relative to that context only one is usually deemed “preferred” in queries. Synonyms are different words with identical meaning.

If multiple terms are used to mean the same thing, one of the terms is identified as the preferred term in the controlled vocabulary and the other terms are listed as synonyms.

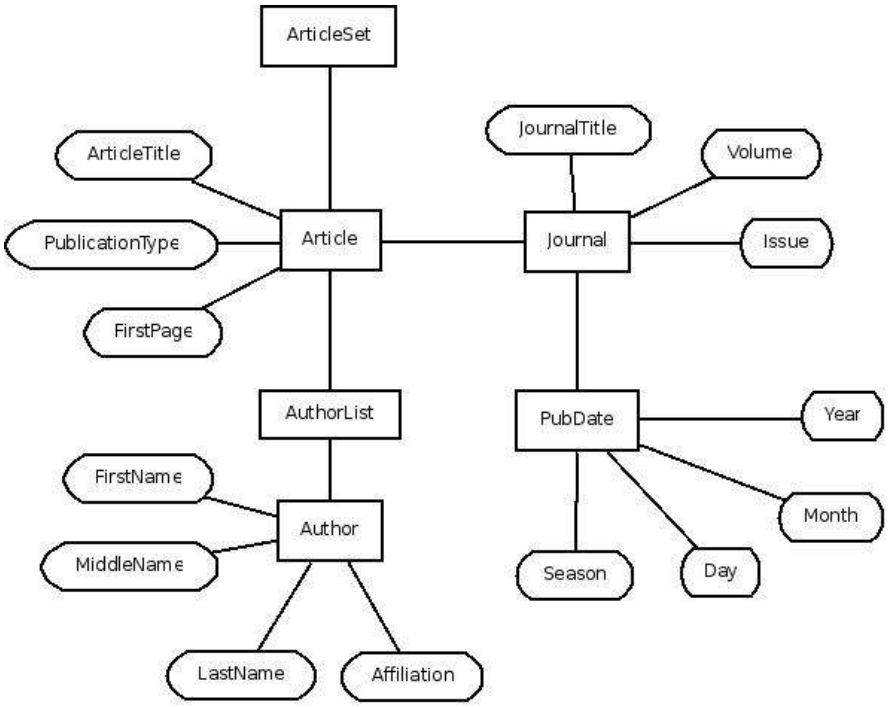
Homonyms

Two words are homonyms if they are pronounced or spelled the same way but have different meanings.

If the same term is commonly used to mean different concepts in different contexts, then this term is a homonym. Homonyms are identical words with different meaning.

2.17. Sketch the major concepts and the conceptual schema of MedLine.

Medline is the NLM’s premier bibliographic database covering the fields of nursing, dentistry, veterinary medicine, the health care system and the pre-clinical sciences. Medline is a virtual databank with two member databanks MEDLINE (Updates) which contains both updates to the data in the main Medline release and new entries since this release and MEDLINE (release).



Unlike using a typical internet search engine, PubMed searching of MEDLINE requires a little investment of time. Using the MeSH database to define the subject of interest is one of the most useful ways to improve the quality of a search. Using MeSH terms in conjunction with limits (such as publication date or publication type), qualifiers (such as adverse effects or prevention and control), and text-word searching is another. Finding one article on the subject and clicking on the "Related Articles" link to get a collection of similarly classified articles can expand a search that yields few results.

Standard searches:

Simple searches on PubMed can be carried out by entering key aspects of a subject into PubMed's search window. PubMed translates this initial search formulation and automatically adds field names, relevant MeSH terms, synonyms, Boolean operators, and 'nests' the resulting terms appropriately, enhancing the search formulation significantly, in particular by routinely combining (using the OR operator) textwords and MeSH terms.

Comprehensive searches:

For comprehensive, optimal searches in PubMed, it is necessary to have a thorough understanding of its core component, MEDLINE, and especially of the MeSH (Medical Subject Headings) controlled vocabulary used to index MEDLINE articles. They may also require complex search strategies, use of field names (tags), proper use of limits and other features, and are best carried out by PubMed search specialists or librarians, who are able to select the right type of search and carefully adjust it for recall and precision. It has been suggested, though, that

even complex literature reviews can be carried out using simple focused search formulations.

Clinical queries/systematic reviews:

A special feature of PubMed is its "Clinical Queries" section, where "Clinical Categories", "Systematic Reviews", and "Medical Genetics" subjects can be searched, with study-type 'filters' automatically applied to identify substantial, robust studies. As these 'clinical queries' can generate small sets of robust studies with considerable precision, it has been suggested that this PubMed section can be used as a 'point-of-care' resource.

Related articles:

A reference which is judged particularly relevant can be marked and "related articles" can be identified. If relevant, several studies can be selected and related articles to all of them can be generated (on PubMed or any of the other NCBI Entrez databases) using the 'Find related data' option. The related articles are then listed in order of "relatedness". To create these lists of related articles, PubMed compares words from the title and abstract of each citation, as well as the MeSH headings assigned, using a powerful word-weighted algorithm.[14] The 'related articles' function has been judged to be so precise that some researchers suggest it can be used instead of a full search.

Mapping to MeSH headings and subheadings:

A strong feature of PubMed is its ability to automatically link to MeSH terms and subheadings. Examples would be: "bad breath" links to (and includes in the search) "halitosis", "heart attack" to "myocardial infarction", "breast cancer" to "breast neoplasms". Where appropriate, these MeSH terms are automatically "expanded", that is, include more specific terms. Terms like "nursing" are automatically linked to "Nursing [MeSH]" or "Nursing [Subheading]". This important feature makes PubMed searches automatically more sensitive and avoids false-negative (missed) hits by compensating for the diversity of medical terminology.

2.18. Explain the differences between OMIM and MedLine.

An entry in OMIM (Online Mendelian Inheritance in Man) is a review focusing on a disease, its phenotypic appearance and the genes involved in the molecular etiology of the disease while Medline is the NLM bibliographic database covering medicine, nursing etc. and it mostly contains abstracts and citation over full articles.

Both OMIM and Medline are literature-based databases. OMIM (Online Mendelian Inheritance in Man) database is a catalogue of human genes and genetic disorders. The database contains

textual information, pictures and reference information. Whereas Medline is NLM's (National Library of Medicine) premier bibliographic database covering the fields of nursing, medicine, dentistry, veterinary.

2.19. What are the three root concepts of GO? EXAM

GO comprises three largely independent sub-ontologies, which describe the following aspects of a gene (or better: the protein expressed from a gene): molecular function, biological process, cellular localization.

- ***Cellular Component Ontology Guidelines***: The cellular component ontology describes locations, at the levels of subcellular structures and macromolecular complexes. The cellular component ontology includes multi-subunit enzymes and other protein complexes, but not individual proteins or nucleic acids. Cellular component also does not include multicellular anatomical terms.
  http://www.geneontology.org/GO.component.guidelines.shtml
- ***Molecular Function Ontology Guidelines***: The functions of a gene product are the jobs that it does or the "abilities" that it has. These may include transporting things around, binding to things, holding things together and changing one thing into another. This is different from the biological processes the gene product is involved in, which involve more than one activity. http://www.geneontology.org/GO.function.guidelines.shtml
- ***Biological Process Ontology Guidelines***: A biological process is a recognized series of events or molecular functions. A biological process is not equivalent to a pathway although some GO terms do describe pathways. (Molecular function, biological process and cellular function).

OR

GO is better represented as three trees. Each representing GO root concept. Molecular function, bio process and cellular localization.
GO helps in the process of exchanging information between functional annotation groups for example a UniProt KB. It helps to find homologous that can have very different names but! GO terms should be the same in very similar.

2.20. What means "annotation"? EXAM

The addition of descriptive information about the function or structure of a molecular sequence to its MOLECULAR SEQUENCE DATA record.

A combination of comments, notations, references, and citations, either in free format or

utilizing a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools.

2.21. Which controlled vocabularies do you know besides GO?

MeSH; IUPAC; HGNC, sequence Ontology, Brenda enzyme source ontology, EMAP, SwissProt keywords.

Trait Ontology (TO): It is a controlled vocabulary to describe each trait as a distinguishable feature, characteristic, quality or phenotypic feature of a developing or mature individual.

Plant Ontology (PO): Gramene is collaborating with The Plant Ontology Consortium (POC) to develop a controlled vocabulary for plant structure (anatomy) and growth stages

Environment Ontology (EO): It represents a controlled vocabulary to describe different types of supplemental environments that have been reported in the experimental profiles of gene expression and phenotype (mutant and QTL) studies on cereal plants. Taxonomy, Thesaurus, Data model, Data dictionary, EC (Enzyme nomenclature) .

2.22 How can a search result be "expanded"? (I refer to the pubmed help, where expansion of search result is a separate point)

2.23 What is meta-information and how is meta-information used?

2.24 What is "Karyn ´s Genomes" and what sort of data do we find there? Are Patent Abstracts part of the scientific literature?

2.26 What is the content of OMIM? Sketch a simple entity model of OMIM comprising the major entity types represented in OMIM! EXAM

Disease name

Molecular information		Medical information	Bibliographic data
Gene name Gene location Cause of disease phenotype Type of hereditary	Link outs into other DB	Ethology Description of phenotype Treatment	Articles where the data was found

Protein function		Role of morbidity and mortality	

2.27 Draw an acyclic directed graph exam 2008

2.28 Describe the process of indexing of MEDLINE entries exam 2008

2.29 What is the "gene ontology" and how does it mediate interoperability of genome annotations? Exam 2008

Genome annotation is the practice of capturing data about a gene product, and GO annotations use terms from the GO ontology to do so. The members of the GO Consortium submit their annotation for integration and dissemination on the GO website, where they can be downloaded directly or viewed online using AmiGO. In addition to the gene product identifier and the relevant GO term, GO annotations have the following data:

- The *reference* used to make the annotation (e.g. a journal article)
- An *evidence code* denoting the type of evidence upon which the annotation is based
- The date and the creator of the annotation

2.30 What is the role of MeSH terms and how daoes MeSH contribute to the computation of "relatedness" of MEDLINE abstracts?

MeSH (Medical subject headings) terms are used in the MEDLINE bibliographical database for indexing articles. They are controlled vocabulary terms containing a set of naming discriptors. These discriptors or heading can retrieve and annotate at various levels of specificity.
The number of occurance of a MeSH term in the MEDLINE abstract provides the MeSH term count. This count is matched in the MeSH database and used to compute the relatedness of abstracts.

6.10Difference between ontology, taxonomy, thesaurus? EXAM

6.11

2.32 **Benefits of gene dictionaries? EXAM**

Unification and standardization of name space for genes.

6.12What is "prove" name? not sure about the question EXAM

It is history of origin of data

3. Nucleotide Databases

3.1. Name three of the most important / most informative entity-types that can be found in EMBL or EntrezGene.

In EMBL each entry corresponds to a single contiguous sequence as contributed to the database or reported in the literature. In some cases, entries have been assembled from several papers reporting overlapping sequence regions. Conversely a single paper often provides data for several entries, as when homologous sequences from different organisms are compared.

. EMBL Nucleotide Sequence Database
. ENSEMBL
. TrEMBL

(OR) Entity-types:

- organism
- molecule
- sequence

3.2. Which objects in biology correspond to these entity-types?

. EMBL: nucleotide sequence
. ENSEMBL: large eukaryotic genome
. TrEMBL: translations of all coding sequences (CDS) present in the EMBL. Nucleotide Sequence (Proteins)

Bio-objects: DNA, RNA, Protein, Lipid, Sugar

Organism: animals, plants, fungi, bacteria, protozoa.

Molecule: DNA, RNA.

Sequence: nucleic acid sequence, i.e. adenine, thymine, guanine, cytosine in case of DNA; adenine, uracil, guanine, cytosine in case of RNA.

3.3. Define a gene and name three attributes

A gene is a molecular unit of heredity of a living organism. It is a name given to some stretches of DNA and RNA that code for a polypeptide or for an RNA chain that has a function in the organism.

(OR) A modern working definition of a gene is "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions ".

(OR) A unit of DNA which performs one function. Usually, this is equated with the production of one RNA or one protein. A gene contains coding regions, introns, untranslated regions and control regions. Attributes: Promotor, sequence, length.

Attributes: Sequence, Organism, Locus in Chromosome, Gene name, Molecular weight, Coding regions, 5' and 3'.

Attributes for example:

Function: hyaluronic acid binding

Process: cell adhesion

Cell location: plasma membrane

It consists: Long strand of DNA (RNA in some viruses), Promoter, Coding sequence.

(OR) It's a unit of heredity and carries inherited information:
a) Gene name
b) Molecular weight
c) 5' and 3'end.

3.4. Who is the owner of an entry in EMBL?

The owner of a sequence submitted to EMBL is the submitter. The submitter has the authority of changing and annotating sequences.

3.5. Name the most important data types / entity types that you have to provide with an entry in EMBL. EXAM

The following fields are mandatory:

- Sequence description - use the examples provided as a guide
- Sequence length - this allows us to ensure that you have provided the whole sequence
- Sequenced molecule - the type of molecule sequenced
- Sequence - must conform to the IUPAC standard. (FASTA or raw sequence formats).

Nucleotide Sequence: The ID line is always the first line of an entry. The general form of the ID

line is: Term ID, entry name, data class, molecule, division, sequence length (Base Pairs). Source Organism.

- Submitter Information
- Release Date Information
- Sequence Data, Description and Source Information
- Reference Citation Information
- Feature Information (e.g. coding regions, regulatory signals etc.)

3.6. What means "vector clipping"?

"Vector clipping" means separating the DNA segment of interest from the vectors DNA, when the DNA of interest is contaminated with the vectors DNA. EBI provides a vector screening service using BLAST algorithm for "vector clipping" procedure.

Vector-clipping: Vector sequences can skew clustering even if a small vector fragment remains in each read. Delete 5' and 3'regions corresponding to the vector used for cloning. Detection of vector sequences is not a trivial task, because they normally lies in the low quality region of the sequence. UniVec is a non-redundant vector database available from NCBI: Use the cross match alignment program to compare each read in file (clone origin fasta) to a fasta database of cloning and sequencing vectors

Vector clipping against a standard database or custom sequences. Clip parts of sequences that does not represent the insert, such as sequencing vectors and adaptors. VecScreen is a system for quickly identifying segments of a nucleic acid sequence that may be of vector origin (but does not remove or mask those vectors). NCBI developed VecScreen to combat the problem of vector contamination in public sequence databases and a web page is designed to help researchers identify and remove any segments of vector origin before sequence analysis or submission.

Is a method for cleaning a sequence from vector contamination? For doing a vector clipping we:

- Analyze each of your sequences for vector contamination using VecScreen at NCBI.

- Remove the vector contaminations from the sequences VecScreen is a system for quickly identifying segments of a nucleic acid sequence that may be of vector origin. NCBI developed VecScreen to combat the problem of vector contamination in public sequence databases. UniVec is a non-redundant vector database available from NCBI: Use the cross match alignment program to compare each

read in file (clone origin fasta) to a fasta database of cloning and sequencing vectors.

3.7. How come that so many genes have more than one name?

Gene names are not standardizing across species. A gene can be present in 2 different animals which produces a protein in both animals with the same function but have different names. Some gene names are based on known functions and genes have several functions, as they are expressed at different times (post translational modification).

There has not been well established and curated gene naming system for a long time. This may have lead to situations where the same gene might have been discovered by different people and thus different names have been assigned. Also a gene may get more than one name in situation where firstly partial gene fragments are discovered and are taken for the whole gene. In this case, after finding out that the partial sequences belong to one gene, several different names of one gene occur.

- Synonyms are a curation issue. Alternative names are generated from a variety of situations: some are different labs stumbling on to the same gene; some are simple differences between characters (MEN1, MEN 1, MEN-1). Sometimes large-scale projects generated some form of coded designation (such as 1200015E08Rik, a mouse example). Alternate names are collected by curation teams at various places. Some come from papers read by curators. Some would have come from computational sources. Most of them are shared around, but there may be some sources that have some name that others don't.

- There are nomenclature standards that are supposed to be used, and there are committees that establish the rules and assign official names and symbols. However, some people refuse to use those names in publications and revert to their preferred name. The official name/symbols shouldn't have duplicates.

- Gene names are not standardized across species. A gene can be present in 2 different animals which produces a protein in both animals with the same function but have different names.

- Some gene names are based on known functions and genes have several functions, as they are expressed at different times (post translational modification).

3.8. Explain the difference between a data repository and a curated database

Data repository is a not curated place to store data. The responsibility for accuracy of the data in a repository lies on the submitter, e.g. EMBL. In a curated database team of specialists check the incoming entries to avoid ambiguities and redundancy, e.g. SwissProt.

(OR) Data Repository is a logical (and sometimes physical) partitioning of data where multiple databases which apply to specific applications or sets of applications reside. DataRepository is a database acting as an information storage facility. Curated databases are databases that are populated and updated with a great deal of human effort. Curated Database is an annotated database created under the supervision of a curator, who makes judgments as data are cleaned up and merged.

(OR) Data Repository is a database acting as an information storage facility. Although often used synonymously with data warehouse, a repository does not have the analysis or querying functionality of a warehouse. Curated databases are databases that are populated and updated with a great deal of human effort. Curated Database is an annotated database created under the supervision of a curator, who makes judgments as data are cleaned up and merged.

3.9. What is a knowledge base?

It is a highly curated database having as much annotations as possible. The data in the knowledge base is believed to be highly accurate and comprising the best current understanding of the matter.

(OR) A knowledge base (abbreviated KB, kb) is a special kind of database for knowledge management. A knowledge base provides a means for information to be collected, organized, shared, searched and utilized. Has accuracy and is not redundant. Machine-readable knowledge bases store knowledge in a computer-readable form, usually for the purpose of having automated deductive reasoning applied to them.

3.10. What are mRNA, hnRNA, cDNA, rRNA and tRNA and how are they represented in EMBL?

- *mRNA* (messenger RNA) is a copy of the information carried by a gene on the DNA. The role of mRNA is to move the information contained in DNA to the translation machinery (ribosomes).
- *hnRNA* is a precursor RNA, i.e. an RNA transcript before it is processed into mRNA, rRNA, tRNA, or other cellular RNA species, any RNA species that is not yet the mature RNA product.

- *cDNA* (complementary DNA) is a piece of DNA copied from a mature mRNA. o *rRNA* is ribosomal RNA. It is a component of the ribosomes, the protein synthetic factories in the cell.
- *tRNA* is a transfer RNA. It transfers an amino acid to the ribosome, so that the amino acid would be added to a polypeptide chain.

In EMBL under SRS interface there is a field molecule, which comprises these values.

mRNA: Messenger RNA is a molecule of RNA that encodes a chemical "blueprint" for a protein product. mRNA is transcribed from a DNA template, and carries coding information to the sites of protein synthesis: the ribosomes.

hnRNA: Precursor mRNA (pre-mRNA) is an immature single strand of messenger ribonucleic acid (mRNA). pre-mRNA is synthesized from a DNA template in the cell nucleus by transcription. Pre-mRNA comprises the bulk of heterogeneous nuclear RNA (hnRNA). The term hnRNA is often used as a synonym for pre-mRNA, although, in the strict sense, hnRNA may include nuclear RNA transcripts that do not end up as cytoplasmic mRNA.

cDNA: In genetics, complementary DNA (cDNA) is DNA synthesized from a messenger RNA (mRNA) template in a reaction catalyzed by the enzyme reverse transcriptase and the enzyme DNA polymerase. cDNA is often used to clone eukaryotic genes in prokaryotes. When scientists want to express a specific protein in a cell that does not normally express that protein (i.e., heterologous expression), they will transfer the cDNA that codes for the protein to the recipient cell. cDNA is also produced by retroviruses which is integrated into its host's genome to create a provirus.

rRNA: Ribosomal ribonucleic acid (rRNA) is the RNA component of the ribosome, the enzyme that is the site of protein synthesis in all living cells. Ribosomal RNA provides a mechanism for decoding mRNA into amino acids and interacts with tRNAs during translation by providing peptidyl transferase activity.

tRNA: Transfer RNA (tRNA) is an adaptor molecule composed of RNA, that is used in biology to bridge the four-letter genetic code (ACGU) in messenger RNA (mRNA) with the twenty-letter code of amino acids in proteins. One end of the tRNA carries the genetic code in a three-nucleotide sequence called the anticodon. The anticodon forms three base pairs with a codon in mRNA during protein biosynthesis. In EMBL they are included under the nucleotide sequence database. Also entities like general information, description, references, cross references, features and sequence are included about the respective query.

3.11. What means “coding sequence” and what is a “non-coding sequence”?

The coding region of a gene, also known as the coding sequence or CDS, is that portion of a gene's DNA or RNA, composed of exons, that codes for protein. The region is bounded nearer the 5' end by a start codon and nearer the 3' end with a stop codon. The coding region in mRNA is bounded by the 5’- UTR and 3’-UTR, which are also parts of the exons. Non-coding sequences: A sequence in a gene that has not been identified as coding for mRNA transcription for protein translation, but is instead responsible for regulatory or other functions.

(OR) That portion of a gene which directly specifies the amino acid sequence of its protein product. Non-coding sequences of genes include control regions, such as promoters, operators and terminators, as well as the intron sequences of certain eukaryotic genes.

(OR) “Coding sequence” is the portion of a gene or an mRNA which actually codes for a protein. Introns are not coding sequences; nor are the 5' or 3' untranslated regions. The coding sequence in a cDNA or mature mRNA includes everything from the ATG (or AUG) initiation codon through to the stop codon, inclusive. “Non-coding sequence” is a sequence that is not translated into protein, e.g. introns, promoters, transcription factor binding sites, all the sites that do not code mRNA.

3.12. How is information on the exon-intron-structure of a gene represented in an EMBL-entry?

In form of annotations, using the FT (Feature Table) lines, which provide a mechanism for the annotation of the sequence data.

Molecule entity which has different values and one of them is other RNA which covers hnRNA and we know that heterogenes RNA contains both intron and exon in RNA structure.

Information about exon-intron structure in EMBL is stored in the “Features” field: “Key” defines intron/exon, “Location” defines the location of intron/exon, e.g. Intron 10..50.

3.13. Give examples for “values” of the entity type “molecule” in EMBL.

Example for the “value” of the entity type “molecule” are linear unassigned DNA, Circular genomic DNA, genomic RNA, mRNA, other DNA, other RNA, pre-RNA, rRNA, snoRNA, snRNA, tRNA, unassigned DNA, unassigned RNA, viral cRNA.

3.14. Who issues an accession number?

Curator if the database is curated, if not, then it is issued automatically. (For EMBL, curator).

(OR) An identifier supplied by the curators of the major biological databases upon submission of a novel entry that uniquely identifies that sequence or other entry.

3.15. What is the difference between an accession number and a database identifier? EXAM

An entry in the database has a unique identifier, however several accession numbers may be related with this entry.

The accession number is assigned to a sequence and should remain linked with it forever, whereas entry names may change; also occasionally a few sequences may be merged to a single sequence that will then inherit all accession numbers. The identifier is a unique identifier for an entry.

Soon after submission of your sequence, you will receive an accession number from the database which you will be able to use in your article to refer to the sequence and it's enough to submit the sequences in one of database because of the data exchange between EMBL,DDBJ and Genebank.

OR

In general, a database identifier is a name that distinguishs each entry (normally based on function). Accession number is a fixed stable alphanumeric code that helps you seek, as they are normally mentioned in the articles.
Identifier can change and be ----- but not accession number!

3.16. Why is EMBL synchronized with DDBJ and NCBI/GenBank?

They are synchronized since submission of the same nucleotide sequence can be done in all of the mentioned databases and the synchronization is a precautionary measure.

The entries in the EMBL, GenBank and DDBJ databases are synchronized on a daily basis, and the accession numbers are managed in a consistent manner between these three centers. Accession number is assigned to a sequence and should remain linked with it forever, whereas entry names may change; also occasionally a few sequences may be merged to a single sequence that will then inherit all accession numbers.

(OR) To make the effective sharing of scientific information possible. To have an access to all the submitted data through the gateways of all three databases. To create a common system, with minimum ambiguity.

3.17. Why is EMBL split in EMBL, EMBL updates and EMBL (whole genome shotgun) at the SRS interface? EXAM

Separating EMBL in 3 databases benefits to the organization and the clarity of information that is stored in it. Having one database that is responsible for the Updated sequence is a good approach in order to know which of the already submitted sequences have been modified (possible errors detected and removed). And having a separation between databases where one is concerned with sequences for part of the genome and one concerned with the whole genome is also a good approach for the querying and retrieval of information.

Secondly, EMBL shotgun sequences are sequences not curated. EMBL and EMBL updates are like this in a sort of versional database.

The EMBL is updated very frequently. In order to maintain fast query engine reindexing must be accomplished when every new entry is submitted. However EMBL comprises many entries reindexing of which takes a lot of time. Thus, the splitting is done in order to decrease time needed for reindexing.

OR

With todays each of sequencing EMBL has huge amount of new sequences and would issue the DB. Would have to ----, EMBL could not function. That is why EMBL Updates was created and its periodical folded bach into EMBL and EMBL WGS was separated from EMBL because of the difference of the data (gene and genome). That way EMBL WGS can show its information in a way more logically to such a big content.

3.18. What is a contig?

Group of clones representing overlapping regions of a genome.

In shotgun DNA sequencing a contig is a set of overlapping DNA segments derived from a single genetic source. A contig in this sense can be used to deduce the original DNA sequence of the source.

(OR) A contig (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA. In bottom-up sequencing projects, a contig refers to overlapping

sequence data (reads); in top-down sequencing projects, contig refers to the overlapping clones that form a physical map of the genome that is used to guide sequencing and assembly. Contigs can thus refer both to overlapping DNA sequence and to overlapping physical segments (fragments) contained in clones depending on the context.

3.19. What means "whole genome assembly"?

Genome assembly refers to the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated.

Whole Genome assembly refers to the process of sequencing a large number of short DNA sequences, all of which were generated by a shotgun sequencing project, and putting them back together to create a representation of the original DNA sequence.

In a shotgun sequencing project, all the DNA from a source (usually a single organism, anything from a bacterium to a mammal) is first fractured into millions of small pieces.

In bioinformatics, sequence assembly refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original sequence. This is needed as DNA sequencing technology cannot read whole genomes in one go, but rather reads small pieces of between 20 and 1000 bases, depending on the technology used. Typically the short fragments, called reads, result from shotgun sequencing genomic DNA, or gene transcript (ESTs). The problem of sequence assembly can be compared to taking many copies of a book, passing them all through a shredder, and piecing the text of the book back together just by looking at the shredded pieces. Besides the obvious difficulty of this task, there are some extra practical issues: the original may have many repeated paragraphs, and some shreds may be modified during shredding to have typos. Excerpts from another book may also be added in, and some shreds may be completely unrecognizable.

3.20. What differentiates EMBL from ENSEMBL?

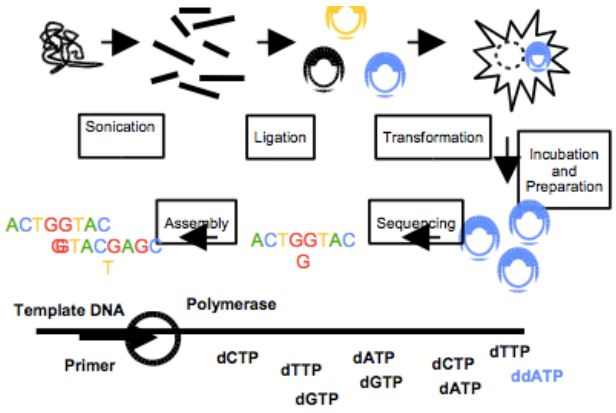
EMBL is sequence (molecule) focused database and ENSEMBL is genome (organism) focused database. The Ensembl project provides automated genome annotation and subsequent visualisation of the annotated genomes.

(OR) EMBL is individual sequence-centric database. ENSEMBL is genome-centric database. Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation of eukaryotic genomes.

3.21. Sketch the process of sequencing and identify possible sources for errors.

Possible sources for errors could be: the fwd and rev primer make a big overlap then the sequence would not be correct (50bp overlap is the allowed limit). Vector (clone) contamination.

Errors: the sample may get contaminated, the primers may not interact efficiently enough, gel (for gel electrophoresis) has non-homogenous areas.



3.22. Name at least three attributes listed under "Features" in a standard EMBL entry.

Molecule type, location, qualifiers for annotations (exons, introns, codons, TATA_signals,...) Source - identifies the biological source of the specified span of the sequence. STS - sequence tagged site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs; Organism scope (for example: eukaryotes and eukaryotic viruses); Molecule scope (DNA,RNA)...

3.23. Sketch the major parts of a typical EMBL entry: what categories does a "normal" EMBL entry have?

- General Information: Primary Accession #, Accession #, Entry ID, MoleculeType etc.
- Description: description, keywords, organism, etc.
- References
- Features
- Sequence

3.24. What are "EST sequences"? And which database comprises information on EST sequences? EXAM

An expressed sequence tag or EST is a short sub-sequence of a cDNA sequence. They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination. GenBank comprises information on EST sequences Because these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. They may be represented in databases as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand.

(OR) EST (expressed sequence tags) are short pieces of cDNA sequence. Tags can be allocated to some certain position (tag markers). ESTs consist of 100-400 base pairs. ESTs are produced via shotgun sequencing. Making many ESTs of one long DNA sequence allows to reconstruct this sequence. UniLib, UniGene comprise EST sequences.

3.24*. What are "EST sequences"? how are they generated, for what purpose and in which DB we find them? EXAM

EST are short sequences based from mRNA. Normally your sequence has a high number of EST from a mRNA extract and with informatics tools you cluster them based on overlapping redundancy to form a contig that should represent the mRNA sequence.
dbEST, UniGene and UniLib, STACK store EST information.

3.25 What is "next generation sequencing" (NGC) and which specific challenges have to be met for the data banking of NGC sequences? EXAM

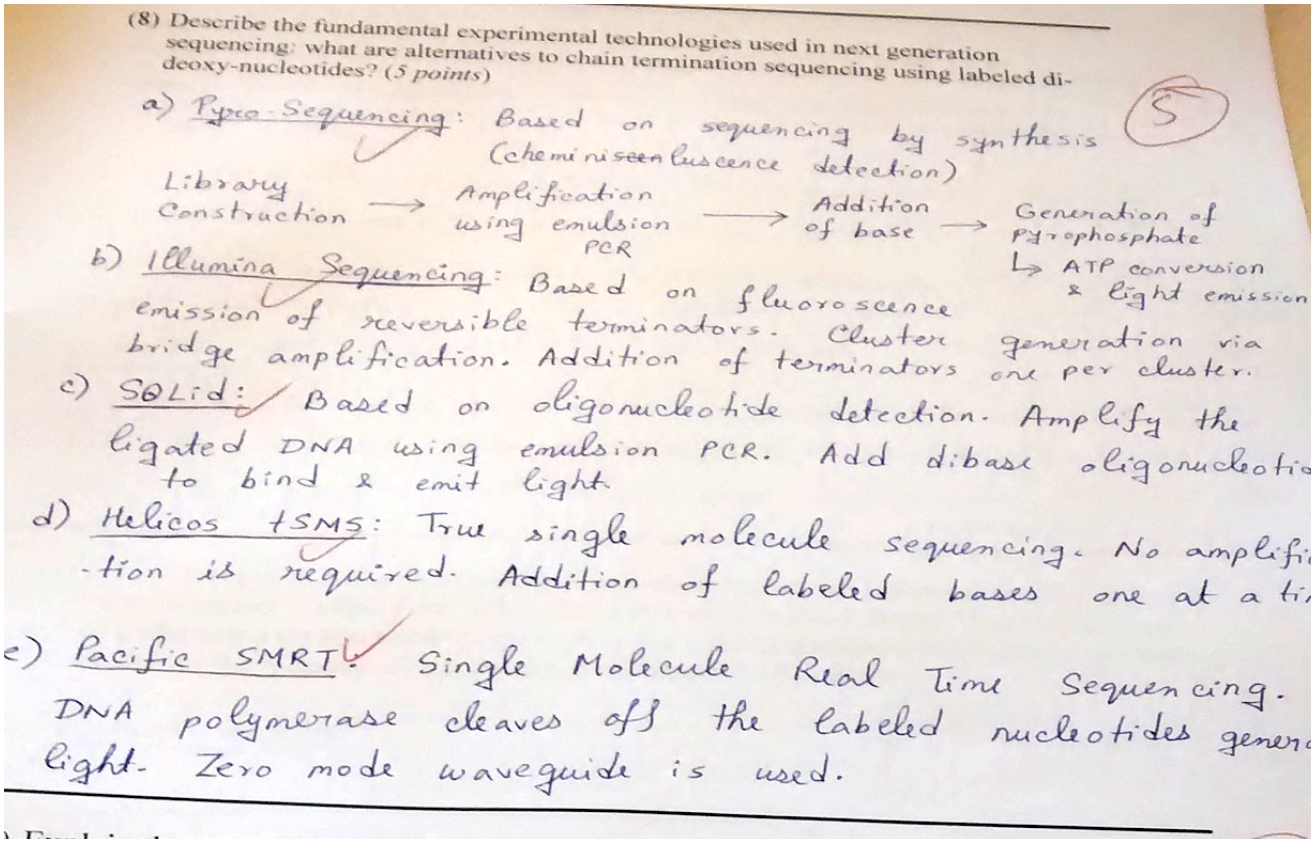
NGS refers to high through technique, that can give hundreds of megabases and even gigabases pairs in short time. Examples: Roches 454, Illumina, SOLID. These techniques generate huge files of small sequences and these need to be used in order to reduce and obtain the sequence.

3.26 What does "assembly" mean when talking about genome sequencing?

3.27 What was the intention to create the RefSeq database?

3.28 How do you submit entire genome sequencing projects to EMBL?

3.29 Describe the fundamental experimental technologies used in next generation sequencing. What are the alternatives to chain termination sequencing using labelled ddNTPs? – exam 2014



(8) Describe the fundamental experimental technologies used in next generation sequencing: what are alternatives to chain termination sequencing using labeled di-deoxy-nucleotides? (*5 points*)

a) Pyro-Sequencing: Based on sequencing by synthesis (chemiluminescence detection)

Library Construction → Amplification using emulsion PCR → Addition of base → Generation of Pyrophosphate ↳ ATP conversion & light emission

b) Illumina Sequencing: Based on fluorescence emission of reversible terminators. Cluster generation via bridge amplification. Addition of terminators one per cluster.

c) SOLiD: Based on oligonucleotide detection. Amplify the ligated DNA using emulsion PCR. Add dibase oligonucleotide to bind & emit light

d) Helicos tSMS: True single molecule sequencing. No amplification is required. Addition of labeled bases one at a time

e) Pacific SMRT: Single Molecule Real Time Sequencing. DNA polymerase cleaves off the labeled nucleotides generating light. Zero mode waveguide is used.

3. 30 How is gene polymorphism-information linked to sequence information in ENSEMBL? Where does SNP- Information come and from and how is it technically linked? – exam 2014

Ensemble is a genome database which provides information about the sequence variants. SNP occurs in the sequence when a single base gets substituted in the sequence. dbSNP is the database where SNP information are stored. Ensemble database has a pointer to dbSNP based on the dbSNP accession number (rs followed by no.). Via this pointer SNP-information comes into Ensemble for a particular gene.

3.31 You are conducting a metagenome sequencing project. Explain, how you submit your sequencing data to the ENA- Database.

We can submit the metagenome sequencing project data in ENA either in form of raw sequencing reads or sequencing reads assembly. User can submit the data via web-based tool WebIn. It will ask the user to fill uo the template. The tool is user-interactive.

While filling up the template user should submit the following information:

1) Sequence details
2) Released date
3) Experimental design
4) Source organism
5) Sample datails
6) Lab and instrumental properties
7) Enviromental conditions

3.32 Flow chart for cloning, sequence alignment EXAM

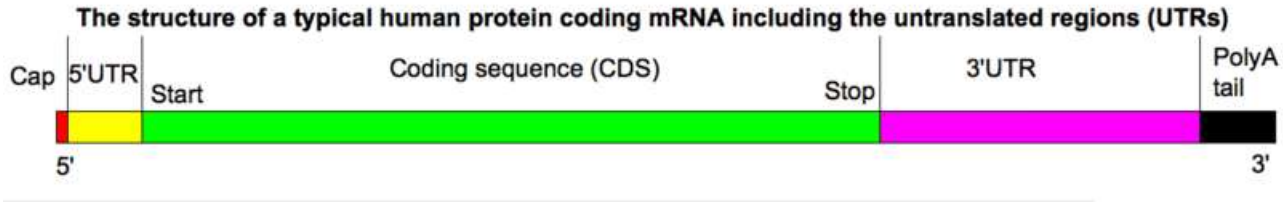
3.33 How to submit data to genome database. EXAM

4. Nucleotide related databases

4.1. What is a 3 ´UTR ?

The three prime untranslated region (3' UTR) is a particular section of messenger RNA (mRNA). An mRNA codes for a protein through translation. The mRNA also contains regions that are not translated. In eukaryotes the 5' untranslated region, 3' untranslated region, cap and polyA tail.

(OR) The untranslated region at the 3'-end of an mRNA, i.e. following the coding region. It contains the polyadenylation signal, as well as binding sites for proteins that affect the mRNA's stability or location in the cell.



(OR) In molecular genetics, an untranslated region (or UTR) refers to either of two sections on each side of a coding sequence on a strand of mRNA. If it is found on the 5' side, it is called the 5' UTR (or leader sequence), or if it is found on the 3' side, it is called the 3' UTR (or trailer sequence).

Features

Key	Location	Qualifier	Value
source	1..255	organism	uncultured eukaryote
		environmental_sample	
		mol_type	mRNA
		country	France:North East France Nievre-Morvan Breuil Chenue forest
		lat_lon	47.18 N 4.4 E
		isolation_source	sandy clay soil from a Picea abies forest stand
		collected_by	Marmeisse R, Debaud JC, Damon C
		collection_date	10-Jul-2007
		clone	AEE0AAA14YJ24CM1
		db_xref	taxon:100272

Various regulatory sequences may lie in the 3'-UTR:

- A polyadenylation signal sequence, which marks the termination of the transcript about 30 base pairs downstream of the signal, followed by a few hundred adenine residues.

- Binding sites for proteins that influence the stability or the transport of the mRNA.

- Binding sites for mi-RNA's.

4.2. What is a transcription factor site and why would you collect information on these sites in a database?

Transcription factors are proteins that interact with DNA and initiate or inhibit the process of transcription upon binding to DNA. A TF site is a region on the DNA to which a TF can bind. It is important to know these sites to understand how (and which) TF's can influence the regulation of specific genes.

The site of the transcription factor that is responsible for binding with the promoter region and initiating transcription process in order to synthesize proteins. We should collect information on these sites in a database because TFSITE provides information on individual regulatory protein binding sites of Eukaryotic genes ranging from Humans to Yeast. Also these sites provide us with the information of artificial sequences resulting from mutagenesis, in-vitro selection procedures from random nucleotide mixtures or from specific theoretical considerations. The consensus binding sequences can also be known by the use of database.

In order for our bodies to have different types of cells, there has to be some mechanism for controlling the expression of our genes. In some cells, certain genes are turned off while in other cells they are *transcribed* and *translated* into proteins. Transcription factors are one of the most common tools that our cells use to control gene expression.

Transcription factors (TFs) are molecules involved in regulating gene expression. They are usually proteins, although they can also consist of short, non-coding RNA. TFs are also usually found working in groups or complexes, forming multiple interactions that allow for varying degrees of control over rates of transcription.

In people (and other eukaryotes), genes are usually in a default "*off*" state, so TFs serve mainly to turn gene expression "*on*". In bacteria, the reverse is often true, and genes are expressed "*constituitively*" until a TF turns it "*off*". TFs work by recognizing certain nucleotide sequences **(motifs)** before or after the gene on the chromosome (up- and downstream).

Eukaryotes often have a **promoter region** upstream from the gene, or **enhancer regions** up or downstream from the gene, with certain specific motifs that are recognized by the various types of TF. The TFs bind, attract other TFs and create a complex that eventually facilitates binding by **RNA polymerase**, thus beginning the process of transcription.

Transcription factors are only one of the means by which our cells express different combinations of genes, allowing for differentiation into the various types of cells, tissues and organs that make up our bodies. However, this mechanism of control is extremely important, especially in light of the findings of

the Human Genome Project; That we actually have less genes in our genome, or on our chromosomes, than originally thought. What this means is different cells have not arisen from differential expression of completely different sets of genes, but are more likely to have varying levels of selective expression of the same groups of genes.

TFs can also control gene expression by creating a "*cascade*" effect, wherein the presence of small amounts of one protein triggers the production of larger amounts of a **second**, which triggers production of *even larger* amounts of a **third**, and so on.

Manipulating TFs to reverse the cell differentiation process is the basis of methods for deriving stem cells from adult tissues.

The site of the transcription factor that is responsible for binding with the promoter region and initiating transcription process in order to synthesize proteins. We are interested in collect this information because transcriptor factors are related to some diseases and play an important role in life development.

Transcription factor sites are modular in structure and contain the following domains:-

DNA-binding domain (DBD), which attach to specific sequences of DNA (enhancer or Promoter: Necessary component for all vectors: used to drive transcription of the vector's transgene promoter sequences) adjacent to regulated genes. DNA sequences that bind transcription factors are often referred to as response elements.

Trans-activating domain(TAD), which contain binding sites for other proteins such as transcription coregulators. These binding sites are frequently referred to as activation functions (AFs).

An optional signal sensing domain (SSD) (e.g., a ligand binding domain), which senses external signals and, in response, transmit these signals to the rest of the transcription complex, resulting in up- or down-regulation of gene expression. Also, the DBD and signal-sensing domains may reside on separate proteins that associate within the transcription complex to regulate gene expression.

4.3. Explain the fundamentals of gene regulation (activation of transcription; features of DNA that mediate and control transcription). and sketch a simple conceptual model of entities that describes this biological process. EXAM

Regulation of gene expression (gene regulation) is the cellular control of the amount and timing of appearance of the functional product of a gene. Although a functional gene product may be an RNA or a protein, the majority of the known mechanisms regulate the expression of protein coding genes. Any step of gene expression may be modulated, from the DNA-RNA transcription step to post-translational modification of a protein. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism.

(OR) For transcription to start RNA polymerase must bind to the promoter (in prokaryotes). In eukaryotes TFs must bind to the TF sites and only then polymerase is able to recognize the promoter region.

(OR) Gene regulation is basically referred as the control of gene expression. It determines the concentration in which the protein encoded by the gene protein in the cell should be present. There are different levels at which regulation can take place: as a "gene expression", the entire process of converting the information contained in the gene in the corresponding gene product known. This process involves several steps. At each of these steps can affect regulatory factors and control the process. The basic principles of gene regulation in all cells are equal, but there are both prokaryotes and eukaryotes, each with special features. Activation of transcription: The very first step, the control of the transcription start is, for most genes of the most important. Hence a general decision is made as to whether the gene is to be expressed or not. This decision is in the control of regulatory sequences. These are regions of DNA that are present in the immediate vicinity of the gene or even further away called as promoter, which are transcribed but not itself. These regulatory sequences may bind proteins that activate or inhibit transcription (repress). These key proteins are called transcription factors , and they allow the cell, genes with a basic mechanism on or off. An activating transcription factor activator, an inhibitory repressor mentioned. After binding of the specific transcription factors to the promoter or enhancer leads to a change in the conformation of the chromatin. This allows other proteins known as basal transcription factors to bind to DNA. The basal transcription factors recruit then the RNA polymerase and transcription of the gene is started. Binding of a repressor to the regulatory regions of DNA prevents the accumulation of more transcription factors and thus prevents an activation of the gene. Another form of repression is the so-called transcriptional interference. Here, in front of the promoter of the gene, a second promoter is present. If active, it binds to this RNA polymerase and synthesizes non-coding RNA. By this transcription, transcription of the actual gene is prevented. Hence, transcription is controlled.

(OR) Regulation of gene expression (or **gene regulation**) includes the processes that cells and viruses use to regulate the way that the information in genes is turned into gene products. Although a functional gene product can be an RNA, the majority of known mechanisms regulate protein coding genes. Any step of the gene's expression may be modulated, from DNA-RNA transcription to the post-translational modification of a protein.

In eukaryotes, the accessibility of large regions of DNA can depend on its chromatin structure, which can be altered as a result of histone modifications directed by DNA methylation, ncRNA, or DNA-binding protein. Hence these modifications may up or down regulate the expression of gene. Certain of these modifications that regulate gene expression are inheritable and are referred to as epigenetic regulation.

--- *Methylation of DNA is a common method of gene silencing*. DNA is typically methylated by methyltransferase enzymes on cytosine nucleotides in a CpG dinucleotide sequence (also called "CpG islands" when densely clustered). Abnormal methylation patterns are thought to be involved in oncogenesis. **CpG islands** or **CG islands** are genomic regions that contain a high frequency of CpG sites but to date objective definitions for CpG islands are limited. In mammalian genomes, CpG islands are typically 300-3,000 base pairs in length. They are in and near approximately 40% of promoters of mammalian genes. About 70% of human promoters have a high CpG content. Given the GC frequency however, the number of CpG dinucleotides is much lower than expected. The "p" in CpG refers to the phosphodiester bond between the cytosine and the guanine, which indicates that the C and the G are next to each other in sequence regardless of being single- or double- stranded.

--- Transcription of DNA is dictated by its structure. *In general, the density of its packing is indicative of the frequency of transcription*. Octameric protein complexes called nucleosomes are responsible for the amount of supercoiling of DNA, and these complexes can be temporarily modified by processes such as phosphorylation or more permanently modified by processes such as methylation. Such modifications are considered to be responsible for more or less permanent changes in gene expression levels.

(OR)

Regulation of transcription thus controls when transcription occurs and how much RNA is created. Transcription of a gene by RNA polymerase can be regulated by at least five mechanisms:

- Specificity factors alter the specificity of RNA polymerase for a given promoter or set of

promoters, making it more or less likely to bind to them (i.e., sigma factors used in prokaryotic transcription).

- **Repressors** bind to non-coding sequences on the DNA strand that are close to or overlapping the promoter region, impeding RNA polymerase's progress along the strand, thus impeding the expression of the gene.

- **General transcription factors** position RNA polymerase at the start of a protein-coding sequence and then release the polymerase to transcribe the mRNA.

- **Activators** enhance the interaction between RNA polymerase and a particular promoter, encouraging the expression of the gene. Activators do this by increasing the attraction of RNA polymerase for the promoter, through interactions with subunits of the RNA polymerase or indirectly by changing the structure of the DNA.

- **Enhancers** are sites on the DNA helix that are bound to by activators in order to loop the DNA bringing a specific promoter to the initiation complex. Enhancers are much more common in eukaryote than prokaryotes, where only a few examples exist (to date)

4.4. Explain how in silico prediction of transcription factor binding sites can be validated through molecular biology experiments. How come that a computer is able to predict the start and the end of a gene? Exam 2008 ????

In silico is an expression used to mean "performed on computer or via computer simulation." The recognition of transcription factor binding sites (TFBSs) is the first step on the way to deciphering the DNA regulatory code. Blocking and activating the TFBS.

You can use pulldown assays to verify predicted binding sites. Attach a magnetic bead to the nucleotide sequence in question, allow it to bind to proteins, pull out the compounds using magnetic force, wash of unbound proteins and check whether the sequence bound to the TF's as predicted (using gel electrophoresis, MS, western plotting, ...). Other methods: Yeast 2 hybrid, site-directed mutagenesis.

4.5. What is a 5 ´UTR?

Sequences on the 5' end of mRNA but not translated into protein is the 5’ UTR. It extends from the transcription start site to just before the ATG translation initiation codon. 5' UTR may contain sequences that regulate translation efficiency or mRNA stability.

(OR) The five prime untranslated region (5' UTR) is a particular section of messenger RNA

(mRNA). A mRNA codes for a protein through translation. The mRNA also contains regions that are not translated: in eukaryotes, the 5' untranslated region, 3' untranslated region, cap and polyA tail. In prokaryotic mRNA the 5' UTR is normally short. Some viruses and cellular genes have unusual long structured 5'UTRs which may have roles in gene expression.

(OR) The untranslated region at the 5'-end of an mRNA. It contains several functional elements, like binding sites for proteins that alter the RNA's stability or location in the cell, as well as sequences that promote the initiation of translation.

(OR) Sequences on the 5' end of mRNA but not translated into protein is the 5' UTR. It extends from the transcription start site to just before the ATG translation initiation codon. 5' UTR may contain sequences.

4.6. How does the transcriptional machinery know about the beginning and the end of a "gene" (a transcript)?

In eukaryotes, RNA polymerase requires the presence of a core promoter sequence in the DNA. Promoters are regions of DNA that promote transcription. Core promoters are sequences within the promoter that are essential for transcription initiation. RNA polymerase is able to bind to core promoters in the presence of various specific transcription factors. (TATA-box) transcription initiation complex Or just: START and STOP codons indicate the begging and the end of a gene transcript. ATG for START UGA for STOP.

(OR) In prokaryotes, RNA polymerase just binds to the 3' end of the gene (the promoter) to initiate transcription. In eukaryotes and archaea, TF's must first mediate the binding of the polymerase to the promoter. After elongation, transcription is terminated either rho-independent (when the synthesized RNA forms a hairpin loop) or rho-dependent (occurrence of the rho protein factor) in bacteria. The termination of transcription in eukaryotes is less well understood.

4.7. Which properties (features) define a class of transcription factors in TF FACTOR?

The classification scheme in TRANSFAC is based on the domain structure of transcription factors. The superclass deals with the basic domain structure and the class deals with the Leucine Zipper factors. The gene transcription factor database contains data on transcription factors, their binding sites and regulated genes.

The properties of the class of TRANSFACTOR are:

- Accession no.
- Identifier (class code)

. Date; author
. Class; link to node in heirarchial classification
. Structure description
. Factors belonging to the class
. Species
. synonyms
. homologues

Hence, it comprises of extensive information on transcription factors, their structures, functions, expression patterns etc.

(OR) The classification scheme is based on the domain structure of the TF's. Superclasses are: Basic Domain, Zinc-coordinating DNA-binding domains, Helix-turn-helix, beta- Scaffold Factors with minor groove contacts, other TF's. Within these supercategories further subdivision takes place.

4.8. Name at least three different classes (types) of transcription factors. EXAM

A transcription factor is a protein that binds to specific DNA sequences, thereby controlling the flow (or transcription) of genetic information from DNA to mRNA. Transcription factors perform this function alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.

Transcription factors may be classified by their:

- Mechanism of action
- Regulatory function
- Sequence homology (and hence structural similarity) in their DNA-binding domains.

(OR) Another classification:

- Homeo domains
- Zinc fingers
- Activator binding sites

(OR) There are three types or classes of transcription factors:

- General transcription factors form a preinitiation complex.

- Upstream transcription factors bind upstream of the initiation site to either stimulate or repress transcription.

- Inducible transcription factors also bind upstream, but must be activated or inhibited.

4.9. How are evidences for the presence of a certain domain or motif represented in TFFACTOR? (I refer to the FT line in TFFACTOR entries).

The presence of certain domains or motifs in TFFACTOR is represented by the classification scheme of TRANSFAC which is based on the domain structure of the transcription factors. The scheme includes Superclass which defines the Basic Domains, Zn-coordinating DNA binding domains, Helix turn helix, Beta-scaffold factors and other transcription factors. The Class also gives an insight into the DNA binding domains.

Usually they are described in the Functional features line in the file. Structural features and interaction factors are also suitable for being linked to references.

4.10 Describe the content and the schema of the TRANSFAC database! EXAM

4.11 Sketch the high level conceptual design of a database representing associations between gene polymorphisms and a disease phenotype! EXAM 2008

4.12 Definition of housekeeping genes EXAM

5. Microarray databases

5.1. What are microarrays?

A collection of microscopic DNA spots attached to a solid surface. They are routinely used to monitor gene expression at the mRNA level. Another name is “gene chip”.

(OR) A DNA microarray (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array for the purpose of expression profiling, monitoring expression levels for thousands of genes simultaneously. Microarrays are solid supports, on which small biomolecules (typically nucleic acids) have been immobilized in a structure way. An area containing exact one defined species of biomolecules is called an element or feature. The immobilization of thousands of features can be done at very high density, allowing to monitor hybridization and binding events of a very high numbers of biomolecules simultaneously.

(OR) A microarray (= gene chip, gene array) is a device for the large-scale, simultaneous measurement of gene expression in a sample of mRNA. It consists of a small solid support (similar to a computer chip, hence the alternative names) onto which a collection of polypeptides has been fixed, chosen in such a way that they selectively hybridize with cDNA of interest. Spots specific to a gene are distributed in an ordered manner over the chip in some sort of array. Several thousands of these spots might be present on just one MA.

5.2. What are alternative gene expression determination technologies?

- northern blot (mRNA quantification)
- RT-PCR followed by qPCR (reverse transcription quantitative polymerase chain reaction)
- SAGE (Serial analysis of gene expression, tag-based)
- MPSS (massively parallel signature sequencing).

One of the important application of Next Generation Sequencing Technology is gene expression analysis. Traditionally, sequencing-based gene expression analysis was done by Expressed Sequence Tag (EST) analysis, Serial Analysis Gene Expression (SAGE), LongSAGE, SuperSAGE and Massively Parallel Sequencing Signatures (MPSS). However, the development of NGS technologies totally changed the way we study gene expression, the structure of the transcriptome, and RNA processing. It is clear that sequencing-based transcriptome analysis in many ways is superior to microarrays, since sequencing- based method is digital, highly accurate, and easy-to-perform, whereas the microarray-generated data are analog and less

accurate, and their acquisition requires specific probe and array designs.

5.3. Explain the microarray workflow in the laboratory and map the major MAGE-OM classes to the workflow. EXAM

1) DNA clones are amplified with PCR, purified and printed onto a solid surface (the microarray) with a robot (classes: Array, BioEvent).

2) cDNA from test and control cells is labeled with different dyes, combined and hybridized to the microarray (classes: BioSource + Treatment = BioSample, LabeledExtract, Hybridization).

3) The microarray is scanned with a laser that reads out the green and red channel separately (class: Feature Extraction).

4) The images are merged and thus the relative intensities/ratio of expression can be seen and analyzed.

Major MAGE-OM classes of superclass Biomaterial are: Biosample, Biosource, Labeled extract, Compound, Bioevent, Treatment, Feature extraction, Bioassay treatment, Bioassay creation, Design element, Feature, Reporter, Composite sequence.

(OR)

1, BioSample preparation: from biosource(untreated biomaterial) to biosample(treated biomaterial, for example cells, tissues)

2, RNA extraction (we get total RNA, which can be purified by poly T+ to get mRNA)

3, Quality Control RNA preparation

4, Labelling of extract (give additional control)

5, Labelling Extract clean-up (remove salts, unincorporated Cy5/Cy3- dNTP and degraded RNA)

6, hybridization (addition of control)

7, Scanning

8, Spot finding (local background definition)

9, quantification (export of tab-delimited text files with raw data)

(OR)

o Extraction of the sample (BioMaterial) from studied tissues of an organism (BioSource) and preparation (BioSample) via a Treatment.

o Extraction of total RNA and purification to mRNA (poly-A tails!)

o Labeling of extract (LabeledExtract) using dyes or other markers (links to

Compound).

o Clean-up of the extract (removal of salts, unincorporated dyes, degraded RNA)

(BioAssayTreatment)

o Hybridisation (also a BioAssayCreation (subclass Hybridization)).

o Scanning (Image), spot finding, quantification (FeatureExtraction producing

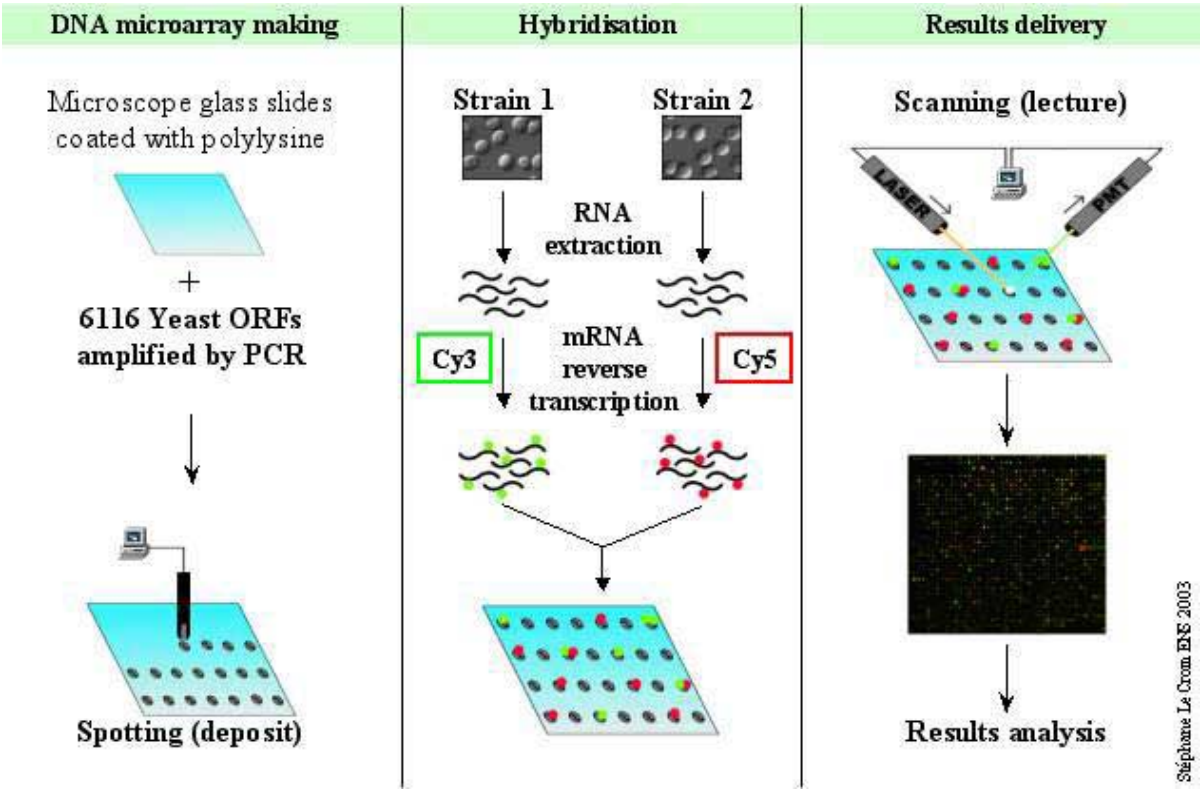
BioAssay and BioAssayData) gives us Features (subclass of DesignElement).

o Further analysis and display (MeasuredBioAssayData).

(OR)

BioSample (**class: BioSample, superclass: BioMaterial**) RNA extraction QC RNA preparation labeling (**class: LabeledExtract, superclass: BioMaterial**) hybridization (**class: Hybridization, superclass: BioAssayCreation**) scanning (**class: Feature Extraction, superclass: BioEvent**) spot finding (**class: Feature, superclass: DesignElement**) quantification

http://www.ebi.ac.uk/microarray/doc/software/schema/MAGE/MAGE.htm



5.4. What is the difference between one-color (one channel) and two-color (two channel) microarray assays? *better to drow the flow chart EXAM

one-colour/channel	**two-colour/channel**
hybridized with cDNA from one sample	hybridized with cDNA from two samples to be compared (e.g. diseased and healthy tissue)
only one dye is used	labeled with different dyes (“colours” Cy3 and Cy5)
arrays provide intensity data for each probe indicating a relative level of hybridization with the labeled target	mixed samples are hybridized to one array, after laser capturing relative intensities are identified
comparison of tow conditions for the same gene requires two separate arrays	relative differences in expression between two conditions is measured rather than absolute level of gene expression
one sample = one array	two samples = one array

Two-colour (two channel) microarrays are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different dyes. Fluorescent dyes commonly used for cDNA labeling include Cy3 (corresponding to the green part of the light spectrum), and Cy5 (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

In One-colour (one channel) microarrays,the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment. Each RNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative. The comparison of two conditions for the same gene requires two separate single-dye hybridizations. Several popular single-channel systems are the Affymetrix "Gene Chip", Illumina "Bead Chip", Agilent single-channel arrays, the Applied Microarrays "CodeLink" arrays, and the Eppendorf "DualChip & Silverquant".

(OR)

One channel microarrays are designed to measure the expression of only one sample at a time. They are designed to match parts of the sequence of the measured mRNA. The oligonucleotides required for this matching can either be synthesized in-situ or deposed by piezoelectric means.

Two channel MA's, on the other hand, measure the expression (using cDNA) of two differently labeled (usually by the means of fluorescent markers) samples at the same time. Hybridization is competitive. Two-channel MA's do not give us any information about the absolute GE levels, but rather about the ratio of expression between the two samples (up- or down-regulated features).

5.5. What are the consequences of one-channel versus two-channel hybridization for normalization and comparison between chip experiments?

One-channel arrays are easier to normalize and serve better for comparisons to other experiments, because it’s just one sample on each array; On the two-channel array the quality of one sample can influence the other and thus the overall array is biased and difficult to

normalize/compare

(OR)

The two-channel approach is good for the comparison between two samples. It will require only minimal normalisation (e.g. to account for differences in the dyes used). However, the results of different MA experiments are hardly comparable at all.

The one-channel approach gives us absolute GE values, which should – theoretically – be well comparable. However, extensive normalisation is necessary to ensure that the values are actually on the same scale. Correction of values can occur on the basis of MA specific error models. It often involves taking the log-values of the actual data since these can be assumed to follow a gaussian distribution and are hence better for statistical analysis.

(OR)

One of the advantages of the One-colour system lies in the fact that an aberrant sample cannot affect the raw data derived from other samples, because each array chip is exposed to only one sample (as opposed to a Two-color system in which a single low-quality sample may drastically impinge on overall data precision even if the other sample was of high quality). Another benefit is that data are more easily compared to arrays from different experiments so long as batch effects have been accounted for. As far as comparison is concerned the drawback of the one-color system is that, when compared to the two-color system, twice as many microarrays are needed to compare samples within an experiment.

5.6. Give an example for a one-color microarray platform:

Affymetrix (Gene Chip), Illumina (Bead Chip), Agilent (single-channel array), Eppendorf (DualChip & Silverquant)

5.7. Give an example for a two-color microarray platform

Agilent (Dual-Mode platform), Eppendorf (DualChip), Spotted Agilent MA's

5.8. Explain the following MAGE-OM (Microarray and gene expression object model) classes:

5.8.1. Array

The physical substrate along with its features and their annotation

5.8.2. Biomaterial

Represents the important substances such as cells, tissues, DNA, proteins etc., can be related to other biomaterial

5.8.3. Bio-Source

Original source material before any treatment

5.8.4. Hybridization

The event when biomaterial is hybridized to an array

5.8.5. Feature

An intended position on the array

5.8.6. Feature extraction

The process by which data is extracted from an image of an array

5.8.7. Compound

Can be a simple compound such as SDS (sodium dodecyl sulfate)

5.8.8. Ontology-entry

A single entry form an ontology or a controlled vocabulary

(OR)

o *Array* The physical substrate along with its features and their annotation.

o *Biomaterial* An abstract superclass of all biologically important substances that can be related to each other, e.g. cells, tissues, DNA, RNA, etc. (subclasses: BioSource, BioSample, LabeledExtract).

o *BioSource* The subclass of BioMaterial. The BioSource is the original source material before any treatment events. It is also a top node of the directed acyclic graph generated by treatments. The association to OntologyEntry allows enumeration of a BioSource ´s inherent properties.

o *Hybridization* An BioEvent, more specific a subclass of BioAssayCreation. As the name suggests, this describes the hybridisation of the MA with the BioMaterial.

o *Feature* An intended position on the MA.

o *Feature extraction* A subclass of BioEvent. Describes the process in which numerical data is extracted from an image of the hybridised MA.

o *Compound* Represents biochemical compounds such as dyes, salts, media, etc., might in turn consist of compounds. LabeledExtract links (n to n) to Compound as the fluorescent dyes used to label nucleic acids for hybridization.

o *Ontology-entry* Refers to an entry in an ontology or controlled vocabulary. Has a category value to describe the type of the relation (e.g. 'species') and a value for the actual referee (e.g. 'homo sapiens').

(OR)

The Microarray And Gene Expression – Object Model is to generate a model for the representation of gene expression data which allows for exchange of data between different gene expression databases.
http://www.ebi.ac.uk/microarray/doc/software/schema/MAGE/MAGE.htm
http://www.ebi.ac.uk/microarray/doc/help/mage-om.html#mage

a. Array

The physical substrate along with its features and their annotation

b. Biomaterial

Superclass, BioMaterial is an abstract class that represents the important substances such as cells, tissues, DNA, proteins, etc... Biomaterials can be related to other biomaterial through a directed acyclic graph (represented by treatment(s)).

c. bioSource

the BioSource is the original source material before any treatment events. It is also a top node of the directed acyclic graph generated by treatments. The association to Ontology Entry allows enumeration of a BioSource's inherent properties.

d. Hybridization

The archetypal bioAssayCreation event, whereby biomaterials are hybridized to an array **e. Feature** An intended position on an array, which derived from: DesignElement

f. Feature extraction

The process by which data is extracted from an image producing a measuredBioAssayData and

a measuredBioAssay

g. Compound A Compound can be a simple compound such as SDS (sodium dodecyl sulfate). It may also be made of other Compounds in proportions using Compound Measurements to enumerate the Compounds and their amounts such as LB (Luria Broth) Media.

i. Labelled Extracts *are special biosamples that have Compounds which are detectable (these are often fluorescent or reactive moieties).*

h. Ontology-entry A single entry from an ontology or a controlled vocabulary. For instance, category could be 'species name', value could be 'homo sapiens' and ontology would be taxonomy database, NCBI.

5.9. What are the key concepts used in the conceptual model of ArrayExpress?

Major components

1) ArrayExpress Experiment Archive (containing data supporting publications)

2) Gene Expression Atlas (semantically enriched database of meta-data)

The database schema is derived from MAGE-OM (a platform independent model developed in UML) Concepts: Identifiable, Array, ArrayDesign, PhysicalArrayDesign, OntologyEntry

(OR)

ArrayExpress is based on MAGE-ML:

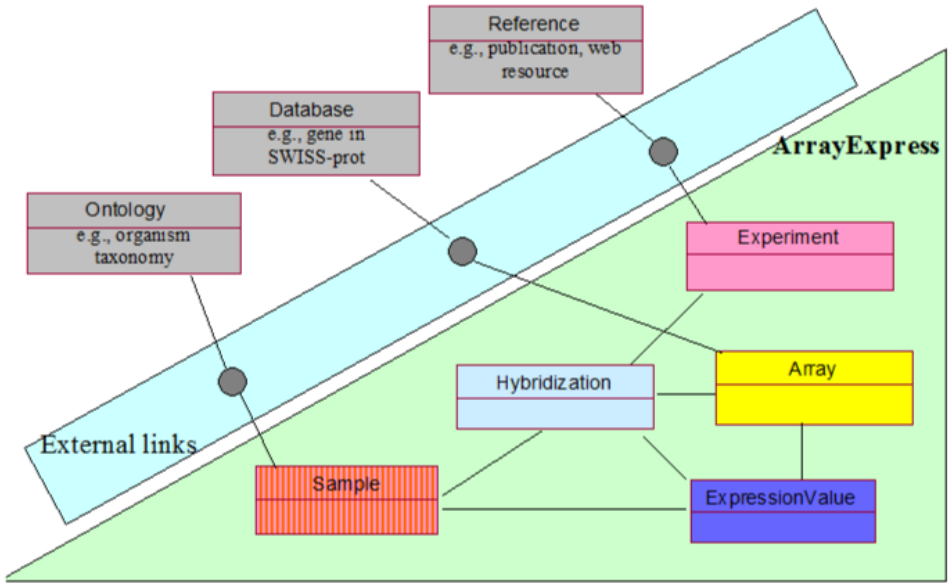
Superclass: BioMaterial; subclasses: BioSample, BioSource, Labeled Extract;

Superclass: BioEvent; subclasses: BioAssay Creation, BioAssay Treatment, Feature Extraction, Treatment;

Class: Compound;

Class: Design Element; subclasses: Feature, Reporter, CompositeSequences;

(OR)



5.10. Why is it essential to capture all data that describe the origin of the bio- sample?

Keeps information about the material type, sampling site/location (where the sample was taken), organism, developmental stage, organism part, sex, targeted cell type etc. which may influence the analysis of the actual data.

(OR)

Microarray experiments provide information about gene expression, which is a dynamic process. Gene expression differs in different types of cells, it also may change in time. Gene expression depends on the biosource and biosample preparation (e.g. what drugs where used, how long the sample was prepared etc). Thus in order to get precise information that could be compared with information attained from other experiments description of the biosample is necessary.

5.11. Outline the major differences between the conceptual design of GEO and ArrayExpress

In ArrayExpress the data is stored as “experiments” GEO you find Series and DataSets as a collection of several experiments. ArrayExpress accepts submissions in MAGE-ML format GEO data is entered as tab delimted ASCII records, it also supports SAGE data.

(OR)

a. Supported data types between GEO and ArrayExpress are different

b. GEO serves as a public repository for a wide range of high throughput experimental data. These data include single and dual channel microarray-based experiments measuring mRNA, miRNA, genomic DNA (arrayCGH, ChIP-chip, and SNP), and protein abundance, as well as non-array techniques such as serial analysis of gene expression (SAGE), mass spectrometry peptide profiling, and various types of quantitative sequence data.

c. arrayExpress focus on microarray-based experiments.

GEO has been designed following a different scope and aiming at a broad representation of all types of gene expression analysis experiments. This includes sage, chip hybridizations, and different types of tag – or signature- sequencing approaches.

(OR)

Both GEO and ArrayExpress are MIAMI compliant and both use similar schemas based on MAGE-OM. However until recently the basic difference between these databases was that in ArrayExpress it was no possibility to search for the data of a gene of interest. GEO has GEO Profiles for that purpose. On the other, hand currently ArrayExpress already has a prototype of such program, thus the difference between GEO and ArrayExpress is decreasing. Another difference is that ArrayExpress contains only data from the microarray experiments, whereas GEO in addition comprises the data from non-array techniques such as serial analysis of gene expression (SAGE) and mass spectrometry proteomic data

(OR)

- In ArrayExpress the data is stored as “experiments”, in GEO you find Series and DataSets as a collection of several experiments.

- GEO serves as a public repository for a wide range of high throughput experimental data. These data include single and dual channel microarray-based experiments measuring mRNA, miRNA, genomic DNA (arrayCGH, ChIP-chip, and SNP), and protein abundance, as well as non-array techniques such as serial analysis of gene expression (SAGE), mass spectrometry peptide profiling, and various types of quantitative sequence data. ArrayExpress focus on microarray-based experiments.

- ArrayExpress accepts submissions in MAGE-ML format, in GEO data is entered as tab delimted ASCII records, it also supports SAGE data.

GEO has been designed following a different scope and aiming at a broad representation of all types of gene expression analysis experiments. This includes sage, chip hybridizations, and different types of tag – or signature- sequencing approaches.

5.12. What are "abundantly expressed" genes? EXAM

!*unsure*! Genes that are highly expressed/over expressed (more than usual) in a sample. Zu wenig)

OR

A small number of genes, each being expressed with in a large number of gene products per cell.

5.13. What is the typical distribution of all mRNA species expressed in a cell? EXAM

We had this in our lecture!!!!!!! The graph!!!!! Lecture 8!

Some mRNAs comprise as much as 3% of the mRNA pool whereas others account for less than 0.01%. These "rare" or "low abundance" messages may have a copy number of only 5–15 molecules per cell. However, these rare species may account for as much as 11,000 different mRNA species, comprising 45% of the mRNA population.

OR

Most mRNA species have only a few copies in a cell, but relatively small number of mRNA species express in large number of copies, such as mRNA of the house keeping genes.

OR

Quantitative distribution: Abundantly-expressed genes' mRNA comprises ~90% of all the quantity of mRNA in the cell, whereas only ~10% of mRNA belongs to regulatory genes. Qualitative distribution: From qualitative point of view, regulatory genes' mRNA has bigger variety in the cell rather then abundantly-expressed mRNA.

???

5.14. Is this distribution cell-type specific?

Yes, because in each cell different genes are expressed.

OR

There are two ways to describe distribution of mRNA in the cell: Quantitative - not cell type specific (always the same shape of the curve, a few genes are expressed a lot, the rest a little); Qualitative - cell type specific (apart from housekeeping-genes, the genes that are expressed are specific to each cell type and internal and external conditions).

5.15. Name at least two foreign keys that could link a microarray database to a nucleotide sequence database or UniProt KB.

accession number, gene/protein name/symbol, RNA sequence

gene, gene product,

5.16. Is there an accession number for microarray data?

Yes, in GEO it is GSExxx (Series) or GDSxxx (DataSet), in ArrayExpress E-"source"-xxx with source indicating where the experiment is coming from.

5.17. Are images taken from microarray scanners part of the database schema of ArrayExpress?

Yes, see database schema. This is very important because the readout from the images is the first step in the analysis where different methods lead to different results.

(OR) Yes, there is a class Image in BioAssay package, which comprises images that are created by an imageAcquisition event, typically by scanning the hybridized array (the PhysicalBioAssay).

5.18. How are experimental series represented in GEO?

Series are supplied by submitters; they link together a group of related samples. They can be viewed in the accession viewer as a table with information about the study and the possibility of downloading the actual data.

(OR) Series records are supplied by submitters. 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 analysis. Each Series record

is assigned a unique and stable GEO accession number (GSExxx). Example Series record

(OR)

GEO is conceptually divided into three components: Platform (for the physical MA), sample (for one hybridization) and series (for the experiment). There's is a 1-to-n relationship from platform to sample and another one from series to sample, hence allowing to easily represent a series of experiments with many hybridizations on the same type of MA (or several types).

5.19. If you ever visited ArrayExpress, you should have read about "Gene Expression Atlas". What does the "Gene Expression Atlas" comprise?

It's a semantically enriched database of meta-analysis based summary statistics which serves queries for condition-specific gene expression patterns. It is based on a subset of the ArrayExpress data.

5.20. 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!

A tiling array consists of overlapping probes from a contiguous region of a genome and is designed to densely represent this genomic region. The region can be a whole human chromosome for example.

5.21. What Boolean operators are allowed for querying ArrayExpress (advanced search)?

AND, OR, NOT

from: http://www.ebi.ac.uk/fg/doc/help/ae_help.html#CombiningTerms

5.22. What fields can be searched in GEO and what fields can be browsed?

searching/query:

- DataSets (keywords, author etc.)
- Gene profiles (gene name, gene symbol etc.)
- accession number (GPLxxx, GSMxxx, GSExxx, GDSxxx)
- BLAST (redirecting to BLAST query)
- DataSets
- GEO accessions (Platforms, Samples, Series)

5.23. What are GEO profiles?

Gene expression profiles derived from curated GEO DataSets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within a DataSet. Profiles have various types of links including internal links that connect genes that exhibit similar behavior, and external links to relevant records in other NCBI databases.

In difference to GEO DataSets which are on a study-level, Profiles are on the gene- level.

from: http://www.ncbi.nlm.nih.gov/geo/info/profiles.html and
http://www.ncbi.nlm.nih.gov/geo/info/overview.html#query

additional possible questions might contain the following topics:

MGED (http://www.mged.org/, http://mged.sourceforge.net/ontologies/index.php)

MIAME (http://www.mged.org/Workgroups/MIAME/miame.html,

http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html)

MAGE-ML (http://www.mged.org/Workgroups/MAGE/mage-ml.html)

5.24 What are the major entity types used in the object model of ArrayExpress?

5.25 Explain the differences between ArrayExpress and GEO

5.26 What is MGED?

5.26* What is MIAME, MAGE, MGED, MGED Ontology? EXAM

MGED – international organisation of biologists, computer scientists and data analysts who work to facilitate biological and biomedical discovery through data integration. They establish standards for data quality, management, annotation and exchange, creation of tools that leverage these standards.

MIAME – international working group which short lists the format description of microarray experiments.

MAGE – aims to provide a set of standard for the representation of microarray expression data the facilitates exchange of microarray information between different data systems.

MGED Ontology – description of key experimental concepts for describing terms such as anatomy, disease, chemical etc.

5.27 What is the purpose of the MGED microarray gene expression experiment ontology?

5.28 Sketch a high level conceptual schema of a microarray database

5.29 Is there an absolute expression unit? How can we compare data from different technical platforms?

5.30 Sketch the workflow from deep sequencing (RNA seq) to submission to ArrayExpress and shed some light on the impact that RNA seq has on standardization of experiment description. Exam 2014

RNAseq

1. Sample extraction from tissue
2. RNA extraection and purification
3. cDNA generation using reversw transcriptase
4. Illumina sequencing
5. Read mapping via Software
6. Statisctical analysis based on raw count

Standartization: it does not depend on the prior knowledge of genome? …dynamic range of expression as low…?

Accuracy

Reproducibility for both replicates information about exon connectivity, tr? SNPs etc.

5.31 How are RNAseq data being integrated in GeneBuilds in ENSEMBL? Exam 2014 !! not full amount of points

1. RNAseq data submitted in ArrayExpress

2. Row sequencing reads in ENA

3. GeneBuilds in ENSEMBL using EnsembleGene:d present in ENA as foreign key

5.32 Which major MAGE-OM classes can be re-used in an experiment, where gene expression is monitored through quantitative PCR? Exam 2008

5.33 Which technology platforms for gene expression analysis do you know and how are they represented in ArrayExpress and GEO? Exam 2008

5.34 Relate microarray DB to ENTREZ feature ??? EXAM

5.35 Single spotted array process (e.g. Illumina) EXAM

5.36 Explain the procedure for Expression databases EXAM

- Build interactions network for interesting protein. Database for PPI – IntAct, String, HPD
- visualise the network to detect interaction esting features
- identify proteins that connect the initial proteins
- highlight protein in the network that are over and under expressed
- predict novel interaction of the protein of interest using interlogues.
- Class networks by -----, biological process

6. Protein databases

6.1. What is the major source of knowledge on proteins?

The UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. UniProt Knowledgebase is a central database of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR; and is most comprehensive catalogue of information of proteins.

6.2. Which parts of UniProt can be distinguished and what is the purpose of the partitioning of UniProt? EXAM

- UniProtKB/Swiss-Prot

manually annotated and reviewed o non-redundant o brings together experimental results, computer features and scientific conclusions.

- UniProtKB/TrEMBL

automatically annotated and not reviewed

high quality computationally analyzed records enriched with automatic annotation and classification.

separated from the UniProtKB/Swiss-Prot entries so that the high quality data of the latter is not diluted

curated data is much more reliable because is supported by evidence (experimental) that we can trace where is coming from (provenance)

OR

Three major parts:

- UniProt Knowledge Base (consisting in turn of SwissProt, TrEMBL and PIR) is the curated DB of all knowledge about proteins (names, sequence, taxonomic and bibliographic data + annotations: e.g. functional info, posttranslational modifications, diseases, structural info, etc.)
  - UniRef (100/90/50): Gives clustered sets of genes (with different clustering thresholds) to speed up searches and find similarities.
  - UniPArc: A comprehensive, non-redundant repository about the history of protein sequences.

Purpose of partitioning: To clearly distinguish between the curated knowledge in the KB and the archive... Each part has its own purpose.

The primary aim of the partitioning of UniProt is to support biological research by maintaining a high quality database that serves a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledge base, with extensive cross-references and querying interfaces freely accessible to the scientific community.

OR

UniProt is comprised of three components, each optimised for different uses. The UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record to speed searches. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

UniProtKB consists of two sections: **UniProtKB/Swiss-Prot** which is manually annotated and is reviewed and **UniProtKB/TrEMBL** which is automatically annotated and is not reviewed. UniProtKB/TrEMBL contains high quality computationally analyzed records that are enriched with automatic annotation and classification. These UniProtKB/TrEMBL unreviewed entries are kept separated from the UniProtKB/Swiss-Prot manually reviewed entries so that the high quality data of the latter is not diluted in any way.

Swiss-Prot is manually annotated and reviewed by curators, TrEMBL is a collection of computer analyzed records expecting to be curated. If we want to access to a protein information, curated data is much more reliable because is supported by evidence (experimental) that we can trace where is coming from (provenance). Swiss-Prot also is non-redundant and brings together experimental results, computer features and scientific conclusions.

These two data sets coexisted during some time with different protein sequence coverage and annotation priorities. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up.

What are the differences between UniProtKB/Swiss-Prot and UniProtKB/TrEMBL?

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

6.3. What are the original root databases of UniProt KB?

- EBI and SIB produced Swiss-Prot and TremBL, and PIR produced PIR-PSD (protein sequence database). Both coexisted with differing in protein sequence coverage and annotation priorities.
- TrEMBL was created because the huge amount of generated data (exceeding Swiss-Prot ´s ability to keep up)
- PIR maintained PIR-PSD, including iProClass, a database of protein sequences and curated families.
- Finally, all of them decided to pool their overlapping resources, efforts and expertise.

(OR)

- SwissProt (the supercurators),
- TrEMBL (computationally translated ORFs from EMBL),
- PIR

6.4. Why is UniProtKB called a "curated" database? EXAM

In Uniprot KB they try to produce a definitive description of proteins (sequence, functions, classification,...). UniProt curators extract biological information from the literature and perform numerous computational analyses.

OR

This is because UniprotKB/Swiss-Prot strives to provide a high level of annotation (such as the description of the functions of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy, expert curators, and a high level of integration with other databases.

OR

Because there are curators to take care of it, doh! Each entry is checked by one or more experts, double-checked with other entries (minimally redundant!) and annotated carefully with state-of-the-art knowledge (.g. functional info, posttranslational modifications, diseases, structural info, cross-references). N.B: a part of the UniProt KB is actually not yet curated, but only annotated computationally (-> TrEMBL)

OR

Curated databases are databases that are populated and updated with a great deal of human effort with references to experimental evidence, data provenance, citations and manual annotations made by qualified experts. The value of curated databases lies in the organization and the quality of the data they contain and, in Uniprot KB they try to produce a definitive description of proteins (sequence, functions, classification,...).

Manual annotation consists of analysis, comparison and merging of all available sequences for a given protein, as well as a critical review of associated experimental and predicted data. UniProt curators extract biological information from the literature and perform numerous computational analyses.

6.5. Why is it called a "knowledge base"?

Uniprot is a knowledge base because it provides to scientific community comprehensive, organized, structured and high-quality data sources about protein sequence and functional information.

OR

A **knowledge base** is a special kind of database (information repositories) for knowledge management. A knowledge base provides a means for information to be collected, organized, shared, searched and utilised. Uniprot is a knowledge base because provides to scientific community comprehensive, organized, structured and high-quality data sources about protein sequence and functional information.

OR

UniProt knowledge base is created by merging the data in Swiss-Prot, TrEMBL and PIR-PSD, which may contain more information than available in any given separate source database. Also,it contains Swiss-Prot, which is recognised as the gold standard of protein annotation, with extensive cross references, literature citations, and computational analysis provided by expert curators.

OR

The information contained in UniProt KB is supposed to represent the height of what we currently know about proteins. What is in there can be considered a fact. Also, since it combines the information from its three member DBs it contains more info.

6.6. What is meant by the field "synonyms" in UniProt KB?

Genes and proteins are often associated with multiple names, and more names are added as new

functional or structural information is discovered. The field simply lists all the names.

OR

This field can be found in the “Protein description” part of UniProt KB entry. The reason for this is that, consistent nomenclature is indispensable for communication, literature searching and entry retrieval. Many species-specific communities have established gene nomenclature committees that try to assign consistent and, if possible, meaningful gene symbols. Other scientific communities have established protein nomenclatures for a set of proteins based on sequence similarity and/or function. But there is no established organization involved in the standardization of protein names, nor are there any efforts to establish naming rules that are valid across the largest spectrum of species possible. So, we can see that one protein can have more then one assigned name. For such alternative names there is a special field “synonyms” in UniProt KB.

OR

Other names by which the protein is known. There can be MANY, since many different naming conventions exists (some naming is done on the basis of species, other w.r.t. to families and still others with respect to function or diseases, etc.).

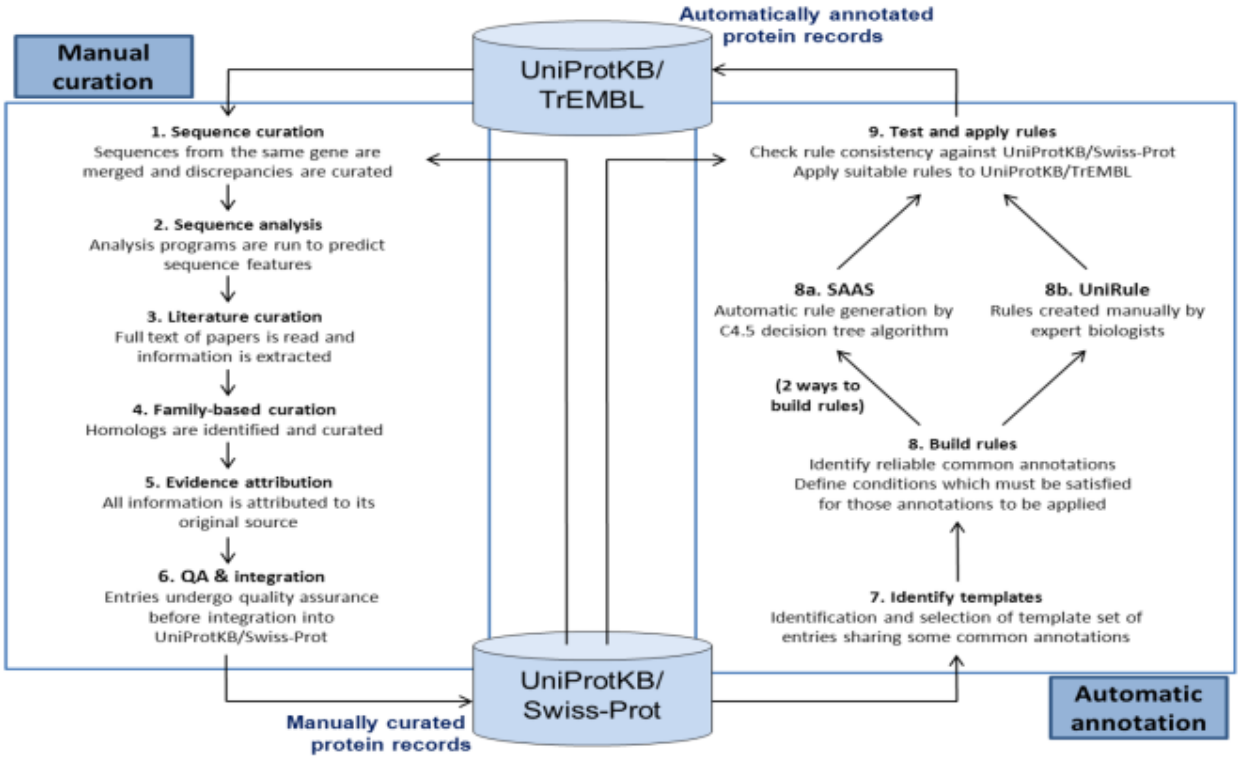
OR

Genes and proteins are often associated with multiple names, and more names are added as new functional or structural information is discovered. Ambiguities regarding gene/protein names are a major problem in the literature and in the sequence databases which tend to propagate the confusion. Because authors often alternate between these synonyms, information retrieval and extraction benefits from identifying these synonymous names. Synonym relationships between gene and protein names are mainly extracted by laborious manual curating and review.

Example: Some gene names have synonyms that are more familiar to biologists of particular breeds. For example, "SELL" means little to immunologists, whereas SELL's alias "CD62L", means rather a lot. Showing the biologist a list of gene names and saying "do any of these ring bells" seems to be an important part of the process of selecting important genes (or, rather, its validation).

6.7. Describe the workflow underlying TREMBL

UniProt annotation flow diagram:



Automatic classification: UniProt uses InterPro to classify sequences at superfamily, family and subfamily levels and to predict the occurrence of functional domains and important sites.

Automatic annotation:

- UniProt has developed two prediction systems, UniRule and the Statistical Automatic Annotation System (SAAS) to automatically annotate UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy UniRule

- Rules are devised and tested by experienced curators using experimental data from manually annotated entries as templates. All predictions are refreshed with each UniProtKB release to ensure the latest state-of- knowledge predictions. SAAS

- SAAS generates automatic rules for functional annotation from expertly annotated entries in UniProtKB/Swiss-Prot using the C4.5 decision tree algorithm. This algorithm uses machine learning to find the most concise rule for an annotation based on the properties of sequence length, InterPro group membership and taxonomy.

OR

UniProtKB/TrEMBL is a computer-annotated protein sequence database complementing the UniProtKB/Swiss-Prot Protein Knowledgebase. UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProtKB/Swiss-Prot. The database is enriched with automated classification and annotation.

OR

Amino acid sequences of the proteins are generated computationally by translating open reading frames from EMBL nucleotide sequence database. Since not a single variant of translation exists, trEMBL data is not as reliable as Swiss-Prot data.

OR

Automatic classification

UniProt uses InterPro to classify sequences at superfamily, family and subfamily levels and to predict the occurrence of functional domains and important sites. InterPro integrates predictive models of protein function, so-called 'signatures', from a number of member databases. InterPro matches are automatically annotated to UniProtKB entries as database cross-references with every InterPro release.

Automatic annotation

UniProt has developed two prediction systems, UniRule and the Statistical Automatic Annotation System (SAAS) to automatically annotate UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy:

- Based on rules
- Rules are created, tested and validated against published experimental data in UniProtKB/Swiss- Prot
- Rules are linked to InterPro member database signatures
- Rules have annotations and conditions
- Rules are reapplied to UniProtKB/TrEMBL every four-weekly release with both automatic and manual QA procedures ensuring they are still valid

UniRule

Rules are devised and tested by experienced curators using experimental data from manually annotated entries as templates. The Unified Rule (UniRule) system is being developed by merging existing manual rule-based systems (HAMAP, PIR name and site rules, and RuleBase

rules) into one system which stores, applies, and evaluates all rules. Although originally developed independently, these rule systems all share a common scientific approach of using protein family membership coupled with additional evolutionary and sequence analysis to accurately identify and annotate protein sequences. UniRule rules can annotate protein properties such as the protein name, function, catalytic activity, pathway membership, and subcellular location, along with sequence specific information, such as the

positions of post-translational modifications and active sites. All predictions are refreshed with each UniProtKB release to ensure the latest state-of-knowledge predictions.

SAAS

SAAS generates automatic rules for functional annotation from expertly annotated entries in UniProtKB/Swiss-Prot using the C4.5 decision tree algorithm. This algorithm uses machine learning to find the most concise rule for an annotation based on the properties of sequence length, InterPro group membership and taxonomy. SAAS employs a data exclusion set that censors data not suitable for computational annotation (such as specific biophysical or chemical properties) and generates human- readable rules for each release. SAAS rules can annotate protein properties such as function, catalytic activity, pathway membership, and subcellular location, but protein names and feature predictions are currently excluded. Generating rules on-the-fly in this way allows rules to evolve along with the content of UniProtKB with little or no manual intervention. It also provides a constant supply of potential “seed rules” which can be further developed by the curators into UniRules

6.8. What is the difference between the Swissprot keywords and GO-terms?

UniProtKB keywords are controlled vocabulary developed according to the need and content of UniProtKB/Swiss-Prot entries. They provide a summary of the entry content and are used to index entries based on 10 categories (Biological process, Cellular component, Coding sequence diversity, Developmental stage, Disease, Domain, Ligand, Molecular function, Post-translation modification, Technical term). Each keyword is attributed manually to UniProtKB/Swiss-Prot entries and automatically to UniProtKB/TrEMBL entries (according to specific annotation rules).

The Gene Ontology project (GO) provides a controlled vocabulary to describe gene and gene product attributes in any organism. This controlled vocabulary is developed independently of any existing databases. There are 3 disjoint categories. An important task is to map the GO terms with the gene and gene products and introduce them into databases via automatic (electronic) or manual annotation.

In UniProtKB, GO terms are manually mapped to keywords, EC number, InterPro matches or

HAMAP family rules and only then transferred automatically to the entries. Direct GO annotation in UniProtKB/Swiss-Prot entries is entirely manual.

OR

The building blocks of the Gene Ontology are the terms. Each entry in GO has a unique numerical identifier of the form GO:nnnnnnn, and a term name, e.g. cell, fibroblast growth factor receptor binding or signal transduction. Each term is also assigned to one of the three ontologies, molecular function, cellular component or biological process. Whereas the keywords in Swiss-Prot are KW (KeyWord) lines which provide information that can be used to generate indexes of the sequence entries based on functional, structural, or other categories. The keywords chosen for each entry serve as a subject reference for the sequence.

OR

GO terms are always a part of one of three sub-ontologies (BP, CC, MF), hence their scope is, in principal, more limited (that should really narrow the applicability in this case, though). They are organised hierachically. SwissProt keywords, on the other hand, are merely indices (along several dimensions, like functional and structural categories).

6.9. Describe the workflow of the generation of a new Swissprot/UniProt KB entry.

OR

The submission of a new protein sequence to UniProtKB can be done by SPIN. SPIN is the web-based tool for submitting directly sequenced protein sequences and their biological annotations to the UniProt Knowledge base. SPIN guides you through a sequence of WWW forms allowing interactive submission. The information required to create a database entry will be collected during this process. The data will need to be reviewed and annotated by the SwissProt experts before publication.

OR

About 85% of the protein sequences provided by UniProtKB come from the translations of coding sequences (CDS) submitted to the EMBL-Bank/GenBank/DDBJ nucleotide sequence resources (International Nucleotide Sequence Database Collaboration (INSDC)). These CDS are either generated by the application of gene prediction programs to genomic DNA sequences or via the translation of cDNAs. These CDS are either generated by gene prediction programs or are experimentally proven.

A protein identifier ("protein_id") is assigned to the translated CDS and can be found in the original EMBL-Bank/GenBank/DDBJ record and in the relevant UniProtKB entry.

The translated CDS sequences are automatically transferred to the TrEMBL section of UniProtKB. The TrEMBL records can be selected for further manual annotation and then integrated into the UniProtKB/Swiss-Prot section.

In addition to translated CDS, UniProtKB protein sequences may come from:

- the PDB database.

- sequences experimentally obtained by direct protein sequencing, by Edman degradation or MS/MS experiments and submitted to UniProtKB/Swiss-Prot. Less than 5% of the UniProtKB/Swiss-Prot entries contain sequence data obtained by direct protein sequencing (list

of entries with the keyword 'Direct protein sequencing').

- sequences scanned from the literature (i.e. PRF or other journal scan project).

- sequences derived from gene prediction, not submitted to EMBL-Bank/GenBank/DDBJ (Ensembl , RefSeq, CCDS, etc). These data are restricted to some organisms, such as homo sapiens.

- sequences derived from in-house gene prediction, in very specific cases.

The 'Protein existence' subsection of the 'Protein attributes' section indicates the evidence for the existence of a given protein, 5 levels of evidence have been defined:

- PE1: evidence at protein level (e.g. clear identification by mass spectrometry)

- PE2: evidence at transcript level (e.g. the existence of cDNA)

- PE3: inferred by homology (a predicted protein which has been assigned membership of a defined protein family in UniProtKB)

- PE4: predicted (a predicted protein which has not yet been assigned membership of a defined protein family in UniProtKB)

- PE5: uncertain (e.g. dubious sequences, such as those derived from the erroneous translation of a pseudogene or non-coding RNA).

Primary source of protein knowledge are journal articles. Information on a given protein is often spread between many different reports. All relevant references used by the annotators to create or update an entry are generally added. This includes references dealing with sequences,

structures, protein-protein interactions, post-translational modifications, subcellular locations, functions, etc.

Reviews are read, especially in the case of extensively studied proteins in order to select the relevant references, but usually not cited, as the UniProtKB/Swiss-Prot relevant information is retrieved from the original publications, which in turn are cited in the appropriate entry. This is particularly important to make sure that the experimental data are reported in the species in which the experiments were performed.

6.10. Why is SwissProt / UniProt KB called a "semantic hub" for molecular biology data? Explain! ??? exam 2008

The UniProt Knowledgebase (UniProtKB) acts as a central hub of protein knowledge by providing a unified view of protein sequence and functional information. Manual and automatic annotation procedures are used to add data directly to the database while extensive cross-referencing to more than 120 external databases provides access to additional relevant information in more specialized data collections. UniProtKB also integrates a range of data from other resources. All information is attributed to its original source, allowing users to trace the provenance of all data. The UniProt Consortium is committed to using and promoting common data exchange formats and technologies, and UniProtKB data is made freely available in a range of formats to facilitate integration with other databases.

OR

It is important to provide the users of bio-molecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections. Swiss-Prot is currently cross-referenced to more than 50 different databases. Cross-references are provided in the form of pointers to information related to Swiss-Prot entries and found in data collections other than Swiss-Prot. This extensive network of cross-references allows Swiss-Prot to play a major role as a focal point of bio-molecular database interconnectivity.

OR

Because of the numerous (>50!) cross-references to and from other DB, e.g. EMBL, PDB, PubMed, ...

6.11. Which sort of references do you know exist in the UniProt KB schema?

References between Swiss-Prot and some bio-molecular databases: organism specific databases; sequence databases; enzyme and pathway databases; family and domain databases; 2D-gel databases; genome annotation databases; 3D structure databases; PTM databases; protein family/group databases.

OR

References to the three types of sequence-related DBs (NA sequences, protein sequences, protein tertiary structures) as well as to specialised data collections. Examples are TF-FACTOR/-SITE, OMIM, EMBL, GeneBank, PDB, BLOCKS, Pfam and many more.

6.12. How is the problem of synonymous names for proteins dealt with in SwissProt / UniProtKB?

They try to attribute a recommended name to all the proteins of UniProtKB/Swiss- Prot, following as far as possible the rules listed in the nameprot.txt document. They have however to be as exhaustive as possible and to also list synonym names (alternative names) for a given protein. These alternative names are manually added if they are found in publication or other databases, including the EMBL database. This is essential for protein sequence retrieval and tracking.

OR

UniProt is constantly striving to further standardize the nomenclature for a given protein across related organisms. This is accomplished via protein family-driven annotation, through both manual and automated pipelines. This also involves the ongoing standardization of all the existing UniProtKB/Swiss-Prot protein names according to our guidelines.

6.13. Where would you find the biologically most relevant information in a UniProt KB entry?

The biologically most relevant information in a UniProt KB entry can be found in the amino acid sequence, protein name or description, taxonomic data and citation information. In addition is important the annotations (such as the description of the function of a protein, its domains structure, post- translational modifications, variants, etc.).

Or

Arguably, the most relevant information, like function and description, similarities, etc., can be

found in the annotations of each entry (which are unfortunately, mostly in a rather free-text-like form and hence computationally complicated to analyse).

6.14. What sort of features described in SwissProt do you know?

The FT (Feature Table) lines provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence.

Some features: Temporal and special context , post-translational modifications of residues, known interactions, transmembrane, intramembrane, peptides, topological domain (mitochondria, cytoplasm), domains(specific combination of secondary structures organized into a characteristic three-dimensional structure or fold), repeats, calcium binding regions, zinc fingers, DNA binding regions, region of interest in the sequence, motifs, bindings sites for a metal ion, binding sites (chemical compounds and amino acid sequence) ...

OR

Protein name and synonyms, description and function, sequence, taxonomic information, literature references, posttranslational modifications, cross-references, protein families, domains and sites, secondary, tertiary and quaternary structure, comments, related diseases.

OR

A curated protein sequence database which strives to provide a high level of annotation (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.

6.15. Sketch a simple schema comprising the major SwissProt entity types (object classes represented in SwissProt / UniProtKB).

- Entry information (status, history, name, accession)
- General annotation (sequence caution, polymorphisms, RNA editing, subunit structure...)
- Names origin (protein names, gene names, encoded on)
- Ontologies (keywords, GO)
- Protein attributes (sequence length, status and processing; protein existence)

- References (literature citations)

- Sequence annotation (see features question before. Conflict, uncertainty, peptide, alternative sequence, disulfide bond, natural variant, modified residue)
- Sequences (Protein sequence data for all described protein isoforms)

6.16. Why are genes and proteins subject of patents?

They can have commercial value, like the polypeptide hormone – insulin. Consider also the gene that encodes erythropoietin (EPO), a kidney hormone that signals the bone marrow to produce more red blood cells and has a $1.5 billion market as a drug to treat anemia.

OR

Gene and protein sequences are subject of patents, because they can have commercial value, like polypeptide hormone – insulin.

OR

Because they can potentially make you rich. Many genes and proteins can be exploited for commercially lucrative purposes, since they are responsible or can be used for the treatment of diseases (most prominent example: insulin).

6.17. What is the purpose of patents?

A patent is a form of intellectual property. A patent is a bargain between the State and an inventor. In return for the inventor describing the invention to the public the State rewards the inventor with a limited monopoly that will prevent unauthorized commercial use of the invention.

The publication of a patent is intended to increase human knowledge, and the inventor has to describe to the skilled reader how to make the invention work. The inventor is granted a monopoly period, usually of 20 years, during which time the inventor can exploit the invention for financial reward.

It is commonly said that the purpose of patent law is to incentivize innovation, to reward people who create a commercialized product. Once published, a patent application and all its information are available to anyone (and you can use it, if you pay for it).

(**In my opinion:** Most patents don't yield innovation. An issued patent gives an inventor a certain status in the marketplace, nothing to do with innovation. Maybe you can argue is true because companies that don ´t want to pay for patents are forced to develop their own new

products and patent them, too. So, purpose of patents is a mechanism for making money that gives benefits to the companies that possess it.)

OR

To allow the patent holder to draw some financial benefit from his findings, while protecting his righteous claim for the intellectual property for the invention (at least for a given time frame). This motivates individuals and companies to make their research public.

OR

A patent system is to provide an advantage to society as a whole by rewarding the development of new inventions. Thus, the patent system has two basic purposes: to promote the advancement of technology and to protect the inventor. The patent system provides a process for the disclosure of valuable information that can stimulate research across the globe. To obtain a patent, an inventor must "teach" the public how to make and use the invention in the best way the inventor knows. Once published, a patent application and all its information are available to anyone. Thus, the patent system greatly stimulates the flow of scientific and technological knowledge.

6.18. Do patents contain sequence information?

Yes, you can find databases with patents that contain sequence information, like Patent Proteins in EBI.

Patents of genes contain nucleic acid sequence information; patents of proteins contain amino acid sequence information.

6.19. What is a “claim” in a patent?

Patent claims are the part of a patent or patent application that defines the scope of protection granted by the patent. The claims define, in technical terms, the extent of the protection conferred by a patent, or the protection sought in a patent application. The claims are of the utmost importance both during prosecution and litigation.

OR

A “Claim” in a patent is a set of phrases following the description of an invention and describing the composition of invention, thus defining the extension of protection provided by the patent, e.g. “Protein alcohol dehydrogenase, comprising two chains...”

OR

Patent claims are usually a series of numbered expressions, following the description of the invention in a patent or patent application, and define the extent of the protection conferred by a patent or by a patent application.

OR

Patent claims are the part of a patent or patent application that defines the scope of protection granted by the patent. The claims define, in technical terms, the extent of the protection conferred by a patent, or the protection sought in a patent application. The claims are of the utmost importance both during prosecution and litigation.
For instance, a claim could read:

"An apparatus for catching mice, said apparatus comprising a base for placement on a surface, a spring member..."

"A chemical composition for cleaning windows, said composition comprising 10–15% ammonia, ..."

6.20. Which types of information from patent literature (patent protein database) is not contained in UniProt KB?

Besides inventor, applicant and classification patent numbers exclusives to the patent, related to proteins we can find these fields

- Molecule Type
- Entry Division
- Entry Data Class
- Sequence Version

OR

UniProt KB and a patent of a protein both contain information about protein's amino acid sequence, protein's name, description, name of the founder (patent owner), publication date. On the other hand, UniProt KB does not contain the information, which is specific for patents, such as patent's number, rights that the patent ensures the patent owner etc.

6.21. What is the purpose of the IPI (International Protein Index) database and what makes it distinct from the UniProt approach?

IPI announces its closure after 10 years of service following the publication of the last release on the 27th September 2011. As a replacement, UniProt provides complete proteome data sets for all IPI species (see UniProt News).

Despite the complete determination of the genome sequence of several higher eukaryotes, their proteomes remain relatively poorly defined. Information about proteins identified by different experimental and computational methods is stored in different databases, meaning that no single resource offers full coverage of known and predicted proteins. IPI (the International Protein Index) was developed to address these issues and offers complete nonredundant data sets representing the human, mouse and rat proteomes, built from the Swiss-Prot, TrEMBL, Ensembl and RefSeq databases.

IPI is an integrated database which clusters protein sequences from different databases to provide non-redundant complete data sets for selected higher eukaryotic organisms. Since it was launched in 2001, IPI has covered the gaps in the gene predictions between different databases, but since then the situation has improved for many of the most-studied genomes. This is due to a close collaboration between Ensembl, RefSeq and UniProt which aims to provide a standard set of gene predictions for the genomes of interest. The new complete proteome sets will therefore provide high coverage complete proteome sets for IPI users.

The UniProt Knowledgebase contains protein data from all species where it is available. This data includes protein sequences determined by direct experiment and derived from the sequencing of individual DNA clones or RNA molecules. It does not, however, necessarily include predictions of protein sequences derived from the complete genome sequence of every organism where this has been determined. This is particularly an issue in higher eukaryotes. Methods for protein prediction in these species are still undergoing improvement and the predictions of groups (such as Ensembl and RefSeq) derived using these methods therefore manifest some instability. Additionally, some years before the sequence of an organism is completed, a preliminary assembly of its genome may become available, from which is it possible to make provisional protein predictions that will subsequently need revision. For these reasons, protein predictions in these species are often not submitted to the EMBL/Genbank/DDBJ nucleotide sequence databases, and do not appear in the UniProt Knowledgebase. IPI protein sets are made for a limited number of higher eukaryotic species whose genomic sequence has been completely determined but where there are a large number of predicted protein sequences that are not yet in UniProt. IPI takes data from UniProt and also from sources of such predictions, and combines them non-redundantly into a comprehensive

proteome set for each species.
OR

The IPI is a database of cross references that provides a top level guide to the main databases that describe the human, mouse and rat proteomes. IPI is created by identifying entries from the constituent databases that represent the same protein; and using these mappings to automatically create a datasets with maximum extent but minimum redundancy. Biodatabase. Lecture I/11/2007.
Pages: 15, 17.
IPI is distinct from the UniProt approach, because its protein sets are made of a limited number of higher eukaryotic species whose genomic sequence has been completely determined but where there are a large number of predicted protein sequences that are not yet in UniProt (no predicted sequence). IPI takes data from UniProt and also from sources of such prediction, and combines them non-redundantly into a comprehensive proteome set for each species.

OR

UniProt aims to provide maximum information about identified proteins (and some predicted proteins from the TrEMBL part) without limiting its scope to certain organisms. It does not try to reflect the complete proteomes for specific organisms. IPI protein sets are made for a limited number of higher eukaryotes which have been completely sequenced and contains many predicted proteins which are not (yet) in
17/29UniProt. It combines the predicted data with confirmed data, e.g. from UniProt, into non-redundant information about what is supposed to be the complete proteome of an organism. It provides a guide to the main databases providing further information about these proteins

6.22. What are the major source databases underlying IPI?

- UniProtKB/Swiss-Prot and UniProtKB/TrEMBL
- RefSeq (Reference Set of protein Sequences)
- Ensembl
- TAIR (The Arabidopsis Information Resource)
- H-InvDB (H-Invitational DataBase)
- Vega (VErtebrate Genome Annotation)

6.23. What is InterPro and in how far is the InterPro scope different from the IPI approach?

InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily,

family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures. InterPro doesn't just contain protein sequence as IPI, besides includes all the respective protein aspects such as families, domains, fingerprint, interactions, structure.

OR

InterPro is the integrated resource for protein domains and functional sites. As the name suggests, it's aim is to provide integrate information from several databases for the functional description of proteins and their classification into groups based on structural properties. Its memeber databases are ProSite, PRINTS, Pfam, ProDom, Smart, TigrFam, PIR, SuperFamily.

IPI does not attempt to create such a categorization nor for the functional inference, but only tries to create an overview of the complete proteome of a limited number of organisms.

OR

InterPro (Integrated Resources of Protein Domain and Functional Site) is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. http://www.ebi.ac.uk/interpro/. InterPro doesn't just contain protein sequence as IPI, besides includes all the respective protein aspects such as families, domains, fingerprint, interactions, structure.

6.24. What is the basis of protein-families? When do we speak of protein families?

A **protein family** is a group of evolutionarily-related proteins, and is often nearly synonymous with
gene family.
Proteins in a family descend from a common ancestor (see homology) and typically have similar three-dimensional structures, functions, and significant sequence similarity. While it is difficult to evaluate the significance of functional or structural similarity, there is a fairly well developed framework for evaluating the significance of similarity between a group of sequences using sequence alignment methods. Proteins that do not share a common ancestor are very unlikely to show statistically significant sequence similarity, making sequence alignment a powerful tool for identifying the members of protein families.
Currently, over 60,000 protein families have been defined, although ambiguity in the definition of *protein family* leads different researchers to wildly varying numbers.
We can talk about a homologous relationship (orthologous or paralogous). We can speak of protein families when the proteins come from similar genes in different species (originated from a common ancestor), or in the same organism.

OR

A protein family is a group of evolutionarily related proteins that are based on the notion of similarity, in other words, in a homologous relationship (orthologous or paralogous). We can speak of protein families when the proteins come from similar genes in different species (originated from a common ancestor), or in the same organism.

OR

The basis of protein-families is that the proteins of one family have common ancestor, i.e. they are evolutionary related. Proteins of one family usually have some functional and structural similarity.

6.25. What types of sequence alignment do you know? And what does PROSITE have to do with this?

Sequence alignment can be performed on a global or a local scale. Global alignment tries to completely align two sequences, while the latter just looks for high-scoring subsequences. ProSite contains patterns and matrices describing motifs with known functional and structural properties. By comparing the a given sequence, we can find out which family a protein belongs to and maybe draw conclusions to its structure and function.

OR

Pairwise: Local, global, semiglobal alignment.

Multiple sequence alignment: ClustalW, TCoffee, Muscle

Sequence against a database: FASTA and BLAST

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. Alignments use information from profiles to find a much better alignment based on chemical or functional properties.

PROSITE is a database of protein families and domains. It is based on the observation that, while there is a huge number of different proteins, most of them can be grouped, on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor.

PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains. Each of these signatures comes with documentation providing background

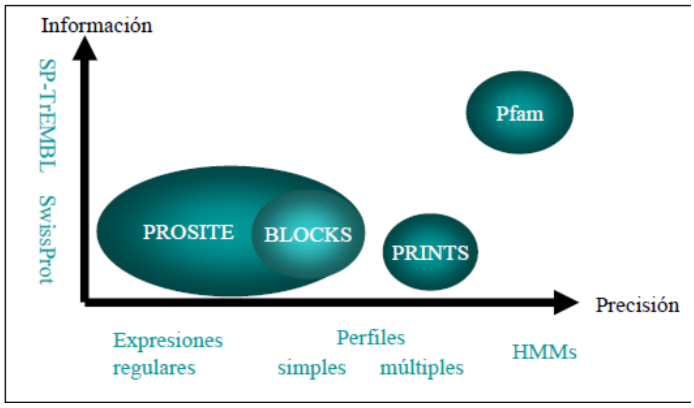
information on the structure and function of these proteins.

OR

1. Local Alignment
2. Global Alignment

PROSITE requires a multiple sequence alignment as input and uses a symbol comparison table to convert residue frequency distributions into weights. The profiles included in the current PROSITE release were generated by this procedure applying recent modifications.
http://www.expasy.org/prosite/prosuser.html#meth1

6.26. What is the difference between PROSITE, BLOCK, PFAM and PRINTS? Which ones of the above contribute to InterPro?



PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

BLOCKS Database makes blocks automatically by looking for the most highly conserved regions in groups of proteins documented in InterPro. Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

PFAM is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs.

InterPro is formed by PROSITE, Pfam and PRINTS.

SMART was the first who started using cartoon like representations. Two modes: Normal – info from SwissProt, trEMBL, Ensembl; Genomic – proteomes of organisms with fully sequenced genomes.

6.27. Define a motif?

A conserved element of a protein sequence that usually correlates with a particular function. Patterns with biological meaning

OR

In proteins of the same family, we can detect small conservative regions, commonly associated to its function : binding sites, active enzyme centers,... but conservation is not perfect (variability).

Due to functional or structural restrictions they are conserved in long evolutionary distances. If they are related to function, we can predict new functions in a new sequence (prediction), and if they are well conserved, we can detect homology. Then, can be used as patterns.
Summarizing: patterns with biological meaning.

6.28. Define a pattern?

Description of structure properties: Deterministic – Decide if the proteins matches it or not
Probabilistic – Assign a value to the match

OR

A pattern is just a structural sequence that is repeated in a set of objects, in our case a sequence. Most used patterns in proteins are motifs.

OR

An element of protein sequence, which has a higher than random probability of occurrence in the proteins. Biological meaning of a pattern is not defined, this is the main difference between a pattern and a motif.

Motifs can be used for finding structural properties. We can use regular expressions (match or not) or we can use HMM (hidden markov models) or profiles to find probabilities of being a

match.

6.29. What role does the annotation of InterPro domains with GO terms play for the prediction of new protein sequences?

OR

We can use GO terms for retrieve information associated to protein sequences similar to our new sequence.
The assignment of GO terms to InterPro entries was done manually by reading the abstract of the entries and annotation of proteins in the protein match table for each entry. An appropriate GO term for an entry is one which applies to the whole protein. The GO terms associated with an InterPro entry applies to all proteins with true hits to the signatures in that entry. The assignments are incomplete and are ongoing due to the dynamic nature of the GO project. Some entries could be mapped to very low level (specific) GO terms, while entries describing wider families or common domains were mapped to higher level terms or could not be mapped at all. If the protein matchlist is completely uncharacterised/unannotated, then no GO terms are assigned. If there are some UniProt matches but they are annotated as hypothetical because the function is not known then they are mapped to the GO term molecular function unknown. InterProScan is a tool that combines different protein signature recognition methods native to the InterPro member databases into one resource with look up of corresponding InterPro and GO annotation.

OR

Annotation of InterPro domains with GO terms allows to categorize a newly discovered protein sequence in terms of molecular function, biological process or cellular localization. It might also allow to reversely draw conclusions on which components of protein families are responsible for a certain function and hence to a better understanding of the process itself.

6.30. What is a regular expression?

A regular expression provides a concise and flexible means for "matching" (specifying and recognizing) strings of text, such as particular characters, words, or patterns of characters.

OR

A formal way to describe arbitrary strings patterns. In biology, it can be used to define and recognize patterns in AA or NA sequences.

6.31. Which experimental procedures produce the data for entries in PDB?

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules.

- X-ray crystallography
- NMR spectroscopy
- cryo-electron microscopy

OR

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

OR

X-ray crystallography, NMR (nucleic magnetic resonance).
Also electron microscopy, atom force microscopy, however very few structures predicted using theses methods are in the PDB.

6.32. Why is it necessary to crystallize proteins in order to obtain structural data?

X-ray scatter from a single molecule is very weak. In a crystal, many molecules are oriented in the same direction, thus making the X-ray scattering stronger (the waves can add up in phase and increase the signal). Therefore, a crystal acts as an amplifier.

OR

Most of the structures included in the PDB archive were determined using X-ray crystallography. For this method, the protein is purified and crystallized, then subjected to an intense beam of X-rays. The proteins in the crystal diffract the X-ray beam into one or another characteristic pattern of spots, which are then analyzed (with some tricky methods to determine the phase of the X-ray wave in each spot) to determine the distribution of electrons in the protein. The resulting map of the electron density is then interpreted to determine the location of each atom.

X-ray crystallography is an excellent method for determining the structures of rigid proteins that form nice, ordered crystals. Flexible proteins, on the other hand, are far more difficult to study by this method because crystallography relies on having many, many molecules aligned in exactly the same orientation, like a repeated pattern in wallpaper. Flexible portions of protein will often be invisible in crystallographic electron density maps, since their electron density will be smeared over a large space.

X-ray scattering from a single molecule would be incredibly weak and extremely difficult to detect above the noise level, which would include scattering from air and water. A crystal arranges huge numbers of molecules in the same orientation, so that scattered waves can add up in phase and raise the signal to a measurable level. In a sense, a crystal acts as an amplifier.

6.32* Limitations of X-Ray crystallography EXAM

1) Crystal structure for all the proteins are not known
2) Resolution should be < 1* 5 A° for better visualization, which is difficult to achieve
3) No X-ray lenses available

6.33. What mathematical approach is taken to reconstruct the 3D structure of a protein from X-ray experiments?

The diffraction pattern is related to the object diffracting the waves through a mathematical operation called the Fourier transform. If you think of the electron density as a mathematical function, then the diffraction pattern is the Fourier transform of that function.

Inverse Fourier transform of the diffraction pattern gives electron density.

6.34. Which type of screening procedure is based on protein structure information?

Virtual screening refers to score, rank, and/or filter a set of structures using one or more computational procedures. Helps decide:

– Which compounds to screen

– Which libraries to synthesize

– Which compounds to purchase from an external source PDB plays a crucial role as a database comprising protein structures used in structure based virtual screening experiments (e.g. docking). In the field of molecular modeling, **docking** is a method which predicts the preferred orientation of one molecule to a second when bound to each

other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions.

STRUCTURE-BASED VIRTUAL SCREENING

1. Protein-Ligand Docking
   - ➢ Aims to predict 3D structures when a molecule "docks" to a protein
     - Need a way to explore the space of possible protein-ligand geometries (poses)
     - Need to score or rank the poses to ID most likely binding mode and assign a priority to the molecules
   - ➢ Problem: involves many degrees of freedom (rotation, conformation) and solvent effects
2. Conformations of ligands in complexes often have very similar geometries to minimum-energy conformations of the isolated ligand

OR

Docking of the proteins. Knowing the protein structure allows predict the properties of protein binding.

6.35. Is PDB a curated database?

Yes (Data from X-ray crystallographic, NMR and cryo-electron microscopic experiments are deposited by scientists from all over the world. Annotators at the RCSB then work with these data to make sure they are represented in the PDB archive in the best possible way. They run a series of checks, make corrections, and correspond with the depositors in an effort to make the data public as quickly and accurately as possible.)

OR

Yes. The submitted data must be validated. The validation report is created automatically, later it is checked by the curator. The submitter and curator discuss the issues of the validation report.

6.36. What types of structure can be deposited to the PDB? (See PDB documentation for details)

A PDB deposition consists of a set of Cartesian coordinates and a set of structure factors or constraints from the refinement which produced the deposited coordinate set. These are accompanied by a description of the macromolecule and the details of the crystallographic or NMR experiment. What types of structure can be deposited to the PDB?

- PDB contains biomolecular polymers including polypeptides, polynucleotides, polysaccharides, and their complexes.
- Polypeptide structures containing 24 or more residues in at least one polymer chain can be deposited to the PDB. Smaller peptides that are complex with a larger polymer chain (greater than the minimum length defined above) are acceptable to the PDB.
- Polynucleotide structures with 4 or more residues are accepted at the PDB.
- Coordinates for the repeating unit of fibrous and amyloid polymers may be deposited to the PDB.
- Polysaccharide structures with 4 or more sugar residues are accepted at the PDB.

NOT ACCEPTED

. Crystal structures of peptides with fewer than 24 residues within any polymer chain such as antibiotics

. NMR structures of such molecules should be submitted to Biological Magnetic Resonance Data Bank (BMRB)

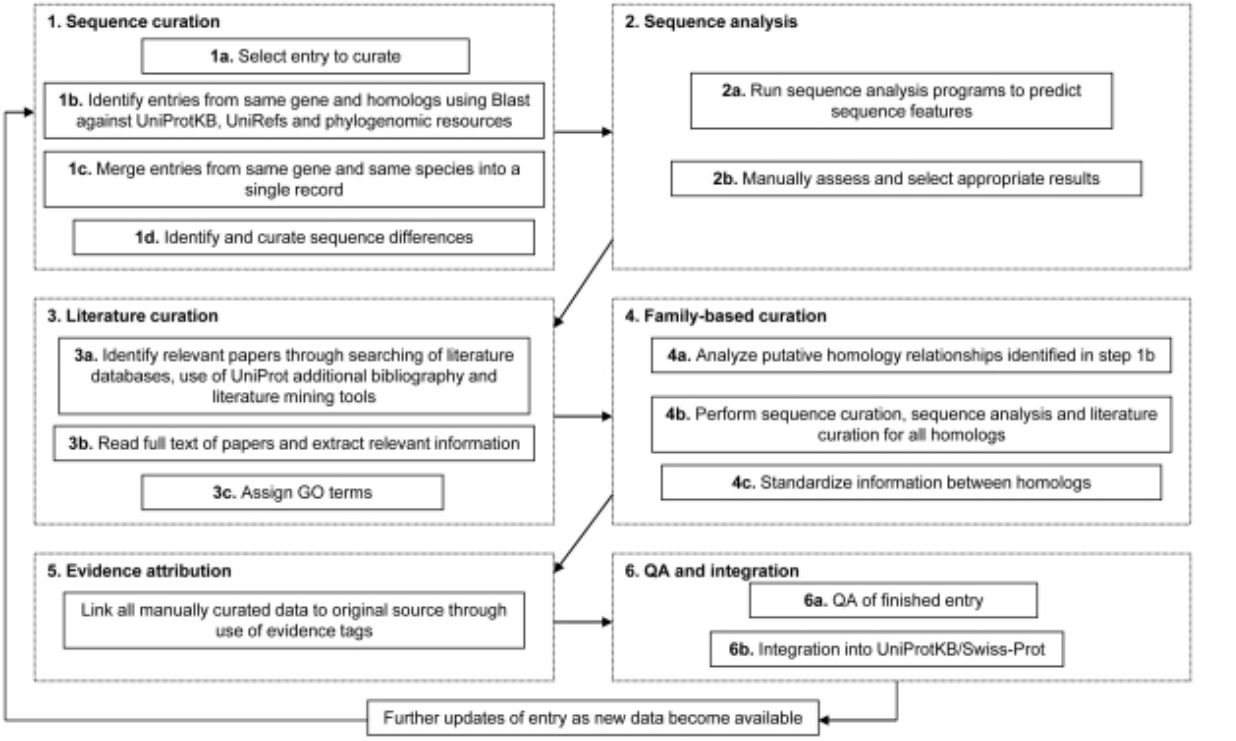
. Smaller oligonucleotides (dinucleotides and trinucleotides)

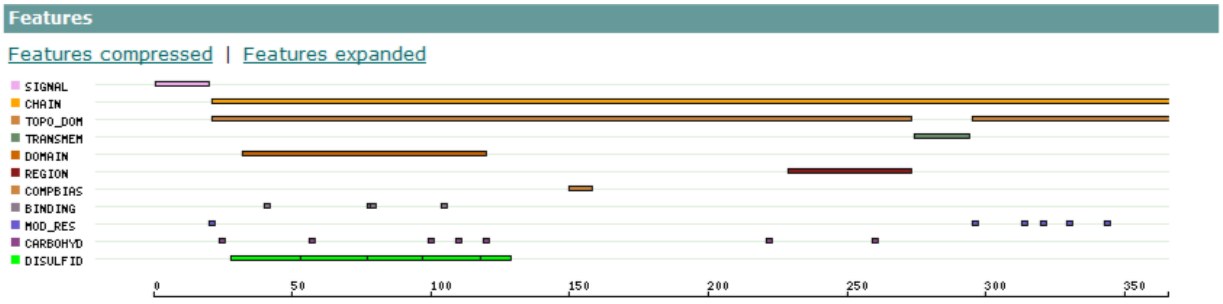
6.37. What is the Chemical Component Dictionary in PDB? What is its purpose?

Dictionary of chemical components (ligands, small molecules and monomers) referred in PDB entries and maintained by the wwPDB. It provides comprehensive search facilities for finding a particular component, or determining components in structure entries or vice versa.

Let ´s guess a new ligand arrives to PDB. Each ligand is assigned a unique ligand code (up to 3 alphanumeric characters). Annotators first check if a ligand already exists in the chemical component dictionary. If the stereochemistry and bond types match an existing ligand, the code for the existing ligand is used in the entry. If there is no match, the new ligand is added to the dictionary with an arbitrarily assigned ID code.

Specialized program is used to generate a chemical component cif file based on the author's deposited coordinates. The ligand's IUPAC-compliant name is predicted based on the deposited coordinates. All atoms in the component cif file (including H-atoms) will have coordinates. Check for ligand code is firstly carried out to verify that whether or not the new ligand exists in the chemical component dictionary.





6.38 Where would you find the biologically most relevant information in a UniProt KB entry? What sort of features described in SwissProt do you know?

6.40 Sketch a high level conceptual schema representing the major entity types of UniProtKB!

6.41 List at least five major feature lines of UniProt KB

6.42 Describe the process of annotation of a protein in UniProtKB: which evidences are used and in which cases are already existing annotations being taken over in a new entry?

6.43 Provide at least four examples for fields in UNI PROT that can be queried EXAM 2014

1. Accession number of the entry

2. protein name

3 Gene name endcoding for the protein

4. Literature reference using PubMed ID

5. Disease associated with the protin

6. Post translational modification

7. Variants

6.44 Explain the term Annotation and describe in brief the annotation of a new protein that is about to be entered in UNI PROT KB/ SwissProt EXAM 2014

Annotation is the combination of notation, reference, vocabulary, to describe the interfered and experimental information about a gene or protein.
When a new protein sequence enters into Swiss Prot from TrEMBL it is manualy annotated like this:

- Sequence curation
- Sequnce analysis
- Literature curator
- Family based curator (homology)
- Evidence attribution
- Quality assessment

6.45 In the InterPRO Tutorial, there is an explanation on the question "why classify proteins". So, why classify proteins and what are the categories that can be used classification? Exam 2014 – not full amount of points

Proteins are classified into families for functional analysis. Proteins which are functionally similar can be classified under the same family. It will be easier to identify the domains and motifs of the protein.

Categories:

1. Domains – part of protein sequence with independent existence
2. Motifs – protein sequence, associated with a particular function
3. Patterns – protein sequence wich no biological meaning
4. Site – protein sequence which acts as the interaction or binding site.
5. Repeats – protein sequence which repats multiple times.

6.46 What are the file formats in UniProt?

- flat text file format
- xml format
- rdf format
- tab delimited gff format
- fasta format

6.47 What is a primary and foreign key in general? What is the primary key in UniProt.

A foreign key is a field in a relational table that matches a candidate key or primary key of another table. The foreign key can be used to cross-reference tables.

The primary key of a relational table is a unique key selecte from the candidate keys. The primary key is used to uniquely defines each record in a relational table.

In UniProt the primary key for each protein is the primary accession number.

Every single entry in UniProt has a unique entry identifier

6.48 Describe the curation process for and entry in UniProt KB. exam 2008 2013!

6.49 How does UniProtKB / SwissProt deal with the heterogeneity of gene names in the scientific literature? Exam 2008

6.50 Sketch an approach to identify meaningful patterns in protein sequences exam 2008

6.51 What are the major entity types required to represent information on protein structures? exam 2008

6.52 Define motifs and patterns as they are used in interpretation of constituent abs and explain how motifs and patterns are detected at the protein level? EXAM 2013

6.53 What is protein gene interaction, protein chemical interaction? EXAM

Protein gene: TF binding to DNA ´s promotor
protein chemical: Drug binding to protein

7. Interaction databases

7.1. Define a simple, conceptual design for a database representing information on protein- protein interactions EXAM

A data model has 3 main components:

- Experiment
- Interaction
- Interactor

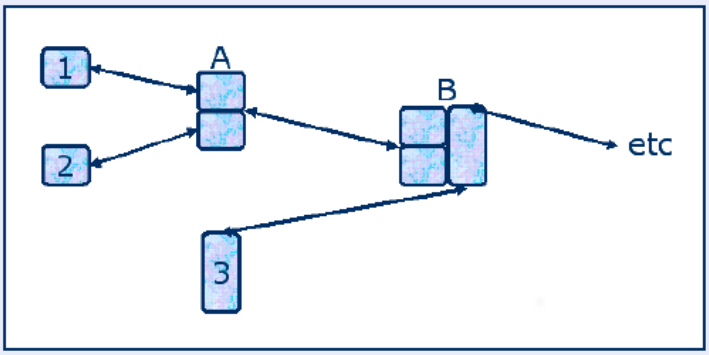
OR

Protein-protein interactions can essentially be broken down into binary interactions. The results of these interactions might act as inputs to further interactions in turn resulting in arbitrarily complex interactions. The design can hence be represented as follows:

Experiment, Interaction, Interactor? Please explain?

An experiment consists of many interactions which contain two or more interactors. An interactor can be either an individual protein or a protein complex (i.e. the result of a previous interaction). Therefore an interaction can consists of two or more protein or complex. Let me explain:

In order to model molecular interactions IntAct has designed three (virtual) objects; Experiment, Interaction and Interactor. For example the two interactions (A & B) can be described by a single Experiment. InteractionA consists of two interactors protein1 and protein2, the two interactors in InteractionB are InteractionA and protein3. Now read the first paragraph again and see if it makes more sense.



An experiment consists of many interactions which contain two or more interactors. An interactor can be either an individual protein or a protein complex, i.e. the result of a previous interaction.

*7.1. Define a simple, conceptual design for a database representing information on protein- protein interactions, involved in a disease – context!!! Exam 2008

7.2. Extend this conceptual design towards any given possible biochemical interaction (e.g. interaction of small molecules (metabolites) and proteins (enzymes))

Interactions of small molecules can be modeled using Petri-Nets. Generally speaking, they provide an intuitive approach of transforming the biological model into a graphical representation which coincides with the qualitative description of this model. Furthermore, they can be easily transformed later for quantitative simulation.

OR

The same schema as in protein-protein interaction can be used, only some additional entity types (e.g. metabolites, enzymes etc) and relationships between them (e.g. binds to, interacts with etc) have to be defined.

7.3. Draft the three major experimental techniques used to investigate protein- protein interactions.

1. Yeast-two-hybrid: Testing for physical interactions (such as binding) between two proteins
2. Co-immuno-precipitation Co-IP works by selecting an antibody that targets a known protein that is believed to be a member of a larger complex of proteins. By targeting this known member with an antibody it may become possible to pull the entire protein complex out of solution and thereby identify unknown members of the complex.
3. 2D-protein gel-electrophoresis and subsequent mass spectrometry 2D-protein gel-electrophoreses separates proteins in two steps, according to two independent properties: the first-dimension is isoelectric focusing (IEF), which separates proteins according to their isoelectric points (pI); the second- dimension is SDS-polyacrylamide gel electrophoresis (SDS-PAGE), which separates proteins according to their molecular weights (MW). In this way, complex mixtures consisted of thousands of different proteins can be resolved and the relative amount of each protein can be determined.

Or

o Yeast two hybrid: The protein we are interested (A) in is coupled to a transcription factor, which requires a co-factor to bind for activation. Proteins whose interactions with A we are interested in are coupled to this co-factor. Consequently, for proteins that interact, the TF will be activated expressing gene products that make it resistant to some sort of poison. We can than remove all yeast organisms that are not resistant and subsequently analyse the binding proteins,

e.g. By gel electrophoresis + MS.

o Immunoprecipation/ Pulldown: Protein A is couple to a magnetic bead or detected via antibodies. It is then possible to extract this specific protein from the cell together with its naturally interacting proteins and to subsequently analyse those.

o Tandem Affinity Purification: The protein of interest is marked with a TAP- tag. First it binds to the immunoglobulin IgG, the protein complex is extracted and the TAP-tag is cleaved. The second binding is to the calmodulin beads. After that the purified protein complex is eluted.

Co-immuno precipitation is considered to be the gold standard assay for protein-protein interactions, especially when it is performed with endogenous (not over expressed and not tagged) proteins.

Fluorescence Resonance Energy Transfer (FRET) is a common technique when observing the interactions of only two different proteins.

The **yeast two-hybrid** screen investigates the interaction between artificial fusion proteins inside the nucleus of yeast.

Chemical cross linking followed by high mass **MALDI mass spectrometry** can be used to analyse intact protein interactions

7.4. Sketch the conceptual model underlying the REACTOME database

REACTOME uses a frame-based knowledge representation. The data model consists of classes (frames) that describe the different concepts (e.g., reaction, simple entity). Knowledge is captured as instances of these classes (e.g., "glucose transport across the plasma membrane", "cytosolic ATP"). Classes have attributes (slots) which hold properties of the instances (e.g., the identities of the molecules that participate as inputs and outputs in a reaction).

OR

REACTOME uses a frame-based data model (that is very similar to a object-oriented class hierarchy). Top-level classes are (*ReferenceEntity*,) PhysicalEntity (subclasses: EntitiyWithAccessionedSequence, GenomeEncodedEntity, SimpleEntity, Complex, EntitySet), CatalystActivity and Event (subclasses: ReactionlikeEvent (Reaction, BlackBoxEvent, Polymerisation, Depolymerisation) and Pathway). It's important to notice that the same molecule in different cellular compartments or differently post- translationally modified variants will be represented by several instances. The class ReferenceEntitiy serves to capture

invariant features of such entities. CatalystActivity uses GO/MF terms to associate two PhysicalEntities.

7.5. Name at least three major classes of interaction types and define two subclasses (could also be instances) for each one of these interaction types.

Class	***Subclasses***
Reaction	*Acylation* *Cleavage*
BlackBoxEvent	*Activation* *Binding*
Polymerisation	*Lattice formation*
Depolymerisation	*Disintegration of the matrix layer*

7.6. Develop a strategy for the comparison of networks of interacting proteins: how would you compare networks?

1) all pairwise comparison

2) clustering

OR

In my opinion, there are two principal ways to compare interaction networks, a semantic one based one functional similarity of inputs and final products and a syntactic one based on the similarity of the reactions involved. While functional similarity probably requires human expert knowledge to come up with meaningful distance measures, but e.g. GO/MF-terms and their distance might be used as a basis. For this purpose InterPro database could be used to find out whether the proteins are similar. Also the intermediate products in the network can be compared. The syntactical approach could maybe be achieved computationally by developing statistical models of reaction types from large amounts of reaction data. For this purpose protein interaction databases should be used, e.g. IntAct, Reactome, in order to check whether the reactions in the networks are similar.

7.7. Does the domain structure of a protein allow predicting its interaction partner?

The domain of a protein is the biologically active part in the protein structure. Therefore, the knowledge of the structure of the domain helps in finding its analog for the prediction of the interacting protein molecule. Moreover, with the help of the structure of the domain we can also predict the type of interaction that is most probable to take place.

OR

Theoretically it is possible to calculate whether the protein with known structure of the domains can interact with another protein, which domain structure is also known. However, such predictions must be treated with precaution, because even if the interaction is possible theoretically, it may never occur in vivo. If these predictions are based on computationally predicted structure, the prediction of interaction is even less reliable, since even slight mistakes in structure prediction might have crucial effects on interactions.

7.8. Give two examples for predicates and rules that can be established for a given interaction of your choice (e.g. protease and substrate; kinase and kinase-substrate).

Rule for the particular type of interaction to take place for instance protease activity. The rule would be the presence of peptide bonds and its hydrolysis.

OR

Predicate Logic might be used to draw conclusions from information retrieved via text-mining.

- E.g. If we find a title “Protease A catalyses the disassociation of B”, we can draw certain conclusions (given an extensive chemical KB):

  protease(A) AND protein(B) AND hasPart(B,B1) AND hasPart(B,B2) protein(B1) AND protein(B2) AND destroyed(B).

o And similarly: kinase(A) AND protein(B) phoshporylated(B)

7.9. Give two examples how protein-protein-interactions are described in scientific text.
EXAM

Proteins bind to each other through a combination of hydrophobic bonding, van der Waals forces, and salt bridges at specific binding domains on each protein. These domains can be small binding clefts or large surfaces and can be just a few peptides long or span hundreds of amino acids, and the strength of the binding is influenced by the size of the binding domain. A common surface domain that facilitates stable protein-protein interactions is the leucine zipper, which consists of α-helices on each protein that bind to each other in a parallel fashion through the hydrophobic bonding of regularly-spaced leucine residues on each α-helix that project between the adjacent helix peptide chains. Because of the tight molecular packing, leucine zippers provide stable binding for multi-protein complexes, although all leucine zippers do not bind identically due to non-leucine amino acids in the α-helix that can reduce the molecular packing and therefore the strength of the interaction.

OR

Protein A and protein B forms a complex C. Protein A binds protein B and influences its activity.

OR

Scientific data describe the basic design of a system for automatic detection of protein-protein interactions extracted from scientific abstracts. By restricting the problem domain and imposing a number of strong assumptions which include pre-specified protein names and a limited set of verbs that represent actions, we show that it is possible to perform accurate information extraction. The performance of the system is evaluated with different cases of real-world interaction networks. Ex-glucokinase (GK) reversibly binds glucokinase regulatory protein (GKRP) to form an inactive complex. Binding is stimulated by fructose 6-phosphate and sorbitol 6-phosphate (hence high concentrations of these molecules tend to reduce GK activity) and inhibited by fructose 1-phosphate (hence a high concentration of this molecule tends to increase GK activity)

Ex-JNK (c-Jun N-terminal Kinase) phosphorylates several transcription factors including c-Jun after translocation to the nucleus.

7.10. Write down all possible terms in scientific text that indicate protein- protein-interactions or other types of molecular interactions

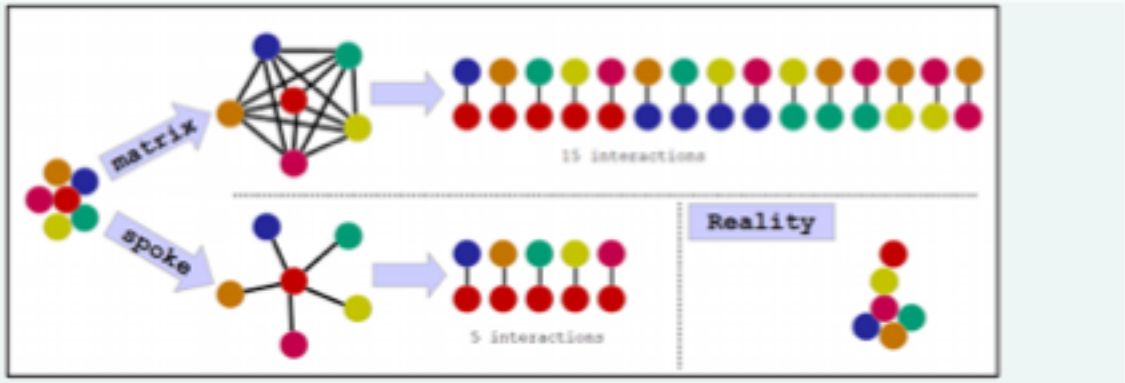
"...X binds Y...", "...X interacts with Y...", "... X phosphorylates Y ...", "... ligands X and Y...", "... X and Y form Z...", "... X inhibits the reaction Y of Z...", etc

Existing expansion algorithm

There are several known algorithm allowing to transform an n-ary interaction into a set of binary. The illustration below present the two well known expansion model and illustrates why they can be incorrect.

Spoke expansion: Links the bait molecule to all prey molecules. If N is the count of molecule in the complex, it generated N-1 binary interactions.

Matrix expansion: Links all molecule to all other molecule present in the complex. If N is the count of molecule in the complex, it generated (N*(N-1))/2 binary interactions.



7.11. What is a SPOKE expansion and how does it differ from a MATRIX expansion network? (Visit the INTACT documentation for the answer)

- Spoke expansion: Links the bait molecule to all prey molecules. If N is the count of molecule in the complex, it generated N-1 binary interactions.

- Matrix expansion: Links all molecules to all other molecules present in the complex. If N is the count of molecule in the complex, it generated (N*(N- 1))/2 binary interactions.

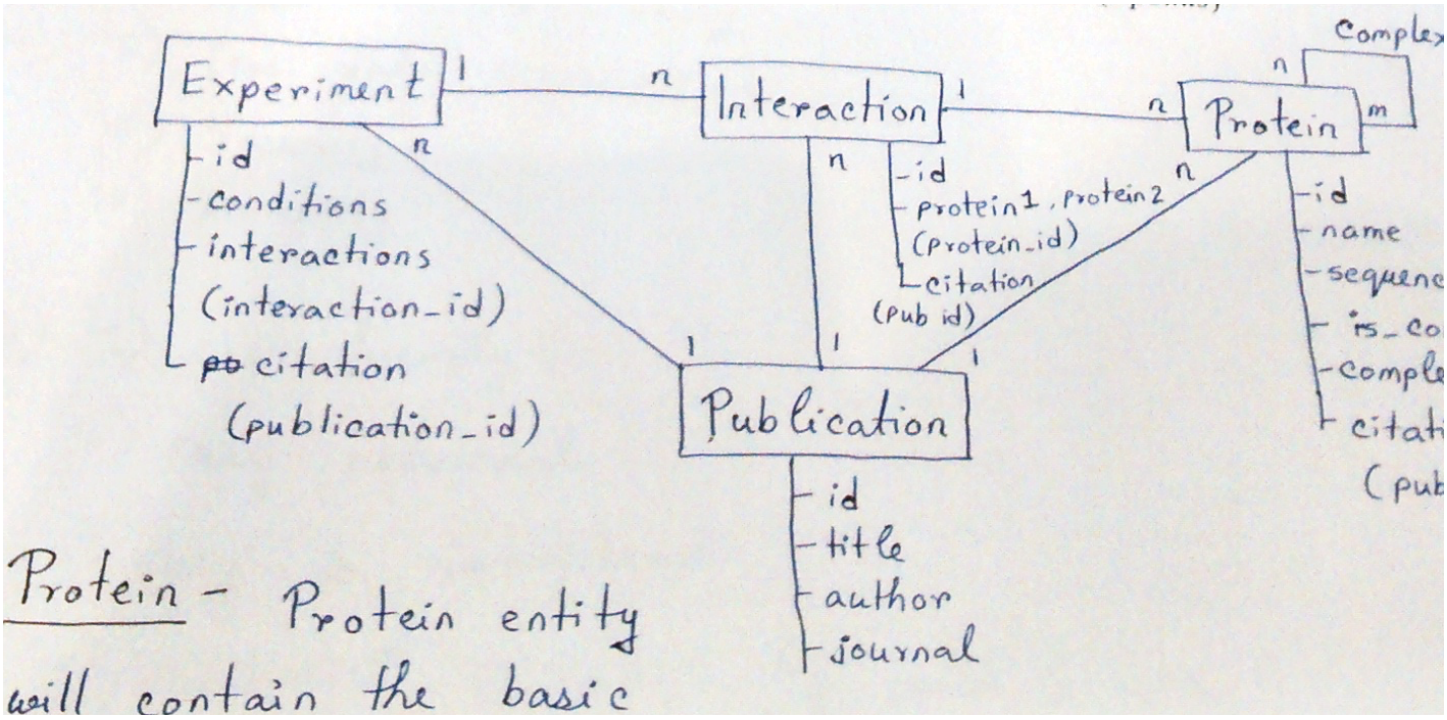
7.12 How would you design an automated program that checks protein-protein-interactions for their plausibility?

7.13 What does the domain composition of a protein have to do with protein-protein-interactions?

7.14 Sketch a high level conceptual model representing all relevant information on protein- protein interactions

7.15 Design a minimal DataBases schema for a PPI database (protein, experiment, publication, interaction). Avoid redundancy! Explain your schema.

Protein entity will contain the basic protein feature details. It will also contain information about forming complex via self join as complex itself. It will also have foreign key to publication format.



7.16 Possible terms in scientific text that indicate protein-protein interaction:

Co-immunoprecipitation, Fluorescence resonance energy transfer, Label transfer, yeast two-hybrid, *In-vivo* crosslinking of protein complexes using photo-reactive amino acid analogs, Tandem affinity purification (TAP), Chemical crosslinking, Quantitative immuno precipitation combined with knock-down (QUICK), Interferometry (DPI), Protein-protein docking, Static Light Scattering (SLS).

7.17 Define a simple, conceptual design for a database representing information on protein-protein interactions?

Biological data in particular protein- protein interaction require conceptual modelling characteristics that are not available in the widely used Entity-Relationship (ER) model.

The goals of conceptual database design representing protein – protein interactions are

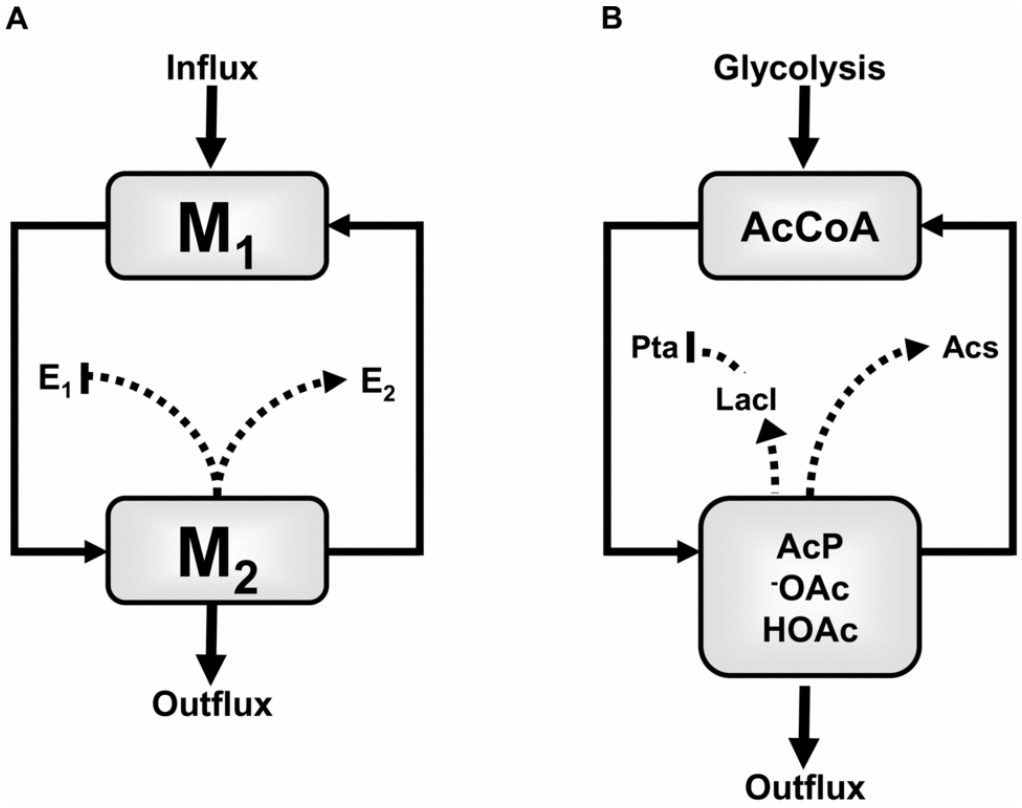
1. Represent the required data and relationships between protein – protein

interaction data.

2. Provide a data model that supports the operations that need to be performed (entering, editing, searching within the protein – protein interaction data).

3. Specify a minimal design that can accomplish these goals with acceptable performance.

7.18 Extend this conceptual design towards any given possible biochemical interaction (e.g. interaction of small molecules (metabolites) and proteins (enzymes))



The conceptual design (**A**) and the actual experimental gene metabolic network (**B**) of the metabolator are shown.

For your Deep understanding of the above figure

Incoming glycolytic flux increases the concentration of AcCoA (M_1), which is then converted into AcP (M_2) by Pta (E_1). When the AcP level increases, it activates the expression of (E_2) to Acs, which converts $^-$OAc (acetate) (M_2) into AcCoA, and , which represses the expression of . The expression of stops the source of AcP, while the expression of decreases the level of AcP. AcP is converted into $^-$OAc by constitutively expressed acetate kinase,

(To your knowledge - MONET (molecular network) ontology - MONET integrates information from metabolic pathways, protein-protein interaction networks for eukaryotes and prokaryotes, and transcription-regulatory interactions for prokaryotes, through a model able to minimize data redundancy and inconsistency.)

7.19 Does the domain structure of a protein allow to predict its interaction partner

Domain structure is that which can evolve, function and exist independently of the rest of the protein structure, it belongs to a specific family of protein. Hence a domain structure of protein helps to formulate hypothesis about its function and predict its interaction partner. (Protein-protein interaction data from similar or related proteins can provide interesting insights into the protein that you are studying. Investigating what proteins interact with another protein related to the one you are studying can give clues to potential interactors as yet unstudied).

7.20 Which types of PPI databases you know in terms of information quality/quantity?

STRING:

known and predicted protein interacitons,

BioGRID:

Model organisms and humans, over 560.000 curated interaction entries

MIPS:

Mammalian PPI database, manually curated high quality data collected from scientific literature

DIP:

combines information from several sources, manually and automatically curated, experimentally determined interactions, consistent set of PPI

IntAct:

open source database, interactions derived from literature curation or direct user submission

7.21 How can you extend or improve PPI network information?

- refine literature curation, extend PPI network based on more PPI data from several sources
- compare networks based on several sources, detect overlaps and differences
- more data => more reliable and precise network!

- text mining to detect interactions in literature and compare them with already existing database entries...

7.22 4 types of biological networks: EXAM

1) PPI Network
2) Gene regulatory network
3) Metabolic network
4) Signalling pathway

7.23 Describe network biology EXAM

Study of complex network that are abundant in biological systems.

7.24 Recommended procedure for building store network for novel unknown interactions EXAM

- start with bench of seed protein
- expands the seed using immediate neighbouring protein
- interconnect the neighbours

7.25 What is gene expression network? EXAM

Network that are built from data directly. Relationships between 2 genes are made if they are co-regulated → same behaviour over time gene regulatory network → Boolean network TF target genes – flow chart of regulation of factors which induce it graph theory

8. Enzyme and metabolic pathway databases

8.1. Give a short explanation of the principles of the Enzyme Classification (EC)

EC numbers do not specify enzymes, but enzyme-catalyzed reactions. If different enzymes (for instance from different organisms) catalyze the same reaction, then they receive the same EC number. By contrast, UniProt identifiers uniquely specify a protein by its amino acid sequence.

OR

The Enzyme Commission number (EC number) is a numerical classification scheme for enzymes, **based on the chemical reactions they catalyze**. As a system of enzyme nomenclature, every EC number is associated with a recommended name for the respective enzyme.

Strictly speaking, EC numbers do not specify enzymes, but enzyme-catalyzed reactions. **If different enzymes** (for instance from different organisms) **catalyze the same reaction, then they receive the same EC number.** By contrast, UniProt identifiers uniquely specify a protein by its amino acid sequence

OR

The *first general principle* of these 'Recommendations' is that names purporting to be names of enzymes, especially those ending in *-ase*, should be used only for single enzymes, *i.e.* single catalytic entities.

The *second general principle* is that enzymes are principally classified and named according to the reaction they catalyse. The chemical reaction catalysed is the specific property that distinguishes one enzyme from another, and it is logical to use it as the basis for the classification and naming of enzymes.

A *third general principle* adopted is that the enzymes are divided into groups on the basis of the type of reaction catalysed, and this, together with the name(s) of the substrate(s) provides a basis for naming individual enzymes. It is also the basis for classification and code numbers.

The first Enzyme Commission, in its report in 1961, devised a system for classification of enzymes that also serves as a basis for assigning code numbers to them. These code numbers, prefixed by EC, which are now widely in use, contain four elements separated by points, with the following meaning:

(i) the first number shows to which of the six main divisions (classes) the enzyme belongs,

(ii) the second figure indicates the subclass,

(iii) the third figure gives the sub-subclass,

(iv) the fourth figure is the serial number of the enzyme in its sub-subclass.

OR

Principles:

o Enzyme names denote only single enzymes, system catalysing a common

process should have the word *system* in their name.

o Enzymes should be named by the reaction they catalyse (but not by a general

phenomenon: translocase).

o Enzymes are divided into groups on the basis of the type of reaction they

catalyse. This plus the names of substrates gouverns naming of individual

enzymes and provides the basis for EC codes and classifications.

Based on these principals, six broad superclasses have been divised to account for all known enzymes: Oxidoreductases, transferases, hydrolases, lysase, isomerases, ligases.

8.2. Describe the difference between KEGG and the ENZYME database

While ENZYME focuses in the information relative to the nomenclature of the enzymes, KEGG PATHWAY is a collection of manually drawn pathway maps, consisting of interacting molecules or genes, and the only information it displays about the enzymes taking part in the pathways are their links to the LIGAND database.

OR

KEGG is a database of biological systems, consisting of genetic building blocks of genes and proteins (KEGG GENES), chemical building blocks of both endogenous and exogenous substances (KEGG LIGAND), molecular wiring diagrams of interaction and reaction networks (KEGG PATHWAY), and hierarchies and relationships of various biological objects (KEGG BRITE). KEGG provides a reference knowledge base for linking genomes to biological systems and also to environments by the processes of PATHWAY mapping and BRITE mapping.

ENZYME is a repository of information relative to the nomenclature of enzymes. It is

primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). The ENZYME data bank contains the following data for each type of characterized enzyme for which an EC number has been provided: EC number, Recommended name, Alternative names, Catalytic activity, Cofactors, Pointers to the Swiss-Prot entrie(s) that correspond to the enzyme, Pointers to disease(s) associated with a deficiency of the enzyme.

OR

ENZYME is just a repository for information regarding enzyme nomenclature. It only contains entries for each EC-number assigned enzyme, along with recommended and alternative names and some information about the catalytic activity, co-factors and links to SwissProt.

KEGG, on the other hand, is a large project comprising a number of different databases for gene and genome related information, enzymatic pathways and bioactive chemicals. As a part of one of its subbranches (KEGG Ligands), KEGG also comprises a EC-number-based nomenclature database, but apart from only giving names and minimal information about enzymes, the KEGG Enzyme database is fully integrated into the other DB's of the project and cross-references to orthologues, genes, structures and other databases.

8.3. Describe the commonalities between both databases

Both databases organize enzymes according to their EC Number, searches based on this number, or even name of the enzyme. At that point, both form the corresponding entry of the LIGAND database to which links PATHWAY, and from the ENZYME entry, cross-references to other databases with further information are common.

OR

They both have entries based on unique enzymes with an EC-number, provide recommended and alternative names, basic information on catalytic activity, co-factors and some DB cross-references (e.g. SwissProt, IUBMB EC).

8.4. What attributes of an enzyme would you need for models of metabolite flux reactions?

1) An idea about the enzyme structure

2) Active sites on the enzyme which bind different substrates

3) Allosteric regulation or other regulatory mechanisms for activation and inhibition of enzyme
4) Its specific activity and Km value for each of the substrates i.e. the lineweaver burke plot for the enzyme.

5) Information about inhibitors- competitive, non competitive and uncompetitive.

6) Knowledge about involvement of the enzyme in different metabolic pathways which may differ acc to cell type.

For modeling the metabolite flux, certain steady state assumptions need to be made and the enzyme kinetics can be used in part to construct models of the metabolite flux.

OR

Metabolic flux = The rate of turnover of molecules through a metabolic pathway or enzyme.

To model metabolite flux reactions one needed to know the substrates of an enzyme and their concentration in the specific part of the cell, co-factors, inhibitors, temperature, pH-values and all kind of other factors influencing the reactivity and reaction speed of the enzyme...

OR

metabolites: intermediates products of metabolism.

Stoichiometry: is a branch of chemistry that deals with the relative quantities of reactants and products in chemical reactions.

Flux balance analysis (FBA) is a mathematical method for analysing metabolism. It is a direct application of *linear programming* to biological systems that **uses the *stoichiometric coefficients*** for each reaction in the system as the set of constraints for the optimization

One of the main strengths of the flux balance approach is that it does not require any knowledge of the metabolite concentrations, or more importantly, the enzyme kinetics of the system; the homeostasis assumption precludes the need for knowledge of nutrients at any time as long as that quantity remains constant. **The stoichiometric coefficients alone are sufficient for the mathematical maximization of a specific objective function**

The objective function is essentially a measure of how each component in the system contributes to the production of the desired product. The product itself depends on the purpose of the model, but one of the most common examples is the study of total biomass.

Enzyme, metabolites composition and the stoichiometric coefficients alone are sufficient for the

model

8.5. Do KEGG or ENZYME comprise the relevant information?

ENZYME is the one which provides the information about the reaction which the enzyme catalysis.

OR

Kegg is a much more comprehensive database which houses multiple databases under it. Since it contains such a set of crosslinked databases it covers a wider range of data about enzymes. For example, while kegg enzymes gives information about the EC nomenclature, basic reaction, substrate and products, it also contains links to the databases kegg reaction, rpair, reaction class and compound. This is in addition to the links to external databases like Brenda and explorenz. Thus it gives much more information than ENZYME. However it does not contain information on enzyme kinetics, catalysis and inhibition.

ENZYME is quite limited, more so than KEGG as it contains only the enzyme nomenclature and external links and none of the above mentioned attributes are discussed.

ENZYME is the one which provides the information about the reaction which the enzyme catalyses.

OR

Neither KEGG nor ENZYME comprise sufficient information to properly model metabolite flux reactions. Partial information could be extracted from the pathways and descriptions of the reactions; however this information would not allow to create a quantitative model.

8.6. What is a “rate limiting step”?

The rate limiting step or rate determining step is the slowest step of a chemical reaction that determines the speed (rate) of the overall reaction. The rate of a reaction depends on the rate of the slowest step.

OR

It is the slowest step in a metabolic pathway, or the step in an enzymatic reaction that requires the greatest amount of energy to initiate.

Or

It's the slowest step in a reaction, the bottle-neck. The whole reaction can not happen quicker then its slowest sub-process.

8.7. What is a "salvage pathway"?

Salvage pathways are used to recover bases and nucleosides that are formed during degradation of RNA and DNA. The salvaged bases and nucleosides can then be converted back into nucleotides.

OR

A parthway in which bases and nucleosides from degraded nucleotide sequences are

being recycled.

OR

It is a recycling metabolic pathway in which biomolecules such as nucleotides are synthesised from intermediates in the degradative pathway for those biomolecules. The intermediate materials would otherwise be waste products.

(OR)
A **salvage pathway** is a pathway in which nucleotides (purine and pyrimidine) are synthesized from intermediates in the degradative pathway for nucleotides. Salvage pathways are used to recover bases and nucleosides that are formed during degradation of RNA and DNA.

OR

Salvage pathways are important especially in organs that cannot carry out de novo synthesis. An example of this pathway is salvaged bases and nucleoside formed from the degradation of RNA and DNA which are used and converted back into Nucleotide.

****De novo synthesis:** refers to the synthesis of complex molecules from simple molecules such as sugars or amino acids, as opposed to their being recycled after partial degradation.

8.8. Which principle approaches towards metabolite network simulation do you know?

Metabolic network reconstruction and simulation allows for an in depth insight into comprehending the molecular mechanisms of a particular organism. A reconstruction breaks down metabolic pathways into their respective reactions and enzymes, and analyzes them within the perspective of the entire network. In simplified terms, a reconstruction involves collecting all of the relevant metabolic information of an organism and then compiling it in a way that makes sense for various types of analyses to be performed.

OR

Comprehending the molecular mechanisms of a particular organism, especially correlating the genome with molecular physiology is the principle approaches towards metabolite network simulation.

The correlation between the genome and metabolism is made by searching gene databases, such as KEGG, GeneDB, etc., for particular genes by inputting enzyme or protein names.

For example, a search can be conducted based on the protein name or the EC number (a number that represents the catalytic function of the enzyme of interest) in order to find the associated.

OR

Deterministic approach: The reactions are described via the systems of differential equations. Rate constants are required. Stochastic/Probabilistic approach. Probabilities are assigned to the reactions (essentially Markov processes). Also Petri nets can be used to simulate a metabolic network.

OR

Metabolic pathways are an essential key to the systemic behaviour of a biological cell, as they describe a multitude of enzymatic reactions carrying out various functions. One possible approach to the study of these large-scale networks is dynamical simulation. But precise mathematical modelling of cellular processes is in most cases impossible because the required kinetic data are missing. Therefore alternative network-based approaches have been developed and have found much interest recently, in particular two very similar mathematical concepts called ***elementary flux modes*** **and** ***extreme pathways*** .In contrast to dynamical simulation approaches, these descriptions only require the knowledge of network topology and stoichiometry, which are well known is many cases. The former obtains one single solution, while the latter develops a solution space. These would include analysis of reactions and interaction data by petri net models

The interaction data can be derived from experiments employing different experimental techniques such as • Yeast - two - hybrid – analysis • Co-immuno-precipitation • 2D-protein gel-electrophoresis and subsequent mass spectrometry • and other variants of proteomics technologies ... Various factors need to be examined such as reaction constants, speed of reaction, concentration of molecules, their localization, states etc and these factors need to be observed over a period of time.

8.9. How would you model the role of a cofactor in an enzymatic reaction simulation?

A cofactor, in an enzymatic reaction simulation is a non-protein molecule which acts as a catalyst. I.e., it lowers the free energy of activation. Maybe: Showing the reaction with and without cofactor. I guess something related with the analysis in the reaction of the enzyme under different concentrations of the cofactor.

OR

Co-factors are required for the proper functioning of enzymes. Depending on their type they can either be integral parts of the enzymes (prosthetic groups) or only loosely bound to it (coenzymes). Either way, one way to model both of them would be as substrates that have to be present for the catalytic reaction to happen, yet remain unchanged, i.e. the list of products will again include these cofactors.

8.10. How would you represent complexes of more than one cofactor, one substrate and one enzyme in a database?

You would have to introduce a container entity type that allows for bundling various entities together. Such a Compound entity would have a 1-to-n relationship to the cofactors, substrates and enzymes. Probably it would make more sense to have one container type for each of them, i.e. CompoundEnzyme, CompoundCofactor and CompoundSubstrate.

8.11. Which multi-enzyme complexes do you know? In which biosynthesis – pathway are they involved?

Multi-enzyme complex: A multi-enzyme with catalytic domains on more than one type of polypeptide chain. Catalytic domain: Any part of a polypeptide chain that possesses a catalytic function. It may contain more than one structural domain.

1) Pyruvate Dehydrogenase complex: citric acid cycle.

2) Fatty acid synthase: fatty acid synthesis.

3) Cytochrome p450 enzyme: oxidation of organic substances

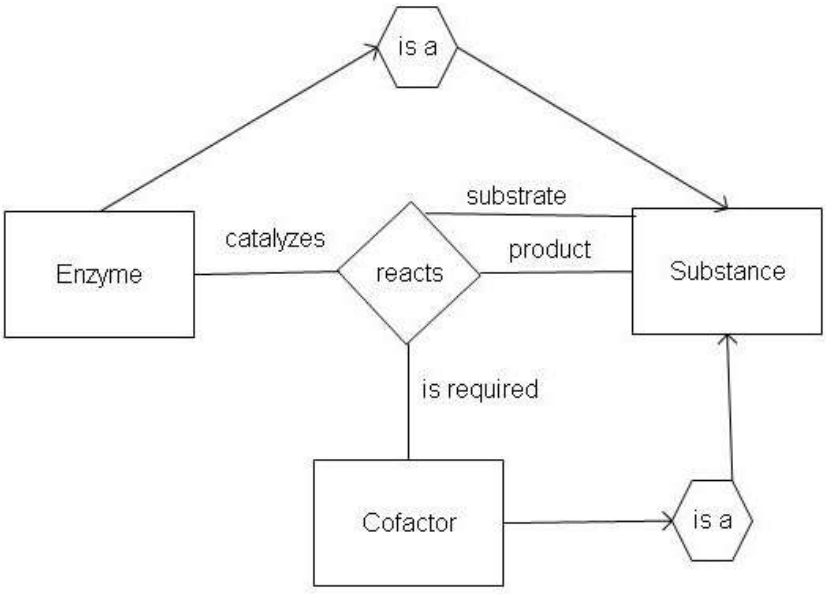
4)Phosphotransferase system in bacteria for the sugar uptake from phosphoenolpyruvate as an energy source.

5)Tryptophane synthesis multi-enzyme complex for tryptophane synthesis.

6) Aminoacyl-tRNA synthetase complex: Signalling pathways.

7) α.-keto acid dehydrogenase multienzyme complex: catalyse the oxidative decarboxylation of pyruvate

8.12. Sketch an ER diagram of a database that represents the citrate-cycle

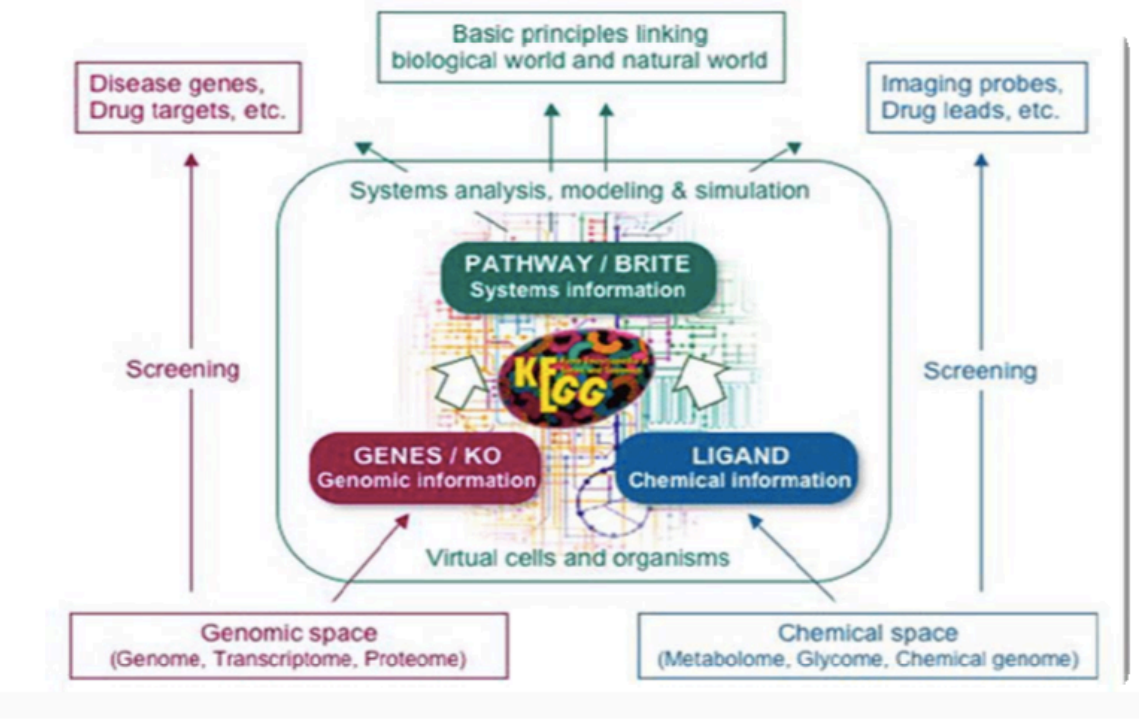


Possible Entity Types: Substrate, Product, Enzyme, Reaction-type

Possible relations: reacts with, reacts to, uses/is catalyzed by

http://www.genome.jp/kegg-bin/show_pathway?map00020

In software engineering, an **entity–relationship model** (**ER model**) is a data model for describing the data or information aspects of a business domain or its process requirements, in an abstract way that lends itself to ultimately being implemented in a database such as a relational database. The main components of ER models are entities (things) and the relationships that can exist among them



8.13. What distinguishes a cartoon – like representation of biochemical reactions and pathways in REACTOME from the representation in KEGG?

The representation in REACTOME is much more interactive than in KEGG. The entities in the pathway maps are coloured, the maps can be zoomed and scrolled. The enzyme names are also displayed on selection which makes it easier to understand. Substrates, products, intermediates and enzymes can be selected and relevant information is displayed in a pane on the left. In KEGG, the enyme classification nos are given, not their names. The pathway window is static and cannot be zoomed. If any entity is clicked on a new window linked to another relevant database under KEGG opens and the information about the entity under consideration can be found there.

OR

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks. Reactome uses a frame-based knowledge representation. The data model consists of classes (frames) that describe the different concepts (e.g., reaction, simple entity). Knowledge is captured as instances of these classes (e.g., "glucose transport across the plasma membrane", "cytosolic ATP"). Classes have attributes (slots) which hold properties of the instances (e.g., the identities of the molecules that

participate as inputs and outputs in a reaction).

OR

In KEGG each pathway links to an independent cartoon-like representation of the pathway, whereas in REACTOME the map of all the pathways is revealed. The specific pathways where the enzyme of interest acts are marked and can be enlarged. Both, KEGG and REACTOME, provide interactive hyperlinked pathway maps.

8.14. If you would have to design a "virtual physiological human" as a form of in silico representation of molecular physiology: which entity-classes would this model contain?

The Virtual Physiological Human (VPH) is a methodological and technological framework that, once established, will enable collaborative investigation of the human body as a single complex system. The collective framework will make it possible to share resources and observations formed by institutions and organizations creating disparate, but integrated computer models of the mechanical, physical and biochemical functions of a living human body. Entity classes: Organism, Organ, Tissue, Cell, Organelle, Interaction, Protein, Cell Signals, Transcript, Gene, Molecule

OR

First, "physiologic models " need to be fed by meaningful and interpretable parameters: that implies the elaboration of virtualdata formation models. Indeed, "physiologic models" must be based on an adequate and accurate observation of the living world. The exploration of the living tissues must be optimised by selecting the best physical sensors adapted to the physical nature of what has to be measured (temperature, pressure, biochemical, electro-magnetic processes, motion, mechanical dynamics, fluid dynamics, metabolism,..) providing data signals and images.

OR

Numerous. Depending on the level of detail, it could start from sub-atomic entities like neutrons and electrons, going up over molecules to macromolecules such as proteins and enzymes. It would also have to contain entity types for sub-cellular compartments, cell types, systems and organs.

8.15. Sketch a strategy for the development of an automated process (a software program) for the extraction of IC50 values from tables in scientific publications. How would you design the problem-solving approach (no need to go into details of software engineering) just describe how an automated approach should work that autonomously extracts IC50 values from tables in biochemistry publications)?

The half maximal inhibitory concentration (IC50) is a measure of the effectiveness of a compound in inhibiting biological or biochemical function. This quantitative measure indicates how much of a particular drug or other substance (inhibitors) is needed to inhibit a given biological process (or component of a process, i.e. an enzyme, cell, cell receptor or microorganism) by half.

OR

The **half maximal inhibitory concentration (IC50)** is a measure of the effectiveness of a compound in inhibiting biological or biochemical function. This quantitative measure indicates how much of a particular drug or other substance (inhibitors) is needed to inhibit a given biological process (or component of a process, i.e. an enzyme, cell, cell receptor or microorganism by half.

8.16. Conceptual model underlying the REACTOME database.

Life on the cellular level is a network of molecular interactions. Molecules are synthesized and degraded, undergo a bewildering array of temporary and permanent modifications, are transported from one location to another, and form complexes with other molecules. Reactome represents all of this complexity as reactions in which input physical entities are converted to output entities

8.17 Three major classes of interaction types and subclasses for these interaction types

a) Experiment: Subclass: BioSource, identification Method, interaction Method.

b) Interaction: Subclass: kD, CvInteractionType

c) Interactor: Subclass: CvInteractorType, BioSource.

8.18 How would you compare networks?

Following software's are used to compare the networks:

APID2NET: unified interactome graphic analyser

APID Agile Protein Interaction Data Analyser

Cytoscape

NAViGaTOR

NetPro

Osprey

8.18* What is Cytoscape software? Explain also Plug ins of Cytoscape and Applications of Cytoscape. EXAM

It is representing a model for analysing networks.
In systems biology we do reverse engineering.

- Guardian of molecular pathways active in human disease
- systematic interpretation of genetic interactions using protein networks
- identifying target proteins for drug synthesis
- determining protein complexes
- annotation of unknown proteins with functions

8.19 Commonalties between KEGG and ENZYME database:

The KEGG(Kyoto Encyclopaedia of Genes and Genomes) database also provides metabolic pathway maps and regulatory pathways maps, which can be viewed in terms of a specific organism. KEGG provides enzyme data with links to pathways, functional data, genes, diseases (OMIM database), motif and EC (Enzyme Classification) accession.

The ENZYME databases comprises of 3 databases namely BRENDA, ERGO and EXPASY. All these databases share similarities with kegg namely:

The BRENDA database is an enzyme database that provides extensive functional data on enzymes. These include details on nomenclature, reactions and specificity, enzyme structure, isolation/preparation and stability

ERGO (formerly WIT) database provides links to information about the functional role of enzymes via links to data in KEGG; links to NCBI Medline entries for each enzyme; link for record of enzymes. The database also provides access to thoroughly annotated genomes within a framework of metabolic reconstructions

Part of the EXPASY database is ENZYME which is a repository of information relative to the

nomenclature of enzymes. It provides information on each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided.

8.20 What attributes of an enzyme would you need for models of metabolite flux reactions? Attributes of an enzyme to model metabolite flux reactions

- the presence or absence of each edge (that is the enzyme), the function (sign) of each edge

- the activity of each valid path.

For example: (for our better understanding, I have given suitable examples)

- the sign of an (enzyme, flux) edge is always positive because the enzyme catalyses the reaction;
- the signs of many (regulator, operon) edges are reported in EcoCyc.
- the presence/absence and functions of (metabolite, regulator) edges,
- the activities of valid paths.
- Other attributes are not given thus have to be inferred from data.
- A complete specification of the values of all attributes in the network is called a model configuration

8.21 Do KEGG and ENZYME contain the relevant information?

Yes. The reason is that both the databases provide related and relevant information. Kegg provides enzyme data with links to many metabolic pathways, gene, diseases which are also present in the enzyme databases thus sharing many similarities with each other.

8.22 Which principle approaches towards metabolite network simulation do you know?

8.23 Sketch a conceptual model representing the interaction of the entity type “substrate” with the entity type “enzyme”

8.24 What types of “pathway databases” can be distinguished? EXAM

8.25 What is Biopax? EXAM

- Biopax – Opensource project → Biopax.org
- Metabolic pathway – small molecules. Enzymes
- molecular interactions – pair-pair interactions
- Technology – Yeast two hybrid, Pooldown
- PPI – Amp – that interactions are stable
- For Signalling pathways – tarnsient interactins

What are the 5 classes of Biopax?

- Entity – Pathway, Interaction, Physical entity, gene
- Relationship – object property, edge of graph level
- Interaction – sort of triple, interaction with nodes
- Aphysical entity or gene is a node in graph
- pathway subclass – series of interactions forming a network

9. Chemical databases

9.1 List at least 5 attributes of a chemical compound.

Chemical formula, 3D structure, names (recommended and alternative), charge, reactivity, molecular mass, ...

OR

Compound ID

Substance ID

Bioassay ID

Molecular weight

Molecular name

9.2 Why is PubChem distinguishing between Substance and Compound?

> Compounds are non-redundant structure-based data, whereas Substances are instances of Compounds, i.e. there is a 1 to n relationship between Compound and Substance.
>
> This allows to have a non-redundant database of chemical compounds and also the information about all the substances submitted is retained, giving maximum coverage, while at the same time guaranteeing high accuracy for the curated part.
>
> OR

ubstance information is electronically submitted to pubchem by depositors. It has information from particular depositor, and information on substances such as natural products extract which may not have associated chemical structure information.

But compound information is non redundant set of standardized and validated

chemical structure.

9.3 Give at least three examples for BioAssays that are of relevance for PubChem

BioAssays in PubChem contain information about the activity of chemical compounds, descriptions of the experimental conditions, protocols of the experiments.

9.4 Explain the workflow of a chemical similarity search

DBs such as PubChem or BRENDA provide graph drawing tools that allow users to draw chemical structures they are interested in. These drawings are internally converted into a non-graphical representation of the chemical compound which is matched to entries in the database. Similar entries (similarity is based on a pre-defined distance measure based on a rule set of chemical knowledge to judge functional and structural similarity) are then returned to the user.

9.5 Which part of PubChem comprises non-redundant structure information?

PubChem Compound.

9.6 What is open access and how does open access relate to PubChem?

Open access is a recently established publishing paradigm under which authors have to pay for the costs related to their publications themselves as opposed to user paying when retrieving the information. Open access data is hence non-proprietary and accessible from everywhere by everybody. PubChem is based on this paradigm, i.e. the data it contains is accessible free of charge.

OR

Open access consists of Links and resources useful for those interested in persuing open access publications or advocating open access to others in the academic community, government bodies.

Open Access and pubchem

- supports NHI's pubchem database
- open access working group defends information resource of pubchem.

9.7 Give a short explanation, why chemical information is more proprietary than biological information.

Chemical information can often be exploited for economic purposes. Drugs mean a lot of money; hence people are generally less generous with information regarding to chemicals of all kinds than with biological information which is (usually) not directly linked to financial profits (though it might be on the long run).

9.8 Provide a short summary on the building principles of the ChEBI ontology

The ChEBI ontology provides a structured provides means for the structured classification of biochemical entities. It's essentially a DAG, but has some inherently cyclic relationships are included.The ontology is divided into four subbranches: Molecular structure, biological role, application and subatomic particle. On the front- end site, it provides two different views: Parent-children based and tree-based. Within the ontology there are unchecked and checked items, the latter have been reviewed by curators, while the earlier have to be regarded as preliminary.

OR

Molecular structure

Molecular entities are classified according to their structure and function

Eg:-hydrocarbons

Biological role Classify entities on the basis of role with a biological content.

Eg:-antibiotic

Application ontology Classify entities on intended use by humans

Eg:-pesticides

Subatomic particle Classify particles smaller than atoms.

Eg:-electron

9.9 How many types of relationships does the ChEBI ontology use?

Nine:

- is a
- is part of
- is conjugate base of
- is conjugate acid of
- is tautomer of
- is enantiomer of
- has functional parent
- has parent hydride
- is substituent group from

9.10 Which are the major source databases for ChEBI?

The main sources are IntEnz, the integrated relational enzyme DB at the EBI following the nomenclature and classification guidelines by the IUBMB, and KEGG Ligand, a composite DB including a collection of biochemical compounds.

OR

– Int Enz (Integrated relational enzyme database of EBI)

- KEGG (Kyoto encyclopedia of genes and genomes ligand database).

- MSD (MSD database developed by MSD group of EBI ligand and small molecule dictionary)

9.11 Do you know commercial databases that provide information similar to ChEBI and PubChem?

- *CAS – Chemical Abstracts Service*.
- ACD/Labs.
- Ligand-protein DB.
- DIP.
- SABIO-RK
- Bio Models database

9.12 List the major entity-types and attributes that a ChEBI entry shows

Name, ID, formula, charge, mass, ontological annotation, IUPAC names, synonyms, CAS registry number, DB cross-references.

OR

Entity:

- Atom
- Molecule
- Ion
- Radical
- Complex

Attributes:

- chEBI identifier (name, ASCII name)
- Structure
- Formula
- Mass
- Charge
- Database links
- IUPAC name

9.13 Give short explanations and examples for:

InChi:

InChi is the new international chemical identifier, which was developed by IUPAC and NIST as a new way of describing chemical structures in text. It is derived directly from chemical structure. Composition, connections and stereochemical information can be represented, but not 3D structure. OR:

- IUPAC international chemical identifier
- String of characters capable of representing a chemical substance.
- Eg:-1/H20/h1H2/p-1

Example: Caffeine

InChI=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChIKey=RYYVLZVUVIJVGH-UHFFFAOYAW

SMILES:

SMILES is the simplified molecular input line entry specification. It includes connectivity but disregards 2D and 3D structure, it also leaves out hydrogens for simplicity. OR Simplified molecular input line entry system is a line notation for entering and representing molecules.

- It is a subset of smart

-Eg:- ethanol CCO Acetic acid CO(=O)O Example: Ethane

Formula:

Gives number of all the atoms in the compound, also provides with some minor information about the bonding between the atoms. OR

- for compound containing discrete molecules this is generally a molecular formula
- It is according to relative molecular mass or structure

Ethane: C2H6 ?

Caffeine: C8H12N4O3

IUPAC:

IUPAC is the international union of pure and applied chemistry. It's the major driving force behind InChi. It provides standards for nomenclature not only for biochemical compounds, but for all kinds of chemicals OR:

- International union of pure and applied chemistry
- Address global issues involving the chemical sciences.
- Contribute to the application of chemistry in the service of mankind
- Eg glyceraldehye-3-phosphate

Caffeine IUPAC name: 1,3,7-trimethylpurine-2,6-dione hydrate

IUPAC name: Ethane

9.14 How many compounds are typically screened in a high throughput screening in the pharmaceutical industry? Compare this number with ChEBI and PubChem entries.

Robotic facilities exist which can test up to 100000 compounds per day! ChEBI currently contains almost 14000 annotated entries, PubChem contains some 38 million substances, 18 million unique compounsd and 710 assays. Consequently, ChEBI evidently does not account for the vast amount of chemical information available today, yet being throughly curated, the information contained is very reliable. The numbers of PubChem entries convey the impression that it reflects state-of-the-art knowledge much more exhaustively, yet one must realize that BioAssays in PubChem are not designed to account for high-throughput screening, but rather to describe assays and the contained compounds.

9.15 Compare the high level structure of PubChem with the conceptual design of ChEBI! Exam 2008

9.16 Small molecule databases? Difference between two small molecules. EXAM