Biomedical Data Bases

Saul Pierotti

January 25, 2020

Introduction

- A database is an organized collection of data
- A DBMS is a piece of sofware that allows DB implementation and data mining
  - It allows data storing, retrieval, to perform backups, to maintain security
- Curation is a synonim of manual annotation
- In the last few years there has been a big effort to standardize DBs
- Even in the best DBs there are errors (!)
- Integration among DBs happens with corss-linking and with merging of DBs
  - Uniprot is an example
- Not all the structures available are on PDB (!)
  - Pharmaceutical companies have their own provate DBs
- Some DB terms
  - A record is a DB entry containing different fields
  - An accession key is a unique identifier of a record
  - A table is a DB file containing many records
    - Different tables can be connected by the same accession key for some records
  - A query is a data request submitted to a DB
- A query can be organized with boolean operator, in order to retrieve the desired result
- The schema of a DB is the logical structure of the data
  - A schema can be written in a flatfile, XML, ecc.
- The instance is the set of actual data
- The schema allows the interpretation of the instance
- Pubmed is based in Betsheda, at the National Library of Medicine
- Many DBs have built-in a way to retrieve data from the command line in a structured way
  - Entrex Direct allows to retrieve results from Entrez in the Unix command line
- The serch engine of PubMed uses automatic term mapping: if I do not specify the search filed, it tries to guess the correct one for each term in the query
- It first looks for the query in the MeSH Translation Table, then for the Journal translation table and then for an Author Index
- MeSH (Medical Subject Headings) is a list of common search terms that are classified in the correct context
  - If I search MeSH directly for a term, it gives me a list of possible meanings
  - Topics in MeSH are organized in a hierarchical wayH
  - I can limit my search to some topics
- Some of the main servers that we will use
  - NCBI (USA) hosts PubMed
  - SIB (Switzerland) hosts the tool ExPASy
    - In ExPASy, we can find SwissModel, UniProt, T-Coffee, LALIGN
    - T-Cofee is a really powerful alignment tool from Prof.Cedric Notredame, who maybe will give us an advanced lecture
  - EBI-EMBL (Germany and England)

    - It started in Heidelberg and then was transferred to Hinxton
    - They provide a lot of training schemes, and freely available data
    - Ensemble is a genome browser from EBI
    - It hosts also InterPro
- Elixir is an european platoform for bioinformatics data resources and tools
  - Casadio is on the board of directors of Elixir

Uniprot

- Uniprot is composed of entries from Swiss-Prot, PIR and TrEMBL, and from external sources
  - PIR was founded but Margaret Dayhoff, the creator of PAM matrices and of the first protein sequence atlas
  - SwissProt was part of SIB and it is the most well-curated part
  - TrEMBL are automatic translations of EMBL DNA sequences
  - It became a central hub for protein information
  - It offers extensive cross-linking
- Uniprot is itself composed of different parts
  - The UniProt Archive (UniParc) is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoided such redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI) making it possible to identify the same protein from different source databases. A UPI is never removed, changed or reassigned. UniParc contains only protein sequences. All other information about the protein must be retrieved from the source databases using the database cross-references. UniParc tracks sequence changes in the source databases and archives the history of all changes. UniParc has combined many databases into one at the sequence level and searching UniParc is equivalent to searching many databases simultaneously.
  - The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.
  - The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records in order to obtain complete coverage of the sequence space at several resolutions while hiding redundant sequences (but not their descriptions) from view. Unlike in UniParc, sequence fragments are merged in UniRef: The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records.
- When we publish something using a database, we need to specify the release that we used (!)
- SwissProt is composed of proteins manually curated entries from TrEMBL
  - The first step of curation is to asses if there is evidence at the protein level
  - GO-terms (and other ontologies), references, isoforms, PTMs (post translational modifications), polimorphisms, crosslinks are added
  - If available, the structure of the protein is reported, modelled or experimental
  - The correct nomenclature is established and reported
- If different isoforms of the protein are identified, the curator searches for the reason of the difference and documents it
- Different isoforms can be due to alternative splicing, natural variations, alternative initiation sites, erroneous prediction, ecc.

- Gene ontology is a classification of a sequence in terms of biological process, molecular function and cellular component
- The TrEMBL part of UniProtKB is mainly redundant, while SwissProt is not-redundant
- How to recognize a SwissProt entry from a TrEMBL entry in Uniprot
  - The accession key for TrEMBL proteins is an alphanumeric string, and the entry name repeats the accession key
  - The accession key for SwissProt proteins is an alphanumeric string, and the entry name is not related to the accession key
  - In the result page, rewied proteins have a golden badge, non reviewed proteins have a blue badge
- In Uniprot we can customize the result list to show the information we need
- Entries in UniProt are huge (!)
- Each entry has an annotation score in 1 to 5
- Functional information is a bit sparse in the entry, it can be in many different fields
- There is also a comment field in the entry
- We did an exercise about how to do an advanced search on UniProt and download the result as an Excel file, that I cannot reproduce here
  - We basically have to use boolean terms for the search, select the needed columns for display and then download in the required file format
- To retrieve the IDs for other databases, we use the Retrieve/ID mapping section
  - Many databases have a mapping feature, but UniProt is the best place for doing ID mapping (!)
  - UniProt mapping can be done also from the Unix command line
- In order to facilitate the retrieval of information from an entry, UniProt creted a visual interface called feature viewer
  - This is required because of the high information density of UniProt entries, which makes it difficult to read all the information in a textual format
- Automatic annotation in TrEMBL uses data from SwissProt to infer information
  - It is composed of 2 main pipelines: Unirule and SAAS
- Unirule uses expert-curated rules
  - For each automatic annotation, we can see the set of rules that generated it
- SAAS uses machine learning for finding the best set of rules, based on sequence features
- HAMAP is used as a partially automated rule source in Unirule
  - It is effective when the protein is part of a well defined family
  - It is used in place of manual annotation when it gives a result at least as good as the manual one
  - It claims to be more error-free than manual annotation
- UniParc is the biggest non-redundant protein sequence database, it contains sequences from various sources
  - It is not annotated
- UniRef provides clusters of sequences from UniProtKB and selected UniParc entry
  - It can be accessed from any entry going to the "similar sequences" link
  - UniRef100 combines identical sequences with at least 11 residues
  - UniRef90 clusters UniRef100 sequences that share at least 90% sequence identity and 80% overlapping with the seed
  - UniRef50 in the same way clusters sequences from UniRef90
  - It is useful because I can use the seed as a representative of the cluster, and so reduce the number of entries that I analyse
    - It can make BLAST searches faster (!)
  - The seed is the longest or best annotated sequence
    - It is considered annotation score, if it is manually annotated, and the source organism
    - If the source organism is a model organism it is preferred
- All the changes that an entry witnessed are stored in UniProt in the "history" link
- In UniProt I can also find complete proteomes
  - They are a collection of all the entries from an organism with a completely sequenced genome
  - Reference proteomes are the ones with best annotated entries and usually belong to model organisms
  - They evolve with time, when new proteins are added

  - They involve a lot of manual curation
- There are 25000 genes, 100000 transcripts and 1000000 proteins in humans
- Uniprot can be accessed programmatically through a python library

Errors in databases

- For a long time sequences have been annotated on the basis of structural and sequence similarity
- This approach has its limits, since not always structural or sequence similarity correspond to similar function
- Burkard Rost published a paper where he says that enzyme function is less conserved than expected
- The current level of misannotation is unknown, and seems to be increasing
- The level of misannotation in enzyme superfamilies that comprise families with different functions is particularly high
- The analysis of the conservation of catalytic residues can improve annotation accuracy

Protein classification databases

- Classifing proteins reduces the amount of data we need to deal with, and allows us to make predictions about protein function
- Interpro and Pfam are databases of protein classification based on sequence
- Proteins can be classified in superfamilies, families, subfamilies
  - The deeper we go, the more we can infer functional features for the members
- A domain is an independently-folding unit of protein sequence which harbours a function
- Sequence features are active sites, binding sites and other charachteristics short sequences found inside domains
- A protein signature is a matemathical model build from multiple sequence allignments
  - Motives, fingerprint, profiles, HMM are signatures
- A signle motif, also called pattern, can be represented by a regular expression
  - Prosite is a DB of patterns
- A fingerprint is composed of multiple motives in a specific arrangement
  - PRINTS is a DB for fingerprints
  - Fingerprints are useful for differentiating subfamilies
- A profile is built from a multiple global allignment and consists of a position-specifc scoring matrix (PSSM)
- HMMs are complex algorithms capable of modelling the prbabilities of residue change, insertion and deletion
  - They are a really powerful protein signature
- The general workflow of building a signature is to do a MSA, generate a draft model, use it against the all Uniprot and develop a mature model
- InterPro is a meta-DB that gathers data about patterns, profiles, fingerprints, HMMs
  - It is an hub of protein classification and annotation
  - It also gathers data from UniProt and PDB
  - It is cited by uniprot in each entry
- Pfam families are made from a seed alignment manually curated, and then are refined by searching the whole uniprot
  - The mature MSA is automatic, while the seed is curated manually
  - A Pfam family is represented by an HMM
  - Superfamilies in Pfam are called clans
  - From 2016 Pfam started to use reference proteomes insted of the whole Uniprot, to decrese redundancy
- The number of protein folds is limited, and classifying protein based on their fold allows to understand their functional and structural relationships
- Topology schemes are useful for identifing the fold of a protein

- We can represent the topology of a protein by drawing its secondary structure as a flat diagram
- Structural classification DBs are SCOP and CATH/Gene3D
- It seems that we have discovered all fold types
    - No new folds have been reported in SCOP and CATH since 2010
- SCOP in itself is dead, but its legacy has been taken by SCOPe
    - The first level classification is alpha, alpha/beta, alpha+beta, all beta, small proteins
    - The classification is class>fold>superfamily>family>sequence
- SCOP2 is a different DB, made from the same people that made SCOP, that uses a different classification system based on networks
- CATH classifies structures from the PDB while Gene3D predicts the location of functional domains in available sequences (Uniprot) using informations from CATH
    - The first level classification is Mainly alpha - Alpha-beta - Mainly beta
    - The classification is class>architecture>topology or fold>homologous superfamily>sequence family
- ECOD is another DB that claims to classify many more domains than CATH and SCOP (it makes a finer classification)
    - Its classification is based on evolutionary relationships
    - It is a newcomer, we will see how it will go

X-ray crystallography

- Bragg's law: $2d \sin\theta = n\lambda$
    - $2d \sin\theta$ is the excess distance traveled by light when hitting a diffraction plane with distance d with an incidence angle θ
    - $n\lambda$ is an integer number of wavelenghts
    - If the additional distance traveled is an integer number of wavelengts, I have constructive interference and therefore a signal
    - The diffracted beam is deviated by 2θ from the incident beam
- The distance of a diffraction spot r from the expected incidence point of the primary beam on a detection screen placed at distance A from the crystal is $A \tan 2\theta$
- The phase of a wave is the distance of the first crest from a reference point
- Sychrotron produce X-rays with really high brillance, so I have an high signal-to-noise ratio
    - They allow to use smaller crystals
- The inverse Fourier transform is equivalent to passing the radiation through a lens
    - It does a Fourier synthesis
    - We do not have materials that can focus X-rays
- The X-rays are scattered by the electron cloud and never reach the nuclei, this is why we get electron densities
- We use crystals instead of single molecules because it amplifies the signal and reduces the noise
- The reciprocal space of a Fourier transform is a graph were each frequency is reported with a corresponding amplitude
    - I always get pairs of frequencies, for example 3 and -3, because I cannot discriminate the direction of a wave
- A 2d wave can be represented in the reciprocal space as points on the plane, were the combinations of x and y frequency is represented
- The coordinates in the reciprocal space are called h and k
- When I sum more 2d waves, in the reciprocal space I just put all the points deriving from them together
- The phase of the different frequencies refers to how they allign toghether
- If I apply a low-pass filter to the reciprocal space, I get back a less detailed image
    - High frequencies are responsible for fine details
- On the opposite, with an high-pass filter I get only the details but not the bulk image
- From a 3d Fourier transfor I get a 3d reciprocal space with coordinates hkl
- From rotating the crystal, I get a series of 2d images that allow to reconstruct the 3d reciprocal space
- The unit cell is the minimal translational repeating unit

- The asymmetric unit is the minimal rotational and translational repeating unit
- The unit cell can be organized as 1 of the 14 possible 3-dimensional Bravais lattices
- The space groups refers to the rotational simmetries of the unit cell
  - It is important for the crystallographer because they determine the rotations that I have to apply to my crystal during data collection
- The space group is represented as $P2_12_12_1$
  - P is the Bravais lattice
  - The numbers refer to the rotational simmetry axes
    - 2 means 180°, 3 means 120°, means 90°
  - The subscript refers to the screw axis
    - 1 means a traslation of 1 unit cell along the rotation axis
- The phase of the diffracted beams cannot be recovered from the diffraction map only
- In multiple isomorphous replacement (MIR) the crystal is immersed in an heavy atom solution, or an heavy atom is co-crystallized with the sample?
  - The addition of the heavy atom should not alter the space group (should be isomorphous)
  - The diffraction maps are collected both with and without the heavy atoms
  - This allows to recover the phases
  - At least 2 isomorphus derivatives must be used to determine univocally the phases
- In multiwavelenght anomalous dispersion (MAD) the X-ray fluorescence of atoms is used to recover the phases
  - Since X-ray fluorescence requires really specific wavelenghts, it can only be done in synchrotrons
  - The re-emitted ray has a predictable phase-shift that allows to recover the phases
- In molecular replacement (MR) heavy atoms are integrated in the structure of the protein
  - Selenocysteine is usually used in place of cystein
- The R factor is the ration between the difference between the observed and calculated electron density and the observed electron density
- R-free is calculated in the same way but uses a subset of data not included in the iterative optimization
  - It avoids overfitting

NMR

- It uses the same principle as MRI
  - A collection of nuclei is distributed on various energy levels dependting on their orientation relative to the magnetic field applied
  - Radiofrequency is applied and the nuclei are excited
  - The nuclei relaxate by releasing energy to the environment (spin-lattice relaxation) or to other nuclei (spin-spin)
  - The relaxation time depends on the molecular environment and produces a signal
- It requires the use of cryomagnet, made of superconductive materials kept at 4K
- The frequencency of a magnet (es. 600 Hz) refers to the frequency of oscillation of a proton in the field it generates
  - It is a way to express field strenght
  - Higher field strenght traslates to higher oscillation frequency
- Proteins are in solution, and need to be in high concentration
  - It cannot be used for membrane proteins
  - Also soluble proteins tend to precipitate in high concentration
- It cannot be used for big proteins (>80kDa)
  - Big proteins require big magnets
- It allows to determine also the structure of mobile regions
  - This is also used for studying protein folding
- 1D and homonuclear 2D NMR allows to reach up to 10 kDa
- Heteronuclear NMR (with C13, N15, H2) allows the determination of bigger proteins
- NMR structures dominate the PDB for small proteins

- The wavelength used is on the microwave range
- The conditions in which it is performed are more similar to those encountered in vivo
- The wavelenght is on the microwave spectrum
- In 1D NMR the relaxiation signals are averaged and then the Fourier transform is applied
  - This allows to recover the typical frequency of each group
- Chemical shift is the variation in magnetic proprieties of nuclei due to their electronic environment
  - The proton of an alifatic compound has a different spectrum than that of an hydrophilic environment
- Scalar coupling is the transfer of magnetization among nuclei through chemical bonds
  - It is measured by COESY spectra
  - It works up to 3 bonds apart
  - If I see a signal between 2 nuclei, I know that they are at most 3 bonds apart
- Dipolar coupling is the interaction of nuclei across space
  - It is measured by NOESY
  - It gives constraints about distance between nuclei
- All these information gives us restraints on dihedral angles, distances
- Structures are determined via simulated annelaing
  - I start with a random model
  - I try to optimize the restraints and iterate
- Since I do simulated annealing, I get different models for each random starting model
  - The models are collected in a bundle in the PDB file
- Model variability can suggest flexibility or uncertainty
- The dynamics of individual atoms can be checked by specific NMR experiments
- An NMR structure is an average of many models
- Sometimes also the single models in the bundle are deposited

CryoEM

- The resolution has increased enormusly in the last 20 years
  - We went from 25 Å in 2005 to 2 Å in 2015
- Lenses for an electron beam are electromagnets
- What allowed a great increase in resolution is cooling the sample, so that it does not get damaged by the electron beam
- We started also to use very low intensity of the beam in order to avoid damage, but this lowers the signal quality
  - In order to increase quality we take many images and average them
- The cooling is done very fast so that water solidifies in an amorphus form
- The microscope sees many different molecules, all oriented randomly
- An algorithm clusters the images by orientation and averages the ones in the same orientation and conformation, so to increase quality
- From the averaged images, I can obtrain a 3d structure
- In the last years a series of small improvements incresed significantly resolution
- CryoEM gives you an electron density (!)
- Size does not matter (!)

PDB file

- The B factor is the mean square deviation of the position, so it is measured in Å^2
- There are several parsing tools for PDB files
- Now data are deposited also in xml format
- RSR (real space R value) measures the fit between a residue and the data
- RSRZ is a normalized RSR relative to residue type and resolution
  - If RSRZ is >2 we have an outlier
- The clashscore refers to atoms bumping into each other

    - I have a clash when 2 atoms are closer that the sum of Van der Waals radii plus a margin
- The RSR-Z (real space r value)
- Poor ranking does not necessarily mean bad quality!

Chimera

- To select a residue I use `select #model:residue@atomtype`
- There is no need to put and among atoms
- To put more that 1 condition, I put &
- Not is done with ~

NCBI

- It is the major bioinformatics hub in the US
- It resides in Bethesda, MD
- It is part of the NIH (National Institute of Health)
- There is much training material on NCBI, such as books and tutorials
- From the homepage, I can perform a search on all the NCBI databases
    - I get a page that redirects me to results in the different databases
- Genbank is the analogue of the ENA (european nucleotide archive)
    - It is a database of genetic sequences
    - The annotation is provided by the sibmitter of data
    - It collects a huge amount of information
- NCBI, EBI and DDBJ formed a consortium called INSDC (international nucleotide sequence database collaboration)
    - They exchange data daily
    - Genbank is the most commonly used portal
- Genbank is not a database, but a portal containing many databases
    - Sequences are not annotated and only updated by submitters
- Sequences are reported with the Genbank flatfile
    - There is an accession number, which is unique
    - A version for the record
    - The GI code is the old accession number format, but it is still present in old entries
        - New entries do not have it!
    - There is a features section
        - It specifies source organism, coding sequences (CDS), protein ID
        - The protein ID entry contains an in silico translation of the CDS
    - The property section collets various properties like source database
- Genebank is organized in 12 traditional divisions and bulk divisions
    - The traditional divisions collect different groups of source organisms and they tend to be well annotated
    - The bulk divisions collect data like EST (expressed sequence tags), and they are less accurate
- Refseq is a collection of reference sequences
    - It is non redundant and contains genomic DNA, mRNA, proteins
    - Entry from Genbank are reviewed and then migrated to Refseq
    - It is like SwissProt for TrEMBL
    - Refseq cannot be searched directly, its entries are inside nucleotide, gene, protein
- NCBI is a bit of a mess for how it is organized, there are collection, databases, things are not so intuitive
- Accession number prefixes in Refseq are widely used also in other databases
    - NM_123456 means mRNA
    - NP_123456 means protein
    - NR_123456 means non coding
    - NW_123456 is a genome

    - XP, XM, XT means the predicted counterpart of NP, NM, NR
- Refseq curation is carachterized by status codes: provisional, validated, reviewed
- Gene
- OMIM is a db of disease genes or diseases
    - It is the reference database on the topic
- dbSNP contains mainly SNPs but also short indels
    - The official name has changed but the acronym was retained
    - Single SNPs have an ss# accession
    - Identical ss# entries are reviewed and consolidated in one rs# entry* BLAST is also hosted in NCBI
    - There is BLASTp, BLASTn, BLASTx, psiBLAST,
    - psiBLAST uses a position-specific scoring matrix
        - This derives from a first MSA generated from a normal BLAST

HMMER

- It is an implementation of HMM algorithms for the search of distant homologs It is an alternative of psiBLAST
- It was slow, but now it is almost as fast as BLAST

Ensemble

- The annotation is first performed automatically for the whole genome (hence the name Ensemble)
- Manual annotation thorugh VEGA and Havana
    - These projects focus solely on vertebrate model organisms
- Gold color is for when Havana and Ensembl agree on a transcript
- Red transcript can be from only Havana, or only Ensembl
    - Their number starts with 20_ for Ensembl and 00_ for Havana
- Blue refers to non-conding transcripts
    - They are only Havana, we do not have tools to detect them automatically (!)
- Green is for genes in the Consensus CDS protein set (CCDS)
    - It means that the gene is consistently predicted by NCBI, EBI, UCSC, Sanger
- Genese in the forward strand are shown above the contig, genes in the reverse strand are below it