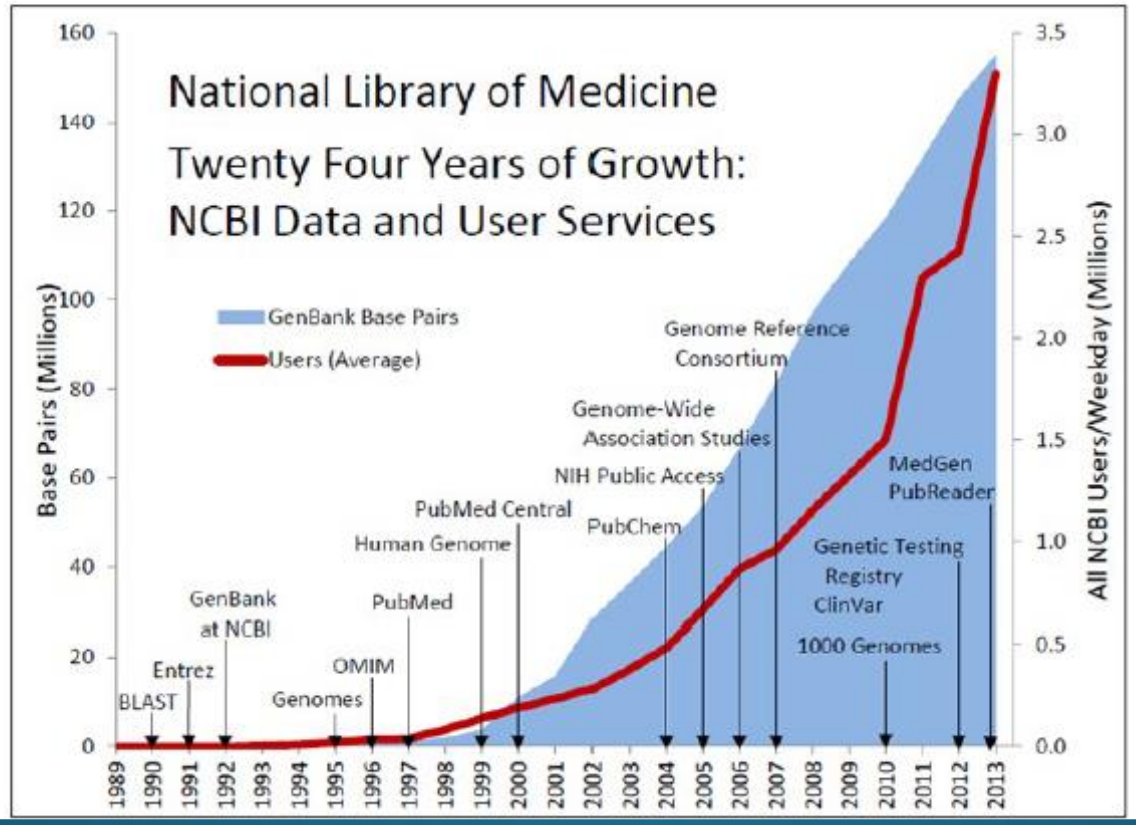
What is Bioinformatics?
National Center for Biotechnology (NCBI) definition

"Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned"

Dr. Manu S Singh

Bioinformatics

Bioinformatics: A rapidly growing discipline



An introduction to biological databases

What is a database ?

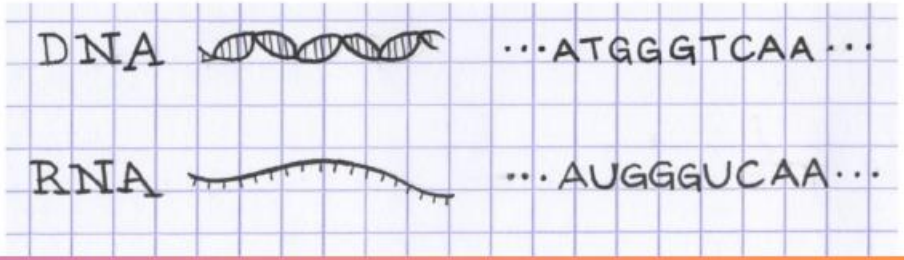
- A collection of
  - structured
  - searchable (index)
  - updated periodically (release)
  - cross-referenced data
- Includes also associated tools (software) necessary for db access/query, db updating, db information insertion, db information deletion....

Why biological databases ?

- Data (genomic sequences, 3D structures....) are no longer published in a conventional manner, but directly submitted to databases.
- Update frequency: daily to annually

The strings of life

- DNA – 3 billion letters (2m. long) in human, size of the alphabet: 4
- RNA – size of the alphabet: 4
- Proteins – size of the alphabet: 20



DNA Sequencing

- DNA sequencing technology enables us to identify the sequence of letters (called nucleotides) that make up the DNA string

In 2001 the first draft of the human genome was released

Differences in DNA make us different

```
Reference genome:   ...ACCGTTACGCGAAAG...

Individual A:       ...AGCGTTACGCGAAAG...

Individual B:       ...ATCGTTACGCGAAAG...

Individual C:       ...ATCGTTA---GAAAG...

Individual D:
...ATCGTTACGCGAAAG...ACCGTTACGCGAAAG...
```

What can we do with this information?

Some applications:

- 1 understand the molecular bases of human variation
- 2 identify the regions of DNA that are highly conserved across healthy individuals (and therefore potentially disease-causing or lethal if mutated)
- 3 filter out common variants that are unlikely to be associated with disease (to increase the power of genome-wide association studies)
- 4 identify population-specific rare variants

But....
Data is only useful if we have the conceptual framework and practical tools to interpret it!

Bioinformatics Databases

The ten important bioinformatics databases

GenBank/DDJB/EMBL	www.ncbi.nlm.nih.gov	Nucleotide sequences
Ensembl	www.ensembl.org **	Human/mouse genome
PubMed	www.ncbi.nlm.nih.gov	Literature references
NR	www.ncbi.nlm.nih.gov	Protein sequences
Swiss-Prot	www.expasy.org	Protein sequences
InterPro	www.ebi.ac.uk	Protein domains
OMIM	www.ncbi.nlm.nih.gov	Genetic diseases
Enzymes	www.expasy.org	Enzymes
PDB	www.rcsb.org/pdb/	Protein structures
KEGG	www.genome.ad.jp	Metabolic pathways

Database 1: nucleotide sequences

- Main nucleic acid sequence databases are –

NCBI **database** (www.ncbi.nlm.nih.gov/)

European Molecular Biology Laboratory (EMBL) **database** (www.ebi.ac.uk/embl/

DNA Database of Japan (DDBJ) **database** (www.ddbj.nig.ac.jp/)

« different views of the same data set »

Ideal minimal content of a sequence database entry

- Sequences !!
- Accession number (AC) (unique identifier, specific to a database)
- Taxonomic data
- References
- ANNOTATION/CURATION
- Keywords
- Cross-references
- Documentation

NCBI

The NCBI entry for an accession contains a lot of information about the sequence, such as papers describing it, features in the sequence, etc.

The ‘DEFINITION’ field gives a short description for the sequence.

The ‘ORGANISM’ field in the NCBI entry identifies the species that the sequence came from.

The ‘REFERENCE’ field contains scientific publications describing the sequence.

The ‘FEATURES’ field contains information about the location of features of interest inside the sequence, such as regulatory sequences or genes that lie inside the sequence.

The ‘ORIGIN’ field gives the sequence itself.

Examples of searches, some of them made by combining search terms using “AND”:

Typed in the search box	Searches for sequences:
NC_001477[AC]	With accession number NC_001477
Nature[JOUR] AND 460[VOL] AND 352[PAGE]	Published in *Nature* **460**:352–358
“Chlamydia trachomatis”[ORGN]	From the bacterium *Chlamydia trachomatis*
“Berriman M”[AU]	Published in a paper, or submitted to NCBI, by M. Berriman
flagellin OR fibrinogen	Which contain the word ‘flagellin’ or ‘fibrinogen’ in their NCBI record
“Mycobacterium leprae”[ORGN] AND dnaA	Which are from *M. leprae*, and contain “dnaA” in their NCBI record
“Homo sapiens”[ORGN] AND “colon cancer”	Which are from human, and contain “colon cancer” in their NCBI record
“Homo sapiens”[ORGN] AND malaria	Which are from human, and contain “malaria” in their NCBI record
“Homo sapiens”[ORGN] AND biomol_mrna[PROP]	Which are mRNA sequences from human
“Bacteria”[ORGN] AND srcdb_refseq[PROP]	Which are RefSeq sequences from Bacteria
“colon cancer” AND srcdb_refseq[PROP]	From RefSeq, which contain “colon cancer” in their NCBI record

NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Vi

Search across databases Chlamydia trachomatis[ORGN] GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

11689	**PubMed:** biomedical literature citations and abstracts	129	**Books:** online books
4716	**PubMed Central:** free, full text journal articles	5	**OMIM:** online Mendelian Inheritance in Man
5	**Site Search:** NCBI web and FTP sites		

35429	**Nucleotide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
none	**EST:** Expressed Sequence Tag records	none	**UniGene:** gene-oriented clusters of transcript seq
148	**GSS:** Genome Survey Sequence records	none	**CDD:** conserved protein domain database
29670	**Protein:** sequence database	none	**UniSTS:** markers and mapping data
22	**Genome:** whole genome sequences	111	**PopSet:** population study data sets

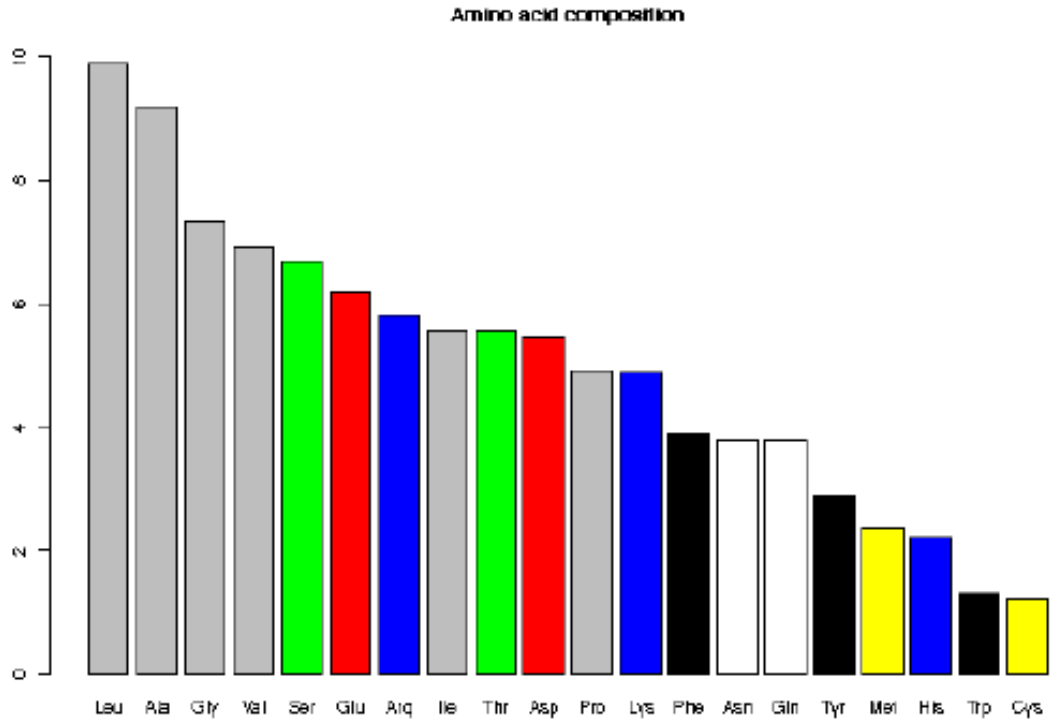
Database 2: Protein

- SwissProt
- Expasy

AMINO ACID COMPOSITION

5.1 Composition in percent for the complete database

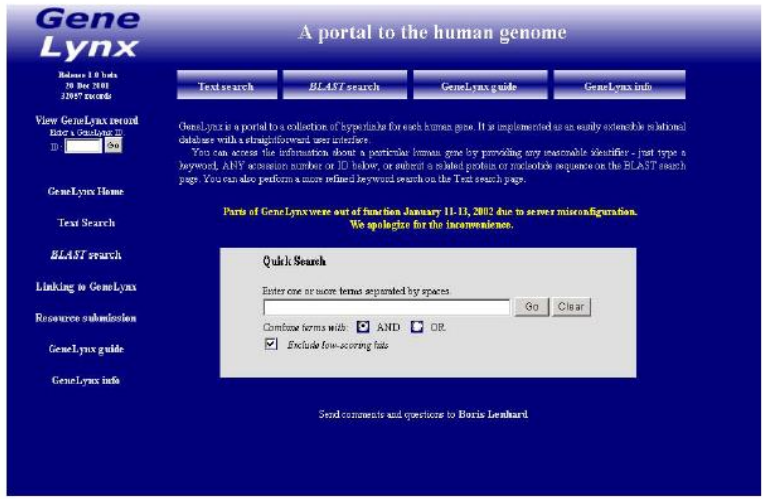
Ala (A) 9.20	Gln (Q) 3.77	Leu (L) 9.90	Ser (S) 6.68
Arg (R) 5.83	Glu (E) 6.20	Lys (K) 4.89	Thr (T) 5.54
Asn (N) 3.77	Gly (G) 7.34	Met (M) 2.35	Trp (W) 1.30
Asp (D) 5.48	His (H) 2.20	Phe (F) 3.89	Tyr (Y) 2.88
Cys (C) 1.23	Ile (I) 5.56	Pro (P) 4.93	Val (V) 6.93



Databases 3: ‘genomics’

- Contain informations on gene chromosomal location (mapping) and nomenclature, and provide links to sequence databases; *has usually no sequence*;
- Exist for most organisms important in life science research; usually species specific.

- Examples: MIM, GDB (human), MGD (mouse), FlyBase (Drosophila), SGD (yeast), MaizeDB (maize), SubtiList (B.subtilis), etc.;



http://www.genelynx.org/

http://homophila.sdsc.edu/

Databases 4: mutation/polymorphism

- Contain informations on sequence variations linked or not to genetic diseases;
- Mainly human but: OMIA - Online Mendelian Inheritance in Animals
- **General db:**
  - OMIM
  - HMGD - Human Gene Mutation db
  - SVD - Sequence variation db
  - HGBASE - Human Genic Bi-Allelic Sequences db
  - dbSNP - Human single nucleotide polymorphism (SNP) db
- **Disease-specific db**: most of these databases are either linked to a single gene or to a single disease;
  - p53 mutation db
  - ADB - Albinism db (Mutations in human genes causing albinism)
  - Asthma and Allergy gene db
  - ....

Database 5: Protein family/ domain

Protein domain/family: some definitions

- Most proteins have « modular » structures
- Estimation: ~ 3 domains / protein
- Protein domains are ideally defined by a specific combination of secondary structures that fold into a characteristic three dimensional (3D) structure.
- Domains not only share a common structure but have also often a similar function that contributes to the global activity of the proteins which contain them.

Pattern-Profile

```
HPT1_HUMAN : NLTTGATLINEQWLLTTAKNA
ACRO_RABIT : YHAQGGVLLNAHWVLTAAHCS
KLKB_HUMAN : RFLQGGALLSGQWVITATHCL
MCT3_SHEEP : SYIQGGFLVREDFVLTAAHCP
TRB2_HUMAN : MHFQGGSLIHPQWVLTAAHCE
PRTC_HUMAN : KLAQGAVLIHPSWVLTAAHCA
EL2_MOUSE  : RHNQGGSLVANNWVLTAAHCH
HPT_CANFA  : NLTSGATLINEQWLMTTAKNV
VSP3_TRIFL : GALQGGTLINQEWVLTASHCL
TMS3_HUMAN : YHLQGGSVITPLHIITAAHCA
TRY2_RAT   : YHPQGGSLINDQWVVSAAHCF
MCT2_RAT   : RVIQGGFLISRQFVLTAAHCF
HPT_MUSSA  : GLTTGATLISDQWLLTTAKNN
TRY4_LUCCU : SHSQGGSVYNSRIIVTAAHCY
PLMN_MACMU : MHFQGGTLISPEWVLTAGHCN
```

- Pattern{LIVM]-[ST]-A-[STAG]-H-C

→ Yes or no

- Profile:

```
ID   TRYPSIN_DOM; MATRIX.
AC   PS50240;
DT   DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); DEC-2001 (INFO UPDATE).
DE   Serine proteases, trypsin domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=234;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=229;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.0169; R2=0.00836256; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=1134; N_SCORE=9.5; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=-1; SCORE=775; N_SCORE=6.5; MODE=1; TEXT='?';
MA   /DEFAULT: M0=-9; D=-20; I=-20; B1=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105;
MA   /I:       B1=0; BI=-105; BD=-105;
MA              A  B  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
MA   /M: SY='I'; M= -8,-29,-34,-26,  3,-34,-24, 34,-26, 19, 15,-24,-21,-21,-24,-19, -8, 25,-19,  3;
MA   /M: SY='N'; M=  0, 14, 10,  1,-22, -1,  6,-23, -4,-26,-17, 20,-14, -1, -6, 13,  2,-20,-34,-15;
MA   /M: SY='E'; M= -4,  4,  7, 14,-26,-13, -7,-23,  3,-22,-16,  2,  7,  3, -3,  2, -2,-21,-30,-18;
MA   /M: SY='R'; M=-12,  5,  5,  7,-23,-17,  3,-24,  8,-20,-12,  7,-16, 10, 12, -2, -6,-21,-27, -9;
MA   /M: SY='W'; M=-16,-33,-35,-27, 13,-22,-24,-11,-18,-13,-13,-31,-27,-20,-18,-30,-21,-18, 97, 25;
MA   /M: SY='V'; M=  1,-29,-31,-28, -1,-30,-29, 31,-22, 13, 11,-27,-27,-26,-22,-12, -2, 41,-27, -8;
MA   /M: SY='L'; M= -8,-29,-31,-22,  9,-30,-21, 23,-27, 37, 20,-28,-28,-21,-20,-25, -8, 17,-20, -1;
MA   /M: SY='T'; M=  2, -1, -9, -9,-11,-17,-19,-10,-10,-13,-11,  1,-11, -9,-10, 23, 43,  0,-32,-12;
MA   /M: SY='A'; M= 45, -9,-19,-10,-20, -2,-15,-11,-10,-11,-10, -9,-11, -9,-19, 10,  1, -1,-21,-18;
MA   /M: SY='A'; M= 40, -9,-17, -8,-21,  5,-18,-14, -9,-13,-12, -8,-11, -9,-16,  9, -2, -5,-21,-21;
MA   /M: SY='H'; M=-18,  0,  0,  1,-21,-19, 89,-29, -8,-21, -1,  9,-19, 11,  0, -7,-17,-29,-30, 16;
MA   /M: SY='C'; M= -9,-18,-28,-29,-20,-29,-29,-29,-29,-20,-19,-18,-39,-29,-29, -9, -9, -9,-49,-29;
MA   /I:       E1=0; IE=-105; DE=-105;
//
```

→ score/threshold

Databases : proteomics

- Contain informations obtained by 2D-PAGE: images of master gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Composed of image and text files
- There is currently no protein Mass Spectrometry (MS) database (not for long…)

Databases : 3D structure

- Contain the spatial coordinates of macromolecules whose 3D structure has been obtained by X-ray or NMR studies

- Proteins represent more than 90% of available structures (others are DNA, RNA, sugars, viruses, protein/DNA complexes...)

- PDB (Protein Data Bank), SCOP (structural classification of proteins (according to the secondary structures)), BMRB (BioMagResBank; RMN results)

  HSSP: Homology-derived secondary structure of proteins.

  SCOP: Structural classification of proteins

  CATH: hierarchical domain classification of protein structures derived from PDB.

- Future: Homology-derived 3D structure db.