Introduction to Database

Nociones Básicas de Bioinformática y Genómica

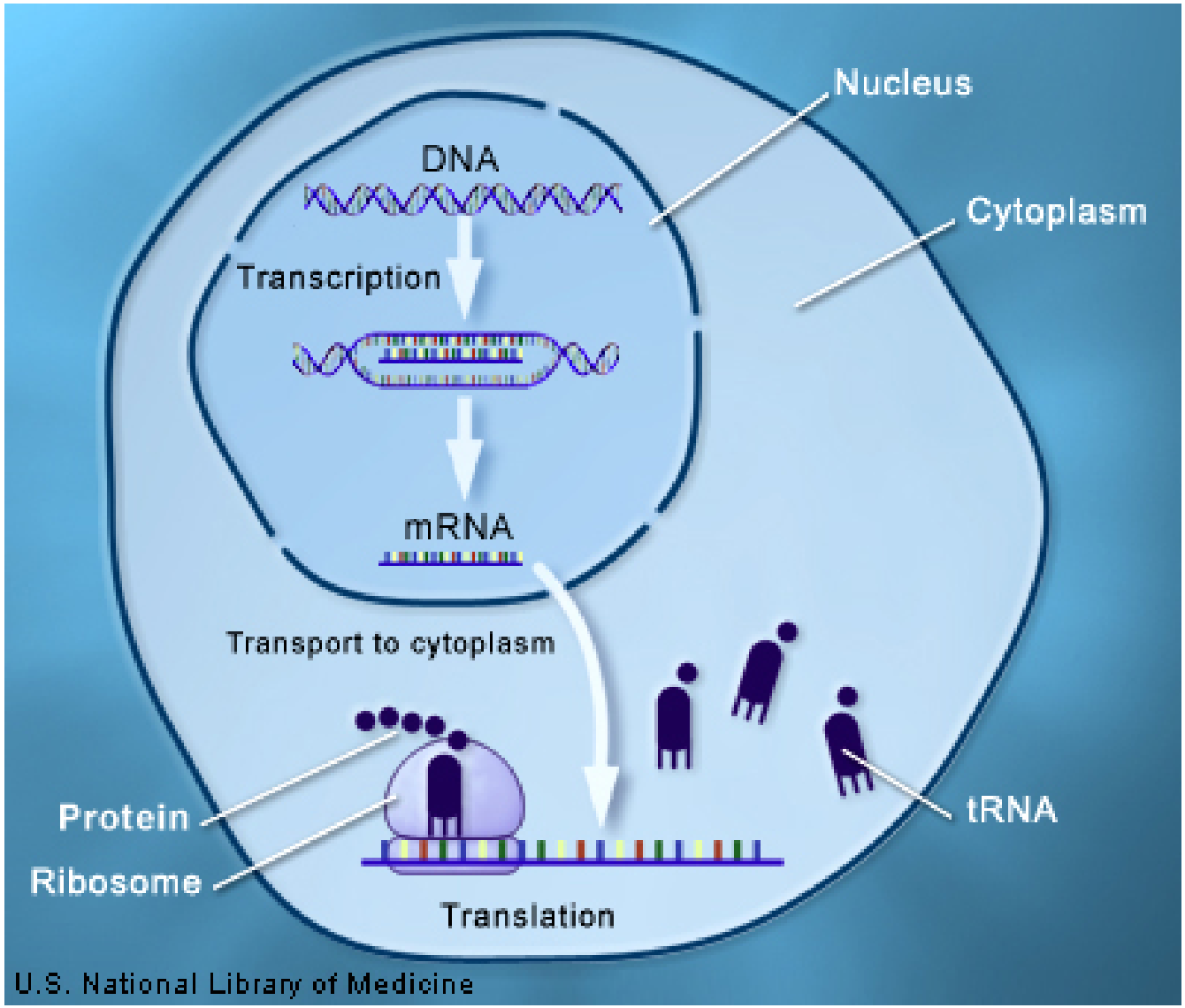
(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2015-02-16

Bioinformatics & Genomics

DNA $\rightarrow$ RNA $\rightarrow$ Protein $\rightarrow$ Biological Function



Why genomics Genomics

- DNA, RNA are interesting on their own.
- Proteins are the interesting functional features.
- DNA, RNA are measurable using **high throughput technologies** such as DNA microarrays and NGS.
- Protein changes can be *inferred* form DNA measurements.
- Protein levels can be *inferred* form RNA measurements.

List of biological databases (Wikipedia) I

1. **Primary nucleotide sequence databases**
2. **Metadatabases**
3. **Genome databases**
4. **Protein sequence databases**
5. Proteomics databases
6. Protein structure databases
7. Protein model databases
8. **RNA databases**
9. Carbohydrate structure databases
10. **Molecular interactions** (protein-protein)

List of biological databases (Wikipedia) II

11. Signal transduction pathway databases
12. **Metabolic pathway databases**
13. **Experimental data repositories** (Microarrays NGS, Sanger)
14. Exosomal databases
15. Mathematical model databases
16. PCR / real time PCR primer databases
17. **Specialized databases**
18. Phenotype databases
19. Taxonomic databases
20. Wiki-style databases
21. Metabolomic Databases

en.wikipedia.org/wiki/List_of_biological_databases

Primary nucleotide sequence databases

Contain any kind of nucleotide sequences, form genes to genomes.
The International Nucleotide Sequence Database (INSD)
Collaboration:

- GenBank
  National Center for Biotechnology Information (NCBI)
- European Nucleotide Archive (ENA)
  European Bioinformatics Institute (EBI)
- DNA Data Bank of Japan (DDBJ)

Primary nucleotide sequence databases: GenBank

- available on the NCBI ftp site:
  http://www.ncbi.nlm.nih.gov/Ftp/
- A new release is made every two months.
- 3 types of entries:
  - CoreNucleotide (the main collection)
  - dbEST (Expressed Sequence Tags)
  - dbGSS (Genome Survey Sequences)

Access:

- Search for sequence identifiers using Entrez Nucleotide:
  http://www.ncbi.nlm.nih.gov/nucleotide/
- Align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool):
  http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Several other e-utilities (see book)

See an example of a GenBank record.

Metadatabases

- Collect and organize data from *primary nucleotide sequence databases* and may other resources.
- Make the information available in a convenient format and provide data handling resources: web pages, application programming interface (API) ...
- Focus on particular species, diseases ...

Examples

- Entrez: searches through almost all NCBI resources
  http://www.ncbi.nlm.nih.gov/sites/gquery
  Queries can be saved if you have a a MyNCBI account
  http://www.ncbi.nlm.nih.gov/.
- GeneCards: provides genomic, proteomic, transcriptomic, genetic and functional information for human genes (known and *predicted*) http://www.genecards.org/

Genome databases

Collect genome sequences and *annotation* (specification about genes) for particular organisms, and try to improve them:

- Data *curation*.
- Complete missing information using *insilico* methods.
- Generate new relational organization.
- Complement feature IDs.
- Provide "easy" access, visualization ...

Examples

- Ensembl: automatic annotation on selected eukaryote genomes.
- UCSC Genome Browser: reference sequence and working draft assemblies for a large collection of genomes
- Wormbase: genome of the model organism C.elegans.

Genome databases: Ensembl

- Ensembl is a joint project between European Bioinformatics Institute (EBI) the European Molecular Biology Laboratory (EMBL) and the Wellcome Trust Sanger Institute.
- Develop a software system which produces and maintains automatic annotation on selected vertebrate and **eukaryote** genomes.
- http://www.ensembl.org

Genome databases: UCSC Genome Browser

- UCSC: University of California, Santa Cruz.
- This site contains the reference sequence and working draft assemblies for a large collection of genomes.
- http://genome.ucsc.edu/
- Complements / formats NCBI

Protein sequence databases

- Most times proteins are the final unit of interest to research.
- There is a direct conversion from DNA/RNA sequences to protein sequences.
- Gene IDs and protein IDs are equivalently used by researchers *(biologists not bioinformaticians)*

Examples

- UniProt: Universal Protein Resource (EBI)
- Swiss-Prot (Swiss Institute of Bioinformatics)
- InterPro Classifies proteins into *families* and predicts the presence of domains and sites.
- Pfam Protein *families* database of alignments and HMMs (Sanger Institute)

RNA databases

- Contain information about RNA molecules.
- Most of them regarding gene *regulatory factors*. (Gene information is usually in other repositories).

Examples

- Ensembl
- mirBase: microRNAs `http://www.mirbase.org/`
- TRANSFAC: transcription factors in eukaryote (Proprietary database).
- JASPAR: transcription factor binding sites for eukaryote (Open access, curated, non-redundant).
  `http://jaspar.genereg.net/`

Protein-protein interactions

- Proteins are the main functional units.
- But they do not work in isolation.
- *Pretty useless at the moment but promising in the future . . .*
- some information is *experimental*, but most of it is generated *insilico*.

Examples

- IntAct: protein–small molecule and protein–nucleic acid interactions.
- BIND: Biomolecular Interaction Network Database.

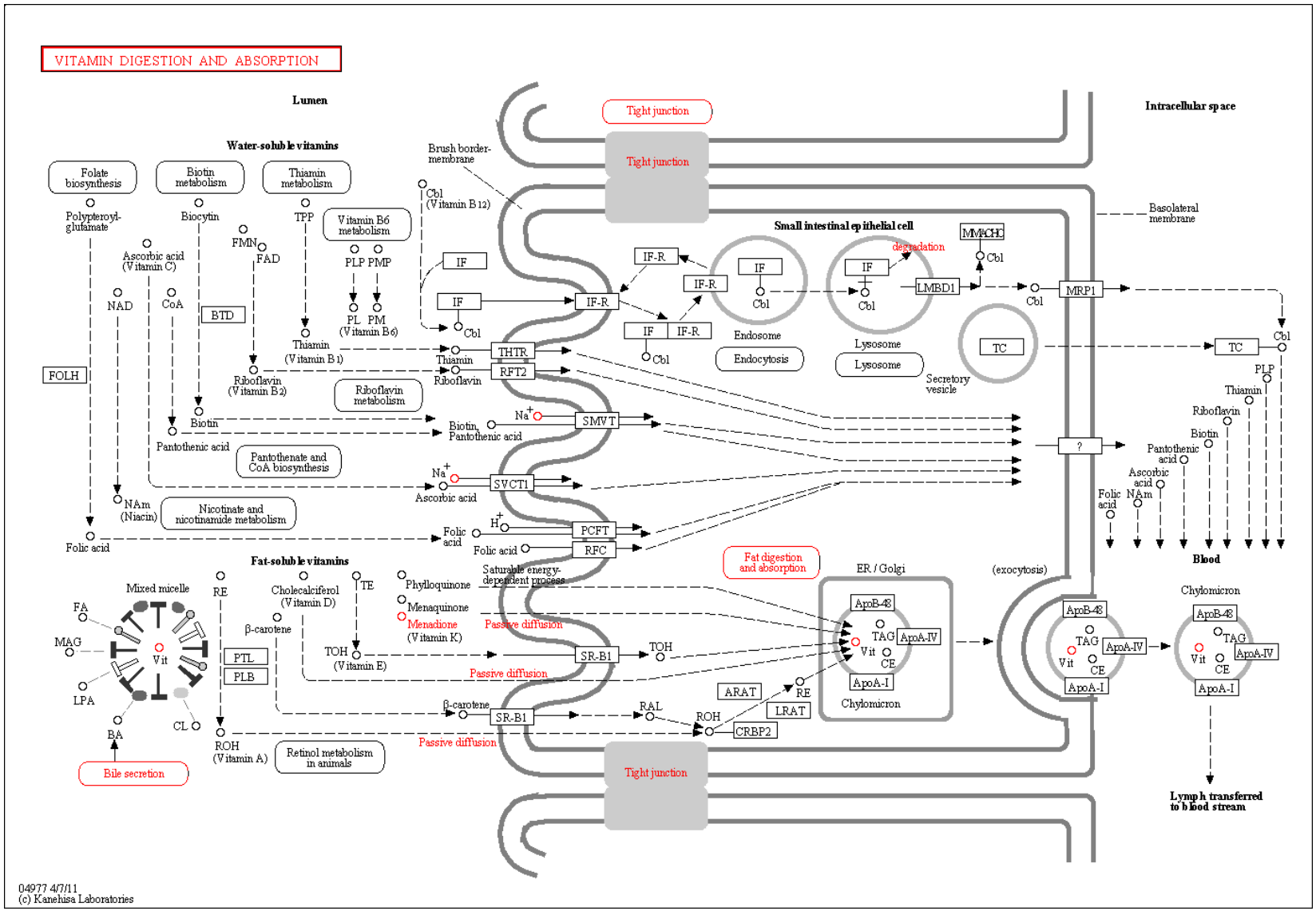
Signal transduction pathway databases & Metabolic pathway databases

- Information about how genes (or proteins) interact among them.
- not only physical interactions ...

Examples

- Reactome: free online database of biological pathways. http://www.reactome.org
- KEGG: Kyoto Encyclopedia of Genes and Genomes. Metabolic pathways. http://www.genome.jp/kegg/pathway.html

KEGG: Metabolic pathway databases



Experimental data repositories

Contain Microarray, NGS, Sanger, and other *experimental* high throughput data.

- GEO: Gene Expression Omnibus (NCBI) http://www.ncbi.nlm.nih.gov/geo/
- ArrayExpress: database of functional genomics experiments including (EBI) http://www.ebi.ac.uk/arrayexpress/
- The Cancer Genome Atlas (TCGA): Data on different cancer related tissues. http://cancergenome.nih.gov/

Specialized databases

- Gene Ontology (GO): standardizes the representation of gene and gene product attributes.
  http://www.geneontology.org/
- OMIM (Online Mendelian Inheritance in Man): Inherited Diseases http://www.ncbi.nlm.nih.gov/omim
- dbSNP: variations in any species and from any part of a genome. http://www.ncbi.nlm.nih.gov/projects/SNP/

Taxonomic databases

The most standard reference at the NCBI

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

Some other taxonomic information in customized databases:
http://www.arb-silva.de/

Wiki-style databases

- http://www.snpedia.com/index.php/SNPedia

Homework

- Quickly explore all the databases of a given category
- See what can be downloaded via FTP or similar
- Spot any interesting tool or related API

Due date: $\sim\sim$ 20 Apr.