Introduction to bioinformatics - 2016

Pongor Sándor

Introduction to bioinformatics - 2016

Course outline

- General intro , core data types, multiple data types (today)
- Core operations
- Sequence alignment
- Multiple alignment, databases
- Phylogenetics
- Systems biology, genomics, next generation sequencing 3 lectures
- Test Dec 13

Concepts, algorithms, databases

This lecture: The basic concepts

- 1. Defining the subject: Bioinformatics, Molecular Biology and Systems Biology
- 2. Theory of biomolecular data: a) Structure, Function, b) logical structure, standard, simplified and annotated descriptions, databases
- 3. Core data types in detail: sequences, 3D, networks/genomes and scientific papers.
- 4. Summary

Part 1

Defining the subject: Bioinformatics

What is bioinformatics?

- Narrow definition:

  Science of biological data. Mostly molecular biology data. Storage (management), analysis and interpretation (visualization) of data. Mostly static.

- Broad definition:

  Science of biological knowledge. All computer applications in (molecular) biology including modeling (simulation of behavior) Also includes dynamics

Note:

- Bioinformatics is historically linked to the “revolution”of DNA sequencing and protein 3D determination 1980’s
- Now broadly extended to all computer uses in biology

What is bioinformatics?

- Computational tools are fundamental in ALL branches of science
- Computer uses in biology:
  - Data management
    - Acquisition
    - Storage, annotation
    - Interpretation, analysis, data-mining
  - Modelling, simulation

Service → Data management

Research, „biocomputing” → Interpretation, analysis, data-mining

Narrow definition

Broad definition

Examples

- A scientist determines a gene sequence (experimental molecular biology, “wet lab”)
- Compares sequences with databases (bioinformatics)
- Predicts function of the gene, based on the comparison (bioinformatics)

- Instead of gene sequence, we can write genome, protein 3D structures, etc.

Why is bioinformatics important?

- "A paradigm shift in biology: from data collection to data processing …."

  Walter Gilbert, *Nature*, 1991

- "Biotechnology is the industrial use of biological information"

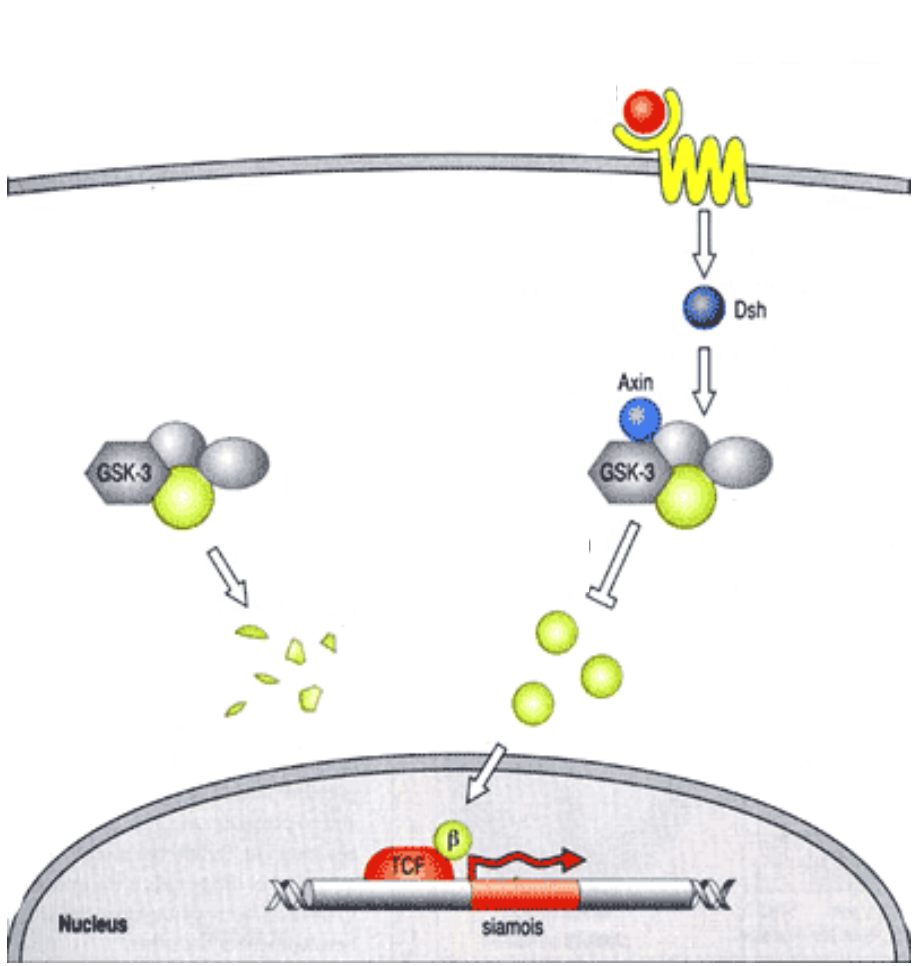
  Lee Hood, in *The Economist*, 1997

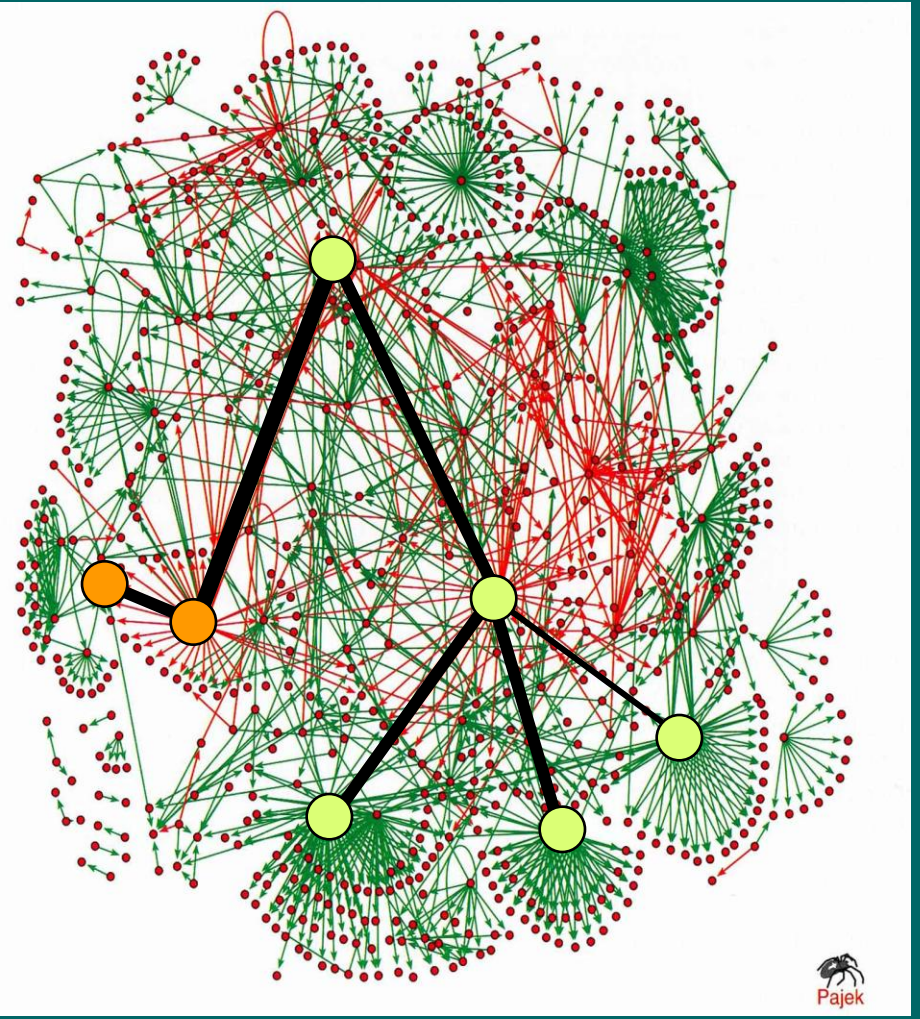
Today: Transition from simple objects to large systems

- Molecular biology/traditional bioinformatics studies single or a few objects.
- Modern experimental approaches can collect data from a large number of objects at the same time → “systems biology”
- Biological systems: a cell, a tissue, an organ, an organism…
- Typical examples: genomes

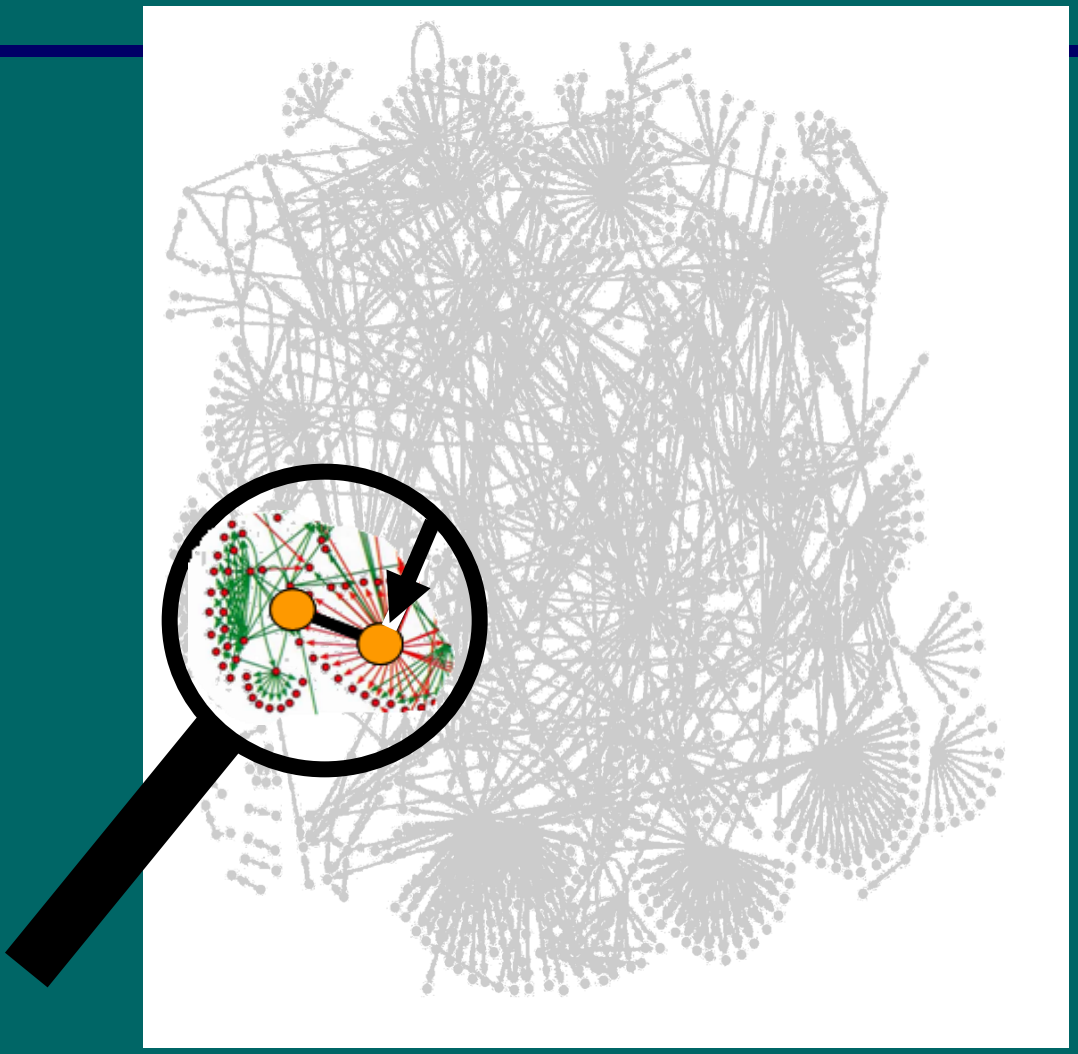
Molecular biology

Systems biology

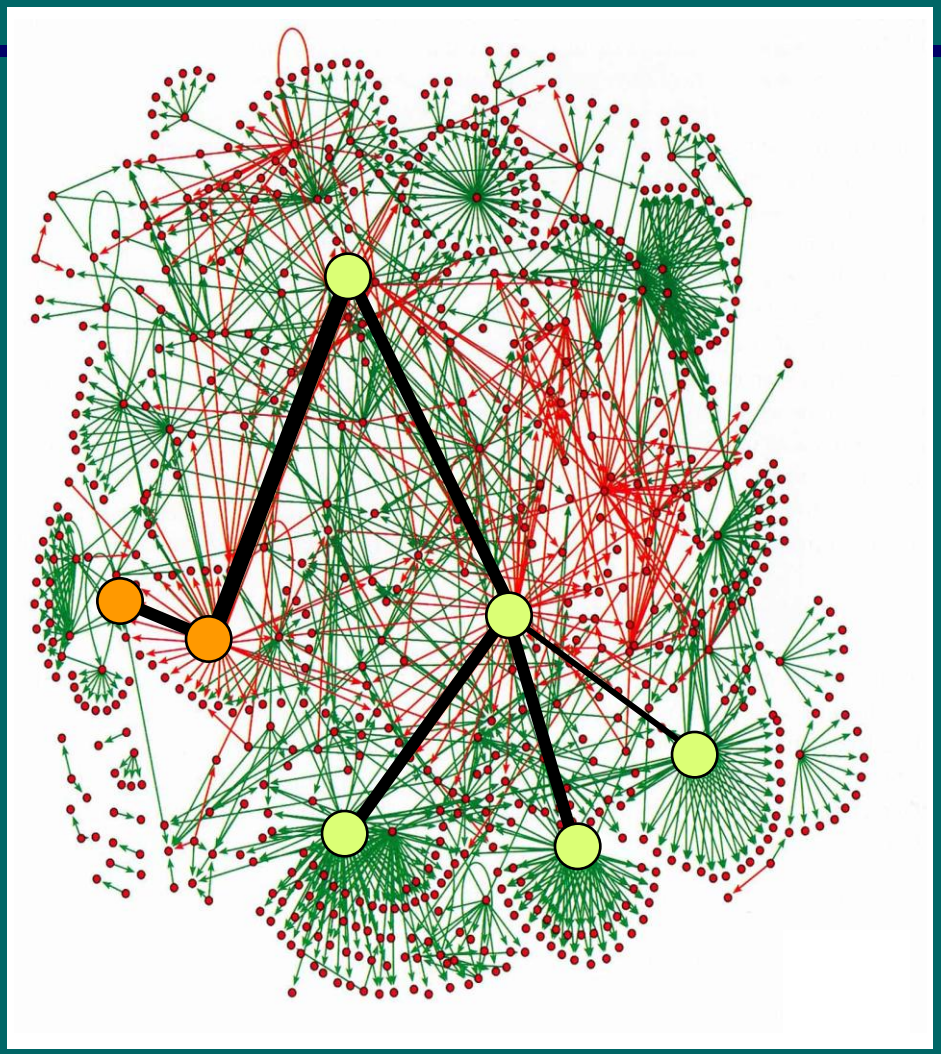




Approaches based on data collection

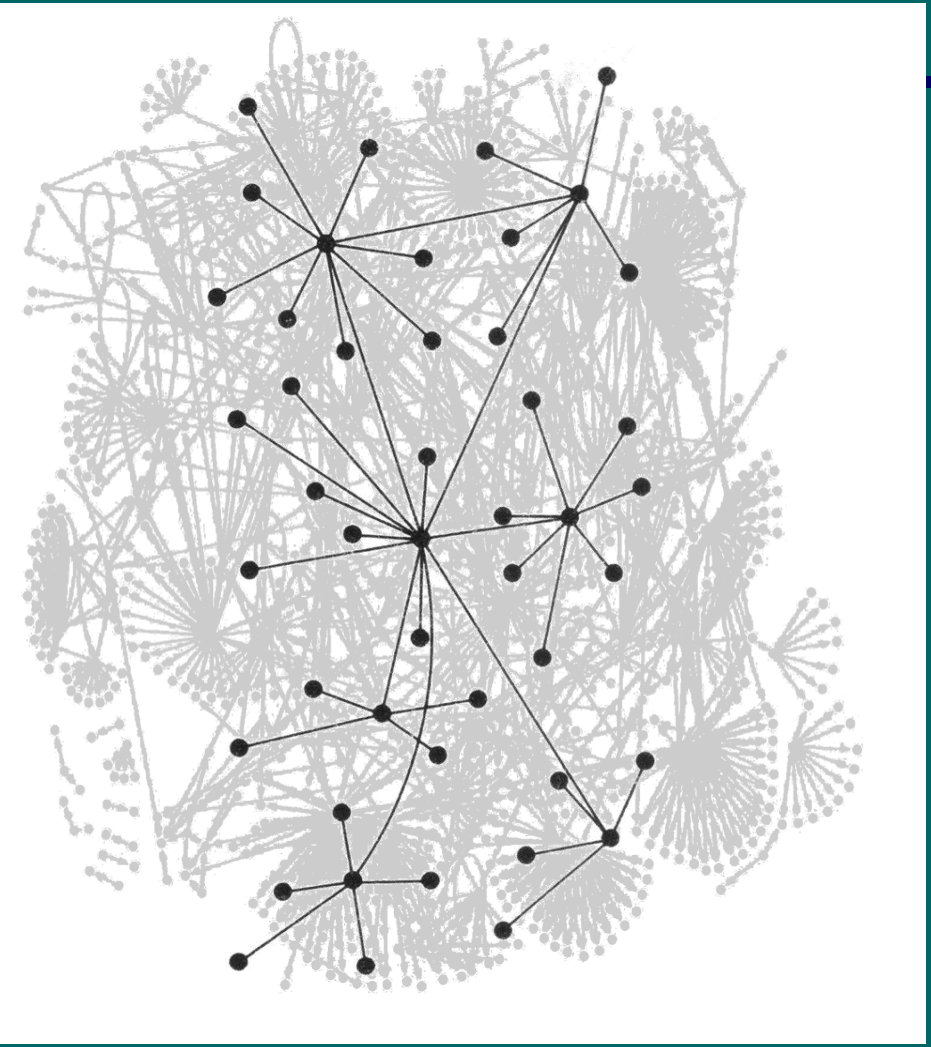


Traditional wet lab methods

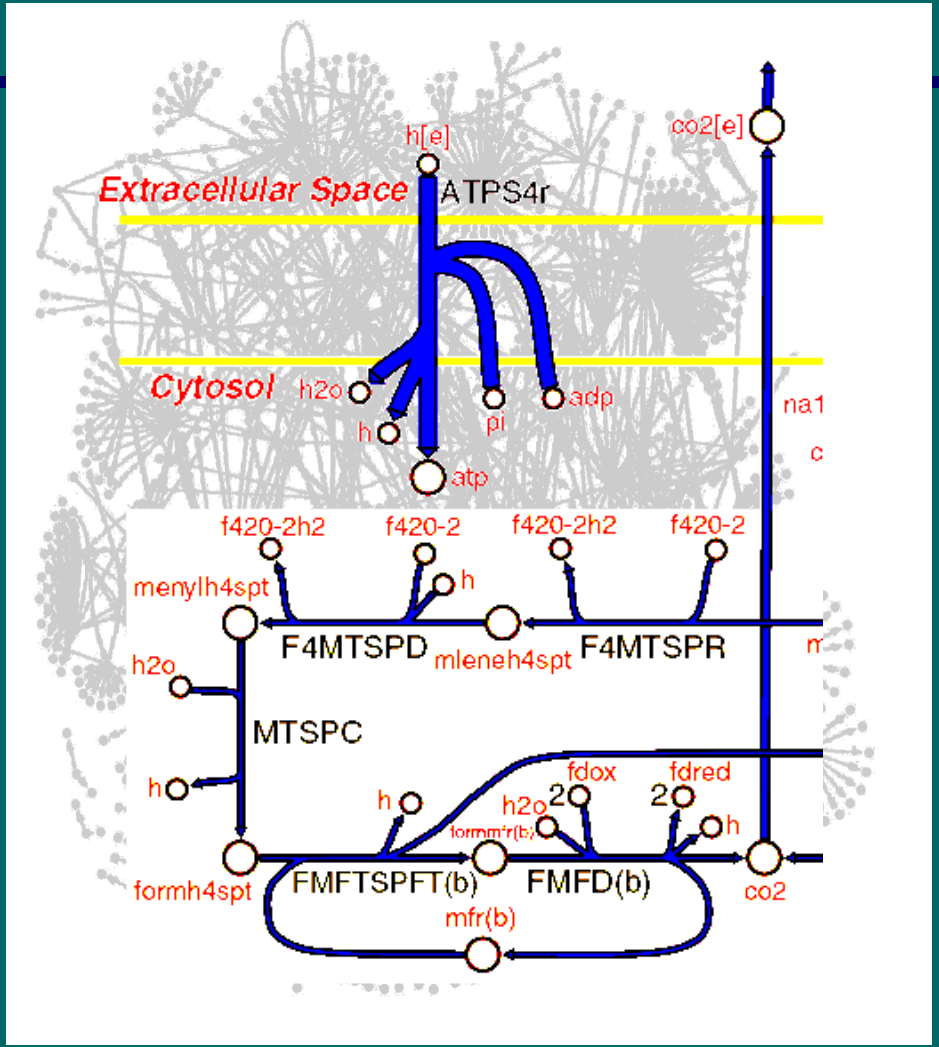


High throughput technologies

Holistic (system-wide) models



Topological models
(networks)



Dynamic models
(flux, transport)

Bioinformatics: molecular vs. systems biology

- Single entities (Molecular biology): Bioinformatics started as computational support to molecular biology, i.e. the molecular studies of simple systems (1-2 genes, 1-2 proteins, etc).
  - Example Predicting gene function via database searching

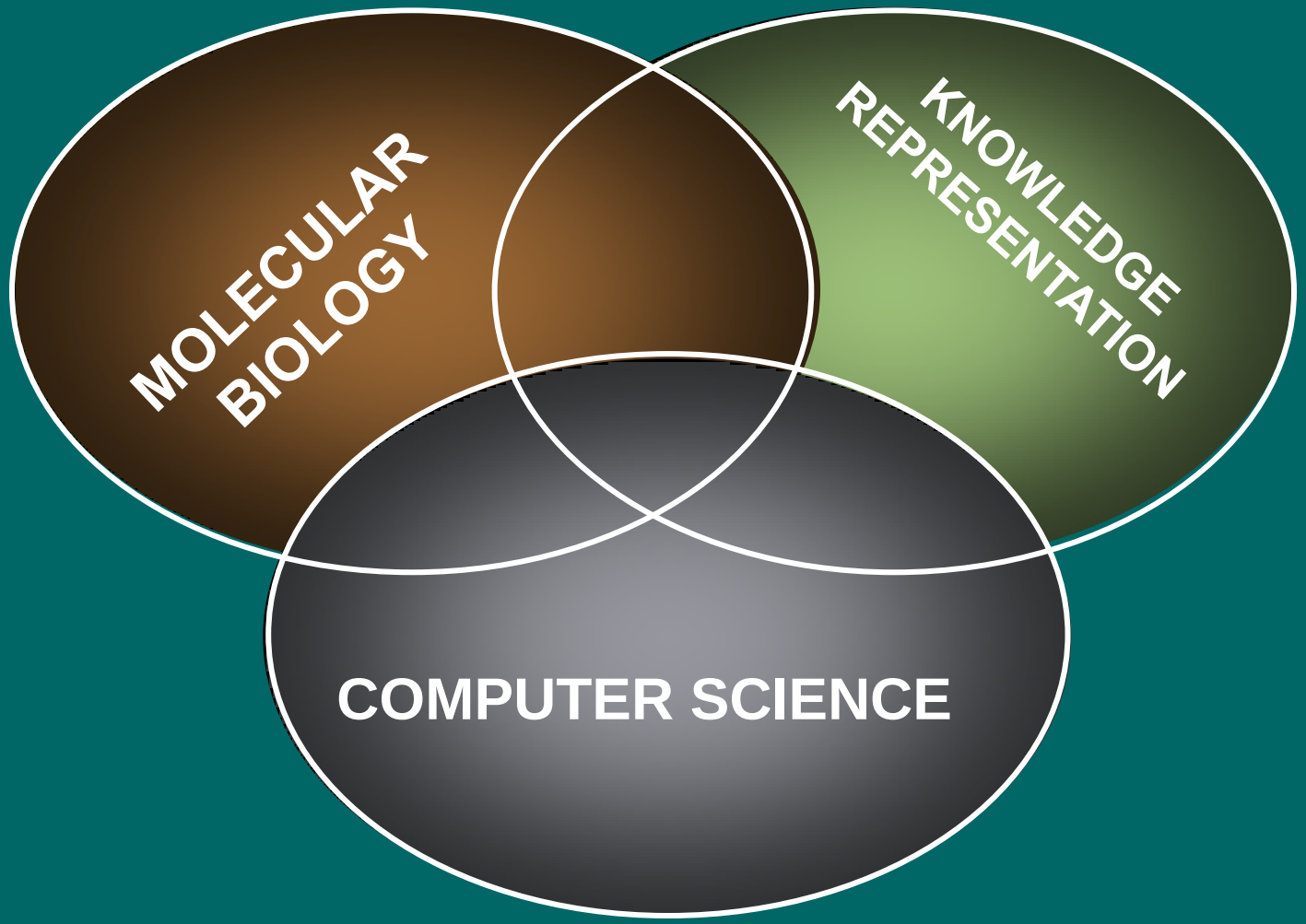
- Large systems (system biology) : As new measuring methods allow the parallel study of many genes and proteins, systems biology emerged as a new field (measuring technique + specific computational approaches).
  - Example: Studying gene expression in a whole genome using next generation sequencing

Modeling and simulation: molecular vs. systems biology

- Single entities (Molecular biology):
  - Modeling the movement of single molecules *in vacuo* or in water (molecular modeling, molecular dynamics)
  - Docking (e,g, pharmacons to their receptor proteins)

- Large systems (system biology):
  - Modeling large molecular assemblies
  - Modeling biological communities (bacteria, animals, human crowds)

Bioinformatics is interdisciplinary



What is particular in bioinformatics?

- The objects: molecular structures, metabolic pathways, regulatory networks AND their databases
- The methods: analysis and use of similarity;
- Complexity of biological knowledge
  (and NOT so much the quantity of data...)

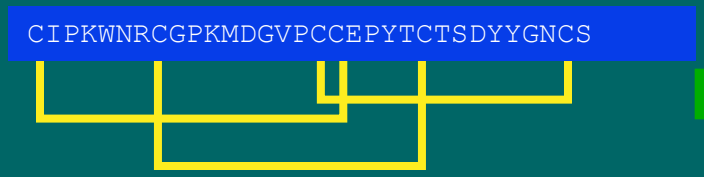
Part 2

Theory of biomolecular data, structure and function (system theory), annotation of data

Molecular structures: many different representations

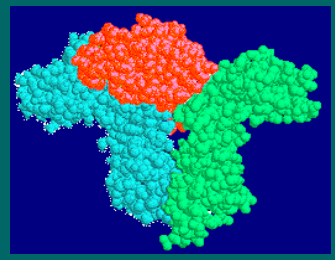
MARTKQTARK
STGGKAPRKQ
LATKAARKSA

Sequences

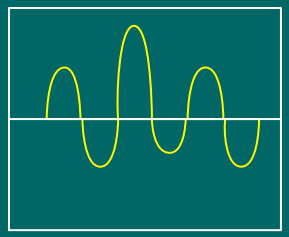


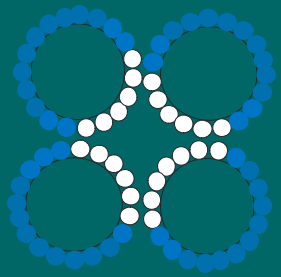
Extended sequences (pl. disulfide topology)

Cartoons of domains or secondary structures

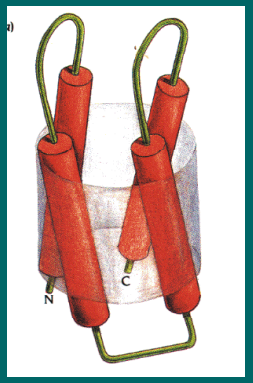


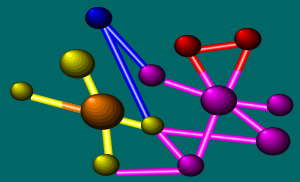
3D structures





Symbolic diagrams (e.g. hydrophobicity plots, helical circle diagrams)

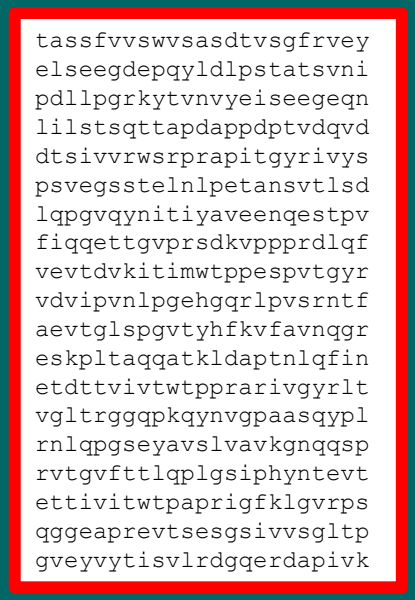


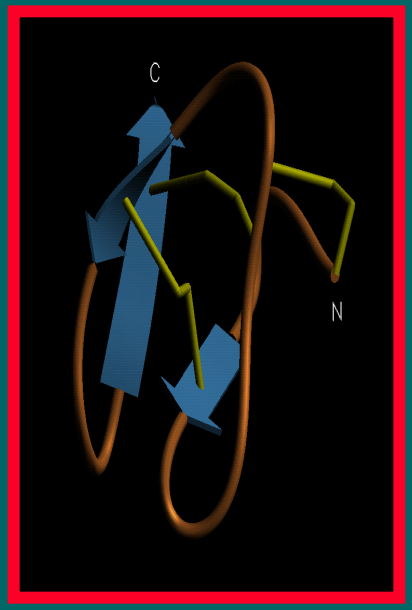


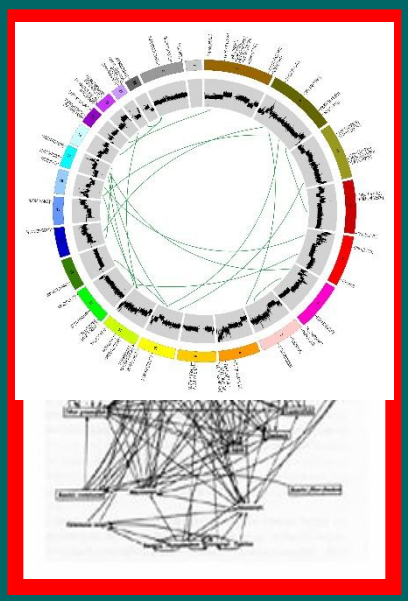
Simplified 3D cartoons

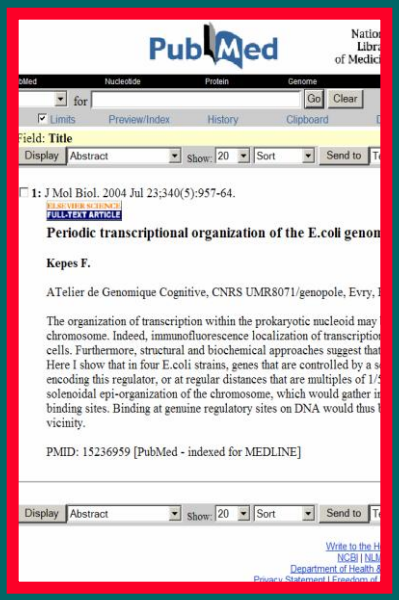
Core data-types

ALL HAVE SIMPLIFIED AND EXTENDED (ANNOTATED) VERSIONS





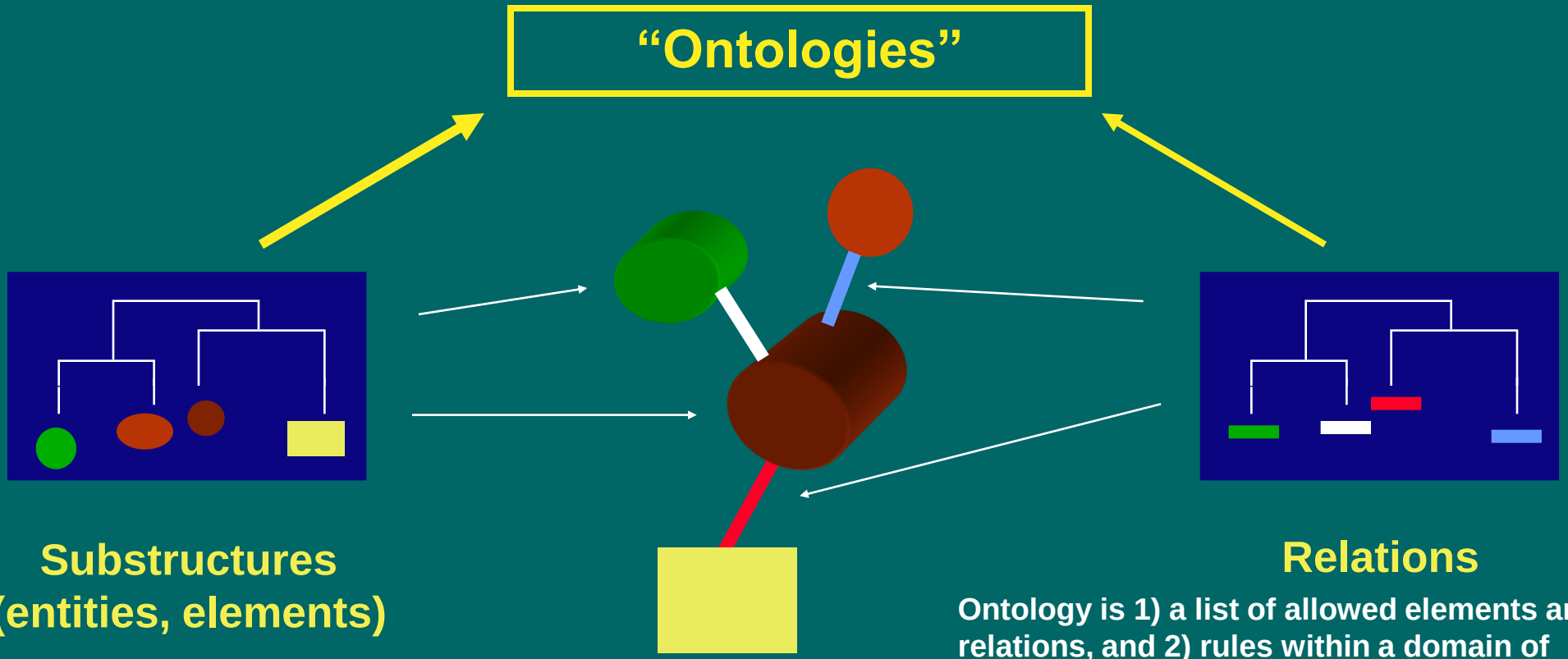




SEQUENCES **3-D** **GENOMES NETWORKS** **TEXT**

WE PUT ALL OF THEM INTO DATABASE RECORDS

Common basis: Structure = (set of elements connected with relationships) named according to conventions



Ontology is 1) a list of allowed elements and relations, and 2) rules within a domain of knowledge (e.g. gene ontologies, 3D structure ontologies etc.)

Informally: controlled vocabulary + rules

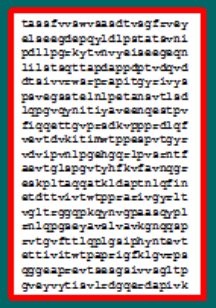
Gene ontology project (GO)

Pongor, Nature, 1987

Examples for entities and relationships
(nodes and edges in a graph)

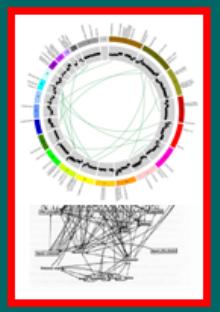
System	Entities (nodes)	Relationships (edges)
a) Conceptual models of natural systems		
Molecules	Atoms	Atomic interactions (chemical bonds)
Assemblies	Proteins, DNA	Molecular contacts
Pathways	Enzymes	Chemical reactions (substrates/products)
Genetic networks	Genes	Co-regulation
b) Structural descriptions		
Protein structure	Atoms	Chemical bonds
Protein structure	Secondary structures	Sequential and topological vicinity
Folds	C_α atoms	Peptide bond
Protein sequence	Amino acid	Sequential vicinity

Rules for selecting a structure type



SEQUENCES

3-D



GENOMES
NETWORKS

TEXT

- Structure definitions are hierarchical (atom – amino acid – protein – pathway – cell – tissue etc.)
- For a given problem it is convenient to choose a standard description or “core structural level”. E.g. DNA sequences are the standard level for molecular biology problems.
- For a standard or core description, we always have an underlying logical structure, plus various additional, simplified and annotated views. (annotation means extending with external information).
- What is annotation? What is function? Explained in the next section.

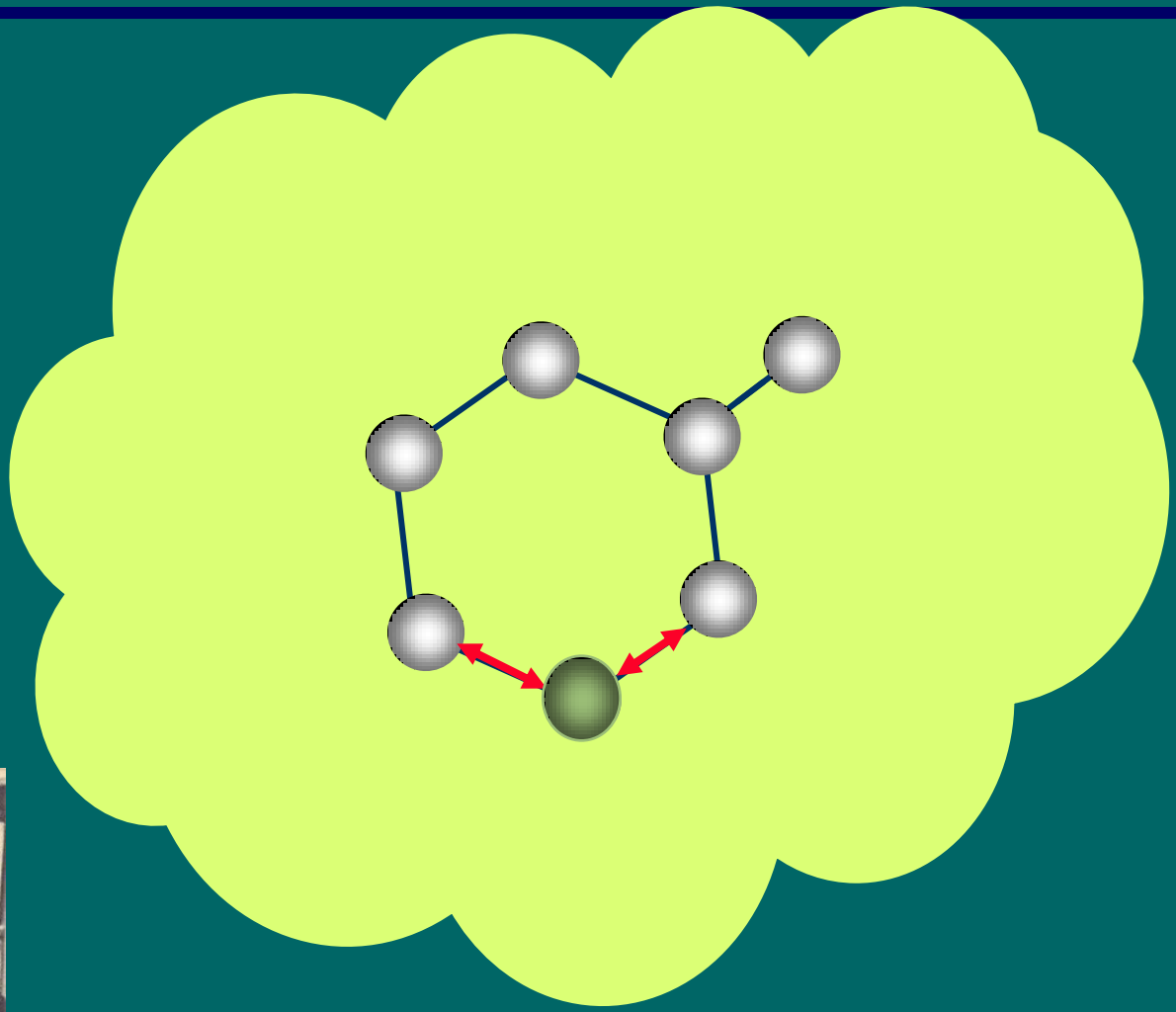
Molecules have structure and function

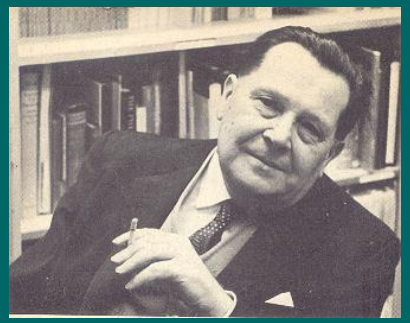
- Structure and function are concepts of systems theory

What are "systems"?

- Any part of reality that can be ~separated from the environment (by a boundary). A community in an environment.
- Consist of interacting parts
- Interact with the environment (inputs, outputs)
- System models are generalizations of reality
- Have a structure that is defined by parts and processes
- Parts have functional as well as structural relationships between each other.

System, stability, structure, function





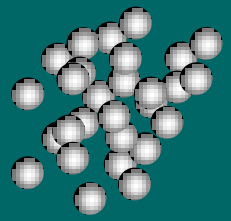
Ludwig von Bertalanffy

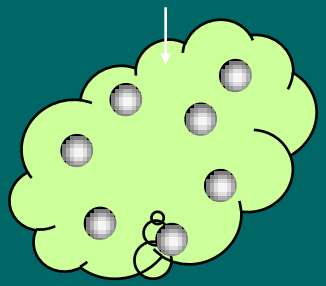
Function is a role within a higher system, a propertty that emerges within a higher system

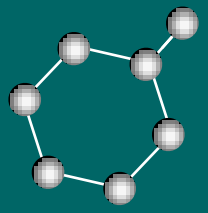
Structure, function

- Structure is a (~constant space-time) arrangement of elements or properties.
- Function is a role played within a system.

We use repeating features for describing systems







Repeats

Between systems

(in time and/or space)

Within systems

Elements

Relationships

Patterns

Symmetry

Proportions, harmonies

This is just to make a funny point: Starting with a structure, we can deduce seemingly disparate concepts like patterns, symmetries or proportions

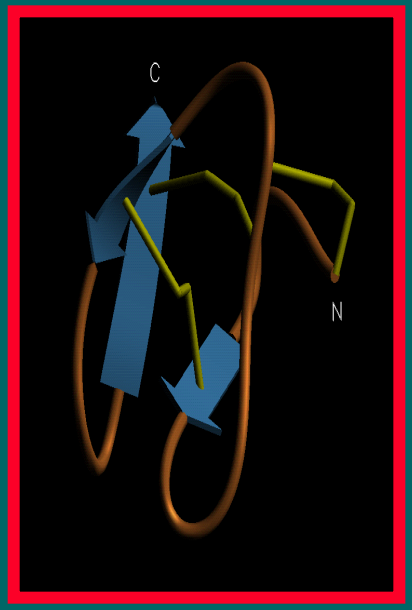
Structural data in brief

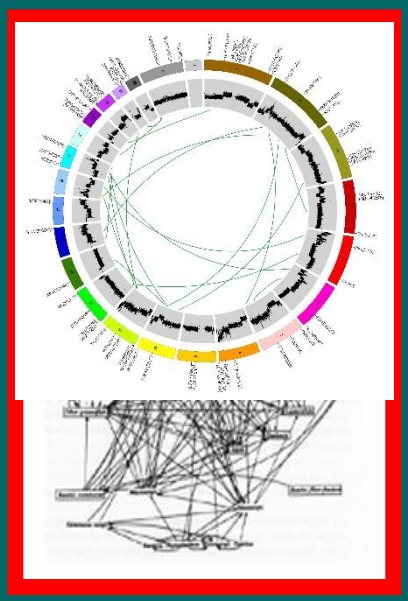
- Structure definitions are hierarchical (atom – amino acid – protein – pathway – cell – tissue etc.)
- For a given problem it is convenient to choose a standard description or “core structural level”. E.g. DNA sequences are the standard level for molecular biology problems.
- For a standard or core description, we always have an underlying logical structure, plus various additional, simplified and annotated views. (annotation means extending with external information).

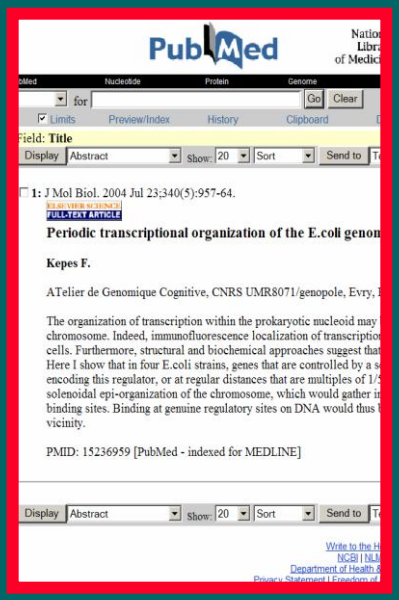
Core data-types

ALL HAVE SIMPLIFIED AND EXTENDED (ANNOTATED) VERSIONS

tassfvvswvsasdtvsgfrvey
elseegdepqyldlpstatsvni
pdllpgrkytvnvyeiseegeqn
lilstsqttapdappdptvdqvd
dtsivvrwsrprapitgyrivys
psvegsstelnlpetansvtlsd
lqpgvqynitiyaveenqestpv
fiqqettgvprsdkvppprdlqf
vevtdvkitimwtppespvtgyr
vdvipvnlpgehgqrlpvsrntf
aevtglspgvtyhfkvfavnqgr
eskpltaqqatkldaptnlqfin
etdttvivtwtpprarivgyrlt
vgltrggqpkqynvgpaasqypl
rnlqpgseyavslvavkgnqqsp
rvtgvfttlqplgsiphyntevt
ettivitwtpaprigfklgvrps
qggeaprevtsesgsivvsgltp
gveyvytisvlrdgqerdapivk







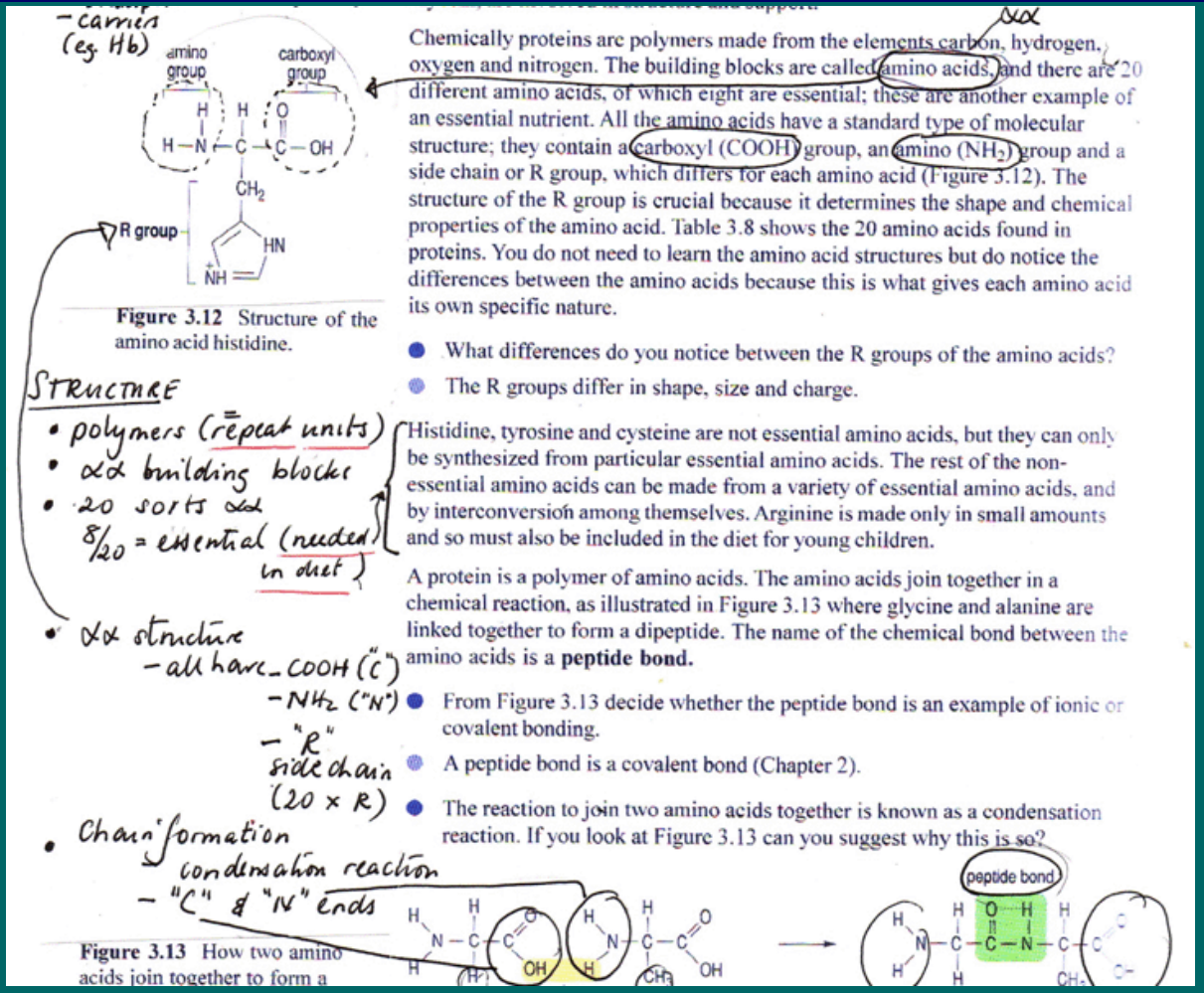
SEQUENCES **3-D** **GENOMES NETWORKS** **TEXT**

WE PUT ALL OF THEM INTO

Part 2/A

Annotations

Annotation: adding notes



– carries (eg. Hb)

amino group

carboxyl group

R group

Figure 3.12 Structure of the amino acid histidine.

Chemically proteins are polymers made from the elements carbon, hydrogen, oxygen and nitrogen. The building blocks are called amino acids, and there are 20 different amino acids, of which eight are essential; these are another example of an essential nutrient. All the amino acids have a standard type of molecular structure; they contain a carboxyl (COOH) group, an amino (NH_2) group and a side chain or R group, which differs for each amino acid (Figure 3.12). The structure of the R group is crucial because it determines the shape and chemical properties of the amino acid. Table 3.8 shows the 20 amino acids found in proteins. You do not need to learn the amino acid structures but do notice the differences between the amino acids because this is what gives each amino acid its own specific nature.

- What differences do you notice between the R groups of the amino acids?
- The R groups differ in shape, size and charge.

Histidine, tyrosine and cysteine are not essential amino acids, but they can only be synthesized from particular essential amino acids. The rest of the non-essential amino acids can be made from a variety of essential amino acids, and by interconversion among themselves. Arginine is made only in small amounts and so must also be included in the diet for young children.

STRUCTURE
- polymers (repeat units)
- αα building blocks
- 20 sorts αα
  8/20 = essential (needed in diet)
- αα structure
  – all have –COOH ("C")
  – NH_2 ("N")
  – "R" side chain (20 × R)
- Chain formation
  condensation reaction
  – "C" & "N" ends

A protein is a polymer of amino acids. The amino acids join together in a chemical reaction, as illustrated in Figure 3.13 where glycine and alanine are linked together to form a dipeptide. The name of the chemical bond between the amino acids is a **peptide bond.**

- From Figure 3.13 decide whether the peptide bond is an example of ionic or covalent bonding.
- A peptide bond is a covalent bond (Chapter 2).
- The reaction to join two amino acids together is known as a condensation reaction. If you look at Figure 3.13 can you suggest why this is so?

peptide bond

Figure 3.13 How two amino acids join together to form a

AD-NOTARE

A textbook with your own notes is valuable…

- Data (e.g. sequence)
- Data on data (annotation, meta-data)
- Data on annotations (ontologies, meta-meta-data: defining the language of annotations)

Anything added to the “standard description” is annotation

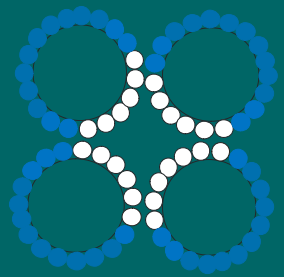
Building a database from raw data + annotations

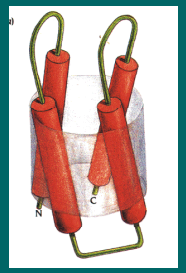
- Put raw data into database records
- Add basic annotations (project name, date etc.)
- Add annotations by similarity. This is called database searching (gives results as: 95% similarity to trypsin → probably trypsin. But only probably!!)
- Add further information based on human knowledge (analysis programs , literature search)

So our notes are partly trivial, partly based on guesses (similarity) or on sophisticated background work.

“Generalized structures” As Database Records

Other db’s



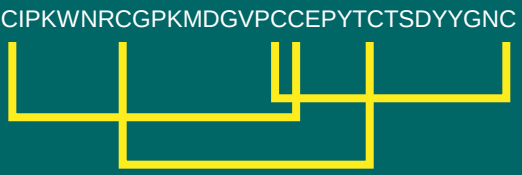


Identification
Name of protein
Organism
Function
Cross-references
...
Domain structure
Sec. structure
Disulphides
….

Sequence (structure)

qfinetdttvivtwtpprarivgyrltvgllseeg
depqyldlpstatsvnipdllpgrkytvnvyeise
egeqnlilstsqttapdappdptvdqvddtsivvr
wsrprapitgyrivyspsvegsstelnlpetansv
tlsdlqpgvqynitiyaveenqestpvfiqqettg
vprsdkvppprdlqfvevtdvkitimwtppespvt
gyrvdvipvnlpgehgqrlpvsrntfaevtglspg
vtyhfkv

ANNOTATIONS

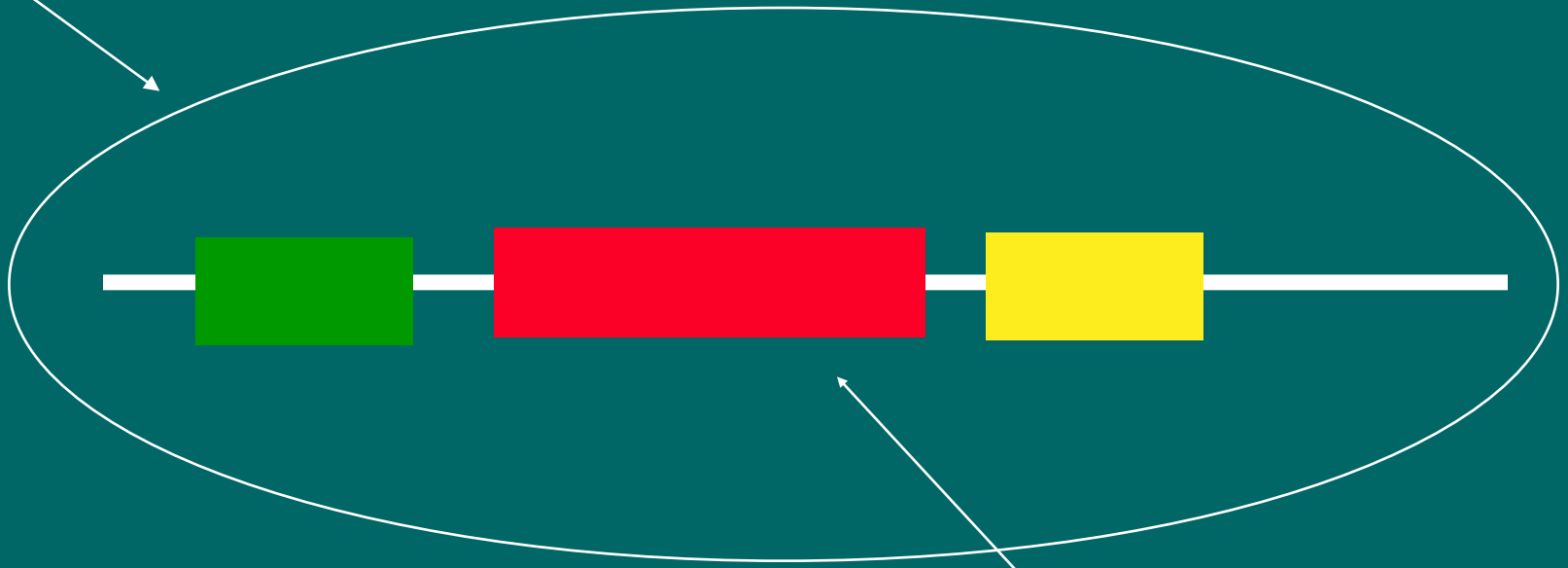


STRUCTURE,
eg. SEQUENCE

Database record, fields

Annotation of (sequence) data

Global descriptors
e.g. function



Annotation requires database searching and knowledge of „biology”

Local (positional) descriptors e.g. domains

Annotation and the World Wide Web

- Traditionally, annotations to a structure are validated and added by humans : authors trying to suggest a function for a new gene, database developers trying to add structural or functional descriptions to molecular data, etc.
- WWW is the biggest annotation system: millions of non-validated links are added to data. Important types include databases (bioinformatics and bibliographic), Wikipedia (community based encyclopedia), specialist wikis, blogs, discussion lists. Google search is a first step…
- Today, database annotation means generating standard language descriptions for data, validated via Internet links and specialized programs. Relies on human intervention.

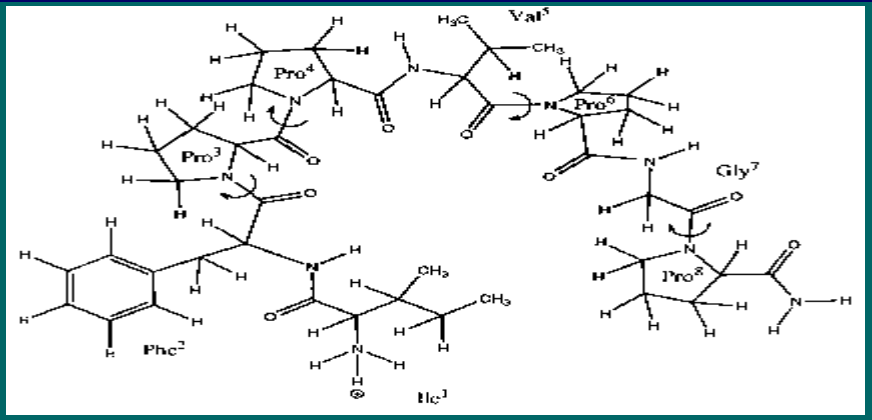
Part 3

The 4 standard data-types in detail

- SEQUENCES

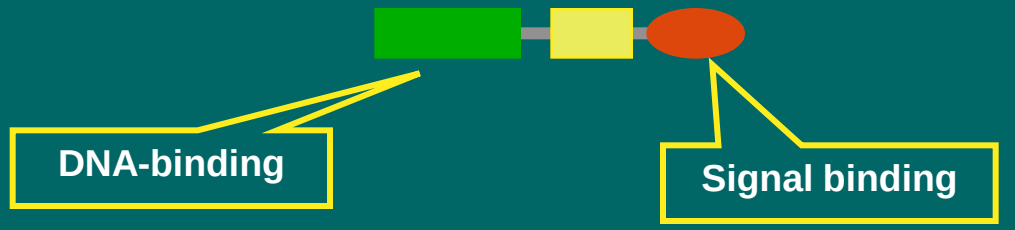
SEQUENCES

- Standard description: Series of characters (denoting amino acids or nucleotides)



- Simplified and/or extended visualization

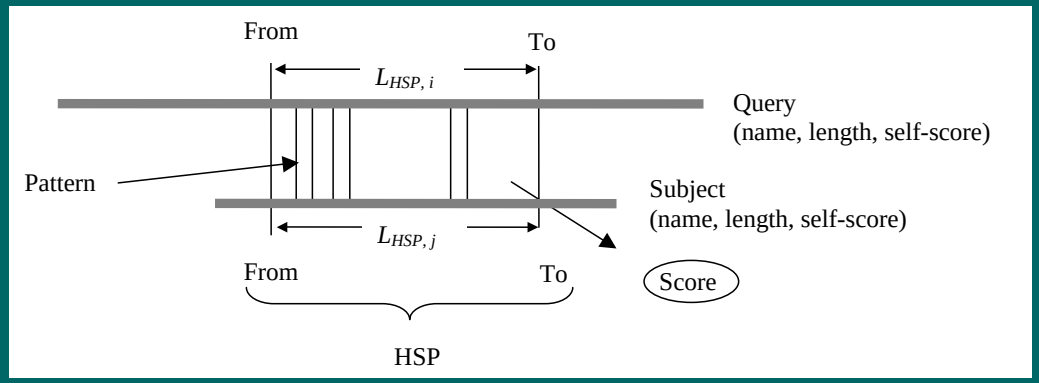
IFPPVPGP



Sequences as language

```
qfinetdttvivtwtpprarivgyrl
tvgllseegdepqyldlpstatsvni
pdllpgrkytvnvyeiseegeqnlil
stsqttapdappdptvdqvddtsivv
rwsrprapitgyrivyspsvegsste
lnlpetansvtlsdlqpgvqynitiy
aveenqestpvfiqqettgvprsdkv
ppprdlqfvevtdvkitimwtppesp
vtgyrvdvipvnlpgehgqrlpvsrn
tfaevtglspgvtyhfkvfavnqgre
skpltaqqatkldaptnlqfinetdt
tvivtwtpprarivgyrltvgltrgg
qpkqynvgpaasqyplrnlqpgseya
vslvavkgnqqsprvtgvfttlqplg
siphyntevtettivitwtpaprigf
klgvrpsqggeaprevtsesgsivvs
gltpgveyvytisvlrdgqerdapiv
kkvvtplspptnlhleanpdtgvltv
swersttpditgyritttptngqqgy
sleevvhadqssctfenlspgleynv
svytvkddkesvpisssfvvswvsas
dtvsgfrveyelseegdepqyldlps
tatsvnipdllpgrkytvnvyeisee
```

Alignments



Character strings, computer-languages, Chomsky et al, etc.

Bssic facts

Sequence codes

Amino Acid	Three-letter code	One letter code	Chemical character
Alanine	Ala	A	nonpolar
Arginine	Arg	R	Basic polar
Asparagine	Asn	N	polar
Aspartic acid	Asp	D	acidic polar
Cysteine	Cys	C	nonpolar
Glutamic acid	Glu	E	acidic polar
Glutamine	Gln	Q	polar
Glycine	Gly	G	nonpolar
Histidine	His	H	Basic polar
Isoleucine	Ile	I	nonpolar
Leucine	Leu	L	nonpolar
Lysine	Lys	K	Basic polar
Methionine	Met	M	nonpolar
Phenylalanine	Phe	F	nonpolar
Proline	Pro	P	nonpolar
Serine	Ser	S	polar
Threonine	Thr	T	polar
Tryptophan	Trp	W	nonpolar
Tyrosine	Tyr	Y	polar
Valine	Val	V	nonpolar

Nucleotide	One letter code	Chemical character
Adenine	A	Purine
Cytosine	C	Pyrimidine
Guanine	N	Purine
Thymine	D	Pyrimidine
Uracil	C	Pyrimidine

- One letter codes are used
- Amino acids: 20-letter alphabet
- Nucleotides 4-letter alphabet (either T (DNA) or U (RNA))

Bssic facts

Sequence formats

- Simple (“FASTA”) format

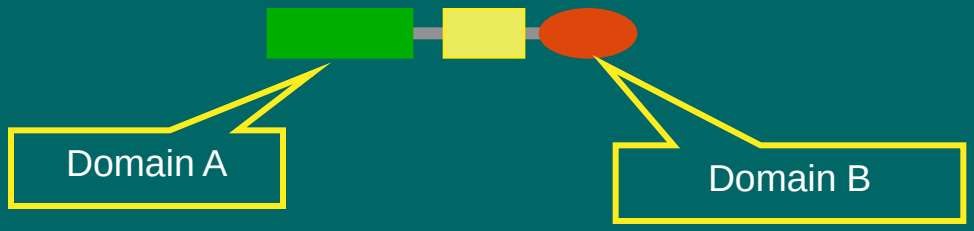
>name
ACAAGTTG

- Multipls “Concatenated FASTA”

>name1
ACAAGTTG
>name1
ACAAGTTG
>name1
ACAAGTTG

mplification +
annotation

PROTEIN SEQUENCE ANNNOTATED WITH DOMAINS



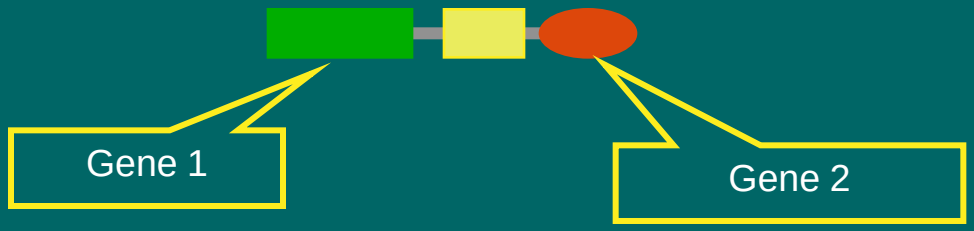

```
001-200    DOMAIN     PROTEASE A
205-230    DOMAIN     TRANSMEMBRANE
250-350    DOMAIN     SIGNAL BINDING
```

TABULAR DESCRIPTION: FEATURE TABLE, PTT TABLE

Sanjar Hudaiberdiev

Simplification + annotation

GENOME SEQUENCE ANNOTATED WITH GENES



Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
2700..3773	+	357	340780747	-	Atc_0003	-	COG1195L	DNA recombination and r
3770..6175	+	801	340780748	-	Atc_0004	-	COG0187L	DNA gyrase subunit B
6240..9710	+	1156	340780749	-	Atc_0005	-	COG0493ER	hypothetical protein
9745..10014	-	89	340780750	-	Atc_0006	-	COG0851D	cell division topologic

Sequence view of a genome

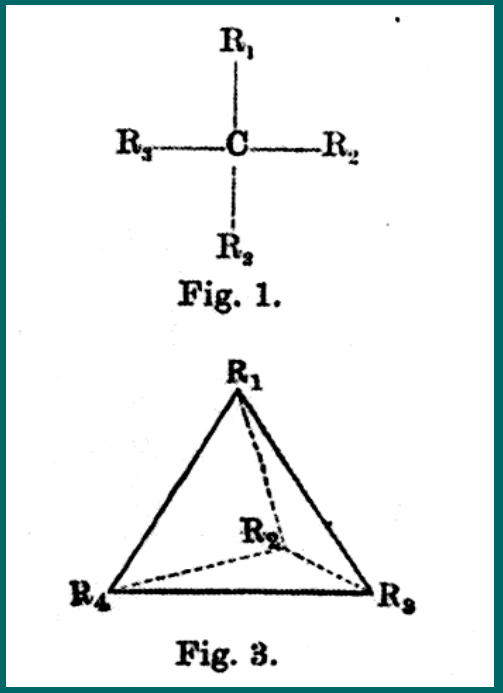
Genome annotation .ptt table

- 3D STRUCTURES

Chimie dans l'espace

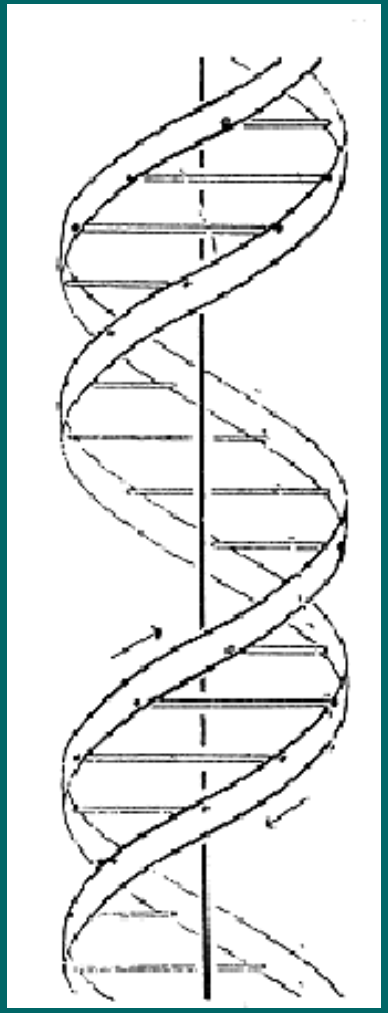
Van t'Hoff

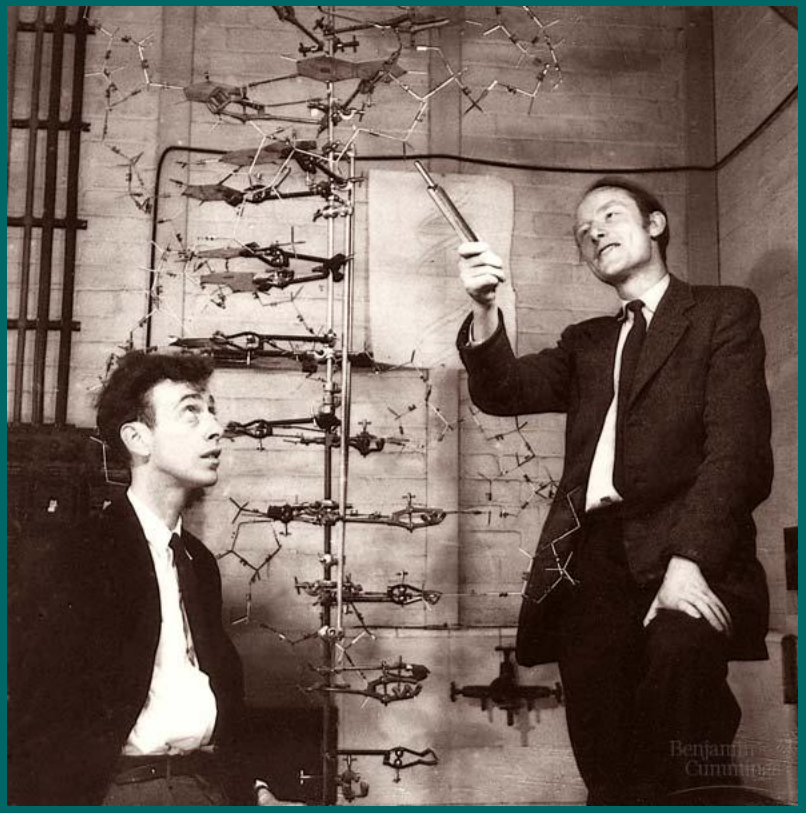
1852-1911



1898

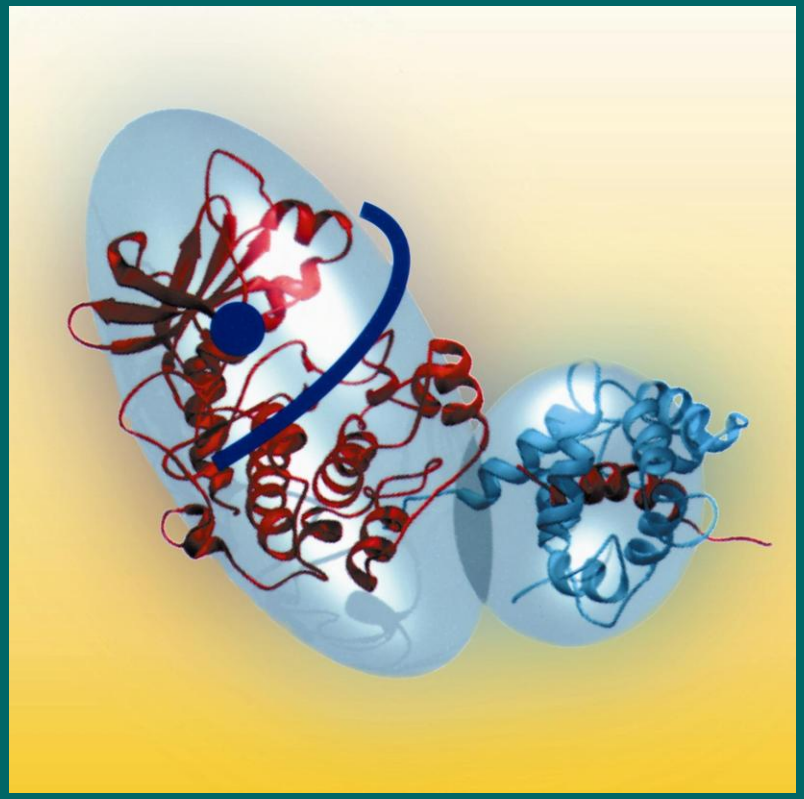
Some molecules are more equal than others…



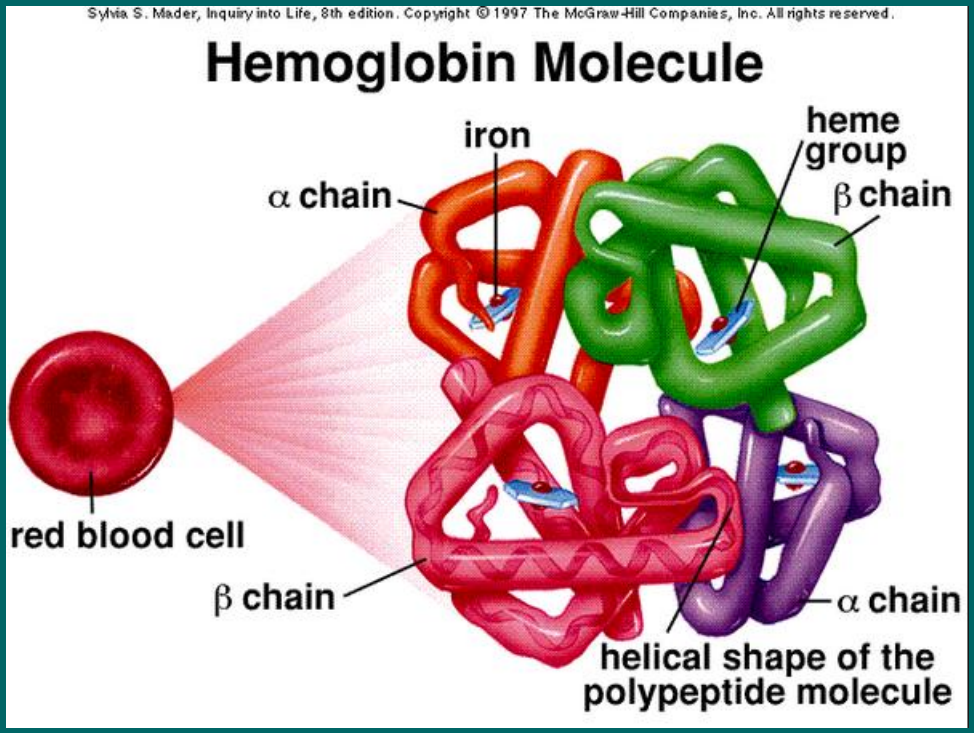


…”This figure is purely diagrammatic. The two ribbons symbolize the phosphate-sugar chains, and the horizontal rods the pairs of the bases holding the chains together. The vertical line marks the fibre axis”

Protein 3D



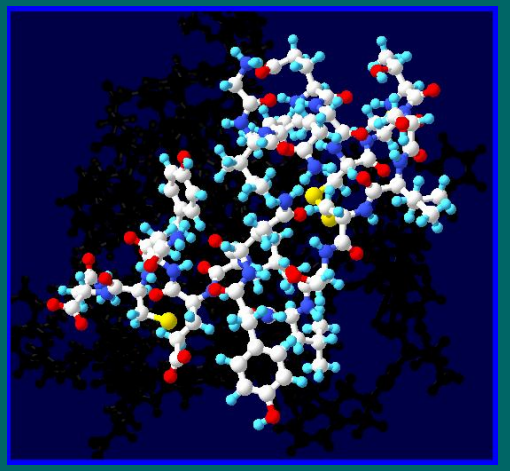
Simplified: 1) surface 2) backbone



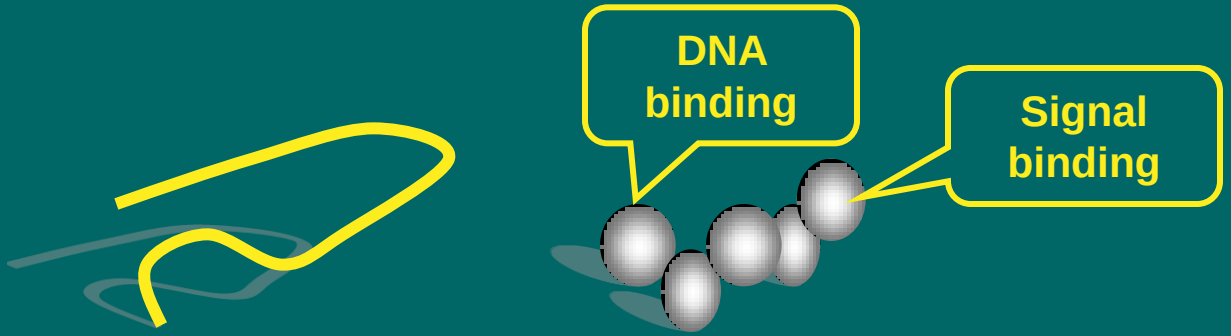
Annotated with structural and functional details

3D structures

- Standard description: 3D coordinates + subunit descriptions (connectivities)
- (atomic, amino acid, nucleotide)
- Simplified and/or extended (annotated) visualization

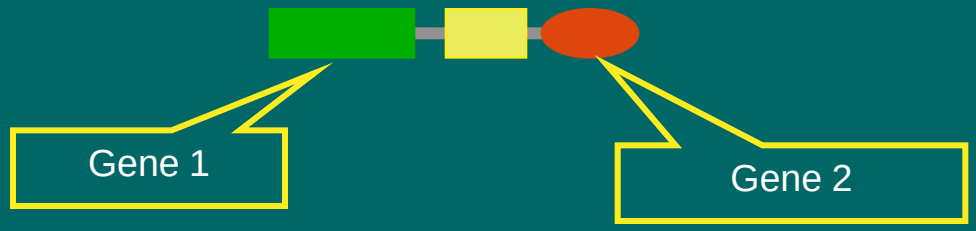


$(x_i, y_i, z_i)_n$



- GENOMES, NETWORKS

REMINDER: ANNOTATED SEQUENCE VIEW OF A GENOME



Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
2700..3773	+	357	340780747	-	Atc_0003	-	COG1195L	DNA recombination and r
3770..6175	+	801	340780748	-	Atc_0004	-	COG0187L	DNA gyrase subunit B
6240..9710	+	1156	340780749	-	Atc_0005	-	COG0493ER	hypothetical protein
9745..10014	-	89	340780750	-	Atc_0006	-	COG0851D	cell division topologic

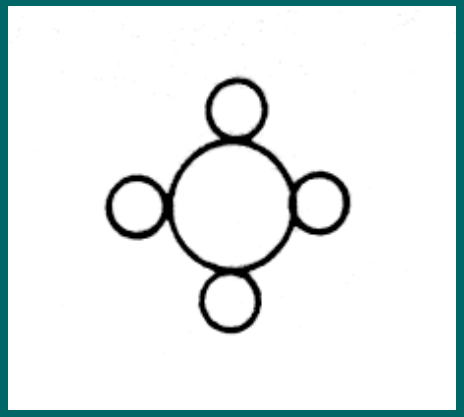
Sequence view of a genome

Genome annotation .ptt table

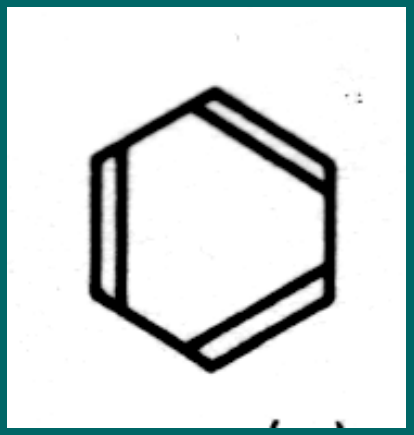
A genome is more than a sequence

- We want to add regulatory links (what regulates what)
- We want to add functional links (e.g. substrates passed between enzymes in a pathway)
- All these are links that define e network of genes, proteins substrates etc.
- “Network” are another core data type.

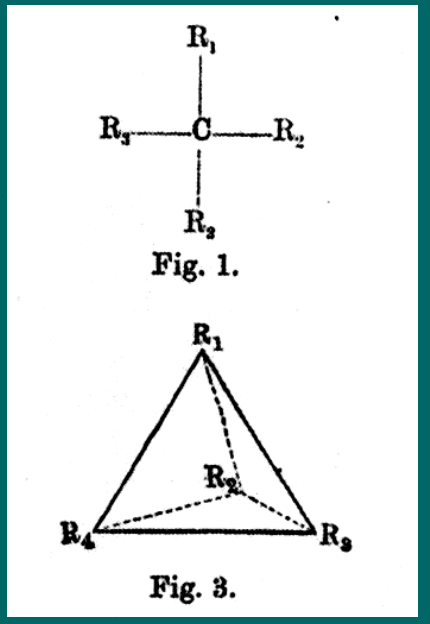
History: small molecules – classical graphs



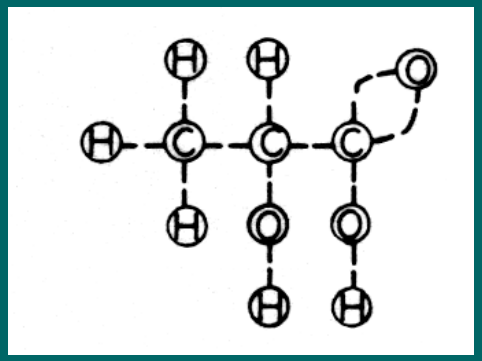
Loschmidt, 1861



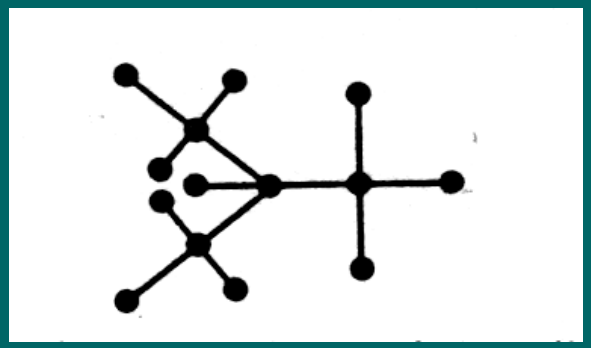
Kekulé, 1865



Van’t Hoff, 1898



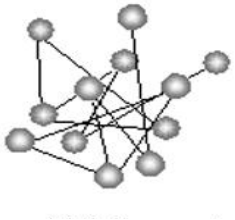
Crum Brown, 1861



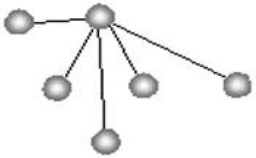
Cayley, 1872

TOPOLOGIES,
GRAPHS

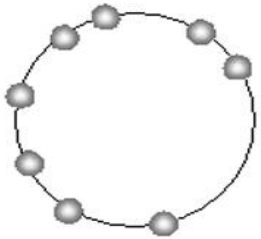
Genomes, assemblies



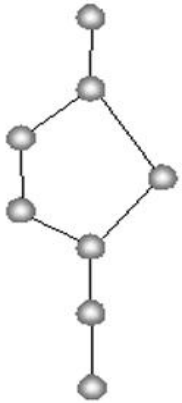
Similarity group



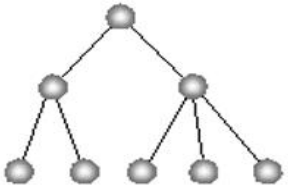
Neighbourhood



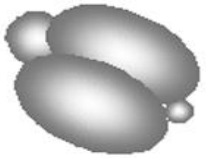
Genome



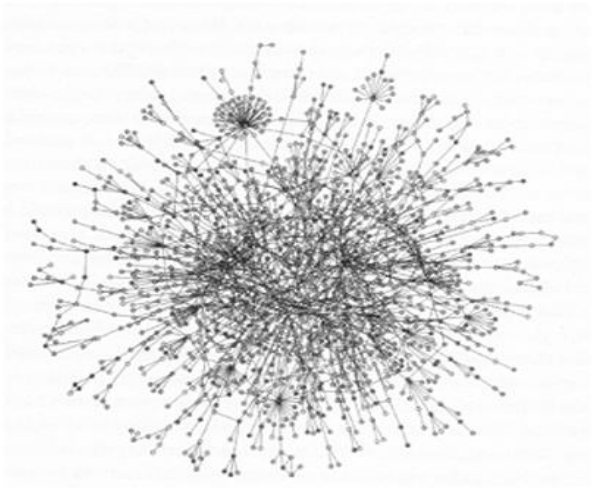
Metabolic
pathway



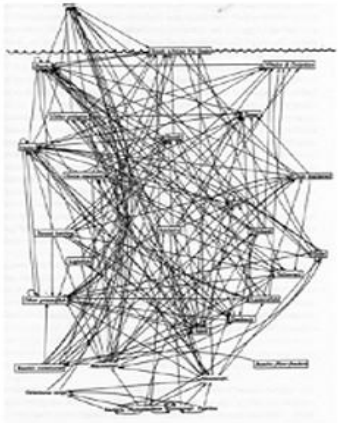
Tree-hierarchy



Complexes



Genetic network

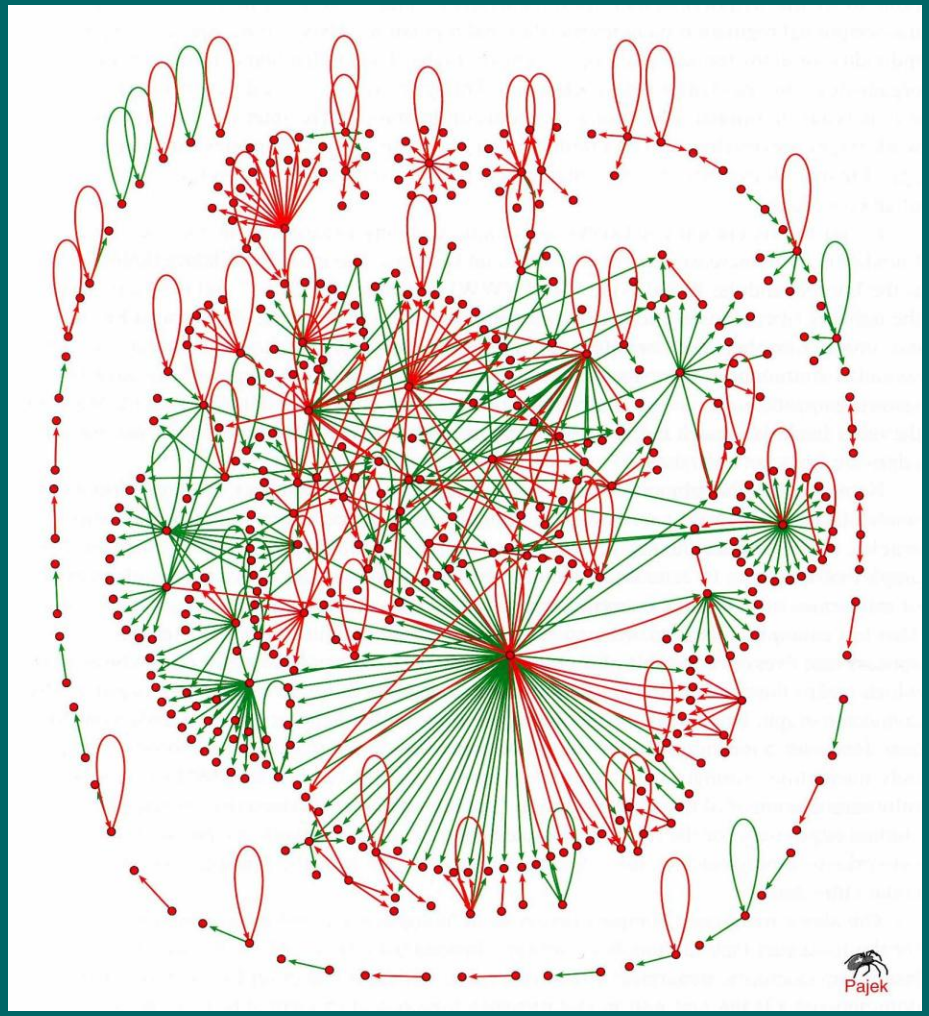


Food network

Entity-relationship models
Topological meta-models

The transcription regulatory networks

+ (up)
- (down)

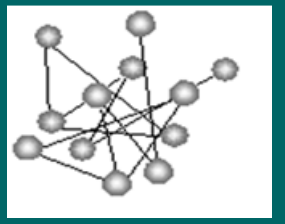


E. coli

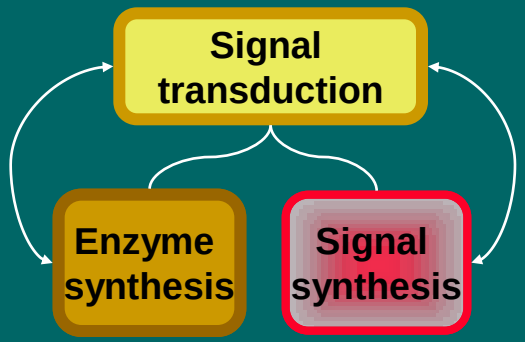
S. cerevisiae

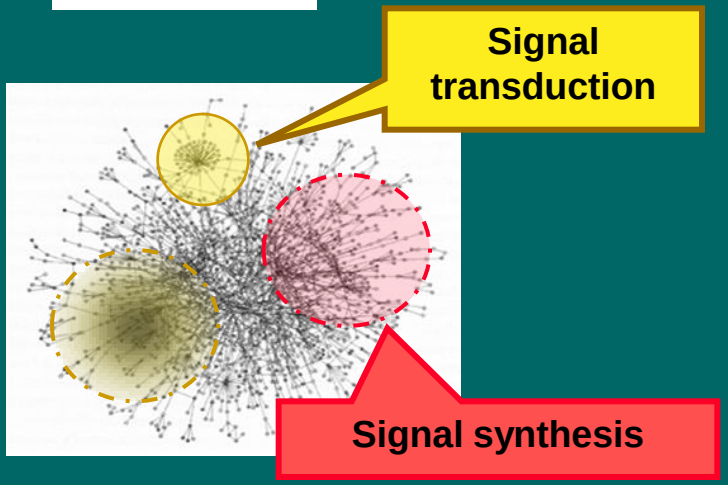
NETWORKS

- Sandard description: Graphs of entities (nodes) and relationships (edges)



- Simplified and/or extended (annotated) visualization



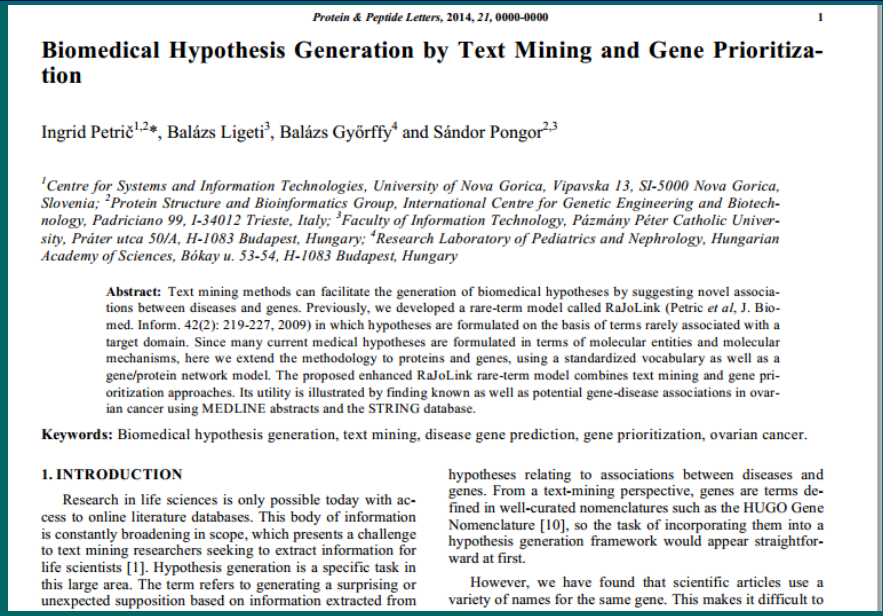


- **TEXTS (article abstracts)**

NATURAL LANGUAGE

Texts: Scientific publications

- A *human message* written in *scientific language* (“special English”, ~fixed vocabulary).
- Like other data, they have logical structure, standard, simplified and extended descriptions and databases
- BUT: messages have an emitter (author) and an audience (reader, reviewer). In other words they are context dependent (unlike, say, sequences or atoms)
- Loosely structured (not as well as molecules). There are ontologies for the language but not for the articles themselves!



Protein & Peptide Letters, 2014, *21*, 0000-0000 1

Biomedical Hypothesis Generation by Text Mining and Gene Prioritization

Ingrid Petrič[1,2]*, Balázs Ligeti[3], Balázs Győrffy[4] and Sándor Pongor[2,3]

[1]*Centre for Systems and Information Technologies, University of Nova Gorica, Vipavska 13, SI-5000 Nova Gorica, Slovenia;* [2]*Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, I-34012 Trieste, Italy;* [3]*Faculty of Information Technology, Pázmány Péter Catholic University, Práter utca 50/A, H-1083 Budapest, Hungary;* [4]*Research Laboratory of Pediatrics and Nephrology, Hungarian Academy of Sciences, Bókay u. 53-54, H-1083 Budapest, Hungary*

Abstract: Text mining methods can facilitate the generation of biomedical hypotheses by suggesting novel associations between diseases and genes. Previously, we developed a rare-term model called RaJoLink (Petric *et al*, J. Biomed. Inform. 42(2): 219-227, 2009) in which hypotheses are formulated on the basis of terms rarely associated with a target domain. Since many current medical hypotheses are formulated in terms of molecular entities and molecular mechanisms, here we extend the methodology to proteins and genes, using a standardized vocabulary as well as a gene/protein network model. The proposed enhanced RaJoLink rare-term model combines text mining and gene prioritization approaches. Its utility is illustrated by finding known as well as potential gene-disease associations in ovarian cancer using MEDLINE abstracts and the STRING database.

Keywords: Biomedical hypothesis generation, text mining, disease gene prediction, gene prioritization, ovarian cancer.

1. INTRODUCTION

Research in life sciences is only possible today with access to online literature databases. This body of information is constantly broadening in scope, which presents a challenge to text mining researchers seeking to extract information for life scientists [1]. Hypothesis generation is a specific task in this large area. The term refers to generating a surprising or unexpected supposition based on information extracted from hypotheses relating to associations between diseases and genes. From a text-mining perspective, genes are terms defined in well-curated nomenclatures such as the HUGO Gene Nomenclature [10], so the task of incorporating them into a hypothesis generation framework would appear straightforward at first.

However, we have found that scientific articles use a variety of names for the same gene. This makes it difficult to

Discipline: Text mining

Example database: PubMed

Texts: Scientific publications

LOGICAL STRUCTURE

- What is the question?
- What is the answer?
- What have we learnt?

- QUESTION
- ANSWER
- CONCLUSIONS

Remark: Not only articles but also their paragraphs and sentences have this structure *in some form.*

Texts: Scientific publications

LOGICAL STRUCTURE

- QUESTION
- ANSWER
- CONCLUSIONS

STANDARD DESCRIPTION

(simple)

- Introduction
- Results (what), Methods (how)
- Discussion

Texts: Scientific publications

STANDARD DESCRIPTION

(complete)

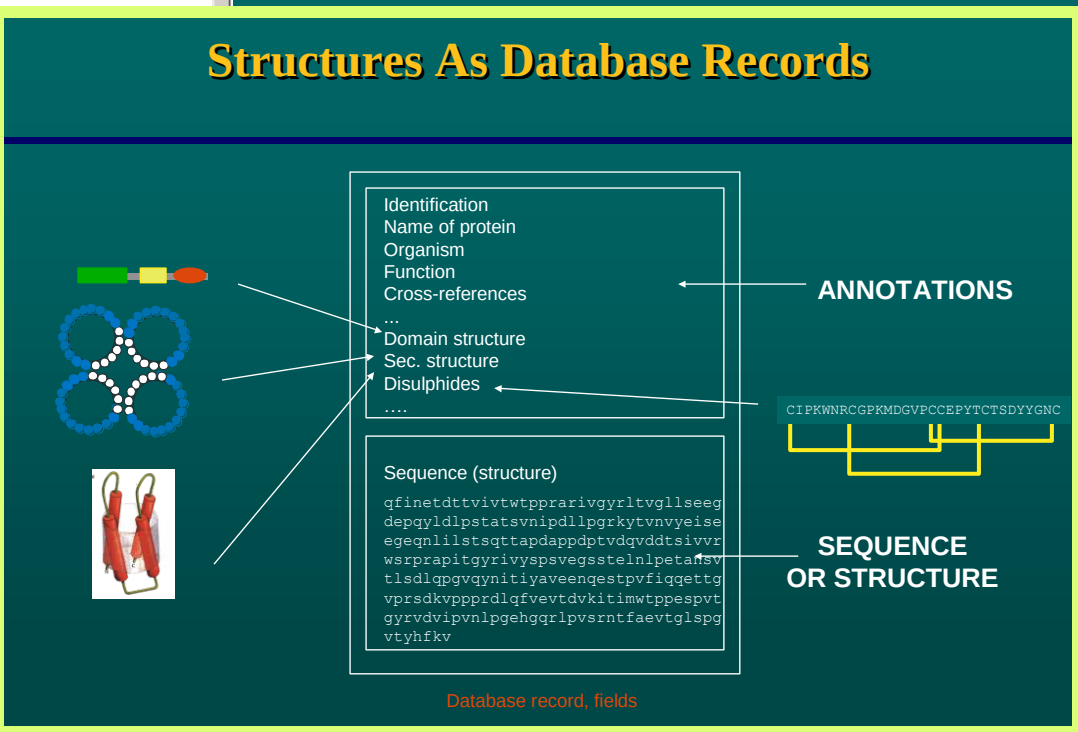
- Title (+ authors etc)
- Abstract
- Introduction
- Results (what), Methods (how)
- Discussion
- References
- Keywords

Simplified descriptions

Annotations

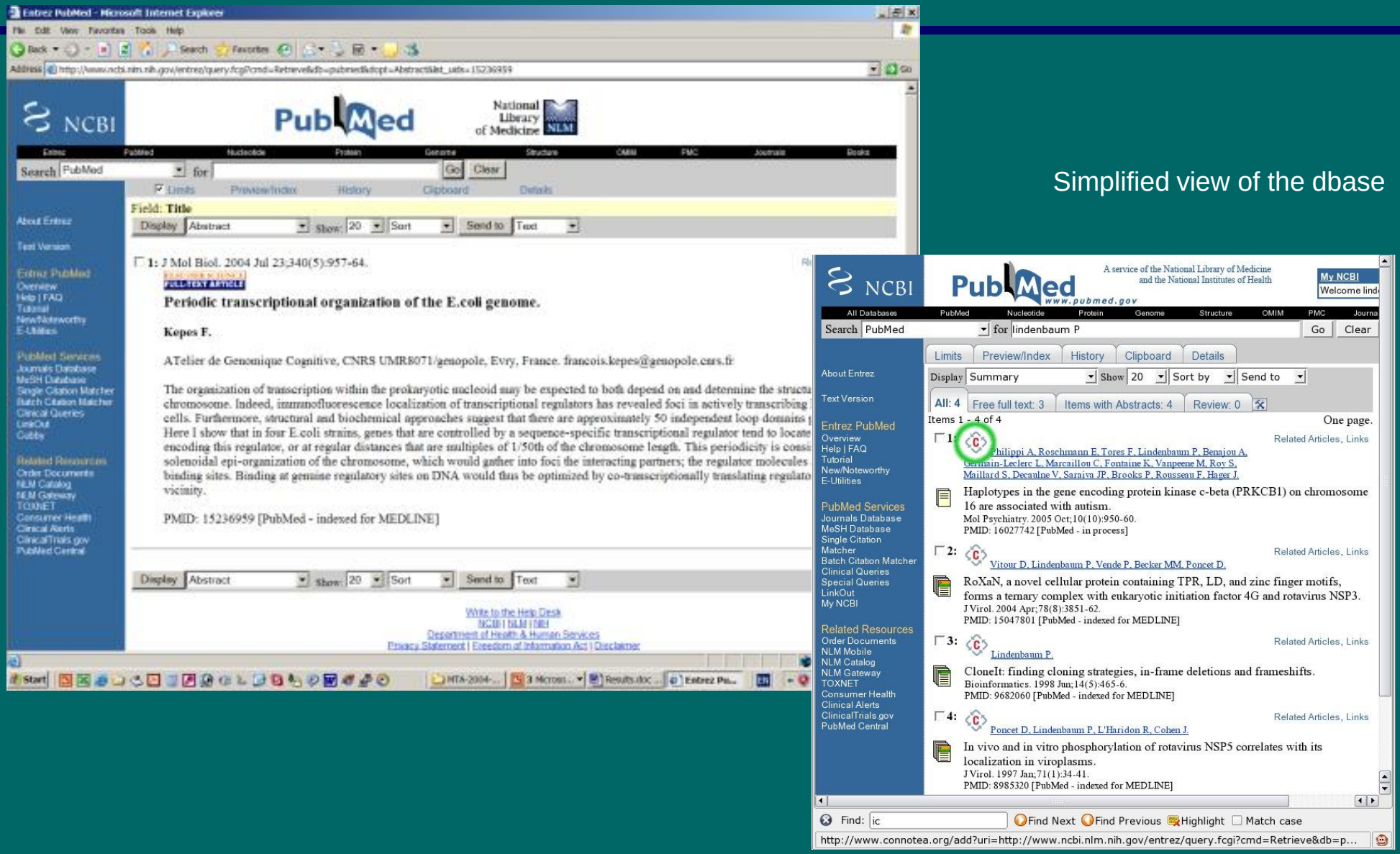
Texts: Scientific publications

Structures As Database Records



Keyword-collecttions, onthologies, etc.

Example: PubMed



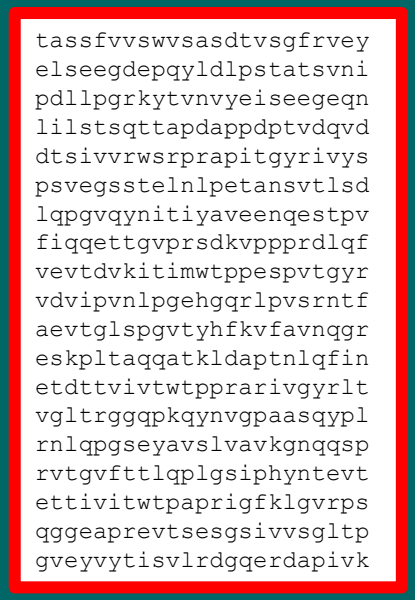
Simplified view of the dbase

Part 5

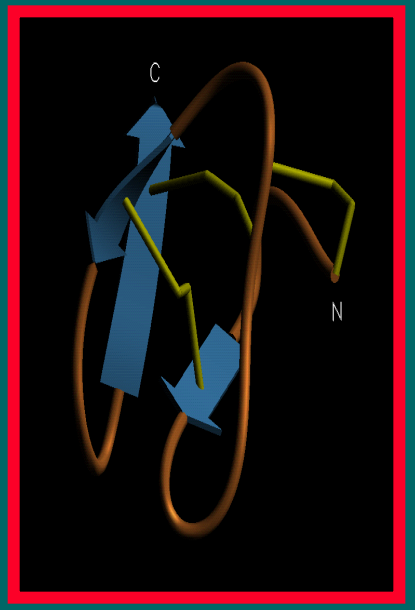
Summary

Core data-types

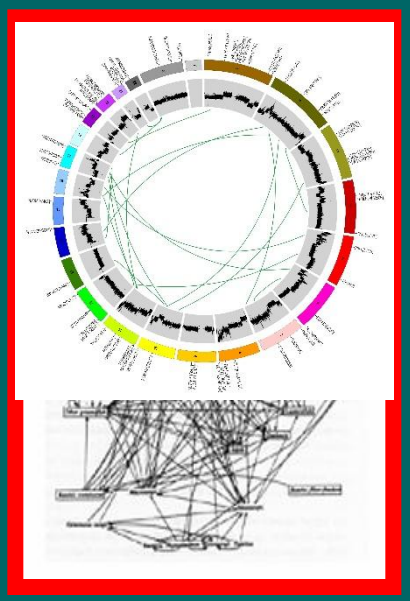
ALL HAVE SIMPLIFIED AND EXTENDED (ANNOTATED) VERSIONS



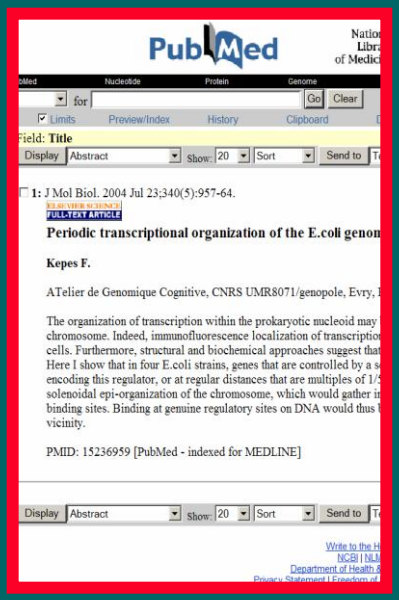
SEQUENCES



3-D



GENOMES NETWORKS



TEXT

WE PUT ALL OF THEM INTO

BASIC CONCEPTS OF BIOINFORMATICS

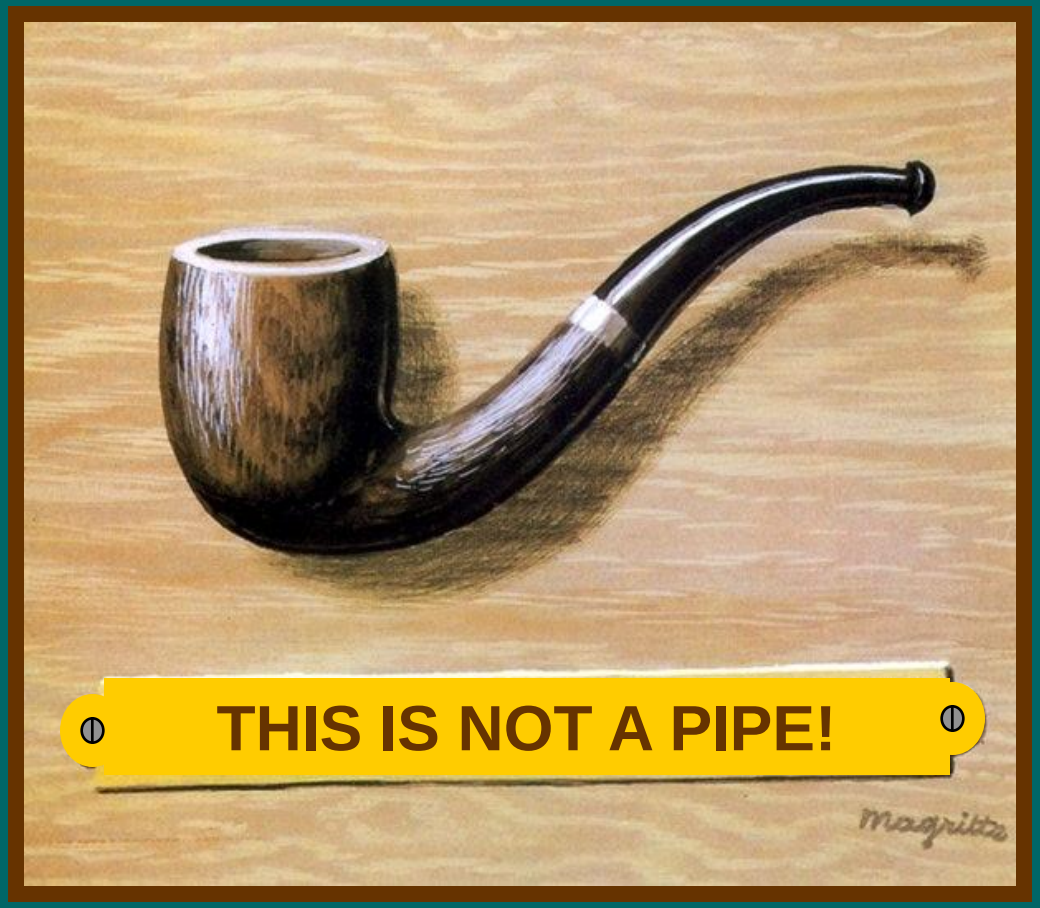
- Biological computer uses include bioinformatics (data management, data-mining) and modeling or simulation
- Concepts of system, structure, function. Structure is an ensemble of elements and relations. Logical structure, simplified and extended (annotated) descriptions.
- 4 core data-types (models): sequence, 3D, network and text
- Models are represented by computers with dedicated data-structures, images or in a textual form.
- Database records contain a variety of data-types in machine-readable and/or human-readable forms.
- Annotation (added human knowledge) is crucial, better if machine readable.

A bit of cultural outlook

The core datatypes reflect basic human situations (paradigms, metaphors etc.)

- Sequences resemble languages („*language metaphor*”)
- 3D structures resemble real life objects („*object metaphore*”)
- Networks resemble social asemblies („*social metaphore*”)
- Scientific papers are messages (”*communication metaphor*”)

Models are human constructs...



Models are human constructs...

THIS IS NOT A MOLECULE

What you should know

- Definition of bioinformatics (narrow sense, broad sense).
- Concepts of system, structure, function. Structure is an ensemble of elements and relations.
- Systems biology deals with parallel characterization (or modeling) of many objects (genes, molecules, cells), hoping to understand large, complex systems. Systems approach to bioinformatics and modeling.
- Traditional (or standard) bioinformatics deals with 4 core data-types : sequence, 3D, network and text. Each has an underlying logical structure, a standard or core description, plus various simplified and/or extended (annotated) descriptions.
- Annotation is adding (textual, sometimes numerical) descriptors to structures or their parts.
- Database records of a molecule (e.g. protein) have a “structural part” that contains the *core-description* (say sequence), and an “annotation part” that is mostly human readable (e.g. bibliography) but may include references to *other structural descriptions* (secondary structure, domain architecture, computed quantities etc.)