

# Computer methods in Molecular Biology

## Trieste June 23 - 30, 2006

- Sequence database searching, theory and practice (Dave Judge and Jack Leunissen)
- Nucleic acid databases, Medline, Pubmed (David Landsman)
- Protein databases, Swissprot, Prosite (Elisabeth Gasteiger)
- EBI Services (Dave Judge)
- Gene discovery (Luciano Milanese)
- Genome analysis (Martin Bishop)
- ICGEB services in domain prediction and DNA structure analysis.



# Bioinformatics

Knowledge-representation in molecular biology

Sándor Pongor

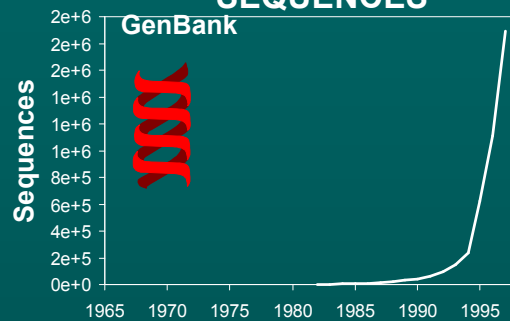
Protein Structure and Bioinformatics, ICGEB, Trieste

# **An overview of bioinformatics**

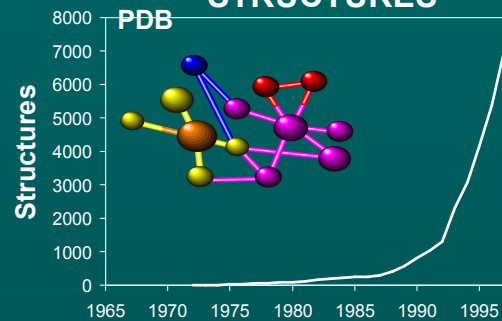
- History and development
- Models:
  - Sequences,
  - 3D structures
  - Networks
- Similarity and classification:
  - database search,
  - consensus descriptions
- Integrated resources

# Representation of biological knowledge

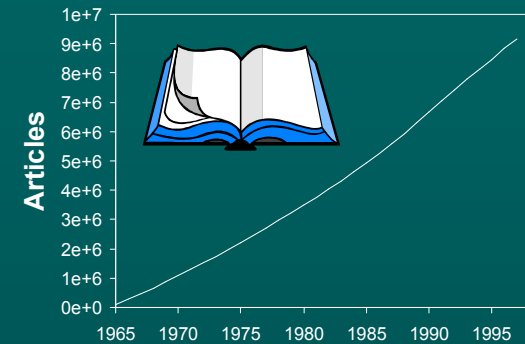
## NUCLEOTIDE SEQUENCES



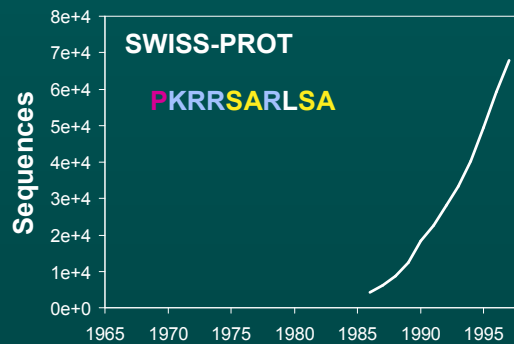
## PROTEIN 3D STRUCTURES



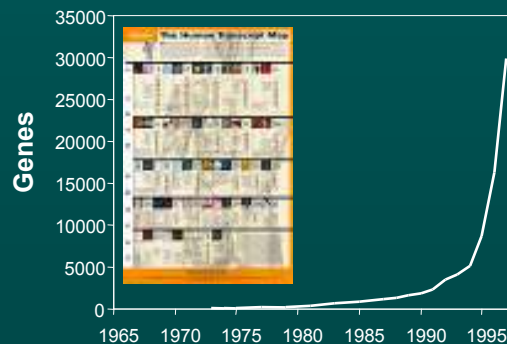
## BIBLIOGRAPHY



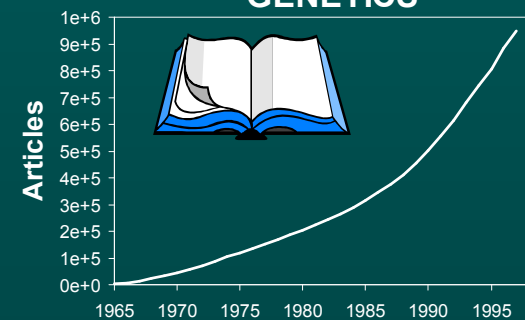
## PROTEIN SEQUENCES



## MAPPED HUMAN GENES



## BIBLIOGRAPHY-GENETICS



Source: NCBI

# Bioinformatics milestones 1

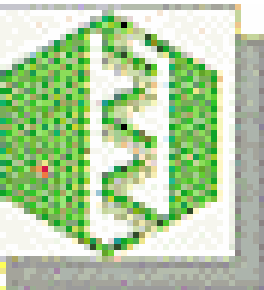
- 1962 - Pauling's theory of molecular evolution
- 1967 - Margaret Dayhoff's Atlas of Protein Sequences
- 1970 - Needleman-Wunsch algorithm
- 1977 - DNA sequencing and software to analyze it (Staden)
- 1981 - The concept of a sequence motif (Doolittle)
- 1982 - Phage lambda genome
- 1983 - Database search (Wilbur-Lipman)
- 1985 - FASTP/FASTN: fast sequence similarity searching
- 1987 - Sequence profiles
- 1987 - EMBL, Genbank, Swiss-Prot databases

# Bioinformatics milestones 2

- 1988 - National Center for Biotechnology Information (US)
- 1988 - EMBnet network for database distribution
- 1990 - BLAST: fast sequence similarity searching
- 1991 - EST: expressed sequence tag sequencing
- 1993 - Sanger Centre, Hinxton, UK
- 1994 - EMBL European Bioinformatics Institute, Hinxton, UK
- 1995 - First bacterial genomes
- 1996 - Yeast genome
- 1997 - PSI-BLAST
- 1998 - Worm (multicellular) genome
- 2000+ The rice and human genomes.
- Microarrays

# The ingredients

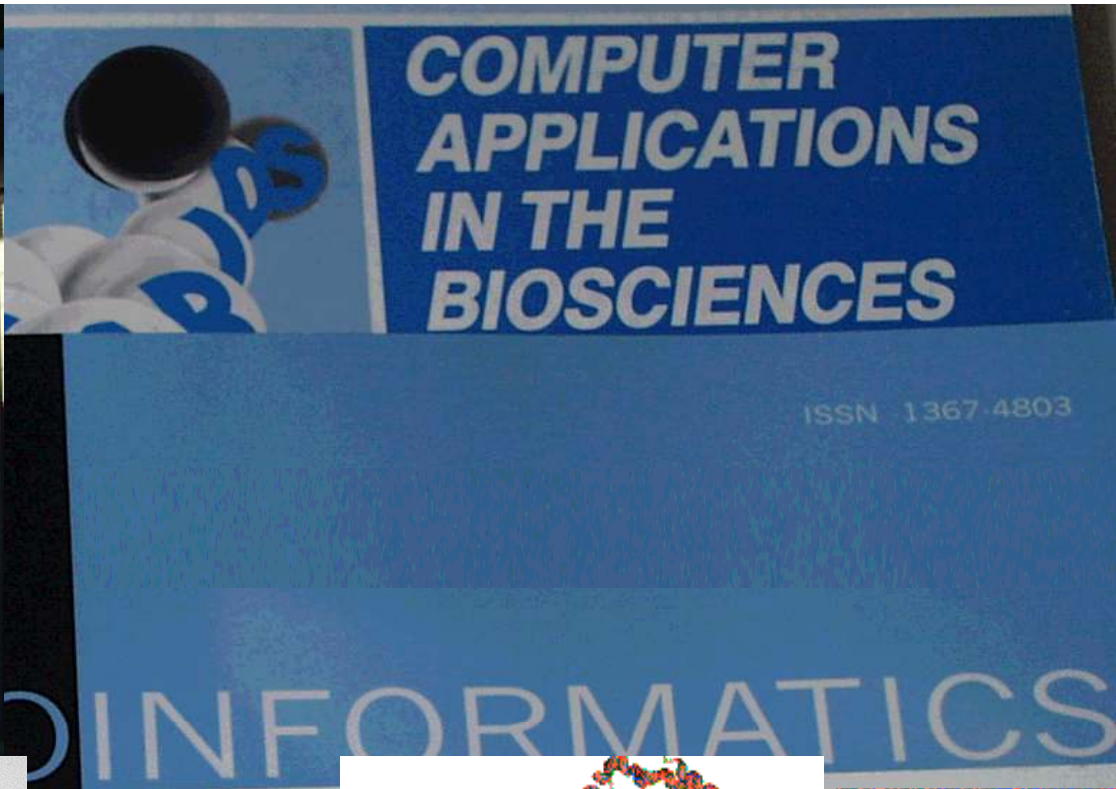
- Data collection techniques (DNA sequencing, protein sequencing, microarrays)
- Theoretical milestones (concepts of DNA structure, protein structure, evolution)
- Algorithms and programs (BLAST, FASTA)
- Databases
- Institutions
- Genomic data



EBI, Hinxton, UK



NCBI, Washington DC

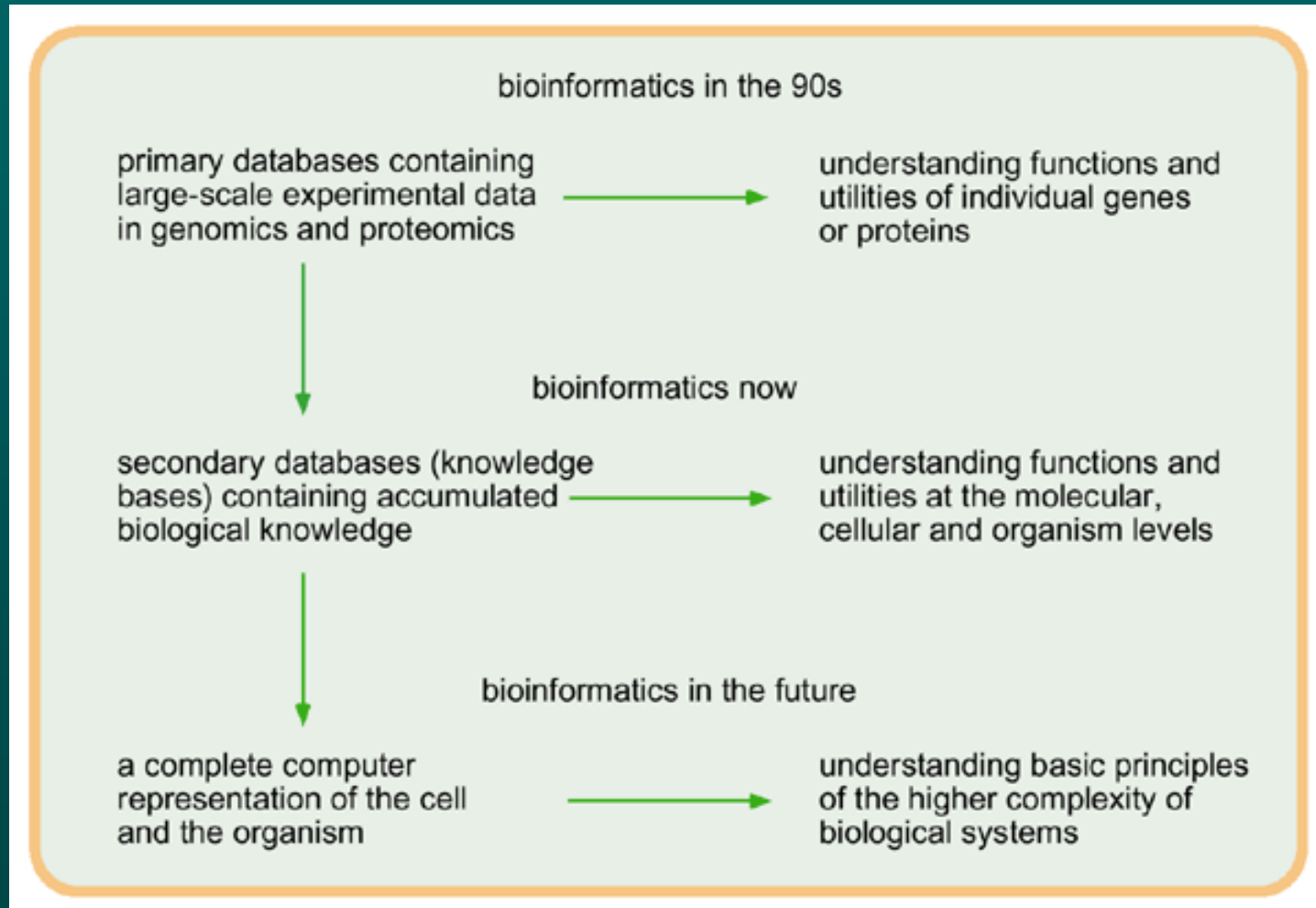


BIOINFORMATICS



Briefings in  
Bioinformatics

# The evolution of bioinformatics



# Bioinformatics is an approach to biology...



bioinformatics now

mathematics

informatics

physics

biology

chemistry

medicine

bioinformatics in the future

mathematics

informatics

physics

biology

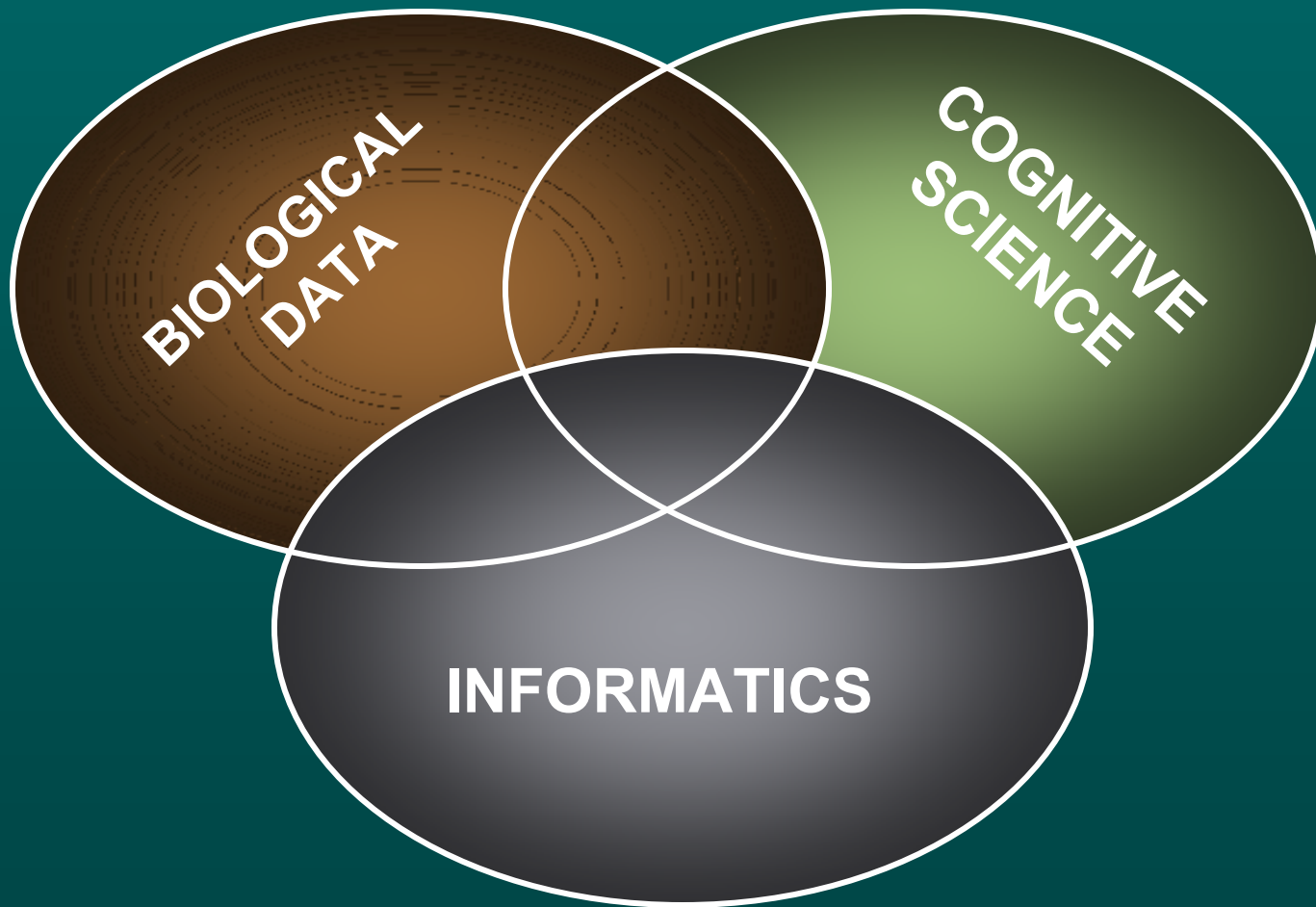
chemistry

medicine

Systems  
theory

Cognitive  
sciences

# BIOINFORMATICS



# MODELS

# Molecular structures

MARTKQTARK  
STGGKAPRKQ  
LATKAARKSA

# Sequences

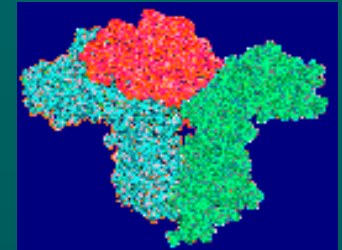
CIPKWNRCGPKMDGVPCCEPYTCTSDYYGNCS



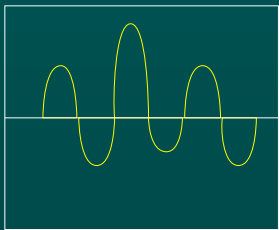
## Extended sequences (e.g. disulphide-topologies)



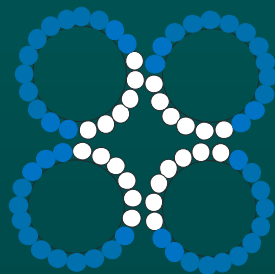
## Domain-cartoons (sec. str. cartoons)



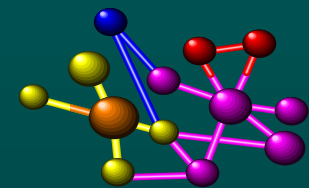
## 3D structures



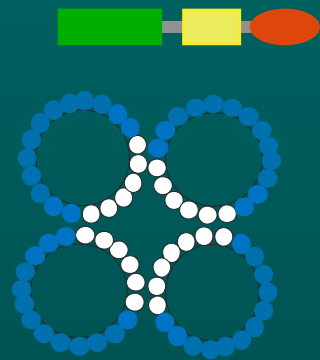
## Diagrams (hydrophobicity plots, helical circles)



## 3D cartoons



# Structures As Database Records



Identification  
Name of protein  
Organism  
Function  
Cross-references

...  
Domain structure  
Sec. structure  
Disulphides  
....

## Sequence (structure)

```
qfinetdttvivtwtpprarivgyrltvglseeg  
depqyldlpstatsvniplpgrkytnvyeise  
egeqnlilstsqttapdapdpdtdqvdtsivvr  
wsrprapitgyrivyspsvegsstelnlpetansv  
tldsdlqpgvqynitivyaveenqestpvfiqqettg  
vprsdkvppprdlqfvevtdvkitimwtppespvt  
gyrvdvipvnlpghehgqrlpvsrntfaevtglspg  
vtyhfkv
```

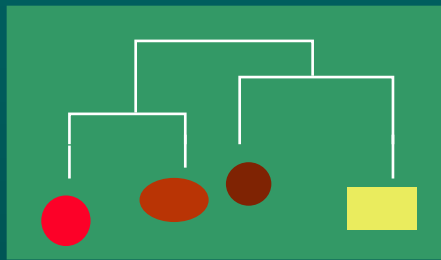
**ANNOTATIONS**

CIPKWNRCGPKMDGVPCCEPYTCTSDYYGNC

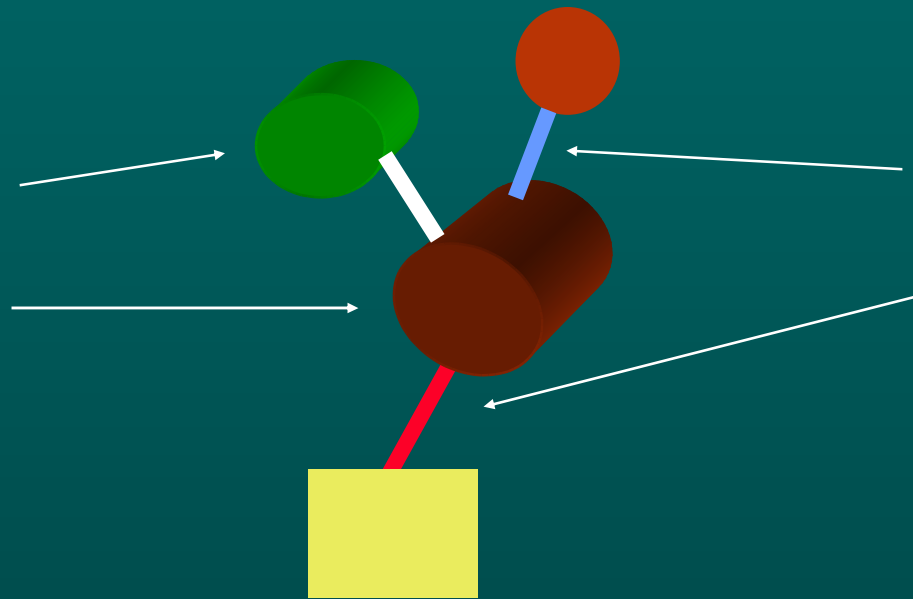
**SEQUENCE  
OR STRUCTURE**

Database record, fields

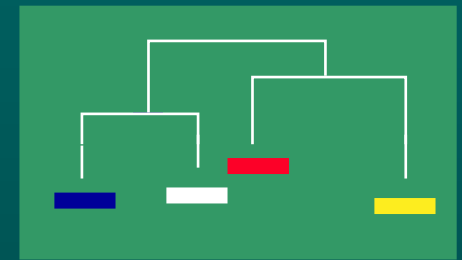
# A structural model



**Substructures**



**Structure**

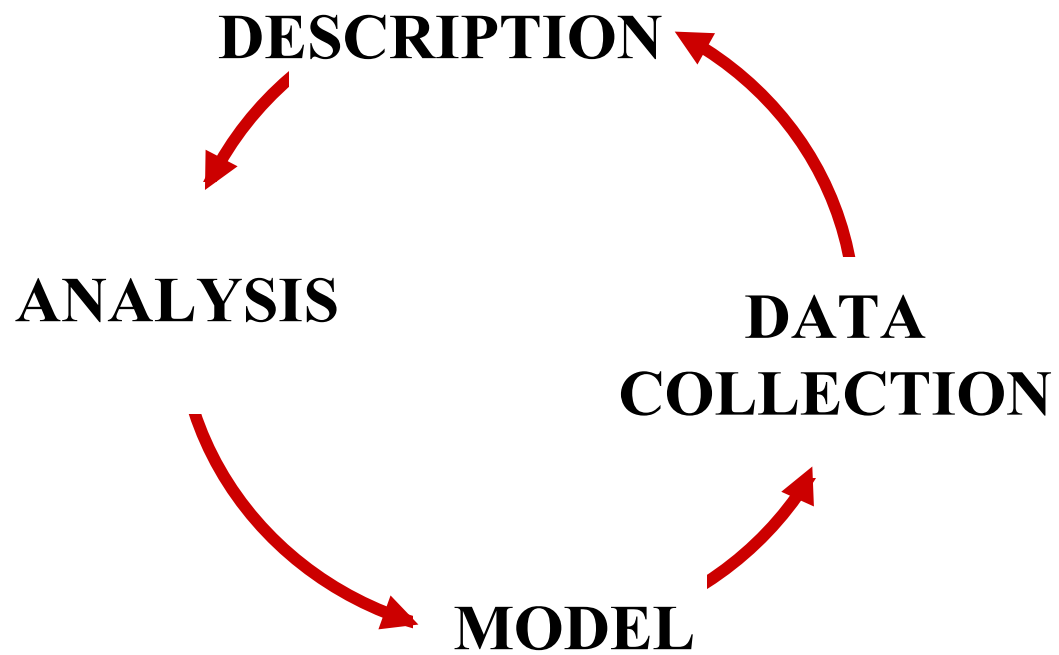


**Relationships**

Substructures, relations, rules = ontology

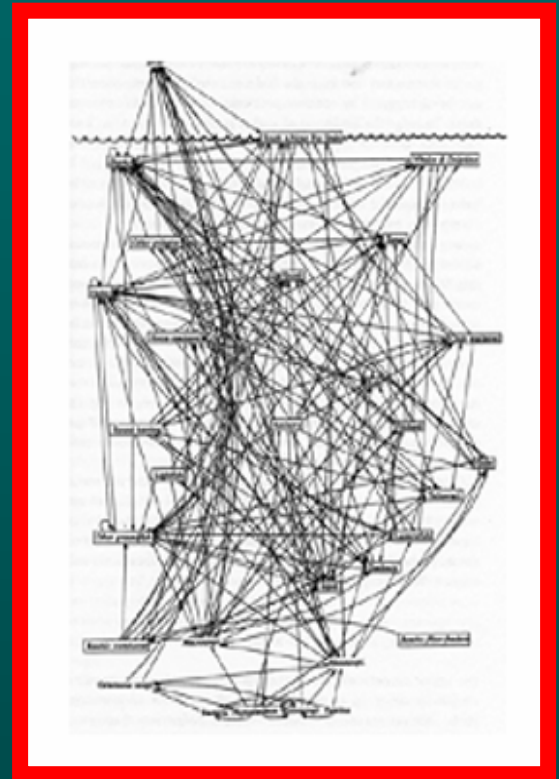
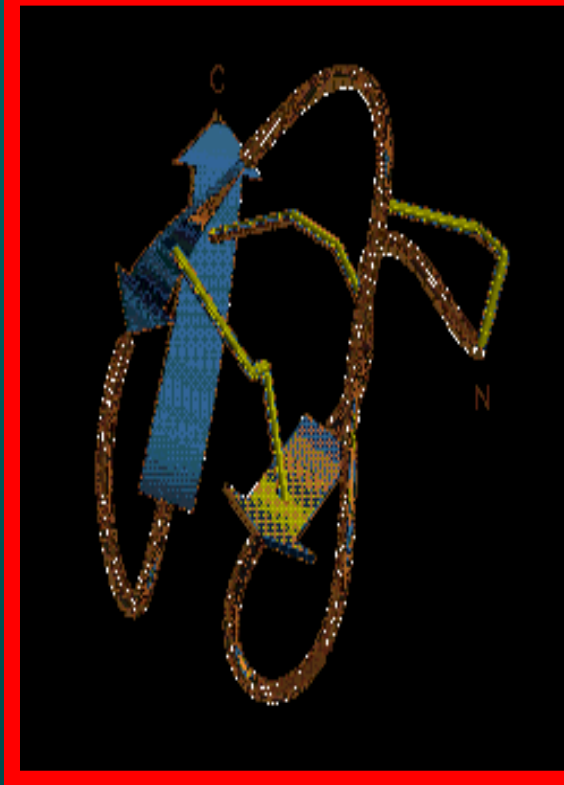
Entity-relationship model  
*Pongor, Nature, 1987*

# Molecules change



# Models

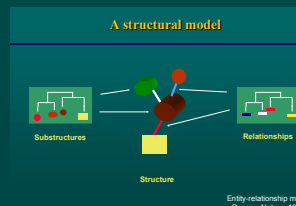
tassfvvswvsasdtvsgfrvey  
elseegdepqyldlpstatsvni  
pdllpgrkytvnvyeiseegeqn  
lilstsqttapdapdptvdqvd  
dtsivvrwsrprapitgyrivys  
psvegsstelnlpetansvtlsd  
lqpgvqynitivyaveenqestpv  
fiqqettgvprsdkvppprdlqf  
vevtdvkitimwtppespvtgyr  
vdvipvnlpgehgqrlpvsrntf  
aevtglspgvtyhfkvfavnqgr  
eskpltaqqatkldaptnlqfin  
etdttvivtwtpprarivgyrlt  
vgltrggqpkqynvgpaasqypl  
rnlqpgseyavslvavkgnqqsp  
rvtgvfttlqplgsiphyntevt  
ettivitwtppaprigfklgvrps  
qggeaprevtsesgsivvsgltp  
gveyvytisvlrdgqerdapivk



SEQUENCES

3D STRUCTURES

NETWORKS



# ■ SEQUENCES

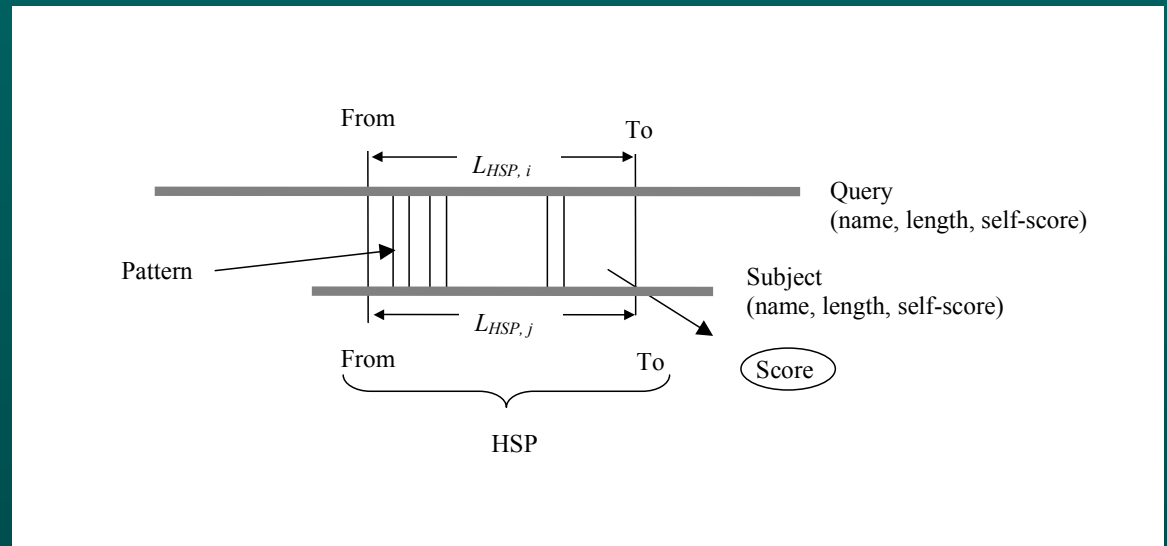
# Sequences as language

```

qfinetdttvivtwtpprarivgyrl
tvglleseegdepqyldlpstatsvni
pdllpgrkytnvveyeiseegeqnlil
stsqttdapdpdptvdqvddtsivv
rwsrprapitgyrivyspsvegsste
lnlpetansvtlsdlqpgvqynitiy
aveenqestpvfiqqettgvprsdkv
ppprdlqfvevtdvkitimwtpesp
vtgyrldvipvnlpghehgqrlpvsrn
tfaevtglspgvtvhfkvfavnqgre
skpltaqqatkldaptnlqfinetdt
tvivtwtpprarivgyrltvgltrgg
qpkqynvgpaasqyplrnlpqgseya
vslvavkgnqgsprvtgvfttlqplg
siphyntevtettivitwtpaprigf
klgvrpsqggeaprevtsesgsivvs
gltpgveyvytisvlrdgqerdapiv
kkvvtplspptnlhleanpdtgvltv
swersttpditgyritttptngqqgy
sleevvhadqssctfenlspgleynv
svytkddkesvpisssfvsvwsas
dtvsgfrveyelseegdepqyldlps
tatsvniplpgrkytnvveyeisee

```

## Alignments



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

# LANGUAGE

## The language of bibliographies

Entrez PubMed - Microsoft Internet Explorer

Address: [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&opt=Abstract&list\\_uids=15236959](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&opt=Abstract&list_uids=15236959)

NCBI PubMed National Library of Medicine NLM

Search PubMed for [Go] [Clear]

Field: Title

Display: Abstract Show: 20 Sort: Send to: Text

1: J Mol Biol. 2004 Jul 23;340(5):957-64.

**Periodic transcriptional organization of the E.coli genome.**

**Kepes F.**

ATelier de Genomique Cognitive, CNRS UMR8071/genopole, Evry, France. francois.kepes@genopole.cnrs.fr

The organization of transcription within the prokaryotic nucleoid may be expected to both depend on and determine the chromosome. Indeed, immunofluorescence localization of transcriptional regulators has revealed foci in actively transcribing cells. Furthermore, structural and biochemical approaches suggest that there are approximately 50 independent loops of DNA in the E. coli nucleoid. Here I show that in four E. coli strains, genes that are controlled by a sequence-specific transcriptional regulator tend to be encoded at regular distances that are multiples of 1/50th of the chromosome length. This periodicity is consistent with a solenoidal epi-organization of the chromosome, which would gather into foci the interacting partners; the regulator would bind to specific sites. Binding at gemine regulatory sites on DNA would thus be optimized by co-transcriptionally translating the binding sites.

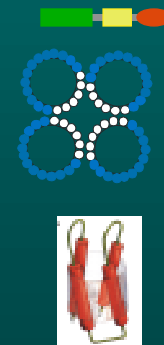
PMID: 15236959 [PubMed - indexed for MEDLINE]

Display: Abstract Show: 20 Sort: Send to: Text

Write to the Help Desk  
NCBI NLM NIH  
Department of Health & Human Services  
Privacy Statement | Freedom of Information Act | Disclaimer

Keyword-collections, ontologies, etc.

## Structures As Database Records



Identification  
Name of protein  
Organism  
Function  
Cross-references

...  
Domain structure  
Sec. structure  
Disulphides  
....

Sequence (structure)

```
qfinetdtvtvwtpprariygyrltvglseeg  
depgyldlpstatsvnipldlpgrkytnvyeise  
egeqnlilstsqttapdapdpdptvdqvdtsivvr  
wsrprapitgyrivyspsvegssteilnlpetafsv  
tlldlpgvgvgnitiyaveengestpvgiqgett  
vprsdkvppprdlqfvevtdvkitimwtppespvt  
gyrvdvipvnlpgehgqrlpvsrntfaevtglspg  
vtyhfkv
```

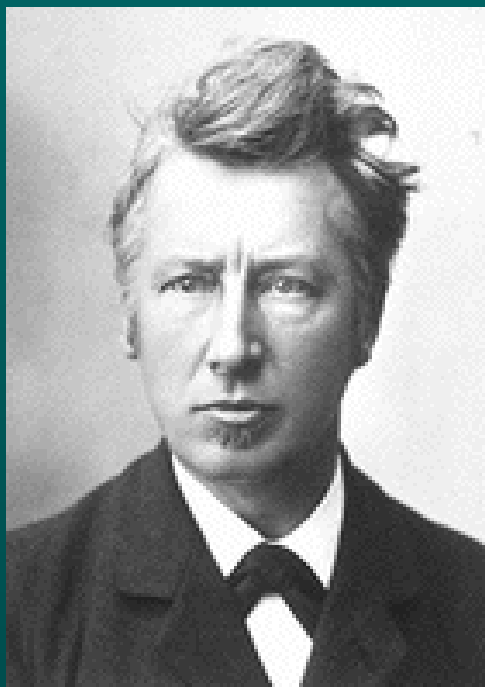
ANNOTATIONS

SEQUENCE  
OR STRUCTURE

Database record, fields

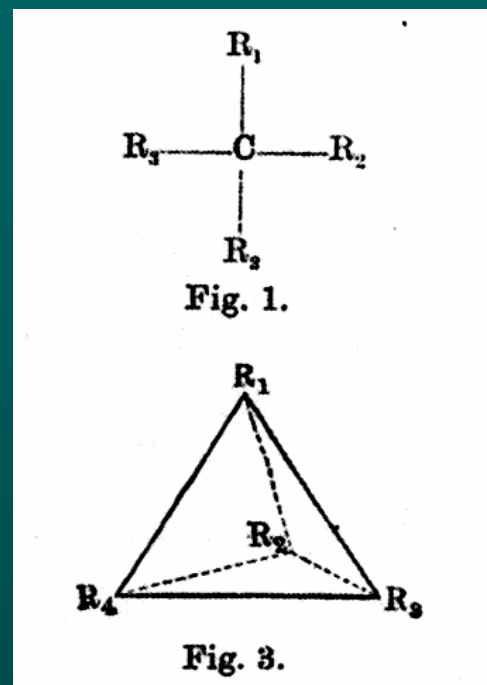
## ■ 3D STRUCTURES

# Chimie dans l'espace



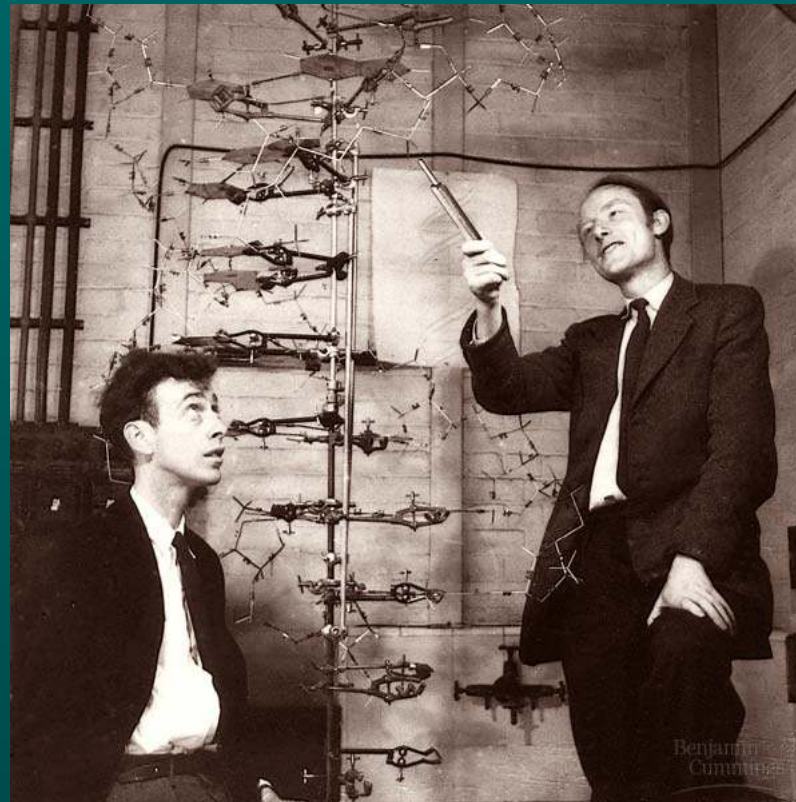
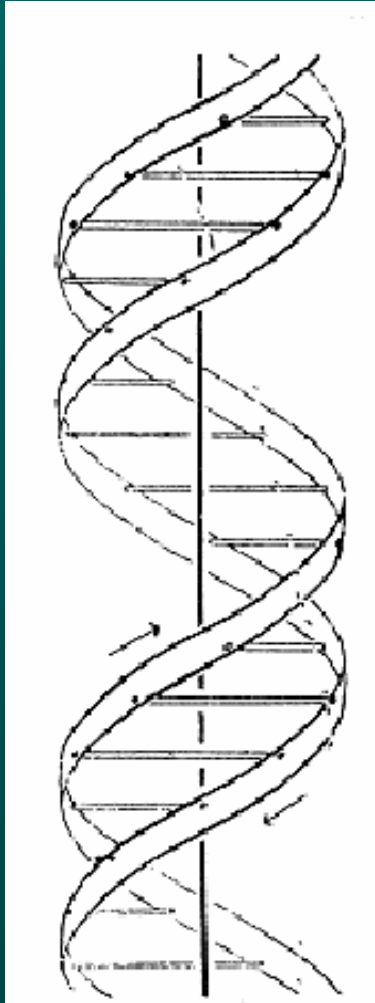
Van t'Hoff

1852-1911



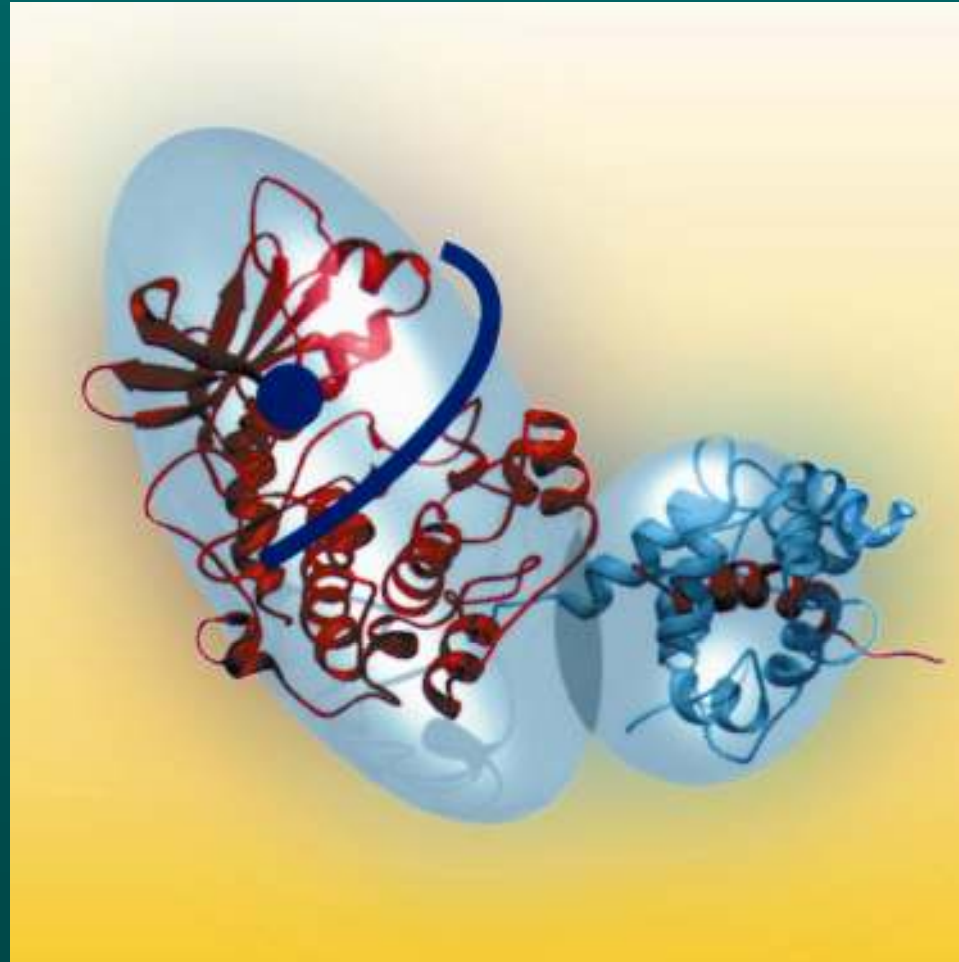
1898

# Some molecules are more equal than others...



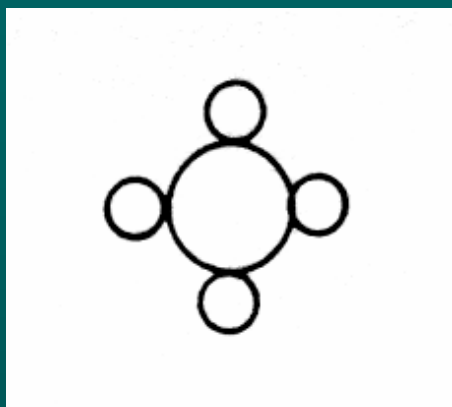
..."This figure is purely diagrammatic. The two ribbons symbolize the the phosphate-sugar chains, and the horizontal rods the pairs of the bases holding the chains together. The vertical line marks the fibre axis"

# Protein models

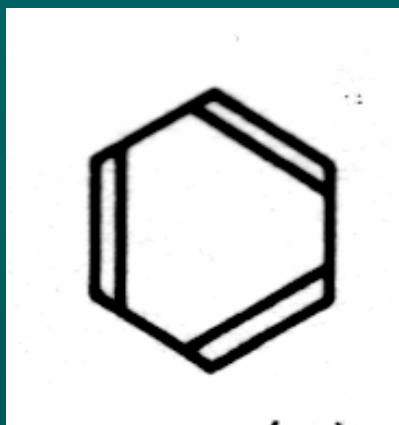


# ■ NETWORKS

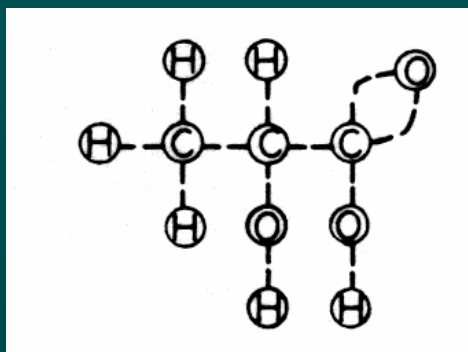
# Small molecules – classical graphs



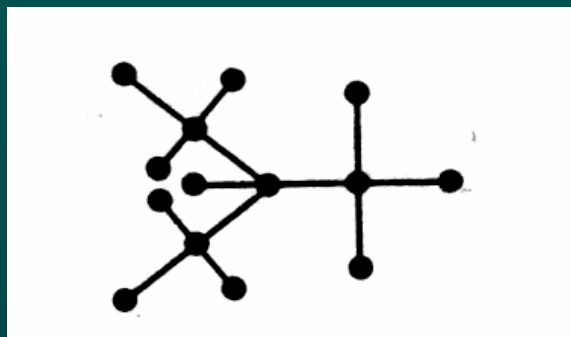
Loschmidt, 1861



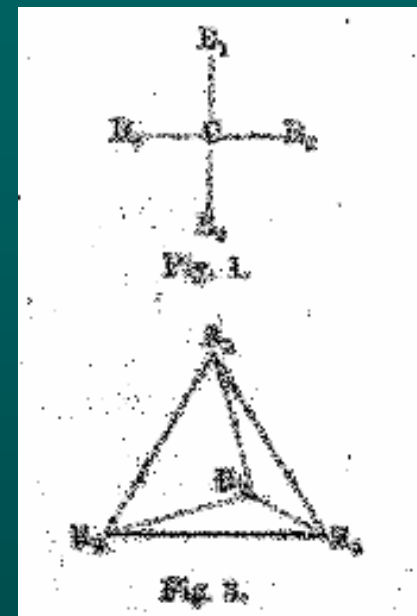
Kekulé, 1865



Crum Brown, 1861



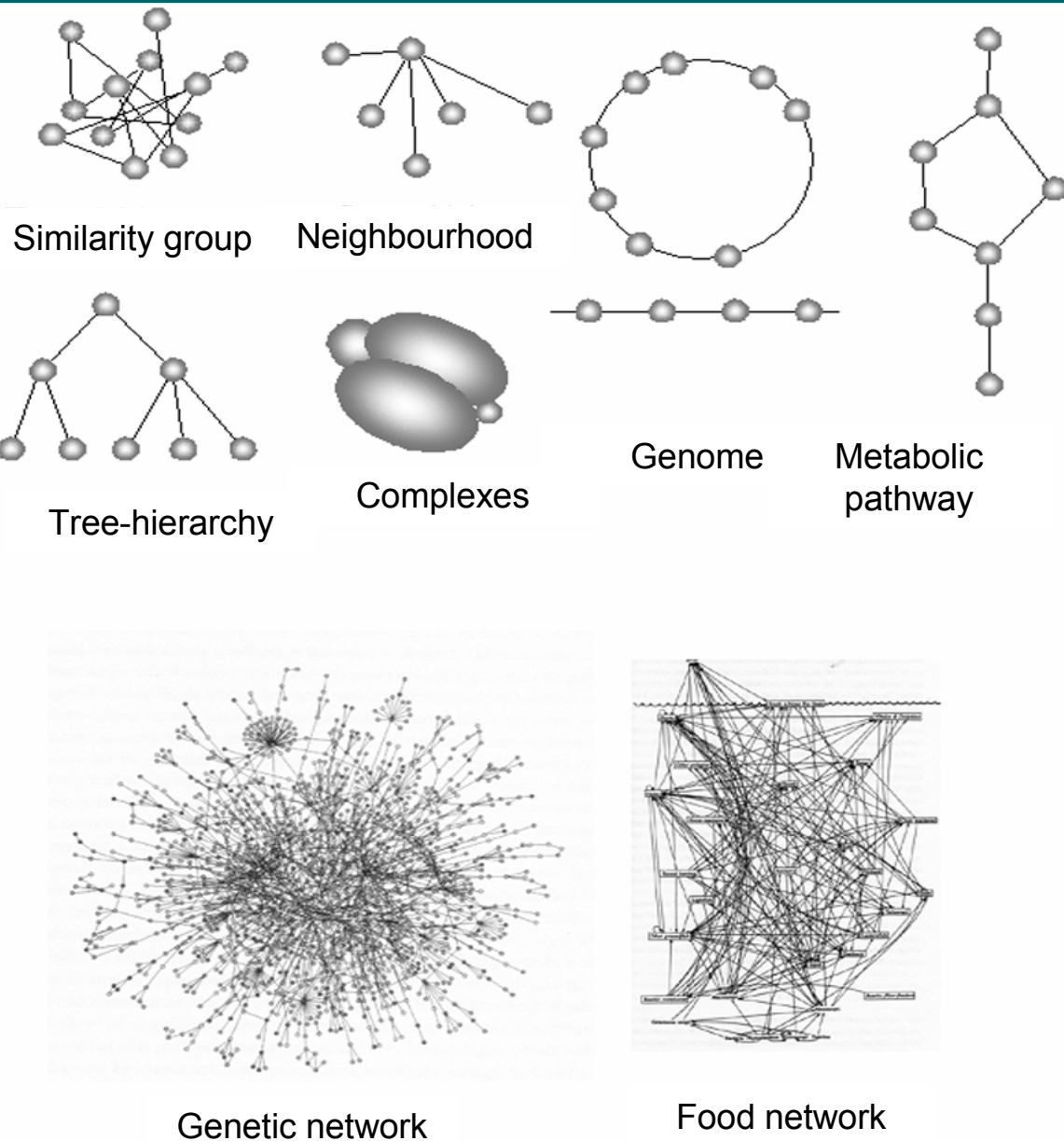
Cayley, 1872



Van't Hoff, 1898

# TOPOLOGIES, GRAPHS

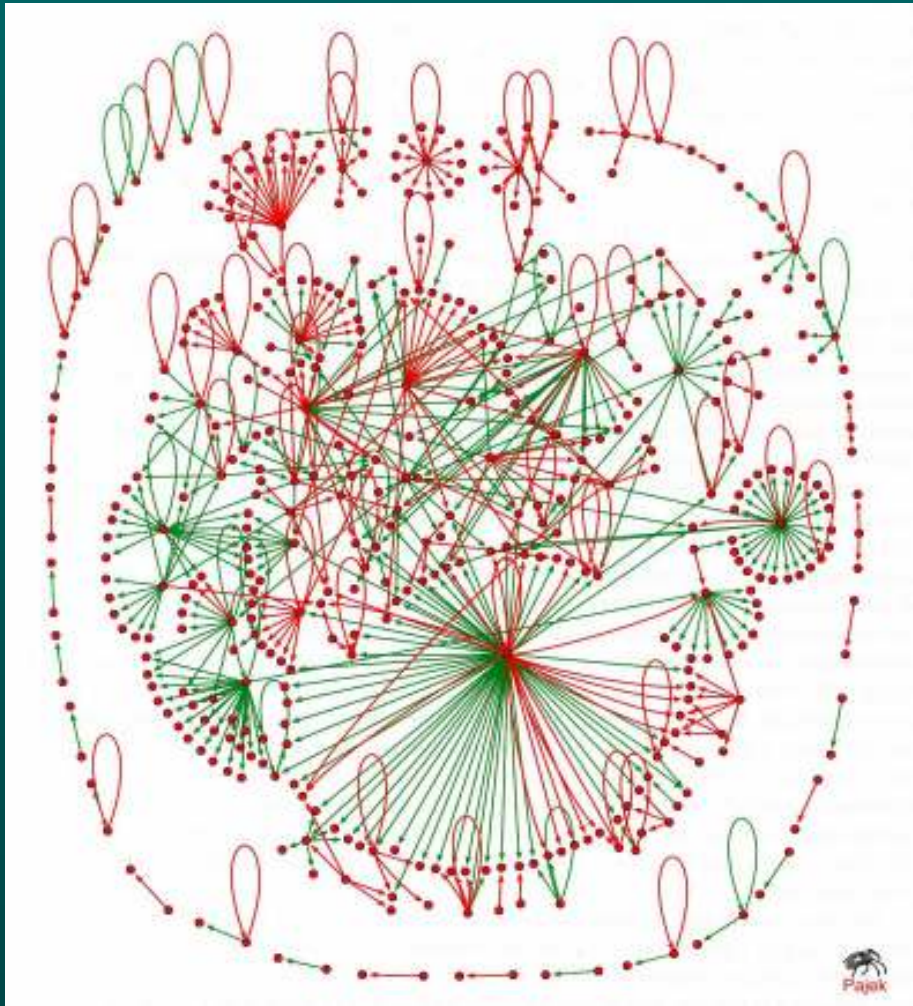
## Genomes, assemblies



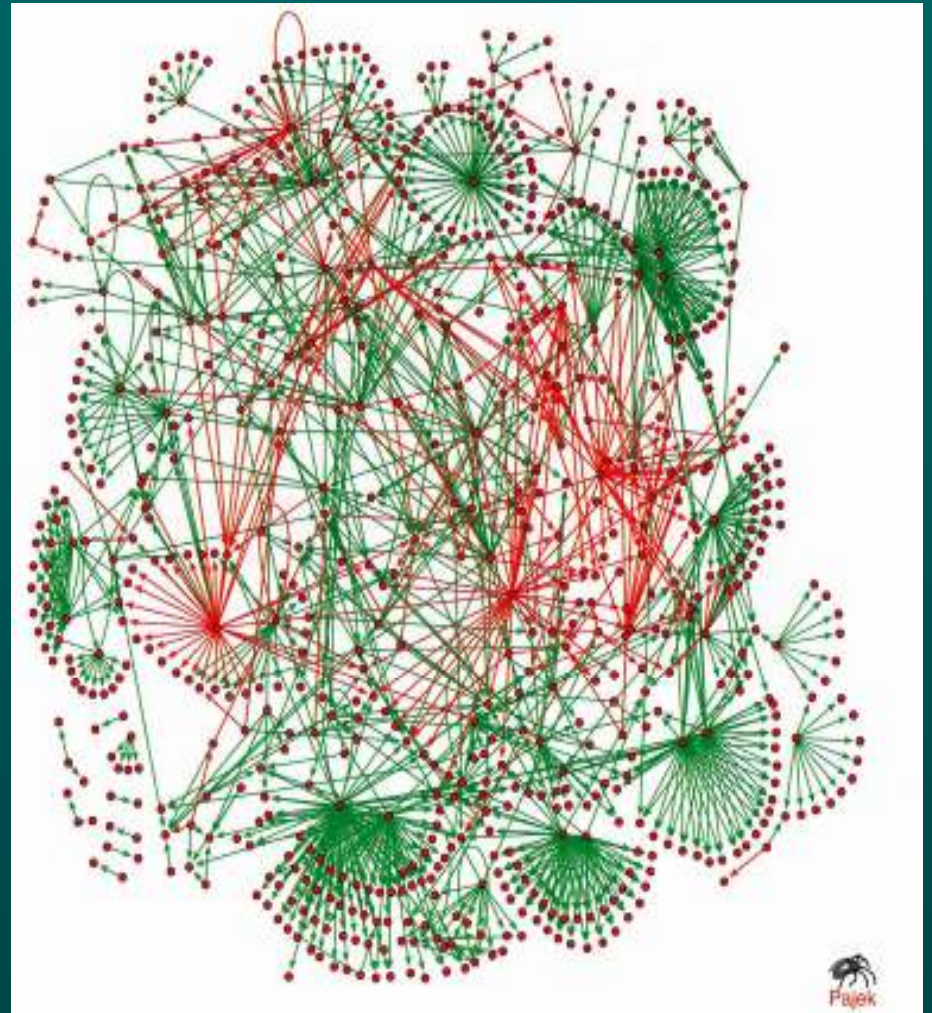
**Entity-relationship models**  
**Topological meta-models**

# The transcription regulatory networks

+ (up)  
- (down)



*E. coli*



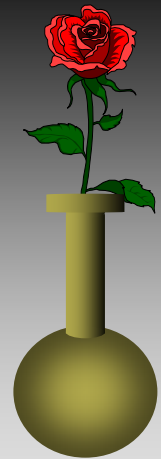
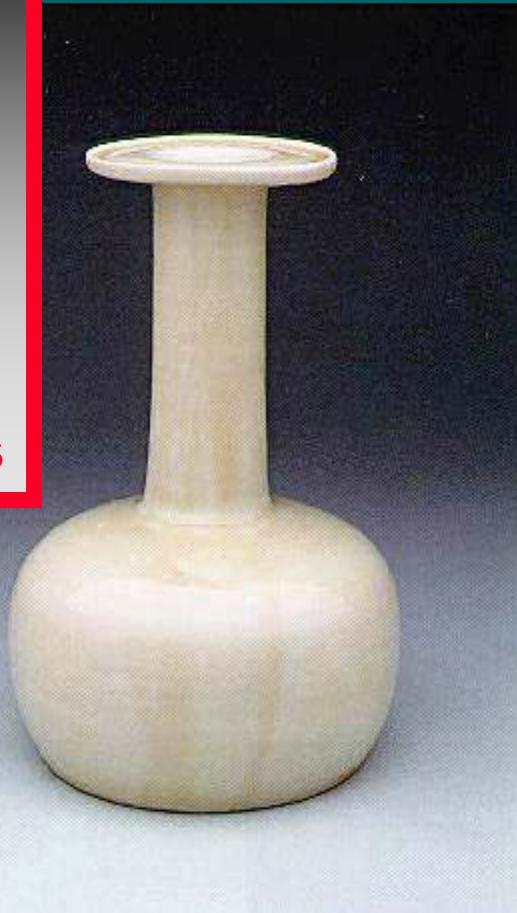
*S. cerevisiae*

# **SIMILARITY, CLASSIFICATION**

# The concept of similarity I



Shared parts

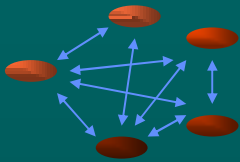


Shared context

...easier if modular

# Multiple Objects

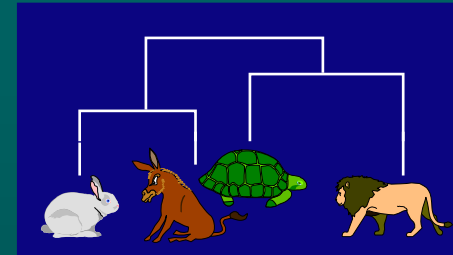
## Structural similarity



Similarity groups  
or neighborhoods

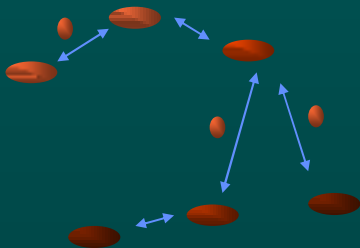
```
CGPK-MDGVPCCEPY  
CGGQNWSGPTCCASG  
CSPTSYN---CCR--  
CSRLMY---DCCT--  
CIPYYL---DCCEPL
```

Multiple alignments

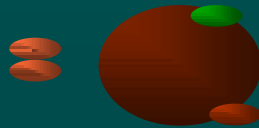


Evolutionary trees

## Context (function)



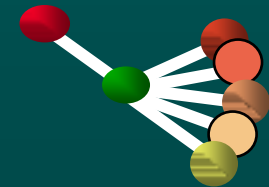
Metabolic pathways



Subunit structures,  
ligands

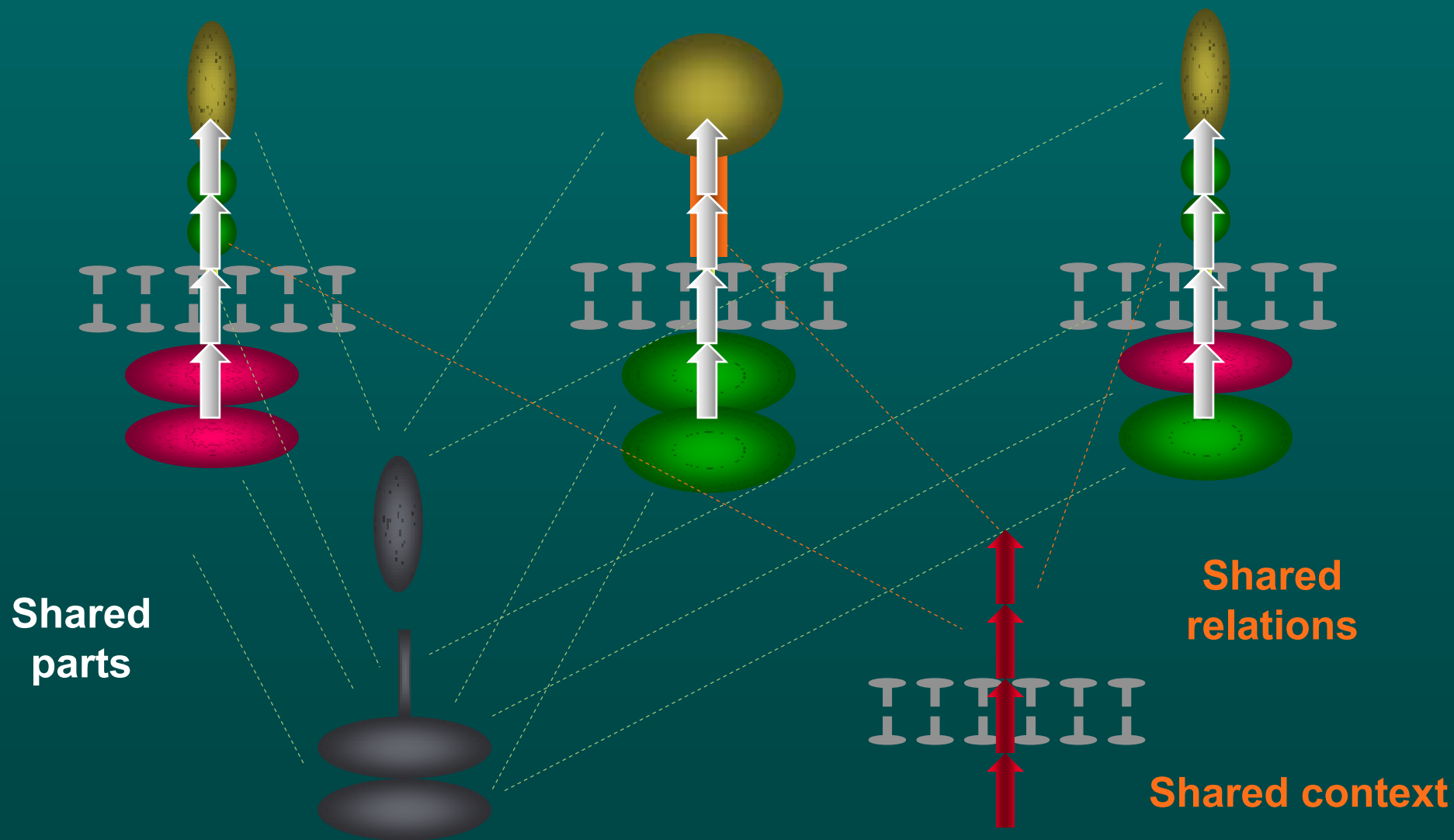


Genomes

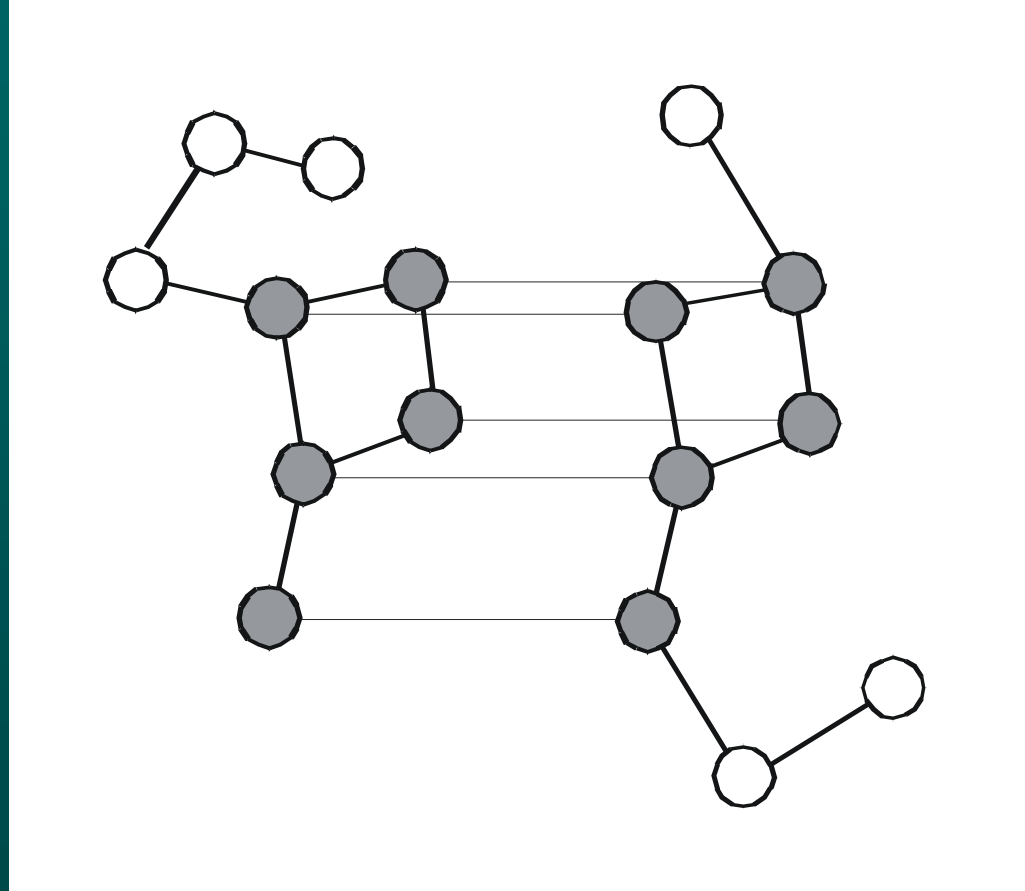


Trajectories

# Similarity of molecules



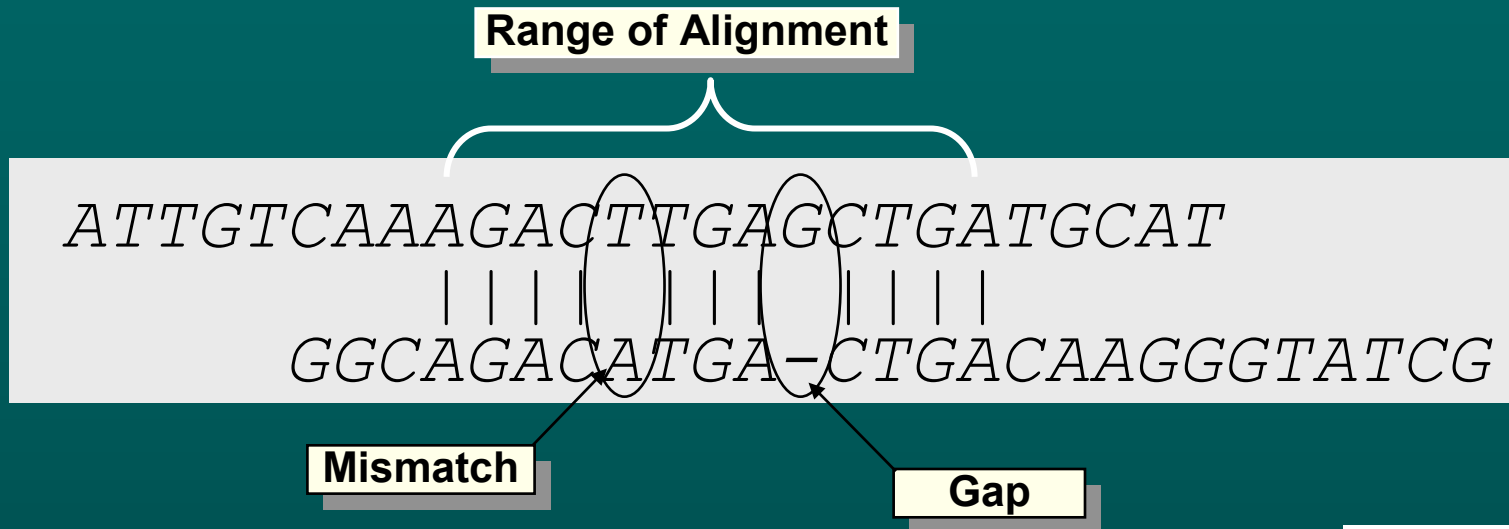
# Substructure identity ~ similarity



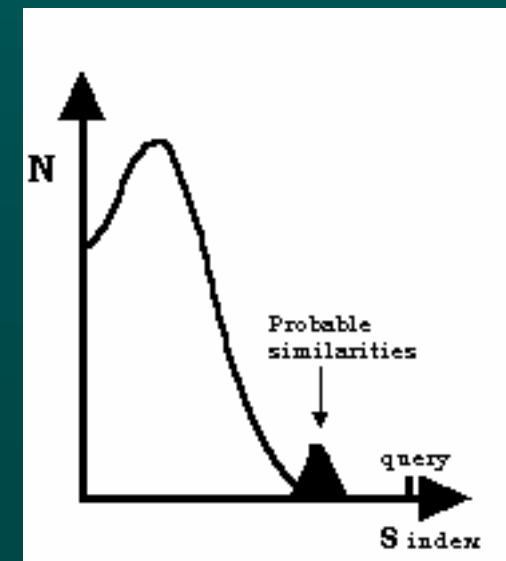
”The similarity of objects can be best described as partial identities of components and relationships

*Erich Goldmeier, The similarity of perceived forms, 1936*

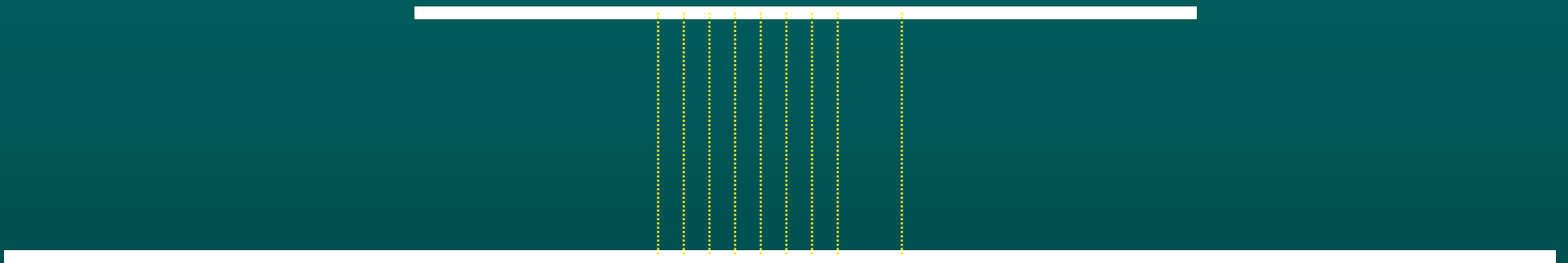
# Quantification of sequence similarity



Score  $\sim$  sg like an edit distance

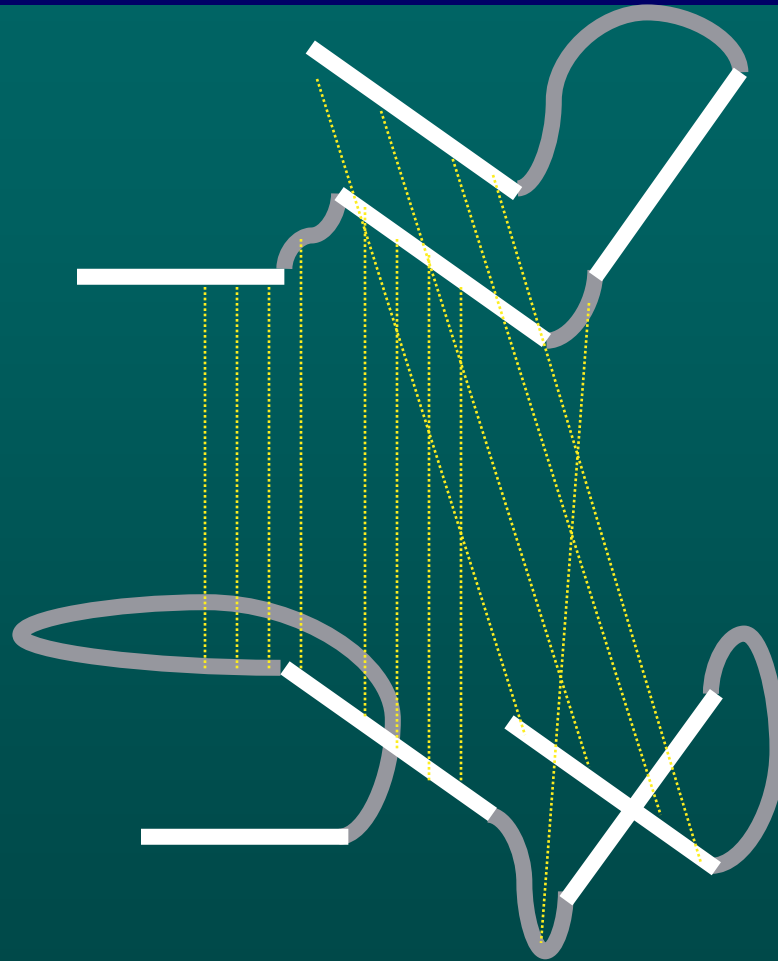


# Sequence comparison



1. Find region of alignment (fast)
2. Calculate similarity score (fast)

# 3-D comparison



1. Find region of alignment (slow)
2. Calculate similarity score (fast)

## Using similarity: Comparing one sequence with a group (database)

# BLAST program

## Similarities??

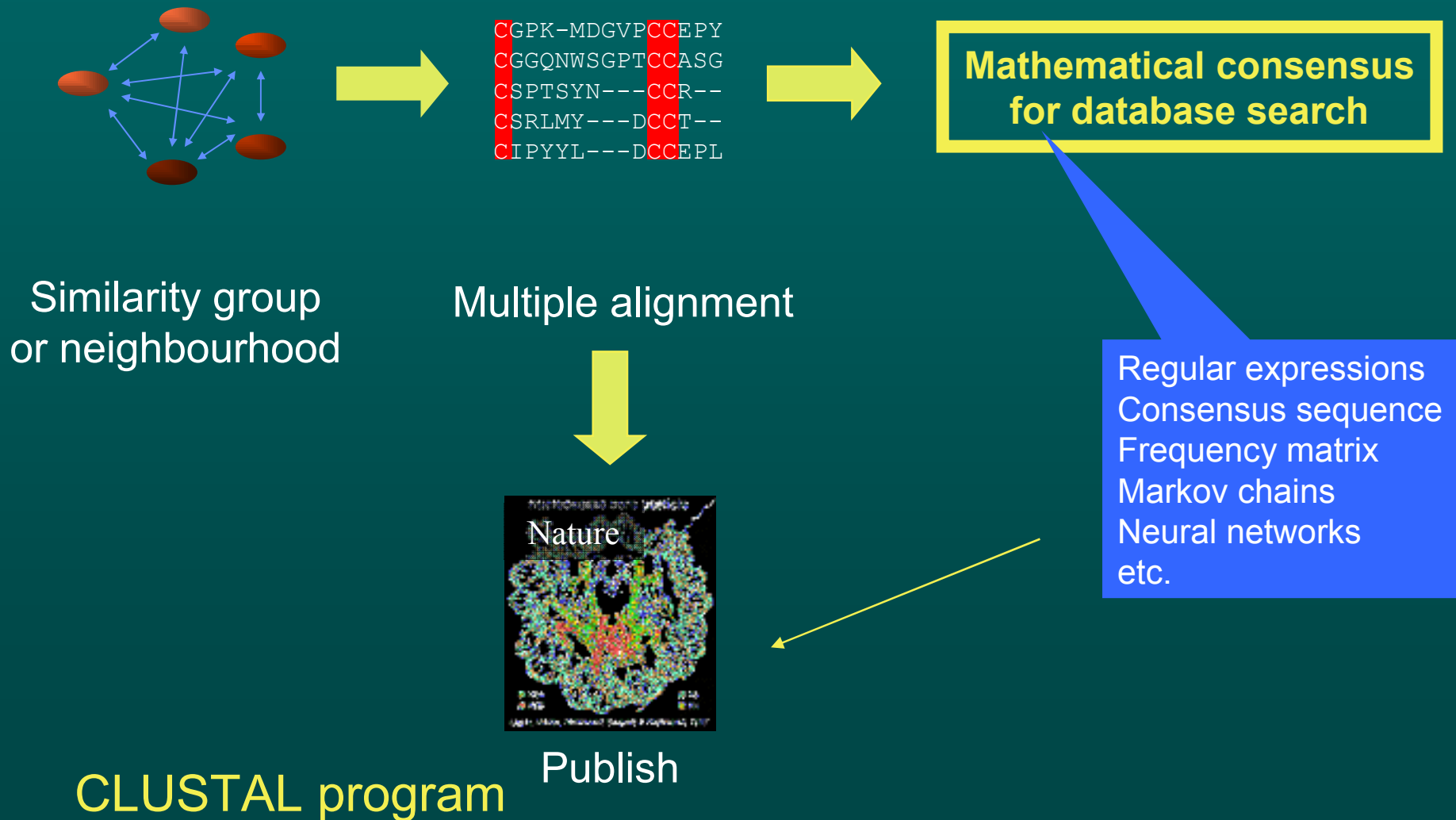
## Ranked list of best similarities

## Twilight zone

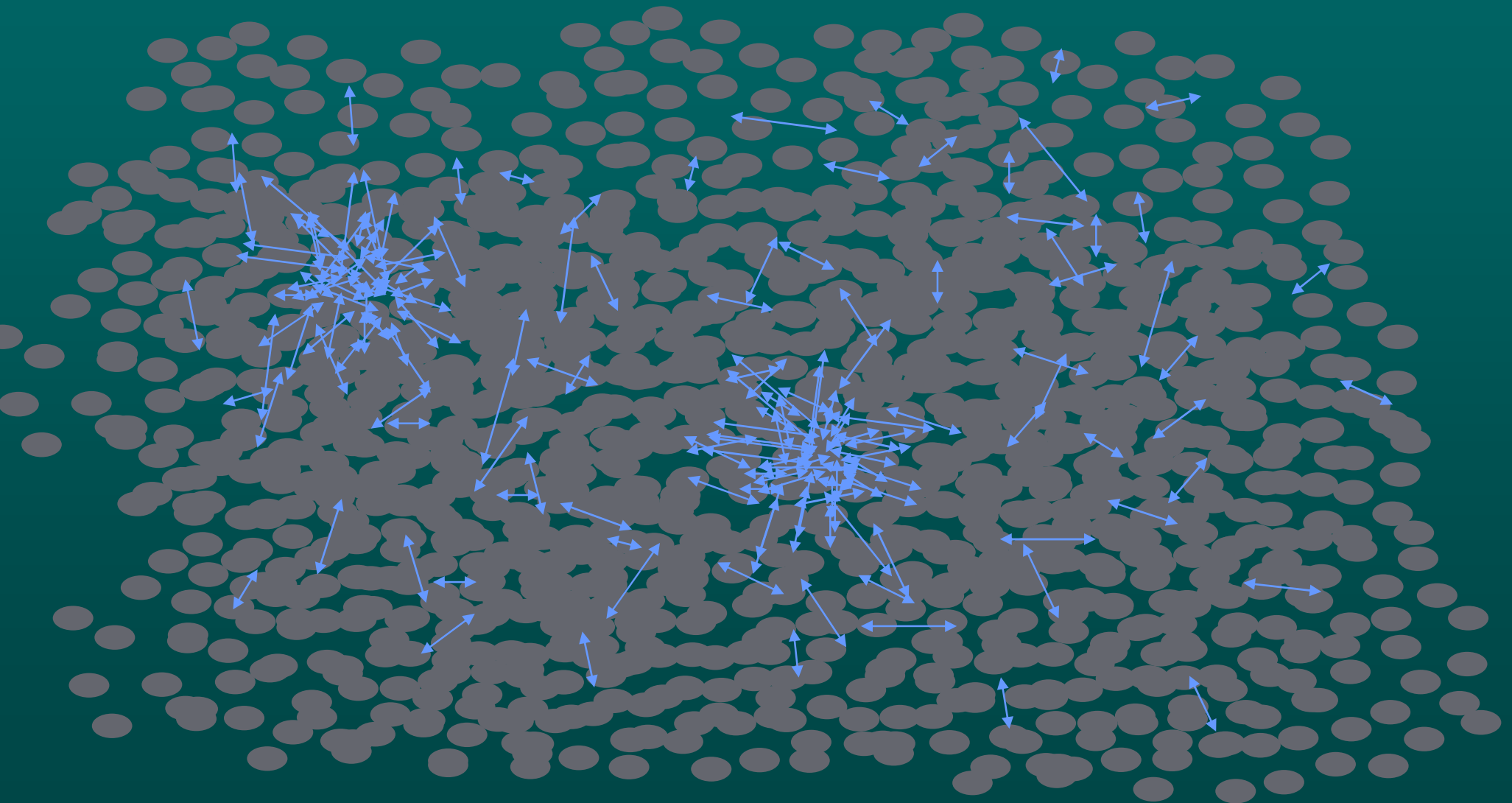
- 1
- 2
- 3
- 4

SEQUENCE SCORE	DESCRIPTION
MISSALT:17 457.36	ALPHA-AMYLASE INHIBITOR A7
MISSALT:0428 152.82	CELLULOSE BINDING PROTEIN
MISSALT:GX 145.77	EXOGALUCANASE I PRECURSOR
MISSALT:Q126 145.66	CELLULASE (EG 3.2.1.91)

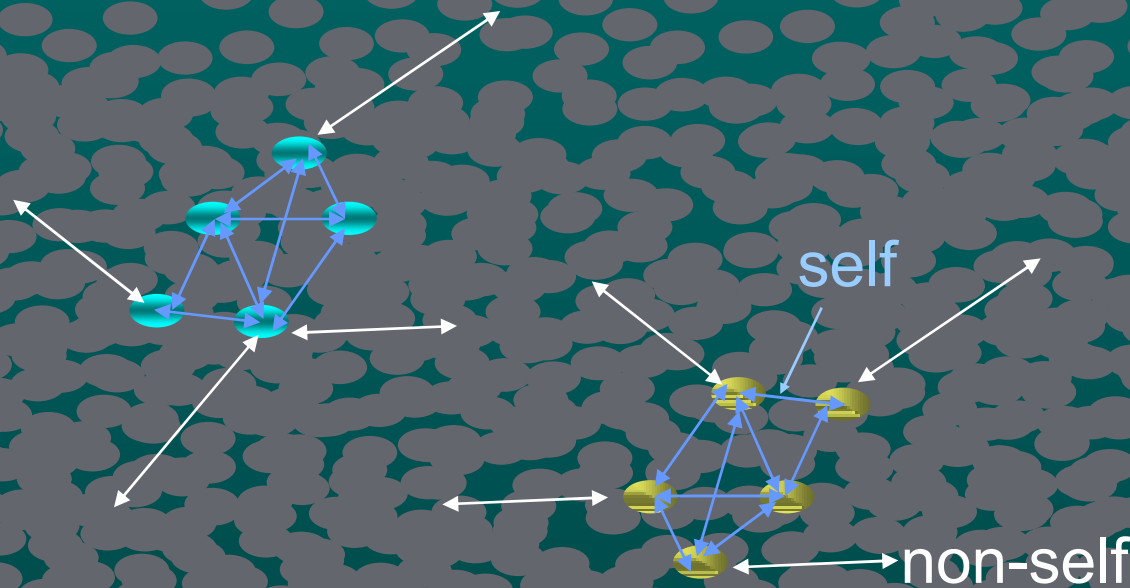
# Using similarity: comparing a group with itself



# The database as network of similarities: A memory network model



# Das Wohltemperierte Database: Similarity network as a knowledge representation



Knowledge-base of meaningful similarities  
Signal and noise distinguished


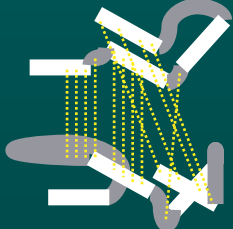

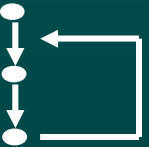
# Similarities: a practical overview



SEQUENCES

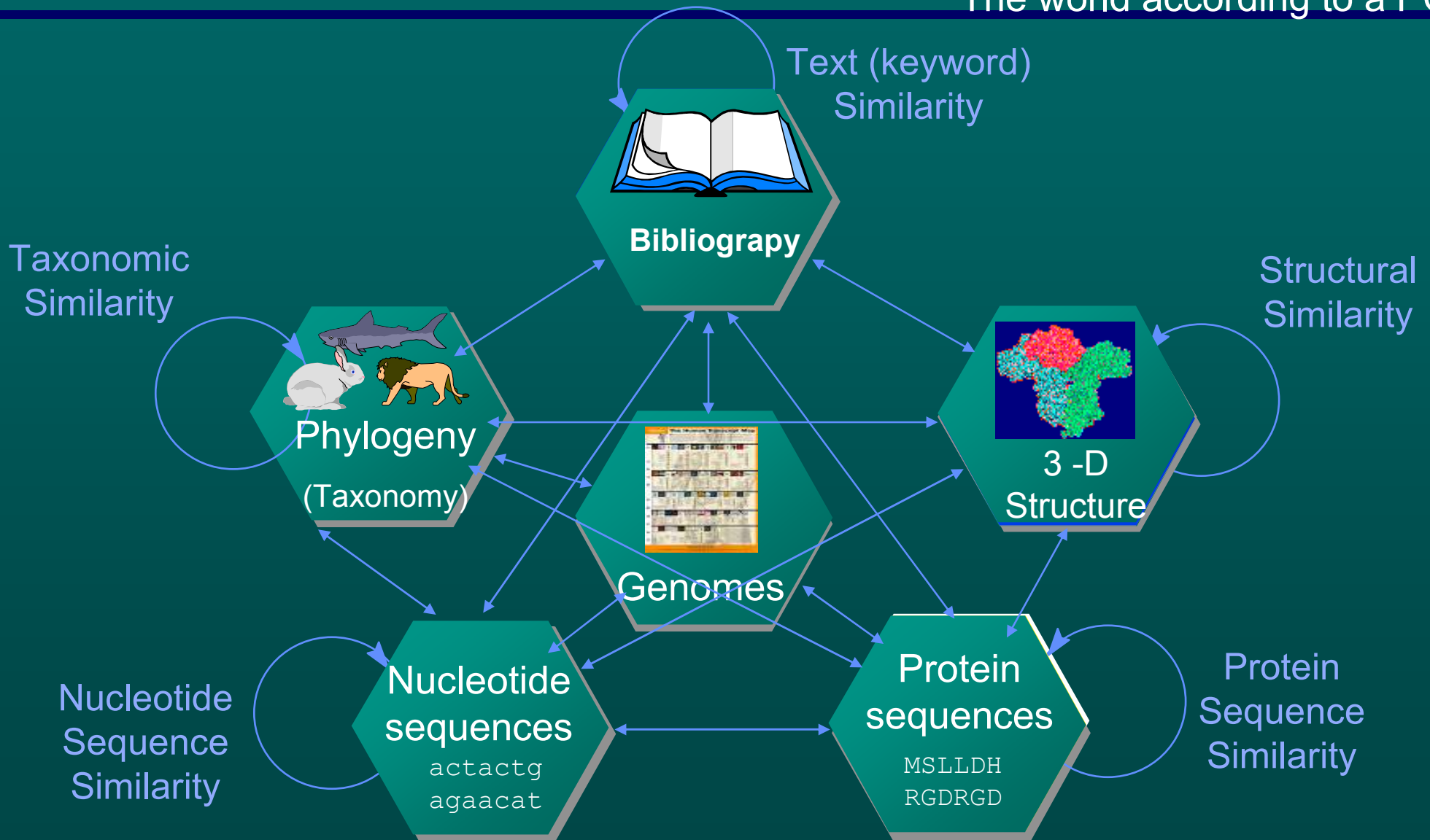
3D STRUCTURES

NETWORKS

Bulk	“Glycine-rich”	“ $\alpha$ -helical”	“scale-free”
Substructure-alignment			???
Motifs	G-RR		

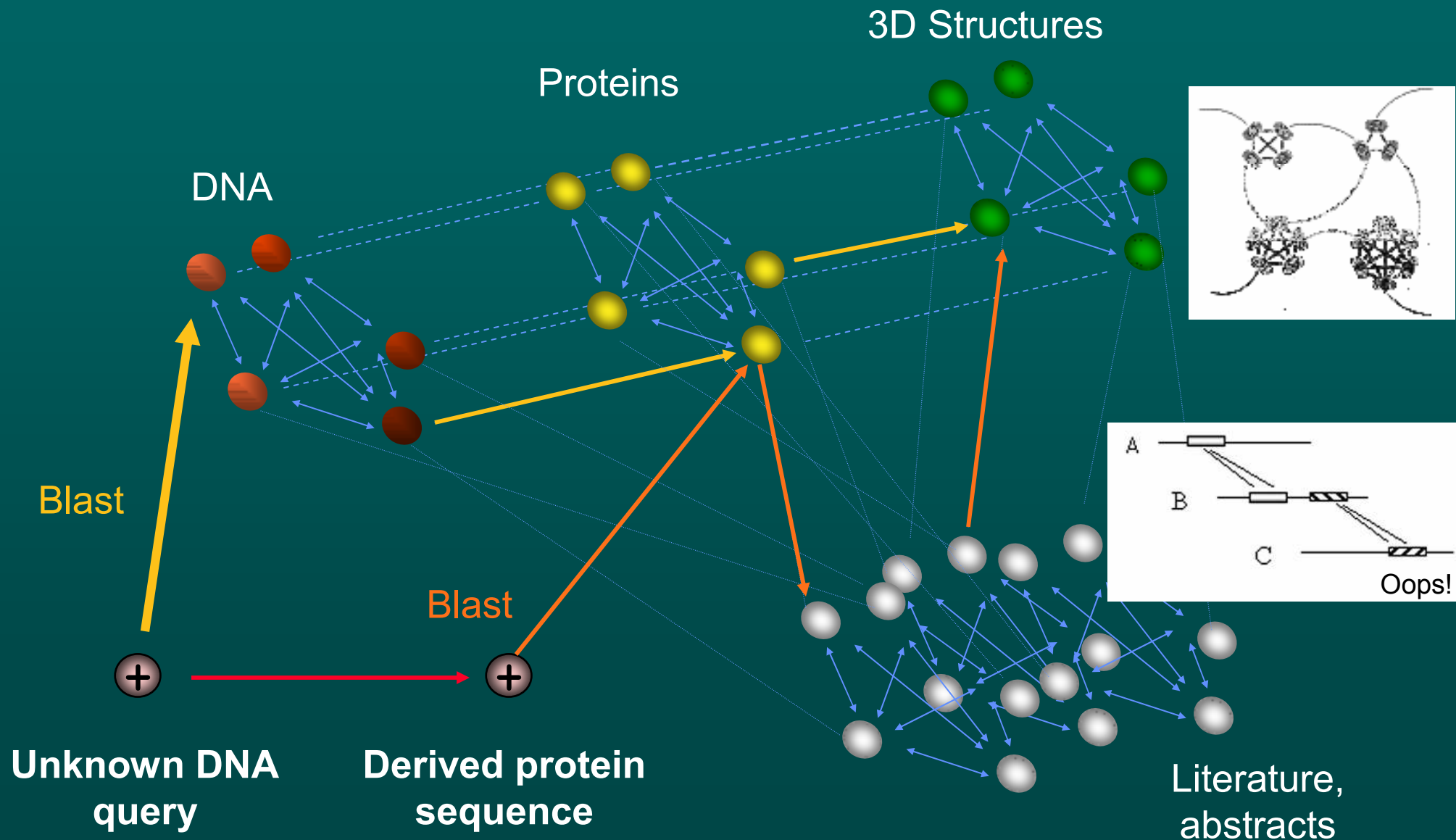
# Biological knowledge as a network of data

The world according to a PC...

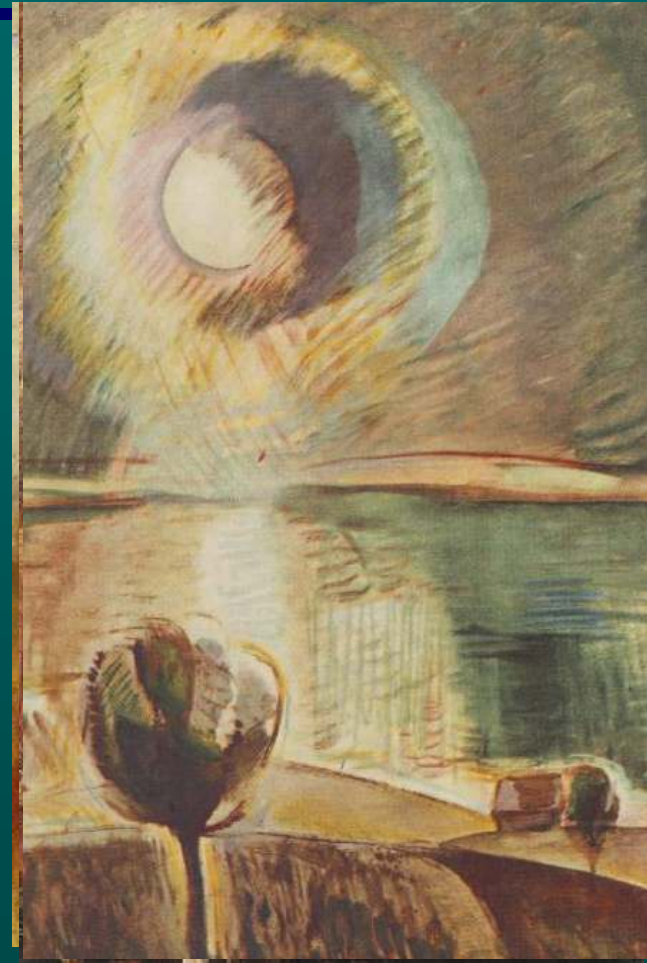


Source: NCBI

# Search on a preprocessed, integrated database: the importance of a good neighbourhood

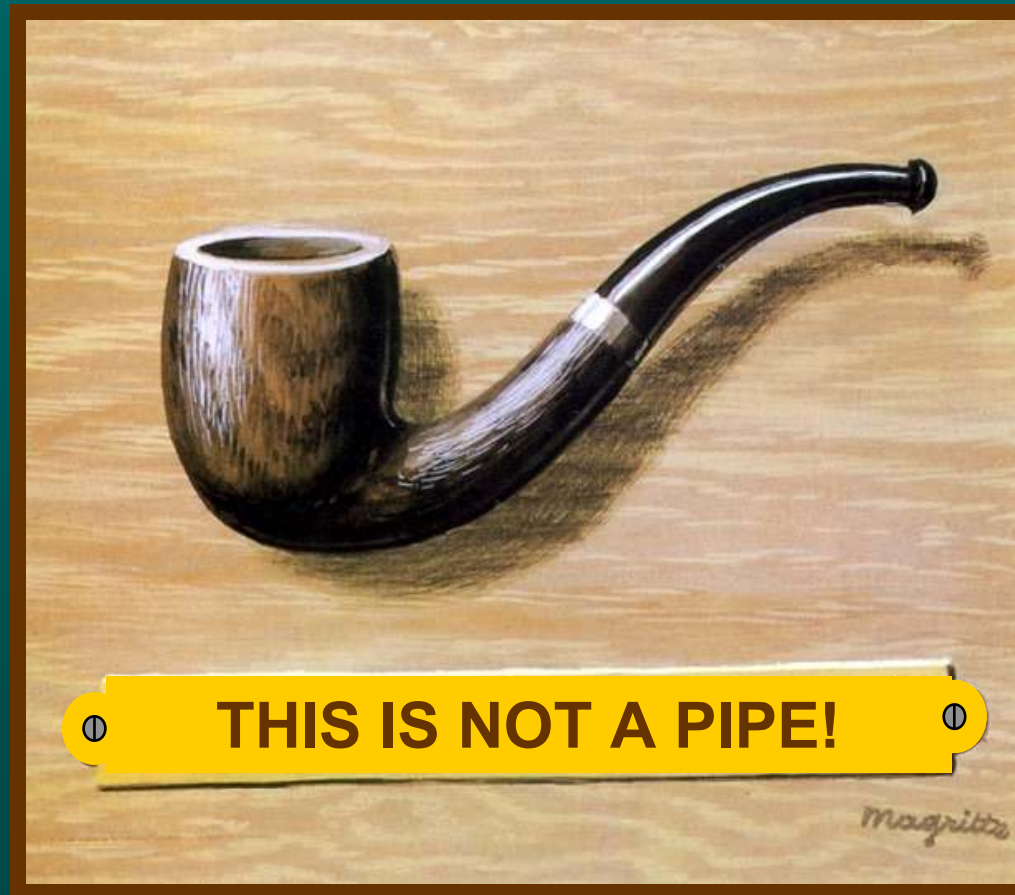


# The concept of similarity II



...Easy for humans, hard for computers

# Models are human constructs...



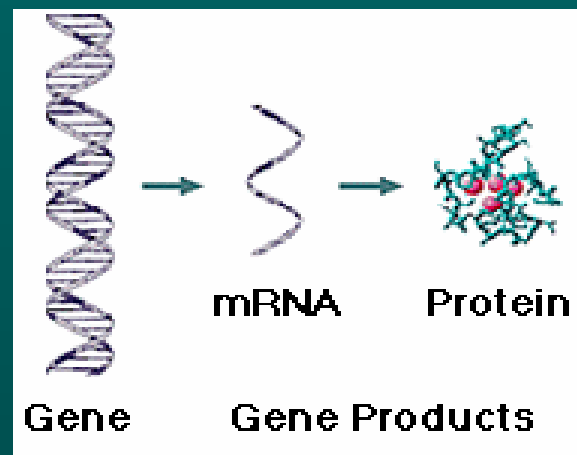
# Models are human constructs...



THIS IS NOT A MOLECULE



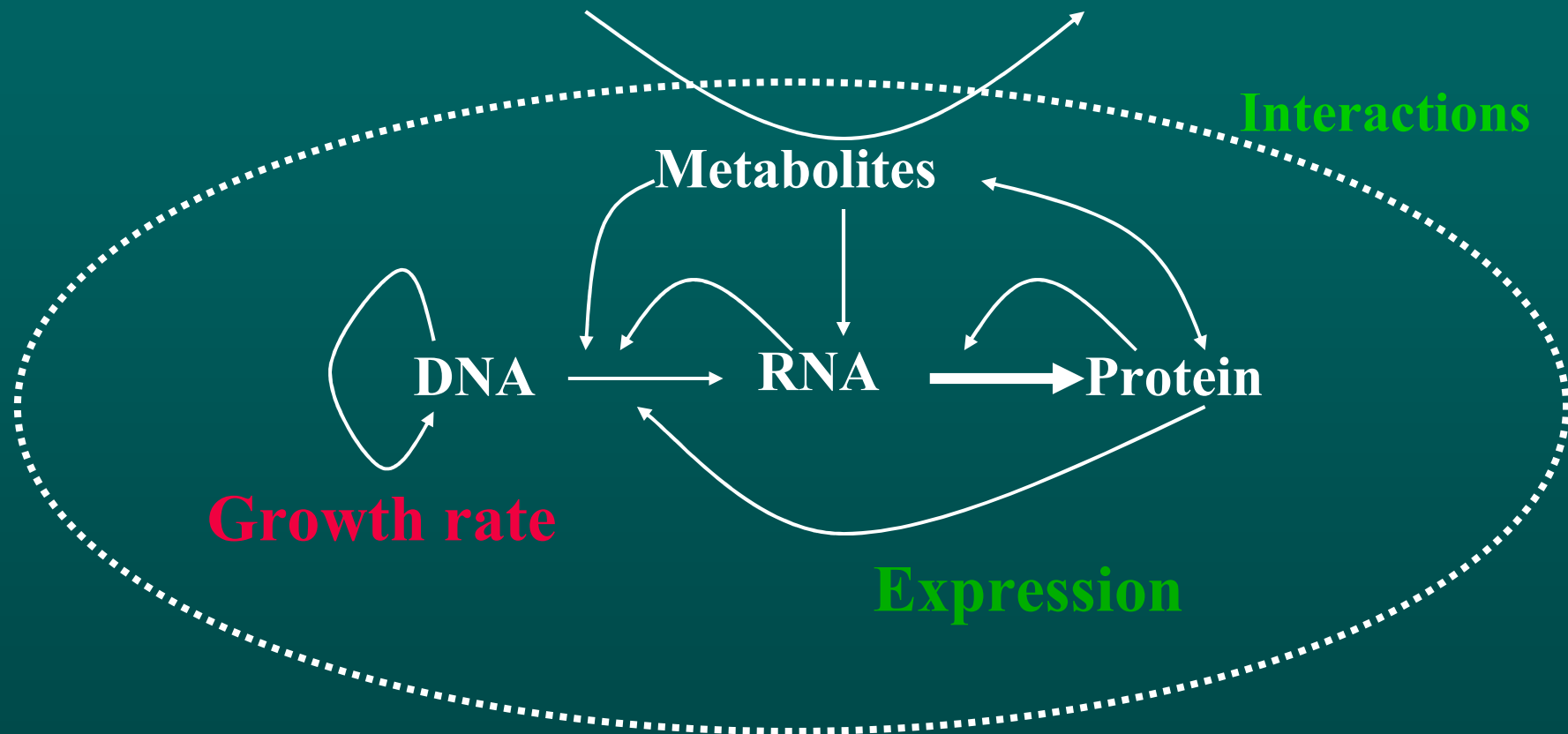
# The central dogma:



DNA → RNA → Protein

Dogma, paradigm, mythology

# New central dogma: Self-assembly, catalysis, replication, networks



**Polymers:** Initiate, elongate, terminate, fold, modify, localize, degrade

**Evolution + Self assembly, Systems biology**

# Summary of topics discussed

- History and development
- Models:
  - sequences,
  - 3D structures
  - Networks
- Similarity and classification:
  - database search,
  - consensus descriptions
- Integrated resources, knowledge integration

# Summary of the introduction

- Bioinformatics is the science of biological information or rather a computer-based approach to biological problems.
- All kinds of biological data are structures defined with entities and relationships (metabolites, genes, networks).
- Typical tasks: Similarity search, categorization and clustering
- Simultaneous handling of many, complex datatypes

# On-line help to this lecture

- Bioinformatics tutorials on-line

<http://www.ebi.ac.uk/2can/home.html>

- ICGEBnet

<http://www.icgeb.org/~netsrv/>

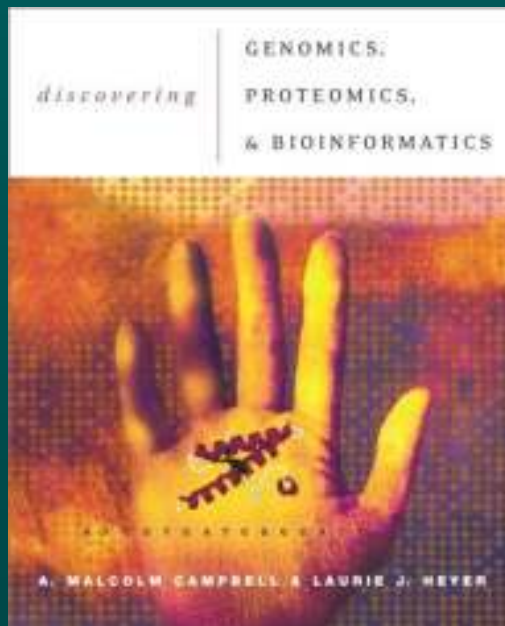
- The Trieste bioinformatics course

<http://www.icgeb.org/~netsrv/netcourse.html>

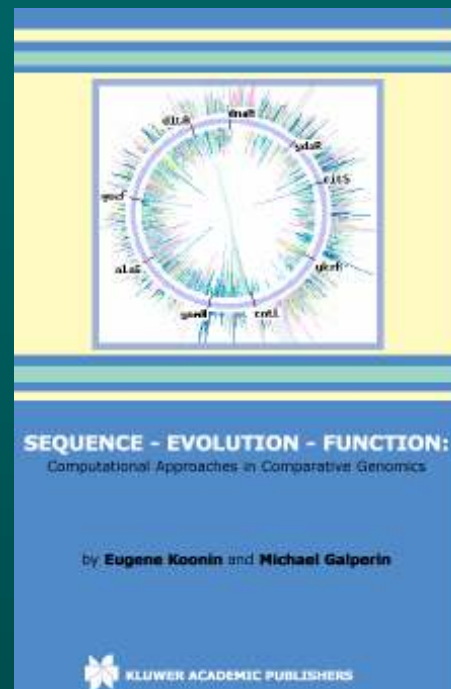
# Reading about bioinformatics



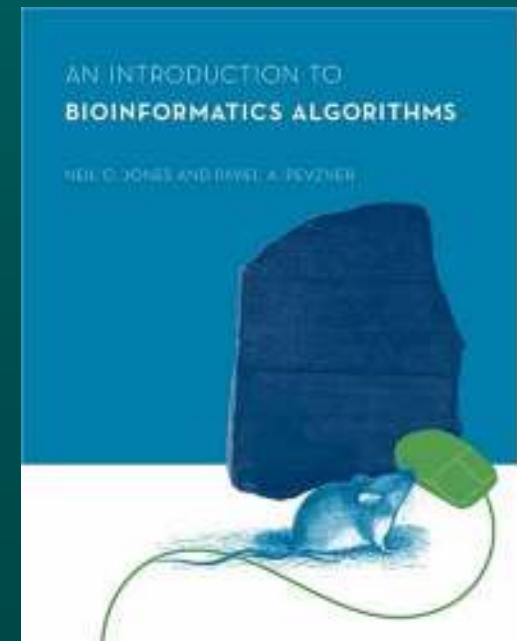
In depth introduction



Genomics research problems



Evolutionary principles



Math principles