Bioinformatics

Knowledge-representation in molecular biology

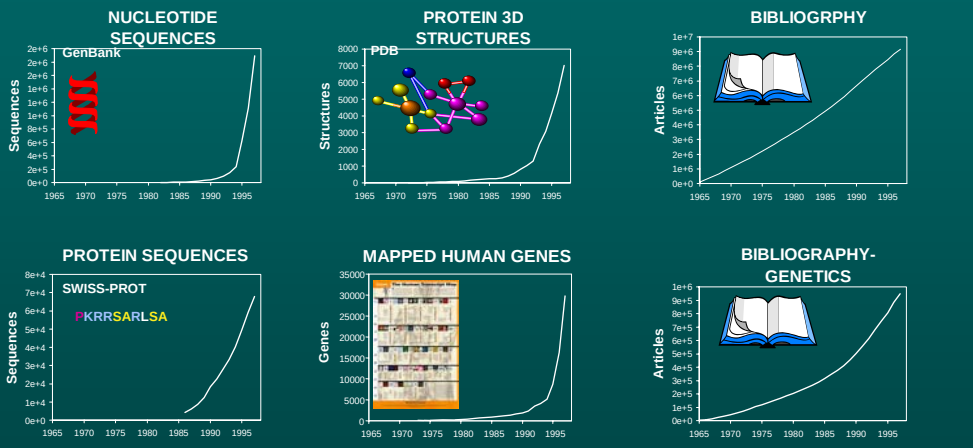
Sándor Pongor

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



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 labmda genome
- 1983 - Database search (Wilbur-Lipman)
- 1985 - FASTP/FASTN: fast sequence similarity searching
- 1987 - Sequence profiles
- 1987 - EMBL, Genbank,Swiss-Prot databases

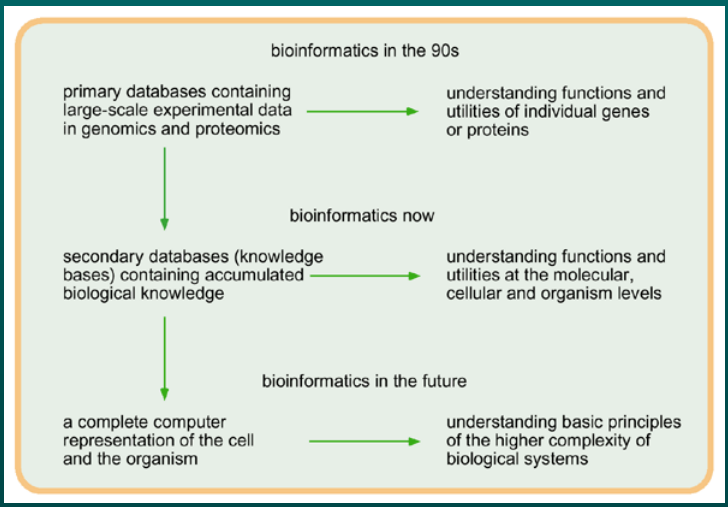
Bioinformatics milestones 2

- 1988 - National Center for Biotechnology Information (US)
- 1988 - EMBnet network for database distribution
- 1990 - BLAST: fast sequence ssimilarity 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

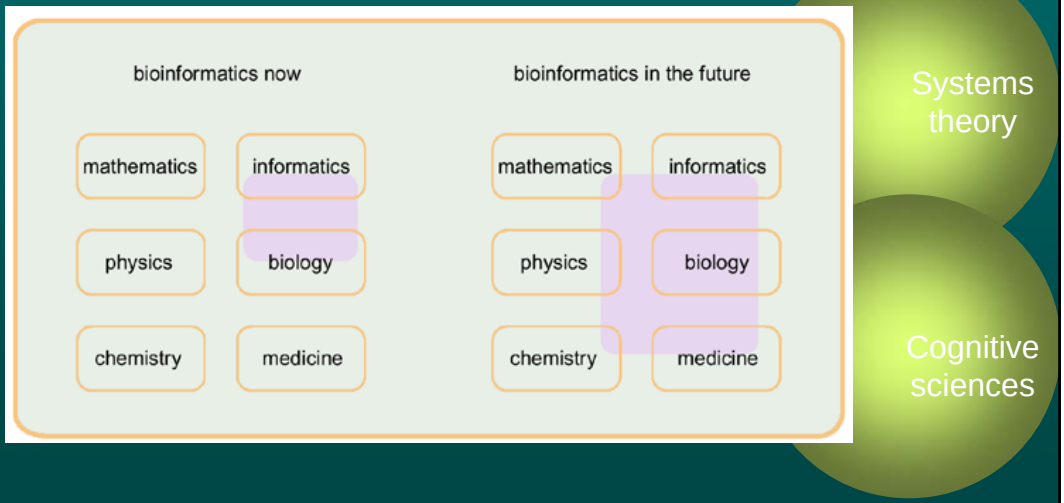
The ingredients

- 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

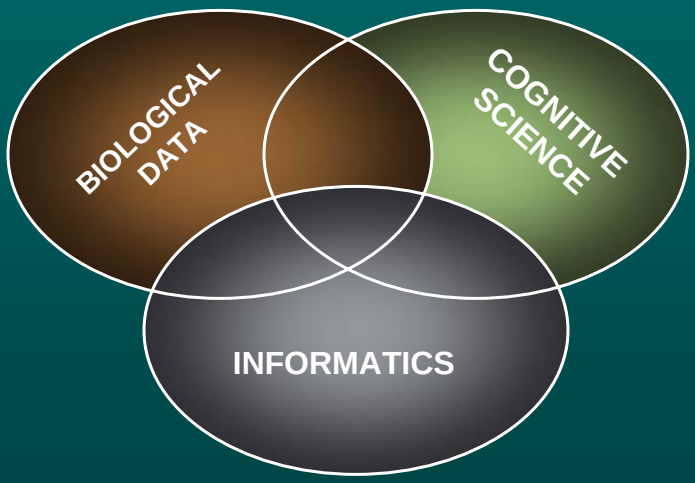
The evolution of bioinformatics



Bioinformatics is an approach to biology…



BIOINFORMATICS



MODELS

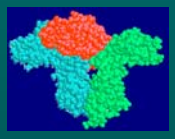
Molecular structures

MARTKQTARK
STGGKAPRKQ
LATKAARKSA

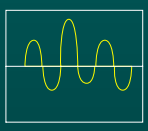
Sequences

Extended sequences
(e.g. disulphide-topologies)

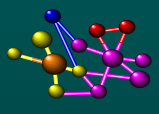
Domain-cartoons
(sec. str. cartoons)



3D structures

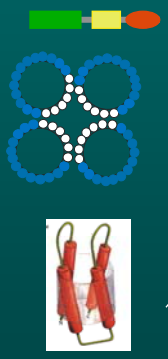


Diagrams (hydrophobicity plots, helical circles)



3D cartoons

Structures As Database Records



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

ANNOTATIONS

CIPKWNRCGPKMDGVPCCEPYTCTSDYYGNC

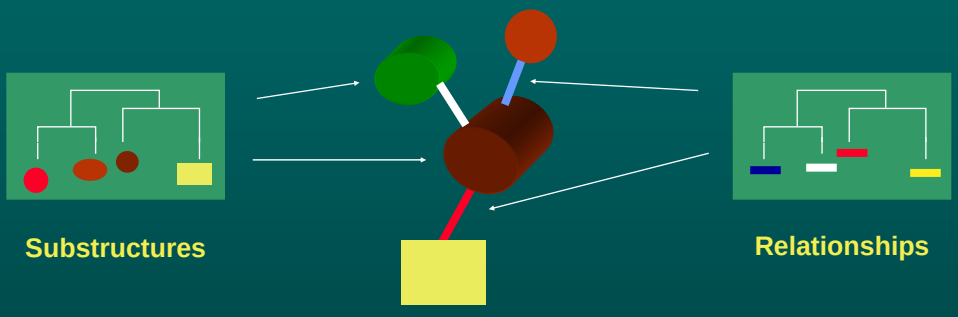
Sequence (structure)

```
qfinetdttvivtwtpprarivgyrltvgllseeg
depqyldlpstatsvnipdllpgrkytvnvyeise
egeqnlilstsqttapdappdptvdqvddtsivvr
wsrprapitgyrivyspsvegsstelnlpetansv
tlsdlqpgvqynitiyaveenqestpvfiqqettg
vprsdkvppprdlqfvevtdvkitimwtppespvt
gyrvdvipvnlpgehgqrlpvsrntfaevtglspg
vtyhfkv
```

SEQUENCE
OR STRUCTURE

Database record, fields

A structural model



Substructures

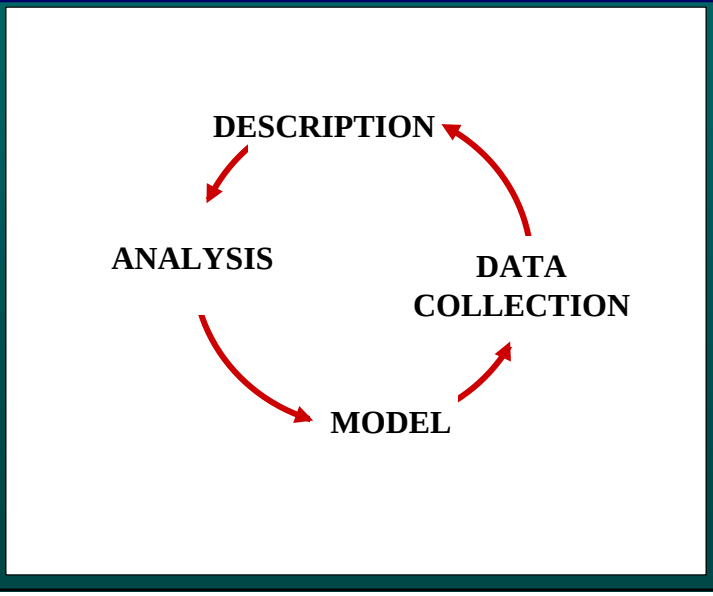
Relationships

Structure

Susbstructures, relations, rules = onthology

Entity-relationship model
Pongor, Nature, 1987

Molecules change



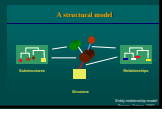
Models

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

SEQUENCES

3D STRUCTURES

NETWORKS



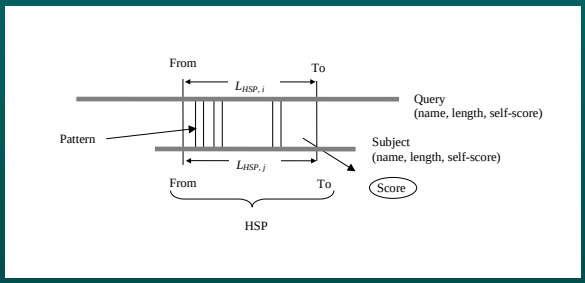
- SEQUENCES

LANGUAGE

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
```

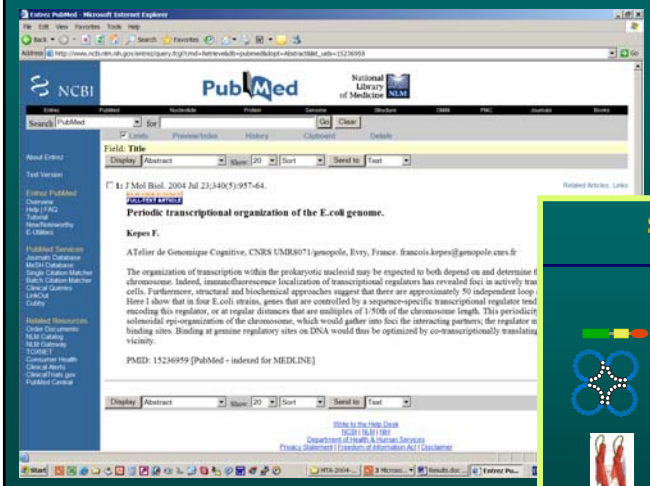
Alignments



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

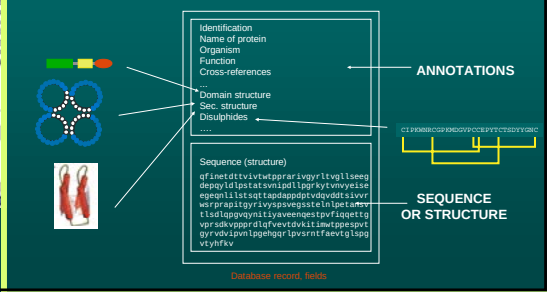
LANGUAGE

The language of bibliographies



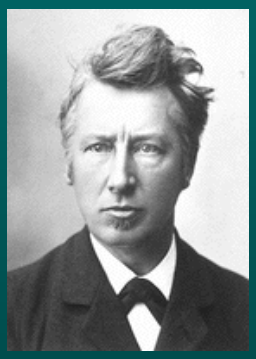
Keyword-collecttions, onthologies, etc.

Structures As Database Records



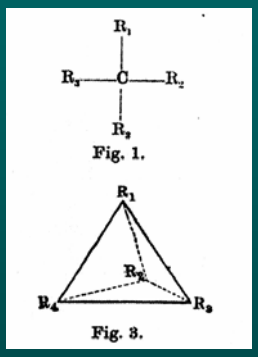
- 3D STRUCTURES

Chimie dans l'espace



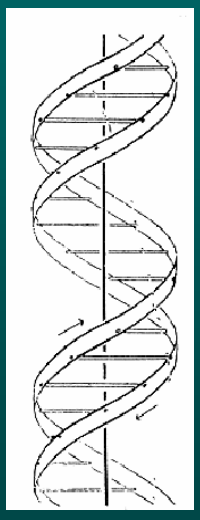
Van t'Hoff

1852-1911



1898

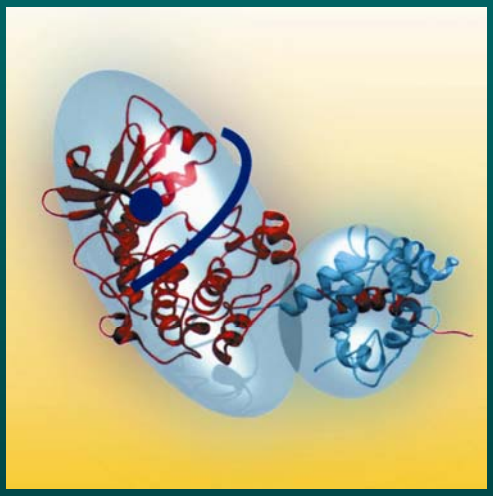
Some molecules are more equal then others…



…"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"

3D OBJECTS

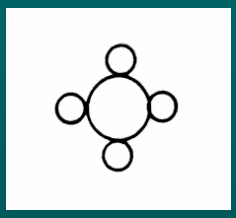
Protein models



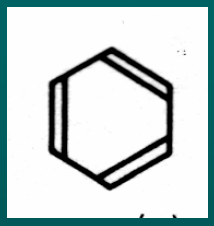
- NETWORKS

TOPOLOGIES, GRAPHS

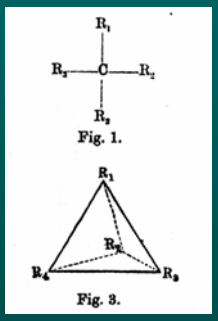
Small molecules – classical graphs



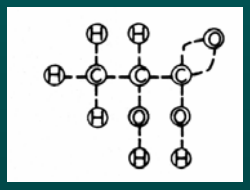
Loschmidt, 1861



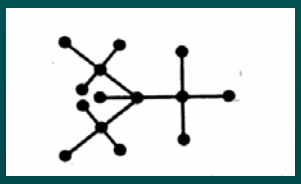
Kekulé, 1865



Van't Hoff, 1898



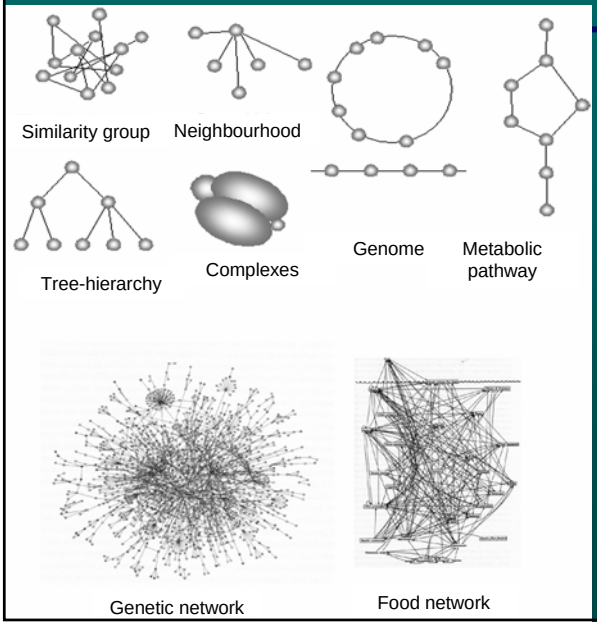
Crum Brown, 1861



Cayley, 1872

TOPOLOGIES, GRAPHS

Genomes, assemblies

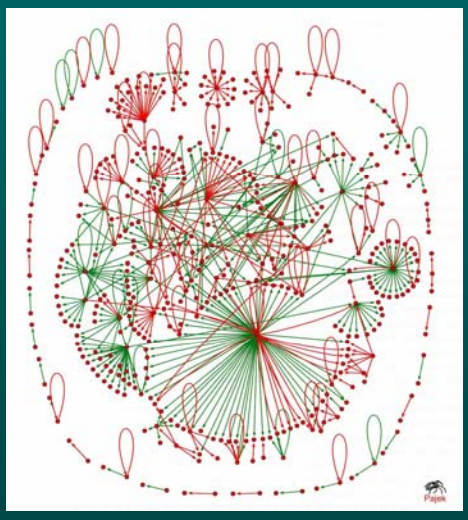


Entity-relationship models
Topological meta-models

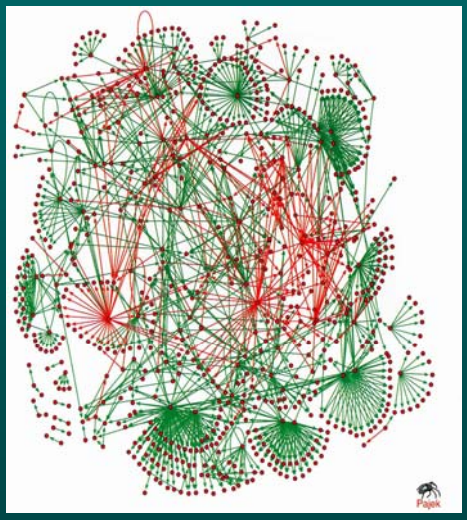
TOPOLOGIES,
GRAPHS

The transcription regulatory networks

+ (up)
- (down)



E. coli



S. cerevisiae

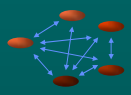
SIMILARITY, CLASSIFICATION

The concept of similarity I

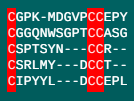
...easier if modular

Multiple Objects

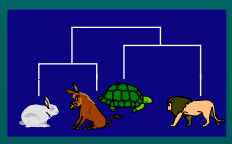
Structural similarity



Similarity groups or neighborhoods

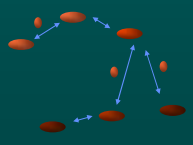


Multiple alignments



Evolutionary trees

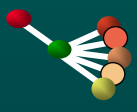
Context (function)



Metabolic pathways

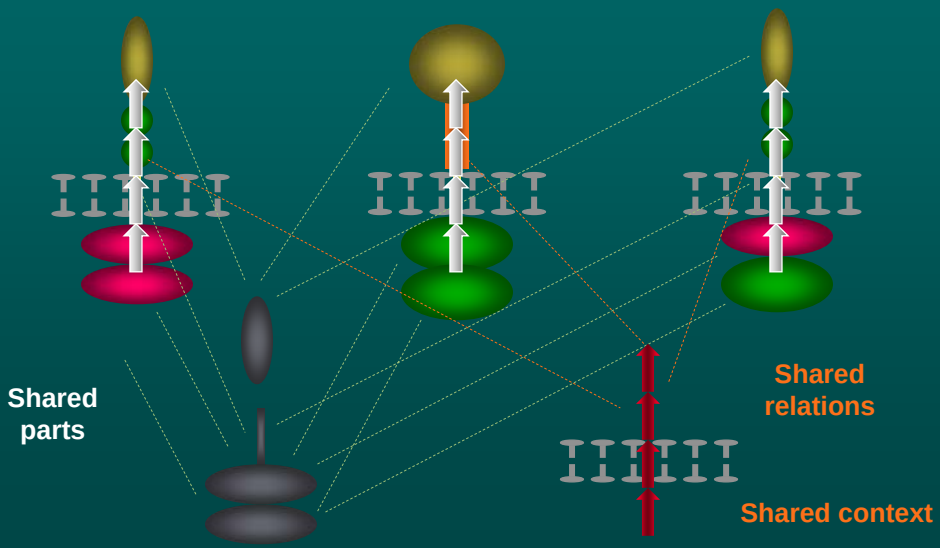
Subunit structures, ligands

Genomes

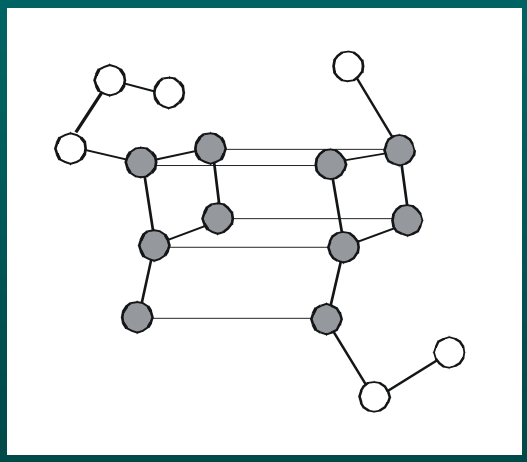


Trajectories

Similarity of molecules



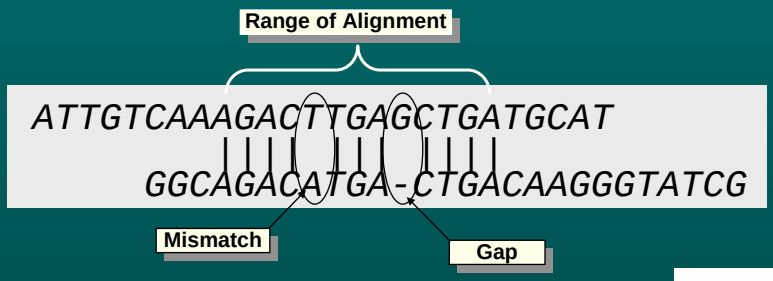
Substructure identity ~ similarity



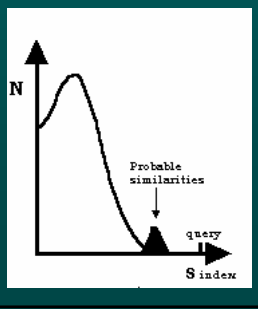
”The similarity of objects can be best described as partial identities of components and relationships

Erich Goldmeier, The similarity of perceived forms, 1936

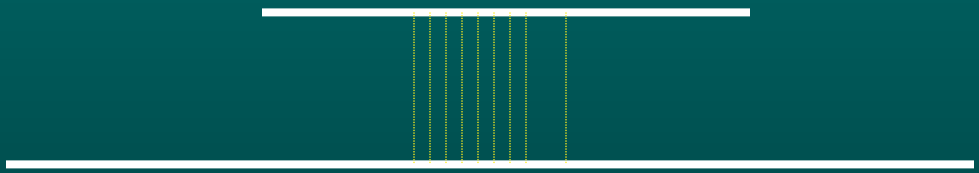
Quantification of sequence similarity



Score ~ sg like an edit distance



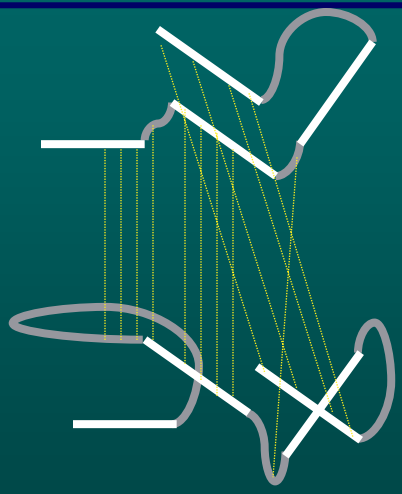
Sequence comparison



1. Find region of alignment (fast)

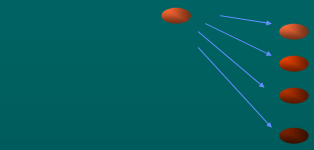
2. Calculate similarity score (fast)

3-D comparison



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

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

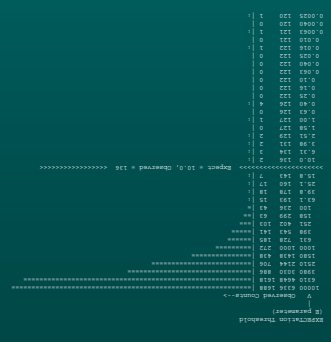


Ranked list of best similarities

Twighlight zone

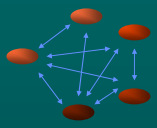
Similarities??

SEQUENCE	SCORE	DESCRIPTION
SWISSALL:1A	457.36	ALPHA-AMYLASE INHIBITOR AAI. 2/95
SWISSALL:O4726	152.82	CELLULOSE BINDING PROTEIN
SWISSALL:GUX	145.77	EXOGLUCANASE I PRECURSOR
SWISSALL:Q12G	145.66	CELLULASE (EC 3.2.1.91)



BLAST program

Using similarity: comparing a group with itself



Similarity group
or neighbourhood

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

Multiple alignment

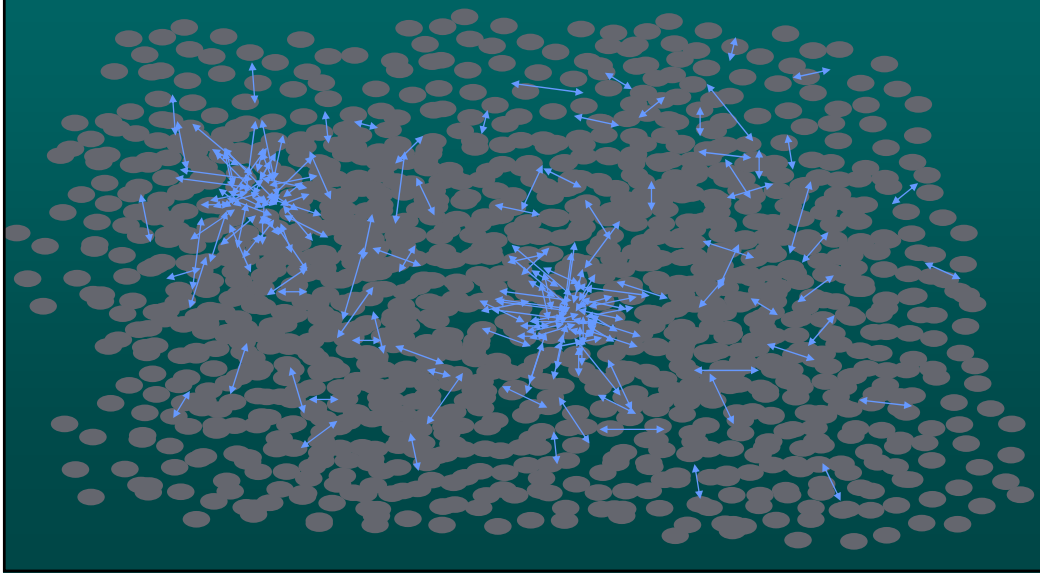
Mathematical consensus for database search

Regular expressions
Consensus sequence
Frequency matrix
Markov chains
Neural networks
etc.

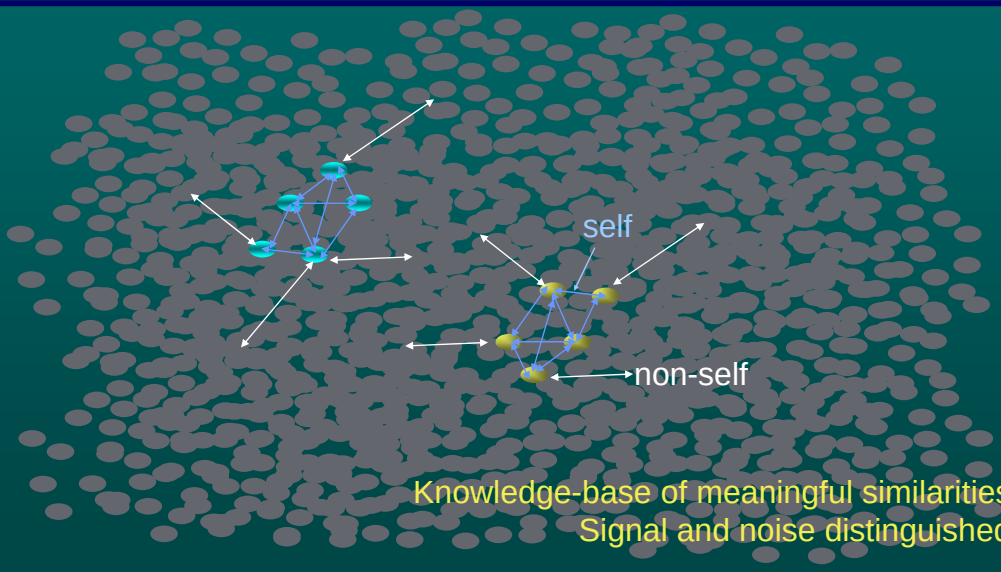
Publish

CLUSTAL program

The database as network of similarities: A memory network model



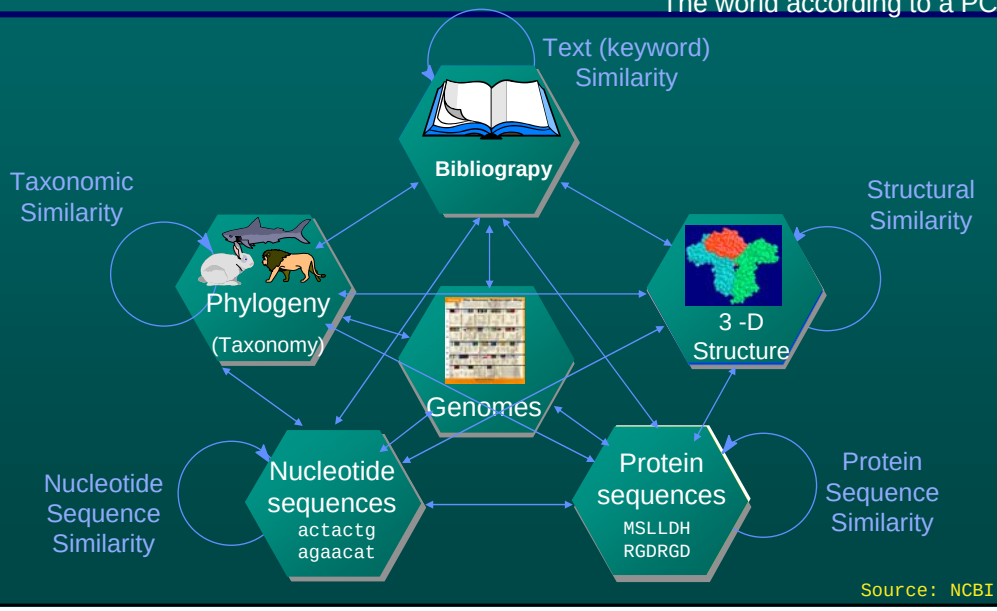
Das Wohltemperierte Database:
Similarity network as a knowledge representation



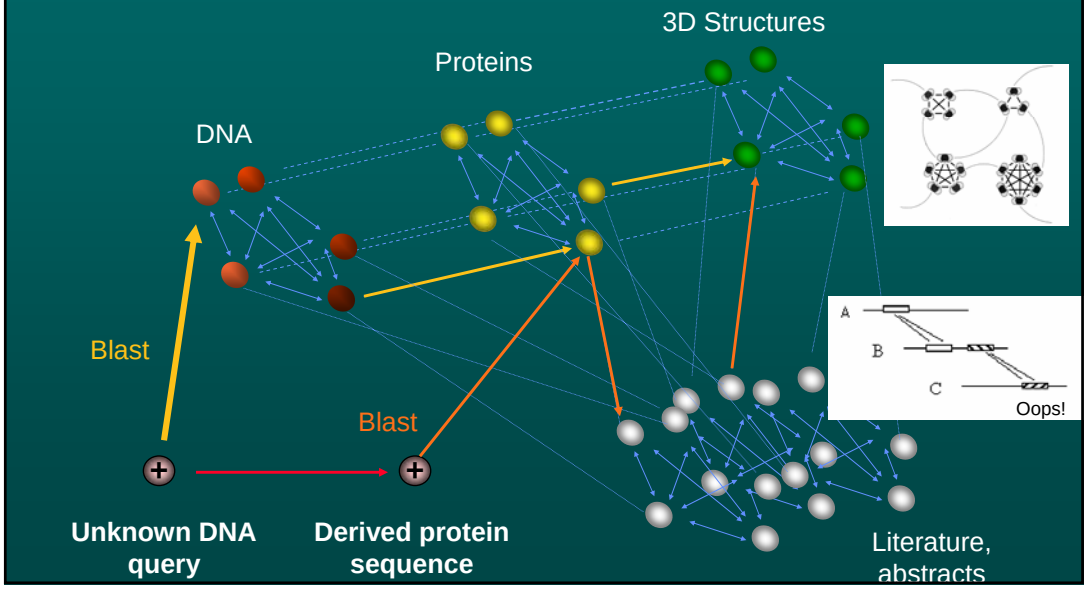
Knowledge-base of meaningful similarities
Signal and noise distinguished

Biological knowledge as a network of data

The world according to a PC...

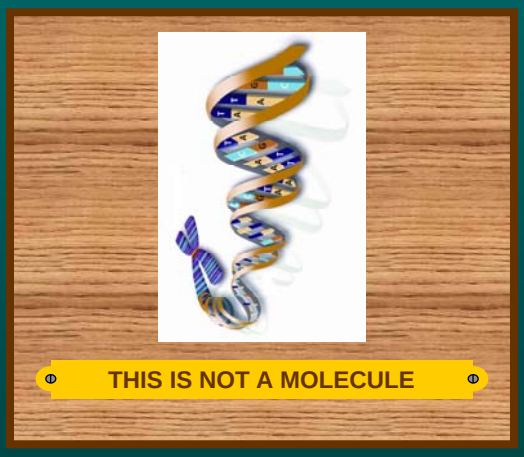


Search on a preprocessed, integrated database: the importance of a good neighbourhood

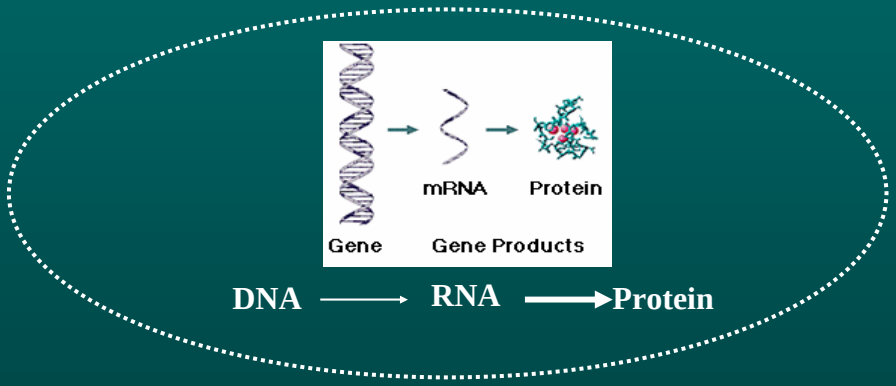


Models are human constructs...

Models are human constructs...

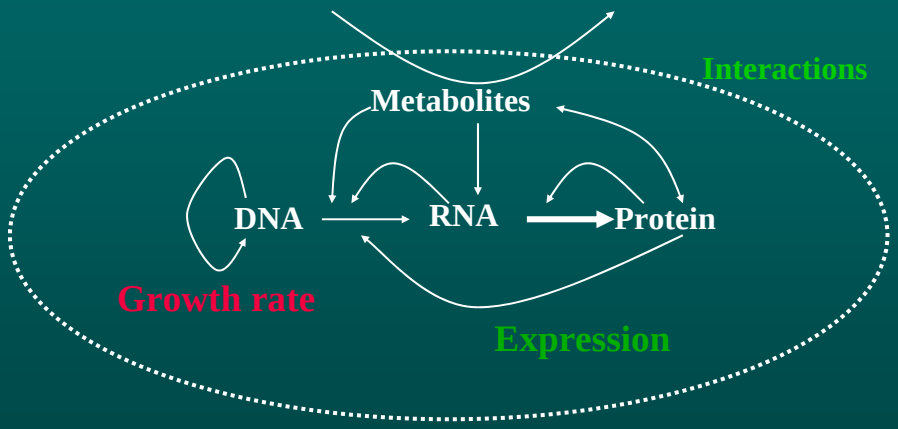


The central dogma:



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

Computer methods in Molecular Biology Trieste June 24 - July 1, 2005

- Sequence database searching, theory and practice (Dave Judge and Jack Leunissen)
- Nucleic acid databases, Medline, Pubmed (David Landsman)
- Protein databases, Swissprot, Prosite (Elisabeth Gasteiger)
- WWW servers, EBI (Kristian Vlahovicek)
- Gene discovery (Luciano Milanesi)
- Genome analysis (Martin Bishop)
- Algorithms, domain similarities (Kristian Vlahovicek, Sándor Pongor)

On-line help to this lecture

- Bioinformatics tutorials on-line

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

- ICGEBnet

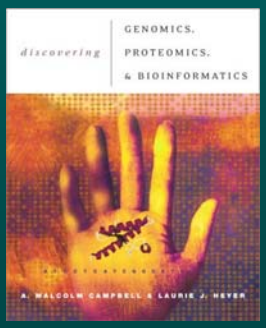
  http://www.icgeb.org/~netsrv/

- The Trieste bioinformatics course

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

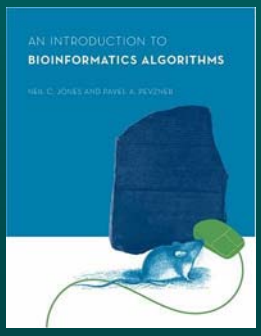
Reading about bioinformatics

In depth introduction



Genomics research problems

Evolutionary principles



Math principles