Introduction to Bioinformatics

Pázmány Péter Catholic University

Faculty of Information Technology

Fall Semester, 2015/16

fides et ratio

Core operations I

Comparison

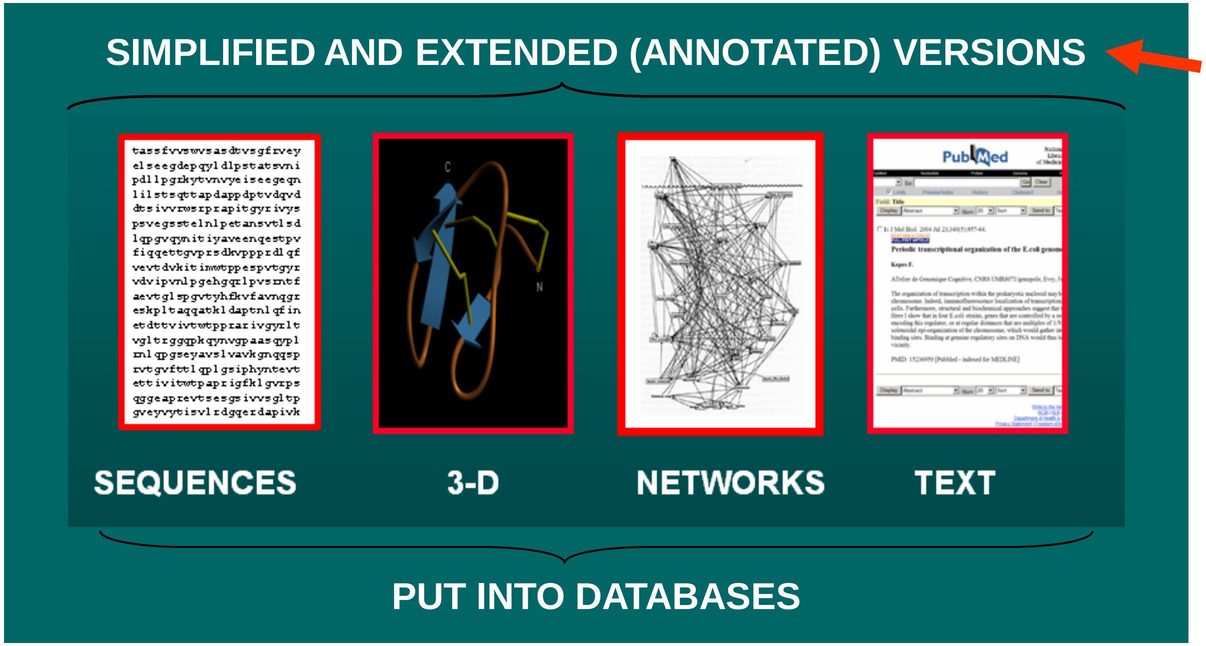
Sándor Pongor

This lecture: Comparison

- Theory: comparison in bioinformatics and in psychology
- Representation in bioinformatics
- Comparison
  - Proximity measures (similarity, distance for unstructured descriptions)
  - Alignments and common patterns (for structured descriptions)

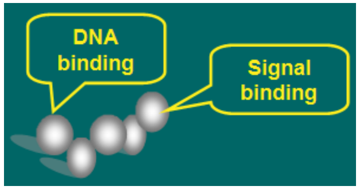
Previous lecture (Core data types)

- Systems, structure, function
- Sequences, 3D-structures, networks, texts
- Standard form, simplified and annotated forms
- Logical structure, data description
- Database records (contain a structured mixture of all this)



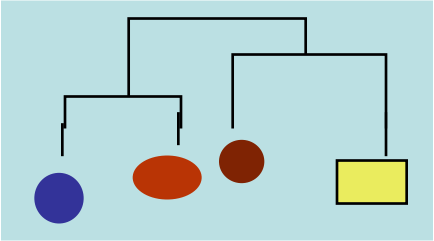
PRESENTATION FORMATS

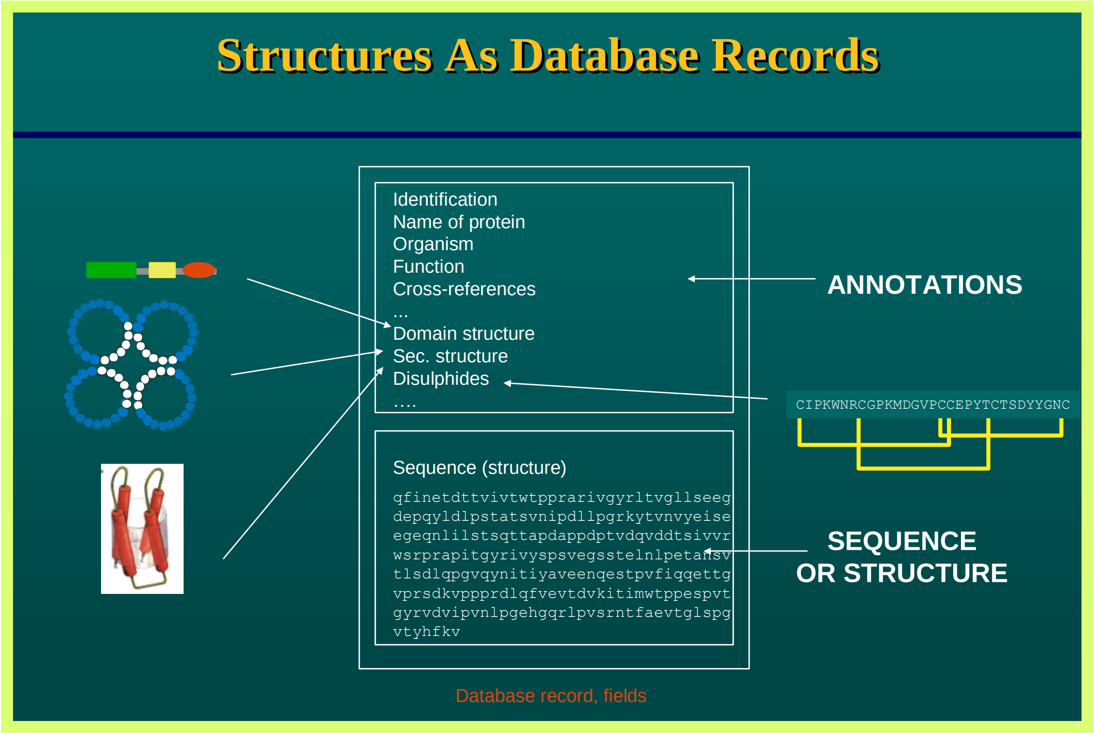
e.g



Previous lecture
(Core data types)

- Database records contain data, metadata (annotations, data on data).
- Rules of data representation and metadata descriptions are in ontologies (definition of concepts = meta-metadata, data on metadata)





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

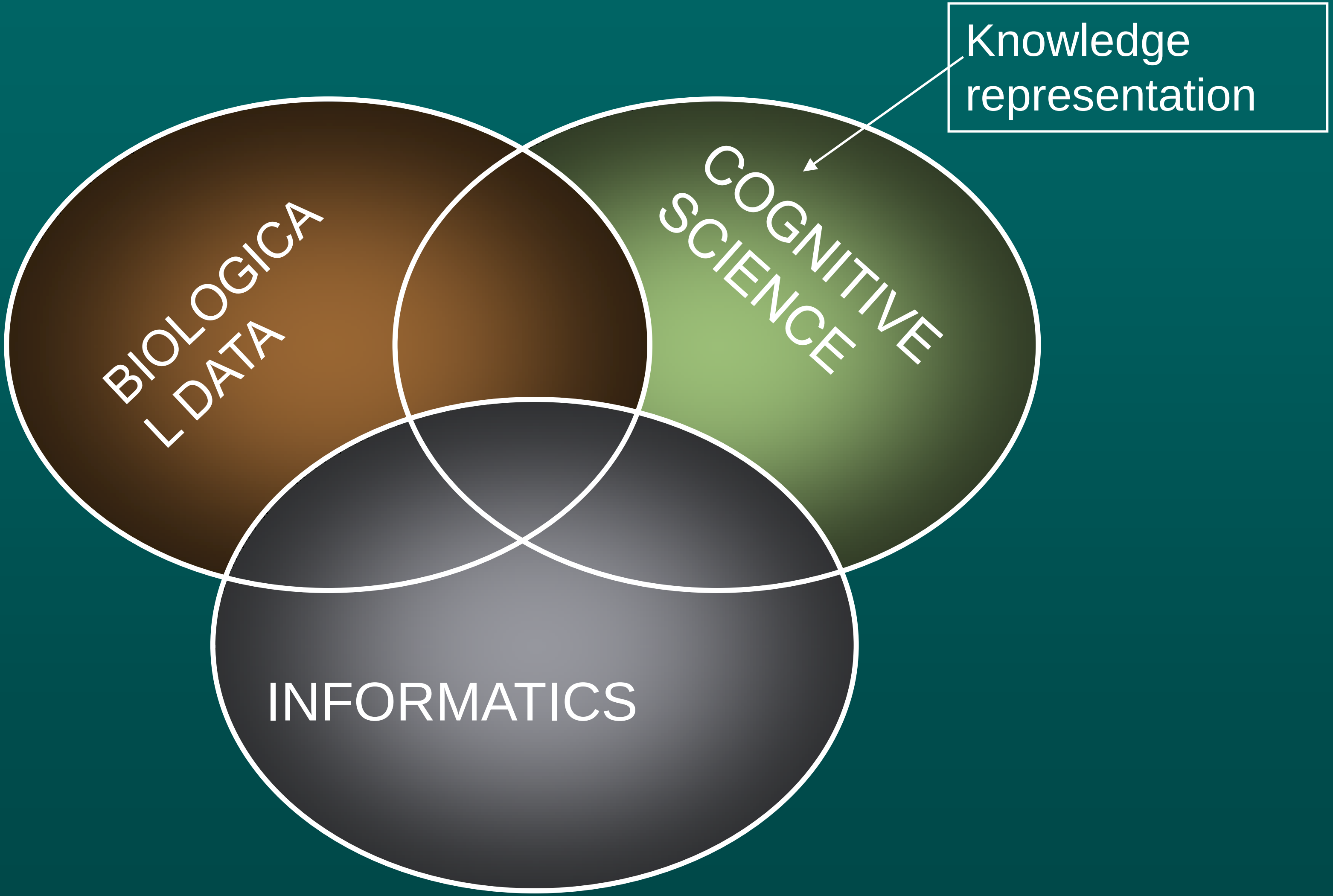
Database record, fields

METADATA

DATA

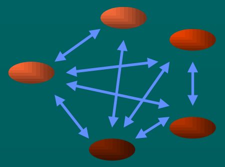
- Understanding data: grouping and classifying, organizing into knowledge items, matching to other knowledge items.
- Humans operate on “logical structures”. Computers operate on descriptions (which are given to them by humans).
- Example1: Humans compare objects by “*similarity of logical structures*”, groups described by “common patterns” (also called motifs = simplified logical structures) and only then by numbers. Car example.
- Example2: Machines compare descriptions first via numerical “*similarity measures computed between descriptions* ” and evaluate significance (probability). Sometimes also by patterns

Classification is based on “similarity” which is **a very human concept**



Understanding is grouping. There are two types:

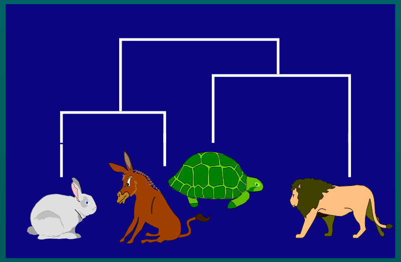
Similarity by structure



Similarity groups or neighborhoods

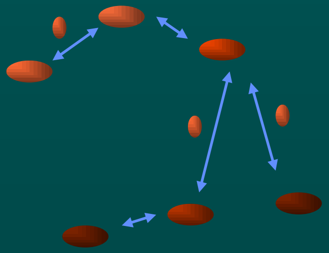
```
CGPK-MDGVPCCEPY
CGGQNWSGPTCCASG
CSPTSYN---CCR--
CSRLMY---DCCT--
CIPYYL---DCCEPL
```

Multiple alignments

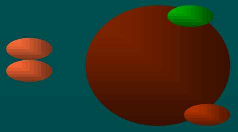


Evolutionary trees

Similarity by context (function)

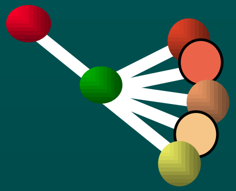


Metabolic pathways



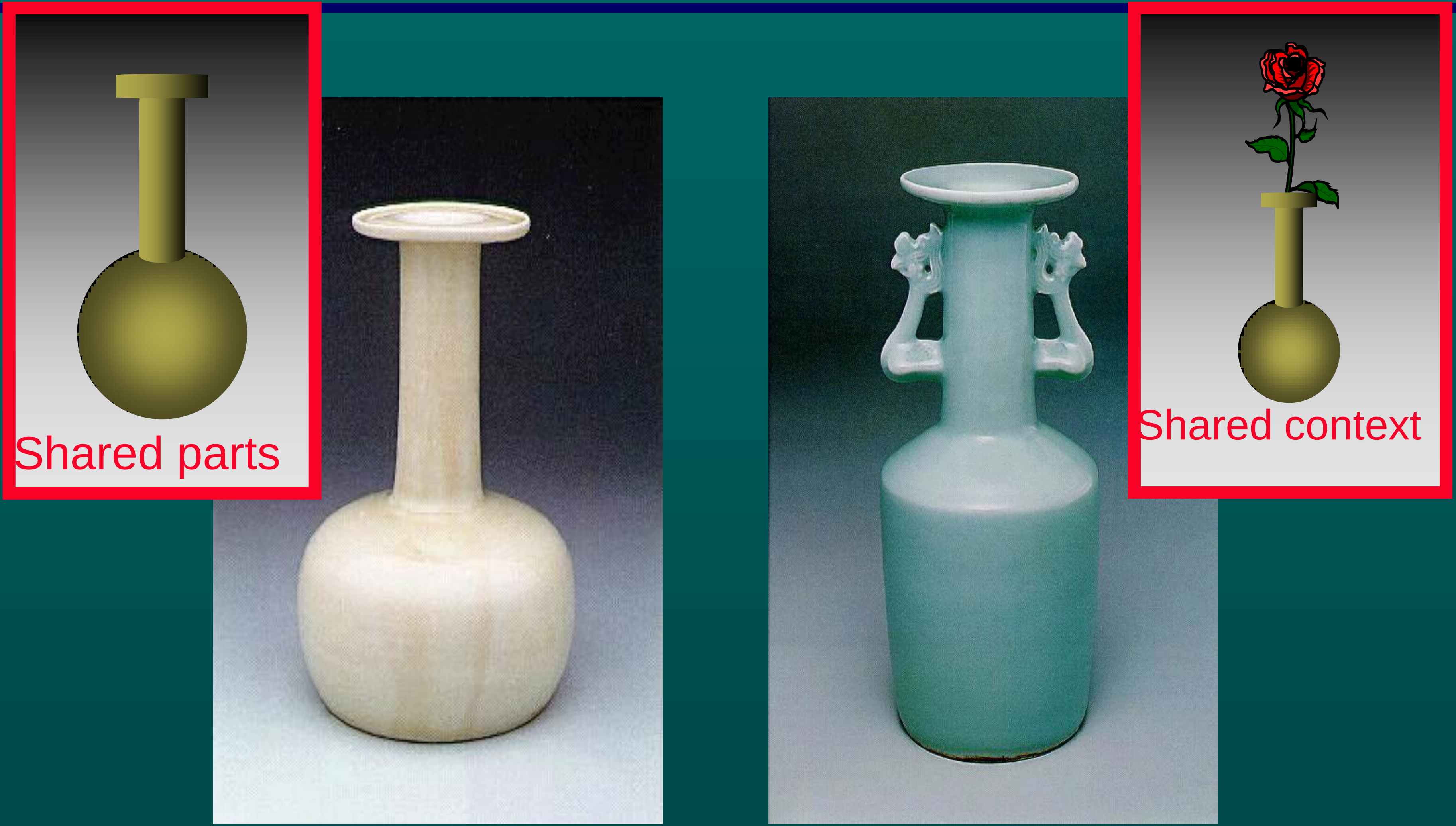
Subunit structures, ligands

Genomes



Trajectories

Similarity for humans

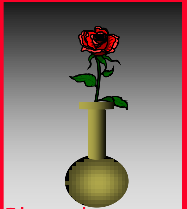


Expressed as pattern or motifs of similarity:
simplified logical structures

Patterns, motifs: simplified logical structures associated from parts

- Patterns (motifs) are associated from parts
- Associations within a context (function)
- Patterns in space
- Patterns in time
- Associated to other patterns/motifs

Shared parts

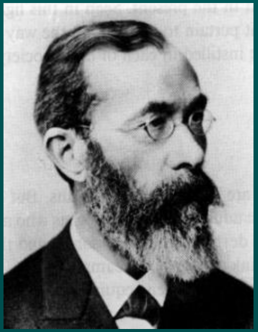


Shared context

Chinese,
British Museum

16th century

Painted by Leonardo
Stolen by Max Schmidt

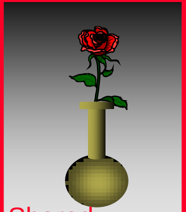


Wilhelm Wundt
1832-1920
Founder of psychology
(structuralism)

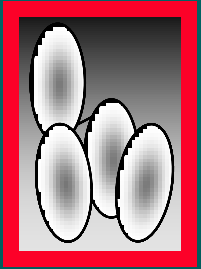
Patterns 2: simplified logical structures in bioinformatics

- Patterns as structures
- Patterns in context (function)
- Patterns in space
- Patterns in time
- Associated to other patterns

Shared parts



Shared context



Glycolysis pathway

Chinese,
British Museum

Cell membrane

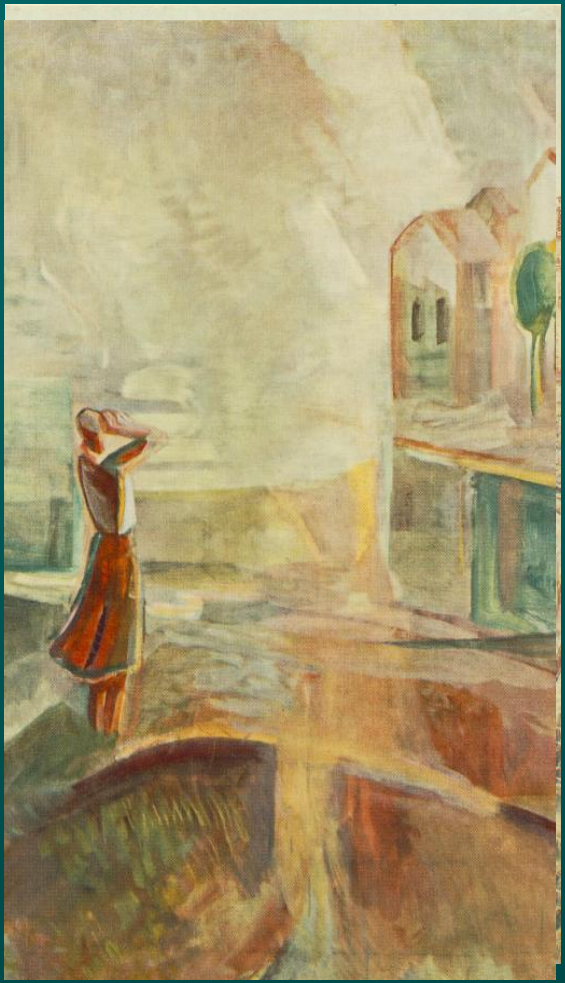
16th century

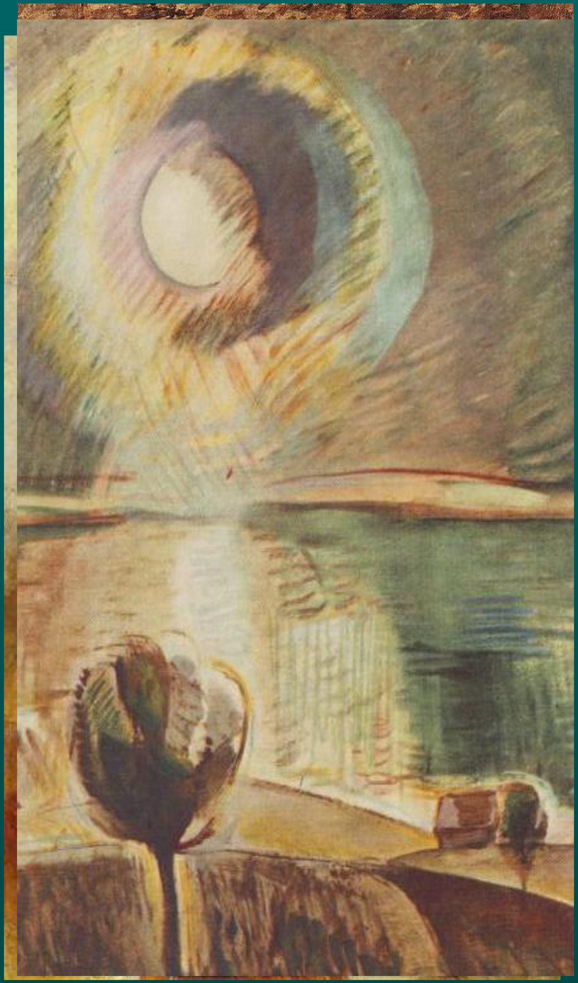
Mitosis phase in cell cycle

Painted by Leonardo
Stolen by Max Schmidt

Discovered by X.Y
Published in Nature

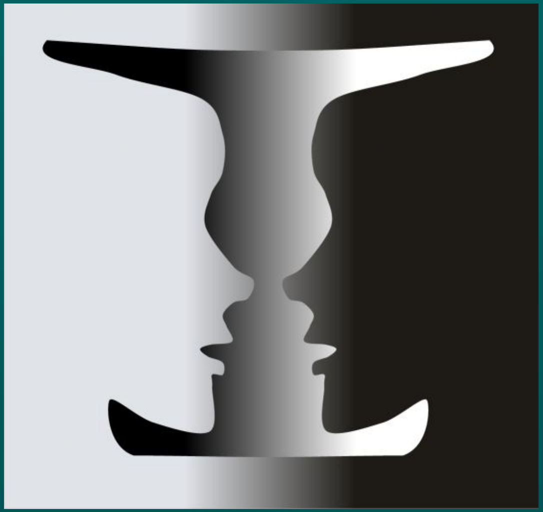
Some patterns are complex - how do we discover them?



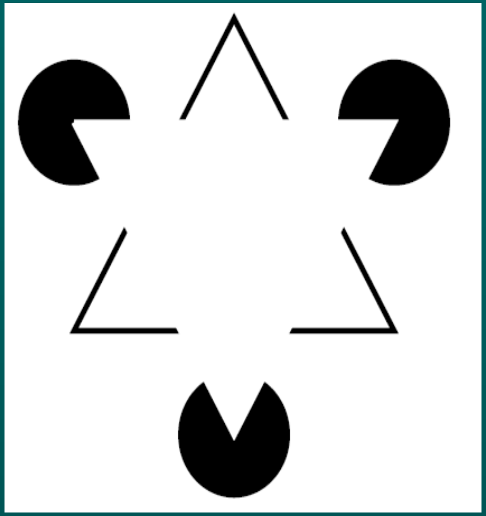


Humans instinctively aggregate any features into “meaningful patterns”

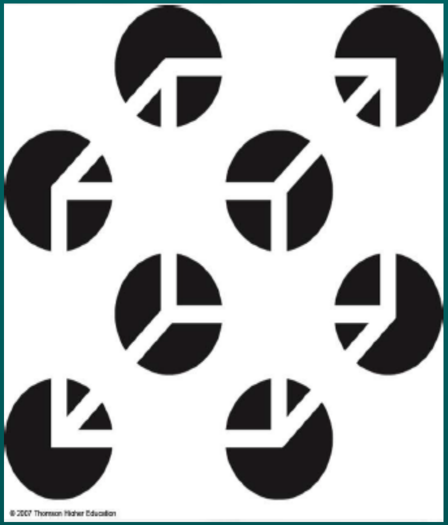
The birth of patterns is explained by Gestalt psychology (1920-1970)



Edgar Rubin's vase
(~1915, Copenhagen)



Kanizsa's Triangle
(~1955, Trieste)

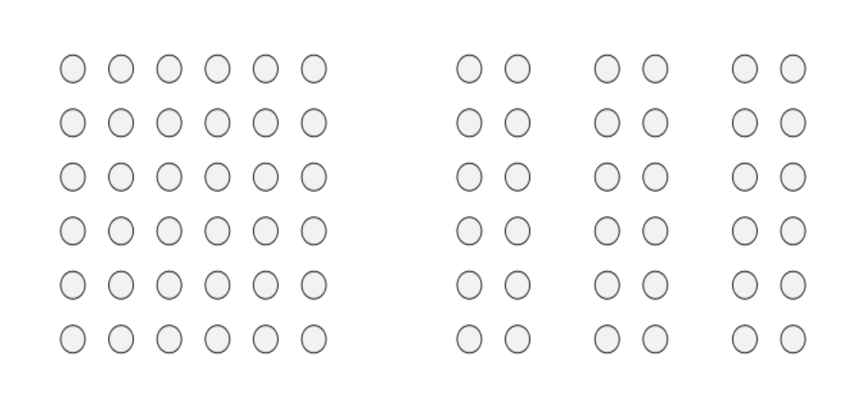


© 2007 Thomson Higher Education

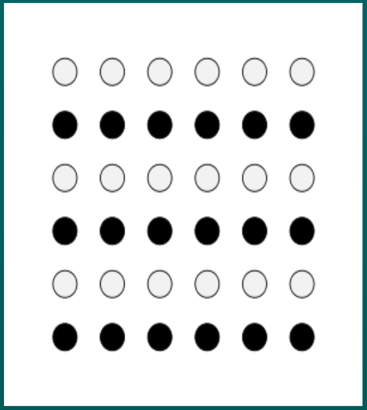
Illusory contours

- Patterns are more than the sum of their parts. (Gestalt: shape, form in German). The emergence of new patterns is not explained by the traditional, structuralist approach

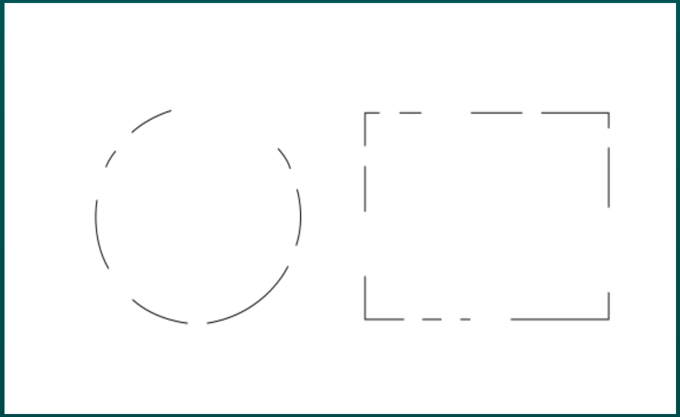
Gestalt psychology principles: How do we aggregate items into patterns?



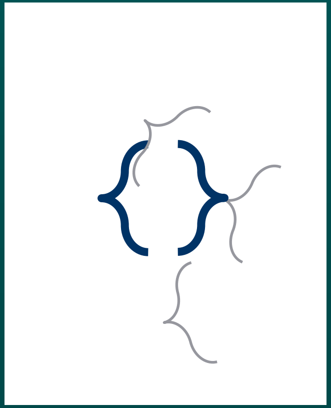
By proximity (nearness)



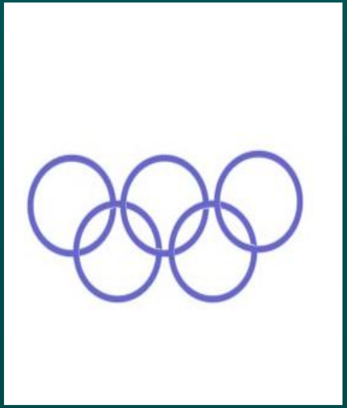
By similarity



By continuity and closedness

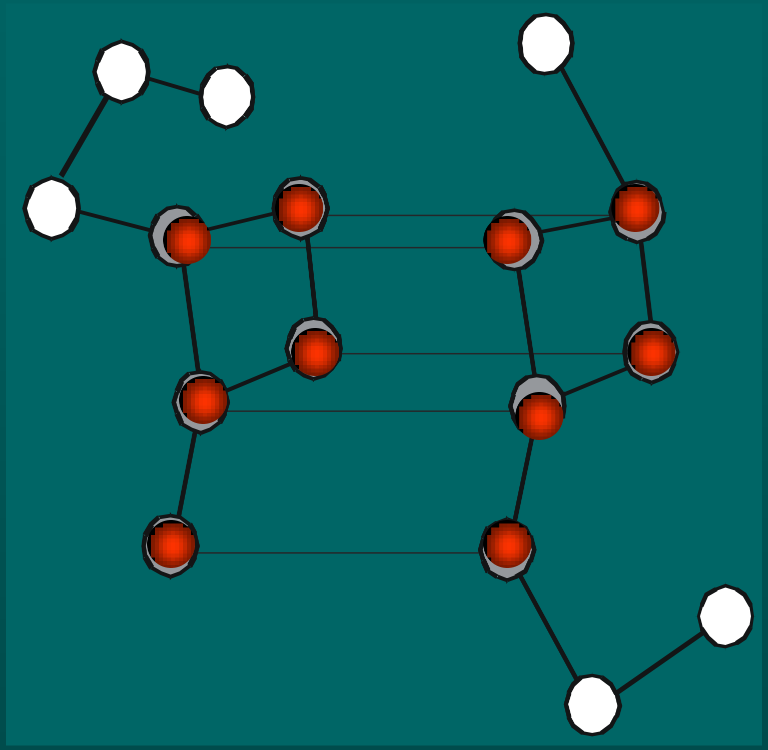


By symmetry



By simplicity (“Pregnänz”)

In bioinformatics: similarity ~ “shared patterns” (like in human psychology)

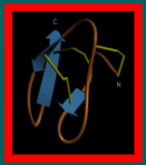


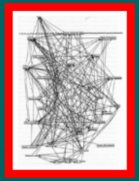
BUT we also assign a score

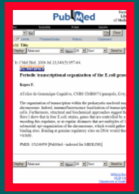
”The similarity of objects can be best described as partial identities of components and relationships

Erich Goldmeier, The similarity of perceived forms, 1936

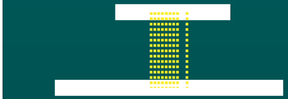
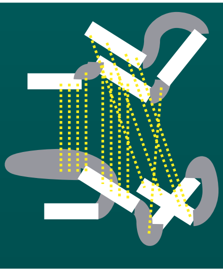
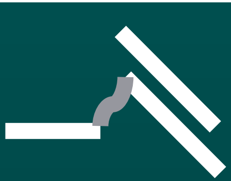
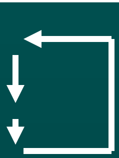
Statements on similarity







SEQUENCES 3D NETWORKS PAPERS

Global	“Glycine-rich”	“α-helical”	“scale-free”	“genomics”
Substructure-alignment			(metabolic pathways)	same author, common references
“known motifs”	G-RR			“Joe Doe, folding”

Two proteins are similar because both are “glycine rich”, “alpha-helical”, “contain motif X”. Motifs can be defined at various levels.

Similarity by
Humans and Machines

- Humans use intuitive patterns, and similarity is defined as a shared pattern (motif).
- Patterns are either instinctive or knowledge based
- The choice and form of the patterns is flexible
- Consensus patterns are in the memory, validated and updated by experience

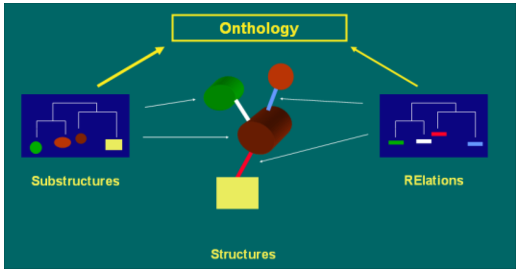
Flexible, qualitative

- Computers use descriptions (vectors, character strings) and a) compute numerical similarity measures („scores"), b) search for predefined patterns
- Descriptions, numerical measures and patterns are all predefined
- Scores and motifs are validated by statistics (significance, predefined algorithms)
- Memory: dbases

Rigid. quantitative

A “representation”

- Entities are described by the EAV scheme (entity, attribute, value). E.g. apple has an attribute “weight”, its value is 150 g. Protein X has a molecular weight of 100,000 daltons
- Relations are described in the same way: A single chemical bond has an attribute “length” which is 1.4 Angstroms. Here we call this a RAV (relation, attribute, value) scheme.
- A “structure” is a structured set of encapsulated EAV substructures



Pongor, Nature, 1987

Quick overview

Sequence descriptions

- ACAACTGG (the sequence itself, structured)
- $A_3C_2G_2T$ (composition, unstructured)
- $(AC)_2(CA)(AA)(CT)(TG)(GG)$ (word composition, “hybrid”)

Remark: Words are structured in themselves, so word composition is partly structured (because the relation between words is not included). Words are “substructures” so this is substructure composition.

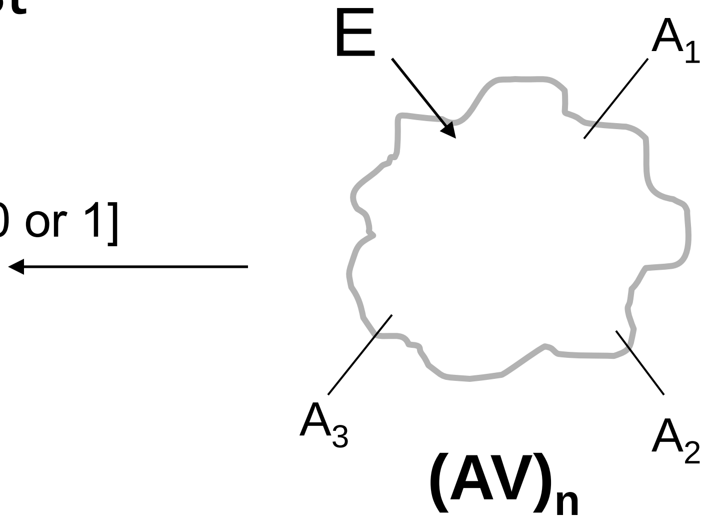
Important properties of representations….

- 2 types: *unstructured* and *structured* (depending if we know/want to use the internal structure)
- Granularity: resolution of the description (e.g. 4 nucleootides, 16 dinucleotides)

Unstructured representations

- We know nothing about internal structure
- Only the properties are known (global descriptors), can be discreet or continuous.
- Best described as vectors (each dimension is an attribute, the contents is the value..). Sometimes a large number of dimensions.
- Vector operations are fast

Binary or “presence/absence” vectors V[0 or 1]
Real valued vectors



Note: From here we use the Entity-Attribute-Value terminology

Vector types

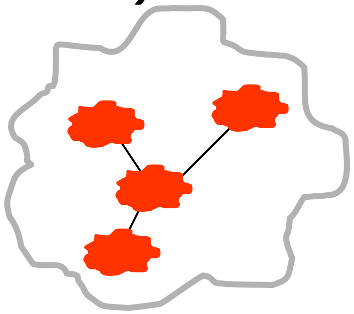
- Binary vectors consist of 0 or 1 values, e.g. 0,1,0,0,1,0. Indicate the presence or absence of attributes.

- Non-binary vectors can contain real or integer-valued components, e.g., 0.5, 0.9, 1.0.

1. Selecting representation

Structured representations

- We know the internal structure in terms or Entities and Relationships (both described in terms of attributes and values→EAV and “RAV”)
- Information-rich, allows detailed comparisons
- Need alignment (matching) for comparison...
- Examples: character strings (sequences), graphs (most molecular structures are like this..)

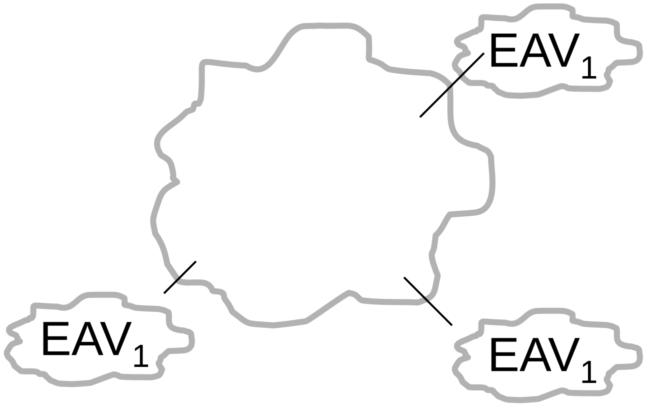


Graphs ← $(\mathrm{EAV})_m$ $(\mathrm{RAV})_n$

1. Selecting representation

Mixed or composition-like descriptions

- We decompose an object to parts of known structure, and count the parts (atomic composition, H_2O, or amino acid composition of proteins).
- The result is a vector, fast operations, alignment (matching) is not necessary
- The information content of the vector depends on the granularity of the parts. Atomic composition of proteins or of people is not informative.



Binary vectors
Real valued vectors

$(EAV)_n$

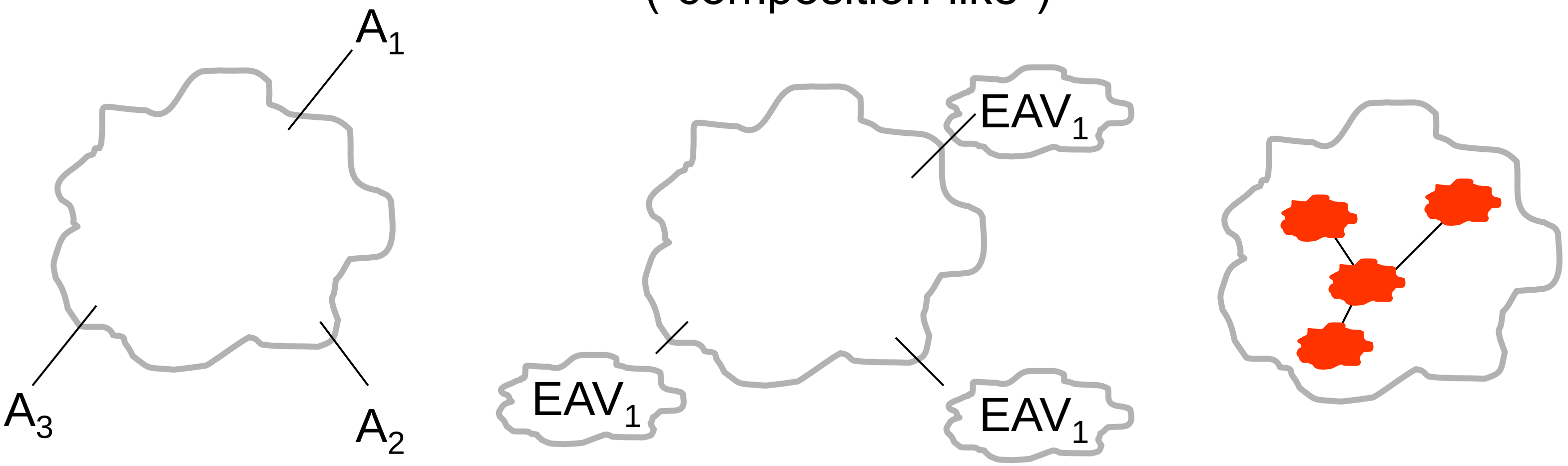
1. Selecting representation

Representations at a glance

Unstructured

Mixed
("composition-like")

Structured



$(\mathbf{AV})_{\mathbf{n}}$

$(\mathbf{EAV})_{\mathbf{n}}$

$(\mathbf{EAV})_{\mathbf{m}}$ $(\mathbf{RAV})_{\mathbf{n}}$

Binary vectors (V[0,1])

Real valued vectors

Graphs

V can be discrete (0 or 1) or real valued

2. Comparison

- Input: Two descriptions
- Output:
  - For unstructured: a score (similarity, distance), is mandatory
  - For structured:

    - a score (like above, mandatory) AND

    - a common pattern (result of matching=alignment), optional

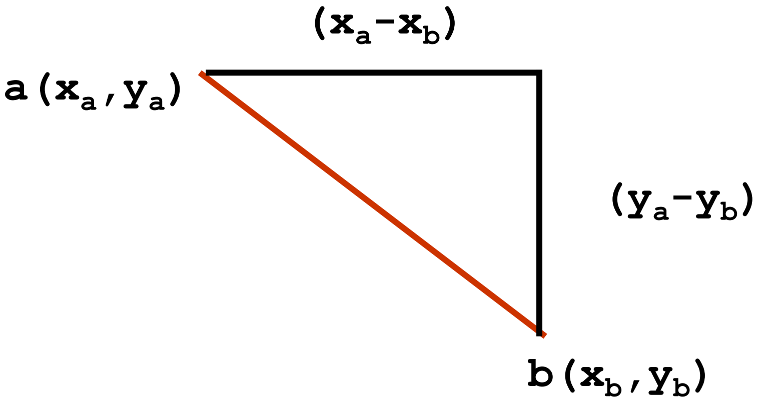
Proximity measures (scores)

- Similarity measures (zero for different objects, large for identical objects)
- Distances (large different objects, zero for identical objects)
- Exist both for vectors and for structures…
- “Well behaved”: if bounded, e.g. [0,1]
- Don’t expect linearity in any sense… (“twice as similar” makes no sense)
- Similarity S~1/D or 1-k*D..

Vector distances

- The concept of proximity is based on the concept of distance.
- The most popular distance of two points, a and b in the plane is the euclidean distance:

$$D_{ab} = \sqrt{(x_a-x_b)^2 + (y_a-y_b)^2}$$



Metric properties:

- 1. Distance is positive D_{ab} >= 0,
- 2. Distance from oneself is zero, D_{aa} =0.
- 3. Distance is the same in both directions, $D_{ab}= D_{ba}$
- 4. Triangular inequality $D_{ab} + D_{bc} > D_{ac}$

2. Comparison

Generalized Distances

- The concept of distance can be extended to *n* dimensions

$$D_{ab} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

- AND it can be extended to exponents other than 2

$$D_{ab} = \left(\sum_{i=1}^{n} |a_i - b_i|^k \right)^{\frac{1}{k}}$$

- The latter are the Minkowski metrices, k= 2 Euclidean, k=1 “city block”, variants extensively used in chemistry, physics, biology….

Similarity measures for vectors

- The dot product or inner product of two vectors is by defined as:

$A.B = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$ or $A.B = \sum_{i=1}^{n} a_i b_i$

- For binary vectors (dimensions zero or one) this is the number of matching nonzero attributes, .

- Vectors of unit length have a dot product [0,1], 1.0 for identical vectors.

Association measures

- Association measures are typically used to measure the similarity of sets, in our case property sets ("presence-absence" descriptions). The Jaccard (or Tanimoto) coefficient [0,1] expresses the similarity of two property sets *a* and *b* of non-zero attributes, respectively as

$$J = \frac{a \cap b}{a \cup b}$$

- J is 1 for identical and zero for completely different sets (or binary vectors).
- Correlation coefficients and related measures can be used for various non-binary vector types.

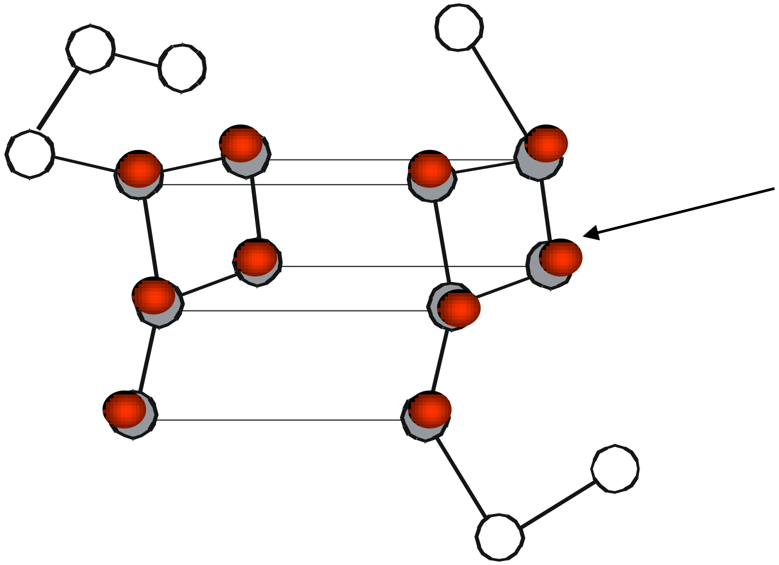
A remark on proximity measures

- There is a very large and ever growing number of proximity measures.
- For easy problems, many of them work equally well… For difficult problems none of them do.
- (So do not get scared if you see unknown proximity measures – neither should you trust them ☺)

Comparing structured descriptions

- Input: 2 structured descriptions (say, sequences)
- Output: 1) a proximity measure (score) and 2) a shared pattern (motif).
- You can use proximity measures if you can turn the description into a vector (see composition-type description).
- In addition, you can match (align) structures that gives a shared pattern. (Alignment and motifs will also be shown in later lectures)

Matching (general) structures

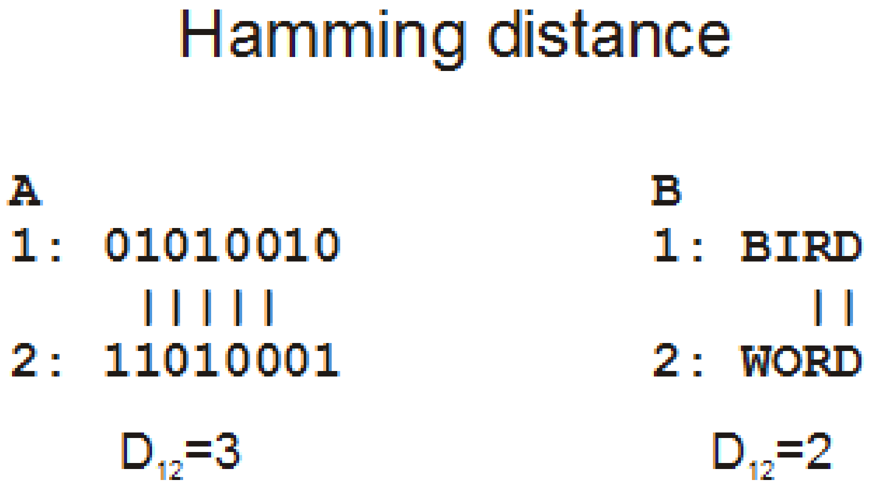


Shared motif)

- Matching graphs consists of finding the largest common subgraph. A computationally hard problem. Finding approximately identical subgraphs is NP complete.

- In the human mind, matching is instinctive (comparing cars…)

Matching bit or character-strings



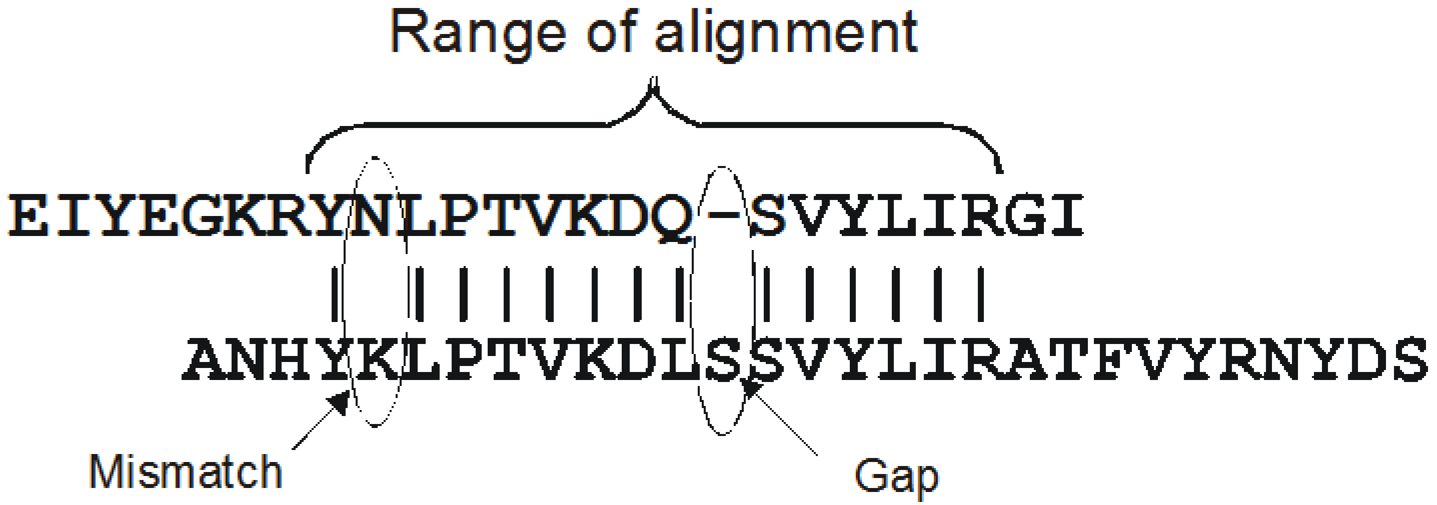
Hamming distance

```
A                          B
1:  01010010               1:  BIRD
     |||||                       ||
2:  11010001               2:  WORD
```

$D_{12}=3$ $D_{12}=2$

- The Hamming distance is the number of exchanges necessary to turn one string of bits or characters into another one (the number of positions not connected with a straight line). The two strings are of identical length and no alignment is done.
- The exchanges in character strings can have different costs, stored in a lookup table. In this case the value of the Hamming distance will be the sum of costs, rather than the number of the exchanges.

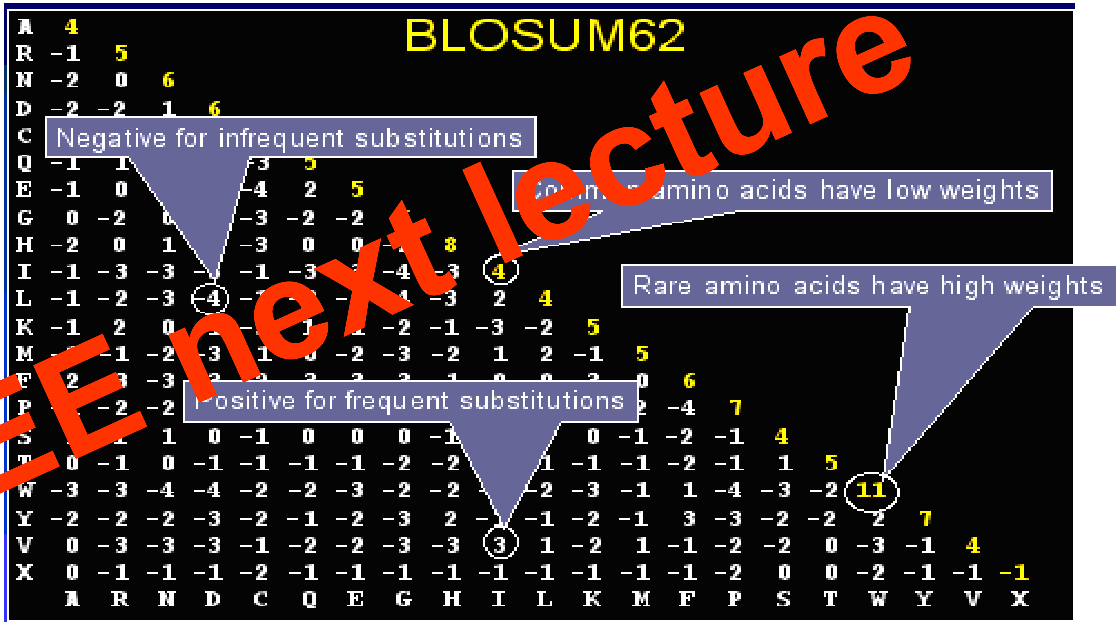
Edit distance between character strings (sequences)



- Also called Levenshtein distance. Defined as a sum of costs assigned to matches, replacements and gaps (= insertions and deletions). The two strings do not need to be of the same length.
- A numerical similarity measure between biological sequences is a maximum value calculated within a *range of alignment*. The maximum depends on the scoring system that includes 1) a lookup table of costs, such as the Blosum matrix for amino acids, and 2) the costing of the gaps. The scores are often not metric, but closed to metricity…

2. Comparison

Example of an amino acid replacement cost matrix: Blosum



BLOSUM62

A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]																
Q	-1	1	[illegible]	[illegible]	-3	5															
E	-1	0	[illegible]	[illegible]	-4	2	5														
G	0	-2	0	[illegible]	-3	-2	-2	[illegible]													
H	-2	0	1	[illegible]	-3	0	0	[illegible]	8												
I	-1	-3	-3	[illegible]	-1	-3	[illegible]	-4	-3	4											
L	-1	-2	-3	-4	[illegible]	[illegible]	[illegible]	[illegible]	-3	2	4										
K	-1	2	0	[illegible]	[illegible]	[illegible]	[illegible]	-2	-1	-3	-2	5									
M	[illegible]	-1	-2	-3	[illegible]	[illegible]	-2	-3	-2	1	2	-1	5								
F	[illegible]	[illegible]	-3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	6							
P	[illegible]	-2	-2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-4	7						
S	[illegible]	[illegible]	1	0	-1	0	0	0	-1	[illegible]	[illegible]	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	[illegible]	[illegible]	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	[illegible]	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	[illegible]	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X

Negative for infrequent substitutions

Common amino acids have low weights

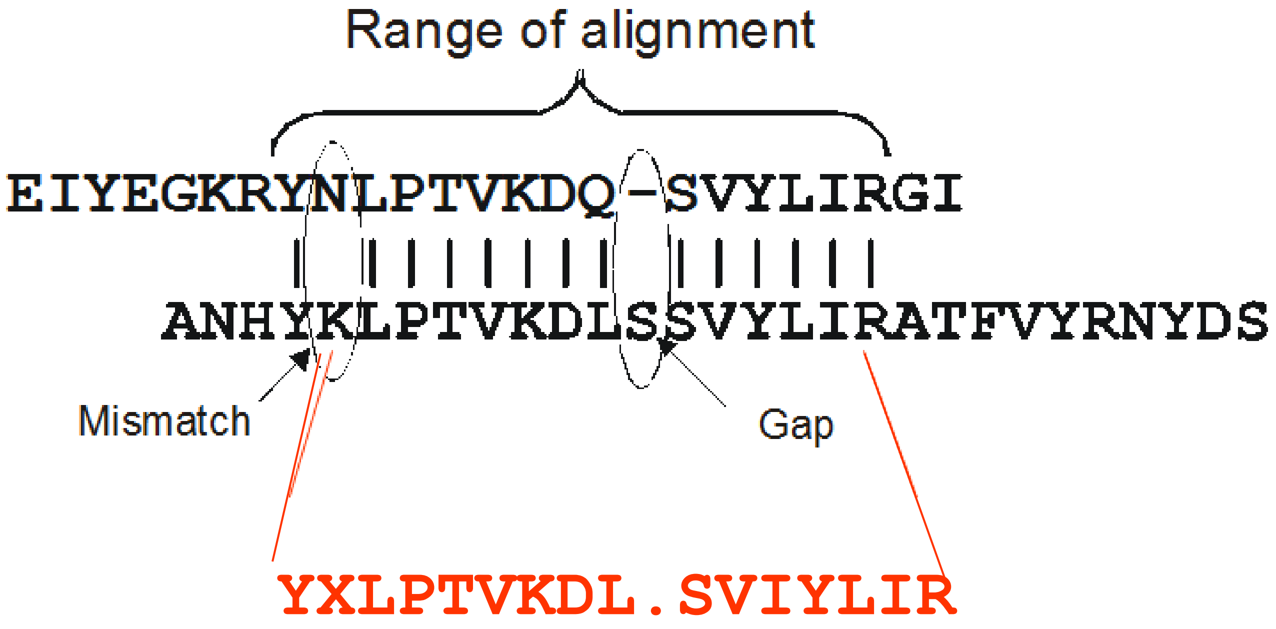
Rare amino acids have high weights

Positive for frequent substitutions

SEE next lecture

- The values can be derived from a large number of aligned sequences.

Motif between aligned sequences



Range of alignment

```
EIYEGKRYNLPTVKDQ-SVYLIRGI
       |  |||||||  ||||||
    ANHYKLPTVKDLSSVYLIRATFVYRNYDS
```

Mismatch

Gap

YXLPTVKDL.SVIYLIR

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

A multiple alignment

-Shared motifs point to evolutionary conservation. More informative than simple sequences

- „What a sequence whispers, an alignment pattern shouts out loud”

Granularity

- *Granularity* is the resolution of a description (e.g. of a vector)
- Too high resolution: all objects seem to be different, no similarity between members of the same group...
- Too low resolution: All objects are equal, no difference between different groups
- Finding the right amount of detail is hard. This is part of „the curse of dimensionality”

Vector example

1. Selecting representation

Granularity: Sequences as words

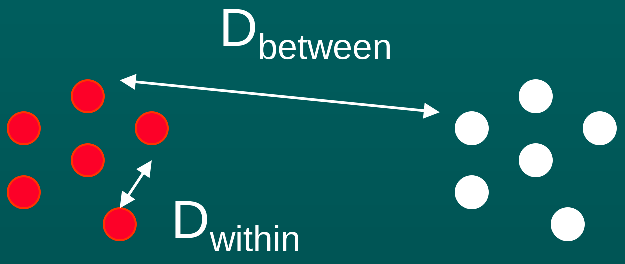
word-size	Alphabet size	
	4 (DNA)	20 (PROTEIN)
1	4	20
2	16	400
3	64	8000
4	256	160000
5	1024	3200000
6	4096	64000000
7	16384	1.28E+09
8	65536	2.56E+10
9	262144	5.12E+11
10	1048576	1.02E+13

- Sequences as vectors of word frequency.
- Di-, tri-, nucleotide, di-tripeptide descriptions
- Nucleotide composition is 4-dimensional, the trinucleotide description is 64 dimensional.
- At low dimensions, all sequences are similar. At very high ones all are different… (why?)

Vector example

Granularity: Is there an optimum?

- At low dimensions, all sequences are similar. At very high ones all are different…
- For a given problem, you can find an optimal resolution. Say you want to find the optimal granularity to separate two sequence groups, using word vectors (one of the groups can be very large a random selected group).

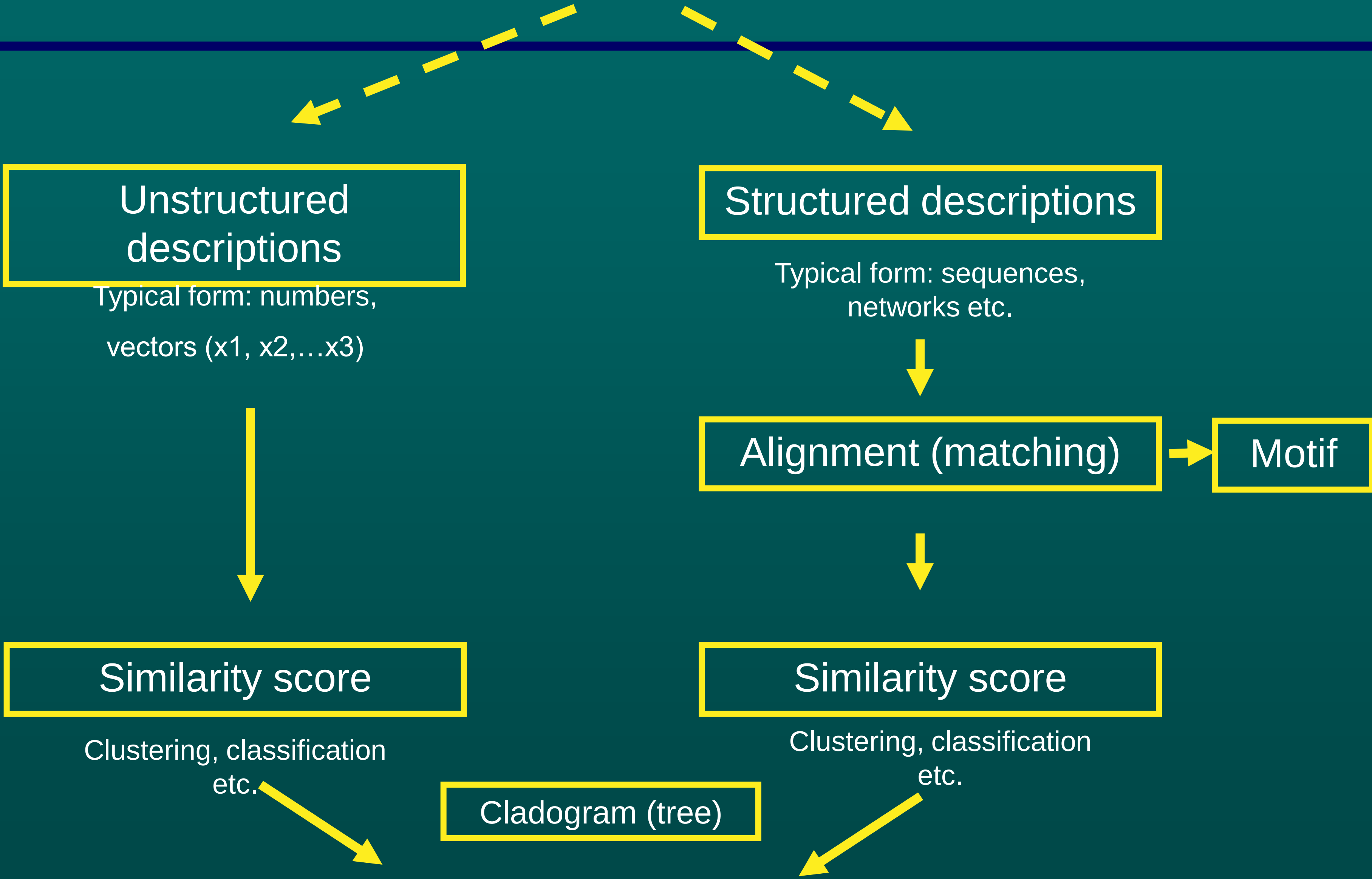


- You can calculate all distances within the groups and between the groups.
- You can compare the two distance sets with standard statistics, such as a t-test etc.
- You do this for all resolutions, and see - from a plot - if there is one which is more significant than the other.
- This is true of course only for normal distributions, but we can suppose that the bias will be the same for all points on the plot.

Sequence comparison (overview)

- We always compare two sequences/motifs. This is pairwise comparison or pairwise alignment. This gives a score and a motif (pattern).
- Two fundamental tasks (discussed in the next lectures):
  - 1) one sequence compared with each member of a database. Ranking hits by score, pick most similar. This is database searching.
  - 2) Members of a group compared with each other in an all-against-all fashion. Here again we have two tasks:
    - 2A find a common motif for the group. This is done by multiple alignment. Gives a common description for the group.
    - 2B Build a cladogram, or tree from the similarity scores. Shows the structure for the group with implications for evolution..

Quantitative comparison



Unstructured descriptions

Typical form: numbers,

vectors (x1, x2,…x3)

Similarity score

Clustering, classification etc.

Structured descriptions

Typical form: sequences, networks etc.

Alignment (matching)

Motif

Similarity score

Clustering, classification etc.

Cladogram (tree)

A math note on similarity and identity

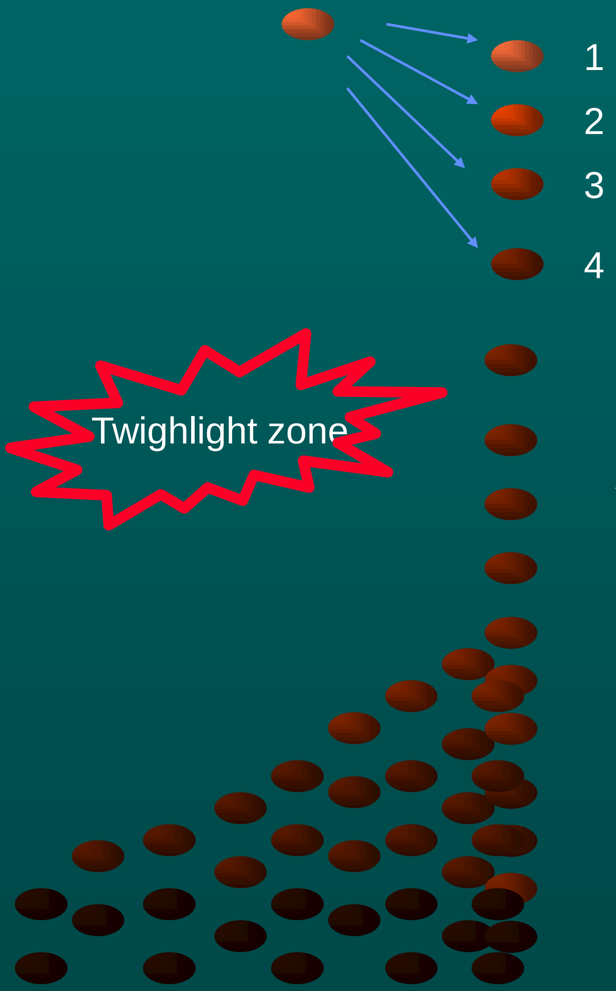
- Identity as a mathematical relation is symmetrical i.e. A~B →B~A, and transitive i.e. A~B~C → A~C

- Similarity is symmetrical and non-transitive A~B →B~A, but A~B~C does not mean A~C.

- Group membership by motif is partial identity (shared substgructure). This is transitive i.e. it is an identity relation.

- Group membership by simple score thresholding can be non-transitive. We can easily err to other groups…

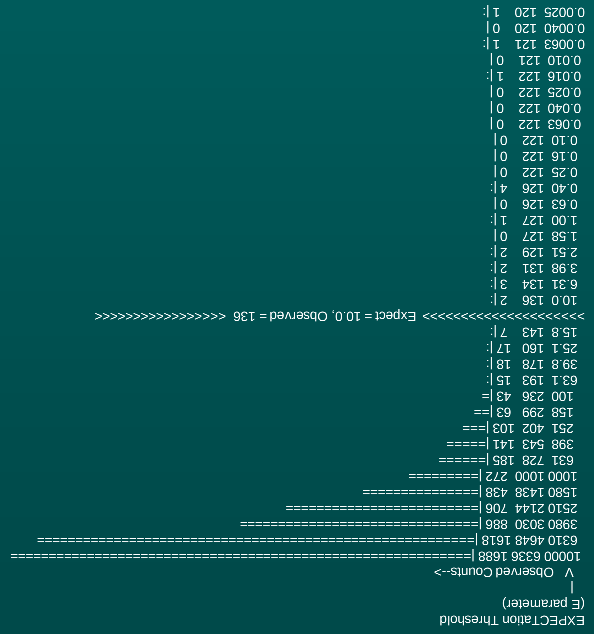
Using comparison 1: Comparing one sequence pairwise with a group (database) –database searching



Ranked list of best similarities

Similarities??

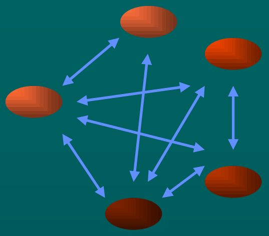
```
SEQUENCE  SCORE   DESCRIPTION

SWISSALL:AAI 457.36    ALPHA-AMYLASE INHIBITOR AAI
2/95

SWISSALL:O426 152.82    CELLULOSE BINDING PROTEIN

SWISSALL:GUX  145.77     EXOGLUCANASE I PRECURSOR

SWISSALL:Q126 145.66    CELLULASE (EC 3.2.1.91)
```



```
0.0025 120   1|:
0.0040 120   0|
0.0063 121   1|:
 0.010 121   0|
 0.016 122   1|:
 0.025 122   0|
 0.040 122   0|
 0.063 122   0|
  0.10 122   0|
  0.16 122   0|
  0.25 122   0|
  0.40 126   4|:
  0.63 126   0|
  1.00 127   1|:
  1.58 127   0|
  2.51 129   2|:
  3.98 131   2|:
  6.31 134   3|:
  10.0 136   2|:
>>>>>>>>>>>>>>>>>>>>>  Expect = 10.0, Observed = 136  <<<<<<<<<<<<<<<<<
  15.8 143   7|:
  25.1 160  17|:
  39.8 178  18|:
  63.1 193  15|:
   100 236  43|=
   158 299  63|==
   251 402 103|===
   398 543 141|=====
   631 728 185|======
 1000 1000 272|=========
 1580 1438 438|===============
 2510 2144 706|=========================
 3980 3030 886|===============================
 6310 4648 1618|=====================================================
10000 6336 1688|=========================================================
  V  Observed Counts-->
  |
(E parameter)
EXPECTation Threshold
```

BLAST program

Using similarity 2A: finding a common motif from an all-against-all comparison of a group.



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

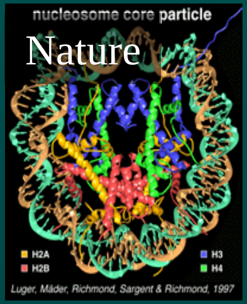
Mathematical consensus for database search

Similarity group or neighborhood

Multiple alignment

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

Find further examples in dbase

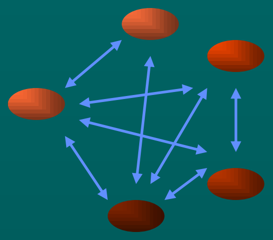


Visual and mathematical motif descriptions

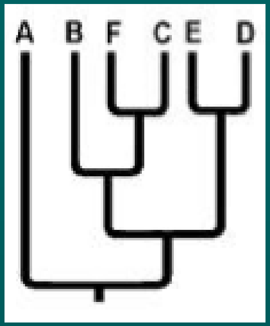
CLUSTAL program

Publis
h

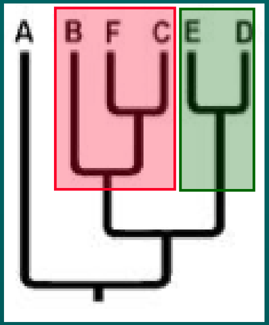
Using similarity 2B: building a cladogram (tree) from an all-against-all comparison of a group.



Similarity group or neighborhood

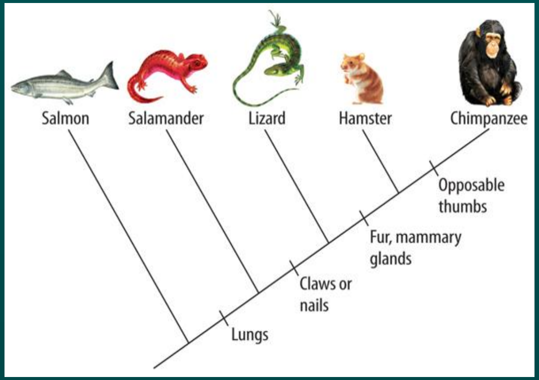


Cladogram



Analyzing subgroups (clades)

E.g. do they correspond functional classes such as cytoplasmic or extracellular versions of the same protein

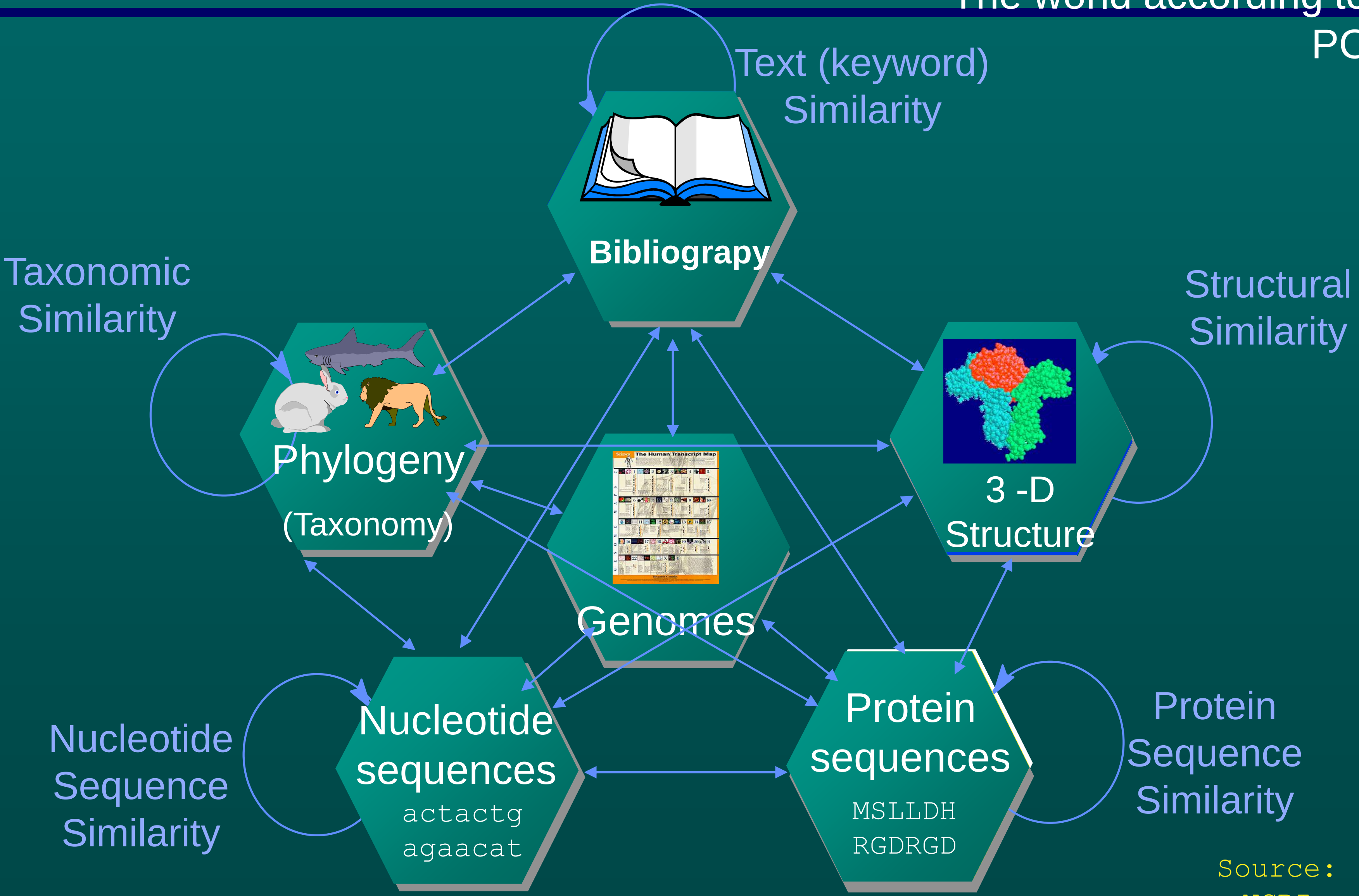


Evolutionary

Phylip package

Biological knowledge as a network of data

The world according to a PC...



Text (keyword) Similarity

Bibliograpy

Taxonomic Similarity

Phylogeny

(Taxonomy)

Structural Similarity

3 -D Structure

Genomes

Nucleotide Sequence Similarity

Nucleotide sequences

actactg
agaacat

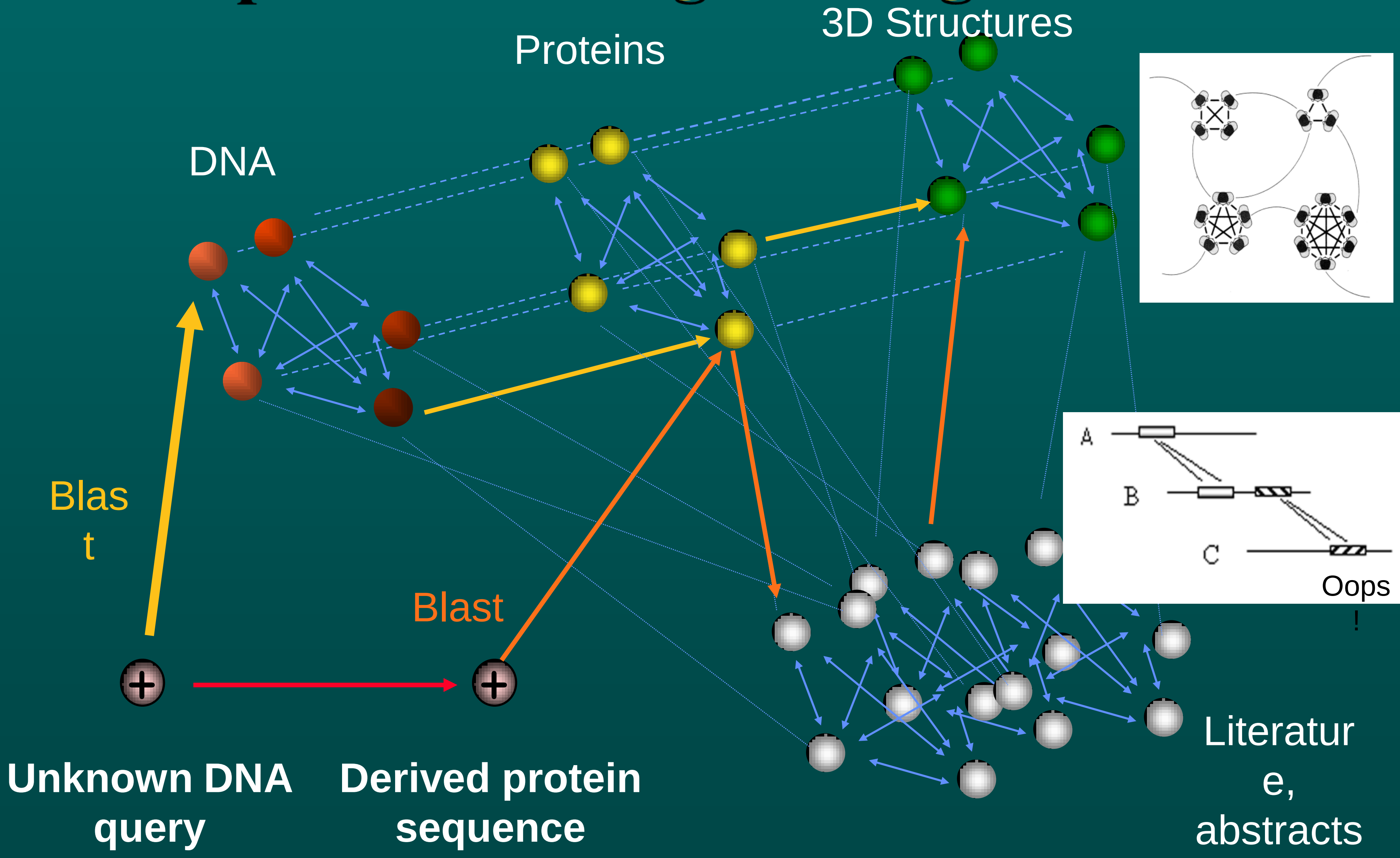
Protein sequences

MSLLDH
RGDRGD

Protein Sequence Similarity

Source: NCBI

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



3D Structures

Proteins

DNA

Blas
t

Blast

Unknown DNA query

Derived protein sequence

Literatur
e,
abstracts

What you should know

- Representations (unstructured, structured, mixed).
- Comparison: 1) Proximity measures (similarities, distances) 2) Motifs (from pairwise and multiple alignment of sequences)
- Main distance and similarity measures
- Granularity problems (in word descriptions)
- Basic applications of comparison: database searching, consensus building for groups (multiple alignment, tree)