Introduction to Bioinformatics

Databases, database searching

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics
Fall Semester, 2016/17

Oct 11, 2016

Last lectures

- 4 datatypes (sequences, structures, networks, texts)
- Sequence alignment (→ score, sequence motif)
- Multiple alignments
- Phylogenetics → tree

Indirect take home message: a sequence group is a cluster, connected by significant similarities, and it can be described with a multiple alignment, a common motif, a tree, a frequency matrix. Etc.

Bioinformatics databases

Sándor Pongor
net.icgeb.org/pongor

PPKE-ITK
ICGEB-Trieste

Outline

- Intro: Databases as a tool of communication, syntax, semantics
- Logical structure (briefly)
- Current dbase formats (even more briefly)
- Database contents
  - Sequence databases (primary/secondary, raw/annotated, comprehensive/specialized)
  - Onthologies (simple keywords, GO)
- Protein sequence clusters (the protein universe)
- Examples
  - UNIPROT
  - PFAM

The place of databases within bioinformatics I

- Algorithms and software are for gathering knowledge.

- Databases store and communicate knowledge.
  A kind of message (collection of messages)

The place of databases within bioinformatics II.

- All messages put (map) knowledge items into a standardized form (phrase structure, diagram, table… database record) and use one or more standardized languages.
- So databases have *syntax* and *semantics*
- Syntax and semantics is often formalized as ontologies (computer science term).

The evolution of messages

- Chemical signals of bacteria
- Cries, speech
- Stone carved inscriptions
- Codices, books
- Journals
- Encyclopedias
- Handbooks
- Scientific databases
- Internet: blogs, discussion lists, bulletin boards

Standardized language?
Standardized form?

...biological databases vs. books

- The classical databases are like *encyclopedias* (esp. UNIPROT)
- Fast development (as new data-types appear). Often contain additional database sections with unorganized data
- Daily/monthly updates (new data)

...Biological databases in brief

1 Searchable, organized (structured), regularly updated datasets

2 Specialize on certain data-types (e.g. protein sequences, DNA sequences, certain genomes) but also contain other types of data and are cross-referenced to may other databases.

3 Contain textual info, written in a standardized language (specified in external ontologies)

4 Often associated with special computational methods (similarity search, visualization etc) → WWW sites, resources.

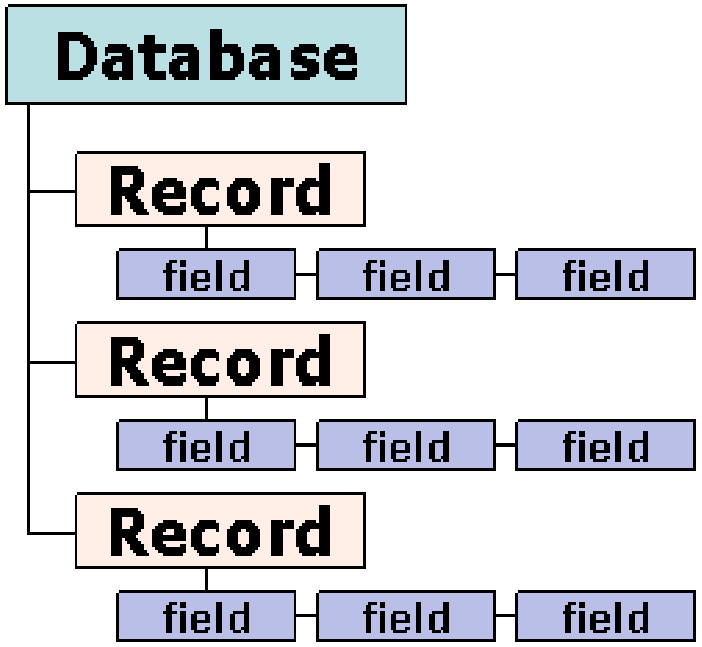
The technical structure of biological databases

- **Units: Records, contain fields subfields etc.**
- **In a flat file, each unit has its delimiter (record starts, record ends).**
- **There are mandatory fields, e.g. ID**
- **Example:**

  Record = one protein

  Field: name of protein

  sequence of protein



This is the same with all, not only biological databases

Current database formats

- Flat files →Human readable
- XML → suitable for linking various databases and visualization through HTML pages.
- RDF → Resource Description Framework (generalized WWW format)
- Relational databases (MySQL, ORACLE). Searchable via SQL (Structured query language). Links given in tables – data integrity guaranteed – generally used by db developers as the core dataset.
- Accessible forms: web page, dynamically generated (usually from a relational dbase)

THERE ARE SEPARATE LANGUAGES FOR THE MAIN DATA-TYPES

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

SEQUENCES

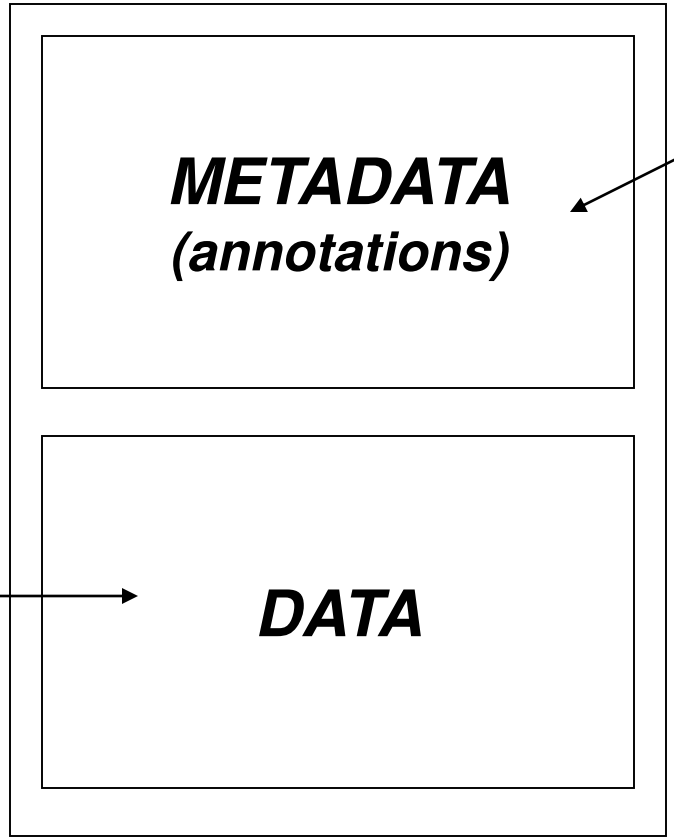
3-D

NETWORKS

TEXT

We focus on sequences

The logical structure of biological databases



Data on data i.e. :

a) Locally stored
b) Cross-references (links) to other dbases

We focus on sequences

This is the same with all, not only biological databases

THE STANDARDIZED LANGUAGE OF BIOLOGICAL SEQUENCES
I. SYNTAX

- DNA, RNA: nucleotide alphabets, 4 main characters, plus wild cards, nucleotide groups, modifications
- Protein: amino acid alphabet, 20 main characters+ wild cards, etc.
- Today dbases use one letter codes, there are other (more or less historical) conventions, sometimes used in arfticles.

http://www.chem.qmul.ac.uk/iupac/misc/naabb.html

http://www.chem.qmul.ac.uk/iupac/AminoAcid/AA1n2.html

THE STANDARDIZED SYNTAX OF BIOLOGICAL SEQUENCES

- One type of relationship: sequential vicinity
- (OK, there are upstream and downstream neighbors)

http://www.chem.qmul.ac.uk/iupac/misc/naabb.html

http://www.chem.qmul.ac.uk/iupac/AminoAcid/AA1n2.html

THE FASTA SEQUENCE FORMAT

Record seaparator

Name, arbitrary form

```
>sp|P12746|LUXR_VIBFI Transcriptional activator
MKNINADDTYRIINKIKACRSNNDINQCLSDMTKMVHCEYYLLAIIY
NYPKKWRQYYDDANLIKYDPIVDYSNSNHSPINWNIFENNAVNKKSE
GFSFPIHTANNGFGMLSFAHSEKDNYIDSLFLHACMNIPLIVPSLVD
NDLTKREKECLAWACEGKSSWDISKILGCSERTVTFHLTNAQMKLNT
GAIDCPYFKN
```

Sequence in one letter code

Used by most programs. A series of such sequences is a concatenated FASTA file

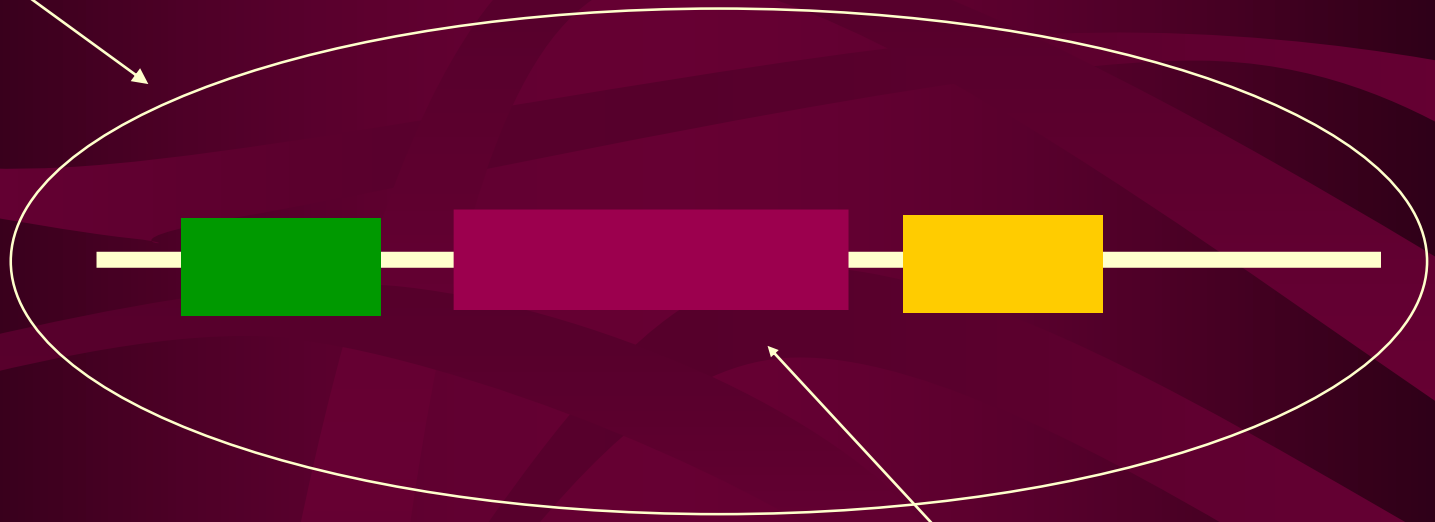
OTHER SEQUENCE FORMATS ARE USED MOSTLY FOR VISUALIZATION

```
        10         20         30         40         50         60
MKNINADDTY RIINKIKACR SNNDINQCLS DMTKMVHCEY YLLAIIYPHS MVKSDISILD

        70         80         90        100        110        120
NYPKKWRQYY DDANLIKYDP IVDYSNSNHS PINWNIFENN AVNKKSPNVI KEAKTSGLIT

       130        140        150        160        170        180
GFSFPIHTAN NGFGMLSFAH SEKDNYIDSL FLHACMNIPL IVPSLVDNYR KINIANNKSN

       190        200        210        220        230        240
NDLTKREKEC LAWACEGKSS WDISKILGCS ERTVTFHLTN AQMKLNTTNR CQSISKAILT
```

Used in some databases, mostly for human use. Similar notations for EMBL, UNIPROT etc databases. FASTA is always offered as an alternative.

The traditional view on sequence data

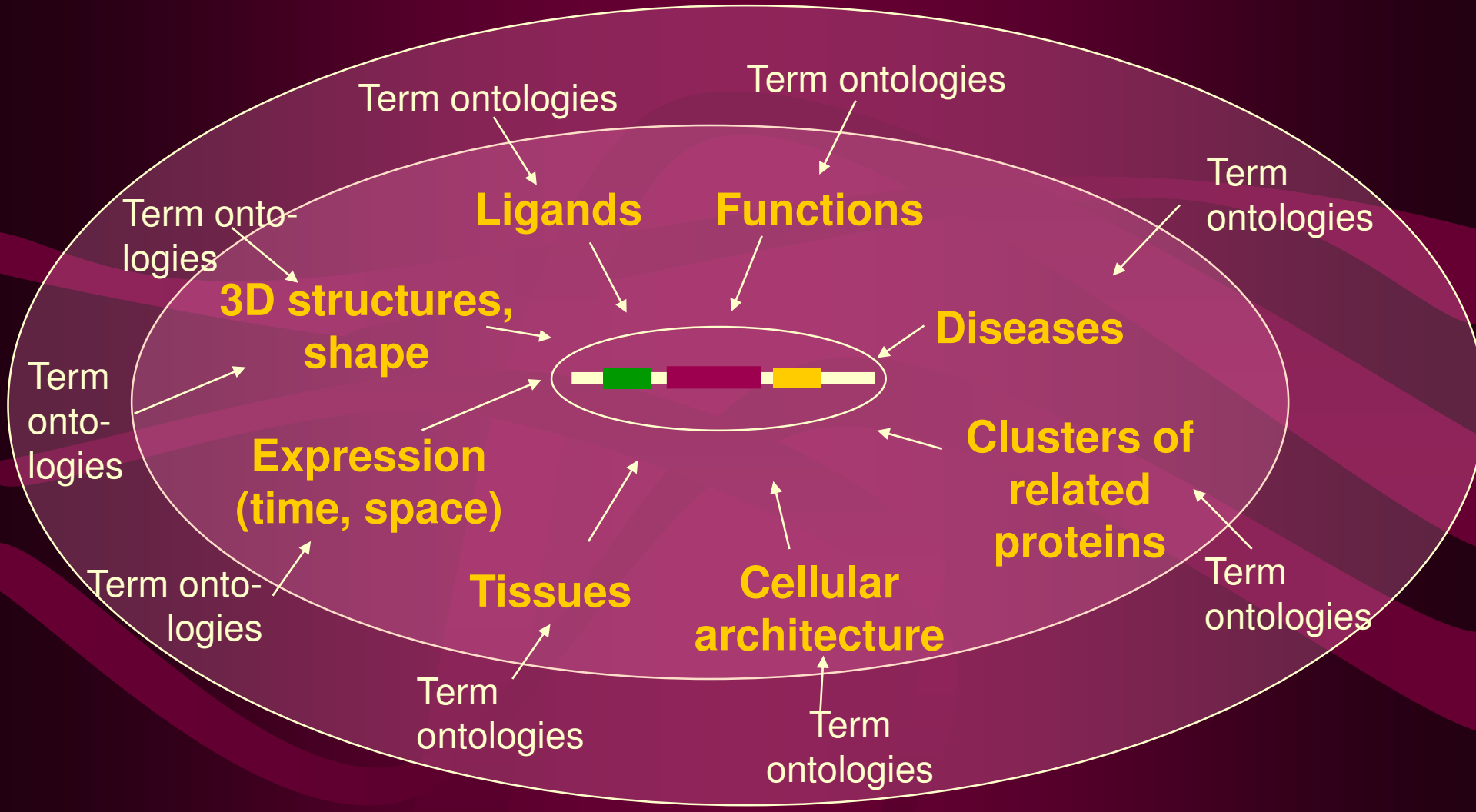
Global descriptors
e.g. function



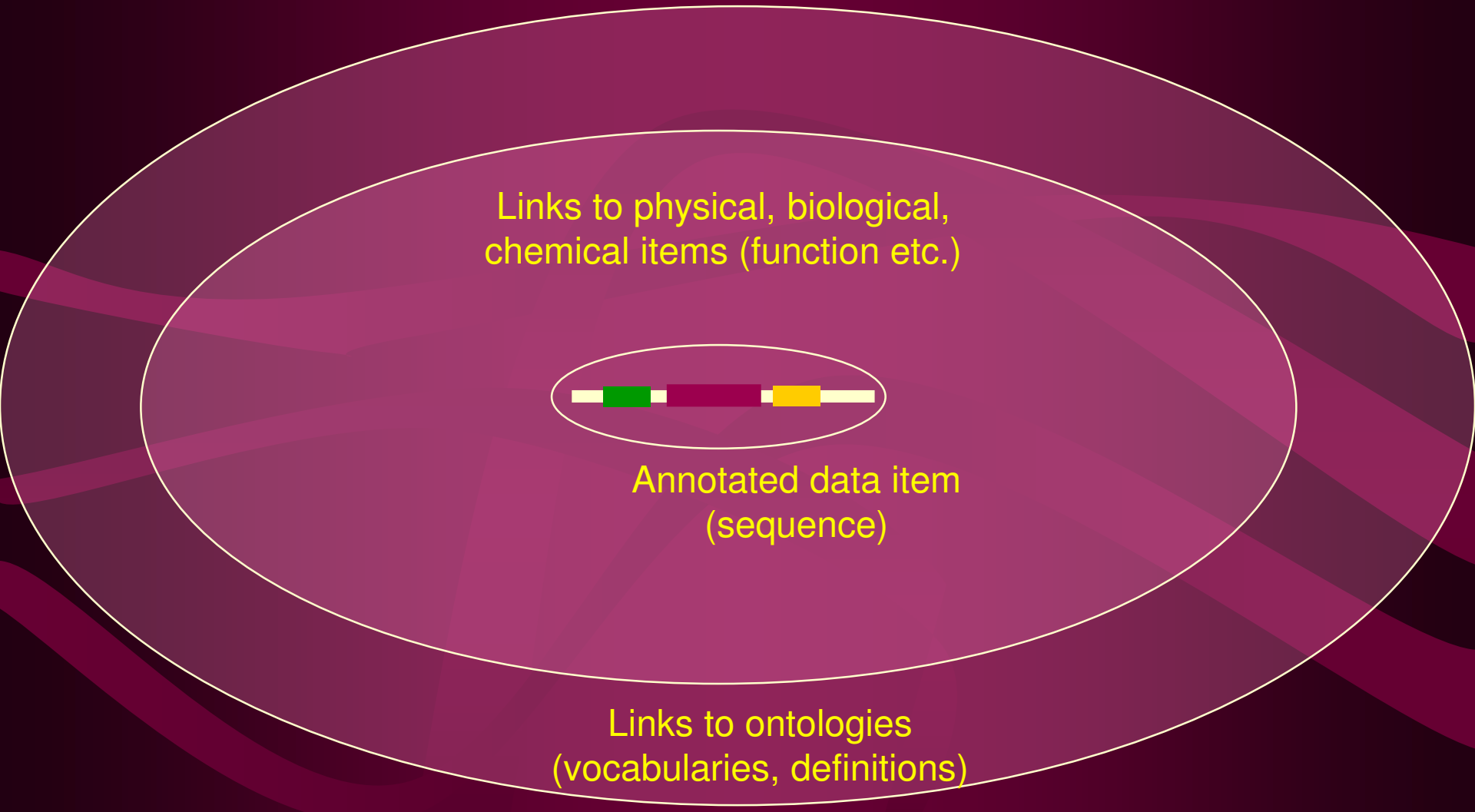
Annotation requires database search and knowledge of biology

Local descriptors e.g. binding sites, domains

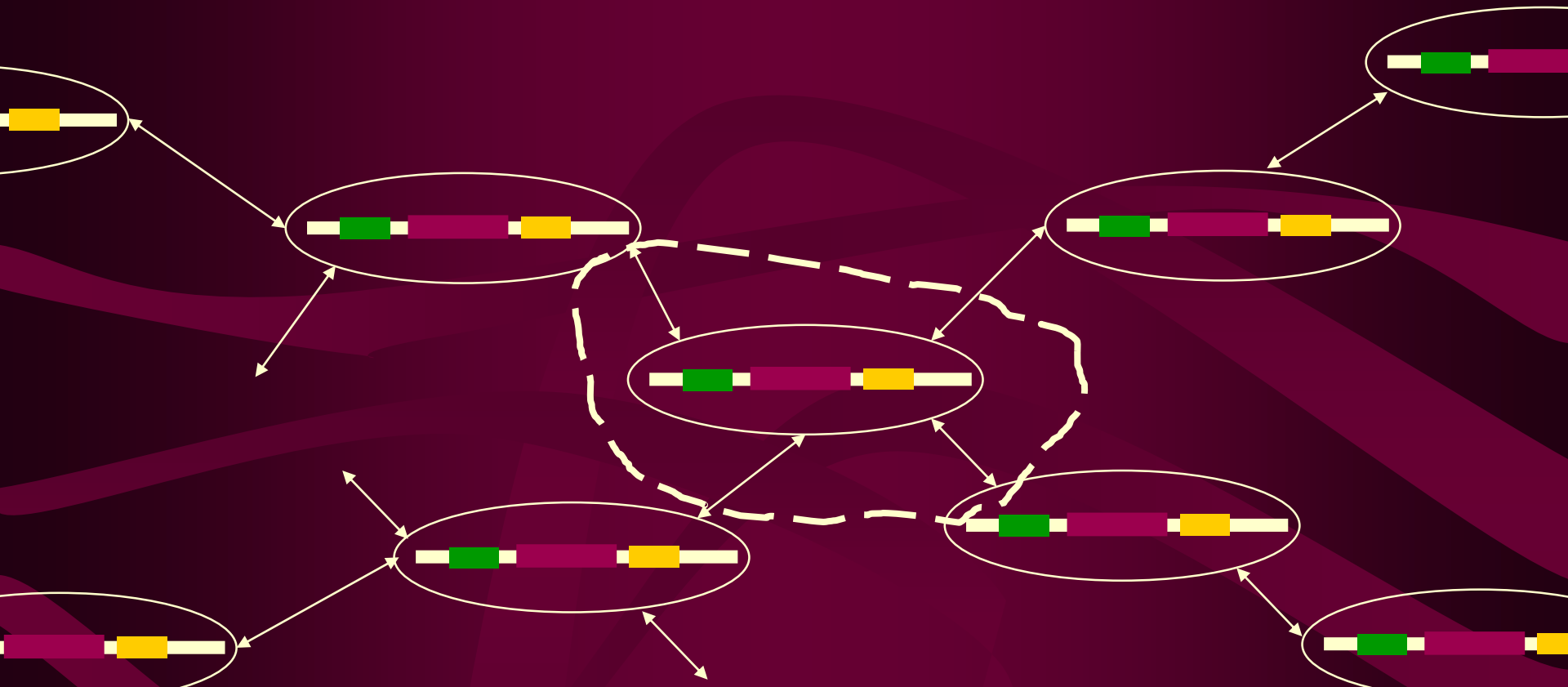
Current view on sequence data 1: systematically annotated data



Current view on sequence data 1: systematically annotated data A

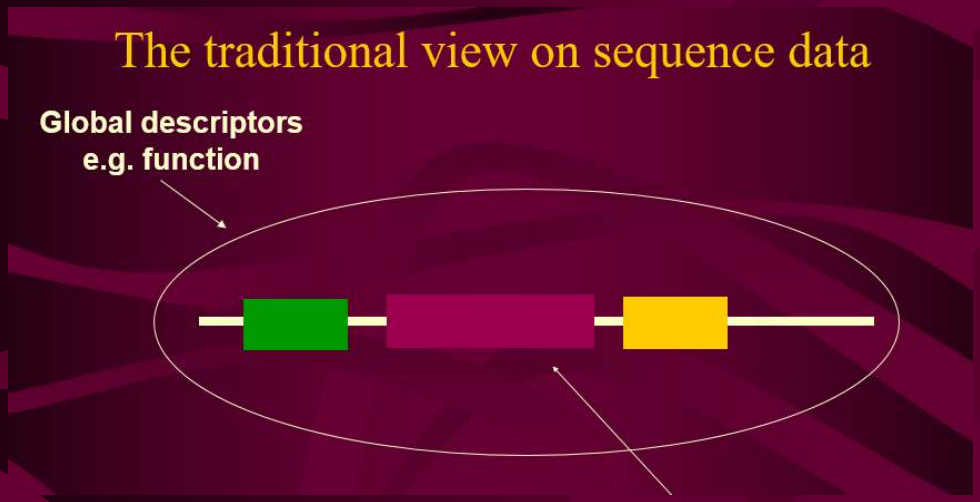


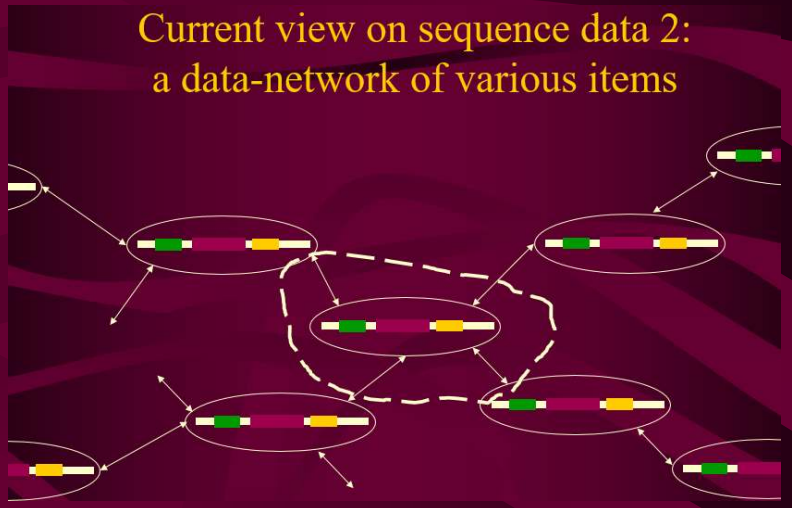
Current view on sequence data 2: a data-network of various items



- A dense network connected by explicite (clickeable) and potential (computable) links
- An item is an arbitrarily delimited subnetwork, not an independent unit.

Traditional and current views on bioinformatics data



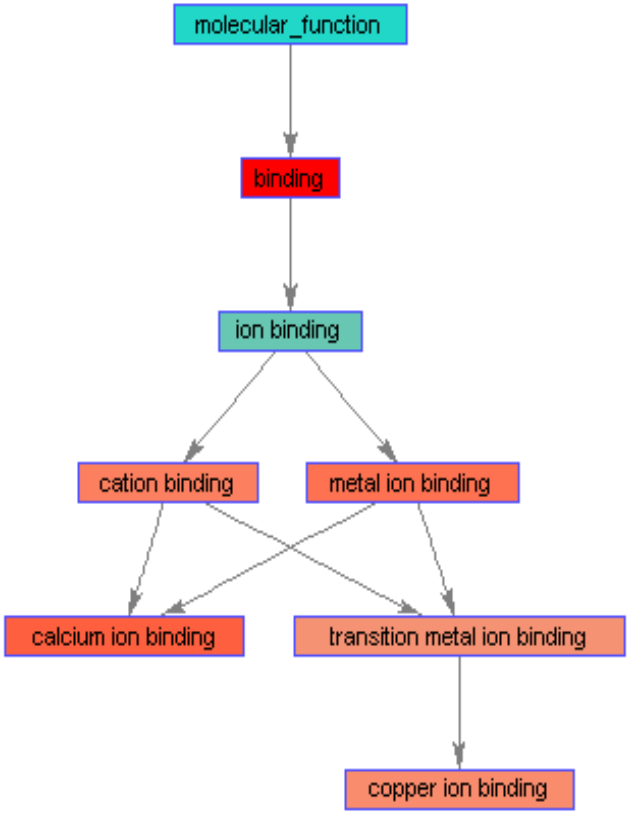


The „data network” is never „complete”. Realistic database architecture include only selected items.

The language of metadata

- Separate (arbitrary) conventions for bibliographic data, for version history, cross references, etc
- Organized languages for function (Gene ontology), for 3D structure (Protein Ontology) etc.
- Organized structures for local descriptors: feature table of GFF format.

II. SEMANTICS: Standardized concepts + standardized language = ontology.



1) Simplest form: added keywords

2) The Gene Onthology (GO) project standardizes the names and functional description of genes and proteins, in the form of concept hierarchies (directed acyclic graphs).

→ Comparable descriptors: "copper binding" is near "ion-binding").

Separate GO ontologies
Molecular function,
Cellular component,
Biological process

www.geneontology.org/

The long and winding road towards machine-readability

Main tasks in database construction

(sequence databases as an example)

1. Data collection, storage
2. Validation (e.g. is the ORF real?)
3. Clustering (Classification, redundance filtering)
4. Annotation (crossrefs + text)
5. Integration, visualization

How many kinds of protein sequence databases?

- Raw (only sequence + some predicted function) vs. annotated (provided with additional info)
- Comprehensive (all organisms) vs. species specific (e.g. human, yeast etc)
- Primary (full sequences) vs. secondary (derived entities, like domain sequences, modification sites etc.)

Notes:

1) These are just the main points, there are countless other, “boutique” databases
2) Dedicated journals: Nucleic Acids Research (NAR) Database issue, NAR Web issue, “Database” Journal

I. How raw protein sequence databases are prepared/updated

Nyers adatbázis előállítása automatikus módszerrel

- Data collection: genome sequencing projects, infidual sequences → DNA sequences → machine translation into open reading frames (ORFs)
- Redundance filtering: preparation of a nearly non-redundant dataset (e.g. using clustering)
- Comparison with previous release - a dataset of known and function-annotated proteins. In case of strong similarities a putative function is assigned ("probable protease") as an ID
- The TrEMBL database (Translated EMBL) is prepared this way.

Identification of Open reading frames (ORFs) in prokaryotes

- An open reading frame in a raw DNA sequence may or may not code a protein sequence. Rough translation according to the codon usage table (CUT) can be misleading.
- Simple programs, like GLIMMER (Steve Salzberg) identify runs of translatable codons (in 6 reading frames), delimited by start and stop codons.
- Start/stop is identified by CUT + short motifs statistically determined previously for a taxonomic group. Probable sequence errors are corrected.
- ORFs above a length threshold are reported. Almost perfect, in practice we find erroneous stops, starts probably due to too low quality sequences.

Example of a first step…

II. Preparing an annotated database

- Annotation is info added to raw data.
- Based on human intervention, literature searching, similarity searching, calculations,
- Main elements
- Literature citations (where was it published, where was the sequence determined, etc.)
- Cross-reference to other databases (3D, DNA, gnomes)
- Known or predicted functions,
- Known mutations, variants.
- Known parts (e.g. domains, active center of enzymes, metal-binding sites etc.).

How sequence databases are organized....

Genomics

Complete genomes

draft genomes

Individual DNAs

Translation

Complete proteomes

Individual protein sequences

MA, HMM

Annotation, validation

Traditional bioinformatics

Protein sequence databases

DNA sequence databases

Group entire sequences or subsequences, or by structure, add extra annotations...

MA, HMM

MA, HMM

Secondary databases for domains, or for functional classes, PFAM, SMART SCOP, CATH,

Secondary databases for organisms, motifs, etc.

Examples

- Comprehensive protein database → Uniprot

http://www.uniprot.org/

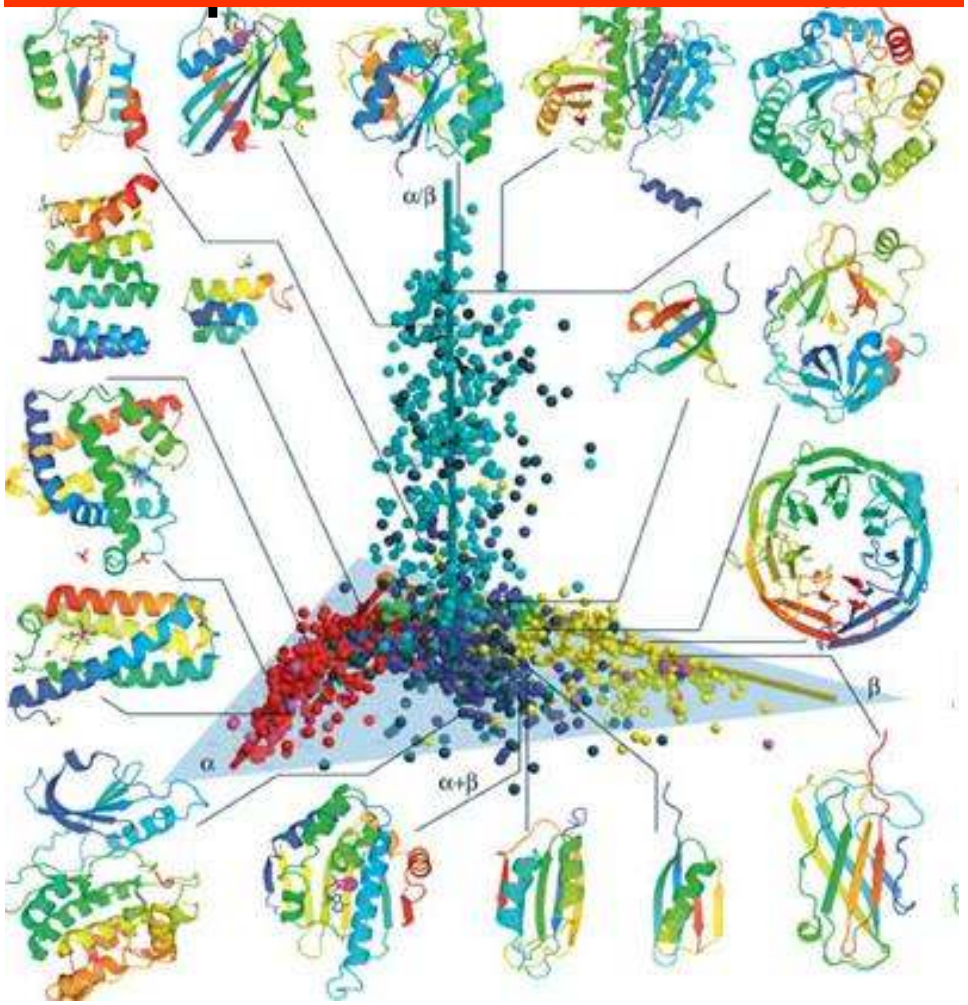
- Protein domain database→ Pfam

http://pfam.janelia.org/

Concepts you need to know for understanding protein databases

- Domains (structural and evolutionary units, parts in larger sequences, amenable to multiple alignment)
- Single domain proteins, multidomain proteins
- COGs (clusters of ortologous proteins): sequences of common evolutionary origin carrying the same function
- Protein families: Like COGs. Emphasis on sequence and structural similarity
- UNIREF clusters: automated sequence clusters, characterized by % identity only.

What are clusters?
The protein Universe

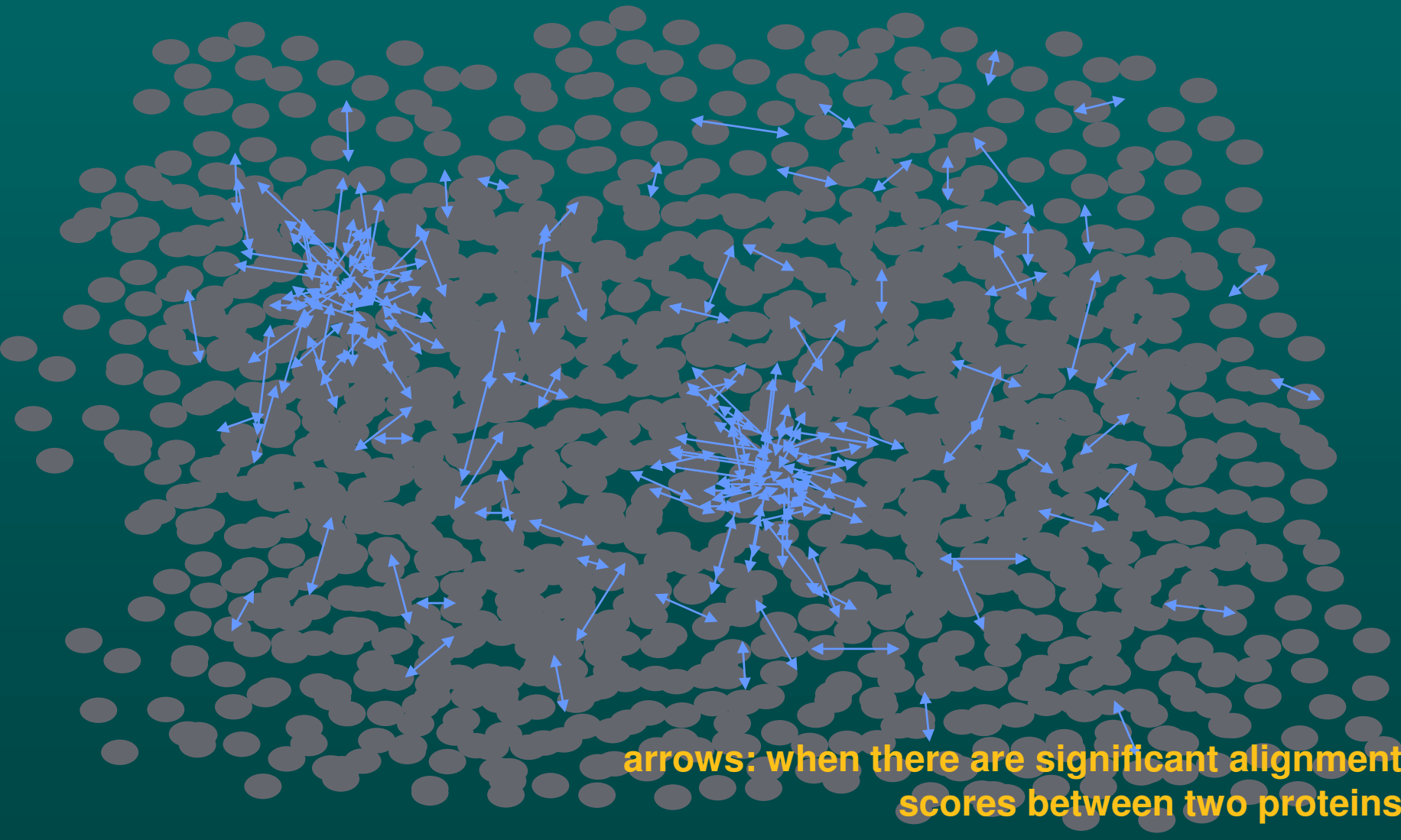


A weighted graph,

nodes = proteins

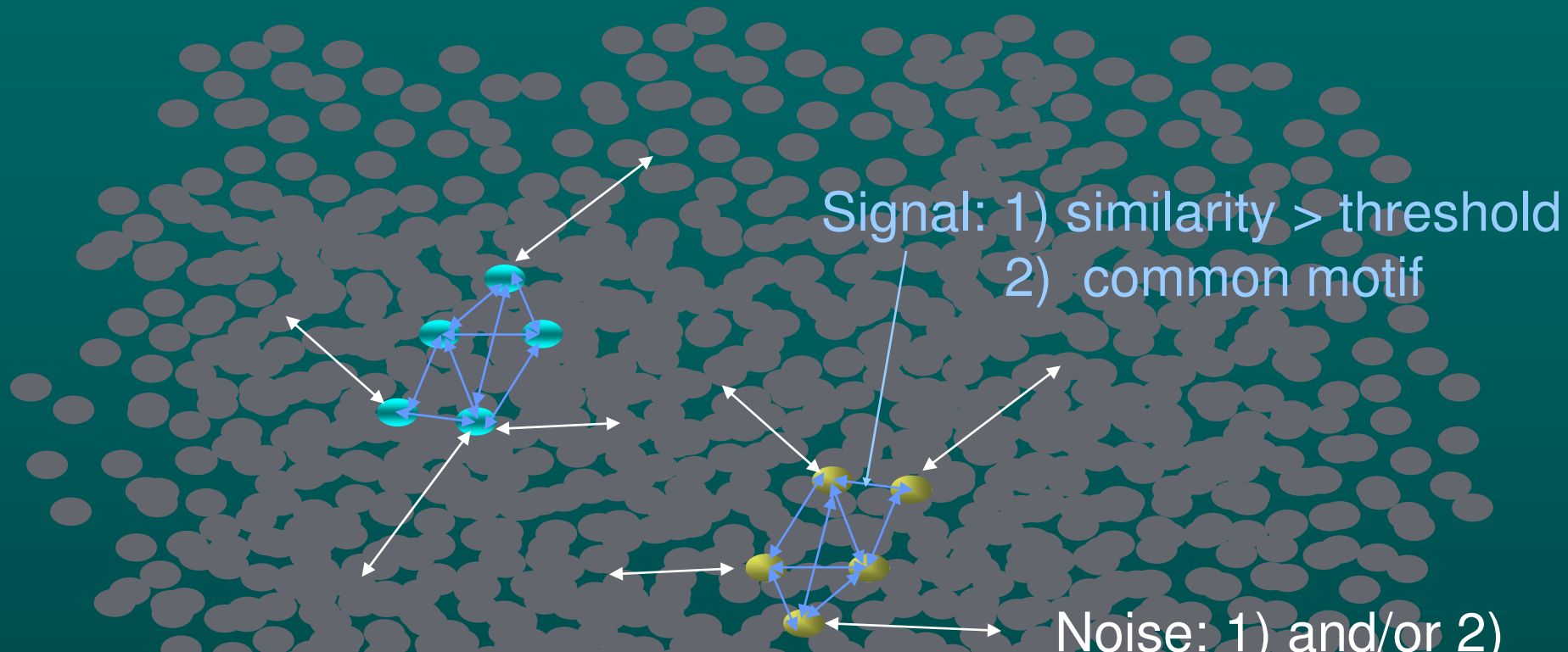
edges = similarities

The (protein sequence) database as network of similarities: clustering



arrows: when there are significant alignment scores between two proteins

From data to knowledge in a protein sequence database: distinguish signal from noise



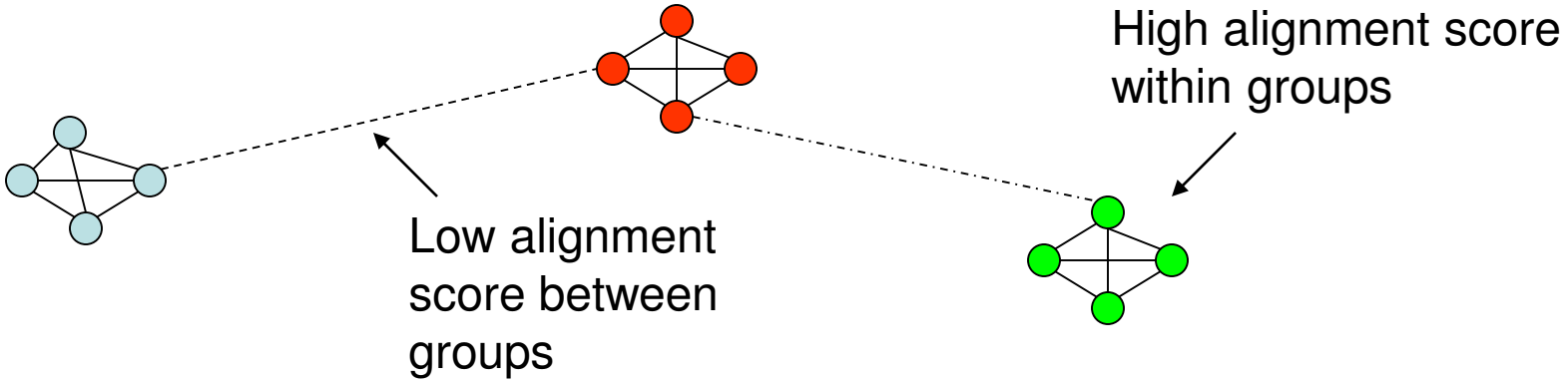
Signal: 1) similarity > threshold
2) common motif

Noise: 1) and/or 2) missing

This is a knowledge-base of meaningful similarities

We know some of the group members in advance (= supervised clustering)

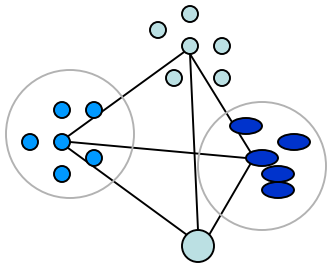
The protein universe (as a sequence similarity network)



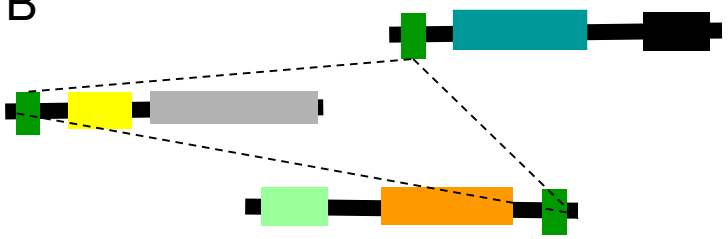
- Proteins of all organisms is “the protein universe”. A cumulative result of evolution…
- Contains many tightly connected “similarity groups” – these are those protein families that share structure and function (orthologs)
- Usually : If a sequence is >90% identical with members of very little similarity between the groups
- Trivial classification is to assign a sequence to a group where it belongs there… Most protein similarities are trivial so classification can be almost always can done by alignment.. The rest is *difficult*!!!

- Proteins consisting of a single domain form +/- clearcut groups (families), but deeper analysis may reveal subroups (A)
- Multi-domain proteins are connected to many different families and are difficult do deal with because of the shared domains (B)

A



B



1) Knowledge-based clusters

- Database of orthologous groups (COG) NCBI – full sequences, best for prokaryotic groups (function-based)

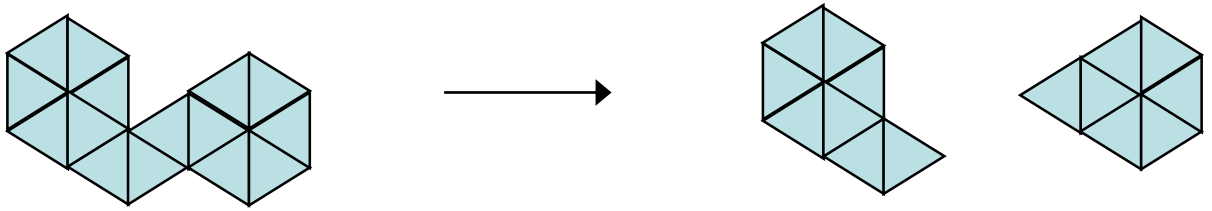
- SBASE library of protein domain sequences – domains, local homology assignments (developed originally by our group)

COG clustering

- Detect triangles of best hits between genomes

- Merge triangles with a common side to form COGs

- Case-by-case 'manual' analysis, examination of large COGs (might be split up)



COGs Categories

INFORMATION STORAGE AND PROCESSING

[J] Translation, ribosomal structure and biogenesis
[A] RNA processing and modification
[K] Transcription
[L] Replication, recombination and repair
[B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING

[D] Cell cycle control, cell division, chromosome partitioning
[Y] Nuclear structure
[V] Defense mechanisms
[T] Signal transduction mechanisms
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[Z] Cytoskeleton
[W] Extracellular structures
[U] Intracellular trafficking, secretion, and vesicular transport
[O] Posttranslational modification, protein turnover, chaperones

ftp://ftp.ncbi.nih.gov/pub/COG/COG/fun.txt

COG/KOG orthologous groups

- ~3000 prokaryotes (bacteria, archaea, viruses)
- ~5000 for eukaryotes

Now the completely human curated COG/KOG is still intensively used, but there are more recent versions, like EGGNOG, which is to a large part machine/created

Eugene Koonin, NCBI

Other approaches to knowledge based protein clustering

- The goal is to find domains, i.e. autonomously evolving units of protein structure
- Early approaches relied on manual work
  - PROSITE: clusters described with sequence motifs (Swiss Bioinformatics Institute)
  - SBASE: clusters defined as searchable sequence collections (our group) (similar to COG)

Other approaches to knowledge based protein clustering

- Modern approaches use clusters (at least one member) with known 3D structures (if available)
- Groups represented as multiple alignments, profile.
- Typical example: PFAM (Sanger Institute, Cambridge, UK)

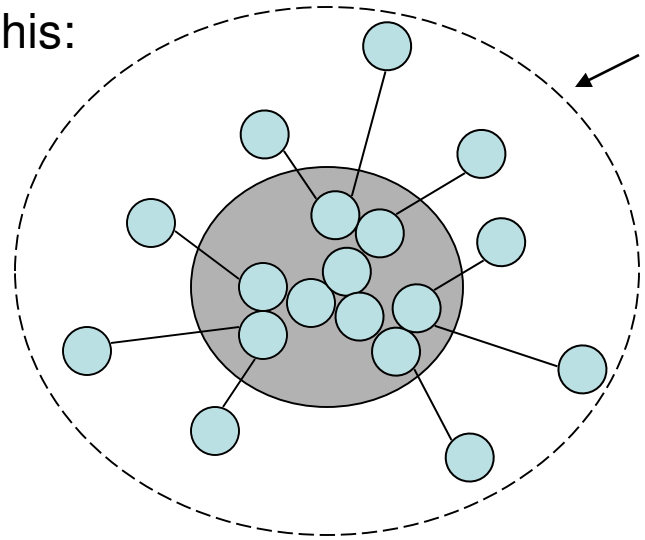
Why are 3D based clusters important?

- Because of evolution. (Most) Proteins evolve by retaining a typical 3D shape
- As genes are duplicated, several independent sequence families arise within the same shape. Then slowly, new shapes emerge.
- So a combined 3D + sequence classification gives insight into long range evolution.
- Typical example: PFAM (Sanger Institute, Cambridge, UK)

Now: unsupervised classification

But how do we automatically find sequence clusters?

Like this:

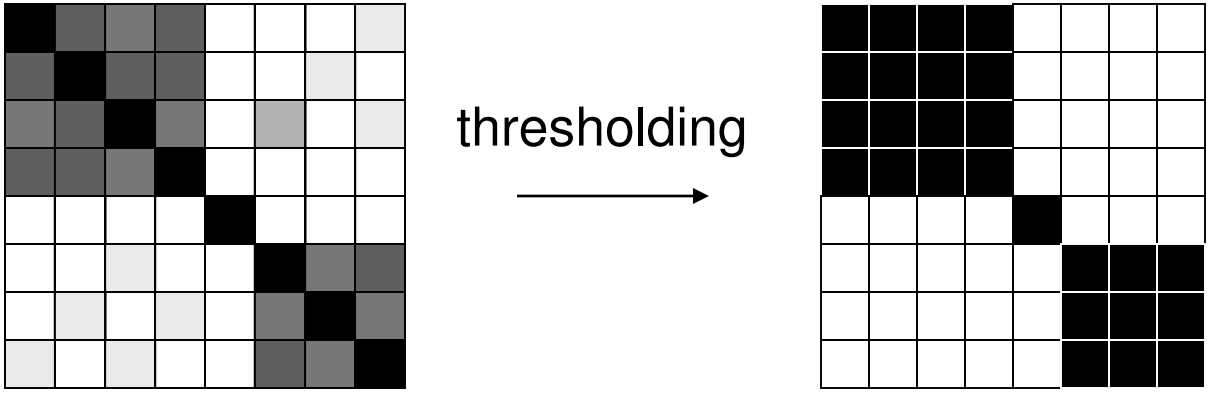


Complete similarity group (difficult, needs manual checking/annotation)

"X% identity group" (Easy, 90% identity indicates functional identity.These are the un-annotated clusters, like UNIREF)

Protein sequence clustering

Heuristic solution to sequence clustering...



- All against all comparison gives a matrix of similarity measures
- Apply a strict threshold: sequences above 90% residue identity belong to the same cluster, otherwise not. (Tarján Róbert's method for finding connected components in graphs)

Protein sequence clustering

What is then the problem? (1)

- Problem 1: Time All-against-all is too expensive to compute with the only suitable algorithm (string alignment with dynamic programming).
- Solution 1: Use an inexpensive description at first, like word presence-absence vectors

Protein sequence clustering

What is the problem? (2)

- Problem 2: Memory Word-frequency vectors AND All-against-all matrices are too big (but fortunately, sparse)
- Solution 2.1: Use a compressed representation for word frequency matrices
  a) hash table (sequence *x*: word1, word2,..)
  or b) index table (word x: seq1, seq2, ...)
- Solution 2.2: Compute all-against-all only for the groups found OK by word frequency..

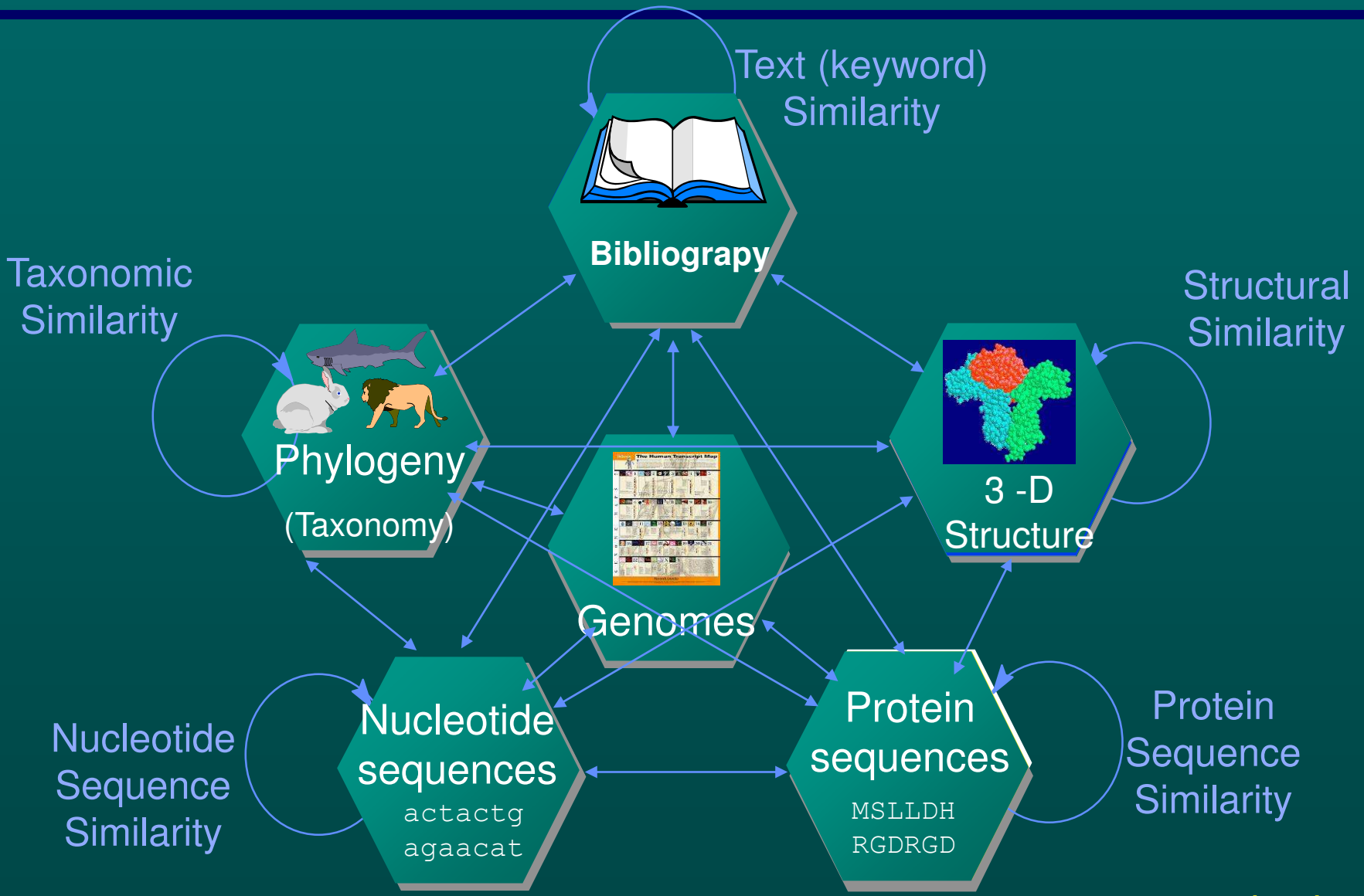
Protein sequence clustering

The (current) solution for protein sequence clustering

- Represent sequences as an index table
- 1) For sequence i=1: retrieve all sequences that share a „sufficient number” of identical words.
- 2) Do an all-against-all comparison for the retrieved group. Those sequences that are above threshold (say, 90 % identity), are recorded in cluster 1, and excluded from further comparison.
- Repeat steps 1 and 2 until there are no more sequences.

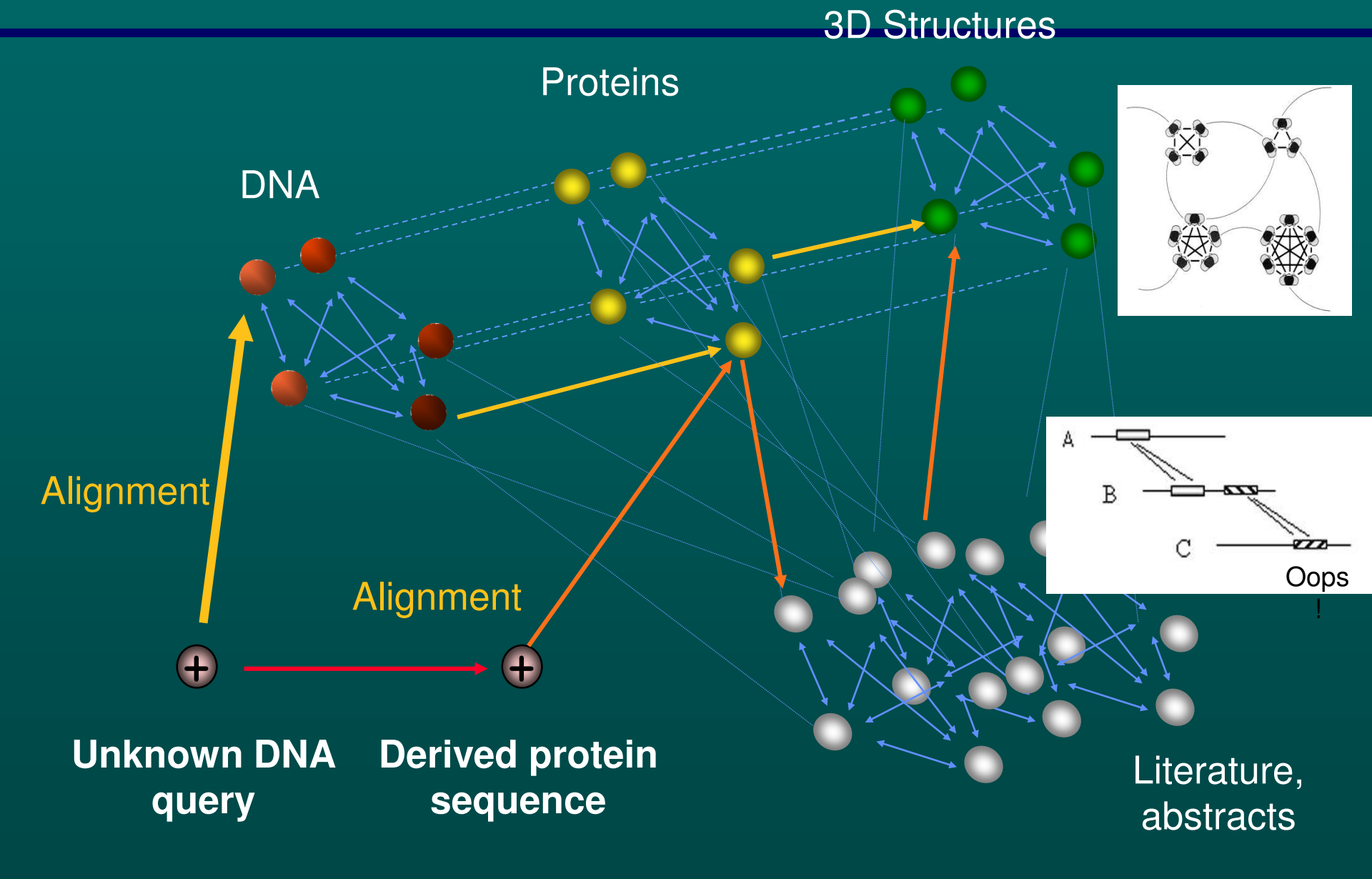
This is how the clusters of the UNIPROT sequence database are prepared (Adam Godzik, CD-hit program)

An integrated database resource at the NCBI: a network of clustered data



Source: PubMed

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



What you should know

!

- Main data types, main tasks
- Logical structure and current dbase formats
- Generation of raw databases and annotated databases
- Types of sequence databases (primary/secondary, raw/annotated, comprehensive/specialized)
- Onthologies
- Protein Universe as Clusters of proteins
- Examples
  - Uniprot (main current sequence dbase)
  - PFAM (main domain dbase)

… will be shown during practicals)