

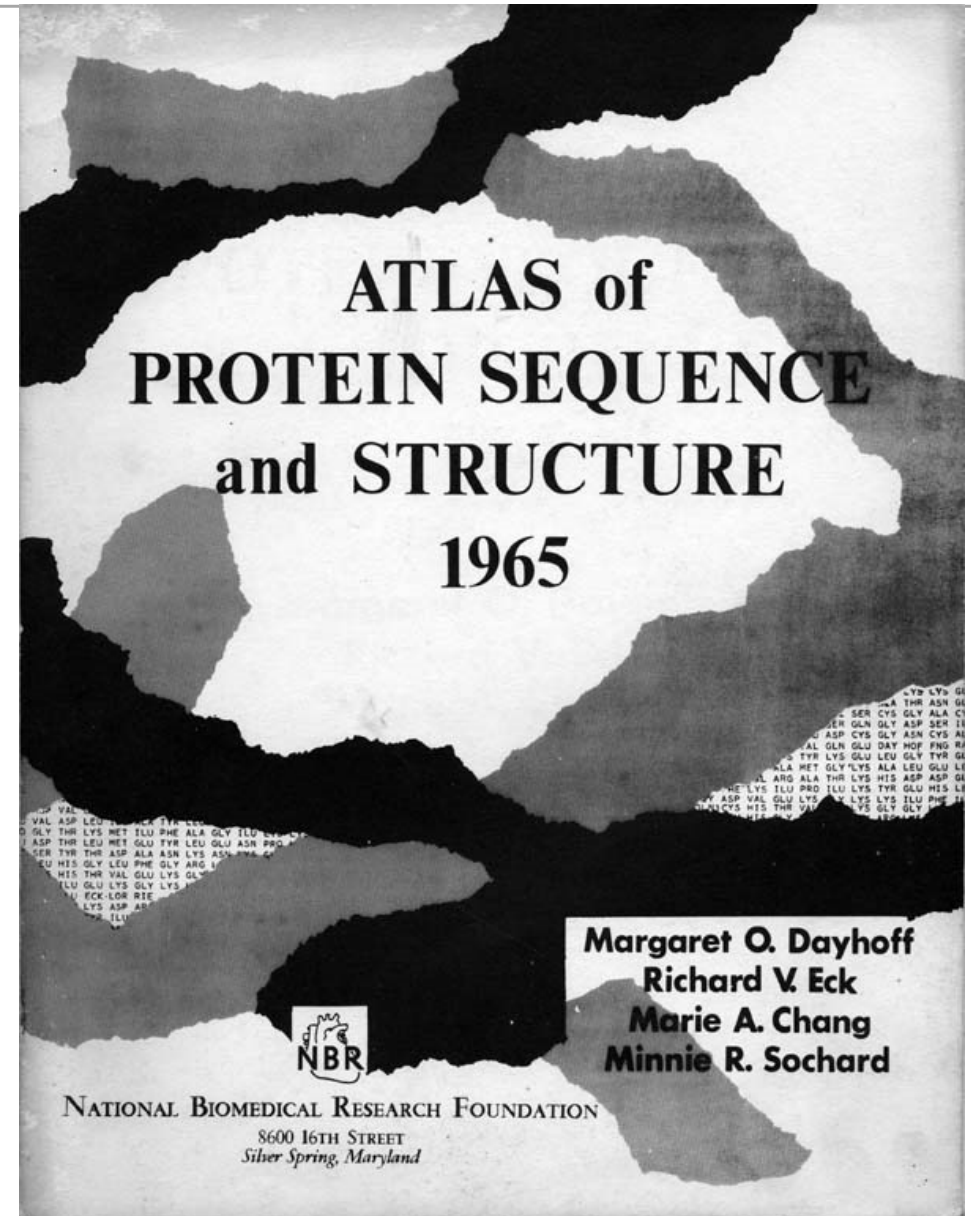
Large scale protein sequence clustering

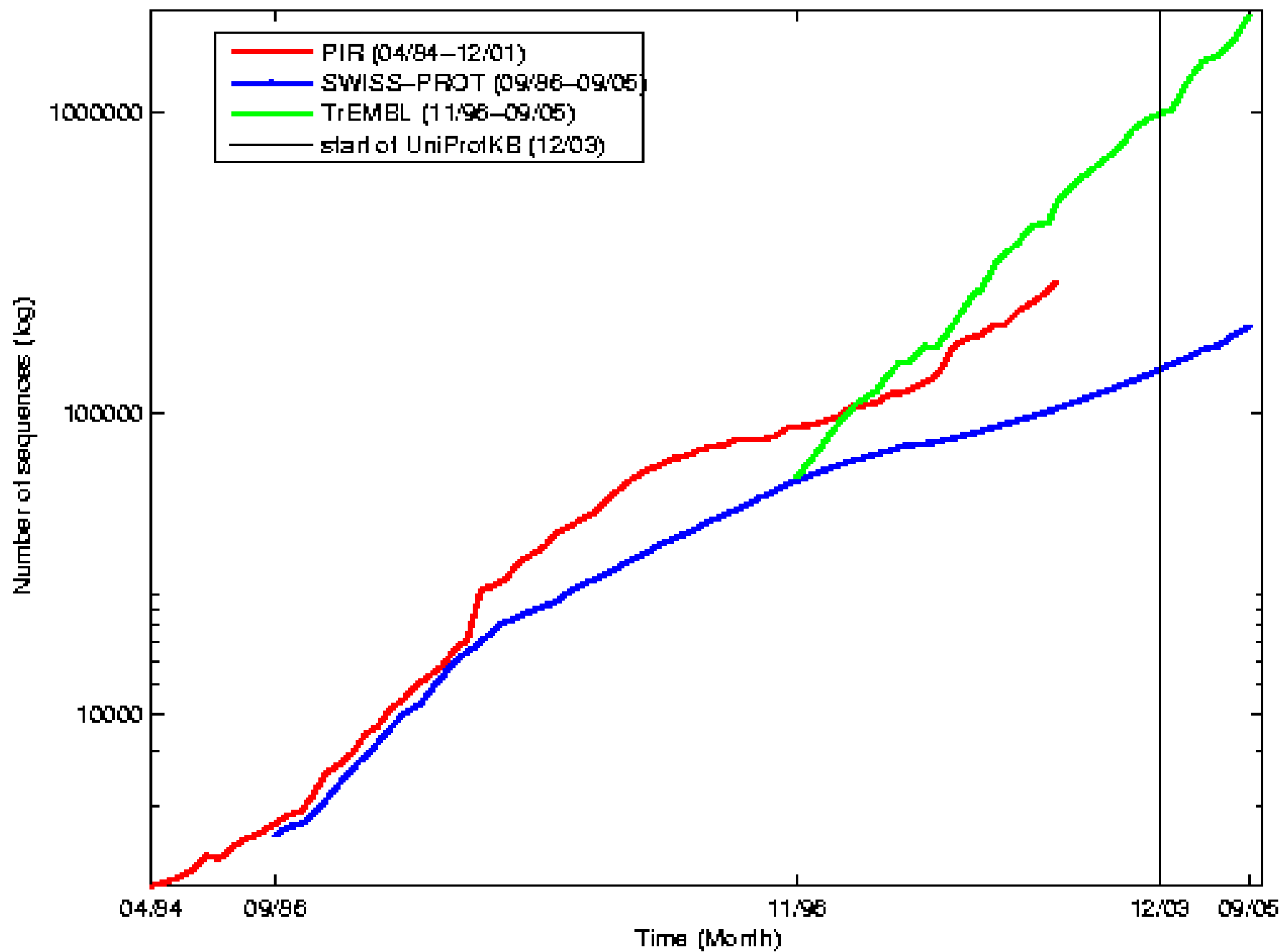
Prof. Dr. Antje Krause
Bioinformatik
FH Bingen
akrause@fh-bingen.de

“Dr. Margaret Oakley Dayhoff (1925-1983) was a pioneer in the use of computers in chemistry and biology, beginning with her PhD thesis project in 1948. Her work was multi-disciplinary, and used her knowledge of chemistry, mathematics, biology and computer science to develop an entirely new field. She is credited today as a founder of the field of Bioinformatics. This field is defined as the use of computers in solving information problems in the life sciences, mainly involving the creation of extensive electronic databases on protein sequences and genomes. Dr. Dayhoff was the first woman in the field of Bioinformatics.”

<http://www.dayhoff.cc/>

- deduce evolutionary relationships of the biological kingdoms, phyla, and other taxa from sequence evidence
- collection of all known protein sequences
- made available to others in 1965 in a small book
- contained sequence information of 65 proteins
- several releases followed
- resulted in the Protein Information Resource (PIR)





Protein sequences



Fachhochschule Bingen
University of Applied Sciences

>O54090|O54090_SULAC Hypothetical protein (Fragment).
MKILDYSDLVFFRKLTKMRDPKTRFDVREFINRGEDYLFNYTNKNVGGVDERRRKFLKS
LIFGMAA

>P70723|P70723_ACIAM Orf-2 (Fragment).
MSKNSLDNLGKALELLKKYPLCDSCLGRCFAKLGYRFANKERGKAIKTYLVLELDRKIK
DHELEDLNEIKEILFNMGKEYLEYLIYLSNEKFQERT

>Q9V2V9|Q9V2V9_PYRAE Rieske iron sulfur protein (ParR).
MV DENRRNTLKI FLGTTAALGAGMLATPLVASVIGSKAGYIKPEPSGAIPVEICKDVDSC
PKDYGVSLDELNRNGPVFKLLKVNMAIPAVFGIVRAKDKEYPVAVVAICTHFGCPVNVS
GGKYLIGFNCPCHGSIFAICNDPNGCPDYNAAFLEMYVSGGPAPRSLRAIKVAVKDGVVY
PLVAYI

>O93973|O93973_MALSM Allergen.
MSNVIKKVFNTDKAEAEAGSKVADAPQEAGHKGEGFLHDAKDRLQGFAGHGHNAQNAASG
VAGSAGAGGAPSVPSANVDVTNPVNDASVQGGVEAPRSWSTQLPQSQSVADTTGATSAGR
NNLTQTTSTGSGVNVAAGNVDQDVQHLAPVTRHVHHRHEIEELLREREHHIHQHHIQHHV
QPVDSEHLAEQIHSRVVPQTTVREHANTDKDAALMRAVAGNPKDTFTQAAIDRSVIDK
GETVREIVHHHIHNIVQPIIEKETHEYHRIRTTIPTTHITHEAPIVHESTAHQPIRKEDF
LKG GGLTSTTRSIEEVGLLN LGNNQRTVEGETYTGGLPLSQ

>Q02039|Q02039_RHYSE NIP1 precursor (NIP1 avirulence protein precursor).
MKFLVLPLSLAFLQIGLVFSTPDRCRYTLCCDGALKAVSACLHESESCLVPGDCCRGKSR
LTLCSYGEGNGGFQCPTGYRQC

>Q873M4|Q873M4_MALSM Manganese superoxide dismutase (Fragment).
PFYPIPSALPFPLPIHSLFSRRTLRFRSRTAARAGTEHTLPPLPYEYNALEPFISADIM
MVHHGKHHQTYVNNLNASTKAYNDAVQAQDVLKQMELLTAVKFNGGGHVNHALFWKTMAP
QSQGGGQLNDGPLKQ AIDKEFGDFEFKAAFTAKALGIQSGWCWLGLSKTGSLLDLVVAK
DQDTLTTHHP IIGWDGWEHAWYLQYKNDKASYLKQWNVVNWSEAESRYSEGLKASL

>Q2V2P9|Q2V2P9_YEAST Protein YDR119W-A.
MFFSQVLRSSARAAPIKRYTGGRIGESWVITEGRRLIPEIFQWSAVLSVCLGWPGAVYFF
SKARKA

MVDENRRNTLKIFLGTTAALGAGMLATPLVASVIGSKAGYIKPEPSGAIIPVEICKDVDSC
PKDYGVSLDELNRNGPVFKLLKVNTMAIPAVFGIVRAKDGKEYPVAYVAICTHFGCPVNS
GGKYLIGFNCPCCHGSIFAICNDPNGCPDYNAAFLEMYVSGGPAPRSLRAIKVAVKDGVVY
PLVAYI



Diseases?



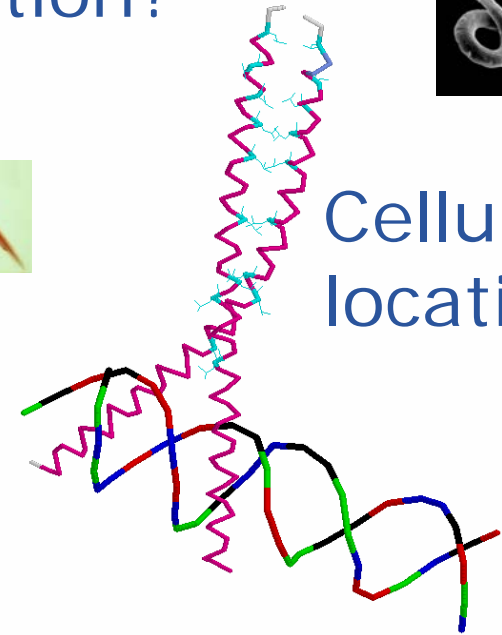
Development?



Function?

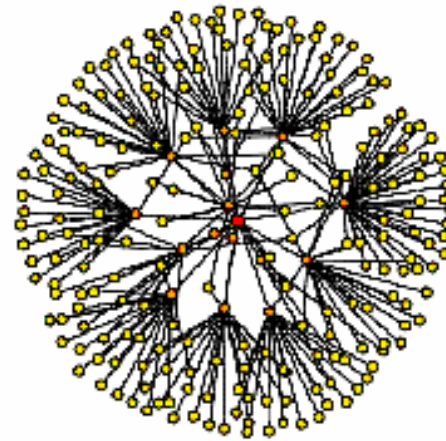


Cellular location?

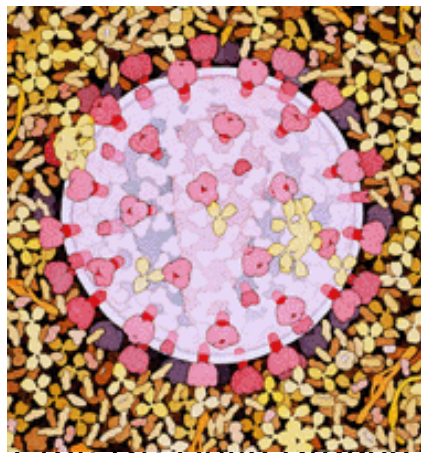


Regulation?

Structure?



Evolutionary history?



Interactions?



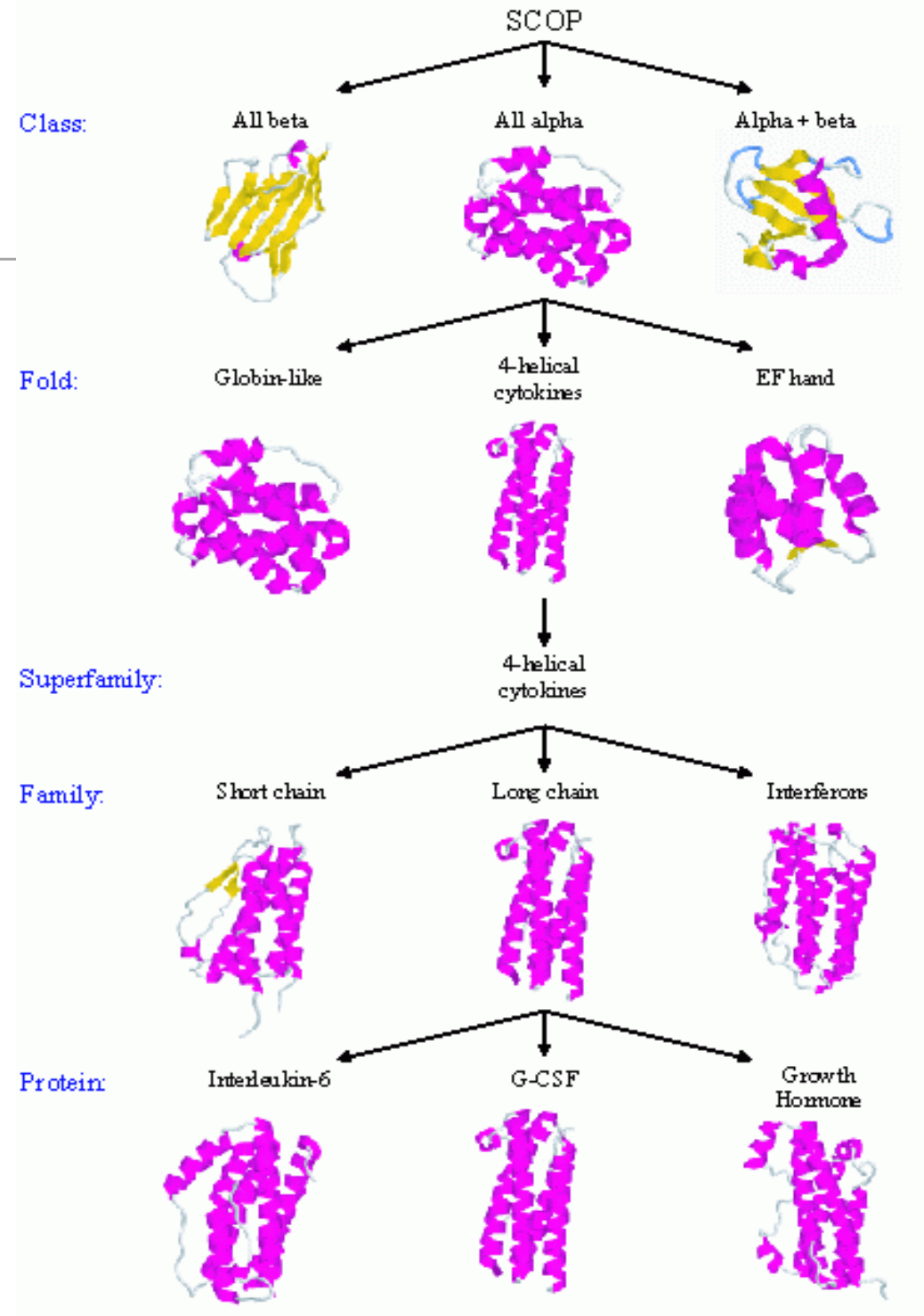
Tissue?



- Prediction of protein structure is still not possible from sequence alone
- Not all mechanisms of protein folding are known
- Experimental protein structure determination
 - is time consuming
 - is very expensive
 - is not always possible (protein must be in crystal structure)
 - results in only one conformation
 - does not show flexible regions
 - does not show the protein in its natural environment
 - can only be done with globular proteins (difficult with transmembrane proteins)

Structural Classification of Proteins (SCOP)

- **Fold:** not necessarily evolutionarily related; with same secondary structure elements and same topology
- **Superfamily:** maybe evolutionarily related; low sequence identity, but similar structure and function
- **Family:** evolutionarily related; mostly $> 30\%$ sequence identity



- Protein **sequence** databases:
 - Information about single proteins
- Protein **structure** databases:
 - Information about single proteins
- Protein **domain** databases:
 - Information about functional domains
- Protein (**sequence**) **family** databases:
 - Information about groups of evolutionarily and functionally related proteins
- Protein (**structure**) **family** databases:
 - Information about structural elements
- **Gene family** databases:
 - Information about groups of evolutionarily and functionally related proteins or genes mainly of completely sequenced species

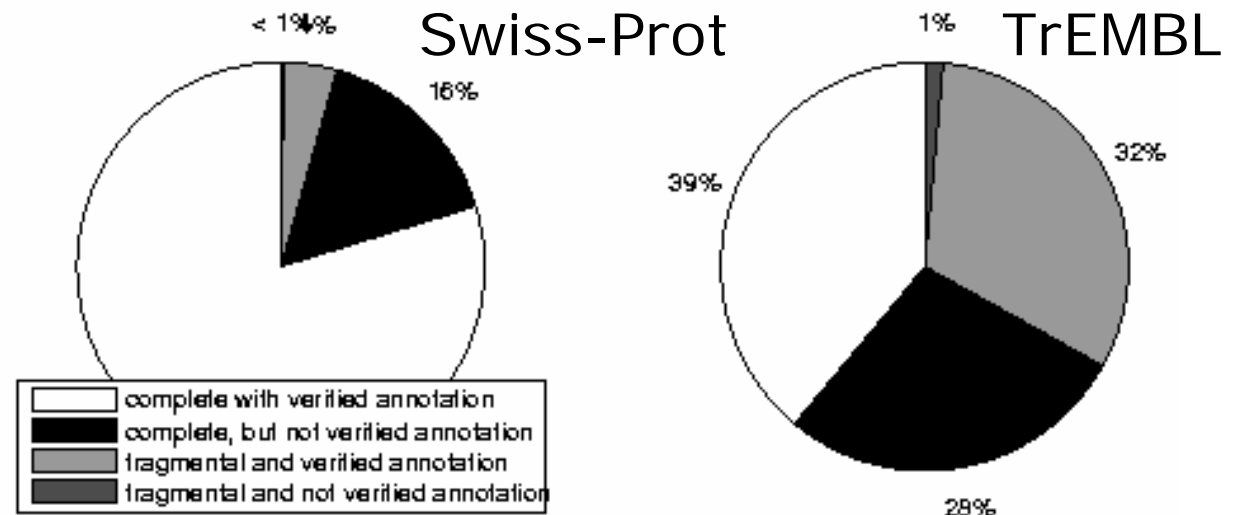
Protein sequence databases

- UniProt = Universal Protein Resource
- Integration of Swiss-Prot/TrEMBL and PIR
- <http://www.uniprot.org>
- central repository of protein sequence and function
- maintained by
 - European Bioinformatics Institute
 - Swiss Institute of Bioinformatics
 - Georgetown University



- contain experimentally verified entries ...
- ... and translated entries from DNA databases, namely EMBL

- predicted proteins
- hypothetical proteins
- putative proteins



- Problem in the past: no clear difference between experimentally verified entries/annotation and predicted entries/annotation

Protein sequence databases (Swiss-Prot/TrEMBL) ⇒ now UniProt!



Fachhochschule Bingen
University of Applied Sciences

ExPASy (<http://www.expasy.ch>)
Expert Protein Analysis System

SIB (<http://www.isb-sib.ch>)
Swiss Institute of Bioinformatics, Geneva, CH

Swiss-Prot (<http://www.expasy.ch/sprot>)
Manually curated protein sequence database

TrEMBL (translated EMBL)
Computer-annotated supplement to Swiss-Prot, contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot

General information about the UniProt/TrEMBL entry

Entry name	Q8TE89_HUMAN
Primary accession number	Q8TE89
Entered in TrEMBL	Release 21, 01-JUN-2002
Sequence was last modified	Release 21, 01-JUN-2002
Annotations were last modified	Release 26, 01-MAR-2004

Protein description

Protein name	Hypothetical protein FLJ23790
Evidences	

Origin of the protein

Gene	Gene name FLJ23790
From	Homo sapiens (Human)[TaxID:9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo.

References

[1]	NUCLEOTIDE SEQUENCE. Kawabata A., Hikiji T., Kobatake N., Inagaki H., Ikema Y., Submitted (FEB-2002) to the EMBL/GenBank/DDBJ databases.
-----	--

Cross-references

EMBL	AK074370; BAB85063.1; -; mRNA.[EMBL/ GenBank/ DDBJ] [CoDingSequence]
Ensembl	ENSG00000176853; Homo sapiens.[Entry/Contig]
InterPro	IPR001032; Leghaemoglobin. Graphical view of the domain structure
PROSITE	PS00208 PLANT_GLOBIN; UNKNOWN_1.



Plant hemoglobins signature

Description:

Leghemoglobins [1] are hemoproteins present in the root nodules of leguminous plants. Leghemoglobins are structurally and functionally related to hemoglobin and myoglobin. By providing oxygen to the bacteroids, they are essential for symbiotic nitrogen fixation.

Structurally related hemoglobins are found in nonsymbiotic plants where they may not function as an oxygen storage or transport proteins, but might act as an oxygen sensors [2].

We have developed a signature pattern that exclusively picks up the sequence of plants hemoglobins. It is centered on an histidine that acts as the heme iron distal ligand.

Last update:

December 2001 / Pattern and text revised.

Technical section:

PROSITE method (with tools and information) covered by this documentation:

General information about the entry

Entry name	PLANT_GLOBIN
Accession number	PS00208
Entry type	PATTERN
Date	APR-1990 (CREATED); DEC-2001 (DATA UPDATE); DEC-2004 (INFO UPDATE).
PROSITE documentation	PDOC00183

Name and characterization of the entry

Description	Plant hemoglobins signature
Pattern	[SN]-P-x-[LV]-x(2)-H-A-x(3)-F.

Numerical results

- UniProtKB/Swiss-Prot release number: **47.1**, total number of sequence entries in that release: **181821**.
- Total number of hits in UniProtKB/Swiss-Prot: **49 hits in 49 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **47 hits in 47 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **2 hits in 2 different sequences**
- Number of known missed hits: **1**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **0**
- Precision (true hits / (true hits + false positives)): **95.92 %**
- Recall (true hits / (true hits + false negatives)): **97.92 %**

Comments

- Taxonomic range: **Eukaryotes**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- 'Interesting' site in the pattern: **6,heme_iron**
- VERSION: **1**

Alignment of matches to PROSITE PS...

File Edit View Go Bookmarks Tools Help

TFHM WEB.DE SP ON SPIEGEL heise

HBL0_PHYPA/78-89	NPkVknHAryvF
HBL1_ARATH/63-74	NPkLkpHAmvF
HBL1_ORYSA/68-79	NPkLktHAmvF
HBL2_ARATH/60-71	NPkLkaHAvkvF
HBL2_BRANA/60-71	NPkLkaHAvkvF
HBL2_GOSHI/57-68	NPkLkaHAvkvF
HBL2_LYCES/57-68	NPkLraHAvkvF
HBL2_ORYSA/71-82	NPkLktHAmvF
HBL3_ORYSA/70-81	NPkLkiHAmavF
HBL4_ORYSA/67-78	NPnLkkHAmavF
HBL_HORVU/64-75	NPkLktHAvsvF
HBL_MAIZE/67-78	NPkLktHAmvF
HBL_TRETO/64-75	NPkLkpHAmvF
HBL_ZEAMP/67-78	NPkLktHAmvF
HBP1_CASGL/56-67	NPkLkaHAavvF
HBP2_CASGL/63-74	NPkLksHAmvF
HBPA_CASGL/56-67	NPkLkaHAavvF
HBPL_PARAD/63-74	NPkLkpHAttvF
HBPL_PARRI/63-74	NPkLkpHAttvF
LGB1_LUPLU/57-68	NPdLqaHAgkvF
LGB1_MEDSA/56-67	SPqLqaHAekvF
LGB1_MEDTR/56-67	SPqLqaHAekvF
LGB1_PEA/54-65	SPkLqaHAeqvF
LGB1_SOYBN/55-66	NPkLtgHAeklf
LGB1_VICFA/55-66	SPkLqaHAekvF
LGB1_VIGUN/56-67	NPkLtgHAeklf
LGB2_LUPLU/57-68	NPeLqaHAgkvF
LGB2_MEDTR/55-66	SPkLqsHAekvF
LGB2_PEA/55-66	SPkLqaHAeqvF
LGB2_PHAVU/55-66	NPkLtaHAeslf
LGB2_SESRO/56-67	NPpLqaHAekvF
LGB2_SOYBN/55-66	NPkLtgHAeklf
LGB2_VIGUN/56-67	NPkLtgHAeklf
LGB3_MEDSA/55-66	SPkLqsHAekvF
LGB3_PEA/54-65	SPkLqaHAekvF
LGB3_SESRO/56-67	NPpLqaHAekvF

Done

Mozilla Firefox

```
>UniProt/TrEMBL|Q8TE89|Q8TE89_HUMAN Hypothetical protein FLJ23790
MNRVQGDYFETLLYKIFVSIIDEHTNVAELANVLEIDLSLVKNAVSMYCR LGFAHKKGQVI
NLDQLHSSWKNVPSVNRLKSTLDPQKMLLSWDGGESRSPVOEASSATD TD TNSQEDPADT
ASVSSLSLSTGHTRKRI AFLFDSTLTAFLMMGNI SPNLKSHAVTMF EVGKLSDESLSFLI
ELEKVQSTGEGEAQRYFDHALTLRNTILFLRHMKDLVAQTAQPDQPNYGFPLDLLRCESL
LGLDPATCSRVLNKNYLLVSMAPLTNEIRPVSSCTPDHIGPAIPEVSSVWFKLYIYHVT
GQGPPSLLLSKGT RLKLPDIFQSYDRLLITSWGHDP EVVPTSNVLTMLNDALTHSAVLI
QGHGLHGIGETVHVPPFDETELQGEFTRVNMGVHKAQILRN RVDLQHL CGYVTMLNAS
SQLADRKLS DASDERGEPDLASGSDVNGSTESFEMVIEEATIDSATKQTSGATTEADWVP
LELCFGIPLFSSSELNRKVKR KIAAHGLCRKESLQNL LHSSRKL SLQVLNFVHSFQEGASI
LDIHTEPSFSSLLSQSSYADMGVPLPAKNLIFKDGVLSEW SGRSPSSLLIANLHLQ
```

Done

Pattern search

Pattern construction

[SN]-P-x-[LV]-x(2)-H-A-x(3)-F.

Multiple Sequence Alignment

Mozilla Firefox

```
>UniProt/Swiss-Prot|P02237|LGB3_SOYBN Leghemoglobin C3
GAFTDKQEALVSSSFEAFKTNIPQYSVVFYTSILEKAPVAKDLFSFLANGVDP TNPKLTG
HAEKLFGLVRDSAGQLKASGTVVIDAALGSIHAQKAITDPQFVVVKEALLKTIREAVGDK
WDELSSAWEVAYDELA AAIKKAF
```

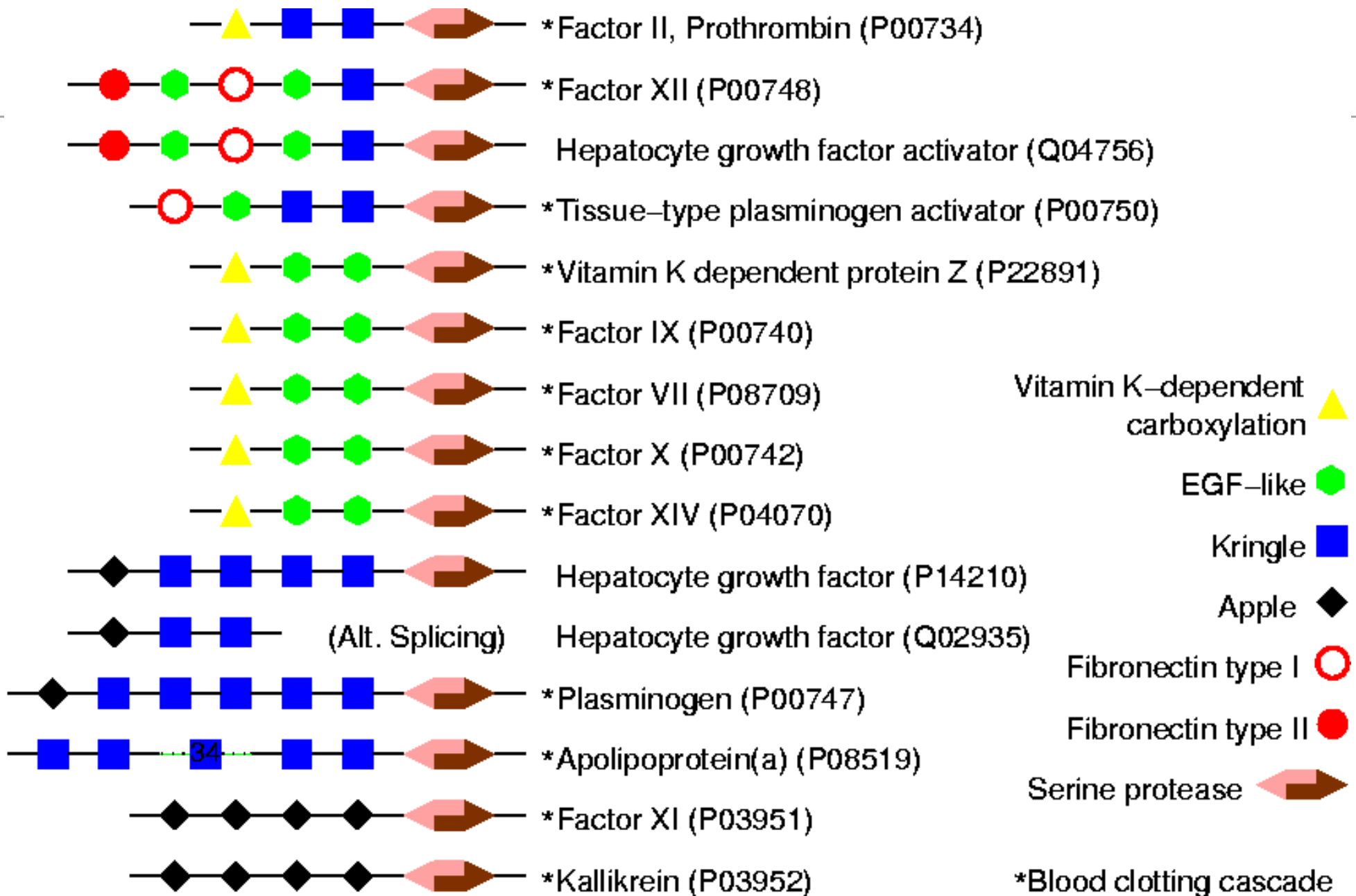
Mozilla Firefox

```
>UniProt/Swiss-Prot|P02239|LGB1_LUPLU Leghemoglobin I
GVLTDVOVALVKSSFEFNANIPKNTHRFFTLVLEIAPGAKDLFSFLKGSSEVPQNPDL
QAHA GKVF KLTYEAAIQLVNGAVASDATLKSLGSVHVS KGVVDAHFVVK EAILKTIKE
VVGDKWSEELNTAWTIAYDELA IAIKKEMKDA A
```

“One or two homologous sequences whisper... a full multiple sequence alignment shouts out loud” (Arthur Lesk, 1996)

```
/NMIMAESPLITICLLGYLLSAECTVFLDHENANKILNRPKRYNSGKLEEFVQGNLERECMEEKCSFEEAREVFENTERTTEFWKQYVDGDQCESNPCLNGGSCKDDINSYECWCPFGFI
VNMIMAESPLITICLL-GYLLSAE--CT-----VFLDHENANKILNRPKRYNSGKLEEFVQGNLERECMEEKCSFEEAREVFENTERTTEFWKQYVDGDQCESNPCLNGGSCK
QALRLLCLLLGLQGCLAAGGVAKASGGETRDMPWKPGPHRVFVTQEEAHGVLHRRRRAN-AFLEELRPGSLERECKEEQCSFEEAREIFKDAERTKLFWISYSDGDQCASSPCQNGGSCK
-----MGSPVQLSLLCVVLASLLLP GK-----GVFINRERANNVLRTRRANS-FFEEFKKGNLERECMEEICSYEEVREIFEDDEKTKEYWTKYKDGDCQCES:
-----MGRPLHLVLLSASLAGLLLLGE-----SLFIRREQANNILARVTRANS-FLEEMKKGHLEECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDCQDET:
-----MAPQLLLCLILTFWLSLSEAES-----NVFLKSKVANRFLQRTKRANS-LFEEFKAGNIERECIEERCSKEEAREAFEDNEKTETFWNVYVDGDQCSSI
-----MAEASGLVTVC-LLGYLLSAEC-----AVFLDRENATKILSRPKRYNSGKLEEFVVRGNLERECIEEKCSFEEAREVFENTEKTTEFWKQYVDGDQCESI
RCLNMIMAEPPGLITIC-LLGYLLGADC-----TVFLDHEDATKVL SRPKRYNSGKLEEFVQGNLERECMEEKCSFEEAREVFENTEKTTEFWKQYVDGDQCESI
RVNMIMAESPLITIC-LLGYLLSAEC-----TVFLDHENANKILNRPKRYNSGKLEEFVQGNLERECMEEKCSFEEAREVFENTERTTEFWKQYVDGDQCESI
-----ALITIF-LLGYLLSTEC-----AVFLDRENATKILTRPKRYNSGKLEEFVVRGNLERECIEERCSFEEAREVFENTEKTTEFWKQYVDGDQCESI
-----MVSQALRL--LCLLLGLQGCLAAGGVAKASGGETRDMPWKPGPHRVFVTQEEAHGVLHRRRRANA-FLEELRPGSLERECKEEQCSFEEAREIFKDAERTKLFWISYSDGDQCAS:
-----MAPQARGLG-LCSLLALQASLAA-----VFITQEEAHSVLRQRANS-FLEELRPGSLERECKEELCSFEEAREVFQSTERTKQFWITYNDGDQCASI
-----MVPQAHGLLLLCFLLQLQGPLGT-----AVFITQEEAHGVLHRQRANS-LLEELWPGSLERECNEEQCSFEEAREIFKSPERTKQFWIVYSYSDGDQCASI
-----MWQLTSLLLFVATWGISGTPAP-----LDSVFSSSERAHQVLRIRKRANS-FLEELRHSSLERECIEEICDFEEAKEIFQNVDDTLAFWSKHVDGDQCLVI
-----MWQLASLSLLL TICGTCSTAAP-----PGSVFSSSES AHQVLRIRKRANS-FLEELRAGSLERECMEEICDFEEAKEIFQNVDDTLAYWSKYVDGDQCAAI
-----MWQFRVFLLLMSTWGISSIPAH-----PDPVFSSSEHAHQVLRVR-RANS-FLEELRPGSLERECMEEICDFEEAQEIFQNVEDTLAFWIKYFDGDQCSAI
-----MAGCVPLLQGLVVLALHRVEP-----SVFLPASKANDVLRWRKRAGSYLLEELFEGNLEKECYEEICVYEEAREVFEVVTDEFWRRYKGGSPCIS(
-----MAGCILLRGRF ILTLILHQVEL-----SVFLPAPKANNVLRWRRRGSSYFLEEIQGNLEKECYEEVCONYEEAREVFEVNDVITDEFWRQYGGGSPCVS(
-----AGSYLLEELFEGHLEKECWEEICVYEEAREVFEDETTDEFWRTYMGGSPCAS(
```

- describe functional regions of proteins (called domains, motifs, signatures...)
- a protein may consist of several and/or different domains (multi-domain-protein)
- domains can be described by
 - patterns (regular expressions)
 - rules
 - profiles
 - Hidden Markov Models



Suppose we have n homologous protein sequences

- What do they have in common?
- What are the functional regions of these proteins?
- Which regions are conserved, which are not conserved?
- How can we characterize these proteins/their functional domains?
- What distinguishes these proteins/their functional domains from others?

Sequences producing significant alignments:			Score	E
			(bits)	Value
070333	POSTSYNAPTIC PROTEIN CRIPT.	[Rattus norvegicus]	175	3e-43
Q9SYB4	T13M11.14 PROTEIN.	[Arabidopsis thaliana]	100	1e-20
Q9VL77	CG4537 PROTEIN.	[Drosophila melanogaster]	98	4e-20
Q93338	C36B1.3 PROTEIN.	[Caenorhabditis elegans]	59	4e-08
087978	PUTATIVE FORMYL TRANSFERASE.	[Bordetella bronchiseptica]	34	0.71
P16374	ZINC FINGER PROTEIN MFG-3.	[Mus musculus]	31	6.1
Q9TXU2	H05C05.1 PROTEIN.	[Caenorhabditis elegans]	31	8.0

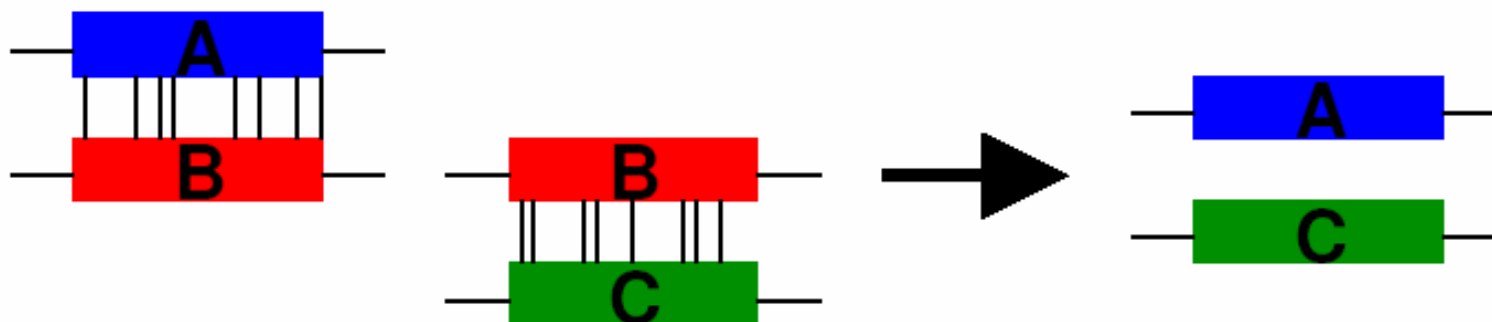
>070333 POSTSYNAPTIC PROTEIN CRIPT. [Rattus norvegicus] Length = 101

Score = 175 bits (438), Expect = 3e-43
Identities = 80/89 (89%), Positives = 81/89 (90%)

```
Query: 26  MVCEKCEKKLGRVITPDTWKDGARNTTESGGRKLNENKALTSKKARFDPYGKNKFSTCRI 85
          MVCEKCEKKLGRVITPDTWKDGARNTTESGGRKLNENKALTSKKARFDPYGKNKFSTCRI
Sbjct: 1   MVCEKCEKKLGRVITPDTWKDGARNTTESGGRKLNENKALTSKKARFDPYGKNKFSTCRI 60
```

```
Query: 86  CKSSVHQPGSHYCQGCATKKASVRCVAKR 114
          CKSSVHQPGSHYCQGCA KK      K+
Sbjct: 61  CKSSVHQPGSHYCQGCAYKKGICAMCGKK 89
```

- use of intermediate sequences to derive knowledge about homology
- if the proteins A and B are homologous and the proteins B and C are homologous, than A and C are homologous, too
- this holds even if there is no sequence similarity detectable between A and C!



Biologically meaningful partitioning of the data:

- Functional annotation
- Gain of information
- Reduction of the search space
- Selection of prototypic or representative sequences
- Phylogenetic analyses
- Gene/protein prediction

etc.

Following M.Dayhoff (1974, refined by W.Barker, 1996)
we can think of a

- **Protein superfamily** as a group of proteins
 - sharing domains
 - being evolutionarily related
 - showing weak sequence similarity
- **Protein family** as a group of proteins
 - being (closely) evolutionarily related
 - (showing at least 50% sequence similarity)
- **Homeomorphic protein family** as a group of proteins
 - having the same domains in the same order

How to “find” protein families and superfamilies with protein sequences?



1. Do sequence similarity searches (e.g. using BLAST) of all sequences in a data set against all sequences in the data set
⇒ *all-against-all comparison*
2. Extract similarity values (score, % identity, Evalue...) and store them in a table ⇒ *distance matrix*
3. Sort sequences in groups according to their similarity values
⇒ *clustering*

Problems:

- Which similarity values are suitable to do the grouping?
- Down (or up) to which value can we reliably detect evolutionary relationships?

- Goal: grouping of elements into disjoint sets, named **clusters**
 - with similar elements in the same cluster (**homogeneity**) and
 - dissimilar elements in different clusters (**seperation**)
- Requires: distance matrix / similarity measure
- Let E be a set of n elements $E_i, i \in \{1, \dots, n\}$
- Let $S: E \times E \rightarrow \mathbb{R}$ be a similarity measure
- Let t be a **threshold** value (or **cutoff** value)
- E_i and E_j are **similar**, if $S(E_i, E_j) > t$

- Nearest Neighbor Clustering
- Sokal & Sneath, 1973
- Two elements E_i and E_j belong to the same cluster, if there is a chain
 $S(E_i, E_k) > t, S(E_k, E_l) > t, \dots, S(E_q, E_r) > t, S(E_r, E_j) > t$
- Similarity graph = connected components
- Problems:
 - Chaining
 - How to choose t ?

- Furthest Neighbor Clustering
- Sokal & Sneath, 1973
- Two elements E_i and E_j belong to the same cluster, if $S(E_i, E_j) > t$
- Similarity graph = maximal complete subgraphs (cliques)
- Problems:
 - in most cases too restrictive
 - finding cliques is NP-hard
 - How to choose t ?

- construct a tree
- cut at a certain level and derive clusters at this level
- clusters at different levels are not disjoint but nested
- **divisive construction** (top-down, splitting):
starts with one cluster which contains all elements and splits clusters in lower levels
- **agglomerative method** (bottom-up, clumping):
starts with n clusters, each containing one element and merges clusters in higher levels

Hierarchical Clustering (agglomerative method)



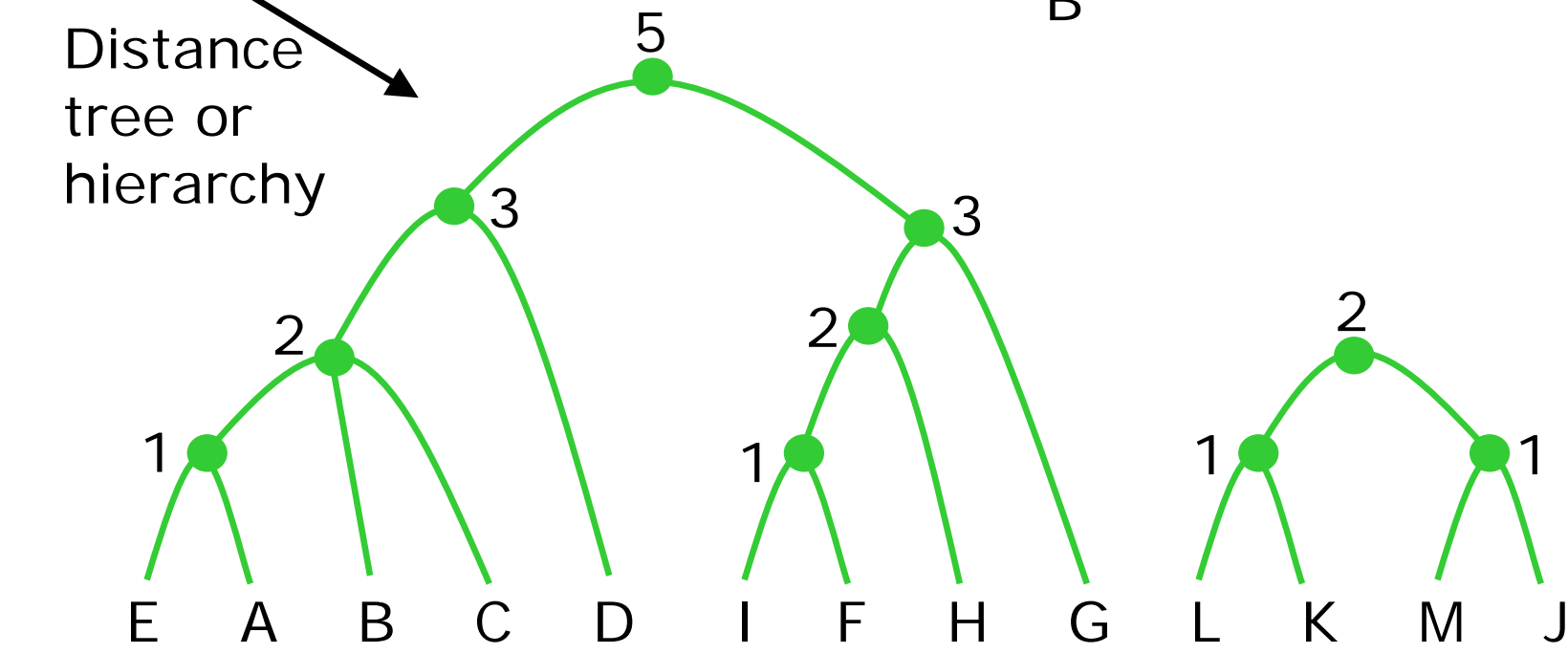
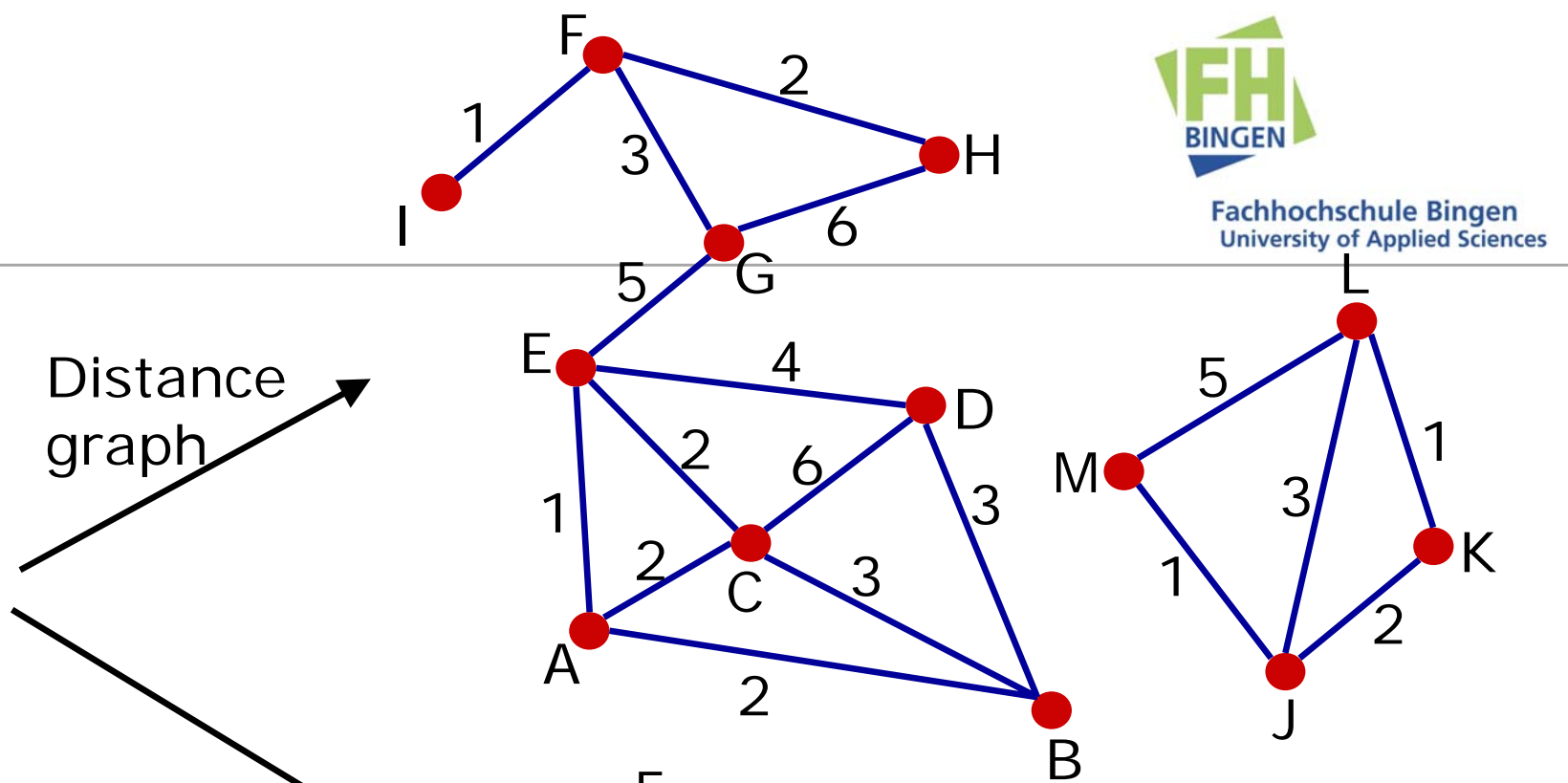
Input: n Elements $E_i, i \in \{1, \dots, n\}$

Output: Hierarchy

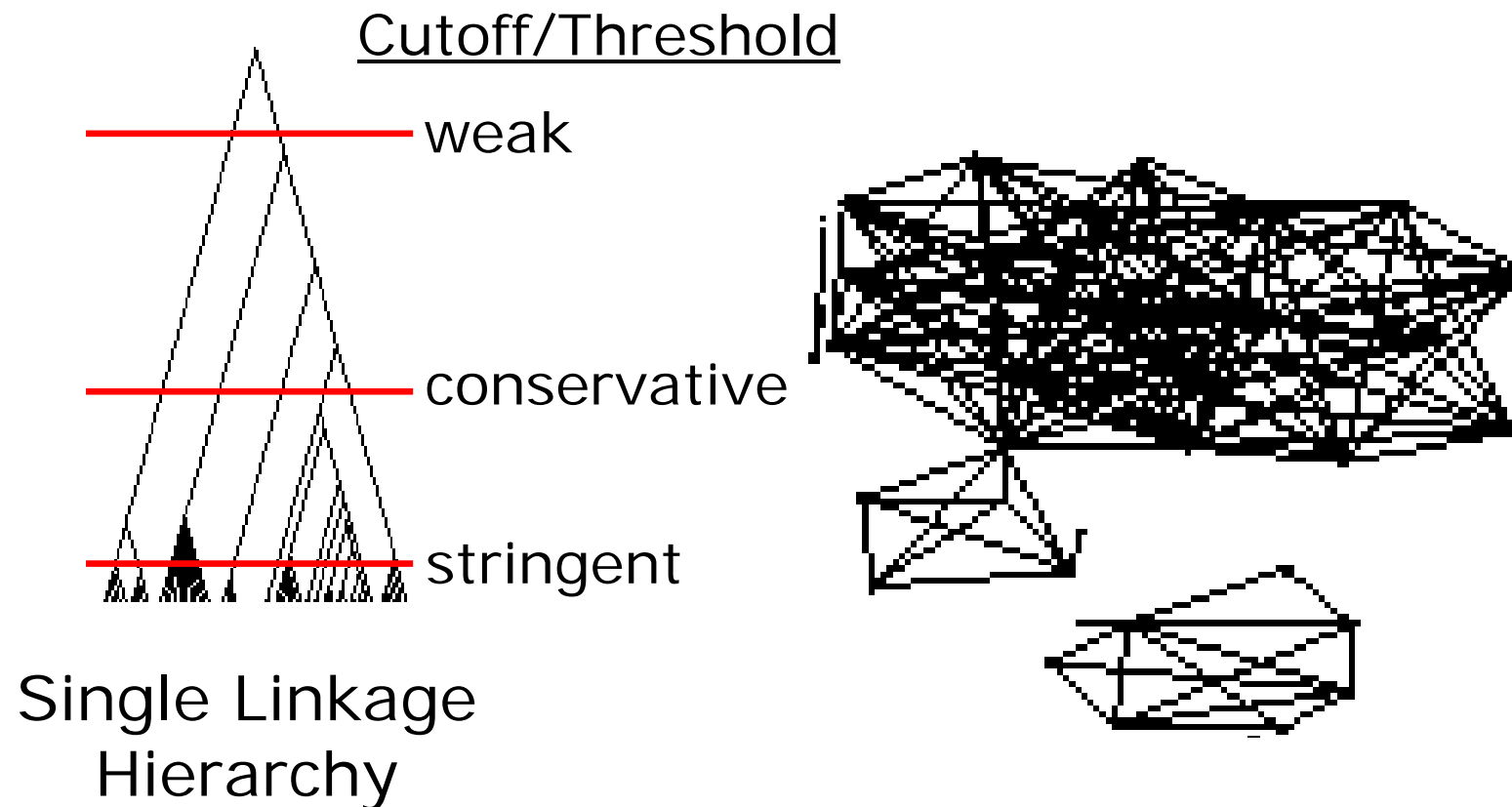
```
1:  $c \leftarrow n$ 
2: for all  $i \in \{1, \dots, n\}$  do
3:    $C_i \leftarrow \{E_i\}$ 
4: end for
5: while  $c > 1$  do
6:    $c \leftarrow c - 1$ 
7:   find next clusters  $C_i$  and  $C_j$ 
8:   merge  $C_i$  and  $C_j$ 
9: end while
```

in case of Single Linkage Clustering the result is a
Minimal Spanning Tree

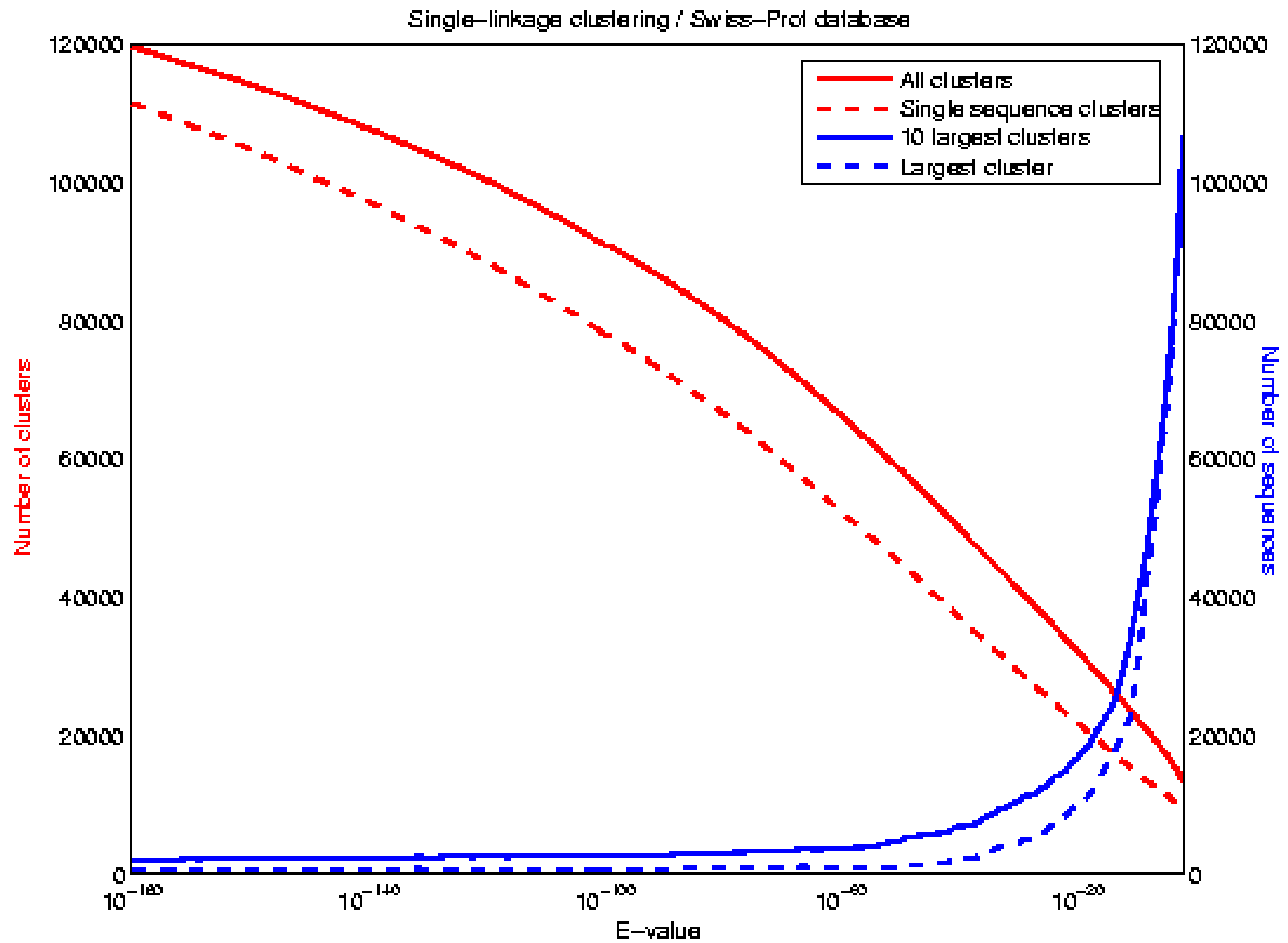
Distances		
I	F	1
L	K	1
M	J	1
E	A	1
F	H	2
K	J	2
E	C	2
A	B	2
A	C	2
B	C	3
F	G	3
L	J	3
D	B	3
E	D	4
M	L	5
E	G	5
G	H	6
C	D	6



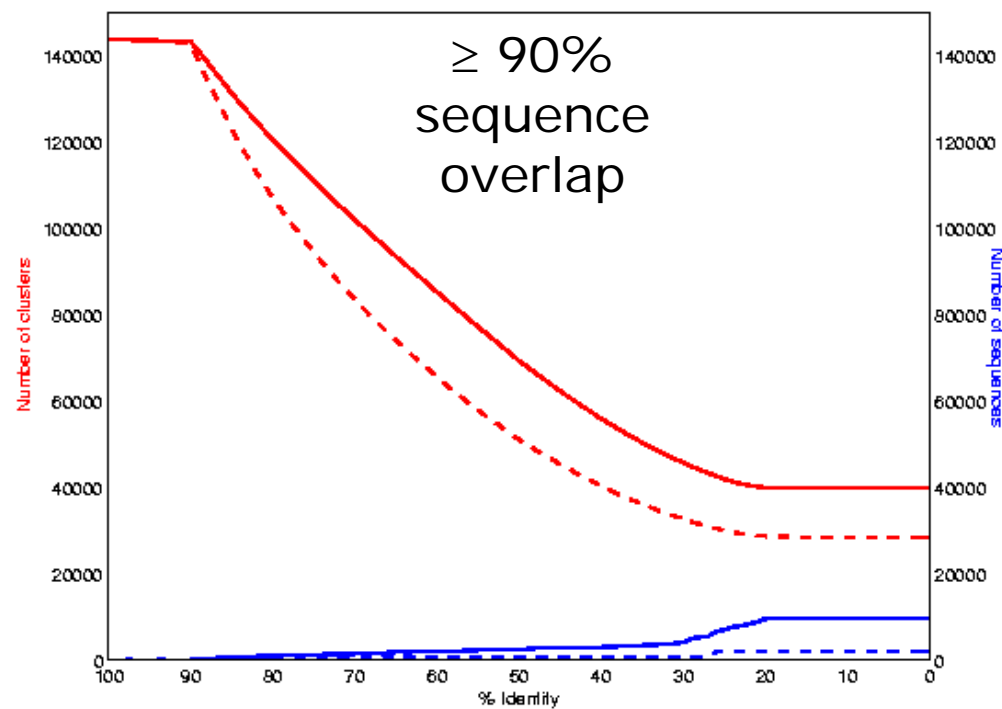
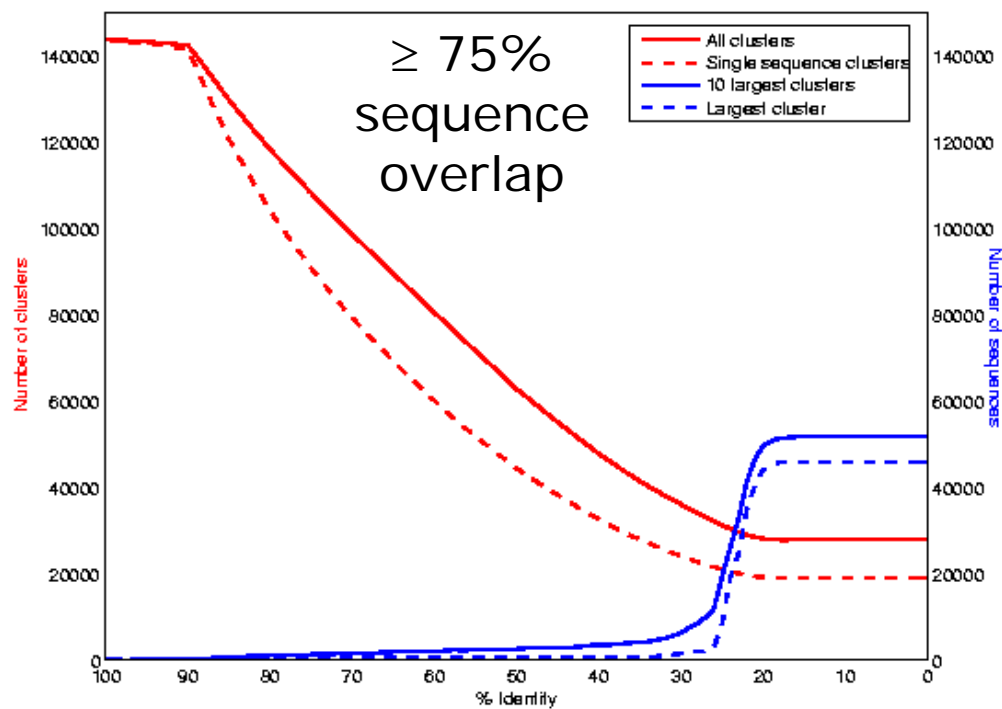
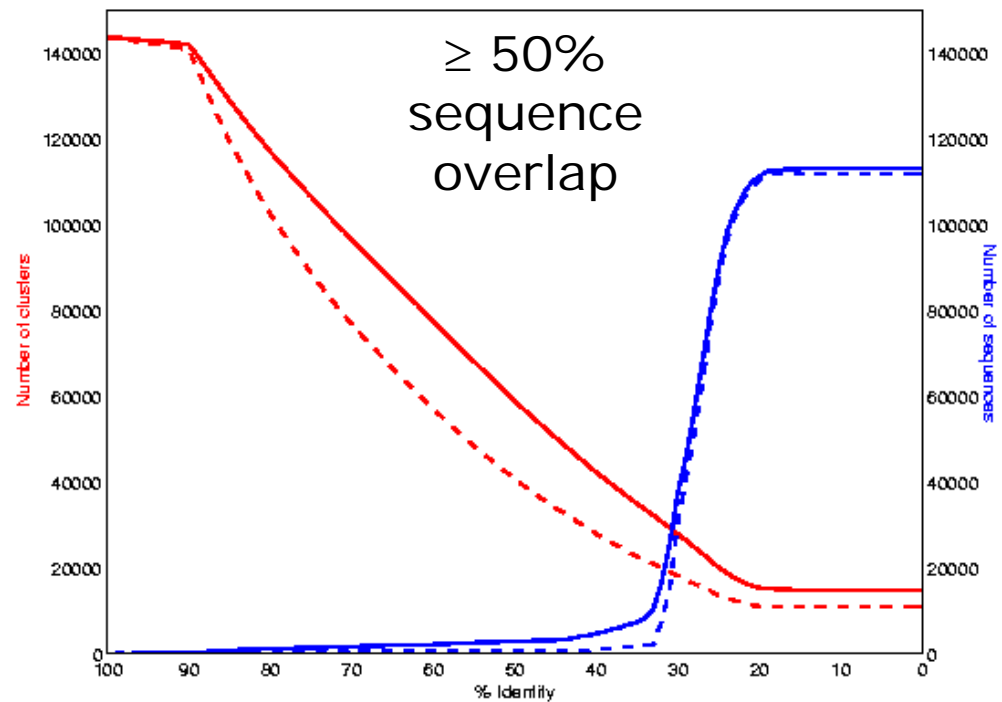
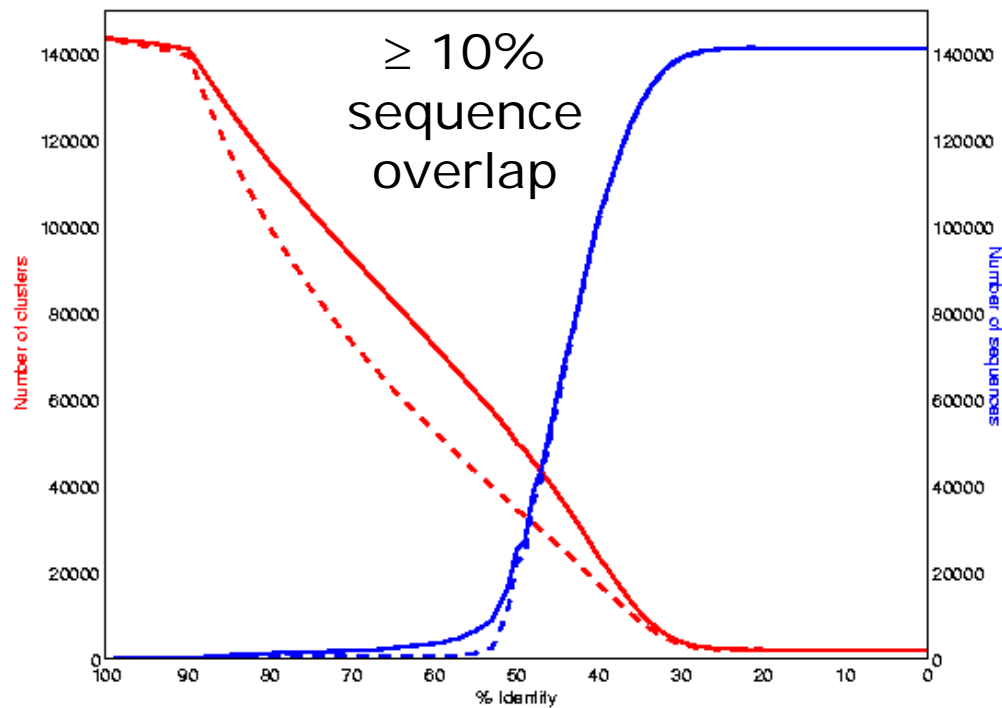
Single Linkage Clustering



- starting with 171,191 redundant sequences from Swiss-Prot
- after all-against-all BLAST database searches:
19,407,137 pairwise values
- after excluding 27,305 fragments (being 90% identical to another sequence over 95% of their sequence length):
13,083,209 pairwise values
- Reminder:
171,191 sequences
⇒ 14,653,093,645 possible pairwise values!
- only 0.132% sequence pairs result in an Evalue < 10!

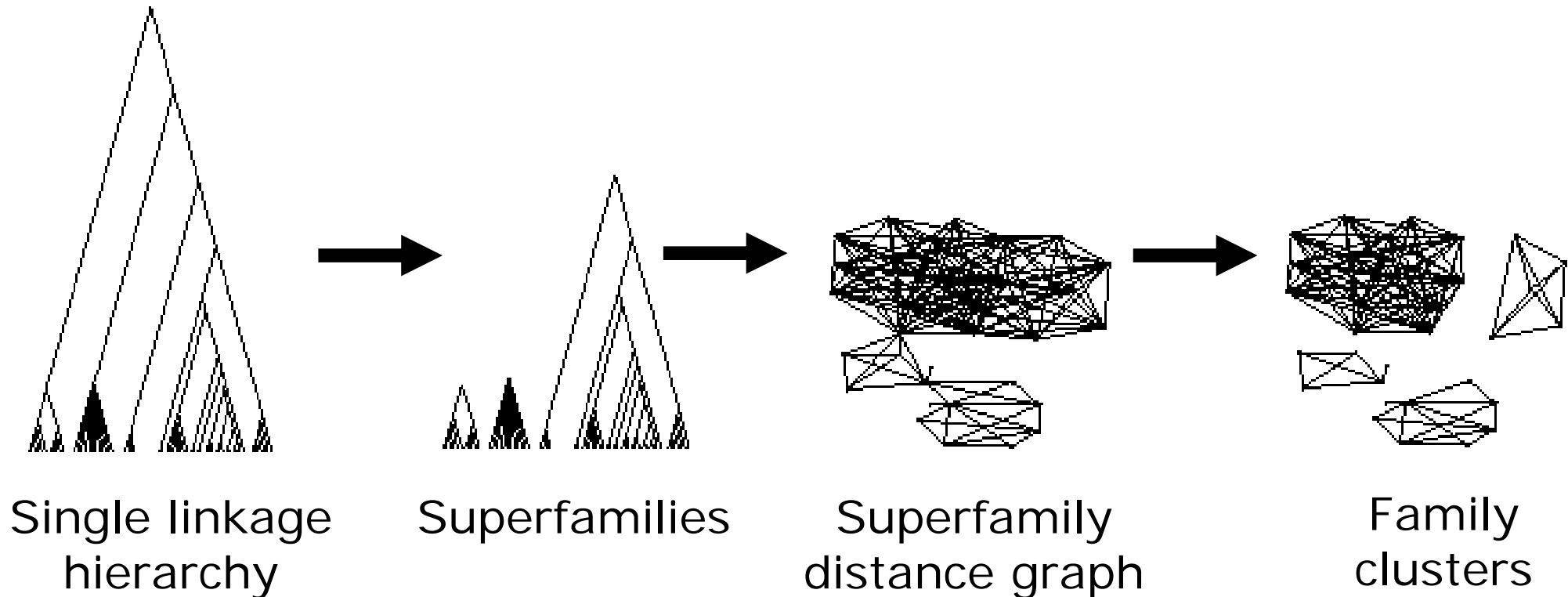


143,886 non-redundant sequences and 13,083,209 pairwise values



- Doing Single Linkage Clustering with this data we can vary on the pairwise results of the BLAST searches, i.e., Evalue, % Identity, length of local alignment, % alignment length of sequence length, score and all combinations!
- With a choice of at least 50% identity we are on the safe side (this was Margaret Dayhoff's original value for a protein family!)
- Unfortunately (but no surprise) nature does not behave in cutoffs ☹️
- There are highly conserved protein families (e.g., histones) and fast evolving protein families (e.g., immunoglobulines)
- Every protein family needs it's own cutoff 😊

SYSTEMS (SYSTEMatic Re-Searching)



Superfamilies as well as family clusters are derived from the structure generated by the data itself
⇒ no need for a user defined static cutoff

- Exploit the self-structuring properties of the data:
 - Determine an individual cutoff for each superfamily based on the single linkage hierarchy
 - Split each superfamily into family clusters based on the superfamily distance graph
- Automated and independent of static user-defined cutoffs
- Results accessible on the Internet
- <http://systems.molgen.mpg.de>

Protein family databases - ProtoNet



Fachhochschule Bingen
University of Applied Sciences

- <http://www.protonet.cs.huji.ac.il/>
- global classification of proteins into hierarchical clusters
- based on Swiss-Prot sequences, with TrEMBL sequences added after clustering
- N. Kaplan *et al.*, NAR, 2005, 33(DB)
- 3 different hierarchical clustering methods available depending on the similarity measure (harmonic-, geometric-, arithmetic average) based on BLAST Evalues

ProtoNet.cs.huji.ac.il
automatic hierarchical classification of proteins

Version 4.0

► Main Page
▷ Navigation Tools
▷ Classify Your Protein
▷ Introduction
▷ Methods
▷ Site Map
▷ Guided Tour
▷ Related Links
▷ ProtoNet Team
▷ Feedback

ProtoNet v4.0 provides global classification of the proteins into hierarchical clusters.

ProtoNet 4.0 features a major database upgrade and improved tools for analysis and visualization of the ProtoNet hierarchy.

ⓘ Please refer to our updated "Site Map", "Guided Tour", and "Introduction" to learn about the details of the new and modified features introduced in version 4.0.

ⓘ We will be most happy to get your comments and suggestions on the "Feedback" page.

- <http://www.ebi.ac.uk/clustr/index.html>
- automatic hierarchical classification of all sequences in UniProt
- uses Z-Score based on Smith-Waterman comparison:
 $Z\text{-Score} = \min(Z(A,B), Z(B,A))$ with
 $Z(A,B) = (\text{Score}(A,B) - M) / \sigma$ with
M: arithmetic mean,
 σ : stand. deviation of all results
- R. Petryszak *et al.*, *Bioinformatics*, 2005, 21(18)
- constructs Single Linkage Hierarchy
- provides a subset of clusters at several different cutoff values



The CluSTr database offers an automatic classification of [UniProt Knowledgebase](#) and [IPI](#) proteins into groups of related proteins. The clustering is based on analysis of all pairwise comparisons between protein sequences. The database provides links to [InterPro](#), which integrates information on protein families, domains and functional sites from [PROSITE](#), [PRINTS](#), [Pfam](#), [ProDom](#), [SMART](#), [TIGRFAMs](#), [Gene3D](#), [SUPERFAMILY](#), [PIR Superfamily](#) and [PANTHER](#).

About CluSTr

To date, CluSTr contains the following information:

- **4.8mln** protein sequences from [UniParc](#) (to be updated to **6.2mln** sequences in May 2006), including
- **2.1mln** sequences from [UniProt Knowledgebase rel. 8.1](#) and
- **155,000** sequences from [IPI](#);
- Over **600mln** similarities;
- **6.4mln** clusters;
- Clustering for **417** organisms with completely sequenced genomes. For the full list of the genomes see [Integr8](#)
- Putative homologues predictions for the above species. For more information see [Homologue Selection at Integr8](#).

- <http://cgg.ebi.ac.uk/services/tribes/>
- uses a Markov Clustering method based on BLAST Evalues
- primarily used for comparing protein sequence sets of completely sequenced genomes, e.g. in ENSEMBL
- clustering software available
- provides one set of protein families
- **more specific** than other methods, but **less sensitive**

	Related	Not related
Found	<i>True positive</i>	<i>False positive</i> ↓
Not found	<i>False negative</i> ↑	<i>True negative</i>



An efficient algorithm for large-scale detection of protein families

A.J.Enright *et al.*, NAR, 2002, 30(7)

relying on database search methods like BLAST for pairwise sequence comparisons:

- low ranking of short sequences in searching large sequence databases
- doing all comparisons is very time consuming
(Smith-Waterman on a GeneMatcher with 1.000.000 sequences took 3 months)
- sequence database searches report only local similarities, i.e., similarity between common domains
- sequence database searches with BLAST are not symmetric
(i.e., score and Evalue)

But wait a moment...



- Why so many databases?
- Which one is “right” which one is “wrong”?
- How can we proof that the results are correct?

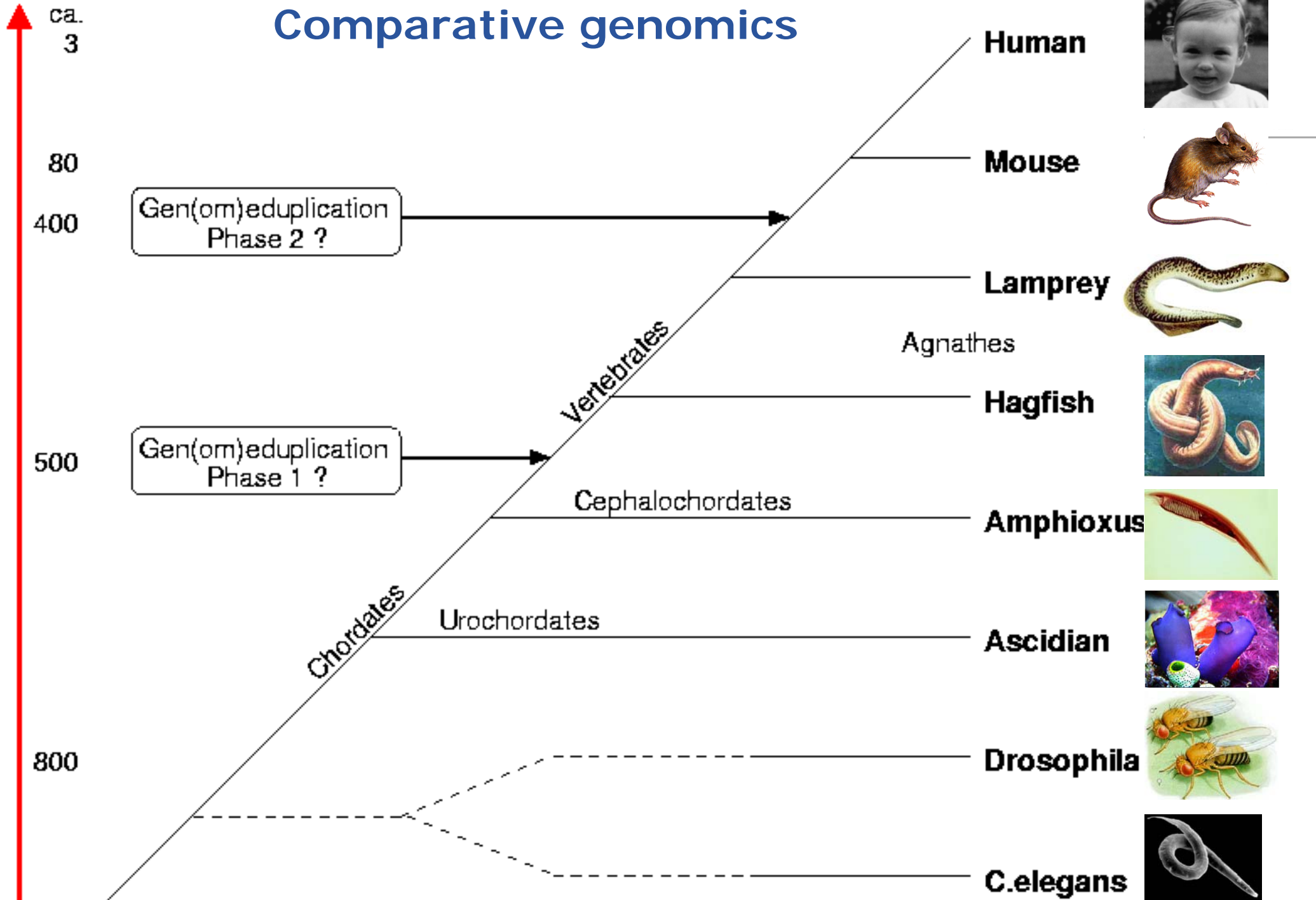
- We want to answer biological questions with these databases
- Different databases are needed to answer different questions
- There is no “right” or “wrong”
- The benefit highly depends on the questions
- The more concise the question, the more beneficial the answer

Suppose we have the gene/protein sequences of 2 completely sequenced species

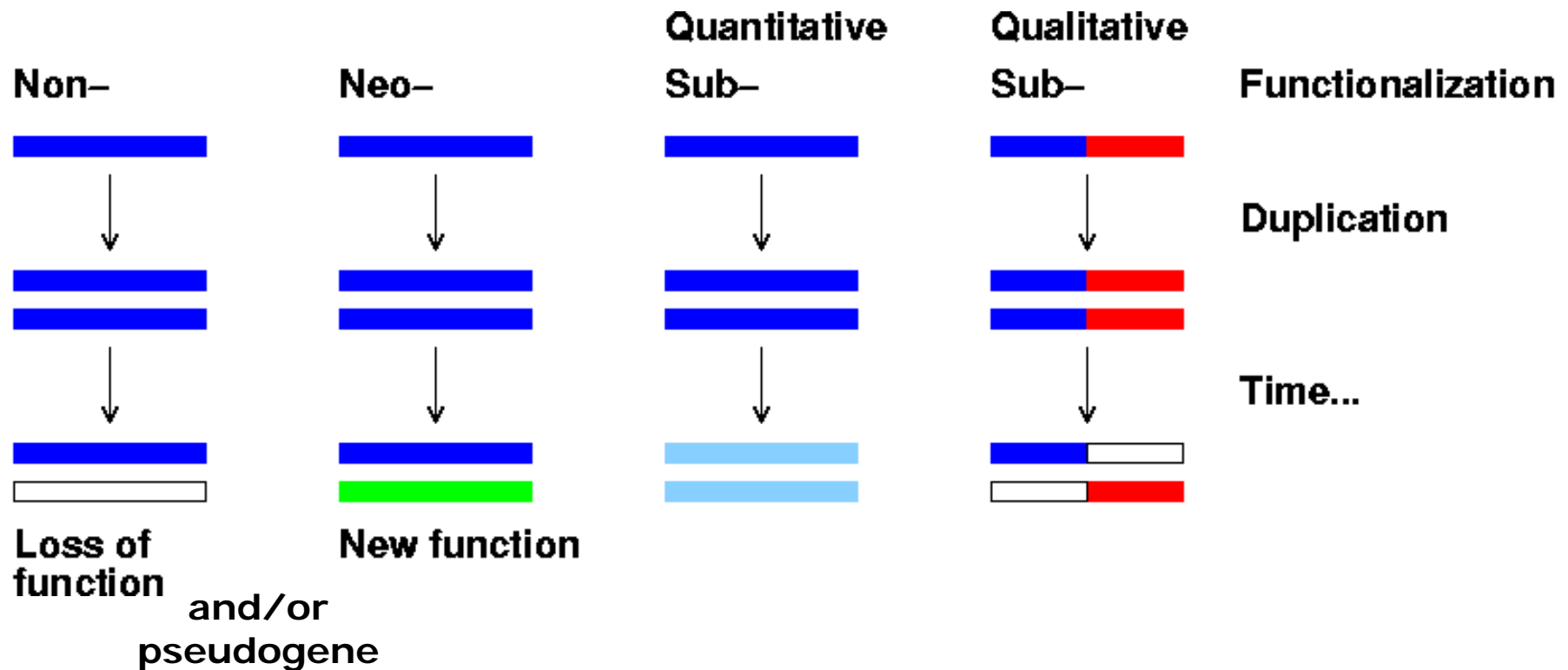
- Which genes/proteins do these species have in common?
- Which genes/proteins are orthologous?
- Where are the differences?
- Which genes/proteins have paralogs in one or the other species?
- Which metabolic pathways do these species have in common, which differ?

Million Years

Comparative genomics

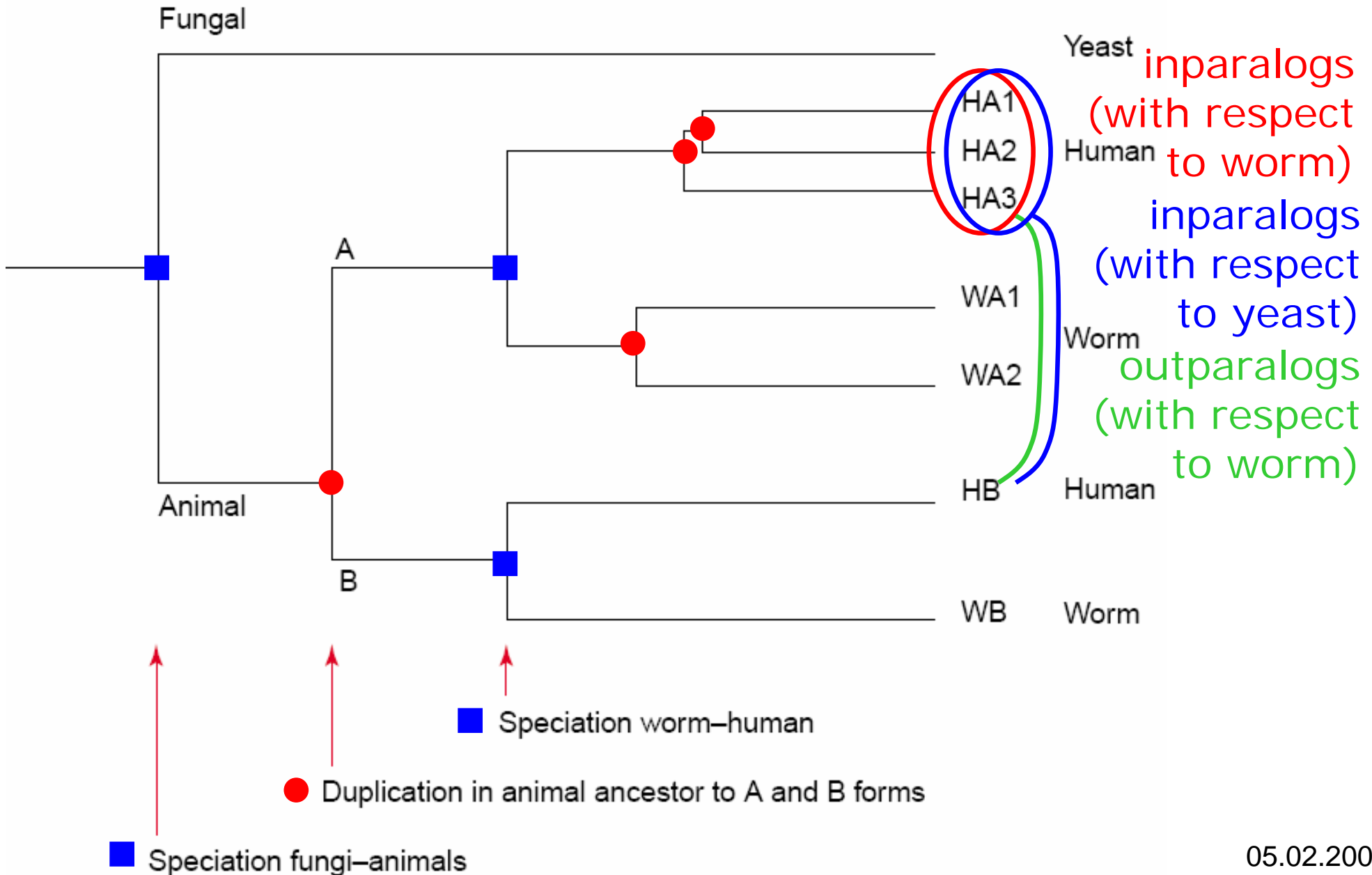


What happens to a duplicated gene?



Duplication-Degeneration-Complementation Model (DDC)
Lynch & Force (Genetics, 1999/2000)

Orthology, paralogy and proposed classification for paralog subtypes



Orthologs:

Genes in the genomes of different species with a common ancestor (resulting from a *speciation* event)

Paralogs:

Genes in the same genome with a common ancestor (resulting from a *duplication* event)

Inparalogs:

Genes in the same genome with a common ancestor (resulting from a *duplication* event **after** the last *speciation* event)

Outparalogs:

Genes in the same genome with a common ancestor (resulting from a *duplication* event **before** the last *speciation* event)

⇒ to determine orthologs and paralogs the easiest way is to analyse only 2 species at a time!

Pairwise-best-hit-method

1. Search with all protein sequences of species **A** against all protein sequences of species **B**

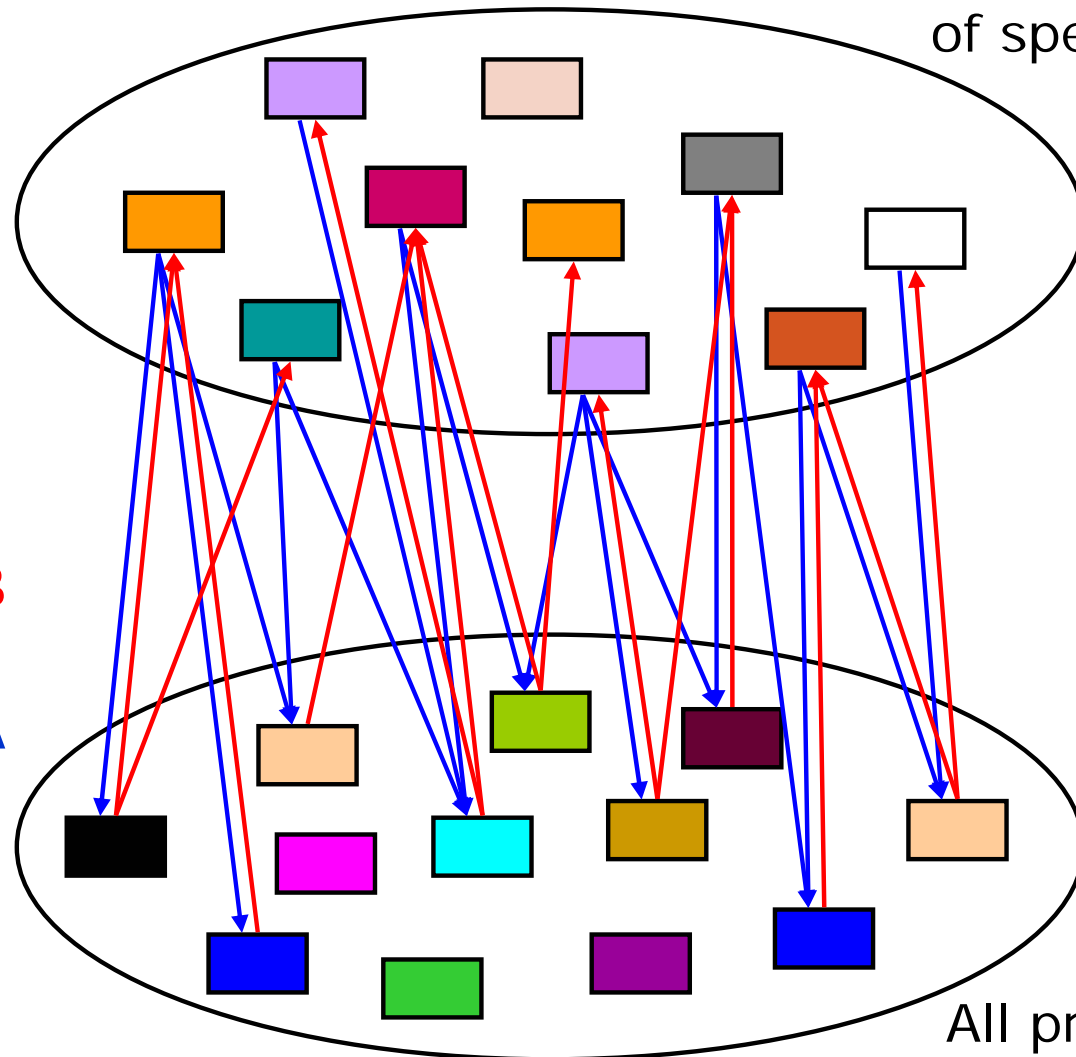
2. Remember only the best hits

3. Search with all protein sequences of species **B** against all protein sequences of species **A**

4. Remember only the best hits

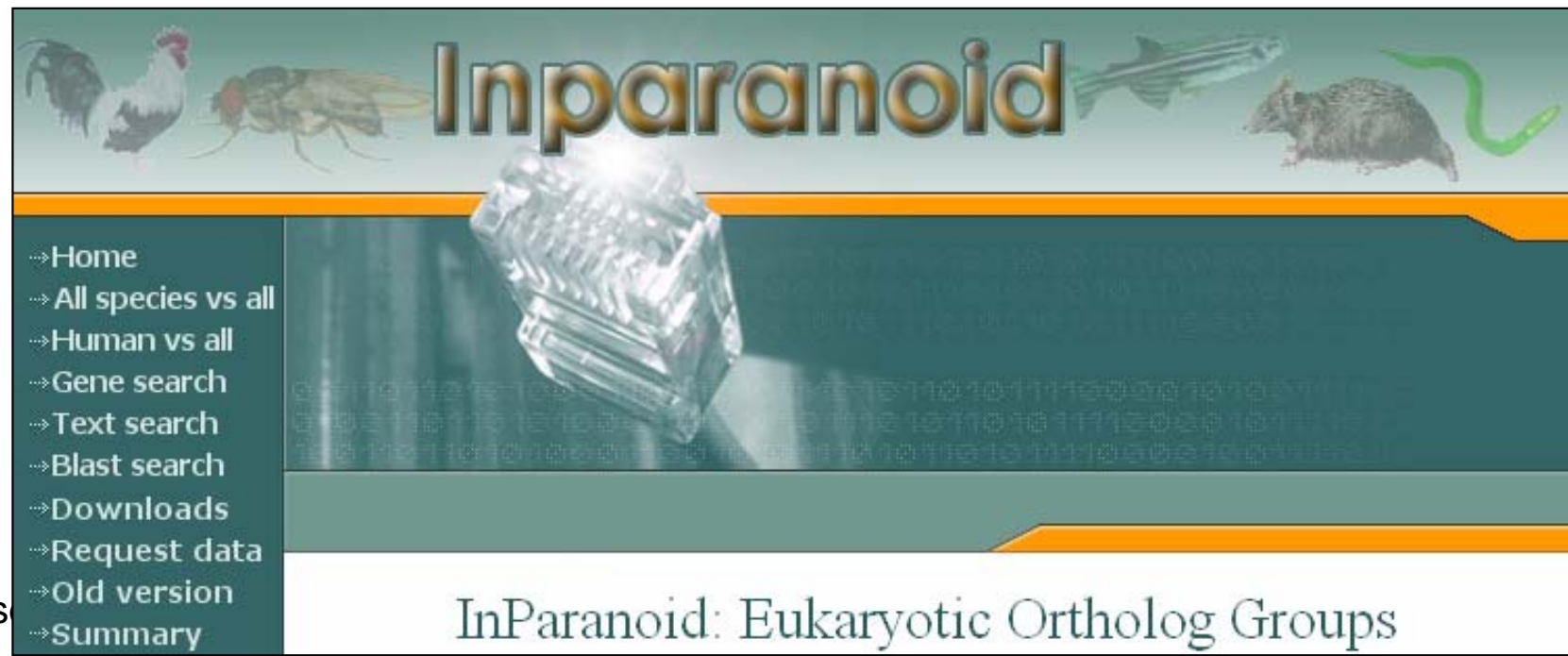
5. All pairwise-best-hits are assumed to be orthologs

All proteins (genes) of species **A**



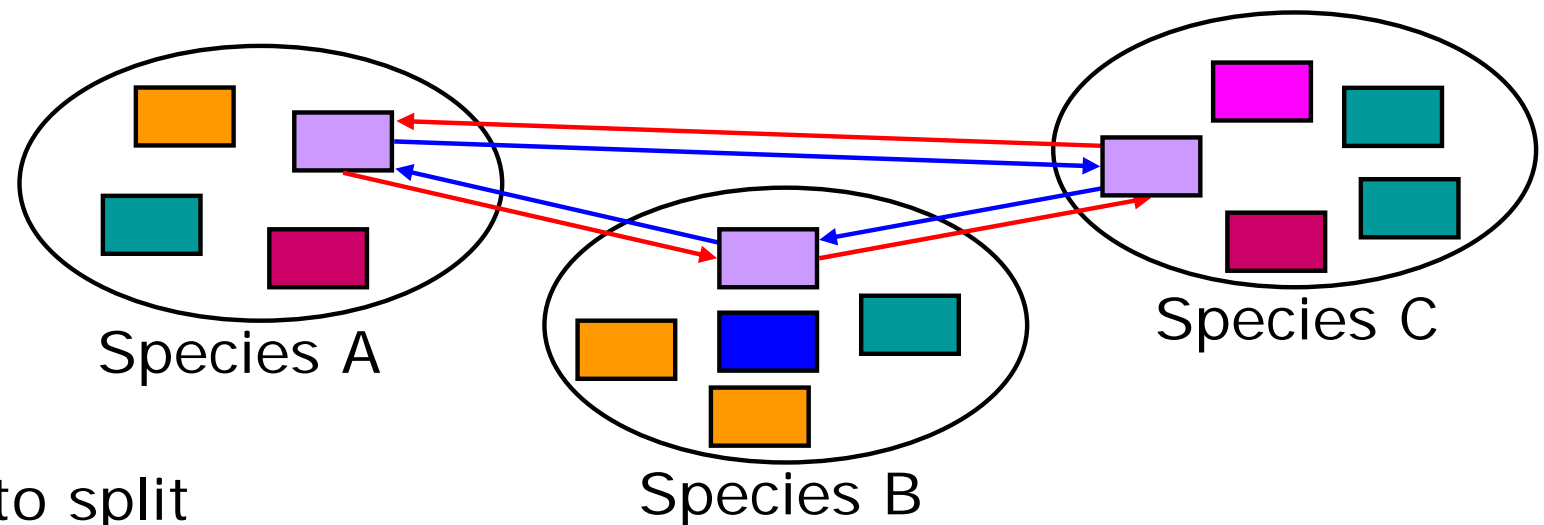
All proteins (genes) of species **B**

- <http://inparanoid.cgb.ki.se/>
- clustering software available (written in Perl, easy to use)
- after determination of main-orthologs inparalogs are added to the groups
- inparalogs: duplicated after speciation event
- outparalogs: speciation event after duplication
- uses BLAST
- K.O'Brien *et al.*, NAR, 2005, 33 (DB)



The screenshot shows the InParanoid website interface. At the top, the word "Inparanoid" is displayed in a large, stylized font, surrounded by illustrations of various animals: a rooster, a fly, a fish, a mouse, and a green worm. Below the title, there is a navigation menu on the left side with the following items: Home, All species vs all, Human vs all, Gene search, Text search, Blast search, Downloads, Request data, Old version, and Summary. The main content area features a large, clear image of a computer keyboard key. At the bottom of the page, the text "InParanoid: Eukaryotic Ortholog Groups" is visible.

- <http://www.ncbi.nlm.nih.gov/COG/>
- Clusters of Orthologous Groups of proteins
- based on all-against-all sequence search
- a protein builds a COG if pairwise-best-hits consist for at least 3 species
- manual post-processing (alignments, trees) of COGs to split COGs of multi-domain-proteins



- R.L.Tatusov
et al., 1997,
Science, 278

