LOSCHMIDT
LABORATORIES

PROTEIN ENGINEERING

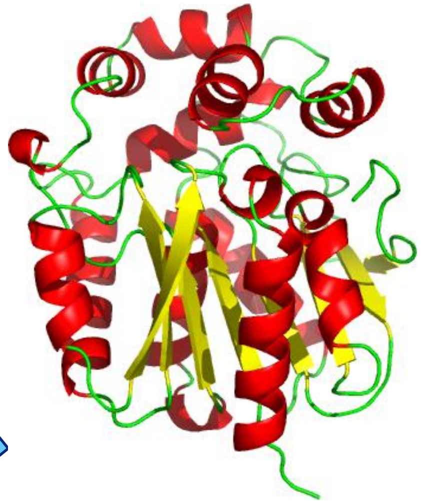
2. *IN SILICO* IDENTIFICATION OF PROTEINS

Outline

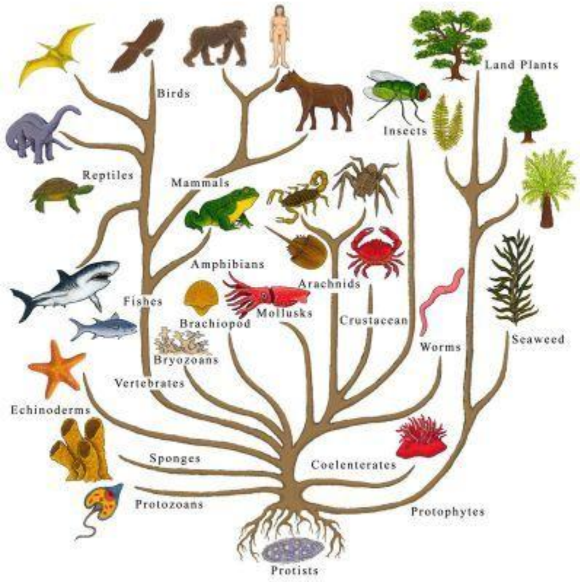
- Why to search for new proteins?
- How to acquire new proteins?
  - traditional approach
  - metagenomic approach
  - bioinformatic approach
- Bioinformatic approach
  - Where to find target sequences?
  - How to find target sequences?
  - How to recognize interesting sequences?
- What to keep in mind?

Strategies for protein optimization

Optimization of protein for applications

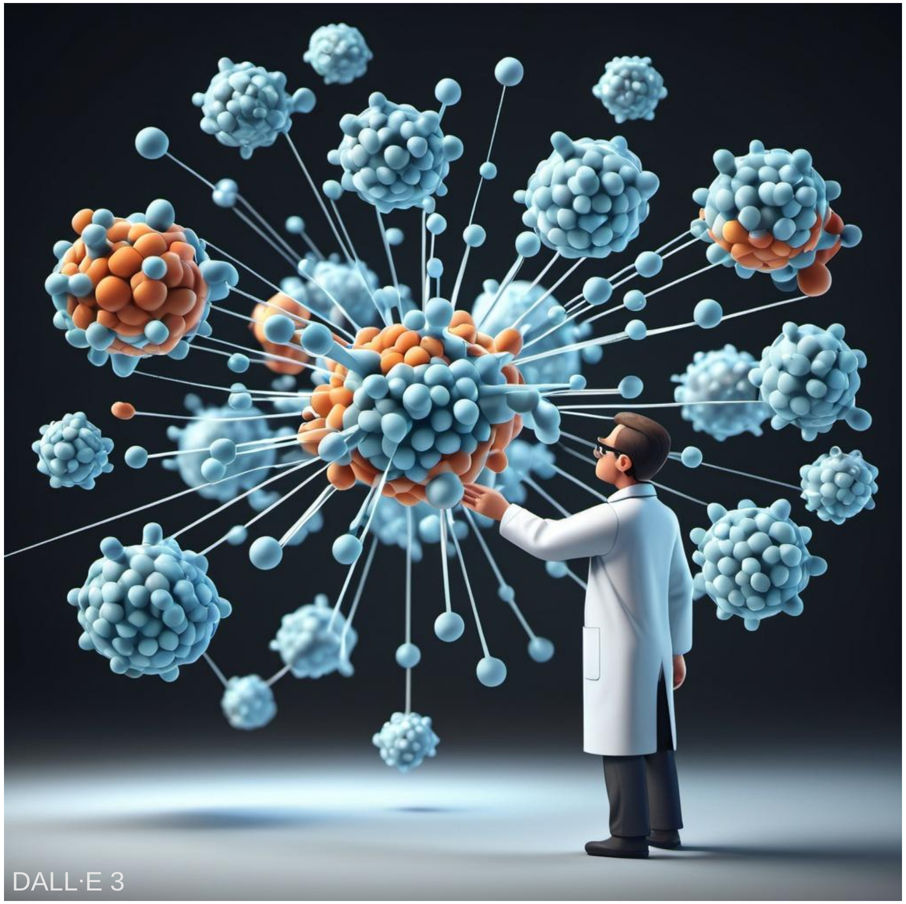


Protein engineering



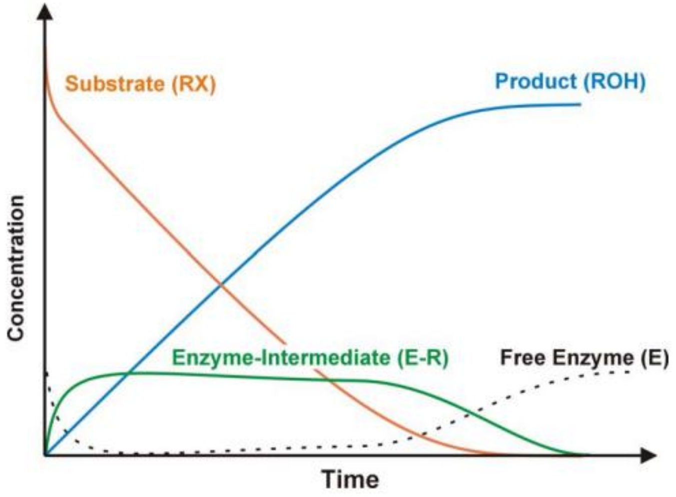
Natural diversity

Why to search for new proteins?



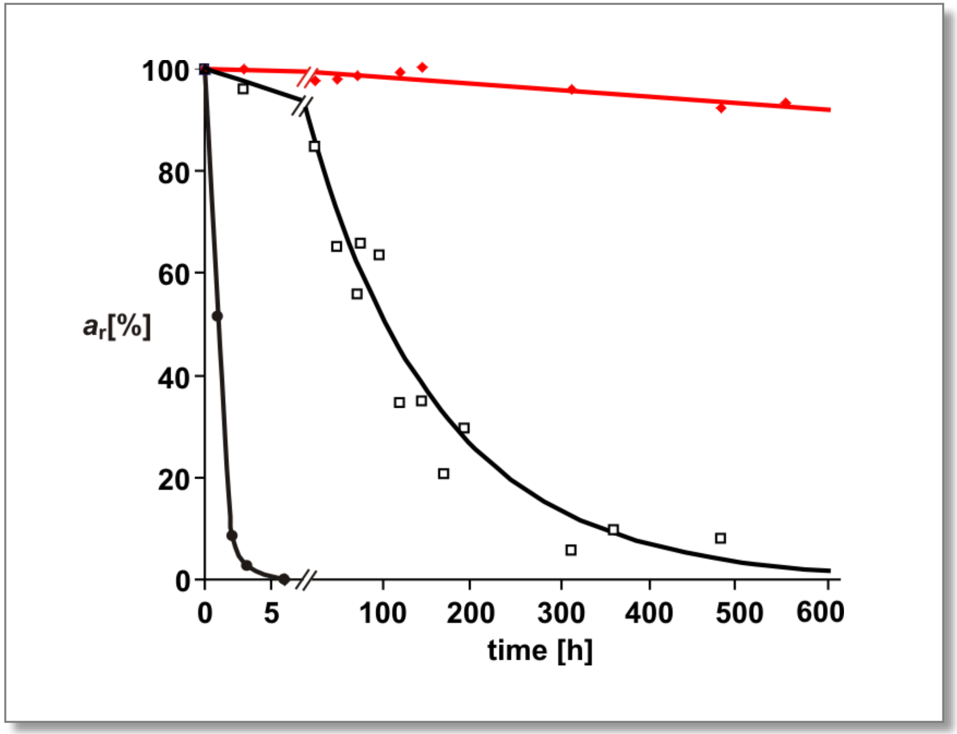
Why to search for new proteins?

- ❑ better understanding of structure-function relationships
  - required for rational design



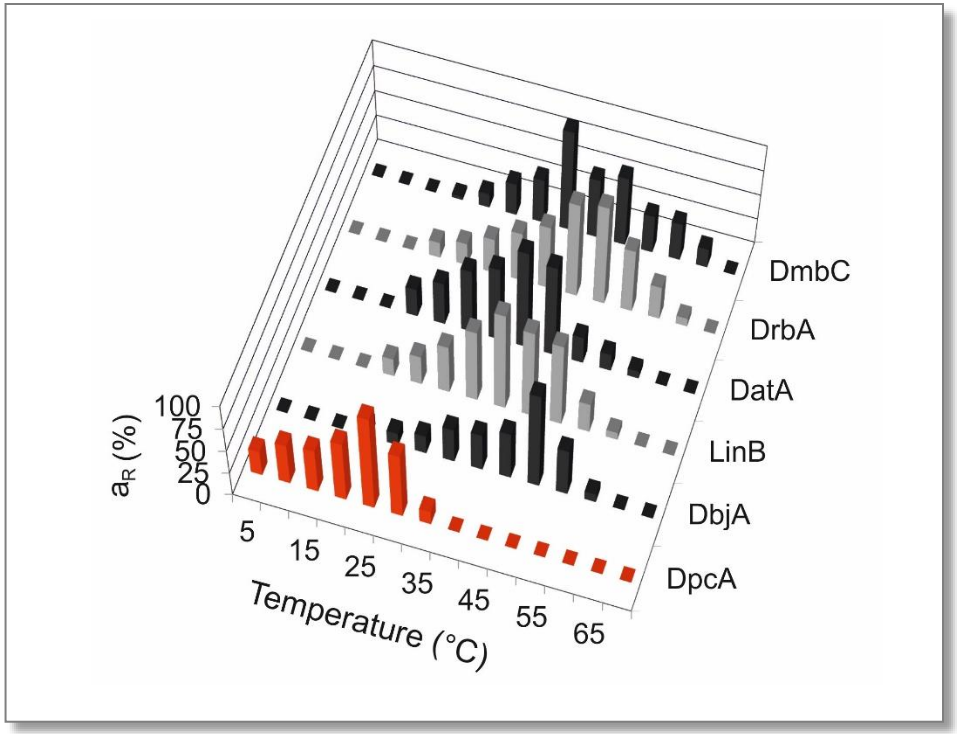
Why to search for new proteins?

- ❑ better understanding of structure-function relationships
- ❑ novel properties
    - ▪ stability



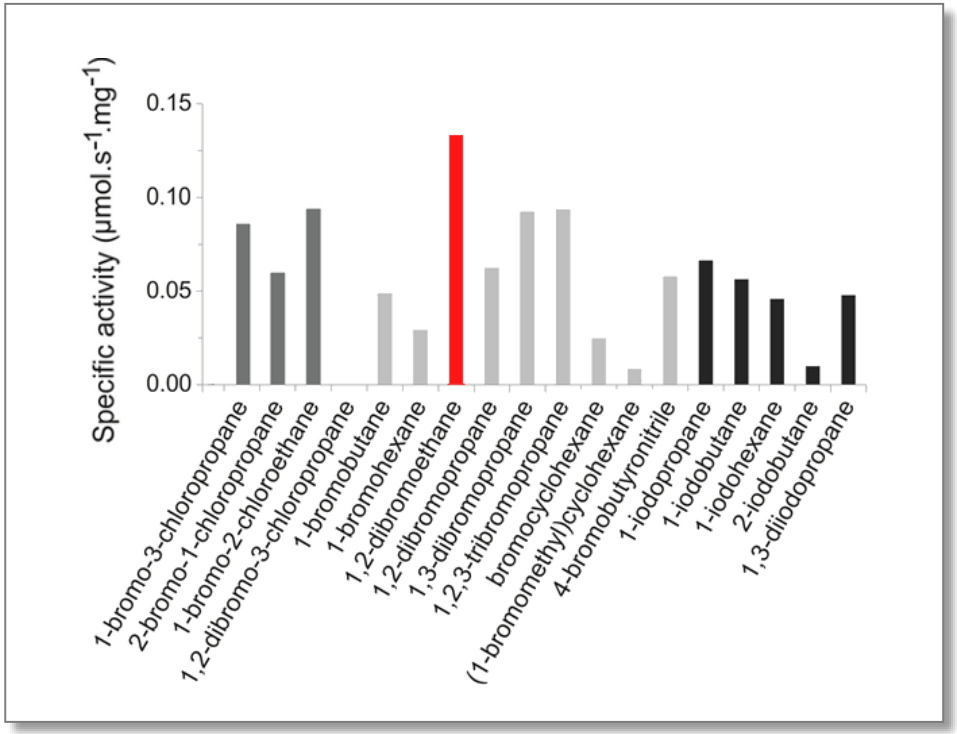
Why to search for new proteins?

- ❑ better understanding of structure-function relationships
- ❑ novel properties
  - ▪ stability
  - ▪ temperature profile



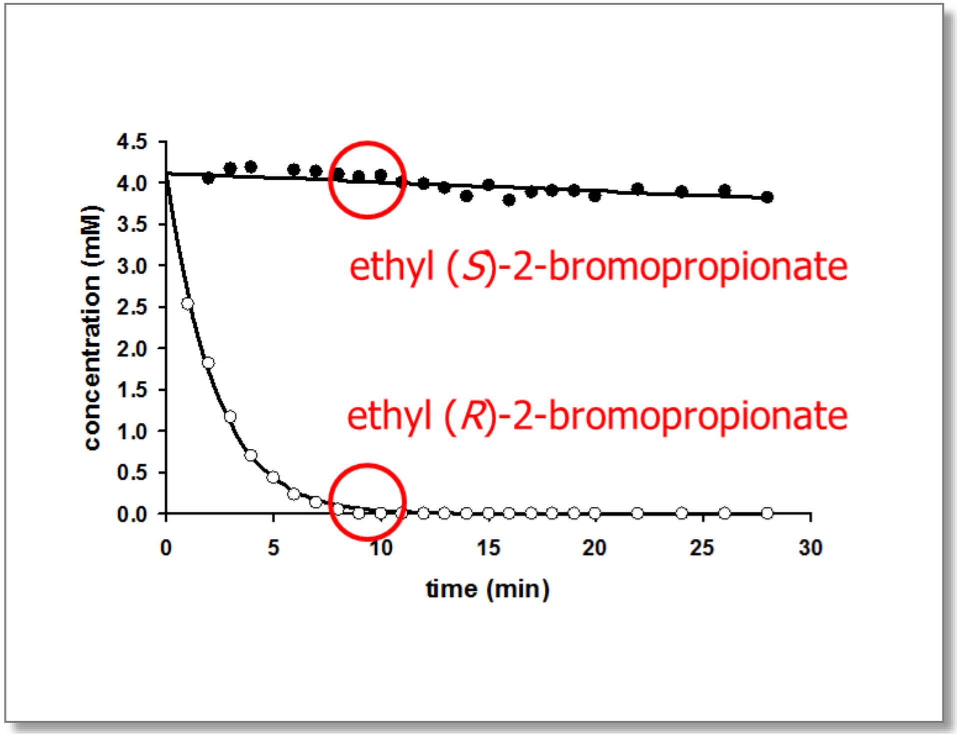
Why to search for new proteins?

- better understanding of structure-function relationships
- novel properties
  - stability
  - temperature profile
  - activity
  - specificity



Why to search for new proteins?

- ❑ better understanding of structure-function relationships
- ❑ novel properties
  - stability
  - temperature profile
  - activity
  - specificity
  - enantioselectivity



Why to search for new proteins?

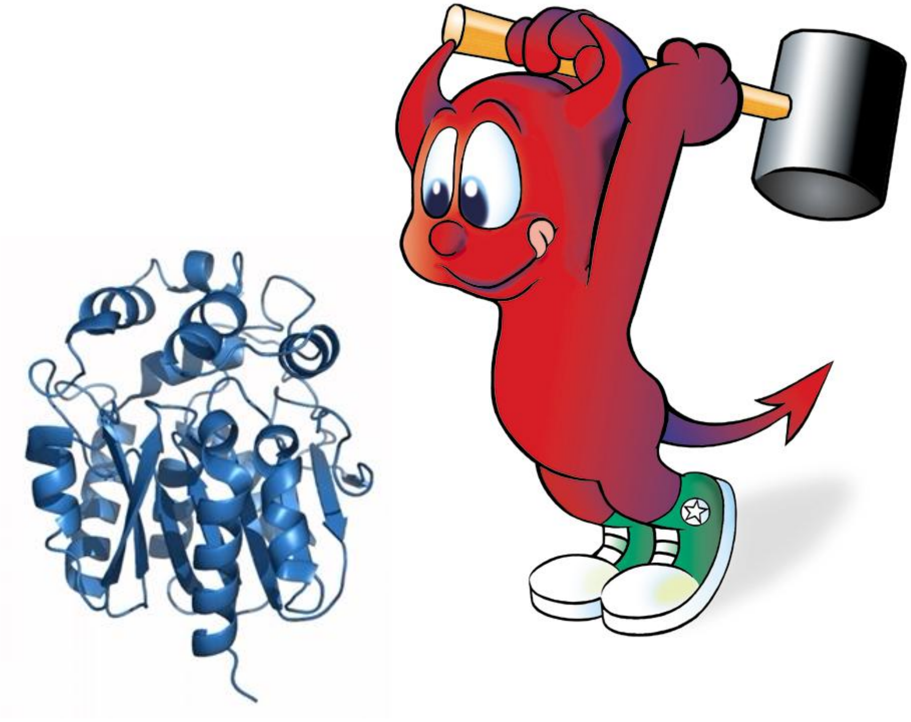
- better understanding of structure-function relationships
- novel properties
  - stability
  - temperature profile
  - activity
  - specificity
  - enantioselectivity
  - …

Why to search for new proteins?

□ better understanding of structure-function relationships

□ novel properties

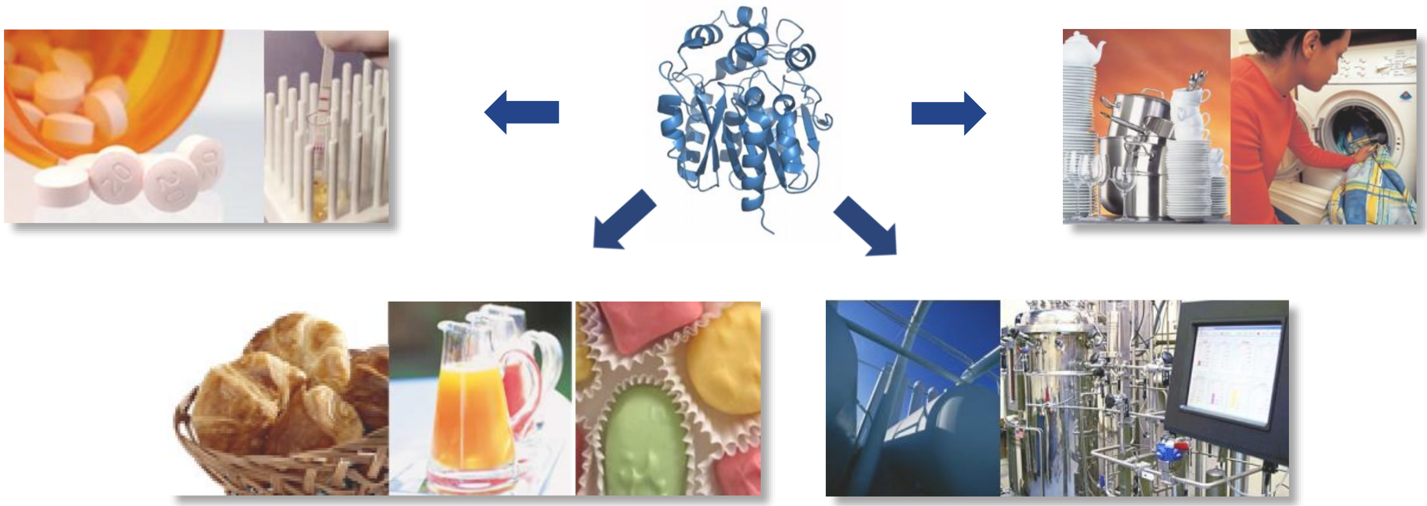
□ better starting points for protein engineering



Why to search for new proteins?

- ❑ better understanding of structure-function relationships
- ❑ novel properties
- ❑ better starting points for protein engineering

→ proteins with desired properties → practical applications

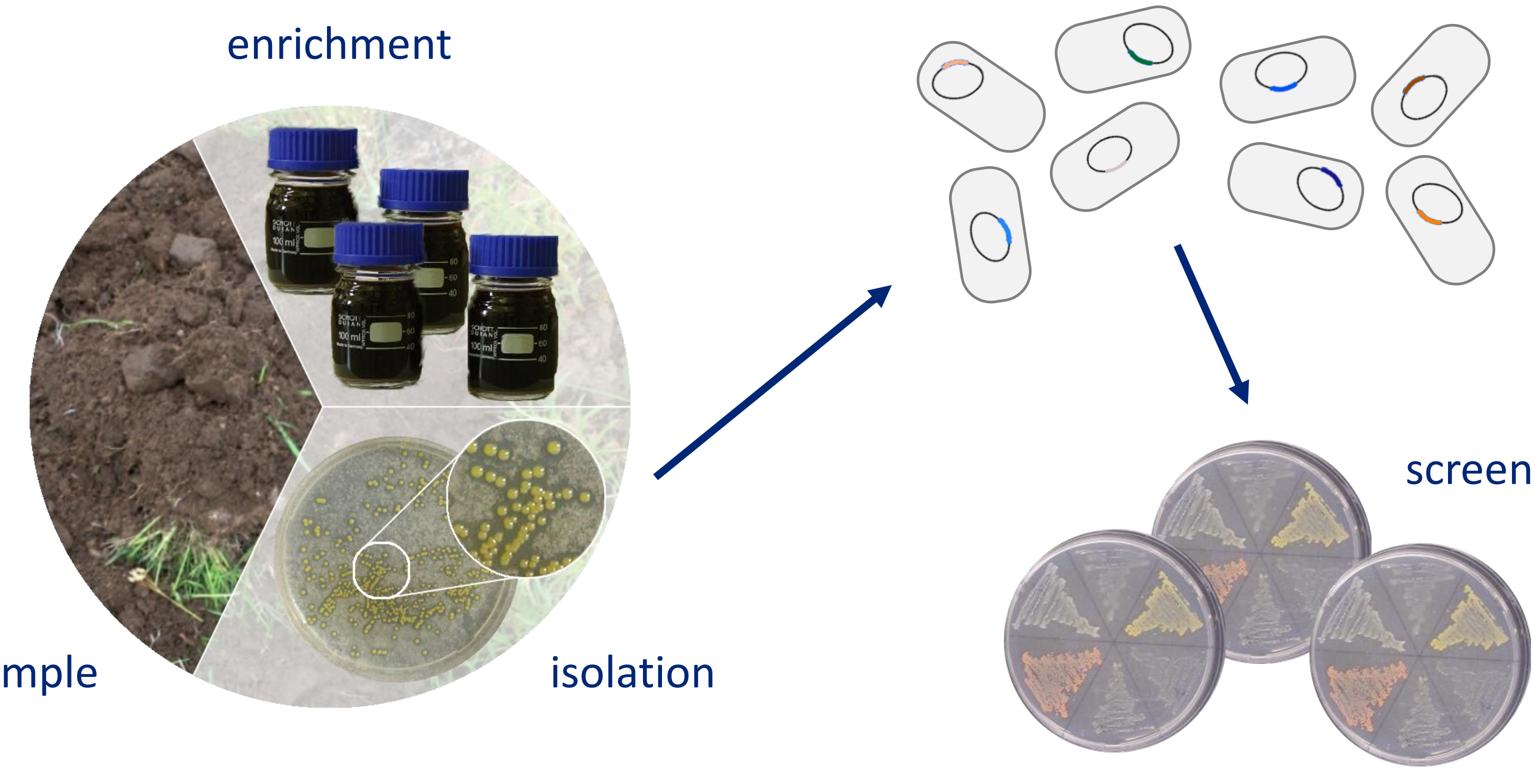


How to acquire new proteins?

- traditional approach
- metagenomic approach
- bioinformatic approach

How to acquire new proteins?

- ❑ traditional approach



enrichment

sample

isolation

DNA library

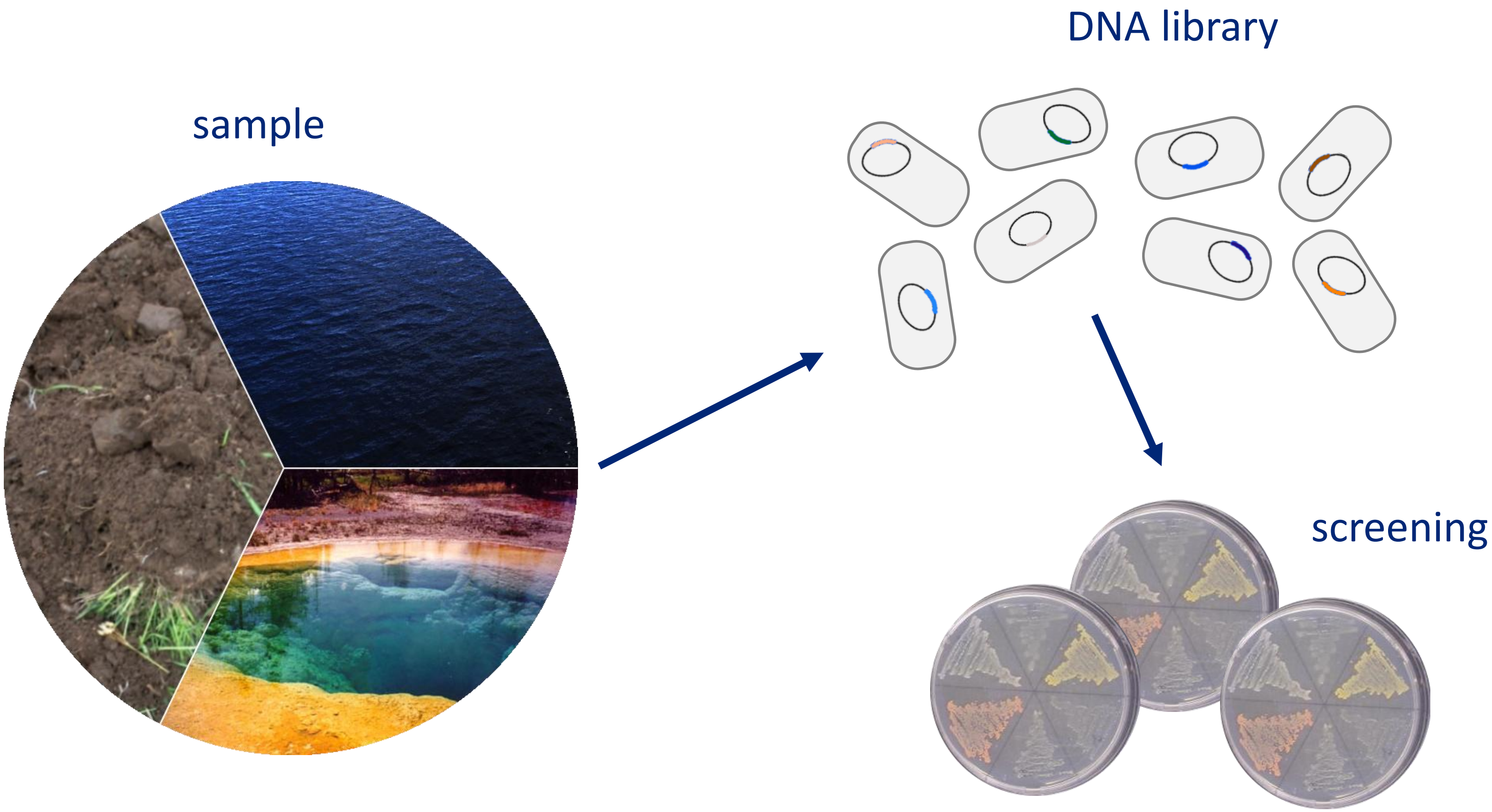
screening

How to acquire new proteins?

- traditional approach
  - microorganisms possessing target activity are enriched from the environment and isolated in pure culture
  - proteins or corresponding genes are recovered from organisms by protein purification, DNA library screening, PCR with specific primers, etc.
  - majority of microorganisms (> 99 %) cannot be cultivated using standard techniques → a large fraction of the microbial diversity in an environment is lost

How to acquire new proteins?

- ❑ metagenomic approach



sample

DNA library

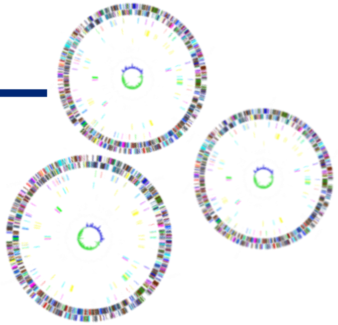
screening

How to acquire new proteins?

- metagenomic approach
    - isolation and cloning of DNA extracted directly from environmental sample (without culturing the present organisms)
    - genes recovered by DNA library screening or PCR with specific primers,...
    - enables to explore biodiversity of uncultured microorganisms

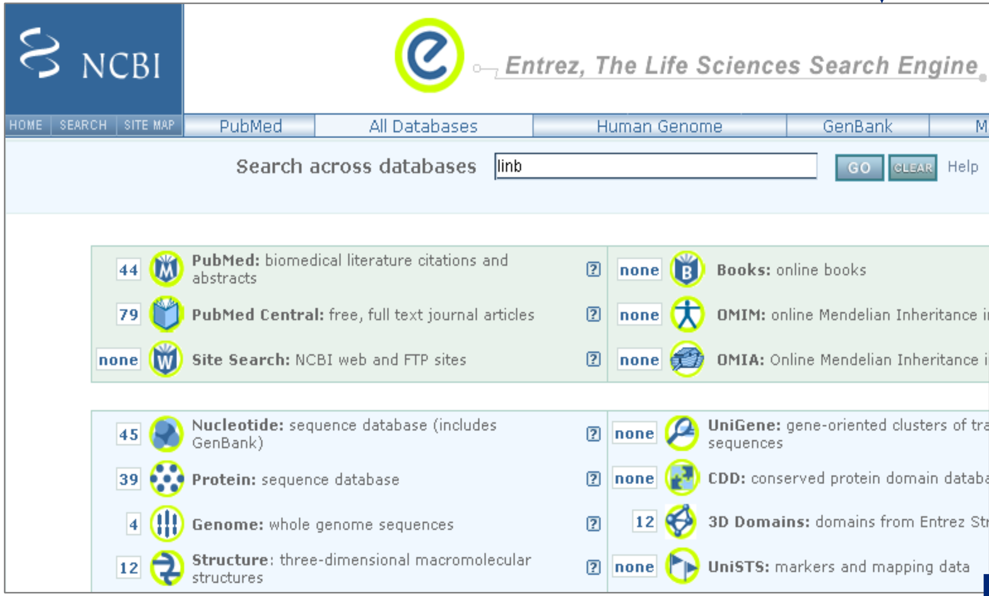
How to acquire new proteins?

- ❑ bioinformatic approach



(meta)genomic sequencing projects

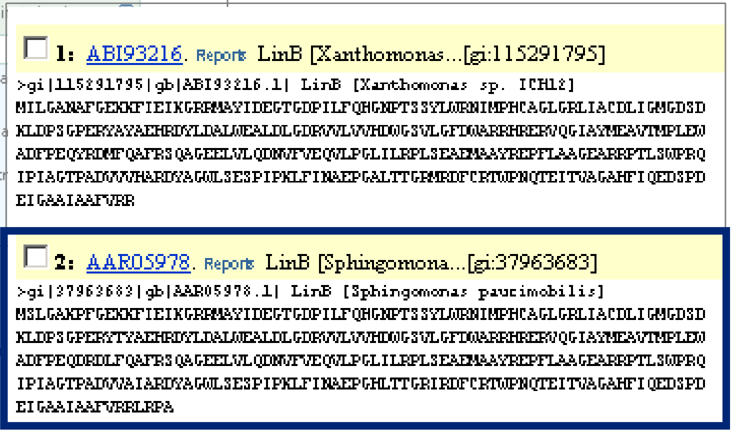
sequence database



Search across databases linb

44	**PubMed:** biomedical literature citations and abstracts	none	**Books:** online books
79	**PubMed Central:** free, full text journal articles	none	**OMIM:** online Mendelian Inheritance in Man
none	**Site Search:** NCBI web and FTP sites	none	**OMIA:** Online Mendelian Inheritance i
45	**Nucleotide:** sequence database (includes GenBank)	none	**UniGene:** gene-oriented clusters of tra sequences
39	**Protein:** sequence database	none	**CDD:** conserved protein domain databa
4	**Genome:** whole genome sequences	12	**3D Domains:** domains from Entrez Str
12	**Structure:** three-dimensional macromolecular structures	none	**UniSTS:** markers and mapping data

in silico "screening"



☐ **1:** ABI93216. Report LinB [Xanthomonas...[gi:115291795]

```
>gi|115291795|gb|ABI93216.1| LinB [Xanthomonas sp. ICH12]
MILGANAFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLIGMGDSD
KLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVVHDWGSVLGFDWARRHRERVQGIAYMEAVTMPLEW
ADFPEQYRDMFQAFRSQAGEELVLQDNVFVEQVLPGLILRPLSEAEMAAYREPFLAAGEARRPTLSWPRQ
IPIAGTPADVVVHARDYAGWLSESPIPKLFINAEPGALTTGRMRDFCRTWPNQTEITVAGAHFIQEDSPD
EIGAAIAAFVRR
```

☐ **2:** AAR05978. Report LinB [Sphingomona...[gi:37963683]

```
>gi|37963683|gb|AAR05978.1| LinB [Sphingomonas paucimobilis]
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLIGMGDSD
KLDPSGPERYTYAEHRDYLDALWEALDLGDRVVLVVHDWGSVLGFDWARRHRERVQGIAYMEAVTMPLEW
ADFPEQDRDLFQAFRSQAGEELVLQDNVFVEQVLPGLILRPLSEAEMAAYREPFLAAGEARRPTLSWPRQ
IPIAGTPADVVAIARDYAGWLSESPIPKLFINAEPGHLTTGRIRDFCRTWPNQTEITVAGAHFIQEDSPD
EIGAAIAAFVRRLRPA
```

gene synthesis, DNA request

How to acquire new proteins?

- bioinformatic approach
  - sequence data from genomic and metagenomic sequencing projects are stored in sequence databases
  - *in silico* searching of sequence databases

    → fast and cheap way to identify novel proteins

    → one cannot find what is not in the database (but there is a lot of data - more than one usually needs)
  - genes are recovered by gene synthesis or obtained from sequencing consortia upon request

Where to find target sequences?

- databases of nucleotide sequences
- databases of protein sequences

Databases of nucleotide sequences

- GenBank
  - http://www.ncbi.nlm.nih.gov/genbank/
  - provided by NCBI (National Center for Biotechnology Information)

- EMBL-BANK
  - http://www.ebi.ac.uk/embl/
  - provided by EBI (European Bioinformatics Institute)

- DDBJ
  - http://www.ddbj.nig.ac.jp/
  - provided by National Institute of Genetics from Japan

Databases of nucleotide sequences

- GenBank, EMBL-Bank, DDBJ
  - annotated collections of all publically available nucleotide sequences
  - freely available to the wide community
  - contain data obtained from genomic centers or research institutions
  - everyday synchronization of new or updated data
  - contain about 250,000,000 sequences (Feb 2024)
  - mostly automatic annotations – lower quality, errors

Databases of protein sequences

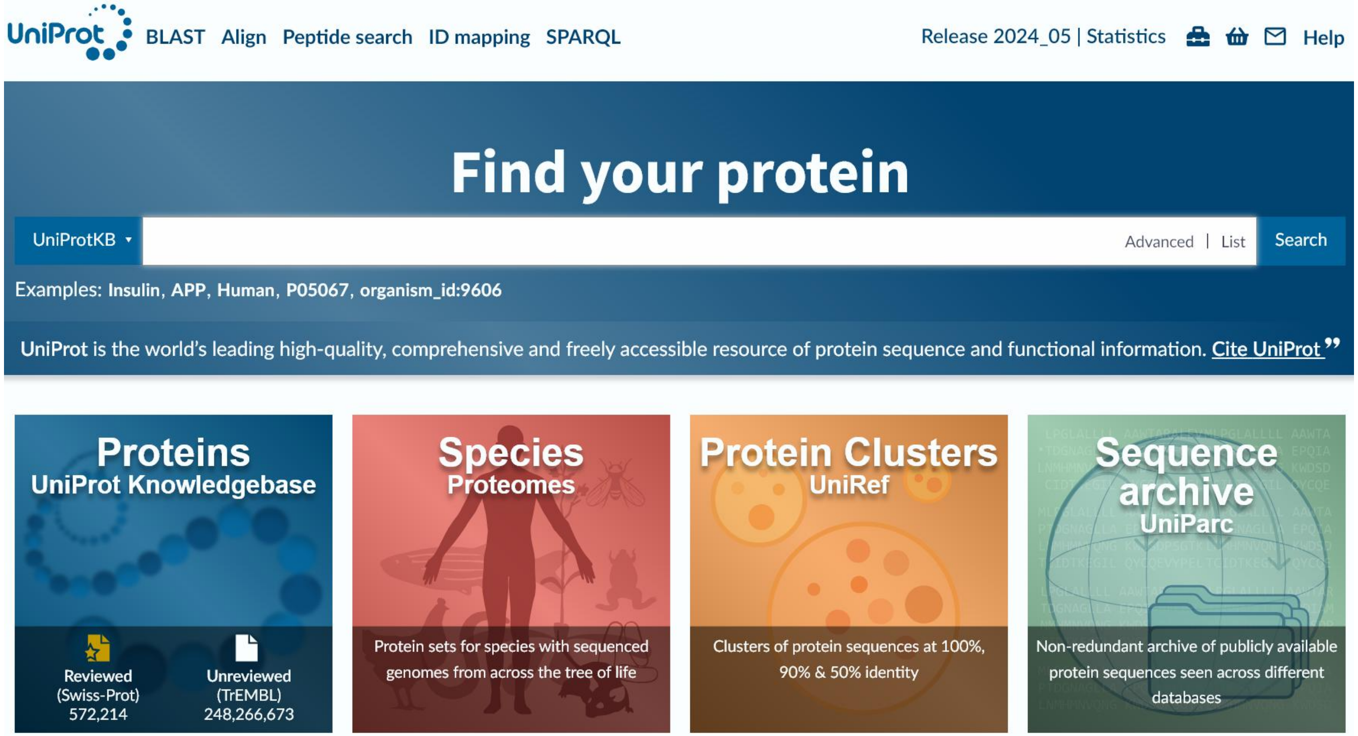
- ❑ UniProtKB
  - http://www.uniprot.org/
  - provided by EBI, Swiss Institute of Bioinformatics and Protein Information Resource
- ❑ nr Protein database
  - http://www.ncbi.nlm.nih.gov/protein/
  - provided by NCBI

Databases of protein sequences

- UniProtKB, nr Protein database
  - annotated collections of publically available protein sequences
  - freely available to wide community
  - contain data obtained by conceptual translation of coding sequences from EMBL-Bank/GenBank/DDBJ or provided by research institutions
  - contain more than 250,000,000 sequences (Feb 2024)
  - mostly automatic annotations – lower quality, errors

Databases of protein sequences

- UniProtKB
  - rich annotations (e.g., information about function of protein and individual amino acids, experimental data, biological ontologies, classifications, ...)
  - clear indication of annotation quality (manual vs. automatic)



UniProt
BLAST
Align
Peptide search
ID mapping
SPARQL
Release 2024_05 | Statistics
Help
Find your protein
UniProtKB
Advanced | List
Search
Examples: Insulin, APP, Human, P05067, organism_id:9606
UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt
Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 572,214
Unreviewed (TrEMBL) 248,266,673
Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life
Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity
Sequence archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases

Databases of protein sequences

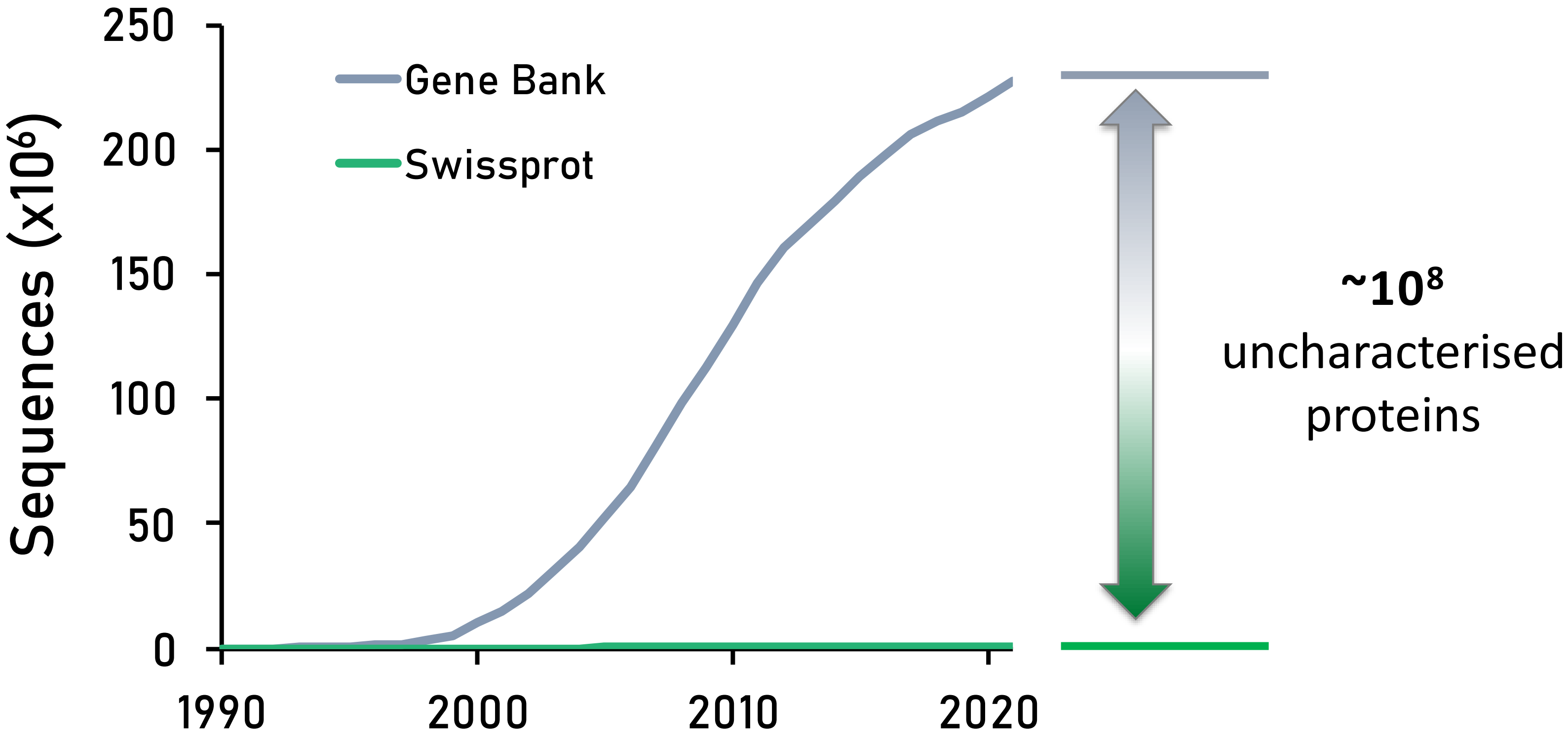
- UniProtKB/Swiss-Prot
  - high-quality annotations, i.e., manually annotated entries or expert-reviewed automatic annotations
  - source of reliable information
  - contains “only” ~ 570,000 sequences (Feb 2025)

- UniProtKB/TrEMBL
  - automatic annotations – lower quality, errors
  - contains ~ 250,000,000 sequences (Feb 2025)

Databases of protein sequences

Number of sequences

Number of characterized proteins



~10^8 uncharacterised proteins

Pitfalls of sequence databases

- large number of errors
  - errors in sequences (wrong base, frameshift errors)
  - wrong positions of genes
  - exon-intron boundary errors
  - errors and inaccuracies in annotations
  - ...

How to find target sequences?

- text-based searches
- sequence-based searches

Data format

- ❑ Fasta sequence:
  - Header starting with “>” followed by the sequence description
  - Sequence data are on the new line

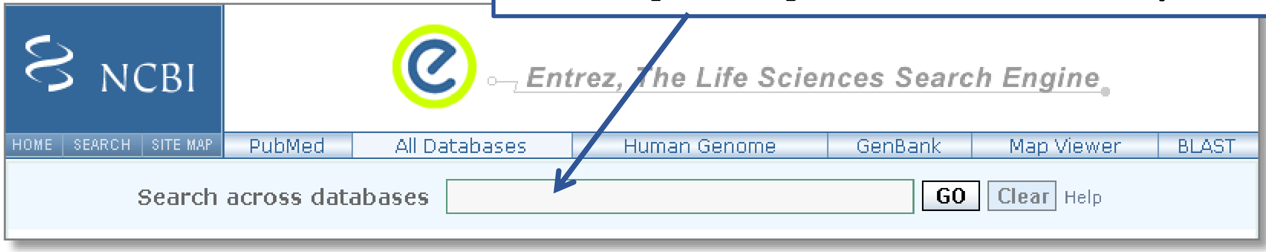
```
>Haloalkane dehalogenase LinB

MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLI

GMGDSDKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVVHDWGSALGFDWARRHRERVQGI

AYMEAIAMPIEWADFPEQDRDLFQAFRSQAGEELVLQD
```

Text-based searches

- database retrieval systems
  - enable quick and easy search of many databases at the same time
  - specification of queries using logical operators (AND, OR, NOT,...)
  - Entrez (NCBI), SRS (EBI)
- results dependent on sequence annotations
  - erroneous, inaccurate or too general annotations
  - synonyms
  - misspellings
  - ...

Text-based searches

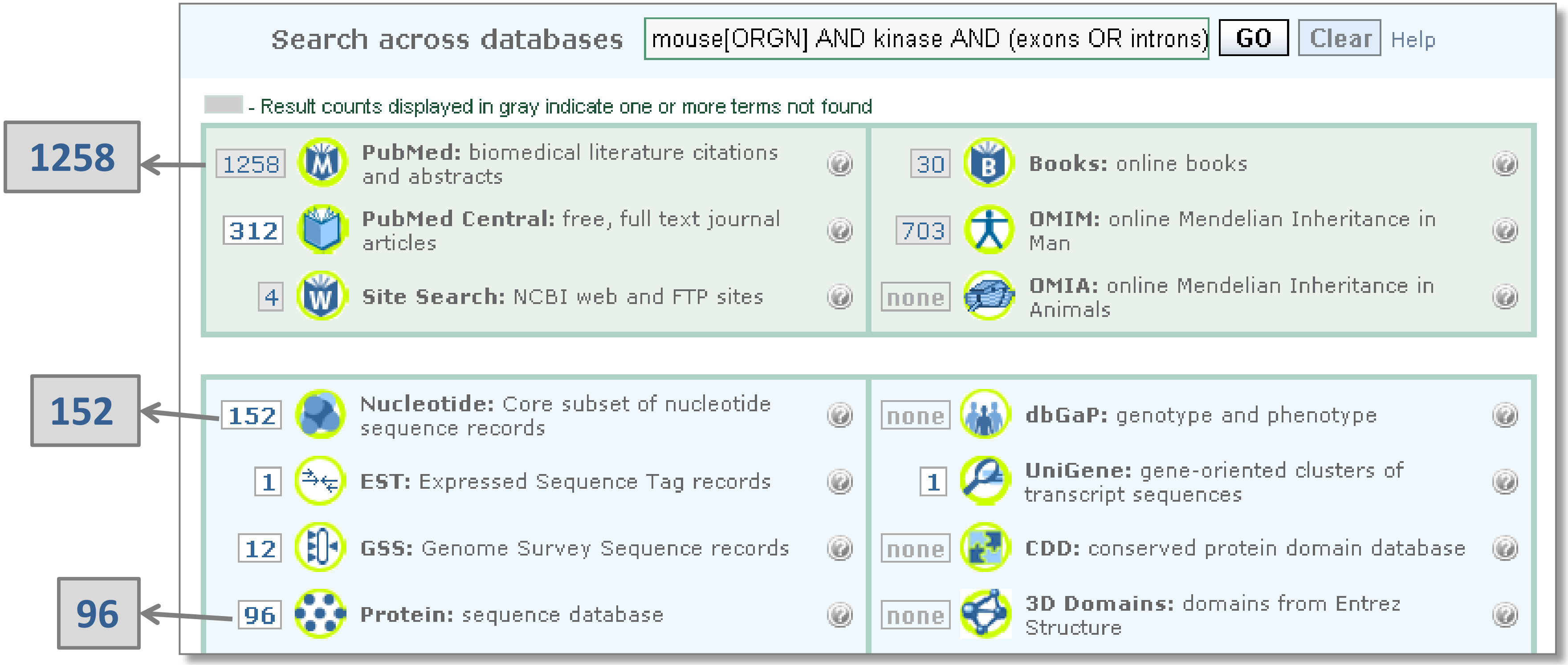
- ❑ database retrieval systems



mouse[ORGN] AND kinase AND (exons OR introns)

Text-based searches

- database retrieval systems



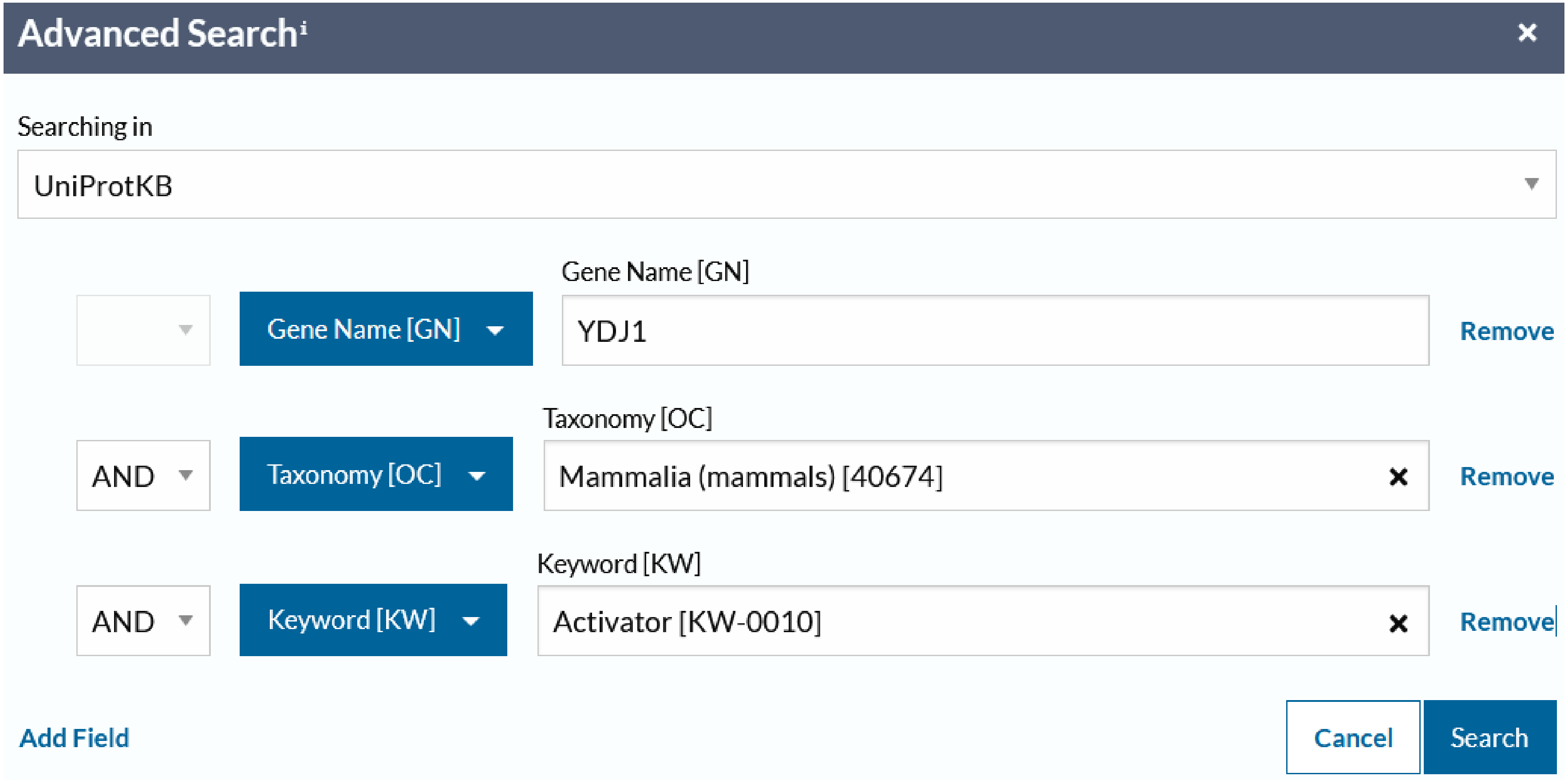
Search across databases mouse[ORGN] AND kinase AND (exons OR introns) GO Clear Help

▨ - Result counts displayed in gray indicate one or more terms not found

1258	**PubMed:** biomedical literature citations and abstracts	30	**Books:** online books
312	**PubMed Central:** free, full text journal articles	703	**OMIM:** online Mendelian Inheritance in Man
4	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals
152	**Nucleotide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
1	**EST:** Expressed Sequence Tag records	1	**UniGene:** gene-oriented clusters of transcript sequences
12	**GSS:** Genome Survey Sequence records	none	**CDD:** conserved protein domain database
96	**Protein:** sequence database	none	**3D Domains:** domains from Entrez Structure

Text-based searches

- ❑ database retrieval systems
- ❑ advanced search options



Searching in: UniProtKB

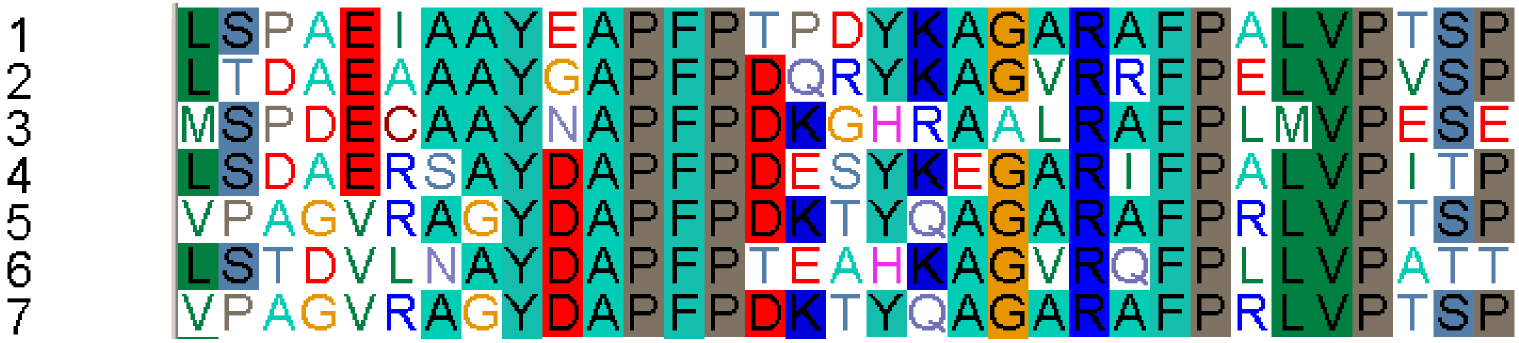
Gene Name [GN]: YDJ1

AND Taxonomy [OC]: Mammalia (mammals) [40674]

AND Keyword [KW]: Activator [KW-0010]

Sequence-based searches

- searches based on sequence similarity
  - results not influenced by sequence annotations
- rely on the assumption that proteins with the same function have similar sequence
  - not always true – close homologs vs. distant homologs vs. analogs



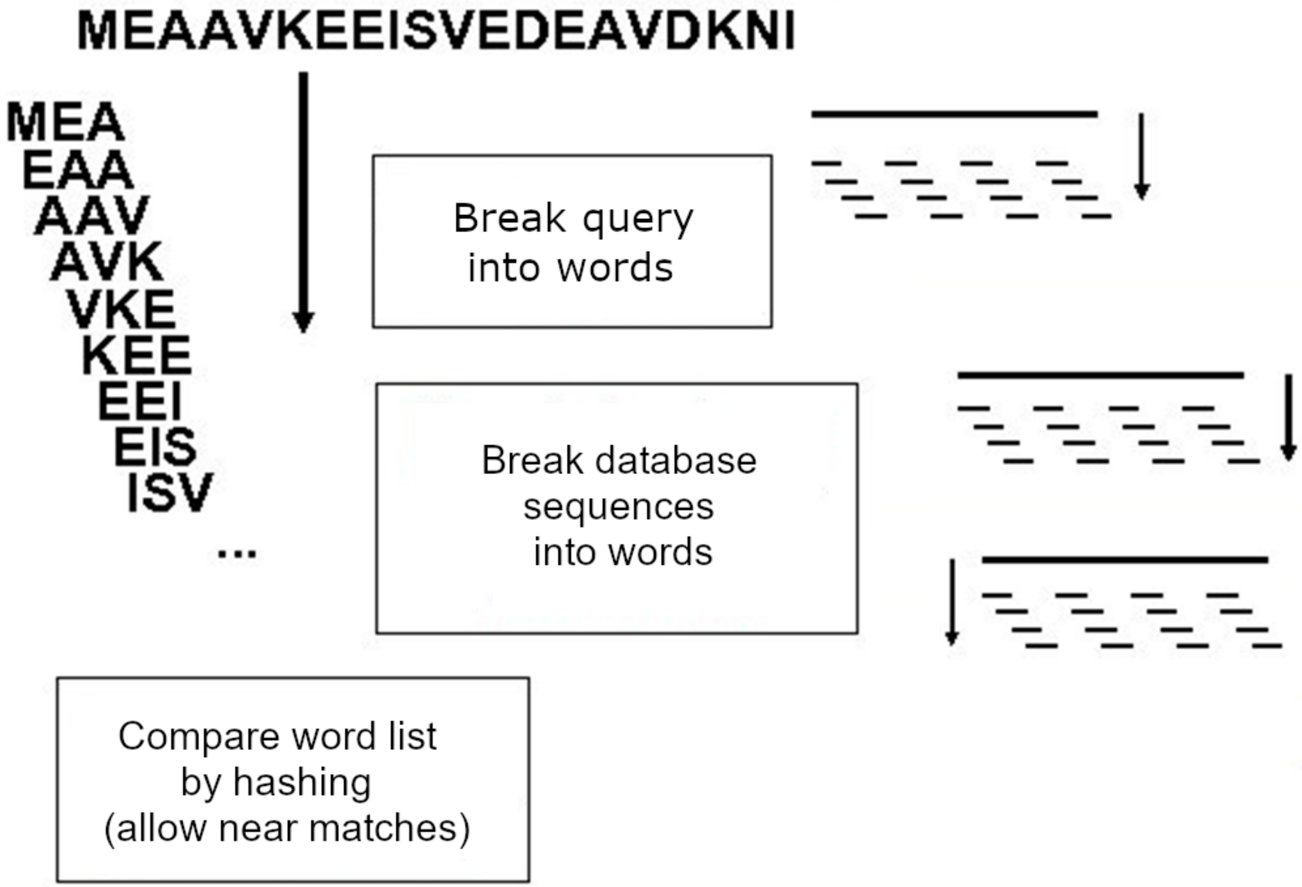
```
1    LSPAEIAAYEAPFPTPDYKAGARAFPALVPTSP
2    LTDAEAAAYGAPFPDQRYKAGVRRFPELVPVSP
3    MSPDECAAYNAPFPDKGHRAALRAFPLMVPESE
4    LSDAERSAYDAPFPDESYKEGARIFPALVPITP
5    VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
6    LSTDVLNAYDAPFPTEAHKAGVRQFPLLVPATT
7    VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
```

Sequence-based searches

- ❑ BLAST
  - Heuristic search of similarity on significant sequences
  - Reasonable sensitivity and good speed
  - Gold standard in sequence search
- ❑ PSI-BLAST
  - “iterative BLAST” making use of multiple sequence alignment
  - more sensitive search to detect weak but biologically significant similarities between sequences
- ❑ HMMER
  - Uses Hidden Markov Models for sensitive detection of remote homologs
  - Slower than BLAST for simple sequence similarity searches
  - Widely used in protein family classification and domain identification

BLAST

- Basic Local Alignment Search Tool



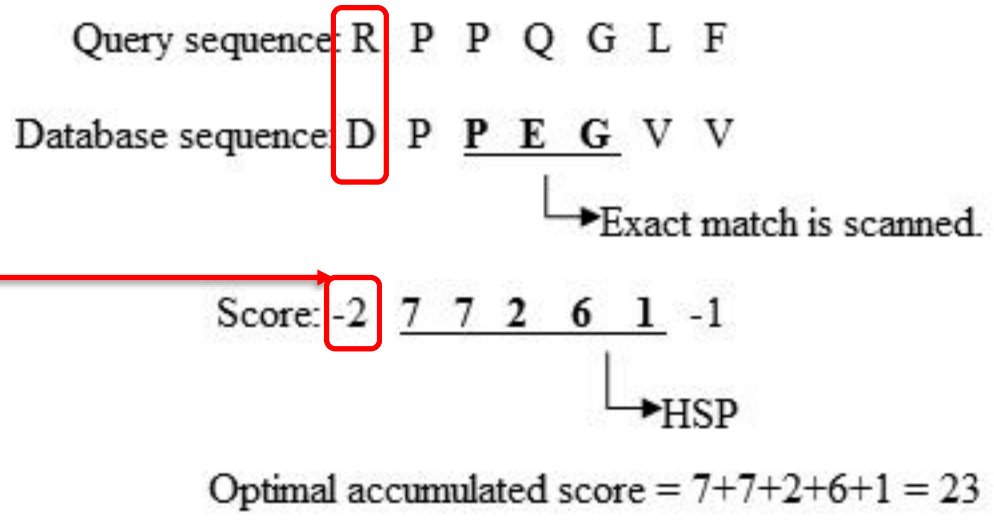
MEAAVKEEISVEDEAVDKNI
MEA
EAA
AAV
AVK
VKE
KEE
EEI
EIS
ISV
...
Break query into words
Break database sequences into words
Compare word list by hashing (allow near matches)

BLAST

- Basic Local Alignment Search Tool

BLOSUM scoring matrix

Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2									
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**



Query sequence: R P P Q G L F

Database sequence: D P **P E G** V V

Exact match is scanned.

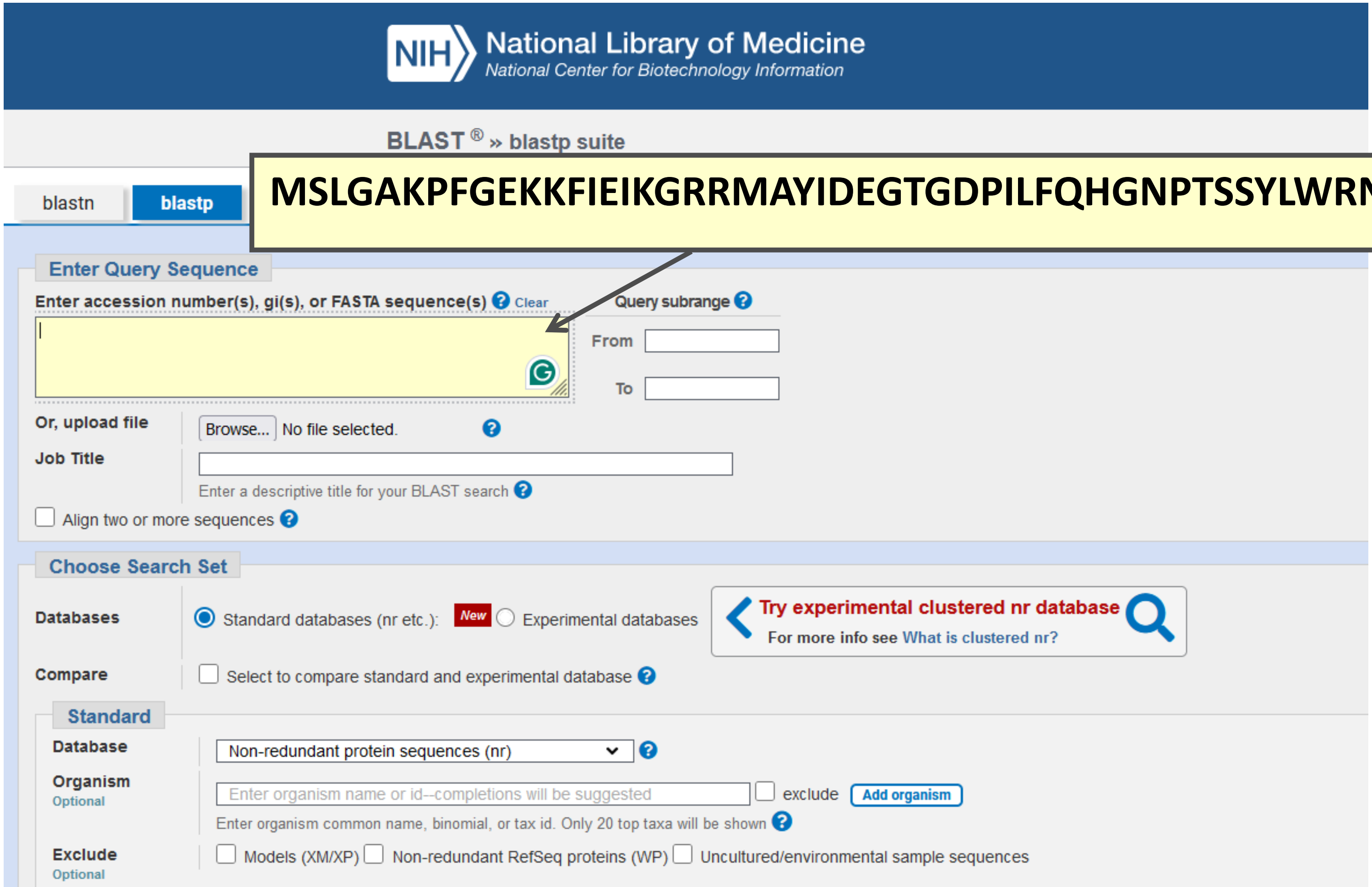
Score: -2 7 7 2 6 1 -1

HSP

Optimal accumulated score = 7+7+2+6+1 = 23

Sequence-based searches

❑ BLAST input



MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNI

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST ® » blastp suite

blastn | blastp

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

Query subrange ?

From

To

Or, upload file: Browse... No file selected. ?

Job Title

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database
For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database ?

Standard

Database: Non-redundant protein sequences (nr) ?

Organism
Optional: Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?

Exclude
Optional: □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

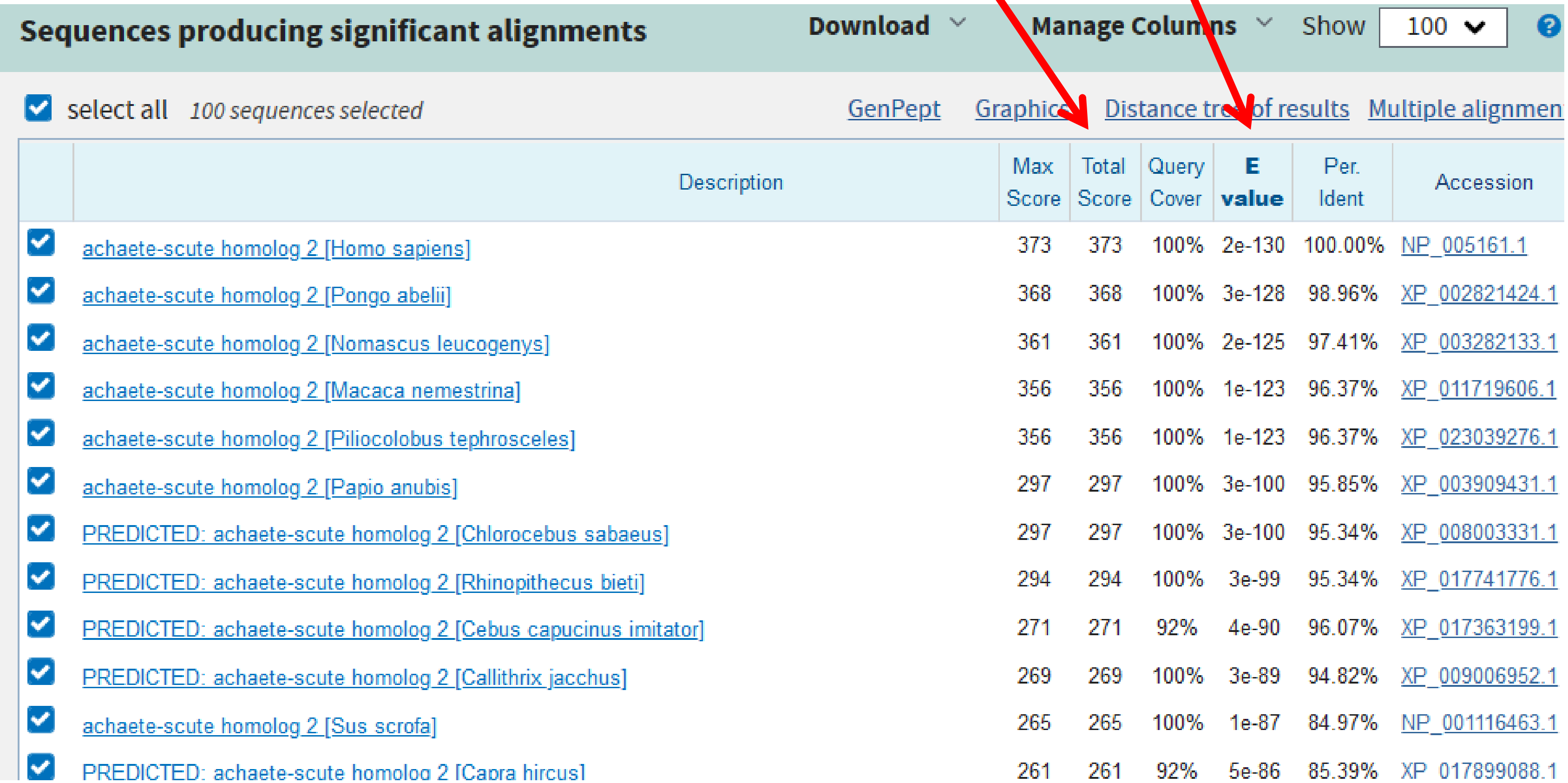
Sequence-based searches

- BLAST results

Score

E-value

hits



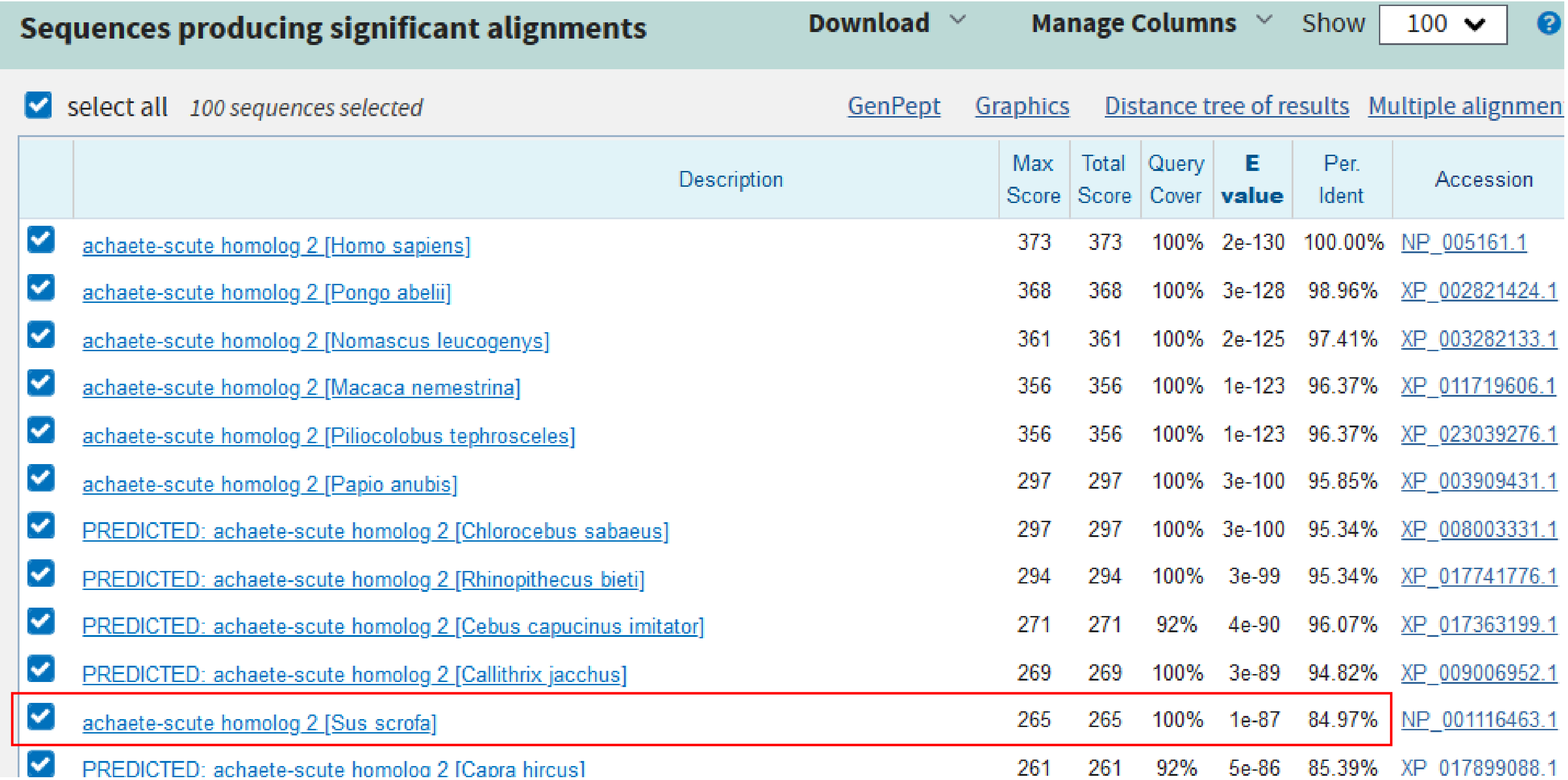
Sequences producing significant alignments **Download** ˅ **Manage Columns** ˅ Show 100 ˅ ?

☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignmen

	Description	Max Score	Total Score	Query Cover	**E value**	Per. Ident	Accession
☑	achaete-scute homolog 2 [Homo sapiens]	373	373	100%	2e-130	100.00%	NP_005161.1
☑	achaete-scute homolog 2 [Pongo abelii]	368	368	100%	3e-128	98.96%	XP_002821424.1
☑	achaete-scute homolog 2 [Nomascus leucogenys]	361	361	100%	2e-125	97.41%	XP_003282133.1
☑	achaete-scute homolog 2 [Macaca nemestrina]	356	356	100%	1e-123	96.37%	XP_011719606.1
☑	achaete-scute homolog 2 [Piliocolobus tephrosceles]	356	356	100%	1e-123	96.37%	XP_023039276.1
☑	achaete-scute homolog 2 [Papio anubis]	297	297	100%	3e-100	95.85%	XP_003909431.1
☑	PREDICTED: achaete-scute homolog 2 [Chlorocebus sabaeus]	297	297	100%	3e-100	95.34%	XP_008003331.1
☑	PREDICTED: achaete-scute homolog 2 [Rhinopithecus bieti]	294	294	100%	3e-99	95.34%	XP_017741776.1
☑	PREDICTED: achaete-scute homolog 2 [Cebus capucinus imitator]	271	271	92%	4e-90	96.07%	XP_017363199.1
☑	PREDICTED: achaete-scute homolog 2 [Callithrix jacchus]	269	269	100%	3e-89	94.82%	XP_009006952.1
☑	achaete-scute homolog 2 [Sus scrofa]	265	265	100%	1e-87	84.97%	NP_001116463.1
☑	PREDICTED: achaete-scute homolog 2 [Capra hircus]	261	261	92%	5e-86	85.39%	XP_017899088.1

Sequence-based searches

- ❑ BLAST results



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	achaete-scute homolog 2 [Homo sapiens]	373	373	100%	2e-130	100.00%	NP_005161.1
☑	achaete-scute homolog 2 [Pongo abelii]	368	368	100%	3e-128	98.96%	XP_002821424.1
☑	achaete-scute homolog 2 [Nomascus leucogenys]	361	361	100%	2e-125	97.41%	XP_003282133.1
☑	achaete-scute homolog 2 [Macaca nemestrina]	356	356	100%	1e-123	96.37%	XP_011719606.1
☑	achaete-scute homolog 2 [Piliocolobus tephrosceles]	356	356	100%	1e-123	96.37%	XP_023039276.1
☑	achaete-scute homolog 2 [Papio anubis]	297	297	100%	3e-100	95.85%	XP_003909431.1
☑	PREDICTED: achaete-scute homolog 2 [Chlorocebus sabaeus]	297	297	100%	3e-100	95.34%	XP_008003331.1
☑	PREDICTED: achaete-scute homolog 2 [Rhinopithecus bieti]	294	294	100%	3e-99	95.34%	XP_017741776.1
☑	PREDICTED: achaete-scute homolog 2 [Cebus capucinus imitator]	271	271	92%	4e-90	96.07%	XP_017363199.1
☑	PREDICTED: achaete-scute homolog 2 [Callithrix jacchus]	269	269	100%	3e-89	94.82%	XP_009006952.1
☑	achaete-scute homolog 2 [Sus scrofa]	265	265	100%	1e-87	84.97%	NP_001116463.1
☑	PREDICTED: achaete-scute homolog 2 [Capra hircus]	261	261	92%	5e-86	85.39%	XP_017899088.1

Sequence-based searches

- BLAST Score
  - raw score normalized on the basis of the scoring method
  - sum of substitution scores and gap penalties
  - higher is better, but does not adequately represent significance of alignment
- BLAST *E*-value
  - number of BLAST alignments with a given or better Score that are expected to be seen simply by chance (with random sequence)
  - indicator of alignment significance (adjusted to the database size)
  - results associated with the lowest *E*-values are the best
  - hits with an *E*-value score > 0.01 belong to the "grey zone" – do not trust them

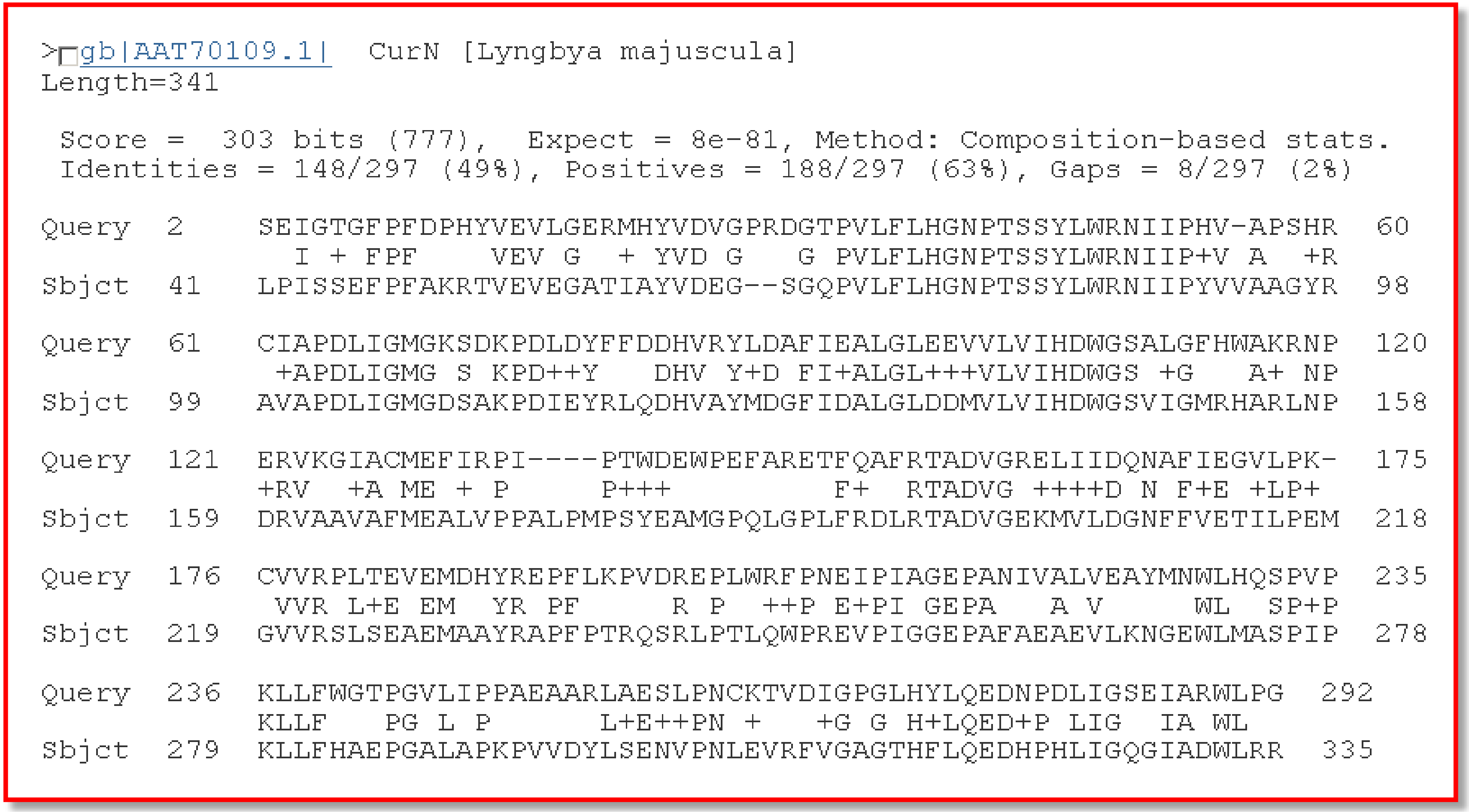
Sequence-based searches

- BLAST alignment
  - identity and similarity level between query and aligned sequence
  - alignment length and coverage of query sequence - the alignment is local, therefore one should always check that the alignment covers a significant portion of the query sequence (e.g., the alignment may involve only few amino acids from the query sequence → not significant hit)

Sequence-based searches

- PSI-BLAST results

alignment



```
>☐gb|AAT70109.1|  CurN [Lyngbya majuscula]
Length=341

 Score =  303 bits (777),  Expect = 8e-81, Method: Composition-based stats.
 Identities = 148/297 (49%), Positives = 188/297 (63%), Gaps = 8/297 (2%)

Query  2     SEIGTGFPFDPHYVEVLGERMHYVDVGPRDGTPVLFLHGNPTSSYLWRNIIPHV-APSHR  60
               I + FPF    VEV G   + YVD G   G PVLFLHGNPTSSYLWRNIIP+V A  +R
Sbjct  41    LPISSEFPFAKRTVEVEGATIAYVDEG--SGQPVLFLHGNPTSSYLWRNIIPYVVAAGYR  98

Query  61    CIAPDLIGMGKSDKPDLDYFFDDHVRYLDAFIEALGLEEVVLVIHDWGSALGFHWAKRNP  120
              +APDLIGMG S KPD++Y   DHV Y+D FI+ALGL+++VLVIHDWGS +G   A+ NP
Sbjct  99    AVAPDLIGMGDSAKPDIEYRLQDHVAYMDGFIDALGLDDMVLVIHDWGSVIGMRHARLNP  158

Query  121   ERVKGIACMEFIRPI----PTWDEWPEFARETFQAFRTADVGRELIIDQNAFIEGVLPK-  175
             +RV  +A ME + P     P+++         F+  RTADVG ++++D N F+E +LP+
Sbjct  159   DRVAAVAFMEALVPPALPMPSYEAMGPQLGPLFRDLRTADVGEKMVLDGNFFVETILPEM  218

Query  176   CVVRPLTEVEMDHYREPFLKPVDREPLWRFPNEIPIAGEPANIVALVEAYMNWLHQSPVP  235
              VVR L+E EM  YR PF     R P  ++P E+PI GEPA   A V     WL  SP+P
Sbjct  219   GVVRSLSEAEMAAYRAPFPTRQSRLPTLQWPREVPIGGEPAFAEAEVLKNGEWLMASPIP  278

Query  236   KLLFWGTPGVLIPPAEAARLAESLPNCKTVDIGPGLHYLQEDNPDLIGSEIARWLPG  292
             KLLF   PG L P      L+E++PN +   +G G H+LQED+P LIG  IA WL
Sbjct  279   KLLFHAEPGALAPKPVVDYLSENVPNLEVRFVGAGTHFLQEDHPHLIGQGIADWLRR  335
```

Optimal search strategy

- text-based search
  - good for finding evolutionary “unrelated” proteins with some specific function
  - a large number of false negatives (missed proteins with target function) and false positives (identified proteins with different function) results due to erroneous or inaccurate annotations

Optimal search strategy

- ❑ text-based search
- ❑ sequence-based search
  - good for finding members of a protein family (i.e., group of evolutionary related proteins sharing some specific function) → not suitable for finding "unrelated" proteins
  - potential false positive results (i.e., proteins belonging to other evolutionary related families)
  - searches using protein sequence queries are generally more sensitive than using nucleotide sequence queries (20 different amino acids vs. 4 different nucleotides)

Optimal search strategy

- ❑ text-based search
- ❑ sequence-based search
- ❑ combination of text-based and sequence-based approaches

  1. text-based search
  2. subdivision of identified sequences into evolutionarily related groups
  3. selection of a few representatives for each group
  4. sequence-based searches using each representative as a query

  - ▪ potential false positive results – should be filtered

How to recognize interesting sequences?

- sequence clustering
- sequence comparison
- information about host organisms
- automated *in silico* enzyme identification
- reconstruction of ancestral proteins

Sequence clustering

- Clustering based on pairwise sequence similarities
  - can be used for a fast and rough classification of sequences in large datasets (thousands of sequences)

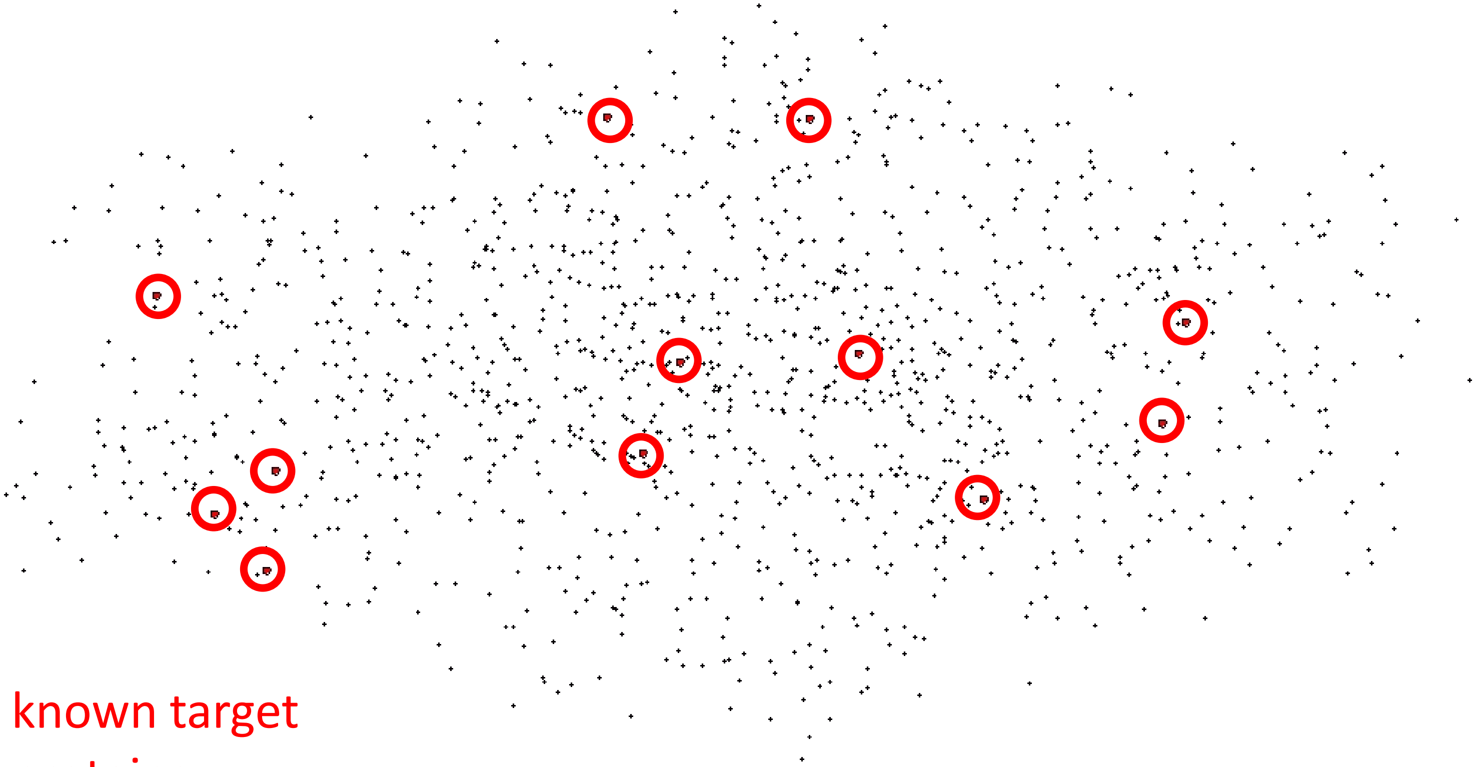
    → effective way to filter results of database searches

    → identification of members of individual protein families

- Tools:
  - CLANS - visualization of pairwise sequence similarities in three-dimensional space → overview of sequence space (https://toolkit.tuebingen.mpg.de/tools/clans)
  - CD-HIT - clustering and comparison of protein or nucleotide sequences (https://sites.google.com/view/cd-hit/)

Sequence clustering

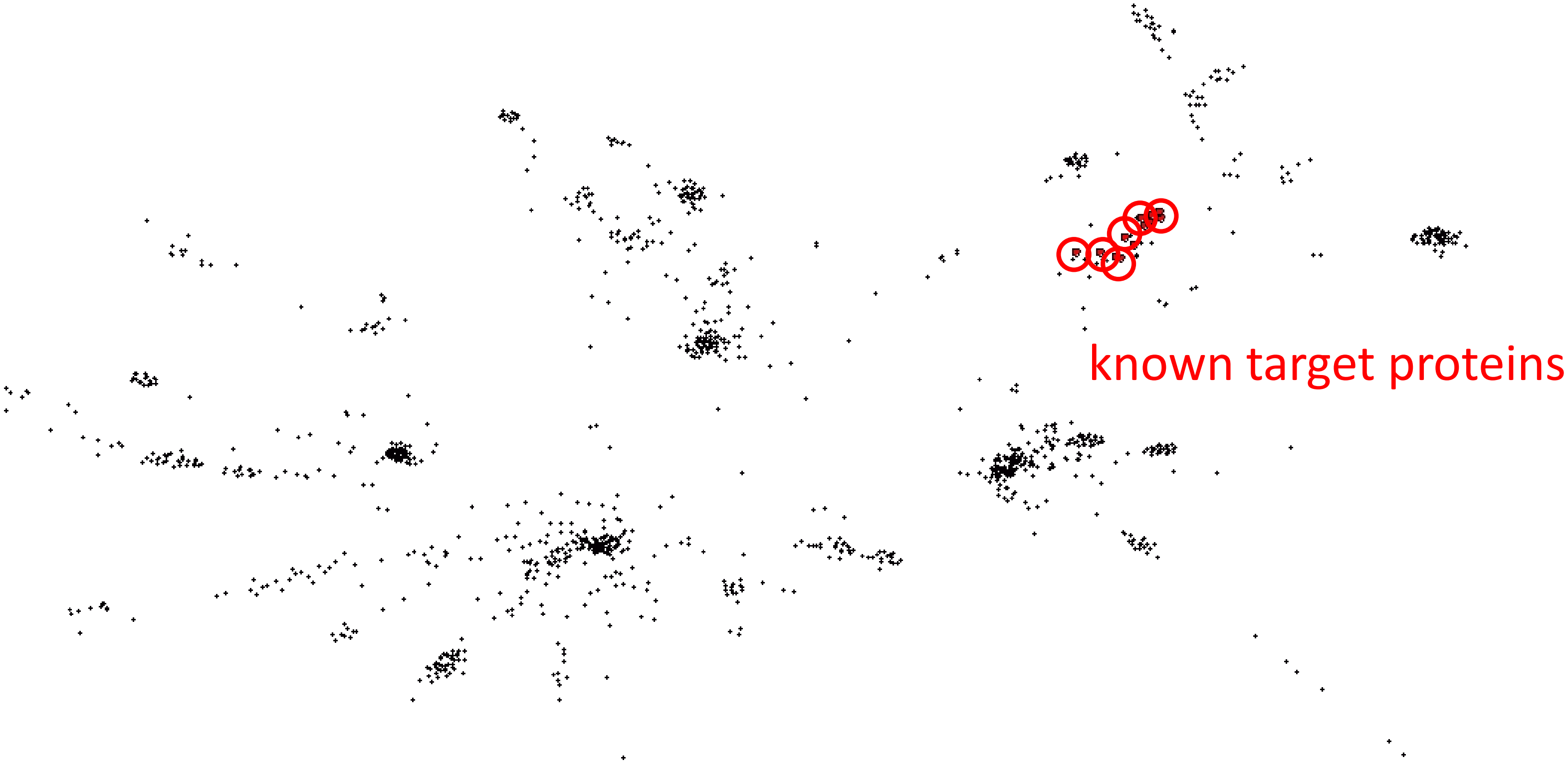
- Clustering based on pairwise sequence similarities



known target
proteins

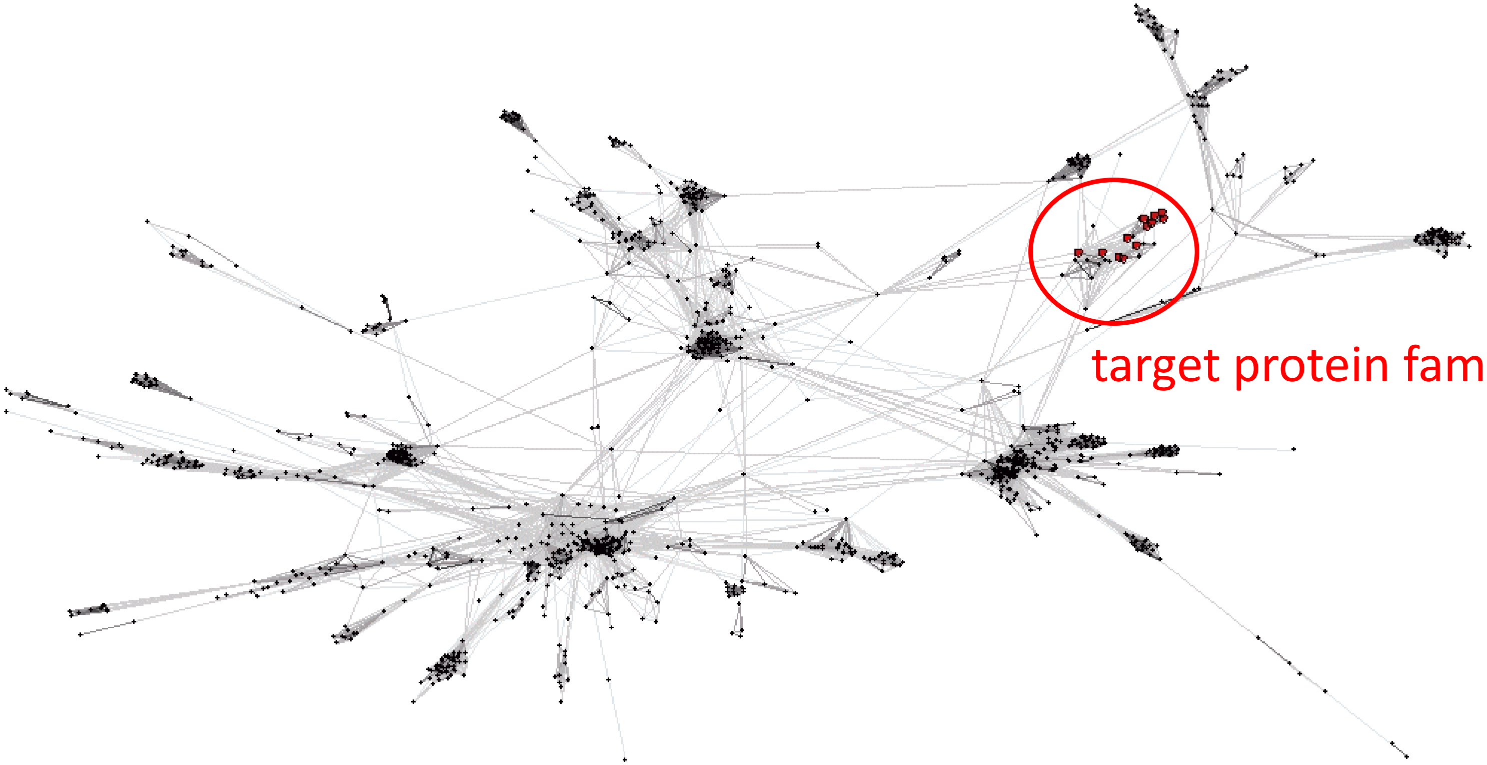
Sequence clustering

- Clustering based on pairwise sequence similarities



Sequence clustering

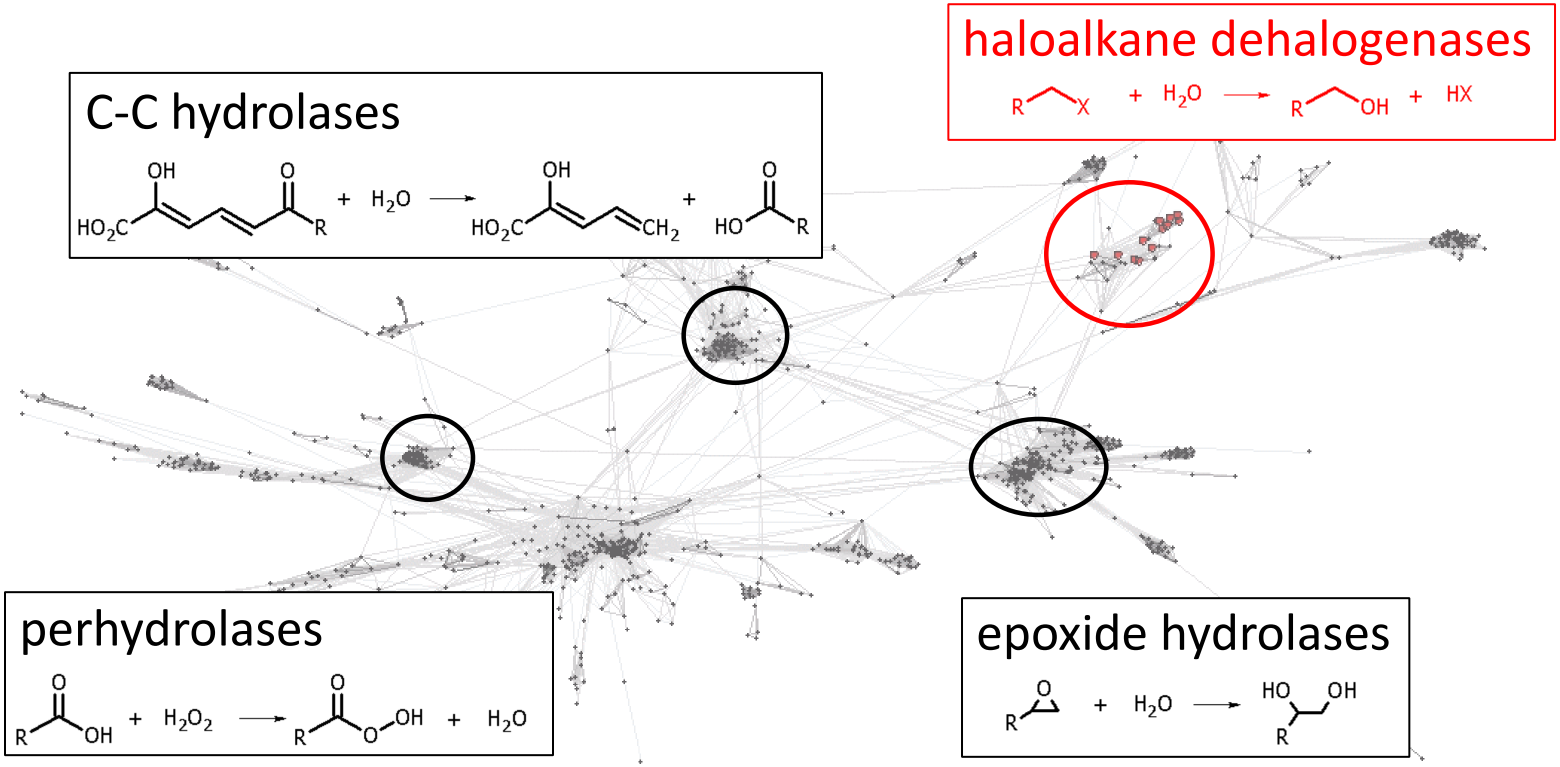
- Clustering based on pairwise sequence similarities



target protein family

Sequence clustering

- ❑ Clustering based on pairwise sequence similarities



C-C hydrolases

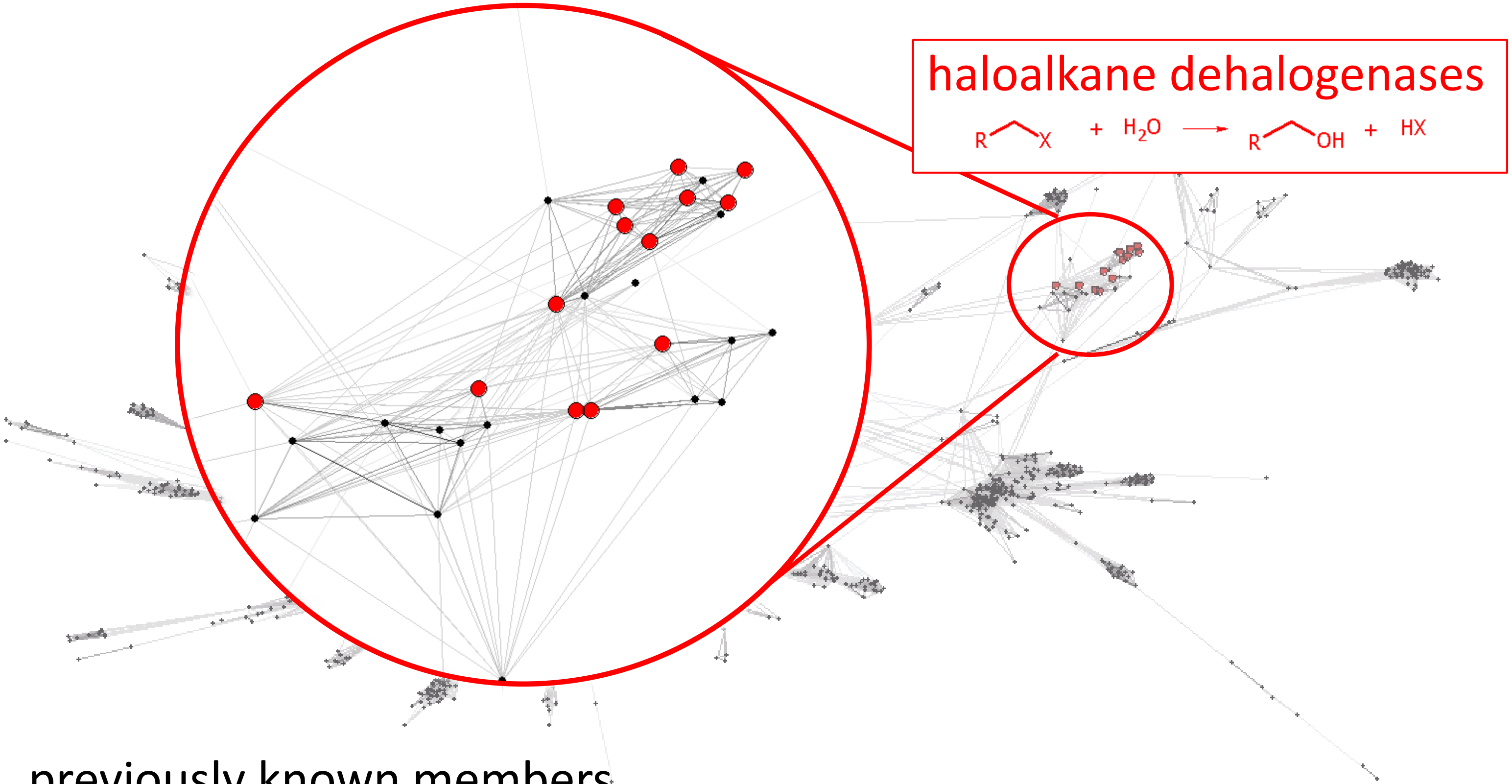
haloalkane dehalogenases

perhydrolases

epoxide hydrolases

Sequence clustering

❑ Clustering based on pairwise sequence similarities

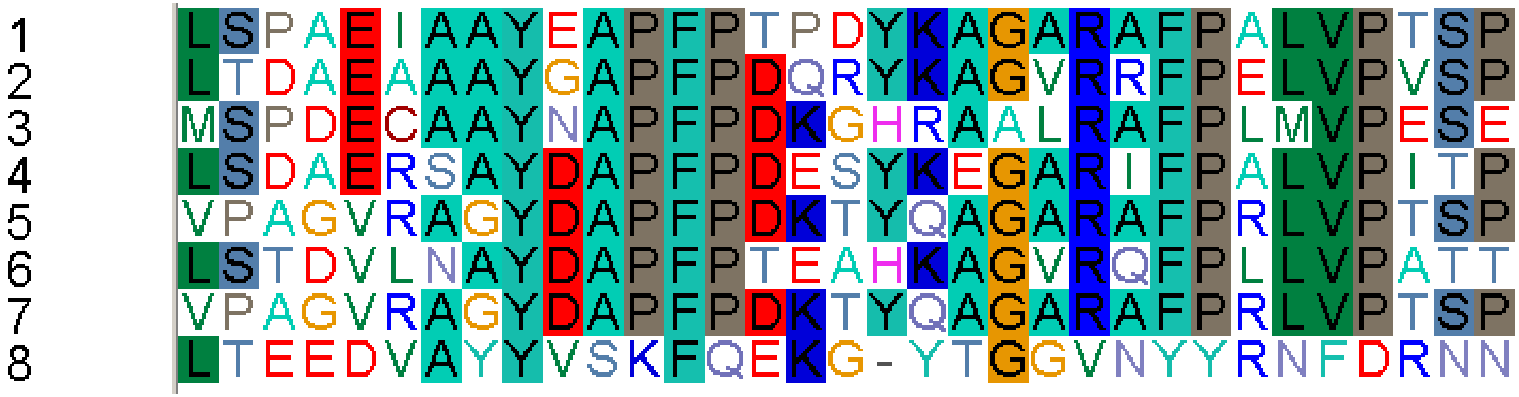


● previously known members

● new family members

Sequence comparison

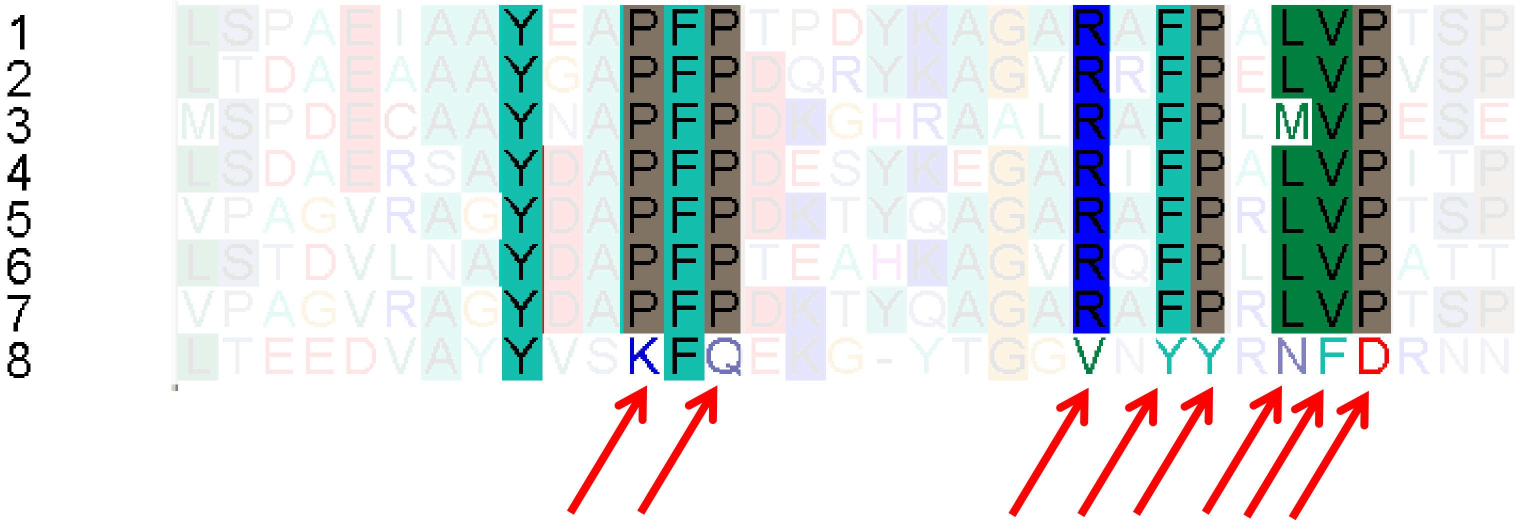
- multiple sequence alignment
  - analysis of conserved residues within protein family → identification of protein family members



```
1   LSPAEIAAYEAPFPTPDYKAGARAFPALVPTSP
2   LTDAEAAAYGAPFPDQRYKAGVRRFPELVPVSP
3   MSPDECAAYNAPFPDKGHRAALRAFPLMVPESE
4   LSDAERSAYDAPFPDESYKEGARIFPALVPITP
5   VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
6   LSTDVLNAYDAPFPTEAHKAGVRQFPLLVPATT
7   VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
8   LTEEDVAYYVSKFQEKG-YTGGVNYYRNFDRNN
```

Sequence comparison

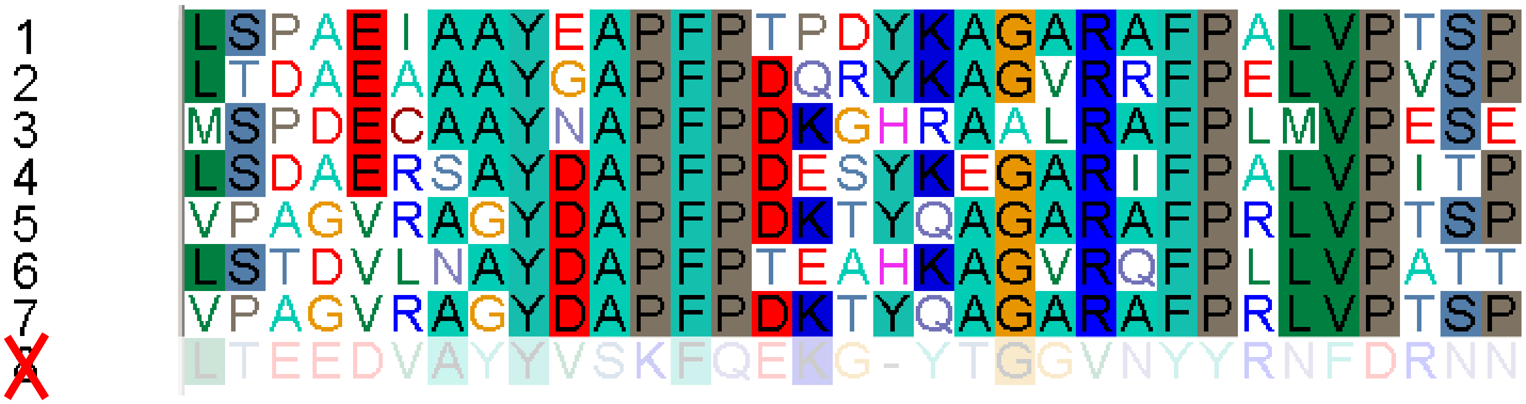
- multiple sequence alignment
  - analysis of conserved residues within protein family → identification of protein family members



```
1  LSPAEIAAYEAPFPTPDYKAGARAFPALVPTSP
2  LTDAEAAAYGAPFPDQRYKAGVRRFPELVPVSP
3  MSPDECAAYNAPFPDKGHRAALRAFPLMVPESE
4  LSDAERSAYDAPFPDESYKEGARIFPALVPITP
5  VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
6  LSTDVLNAYDAPFPTEAHKAGVRQFPLLVPATT
7  VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
8  LTEEDVAYYVSKFQEKG-YTGGVNYYRNFDRNN
```

Sequence comparison

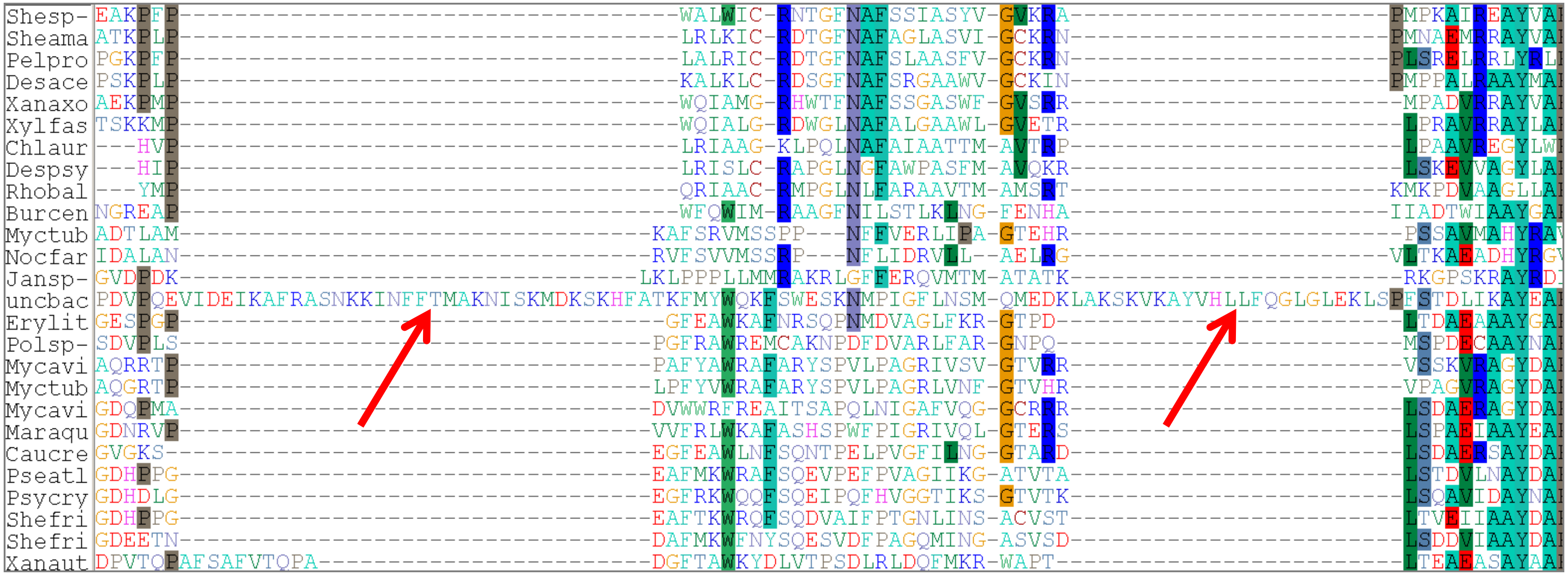
- multiple sequence alignment
  - analysis of conserved residues within protein family → identification of protein family members



```
1 LSPAEIAAYEAPFPTPDYKAGARAFPALVPTSP
2 LTDAEAAAYGAPFPDQRYKAGVRRFPELVPVSP
3 MSPDECAAYNAPFPDKGHRAALRAFPLMVPESE
4 LSDAERSAYDAPFPDESYKEGARIFPALVPITP
5 VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
6 LSTDVLNAYDAPFPTEAHKAGVRQFPLLVPATT
7 VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
8 LTEEDVAYYVSKFQEKG-YTGGVNYYRNFDRNN
```

Sequence comparison

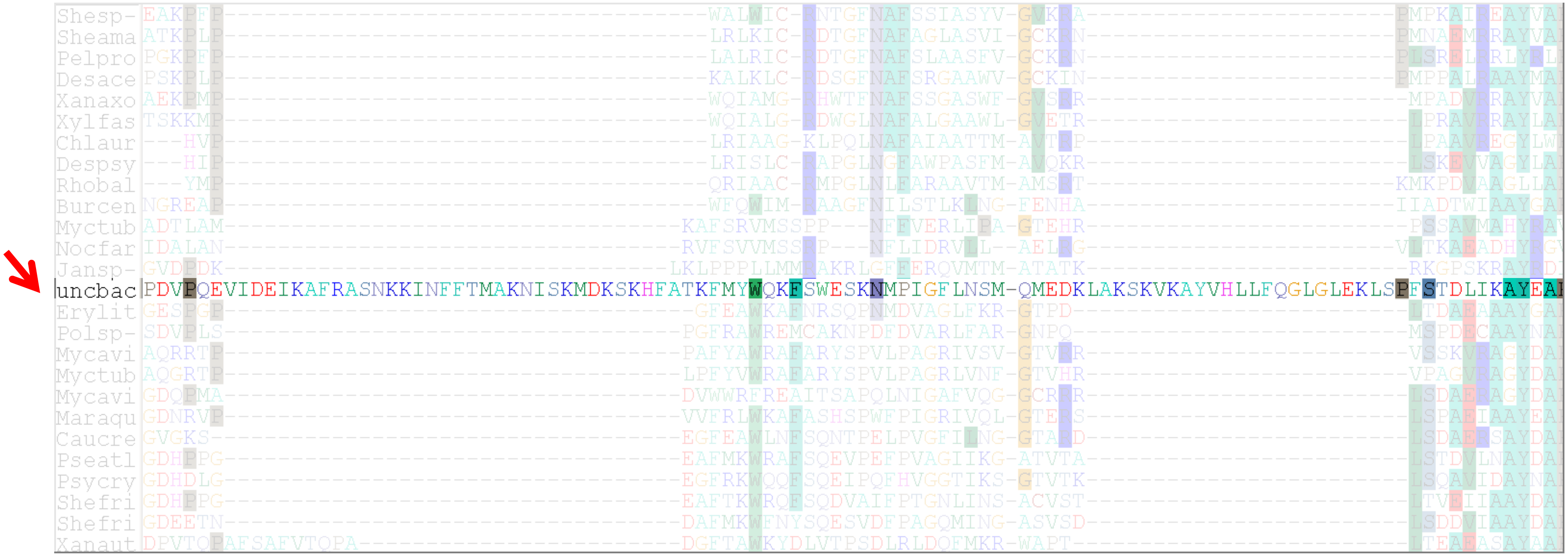
- multiple sequence alignment
  - identification of sequences with unique features → proteins with potentially novel characteristics or problematic for production



```
Shesp- EAKPFP------------------------------------WALWIC-RNTGFNAFSSIASYV-GVKRA-----------------------PMPKAIREAYVA
Sheama ATKPLP------------------------------------LRLKIC-RDTGFNAFAGLASVI-GCKRN-----------------------PMNAEMRRAYVA
Pelpro PGKPFP------------------------------------LALRIC-RDTGFNAFSLAASFV-GCKRN-----------------------PLSRELRRLYRL
Desace PSKPLP------------------------------------KALKLC-RDSGFNAFSRGAAWV-GCKIN-----------------------PMPPALRAAYMA
Xanaxo AEKPMP------------------------------------WQIAMG-RHWTFNAFSSGASWF-GVSRR------------------------MPADVRRAYVA
Xylfas TSKKMP------------------------------------WQIALG-RDWGLNAFALGAAWL-GVETR------------------------LPRAVRRAYLA
Chlaur ---HVP------------------------------------LRIAAG-KLPQLNAFAIAATTM-AVTRP------------------------LPAAVREGYLW
Despsy ---HIP------------------------------------LRISLC-RAPGLNGFAWPASFM-AVQKR------------------------LSKEVVAGYLA
Rhobal ---YMP------------------------------------QRIAAC-RMPGLNLFARAAVTM-AMSRT-----------------------KMKPDVAAGLLA
Burcen NGREAP------------------------------------WFQWIM-RAAGFNILSTLKLNG-FENHA-----------------------IIADTWIAAYGA
Myctub ADTLAM----------------------------------KAFSRVMSSPP---NFFVERLIPA-GTEHR------------------------PSSAVMAHYRA
Nocfar IDALAN----------------------------------RVFSVVMSSRP---NFLIDRVLL--AELRG-----------------------VLTKAEADHYRG
Jansp- GVDPDK---------------------------------LKLPPPLLMMRAKRLGFFERQVMTM-ATATK------------------------RKGPSKRAYRD
uncbac PDVPQEVIDEIKAFRASNKKINFFTMAKNISKMDKSKHFATKFMYWQKFSWESKNMPIGFLNSM-QMEDKLAKSKVKAYVHLLFQGLGLEKLSPFSTDLIKAYEA
Erylit GESPGP-----------------------------------GFEAWKAFNRSQPNMDVAGLFKR-GTPD-------------------------LTDAEAAAYGA
Polsp- SDVPLS----------------------------------PGFRAWREMCAKNPDFDVARLFAR-GNPQ-------------------------MSPDECAAYNA
Mycavi AQRRTP----------------------------------PAFYAWRAFARYSPVLPAGRIVSV-GTVRR------------------------VSSKVRAGYDA
Myctub AQGRTP----------------------------------LPFYVWRAFARYSPVLPAGRLVNF-GTVHR------------------------VPAGVRAGYDA
Mycavi GDQPMA----------------------------------DVWWRFREAITSAPQLNIGAFVQG-GCRRR------------------------LSDAERAGYDA
Maraqu GDNRVP----------------------------------VVFRLWKAFASHSPWFPIGRIVQL-GTERS------------------------LSPAEIAAYEA
Caucre GVGKS-----------------------------------EGFEAWLNFSQNTPELPVGFILNG-GTARD------------------------LSDAERSAYDA
Pseatl GDHPPG----------------------------------EAFMKWRAFSQEVPEFPVAGIIKG-ATVTA------------------------LSTDVLNAYDA
Psycry GDHDLG----------------------------------EGFRKWQQFSQEIPQFHVGGTIKS-GTVTK------------------------LSQAVIDAYNA
Shefri GDHPPG----------------------------------EAFTKWRQFSQDVAIFPTGNLINS-ACVST------------------------LTVEIIAAYDA
Shefri GDEETN----------------------------------DAFMKWFNYSQESVDFPAGQMING-ASVSD------------------------LSDDVIAAYDA
Xanaut DPVTQPAFSAFVTQPA------------------------DGFTAWKYDLVTPSDLRLDQFMKR-WAPT-------------------------LTEAEASAYAA
```

Sequence comparison

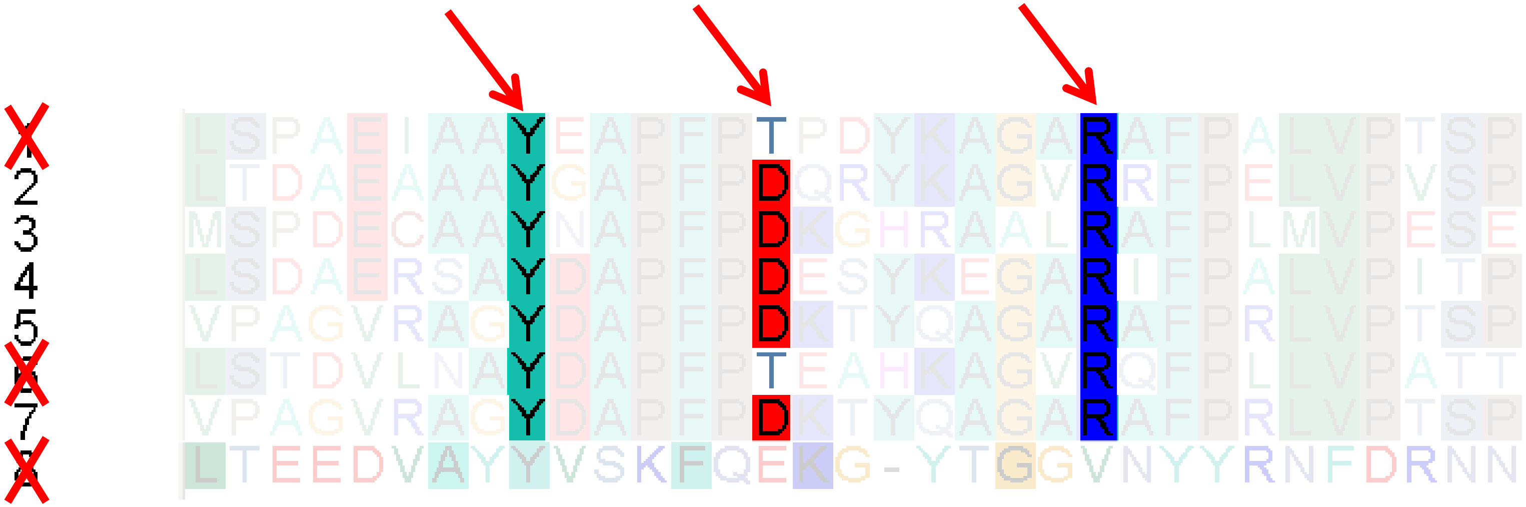
- multiple sequence alignment
  - identification of sequences with unique features → proteins with potentially novel characteristics or problematic for production



```
Shesp- EAKPFP------------------------------------WALWIC-RNTGFNAFSSIASYV-GVKRA-----------------------PMPKAIREAYVAI
Sheama ATKPLP------------------------------------LRLKIC-RDTGFNAFAGLASVI-GCKRN-----------------------PMNAEMRRAYVAI
Pelpro PGKPFP------------------------------------LALRIC-RDTGFNAFSLAASFV-GCKRN-----------------------PLSRELRRLYRLI
Desace PSKPLP------------------------------------KALKLC-RDSGFNAFSRGAAWV-GCKIN-----------------------PMPPALRAAYMAI
Xanaxo AEKPMP------------------------------------WQIAMG-RHWTFNAFSSGASWF-GVSRR------------------------MPADVRRAYVAI
Xylfas TSKKMP------------------------------------WQIALG-RDWGLNAFALGAAWL-GVETR------------------------LPRAVRRAYLAI
Chlaur ---HVP------------------------------------LRIAAG-KLPQLNAFAIAATTM-AVTRP------------------------LPAAVREGYLWI
Despsy ---HIP------------------------------------LRISLC-RAPGLNGFAWPASFM-AVQKR------------------------LSKEVVAGYLAI
Rhobal ---YMP------------------------------------QRIAAC-RMPGLNLFARAAVTM-AMSRT-----------------------KMKPDVAAGLLAI
Burcen NGREAP------------------------------------WFQWIM-RAAGFNILSTLKLNG-FENHA-----------------------IIADTWIAAYGAI
Myctub ADTLAM----------------------------------KAFSRVMSSPP---NFFVERLIPA-GTEHR------------------------PSSAVMAHYRAV
Nocfar IDALAN----------------------------------RVFSVVMSSRP---NFLIDRVLL--AELRG-----------------------VLTKAEADHYRGV
Jansp- GVDPDK---------------------------------LKLPPPLLMMRAKRLGFFERQVMTM-ATATK------------------------RKGPSKRAYRDI
uncbac PDVPQEVIDEIKAFRASNKKINFFTMAKNISKMDKSKHFATKFMYWQKFSWESKNMPIGFLNSM-QMEDKLAKSKVKAYVHLLFQGLGLEKLSPFSTDLIKAYEAI
Erylit GESPGP-----------------------------------GFEAWKAFNRSQPNMDVAGLFKR-GTPD-------------------------LTDAEAAAYGAI
Polsp- SDVPLS----------------------------------PGFRAWREMCAKNPDFDVARLFAR-GNPQ-------------------------MSPDECAAYNAI
Mycavi AQRRTP----------------------------------PAFYAWRAFARYSPVLPAGRIVSV-GTVRR------------------------VSSKVRAGYDAI
Myctub AQGRTP----------------------------------LPFYVWRAFARYSPVLPAGRLVNF-GTVHR------------------------VPAGVRAGYDAI
Mycavi GDQPMA----------------------------------DVWWRFREAITSAPQLNIGAFVQG-GCRRR------------------------LSDAERAGYDAI
Maraqu GDNRVP----------------------------------VVFRLWKAFASHSPWFPIGRIVQL-GTERS------------------------LSPAEIAAYEAI
Caucre GVGKS-----------------------------------EGFEAWLNFSQNTPELPVGFILNG-GTARD------------------------LSDAERSAYDAI
Pseatl GDHPPG----------------------------------EAFMKWRAFSQEVPEFPVAGIIKG-ATVTA------------------------LSTDVLNAYDAI
Psycry GDHDLG----------------------------------EGFRKWQQFSQEIPQFHVGGTIKS-GTVTK------------------------LSQAVIDAYNAI
Shefri GDHPPG----------------------------------EAFTKWRQFSQDVAIFPTGNLINS-ACVST------------------------LTVEIIAAYDAI
Shefri GDEETN----------------------------------DAFMKWFNYSQESVDFPAGQMING-ASVSD------------------------LSDDVIAAYDAI
Xanaut DPVTQPAFSAFVTQPA------------------------DGFTAWKYDLVTPSDLRLDQFMKR-WAPT-------------------------LTEAEASAYAAI
```

Sequence comparison

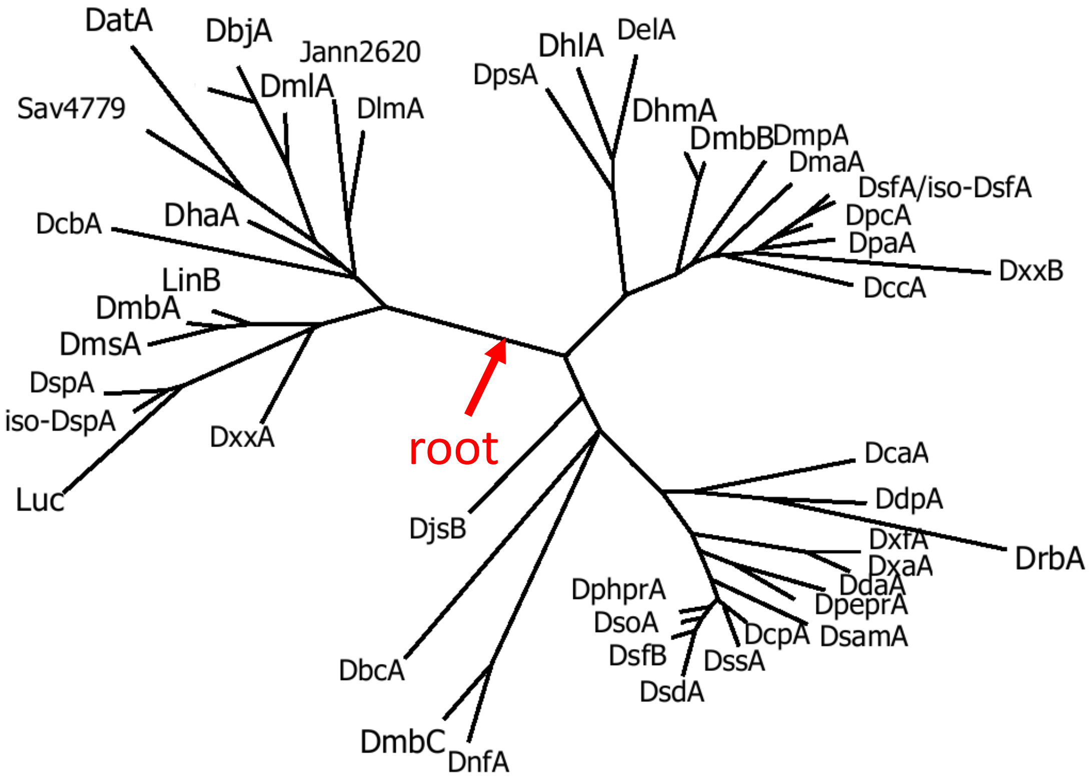
- ❑ multiple sequence alignment
- ❑ essential residues
  - analysis of conserved residues important for function (catalytic, binding, coordinating residues, etc.)
  - In UniProt/SwissProt -> Function -> Features -> list of active or binding site residues



```
1  LSPAEIAAYEAPFPTPDYKAGARAFPALVPTSP
2  LTDAEAAAYGAPFPDQRYKAGVRRFPELVPVSP
3  MSPDECAAYNAPFPDKGHRAALRAFPLMVPESE
4  LSDAERSAYDAPFPDESYKEGARIFPALVPITP
5  VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
6  LSTDVLNAYDAPFPTEAHKAGVRQFPLLVPATT
7  VPAGVRAGYDAPFPDKTYQAGARAFPRLVPTSP
8  LTEEDVAYYVSKFQEKG-YTGGVNYYRNFDRNN
```

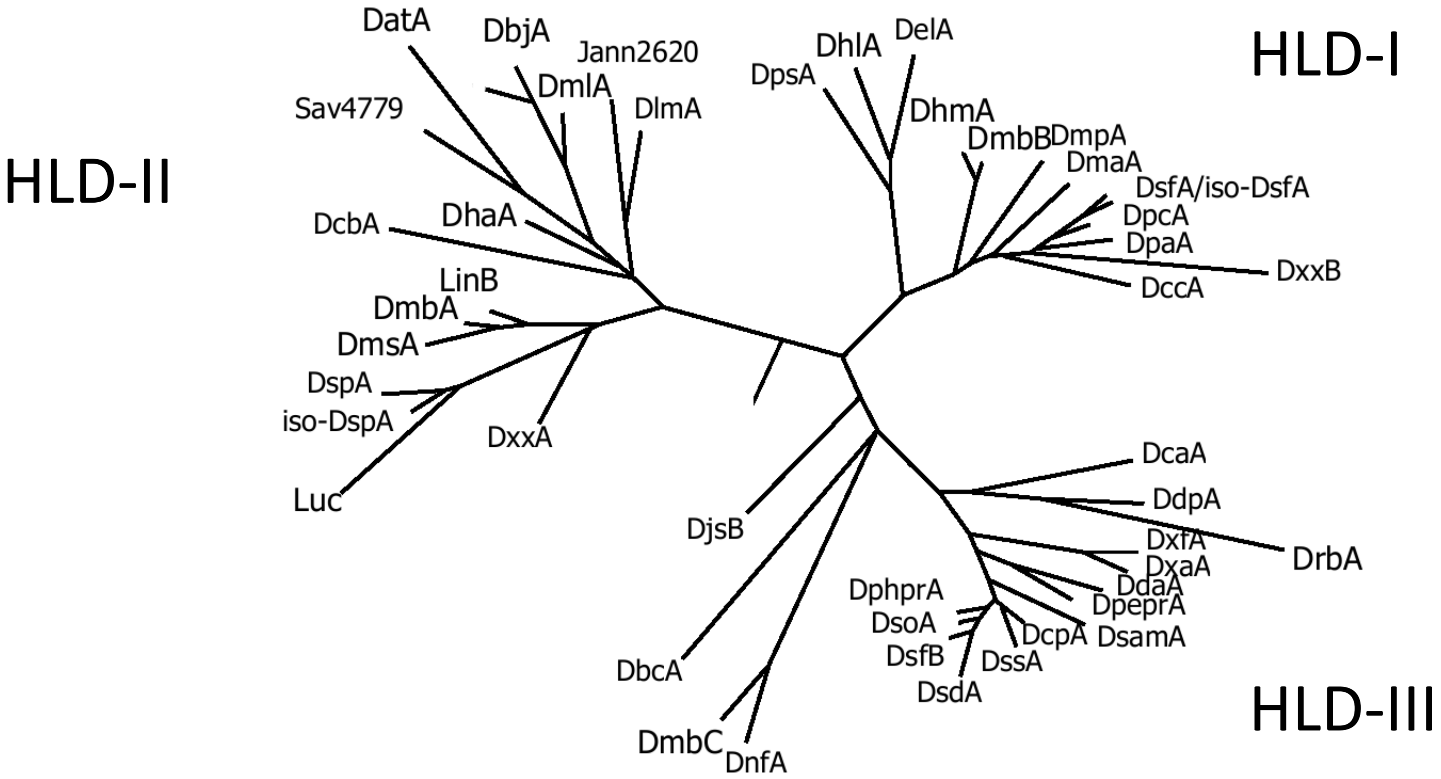
Sequence comparison

- phylogenetics
  - establishment of evolutionary relationships among sequences



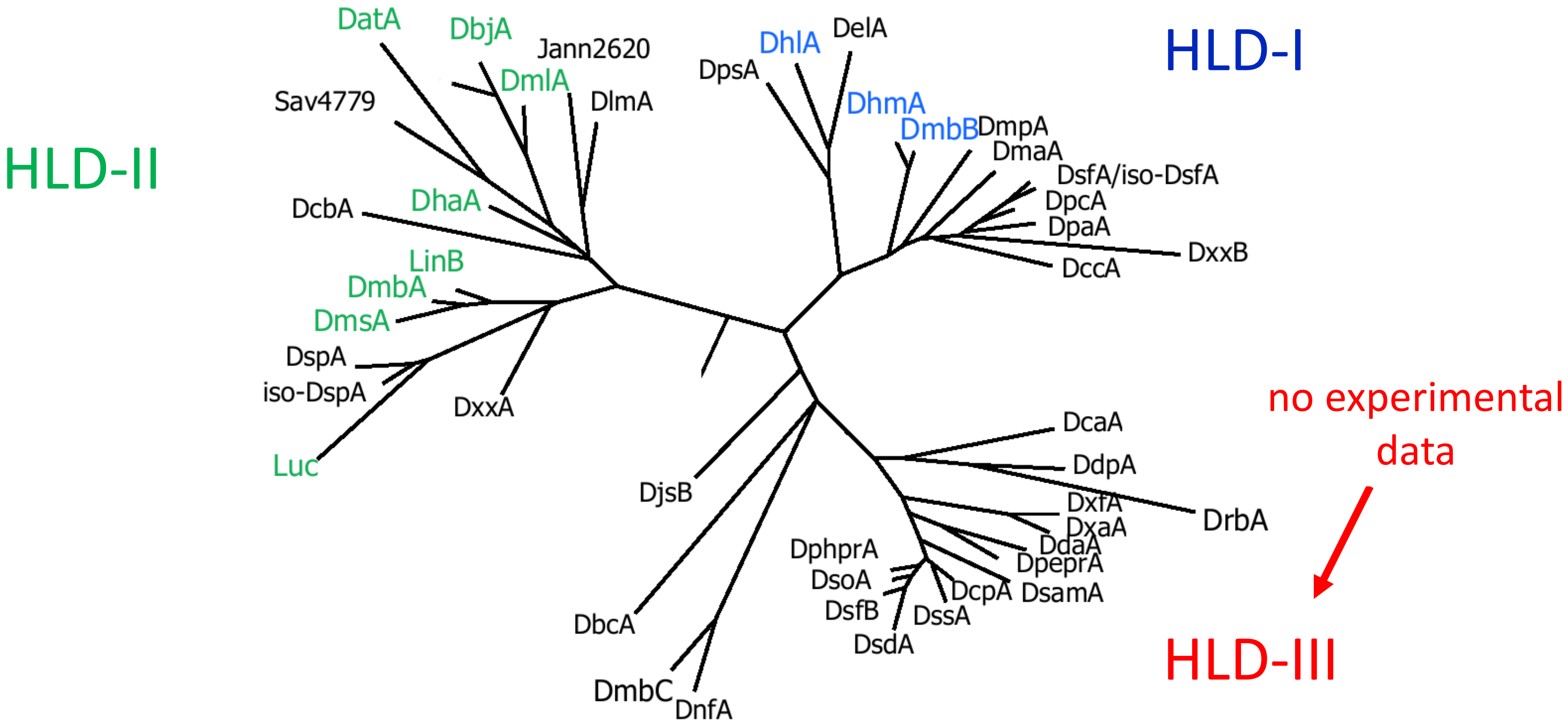
Sequence comparison

- phylogenetics
  - classification of sequences



Sequence comparison

- phylogenetics
  - information about experimental data→ selection of novel proteins



no experimental data

Information about host organisms

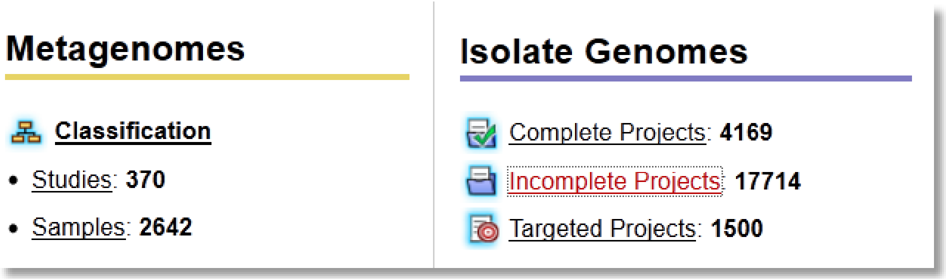
- extremophiles - microorganisms living in extreme conditions
    - geochemical extremes (pH, salinity)
    - physical extremes (temperature, pressure)
- proteins from extremophiles
    - often adapted to extreme conditions → unique characteristics, useful for practical applications

Information about host organisms

- Genomes OnLine Database (GOLD)
  - http://www.genomesonline.org/
  - list of complete (>36,000) and ongoing (> 115,000) information about individual projects and source organisms (Feb 2024)
- Entrez Genome
  - http://www.ncbi.nlm.nih.gov/sites/genome
  - provided by NCBI
  - ~700,000 genome information by organism
  - information about genome, source organism, genes, encoded proteins, graphical representations, ...

Information about host organisms

- GOLD



Metagenomes

Classification

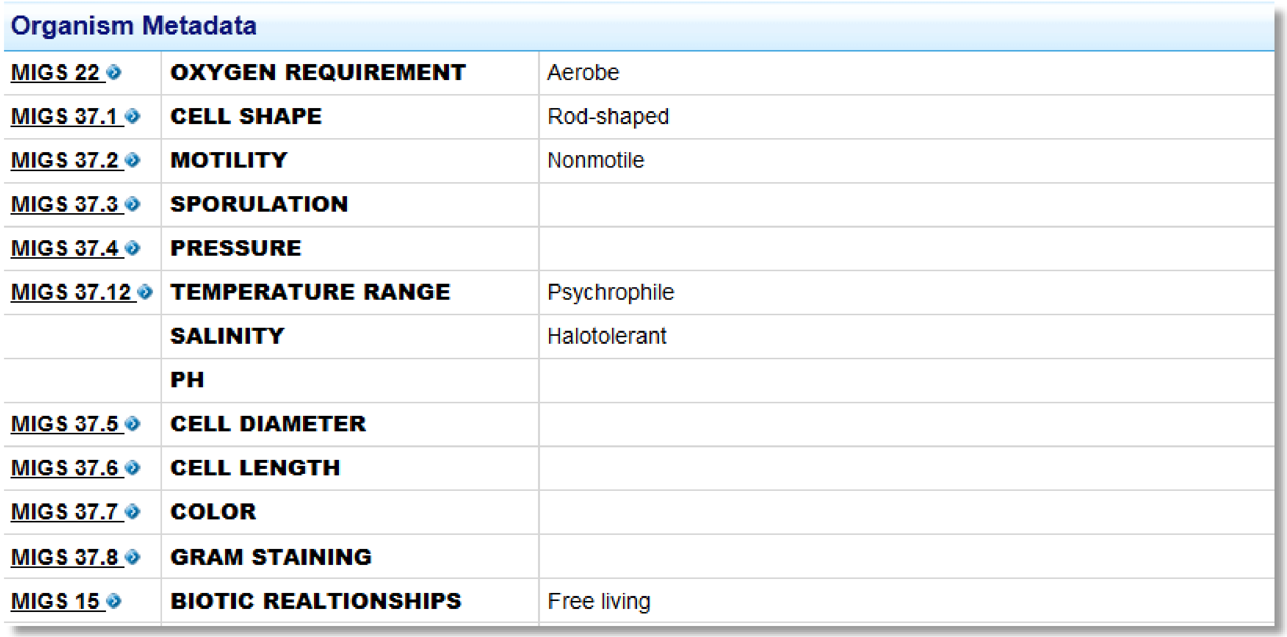
- Studies: 370
- Samples: 2642

Isolate Genomes

Complete Projects: 4169

Incomplete Projects: 17714

Targeted Projects: 1500

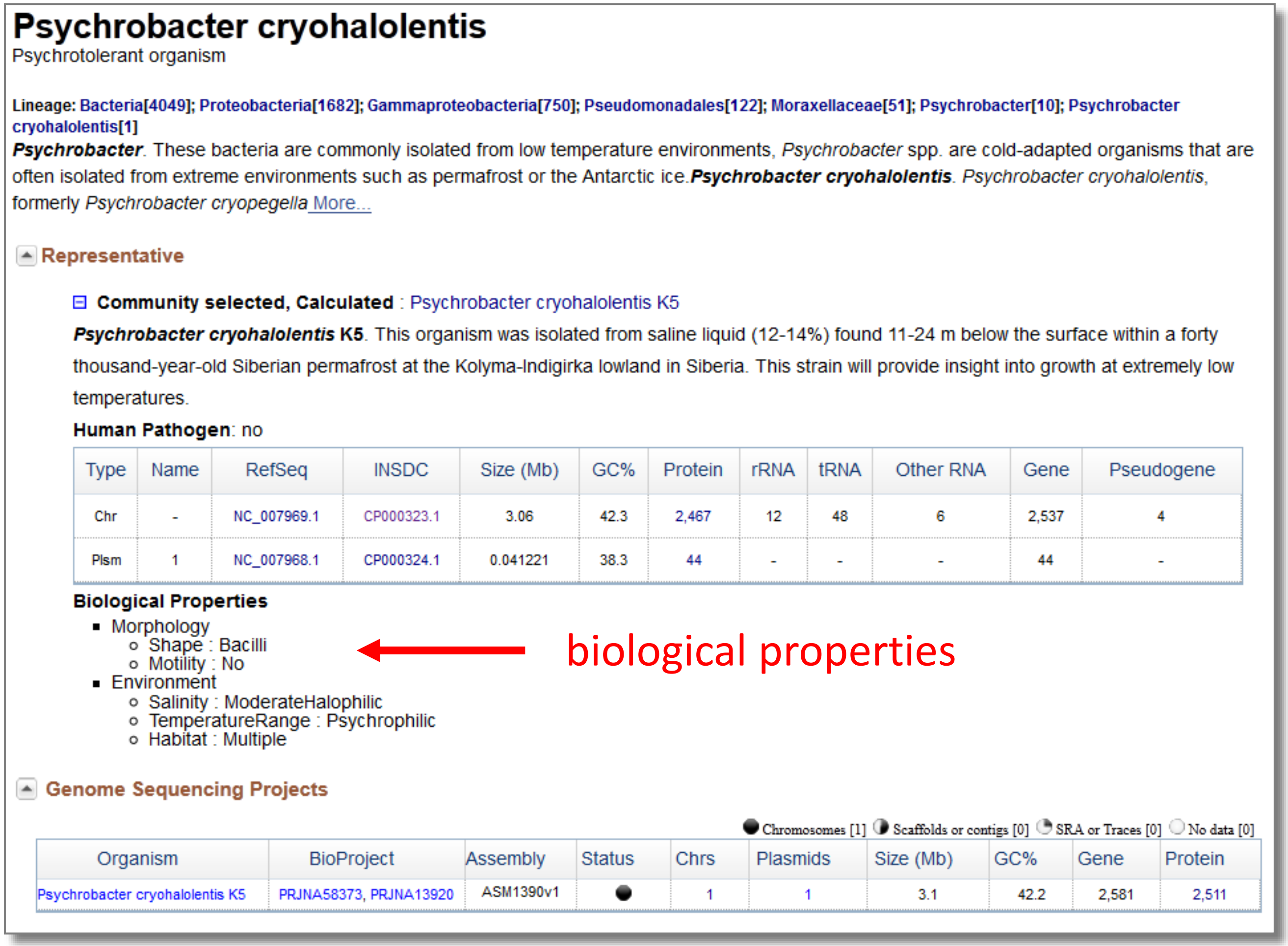


Organism Metadata

MIGS 22	OXYGEN REQUIREMENT	Aerobe
MIGS 37.1	CELL SHAPE	Rod-shaped
MIGS 37.2	MOTILITY	Nonmotile
MIGS 37.3	SPORULATION	
MIGS 37.4	PRESSURE	
MIGS 37.12	TEMPERATURE RANGE	Psychrophile
	SALINITY	Halotolerant
	PH	
MIGS 37.5	CELL DIAMETER	
MIGS 37.6	CELL LENGTH	
MIGS 37.7	COLOR	
MIGS 37.8	GRAM STAINING	
MIGS 15	BIOTIC REALTIONSHIPS	Free living

Information about host organisms

- ❑ Entrez Genome



Psychrobacter cryohalolentis

Psychrotolerant organism

Lineage: Bacteria[4049]; Proteobacteria[1682]; Gammaproteobacteria[750]; Pseudomonadales[122]; Moraxellaceae[51]; Psychrobacter[10]; Psychrobacter cryohalolentis[1]

Psychrobacter. These bacteria are commonly isolated from low temperature environments, *Psychrobacter* spp. are cold-adapted organisms that are often isolated from extreme environments such as permafrost or the Antarctic ice.***Psychrobacter cryohalolentis***. *Psychrobacter cryohalolentis*, formerly *Psychrobacter cryopegella* More...

Representative

⊟ **Community selected, Calculated** : Psychrobacter cryohalolentis K5

***Psychrobacter cryohalolentis* K5**. This organism was isolated from saline liquid (12-14%) found 11-24 m below the surface within a forty thousand-year-old Siberian permafrost at the Kolyma-Indigirka lowland in Siberia. This strain will provide insight into growth at extremely low temperatures.

Human Pathogen: no

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_007969.1	CP000323.1	3.06	42.3	2,467	12	48	6	2,537	4
Plsm	1	NC_007968.1	CP000324.1	0.041221	38.3	44	-	-	-	44	-

Biological Properties

- Morphology
  - Shape : Bacilli
  - Motility : No
- Environment
  - Salinity : ModerateHalophilic
  - TemperatureRange : Psychrophilic
  - Habitat : Multiple

Genome Sequencing Projects

● Chromosomes [1] ◑ Scaffolds or contigs [0] ◔ SRA or Traces [0] ○ No data [0]

Organism	BioProject	Assembly	Status	Chrs	Plasmids	Size (Mb)	GC%	Gene	Protein
Psychrobacter cryohalolentis K5	PRJNA58373, PRJNA13920	ASM1390v1	●	1	1	3.1	42.2	2,581	2,511

biological properties

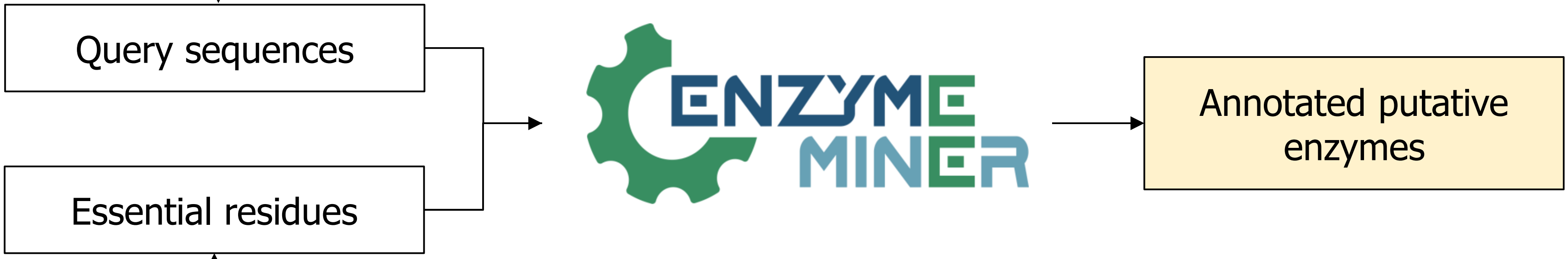
Automated *in silico* enzyme identification

- EnzymeMiner
  - https://loschmidt.chemi.muni.cz/enzymeminer/
  - Search for novel enzymes with particular activity
  - Only input is fasta sequence and essential residues
  - Filtering of sequences using catalytic residues, MSA, clustering
  - Annotations of sequences based on bioinformatics predictions, information available in sequence and genome databases
  - 2D space of sequence similarity network

Automated *in silico* enzyme identification

❑ EnzymeMiner

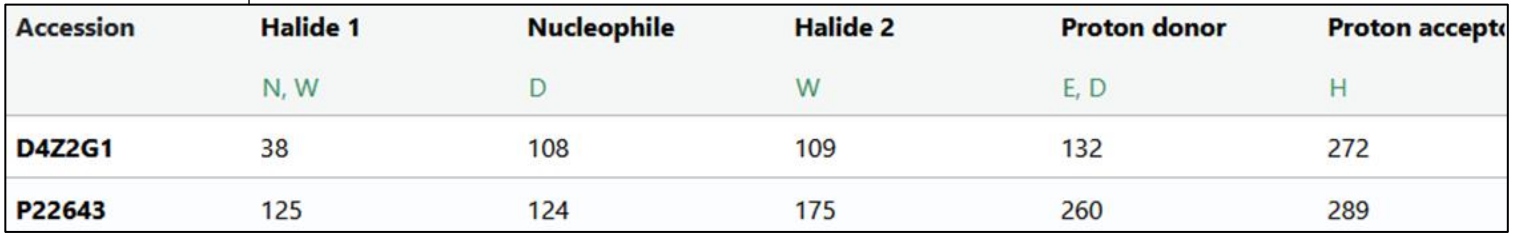
```
> LinB
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLIGMGDSDKLDPSGPERYAYAEHRD
YLDALWEALDLGDRVVLVVHDWGSALGFDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRSQAGEELVLQDNVFVE
QVLPGLILRPLSEAEMAAYREPFLAAGEARRPTLSWPRQIPIAGTPADVVAIARDYAGWLSESPIPKLFINAEPGALTTGRMRDFCRT
WPNQTEITVAGAHFIQEDSPDEIGAAIAAFVRRLRPA
```



Query sequences

Essential residues

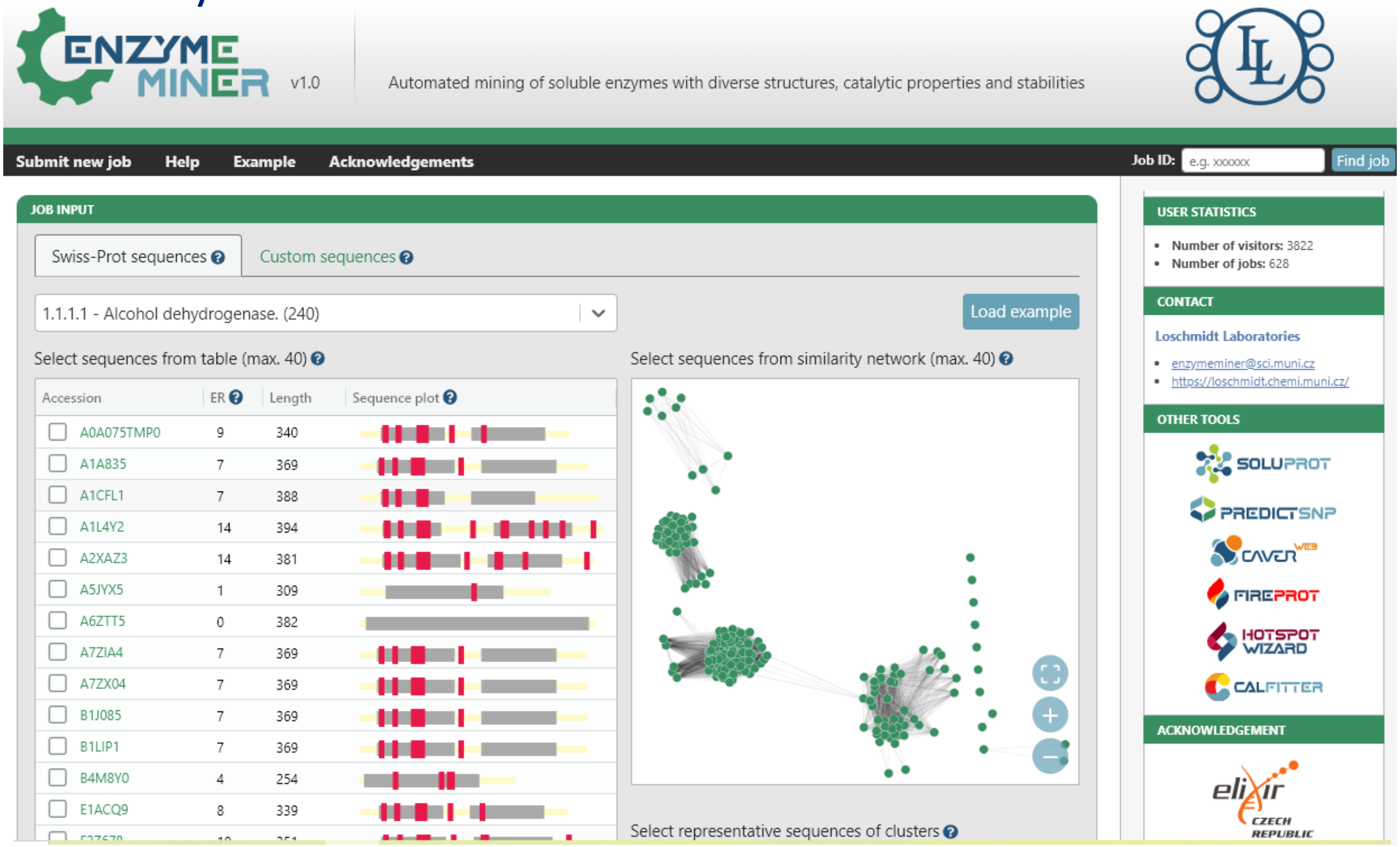
Annotated putative enzymes



Accession	Halide 1	Nucleophile	Halide 2	Proton donor	Proton accept
	N, W	D	W	E, D	H
D4Z2G1	38	108	109	132	272
P22643	125	124	175	260	289

Automated *in silico* enzyme identification

❑ EnzymeMiner



Automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities

JOB INPUT

1.1.1.1 - Alcohol dehydrogenase. (240)

Select sequences from table (max. 40) ?

Accession	ER ?	Length	Sequence plot ?
☐ A0A075TMP0	9	340	
☐ A1A835	7	369	
☐ A1CFL1	7	388	
☐ A1L4Y2	14	394	
☐ A2XAZ3	14	381	
☐ A5JYX5	1	309	
☐ A6ZTT5	0	382	
☐ A7ZIA4	7	369	
☐ A7ZX04	7	369	
☐ B1J085	7	369	
☐ B1LIP1	7	369	
☐ B4M8Y0	4	254	
☐ E1ACQ9	8	339	

Select sequences from similarity network (max. 40) ?

Select representative sequences of clusters ?

USER STATISTICS

- Number of visitors: 3822
- Number of jobs: 628

CONTACT

Loschmidt Laboratories

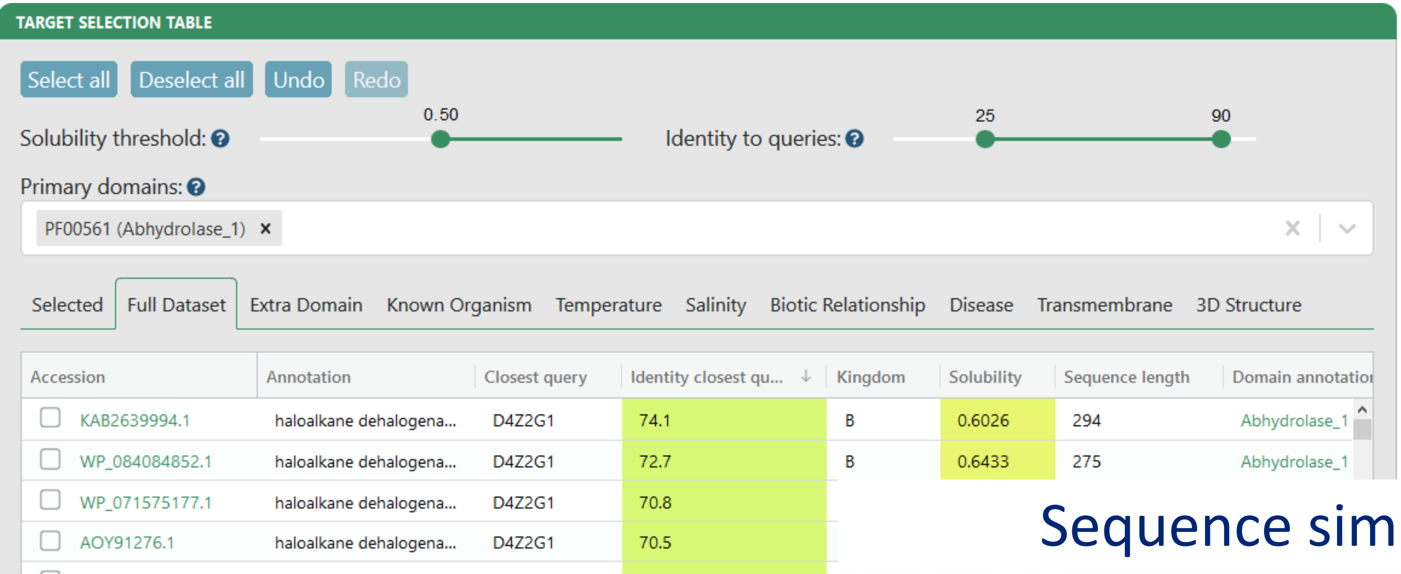
- enzymeminer@sci.muni.cz
- https://loschmidt.chemi.muni.cz/

OTHER TOOLS

ACKNOWLEDGEMENT

Automated *in silico* enzyme identification

Target selection table



TARGET SELECTION TABLE

Select all | Deselect all | Undo | Redo

Solubility threshold: 0.50

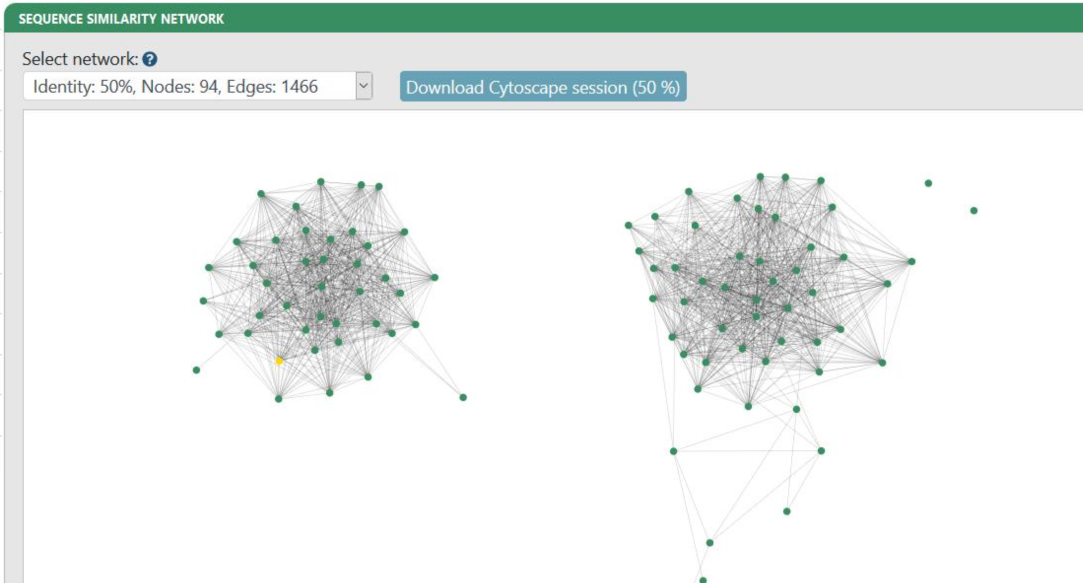
Identity to queries: 25 – 90

Primary domains: PF00561 (Abhydrolase_1)

Selected | Full Dataset | Extra Domain | Known Organism | Temperature | Salinity | Biotic Relationship | Disease | Transmembrane | 3D Structure

	Accession	Annotation	Closest query	Identity closest qu... ↓	Kingdom	Solubility	Sequence length	Domain annotatio
□	KAB2639994.1	haloalkane dehalogena...	D4Z2G1	74.1	B	0.6026	294	Abhydrolase_1
□	WP_084084852.1	haloalkane dehalogena...	D4Z2G1	72.7	B	0.6433	275	Abhydrolase_1
□	WP_071575177.1	haloalkane dehalogena...	D4Z2G1	70.8				
□	AOY91276.1	haloalkane dehalogena...	D4Z2G1	70.5				
□	TMJ55042.1	haloalkane dehalogena...	D4Z2G1	70.3				
□	WP_071068776.1	haloalkane dehalogena...	D4Z2G1	70.2				
□	WP_066929894.1	haloalkane dehalogena...	D4Z2G1	70.1				
□	WP_096502050.1	haloalkane dehalogena...	D4Z2G1	69.9				
□	WP_071011817.1	haloalkane dehalogena...	D4Z2G1	69.8				
□	WP_015306650.1	haloalkane dehalogena...	D4Z2G1	69.8				
□	WP_110315832.1	haloalkane dehalogena...	D4Z2G1	69.8				
□	WP_064949090.1	haloalkane dehalogena...	D4Z2G1	69.7				
□	WP_083164861.1	haloalkane dehalogena...	D4Z2G1	69.7				
□	WP_057374253.1	haloalkane dehalogena...	D4Z2G1	69.6				
□	WP_015290793.1	haloalkane dehalogena...	D4Z2G1	69.6				
□	2QVB_A	Chain A, Crystal Struct...	D4Z2G1	69.5				

Sequence similarity network

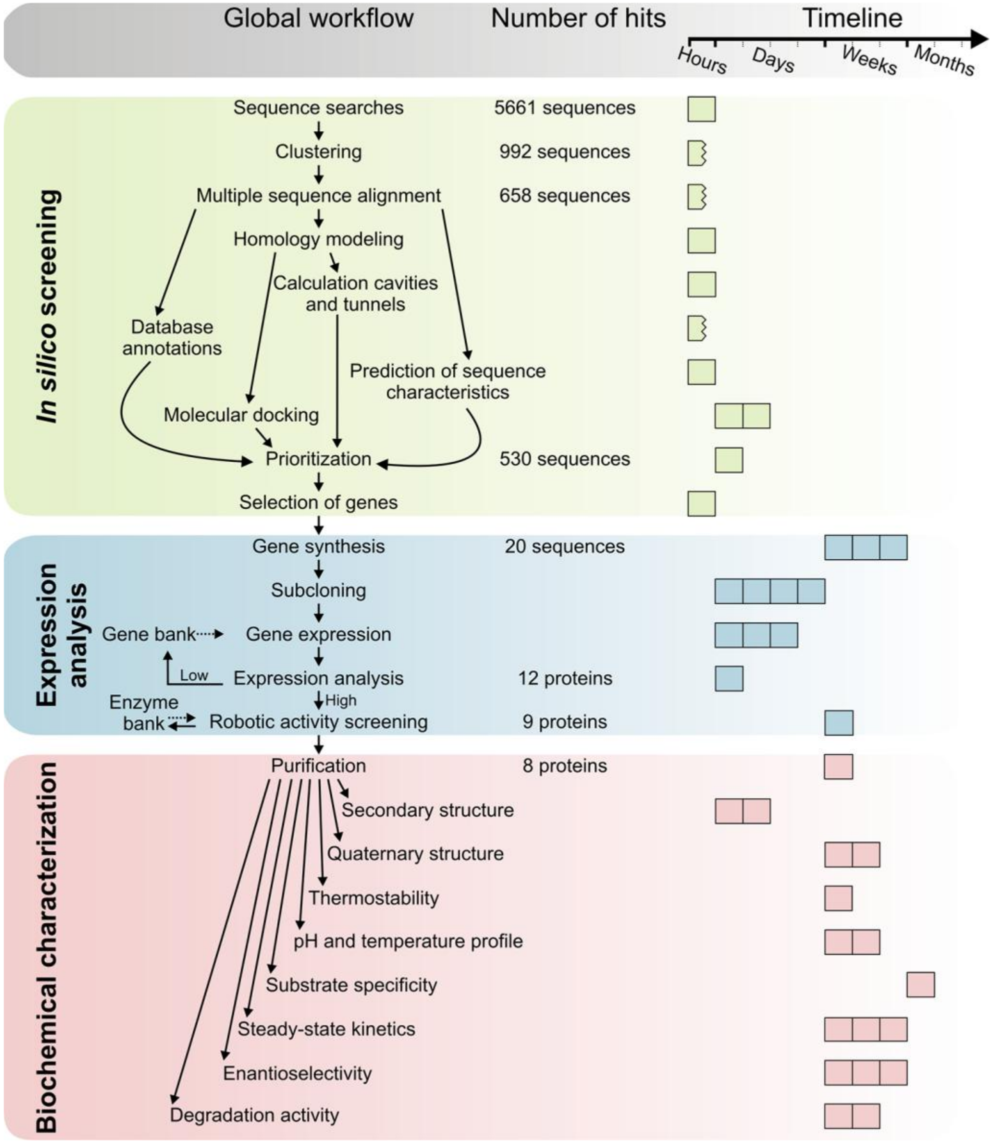


SEQUENCE SIMILARITY NETWORK

Select network: Identity: 50%, Nodes: 94, Edges: 1466

Download Cytoscape session (50 %)

EnzymeMiner use case



Putative dehalogenases

2,905 sequences

Gene synthesis

67 sequences

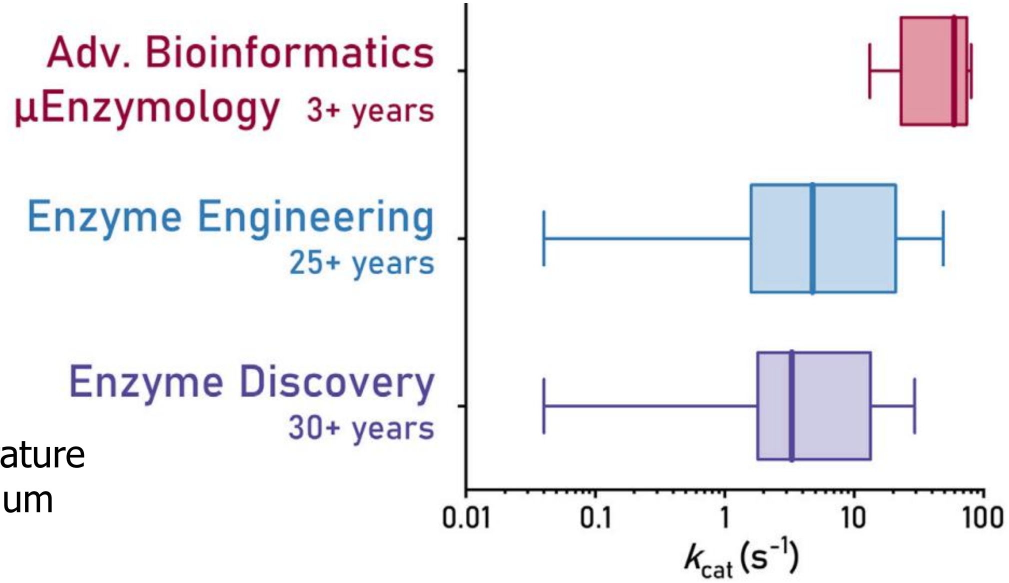
Characterization

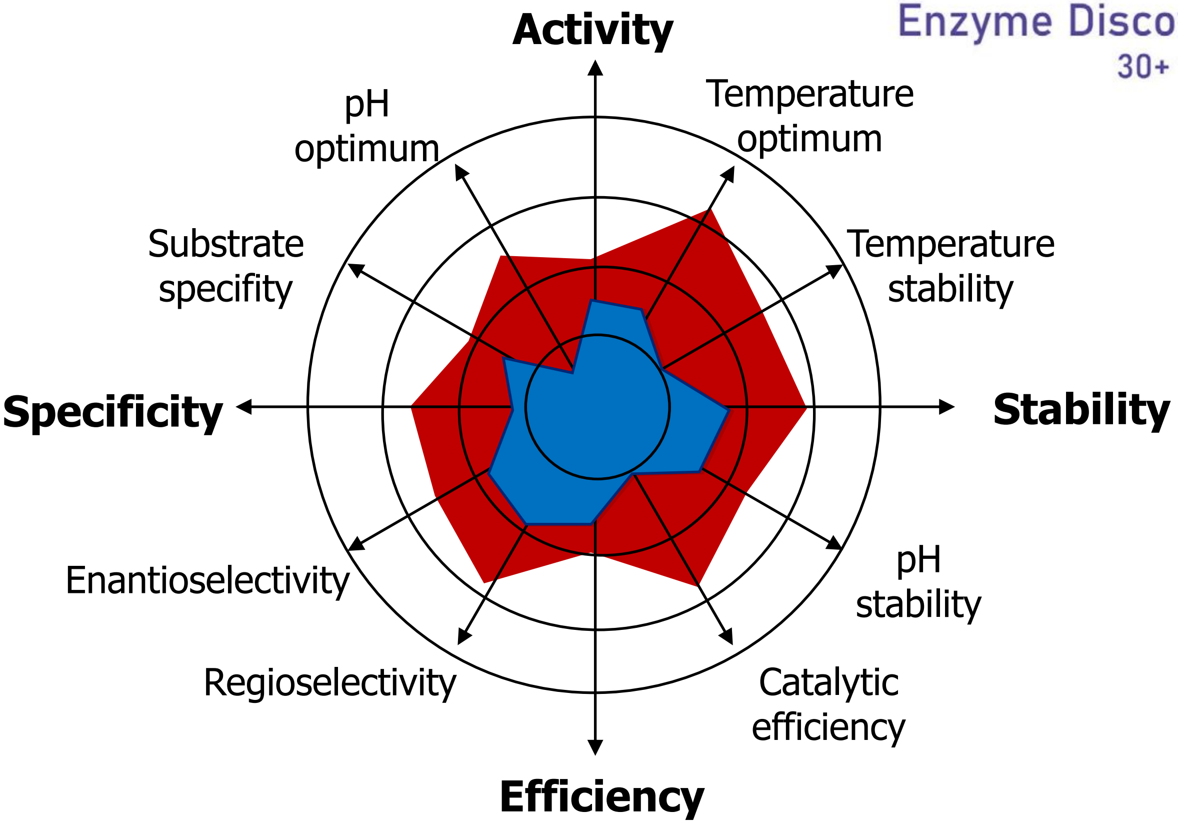
35 proteins

EnzymeMiner use case

The most thermostable

enzyme T_m = 71°C





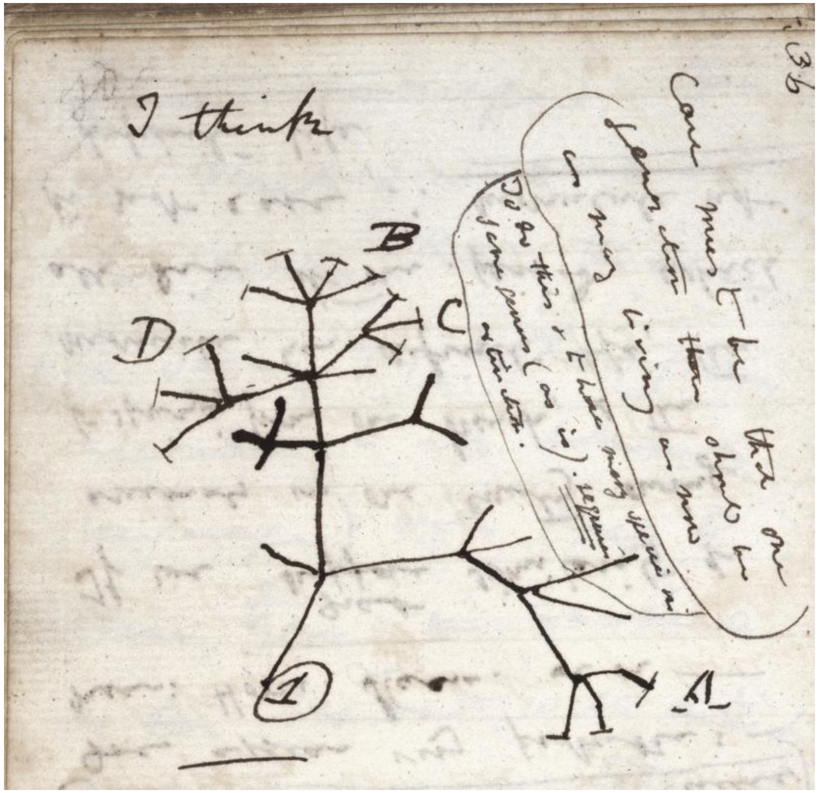
The most catalytically

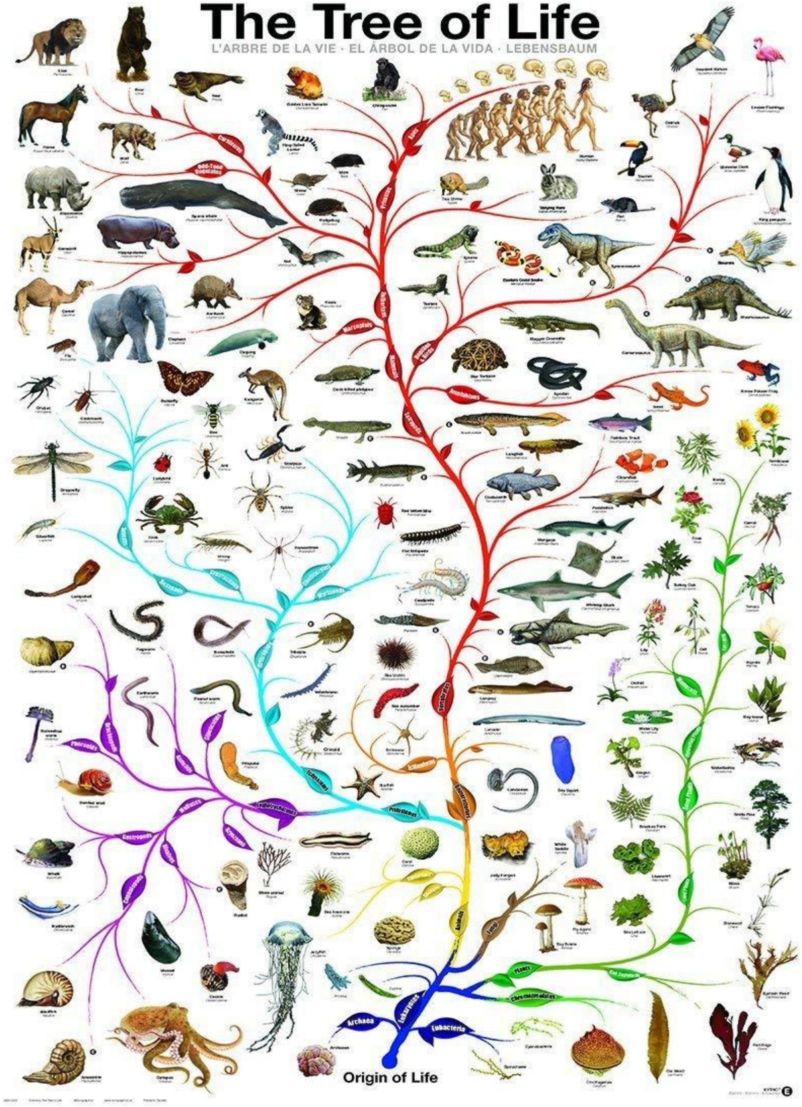
efficient enzyme (100x)

Enzymes active at near-to-

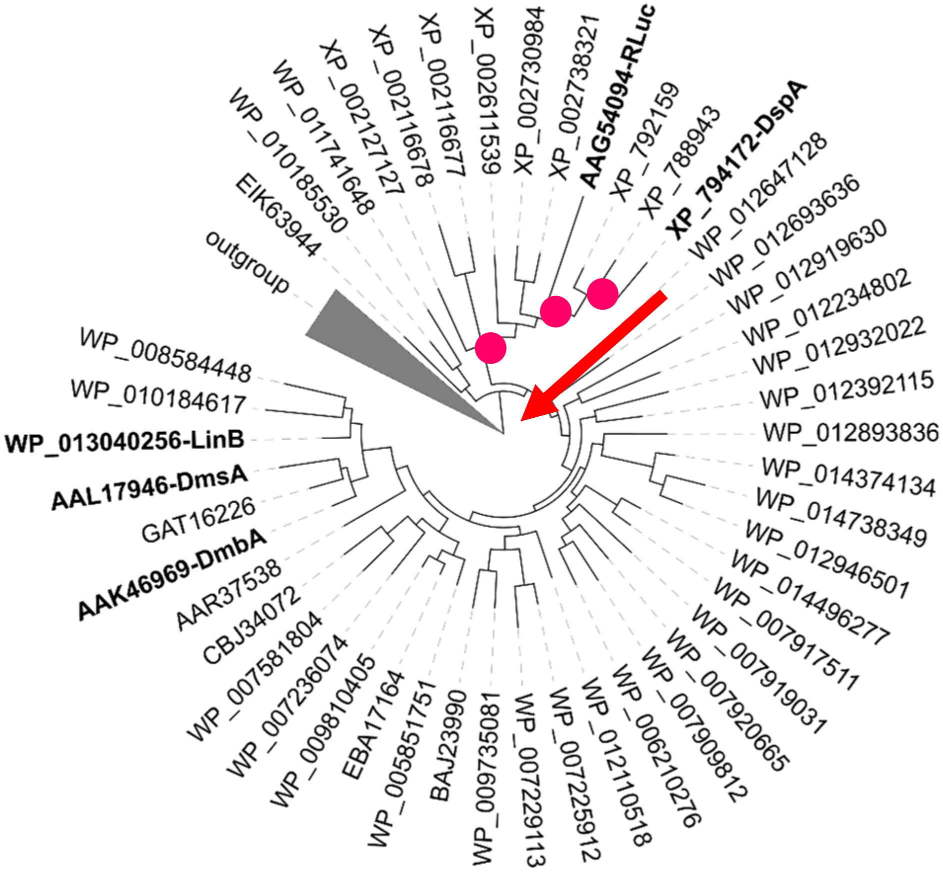
zero temperature

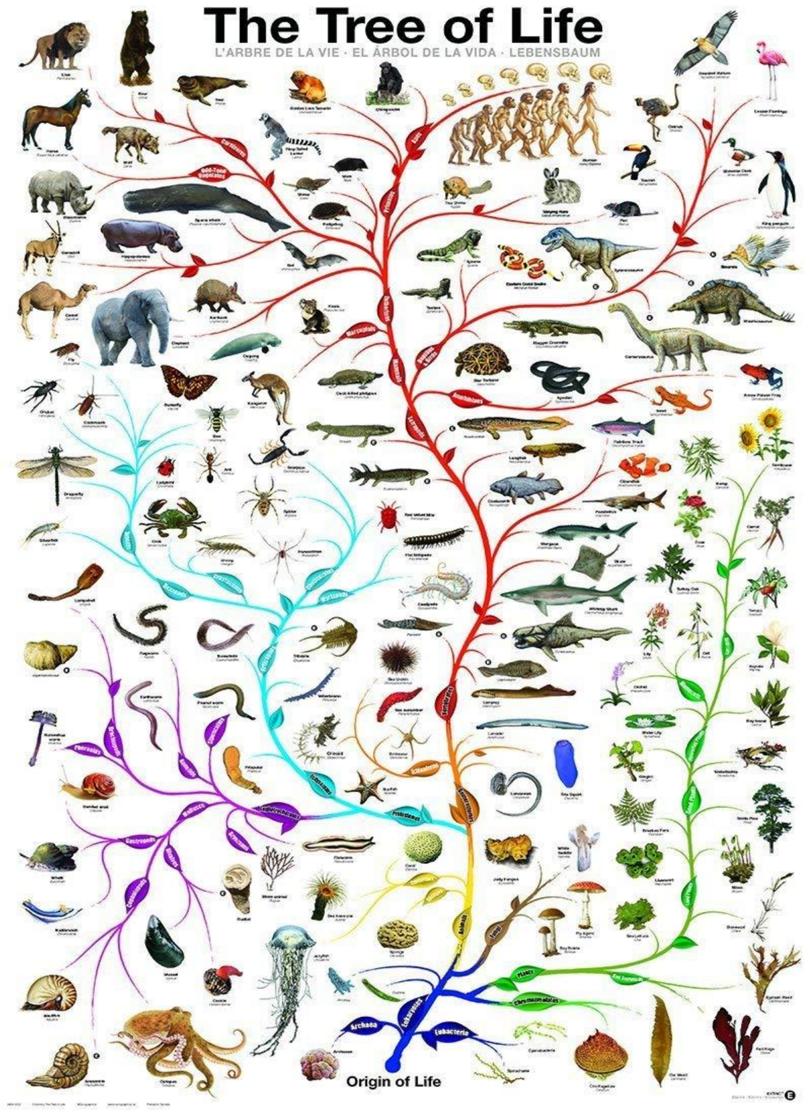
Reconstruction of ancestral proteins


36
I think
Case must be that one
generation then should be
as many living as now
To do this & to have many species
in same genus (as is). requires
extinction.
B
C
D
1
A



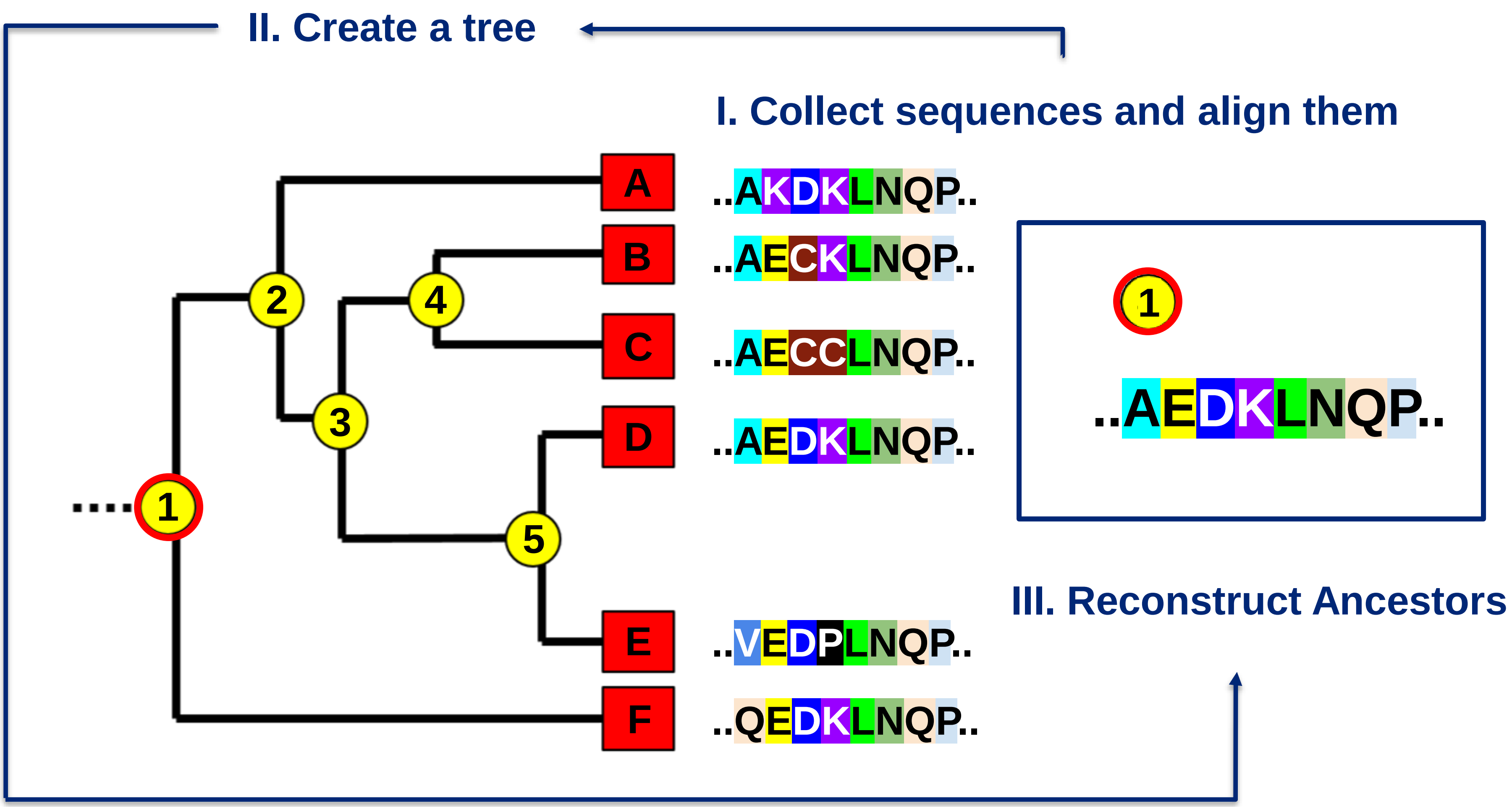
Reconstruction of ancestral proteins





millions years

Reconstruction of ancestral proteins



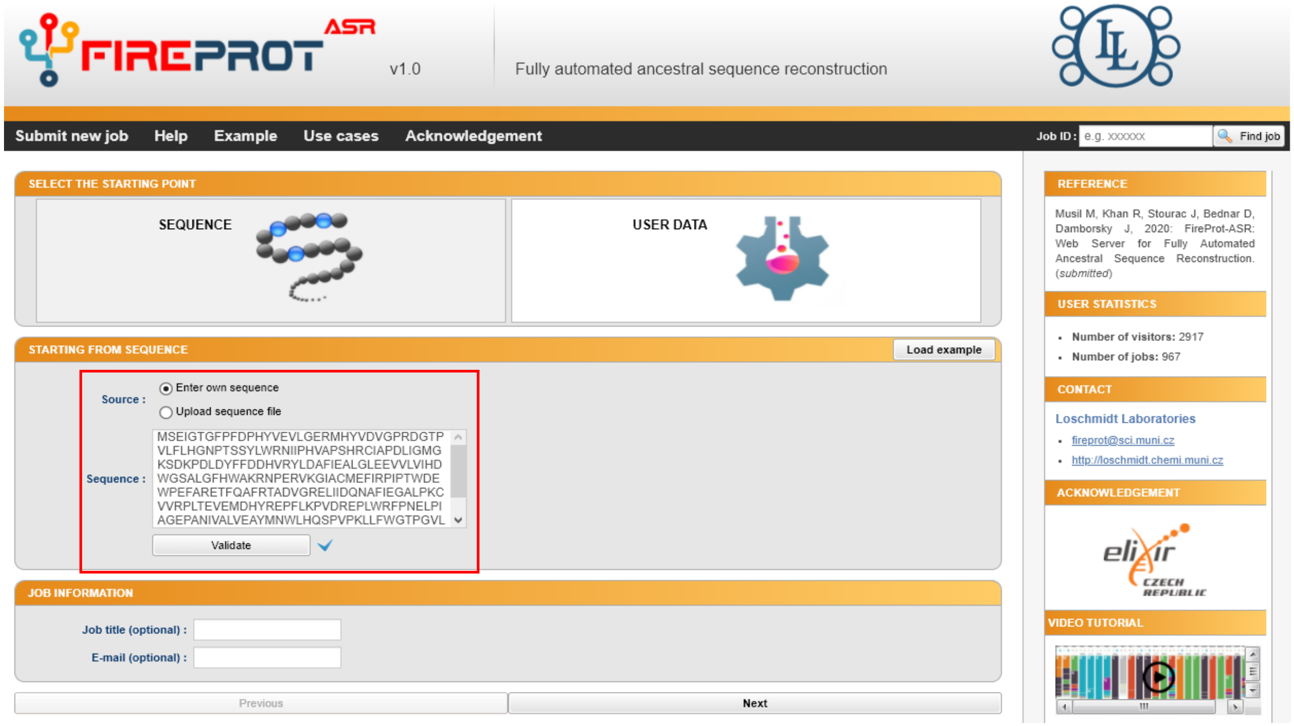
II. Create a tree

I. Collect sequences and align them

III. Reconstruct Ancestors

Reconstruction of ancestral proteins

- FireProt^{ASR}
  - https://loschmidt.chemi.muni.cz/fireprotasr/
  - Web server for automated ancestral sequence reconstruction
  - Ancestral reconstruction, successor prediction, latent space analysis
  - Design of stable proteins with good yields and broad specificity



Source : ● Enter own sequence ○ Upload sequence file

Sequence :

```
MSEIGTGFPFDPHYVEVLGERMHYVDVGPRDGTP
VLFLHGNPTSSYLWRNIIPHVAPSHRCIAPDLIGMG
KSDKPDLDYFFDDHVRYLDAFIEALGLEEVVLVIHD
WGSALGFHWAKRNPERVKGIACMEFIRPIPTWDE
WPEFARETFQAFRTADVGRELIIDQNAFIEGALPKC
VVRPLTEVEMDHYREPFLKPVDREPLWRFPNELPI
AGEPANIVALVEAYMNWLHQSPVPKLLFWGTPGVL
```

✓

Job title (optional) :

E-mail (optional) :

REFERENCE

Musil M, Khan R, Stourac J, Bednar D, Damborsky J, 2020: FireProt-ASR: Web Server for Fully Automated Ancestral Sequence Reconstruction. (*submitted*)

USER STATISTICS

- **Number of visitors:** 2917
- **Number of jobs:** 967

CONTACT

Loschmidt Laboratories

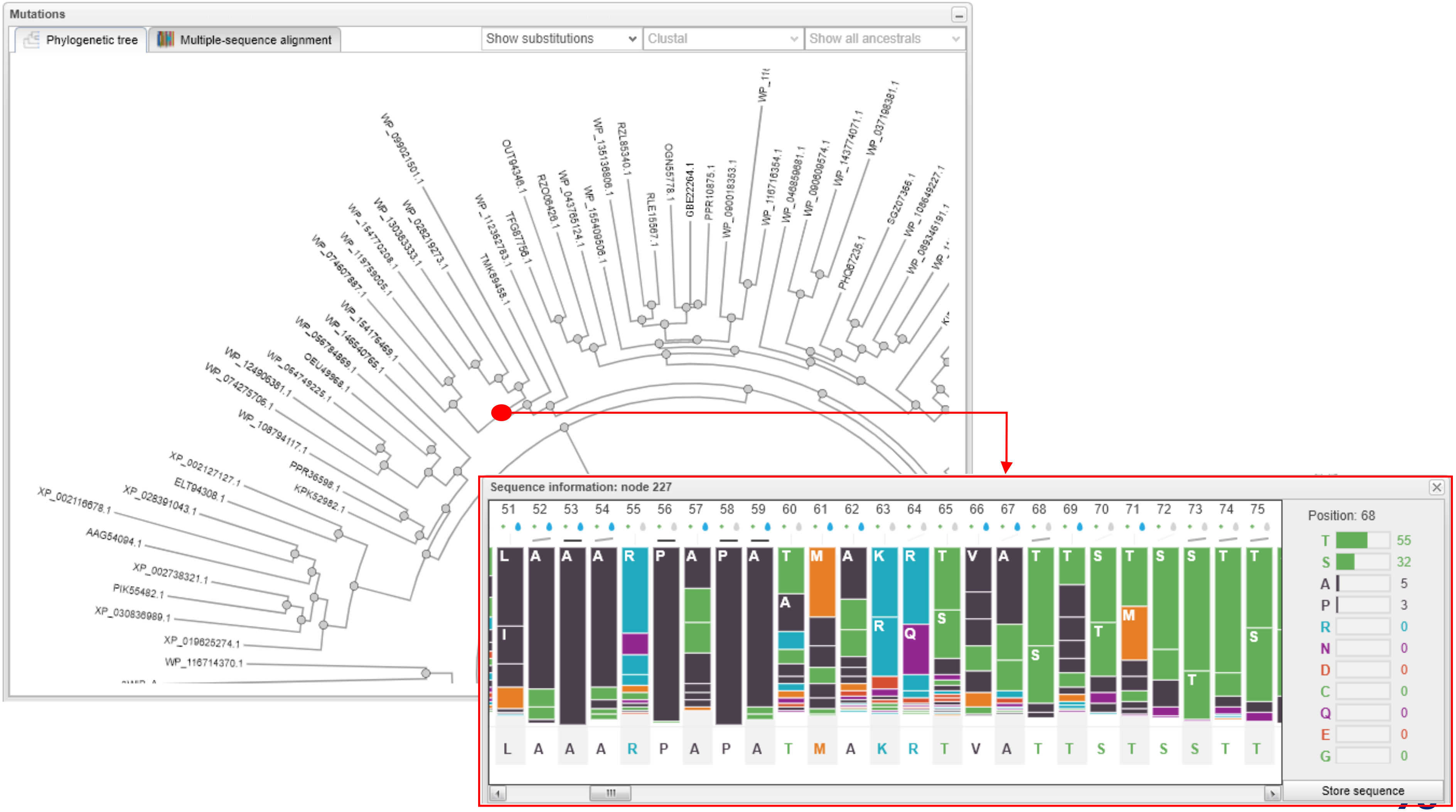
- fireprot@sci.muni.cz
- http://loschmidt.chemi.muni.cz

ACKNOWLEDGEMENT

VIDEO TUTORIAL

Reconstruction of ancestral proteins

- FireProt^{ASR}



Sequence information: node 227

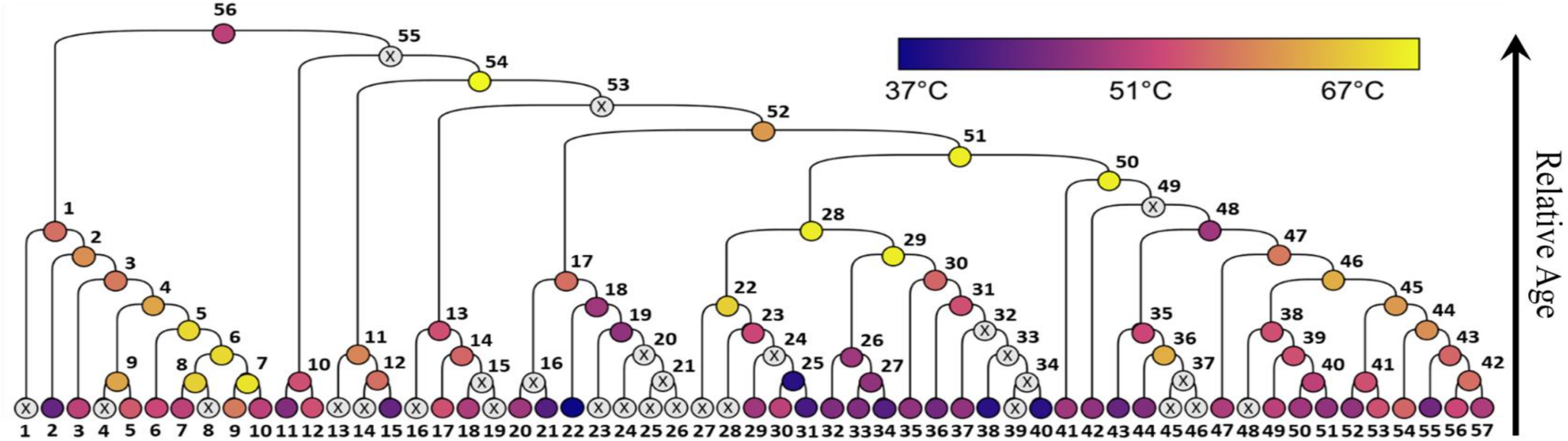
51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
L	A	A	A	R	P	A	P	A	T	M	A	K	R	T	V	A	T	T	S	T	S	S	T	T

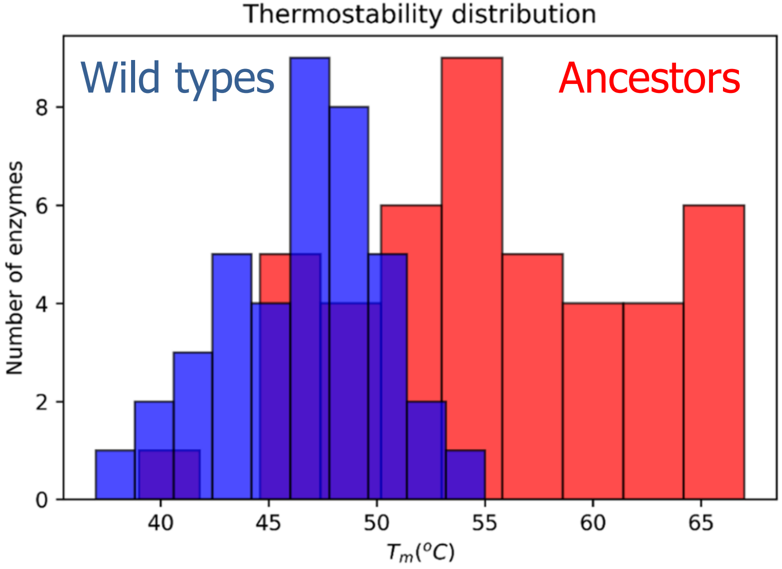
Position: 68

T	55
S	32
A	5
P	3
R	0
N	0
D	0
C	0
Q	0
E	0
G	0

FireProt^{ASR} use case

57 wild types and 56 ancestors tested





Ancestors:

- Average T_m +9 °C
- 60 % have higher T_m than the best WT
- enzymes with top activity
- Slightly better yields

What to keep in mind?

What to keep in mind?

- sequence databases
  - **nucleotide**: GenBank, EMBL-BANK, DDBJ; **protein**: UniProtKB, nr Protein database
  - **errors** in sequences and annotations
- database searches
  - **text-based:** results influenced by sequence annotations
  - **sequence-based:** identification of family members - BLAST, PSI-BLAST - *E*-value
  - false positive results: sequences should be filtered
- selection of proteins for experimental characterization
  - **clustering:** classification and filtering of hits from database searches
  - **sequence comparison:** classification and identification of unique sequences
  - sequences from **extremophiles:** potentially adapted to extreme conditions
  - **EnzymeMiner:** automated identification of interesting catalysts
  - **FireProt^{ASR}:** design of stable, soluble, and broad specificity proteins

What to keep in mind?

- *in silico* identification and analysis of sequences - fast and cheap way to identify new proteins

References

- Xiong, J. (2006). **Essential Bioinformatics**. Cambridge University Press, New York, p. 352.
- Claverie, J-M. and Notredame, C. (2006). **Bioinformatics For Dummies** (2nd ed.). Wiley Publishing, Hoboken, p. 436.
- Steele, H.L. *et al.* (2009). Advances in Recovery of Novel Biocatalysts from Metagenomes. *Journal of Molecular Microbiology and Biotechnoly* **16**: 25–37.
- NCBI Resource Coordinators (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **41**: D8-D20.
- Magrane, M. and Consortium U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**: bar009.
- Frickey, T. and Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**: 3702-3704.
- Pagani, I. *et al.* (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **40**, D571-579.
- Van den Burg, B. (2003). Extremophiles as a source for novel enzymes. *Current Opinion in Microbiology* **6**: 213-218.

LOSCHMIDT LABORATORIES

Protein production

Protein Engineering Lecture #3
Michal Vašina

DALL·E 3