LOSCHMIDT
LABORATORIES

PROTEIN ENGINEERING

2. *IN SILICO* IDENTIFICATION OF PROTEINS

Loschmidt Laboratories
Department of Experimental Biology
Masaryk University, Brno

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?

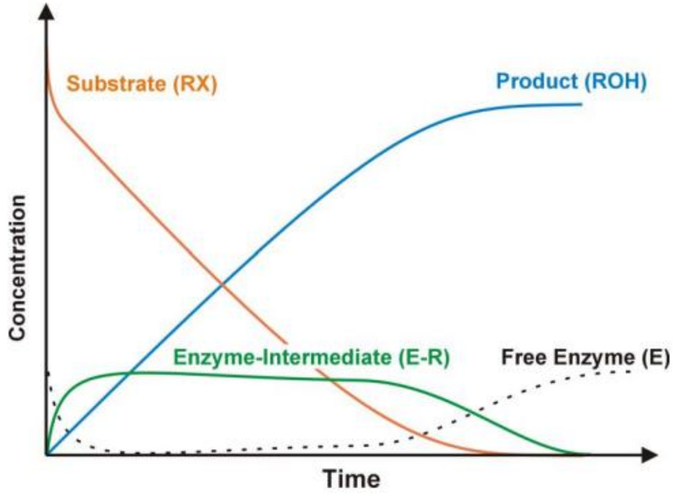
Why to search for new proteins?

Why to search for new proteins?

- plenty of reasons

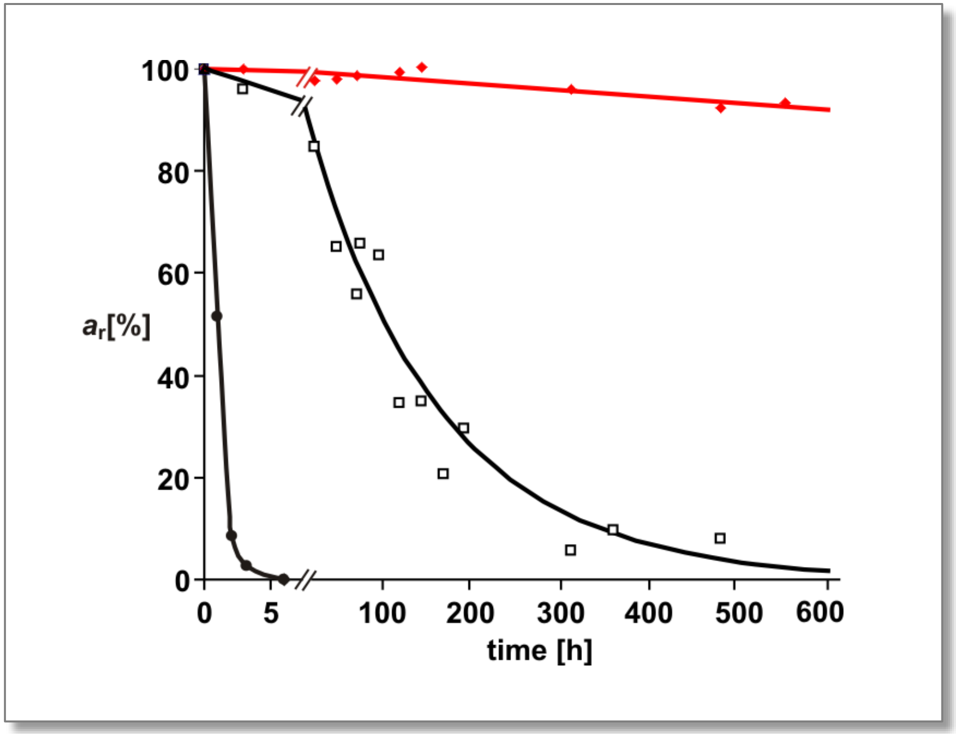
Why to search for new proteins?

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



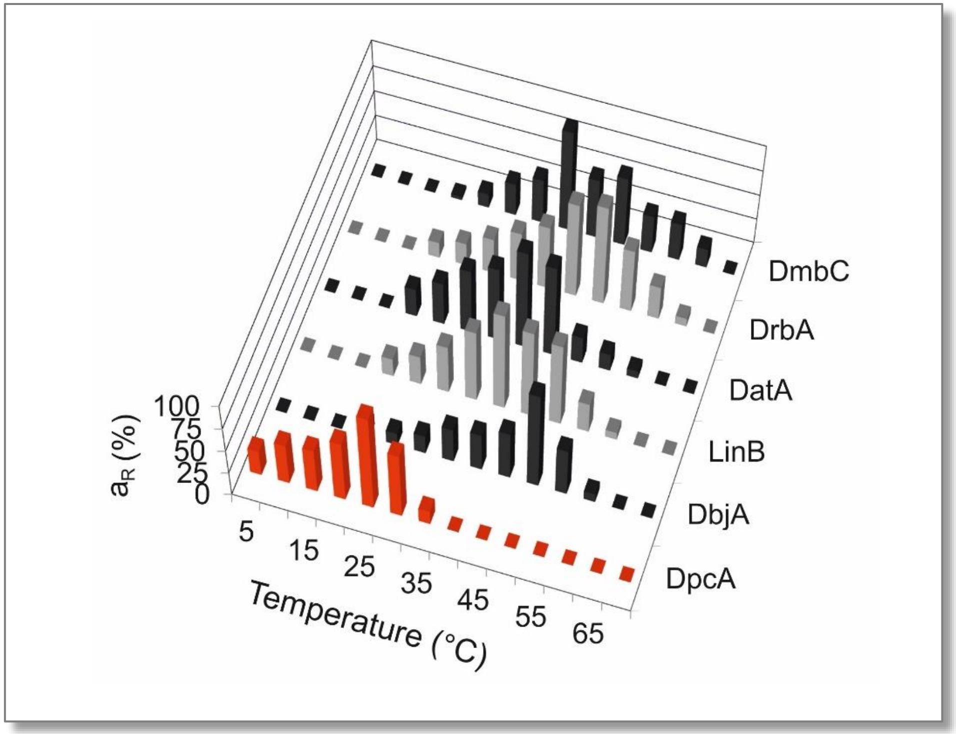
Why to search for new proteins?

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



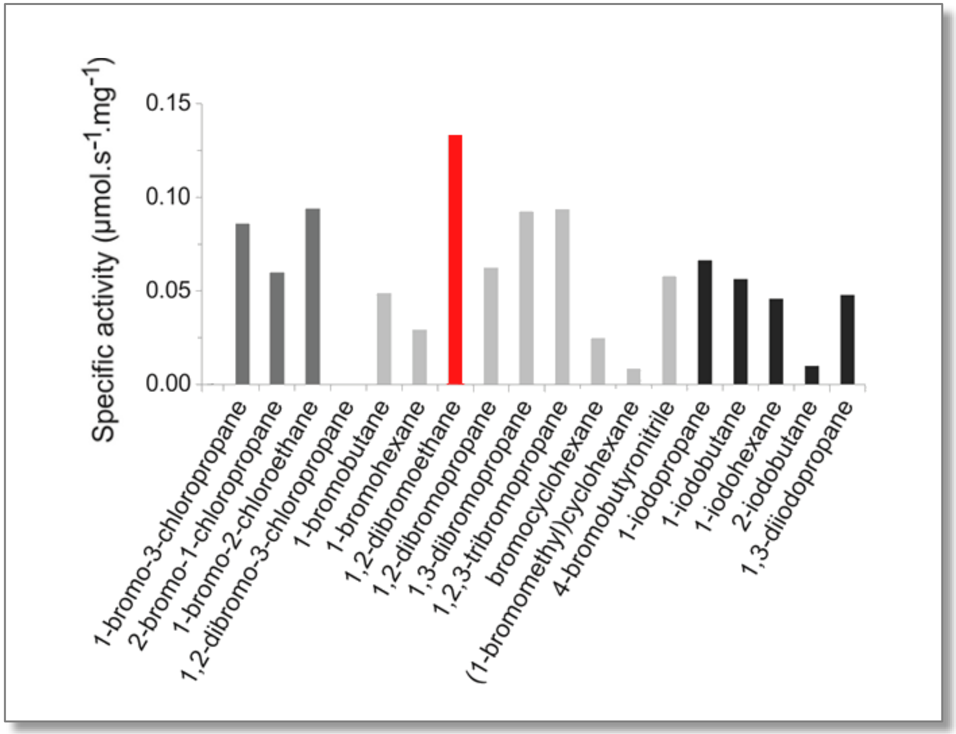
Why to search for new proteins?

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



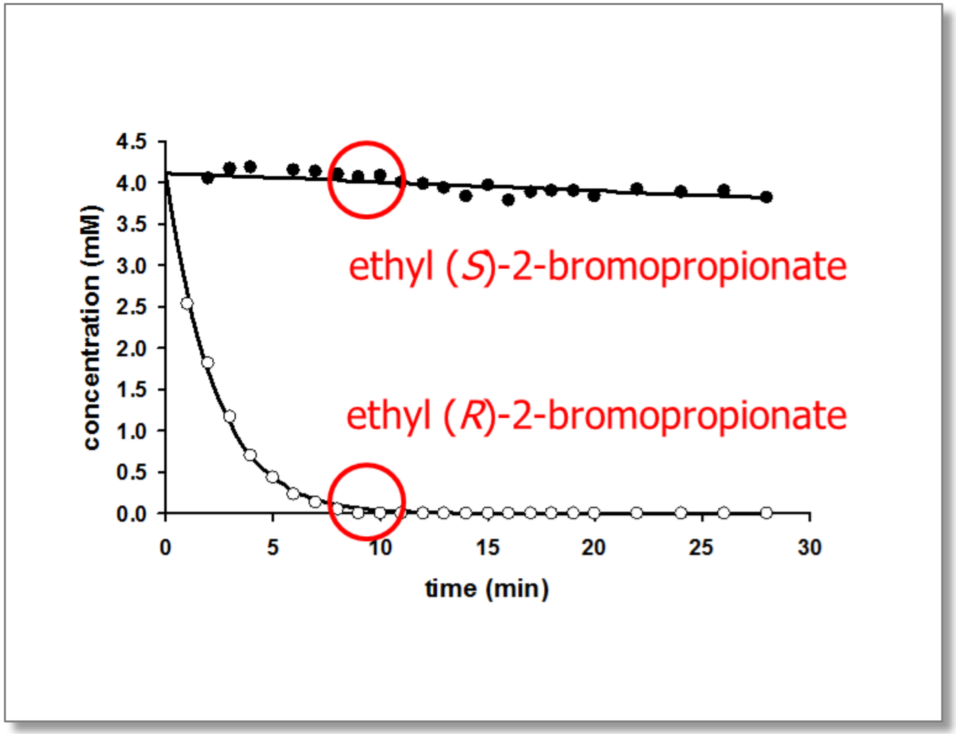
Why to search for new proteins?

- 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

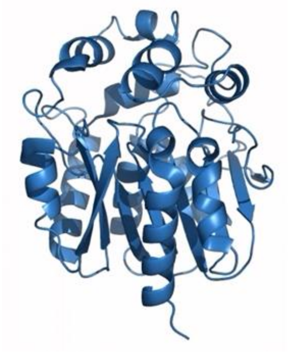


Why to search for new proteins?

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

Why to search for new proteins?

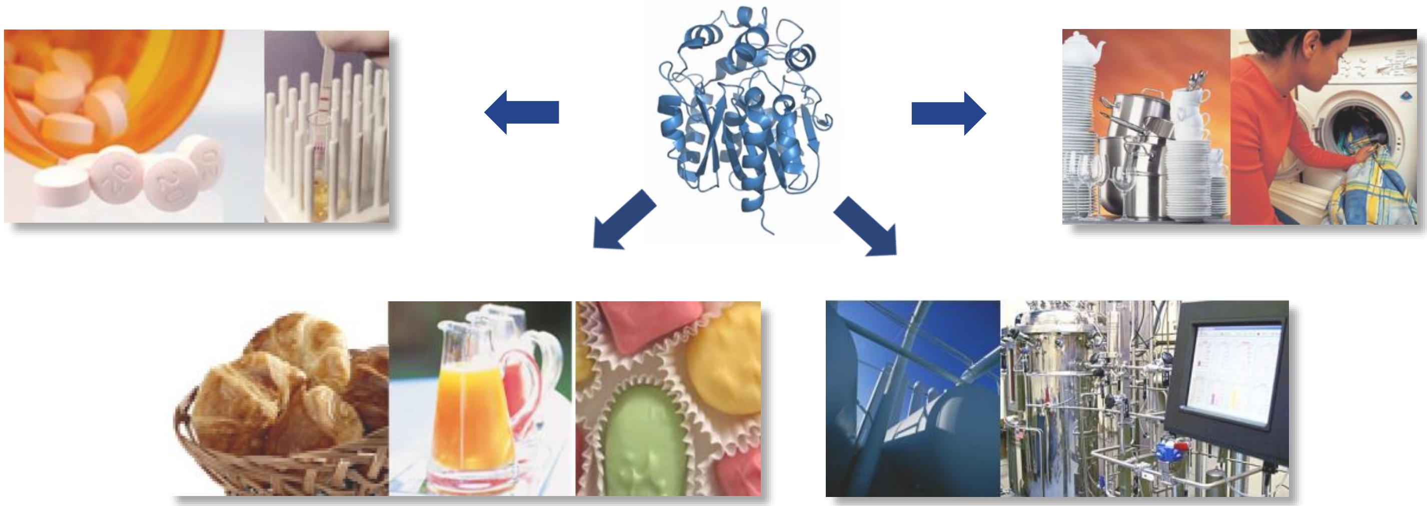
- ❑ better understanding of structure-function relationships
- ❑ novel properties
- ❑ better starting points for protein engineering



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



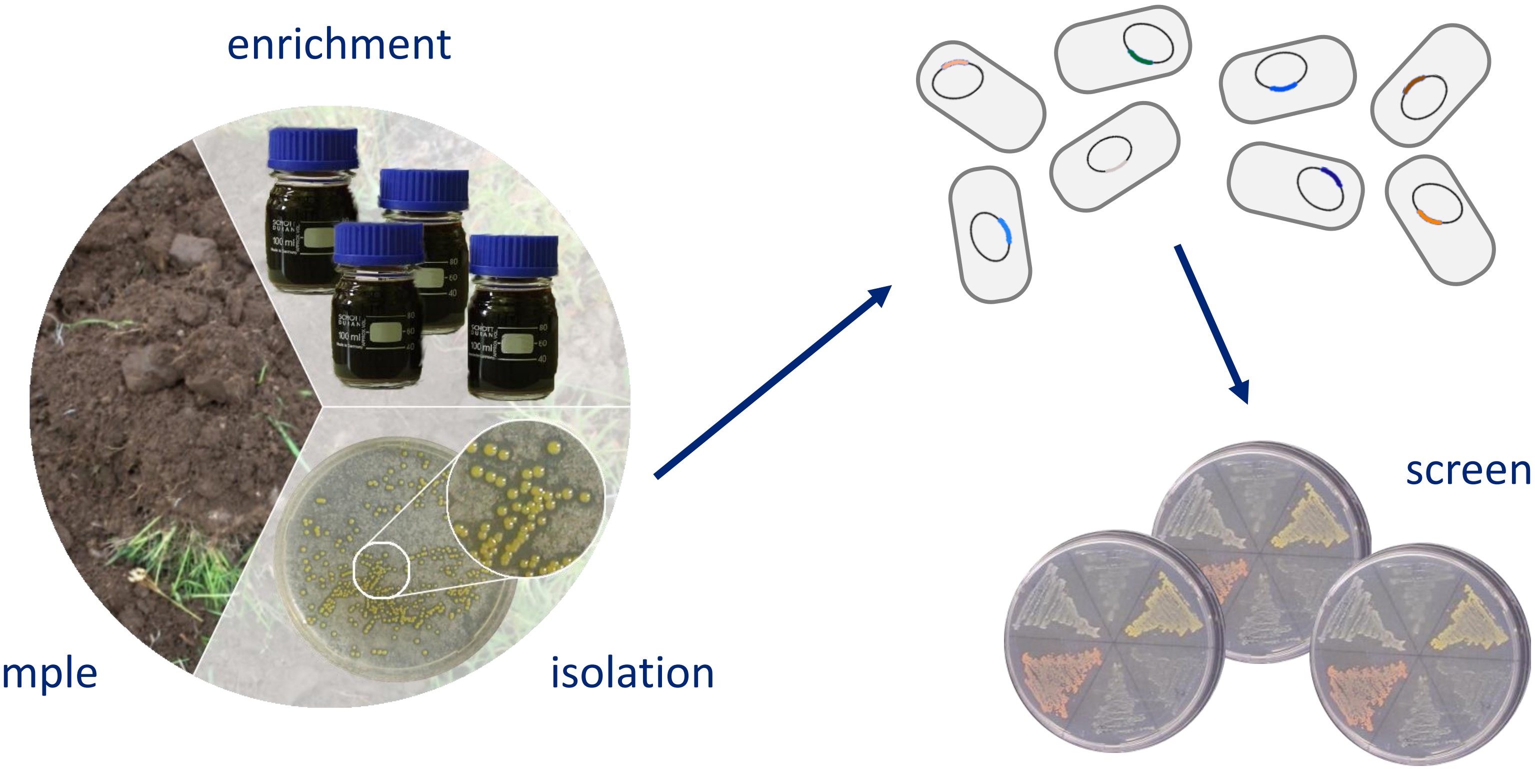
How to acquire new proteins?

How to acquire new proteins?

- traditional approach
- metagenomic approach
- bioinformatic approach

How to acquire new proteins?

- traditional approach



enrichment

sample

isolation

DNA library

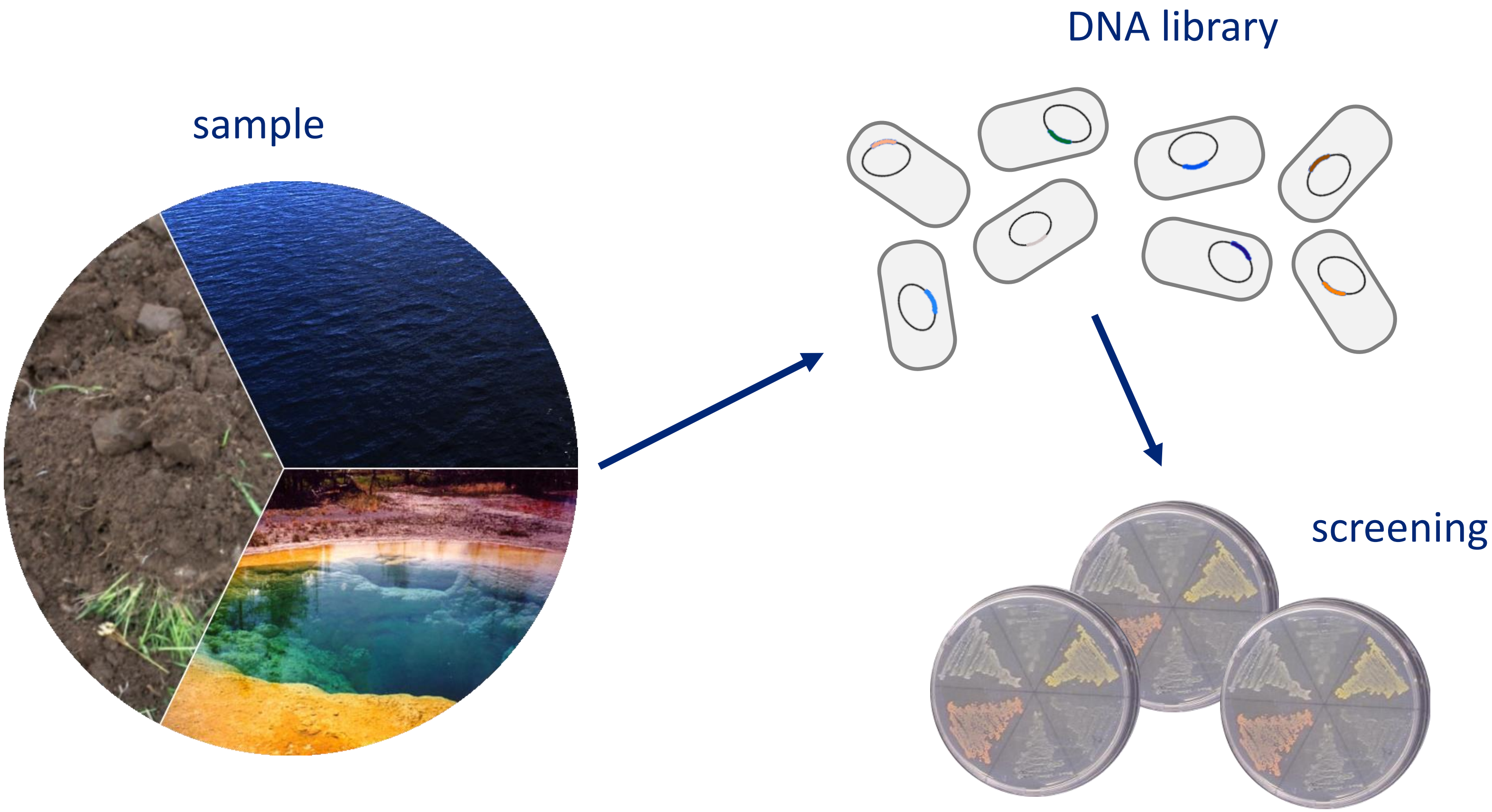
screening

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,...
  - ☹ 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?

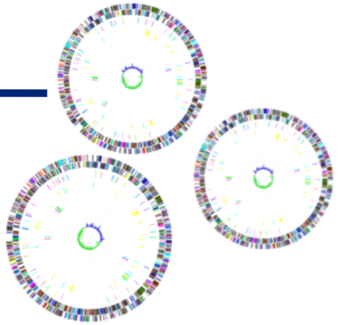
- ❑ metagenomic approach
  - isolation and cloning of DNA extracted directly from environmental sample (without culturing the present organisms)
  - genes recovered by DNA library screening or PCR with specific primers,...
  - ☺ enables to explore biodiversity of uncultured microorganisms

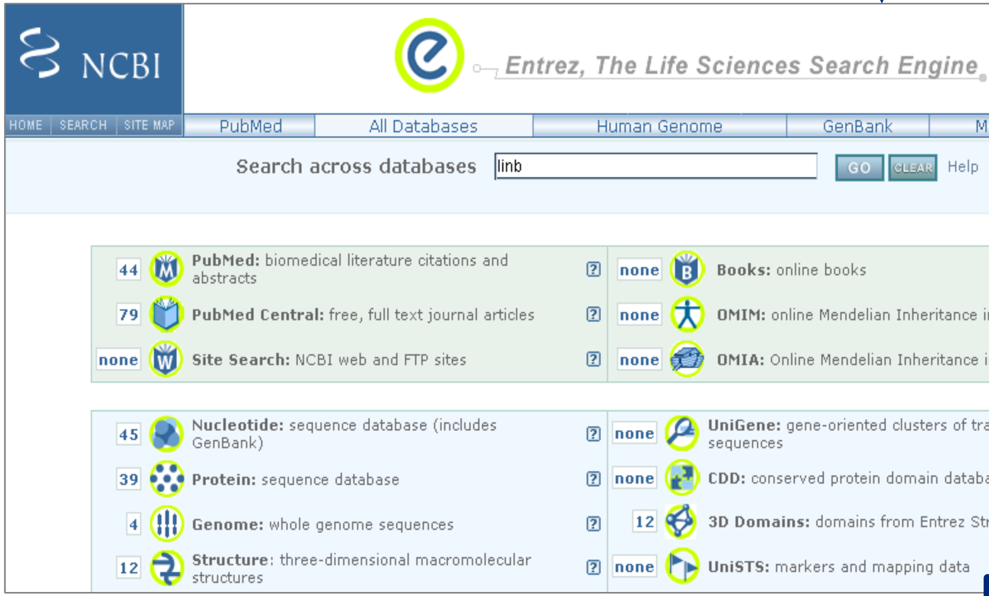
How to acquire new proteins?

- ❑ bioinformatic approach

sequence database

(meta)genomic sequencing projects



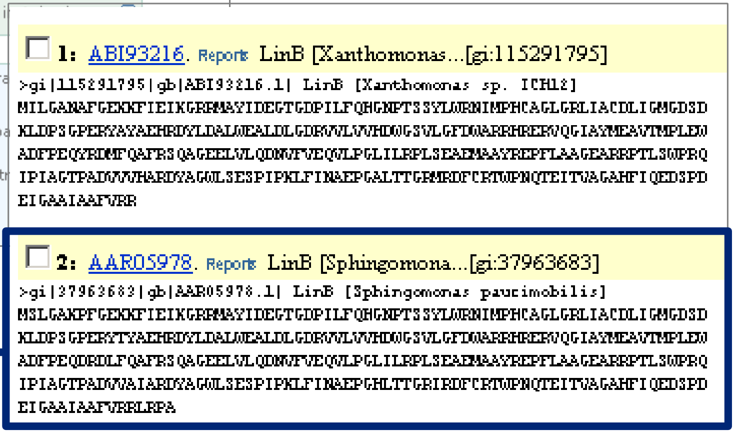

NCBI
HOME | SEARCH | SITE MAP
Entrez, The Life Sciences Search Engine
PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST
GO | CLEAR | Help

Search across databases: linb

44 PubMed: biomedical literature citations and abstracts
79 PubMed Central: free, full text journal articles
none Site Search: NCBI web and FTP sites
none Books: online books
none OMIM: online Mendelian Inheritance in Man
none OMIA: Online Mendelian Inheritance i

45 Nucleotide: sequence database (includes GenBank)
39 Protein: sequence database
4 Genome: whole genome sequences
12 Structure: three-dimensional macromolecular structures
none UniGene: gene-oriented clusters of tra sequences
none CDD: conserved protein domain databa
12 3D Domains: domains from Entrez Str
none UniSTS: markers and mapping data

in silico "screening"



□ 1: ABI93216. Reports LinB [Xanthomonas...[gi:115291795]

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

□ 2: AAR05978. Reports 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?

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)

NCBI

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

EMBL
NUCLEOTIDE
SEQUENCE
DATABASE

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

DDBJ
DNA Data Bank of Japan

Databases of nucleotide sequences

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

Databases of protein sequences

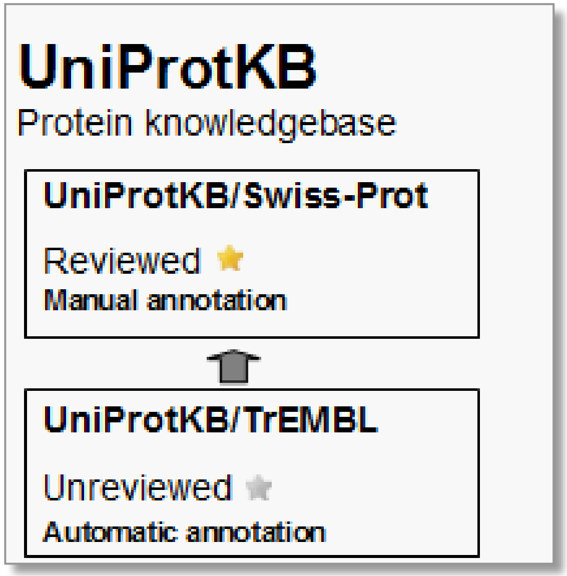
- UniProtKB
  - http://www.uniprot.org/
  - provided by EBI, Swiss Institute of Bioinformatics and Protein Information Resource
- nr Protein database
  - http://www.ncbi.nlm.nih.gov/protein/
  - provided by NCBI

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 100,000,000 sequences
  - ☹ 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)



UniProtKB

Protein knowledgebase

UniProtKB/Swiss-Prot

Reviewed ★

Manual annotation

UniProtKB/TrEMBL

Unreviewed ★

Automatic annotation

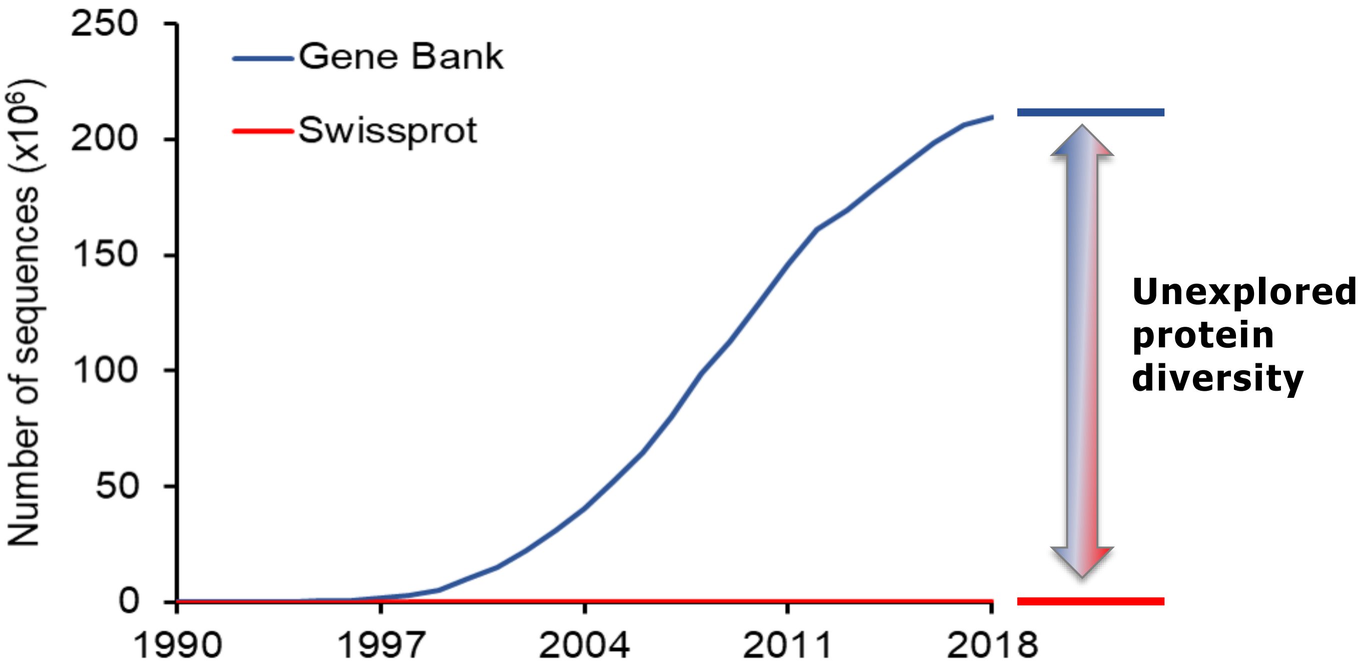
Databases of protein sequences

- ❑ UniProtKB/Swiss-Prot
  - high quality annotations, i.e., manually annotated entries or expert-reviewed automatic annotations
  - ☺ source of reliable information
  - ☹ contains “only” ~ 600,000 sequences

- ❑ UniProtKB/TrEMBL
  - ☹ automatic annotations – lower quality, errors
  - ☺ contains ~ 180,000,000 sequences

Unexplored protein diversity

- Number of sequences
- Number of characterized 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?

How to find target sequences?

- text-based searches
- sequence-based searches

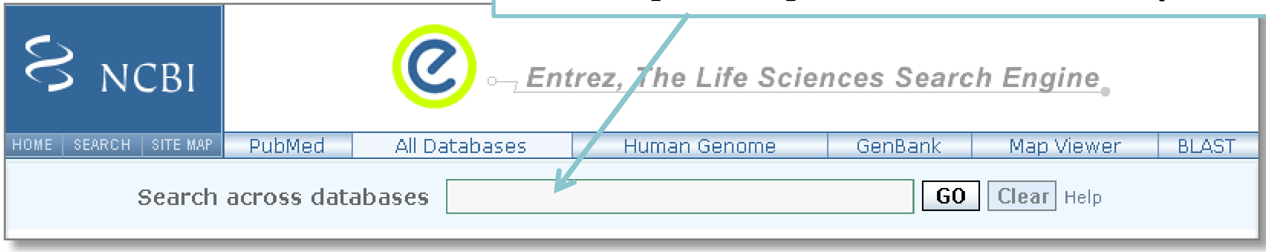
Text-based searches

- ❑ 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

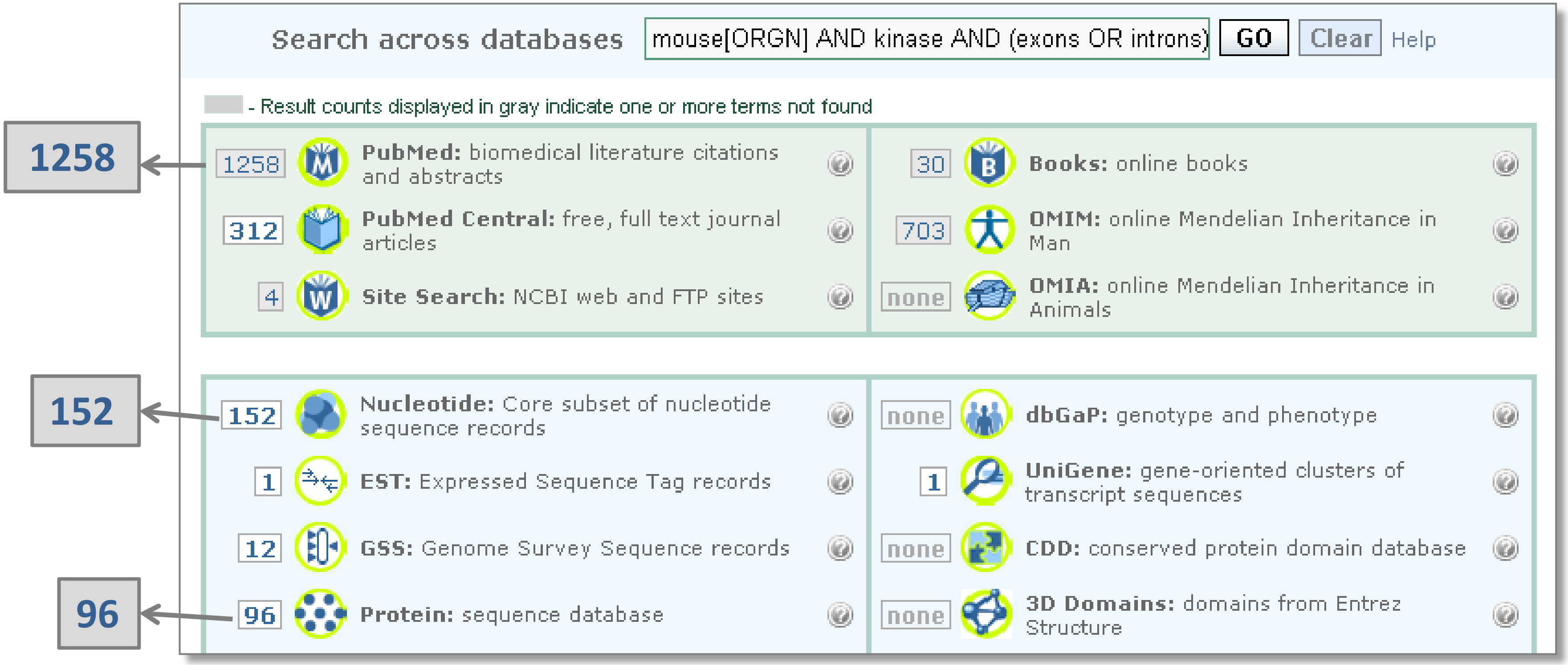
- database retrieval systems

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

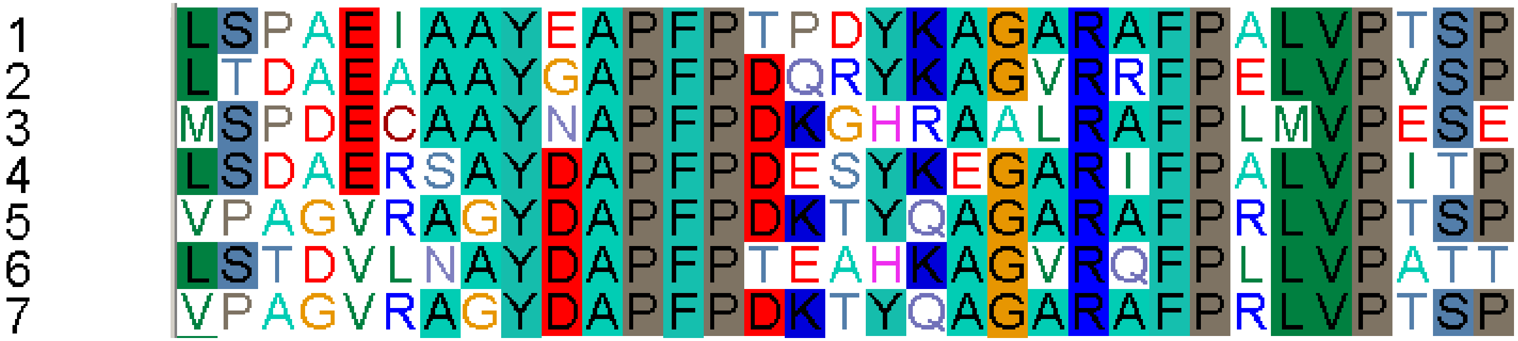
1258

152

96

Sequence-based searches

- searches based on sequence similarity
  - ☺ results not influenced by sequence annotations
- rely on 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

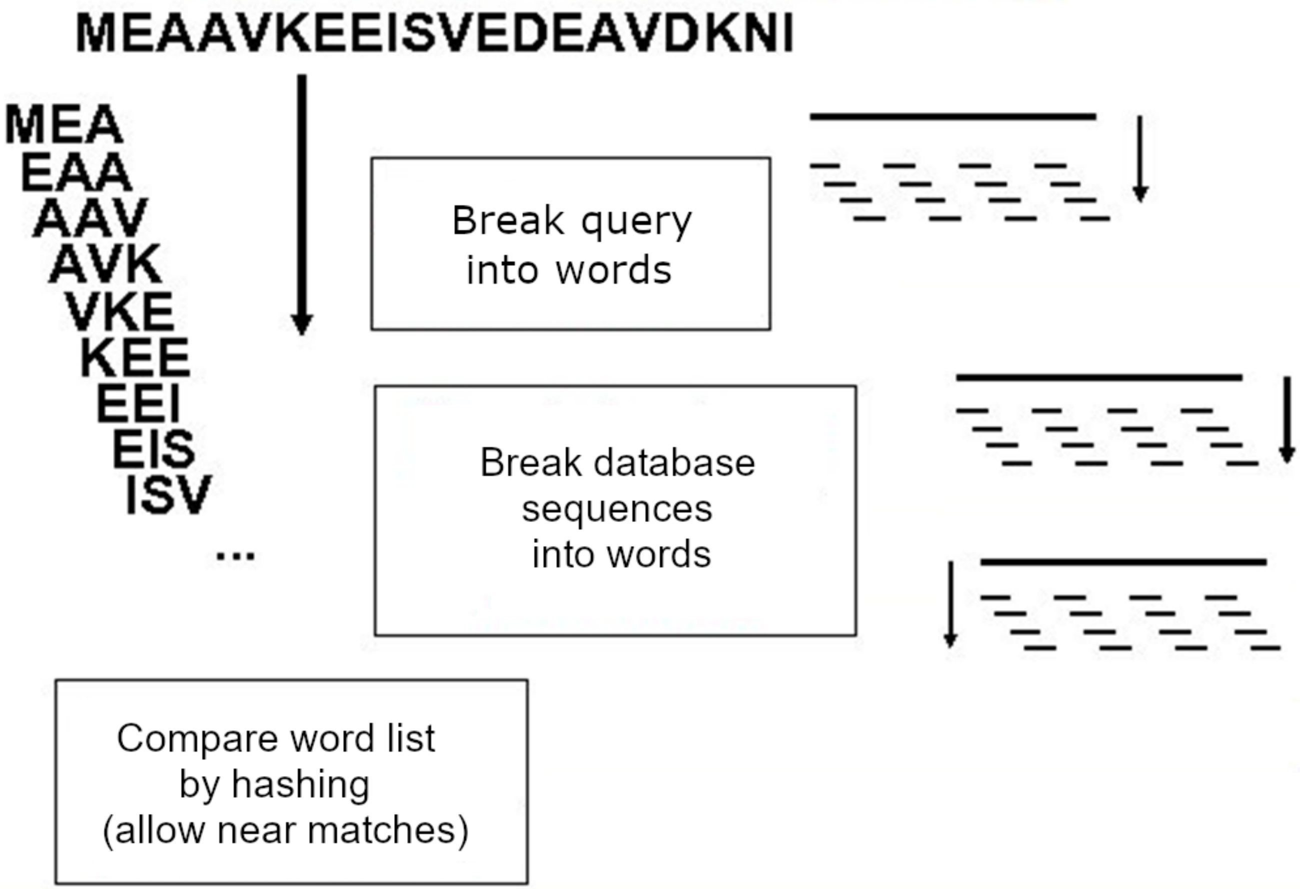
- BLAST
  - based on local pairwise alignment

- PSI-BLAST
  - “iterative BLAST” making use of multiple sequence alignment
  - very sensitive search strategy to detect weak but biologically significant similarities between sequences

- ...

BLAST

- Basic Local Alignment Search Tool



MEAAVKEEISVEDEAVDKNI

MEA
EAA
AAV
AVK
VKE
KEE
EEI
EIS
ISV
...

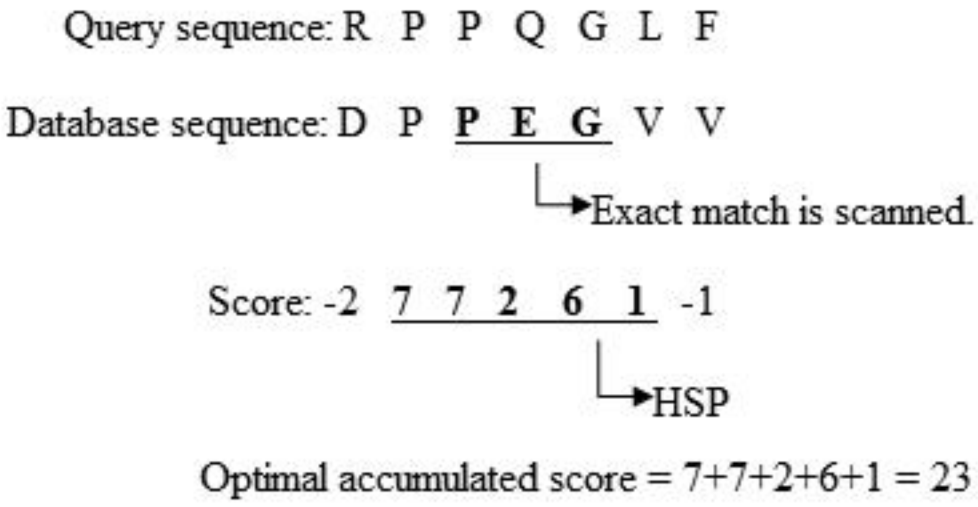
Break query into words

Break database sequences into words

Compare word list by hashing (allow near matches)

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool



Query sequence: R P P Q G L F

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

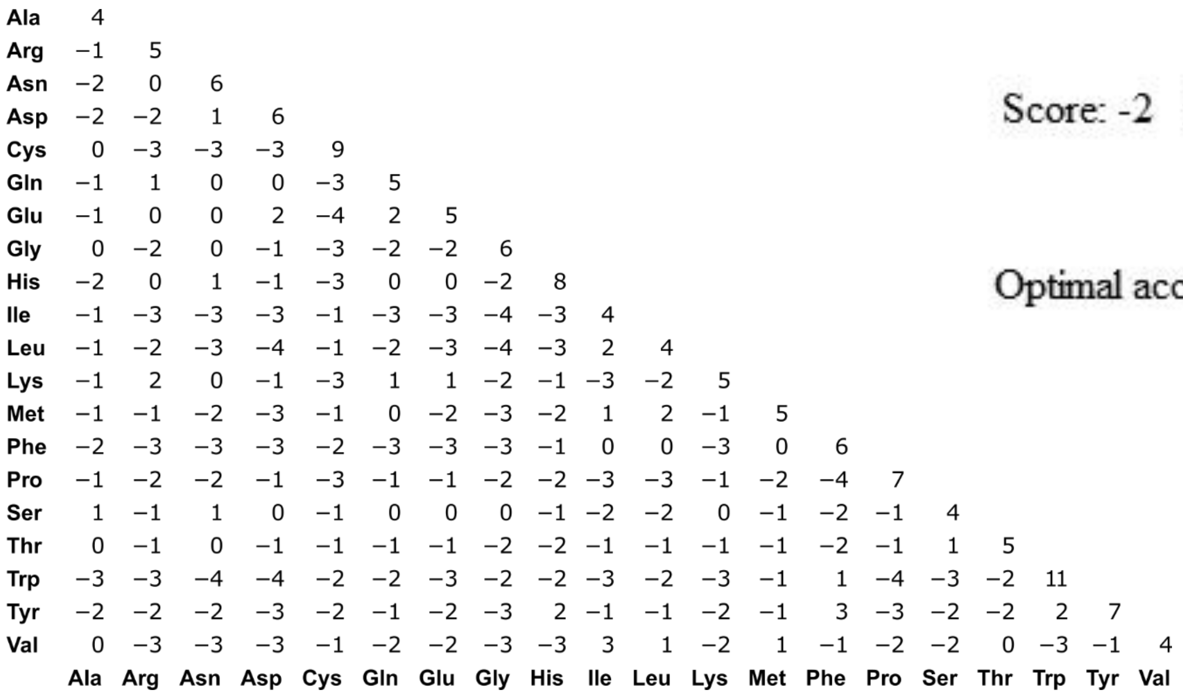
Exact match is scanned.

Score: -2 7 7 2 6 1 -1

HSP

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

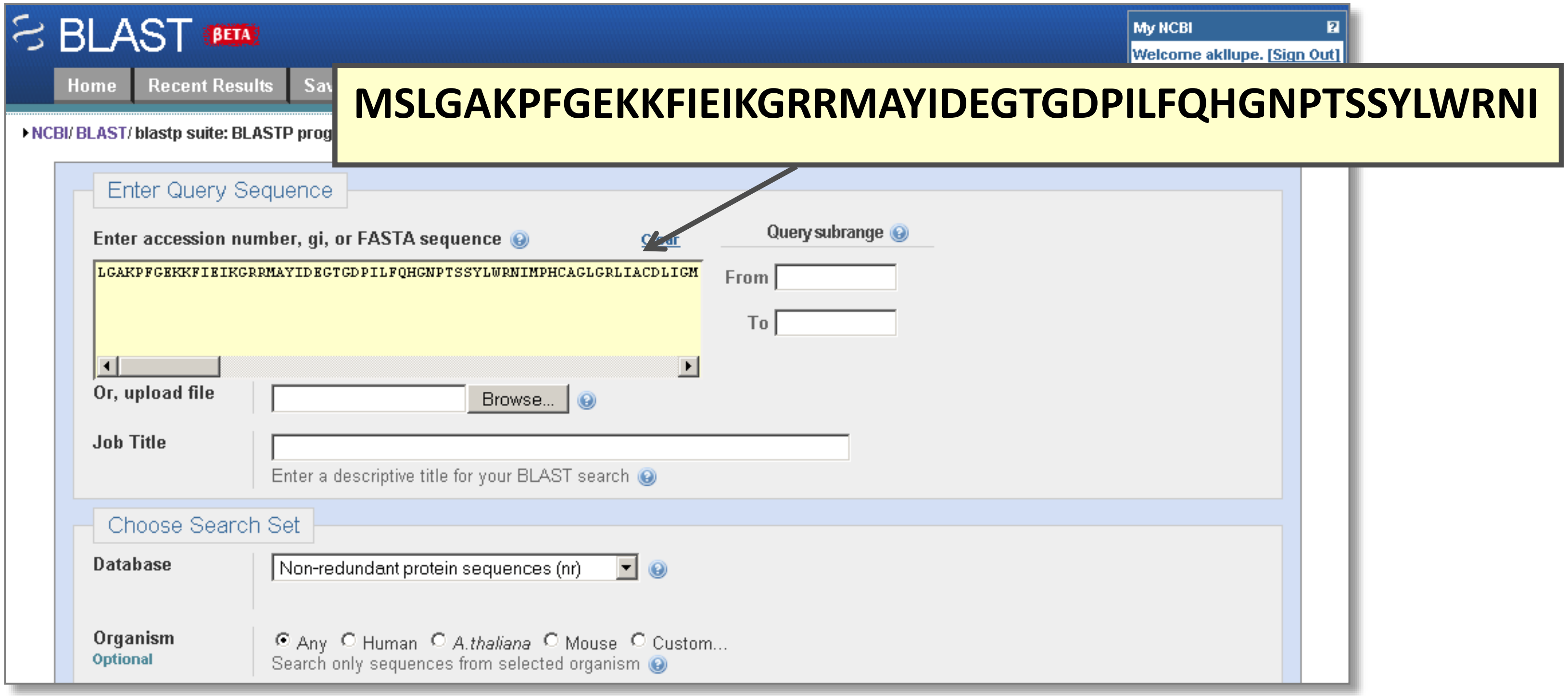
BLOSUM scoring matrix



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	5								
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**

Sequence-based searches

- PSI-BLAST input



MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNI

Enter accession number, gi, or FASTA sequence

LGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLIGM

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

Database: Non-redundant protein sequences (nr)

Organism (Optional): ● Any ○ Human ○ *A.thaliana* ○ Mouse ○ Custom...

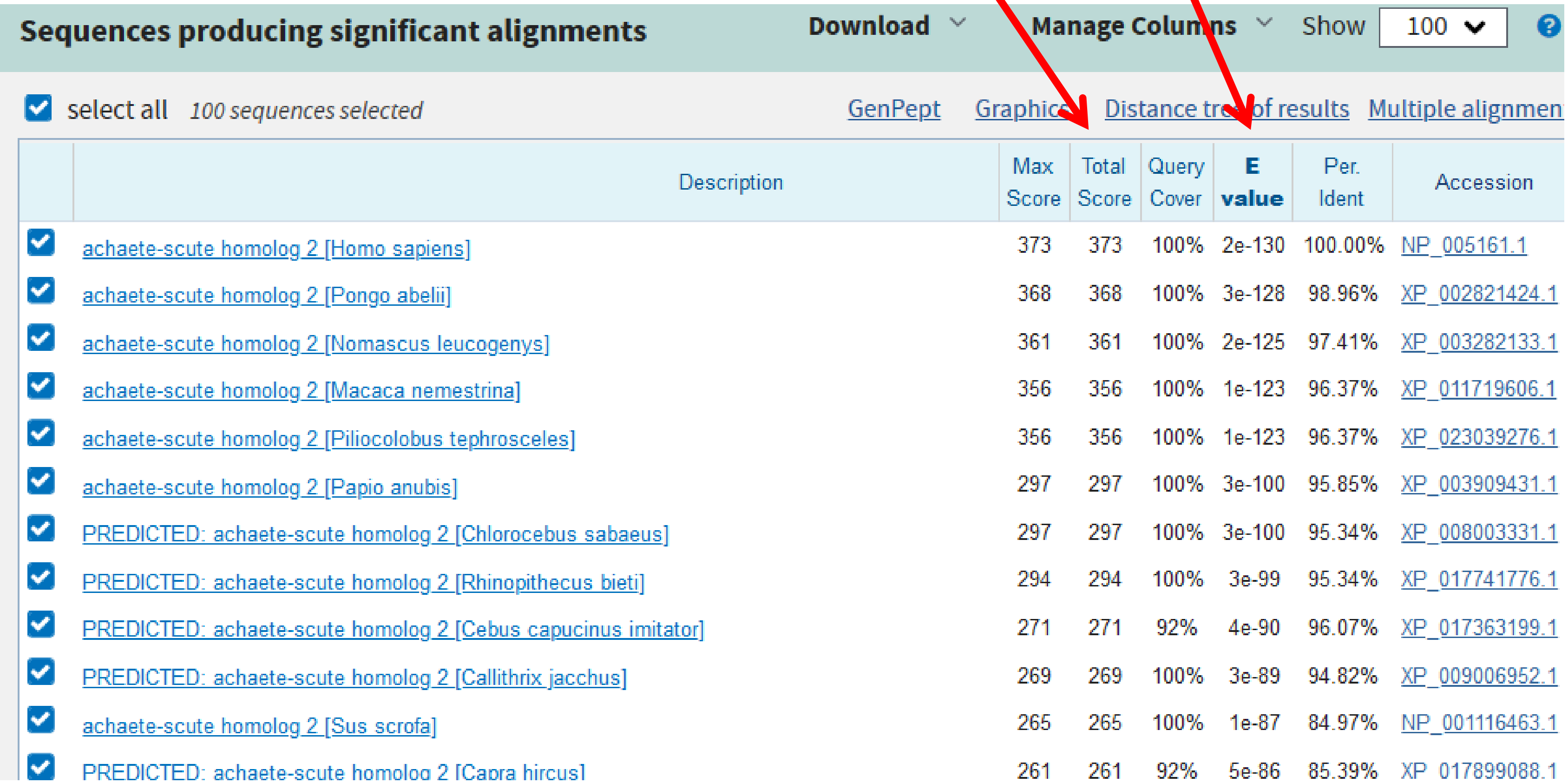
Search only sequences from selected organism

Sequence-based searches

- PSI-BLAST results

Score

E-value



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

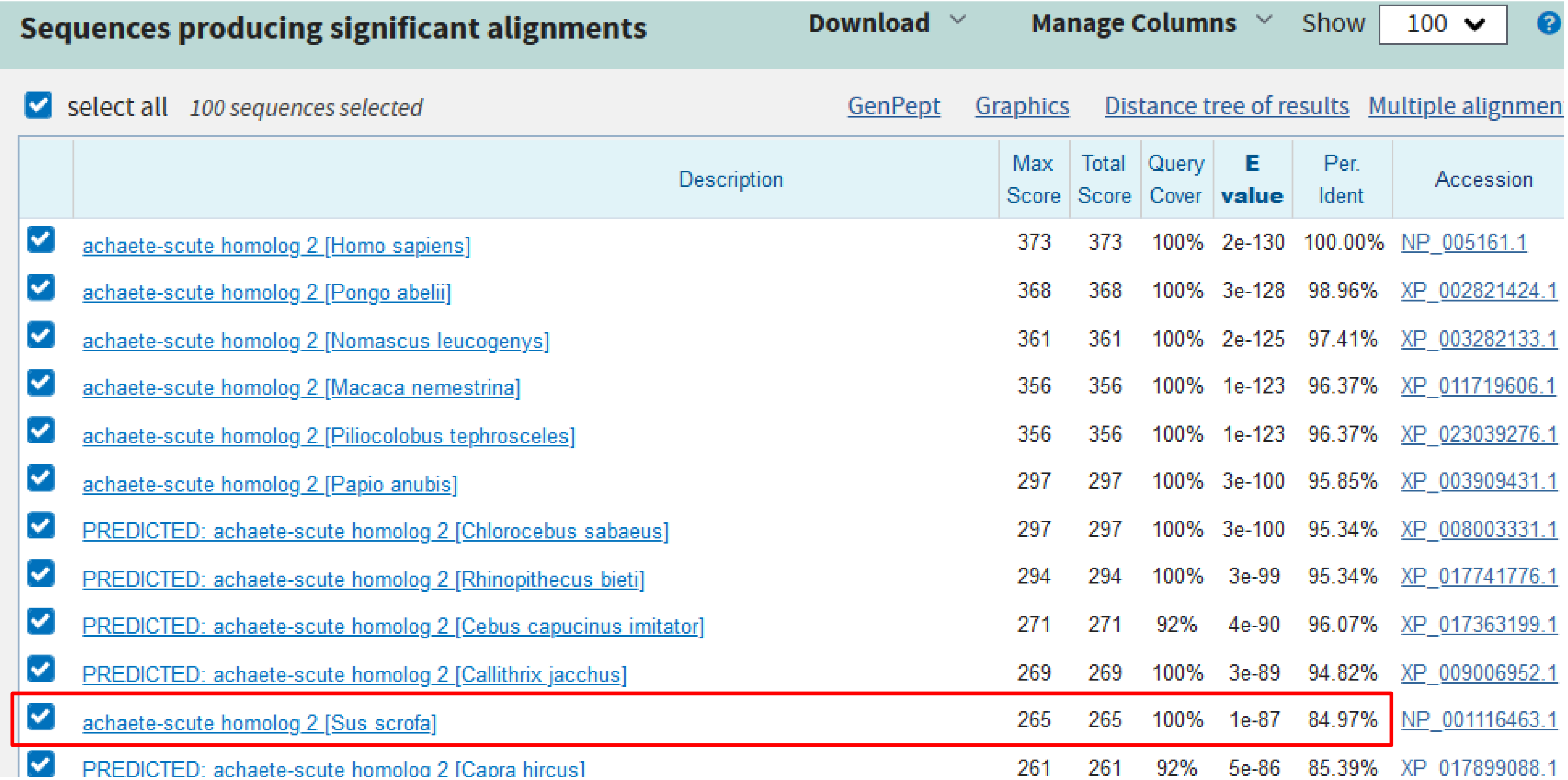
☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignmen

☑	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

hits

Sequence-based searches

❑ PSI-BLAST results



Sequences producing significant alignments

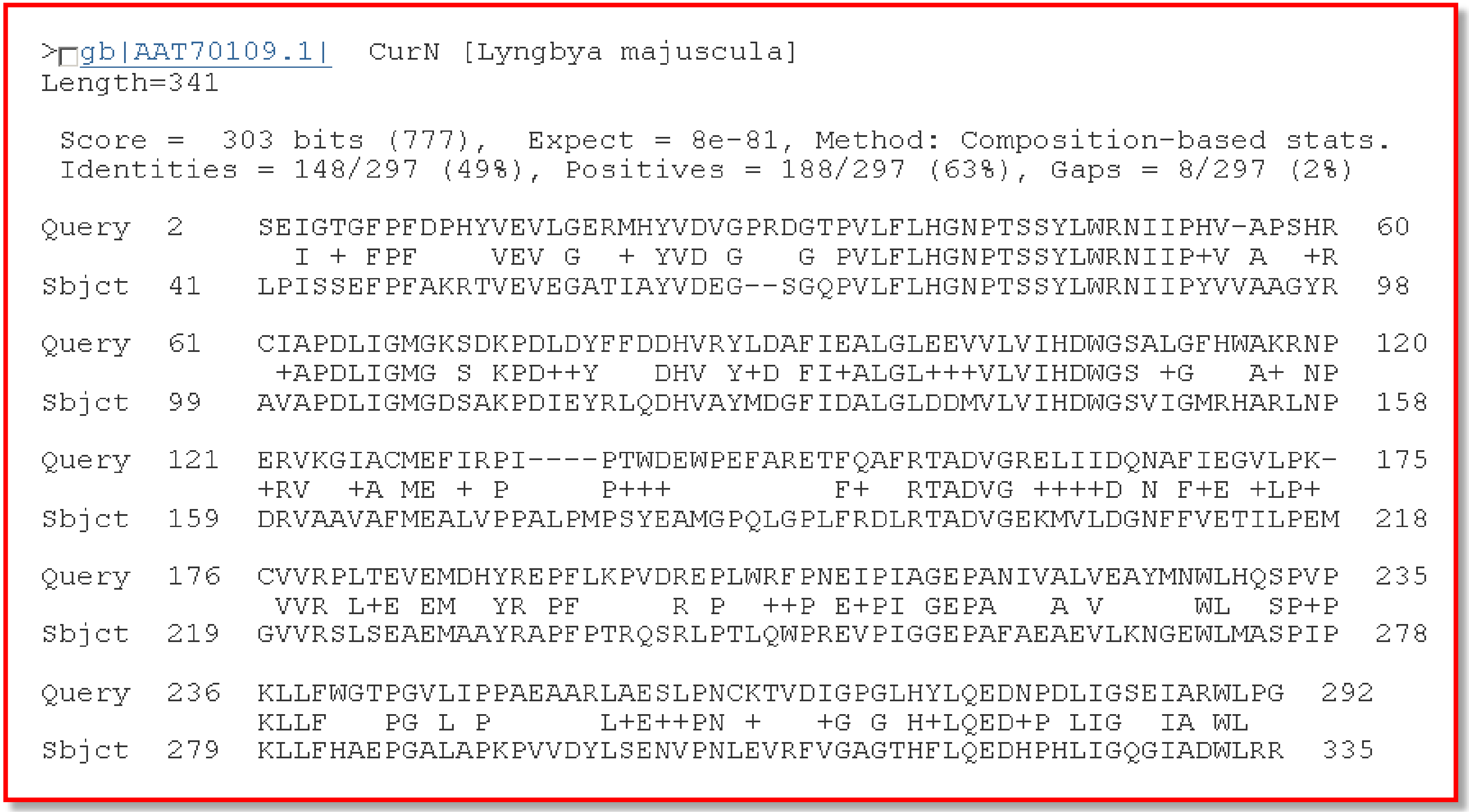
☑ select all *100 sequences selected*

	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

- 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
```

Sequence-based searches

- BLAST Score
  - normalized raw score
  - raw score = sum of substitution scores and gap penalties
  - higher is better, but does not adequately represent significance of alignment
- BLAST *E*-value
  - equal to the number of BLAST alignments with a given Score that are expected to be seen simply by a chance
  - indicator of alignment significance
  - 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)

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 evolutionary related groups
  3. selection of 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?

How to recognize interesting sequences?

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

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)

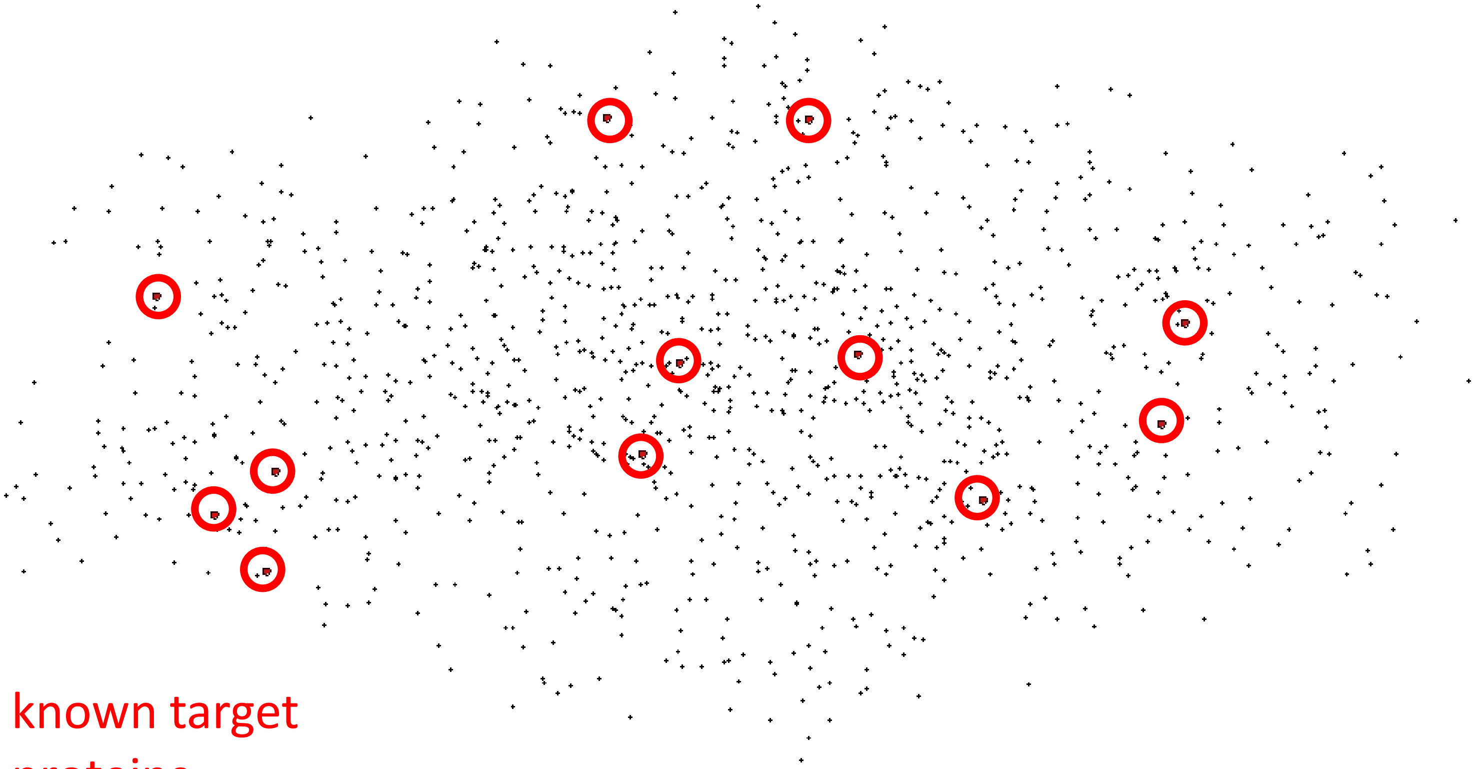
    → effective way to filter results of database searches

    → identification of members of individual protein families

  - CLANS - visualization of pairwise sequence similarities in three-dimensional space → overview of sequence space

Sequence clustering

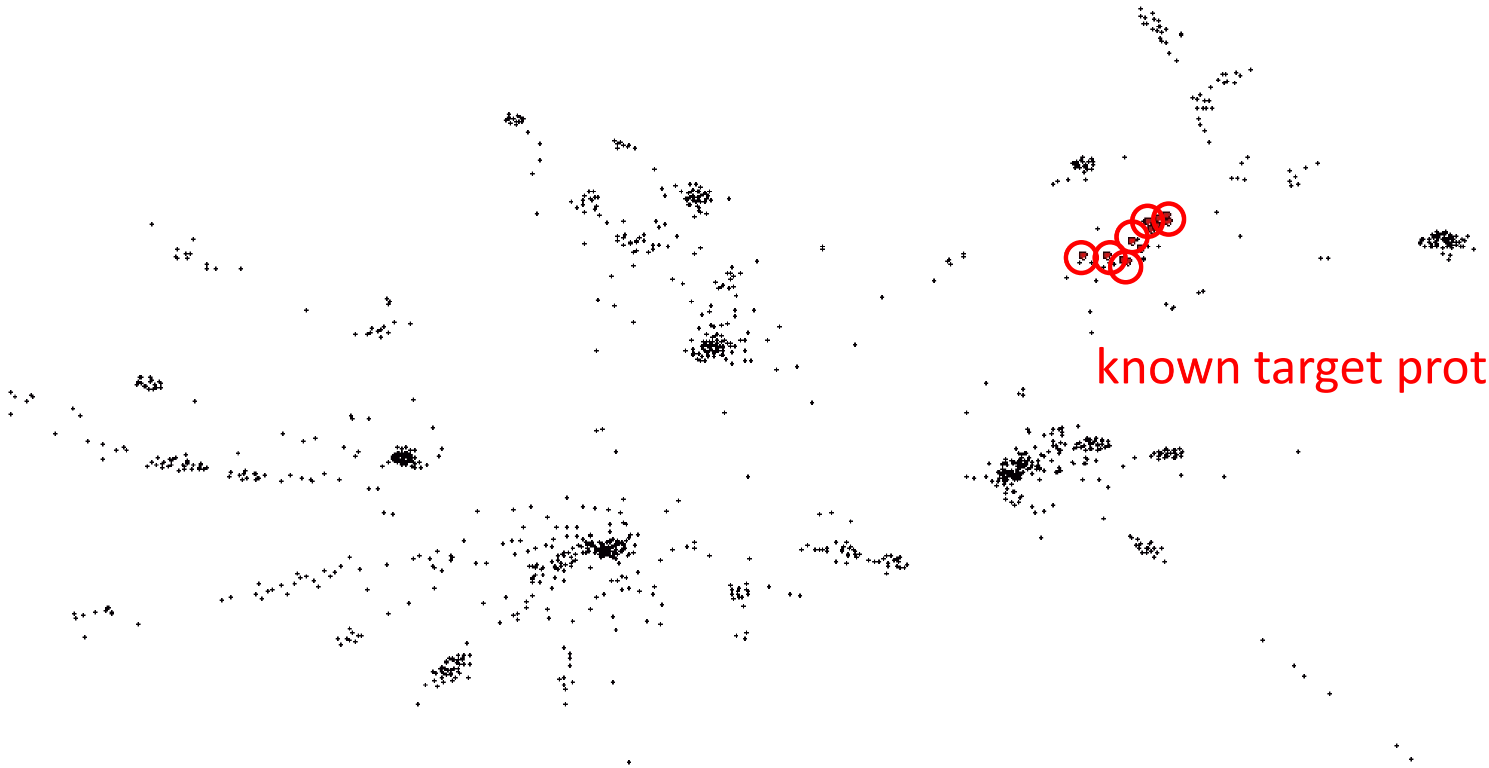
- clustering based on pairwise sequence similarities



known target proteins

Sequence clustering

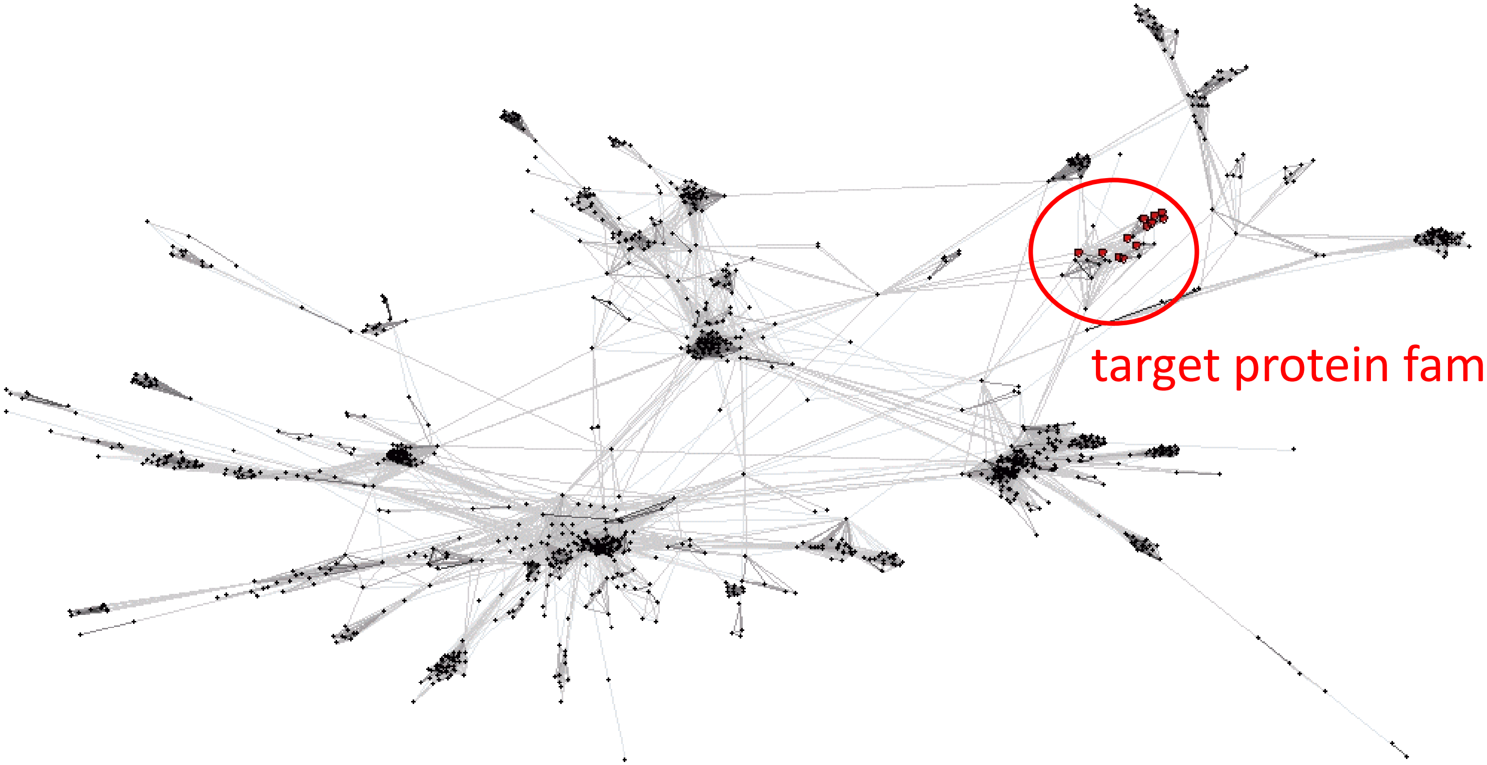
- clustering based on pairwise sequence similarities



known target proteins

Sequence clustering

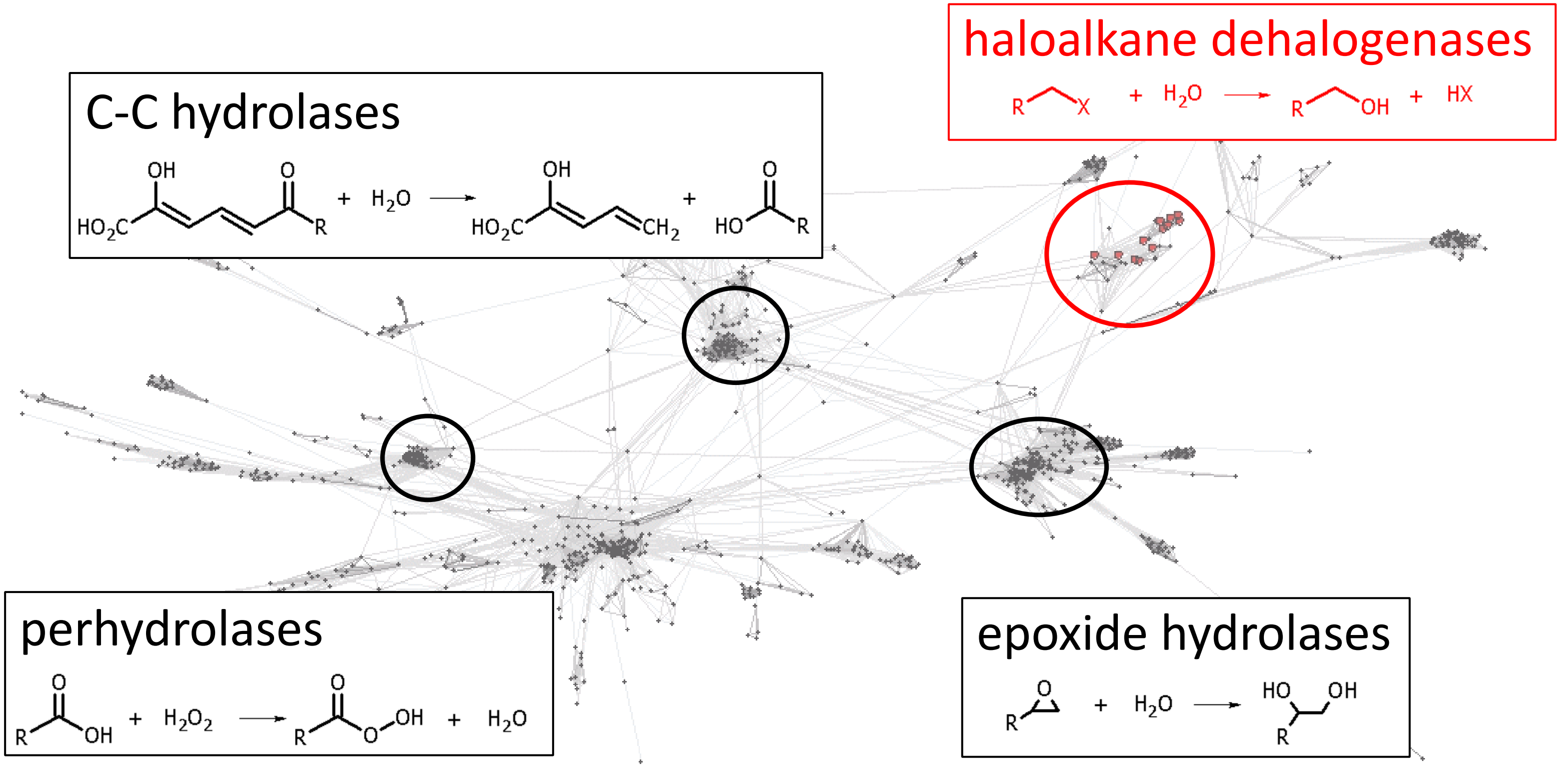
- clustering based on pairwise sequence similarities



target protein family

Sequence clustering

- clustering based on pairwise sequence similarities



C-C hydrolases

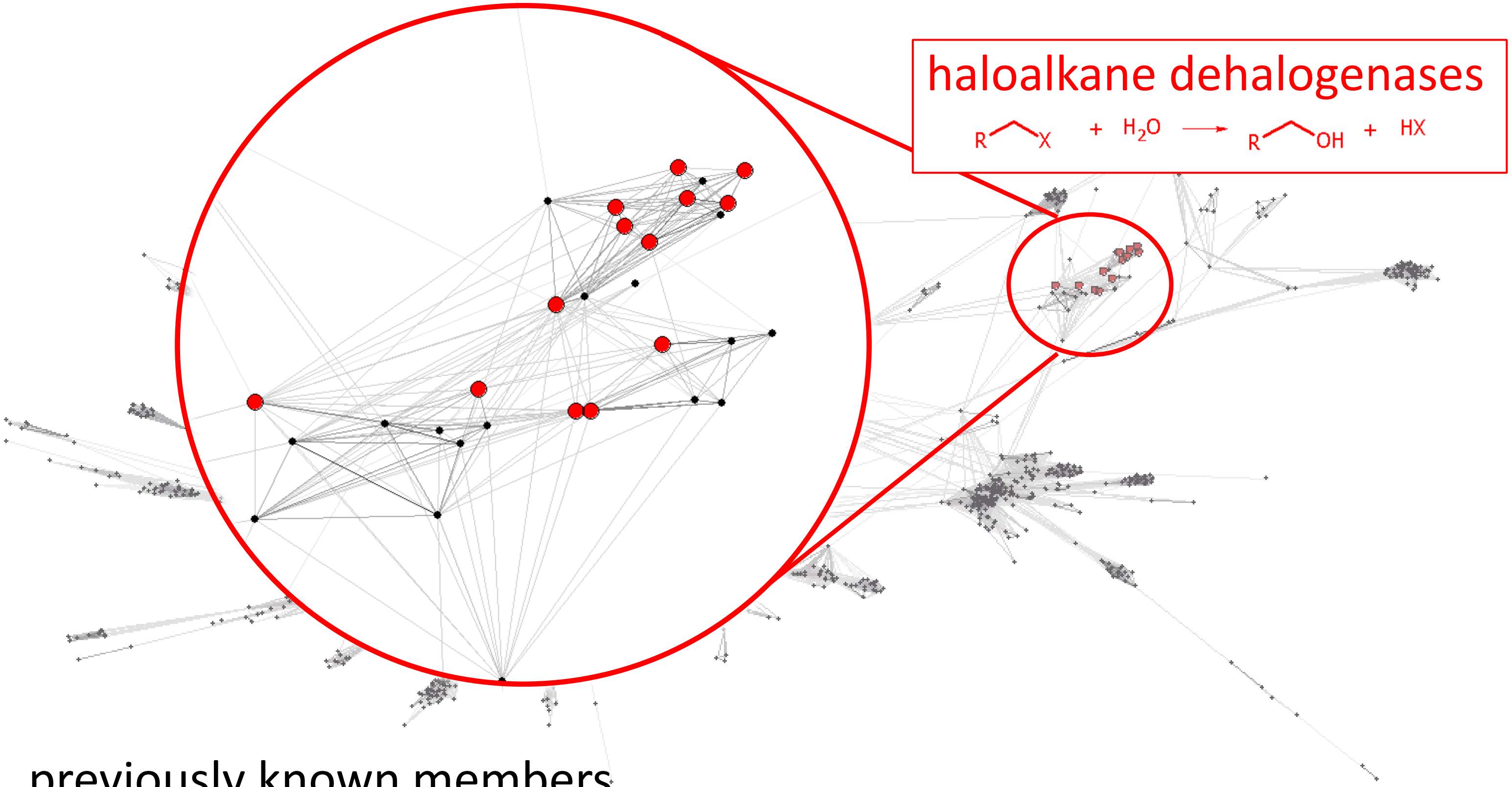
haloalkane dehalogenases

perhydrolases

epoxide hydrolases

Sequence clustering

- clustering based on pairwise sequence similarities



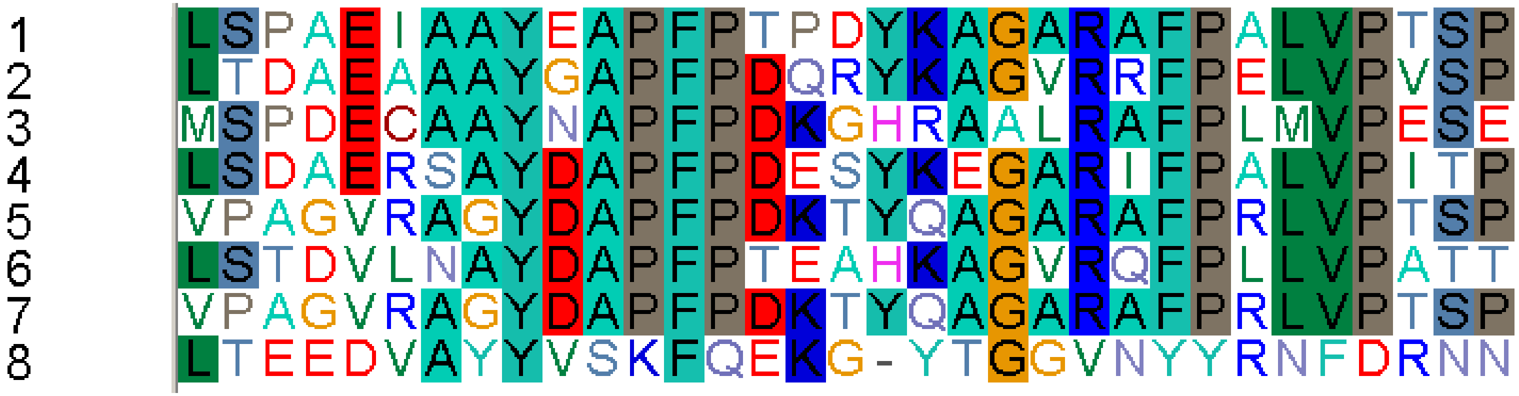
haloalkane dehalogenases

🔴 previously known members

● new family members

Sequence comparison

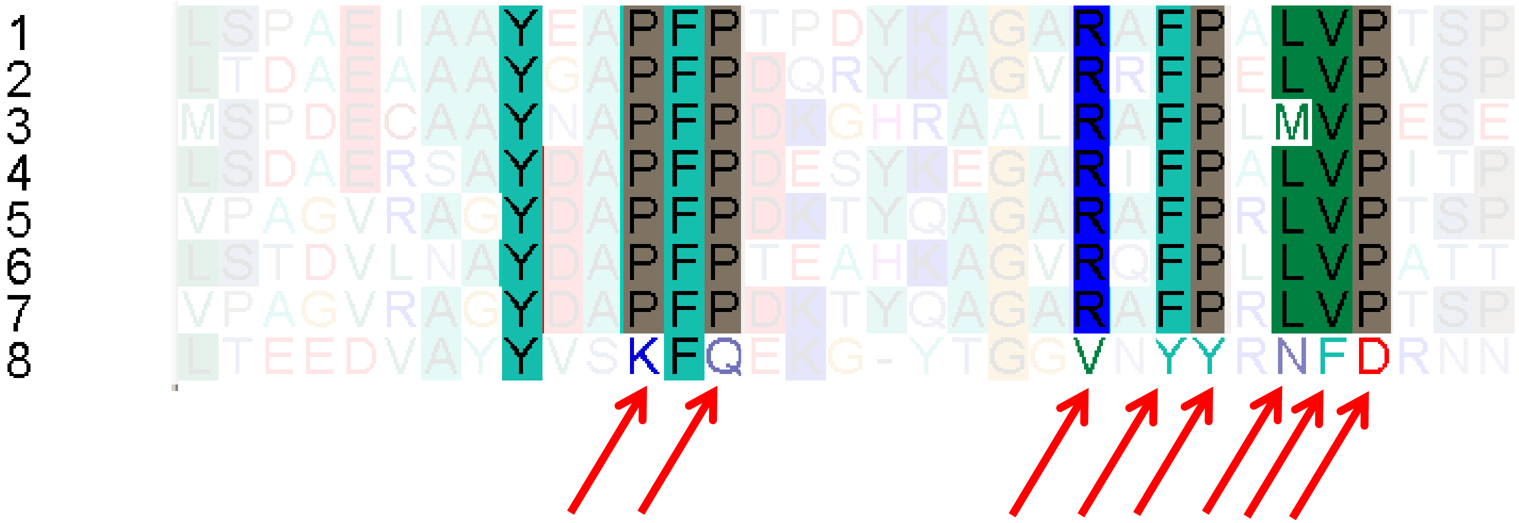
- multiple sequence alignment
    - analysis of conserved residues within protein family → identification of protein family members



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

Sequence comparison

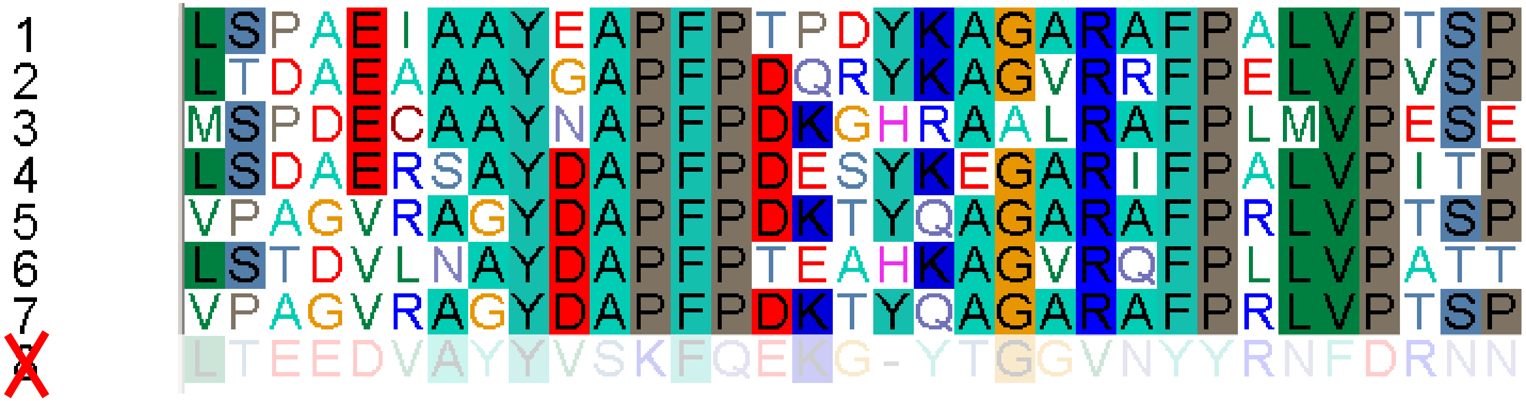
- multiple sequence alignment
  - analysis of conserved residues within protein family → identification of protein family members



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

Sequence comparison

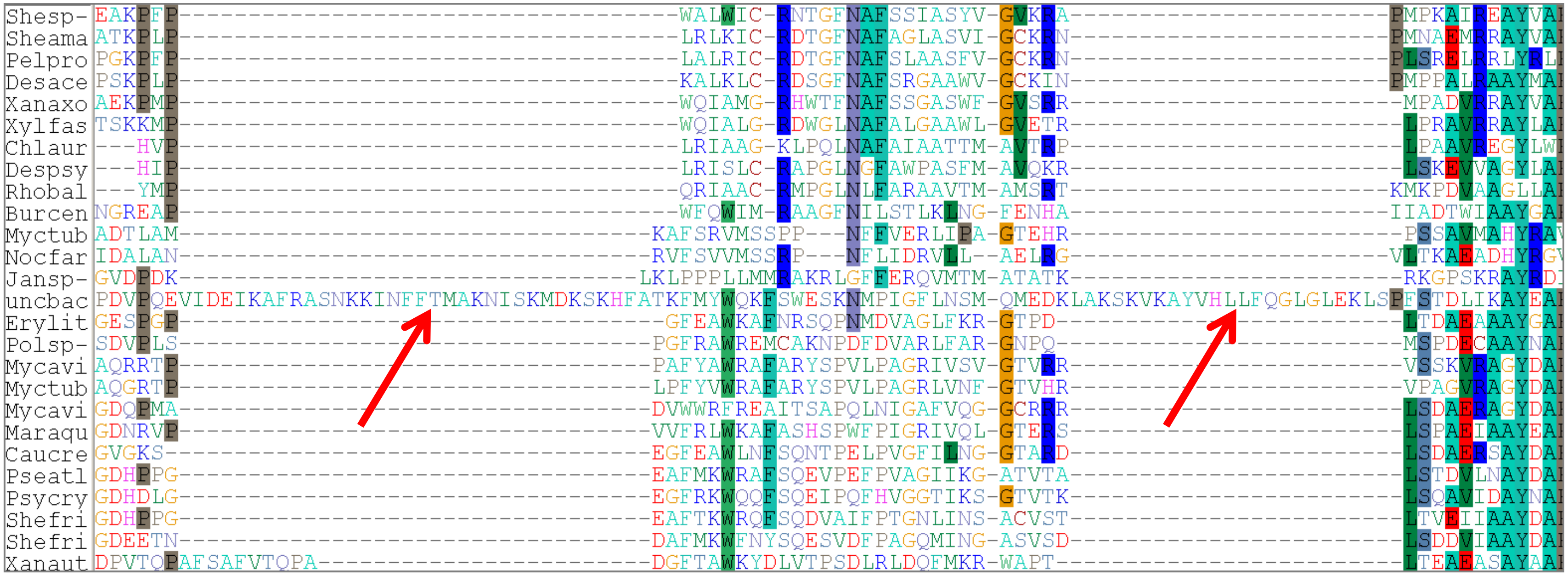
- multiple sequence alignment
  - analysis of conserved residues within protein family → identification of protein family members



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

Sequence comparison

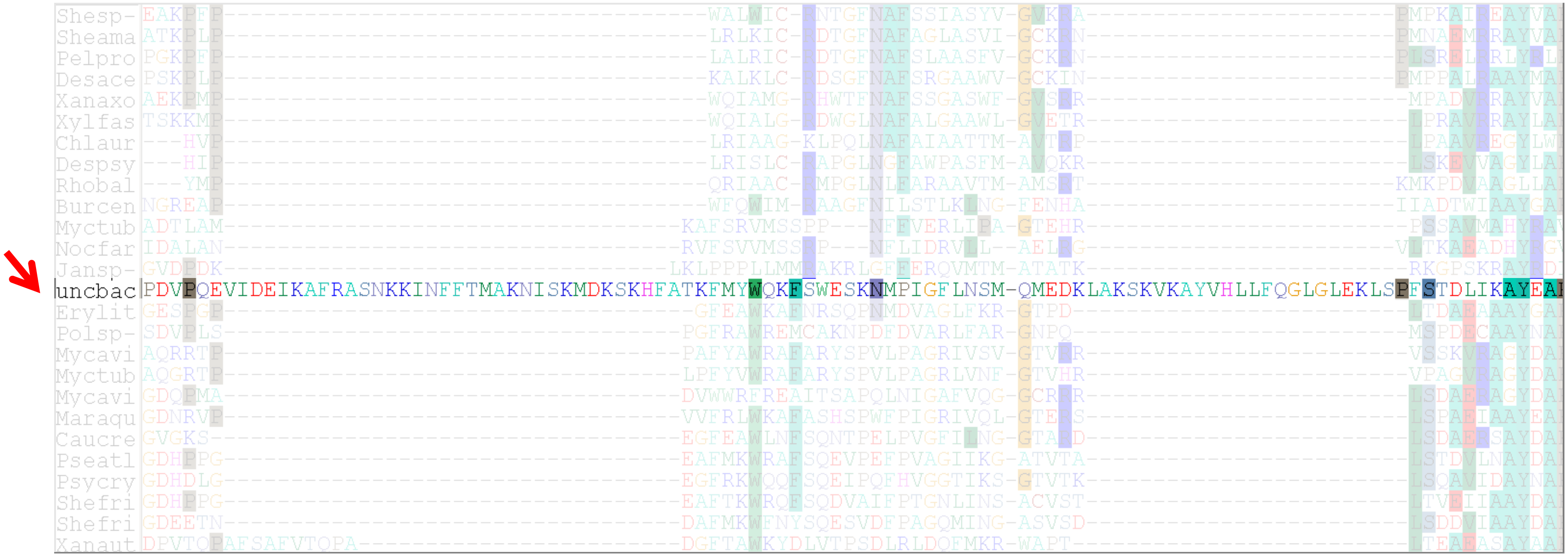
- multiple sequence alignment
  - identification of sequences with unique features → proteins with potentially novel characteristics



```
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

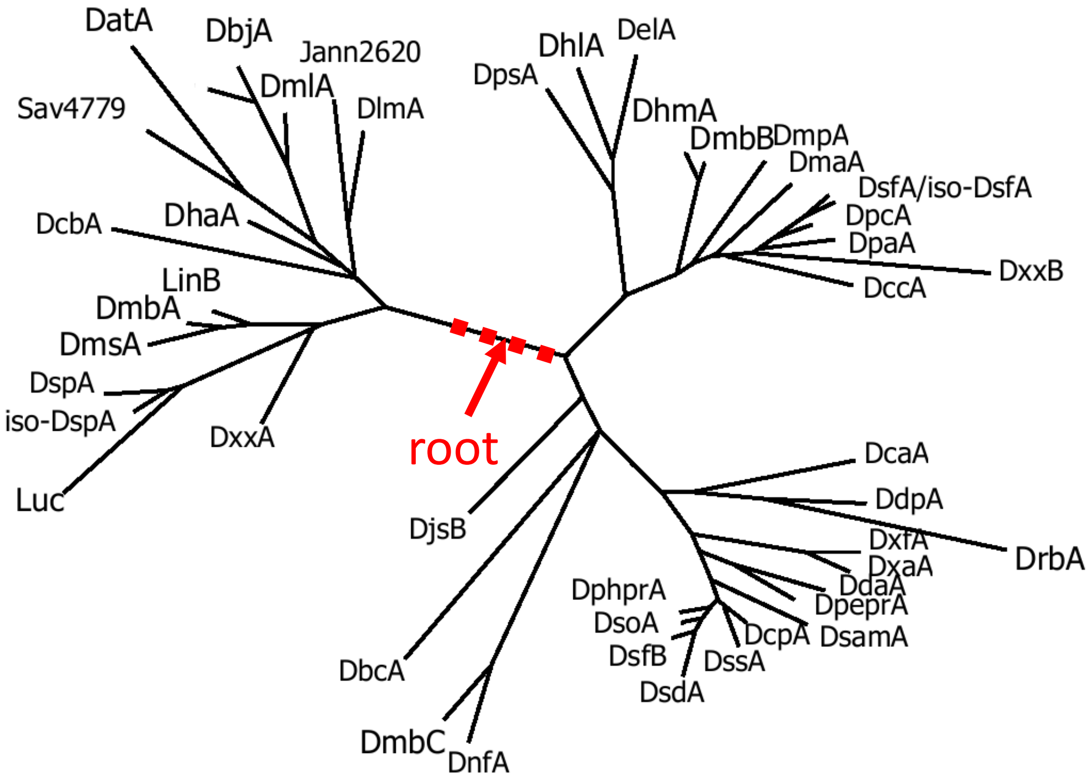
- multiple sequence alignment
  - identification of sequences with unique features → proteins with potentially novel characteristics



```
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
```

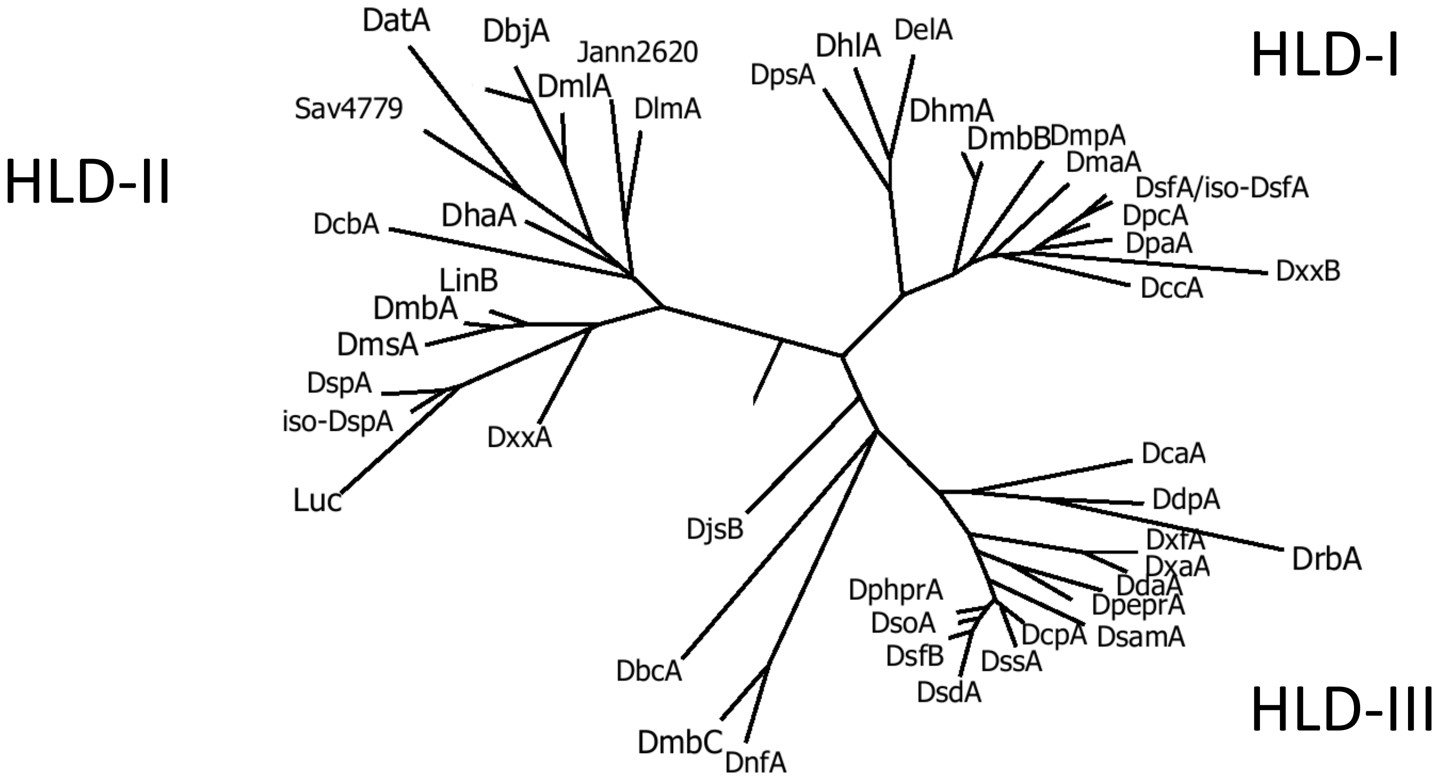
Sequence comparison

- phylogenetics
  - establishment of evolutionary relationships among sequences



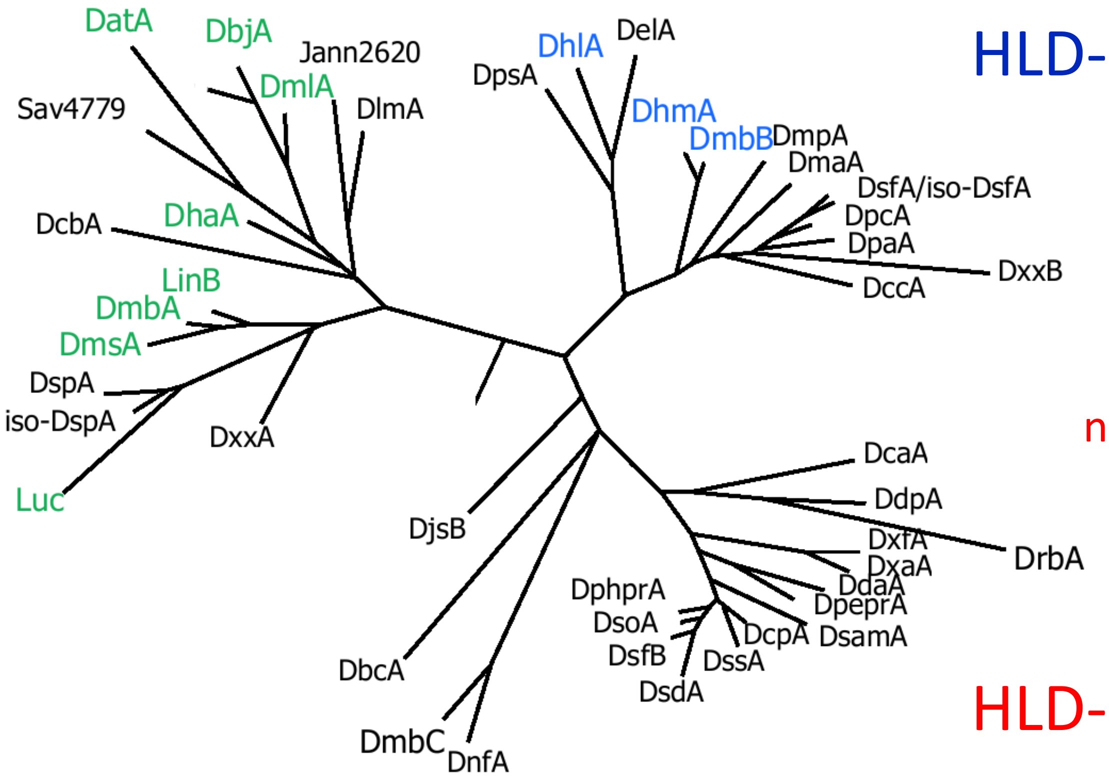
Sequence comparison

- phylogenetics
  - classification of sequences



Sequence comparison

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



HLD-II

HLD-I

no experimental data

HLD-III

Information about host organisms

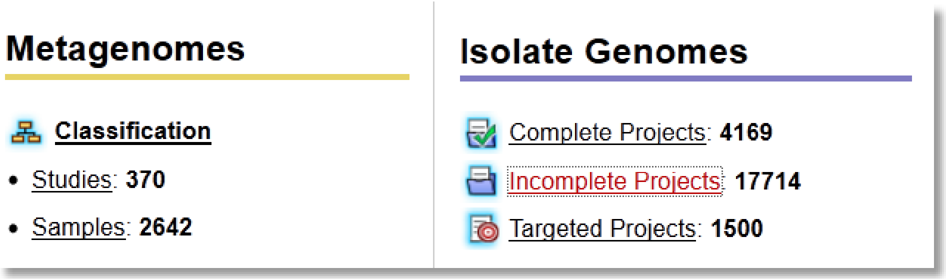
- extremophiles - microorganisms living in extreme conditions
  - geochemical extremes (pH, salinity)
  - physical extremes (temperature, pressure)
- proteins from extremophiles
  - often adapted to extreme conditions → unique characteristics, useful for practical applications

Information about host organisms

- Genomes OnLine Database (GOLD)
  - http://www.genomesonline.org/
  - list of complete (>6,000), ongoing (> 27,000) and targeted genome (>1,000) projects
  - information about individual projects and source organisms
- Entrez Genome
  - http://www.ncbi.nlm.nih.gov/sites/genome
  - provided by NCBI
  - data from more than 20,000 finished or ongoing genome projects (includes almost 10,000 organisms)
  - information about genome, source organism, genes, encoded proteins, graphical representations, ...

Information about host organisms

- GOLD



Metagenomes

Classification

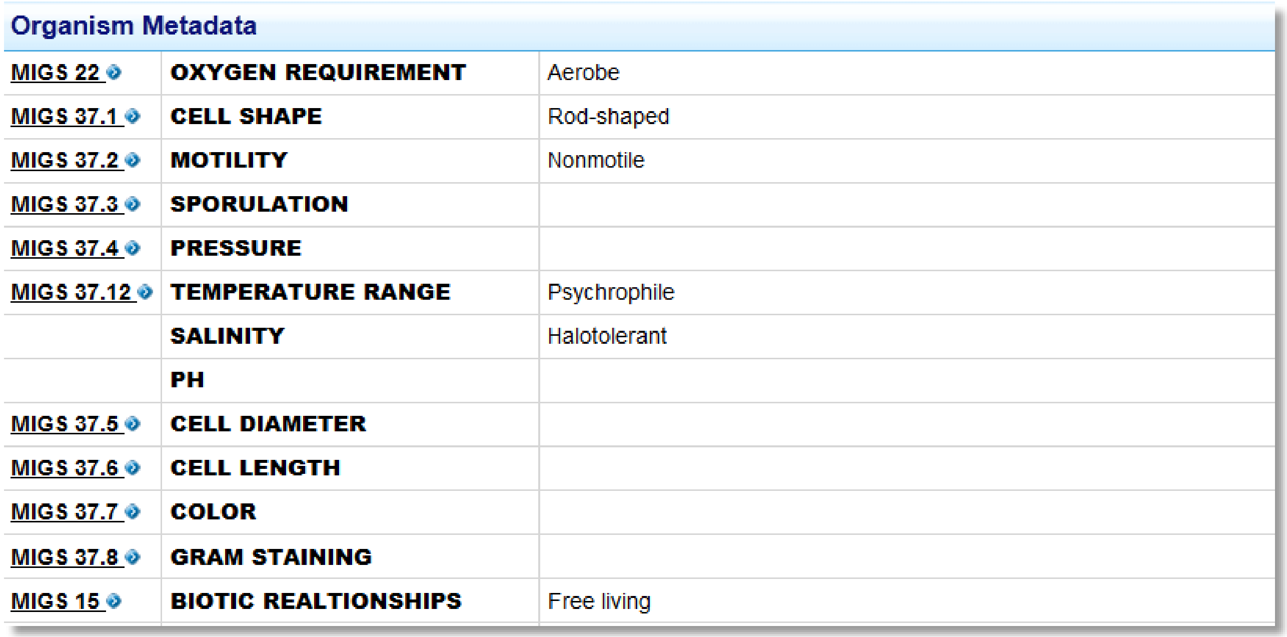
- Studies: **370**
- Samples: **2642**

Isolate Genomes

Complete Projects: **4169**

Incomplete Projects: **17714**

Targeted Projects: **1500**

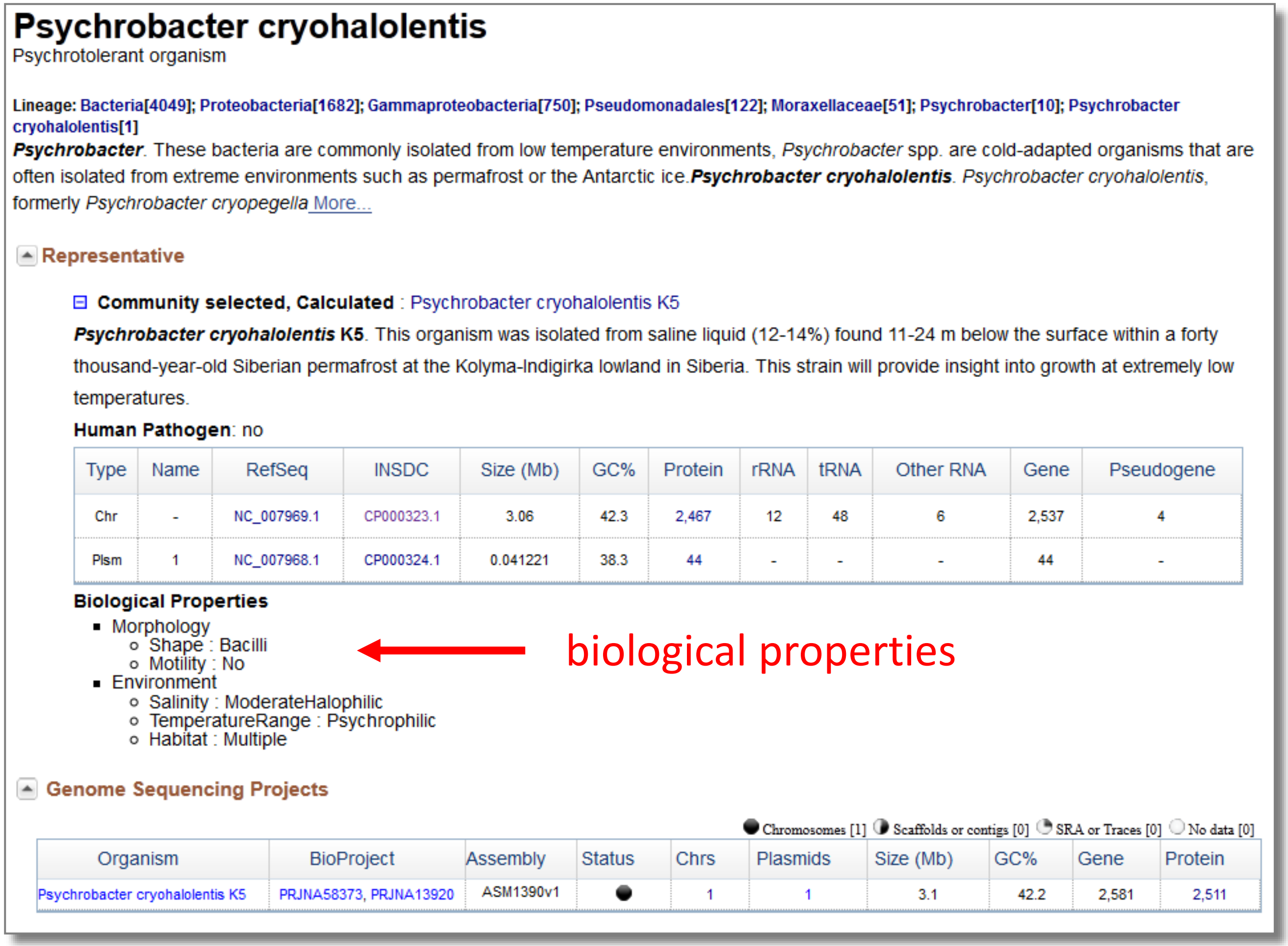


Organism Metadata

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

biological properties

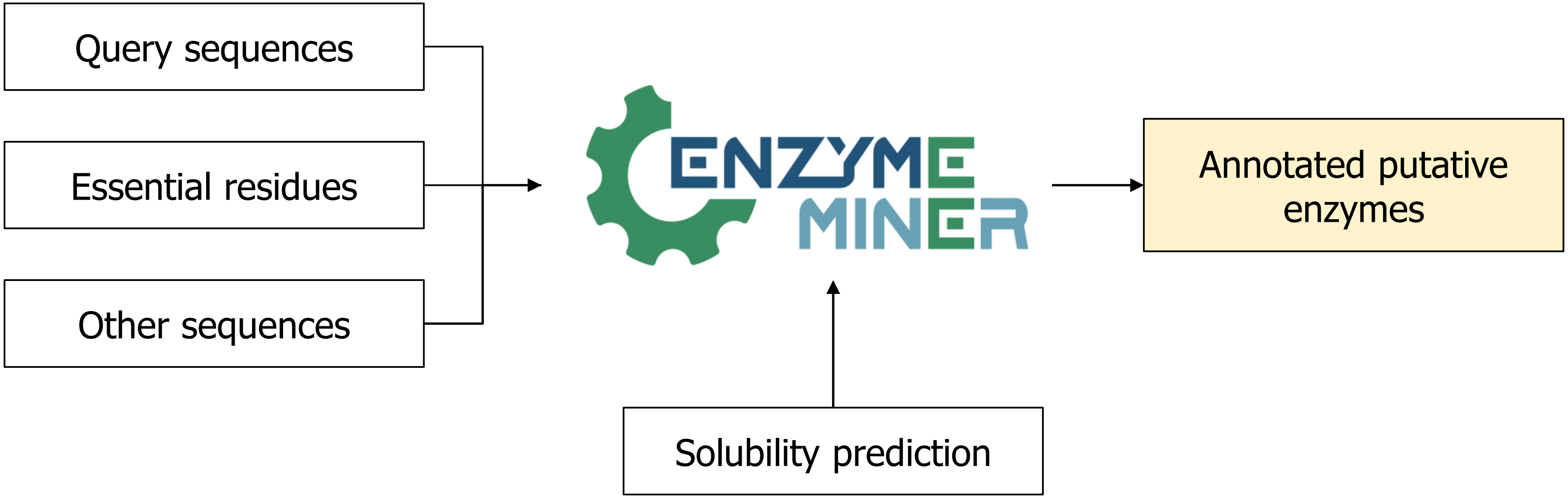
▲ **Genome Sequencing Projects**

● 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

Automated *in silico* enzyme identification

- ❑ Enzyme Miner
  - https://loschmidt.chemi.muni.cz/enzymeminer/



Query sequences

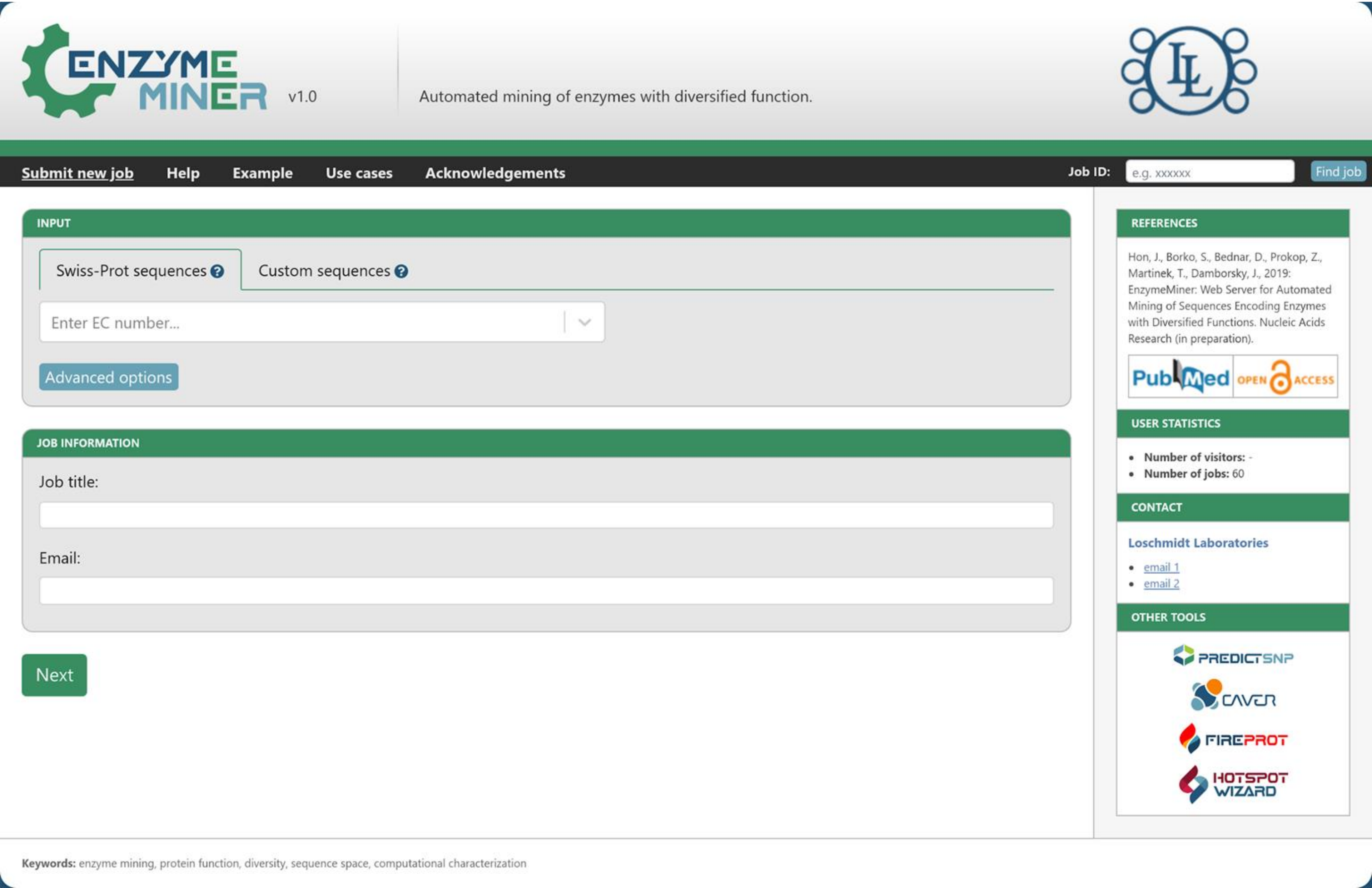
Essential residues

Other sequences

Annotated putative enzymes

Solubility prediction

Automated *in silico* enzyme identification



ENZYME MINER v1.0

Automated mining of enzymes with diversified function.

INPUT

Swiss-Prot sequences ? | Custom sequences ?

Enter EC number...

JOB INFORMATION

Job title:

Email:

REFERENCES

Hon, J., Borko, S., Bednar, D., Prokop, Z., Martinek, T., Damborsky, J., 2019: EnzymeMiner: Web Server for Automated Mining of Sequences Encoding Enzymes with Diversified Functions. Nucleic Acids Research (in preparation).

USER STATISTICS

- **Number of visitors:** -
- **Number of jobs:** 60

CONTACT

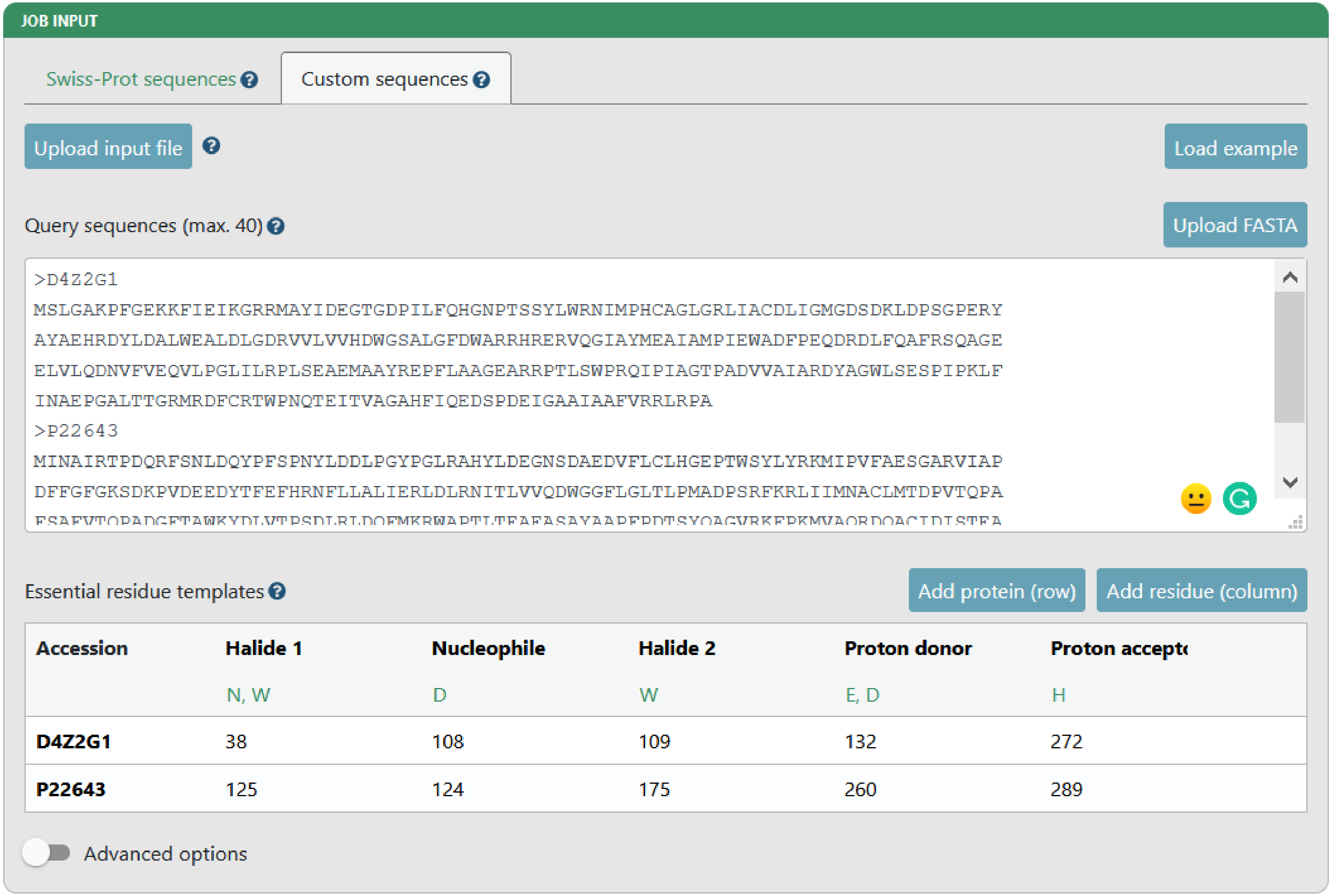
Loschmidt Laboratories

- email 1
- email 2

OTHER TOOLS

Keywords: enzyme mining, protein function, diversity, sequence space, computational characterization

Automated *in silico* enzyme identification



Query sequences (max. 40)

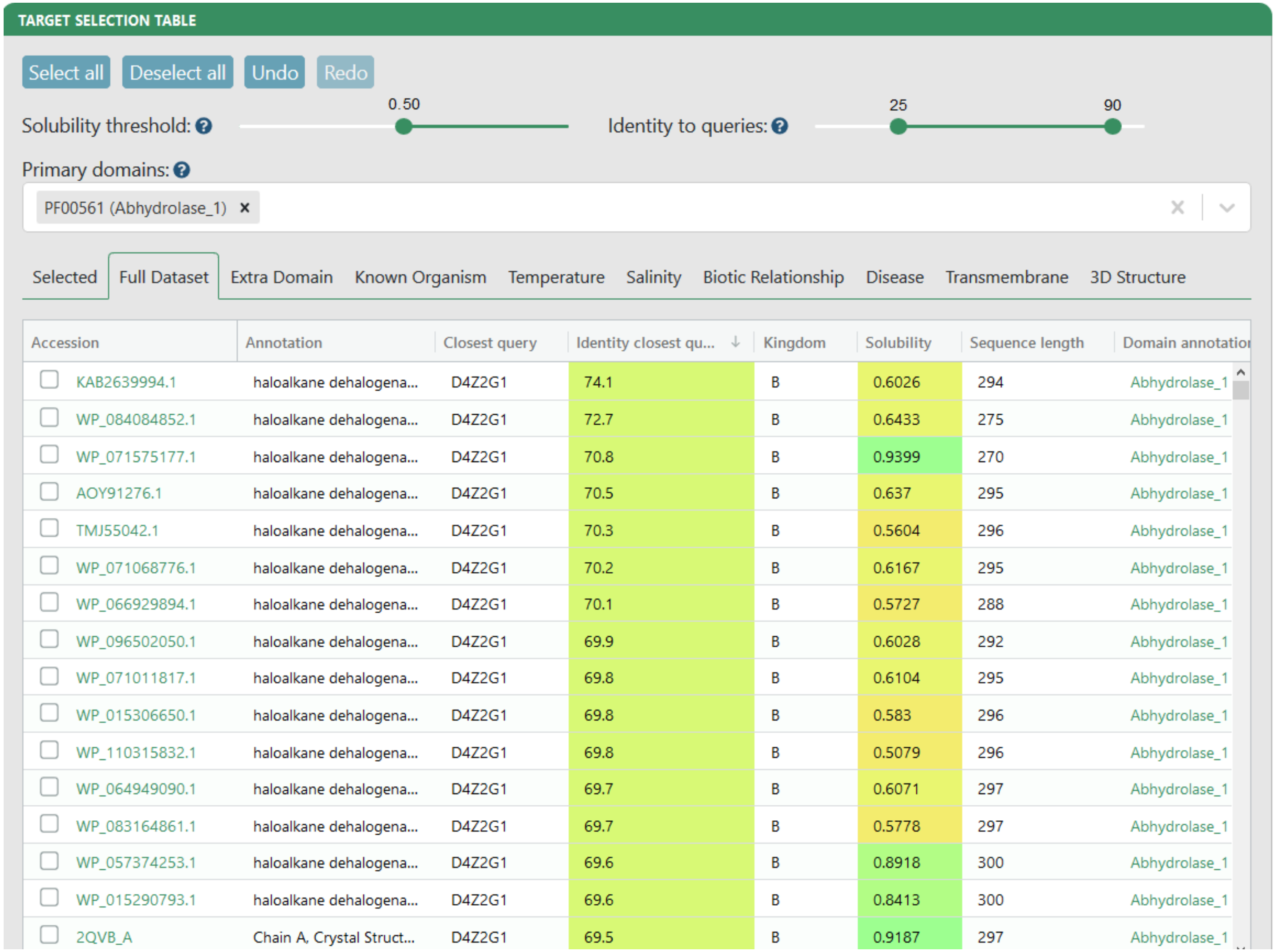
```
>D4Z2G1
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGRLIACDLIGMGDSDKLDPSGPERY
AYAEHRDYLDALWEALDLGDRVVLVVHDWGSALGFDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRSQAGE
ELVLQDNVFVEQVLPGLILRPLSEAEMAAYREPFLAAGEARRPTLSWPRQIPIAGTPADVVAIARDYAGWLSESPIPKLF
INAEPGALTTGRMRDFCRTWPNQTEITVAGAHFIQEDSPDEIGAAIAAFVRRLRPA
>P22643
MINAIRTPDQRFSNLDQYPFSPNYLDDLPGYPGLRAHYLDEGNSDAEDVFLCLHGEPTWSYLYRKMIPVFAESGARVIAP
DFFGFGKSDKPVDEEDYTFEFHRNFLLALIERLDLRNITLVVQDWGGFLGLTLPMADPSRFKRLIIMNACLMTDPVTQPA
FSAFVTQPADGFTAWKYDLVTPSDLRLDQFMKRWAPTLTEAEASAYAAPFPDTSYQAGVRKFPKMVAQRDQACIDISTEA
```

Essential residue templates

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

Advanced options

Automated *in silico* enzyme identification



TARGET SELECTION TABLE

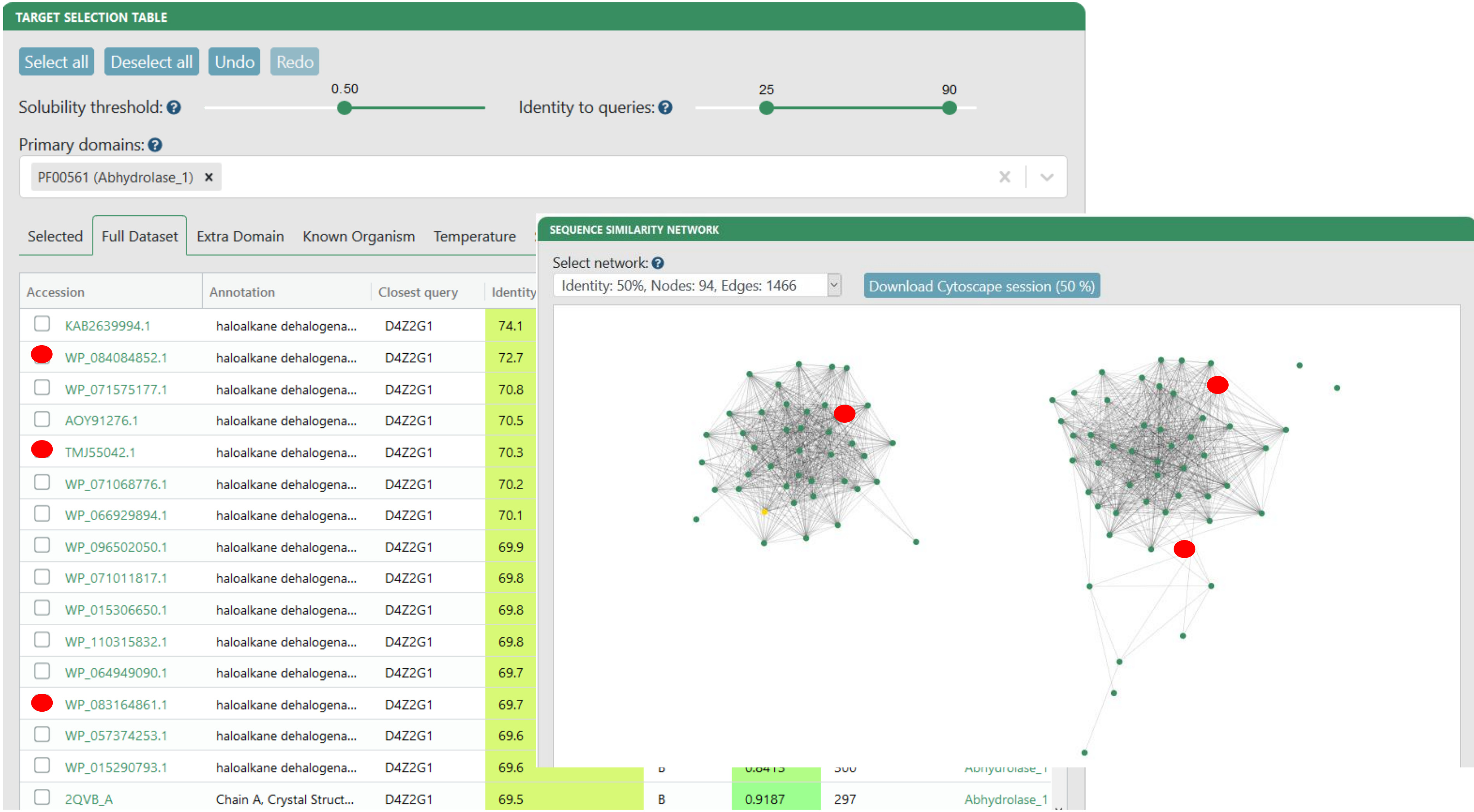
Solubility threshold: 0.50

Identity to queries: 25 – 90

Primary domains: PF00561 (Abhydrolase_1) ×

	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	B	0.9399	270	Abhydrolase_1
□	AOY91276.1	haloalkane dehalogena...	D4Z2G1	70.5	B	0.637	295	Abhydrolase_1
□	TMJ55042.1	haloalkane dehalogena...	D4Z2G1	70.3	B	0.5604	296	Abhydrolase_1
□	WP_071068776.1	haloalkane dehalogena...	D4Z2G1	70.2	B	0.6167	295	Abhydrolase_1
□	WP_066929894.1	haloalkane dehalogena...	D4Z2G1	70.1	B	0.5727	288	Abhydrolase_1
□	WP_096502050.1	haloalkane dehalogena...	D4Z2G1	69.9	B	0.6028	292	Abhydrolase_1
□	WP_071011817.1	haloalkane dehalogena...	D4Z2G1	69.8	B	0.6104	295	Abhydrolase_1
□	WP_015306650.1	haloalkane dehalogena...	D4Z2G1	69.8	B	0.583	296	Abhydrolase_1
□	WP_110315832.1	haloalkane dehalogena...	D4Z2G1	69.8	B	0.5079	296	Abhydrolase_1
□	WP_064949090.1	haloalkane dehalogena...	D4Z2G1	69.7	B	0.6071	297	Abhydrolase_1
□	WP_083164861.1	haloalkane dehalogena...	D4Z2G1	69.7	B	0.5778	297	Abhydrolase_1
□	WP_057374253.1	haloalkane dehalogena...	D4Z2G1	69.6	B	0.8918	300	Abhydrolase_1
□	WP_015290793.1	haloalkane dehalogena...	D4Z2G1	69.6	B	0.8413	300	Abhydrolase_1
□	2QVB_A	Chain A, Crystal Struct...	D4Z2G1	69.5	B	0.9187	297	Abhydrolase_1

Automated *in silico* enzyme identification



TARGET SELECTION TABLE

Solubility threshold: 0.50

Identity to queries: 25 – 90

Primary domains: PF00561 (Abhydrolase_1)

	Accession	Annotation	Closest query	Identity				
□	KAB2639994.1	haloalkane dehalogena...	D4Z2G1	74.1				
●	WP_084084852.1	haloalkane dehalogena...	D4Z2G1	72.7				
□	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	B	0.8415	300	Abhydrolase_1
□	2QVB_A	Chain A, Crystal Struct...	D4Z2G1	69.5	B	0.9187	297	Abhydrolase_1

SEQUENCE SIMILARITY NETWORK

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

Ancestral sequence reconstruction

- Proteins with exceptional properties
  - improved stability, yields, specificity…
- Resurrection of the most probable protein sequences from past
  - selection of homologous sequences
  - multiple sequence alignment
  - construction of phylogenetic tree
  - reconstruction of ancestral sequences

Ancestral sequence reconstruction

I. Collect sequences and align them

..AKDKLNQP..

..AECKLNQP..

..AECCLNQP..

..AEDKLNQP..

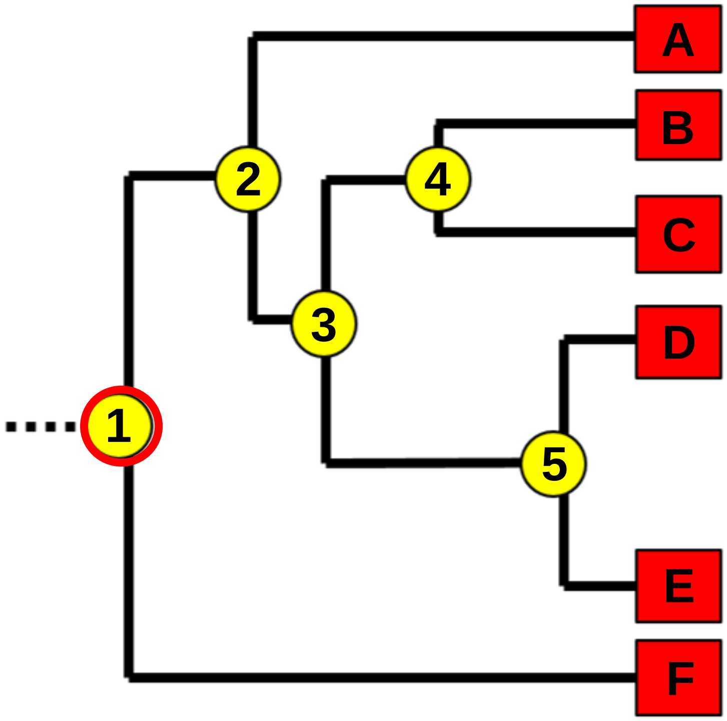
..VEDPLNQP..

..QEDKLNQP..

Ancestral sequence reconstruction

II. Create a tree

I. Collect sequences and align them



..AKDKLNQP..

..AECKLNQP..

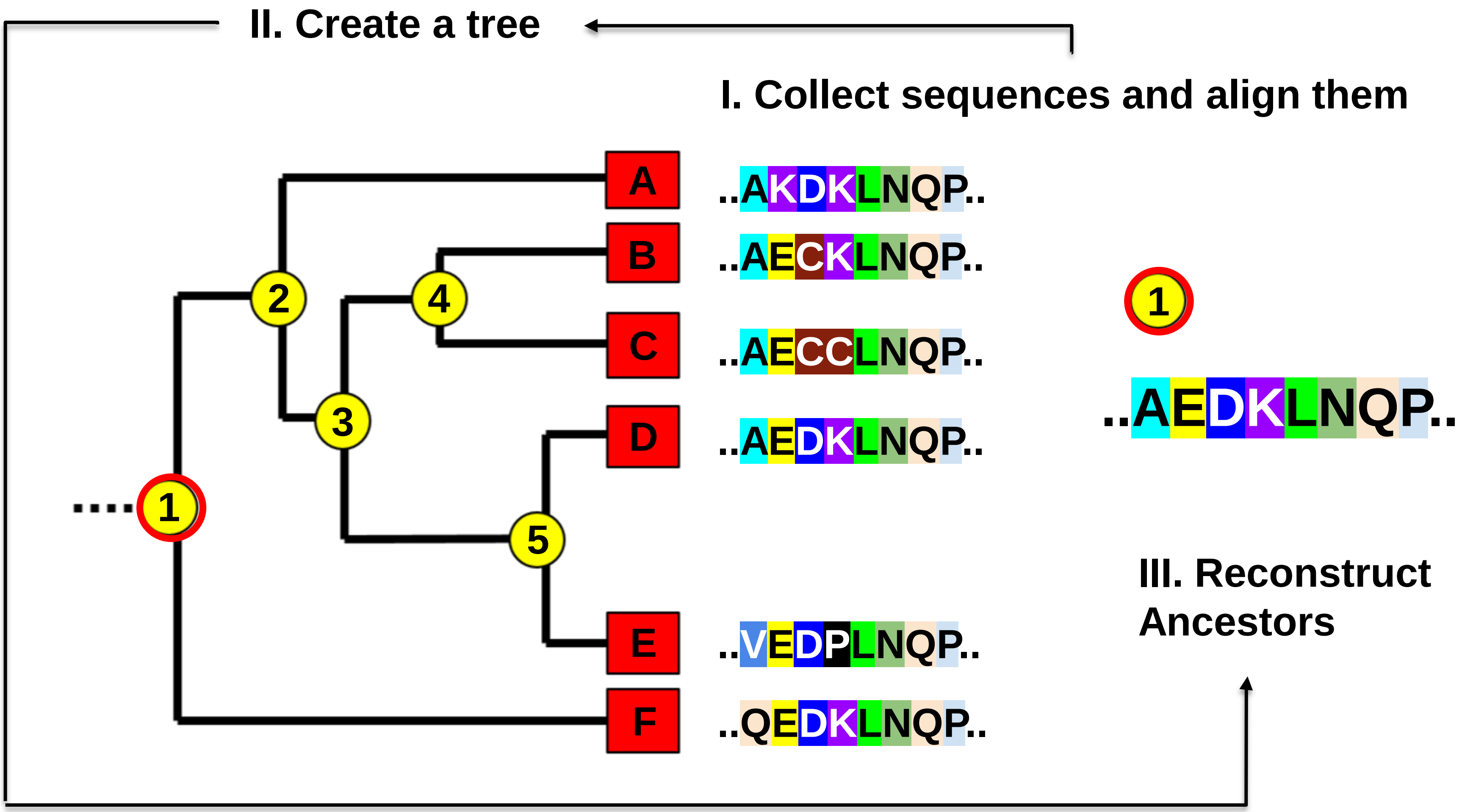
..AECCLNQP..

..AEDKLNQP..

..VEDPLNQP..

..QEDKLNQP..

Ancestral sequence reconstruction



II. Create a tree

I. Collect sequences and align them

III. Reconstruct Ancestors

..AKDKLNQP..

..AECKLNQP..

..AECCLNQP..

..AEDKLNQP..

..VEDPLNQP..

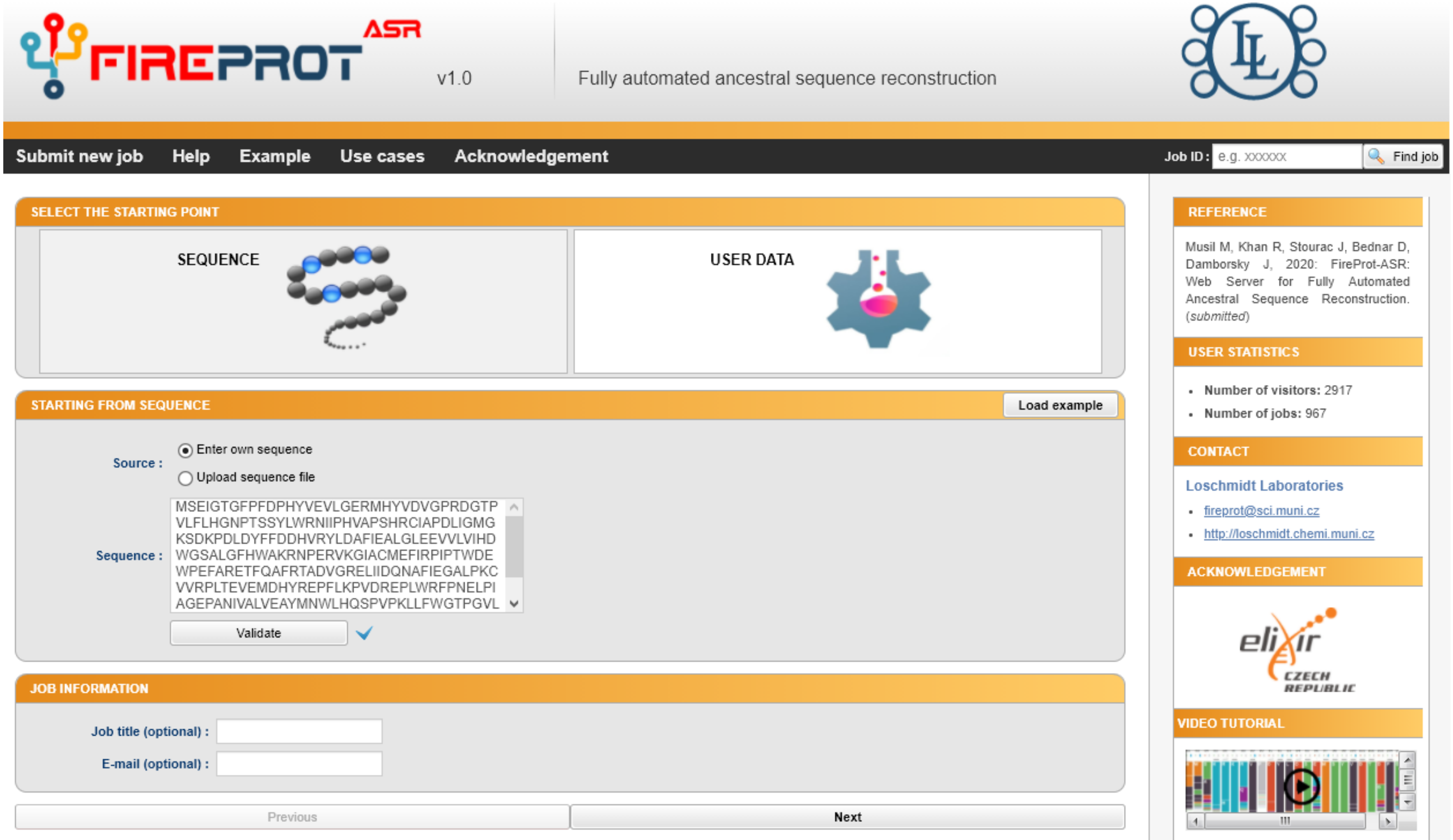
..QEDKLNQP..

1

..AEDKLNQP..

Ancestral sequence reconstruction

- FireProt^{ASR} web server
  - https://loschmidt.chemi.muni.cz/fireprotasr/



Fully automated ancestral sequence reconstruction

SELECT THE STARTING POINT

STARTING FROM SEQUENCE

Source : ● Enter own sequence ○ Upload sequence file

Sequence :

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

JOB INFORMATION

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

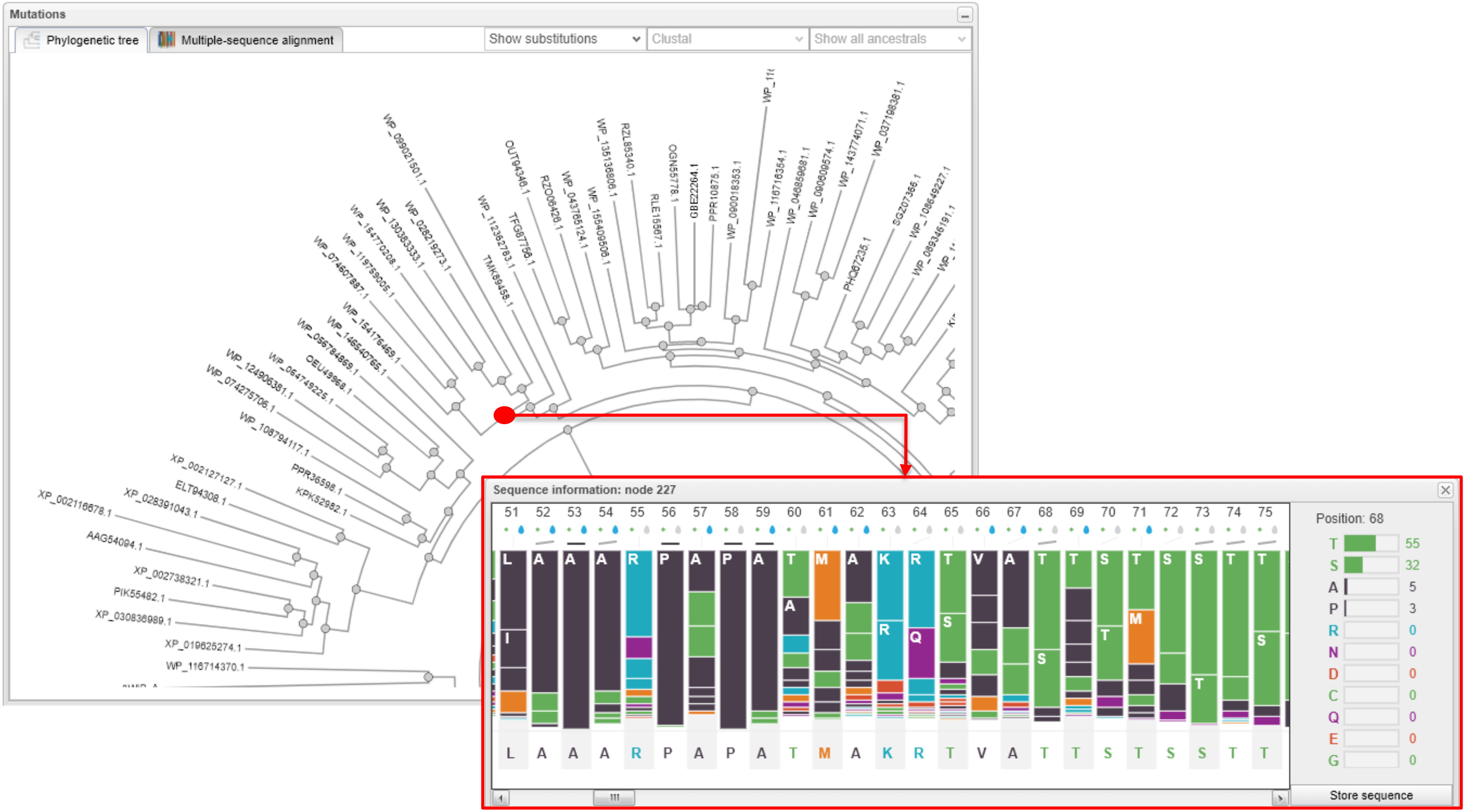
ACKNOWLEDGEMENT

elixir CZECH REPUBLIC

VIDEO TUTORIAL

Ancestral sequence reconstruction

- FireProt^{ASR} web server
  - https://loschmidt.chemi.muni.cz/fireprotasr/



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

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
  - **combination** of both approaches: optimal strategy to filter false positives
- selection of proteins for experimental characterization
  - **clustering:** classification and filtering of hits from database searches - CLANS
  - **sequence comparison:** classification and identification of unique sequences
  - sequences from **extremophiles:** potentially adapted to extreme conditions
  - **Enzyme Miner:** automated identification of interesting catalysts
  - **Ancestral** protein sequences with interesting properties

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 ENGINEERING

3. PREPARATION OF RECOMBINANT PROTEINS, PROTEIN EXPRESSION AND PURIFICATION

Loschmidt Laboratories
Department of Experimental Biology
Masaryk University, Brno