

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?

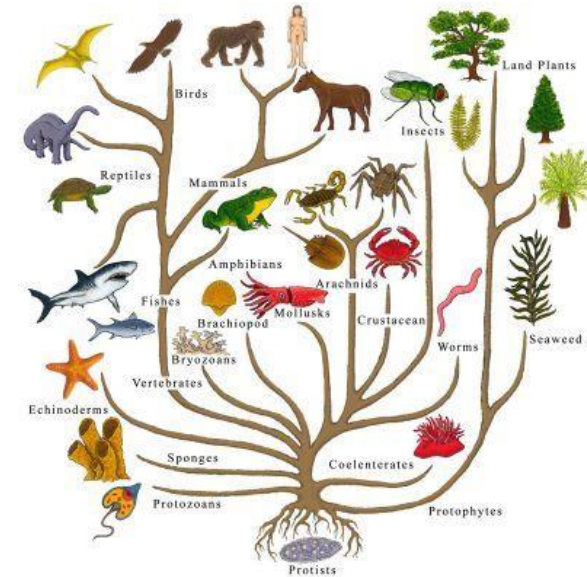
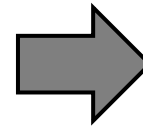
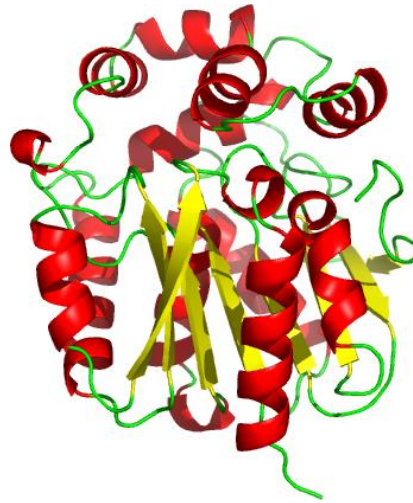
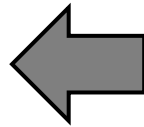
- plenty of reasons

# Why to search for new proteins?

## Optimization of protein for application



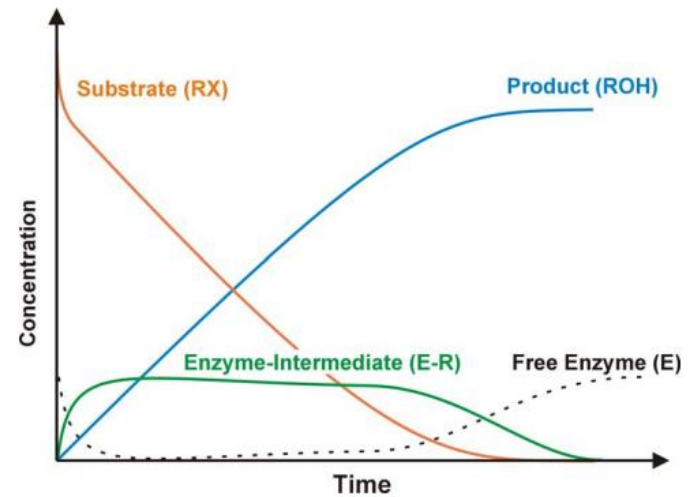
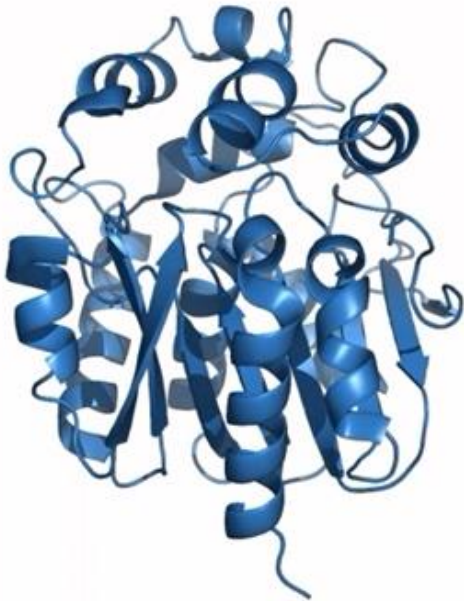
Protein engineering



Natural diversity

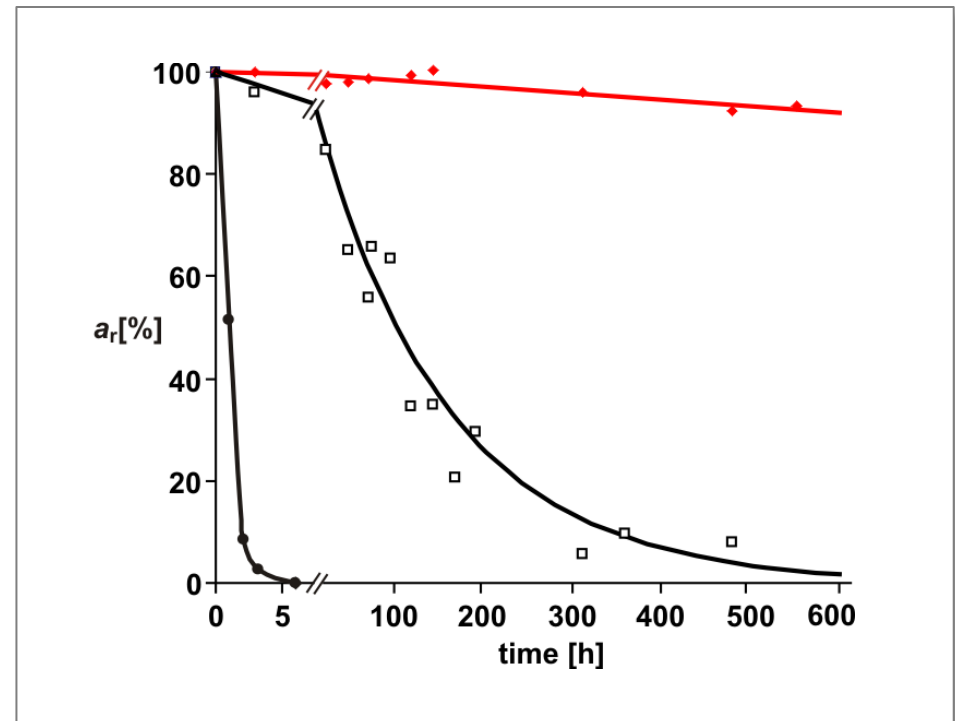
# Why to search for new proteins?

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



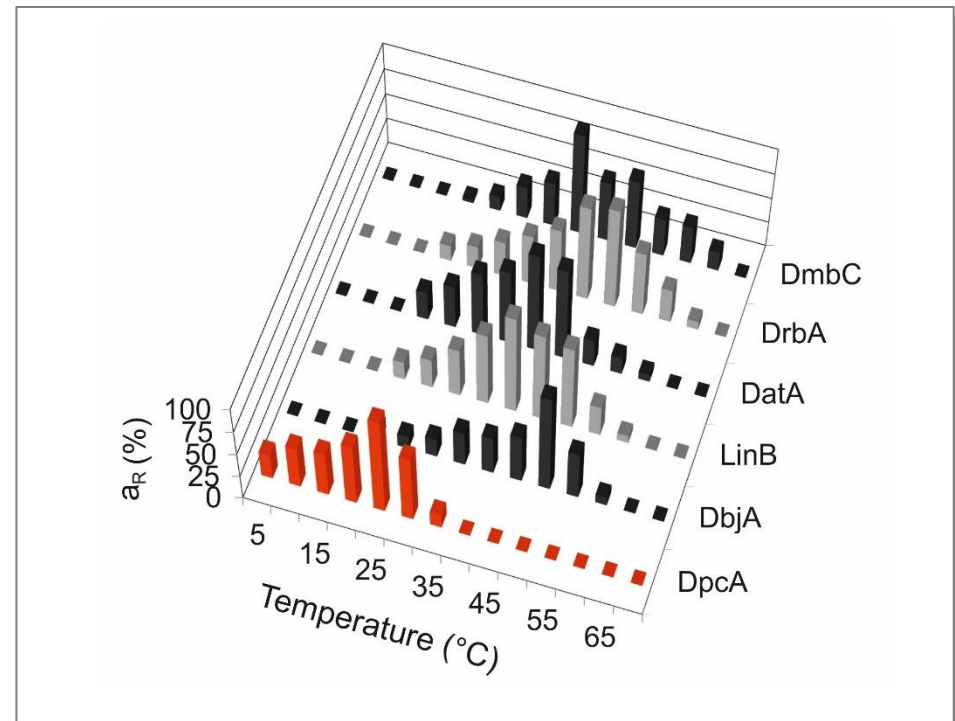
# Why to search for new proteins?

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



# Why to search for new proteins?

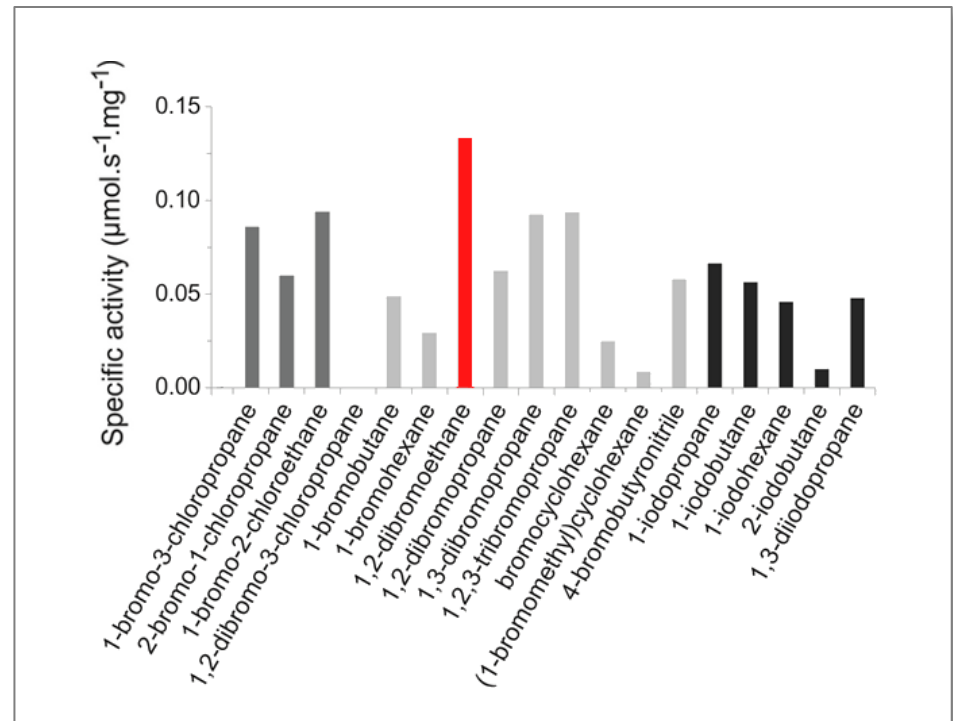
- better understanding of structure-function relationships
- novel properties
  - stability
  - temperature profile





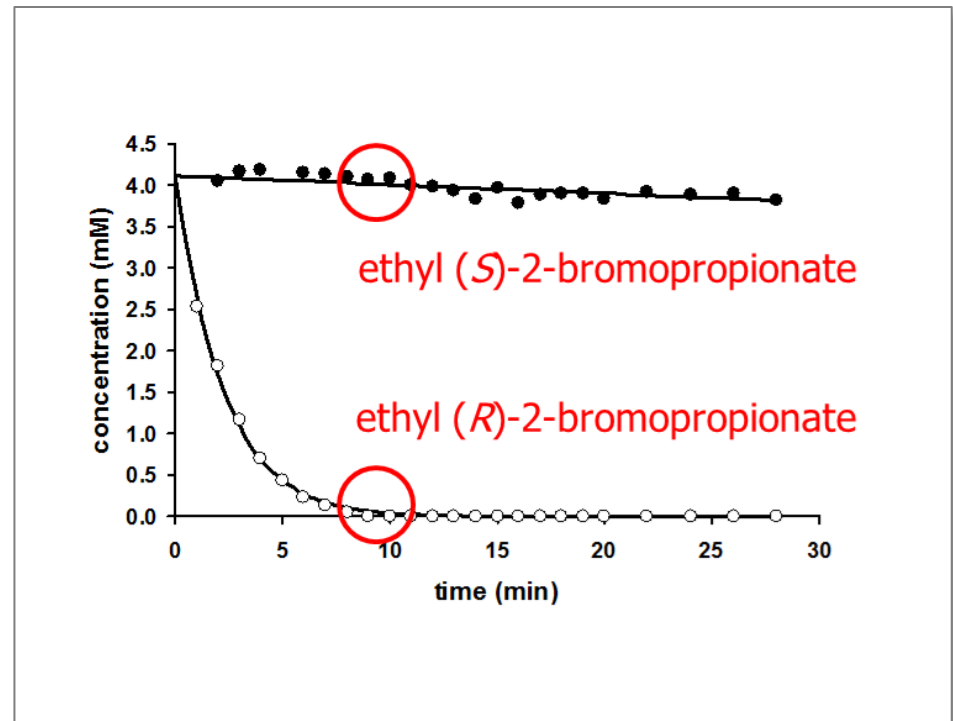
# Why to search for new proteins?

- 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**

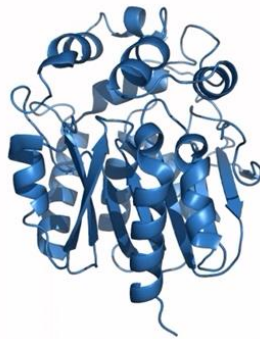


# Why to search for new proteins?

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

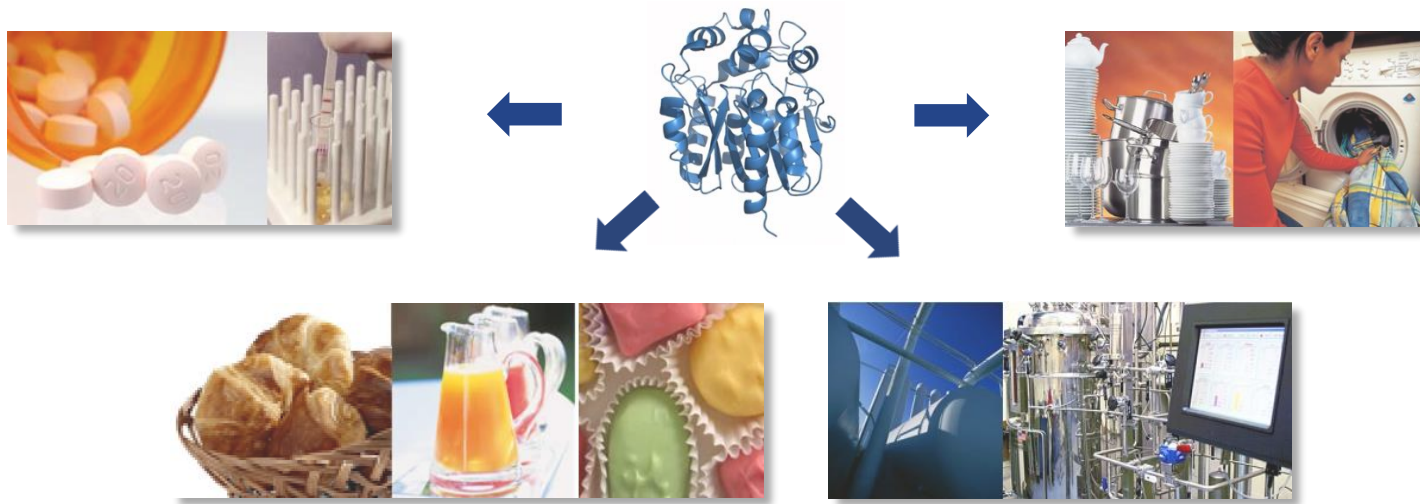
# Why to search for new proteins?

- ❑ 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?

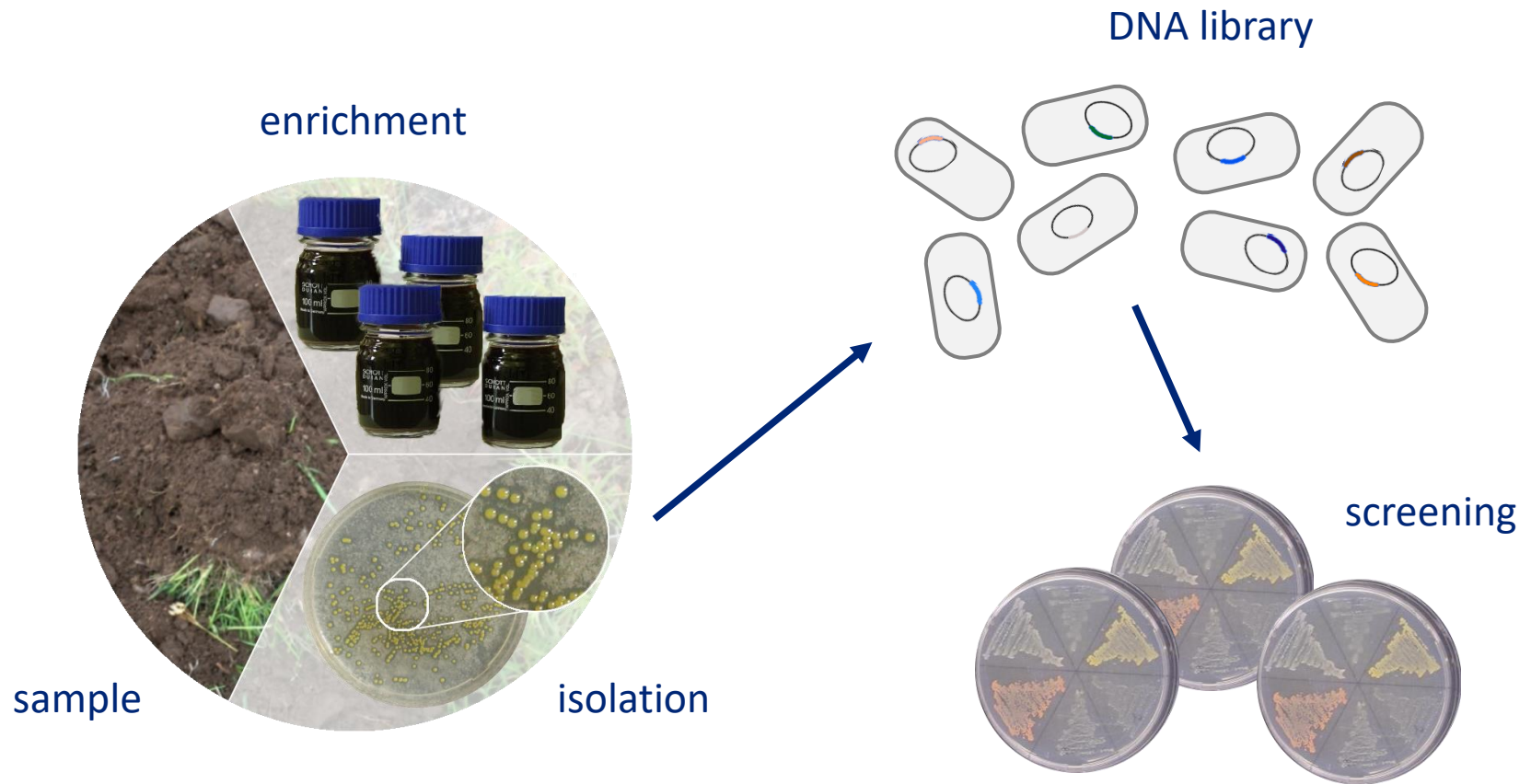


## How to acquire new proteins?

- traditional approach
- metagenomic approach
- bioinformatic approach

# How to acquire new proteins?

- traditional approach





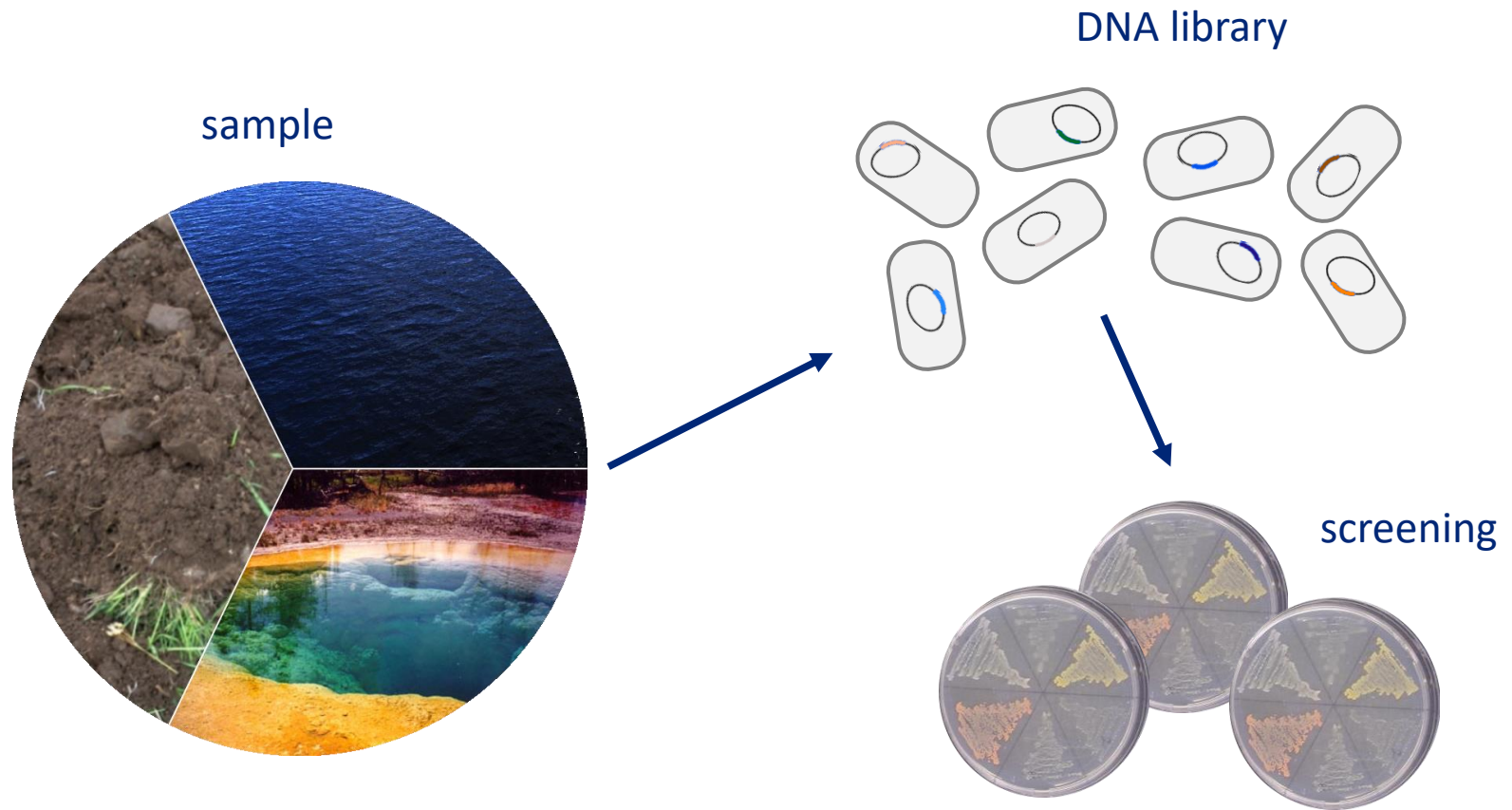
# How to acquire new proteins?

## ❑ 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



# How to acquire new proteins?

## □ metagenomic approach

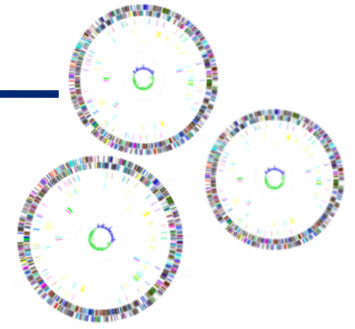
- isolation and cloning of **DNA** extracted directly **from environmental sample** (without culturing the present organisms)
- genes recovered by DNA library screening or PCR with specific primers,...
- 😊 enables to explore biodiversity of uncultured microorganisms

# How to acquire new proteins?

## bioinformatic approach

sequence database

(meta)genomic sequencing projects



*in silico* "screening"

1: [ABI93216](#). Reports: LinB [Xanthomonas...[gi:115291795]

```
>gi|115291795|gb|ABI93216.1| LinB [Xanthomonas sp. ICHL2]
MILGANPFGKGFIEIKGRMAYIDEGTGDPILFQHGNETS SYLGRNIMPHCAGLGELIACDLIGMGDSD
KLDPSGPEKTYAEHRDYLDAWEALDLGDWLVVVDGQSVLGFDAERHREEQGIAZYMEVITMPLDQ
ADFPQQRDLFQAFRSQAGEELVLQDNVFEQLPLGLILRPLSEADMAAYREPLAAGEAREPTLSQVFRQ
IPIAGTPADQWAIARDYAGOLSESPIPKLFINAEPGALTTGRMEDFCRTWPNQTEITVAGAHFIQEDSPD
EIGAAIAAFVRR
```

2: [AAR05978](#). Reports: LinB [Sphingomonas...[gi:37963683]

```
>gi|37963683|gb|AAR05978.1| LinB [Sphingomonas paucimobiliis]
MSLGANPFGKGFIEIKGRMAYIDEGTGDPILFQHGNETS SYLGRNIMPHCAGLGELIACDLIGMGDSD
KLDPSGPEKTYAEHRDYLDAWEALDLGDWLVVVDGQSVLGFDAERHREEQGIAZYMEVITMPLDQ
ADFPQQRDLFQAFRSQAGEELVLQDNVFEQLPLGLILRPLSEADMAAYREPLAAGEAREPTLSQVFRQ
IPIAGTPADQWAIARDYAGOLSESPIPKLFINAEPGALTTGRIRDFCRTWPNQTEITVAGAHFIQEDSPD
EIGAAIAAFVRELRPA
```

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)



## □ 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 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



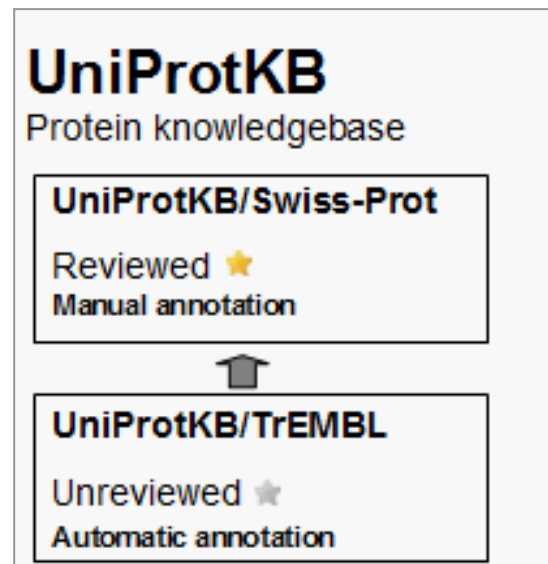
# Databases of protein sequences

- ❑ 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)



# 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” ~ **560,000** sequences

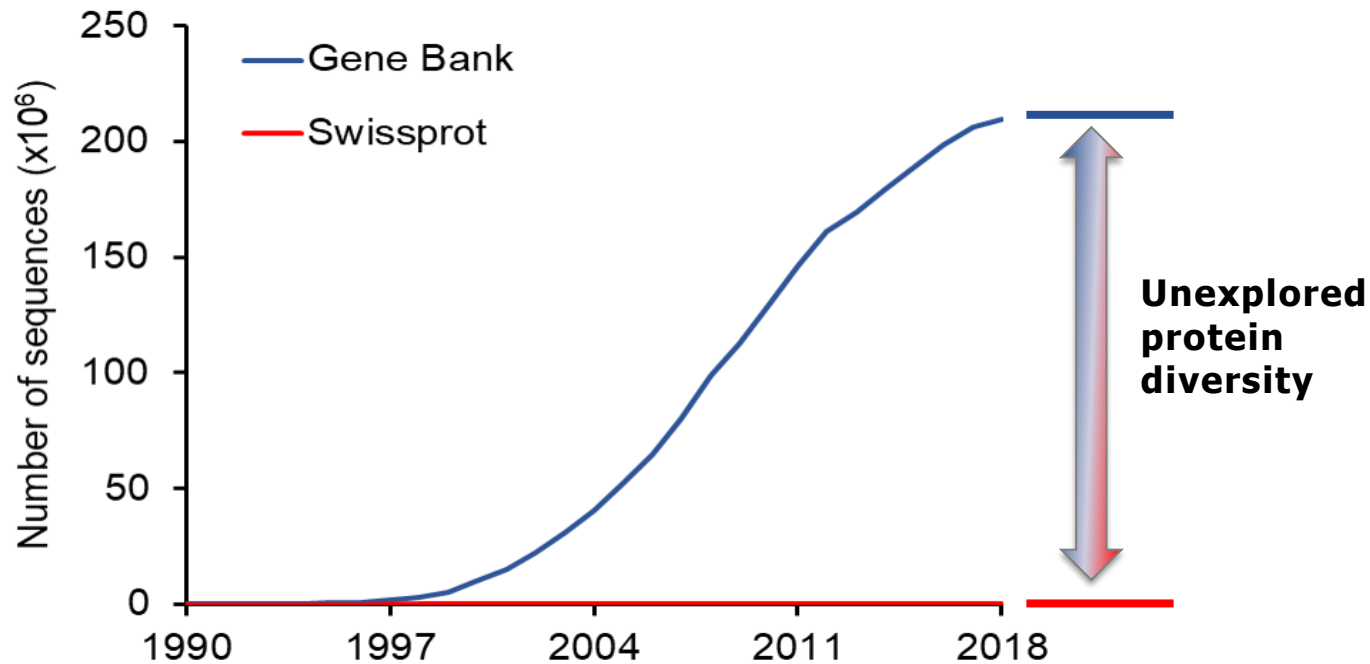
## □ UniProtKB/TrEMBL

- 😞 **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?

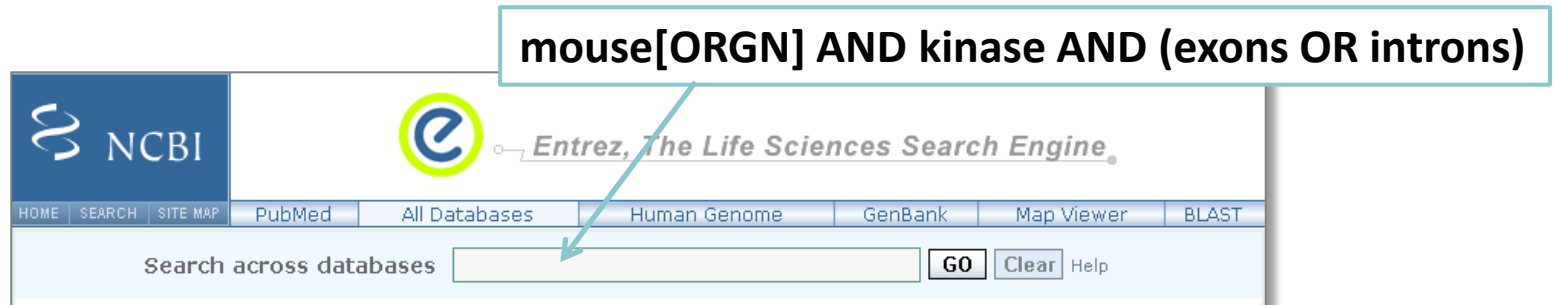
- text-based searches
- sequence-based searches

# 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



The image shows a screenshot of the NCBI Entrez search engine interface. At the top left is the NCBI logo. To its right is the Entrez logo and the text "Entrez, The Life Sciences Search Engine". Below this is a navigation bar with links for HOME, SEARCH, SITE MAP, PubMed, All Databases, Human Genome, GenBank, Map Viewer, and BLAST. The main search area contains the text "Search across databases" followed by a search input field, a "GO" button, a "Clear" button, and a "Help" link. A callout box with a light blue border and black text contains the search query: "mouse[ORGN] AND kinase AND (exons OR introns)". A light blue arrow points from this callout box to the search input field.

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

NCBI

Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases  GO Clear Help

# Text-based searches

## ❑ database retrieval systems

Search across databases    [Help](#)

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

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

# 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

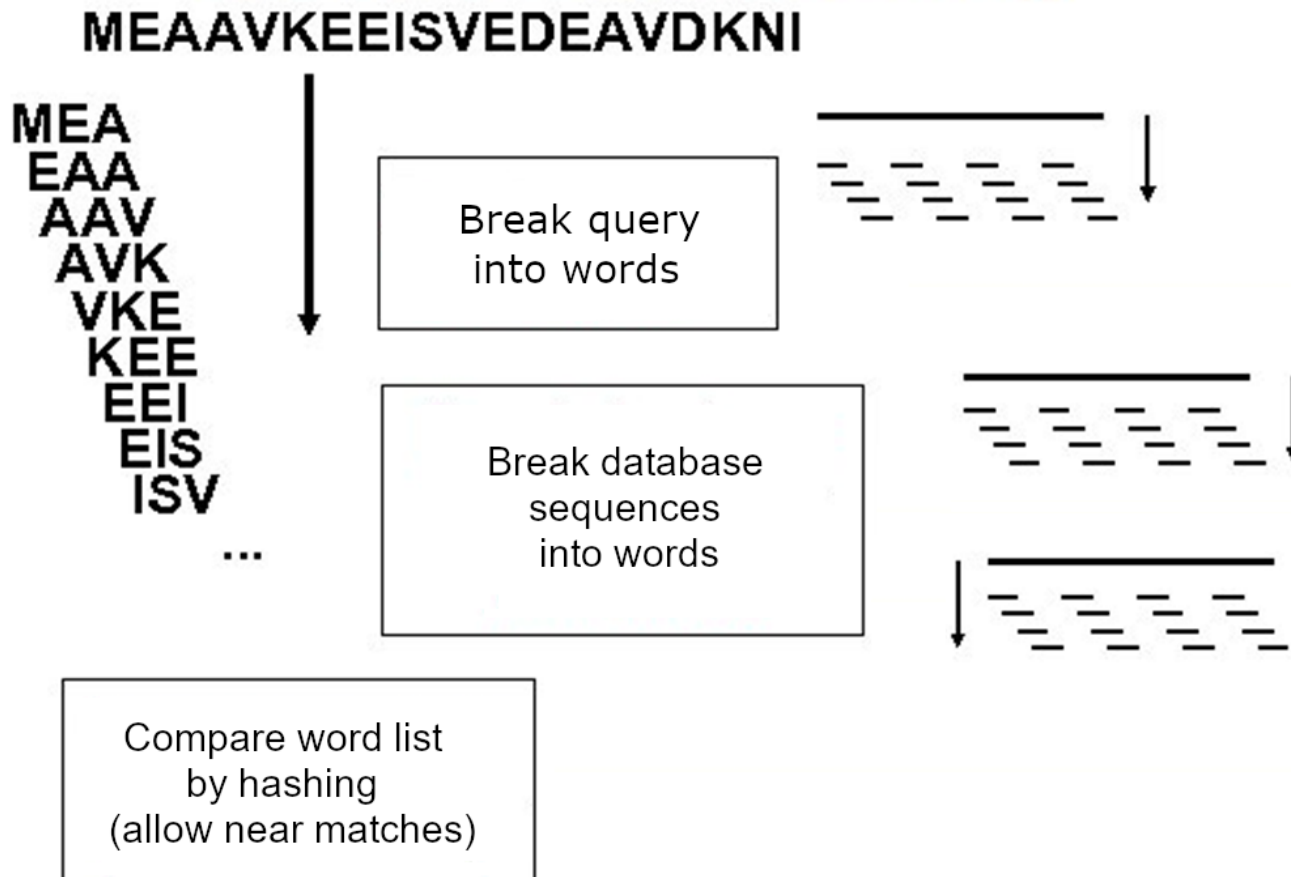
1	L	S	P	A	E	I	A	A	Y	E	A	P	F	F	T	P	D	Y	K	A	G	A	R	A	F	P	A	L	V	P	T	S	P
2	L	T	D	A	E	A	A	A	Y	G	A	P	F	F	D	Q	R	Y	K	A	G	V	R	R	F	P	E	L	V	P	V	S	P
3	M	S	P	D	E	C	A	A	Y	N	A	P	F	F	D	K	G	H	R	A	A	L	R	A	F	P	L	M	V	P	E	S	E
4	L	S	D	A	E	R	S	A	Y	D	A	P	F	F	D	E	S	Y	K	E	G	A	R	I	F	P	A	L	V	P	I	T	P
5	V	P	A	G	V	R	A	G	Y	D	A	P	F	F	D	K	T	Y	Q	A	G	A	R	A	F	P	R	L	V	P	T	S	P
6	L	S	T	D	V	L	N	A	Y	D	A	P	F	F	T	E	A	H	K	A	G	V	R	Q	F	P	L	L	V	P	A	T	T
7	V	P	A	G	V	R	A	G	Y	D	A	P	F	F	D	K	T	Y	Q	A	G	A	R	A	F	P	R	L	V	P	T	S	P

# 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







# Sequence-based searches

- ❑ PSI-BLAST input

BLAST **BETA** My NCBI Welcome aklupe. [Sign Out]

Home Recent Results Save

NCBI/BLAST/blastp suite: BLASTP prog

**MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPIILFQHGNPTSSYLWRNI**

Enter Query Sequence

Enter accession number, gi, or FASTA sequence  Clear

Query subrange From  To

Or, upload file  Browse...

Job Title   
Enter a descriptive title for your BLAST search

Choose Search Set

Database

Organism  Any  Human  *A.thaliana*  Mouse  Custom...  
Optional 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 alignment

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

hits

# Sequence-based searches

## ❑ PSI-BLAST results

Sequences producing significant alignments Download Manage Columns Show  ?

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

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

# 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 SEIGTGFPFDPHYVEVLGERMHYVDVGPRDGTPVLFHLGNPTSSYLWRNIIPHV-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 ERVKGIAACMEFIRPI----PTWDEWPEFARETFFQAFRTADVGRELIIDQNAFIEGVLPK- 175
      +RV +A ME + P P+++ F+ RTADVG ++++D N F+E +LP+
Sbjct 159 DRVAAVAFMEALVPPALPMPSEYAMGPQLGPLFRDLRTADVGEKMLDGNFFVETILPEM 218

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

Query 236 KLLFWGTPGVLIPPAEARLAESLPNCKTVDIGPGLHYLQEDNPDIGSEIARWLPG 292
      KLLF PG L P L+E++PN + +G G H+LQED+P LIG IA WL
Sbjct 279 KLLFHAEPGALAPKPVVDYLSENVNLEVRVFGAGTHFLQEDHPHLIGQGIADWLRR 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

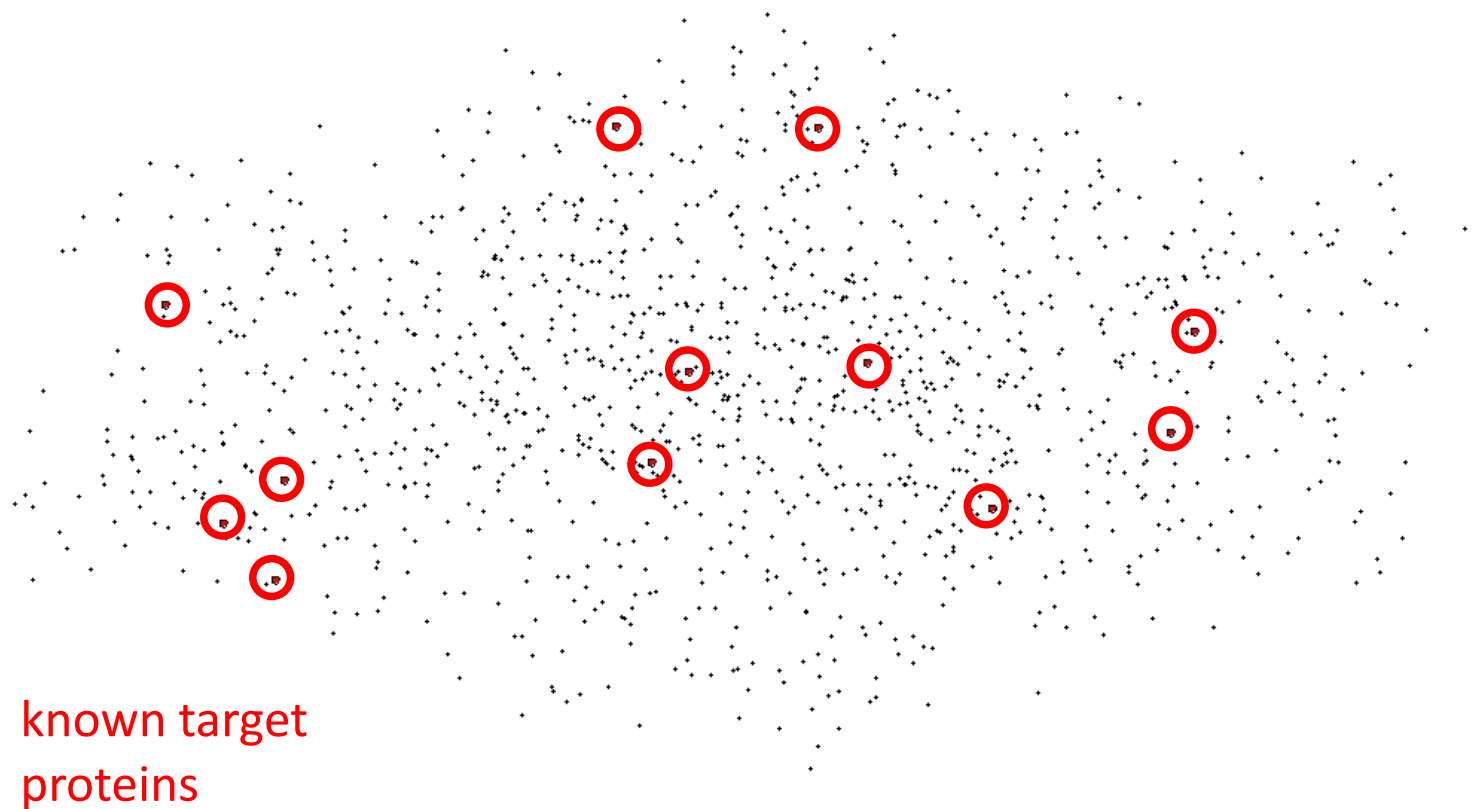
# Sequence clustering

- 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



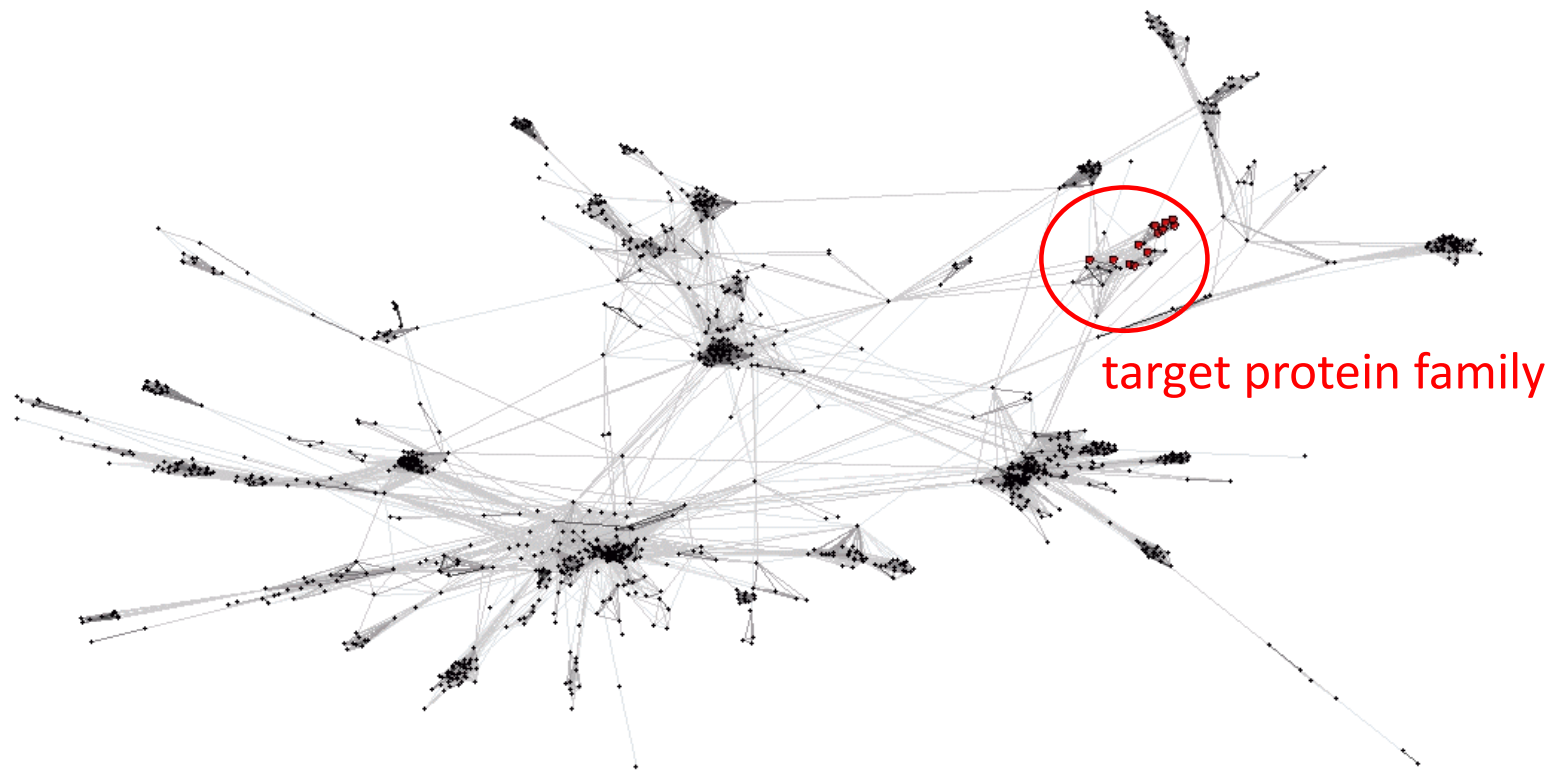
# Sequence clustering

- clustering based on pairwise sequence similarities



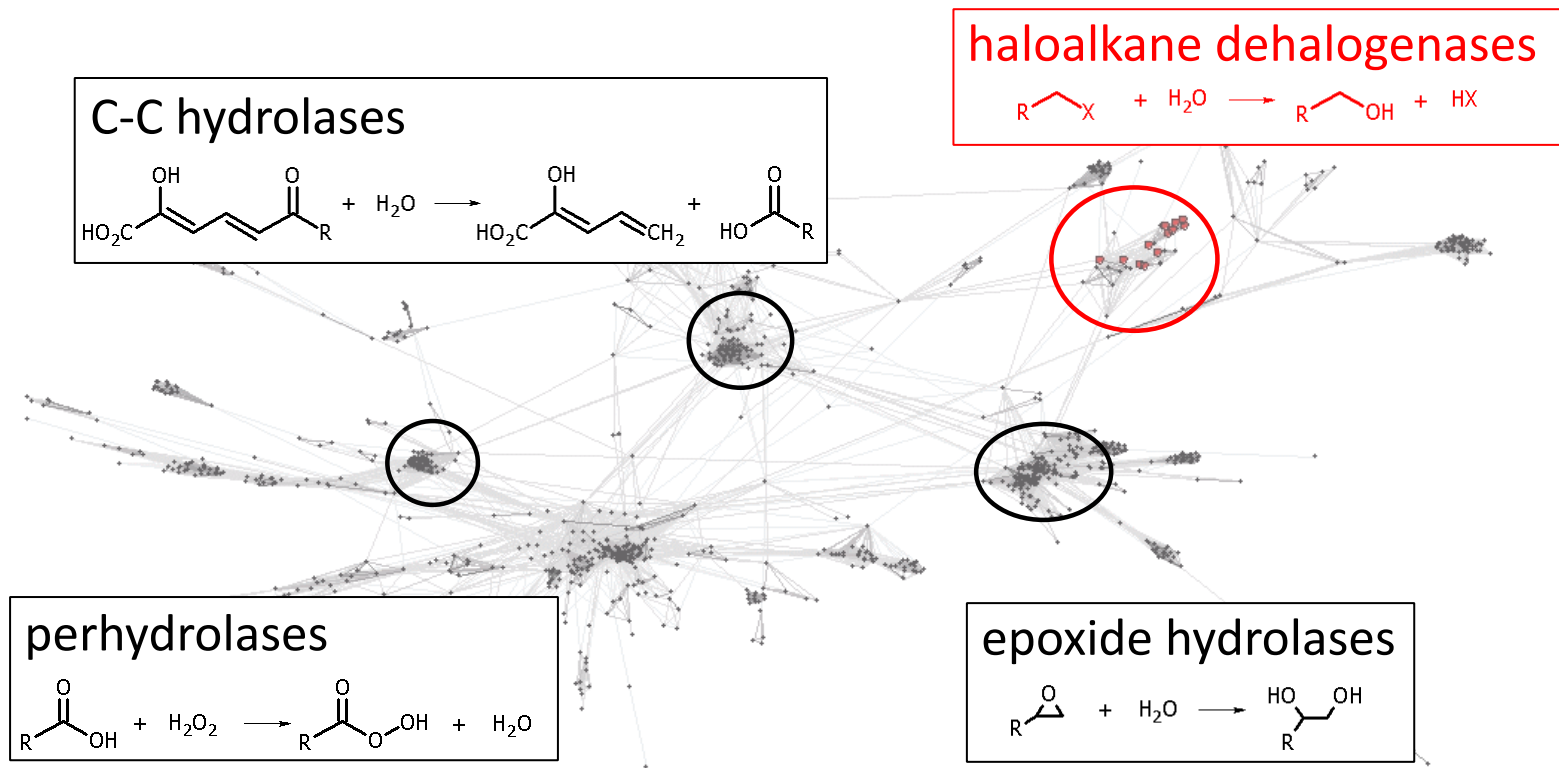
# Sequence clustering

- clustering based on pairwise sequence similarities



# Sequence clustering

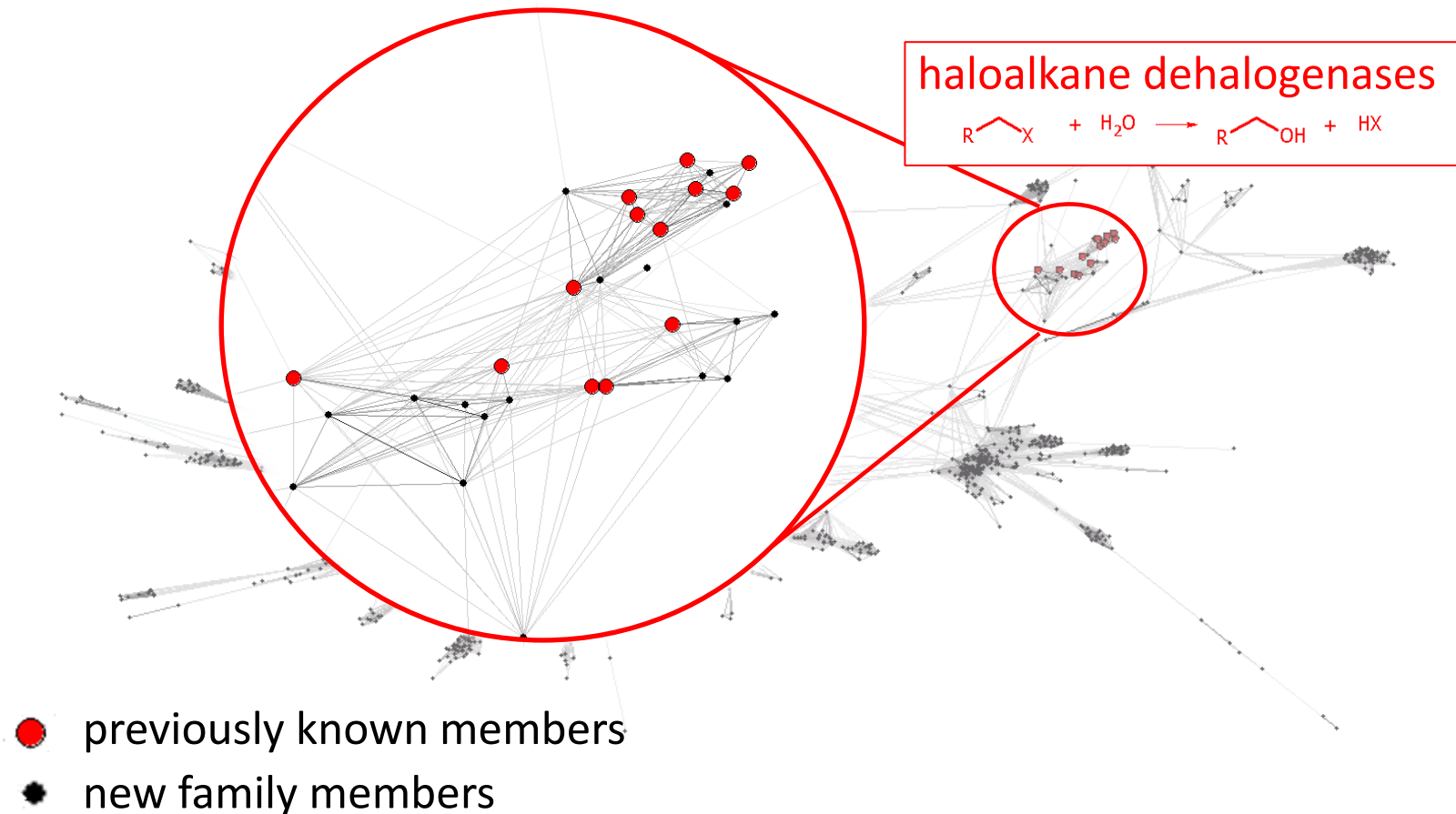
- clustering based on pairwise sequence similarities





# Sequence clustering

- clustering based on pairwise sequence similarities



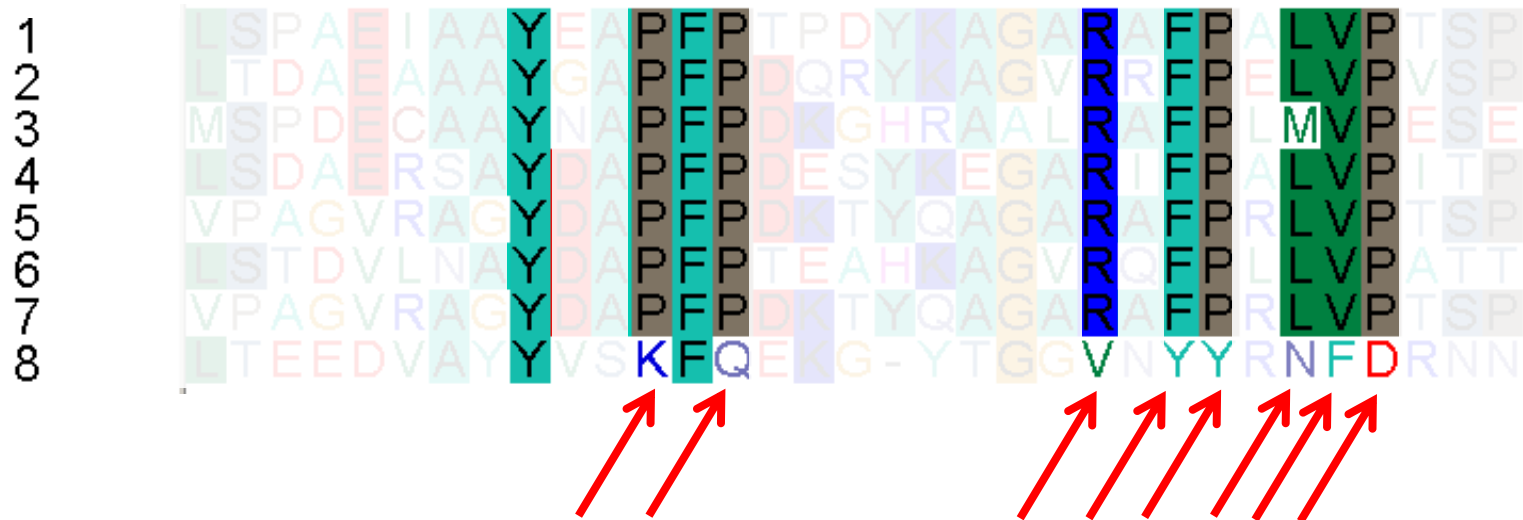
# Sequence comparison

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

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

# Sequence comparison

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



# Sequence comparison

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



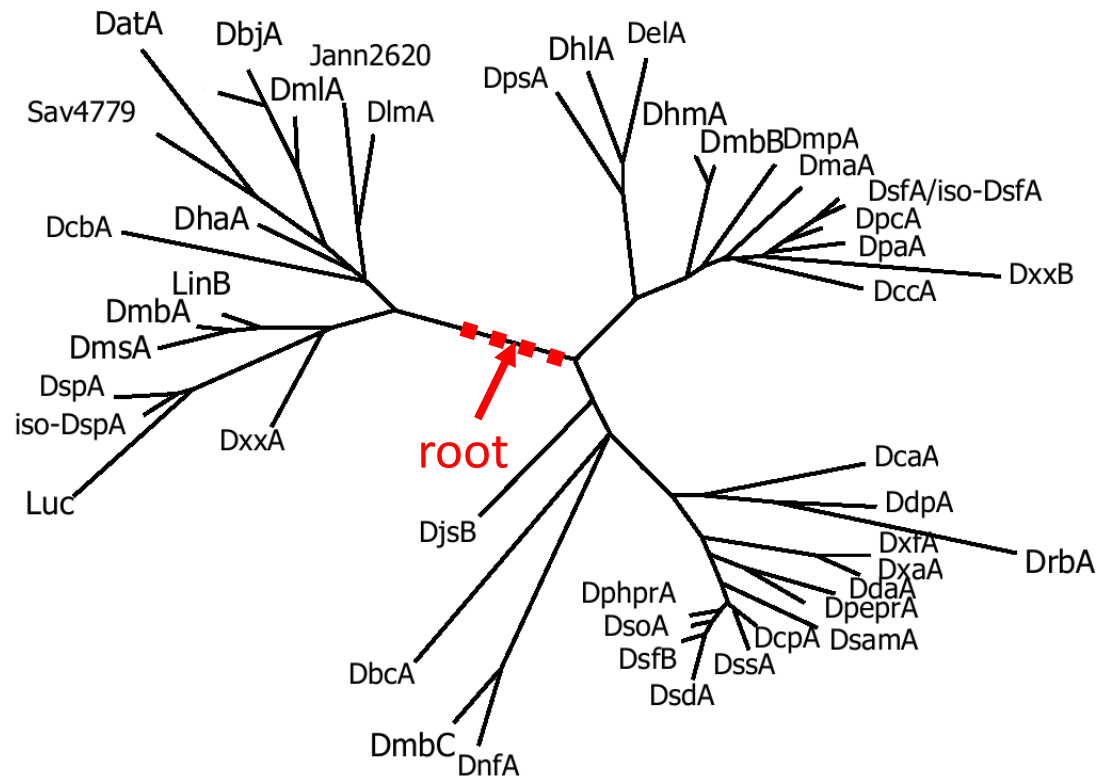




# Sequence comparison

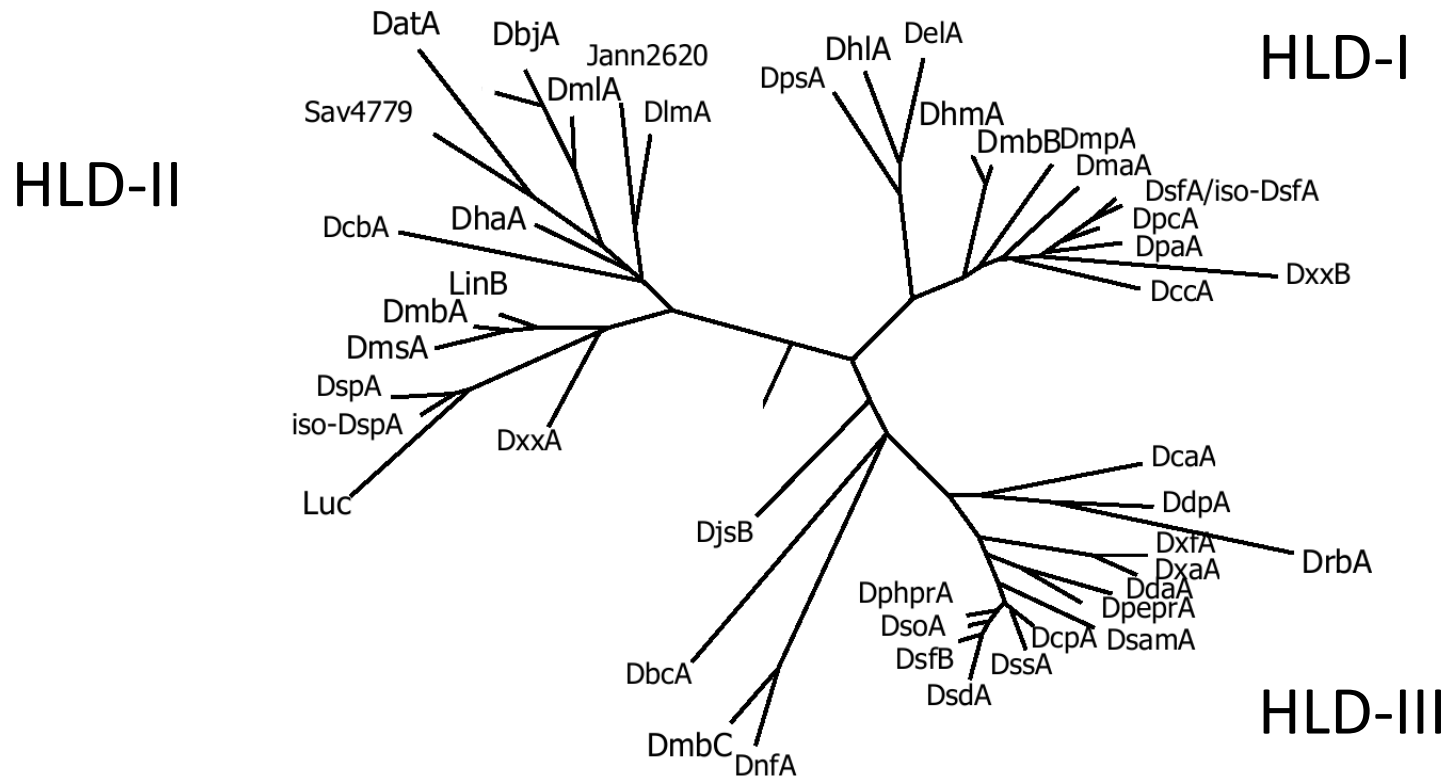
## □ phylogenetics

- establishment of **evolutionary relationships** among sequences



# Sequence comparison

- phylogenetics
  - classification of sequences

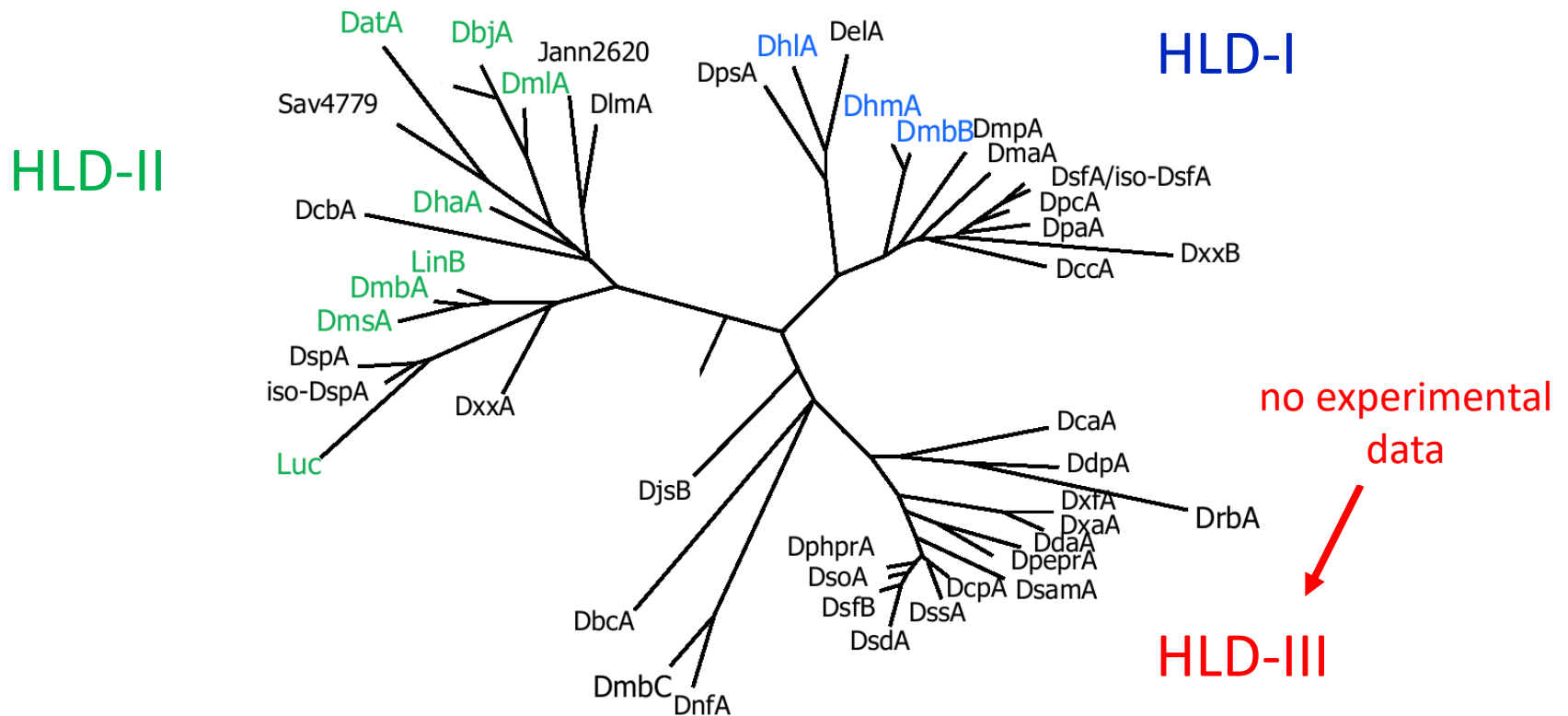




# Sequence comparison

- phylogenetics

- information about experimental data → selection of novel proteins



# 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

<a href="#">MIGS 22</a> 	<b>OXYGEN REQUIREMENT</b>	Aerobe
<a href="#">MIGS 37.1</a> 	<b>CELL SHAPE</b>	Rod-shaped
<a href="#">MIGS 37.2</a> 	<b>MOTILITY</b>	Nonmotile
<a href="#">MIGS 37.3</a> 	<b>SPORULATION</b>	
<a href="#">MIGS 37.4</a> 	<b>PRESSURE</b>	
<a href="#">MIGS 37.12</a> 	<b>TEMPERATURE RANGE</b>	Psychrophile
	<b>SALINITY</b>	Halotolerant
	<b>PH</b>	
<a href="#">MIGS 37.5</a> 	<b>CELL DIAMETER</b>	
<a href="#">MIGS 37.6</a> 	<b>CELL LENGTH</b>	
<a href="#">MIGS 37.7</a> 	<b>COLOR</b>	
<a href="#">MIGS 37.8</a> 	<b>GRAM STAINING</b>	
<a href="#">MIGS 15</a> 	<b>BIOTIC REALTIONSHPIS</b>	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
Plasm	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

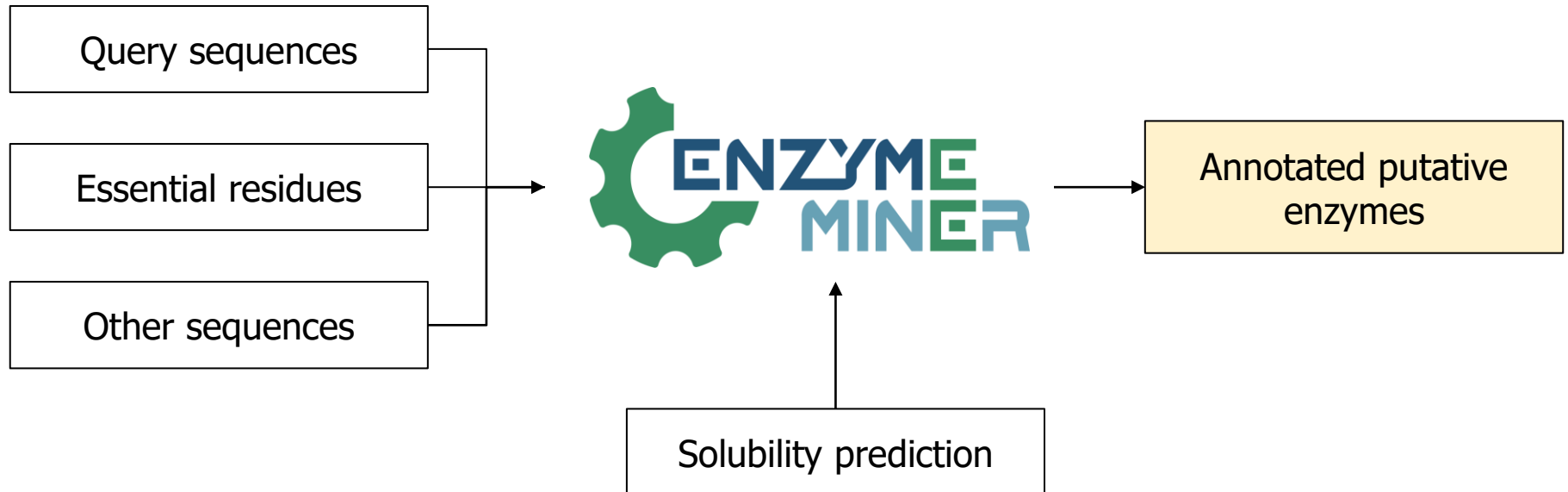
● Chromosomes [1] ● Scaffolds or contigs [0] ● SRA or Traces [0] ○ No data [0]

Organism	BioProject	Assembly	Status	Chrs	Plasmids	Size (Mb)	GC%	Gene	Protein
<a href="#">Psychrobacter cryohalolentis K5</a>	<a href="#">PRJNA58373, PRJNA13920</a>	ASM1390v1	●	1	1	3.1	42.2	2,581	2,511

# Automated *in silico* enzyme identification

## □ Enzyme Miner

- <https://loschmidt.chemi.muni.cz/enzymeminer/>



# Automated *in silico* enzyme identification



Automated mining of enzymes with diversified function.



[Submit new job](#) [Help](#) [Example](#) [Use cases](#) [Acknowledgements](#)

Job ID:

[Find job](#)

## INPUT

Swiss-Prot sequences ?

Custom sequences ?

Enter EC number...



[Advanced options](#)

## JOB INFORMATION

Job title:

Email:

[Next](#)

## 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



Swiss-Prot sequences ?

Custom sequences ?

Load saved input ?

Query sequences: ?

```
>DrbA
MSCRLSSNRRGSSKLAAMTNLASDLFPHPSSELSIDGHTLRYIDTAASSDIPSSAVGSSD
GEPTFLCVHGNPTWSFYRRIIERYGKQQRVIAVDHIGCGRSDKPSSEDFPYTMAAHRDN
LIRLVDELDLKNVILIAHDWGGAIGLSAMHARRDRLAGIGLLNTAAFPPYPMPQRIACR
MPVI.GTPAVRGI.NL.FARAAVTMAMSR TKMKPDVAAGI.LAPYDNWKNRVAIDRFVVDIPLN
```

Load from file:

 Soubor nevybrán

Other known sequences: ?

```
>DmbC
MSIDFTPDPQLYPFESRWFSSRGRHIVDEGTGPPILLCHGNPTWSFLYRDIIVALRDR
FRCVAPDYLGFGLSERPSGFGYQIDEHARVIGEFVDHLGLDRYLSMGQDWGGPISMAVAV
ERADRVRGVVLGNTWFWPADTLAMKAFSRVMSPPVQYAILRRNFFVERLIPAGTEHRPS
SAVMAHYRAVOPNAAAARRGVAFMPKOTI.AARPI.LARLAREVPATL.GTKPTLI.TWGMKDVA
```

Load from file:

 Soubor nevybrán

Essential residue templates: ?

Add protein (row)

Add residue (column)

Accession	nucleophile	acid1	acid2	base	halide1	halide2	halide3
	D	D, E	D, E	H	H, N, Q, W, Y	H, N, Q, W, Y	H, N, Q, W, Y
<b>DrbA</b>	139	Enter position	272	300	71	140	Enter position
<b>DmbB</b>	123	Enter position	250	279	Enter position	124	164
<b>DmbC</b>	109	Enter position	238	267	43	110	Enter position



## JOB OUTPUT INFORMATION

ID: example

Title: Example

Time: 18 12 2019 14:04

Status: Done

Download input file

Re-run job

## DOWNLOAD RESULTS

Result table (xlsx)

Result table (tsv)

Raw results

## TARGET SELECTION TABLE

Select all

Deselect all

Undo

Redo

Solubility threshold: ?

0.00

Primary domains: ?

PF00561 (Abhydrolase\_1) x

x

v

Selected

Full Dataset

Extra domain

Known Organism

Temperature

Salinity

Biotic Relationship

Disease

Transmembrane

With Structure

Accession	Annotation	Closest query	Identity closest query	Kingdom	Solubility	Sequence length	Domain
<input type="checkbox"/> 2PSD_A	Chain A, Crystal Stru...	D4Z2G1	41.5	E	0.9735	318	Abt
<input type="checkbox"/> 2PSF_A	Chain A, Crystal Stru...	D4Z2G1	41.5	E	0.9615	310	Abt
<input type="checkbox"/> 2PSJ_A	Chain A, Crystal Stru...	D4Z2G1	41.5	E	0.9614	319	Abt
<input type="checkbox"/> 2PSH_A	Chain A, Crystal Stru...	D4Z2G1	41.2	E	0.9586	319	Abt
<input type="checkbox"/> WP_071575177.1	haloalkane dehalog...	D4Z2G1	70.8	B	0.9399	270	Abt
<input type="checkbox"/> 3SK0_A	Chain A, structure o...	D4Z2G1	46.2	B	0.9393	311	Abt
<input type="checkbox"/> 4BRZ_A	Chain A, Haloalkane...	D4Z2G1	61.7		0.9357	290	Abt



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
  - false positive results: sequences should be filtered
- ❑ 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

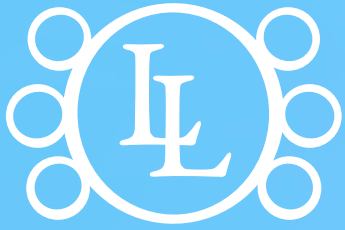
# What to keep in mind?

- ❑ *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** (2<sup>nd</sup> 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 Biotechnology* **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