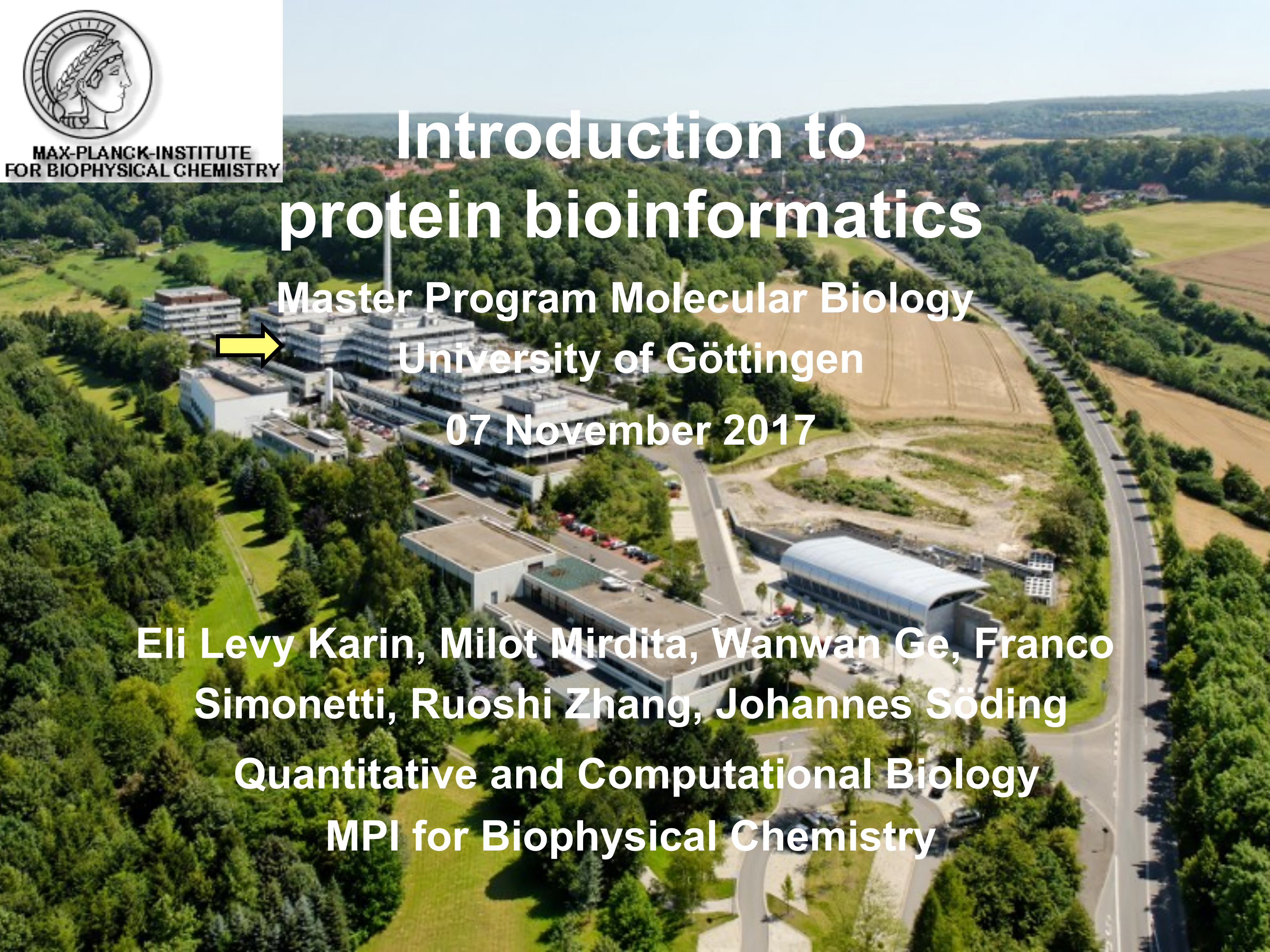
MAX-PLANCK-INSTITUTE
FOR BIOPHYSICAL CHEMISTRY



Introduction to protein bioinformatics

Master Program Molecular Biology

University of Göttingen

07 November 2017

Eli Levy Karin, Milot Mirdita, Wanwan Ge, Franco Simonetti, Ruoshi Zhang, Johannes Söding

Quantitative and Computational Biology

MPI for Biophysical Chemistry

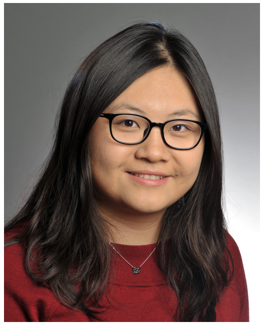
Söding lab in November 2019

Tools for metagenomics, protein structure & function

Milot Mirdita

Annika Seidel

Dr. Eli Levy Karin


Ruoshi Zhang

Christian Roth

Transcription / quant. medicine

Wanwan Ge

Salma Sohrabi-Jahromi

Dr. Franco Simonetti

Dr. Saikat Banerjee

Master/Bachelor:

Jonas Huegel
Vlad Dmbrovskyi
Kira Detrois
Hans-Georg Sommer

Recent alumni:

Niko Papadopoulos
Martin Steinegger
Clovis Galiez
Gonzalo Parra

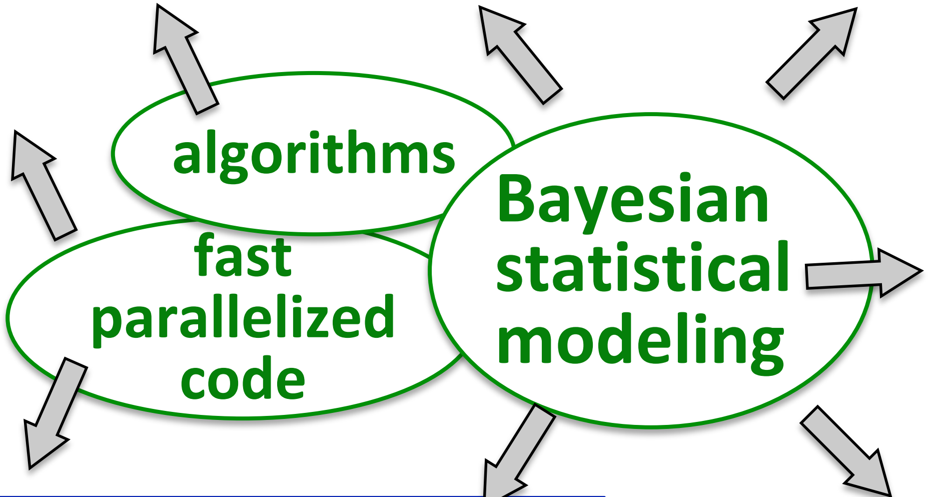
Tools for big data in biomedicine

Computational metagenomics

- Fast seq. searching & clustering methods
- Large-scale binning & X-assembly
- Viral and eukaryotic metagenomics
- Functional module discovery

Transcriptional regulation

- Biomolecular condensates
- RNA-protein binding
- Regulatory motif discovery



algorithms

fast parallelized code

Bayesian statistical modeling

Single-cell transcriptomics

- De-noising scRNA-seq data
- Reconstruction of cellular lineage trees

Protein function & structure

- Protein structure & function prediction (HHpred / HH-suite)
- Statistical method for residue contact prediction ⇒ protein structure pred.

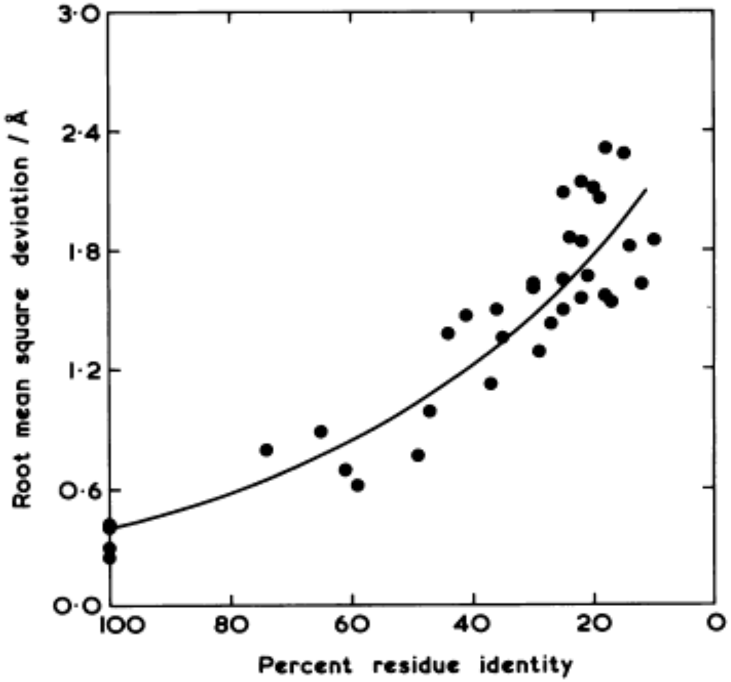
Origin of complex diseases

- Find genes/pathways which, when more highly expressed, confer higher risk for a complex disease.
- ⇒ Probabilistic models that integrate massive genome, GWAS & eQTL data

Goals for next 1 ½ days

- Understand principles of homology-based inference and sequence similarity searches
- Understand sequence alignment and the role of algorithms in bioinformatics
- Sequence profiles; Information is power!
- Learn basic analysis of metagenomics dataset
- Perform/understand secondary structure prediction, disorder prediction, transmembrane helices,…
- Understand principle of homology modeling and its limitations

Protein structure is highly conserved even without obvious sequences similarity



Sequence identity	RMSD in conserved core	Fraction of aa's in conserved core
60%	0.85 Å	95%
50%	1.0 Å	90%
40%	1.2 Å	80%
30%	1.5 Å	70%
20%	1.8 Å	55%

Sequence-structure relationship for 32 **homologous** protein pairs
[Chothia & Lesk 1986]

Structure prediction based on template with known structure can yield useful 3D models even below the twilight zone (20%)
But: quality of alignment will be crucial!

Protein sequence determines structure!

In the early seventies, Anfinsen made a fundamental discovery: The sequence of amino acids of a protein determines its native structure (with few exceptions). If all the information on a protein's structure is contained in its sequence, we should in principle be able to predict its structure from its sequence!

Since Anfinsen's discovery to the present day, this challenge has kept *computational chemists* working hard to uncover the rules of protein folding from first physical principles.

Do you know "exceptions" to Anfinsen? (3)   Allostery; misfolded proteins (Alzheimer's, prions); chaperones (GroEL, Hsp70)

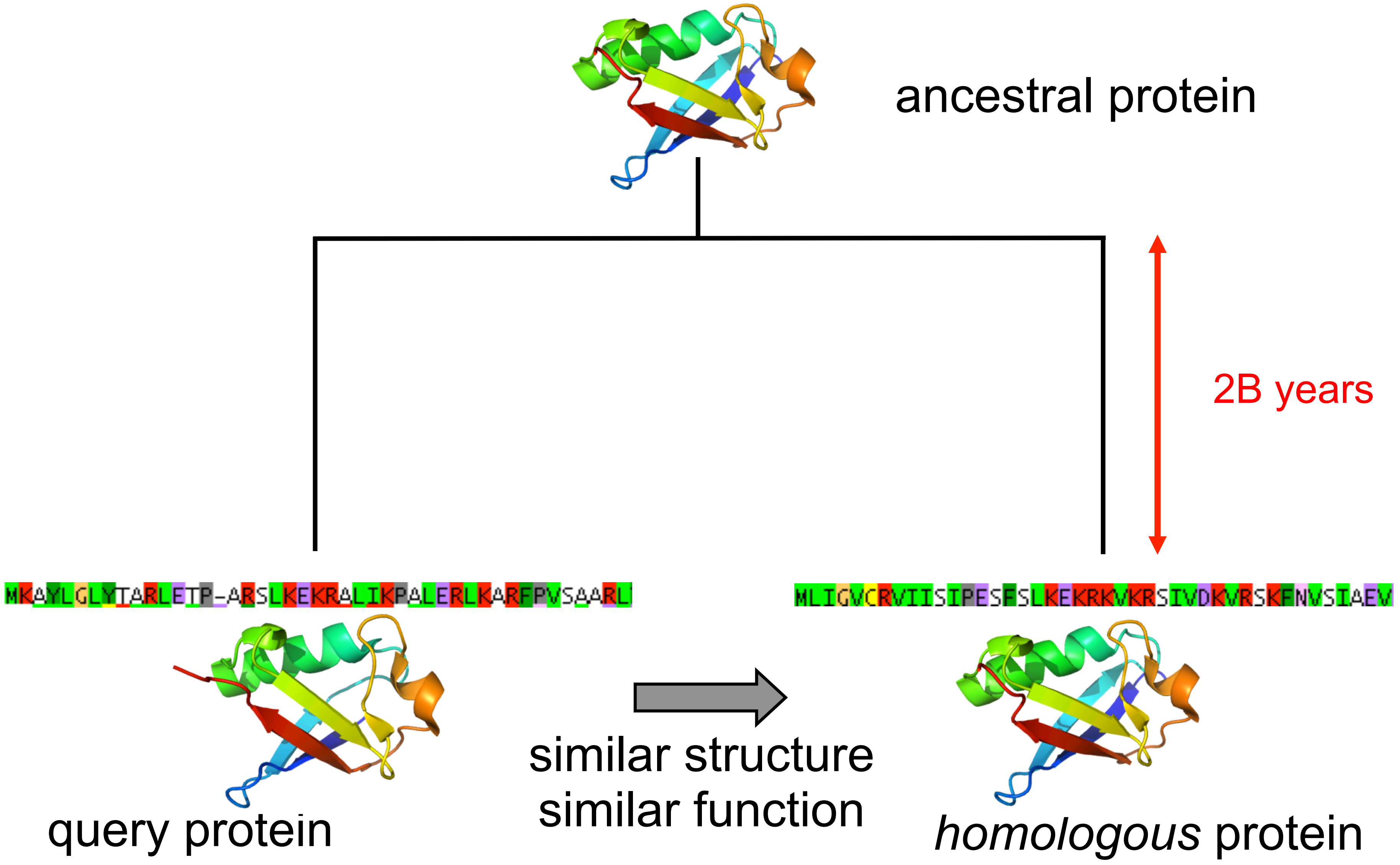
The triumph of comparative modeling

In parallel to these endeavors to predict the structures of proteins, biochemists and bioinformaticians developed a more modest, pragmatic approach: *comparative modeling*.

It relies on the fact that *homologous* proteins (those related by common ancestry) usually have very similar structures. If a protein with known structure can be found that has sufficiently high sequence similarity, the two are likely to be *homologous*, and the unknown structure can be modeled using the known structure as a *template*.

Comparative modeling is the mainstay of protein structure prediction. Due to improvements in the methods to detect and align ever more remotely related protein templates, comparative modeling can now predict the structures of about half of all known proteins from their sequence.

Homologous =
descended from common ancenstor



ancestral protein

2B years

MKAYLGLYTARLETP-ARSLKEKRALIKPALERLKARFPVSAARL

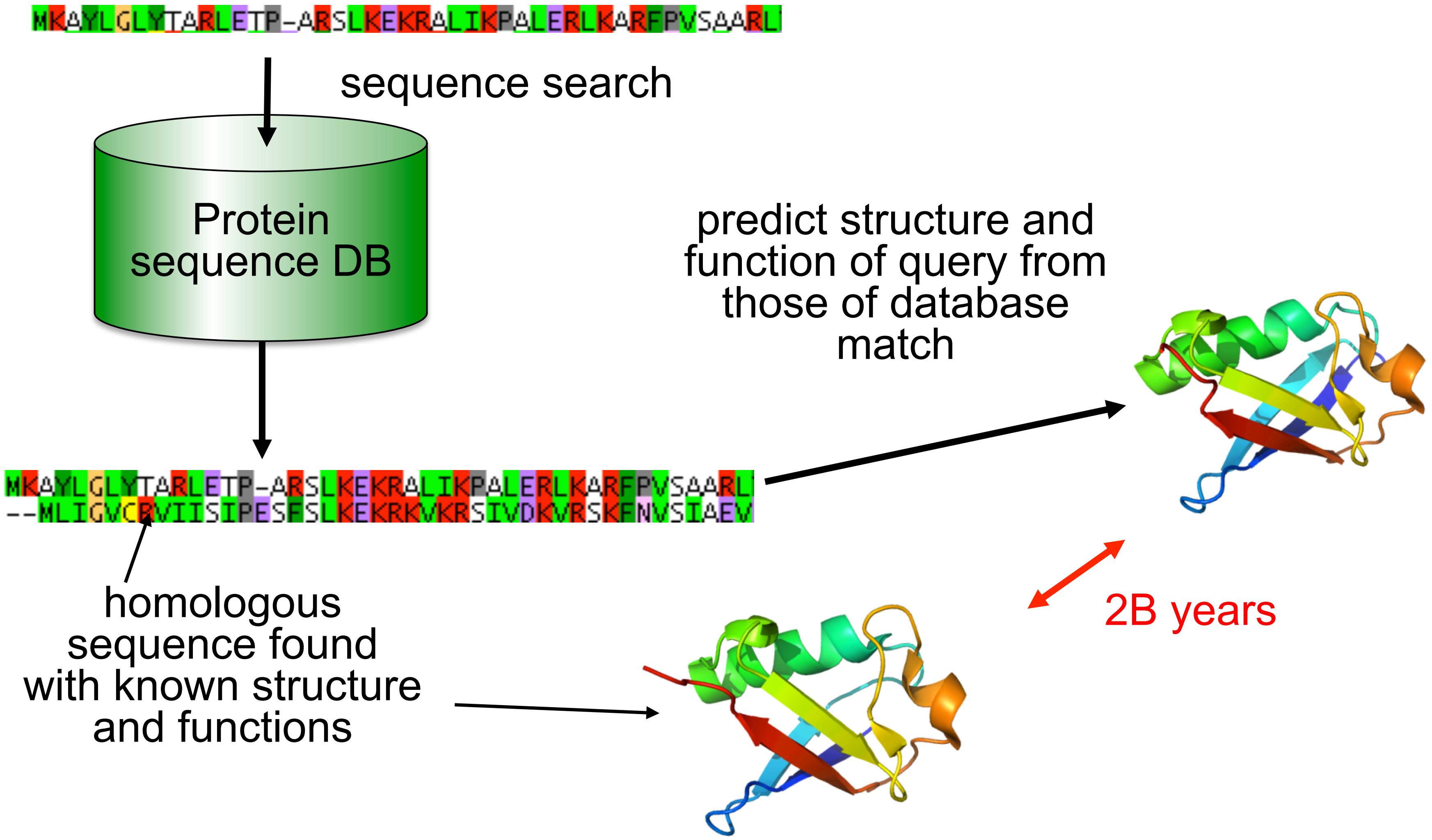
MLIGVCRVIISIPESFSLKEKRKVKRSIVDKVRSKFNVSIAEV

query protein

similar structure
similar function

homologous protein

Homology-based inference of protein structure and function


MKAYLGLYTARLETP-ARSLKEKRALIKPALERLKARFPVSAARL
Protein sequence DB
MKAYLGLYTARLETP-ARSLKEKRALIKPALERLKARFPVSAARL
--MLIGVCRVIISIPESFSLKEKRKVKRSIVDKVRSKFNVSIAEV

query protein

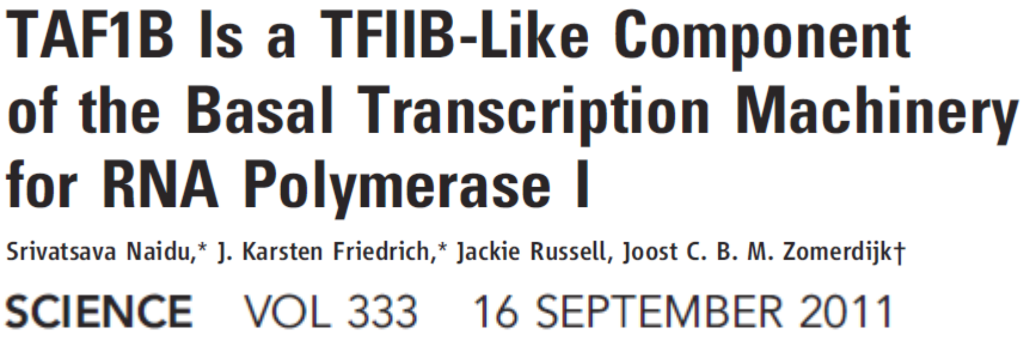
sequence search

predict structure and function of query from those of database match

homologous sequence found with known structure and functions

2B years

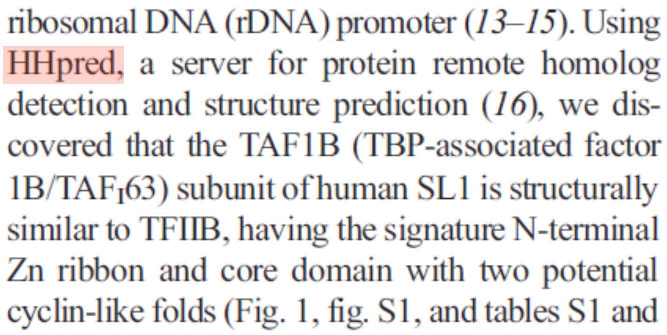
Distant homology can predict function



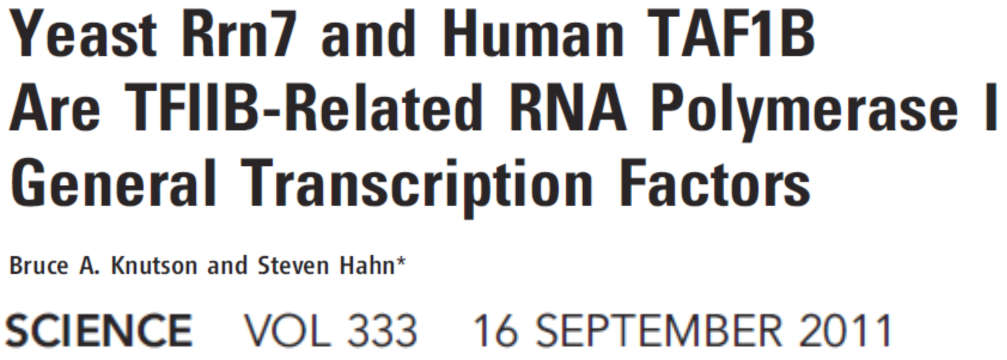
TAF1B Is a TFIIB-Like Component of the Basal Transcription Machinery for RNA Polymerase I

Srivatsava Naidu,* J. Karsten Friedrich,* Jackie Russell, Joost C. B. M. Zomerdijk†

SCIENCE VOL 333 16 SEPTEMBER 2011



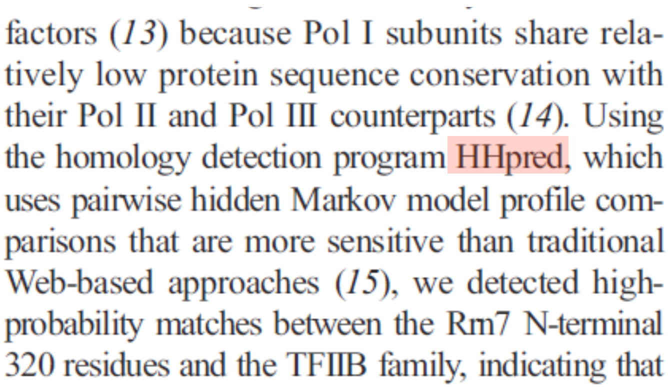
ribosomal DNA (rDNA) promoter (*13*–*15*). Using HHpred, a server for protein remote homolog detection and structure prediction (*16*), we discovered that the TAF1B (TBP-associated factor 1B/TAF$_I$63) subunit of human SL1 is structurally similar to TFIIB, having the signature N-terminal Zn ribbon and core domain with two potential cyclin-like folds (Fig. 1, fig. S1, and tables S1 and



Yeast Rrn7 and Human TAF1B Are TFIIB-Related RNA Polymerase I General Transcription Factors

Bruce A. Knutson and Steven Hahn*

SCIENCE VOL 333 16 SEPTEMBER 2011



factors (*13*) because Pol I subunits share relatively low protein sequence conservation with their Pol II and Pol III counterparts (*14*). Using the homology detection program HHpred, which uses pairwise hidden Markov model profile comparisons that are more sensitive than traditional Web-based approaches (*15*), we detected high-probability matches between the Rrn7 N-terminal 320 residues and the TFIIB family, indicating that

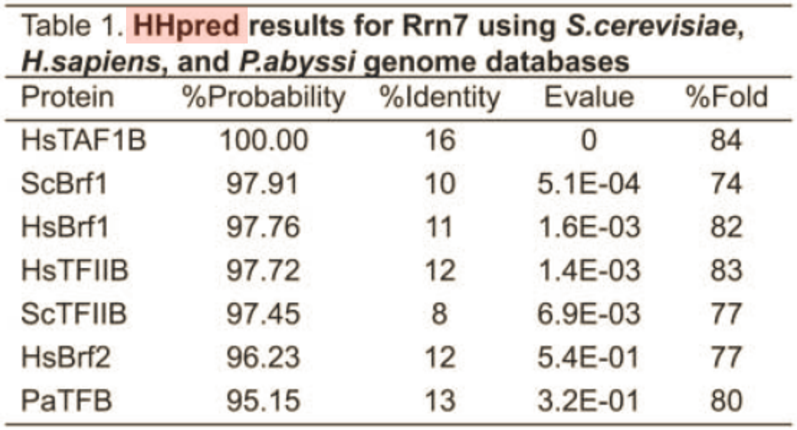
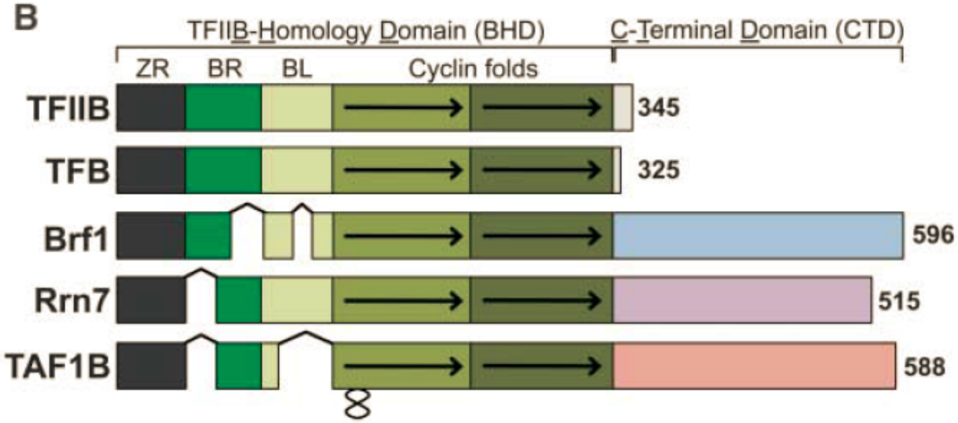


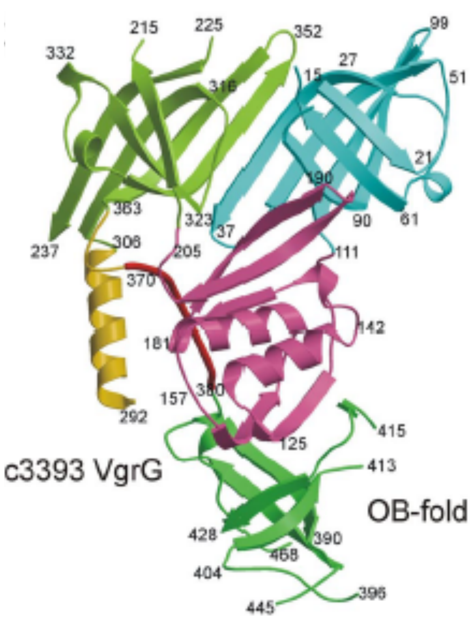
Table 1. **HHpred results for Rrn7 using *S.cerevisiae*, *H.sapiens*, and *P.abyssi* genome databases**

Protein	%Probability	%Identity	Evalue	%Fold
HsTAF1B	100.00	16	0	84
ScBrf1	97.91	10	5.1E-04	74
HsBrf1	97.76	11	1.6E-03	82
HsTFIIB	97.72	12	1.4E-03	83
ScTFIIB	97.45	8	6.9E-03	77
HsBrf2	96.23	12	5.4E-01	77
PaTFB	95.15	13	3.2E-01	80

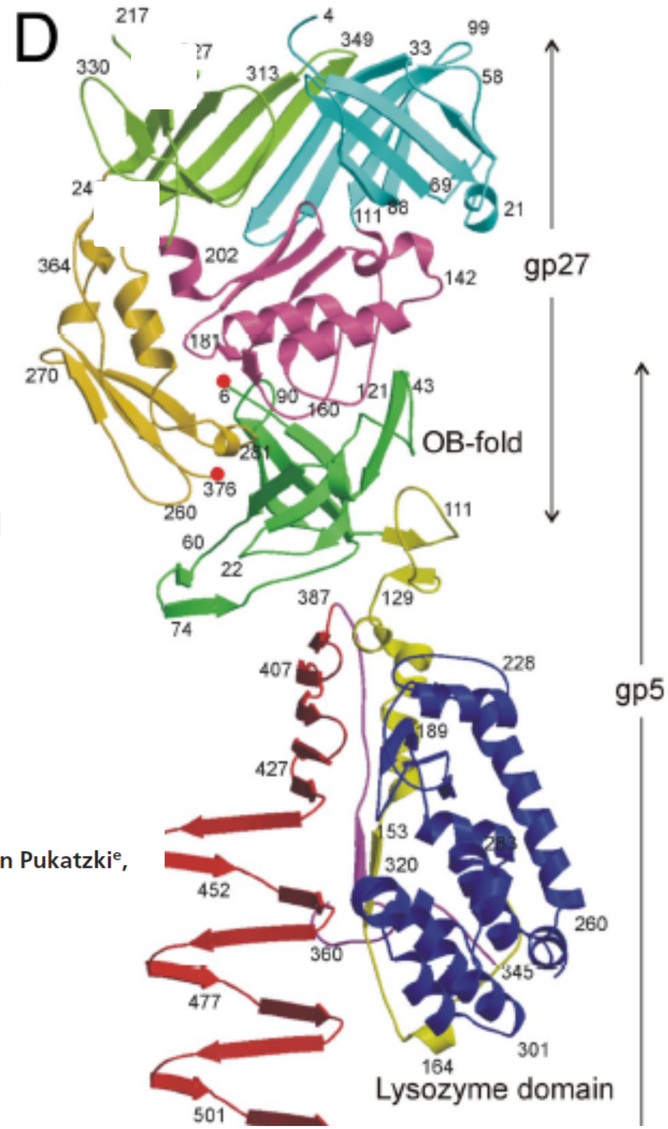


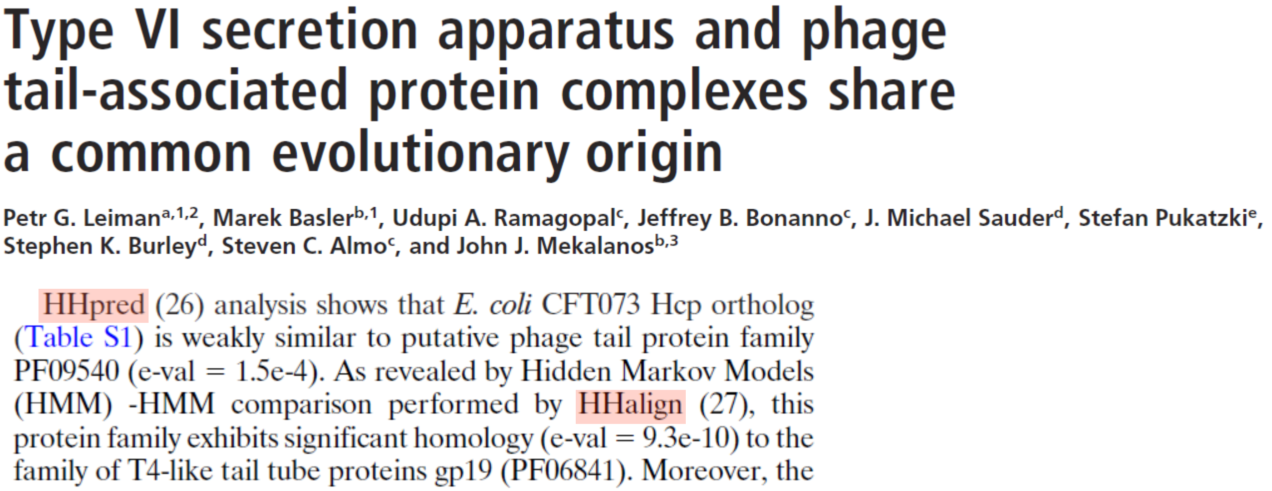
Distant homology can predict function

Type VI secretion
(trimeric unit)



phage T4
needle and spike





Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin

Petr G. Leiman[a,1,2], Marek Basler[b,1], Udupi A. Ramagopal[c], Jeffrey B. Bonanno[c], J. Michael Sauder[d], Stefan Pukatzki[e], Stephen K. Burley[d], Steven C. Almo[c], and John J. Mekalanos[b,3]

HHpred (26) analysis shows that *E. coli* CFT073 Hcp ortholog (Table S1) is weakly similar to putative phage tail protein family PF09540 (e-val = 1.5e-4). As revealed by Hidden Markov Models (HMM) -HMM comparison performed by HHalign (27), this protein family exhibits significant homology (e-val = 9.3e-10) to the family of T4-like tail tube proteins gp19 (PF06841). Moreover, the

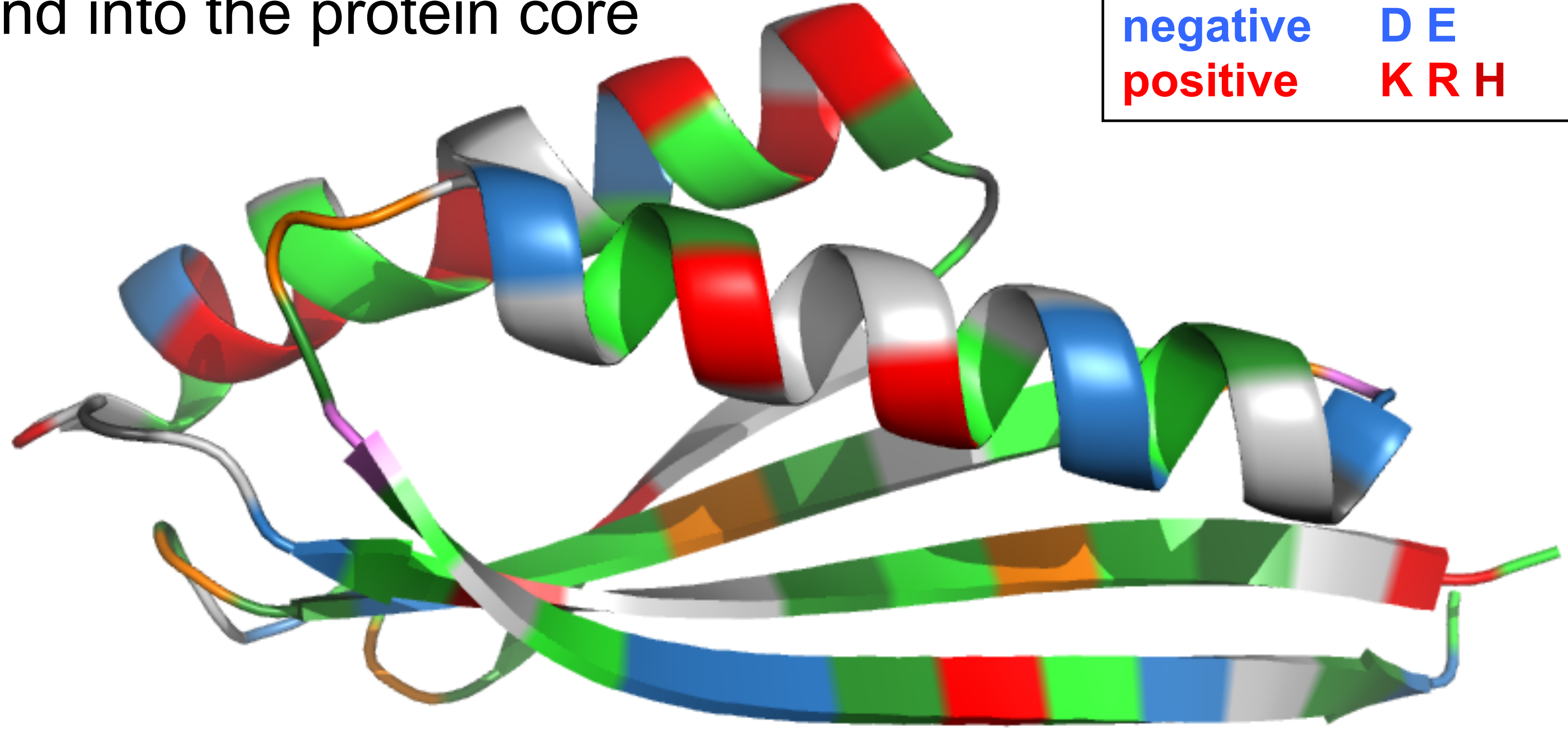
How can we infer common descent over time spans of billions of years?

Hydrophobic residues form the domain cores

Example: protein with a ferredoxin fold

Most hydrophobic side chains extend into the protein core

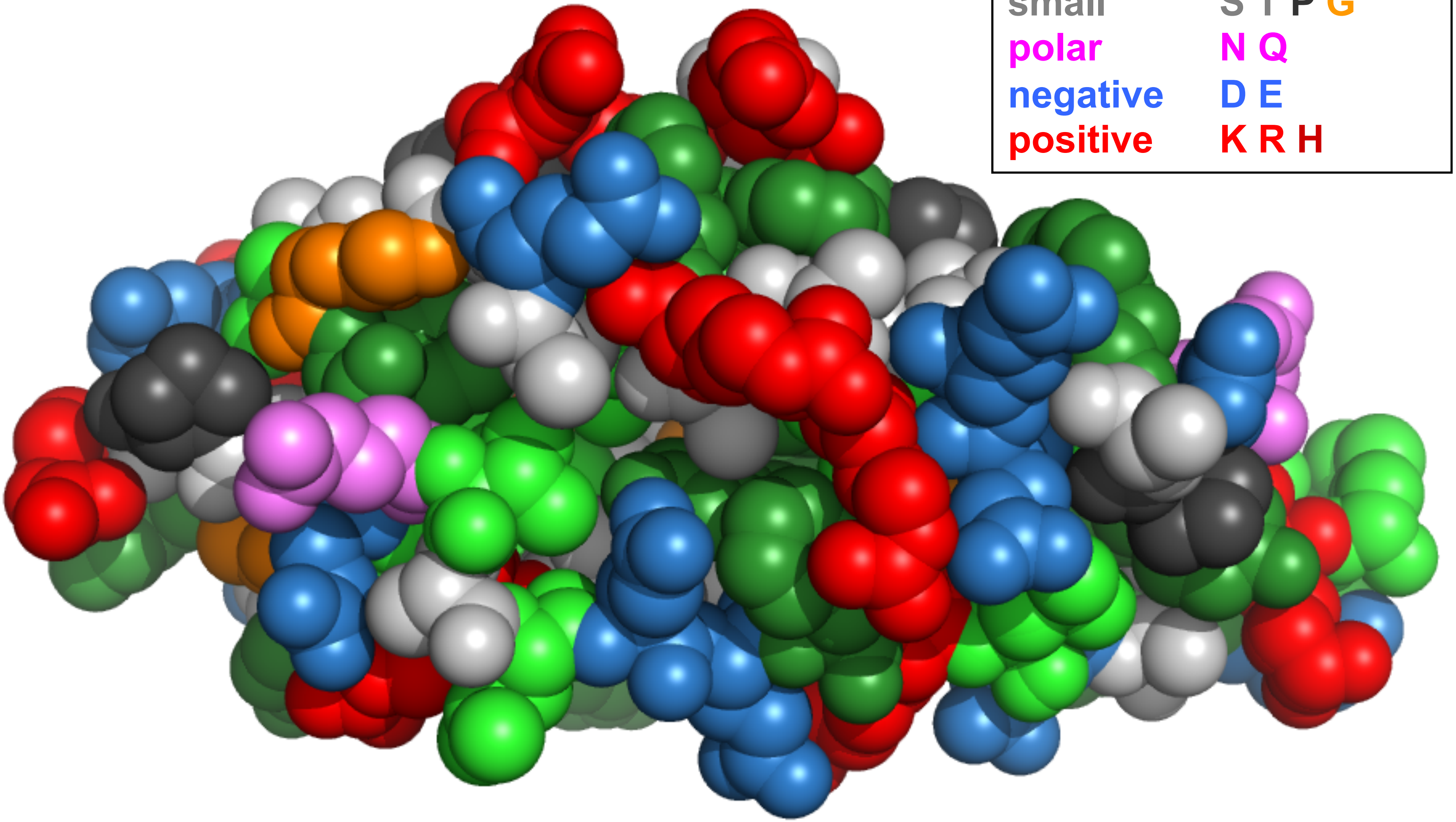
aliphatic	V L I M A C
aromatic	F W Y
small	S T P G
polar	N Q
negative	D E
positive	K R H



Hydrophobic residues form the domain cores

The protein core is tightly packed…

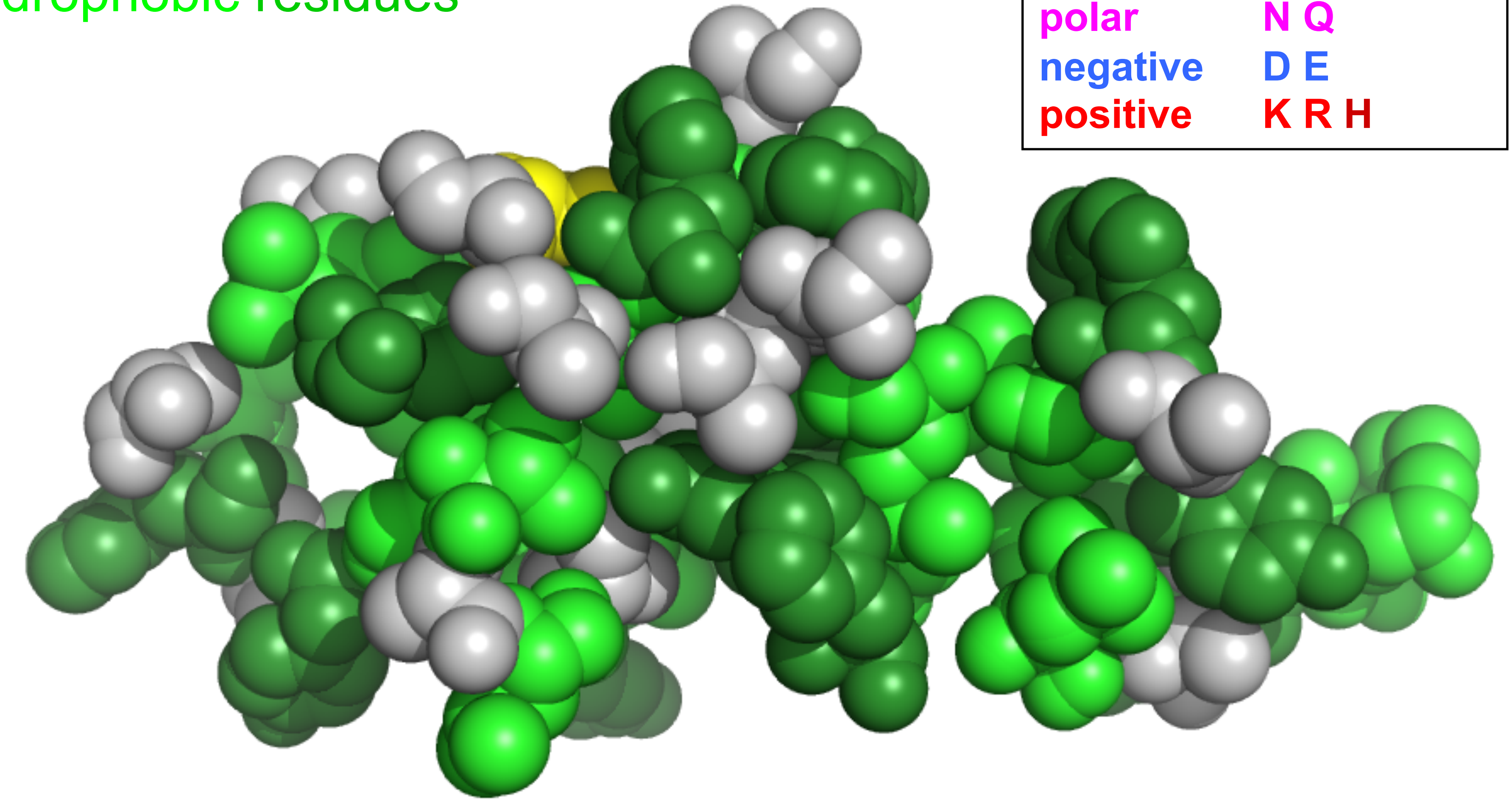
aliphatic	V L I M A C
aromatic	F W Y
small	S T P G
polar	N Q
negative	D E
positive	K R H



Hydrophobic residues form the domain cores

The protein core is tightly packed with mainly hydrophobic residues

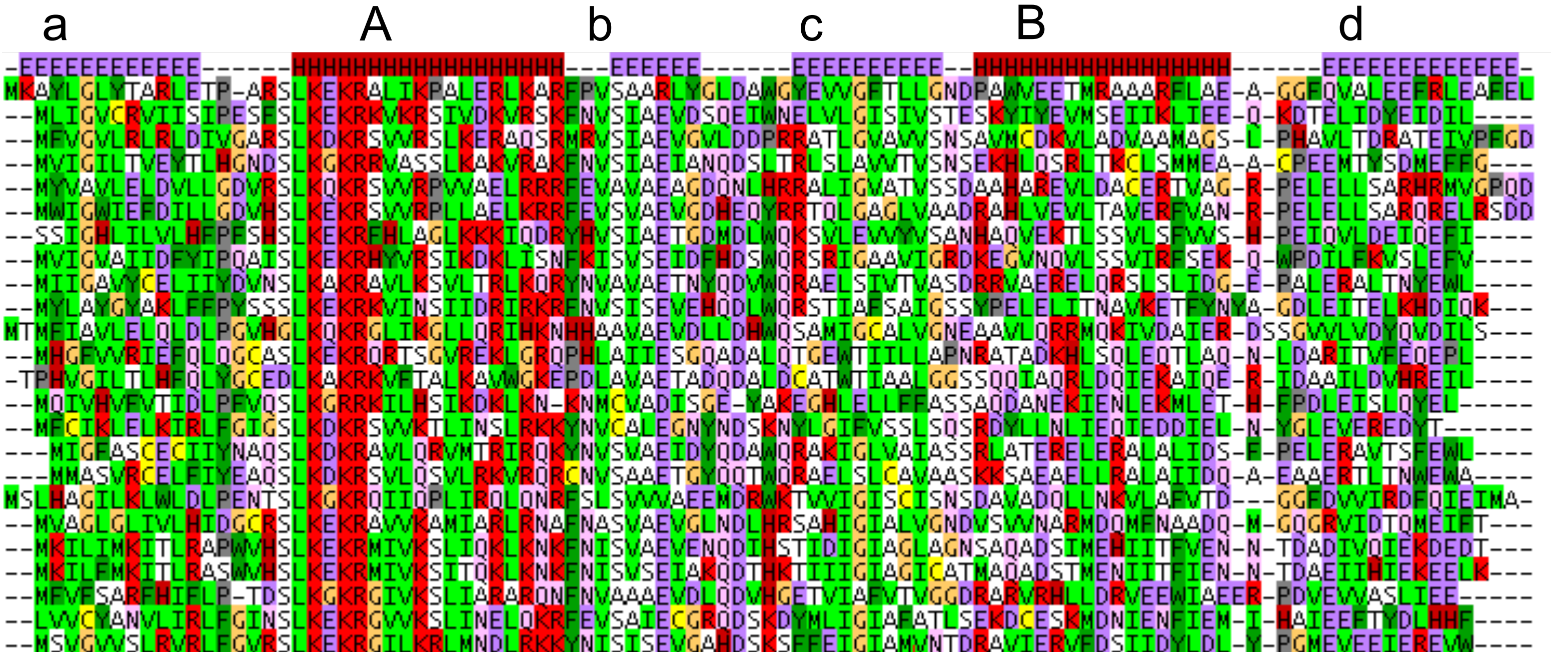
aliphatic	V L I M A C
aromatic	F W Y
small	S T P G
polar	N Q
negative	D E
positive	K R H

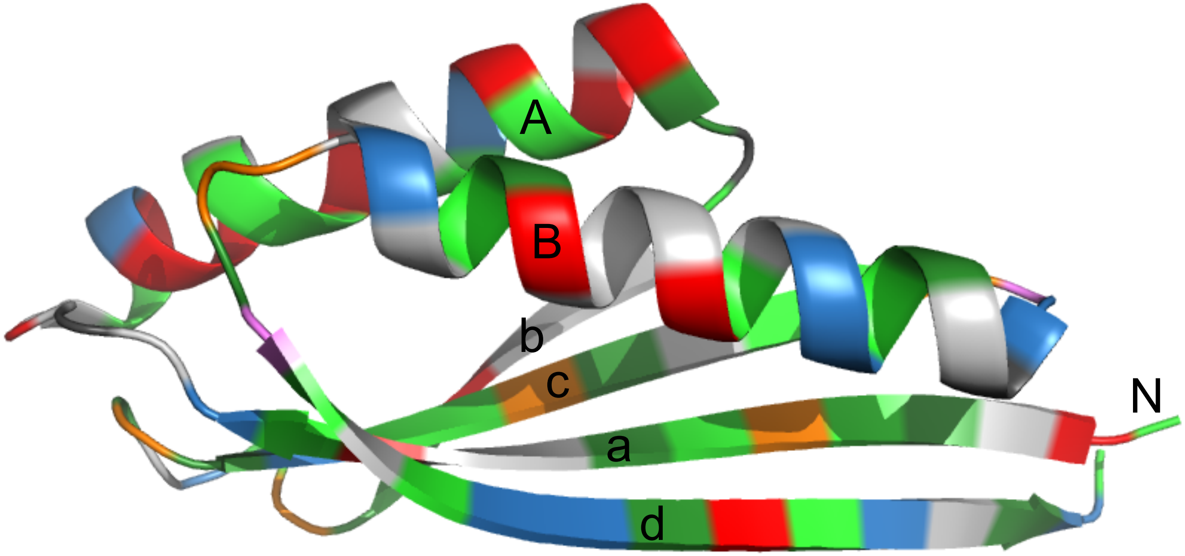


Molecular 3D Puzzle

Core residues are often well conserved

Multiple sequence alignment





Note the conserved hydrophobic columns in strands and helices.

How is it that we can infer common descent over time spans of billions of years?

- Sequence evolution is highly constrained by the requirement of a stable structural core

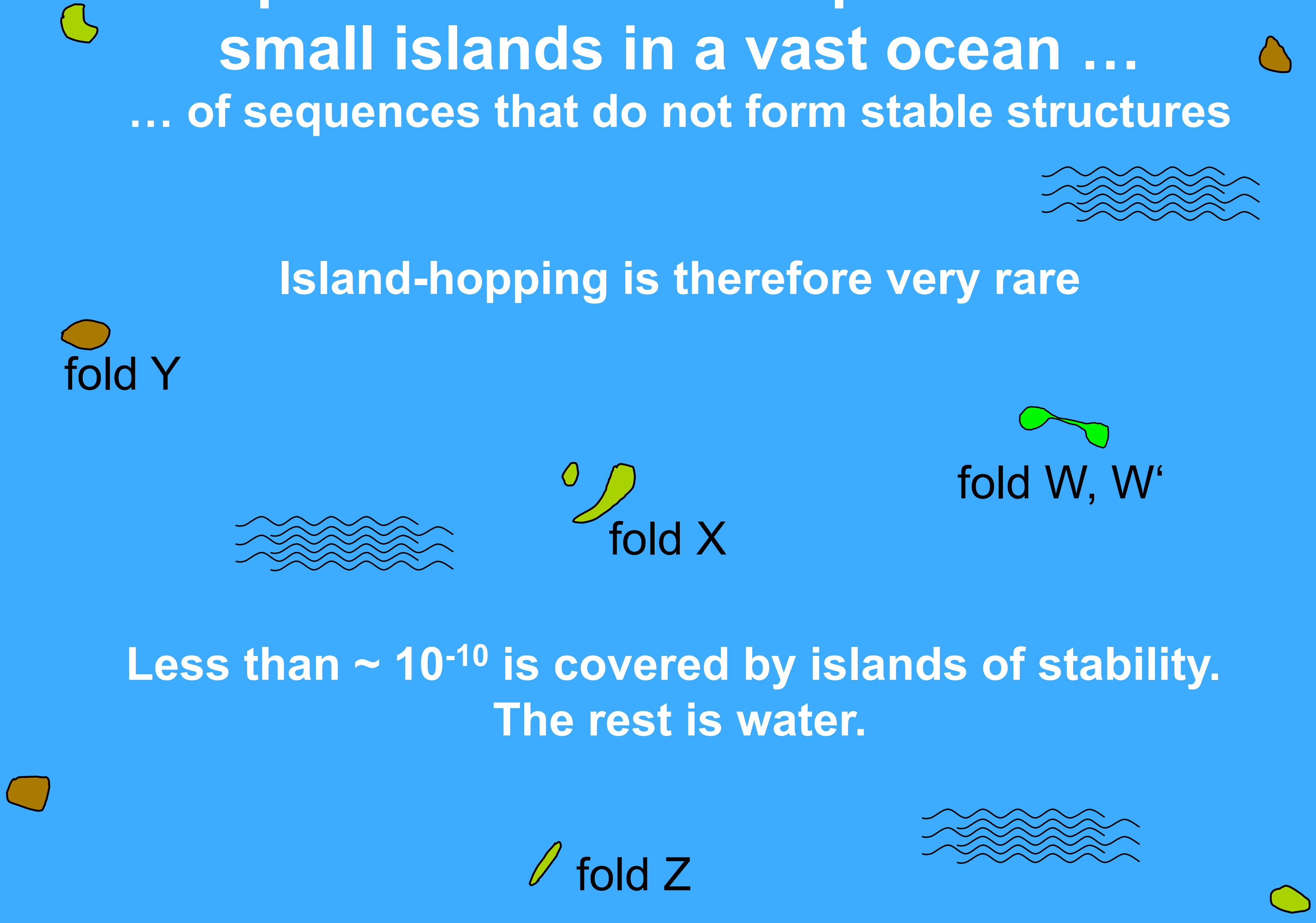
- Every fold has a specific 3D jig-saw puzzle logic of how its side-chains interlock that is highly conserved

- This logic is reflected in a protein‘s multiple sequence alignment: in pattern of conserved hydrophobicity and amino acid properties

- By **comparing multiple alignments** we can detect similar patterns that indicate the same 3D folding logic

The space of foldable sequences is like small islands in a vast ocean …

… of sequences that do not form stable structures



Island-hopping is therefore very rare

Less than ~ 10^{-10} is covered by islands of stability.
The rest is water.

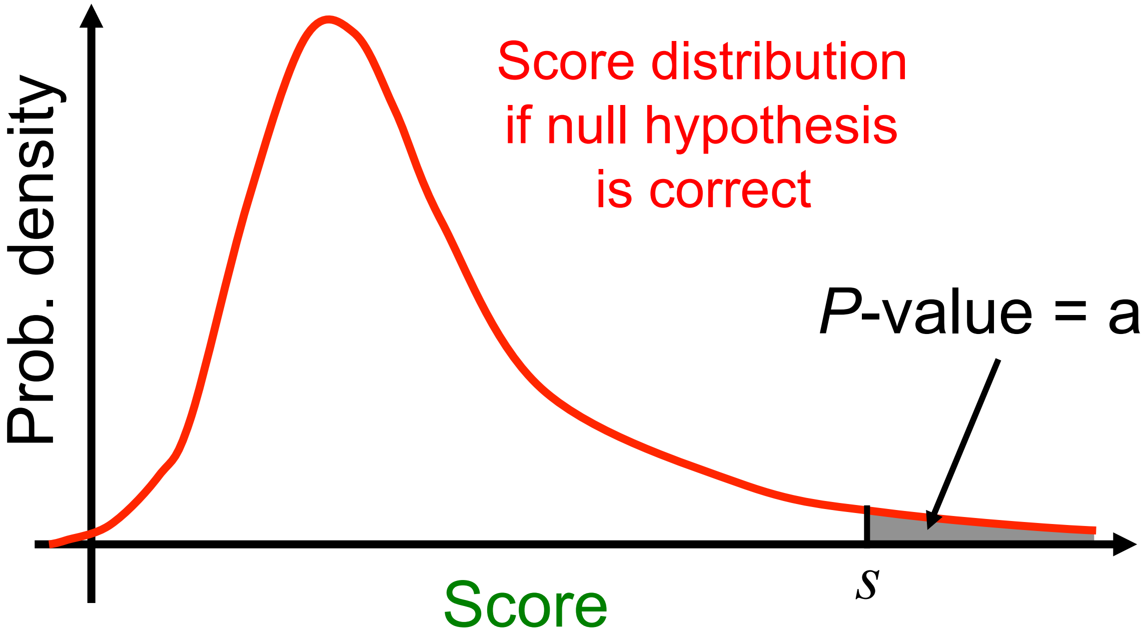
P-values quantify plausability of *null hypothesis*

Given: a *null hypothesis* (boring “hypothesis of randomness”) and a score (“test statistic”) with known distribution under the null hypothesis

Goal: find interesting cases for which the null hypothesis can be rejected

***P*-value** = the probability to obtain a score as observed *or more extreme*, under the null hypothesis.

A small *P*-value (e.g. < 0.01) indicates the null hypothesis can be rejected.



Why „or more extreme“?

***P*-value** = the probability to obtain a score as observed **or more extreme** under the null hypothesis

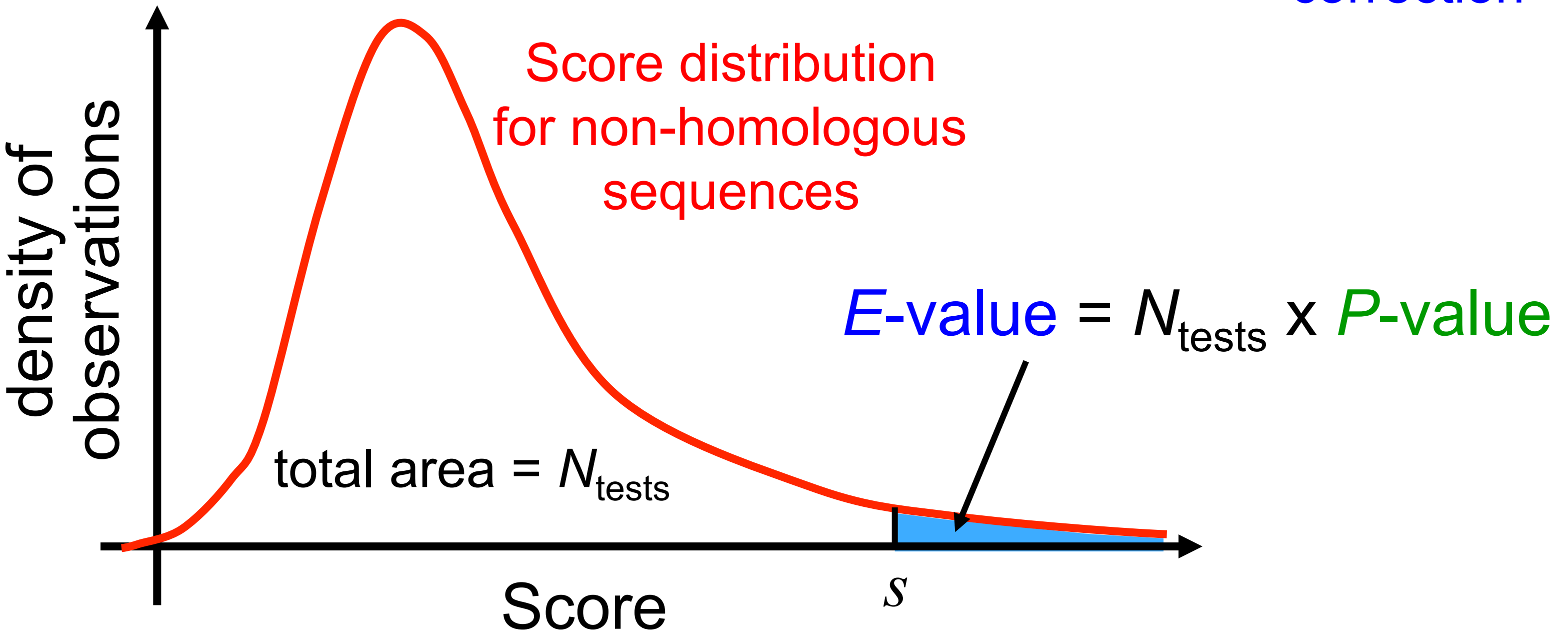
E-value = expected number of observations more extreme than the one observed

① P-value = Probability for event with score ≥ s under the null hypothesis

② E-value = *Expected number of events out of N_{tests} trials* with score ≥ S under the null hypothesis

E-value = N_{tests} x *P*-value

similar to Bonferroni multiple testing correction



Score distribution for non-homologous sequences

E-value = N_{tests} x *P*-value

P-values assess the plausibility of a null hypothesis

The P-value is the probability to obtain a result as observed *or more extreme*, given the *null hypothesis* (often a “hypothesis of randomness”). A small P-value (e.g. < 0.05) indicates the null hypothesis can be rejected.

Suppose we suspect a die to be loaded. We throw it 30 times and obtain a six only once. Can we conclude that the die is loaded?

Exercise: Compute the P-value for the *null hypothesis* that the die is fair. What do you conclude from it?

The probability to obtain a six only zero or one times, given the die is not loaded (the null hypothesis), is

$$P(k \leq 1 \text{ six out of } 30 \,|\, p_{\text{six}} = 1/6) = \sum_{k=0}^{1} \binom{30}{k} (1/6)^k (5/6)^{30-k}$$

$$= \binom{30}{0} (1/6)^0 (5/6)^{30} + \binom{30}{1} (1/6)(5/6)^{29} \quad = 0.0042 + 0.0253 = 0.029$$

We can reject the null-hypothesis that the die is fair with a P-value of 3%.

Structure and function of protein domains are often conserved over billions of years

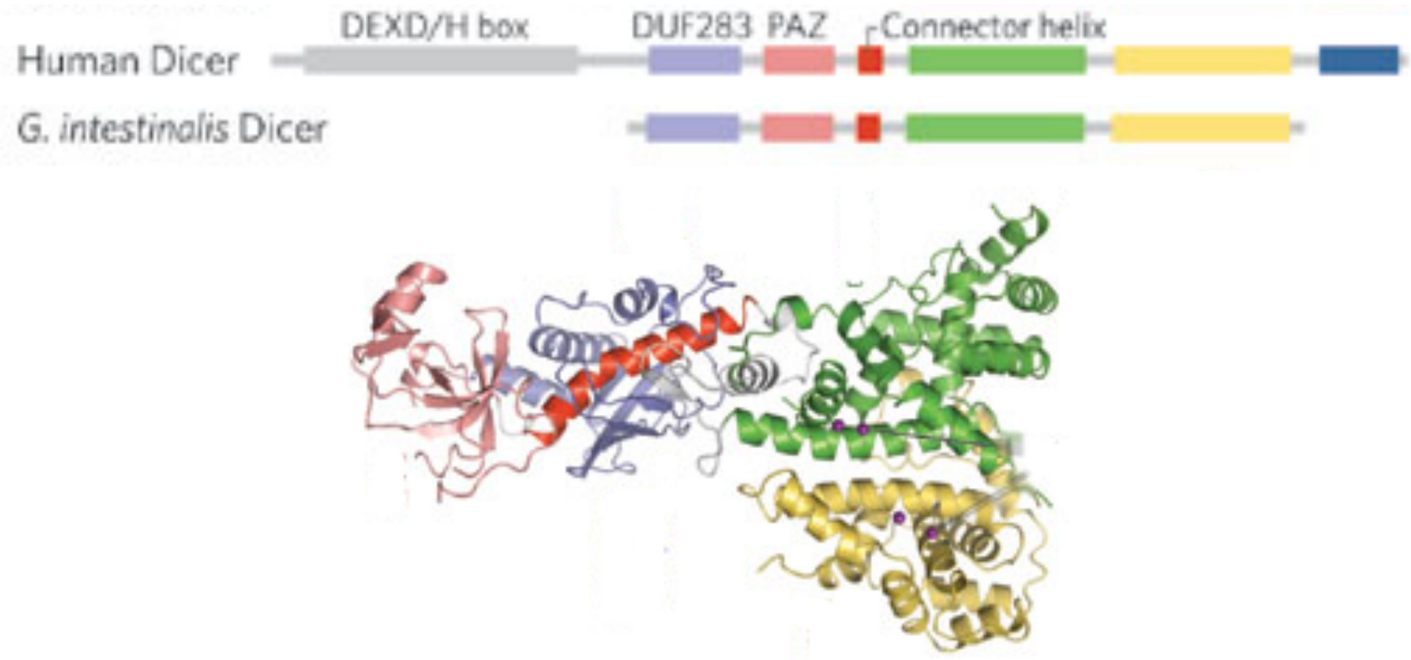
Sequences are diverged beyond recognition at those time scales

We develop tools to reliably uncover homologous relationships by comparing multiple sequence alignments of closer homologs

Domains are the building blocks of proteins

- their structural, functional, and evolutionary units

- Most eurkaryotic proteins have multiple structural domains
- Domains have often been duplicated and rearranged during evolution



We can often formulate hypotheses about protein function based on its domains

Many parts in eukaryotic proteins are *disordered* (or *natively unfolded*)

Fraction of proteins with predicted natively unfolded region longer than 50 residues: [Dunker et al., Genome Informatics (2000)]

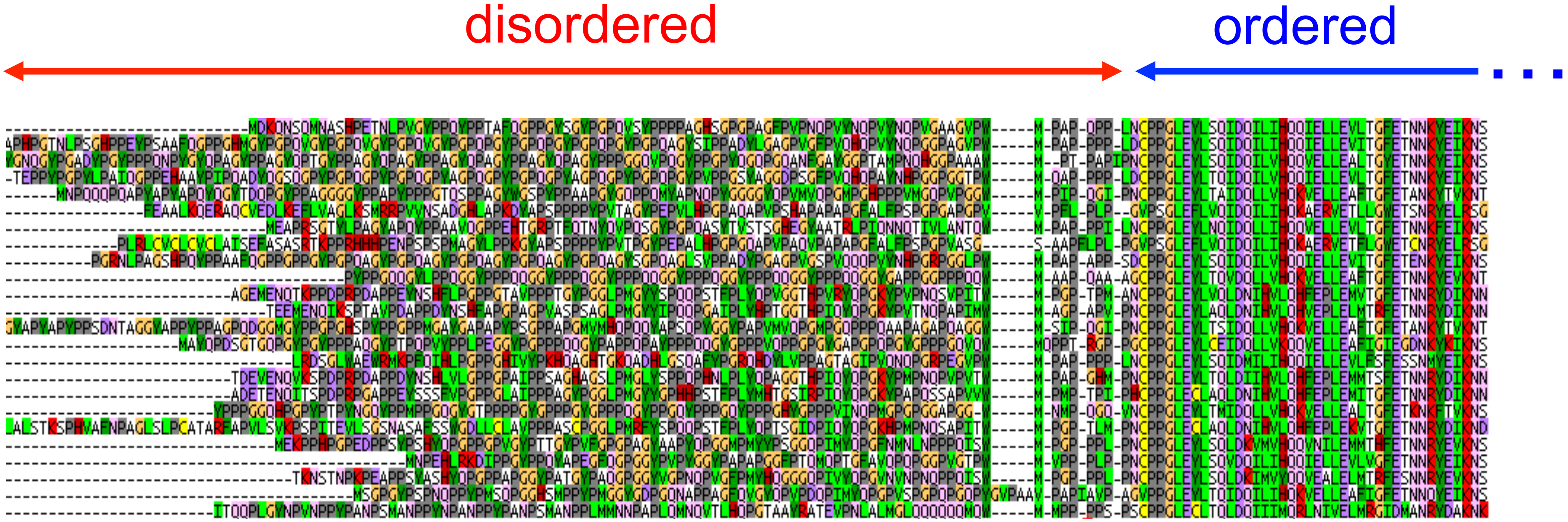
- >50% in humans
- ~30% in *C. elegans*, *A. thaliana* and *S. cerevisiae*
- 3%-25% in bacteria and archaea

What do they do?

Disordered regions often diverge quickly

No selection for folding ⇒ less conservation

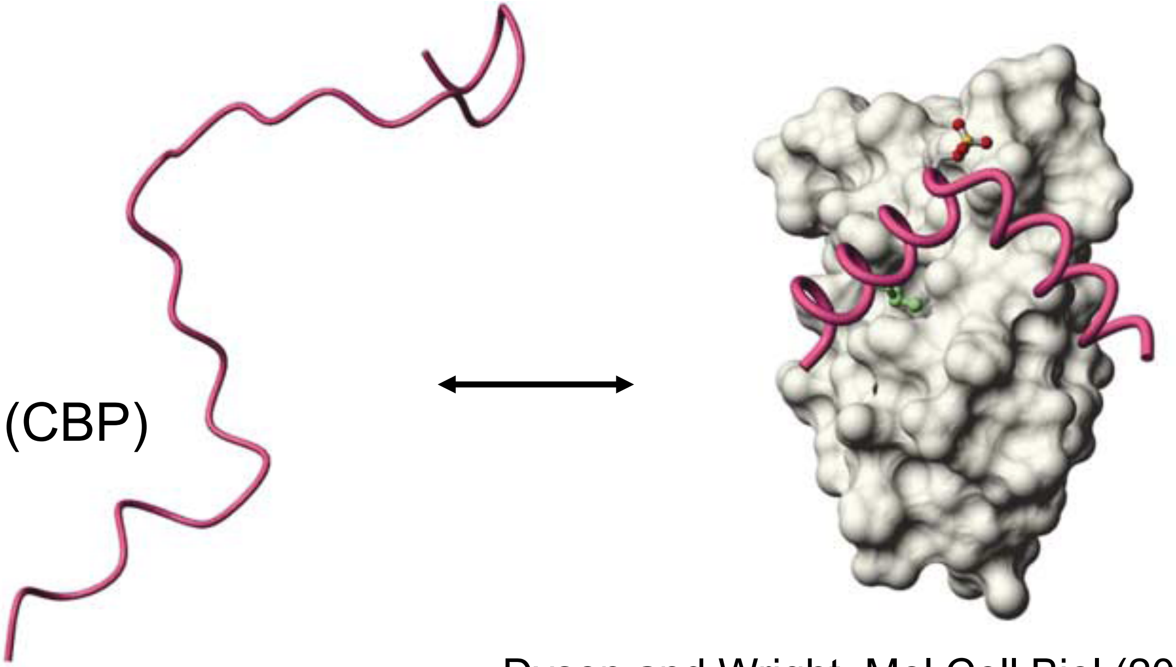
Disordered activation domain of Phospholipid scramblase 1



searches with sequences containing disordered regions tend to generate false positive matches!

Disordered regions are interspersed with **short linear motifs** that can bind to specific target domains

pKID domain of CREB
binding to KIX domain
of CREB-binding protein (CBP)

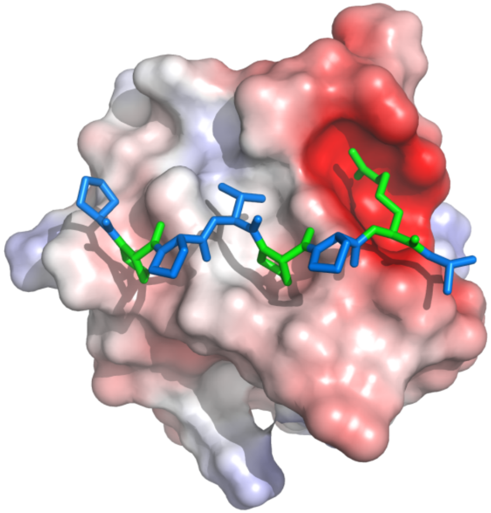


Dyson and Wright, Mol Cell Biol (2005)

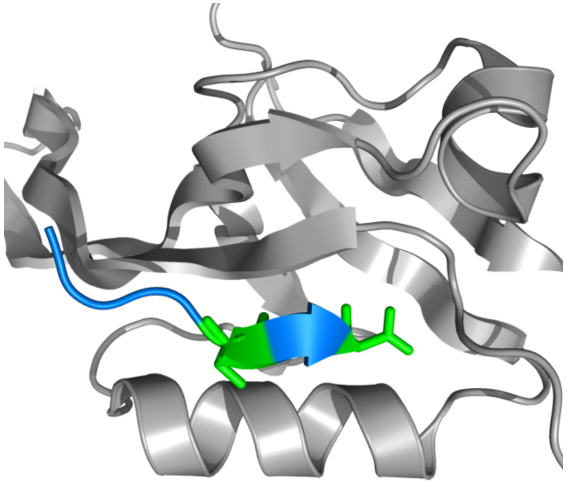
Short linear motifs fold upon binding to their target domain

Short linear motifs mediate regulatory protein-protein interactions

SH3 domain ↔ PxxPx[KR]



PDZ domain ↔ [ST]x[VIL]$



- **Dominant mechanism for transitory protein-protein interactions** (intracellular signaling, protein recruitment, targeted transport,…)
- Motif accessibility often regulated by post-translational modifications (phosphorylation, methylation, dimerization etc.)
- **Low affinity: 5 to 150 μM !** ⇒ hard to discover experimentally
- Partial conservation
- Key interactions involved in **liquid-liquid phase separation**

Liquid-liquid phase separation – a long-known phenomenon now revolutionizing cell biology

Tutorial:

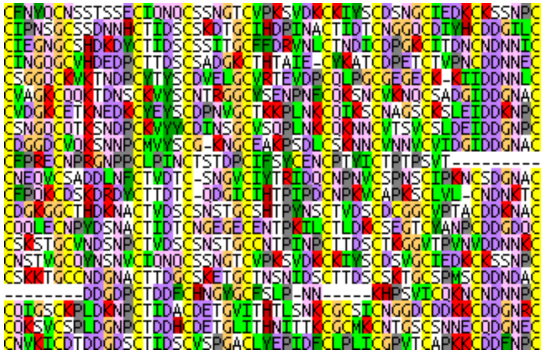
- Get familiar with Uniprot, BLAST, PDB
- Search for Pfam domains
- Build multiple sequence alignment with HHblits
- Check alignment visually with JalView
- Build model using Modeller

The linux command line (bash)

1. Don’t forget spaces
2. Everything in linux is case-sensitive (filenames, commands,..)
3. A filename consists of a directory path and a basename:
/usr/local/soeding/my_file.txt
You can give only the basename *if the file is in the current directory*

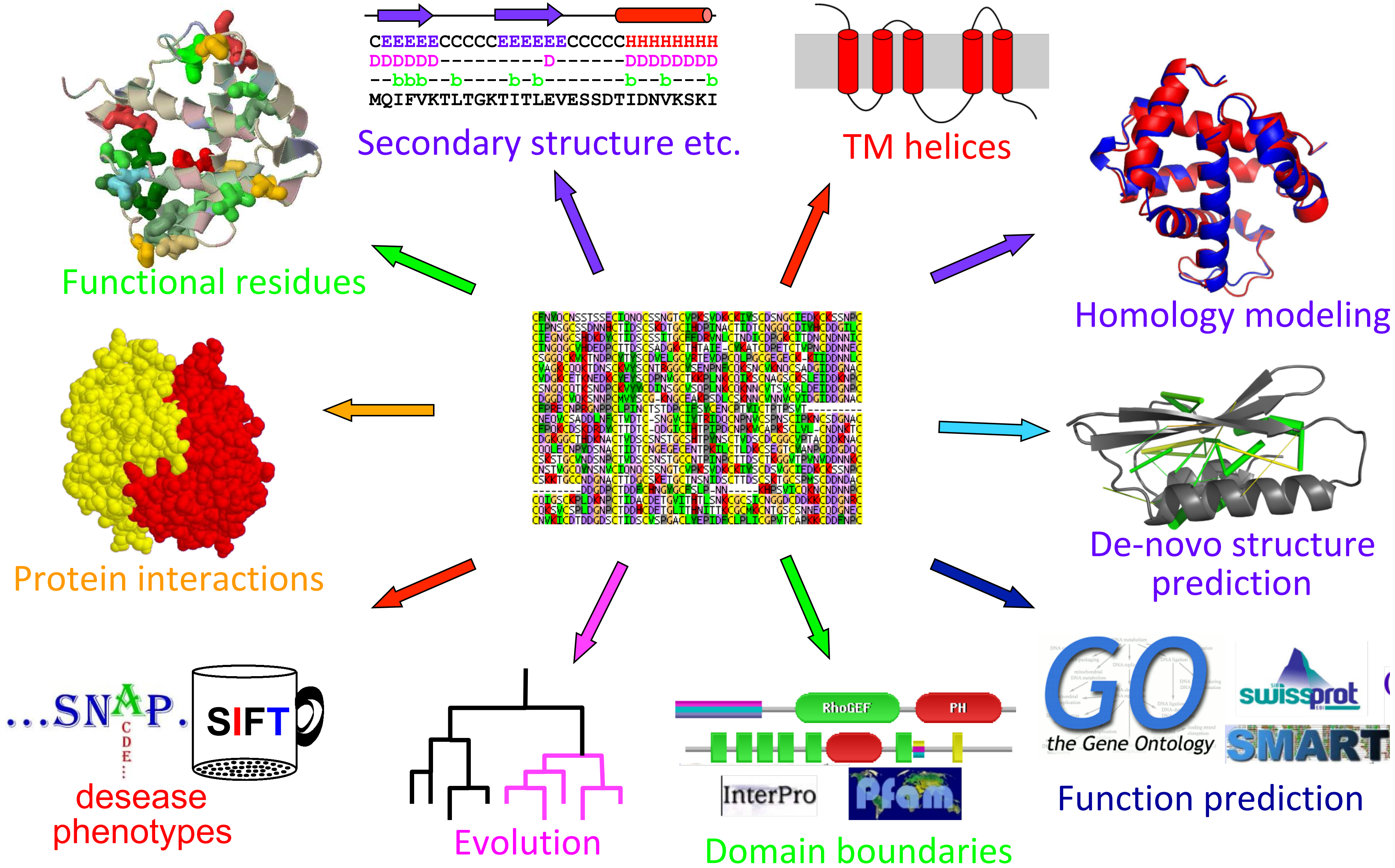
ls	list content of current directory
ls -ltrF	ls in long format, time-sorted in reverse order, with Filetype
cd <path/dir>	change to directory <path/dir>
cd ..	go up 1 step in directory hierarchy
gedit <file>	open file in editor
gedit <file> &	open file in editor *in background*
less <file>	look at file (to quit type q); works for huge files
cp <file> <dest>	copy file to destination directory (cp file.txt ~/molbiol/day1/)
mv <file> <dest>	move file to destination directory
rm <file>	remove file (careful!)
mkdir <dir>	create new directory (remove with rmdir <dir>)
info ls, man ls	show info / manual page of ls command

Sequence searching



```
CFNYQCNSSTSSECIQNQCSSNGTCVPKSVDKCKIYSCDSNGCIEDKCKSSNPC
CIPNSGCSSDNNHCTIDSCSKDTGCIHDPINACTIDTCNGGQCDIYHCDDGILC
CIEGNGCSHDKDYCTIDSCSSITGCFFDRVNLCTNDICDPGKCITDNCNDNNIC
CINGQGCVHDEDPCTTDSCSADGKCTHTAIE-CYKATCDPETCTVPNCDDNNEC
CSGGQCKVKTNDPCVTYSCDVELGCVRTEVDPCQLPGCGEGECK-KIIDDNNLC
CVAGKCQQKTDNSCKVYSCNTRGGCVSENPNFCQKSNCVKNQCSADGIDDGNAC
CVDGKCETKNEDKCYEYSCDPNVGCTKKPLNKCQIKSCNAGSCKSLEIDDKNPC
CSNGQCQTKSNDPCKVYYCDINSGCVSQPLNKCQKNNCVTSVCSLDEIDDGNPC
CDGGDCVQKSNNPCMVYSCG-KNGCEAKPSDLCSKNNCVNNVCVIDGIDDGNAC
CFPRECNPRGNPPCLPINCTSTDPCIFSYCENCPTYICTPTPSVT---------
CNEQVCSADDLNFCTVDTC-SNGVCIYTRIDQCNPNVCSPNSCIPKNCSDGNAC
CFPQKCDSKDRDYCTTDTC-QDGICIHTPIPDCNPKVCAPKSCLVL-CNDNKTC
CDGKGGCTHDKNACTVDSCSNSTGCSHTPYNSCTVDSCDCGGCVPTACDDKNAC
CQQLECNPYDSNACTIDTCNGEGECENTPKILCTLDKCSEGTCYANPCDDGDQC
CSKSTGCVNDSNPCTVDSCSNSTGCCNTPINPCTTDSCTKGGVTPVNVDDNNKC
CNSTVGCQYNSNVCIQNQCSSNGTCVPKSVDKCKIYSCDSVGCIEDKCKSSNPC
CSKKTGCCNDGNACTTDGCSKETGCTNSNIDSCTTDSCSKTGCSPMSCDDNDAC
--------DDGDPCTDDFCHNGVGCFSLP-NN-----KHPSVICQKNCNDNNPC
CQIGSCKPLDKNPCTIDACDETGVITHTLSNKCGCSICNGGDCDDKKCDDGNRC
CQKSVCSPLDGNPCTDDHCDETGLITHNITTKCGCMKCNTGSCSNNECQDGNEC
CNVKICDTDDGDSCTIDSCVSPGACLYEPIDFCLPLICGPVTCAPKKCDDFNPC
```

Sequence searches are at the basis of most of protein bioinformatics



Functional residues

Secondary structure etc.

TM helices

Homology modeling

Protein interactions

De-novo structure prediction

desease phenotypes

Evolution

Domain boundaries

Function prediction

CEEEEECCCCCEEEEEECCCCCHHHHHHHH
DDDDDD---------D------DDDDDDDD
--bbb--b----b-b-------b--b---b
MQIFVKTLTGKTITLEVESSDTIDNVKSKI
...SNAP.
C
D
E
:
SIFT
RhoGEF
PH
InterPro
Pfam
GO
the Gene Ontology
swissprot
SMART

Sequence-sequence comparison

- A sequence alignment groups similar residues into same column. These residues are assumed to occupy homologous positions in the proteins

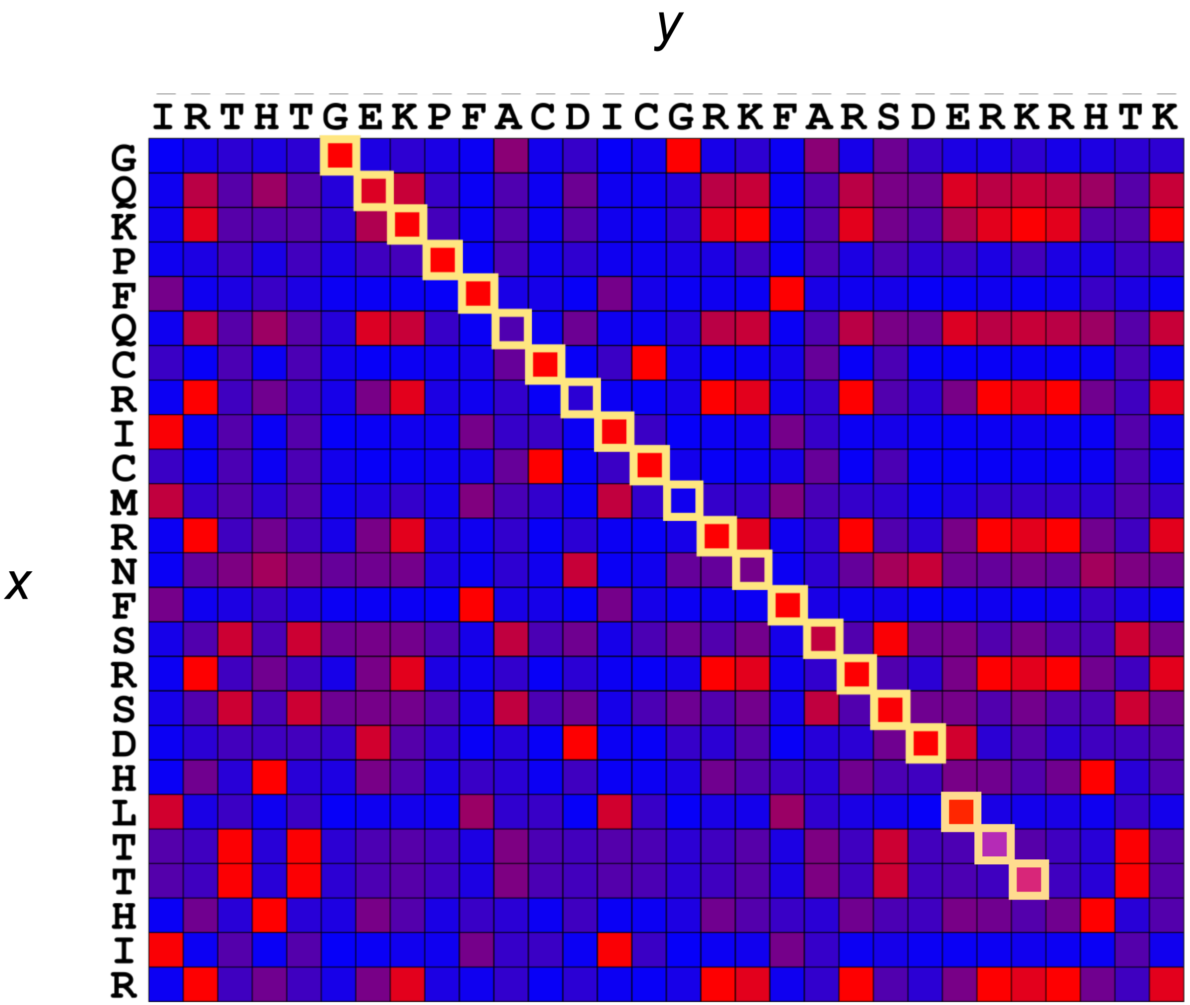
```
HBA_human   ... VKAAWGKVGA--HAGEYGAE ...

GLB1_glydi  ... IAATWEEIAGADNGAGVGKD ...
```

- Alignment score = sum of **similarity** **scores** − gap penalties:

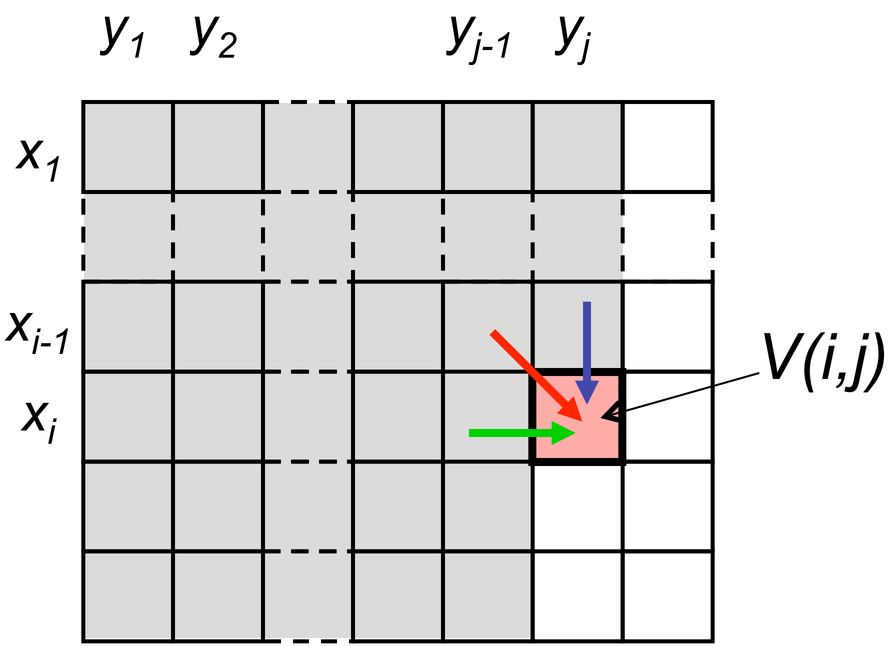
  Score = S(V,I)+…+S(V,I)+…+S(E,G)+…+S(G,G) – d – e

- Find alignment with maximum score, rank by score

Sequence alignment: maximize sum of amino acid similarity scores



Dynamic programming finds the sequence-sequence alignment with highest score



alignment ending in:

$. . . \; x_{i-1} \; x_i$
$. . . \; y_{j-1} \; y_j$

$. . . \; x_i \; -$
$. . . \; y_{j-1} \; y_j$

$. . . \; x_{i-1} \; x_i$
$. . . \; y_j \; -$

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i,y_j) \\ V(i,j-1) - gap.penalty \\ V(i-1,j) - gap.penalty \end{cases}$$

substitution matrix

Exercise: find the alignment with highest score by dynamic programming!

	G	A	A	T	T	C	A	G	T	T
A	0	1	1	0	0	0				
T	0	0	0	2	1	0				
T	0	0	0	1						
A	0	1	1	0						
G	1	0	0	0						
G	1	0	0	0						
T	0	0	0	1						
T	0	0	0	1						
T	0	0	0	1						

match = +1
mismatch = -1
gap.penalty= -1

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i,y_j) \\ V(i,j-1) - \text{gap.penalty} \\ V(i-1,j) - \text{gap.penalty} \end{cases}$$

Exercise: find the alignment with highest score by dynamic programming!

	G	A	A	T	T	C	A	G	T	T
A	0	1	1	0	0	0	1	0	0	0
T	0	0	0	2	1	0	0	0	1	1
T	0	0	0	1	3	2	1	0	1	2
A	0	1	1	0	2	2	3	2	1	1
G	1	0	0	0	1	1	2	4	3	2
G	1	0	0	0	0	0	1	3	3	2
T	0	0	0	1	1	0	0	2	4	4
T	0	0	0	1	2	1	0	1	3	5
T	0	0	0	1	2	1	0	0	2	4

match = +1

mismatch = -1

gap.penalty= -1

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i,y_j) \\ V(i,j-1) - \text{gap.penalty} \\ V(i-1,j) - \text{gap.penalty} \end{cases}$$

Exercise: find the alignment with highest score by dynamic programming!

	G	A	A	T	T	C	A	G	T	T
A	0	1	1	0	0	0	1	0	0	0
T	0	0	0	2	1	0	0	0	1	1
T	0	0	0	1	3	2	1	0	1	2
A	0	1	1	0	2	2	3	2	1	1
G	1	0	0	0	1	1	2	4	3	2
G	1	0	0	0	0	0	1	3	3	2
T	0	0	0	1	1	0	0	2	4	4
T	0	0	0	1	2	1	0	1	3	**5**
T	0	0	0	1	2	1	0	0	2	4

match = +1

mismatch = -1

gap.penalty= -1

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i,y_j) \\ V(i,j-1) - \text{gap.penalty} \\ V(i-1,j) - \text{gap.penalty} \end{cases}$$

Exercise: find the alignment with highest score by dynamic programming!

	G	A	A	T	T	C	A	G	T	T
A	0	1	**1**	0	0	0	1	0	0	0
T	0	0	0	**2**	1	0	0	0	1	1
T	0	0	0	1	**3**	**2**	1	0	1	2
A	0	1	0	0	2	2	**3**	2	1	1
G	1	0	0	0	1	1	2	**4**	3	2
G	1	0	0	0	0	0	1	**3**	3	2
T	0	0	0	1	1	0	0	2	**4**	4
T	0	0	0	1	2	1	0	1	3	**5**
T	0	0	0	1	2	1	0	0	2	4

match = +1

mismatch = -1

gap.penalty= -1

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i,y_j) \\ V(i,j-1) - \text{gap.penalty} \\ V(i-1,j) - \text{gap.penalty} \end{cases}$$

```
GAATTCAG-TT-
--ATT-AGGTTT
```

Exercise: find the alignment with highest score by dynamic programming!

	G	A	A	T	T	C	A	G	T	T
A	0	1	1	0	0	0	1	0	0	0
T	0	0	0	2	1	0	0	0	1	1
T	0	0	0	1	3	2	1	0	1	2
A	0	1	0	0	2	2	3	2	1	1
G	1	0	0	0	1	1	2	4	3	2
G	1	0	0	0	0	0	1	3	3	2
T	0	0	0	1	1	0	0	2	4	4
T	0	0	0	1	2	1	0	1	3	5
T	0	0	0	1	2	1	0	0	2	4

match = +1
mismatch = -1
gap.penalty= -1

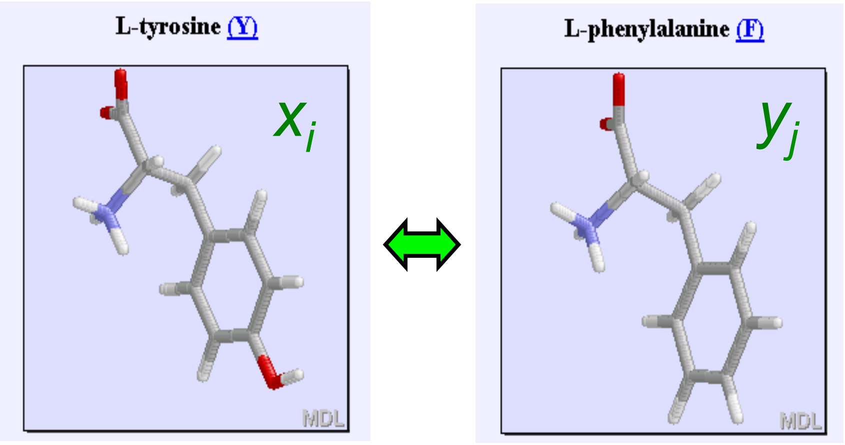
$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i, y_j) \\ V(i,j-1) - \text{gap.penalty} \\ V(i-1,j) - \text{gap.penalty} \end{cases}$$

```
GAATTCA-GTT-
--ATT-AGGTTT
```

Point mutations between similar amino acids may not disturb protein structure or function

$$S(x_i,y_j) = \log \frac{P(x_i,y_j)}{P(x_i)\,P(y_j)}$$

↑ Log-odds score

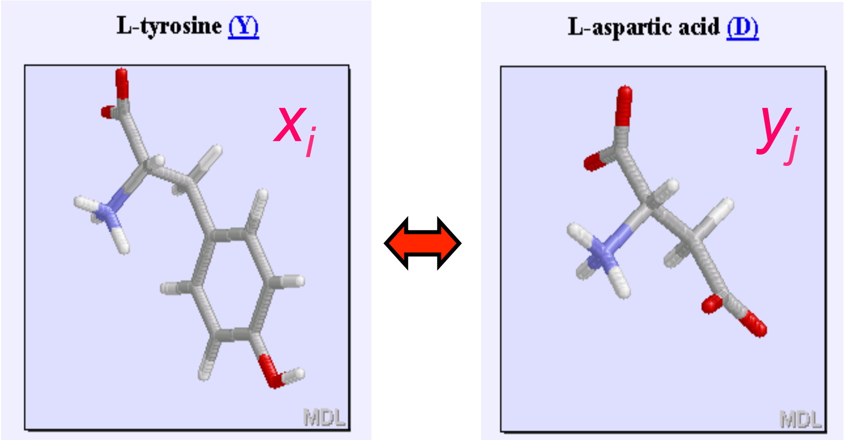


Frequent mutations get positive substitution matrix scores

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Point mutations between dissimilar amino acids often damage the protein

$$S(x_i,y_j) = \log \frac{P(x_i,y_j)}{P(x_i)\,P(y_j)}$$



Rare substitutions get negative substitution matrix scores

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

When searching for homologous proteins, search with the protein sequence, not the DNA sequence!

Why?

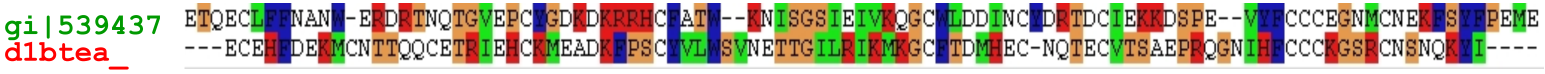
Selection of mutations in *coding regions* **acts on the level of codons and amino acids**, not on the level of nucleotides.

When comparing nucleotides sequences we ignore the differences in selection pressure between

- silent mutations (which don‘t change the amino acid),
- conservative muitations (which lead to substitution with a similar amino acid)
- Non-conservative mutations (which lead to substitution with a dissimilar amino acid) and
- Nonsense mutations (which introduce a stop codon)

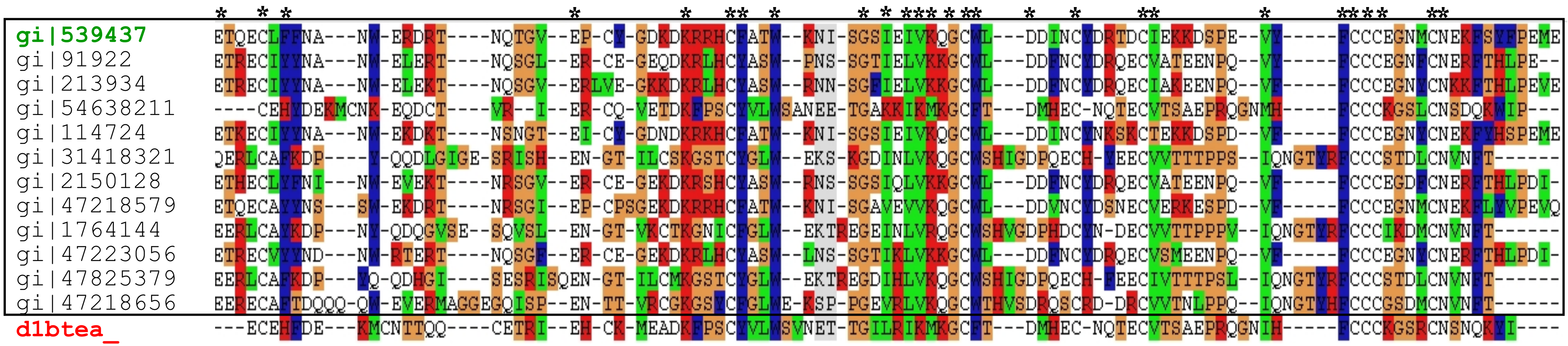
Key message:
Information is power. Use it!

Are these sequences homologous?



```
gi|539437  ETQECLFFNANW-ERDRTNQTGVEPCYGDKDKRRHCFATW--KNISGSIEIVKQGCWLDDINCYDRTDCIEKKDSPE--VYFCCCEGNMCNEKFSYFPEME
d1btea_    ---ECEHFDEKMCNTTQQCETRIEHCKMEADKFPSCYVLWSVNETTGILRIKMKGCFTDMHEC-NQTECVTSAEPRQGNIHFCCCKGSRCNSNQKYI----
```

BLAST E-value = 0.2



```
              *   * *                         *         *   **  *       * * *** ***    *   *     **         *      ****    **
gi|539437     ETQECLFFNA---NW-ERDRT----NQTGV--EP-CY-GDKDKRRHCFATW--KNI-SGSIEIVKQGCWL---DDINCYDRTDCIEKKDSPE--VY-----FCCCEGNMCNEKFSYFPEME
gi|91922      ETRECIYYNA---NW-ELERT----NQSGL--ER-CE-GEQDKRLHCYASW--PNS-SGTIELVKKGCWL---DDFNCYDRQECVATEENPQ--VY-----FCCCEGNFCNERFTHLPE--
gi|213934     ETRECIYYNA---NW-ELEKT----NQSGV--ERLVE-GKKDKRLHCYASW--RNN-SGFIELVKKGCWL---DDFNCYDRQECIAKEENPQ--VF-----FCCCEGNYCNKKFTHLPEVE
gi|54638211   ----CEHYDEKMCNK-EQDCT----VR--I--ER-CQ-VETDKFPSCYVLWSANEE-TGAKKIKMKGCFT---DMHEC-NQTECVTSAEPRQGNMH-----FCCCKGSLCNSDQKWIP---
gi|114724     ETKECIYYNA---NW-EKDKT----NSNGT--EI-CY-GDNDKRKHCFATW--KNI-SGSIEIVKQGCWL---DDINCYNKSKCTEKKDSPD--VF-----FCCCEGNYCNEKFYHSPEME
gi|31418321   QERLCAFKDP----Y-QQDLGIGE-SRISH--EN-GT-ILCSKGSTCYGLW--EKS-KGDINLVKQGCWSHIGDPQECH-YEECVVTTTPPS--IQNGTYRFCCCSTDLCNVNFT------
gi|2150128    ETHECLYFNI---NW-EVEKT----NRSGV--ER-CE-GEKDKRSHCYASW--RNS-SGSIQLVKKGCWL---DDFNCYDRQECVATEENPQ--VF-----FCCCEGDFCNERFTHLPDI-
gi|47218579   ETQECAYYNS---SW-EKDRT----NRSGI--EP-CPSGEKDKRRHCFATW--KNI-SGAVEVVKQGCWL---DDVNCYDSNECVERKESPD--VF-----FCCCEGNMCNEKFLYVPEVQ
gi|1764144    EERLCAYKDP---NY-QDQGVSE--SQVSL--EN-GT-VKCTKGNICFGLW--EKTREGEINLVRQGCWSHVGDPHDCYN-DECVVTTPPPV--IQNGTYRFCCCIKDMCNVNFT------
gi|47223056   ETRECVYYND---NW-RTERT----NQSGF--ER-CE-GEKDKRLHCYASW--LNS-SGTIKLVKKGCWL---DDFNCYDRQECVSMEENPQ--VF-----FCCCEGNYCNERFTHLPDI-
gi|47825379   EERLCAFKDP---YQ-QDHGI----SESRISQEN-GT-ILCMKGSTCYGLW--EKTREGDIHLVKQGCWSHIGDPQECH-FEECIVTTTPSL--IQNGTYRFCCCSTDLCNVNFT------
gi|47218656   EERECAFTDQQQ-QW-EVERMAGGEGQISP--EN-TT-VRCGKGSYCFGLWE-KSP-PGEVRLVKQGCWTHVSDRQSCRD-DRCVVTNLPPQ--IQNGTYHFCCCGSDMCNVNFT------
d1btea_       ---ECEHFDE---KMCNTTQQ----CETRI--EH-CK-MEADKFPSCYVLWSVNET-TGILRIKMKGCFT---DMHEC-NQTECVTSAEPRQGNIH-----FCCCKGSRCNSNQKYI----
```

PSI-BLAST E-value = 1E-17

Yes they are!

Sequence profiles are a condensed representation of multiple alignments

```
HBA_human   ... W  G  K  V  G  A  H  A  G  E ...
HBB_human   ... W  G  K  V  -  -  N  V  D  E ...
MYG_phyca   ... W  G  K  V  E  A  D  V  A  G ...
LGB2_luplu  ... W  E  E  F  N  A  N  I  P  K ...
```

The profile contains scores quantifying how frequent the 20 amino acids are in each column of the multiple sequence alignment:

Score = log[$p_j(aa)/p_{av}(aa)$]

p(aa) = frequency of aa in column, incl. pseudo-counts

f_{av}(aa) = freq. of aa in db

		W	G	K	V	G	A	H	A	G	E	
A	...	-3,2	-1,9	-2,1	-2,2	-2,0	3,4	-2,1	1,4	1,5	-2,0	...
C	...	-2,3	-2,8	-2,9	-2,1	-2,7	-1,8	-2,7	-2,1	-2,6	-2,9	...
D	...	-3,7	-1,6	-1,6	-3,1	-1,4	-2,1	2,0	-2,8	1,6	-1,5	...
E	...	-3,4	2,1	2,1	-2,8	2,1	-2,0	-1,6	-2,5	-1,9	2,5	...
F	...	-0,8	-3,6	-3,2	2,9	-3,3	-2,8	-2,8	-2,0	-3,2	-3,3	...
G	...	-3,3	2,9	-2,3	-3,3	1,9	-1,8	-2,0	-2,8	1,5	1,6	...
H	...	-2,3	-2,2	-1,8	-2,4	-1,9	-2,3	2,4	-2,6	-2,3	-2,0	...
I	...	-2,6	-3,3	-2,8	-1,2	-3,1	-2,3	-3,0	2,4	-2,9	-3,0	...
K	...	-3,2	-2,1	3,2	-2,7	-1,9	-2,1	-1,8	-2,5	-2,1	2,1	...
L	...	-2,2	-3,3	-2,8	-1,4	-3,1	-2,4	-3,0	-1,5	-2,9	-3,0	...
M	...	-2,3	-3,0	-2,5	-1,5	-2,8	-2,2	-2,7	-1,5	-2,7	-2,7	...
N	...	-3,2	-1,8	-1,7	-2,8	2,8	-2,1	3,3	-2,6	-1,9	-1,8	...
P	...	-3,7	-2,4	-2,2	-2,8	-2,3	-1,9	-2,3	-2,5	2,6	-2,3	...
Q	...	-2,9	-2,0	-1,5	-2,6	-1,8	-2,1	-1,7	-2,4	-2,0	-1,6	...
R	...	-2,5	-2,2	-1,3	-2,8	-2,0	-2,2	-1,9	-2,6	-2,2	-1,7	...
S	...	-3,1	-1,9	-2,0	-2,5	-1,8	-1,6	-1,8	-2,2	-1,8	-1,9	...
T	...	-3,2	-2,2	-2,0	-2,2	-2,0	-1,8	-1,9	-2,0	-2,0	-2,1	...
V	...	-2,9	-2,9	-2,6	2,9	-2,8	-2,0	-2,8	2,3	-2,6	-2,7	...
W	...	6,1	-3,4	-3,2	-1,9	-3,3	-3,2	-3,0	-2,8	-3,5	-3,3	...
Y	...	-0,6	-3,2	-2,8	-1,4	-2,8	-2,7	-2,6	-2,4	-3,0	-2,9	...

Sequence profiles are also called „position-specific substitution matrices“. Why?

Profiles-sequence comparison

Query profile

```
HBA_human   ... W  G  K  V  G  A  -  -  H  A  G  E ...
HBB_human   ... W  G  K  V  -  -  -  -  N  V  D  E ...
MYG_phyca   ... W  G  K  V  E  A  -  -  D  V  A  G ...
LGB2_luplu  ... W  K  D  F  N  A  -  -  N  I  P  K ...
GLB1_glydi  ... W  E  E  I  A  G  A  D  N  G  A  G ...
```

Matched database sequence

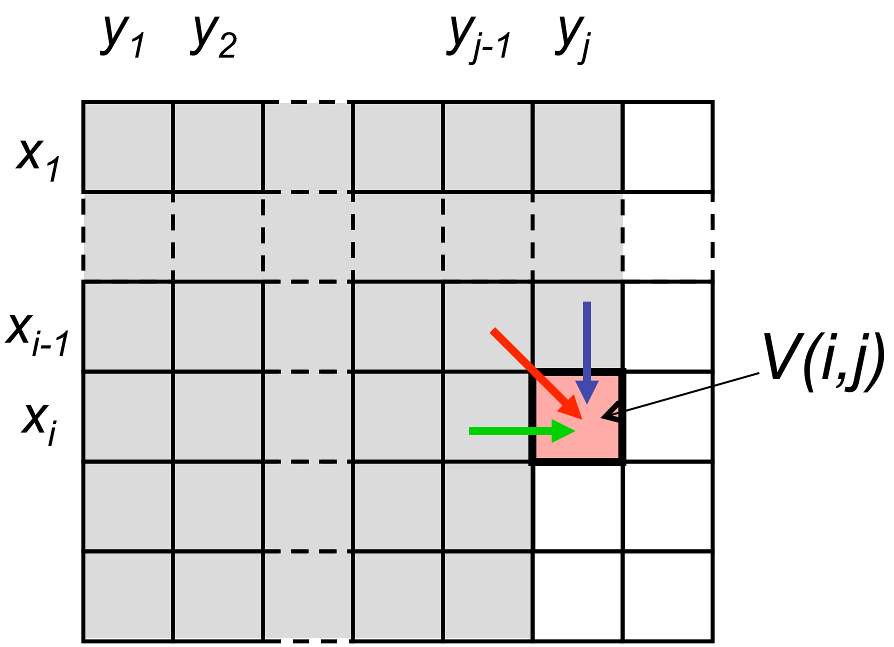
		W	G	K	V	G	A			H	A	G	E	
A	…	-3,2	-1,9	-2,1	-2,2	**-2,0**	3,4			-2,1	1,4	**1,5**	-2,0	…
C	…	-2,3	-2,8	-2,9	-2,1	-2,7	-1,8			-2,7	-2,1	-2,6	-2,9	…
D	…	-3,7	-1,6	-1,6	-3,1	-1,4	-2,1			2,0	-2,8	1,6	-1,5	…
E	…	-3,4	**2,1**	**2,1**	-2,8	2,1	-2,0			-1,6	-2,5	-1,9	2,5	…
F	…	-0,8	-3,6	-3,2	2,9	-3,3	-2,8			-2,8	-2,0	-3,2	-3,3	…
G	…	-3,3	2,9	-2,3	-3,3	1,9	**-1,8**			-2,0	**-2,8**	1,5	**1,6**	…
H	…	-2,3	-2,2	-1,8	-2,4	-1,9	-2,3			2,4	-2,6	-2,3	-2,0	…
I	…	-2,6	-3,3	-2,8	**-1,2**	-3,1	-2,3			-3,0	2,4	-2,9	-3,0	…
K	…	-3,2	-2,1	3,2	-2,7	-1,9	-2,1			-1,8	-2,5	-2,1	2,1	…
L	…	-2,2	-3,3	-2,8	-1,4	-3,1	-2,4			-3,0	-1,5	-2,9	-3,0	…
M	…	-2,3	-3,0	-2,5	-1,5	-2,8	-2,2			-2,7	-1,5	-2,7	-2,7	…
N	…	-3,2	-1,8	-1,7	-2,8	2,8	-2,1			**3,3**	-2,6	-1,9	-1,8	…
P	…	-3,7	-2,4	-2,2	-2,8	-2,3	-1,9			-2,3	-2,5	2,6	-2,3	…
Q	…	-2,9	-2,0	-1,5	-2,6	-1,8	-2,1			-1,7	-2,4	-2,0	-1,6	…
R	…	-2,5	-2,2	-1,3	-2,8	-2,0	-2,2			-1,9	-2,6	-2,2	-1,7	…
S	…	-3,1	-1,9	-2,0	-2,5	-1,8	-1,6			-1,8	-2,2	-1,8	-1,9	…
T	…	-3,2	-2,2	-2,0	-2,2	-2,0	-1,8			-1,9	-2,0	-2,0	-2,1	…
V	…	-2,9	-2,9	-2,6	2,9	-2,8	-2,0			-2,8	2,3	-2,6	-2,7	…
W	…	**6,1**	-3,4	-3,2	-1,9	-3,3	-3,2			-3,0	-2,8	-3,5	-3,3	…
Y	…	-0,6	-3,2	-2,8	-1,4	-2,8	-2,7			-2,6	-2,4	-3,0	-2,9	…

gap penalties

Score = 6.1 +2.1 +2.1 –1.2 –2.0 –1.8 – 5.0 – 0.5 +3.3 –2.8 +1.5 +1.6

⇨ Find alignment with maximum score

Dynamic programming finds sequence-sequence alignment with highest score



alignment ending in:

$. . . \; x_{i-1} \mid x_i$
$. . . \; y_{j-1} \mid y_j$

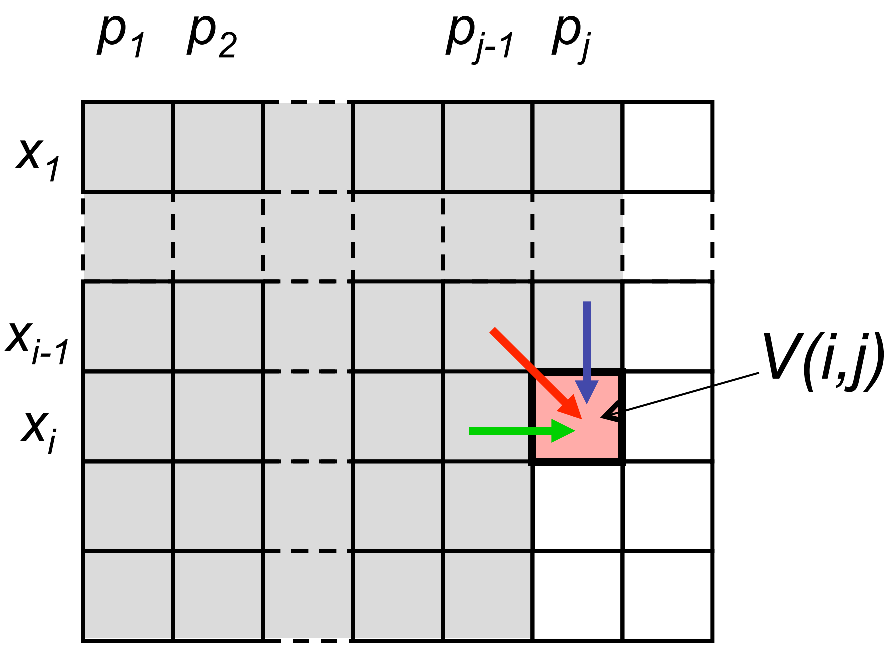
$. . . \; x_i \mid -$
$. . . \; y_{j-1} \mid y_j$

$. . . \; x_{i-1} \mid x_i$
$. . . \; y_j \mid -$

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + S(x_i,y_j) \\ V(i,j-1) - gap.penalty \\ V(i-1,j) - gap.penalty \end{cases}$$

substitution matrix

Dynamic programming used to find profile-sequence alignment with highest score



alignment ending in:

$\ldots x_{i-1} \; x_i$
$\ldots p_{j-1} \; p_j$

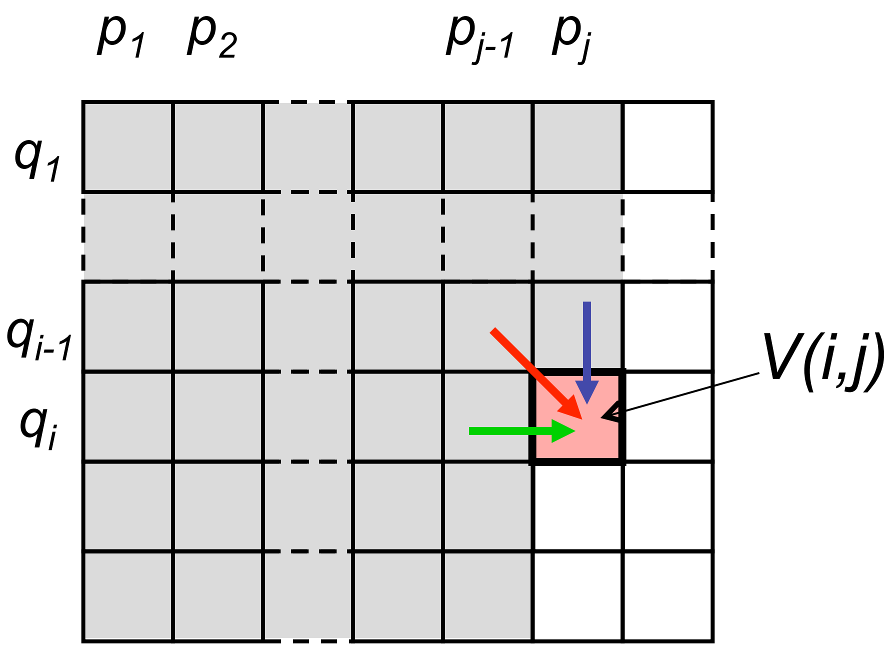
$\ldots x_i \; -$
$\ldots p_{j-1} \; p_j$

$\ldots x_{i-1} \; x_i$
$\ldots p_j \; -$

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + \log \dfrac{p_j(x_i)}{P(x_i)} \\ V(i,j-1) - gap.penalty \\ V(i-1,j) - gap.penalty \end{cases}$$

Profile score

Dynamic programming used to find profile-profile alignment with highest score



alignment ending in:

$\ldots\ q_{i-1}\ |\ q_i$
$\ldots\ p_{j-1}\ |\ p_j$

$\ldots\ q_i\ |\ -$
$\ldots\ p_{j-1}\ |\ p_j$

$\ldots\ q_{i-1}\ |\ q_i$
$\ldots\ p_j\ |\ -$

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + \log \sum_{a=1}^{20} \dfrac{q_i(a)\, p_j(a)}{P(a)} \\ V(i,j-1) - gap.penalty \\ V(i-1,j) - gap.penalty \end{cases}$$

Profile-profile score

Profile-profile comparison

```
HBA_human   ... W  G  K  V  G  A  -  -  H  A  G  E ...
HBB_human   ... W  G  K  V  -  -  -  -  N  V  D  E ...
MYG_phyca   ... W  G  K  V  E  A  -  -  D  V  A  G ...
LGB2_luplu  ... W  E  E  F  N  A  -  -  N  I  P  K ...
```

```
GLB1_glydi  ... W  K  D  I  A  G  A  D  N  G  A  V ...
GLB3_chitp  ... F  D  K  V  K  G  -  -  -  -  -  N ...
GLB5_petma  ... W  A  P  V  Y  S  A  N  T  Y  E  T ...
```

		W	G	K	V	G	A			H	A	G	E	
A	…	-3,2	-1,9	-2,1	-2,2	-2,0	3,4			-2,1	1,4	1,5	-2,0	…
C	…	-2,3	-2,8	-2,9	-2,1	-2,7	-1,8			-2,7	-2,1	-2,6	-2,9	…
D	…	-3,7	-1,6	-1,6	-3,1	-1,4	-2,1			2,0	-2,8	1,6	-1,5	…
…	…	…	…	…	…	…	…			…	…	…	….	…
V	…	-2,9	-2,9	-2,6	2,9	-2,8	-2,0			-2,8	2,3	-2,6	-2,7	…
W	…	6,1	-3,4	-3,2	-1,9	-3,3	-3,2			-3,0	-2,8	-3,5	-3,3	…
Y	…	-0,6	-3,2	-2,8	-1,4	-2,8	-2,7			-2,6	-2,4	-3,0	-2,9	…

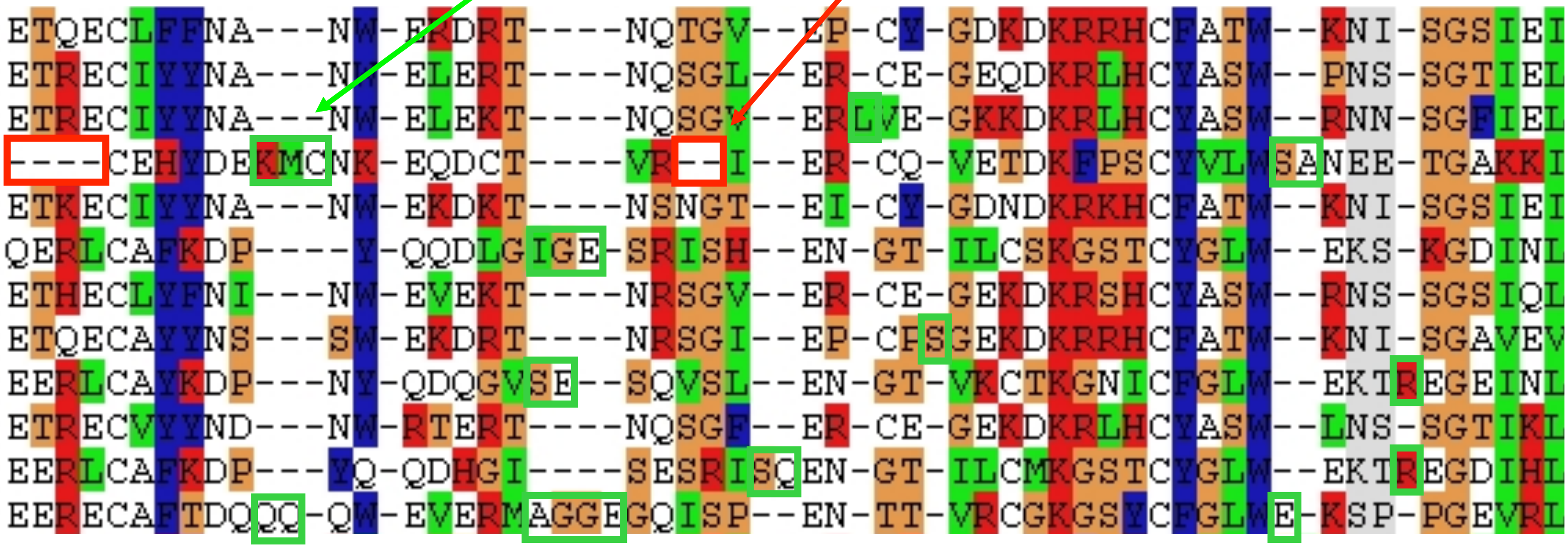
Compare amino acid distributions

		W	K	D	I	A	G	A	D	N	G	A	V	
A		-3,1	1,8	-2,0	-2,1	2,2	-1,8	3,4	-2,1	-2,0	-2,2	2,5	-1,8	…
C		-2,3	-2,5	-3,0	-2,1	-2,2	-2,4	-1,8	-3,1	-2,4	-2,4	-2,2	-2,4	…
D		-3,7	2,0	2,7	-3,1	-2,2	-1,9	-2,1	3,9	-1,6	-2,3	-1,6	-2,0	…
…	…	…	…	…	…	…			…	…	…	….	….	…
V		-2,6	-2,4	-2,7	2,7	-2,2	-2,8	-2,0	-3,0	-2,4	-2,7	-2,2	-2,5	…
W		5,6	-3,3	-3,5	-2,7	-1,8	-3,2	-3,2	-3,7	-3,2	-1,5	-3,3	-3,2	…
Y		-0,5	-2,8	-2,9	-2,3	2,7	-3,1	-2,7	-2,9	-2,5	3,2	-2,8	-3,0	…

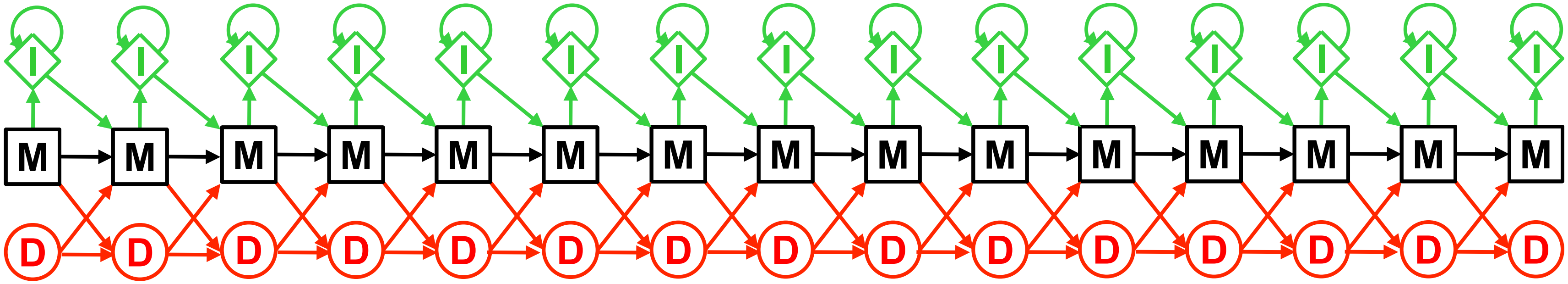
Various ad-hoc measures of column similarity are used, e.g. Score = $\sum_{a=1}^{20} q_{ia} \, p_{ia}$

A profile HMM is a sequence profile extended by position-specific gap penalties

Record probability of insertions and deletions **at each position**

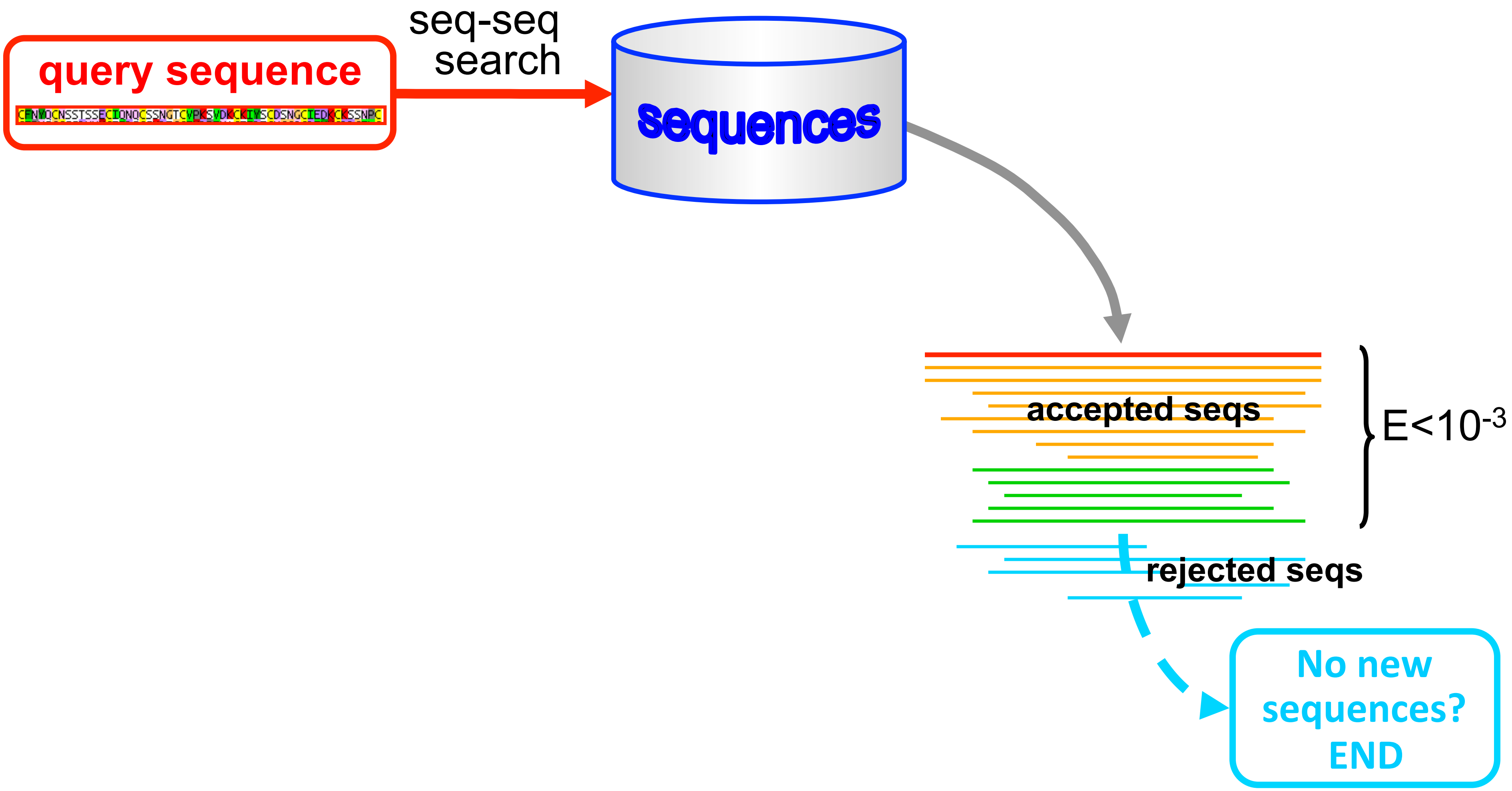


```
ETQECLFFNA---NW-ERDRT----NQTGV--EP-CY-GDKDKRRHCFATW--KNI-SGSIEI
ETRECIYYNA---NW-ELERT----NQSGL--ER-CE-GEQDKRLHCYASW--PNS-SGTIEL
ETRECIYYNA---NW-ELEKT----NQSGV--ERLVE-GKKDKRLHCYASW--RNN-SGFIEL
----CEHYDEKMCNK-EQDCT----VR--I--ER-CQ-VETDKFPSCYVLWSANEE-TGAKKI
ETKECIYYNA---NW-EKDKT----NSNGT--EI-CY-GDNDKRKHCFATW--KNI-SGSIEI
QERLCAFKDP----Y-QQDLGIGE-SRISH--EN-GT-ILCSKGSTCYGLW--EKS-KGDINL
ETHECLYFNI---NW-EVEKT----NRSGV--ER-CE-GEKDKRSHCYASW--RNS-SGSIQL
ETQECAYYNS---SW-EKDRT----NRSGI--EP-CFSGEKDKRRHCFATW--KNI-SGAVEV
EERLCAYKDP---NY-QDQGVSE--SQVSL--EN-GT-VKCTKGNICFGLW--EKTREGEINL
ETRECVYYND---NW-RTERT----NQSGF--ER-CE-GEKDKRLHCYASW--LNS-SGTIKL
EERLCAFKDP---YQ-QDHGI----SESRISQEN-GT-ILCMKGSTCYGLW--EKTREGDIHL
EERECAFTDQQQ-QW-EVERMAGGEGQISP--EN-TT-VRCGKGSYCFGLWE-KSP-PGEVRL
```



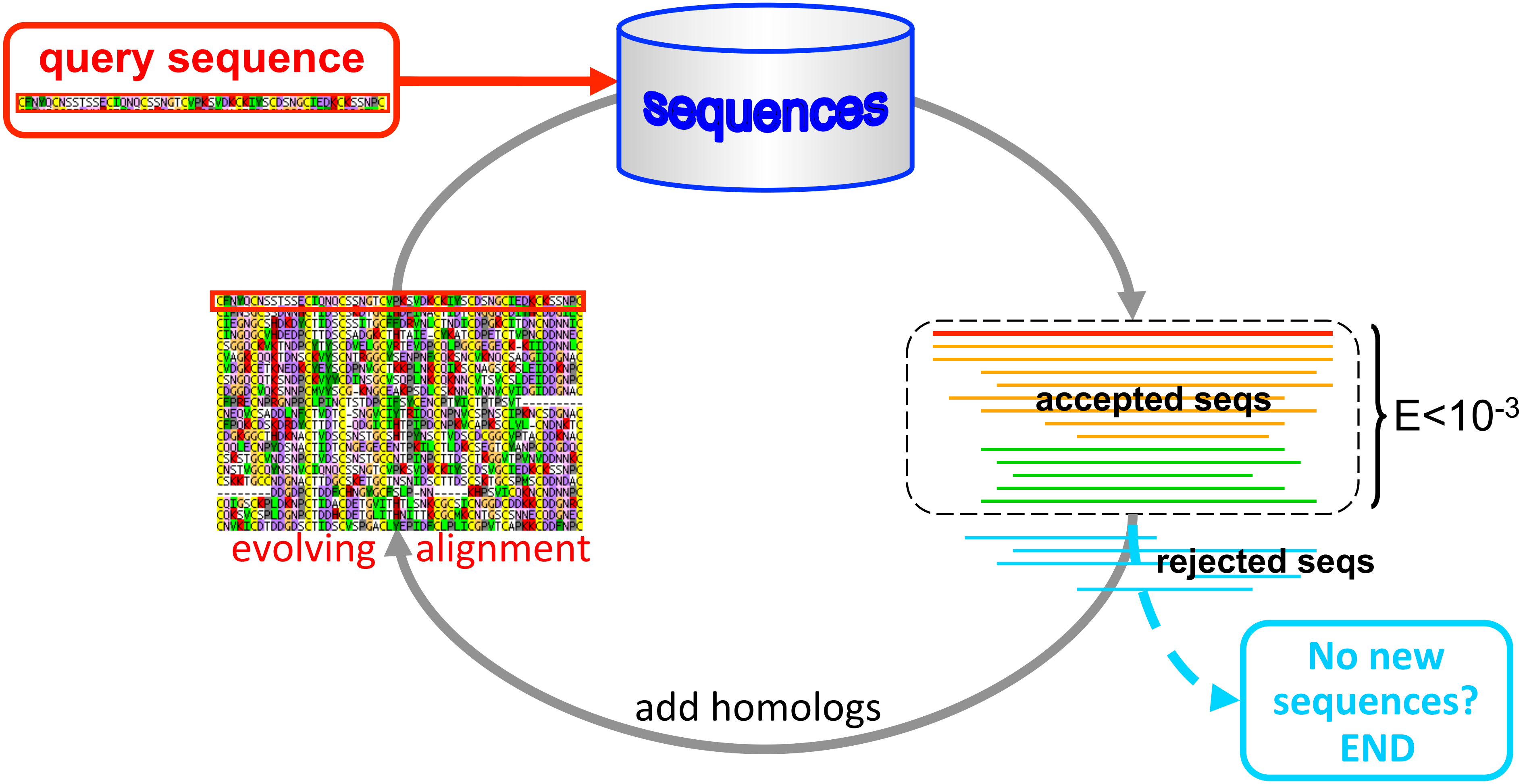
BLAST

Search with single sequence through sequence database



PSI-BLAST

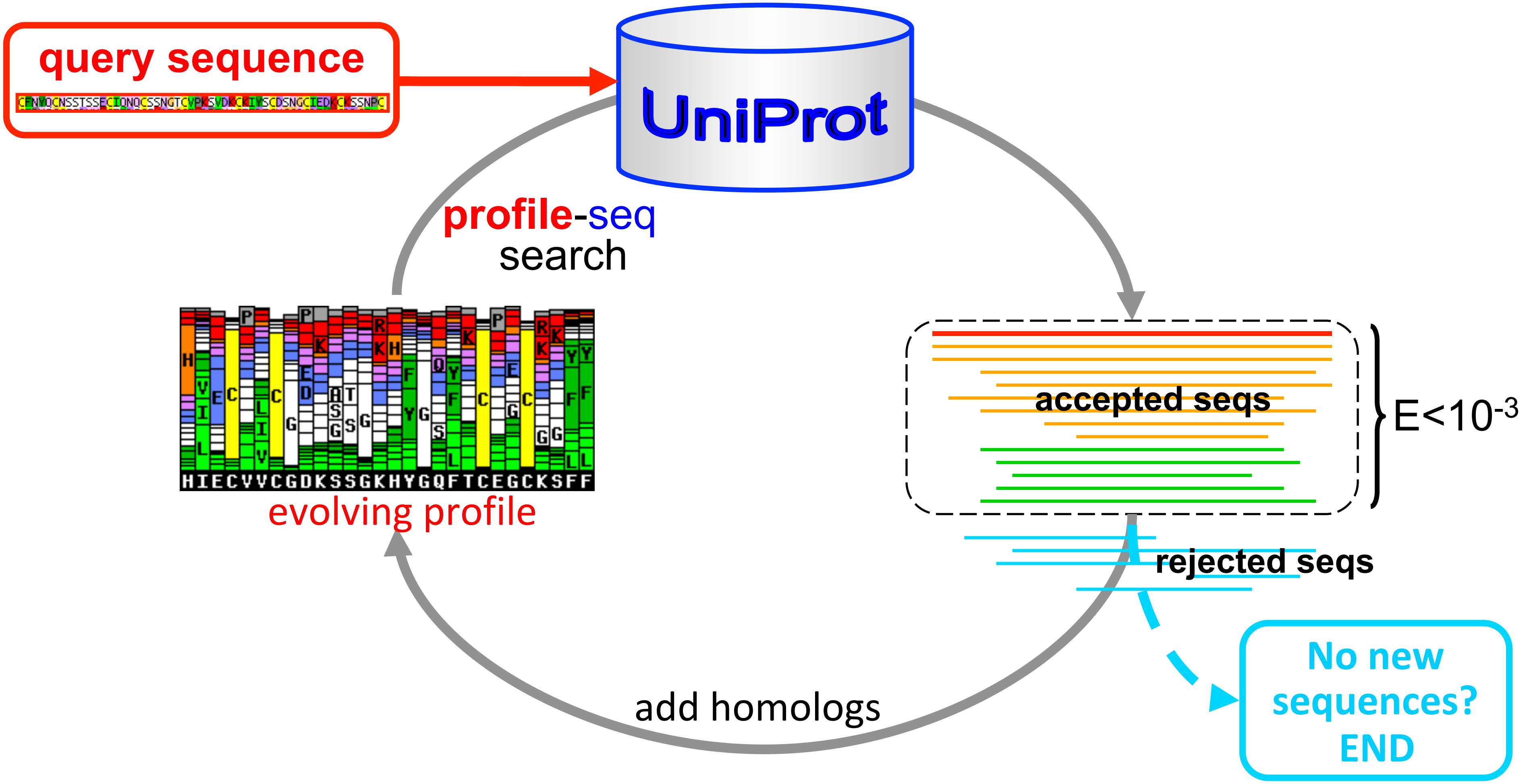
Iterative search with sequence profile through sequence db



Much more sensitive than BLAST

PSI-BLAST, MMseqs2

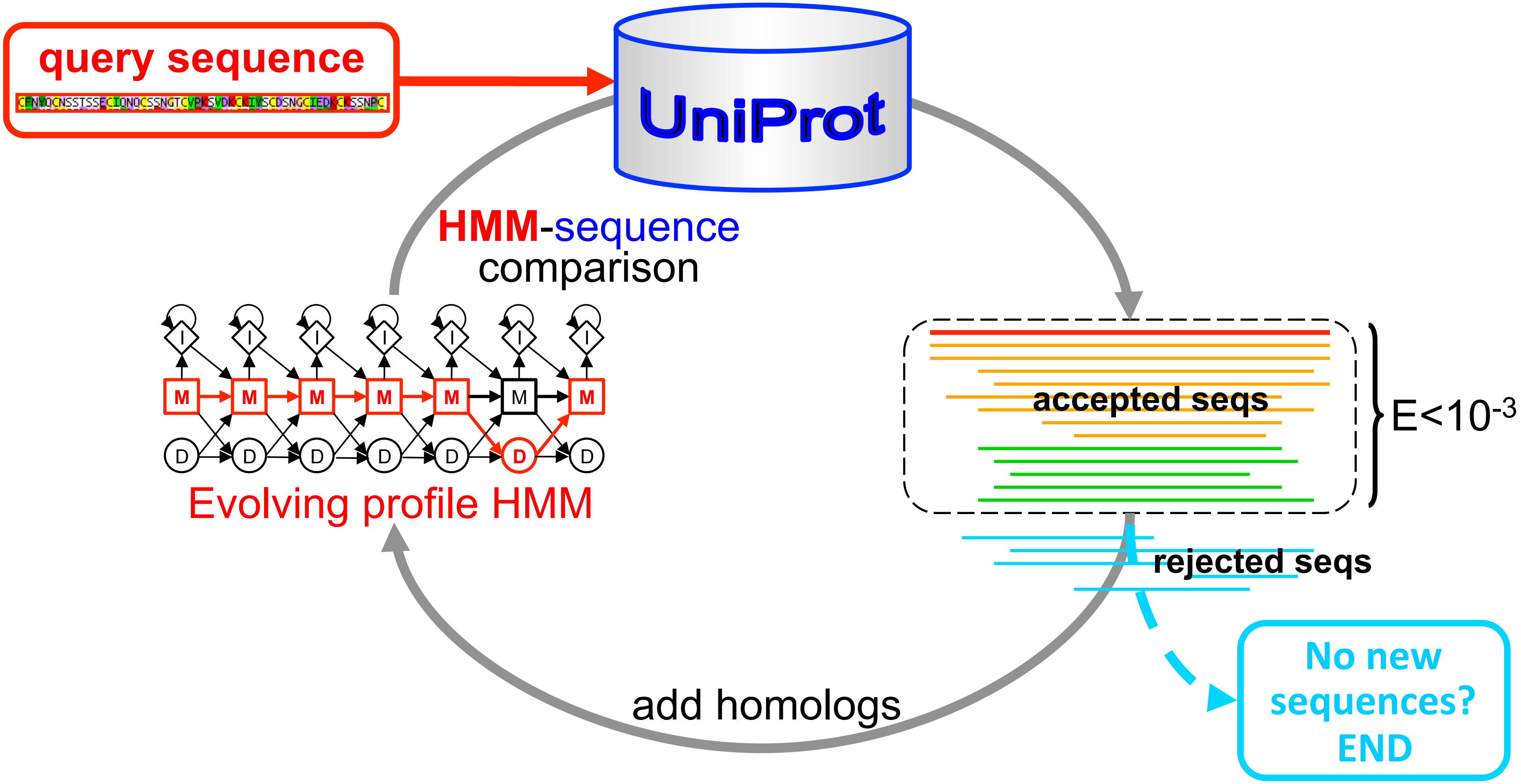
Iterative search with sequence profile through sequence db



Much more sensitive than BLAST

HMMER3

Iterative search with profile HMM through sequence db



query sequence

HMM-sequence comparison

Evolving profile HMM

accepted seqs

$E<10^{-3}$

rejected seqs

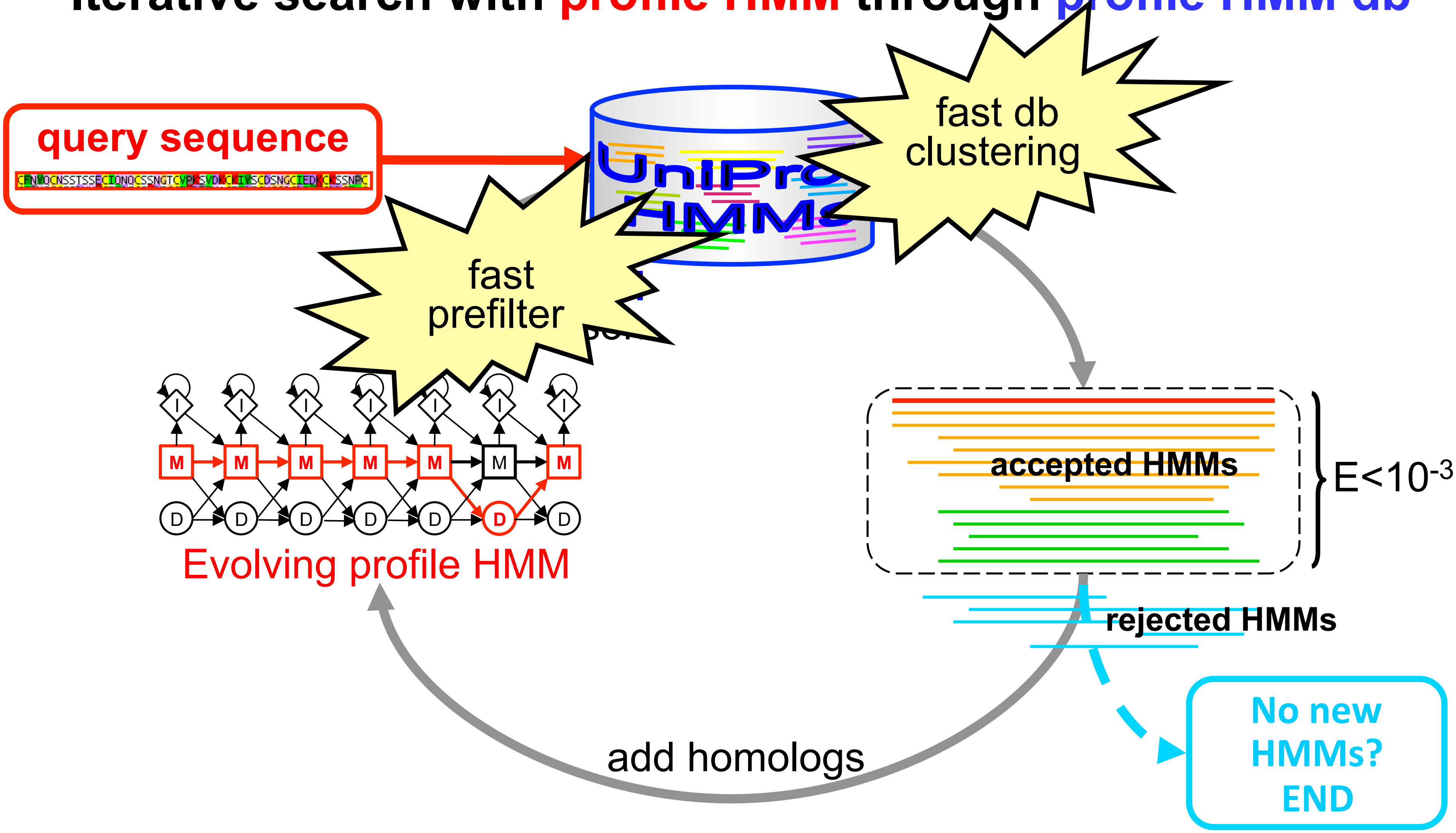
No new sequences? END

add homologs

More sensitive & better alignments than PSI-BLAST

HHblits

Iterative search with profile HMM through profile HMM db



Best sensitivity, alignment quality, and speed

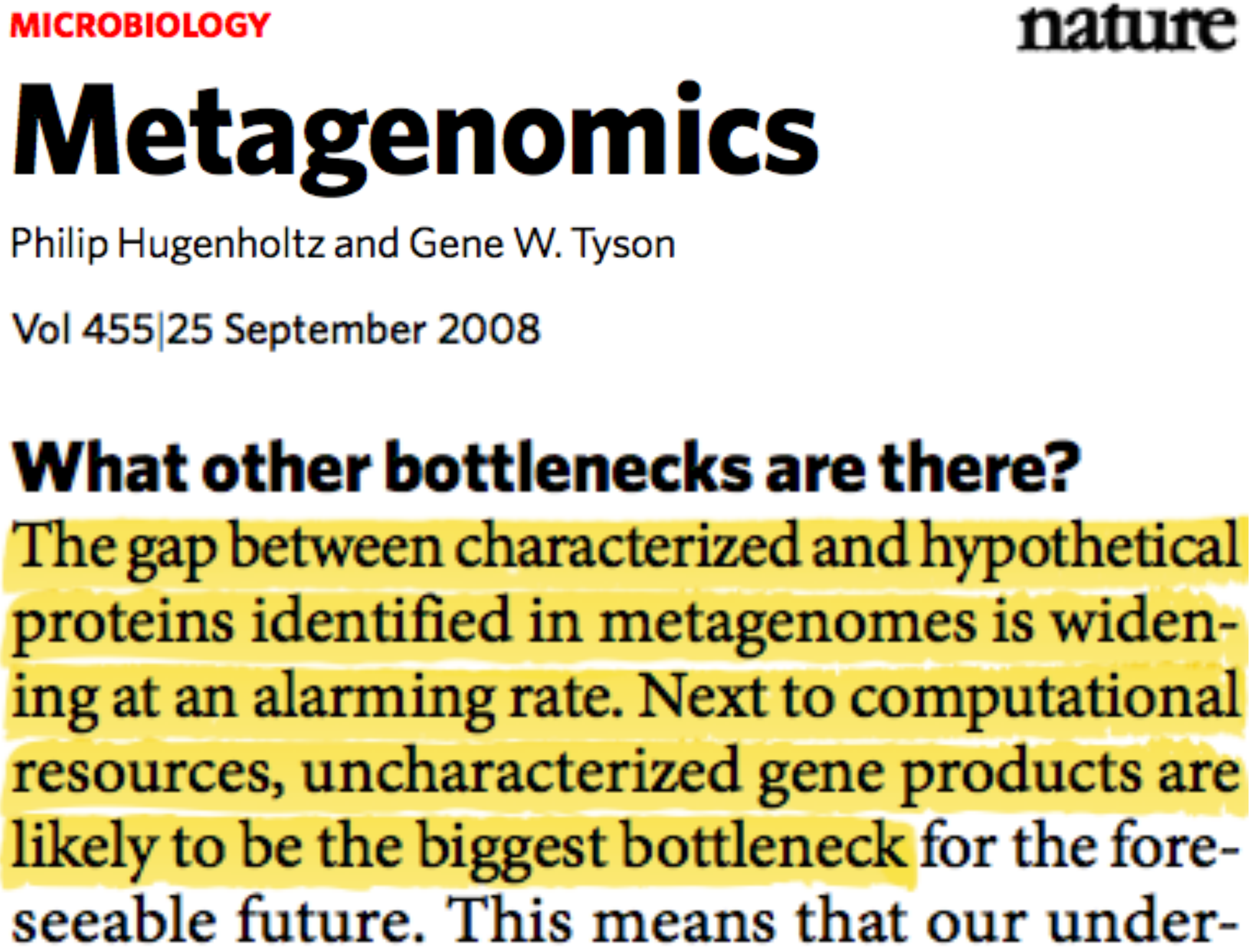
Remmert et al., Nature Methods 2011

MMseqs2

Ultrafast and sensitive sequence and profile searches

Martin Steinegger

with Milot Mirdita, Eli Levy Karin, Clovis Galiez, Ruoshi Zhang



MICROBIOLOGY

nature

Metagenomics

Philip Hugenholtz and Gene W. Tyson

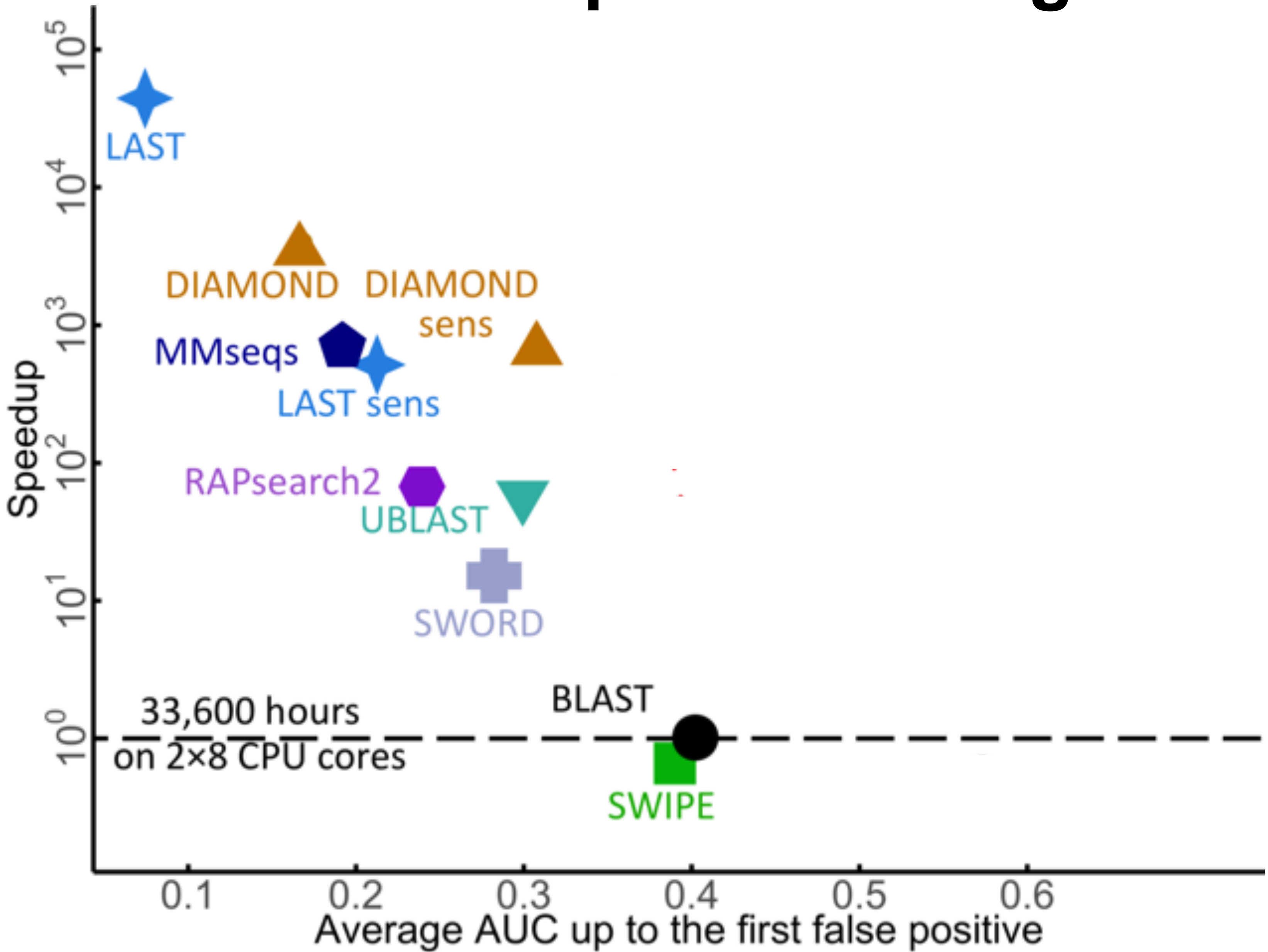
Vol 455|25 September 2008

What other bottlenecks are there?

The gap between characterized and hypothetical proteins identified in metagenomes is widening at an alarming rate. Next to computational resources, uncharacterized gene products are likely to be the biggest bottleneck for the foreseeable future. This means that our under-

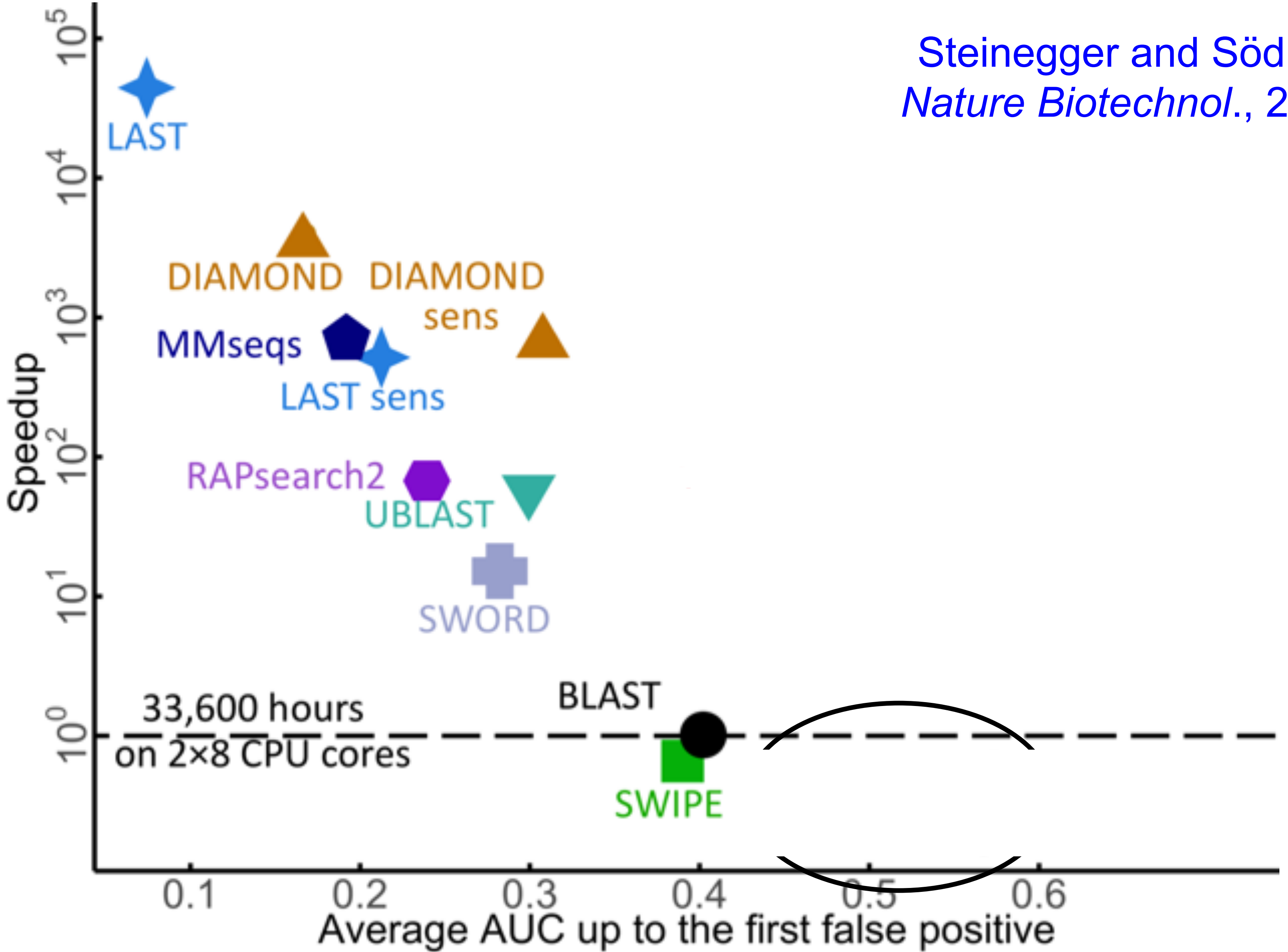
Often, 50%-90% of ORFs remain unannotated: no function, no taxon

Faster but less sensitive search tools have been developed for metagenomics



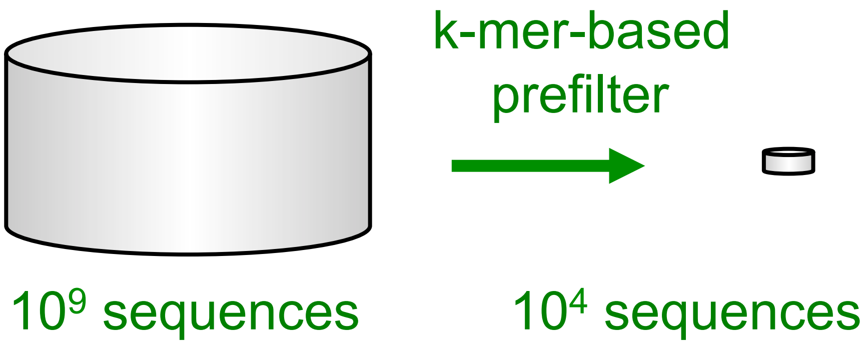
MMseqs profile searches 300 times faster and more sensitive than PSI-BLAST

Steinegger and Söding, *Nature Biotechnol*., 2017.



Fast and sensitive prefilter is most critical part for search performance

Reduces search space 10^5-fold while losing few true positives

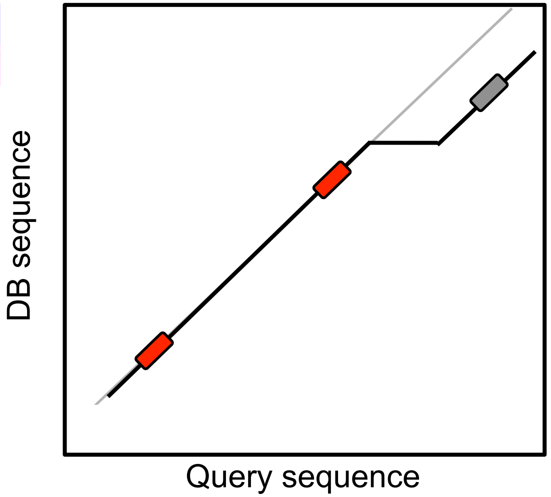


Key ideas for prefilter in MMseqs

- Match *long* & *similar* *k*-mers

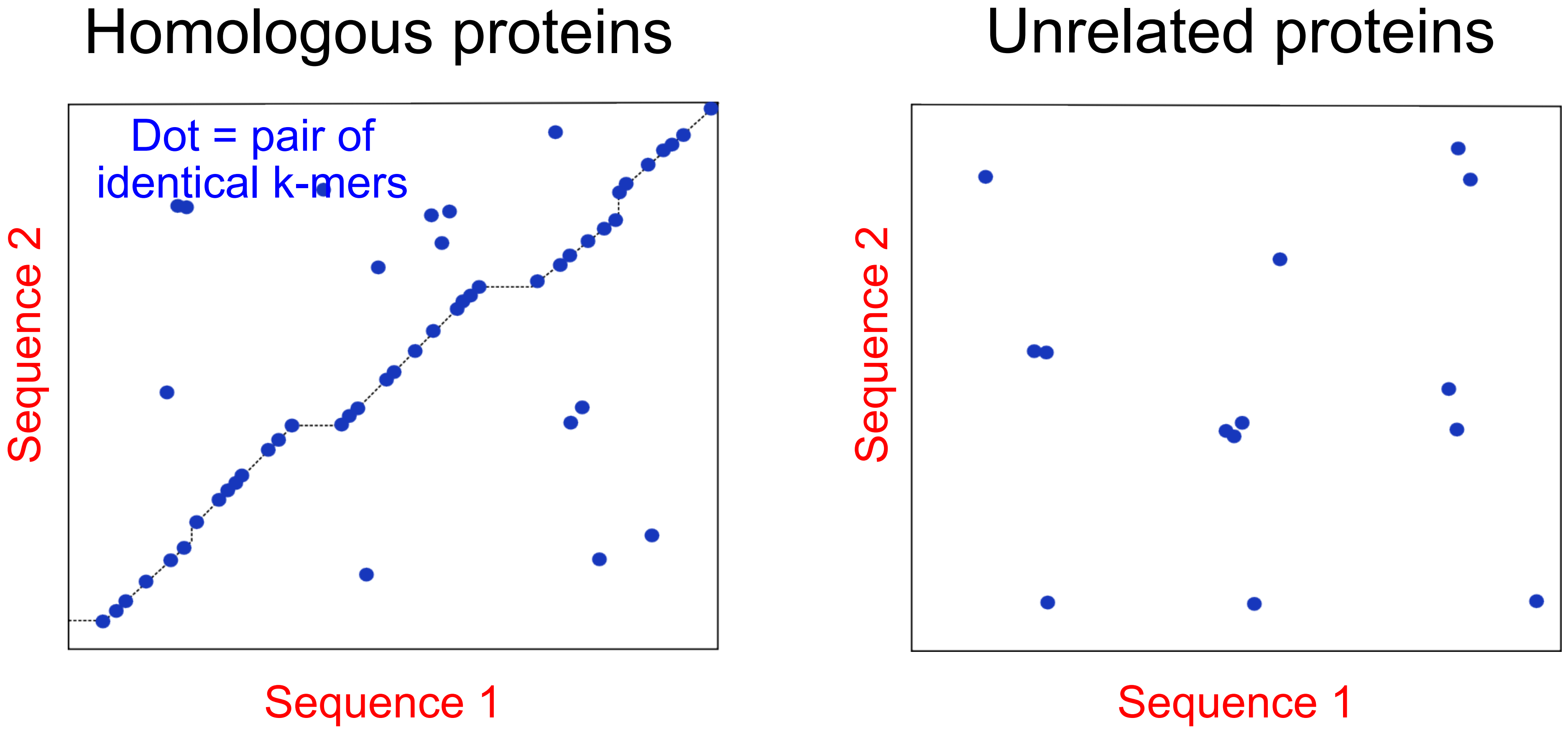
VRLSLCW
IRMTVCF
PLCYAGD
PVCYSGN

- Two *k*-mer matches without gap in-between
- Sequence *profiles*!
- No random memory access in innermost loop



Conventional alignment-free comparison: count identical k-mers

Sequence 1 . . . VRLS . . . PLCW . . . YAGD . . .
Sequence 2 . . . VRLS . . . PLCW . . . YAGD . . .



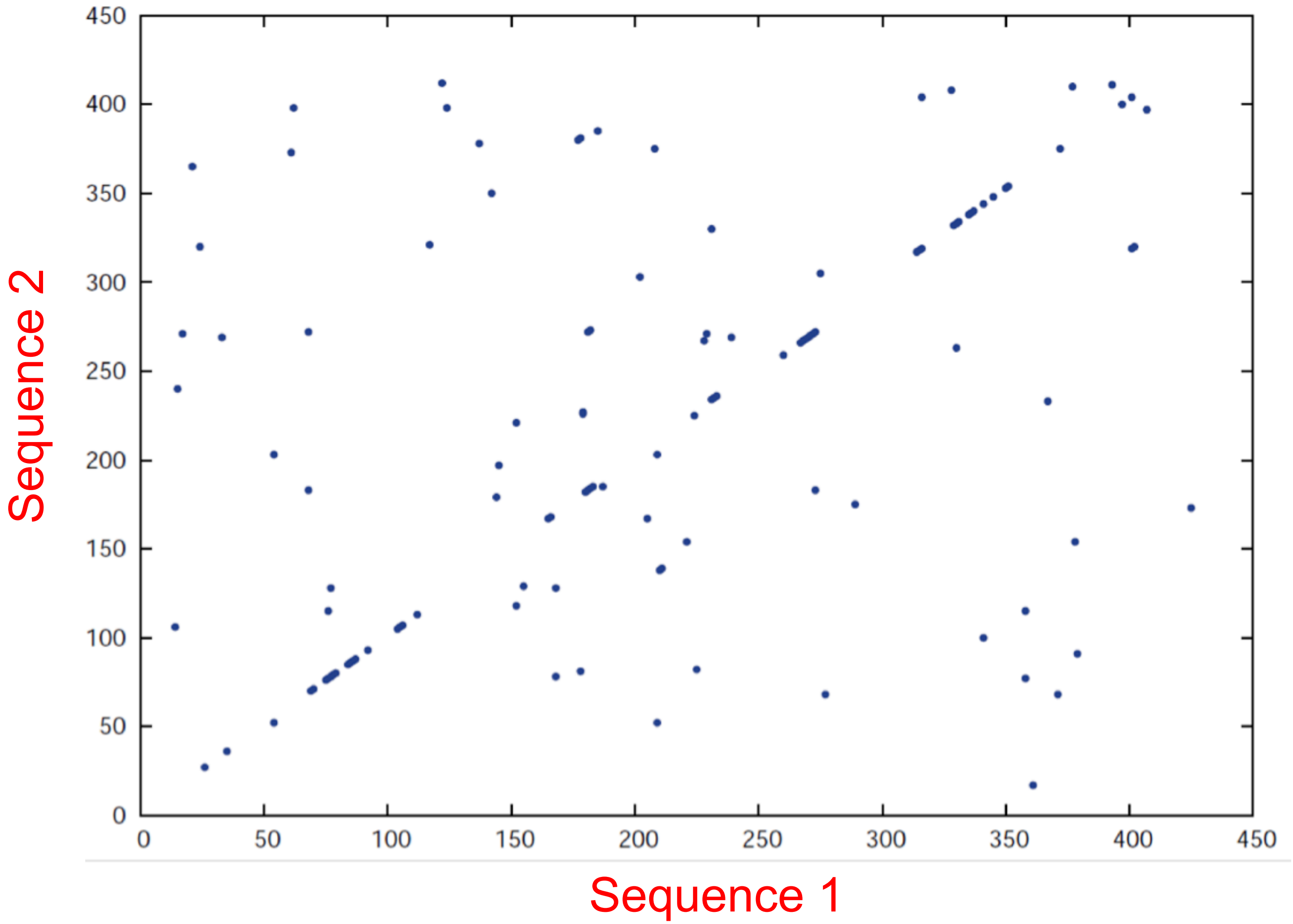
Homologous proteins

Unrelated proteins

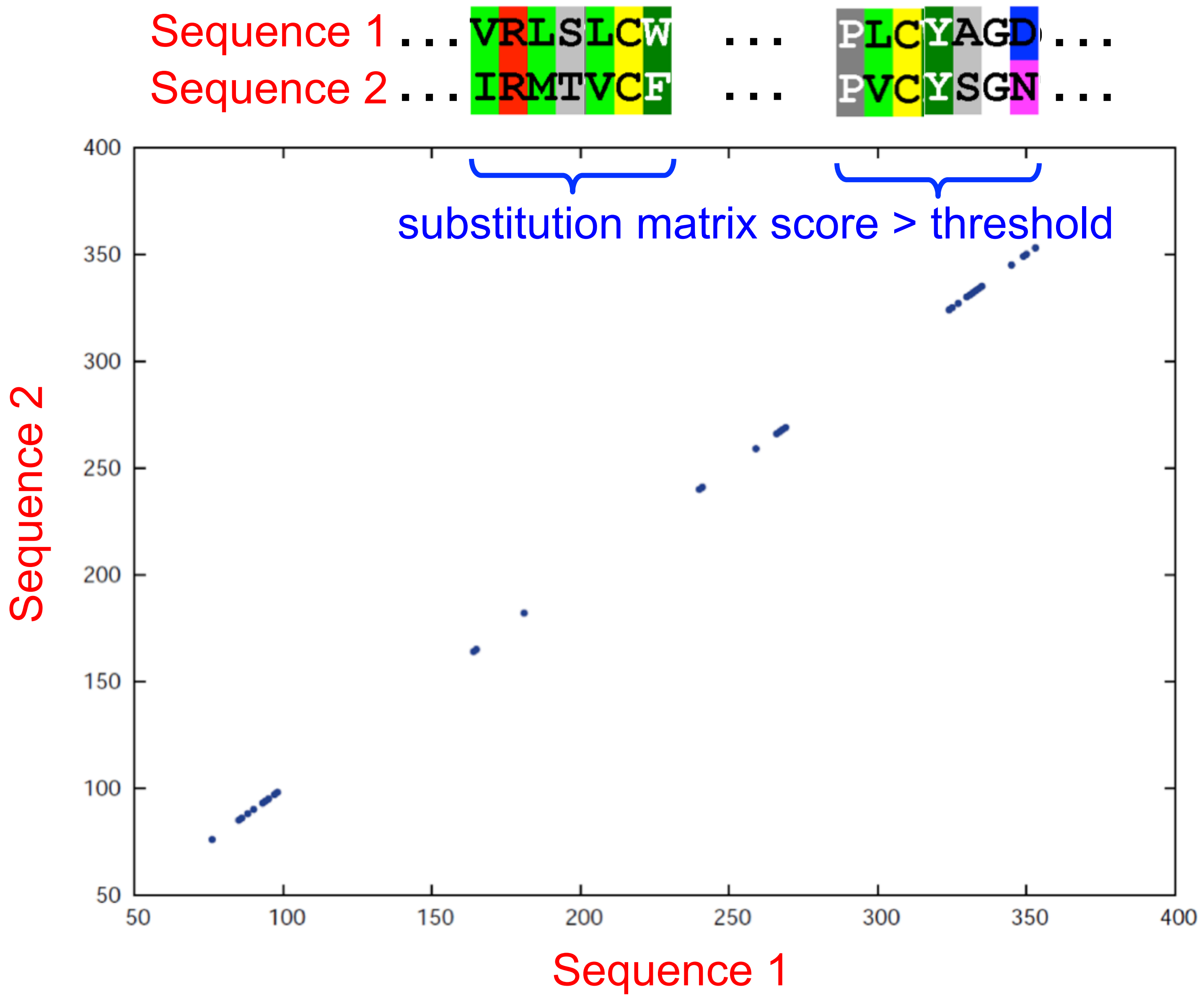
Most 3-mer matches occur by chance

Sequence 1 ...RLS ... PLC ... YAG ...

Sequence 2 ...RLS ... PLC ... YAG ...



MMseqs: sum scores of similar 7-mers



Sequence 1 ...VRLSLCW ... PLCYAGD...

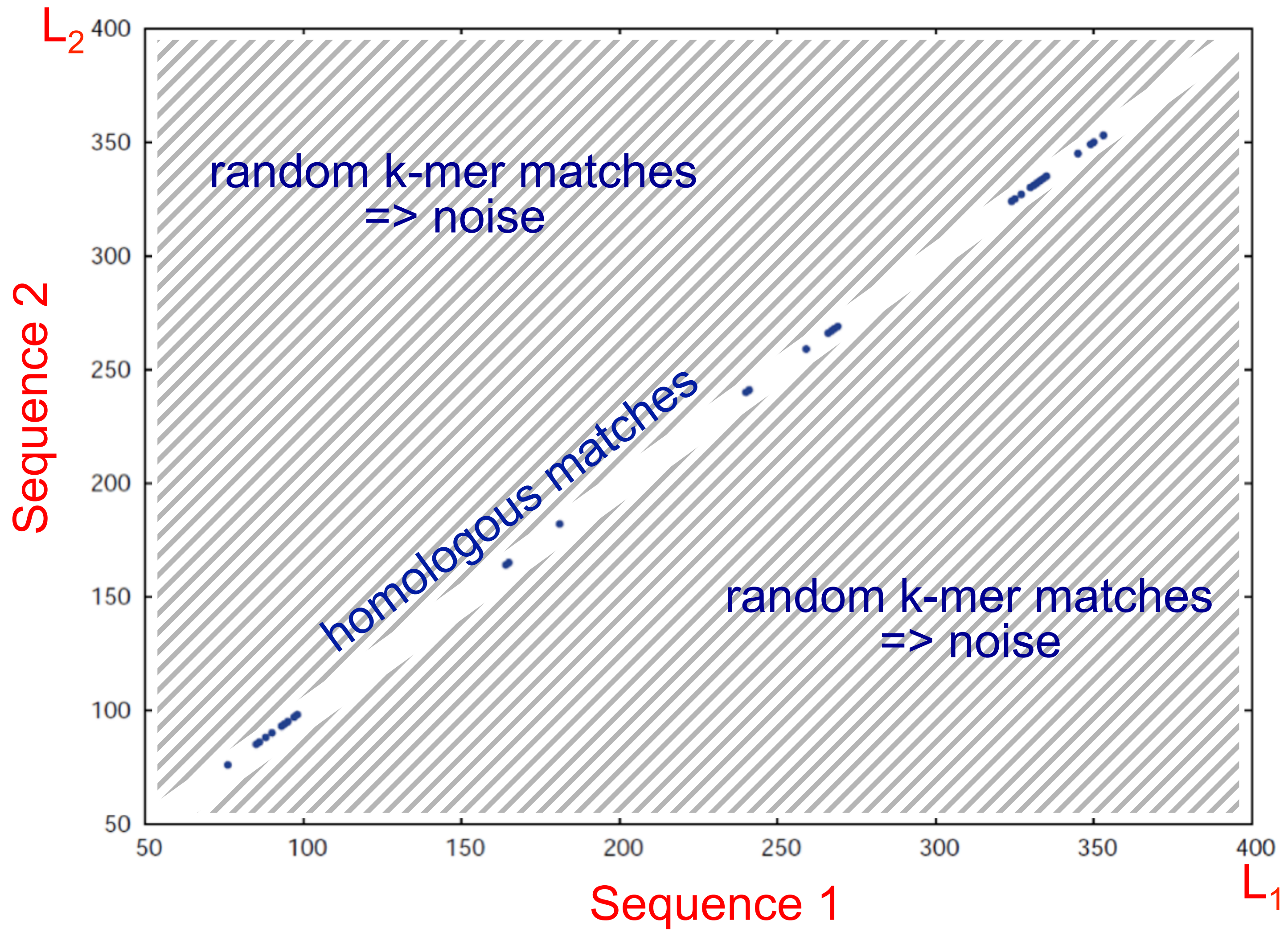
Sequence 2 ...IRMTVCF ... PVCYSGN...

substitution matrix score > threshold

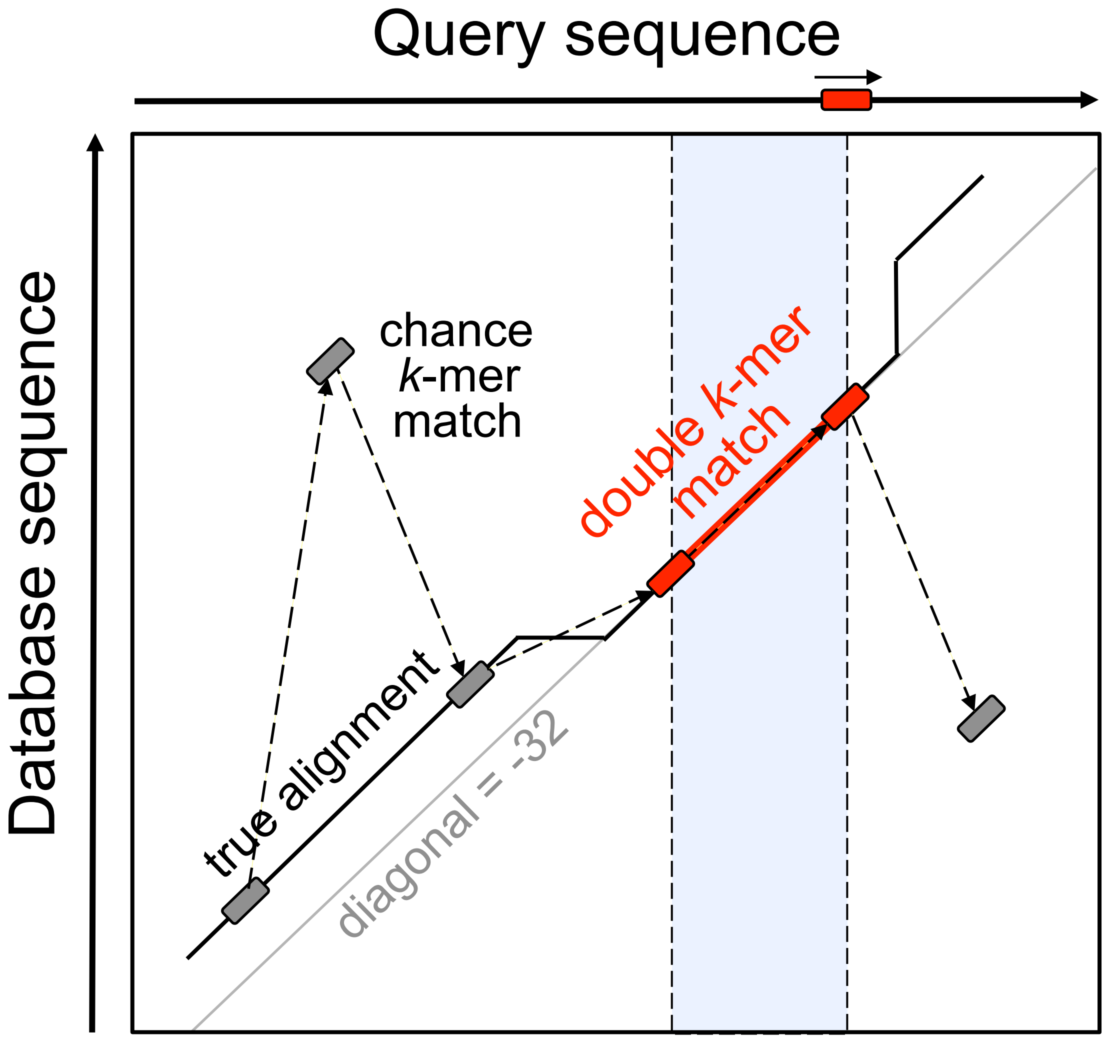
But: random matches scale as L_1L_2 and signal only as $\min\{L_1, L_2\}$

Sequence 1 . . . VRLSLCW . . . PLCYAGD . . .

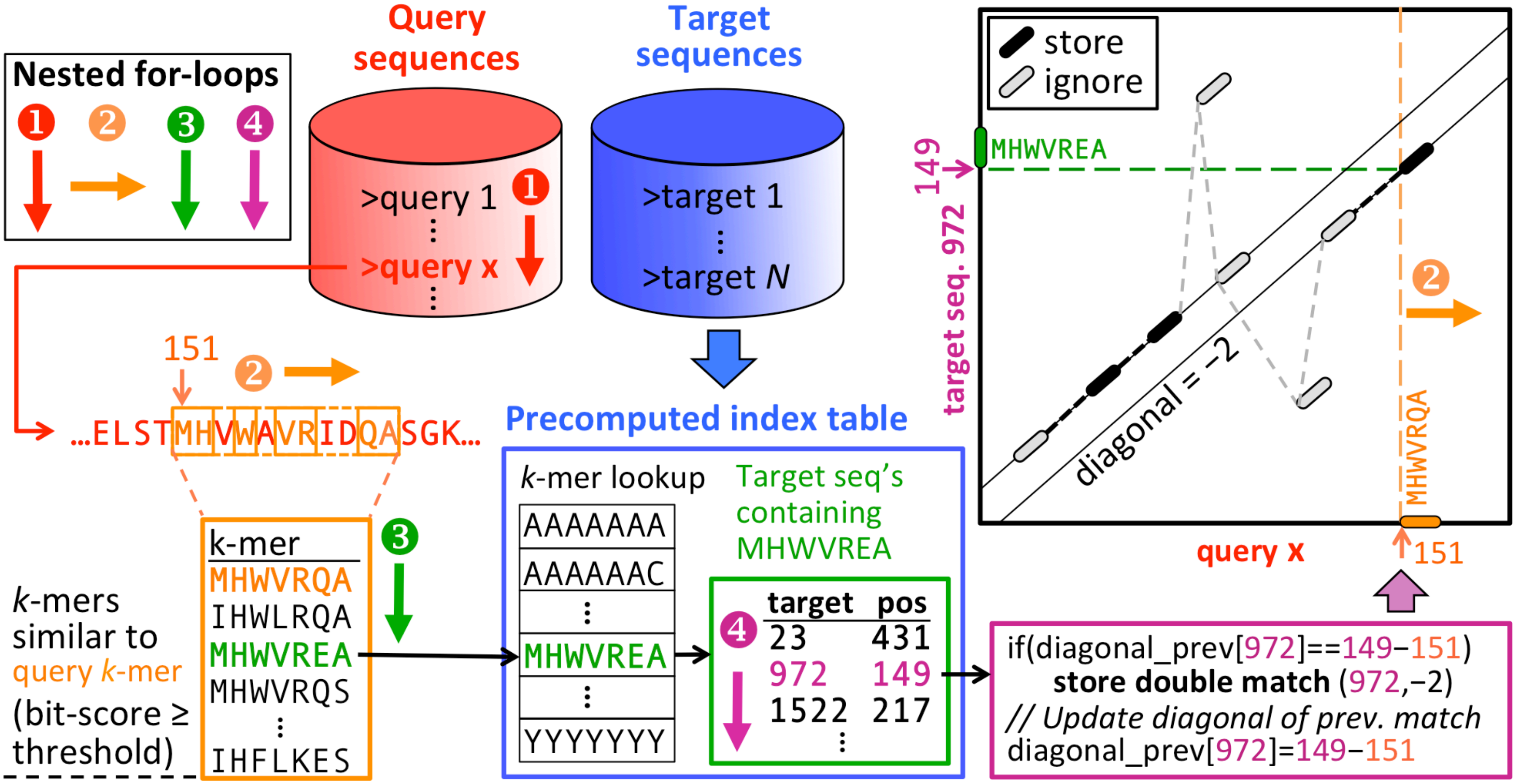
Sequence 2 . . . IRMTVCF . . . PVCYSGN . . .



Find db sequences with 2 consecutive *k*-mer matches on same diagonal



Find 2 consecutive *k*-mer matches on same diagonal



k-mers similar to query *k*-mer (bit-score ≥ threshold)

k-mer
MHWVRQA
IHWLRQA
MHWVREA
MHWVRQS
⋮
IHFLKES

Precomputed index table

k-mer lookup
AAAAAAA
AAAAAAC
⋮
MHWVREA
⋮
YYYYYYY

Target seq's containing MHWVREA

target	pos
23	431
972	149
1522	217
⋮	

```
if(diagonal_prev[972]==149−151)
    store double match (972,−2)
// Update diagonal of prev. match
diagonal_prev[972]=149−151
```

CPU cache optimized!

Executive summary

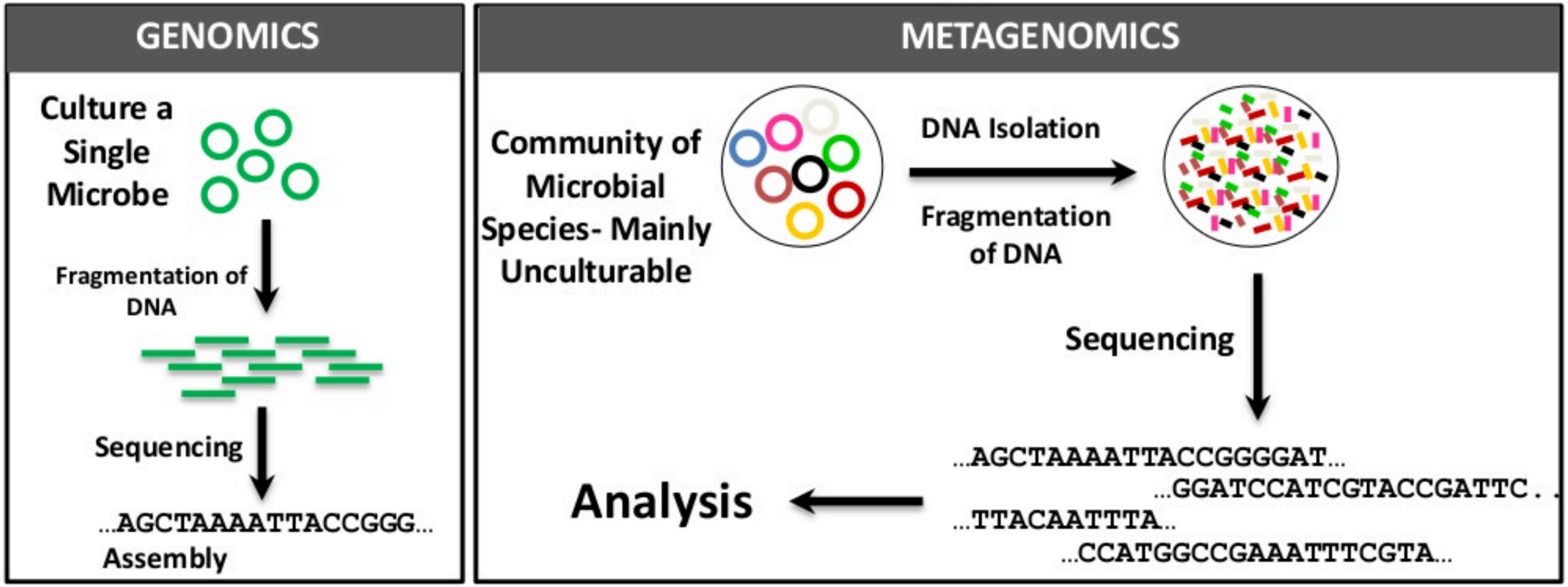
1990

PSI-BLAST

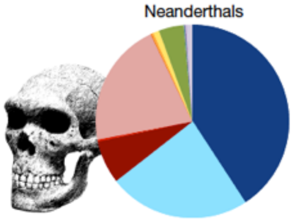
2019

MMseqs2

With metagenomics we can study the ~99% uncultivable microbes by sequencing their DNA directly from environment



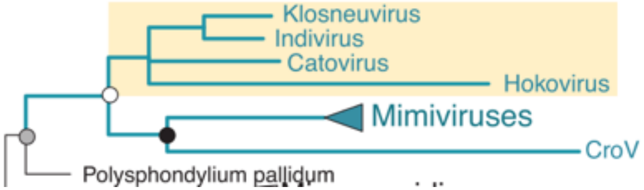
Metagenomics age of enlightenment



Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus

Nature 2017, Apr 20

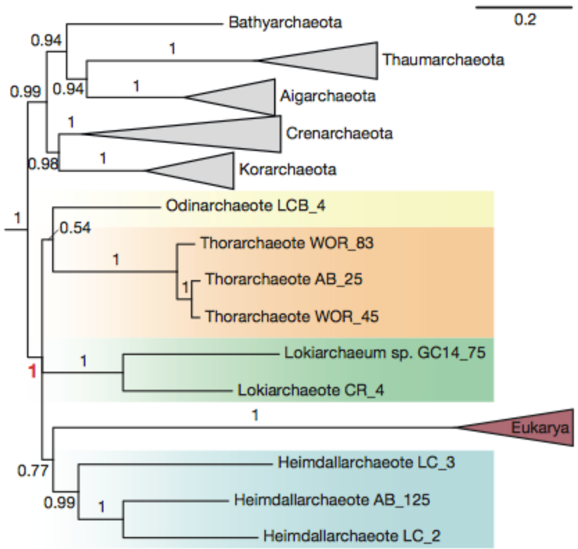
Giant viruses with an expanded complement of translation system components



Science 2017, Apr 7

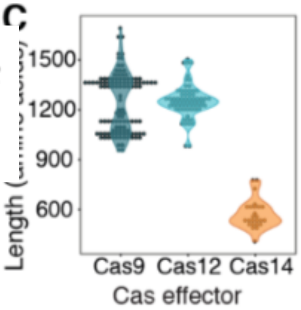
Asgard archaea illuminate the origin of eukaryotic cellular complexity

Nature 2017, Jan 19



Programmed DNA destruction by miniature CRISPR-Cas14 enzymes

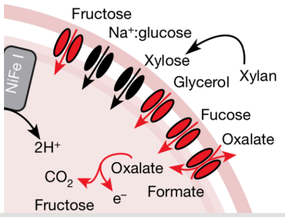
Science 2018, Nov 16



Metagenomics age of enlightenment

Genome-centric view of carbon processing in thawing permafrost

Nature 2018, Aug 02



Structure a
microbiome

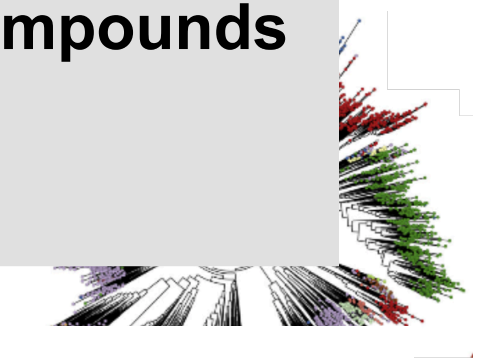
Nature 2018, Aug

Extensive
Revealed b
Metagenor
Lifestyle

Cell 2019, Ja

Applications:

- **Human health (gut, skin, ...)**
- **Ecology & climate**
- **Enzymes for biotechnology**
- **New drugs and natural compounds**
- **Evolution, tree of life**
- **...**

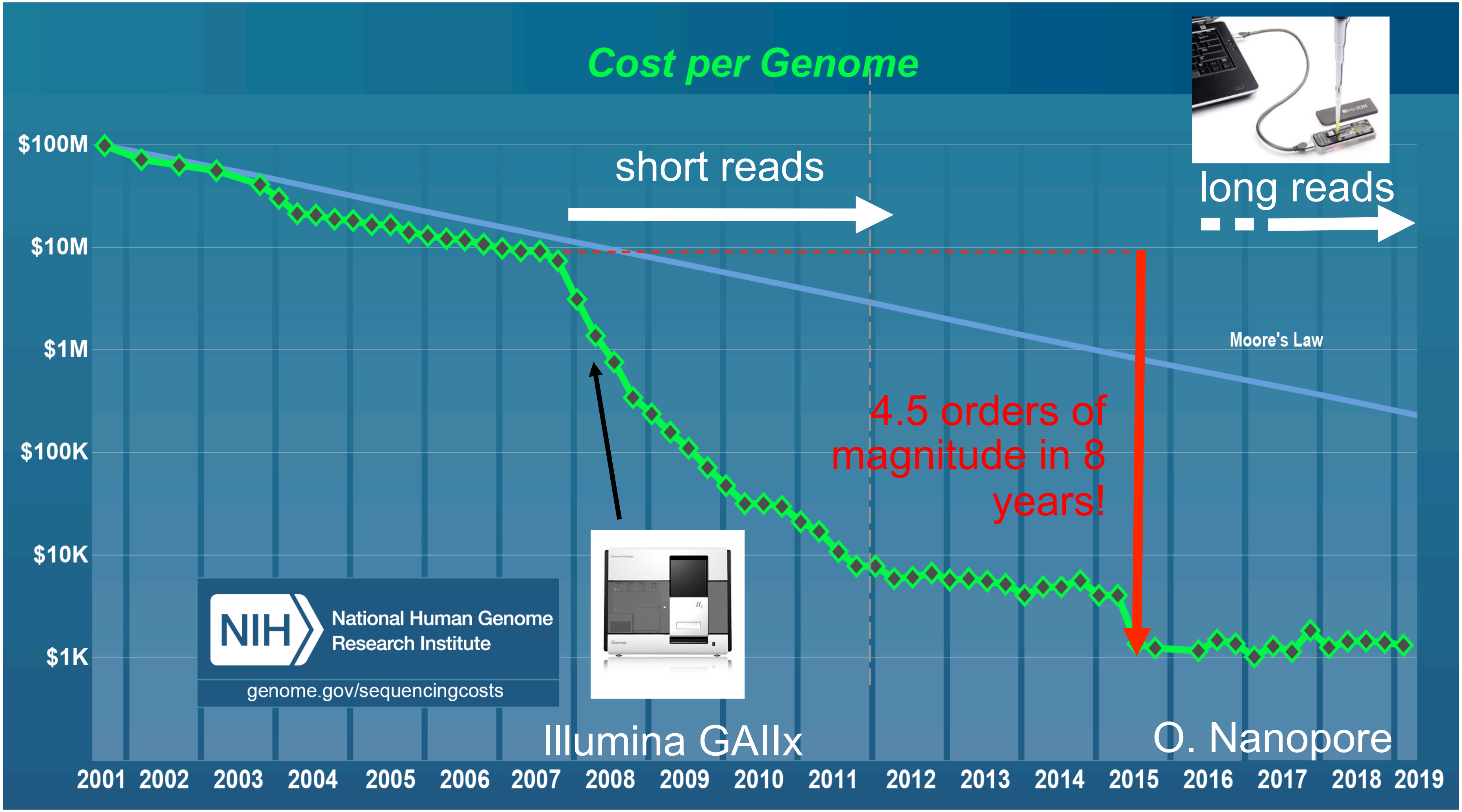


nature

Article | Published: 23 October 2019

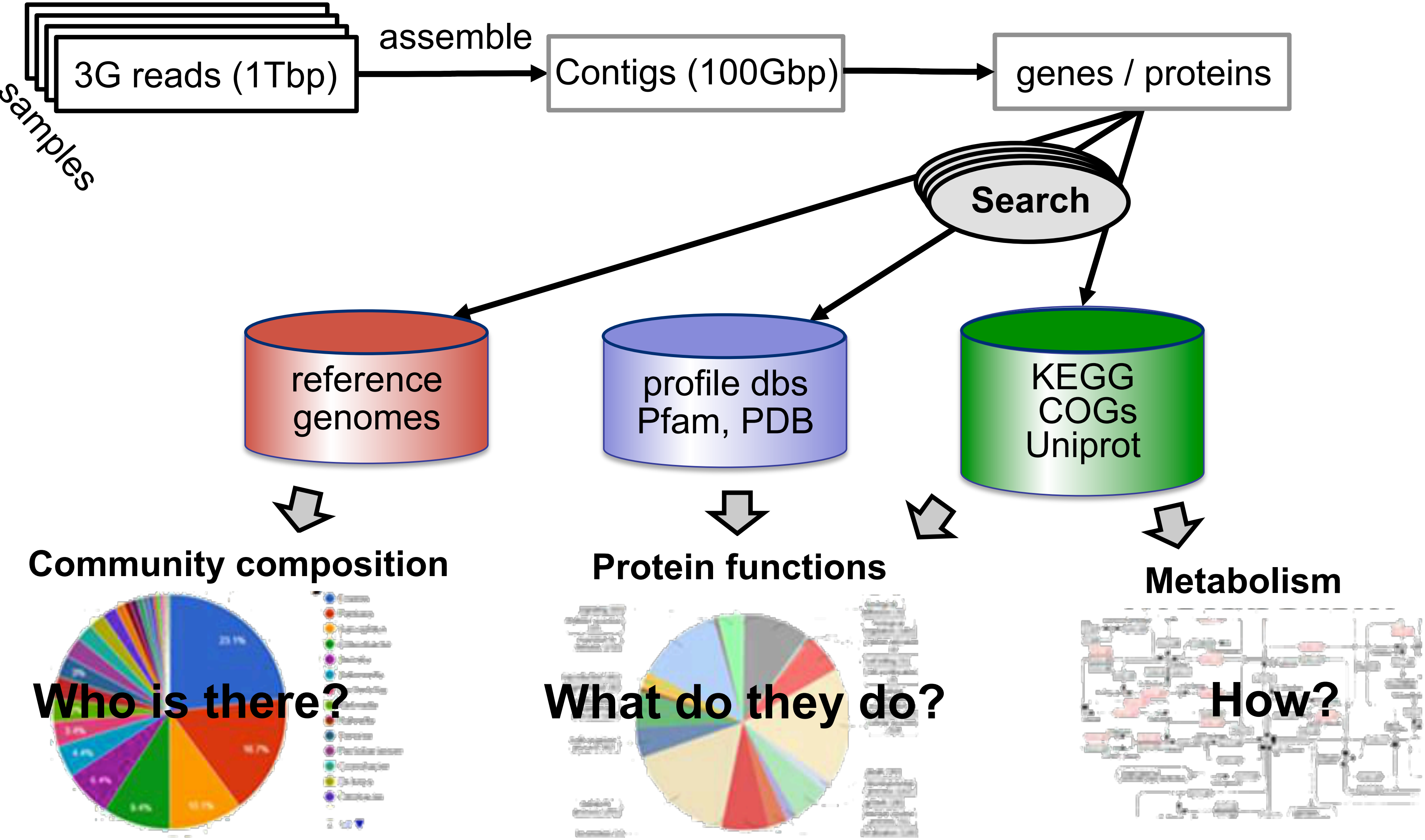
The microbiota regulate neuronal function and fear extinction learning

Metagenomics is driven by fast-decreasing sequencing costs



- **Costs for computing by far exceed sequencing costs**
- **Bottleneck: sequence searches**

Shotgun metagenomics data analysis



samples

3G reads (1Tbp)

assemble

Contigs (100Gbp)

genes / proteins

Search

reference genomes

profile dbs Pfam, PDB

KEGG COGs Uniprot

Community composition

Who is there?

Protein functions

What do they do?

Metabolism

How?