



MAX-PLANCK-INSTITUTE
FOR BIOPHYSICAL CHEMISTRY

Introduction to protein bioinformatics

IMPRS Molecular Biology

University of Göttingen

08 November 2021

Ruoshi Zhang, Venket Raghavan, Michel van
Kempen, Sasha Kolodyazhnaya, Johannes Söding

Quantitative and Computational Biology

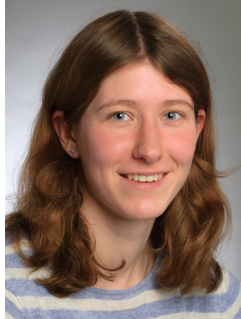
MPI for Biophysical Chemistry

Söding lab in November 2021

Tools for metagenomics, protein structure & function



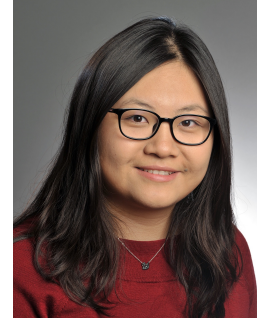
Milot
Mirdita



Annika
Seidel



Venket
Raghavan



Ruoshi
Zhang



Sasha
Kolodyazhnaya



Hong Su

Transcription / quant. medicine



Etienne
Morice



Amelie Hilger



Michel van
Kempen



Yazhini A

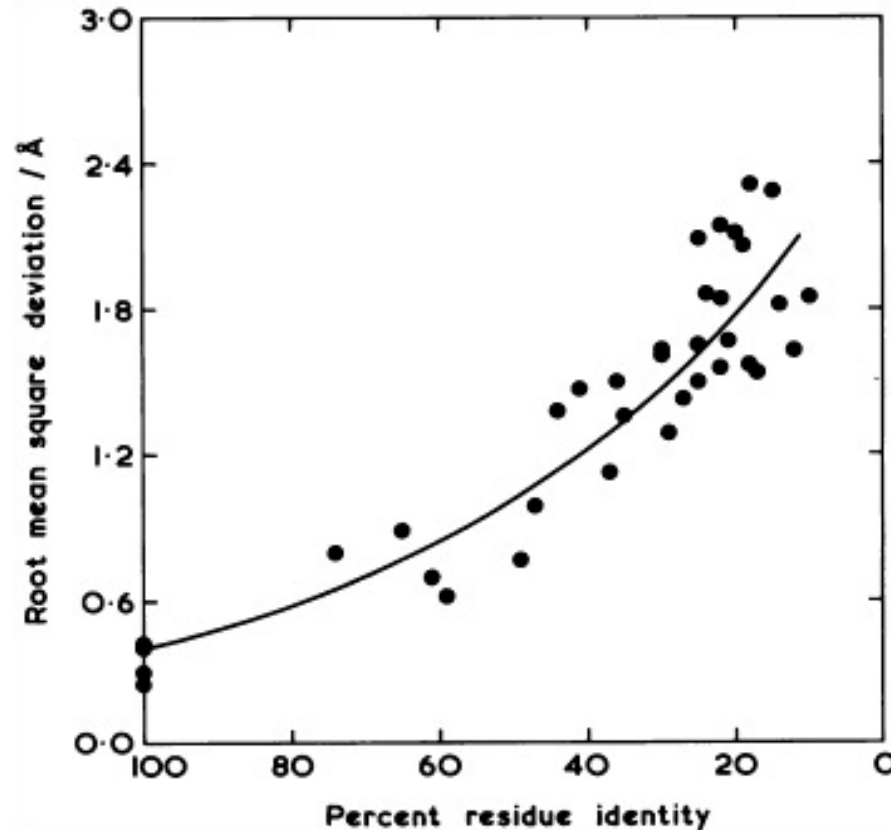


Louis Kraft

Goals for next 1 ½ days

- Protein structure and sequence conservation
- Homology-based inference and sequence similarity searches
- P- and E-value
- Sequence alignment (dynamic programming)
→ Role of algorithms in bioinformatics
- Sequence profiles: information is power!
- MMseqs, basic analyses of metagenomics dataset
- (Genome assembly)
- Structure databases
- AlphaFold

Protein structure is highly conserved even without obvious sequences similarity



[Chothia & Lesk 1986]

Sequence identity

RMSD in conserved core

Fraction in core

60%

0.85 Å

95%

40%

1.2 Å

80%

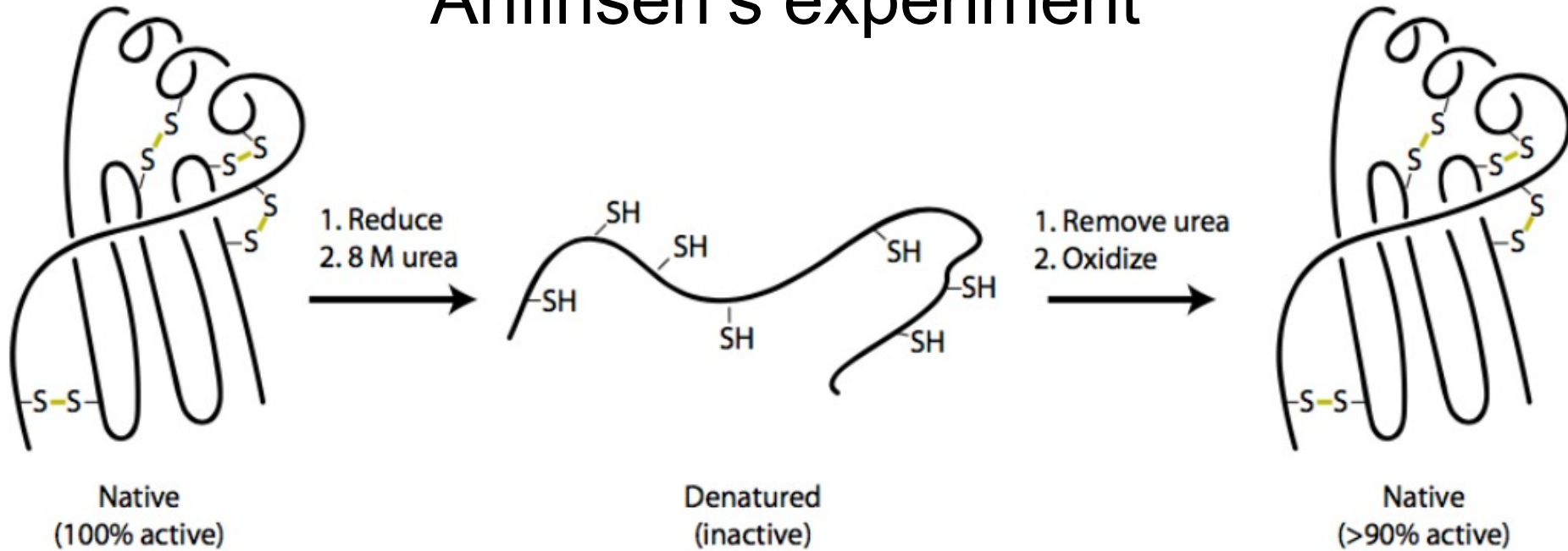
20%

1.8 Å

55%

Protein sequence determines structure!

Anfinsen's experiment



Anfinsen CB. "Principles that govern the folding of protein chains". Science 1973

If all the information to correctly fold a protein is contained in its amino acid sequence, we should be able to predict its structure from its sequence!

Computational chemistry: uncover the rules of protein folding from first physical principles

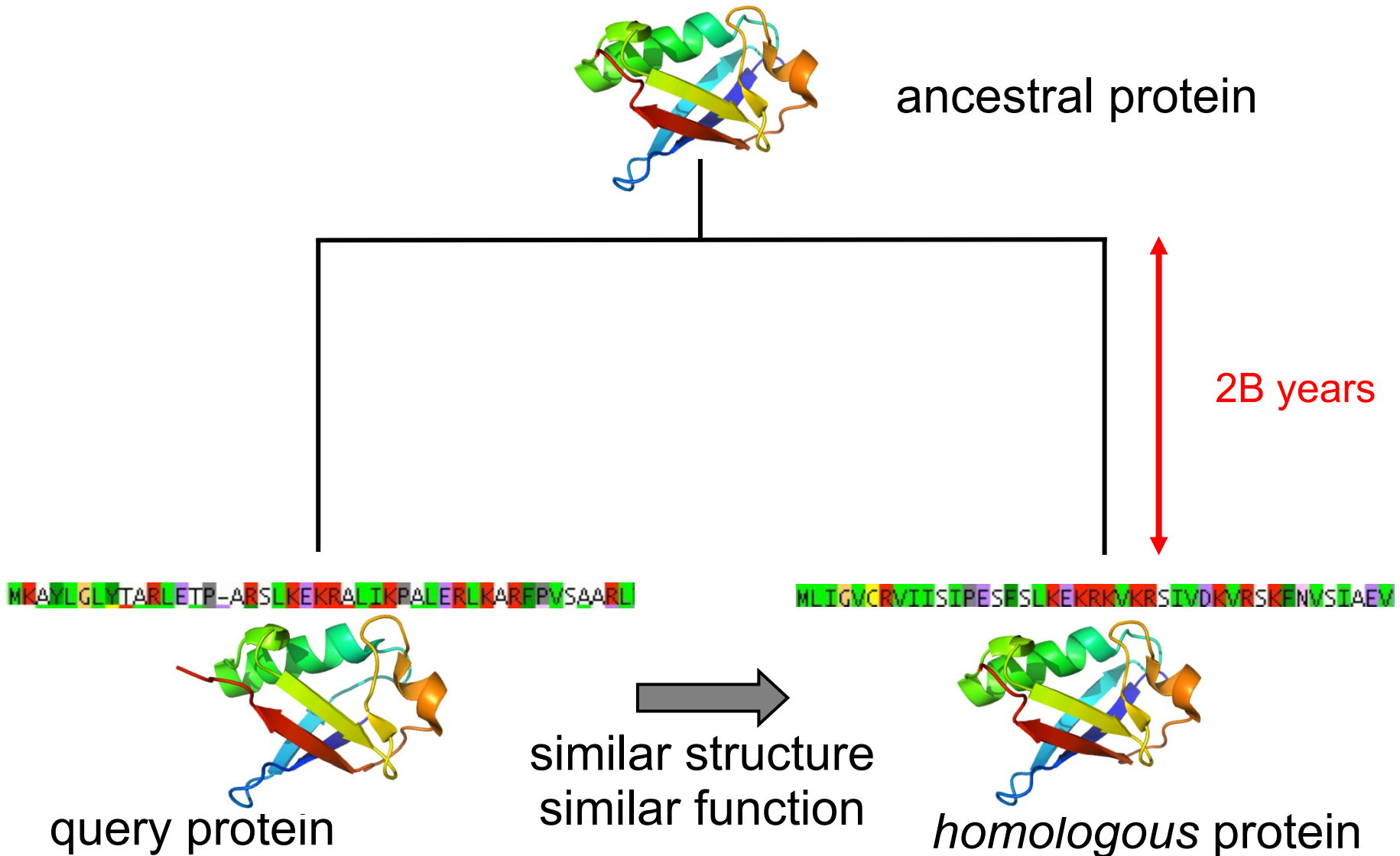
Do you know "exceptions" to Anfinsen? (3) Allostery; misfolded proteins (Alzheimer's, prions); chaperones (GroEL, Hsp70, Hsp90, ...)

From comparative protein structure modeling to deep learning and AlphaFold

Comparative modeling has been the mainstay of protein structure prediction up to now. It relied 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 now superseded by **deep neural networks** (transformers) such as **AlphaFold**, trained on all ~160k protein structures.

Homologous = descended from common ancestor

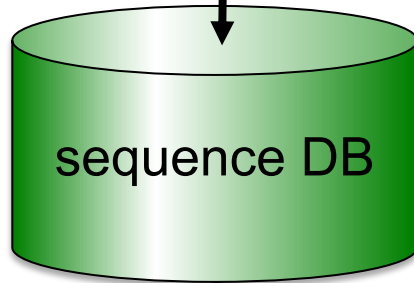


Homology-based inference of protein structure and function

query protein

MKAVLGLYIARLETP-ARSLKEKRALIKPALERLKARFPVSAARL

sequence search



sequence DB

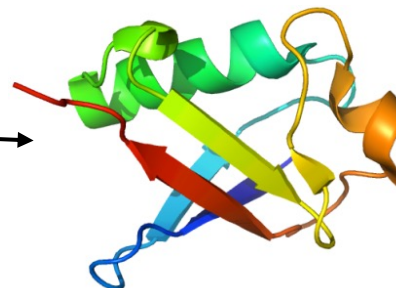
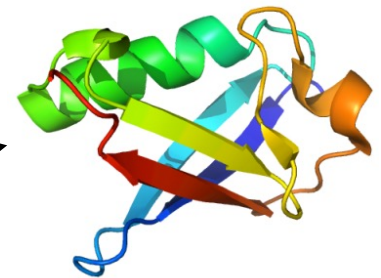
MKAVLGLYIARLETP-ARSLKEKRALIKPALERLKARFPVSAARL
--MLIGVCRVITISIPESHSLKEKRKWRSTVDKVRSKENVSTAEM

homologous
sequence found
with known structure
or functions

When are two sequences
similar enough to ascertain
homology?

→ E-value < 0.01

predict structure
and/or function of
query from those of
database match



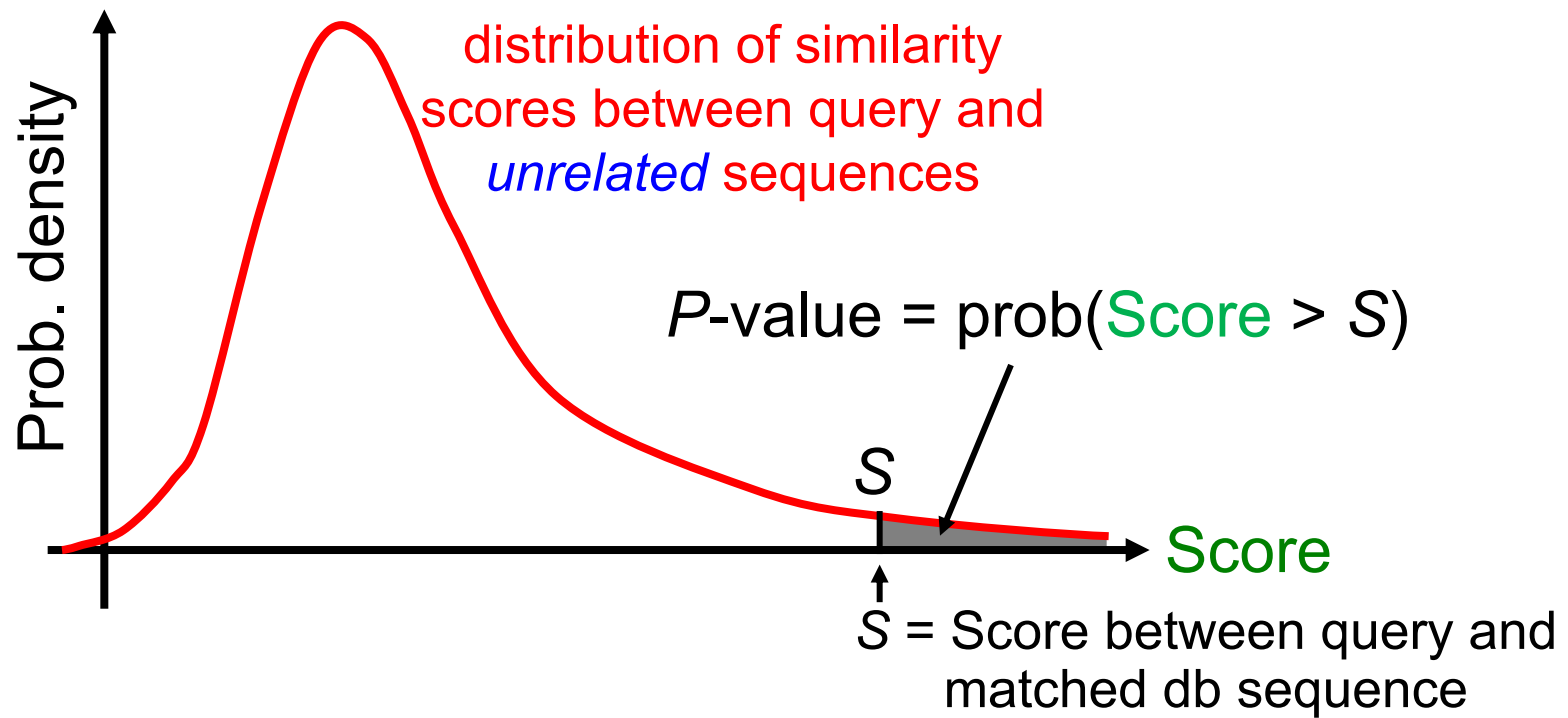
2B years

When are two sequences similar enough to ascertain homology?

Null hypothesis (boring “hypothesis of randomness”): query sequence is not in any way related to database sequence, similarity score is “random”.

Can we reject this null hypothesis (assume the db sequence is homologous)?

The sequence similarity score (our “test statistic”) has a **distribution** with only two parameters which we can compute. 😊



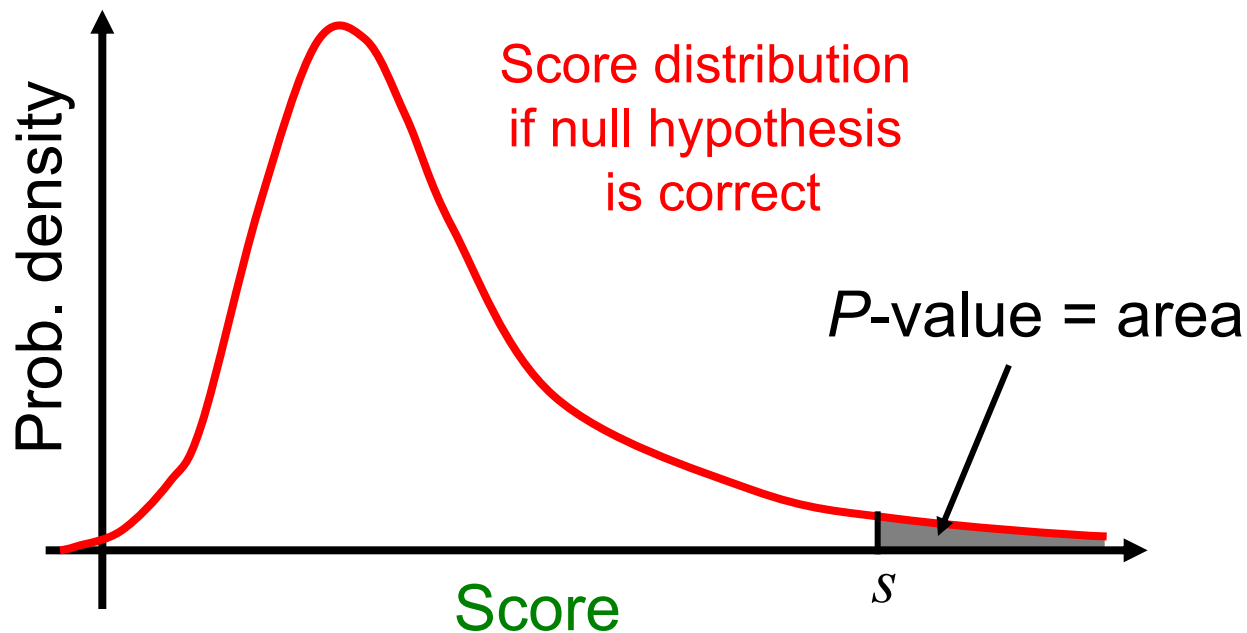
Small P-value: reject null hypothesis

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.



E-values

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

Suppose you searched a sequence database with a query sequence and you obtained a match with a P-value = $1\text{E-}6$. Can you trust this matched sequence to be homologous to your query?

Suppose your sequence database contains 10^8 sequences. Can you trust the matched sequence with a P-value = $1\text{E-}6$ to be homologous to your query?

No! Each db sequence has a probability of $1\text{E-}6$ to have a P-value $< 1\text{E-}6$ *by pure chance alone*. So the expected number of db sequences to achieve a P-value $< 1\text{E-}6$ is

$$E = 10^8 \times 1\text{E-}6 = 100 !$$

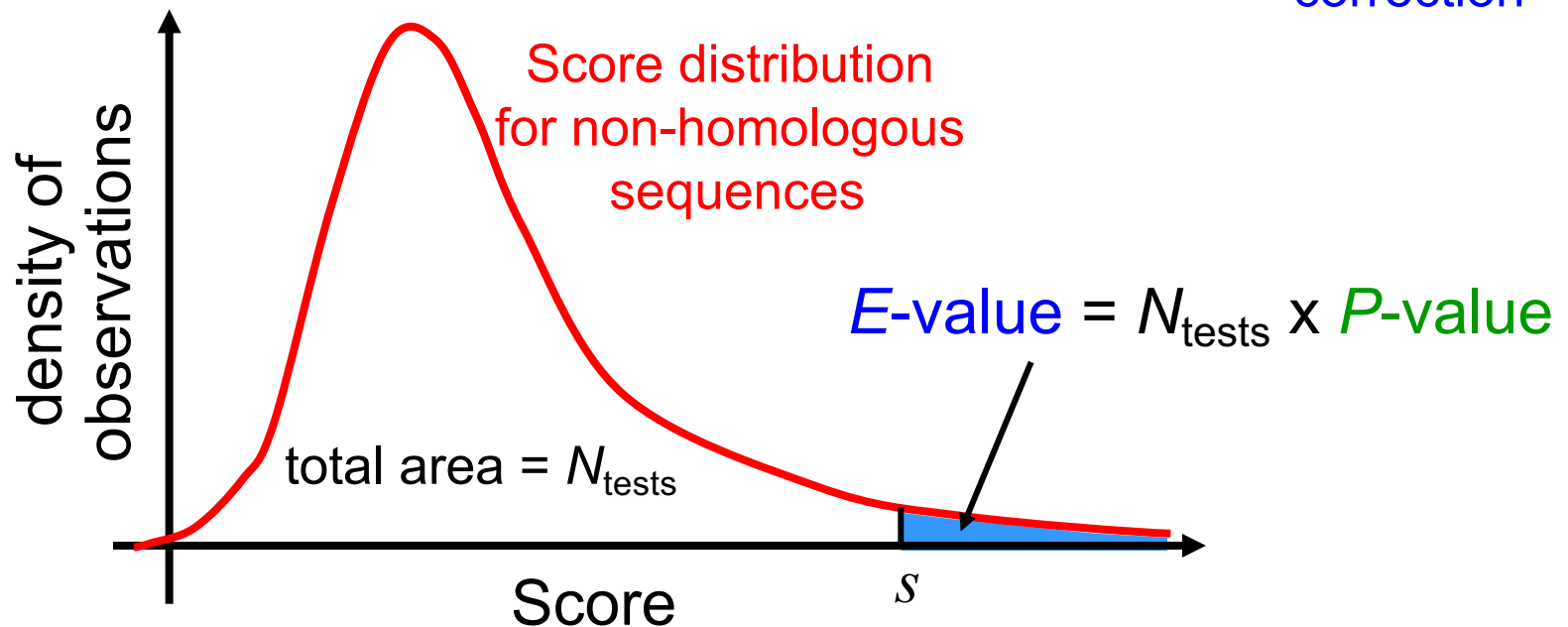
Therefore, the match is not at all trustworthy.

E-value = expected number of observations at least as extreme as the one observed

- ① **P-value** = Probability for event with score $\geq s$ under the null hypothesis
- ② **E-value** = Expected number of events out of N_{tests} trials with score $\geq S$ under the null hypothesis

$$E\text{-value} = N_{\text{tests}} \times P\text{-value}$$

similar to
Bonferroni
multiple testing
correction



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

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

Bruce A. Knutson and Steven Hahn*

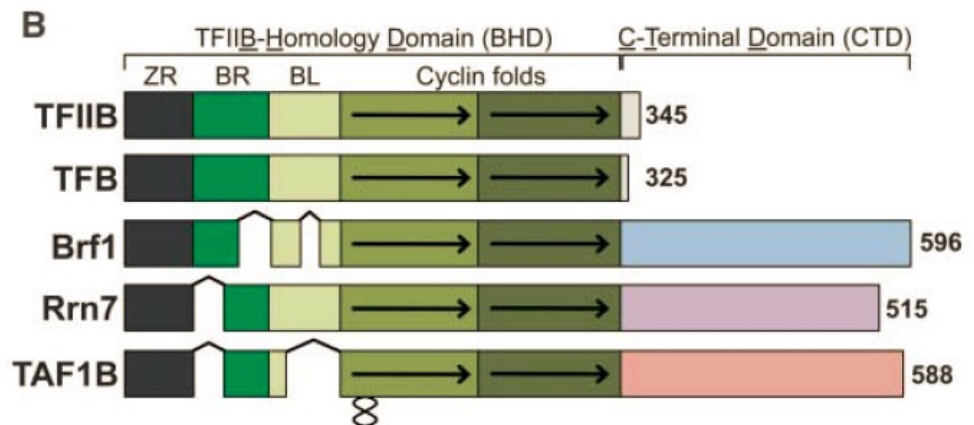
SCIENCE VOL 333 16 SEPTEMBER 2011

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/TAF₁₆₃) 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

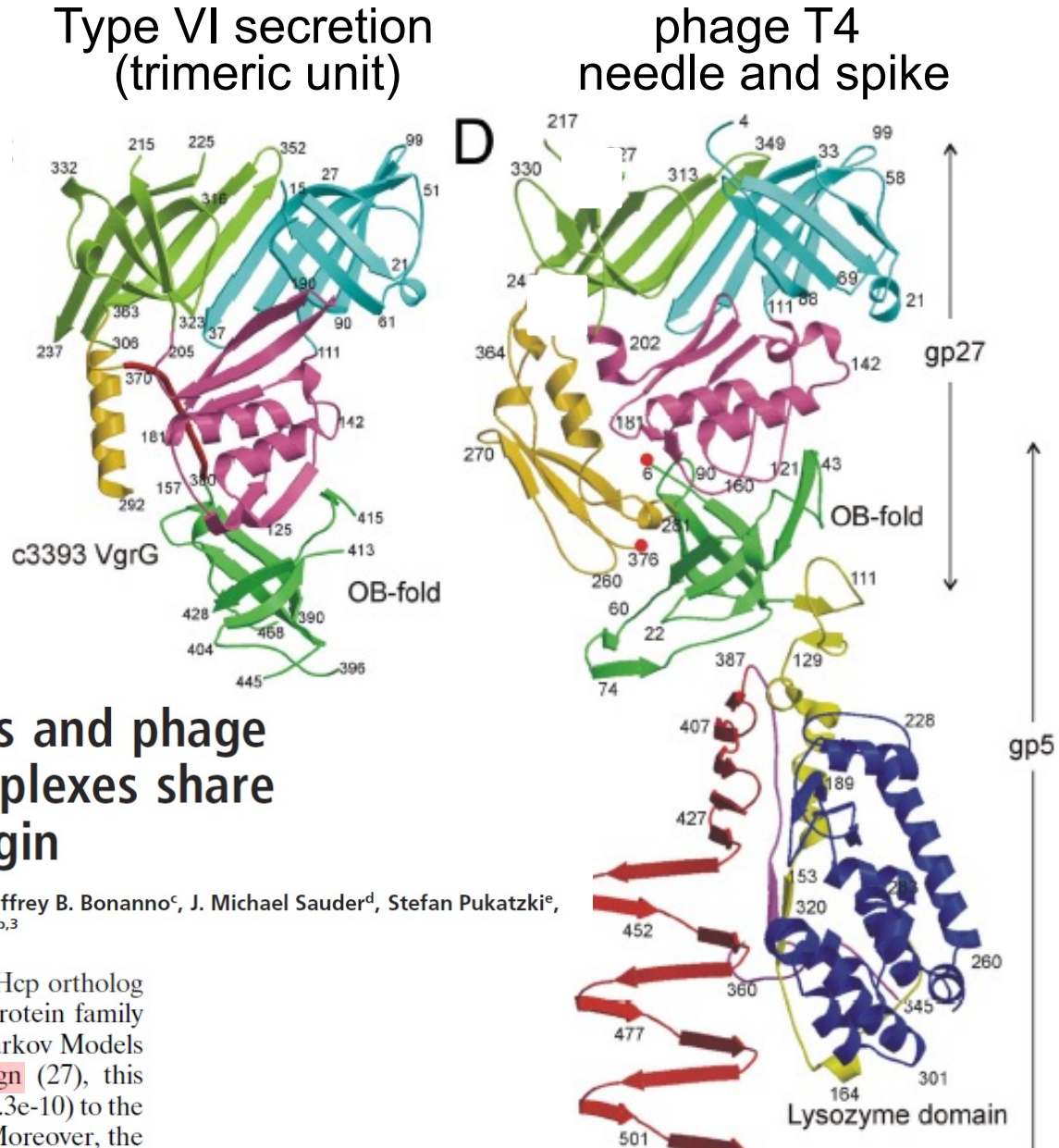
factors (13) because Pol I subunits share relatively low protein sequence conservation with their Pol II and Pol III counterparts (14). Using the homology detection program **HHpred**, which uses pairwise hidden Markov model profile comparisons that are more sensitive than traditional Web-based approaches (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.abysssi* 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 apparatus and phage tail-associated protein complexes share a common evolutionary origin

Petr G. Leiman^{a,1,2}, Marek Basler^{b,1}, Udipi 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 HAlign (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

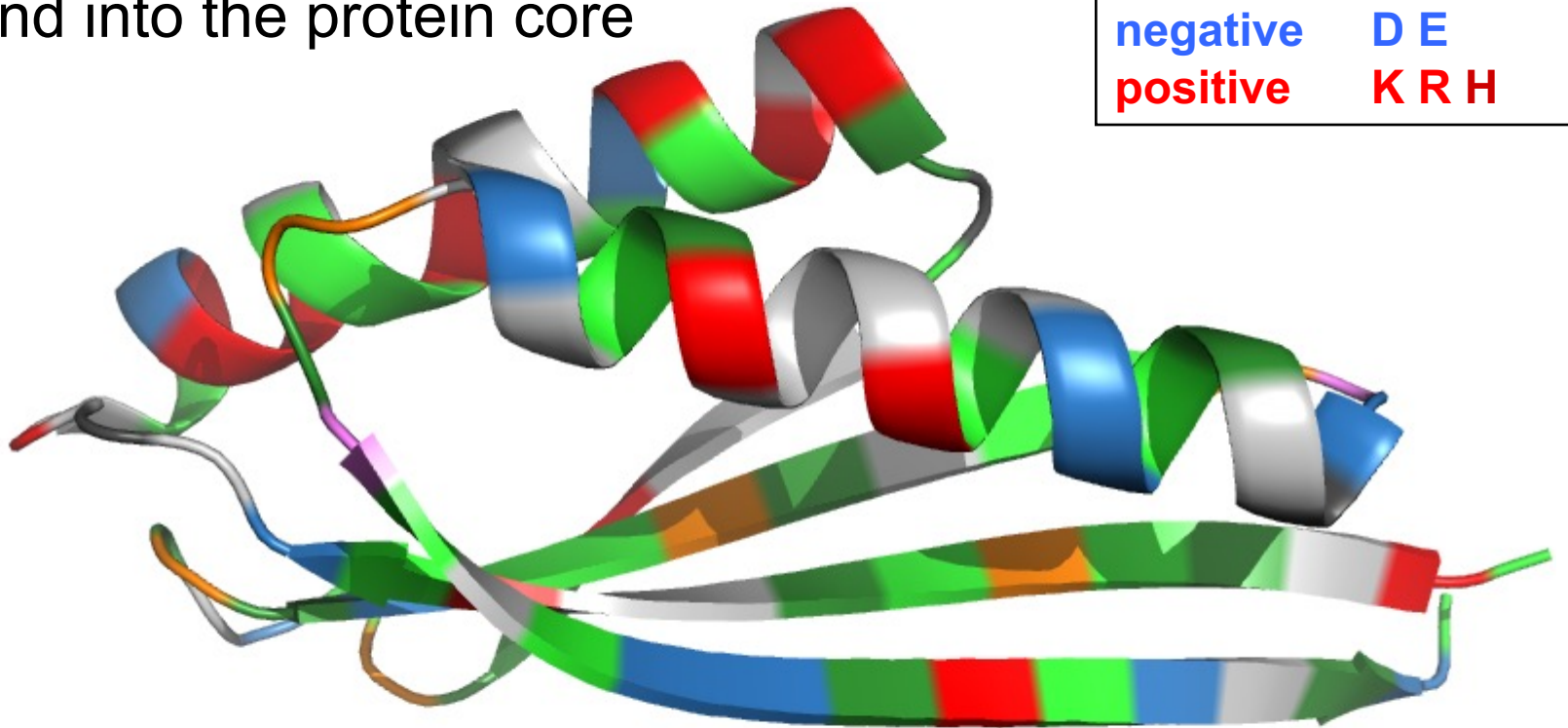
**How can we infer common descent
over time spans of billions of years?**

Hydrophobic residues form the domain cores

Example: protein with a ferredoxin fold

Most hydrophobic side chains extend into the protein core

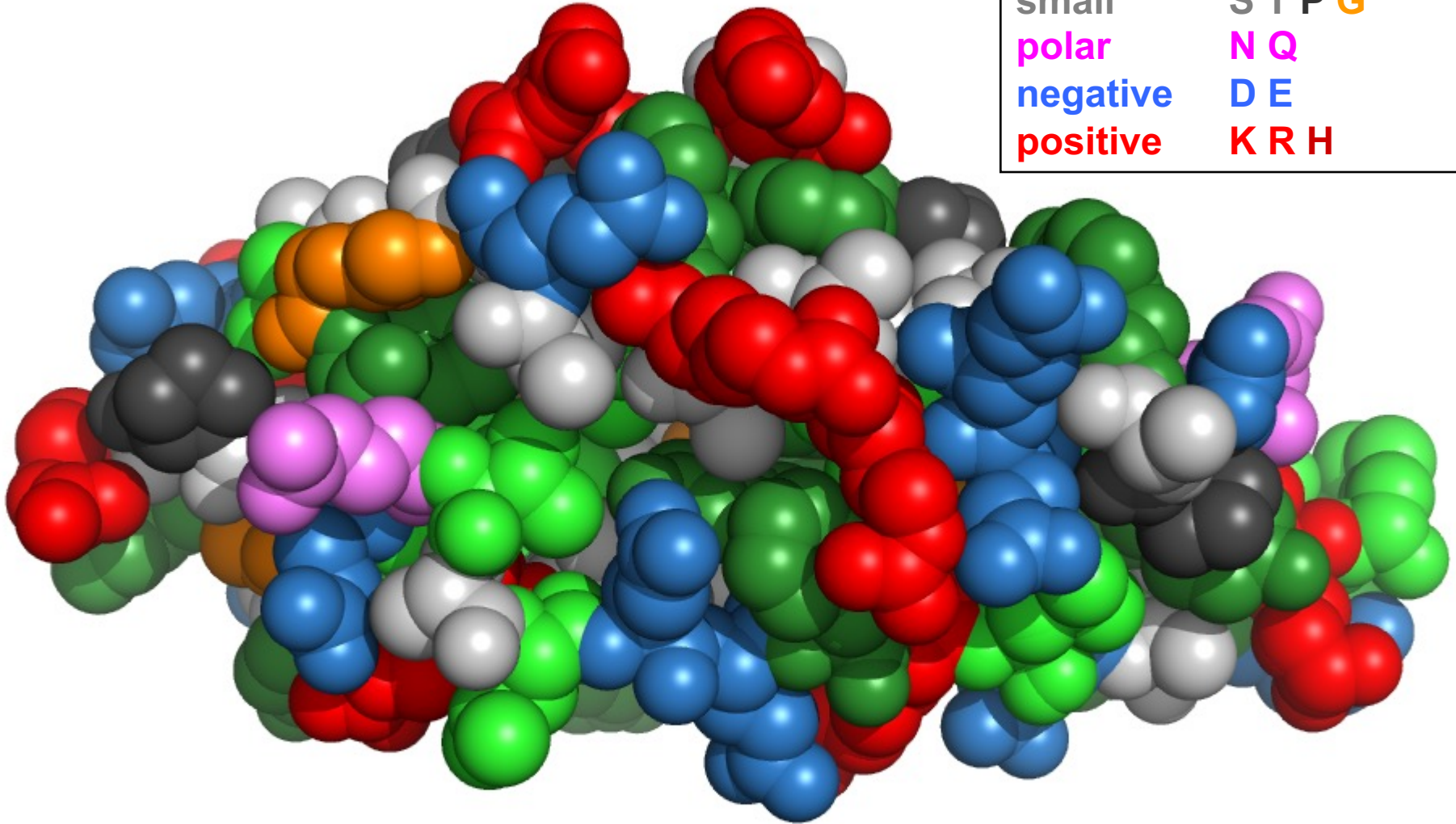
aliphatic	V L I M A C
aromatic	F W Y
small	S T P G
polar	N Q
negative	D E
positive	K R H



Hydrophobic residues form the domain cores

The protein core is tightly packed...

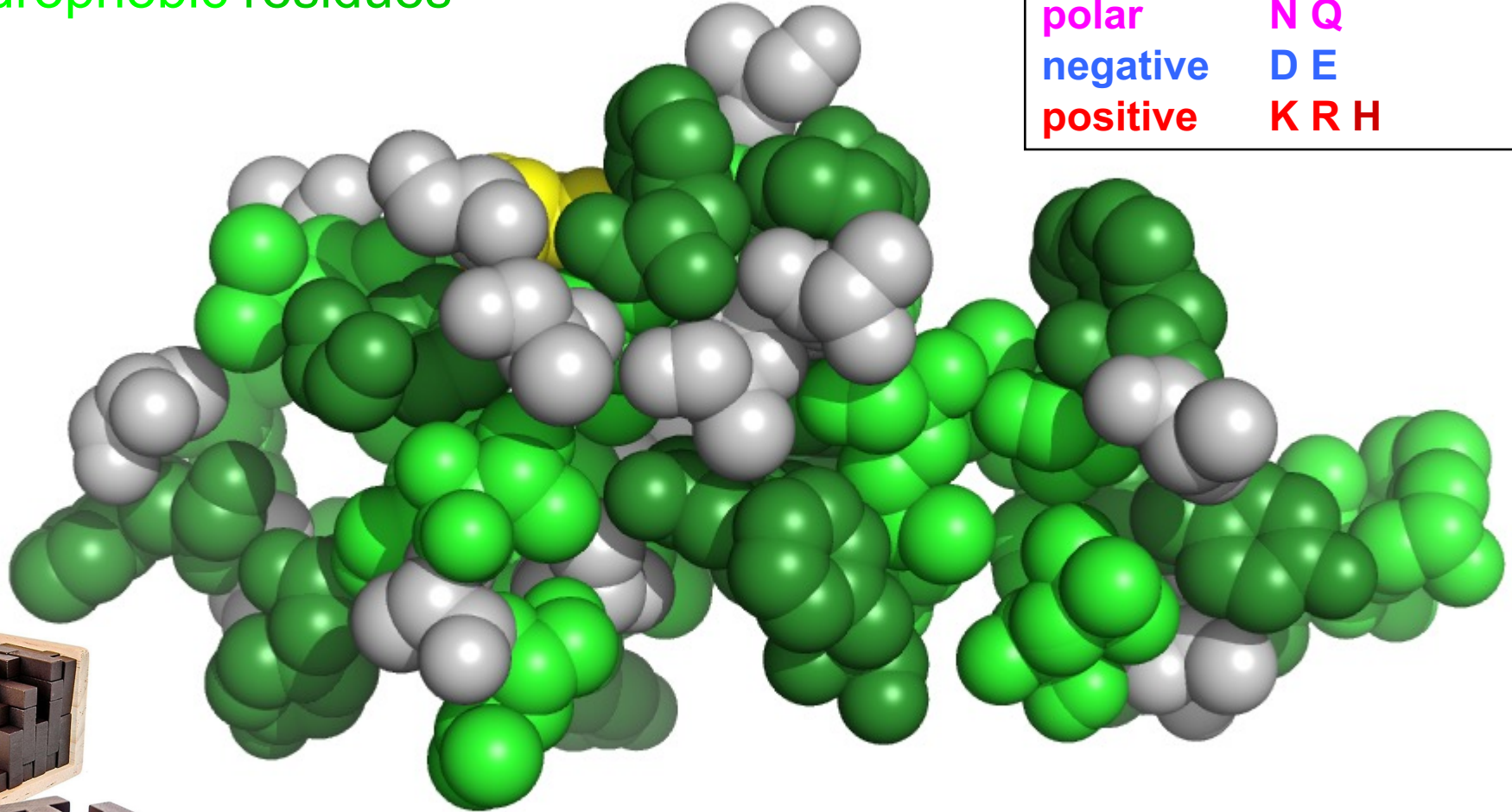
aliphatic	V L I M A C
aromatic	F W Y
small	S T P G
polar	N Q
negative	D E
positive	K R H



Hydrophobic residues form the domain cores

The protein core is tightly packed **with mainly hydrophobic residues**

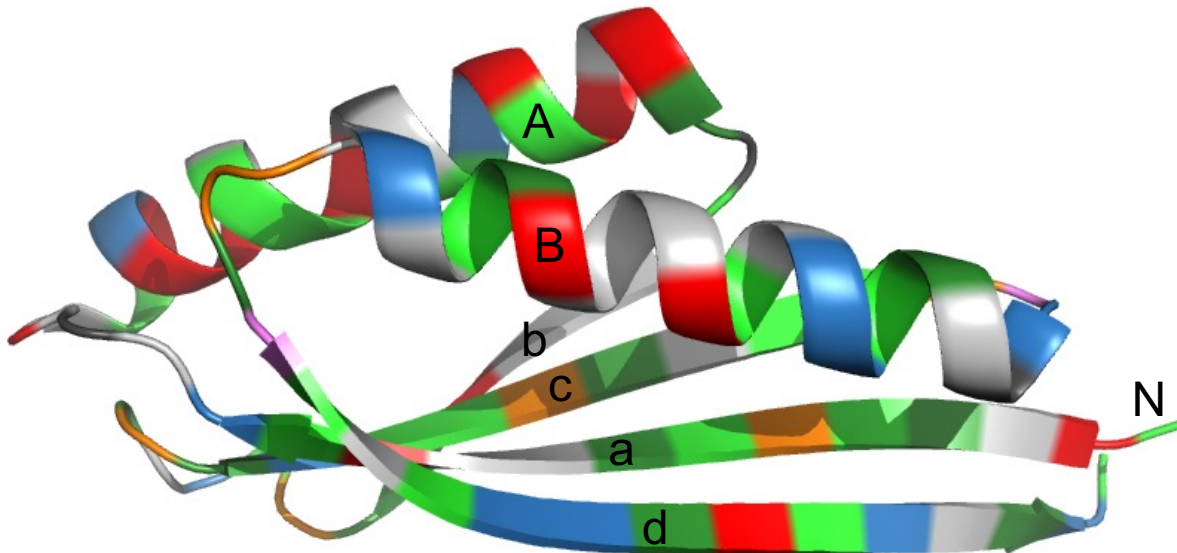
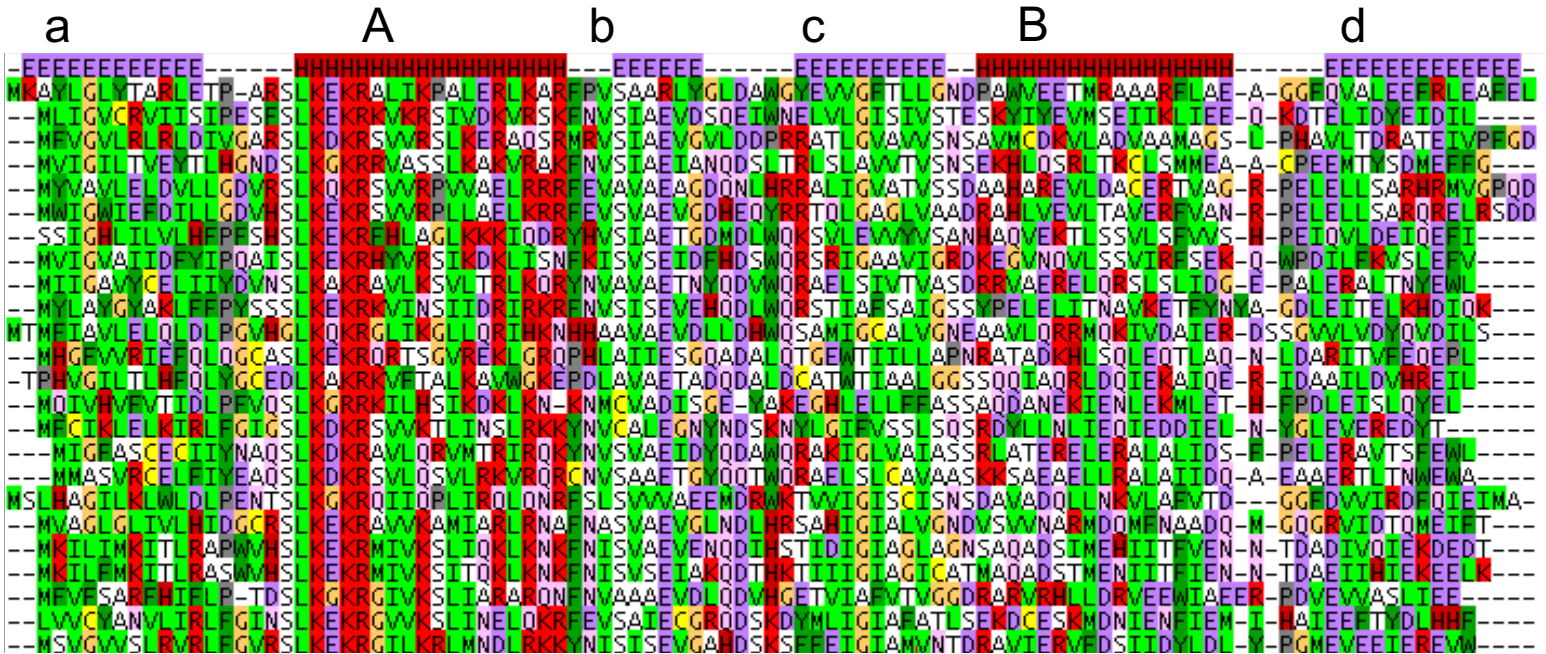
aliphatic	V L I M A C
aromatic	F W Y
small	S T P G
polar	N Q
negative	D E
positive	K R H



Molecular 3D Puzzle

Core residues are often well conserved

Multiple sequence alignment



Note the conserved hydrophobic columns in strands and helices.


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


fold Y




fold X


fold W, W'

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

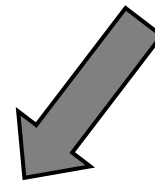
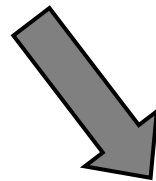
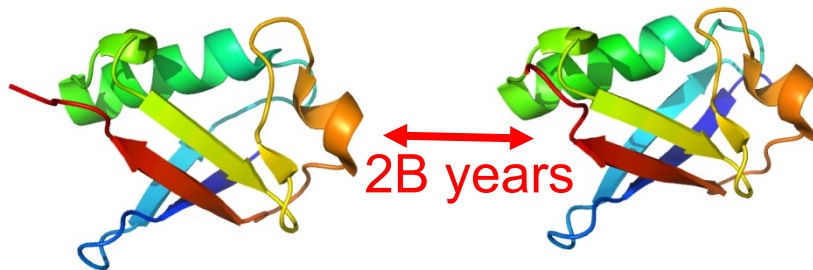

fold Z



Homology-based protein structure and function prediction is powerful

Structures and functions of proteins may be conserved over billions of years

Homology (common descent) can often be predicted from sequence similarity

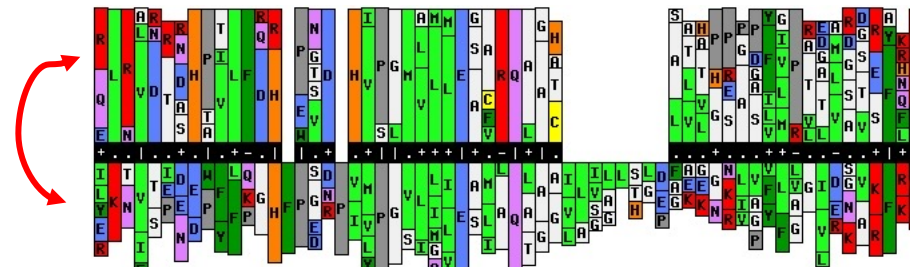
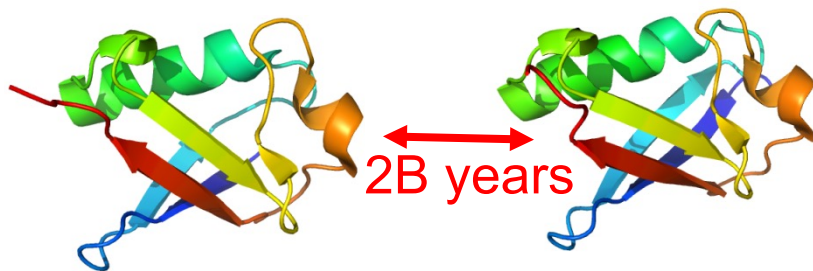


We can predict the structure and function of proteins based on sequence similarity to *homologous* proteins

Homology-based protein structure and function prediction is powerful

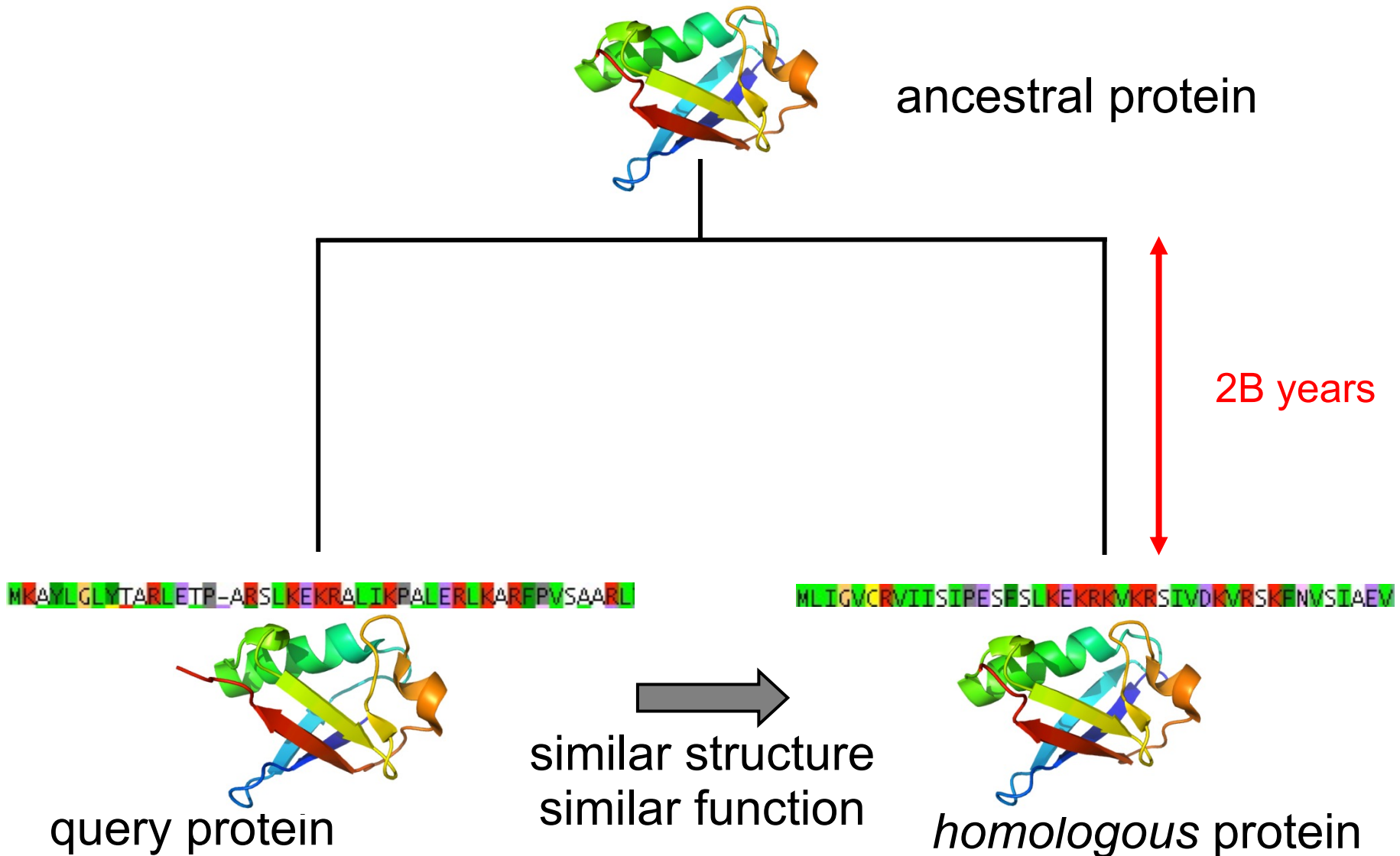
Structures and functions of proteins may be conserved over billions of years

Homology (common descent) can often be predicted by aligning sequence profiles built from closer homologs



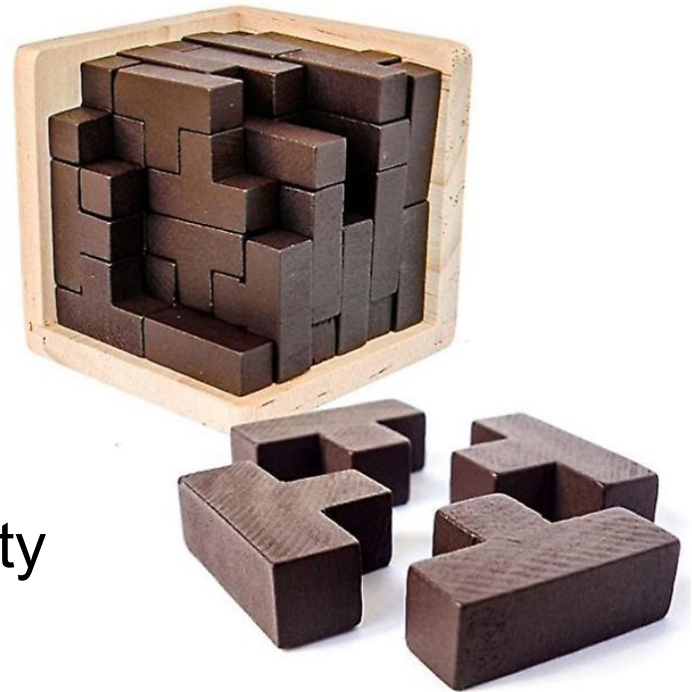
We can predict the structure and function of proteins based on sequence similarity to *homologous* proteins

Homologous = descended from common ancestor



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, which 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



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

Sequences are diverged beyond recognition at
those time scales

We and others develop tools to build and
compare multiple sequence alignments of closer
homologs

From the similarity score we obtain an E-value.
When $E < 0.01$, homology is likely.

Domains are the building blocks of proteins

– their **structural**, **functional**, and **evolutionary** units

- Most eukaryotic 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*)

What do they do?

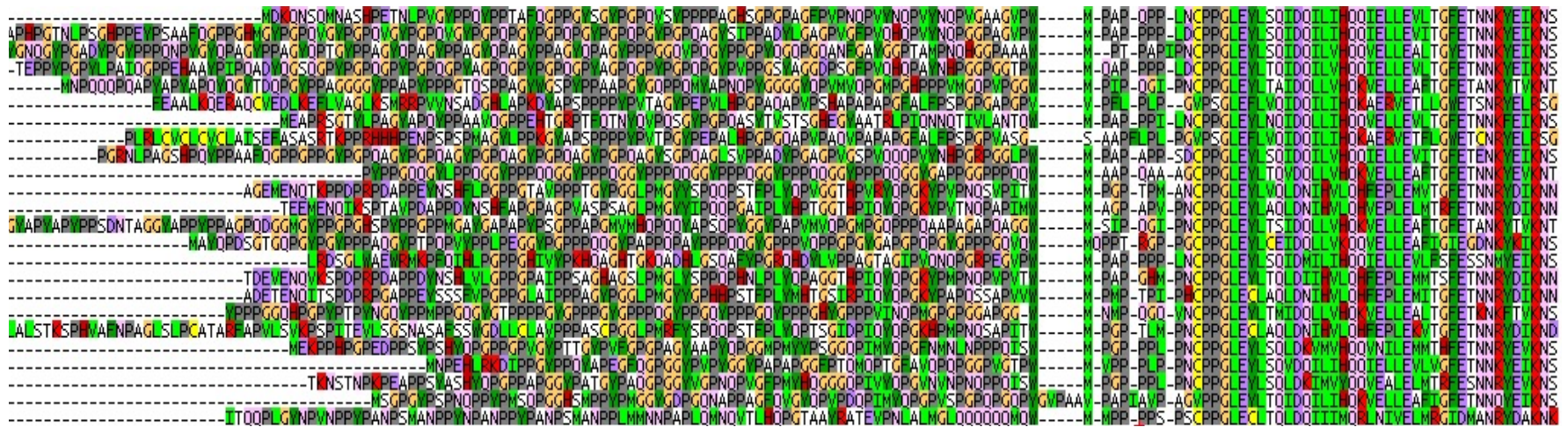
Natively unfolded residues in human proteome: 37% - 50%

Fewer in simpler eukaryotes

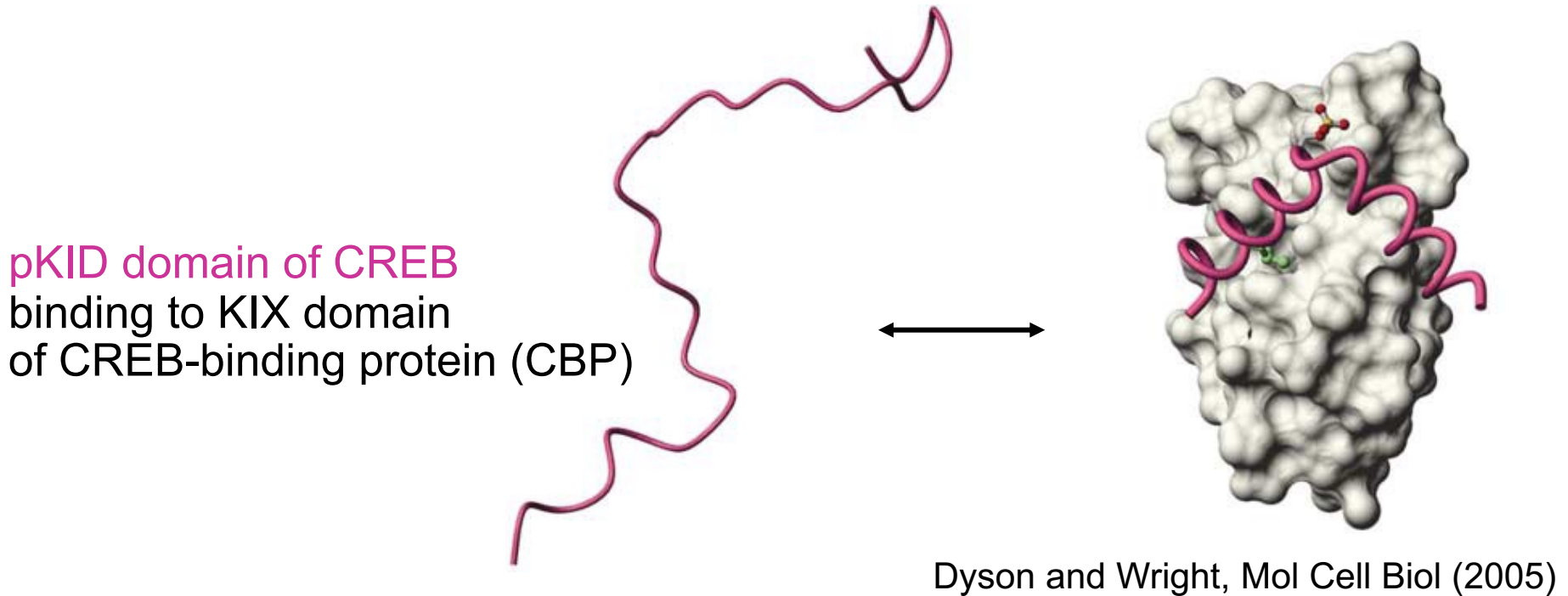
Much fewer in bacteria and archaea (only 3%-25% of their proteins contain disordered regions > 50 aa)

disordered

ordered

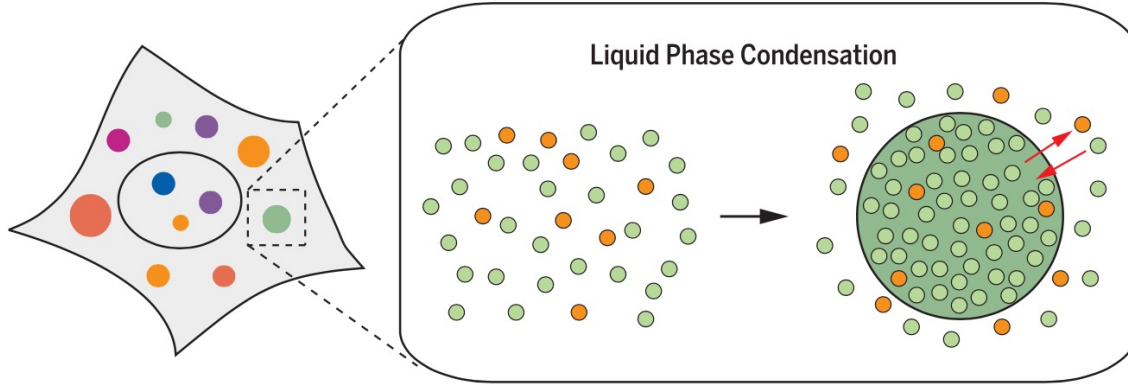


Disordered regions are interspersed with **short linear motifs** that can bind to specific target domains



Short linear motifs **fold upon binding** to their target domain

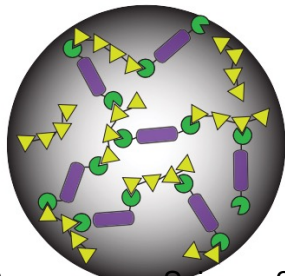
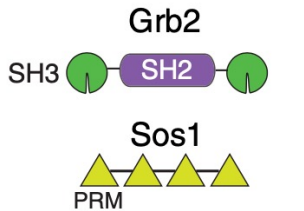
Liquid-liquid phase separation – a long-known phenomenon now revolutionizing cell biology



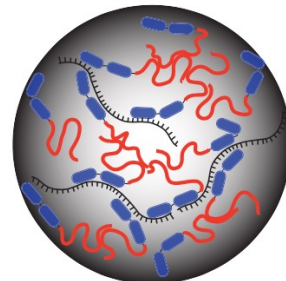
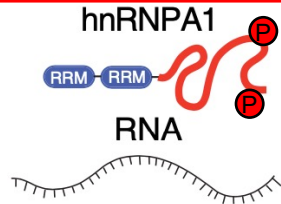
Many types of membraneless droplets exist in cytosol and nucleus of eukaryotic cells: nucleolus, stress granules, P-bodies, splicing speckles,...

Multivalent weak interactions

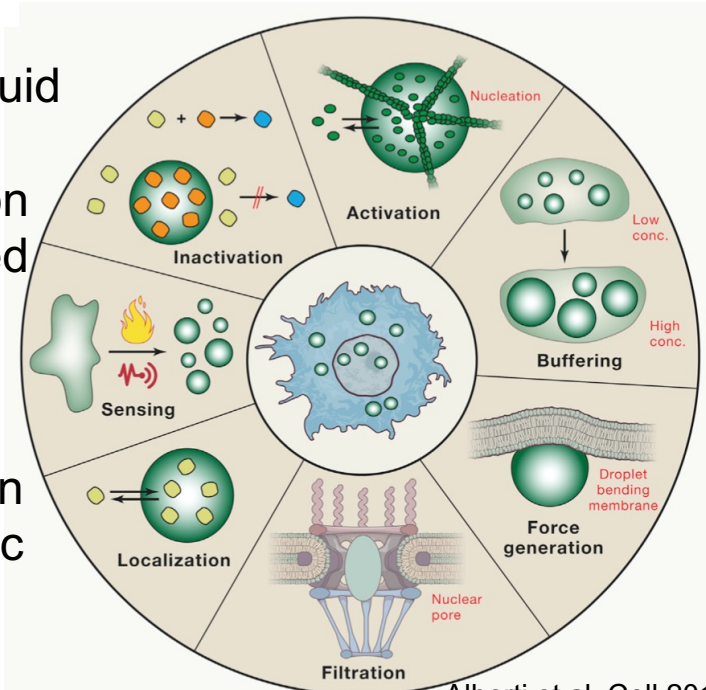
Multivalent Proteins



Disordered Proteins



Liquid-liquid phase separation is involved in almost every cellular process in eukaryotic cells





5 minutes 😊

Sequence searching

```
CFNIDCNSSSTSSFCIDQNCSSNGTCVPIKSVDRCLTYSKDSNGCTEDICLSSNPK  
CIPNSGCSDDNNICTIDSCSNDTGCTADPINACTIONDTGNGGCDIYICDDGILC  
CTEGNGCSIDIDCTIDSCSSITGCFEDVNICTNDICDPGRCLTDNCNDNNIC  
CTNGGGCVDEDICTTDSCSADGCTITATE-CAATCDPETCTVPCDDNNEC  
CSGGDCVMTNDPCTTSCDVEIGCVTEVDPCLPGCGEGECK-ITIDDNNEC  
CWAGICQQTDNSCMVSCNTGGCSENPNFCOISNCVINOCSADGIDDGNAC  
CVDGICETINEDICVEISCDPNVGGCTKPLNICOISCNAGSCSLEIDDNPK  
CSNGGCOITSNDRCKVYCDINSGCVSOPLNICOINNCVTSVCSLEIDDGNPK  
CDGGDCQISNNPKMWSGGLNGCEAPSDLCSNNCCVNNVCTIDGIDDGNAC  
CFPIECNPKGNPPCLPINCTSTDPCTISNENCRITICTRPSVT-----  
CNEQVCSADDINFTVDTCS-SNGVCIITRIDCCNPNVCSPNSCIPNCSIDGNAC  
CFPQICDSIDIDCTTDTCS-ODGICTITPDPCCNPKMCAPIKSLVLCNDNITC  
CDGAGGCTIDINACTIONACTVDSCSNSTGCSITPVNSCTVDSDCGGCVPTACDDINAC  
CQILECNPKVDSNACTIONDTGNGEGECENTPKITCTLDICSEGTCKANBCDDGDDC  
CSISTGCVNDSNPKCTVDSCSNSTGCCNTRINPKCTTDSCTGGVTPVNDNNPK  
CNSTVGGCOINSNFCIDQNCSSNGTCVPIKSVDRCLTYSKDSNGCTEDICLSSNPK  
CSNITGCCNDGNACTTDGCSRETGGCTNSNIDSTTDSCTTGGCSPNSCDDNDAC  
-----DDGDPCTDDEICNGVGCESLP--NN-----KPSVLCQINCNNDNPK  
COTGSCPLDINPKCTIDACDETGVIITITLSNCGGCTCNGGDCDDKICDDGNPK  
CQISVCSPLDGNPKCTDDICDETGVIITITLSNCGGCTCNGGDCDDKICDDGNPK  
CNMILCDTDDGDSCTIDSCISPGACVKEPTDCLPLTCGPMTCAPKICDDENPK
```

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 = $S(V,I) + \dots + S(V,I) + \dots + S(E,G) + \dots + S(G,G) - d - e$
- Find alignment with maximum score, rank by score

Goal of sequence alignment: maximize alignment score

Alignments correspond 1:1 to paths in *dynamic progr.* matrix

	G	A	A	T	T	C	A	G	T	T
A	-1	1	1	-1	-1	-1	1	-1	-1	-1
T	-1	-1	-1	1	1	-1	-1	-1	1	1
T	-1	-1	-1	1	1	-1	-1	-1	1	1
A	-1	1	1	-1	-1	-1	1	-1	-1	-1
G	1	-1	-1	-1	-1	-1	-1	1	-1	-1
G	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
T	-1	-1	-1	1	1	-1	-1	-1	1	1
T	-1	-1	-1	1	1	-1	-1	-1	1	1
T	-1	-1	-1	1	1	-1	-1	-1	1	1

Scores:

match = +1 ↘

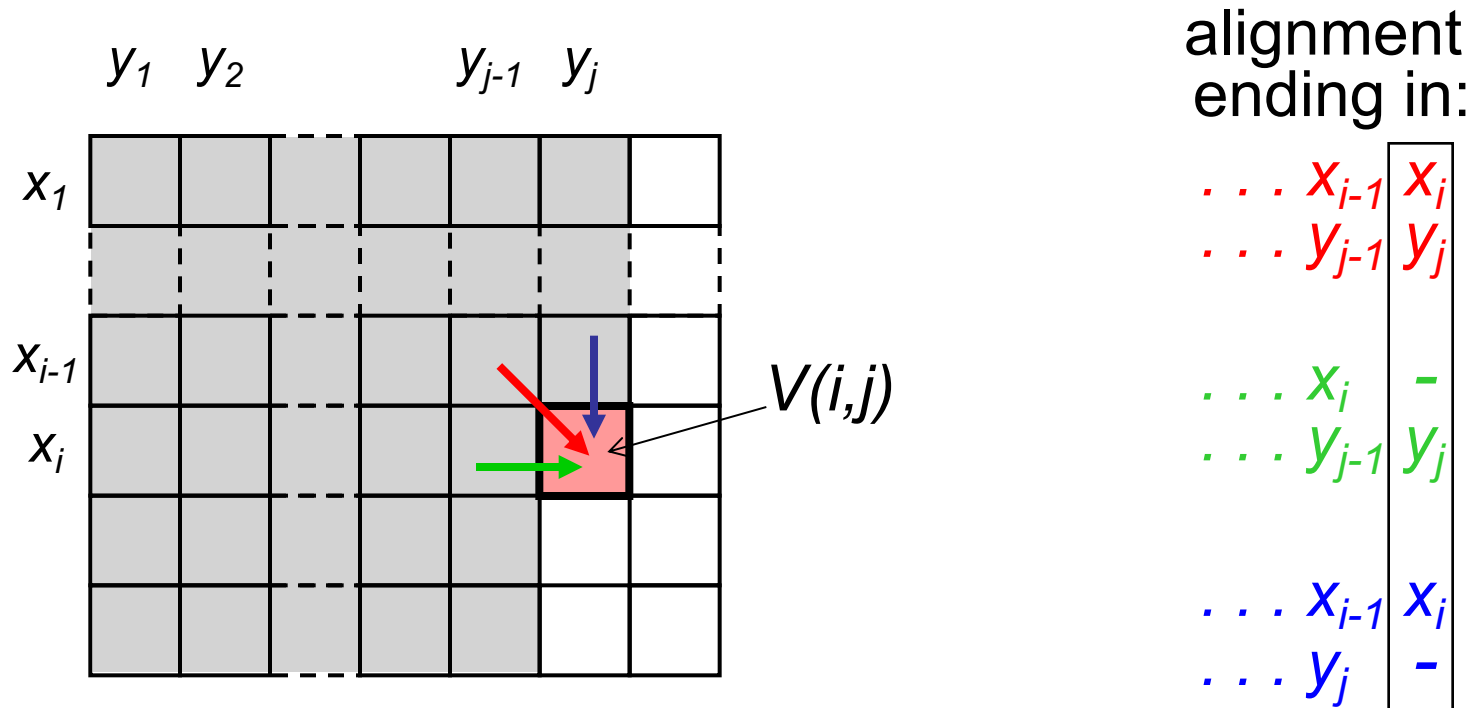
mismatch = -1 ↘

Gap = -1 ↓→

Corresponding
alignment:

GAATT CAG - TT -
- - ATT - AGGTTT

Dynamic programming finds the sequence-sequence alignment with highest score



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

similarity score

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						

similarity scores:

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				
T	0	0	0	2	1	0				
T	0	0	0	1	3	2				
A	0	1	1	0	2	2				
G	1	0	0	0	1					
G	1	0	0	0						
T	0	0	0	1						
T	0	0	0	1						
T	0	0	0	1						

similarity scores:

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

similarity scores:

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

similarity scores:

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

similarity scores:

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

similarity scores:

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

similarity scores:

match = +1

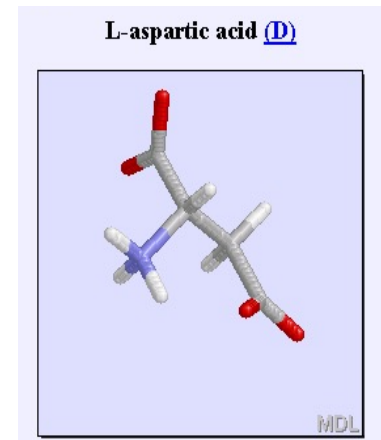
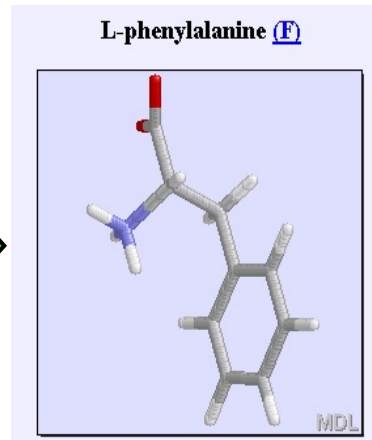
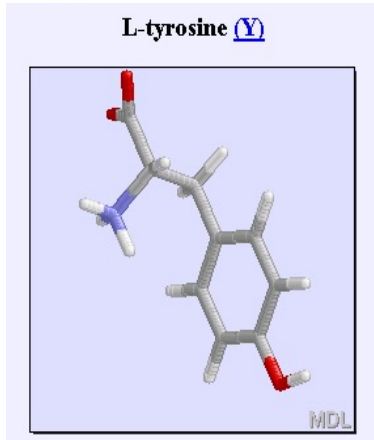
mismatch = -1

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

Substitution matrices score the similarity between amino acids



Similar amino acids can frequently substitute for each other since without fitness loss

Dissimilar amino acids can rarely substitute for each other without fitness loss

How to “measure” similarity between amino acids?

Count how often each pair of amino acids a,b is aligned together

Log-odds score

$$S(a,b) = \log \frac{P(a,b)}{P(a)P(b)}$$



Log odds $P(a,b) / P(a)P(b)$ measures how much more frequently a and b are found aligned than by random chance

$$S(a,b) = \log \frac{P(a,b)}{P(a) P(b)}$$

← Probability for finding (a,b) among aligned residue pairs (model prob.)

← Probability for finding (a,b) among randomly drawn amino acids (null prob.)

Examples:

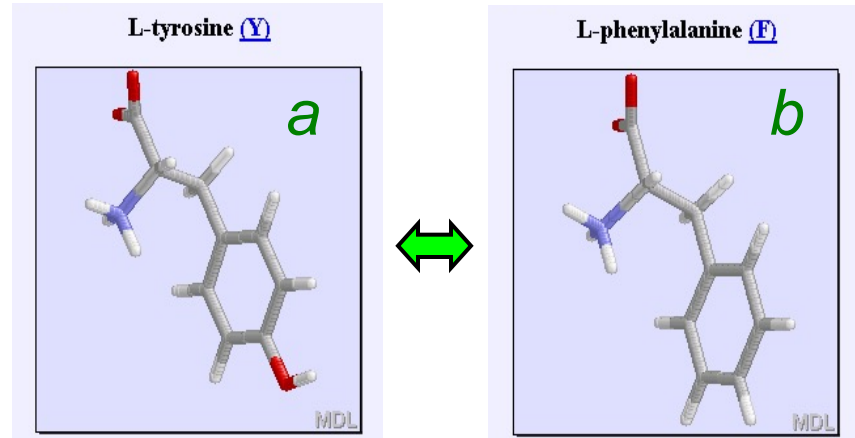
$$S(Y,F) = \log_2 \frac{P(Y,F)}{P(Y) P(F)} = \log_2 \frac{3.7E-3}{3.3E-2 \times 4.0E-2} = \log_2 2.9 = 1.5$$

$$S(W,D) = \log_2 \frac{P(W,D)}{P(W) P(D)} = \log_2 \frac{1.9E-4}{1.3E-2 \times 5.9E-2} = \log_2 0.25 = -2.0$$

Substitutions between **similar** amino acids have $P(a,b) > P(a)P(b) \Rightarrow$ positive score

$$S(a,b) = \log \frac{P(a,b)}{P(a)P(b)}$$

↑
Log-odds score



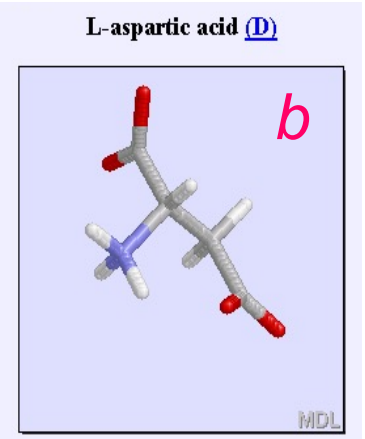
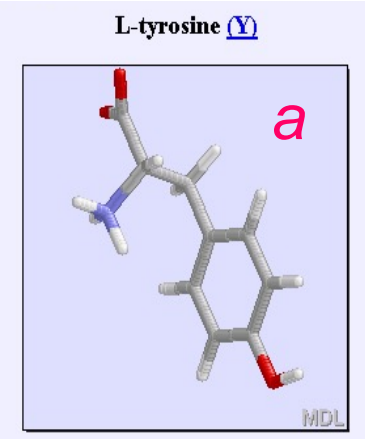
Frequent mutations get positive substitution matrix scores

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
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Substitutions between **dissimilar** amino acids have $P(a,b) < P(a)P(b) \Rightarrow$ negative score

$$S(a,b) = \log \frac{P(a,b)}{P(a)P(b)}$$

↑
Log-odds score



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
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Rare substitutions get negative substitution matrix scores



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 mutations (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
dlbtea_ ETQECLFFNAN--EEDRITNQTGVVFCVGDNDKRRRHCATW--FNISGSIEIVKQGCMLDDINCDDPDCIEKKDSPE--VVFCCCEGNMCNEKFSYPPEME
---ECEHDEKMCNNTTQQCETRIEHCMEADKFPSCVLSVNETTGILIKMKGCGTDMHEC-NQTECVTSAPFQGNHIFCCCKGSPCNSENQKVI----

BLAST E-value = 0.2

	* * *		*		*	** *		*	*	** ** *	*		**		*		***	**							
gi 539437	ETQECLFFNAN--N--EEDRT---NQTV--EP-CVGDNDKRRRHCATW--FNISGSIEIVKQGCMLDDINCDDPDCIEKKDSPE--VVFCCCEGNMCNEKFSYPPEME	gi 91922	ETHECLFFNAN--N--ELERT---NQSL--EP-CEGEQDKRLHCASW--FNS-SGTELVKKGCM---DDINCDDPQECVATEENPO--VF---CCCEGNMCNERFTHLPE--	gi 213934	ETHECLFFNAN--N--ELEKT---NQSGV--ERLVEGKDKRLHCASW--FNN-SGTELVKKGCM---DDINCDDPQECTAKEENPO--VF---CCCEGNMCNERFTHLPEVE	gi 54638211	---CEHDEKMCNK--EQDCT---VF--I--EFCQVEIDKFPSCVLSANEETGAKKIKMKGCGT---DMHEC-NQTECVTSAPFQGNMHI---CCCKGSLCNSDQKVI	gi 114724	ETHECLFFNAN--N--EEDKT---NSNGT---EIV-CVGDNDKRRRHCATW--FNISGSIEIVKQGCMLDDINCDDPDCIEKKDSPE--VVFCCCEGNMCNEKFSYPPEME	gi 31418321	PERICAKKDE---VQQDLGIGE-SRISH--EN-GT-ILCSKSTCGGLW--EKS-KGDINLVKQGCMSHIGDPQECH--EECVVTTTPPS--IQNGTVRCCSTDLCNVNNT	gi 2150128	ETHECLFFNAN--N--EEVET---NSGV--EP-CEGEQDKRSHCASW--FNS-SGTELVKKGCM---DDINCDDPQECVATEENPO--VF---CCCEGDVNERFTHLPDI	gi 47218579	ETQECATNS---SV--EEDRT---NSGI--EP-CVSGEVDKRRRHCATW--FNISGAVVVKQGCMLDDVNCDDSNELVERKESPD--VF---CCCEGNMCNEKFLVPEVQ	gi 1764144	EEERICAKKDE---N--QDQGVSE--SQVSL--EN-GT-VKCTKGNICGLW--EKTREGDINLVKQGCMSHVGDVPHDCN-DECVVTTTPPV--IQNGTVRCCCKDKMCNVNNT	gi 47223056	ETHECVYIND---N--TEERT---NQSGV--EP-CEGEQDKRLHCASW--FNS-SGTEIKLVKKGCM---DDINCDDPQECVSMEEENPO--VF---CCCEGNMCNERFTHLPDI	gi 47825379	EEERICAKKDE---VQ-QDHGI---SESRIISQEN-GT-ILCMKSTCGGLW--EKTREGDINLVKQGCMSHIGDPQECH--EECLVTTTPSL--IQNGTVRCCSTDLCNVNNT	gi 47218656	EEECATDQQQ-Q--EVERMAGGEGQISF--EN-TT-VCCKGSGNCGLWE--KSP-PGEVRLVQGCMTVSDRQSCD-DCVVTNLPPQ--IQNGTVRCCCGSDMCNVNNT	dlbtea_	---ECEHDEKMCNNTTQQCETRIEHCMEADKFPSCVLSVNETTGILIKMKGCGTDMHEC-NQTECVTSAPFQGNHIFCCCKGSPCNSENQKVI----

PSI-BLAST E-value = 1E-17

Yes they are!

Sequence-sequence alignment uses substitution matrix scores

```
gi|539437  ETQECLFNAN--EEDFINQTGVEPCGDIDRRHCAT--INISGSIEIVKQGCILDDINCDEIDCLEKSDSPE--VYFCCEGNMCNEKSYFPEME  
dlbtea_    ---ECEHDEKMCNNTTQQCETRIEICKMEADKIPSCVLSVNETTGILIKKMGCTDMIEC-NQTECVTSAEPPQGNIFCCCKGSPCNQKVI-----
```

Sequence-sequence alignment uses substitution matrix scores

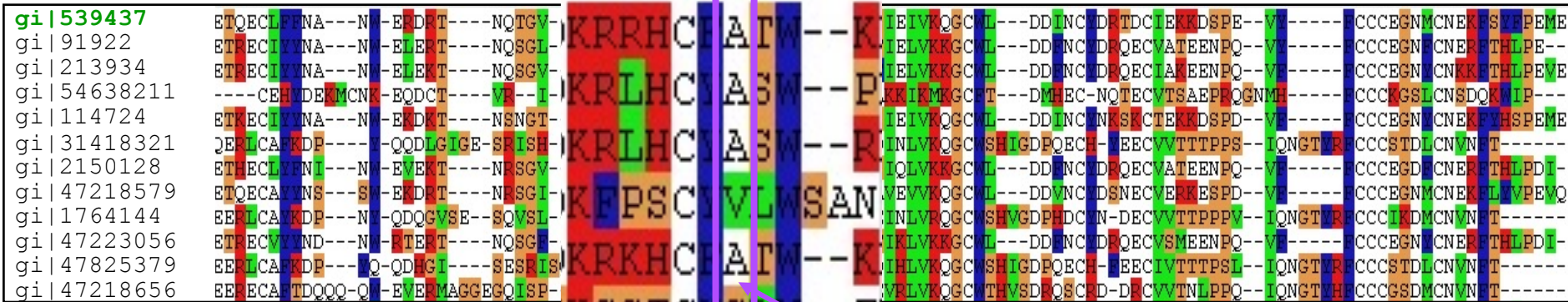
gi|539437 dlbtea_ ETKQECI DKRRRHCEATW--KNIS EKKDSPE--VIFCCCEGNMCNEKSYFPEME
 ---ECE TSAEFPQGNIFCCCKGSPCNENQKVI----

A	4
R	-1
N	-2
D	-2
C	0
Q	-1
E	-1
G	0
H	-2
I	-1
L	-1
K	-1
M	-1
F	-2
P	-1
S	1
T	0
W	-3
Y	-2
V	0
aa	S(A,aa)

A-column of substitution matrix contains scores for substituting A (alanine) with each of the 20 amino acids “aa”

$$S(A,aa) = \log \frac{P(A,aa)}{P(A) P(aa)}$$

What score to use for aligning an MSA with a sequence?



Scores for finding each of the 20 amino acids "*aa*" in this position ?

Count how often amino acid *aa* appears in MSA column *j*!

Log-odds *profile score*

$$S_j(aa) = \log \frac{P_j(aa)}{P(aa)}$$

Sequence profiles are a condensed representation of multiple alignments

They contain *position-specific* amino acid substitution scores

```

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 scores quantify how much more frequent each amino acid *aa* is in column *j* of the MSA than its average frequency in the db:

$$S_j(aa) = \log \frac{p_j(aa)}{p_{av}(aa)}$$

↑
log-odds score

$p_j(aa)$ = frequency of *aa* in column (incl. pseudo-counts)

		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	...

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

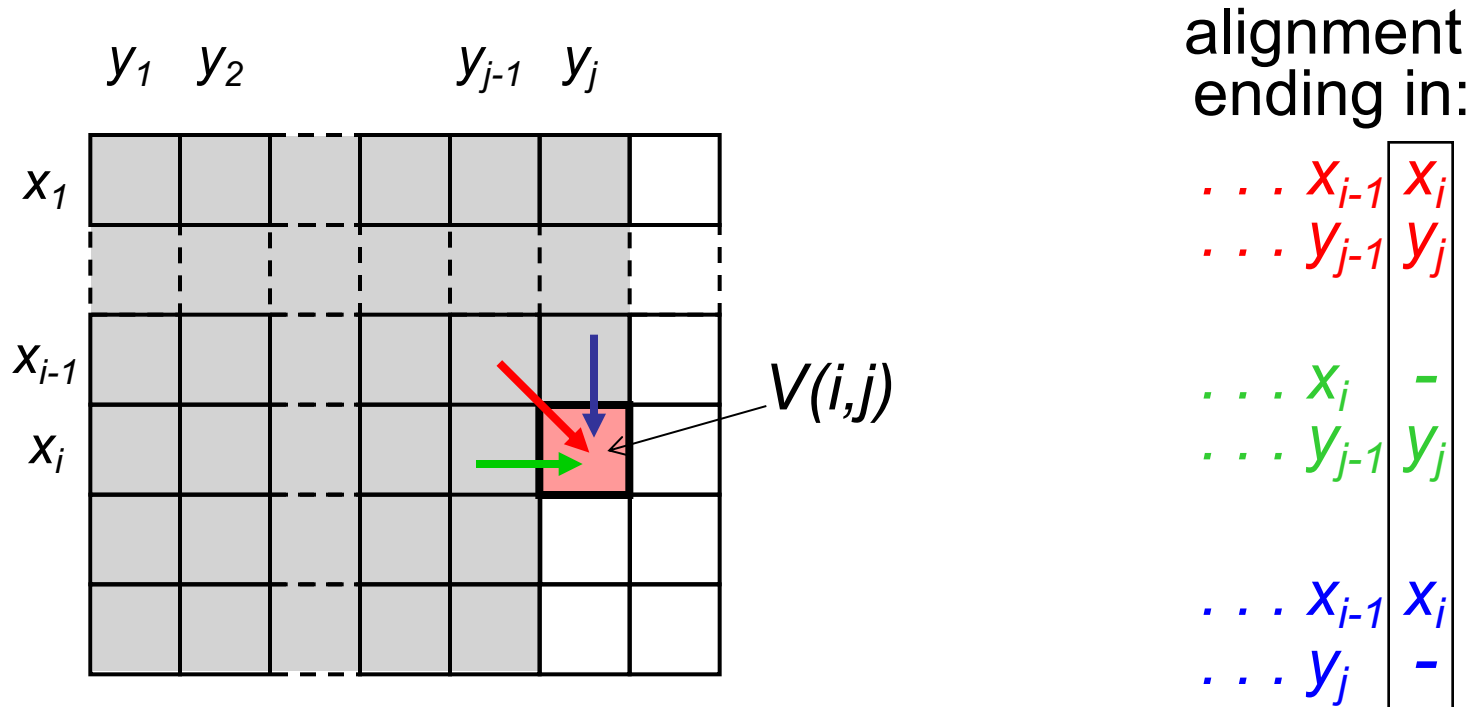
		W	G	K	V	G	A			H	A	G	E	
A	...	-3,2	-1,9	-2,1	-2,2	-2,0	3,4			-2,1	1,4	1,5	-2,0	...
C	...	-2,3	-2,8	-2,9	-2,1	-2,7	-1,8			-2,7	-2,1	-2,6	-2,9	...
D	...	-3,7	-1,6	-1,6	-3,1	-1,4	-2,1			2,0	-2,8	1,6	-1,5	...
E	...	-3,4	2,1	2,1	-2,8	2,1	-2,0			-1,6	-2,5	-1,9	2,5	...
F	...	-0,8	-3,6	-3,2	2,9	-3,3	-2,8			-2,8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

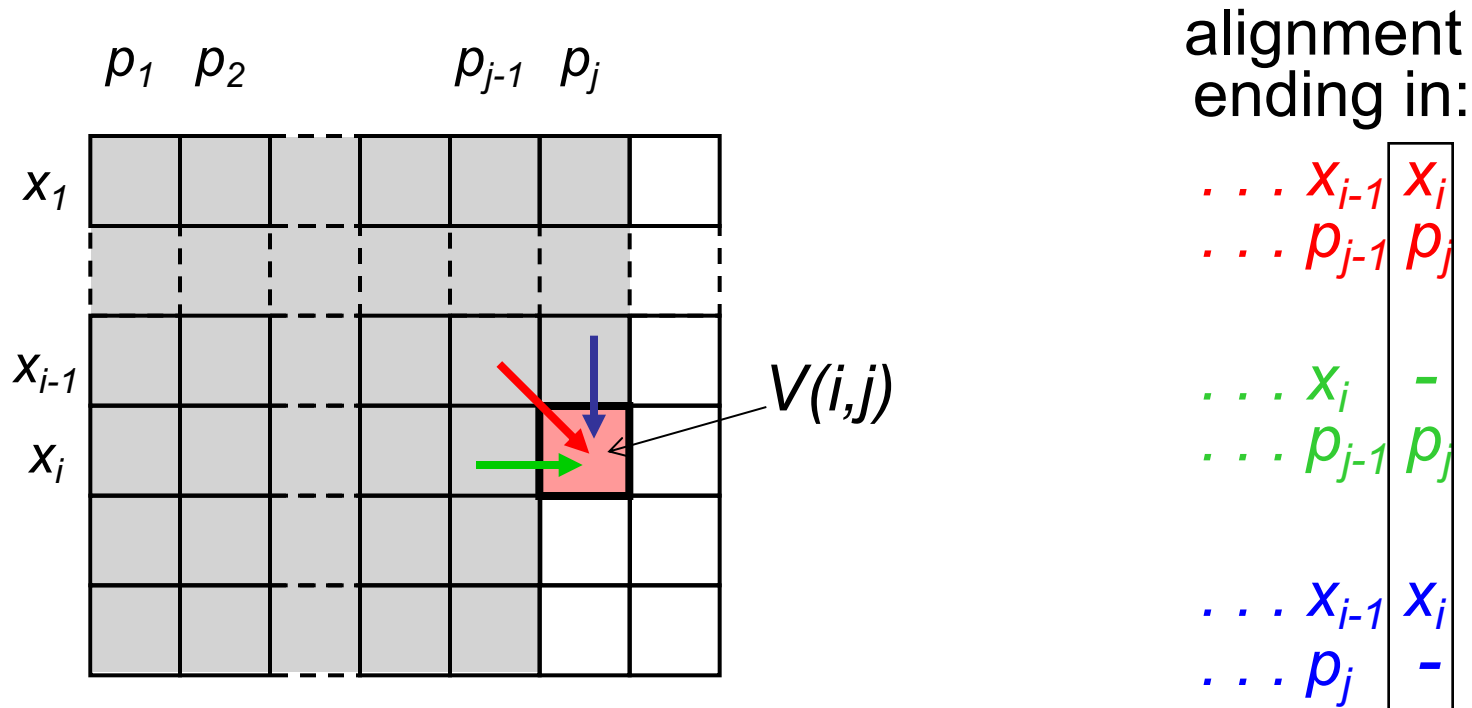
Dynamic programming finds sequence-sequence alignment with highest score



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

substitution matrix

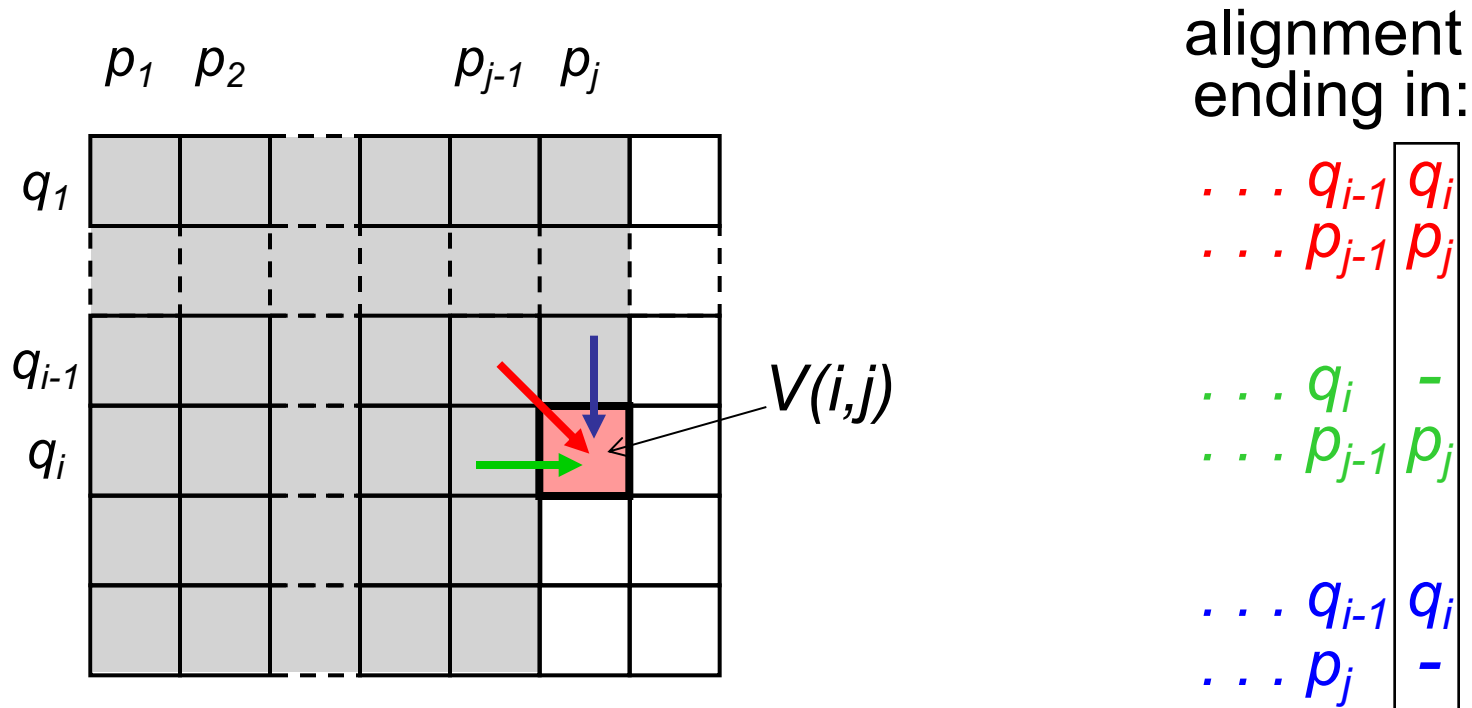
Dynamic programming finds profile-sequence alignment with highest score



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

Profile score

Dynamic programming finds profile-profile alignment with highest score



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

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	...

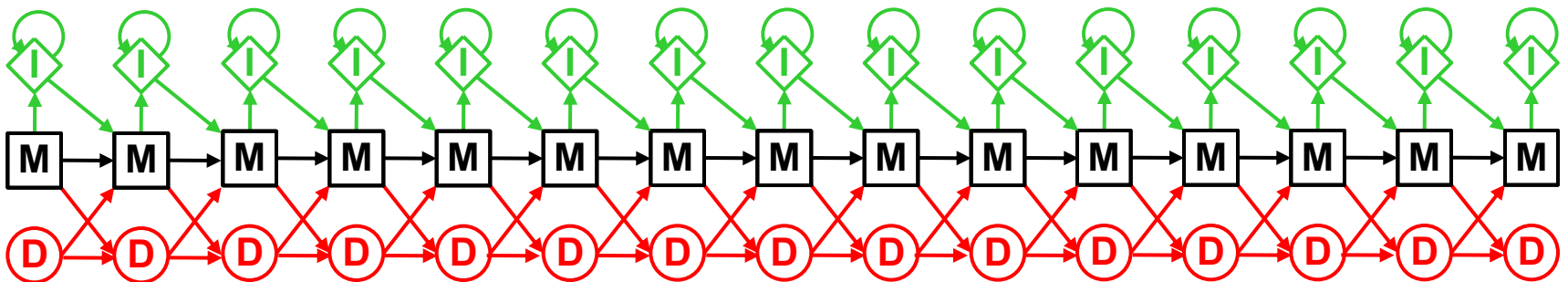
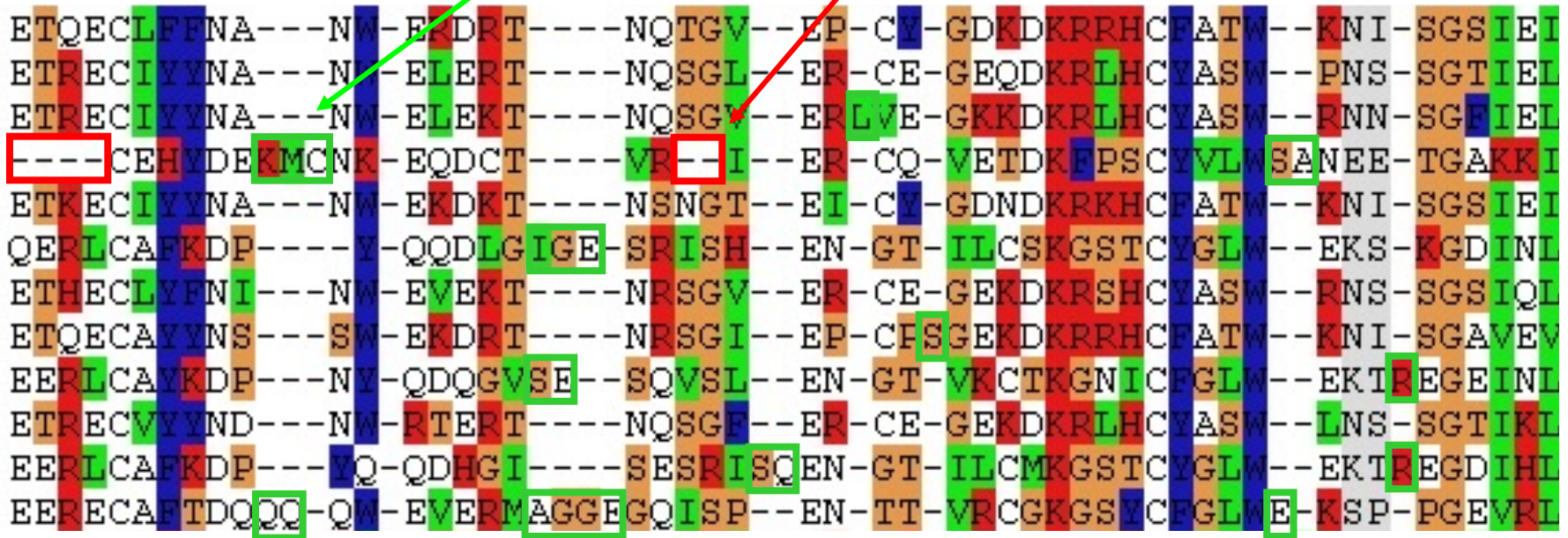
		W	K	D	I	A	G	A	D	N	G	A	V	
A		-3,1	1,8	-2,0	-2,1	2,2	-1,8	3,4	-2,1	-2,0	-2,2	2,5	-1,8	...
C		-2,3	-2,5	-3,0	-2,1	-2,2	-2,4	-1,8	-3,1	-2,4	-2,4	-2,2	-2,4	...
D		-3,7	2,0	2,7	-3,1	-2,2	-1,9	-2,1	3,9	-1,6	-2,3	-1,6	-2,0	...
...
V		-2,6	-2,4	-2,7	2,7	-2,2	-2,8	-2,0	-3,0	-2,4	-2,7	-2,2	-2,5	...
W		5,6	-3,3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

Compare amino acid distributions

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

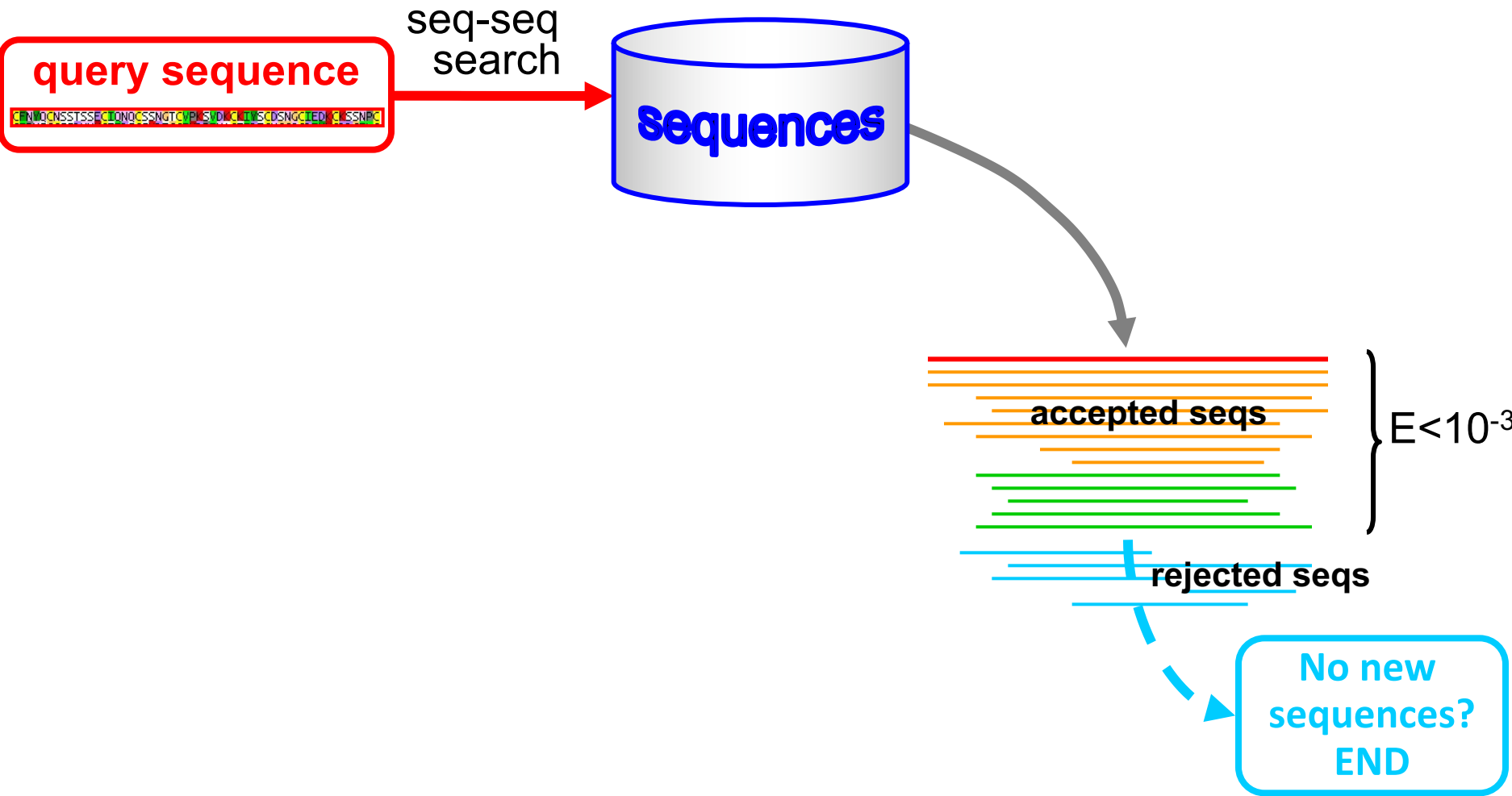
A profile HMM is a sequence profile extended by position-specific gap penalties

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



BLAST

Search with **single sequence** through **sequence database**



PSI-BLAST

Iterative search with **sequence profile** through **sequence db**

query sequence

```
CINLDCNSSTSSFCINQDCSSNGTCLPISVLDCLTNSCDSNGCLEDIQLSSNPLC
```

Sequences

```
CINLDCNSSTSSFCINQDCSSNGTCLPISVLDCLTNSCDSNGCLEDIQLSSNPLC
CIEGNGCSLDIDCTTDCSSITGCEBQINDIENDICDPRCTTDCNDNNEC
CINGGQCVLDEDICTTDCSADGCTITALE-CHVATCPETCTVPCDDNNEC
CSGGCCVMTNDRCTTDCSADGCTITALE-CHVATCPETCTVPCDDNNEC
CAGKCNQITDNCSCNNTAGGCSFNPNFCVLSNCHNOCVADGIDDGNAC
CNDGKCEI-NEDVLESCDPRGCTTIPNLCQVSCNAGSCVLEIDDDNPLC
CSNGGCVT-SNDRPCLVDCDINSQCSPNPNCCVNNCTSVCSIDEIDDDGNPC
CDGGCVQV-SNDRPCLVDCDINSQCSPNPNCCVNNCTSVCSIDEIDDDGNPC
CPRFCNPRGNPRLPINCSTDPCTESVENCRTVICTTBSVT-----
CNEQVCSADDINECTVDTTC-SNGVCTITPDCNPNVCSFNSCIPNCSVGNAC
CRPDCBSVLDIACTTDTTC-DDGLCTITPDCNPNVCSFNSCIPNCSVGNAC
CDGCGGCTV-DNACTVDCSNSTGCSVTPNNSCTVDCGGGVVTA-CDDNINAC
CQGLENPVDSNACTVDTTNGEGECENTP-ETLTLVDCVSGFTV-ANPCDDGDDC
CSVSTGCVNDSNACTVDCSNSTGCSVTPNNSCTVDCGGGVVTA-CDDNINAC
QNSTVGCNDSNACTVDCSNSTGCSVTPNNSCTVDCGGGVVTA-CDDNINAC
CSVSTGCVNDSNACTVDCSNSTGCSVTPNNSCTVDCGGGVVTA-CDDNINAC
CSVSTGCVNDSNACTVDCSNSTGCSVTPNNSCTVDCGGGVVTA-CDDNINAC
-----DDGPRCTDDCINGVCCSLP-NN-----RPSVTCQVNCNDNPLC
CQVGSCLPVDINPCTVDADETVGLTTLVSNVGGCSVINGGDDDDV-CDDGNPC
CQVSNVCLPVDINPCTVDADETVGLTTLVSNVGGCSVINGGDDDDV-CDDGNPC
CQVSNVCLPVDINPCTVDADETVGLTTLVSNVGGCSVINGGDDDDV-CDDGNPC
CQVSNVCLPVDINPCTVDADETVGLTTLVSNVGGCSVINGGDDDDV-CDDGNPC
```

evolving alignment

accepted seqs

$E < 10^{-3}$

rejected seqs

No new sequences?
END

add homologs

Much more sensitive than BLAST

PSI-BLAST, MMseqs2

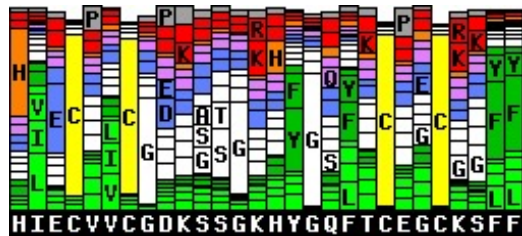
Iterative search with **sequence profile** through **sequence db**

query sequence

CINLQCSSTSSFCIQNQCSNQTCPVLSVQICLISCDISNGCLEDCSSNPR

UniProt

profile-seq
search



evolving profile

accepted seqs

$E < 10^{-3}$

rejected seqs

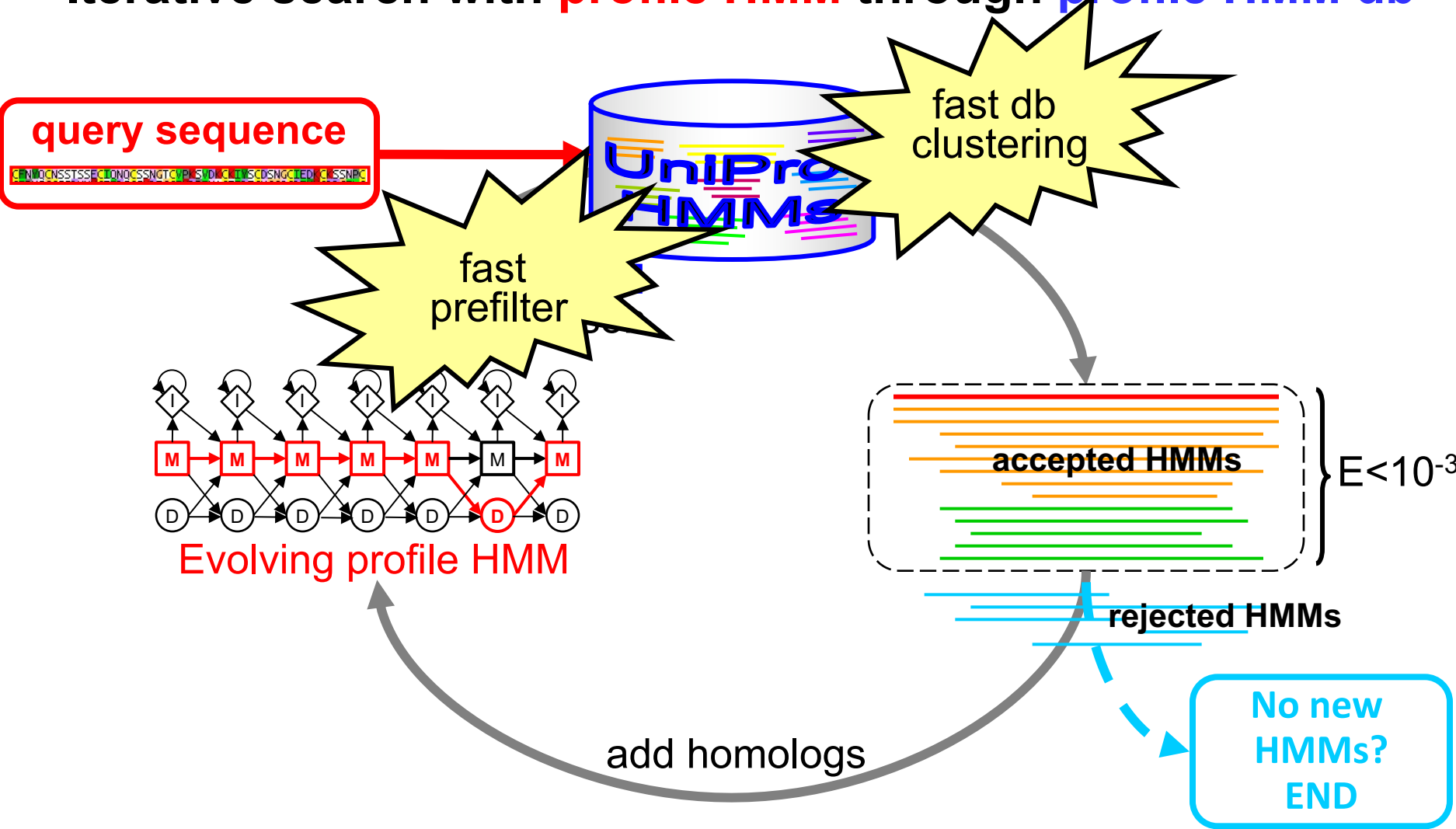
No new
sequences?
END

add homologs

Much more sensitive than BLAST

HHblits

Iterative search with **profile HMM** through **profile HMM db**



Best sensitivity, alignment quality, and speed



See you back at 13:30h 😊

Student feedback after lecture

- I think the speed was good, the explanations were clear but *I needed more breaks*, specially for the first part of the presentations. *Small 5 min breaks would be fine*
- alignment for the section of *profile based dynamic programming*
- I found the *part of HMM a bit difficult to comprehend*. I liked the dynamic programming exercise.
- maybe could introduce a little more about the principles of clustering?
- I was following the local alignment algorithm explanation until the part where *position specific score* was introduced. Then started getting a bit confused
- Everything was great. Actually, I would prefer to have more days of practice and more command line exercises but I am not sure that you can change it. Thanks a lot!
- The parts where sequence profiles are explained can be explained a bit more in detail. Also, overall the presentation has gone fast for me as not everyone has the same background in bioinformatics.
- The slides were presented well but it was a bit fast and sometimes was difficult to analyze some technical stuff. otherwise the basics were very well explained. Thanks
- I think the hardest topic was the matrix of similarity calculations. It was well explained, but I felt it a little fast. Some more examples and exercises would have helped
- explanation of the *log odds score*
- The details of MMseqs2 was difficult to understand
- The course was great, I really liked the tutors were very responsive to questions. Also the organization of the course is very nice, with the breakout rooms, the breaks, and the general meetings. Many of the info were new to me, but I finished the course feeling like I understood it very well. I liked the fact that some exercises were put in the middle to help us figure things out ourselves. Thank you for your efforts!
- I feel that we could *go slower on the topic of HMM- profile, profile-profile comparison* as there are many complex things to understand and visualize.
- I did not understand too much the part that covered *MMSeq, BLAST, HHMER3 and HHblits*. I think it was too quick. In general, *there are times where I think the professor spoke too fast*. Everything else was great!

Small P-value: reject 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 never observed a 6. Can we conclude that the die is loaded?



Exercise: Compute the P-value for the *null hypothesis* that the die is fair. What if we observed a 6 only once out of 30 throws?

What do you conclude from the P-value?

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

$$\begin{aligned} P(k \leq 1 \text{ six out of } 30 | p_{\text{six}} = 1/6) &= \sum_{k=0}^1 \binom{30}{k} (1/6)^k (5/6)^{30-k} \\ &= \binom{30}{0} (1/6)^0 (5/6)^{30} + \binom{30}{1} (1/6) (5/6)^{29} = 0.0042 + 0.0253 = 0.029 \end{aligned}$$

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

Why „or more extreme“?

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

Suppose we throw a die $6N$ times and observe a six N times. What do you guess is the P-value?

Why “or more extreme” in the definition of the P-value?

Please type in your answers at

<https://forms.google.com/???>

The linux command line (bash)

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

<code>ls</code>	list content of current directory
<code>ls -ltrF</code>	ls in <u>l</u> ong format, <u>t</u> ime-sorted in <u>r</u> everse order, with <u>F</u> iletype
<code>cd <path/dir></code>	change to directory <path/dir>
<code>cd ..</code>	go up 1 step in directory hierarchy
<code>gedit <file></code>	open file in editor
<code>gedit <file> &</code>	open file in editor <i>in background</i>
<code>less <file></code>	look at raw file (q: quit, b: back, /: find); works for huge files
<code>cp <file> <dest></code>	copy file to destination directory (cp file.txt ~/molbiol/day1/)
<code>mv <file> <dest></code>	move file to destination directory
<code>rm <file></code>	remove file (careful!)
<code>mkdir <dir></code>	create new directory (remove with <code>rmdir <dir></code>)
<code>info ls, man ls</code>	show info / manual page of ls command
<code>chmod +x <file></code>	change settings of file to make it ex ecutable