# **Computational Biology** (BIOSC 1540)

### Lecture 11A

### Protein structure prediction

Foundations

Mar 25, 2025





# Announcements

#### Assignments

- P02B is due Mar 28
  - PO2C is due Mar 28
  - P03A is due Apr 4

• Quiz 04 will be on Apr 8 and cover L09A to L12A

#### **Final exam**

 The final exam is on Monday, Apr 28, at 4:00 pm in 244 Cathedral of Learning

# After today, you should have a better understanding of

Identify what makes structure prediction challenging

# Protein structure is essential for understanding biological function

Proteins are molecular machines; their 3D shape determines how they interact with substrates, DNA, other proteins, etc.



# Experimental methods for structure determination are powerful but limited

X-ray crystallography, NMR, and cryo-EM provide high-resolution data.

However, these methods are time-consuming, expensive, and often fail for specific proteins.



As of 2021, >200 million protein sequences exist, but <200,000 structures are known

# Protein structure prediction fills a crucial gap in biological discovery

Enables structural understanding of sequences with no experimental structure. This accelerates many research fields and democratizes access to atomistic insights



**Example:** Our collaborators (Dr. Cahoon) crystallized *Lm* PrsA1 in 2016, but we need a structural model of PrsA2. Instead of potentially years, AlphaFold 3 gives us a decent prediction within minutes

Protein folding is computationally hard due to the vast conformational space

**Levinthal's Paradox:** A protein can't sample all conformations in a biologically reasonable time, yet it folds quickly

**Example:** A protein with 100 amino acids, each capable of adopting about 3 torsion angles, results in  $\sim 3^{100}$  possible conformations

Proteins fold in milliseconds—implying nature doesn't sample all conformations.



# Proteins fold into their native structure by minimizing free energy

#### A potential energy surface (PES)

represents the energy of a system as a function of the positions of its atoms

Allows us to understand how the system's energy changes upon reactions or movements

Proteins adopt conformations that minimize thermodynamic free energy

Scoring functions attempt to model this with statistical or physics-based potentials



# Flexible and disordered regions add complexity to structure prediction

Many proteins exist in ensembles of structures or are natively disordered.

Their function may depend on transient interactions or induced folding.





# Environmental context dramatically impacts protein folding

Proteins fold differently in different environments

7MHX

Predictions need to capture interactions with solvent molecules, ions, and cofactors



**Example:** Predicting transmembrane protein structures, where the lipid bilayer plays a key role in folding, is particularly complex.

# Environmental context dramatically impacts protein folding

PTMs such as phosphorylation, glycosylation, and methylation can alter protein folding and function

**Example:** eIF4E is a eukaryotic translation initiation factor involved in directing ribosomes to the cap structure of mRNAs

Ser209 is phosphorylated by MNK1

AlphaFold 3 accurately predicts these changes when they are already known



After today, you should have a better understanding of

#### Homology modeling

**Hidden Markov Model alignments** 

# Homology modeling predicts structure using evolutionary relationships

Template Results 6

# Based on the principle that **proteins with similar sequences tend to adopt similar structures**.

Often the first modeling strategy attempted due to simplicity and reliability.

Requires a *template*—a known structure with detectable sequence similarity.

	_							
emplates	Quaternary Structure	Sequence Similarity	Alignment			Build Models	5	
ore 🗸						Clear Selection		
Target MT	LSILVAHDLQRVIGFEN	QLPWHL - PNDLKHVKK	LST	37				
e4e.1.AMT	LSILVAHDLQRVIGPEN	DLPWHL PNDLKHVKK	L S T	60				
or6.1.A 🕅	LSILVAHDLQRVIGPEN	DLPWHL PNDLKHVKK	L S T	37				
sqy.1.A 🕅	LSILVAHDLQRVIGPEN	DLPWHL PNDLKHVKK	L S T	37				
fyw.1.A - T	LSILVAHDLQRVIGPEN	DLPWHL PNDLKHVKK	L S T	36				
or8.1.A - 👖	LSILVAHDLQRVIGPEN	DLPWHL PNDLKHVKK	L S T	36			0	
dga.1.A - 🔣	KNEVFNNYTF <b>R</b> GL <b>G</b> KG	<b>LPW</b> KCNSLDMKYFCA	VT <b>T</b> )YVN(ESKYEKLKY	75		2000	3	
Target		GHTLVMGRKTFE	SIGKPLPNRRNV	61		<b>S</b>	Ď	
e4e.1.A		GHTLVMGRKTFE	SIGKPLPNRRNV	84		A ATO		
or6.1.A		GHTLVMGRKTFE	SIGKPLPNRRNV	61				
sqy.1.A		GHTLVMGRKTFE	SIGKPLPNRRNV	61				
fyw.1.A		GHTLVMGRKTFE	SIGKPLPNRRNV	60		₩.C		
or8.1.A		GHTLVMGRKTFE	SIGKPLPNRRNV	60				
dga.1.A KR	CKYL <mark>NKE</mark> TVDNVNDMPN	SKKLQNVVVMGRTSWE	SIPKKFKPLSNRINV	125	A Co	rtoon .		C
Target VL	TSDTSFN VEGVDVIH	SIEDIYQL PGH	VFIFGGQTLFEEMID	104	₩ Ca			-
e4e.1.A VD	TSDTSFNVEGVDVDH	SIEDIYQL PGH	VFDFGGQTLFEEMID	127				
or6.1.A 🚺	TSDTSFNVEGVDVIH	SIEDIYQL PGH	VFIFGGQTLFEEMID	104	6e4e.1.A			*
5qy.1.A 🚺	TSDTSFNVEGVDVIH	SIEDIYQL PGH	VFIFGGQTLFEEMID	104	6pr6.1.A			×
fyw.1.A VL	TSDTSFNVEGVDVIH	SIEDIYQL PGH	VFIFGGQTLFEEMID	103				
or8.1.A VL	TSDTSFNVEGVDVIH	SIEDIYQL PGH	VFIFGGQTLFEEMID	103	3sqy.1.A			×
dga . 1 . A 竝	SRTLKKEDFDEDVYIDN	K <u>vedlivl</u> lgk)lnyyk	CFIIGGSVVYQEFLE	175	3fyw.1.A			×
Target	KVDDMYITVIEGKFRGD	TFFPPYTFEDWEVASS	VEGKLDEKNTIPHTF	152	6pr8 1 A			*
e4e.1.A	KVDDMYITVDEGKFRGD	TPFPPYTFEDWEVASS	VEGKLDEKNTIPHTF	175	0010.1.7			
or6.1.A	KVDDMYITVDEGKFRGD	TFFPPYTFEDWEVASS	VEGKLDEKNTIPHTF	152	3dga.1.A			×
sqy.1.A	KVDDMYITVDEGKFRGD	TFFPPYTFEDWEVASS	VEGKLDEKNTIPHTF	152				
fyw.1.A	KVDDMYITVDEGKFRGD	TFFPPYTFEDWEVASS	VEGKLDEKNTIPHTF	151				
or8.1.A	KVDDMYITVDEGKFRGD	TFFPPYTFEDWEVASS	VEGKLDEKNTIPHTF	151				
dga.1.A <mark>KK</mark>	LIKKIYFTRINSTOECD	FPEINENEYQIIS	SDVYDSNNTTLDF	223				
Target LH	LIRKK			159				
e4e.1.ALH	LIBKK			182				
or6.1.ALH	LIBKK			159				
sqy.1.A	LIBKK			159				
fyw.1.A	LIBKK			158				
or8.1.ALH	LIBKK			158				
lga. 1 . A 🛄	YKOTN			230				

# The first step in homology modeling is to search for similar sequences

You begin with a **query sequence** (the protein you want to model)

> PrsA2

CGGGGDVVKTDSGDVTKDELYDAMKDKYGSEFVQQLTFEKILGDKYKVSDE DVDKKFNEYKSQYGDQFSAVLTQSGLTEKSFKSQLKYNLLVQKATEANTDT SDKTLKKYYETWQPDITVSHILVADENKAKEVEQKLKDGEKFADLAKEYST DTATKDNGGQLAPFGPGKMDPAFEKAAYALKNKGDISAPVKTQYGYHIIQM DKPATKTTFEKDKKAVKASYLESQLTTENMQKTLKKEYKDANVKVEDKDLK DAFKDFDGSSSSDSDSSK

# Sequence alignment maps residues from the target onto the template structure



This is an MSA of *Listeria monocytogenes* PrsA2 to related proteins

# Basic alignment algorithms are too simplistic for distant homolog detection

# Methods like **Smith-Waterman** use direct pairwise alignment based on similarity scores (e.g., BLOSUM62).

They do not consider evolutionary variation, insertions, or residuelevel probabilities.

		1064		1074	1084	
PrsA2/1-271	 FADLA	KEYST		DN	GGOLAP	
IniRef100_A0A1/1-271	 FAKLA	KSDSC		PN	GGKAEP-	
IniRef100_A0A0/1-271	 FAKLA	Α <mark>κ</mark> κ <mark>γ</mark> ςτ	Ο Τ Α <mark>Τ Κ</mark>	NK	GGKLPA-	
IniRef100_A0A0/1-301	 FAKLA	A <mark>K E Y S</mark> T	DPG <mark>S</mark> K	D K	GGDLGF-	· - ·
IniRef100_A0A0/1-299	 FAALA	A <mark>k e y s</mark> e	ED PG <mark>S</mark> K	D <mark>N</mark>	GG L <mark>Y</mark> T	· - ·
IniRef100_A0A0/1-339	 FAKLA	A K H S C	DTGSK	D K	<mark>GGDL</mark> G F -	· - ·
IniRef100_A0A0/1-283	 FAALA	A <mark>k e y s</mark> e	D PG <mark>S</mark> K	D <mark>N</mark>	GGLYED-	
IniRef100_A0A7/1-256	 FADVA	A KEL SO	D T Q <mark>T K</mark>	D <mark>N</mark>	<mark>GGDLG</mark> F -	
IniRef100_A0A1/1-260	 FAKLA	A <mark>d e y</mark> s e	<mark>ed p</mark> gn v	DKGKKK	( <mark>GGDLG</mark> <mark>W</mark> -	· - ·
IniRef100_A0A0/1-301	 FAELA	A <mark>k e y s</mark> d	DEGSK	D <mark>N</mark>	GGDLG Y -	· - ·
IniRef100_A0A0/1-355	 FAELA	A <mark>k</mark> kn <mark>s</mark> c	D PG S A	A <mark>N</mark>	<mark>GGDLG</mark> F-	· - ·
IniRef100_A0A3/1-281	 FAALA	A <mark>K E H S</mark> L	. DIE <mark>S</mark> A	PL	<mark>GGDL</mark> N W -	· - ·
IniRef100_A0A0/1-291_	 FAELA	A <mark>k</mark> k <mark>ys</mark> e	D - V <mark>S</mark> A	SS	<mark>GGDLG</mark> F -	· - ·
Conservation	u		l de		La l	
	 **39*	342*4	0384	43	**3530-	
Quality					di .	
Consensus					<b>L</b> ,	
	 FAKLA	AKEYSÇ	) D + G S K	DNGKKK	GGDLG F -	
Occupancy						

We need methods that detect **evolutionarily distant but structurally conserved** relationships.

# Profile-based aligners improve sensitivity by modeling residue variability at each position

A **profile** captures how conserved each position is across an MSA



Instead of a single residue, each position becomes a **probability distribution** over all 20 amino acids.

# Profile-based aligners improve sensitivity by modeling residue variability at each position

A **profile** captures how conserved each position is across an MSA



Instead of a single residue, each position becomes a **probability distribution** over all 20 amino acids.

# HHblits starts by converting the query sequence or MSA into a profile HMM

A **profile HMM** models the amino acid probabilities at each position, plus insertion and deletion likelihoods.

Sequence likelihood can be computed by walking along the profile HMM

The result is a probabilistic model that captures both **conservation** and **structural variability**.

#### Multiple sequence alignment

equence 1:	F	K	L	L	s	H	C	L	L	v
equence 2:	F	K	A	F	G	Q	т	М	F	Q
equence 3:	Y	P	I	v	G	Q	E	ь	Г	G
equence 4:	F	P	v	v	К	E	A	I	г	K
equence 5:	F	K	v	L	A	A	v	I	A	D
lequence 6:	L	Е	F	I	S	Е	С	I	I	Q
equence 7:	F	K	L	L	G	N	v	г	v	с



# HHblits performs accurate HMM-HMM alignments on the best candidates

Full alignments are done using the **Viterbi algorithm** to find the best path through the HMM state space.

A **maximum accuracy (MAC)** alignment is also computed to optimize for correct residue–residue matches.

These alignments return **E-values** to estimate match confidence.

Only statistically significant hits (e.g., E < 1e-3) are retained for the next iteration.

#### The alignment is represented as red path through both HMMs



### Homology modeling is most accurate when sequence identity is >30%

**>50% identity:** high-accuracy models (~1 Å RMSD) are achievable

**Between 30–50%:** moderate accuracy; errors appear in loops, side chains

<**30% identity:** The "twilight zone" where structural similarity is uncertain

Target	CGGGGDVVKTDSGD	VTKDELYDAMKD	КҮ	GSEFVQQLTF	EKI L	GDKY KVSDEDV	/DKKFNEYKSQYG	DQFSAVLTQSGLT	- EKSFKSQLKY	NLLVQKATE
5htf.1.A	CG-SSAVIKTDAGS	VTODELYEAMKT	ΤΥ	GNEVVQQLTF	KKI L	EDKY TVTEKEV	/NAEYKKYEEQYG-	DSFESTLSSNNLT	KTSFKENLEY	NLLVQKATE
6vj6.2.A	DNIVDTKSGS	ISESDENKKLKE	NY	GKQNLSEMVV	EK V L	HDKY KVTDEEV	TKQLEELKDKMG-	DNFNTYMESNGVK	NEDQLKEKLKL	TFAFEKAIK
5tvl.1.0	ADLISKGDV	ITEHOFYEOVKN	NPS	AQQVLLNMTI	QKVFEK	QY)GSELD(DKEV	/DDTIAEEKKQYG-	ENYORVLSQAGMT	-LETRKAQIRT	SKLVELAVK
8qpv.1.F	DCVAAVWNGV	VESDVDGLMQS	VKLNAAQARQQLI	PDDATLRHQIMER <b>L</b> IM	DQIILQMGQ	KM)GV KISDEQI	DQAIANIAKQNNM	IT(LDQMRSRLAYD)GLM	-YNTYRNQIRK	EMIISEVRNNEV
8pzu.1.F	CVAAVWNGV	VESDVDGLMQS	VKENAAQARQQLI	PDDATLRHQIMER <b>L</b> IM	DQI ILQMGQ	K M G V K I S D(E Q I	DQAIANIAKQNNM	ITLDQMRSRLAYDG	-YNTYRNQIRK	EMIISEVRN
8pz2.1.F				QIMERLIN	DQIILQM	GQKMGVKISDEQL	DQAIANIA)KQN <mark>NM</mark>	ITLDQMRSRLAYDG	-YNTYRNQIRK	EMIISEVRN
7efj.1.A										
Target		IKKVVETW	0.P		DITVSHI	VA	n			
5htf 1 4	A NMDVSESK	LKAYYKTW	FP		DITVBHIL	vo.	D			
6vi6 2 A	ATVIEKD	TROHYK	P		KLOVSHIL	U.K.	D			
5tyl 1 0	KVA FAELTDEA	TROUTR	TP		DVTAOTTR	NU CONTRACTOR	N			
800V.1.E	RR. RITTI POF	VESLADOV	GNONDAS	ST.	FINISHTI	TP		AFSOARATVDOARNO	ADEGKLATAHS	ADOOALNGGOMG
8pzu.1.F	-NEVERITIE POE	VESLADOV	GNONDAS	ST.	FLNI SHTL	TP.		AESOARATVDOARNO	ADEGKLATAHS	A DO O ALLNG GOMG
8pz2.1.F	-NEVRRITI	LPOEVESLADOV	GNONDAST		ELNISHTI	IPLPENPTSDOV	IFA			
7efi.1.A	KANSESS	GRVYYENH	TTNASO	WERPS GNSSSGGKNGO	GEPARVRCSHLL	<b>йж</b>	<del>  </del>			
	The second s	and a state of the	*	on so so and a	our method man				-	
larget						<u>E</u>		- NKAKEVEQKLK - DG	- E - KFADLAKE	YSTDIATKDNGG
5ntt.1.A								- ATAKELUTKLK-NG	- E - KFIDLAKE	YSIDIAISINGG
6V]6.2.A						E		-KIAKEIKEKLN-SG	- E - DEAALAKO	YSEDPGSKEKJGG
5tvl.1.0								- DKAKEVLEKAKAEG	- A - DIFAQLAKD	NSTDEKTKENGG
8qpv.1.F	WGRIQELPGIFAQA	LSTAKKGDIVGP	TRSGVGFHILKVI	NULRGESKNISVIEV	ARHILLKPSPIM	TDEQAR		-VKLEQIAADIK-SG		FSUDPGSANUGG
8pzu.1.F	WORLDELPGIFAUA	LSTAKKGUIVGP	THEORETHICK	NULRGESKNISVIEVF	ARHILLKPSPIM	IDEUAR		- VKLEQIAADIK-SG		FSUDPGSANUGG
6pz2.1.F						COCODOC		- SUARAIVDUAR - NG	- A- DFGKLAIA	HSADQUAL - NGG
/etj.1.A							WRUEKIIKINEEA	LELINGTIQKIN- 50	EE-D(FESLASU	DSDUS-SANAKG
Target	QLAPFGPG - KMD	PAFEKAAYALKN	KGDISAPVKT	QYGYHIIQMDK	PATKTTFEKDK	KAVKASYLESQL	TT ENMQKTLKKE	YKDANVKVEDKDLKD	AFKDFDGSSSS	DSDSSK
5htf.1.A	LDPFGPG - EMD	ETFEKAAYALEN	KDDDSGIVKS	TYGYHLIQLVK	KTEKGT YAKEK	ANVKAAYIKSQLT	IS ENMTAALKKE	LKAANID DKDSDLKD	AFADYTSTSST	SS
6vj6.2.A	DLSEDGPG-MMV	KEFEDAAYKLE-	VGQDSEPVKS	SFGYHIIKLTD	- KELKPYEBEK	ENIRKELEQORIO	)) D(P - Q F H <b>Q</b> Q V T R D L	LKNADIKVSDKDLKD	TFKEL	
5tvl.1.C	EIT DSASTEVP	EQVKKAAFALD-	V D G <b>V S</b> D V I D A T G <sup>-</sup>	TQAYSSOYYIVK	- TEKSSN(IDDY <b>K</b>	EKL <b>K</b> TVI <b>L</b> TQ-KO	ND(STFVQSIIGKE	LQAANIKVKDQAFQA	IFTQYIGGGDS	S S
8qpv.1.F	DLGWATPD-IFD	PAFRDALTRLN	KGQMSAPVHS	SFGWHLIELLDT	RNVDKTD - AA <mark>QK</mark>	DRAYRMLM - NRK	SEEAASWMQEQ	RASAYVKDLSNXXX	X	
8pzu.1.F	DLGWATPD-IFD	PAFRDALTRLN-	KGQMSAPVHS	SFGWHLIELLDT	RNVDKTD-AAQK	DRAYRML - MNRK	SEEAASWMQEQ	RASAYWKILS		
8pz2.1.F	QMGWG-RIQEL	PGIFAQALSTAK	KGDIVGPIRS	GVGFHILKVN						
7efj.1.A	DLGAESRG-QMQ	KPFEDASFALR-	TGESGPVET	DSGIHIILBT						

After today, you should have a better understanding of

#### Homology modeling

**Template building** 

The model is built by copying the template structure and modeling variable regions

**Conserved regions**: Backbone atoms are copied from template directly.

**Variable regions** (loops, inserts): Built using fragment libraries or loop modeling algorithms.

**Side chains** are adjusted using rotamer libraries to fit target sequence.



# Model refinement improves geometry and resolves steric clashes

After model construction, the structure often contains **bad bond angles**, **clashes**, or **unrealistic torsions**.

Refinement includes **energy minimization** using force fields or statistical potentials.

Some tools use **molecular dynamics** or **Monte Carlo sampling**.



# Model validation helps assess confidence and detect errors

# **Ramachandran plots** visualize backbone torsion angles.

**Statistical scores** (e.g., DOPE, QMEAN, GA341) evaluate nativeness.

**Residue-by-residue assessment** helps identify weak regions (e.g., VERIFY3D, ERRAT).

Good models have:

- Most residues in favored Ramachandran regions
- Low-energy scores
- No large clashes





# Homology modeling works best when you iterate and re-evaluate

If a model fails validation, revisit earlier steps:

- Try a different template
- Refine the alignment
- Adjust loop modeling parameters

**Multiple models** are often built and ranked—choose the one with the best validation metrics.

# After today, you should have a better understanding of



Know when to use threading instead of homology modeling

### Why Use Threading?

In cases where sequence similarity to known structures is low (< 30%), homology modeling becomes unreliable

Threading **matches sequences to known structural folds** based on structural rather than sequence similarity

**Phyre2**, **RaptorX**, **MUSTER**, and **I-TASSER** are commonly used for threading and takes much longer than homology modeling

# Identifying the Right Fold



# After today, you should have a better understanding of



#### Interpret a contact map for protein structures

# Contact Maps Visualize Residue Interactions in Proteins

A contact map is a 2D representation of which residues are in close proximity

Each point on the map corresponds to two residues that are close in 3D space

mapiya.lcbio.pl



# Contact Maps Represent Spatial Proximity, Not Sequence Order

Contacts are determined by spatial proximity, typically within a certain distance threshold



Residues far apart in the sequence can still be close in the 3D structure, reflected in the contact map



# Residues on the diagonal are adjacent in sequence (and spatially)



33

# After today, you should have a better understanding of

#### Comprehend how coevolution provides structural insights

# After today, you should be able to

Comprehend how coevolution provides structural insights

# The Rise of Machine Learning in Structural Biology

Traditional methods like **homology modeling** and threading rely on **templates and known structures** 

ML predicts 3D structures only from sequence data

**AlphaFold** (DeepMind) and **RosettaFold** (Baker Lab) lead the charge in this area

# What is AlphaFold?

Developed by DeepMind, **AlphaFold** predicts protein structures with atomic accuracy by using deep learning models trained on large structural datasets

#### Breakthroughs

- AlphaFold 2 achieved near-experimental level accuracy in the 2020 **CASP14** competition (Critical Assessment of protein Structure Prediction)
- **AlphaFold 3** (2024) predicts proteins, DNA, RNA, ligands, and posttranslational modifications

### Coevolving residues mutate in a correlated manner

Mutations in one residue often result in **compensatory mutations** in its interacting partner

This is observed across species through **analysis of homologous protein sequences** 

#### **Correlated mutations** indicate **functionally significant** residue pairs

**Evolution** 

	$\bigcirc \bigcirc $	с
Arg (positive)	A T R L T L T A K K D G P C D A T R L T L T A K K D G P C D A T R L T L T A K K D G P C D	Asp (negative)
Lys (positive)	$\begin{array}{c} \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{L} \ \mathbf{C} \ \mathbf{L} \ \mathbf{T} \ \mathbf{A} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{P} \ \mathbf{K} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{T} \ \mathbf{A} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{P} \ \mathbf{K} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{T} \ \mathbf{A} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{P} \ \mathbf{K} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{K} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{P} \ \mathbf{K} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{P} \ \mathbf{K} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{C} \ \mathbf{C} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{C} \ \mathbf{C} \ \mathbf{D} \\ \mathbf{K} \ $	Glu (negative)
Trp (hydrophobic)	$\begin{array}{c} \mathbf{A} \ \mathbf{T} \ \mathbf{K} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{G} \ \mathbf{G} \ \mathbf{K} \ \mathbf{K} \ \mathbf{E} \ \mathbf{G} \ \mathbf{G} \ \mathbf{C} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{W} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{T} \ \mathbf{A} \ \mathbf{K} \ \mathbf{K} \ \mathbf{V} \ \mathbf{G} \ \mathbf{P} \ \mathbf{C} \ \mathbf{D} \\ \mathbf{A} \ \mathbf{T} \ \mathbf{W} \ \mathbf{L} \ \mathbf{T} \ \mathbf{L} \ \mathbf{T} \ \mathbf{A} \ \mathbf{K} \ \mathbf{K} \ \mathbf{V} \ \mathbf{G} \ \mathbf{P} \ \mathbf{C} \ \mathbf{D} \\ \end{array}$	Val (hydrophobic

### **Evolutionary Analysis Reveals Structural Insights**

Coevolution analysis helps predict which residues are close in the 3D structure

Residues showing correlated mutations are likely to be spatially close in the folded protein

This is particularly useful when no experimental structure is available



# Multiple Sequence Alignments Enable Coevolution Detection

Coevolution is detected using large MSAs from homologous proteins

The more diverse the sequences in the MSA, the better the resolution of coevolving residues

Evolutionary information from MSAs guides predictions for residue-residue contacts

Ξ	27659 sequences, 159 columns		Show	/ Hide 🗸	Sort ∨	Group $\lor$	Filter ∨
	ID 🌲	• 46 •	50	. 9		20	
	Reference TARGET/1-159	(	ESIG <mark>K</mark>	PLP <mark>NRR</mark> N	V V L T S D T S	FNVEGV	
	Consensus Sequence Logo	<mark>ĸĸぃт</mark> манрутмаккт К₿┶Т₩Ġ₭₽¥IJŇĠ₽КТ₩	ESIGR			YQAEGA	
	Coverage	. 85912855470VD81X	. ₩ ₩ ₩ ₩ ₩	5 L 5 R 1 5 H	Y I Q S H H H		
	Conservation	had a second block		din La	distant.	= = =	
1	TARGET/1-159	<pre>/KKLSTGHTLVMGRKTF</pre>	ESIG <mark>K</mark>	PLP <mark>NRR</mark> N	VVLTSDTS	FNVEGV	
2	UniRef90_A0A2J8ADC8/2-178	R S I T A G G G V I M G R T T F	DSIPR	PLQGRLN	VVLTTSAD	DLMNSNI	
3	UniRef90_A0A2J8ADC8/438-608	<b>R</b> S I T A G G G V I M G <mark>R K</mark> T <mark>F</mark>	DSIPR	PLKGRLN	V V L T <mark>R</mark> S S A	DLDPNI	
4	UniRef90_A0A2J8ADC8/906-1076	- <mark>R</mark> S I T A G G G V I M G <mark>R K</mark> T <mark>F</mark>	DSIPR	PLKGRLN	V V L T <mark>R</mark> S S A	A D L D P N I	
5	UniRef90_A0A2U1P121/436-611	<b>K K L T M S N A V I M G R K T W</b>	QSIP R	PLPDRLN	V V L T <mark>R</mark> S T I	FDAENT	
6	UniRef90_A0A2U1P121/715-890	• K K L T M S N A V I M G R K T W	Q S I P <mark>R</mark>	P L P G <mark>R</mark> L N	V V L T <mark>R</mark> S T A	FDSDNV	
7	UniRef90_A0A6J8E626/8-185	K K I T M E N V V I M G R K T W	IFSIPR	PLPKRIN	IILSREM	EAPSGV	
8	UniRef90_A0A6J8E626/194-373	<b>' T Q K </b> C S P T V V I <b>K</b> G <b>R</b> M T W	ECTKR	TNPGYLN	VIIS <mark>H</mark> SKF	O Q D E Y V	目的開始
9	UniRef90_UPI0022B18839/37-217	<b>K K M T T Q N </b> V V I M G <b>R K T W</b>	NSIPR	PLPKRIN	IILS <mark>K</mark> TMS	H A P T G A	
10	UniRef90_UPI0022B18839/246-424	' M E L T S <mark>R</mark> C V N I Q G <mark>R K</mark> T W	EGTGK	QRQSVYN	IVI <mark>T</mark> RDE(	R R D P D V	
11	UniRef90_UPI001457FF82/6-184	K R I T T E N V V I M G R K T W	VSIPR	PLPRRIN	IILSRTM	IETPTGT	
12	UniRef90_UPI001457FF82/192-372	' M D L T S <mark>R</mark> C V N I K G R V T W	Q C T C K	ARGSIIN	IVIS <mark>RN</mark> PS	EEDPYV	
13	UniRef90_UPI00234EC90F/4-183	<b>K K</b> I T S <b>D N</b> A V I M G <b>R K</b> T W	VSIPR	PLKGRVN	IVLSREL	EVPEGV	
14	UniRef90_UPI00234EC90F/225-406	' S A <b>H</b> V S T I I <b>Q</b> I <b>R</b> G <mark>R</mark> L T W	L S A M <mark>R</mark>	ΝΑΡΝΥΥΤ	IIVSGTW	ERDPRV	
15	UniRef90_A0A8B7ZNQ9/10-182	' L N C V <mark>N K N</mark> A M I V G <mark>R</mark> L T -	ESISE	R K P H K L F	FVLSKTL	ELPPKA	
16	UniRef90_A0A8B7ZNQ9/197-376	S	LSIPR	PLPNRVN	V V L S <mark>K</mark> T L S	SECPADA	
17	UniRef90_R7UK12/6-185	RKITSEN VVLMGRKTW	ESIP R	PLPNRIN	V V L S A S L I	EAPQGS	
18	UniRef90_R7UK12/195-373	' S Q L T Q G V V N I K G R A T W	QDTGK	ARPNVIT	IIISKTLS	QLPEGA	
19	UniRef90_T1G9P0/5-182	RKLTSENAILMGRKTW	DSIPK	PLKNRLN	V V I S <mark>R</mark> T L E	C P D G R L	
20	UniRef90_T1G9P0/198-374		E S M K R	HIEGAVY	I V V G S <mark>K K</mark> H	ILLSYPD	
21	UniRef90_A0A7J7JG58/7-183	<b>R K L T T <mark>K N</mark> A V I M G <mark>R K T Y</mark></b>	(FSIPR	PLKNRVN	I V L S <mark>R</mark> A S 1	LDIESV	
22	UniRef90_A0A7J7JG58/191-368	<sup>•</sup> N S V I S P V C L I E G <mark>R</mark> L S Y	QEAIV	DKPGFIT	V V L S S D P S	S R V P S P H	
23	UniRef90_A0A210PN44/7-170	K R I T T E N V L I M G R K T W	T S I P <mark>R</mark>	PLP <mark>KR</mark> IN	IILSRTMI	TETPTGT	
24	UniRef90_A0A210PN44/180-361	M D L T S K C V N I K G R V T W	Q C T C K	SRDSIIN	IVIS <mark>R</mark> NPS	E E D P Y V	
25	UniRef90_A0A9N7VKC0/13-190	L	IFSHPF	PLA <mark>NT</mark> LH	V V L S T <mark>K</mark> L I	(KVPDHA	
26	UniRef90_A0A9N7VKC0/214-371	L N T V T R N M M V W G K L C W	IFSHPF	PLANILH	V V L <mark>N T K</mark> L N	EVPDHA	
27	UniRef90_A0A8B6BLZ4/22-159	<mark>K K N</mark> V V I M G <b>R K</b> T <mark>W</mark>	VSIPR	PLPKRIN	IVLSREM	EAPPGV	
28	UniRef90_A0A8B6BLZ4/167-346	T	ESTKR	THPGYLN	V I I S <mark>H</mark> S <mark>K</mark> F	DQLESD	
29	UniRef90_A0A9J2PNF8/135-309	. E Y L T T K N A V L M G R K V W	ESYPR	PLEDRLN	VVLSETM	D P A E S F	
30	UniRef90_A0A940DWL2/2-156	• K K V T <mark>Y</mark> G H P V V M G R K T Y	ESIGK	PLPG <mark>REN</mark>	IVVTRNKE	Y Q P E G V	
31	UniRef90_H8KUH0/6-166	KKLTTGNTIIMGRKTY	DSIGR	P L P <mark>N R R N</mark>	VIISRNK	D L K I E G C	
32	UniRef90_A0A2D5XDZ3/1-161	KKLTLG <mark>K</mark> PIIMG <mark>RK</mark> T <mark>Y</mark>	ETIGK	PLPNRKN	IIITRDQD	YKAEGC	

#### evcouplings.org

### **Coevolution example: DHFR**

Residues with a high Score (i.e., coevolve) are near each other in the protein's structure (i.e., small distance)



# Coevolutionary signals can be noisy

Not all correlated mutations are due to direct physical interactions; some may be indirect

Noise in the data can come from random mutations or insufficient evolutionary diversity.

Large and diverse sequence data sets are needed for reliable coevolution predictions.

Ξ	23609 sequences, 190 columns		Show / Hide $\lor$	Sort ∨	Group $\lor$ Filter $\lor$
	ID 🌲	. 8 . 8	- 9	. 03	
	Reference				
	TARGET/1-190	<b>i f</b> valntlvvavyf <b>i</b> etad	1 e <b>q</b> v v <mark>y</mark> g <mark>k n n 1 n</mark>	<b>d k</b> T 1 <b>d</b> 1 <mark>k d</mark> g t	ygtepL <u>zeigezi</u>
	Consensus				· · · · · · · · · · · · · · · · · · ·
	Coguenes Lega				
	Sequence Logo				
	Coverage				
	Conservation				<b>3</b>
1	TARGET/1-190	ifvalntlvvavyf <mark>r</mark> etac	deqvv <mark>y</mark> g <mark>k</mark> nnin	<b>q k</b> l i <b>q</b> l <mark>k d</mark> g t	ygfep 👬
2	UniRef90_UPI0005CC9E4F/90-233		vv <mark>eqld</mark> lp	etieqlne <mark>k</mark> i	tvnse
3	UniRef90_UPI00203FAC45/90-235		v s <mark>k</mark> e l d l n	mtaaql <mark>n</mark> g <mark>k</mark> i	tvqse
4	UniRef90_A0A023CJ72/79-235	vii <mark>k</mark> spti	ile <mark>k</mark> v <mark>kk</mark> elnld	<b>r</b> tidelneqi	qvsse 👘
5	UniRef90_A0A1Q5SWQ9/132-273	ii <mark>k</mark> spti	ile <mark>k</mark> v <mark>k</mark> del <mark>h</mark> ld	<b>r</b> tldelneqi	qvsse 👘 👘
6	UniRef90_A0A1I0TB73/79-231	vii <mark>k</mark> spti	ile <mark>k</mark> v <mark>k</mark> eelnld	<b>r</b> s v d e l n e q i	qvsse 🚔 👘
7	UniRef90_A0A3A1R1Y0/89-237		<mark>k</mark> vsqnldld	m t s s q l n g <mark>k</mark> i	tvgse
8	UniRef90_A0A1W6VWN6/80-223	iikspti	le <mark>k</mark> v <mark>k</mark> delqld	qtldelneqi	qvsse 👘
9	UniRef90 A0A7M1TW98/79-227	ikspti	ile <mark>k</mark> vknelnld	qtldelneqi	q v s s e
10	UniRef90 A0A150M3I9/80-233	iikspti	ile <mark>k</mark> v <b>r</b> eelnld	<b>r</b> tvdelneqi	qvtse 👘
11	UniRef90 UPI00077CAD76/80-232	iikspti	il <b>dq</b> vi <b>k</b> gldld	m t t s q l n e k i	tvgse 👘
12	UniRef90_UPI0007D0AEB9/85-236		ile <mark>k</mark> vvseldln	mttsdlng <mark>k</mark> i	tvgse 👘 👘
13	UniRef90 UPI001FAEEC2B/87-237		ld <mark>k</mark> vid <b>r</b> mnid	e s v e s l n e <mark>k</mark> i	tvnse
14	UniRef90 A0A0D8BY29/84-225		ile <mark>k</mark> vknelrld	<b>r</b> tldelneqi	qvsse
15	UniRef90 UPI001BCA202F/90-231		vikeldld	mtaqqlnq <mark>k</mark> i	tvqse
16	UniRef90 A0A7Y5AXW0/90-223		v <mark>kntldld</mark>	mtteqlnski	tvasa 👘
17	UniRef90 A4ITF4/83-224		ile <mark>k</mark> vkgel <b>g</b> ld	qtldelneqi	qvsse a
18	UniRef90 A0A428JAA2/90-231		v <mark>k</mark> eeldln	m s t t e l n s <mark>k</mark> l	tvqse 🚽
19	UniRef90 A0A327YHM8/82-234		ile <mark>k</mark> v <mark>q</mark> eeleie	<b>r</b> sieqlneqi	qvase 🕺 🕅
20	UniRef90 UPI002148361B/71-232	vemintyviikspai	ilegviqeldln	qnvdqlnesi	s v n s e
21	UniRef90 UPI0025A04A00/71-227	vqmvntyviikspai	ils <mark>k</mark> vidnldln	ttvealns <mark>k</mark> l	svnse
22	UniRef90 A0A9E8FEN0/69-230	.qtnvqmvntyviikspai	ils <mark>k</mark> vi <mark>qh</mark> ldlk	q s v a d l n n q l	tinse
23	UniRef90 A0A366XUS2/88-219		.e <mark>k</mark> videldln	k t a s q l n s q i	svsav 👘
24	UniRef90 A0A165WR30/89-233		lvieelgln	etvdelne <mark>k</mark> i	tvgse 🔤 👘
25	UniRef90 A0A6M6DVP5/82-231		ld <mark>k</mark> vkselnld	<b>r</b> tvedlnsqi	tvssa
26	UniRef90 UPI001CBAC555/85-235		ild <mark>k</mark> viqqldln	m t s g q l n e n v	avese
27	UniRef90_A0A6I2MA70/82-234		ild <mark>k</mark> viq <b>r</b> ldln	m t s g q l n e n v	avase 👘
28	UniRef90_UPI001AEDD1CF/90-234		vi <mark>k</mark> emnld	mtaaqlne <mark>k</mark> i	tvgse -
29	UniRef90 A0A1B1Z946/88-232		<mark>k</mark> via <b>q</b> m <mark>nld</mark>	ttvgqineqi	tvssk
30	UniRef90_UPI0006A7EC86/90-231		vikqldle	qtfqqlneki	Tvase
31	UniRef90 UPI00207AA881/82-236		ild <mark>k</mark> vieqldld	m s s g q l n a n v	tvgse
32	UniRef90_UPI001C1F2282/88-232		.d <mark>k</mark> viaeldln	<b>r</b> s <mark>y</mark> g a l n s s v	tvgsa
33	UniRef90_UPI00158DFFD5/87-235		le <mark>k</mark> visdldld	itasqlnekl	tvase
34	UniRef90 UPI001CCEBACC/99-230			. n f s q l n g k i	tvatk

# Machine learning leverages coevolution for high-accuracy predictions

AlphaFold and RosettaFold utilize coevolutionary data from MSAs to predict residue interactions

These models incorporate evolutionary information along with structural features, leading to highly accurate predictions



# After today, you should have a better understanding of

#### Explain why ML models are dominate

protein structure prediction

### AlphaFold pipeline, simplified

#### Given the following data

#### Input sequence

MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKK LSTGHTLVMGRKTFESIGKPLPNRRNVVLTSDTS FNVEGVDVIHSIEDIYQLPGHVFIFGGQTLFEEM IDKVDDMYITVIEGKFRGDTFFPPYTFEDWEVAS SVEGKLDEKNTIPHTFLHLIRKK

#### Multiple Sequence Alignment

1	TARGET/1-159	<pre>/KKLSTGHTLVMGRKTFESIGKPLPNRRNVVLTS</pre>	DTSFNV
2	UniRef90_A0A2J8ADC8/2-178	R S I T A G G G V I M G R T T F D S I P R P L Q G R L N V V L T T	SADLMN
3	UniRef90_A0A2J8ADC8/438-608	<b>R</b> S I T A G G G V I M G <b>R K</b> T <b>F</b> D S I P <b>R</b> P L <b>K</b> G <b>R</b> L <b>N</b> V V L T <b>R</b>	SSADLD
4	UniRef90_A0A2J8ADC8/906-1076	R S I T A G G G V I M G R K T F D S I P R P L K G R L N V V L T R	SSADLD
5	UniRef90_A0A2U1P121/436-611	K K L T M S N A V I M G R K T W Q S I P R P L P D R L N V V L T R	STNFDA
6	UniRef90_A0A2U1P121/715-890	K K L T M S N A V I M G R K T W Q S I P R P L P G R L N V V L T R	STAFDS
7	UniRef90_A0A6J8E626/8-185	KKITMENVVIMGRKTWFSIPRPLPKRINIILSR	EMKEAP
8	UniRef90_A0A6J8E626/194-373	T Q K C S P T V V I K G R M T WE C T K R T N P G Y L N V I I S H	SKRDQD
9	UniRef90_UPI0022B18839/37-217	K K M T T O N V V I M G R K T W N S I P R P L P K R I N I I L S K	TMSHAP
10	UniRef90_UPI0022B18839/246-424	MELTSRCVNIQGRKTWEGTGKQRQSVYNIVITR	DEGRRD
11	UniRef90 UPI001457FF82/6-184	KRITTENVVINGRKTWVSIPRPLPRRINIILSR	TMNETP
12	UniRef90 UPI001457FF82/192-372	MDLTSRCVNIK GRVTWQCTCKARGSIINIVISR	NPSEED
13	UniRef90_UPI00234EC90F/4-183	K K I T S D N A V I M G R K T W V S I P R P L K G R V N I V L S R	ELKEVP
14	UniRef90_UPI00234EC90F/225-406	SAHVSTIIQIRGRLTWLSAMRNAPNVYTIIVSG	TWTERD
15	UniRef90 A0A8B7ZNQ9/10-182	LNCVNKNAMIVGRLT-ESISERKPHKLFFVLSK	TLKELP
16	UniRef90_A0A8B7ZNQ9/197-376	S A Q T S K N A V I M G R K T W L S I P R P L P N R V N V V L S K	TLSECP
17	UniRef90_R7UK12/6-185	R K I T S E N V V L M G R K T W E S I P R P L P N R I N V V L S A	SLKEAP
18	UniRef90 R7UK12/195-373	S Q L T Q G V V N I K G R A T W Q D T G K A R P N V I T I I I S K	TLSQLP
19	UniRef90_T1G9P0/5-182	RK LTSENAILMGRKTWDSIPKPLKNRLNVVISR	TLECPD
20	UniRef90_T1G9P0/198-374	ISLLOGTAIIVGRLTWESMKRHIEGAVYIVVGS	KKHLLS
21	UniRef90 A0A7J7JG58/7-183	RKLTTKNAVIMGRKTYFSIPRPLKNRVNIVLSR	ASTLDI
22	UniRef90 A0A7J7JG58/191-368	N S V I S P V C L I E G R L S Y Q E A I V D K P G F I T V V L S S	DPSRVP
23	UniRef90_A0A210PN44/7-170	KRITTENVLIMGRKTWTSIPRPLPKRINIILSR	ТИТЕТР
24	UniRef90_A0A210PN44/180-361	M D L T S K C V N I K G R V T W Q C T C K S R D S I I N I V I S R	NPSEED
25	UniRef90 A0A9N7VKC0/13-190	L K T V T R N M M V W G K L C W F S H P F P L A N T L H V V L S T	KLKKVP
26	UniRef90 A0A9N7VKC0/214-371	LNTVTRNMMVWGKLCWFSHPFPLANILHVVLNT	KLNEVP
27	UniRef90_A0A8B6BLZ4/22-159	KKNVVIMGRKTWVSIPRPLPKRINIVLSR	ЕМКЕАР
28	UniRef90_A0A8B6BLZ4/167-346	TOKCSPTVVIKGRMTWESTKRTHPGYLNVIISH	SKRDQL
29	UniRef90_A0A9J2PNF8/135-309	EYLTTKNAVLMGRKVWESYPRPLEDRLNVVLSE	TMDDPA
30	UniRef90_A0A940DWL2/2-156	K K V T Y G H P V V M G R K T Y E S I G K P L P G R E N I V V T R	NKEYQP
31	UniRef90_H8KUH0/6-166	KKLTTGNTIIMGRKTYDSIGRPLPNRRNVIISR	NKDLKI
32	UniRef90_A0A2D5XDZ3/1-161	KKLTLGKPIIMGRKT <mark>Y</mark> ETIGKPLPNRKNIII <b>T</b> R	D Q D <mark>Y K</mark> A







Atomistic structure

### AlphaFold 2 pipeline: Evoformer

Using MSAs and contact maps, DeepMind trained a model to predict protein structures



### Contact maps are converted into dihedral angles



#### **AF2 iterations**



Recycling iteration 0, block 01 Secondary structure assigned from the final prediction

### What is new in AlphaFold 3?

Biggest change is the use of a **diffusion model** 

Diffusion models essentially learn to **unscramble atoms into a structure** 



# AlphaFold 3 is supercharged for any biomolecule

Proteins, DNA, RNA, ligands, PTMs, protein-proteins, etc.



b



### AlphaFold 3



MTLSILVAHDLQRVIGFENQLPWHLPNDLKHVKKLSTGHTL VMGRKTFESIGKPLPNRRNVVLTSDTSFNVEGVDVIHSIED IYQLPGHVFIFGGQTLFEEMIDKVDDMYITVIEGKFRGDTF FPPYTFEDWEVASSVEGKLDEKNTIPHTFLHLIRKK

#### DHFR (UniProt)

#### •••

MGKKEVILLFLAVIFVALNTLVVAVYFRETADEQVVYGK NNINQKLIQLKDGTYGFEPALPHVGTFKVLDSNRVPQIA QEIIRNKVKRYLQEAVRIEGTYPIVDGLVNAKYTVANPN NLHGYEGFLFKDNVPLTYPQEFILSNLDGKVRSLQNYDY DLDVLFGEKEEVKSEILRGLYYNTYTRAFSPYKL

> Novel protein (ChatGPT)

# <text>

AlphaFold 3 model is a Google DeepMind and Isomorphic Labs collaboration

#### How does AlphaFold Server work?

AlphaFold Server is a web-service that can generate highly accurate biomolecular structure predictions containing proteins, DNA, RNA, ligands, ions, and also model chemical modifications for proteins and nucleic acids in one platform. It's powered by the newest AlphaFold 3 model.

#### alphafoldserver.com

# AlphaFold 3 is a breakthrough, not the final solution



← Back

← Back	🛃 Download	Clone and reuse	E Feedback on structure	
Very high (pl[	ODT > 90)	Confident (90 > pIDDT > 70)	Low (70 > plDDT > 50)	Very low (pIDDT < 50)
		ipTM = - pTM = 0.2	learn more	



### Caveat: Proteins are dynamic

https://www.youtube.com/embed/AjcUmxT-QEA?si=gupgTpuV5IvOB\_ut&start=43&enablejsapi=1

### What about intrinsically disordered proteins?

At least 40% of proteins have disordered regions

AlphaFold (and all other methods) struggle with disordered regions





### Before the next class, you should

#### Lecture 11A:

Protein structure prediction -Foundations

#### Lecture 11B:

Protein structure prediction -Methodology



• Work on P02B, P02C, and P03A