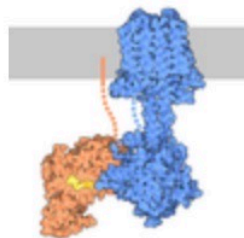
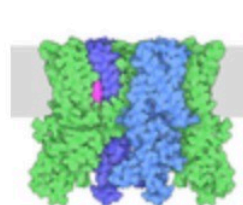




December 2020
Hepatitis C Virus
Protease/Helicase



November 2020
Adenylyl Cyclase

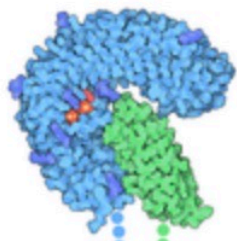


October 2020
Capsaicin Receptor TRPV1

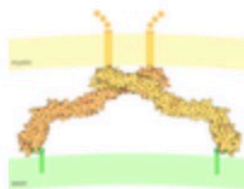


September 2020
SARS-CoV-2 RNA-dependent
RNA Polymerase

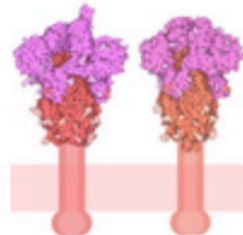
Proteins



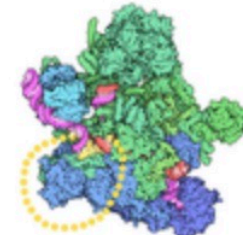
August 2020
Phytosulfokine Receptor



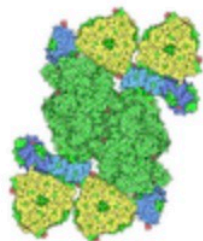
July 2020
Myelin-associated
Glycoprotein



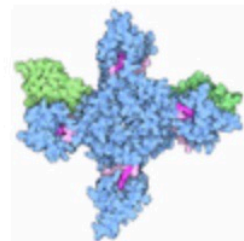
June 2020
SARS-CoV-2 Spike



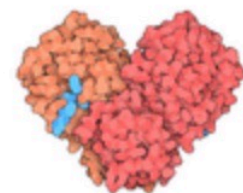
May 2020
Spliceosomes



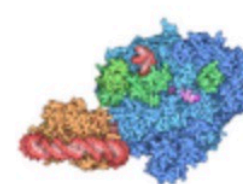
April 2020
Photosynthetic
Supercomplexes



March 2020
Voltage-gated Sodium
Channels

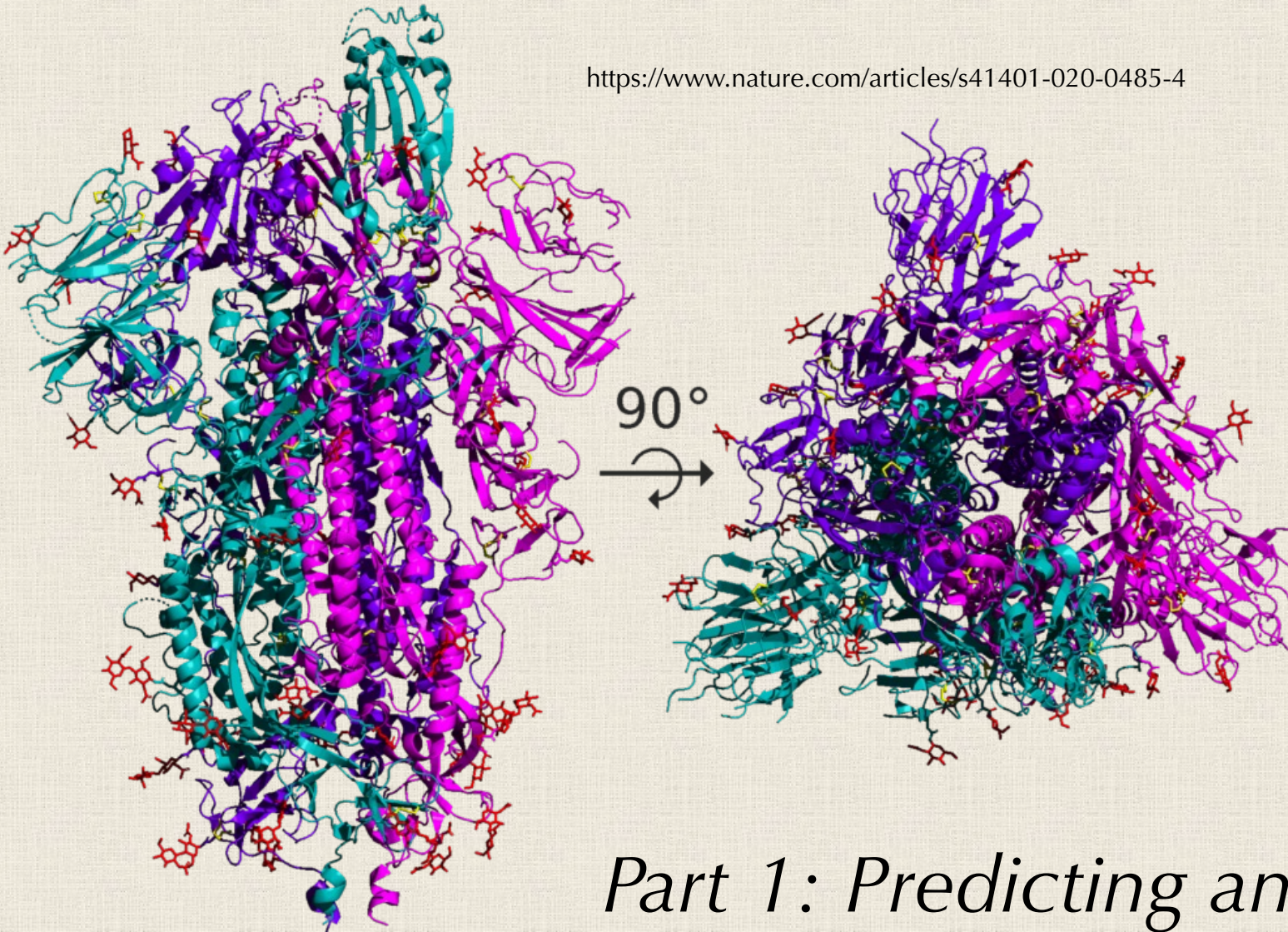


February 2020
Coronavirus Proteases



January 2020
Twenty Years of Molecules

<https://www.nature.com/articles/s41401-020-0485-4>



Part 1: Predicting and Analyzing Protein structures

AN INTRODUCTION TO PROTEIN STRUCTURE PREDICTION

Comparing SARS-CoV and SARS-CoV-2

Recall that after discussing alignment, we aligned the SARS-CoV and SARS-CoV-2 genomes.

Comparing SARS-CoV and SARS-CoV-2

Recall that after discussing alignment, we aligned the SARS-CoV and SARS-CoV-2 genomes.

One of the most critical regions encodes the **spike protein**, which coats the surface of the virus and binds to receptors on the human ACE2 enzyme.

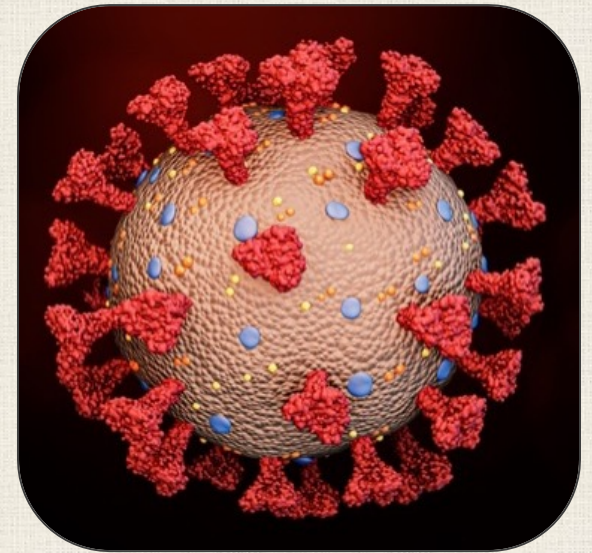


Image Credit: MattLphotography/Shutterstock.com

What exactly does the spike protein do?

**Model of Membrane Fusion by
SARS CoV-2 Spike Protein**

<https://www.youtube.com/watch?v=e2Qi-hAXdJo&t=18s>

Let's align the spike proteins!

SARS-CoV genome has accession ID NC_004718.3.

Spike protein ranges from position 21492 to 25259

SARS-CoV-2 genome accession ID NC_045512.2.

Spike protein ranges from position 21563 to 25384.

Great free tool to translate gene from DNA to protein at <https://web.expasy.org/translate/>.

Let's align the spike proteins!

MFIFLLFLTLTSGSDLRCTTFDDVQAPNYTQHTSSMRGVYYPDEIFRSDTLYLTQDLFLPFYSNVTGFHTINHTFGNPVIPFKDGIYFAATEKSNVVR
GWVFGSTMNKSQSVIIINNSTNVVIRACNFELCDNPFFAVSKPMGTQHTMIFDNAFNCTFEYISDAFSLDVSEKSGNFKHLREFVFNKNDGFLVYVK
GYQPIDVVRDLPSGFNTLKPIFKLPLGINITNFRAILTAFAQAQDIWGTSAAYFVGYLKPTTFMLKYDENGITDAVDCSQNPLAELKCSVKSFEIDK
GIYQTSNFRVVPDGDVVRFPNITNLCPFGEVFNATKFPVYAWERKKISNCVADYSVLNSTFFSTFKCYGVSATKLNLDLCSNVYADSFVVKGDDVRQ
IAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCPPALNCYWPLNDYGFYTTTGIGYQPYR
VVVLSFELLNAPATVCGPKLSTDLIKNQCVNFNFNGLTGTGVLTPSSKRFPQFQFGRDVSDFDTSVRDPKTSEILDITSPCAFGGVSITPGTNASSEV
AVLYQDVNCTDVSTAIHADQLTPAWRIYSTGNVVFQTAGCLIGAHEVDTSYECDIPIGAGICASYHTVSLRSTSQKSIVAYTMSLGADSSIAYSNNT
IAIPTNFSISITTEVMPVSMAKTSVDCNMYICGDSTECANLLQYGSFCTQLNRALSGIAAEQDRNTREVFQVKQMYKTPTLKYFGGFNFSQILPDPL
KPTKRFSIEDLLFNKVTLADAGFMKQYGECLGDINARDLICAQKFNGLTVLPPLLTDMMIAAYTAALVSGTATAGWTFGAGAALQIPFAMQMAYRFNGI
GVTQNVLYENQKQIANQFNKAIKSIQESLTTTSTALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYV
TQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLSFPQAAPHGVVFLHVTYVPSQERNFTTAPAICHEGKAYFPREGVVFVNGTSWFITQR
NFFSPQIITDNTFVSGNCDVVIGIINNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVUNIQKEIDRLNEVAKNLNESLIDLQELGKY
EQYIKWPWYVWLGFIAGLIAIVMVTILLCCMTSCCSCLKGACSCGSCCKFDEDDSEPVKGVKLYHT

SARS-CoV

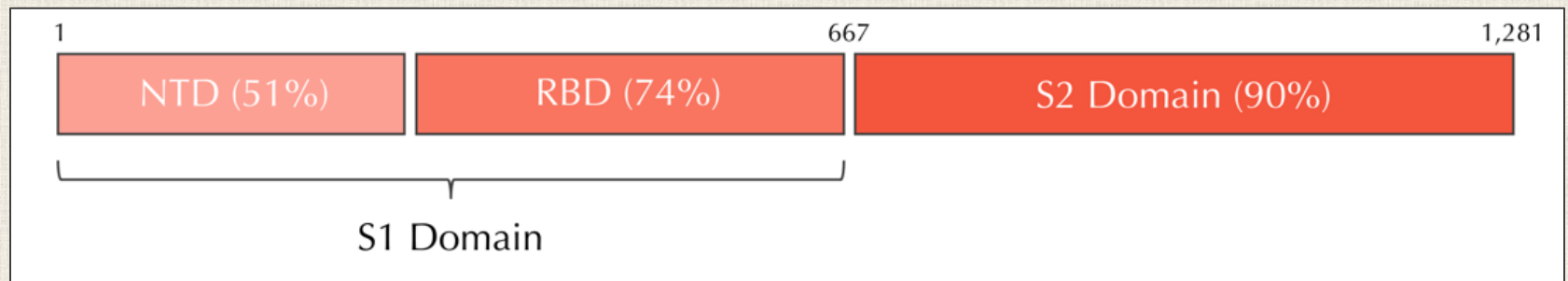
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSN
IIRGWIFGTTLDKSTQSLILVNNATNVVIVKVECFQFCNDPFLGVYHKNKSWMESEFRVYSSANNCTFEYVSQPFLLMDLEGGKQGNFKNLREFVFNKID
GYFKIYSKHTPINLVRDL PQGFSALEPLVDLPIGINITRFQTL LALHRSYLPDSSSSGWTAGAAAYVGYLQPRTFLLKYNENGTITDAVDCALDPLS
ETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLNYSASFSTFKCYGVSPTKLNLDLCTNVY
ADSFVIRGDEVQRAPGQTGKIADYNYKLPDDFTGCVIAWNSNNDLSKVGNYNYLYRLFRKSNLKPFERDISTEYQAGSTPCNGVEGFNCYFPLQSY
GFQPTNGVGYQPYR VVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQFGRDIADTTDAVRDPQTLEILDITPCSF GG
VSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGNSVVFQTRAGCLIGAHEVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSII
AYTMSLGAENSVAYSNNIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECNSLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTP
PIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIARDLICAQKFNGLTVLPPLLTDMMIAQYTSALLAGTITSGWTFGAG
AALQIPFAMQMAYRFNGIGVTQNVLYENQKQIANQFNKAIKSIQESLSSASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAE
VQIDRLITGRLQSLQTYVVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLSFPQSAAPHGVVFLHVTYVPAQEKNTTAPAICHGDKAHF
PREGVFSNGTHWFVTQRNFYEQIITDNTFVSGNCDVVIGIINNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVUNIQKEIDRLNE
VAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVKGVKLYHT

SARS-CoV-2

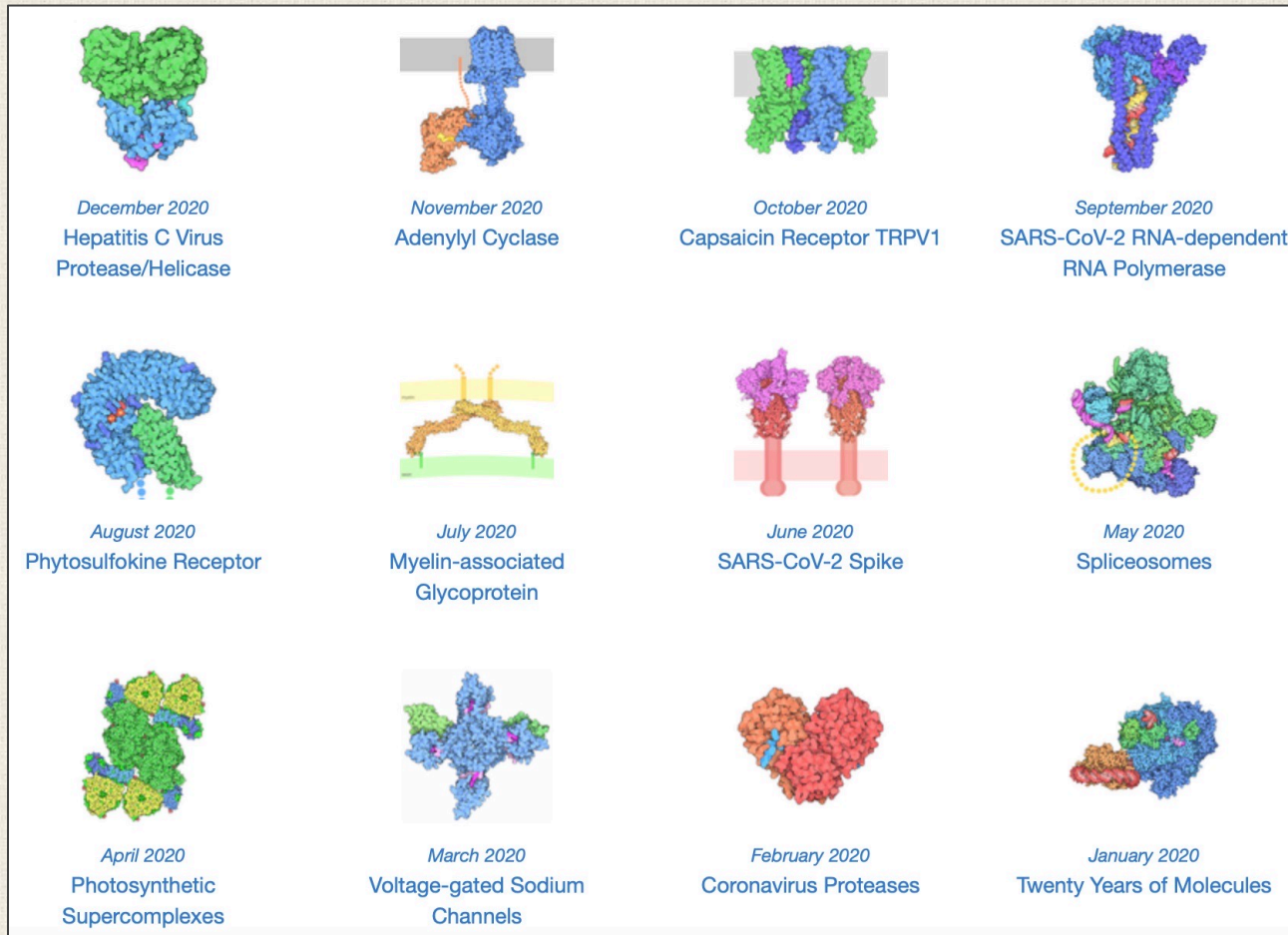
https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle

Let's align the spike proteins!

The spike proteins are *extremely* variable in some regions. These have been primary focus in determining why SARS-CoV-2 was more infectious.

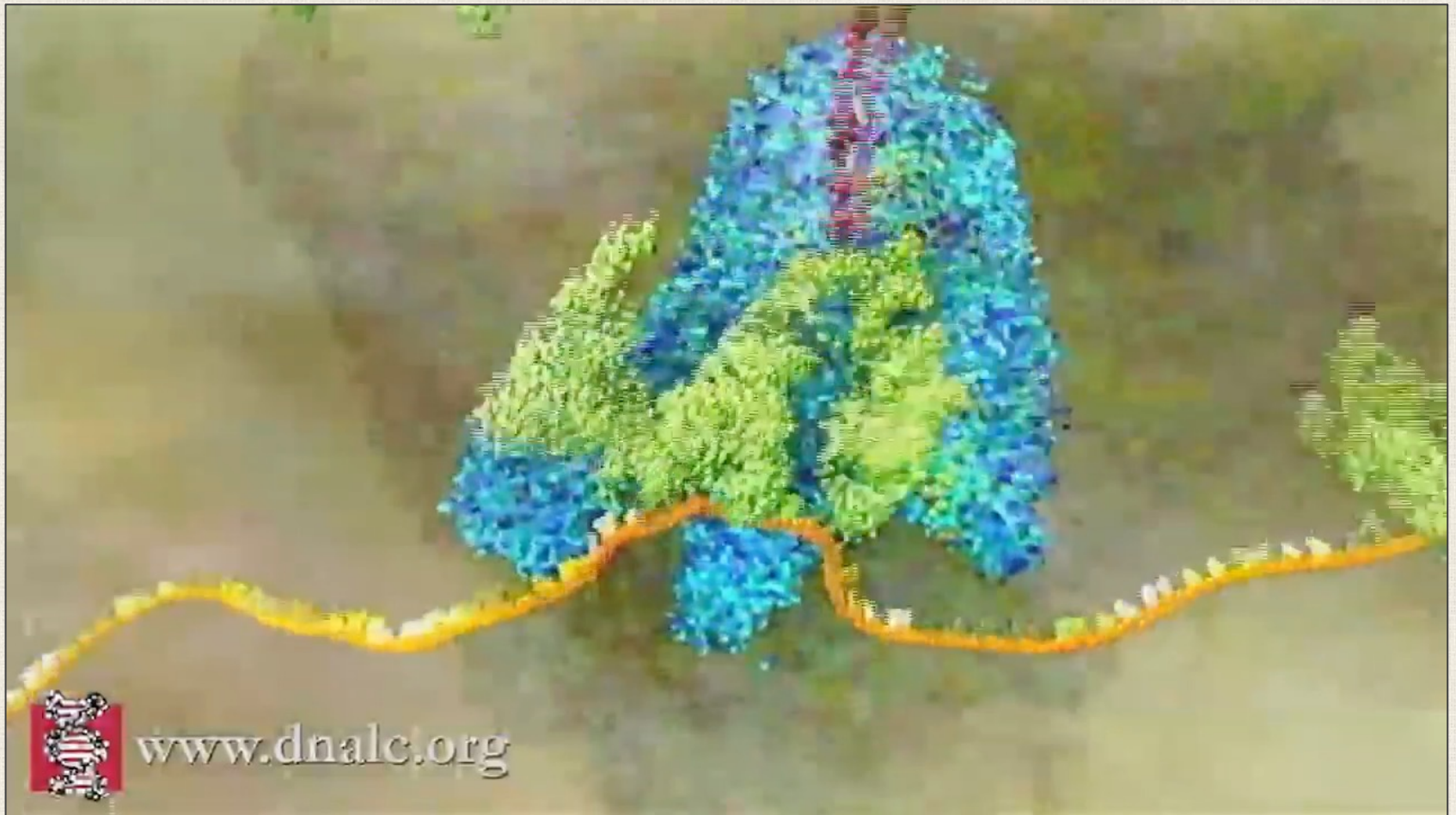


Proteins Come in All Different Shapes



<https://pdb101.rcsb.org/motm/motm-by-date>

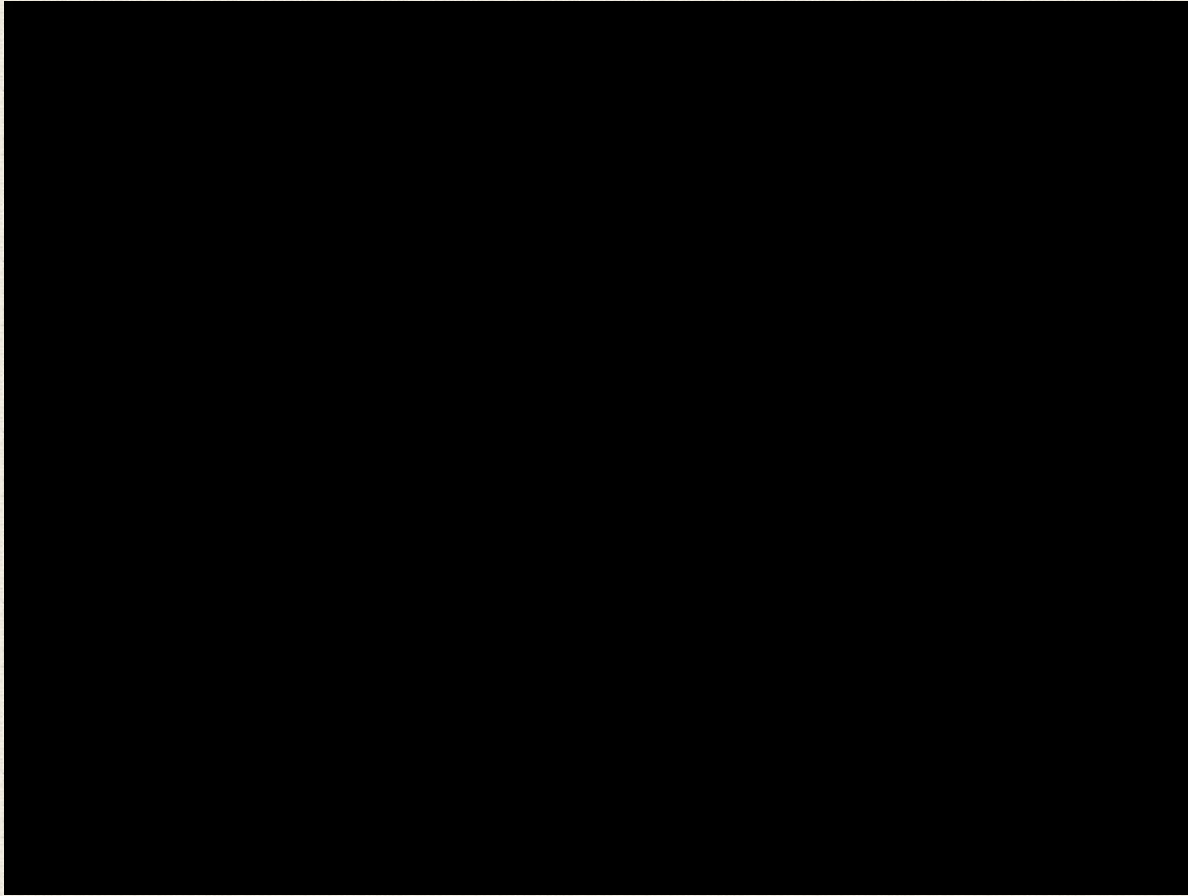
The Shape of a Protein Influences Its Function



https://youtu.be/TfYf_rPWUdY

Ribosome in action

A protein typically folds into the same shape every time



<https://www.youtube.com/watch?v=yZ2aY5lxEGE>

The *Biological* Problem is Clear

Protein Structure Prediction Problem

- **Input:** An amino acid string corresponding to a protein.
- **Output:** The 3-D shape of the protein.

Nature has devised a “**magic algorithm**” solving this biological problem. Can we reverse engineer this algorithm?

The Russian Academy of Sciences' Protein Institute...



Институт белка Российской академии наук

[Руководство](#) [Директора](#) [Лаборатории](#) [Об Институте](#) [Материалы](#) [Аспирантура](#) [Учебный центр](#) [ЦКП](#) [Контакты](#)

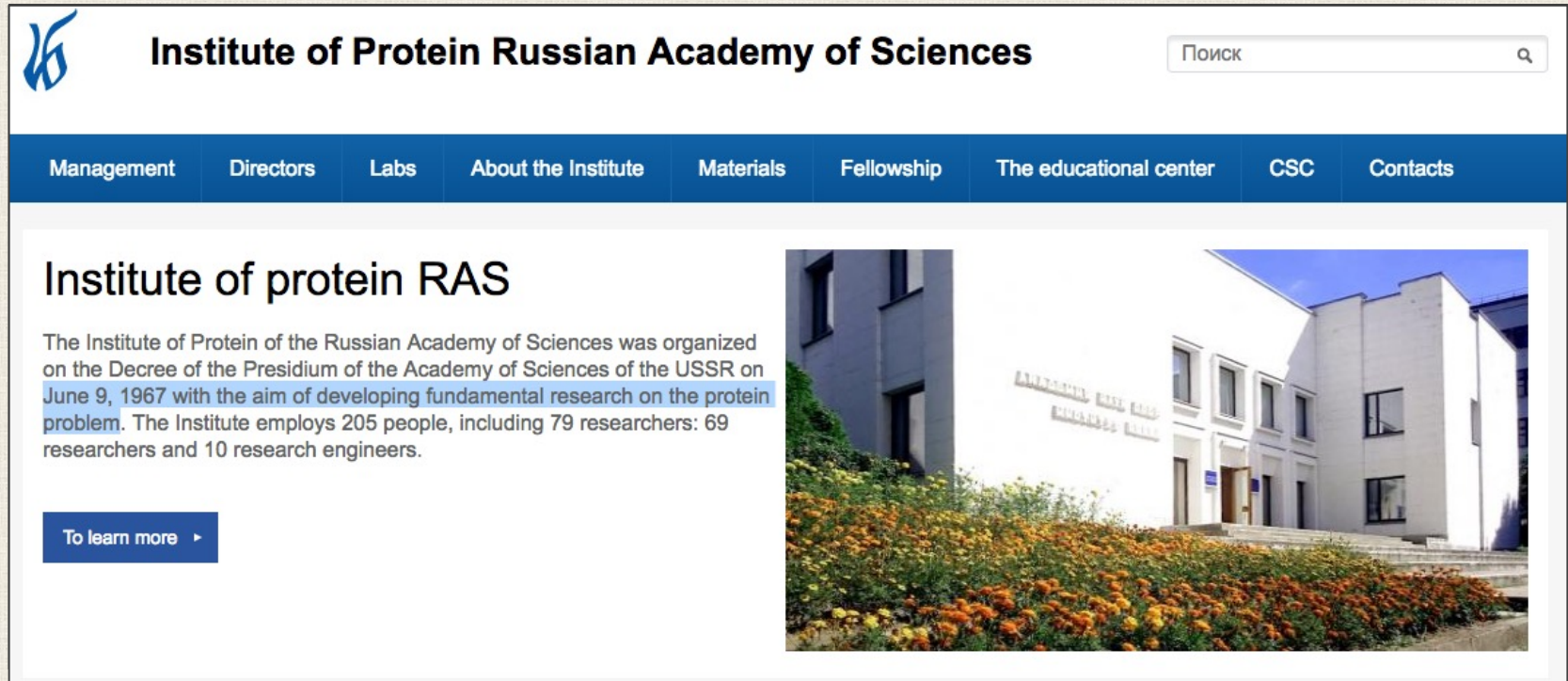
Институт белка РАН

Институт белка РАН организован по Постановлению Президиума Академии наук СССР 9 июня 1967 г. с целью развертывания фундаментальных исследований по проблеме белка. В Институте трудится 205 человек, из них 79 исследователей: 69 научных сотрудников и 10 инженеров-исследователей.

[Узнать больше ▶](#)



...has tried to solve this problem for over
50 years!



Institute of Protein Russian Academy of Sciences


Поиск

Management Directors Labs About the Institute Materials Fellowship The educational center CSC Contacts

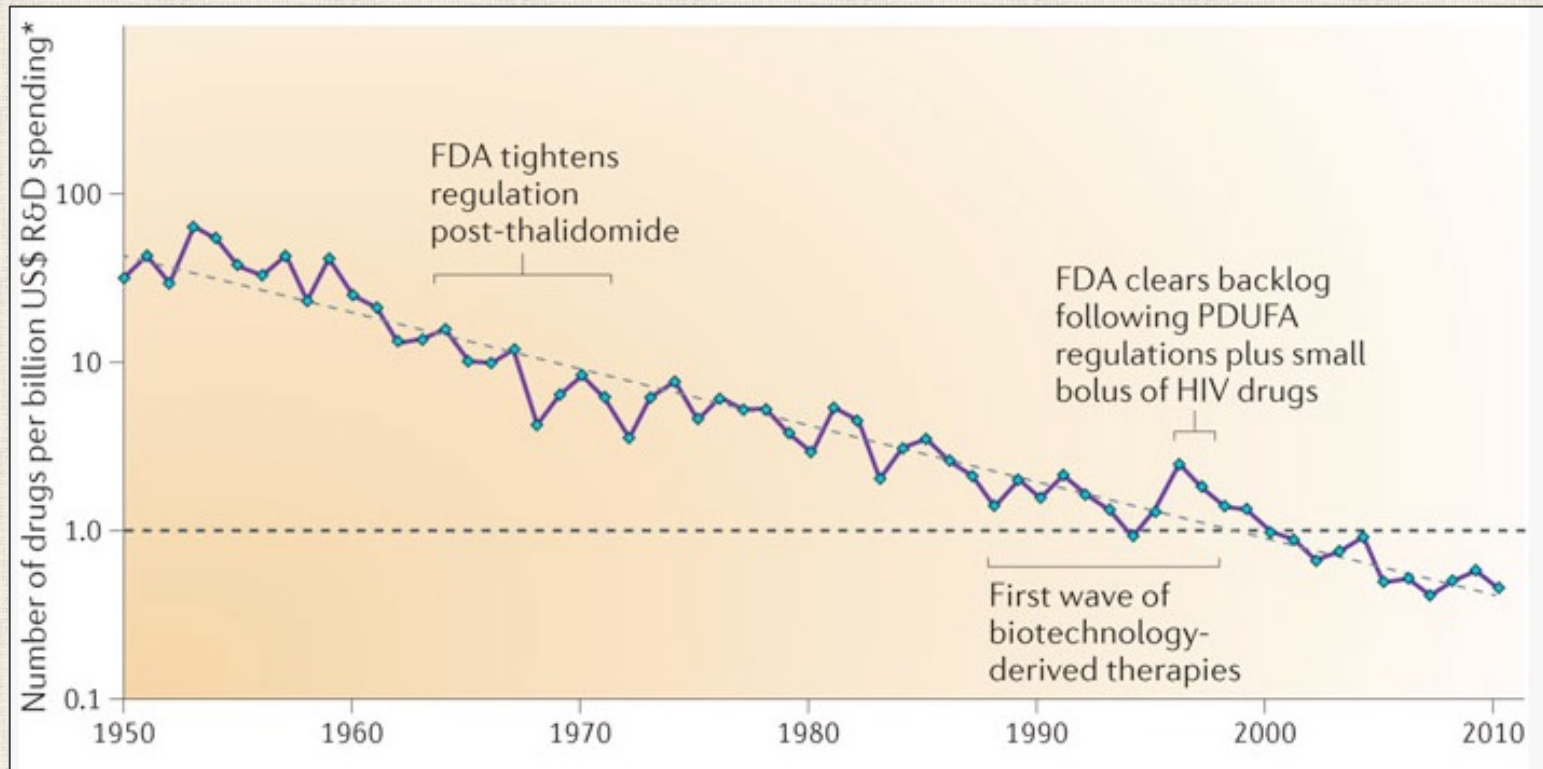
Institute of protein RAS

The Institute of Protein of the Russian Academy of Sciences was organized on the Decree of the Presidium of the Academy of Sciences of the USSR on June 9, 1967 with the aim of developing fundamental research on the protein problem. The Institute employs 205 people, including 79 researchers: 69 researchers and 10 research engineers.

To learn more ▶



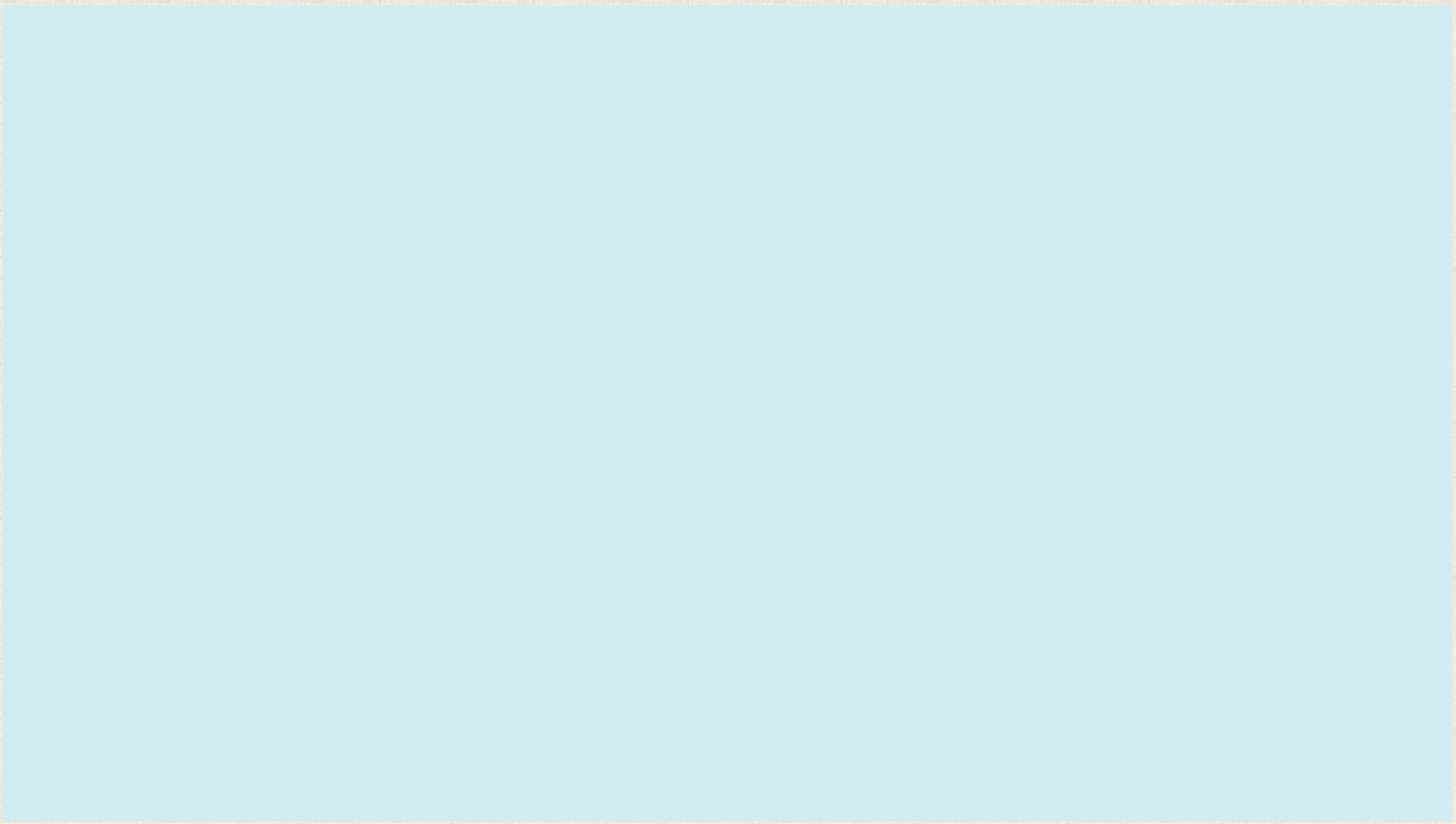
Drug discovery often relies on finding drugs that will bind to protein of interest



https://blogs.sciencemag.org/pipeline/archives/2012/03/08/erooms_law

EROOM'S LAW
MOORE'S LAW

We *can* determine the shape of a protein experimentally



<https://www.youtube.com/watch?v=Qq8DO-4BnIY>

So ... why not use cryo-EM for all proteins?

The electron microscope needed can cost \$5M or more and cost a fortune to run.

So ... why not use cryo-EM for all proteins?

The electron microscope needed can cost \$5M or more and cost a fortune to run.

And remember that just for humans, there are between 600,000 and 6 million isoforms!

So ... why not use cryo-EM for all proteins?

The electron microscope needed can cost \$5M or more and cost a fortune to run.

And remember that just for humans, there are between 600,000 and 6 million isoforms!

Key point: with today's technology, we will never be able to experimentally determine the structure of all proteins.



Google

tnt recipe

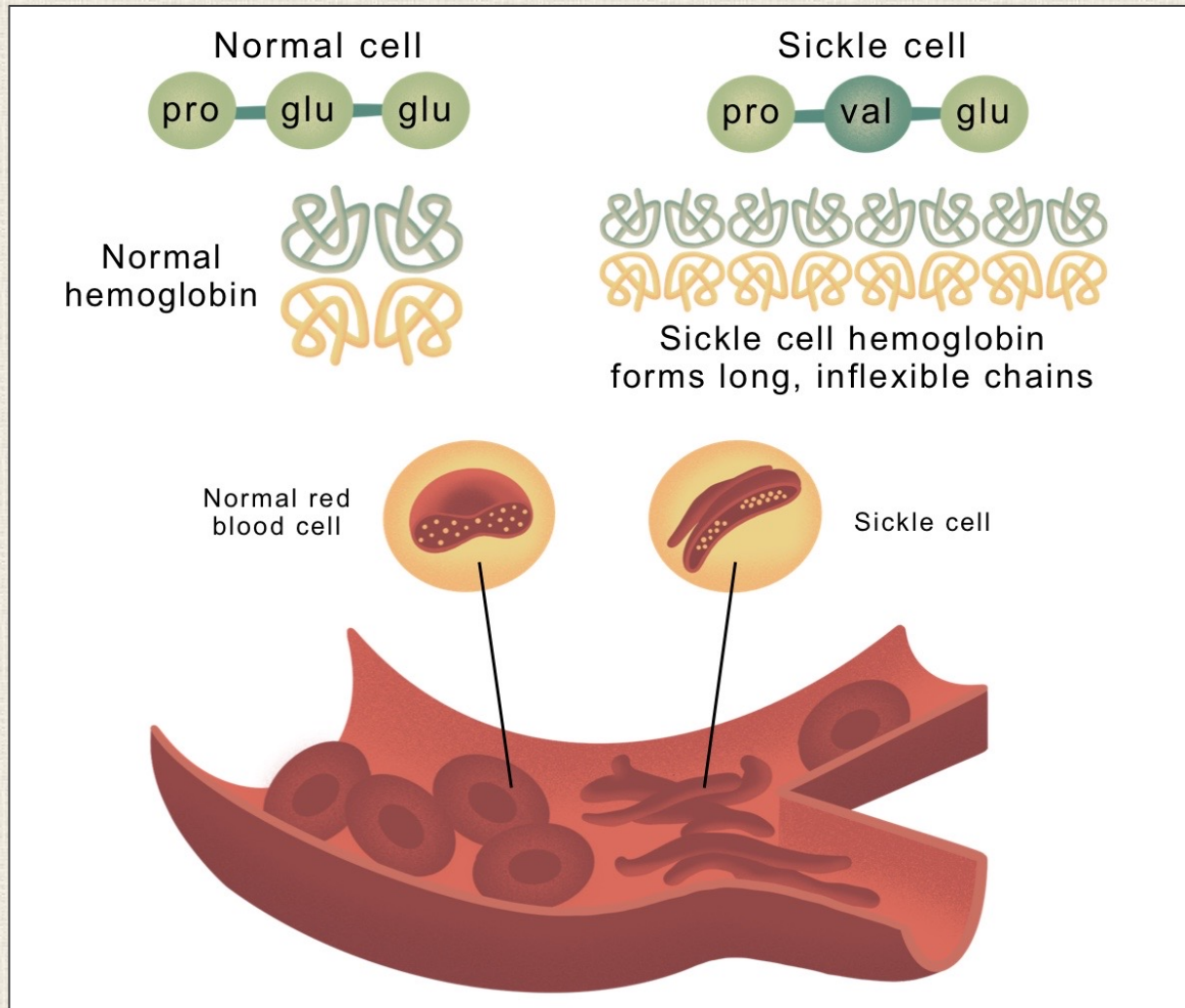
In English, some small changes can cause enormous differences in meaning ...



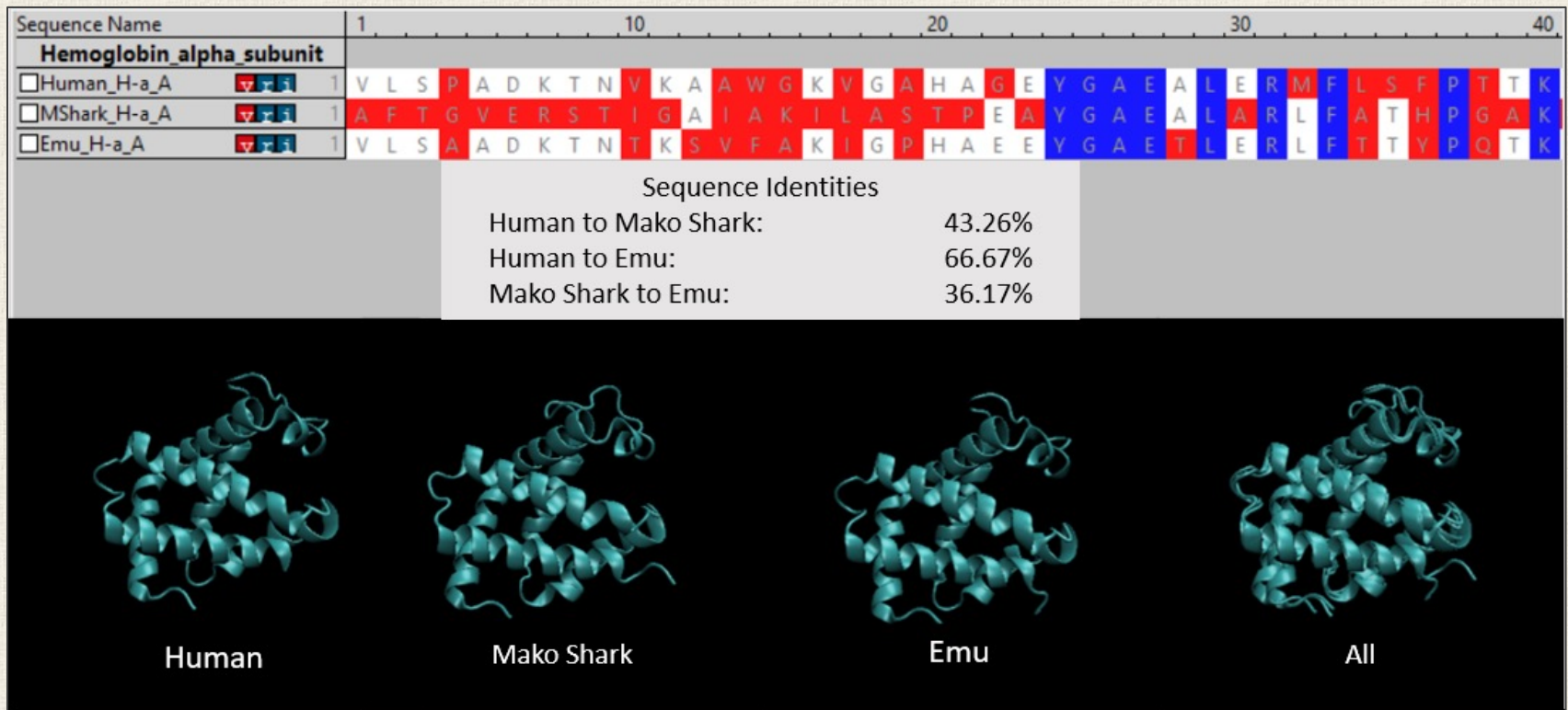
Google

tnt recipe minecraft

In proteins, some small mutations can cause enormous structural changes ...



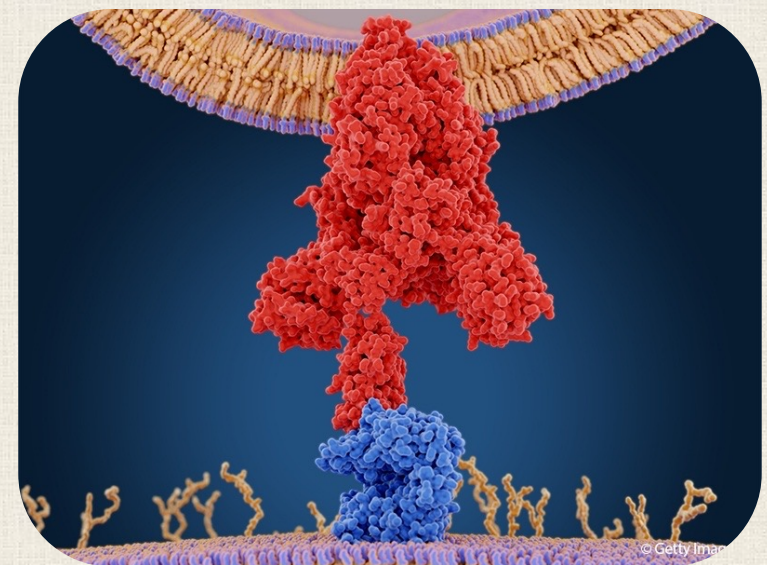
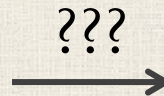
... and yet some similar structures have very different sequences!



Two Big Picture Questions

Question 1: What is the 3-dimensional protein corresponding to a string of amino acids?

MFVFLVLLPLVSSQCVNLTRRTQLPPAYTNSFTRGVVYPDKVFRSSVLHSTQDLFLP
FFSNVTWFHAIHVSNGTKRFNDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSTK
QSLIIVNNATNVVIVKVFCEQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYV
SQPFLMDLEGKQGFKNLREFVFKNIDGYFKIYSKHTP.INLVRDLPPQGFSALEPLVD
LP.IGINITRFQTLALHRSYLTPGDSSSGWTAGAAAYVGYLQPRTFLLKYNENGTI
TDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVF
NATRFASVYAWNRRKISNCVADYSVLVNSASFSTFKCYGVSPTKLNLDLCTNRYADS
FVIRGDEVQRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLRYLF
RKSNLKPFERDISTEIQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVL
SFELLHAPATVCGPKKSTNLVKNKCVNFNENGLTGTGVLTESNKKFLPFQGFGRDIA
DTTDAVRDPQTLLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHAD
QLTPTWRVYSTGSNVFQTRAGCLIGAHEVNNSEYCDIPIGAGICASYQTQTNSPRRA
RSVASQSIIAYTMSLGAENSVAYSNNLSIAIPTNFTISVTTEILPVSMTKTSVDCTMY
ICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVQKIYKTPPIKDFG
GFNFSQILPDPSPKSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDI AARDL ICAQKF
NGLTVLPPLLTDEMQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVT
QNVLYENQKLIANQFNSAIGKIQDSLSTASALGKLQDVVNQNAQALNTLVKQLSSN
FGAISSVLNDILSRDKVEAEVQIDRLITGRLQSLQTYVTQQQLIRAAEIRASANLAA
TKMSECVLGQSKRVDFCGKGYHLSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICH
DGKAHFREGVFSVNGTHWFVTQRNFYEPQIITDNTFVSGNCDVVIGIVNNTVYDP
LQPELDSFKEELDKYFKNHTSPDVLGDISGINASVVNIQKEIDRLNEVAKNLNESL
IDLQELGKYEYIKWPWYIWLGFIAGLIAIVMVTIMLCMTSCCSCLKGCCSCGSSC
KFDEDDSEPVKGVKLVHHT

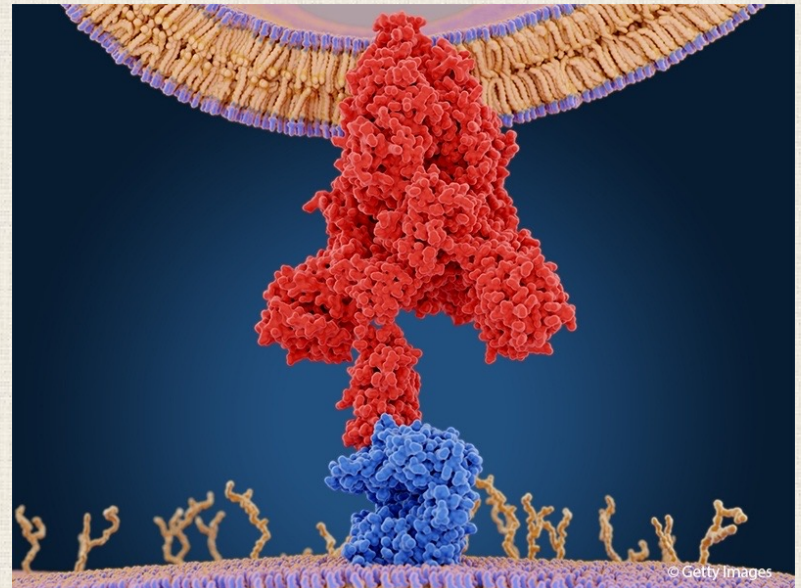


<https://www.cas.org/blog/covid-19-spike-protein>

Two Big Picture Questions

Question 2: How can we compare two (similar) proteins on the level of *structure*?

Key Point: We want to make conclusions about how a change in the structure of a protein (e.g., spike protein) affects the *function* of the protein.



SOME NECESSARY BIOCHEMISTRY

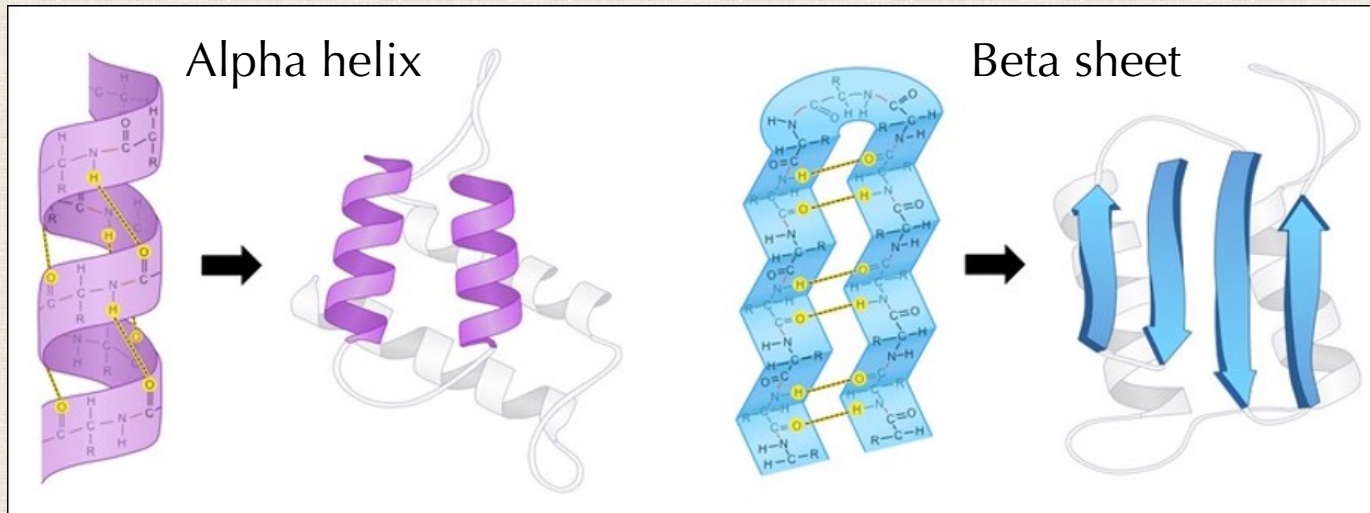
What do we mean by "structure"?

A protein's **primary structure** refers to the amino acid sequence of its **polypeptide** chain.

MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVVYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFA
STEKSNIIRGWIFGTTLDSKTQSLIIVNNATNVVIKVFCEFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNF
KNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLALHRSYLTGDSGWTAGAAAYVGYLQPRTFLKLY
NENGTITDAVDCALDPLSEKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSA
SFSTFKCYGVSPTKLNLCFTNVYADSFVIRGDEVQRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGNYNYLYRFRKSNLKPFE
RDISTEIQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKK
FLPFQQFGRDIADTTDAVRDPQTEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLI
GAEHVNSYECDIPIGAGICASYQTQTNPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICG
DSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSPKPSKRSFIEDLLFNKVTLADAGFIKQY
GDCLGDI AARDL ICAQKFNGLTVLPPLL TDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFN
SIAIGKIQDSLSSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASAN
LAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNF TAPAICHGKAHFPREGVFVSNGTHWFVTQRNFYEPQIIT
TDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQY
IKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLLKGVKLYHT

What do we mean by "structure"?

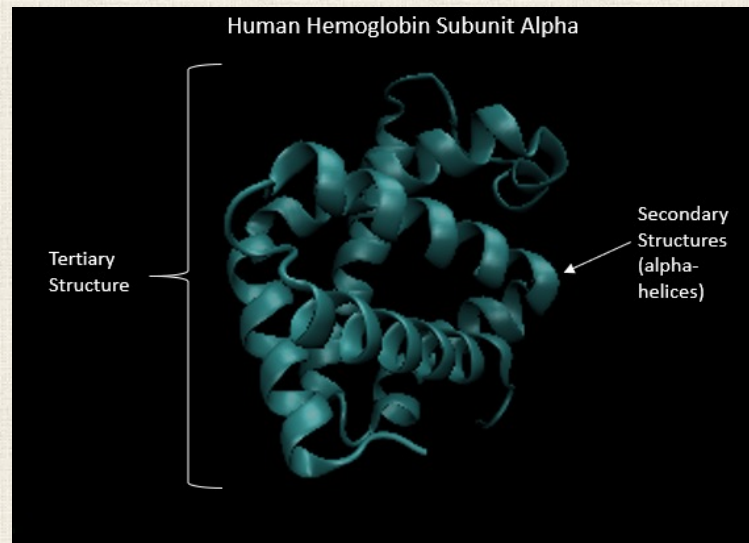
A **secondary structure** is a repeating substructure that forms as a substructure of the overall folded protein.



<https://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/73-translation/protein-structure.html>

What do we mean by “structure”?

A protein’s **tertiary structure** describes its final 3D shape after the polypeptide chain has folded and is chemically stable. This is what we most commonly refer to as the “structure” of a protein.

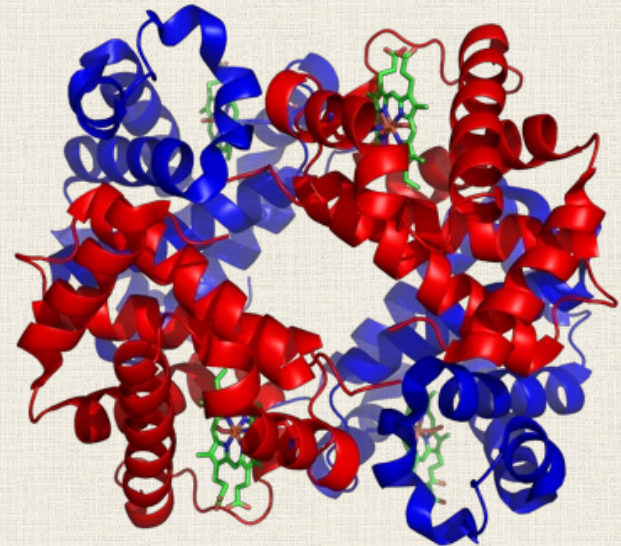


<https://www.rcsb.org/structure/1SI4>

What do we mean by "structure"?

Some proteins have a **quaternary structure**, which describes the protein's interaction with other copies of itself to form a single functional unit, or a **multimer**.

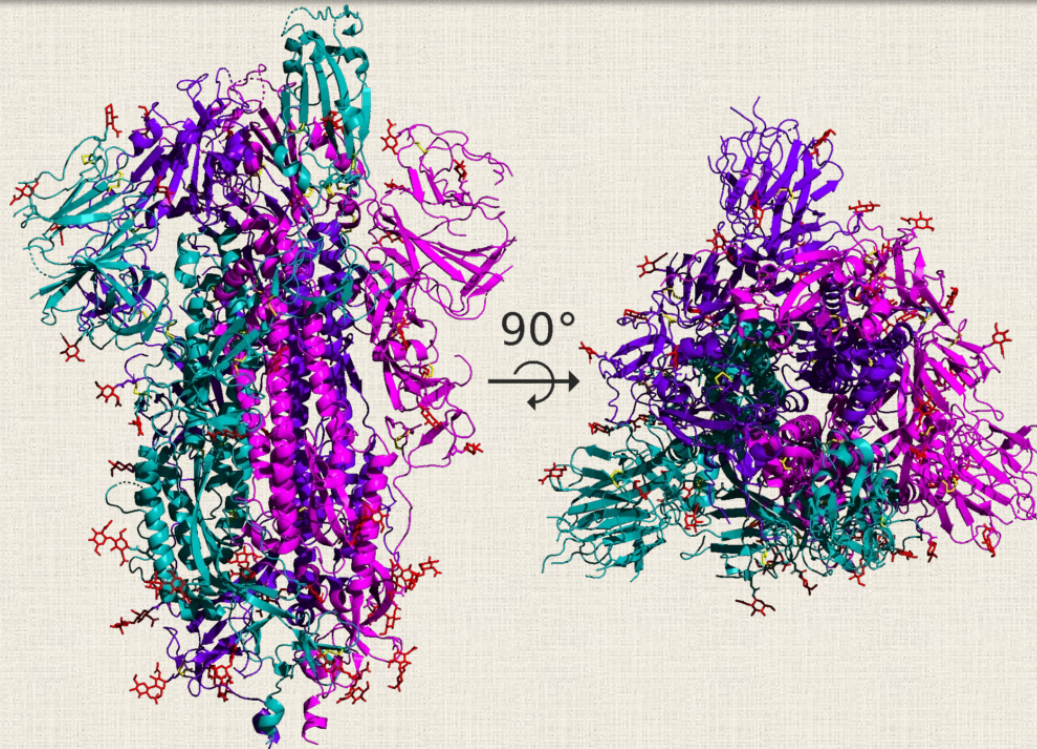
Hemoglobin is a multimer consisting of two alpha subunits and two beta subunits.



https://commons.wikimedia.org/wiki/File:1GZX_Haemoglobin.png

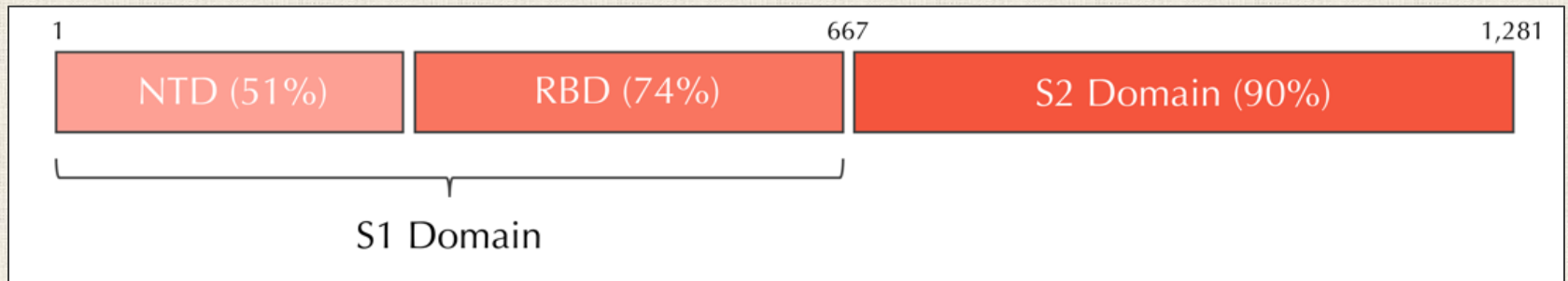
A note on the spike protein

The spike protein is a **homotrimer**, formed of three essentially identical units called **chains**, each one translated from the same genome region.



A note on the spike protein

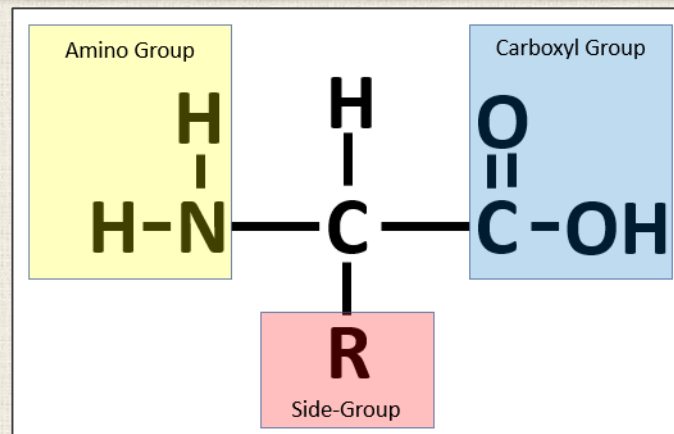
And each chain is formed of two subunits that itself is formed of independently folding **domains** that are each responsible for a specific interaction or function.



A bit more biochemistry

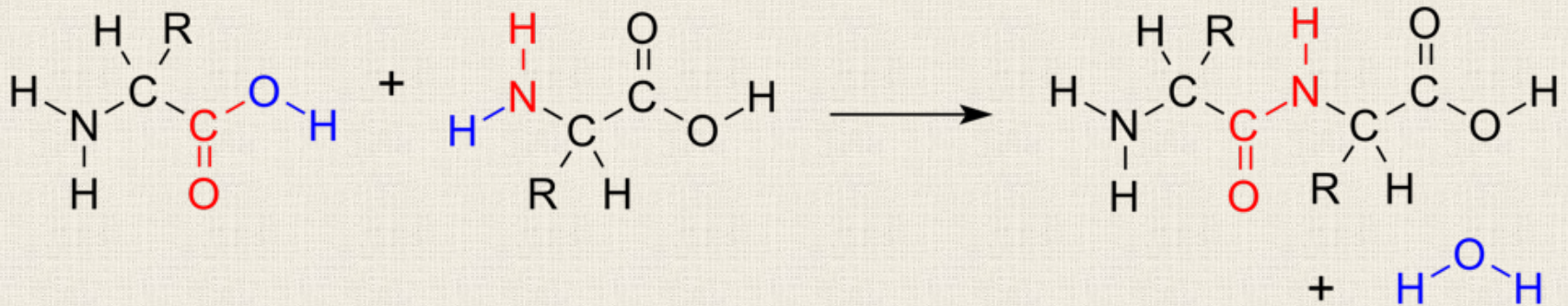
An amino acid's central **alpha carbon** atom is connected to four different molecules:

1. a hydrogen atom (H)
2. a carboxyl group ($-\text{COOH}$)
3. an amino group ($-\text{NH}_2$)
4. a **side chain** (denoted "R"), which differs between amino acids.



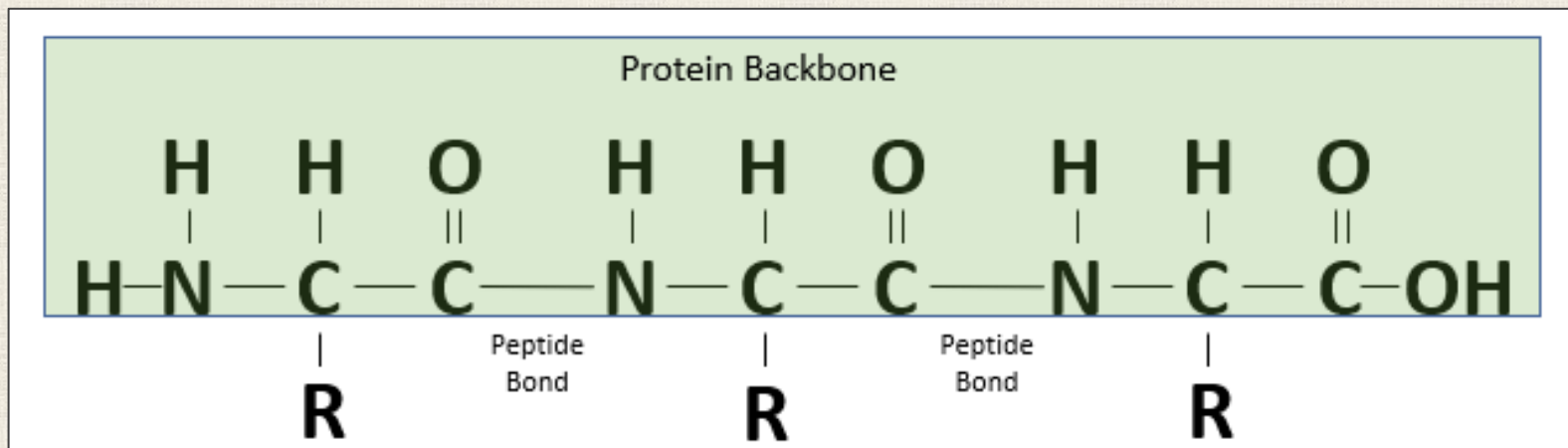
A bit more biochemistry

To form a polypeptide chain, consecutive amino acids are linked together during a condensation reaction in which the amino group of one amino acid is joined to the carboxyl group of another, while a water molecule (H_2O) is expelled.



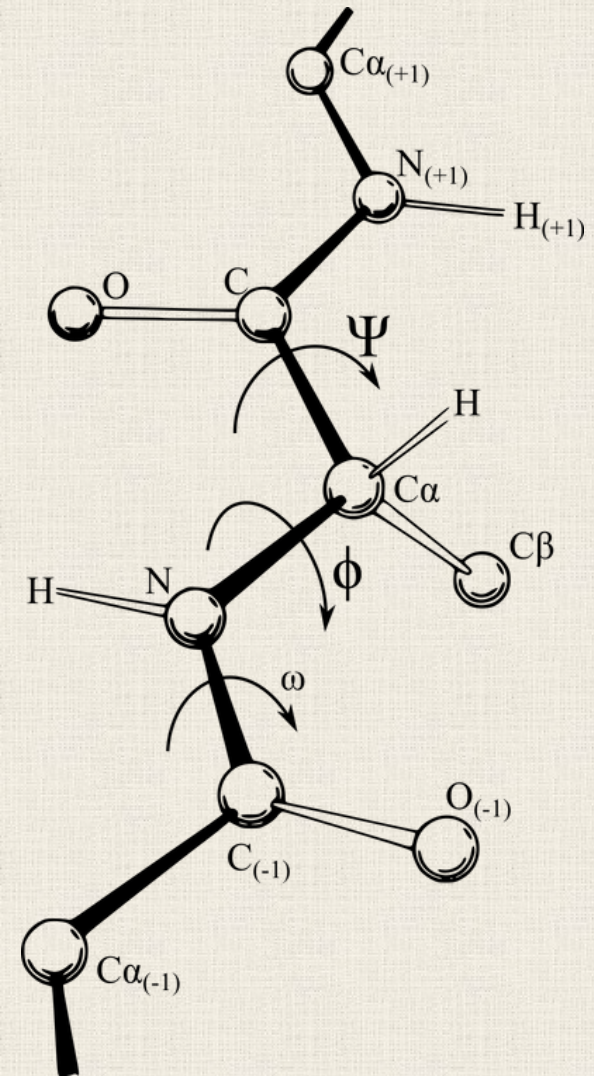
A bit more biochemistry

The resulting N-C bond that is produced, called a **peptide bond**, is very strong. The peptide has very little rotation around this bond, which is almost always locked at 180° . The polypeptide chain is formed of consecutive peptide bonds.



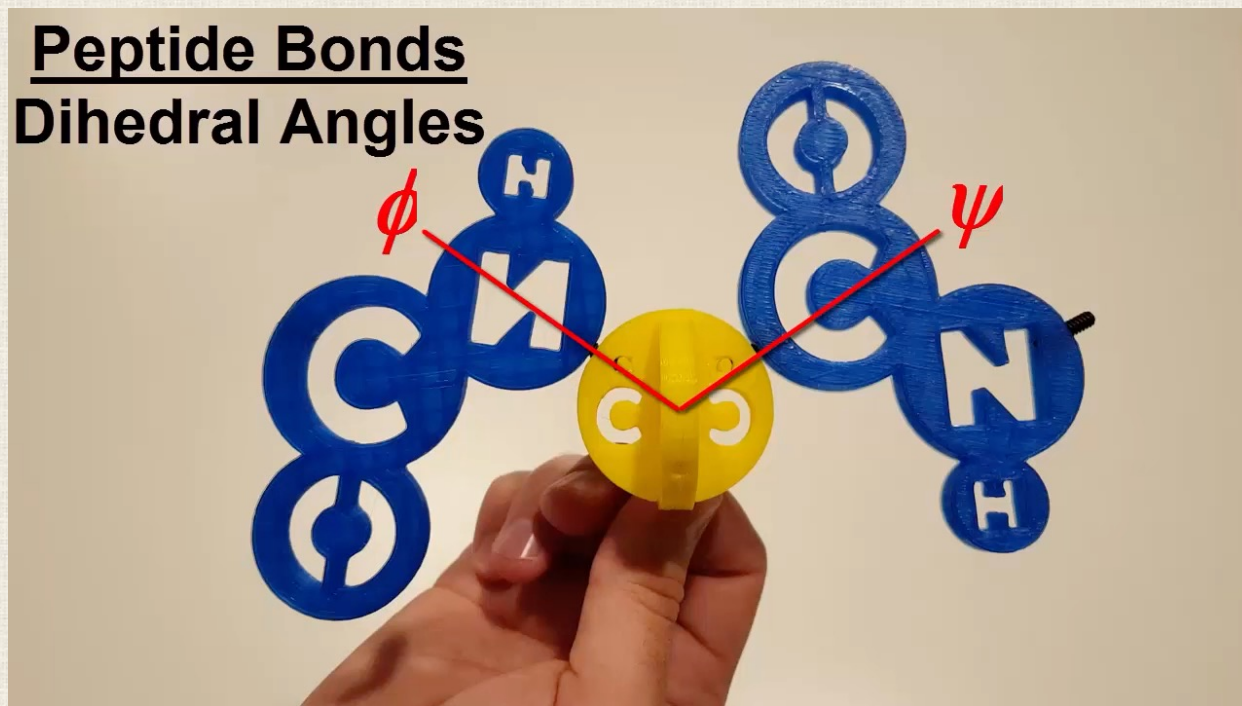
A bit more biochemistry

The bonds *within* an amino acid are not as rigid. The polypeptide is free to rotate around these two bonds. This rotation produces two angles of interest, called the **phi angle (ϕ)** and **psi angle (ψ)**, where the alpha carbon connects to its amino group and carboxyl group, respectively.



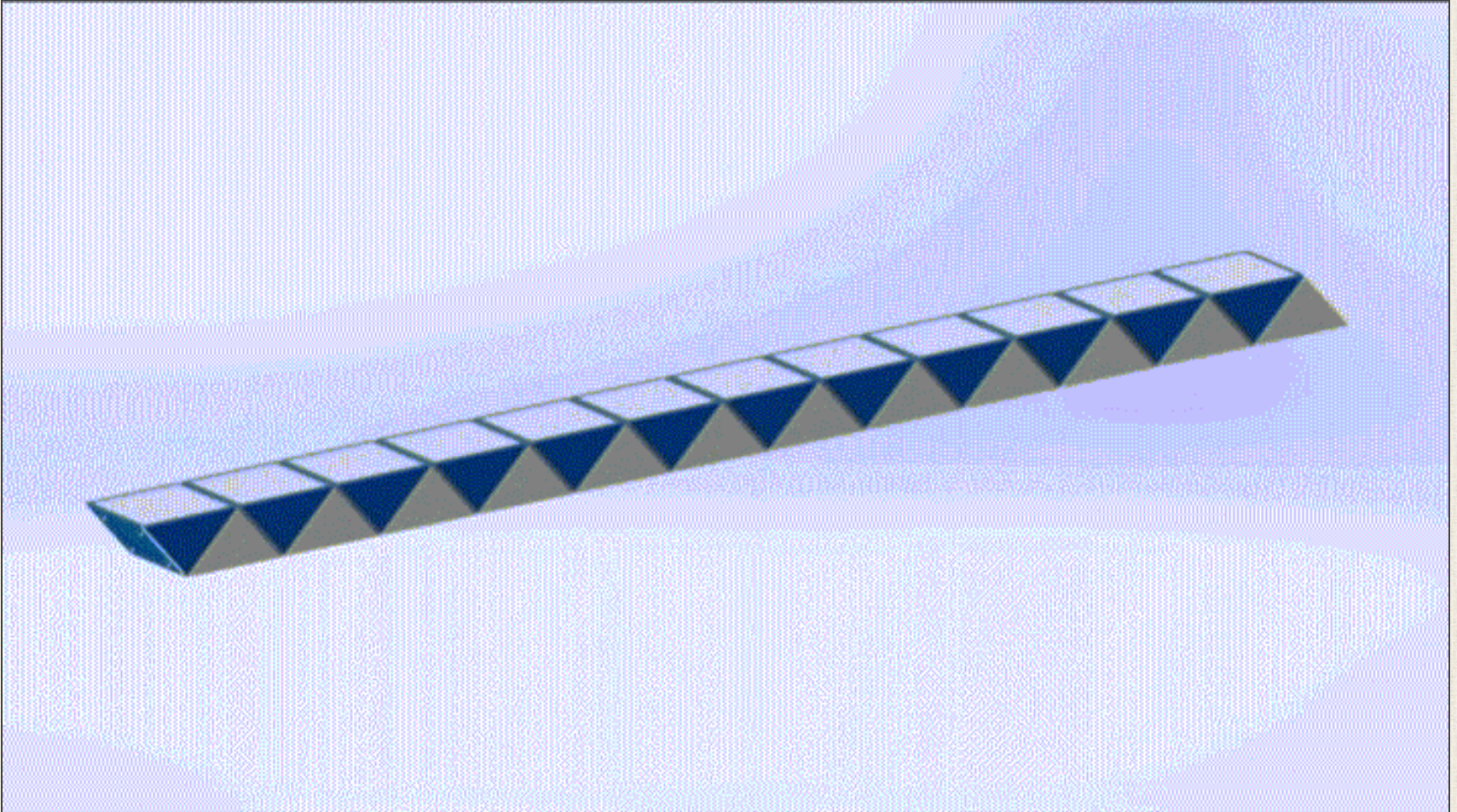
Proteins are flexible and can therefore form a huge number of shapes

This video illustrates how changing ϕ and ψ at an amino acid can drastically change a protein's shape.



Courtesy: Jacob Elmer, https://youtu.be/1usemtlYe_s

A good analogy for polypeptide flexibility is the “Rubik’s Twist” puzzle



Proteins are flexible and can therefore form a huge number of shapes

A polypeptide with n amino acids has $n - 1$ peptide bonds, meaning $n - 1$ ϕ angles and $n - 1$ ψ angles.

Proteins are flexible and can therefore form a huge number of shapes

A polypeptide with n amino acids has $n - 1$ peptide bonds, meaning $n - 1$ ϕ angles and $n - 1$ ψ angles.

If each bond has k stable conformations, then the polypeptide has k^{2n-2} potential structures!

Proteins are flexible and can therefore form a huge number of shapes

A polypeptide with n amino acids has $n - 1$ peptide bonds, meaning $n - 1$ ϕ angles and $n - 1$ ψ angles.

If each bond has k stable conformations, then the polypeptide has k^{2n-2} potential structures!

The ability for the magic algorithm to find a single conformation despite such an enormous number of potential shapes is called **Levinthal's paradox**.

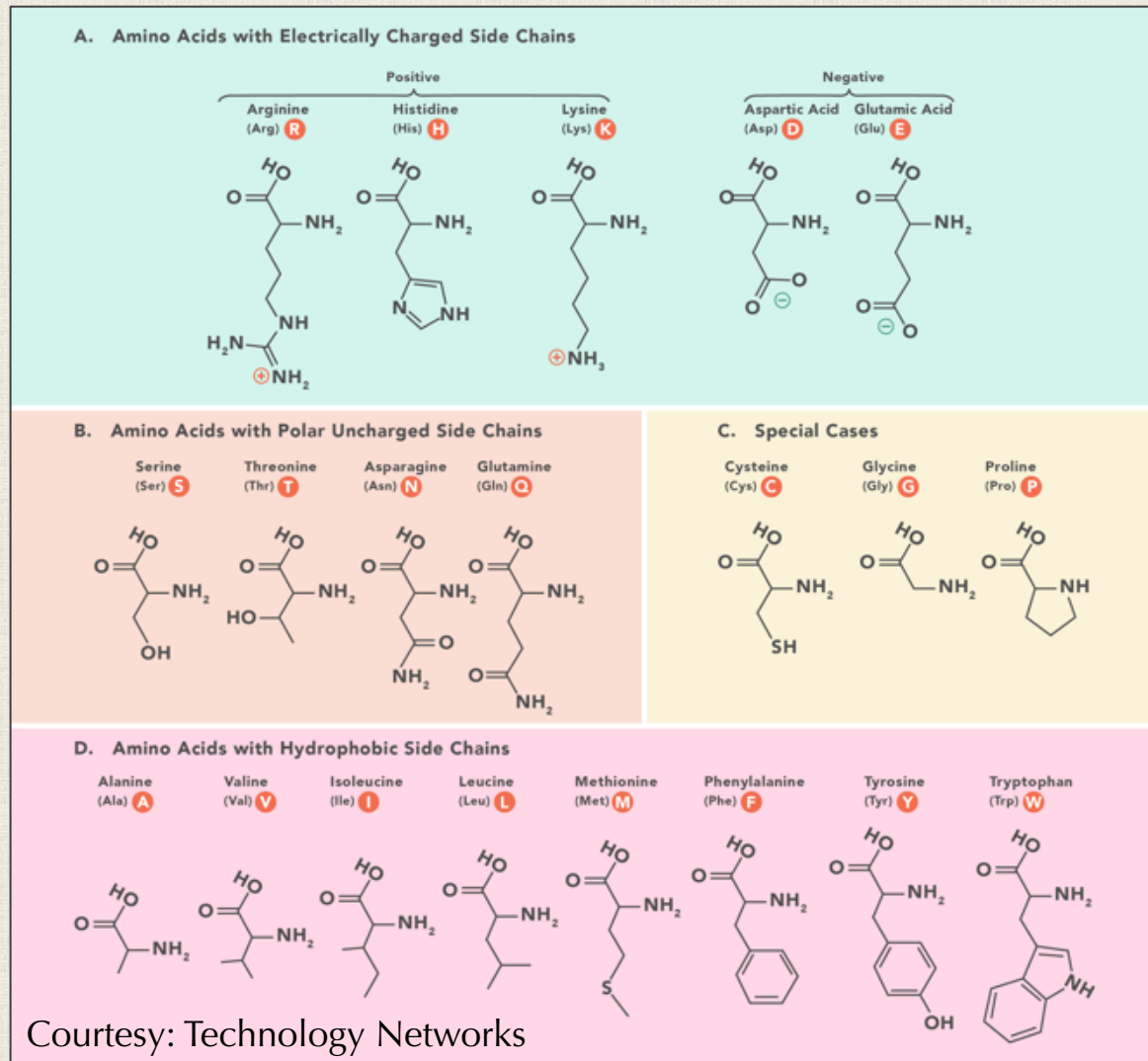
Mutations aren't made alike

We already know from scoring alignments that some amino acid mutations may be more favorable than others.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	-
A	2	-2	0	0	-3	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3	-8
C	-2	12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0	-8
D	0	-5	4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4	-8
E	0	-5	3	4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4	-8
F	-3	-4	-6	-5	9	-5	-2	1	-5	2	0	-3	-5	-5	-4	-3	-3	-1	0	7	-8
G	1	-3	1	0	-5	5	-2	-3	-2	-4	-3	0	0	-1	-3	1	0	-1	-7	-5	-8
H	-1	-3	1	1	-2	-2	6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0	-8
I	-1	-2	-2	-2	1	-3	-2	5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1	-8
K	-1	-5	0	0	-5	-2	0	-2	5	-3	0	1	-1	1	3	0	0	-2	-3	-4	-8
L	-2	-6	-4	-3	2	-4	-2	2	-3	6	4	-3	-3	-2	-3	-3	-2	2	-2	-1	-8
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6	-2	-2	-1	0	-2	-1	2	-4	-2	-8
N	0	-4	2	1	-3	0	2	-2	1	-3	-2	2	0	1	0	1	0	-2	-4	-2	-8
P	1	-3	-1	-1	-5	0	0	-2	-1	-3	-2	0	6	0	0	1	0	-1	-6	-5	-8
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4	1	-1	-1	-2	-5	-4	-8
R	-2	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6	0	-1	-2	2	-4	-8
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2	1	-1	-2	-3	-8
T	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3	0	-5	-3	-8
V	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4	-6	-2	-8
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	0	-8
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10	-8
-	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8

PAM250 matrix

Amino acids' side chain variety produces different chemical properties



Courtesy: Technology Networks

Proteins seek the lowest potential energy conformation

We can view protein folding as finding the tertiary structure that is the most *stable* given a polypeptide's primary structure (i.e., has lowest potential energy).

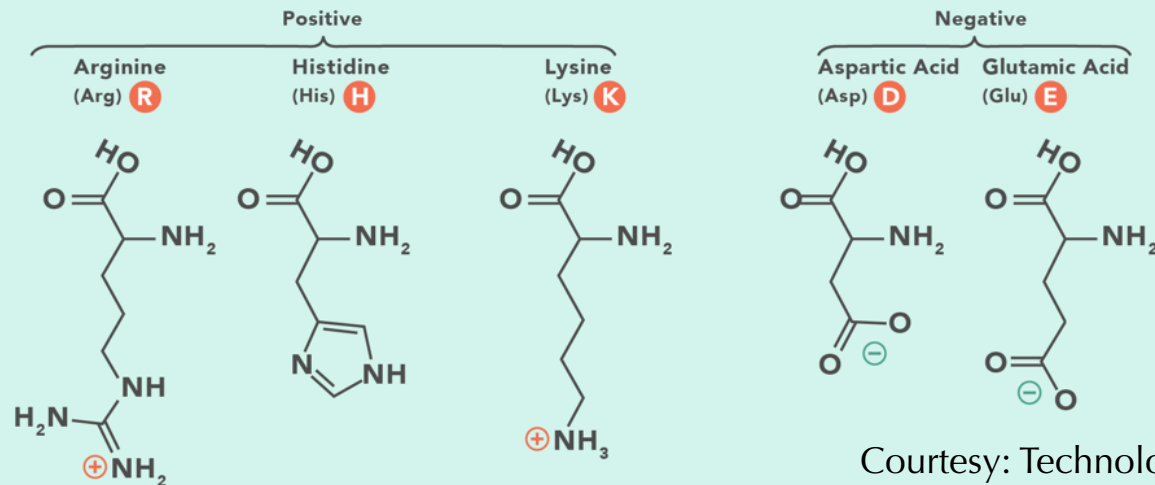
Proteins seek the lowest potential energy conformation

We can view protein folding as finding the tertiary structure that is the most *stable* given a polypeptide's primary structure (i.e., has lowest potential energy).

The **potential energy** (a.k.a. **free energy**) of a protein is the energy stored within it due to its position, state, and arrangement. It derives from the protein's bonds as well as non-bonded energy (e.g., **electrostatic interactions** and **van der Waals forces**).

Electrostatic interactions occur between amino acids of opposite charge

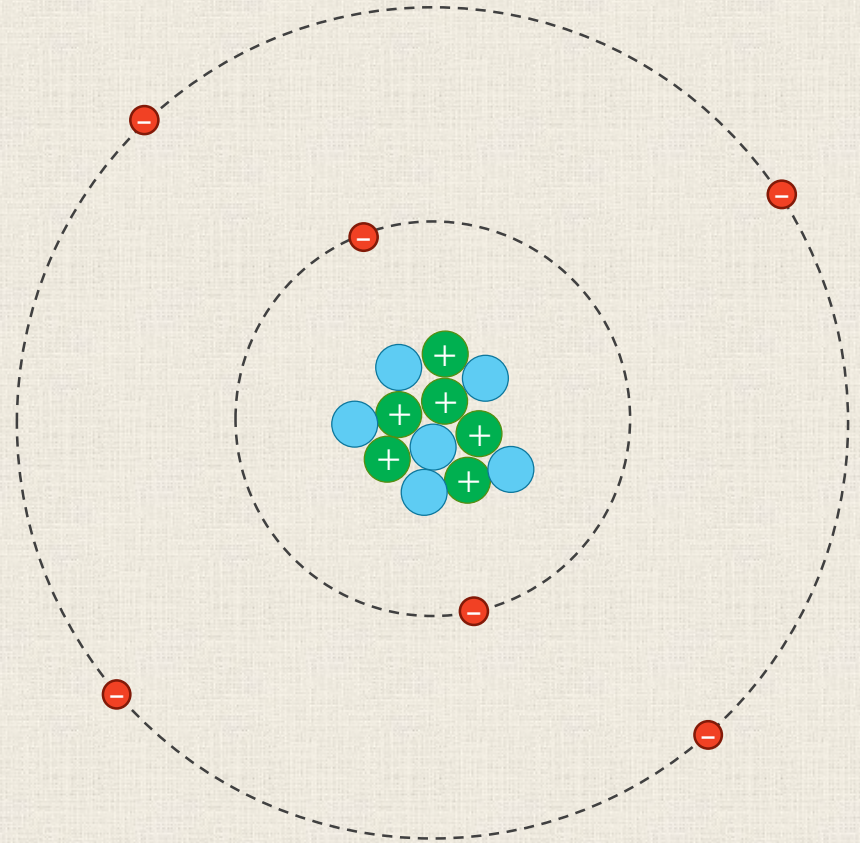
A. Amino Acids with Electrically Charged Side Chains



Courtesy: Technology Networks

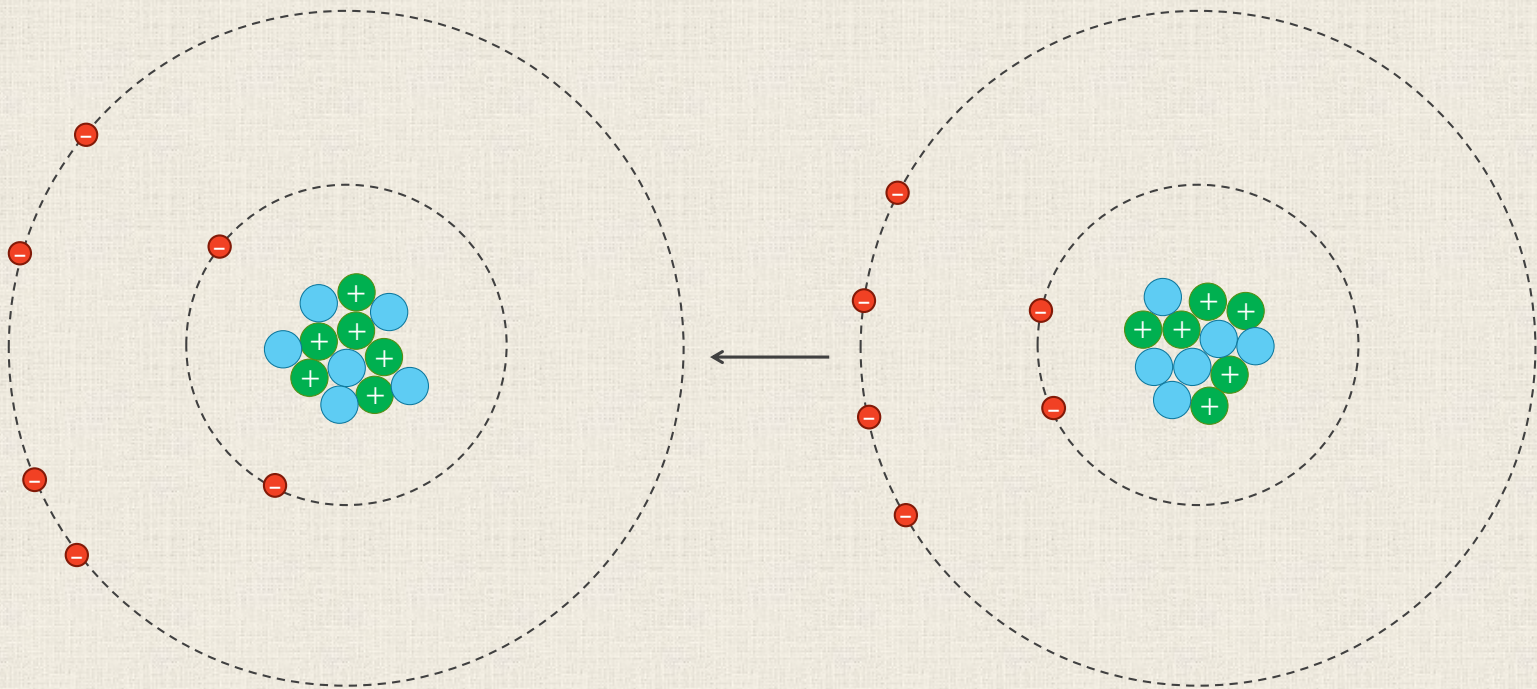
What are van der Waals forces?

Atoms are dynamic systems, with electrons constantly buzzing around the nucleus. At any given moment, they are probably relatively uniform.



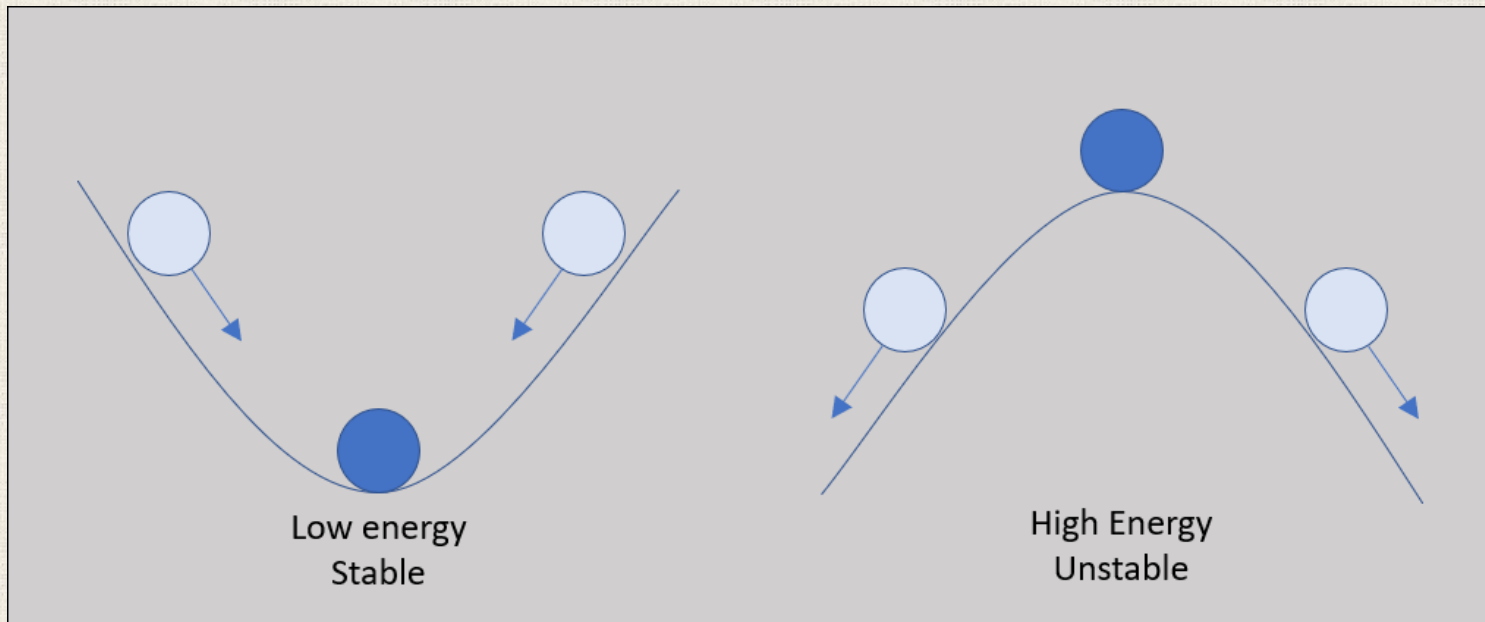
What are van der Waals forces?

Due to random chance, electrons may accumulate on one side of an atom, creating a temporary “pole” that causes this effect in nearby atoms as well.



A classic analogy of proteins finding lowest energy conformation

Imagine a ball on a slope; gravity causes it to tend to move down the slope. Similarly, a polypeptide tends toward lower energy conformations.



***AB INITIO* PROTEIN STRUCTURE PREDICTION**

ab initio Protein Structure Prediction

Biochemists have produced scoring functions called **force fields** that compute the potential energy of a candidate protein structure.

***ab initio* Protein Structure Prediction Problem**

- **Input:** An amino acid polypeptide and a force field.
- **Output:** The tertiary structure for this polypeptide having minimum potential energy, given this force field.

ab initio Protein Structure Prediction

Unfortunately, even simple versions of this problem wind up being *NP*-Hard ...

***ab initio* Protein Structure Prediction Problem**

- **Input:** An amino acid polypeptide and a force field.
- **Output:** The tertiary structure for this polypeptide having minimum potential energy, given this force field.

ab initio Protein Structure Prediction

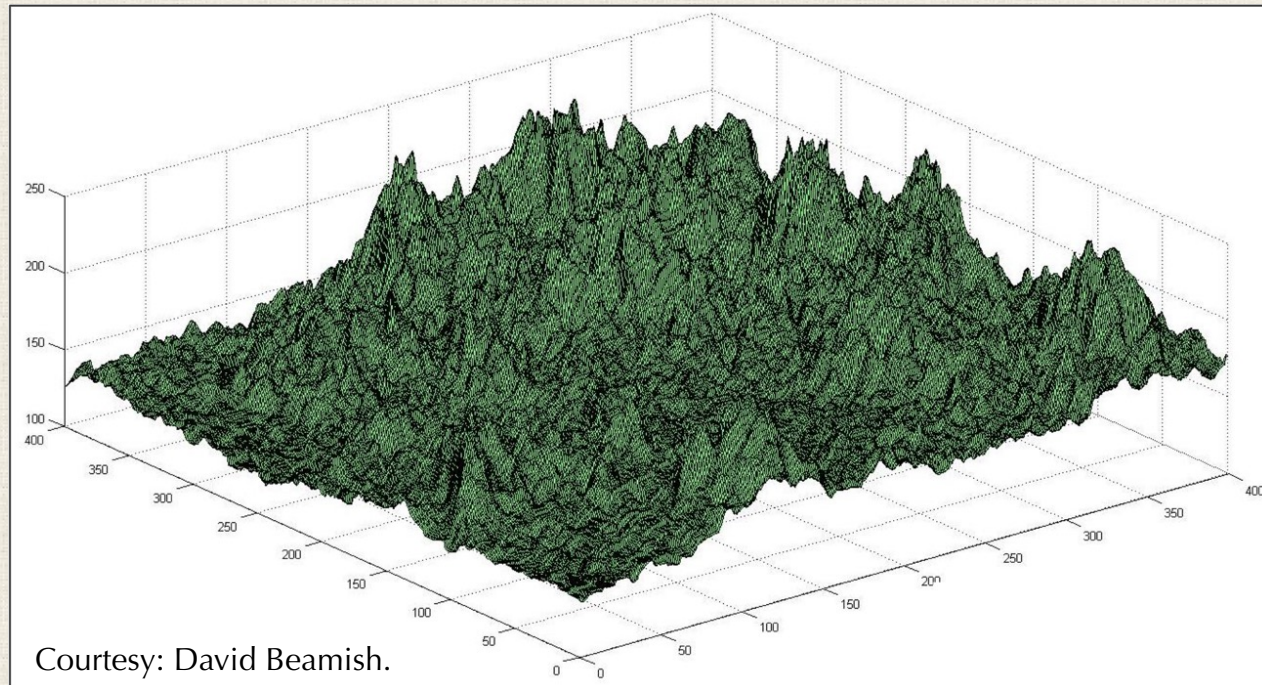
STOP: What does this problem remind us of?

***ab initio* Protein Structure Prediction Problem**

- **Input:** An amino acid polypeptide and a force field.
- **Output:** The tertiary structure for this polypeptide having minimum potential energy, given this force field.

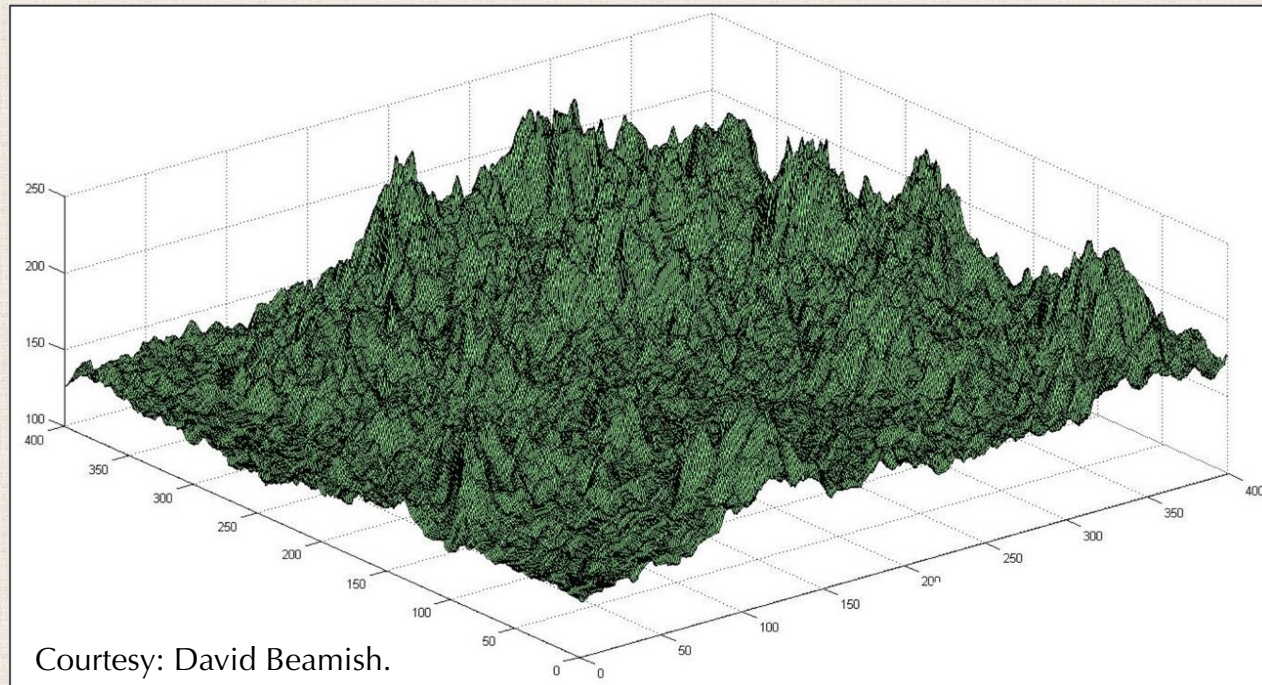
ab initio Protein Structure Prediction

Answer: This is an optimization problem, and the search space is all conformations of the polypeptide.



ab initio Protein Structure Prediction

STOP: What algorithm for *ab initio* structure prediction might you use?



A “Local Search” Algorithm for Protein Structure Prediction

1. Start with an arbitrary protein conformation.
2. Make slight changes to the structure in a variety of ways to produce “neighbors”.
3. Consider the neighbor with optimal score. Is its score better than the current structure?
 - If “yes”, update the current structure to this neighbor and iterate at step 2.
 - If “no”, return the current structure.



A “Local Search” Algorithm for Protein Structure Prediction

1. Start with an arbitrary protein conformation.
2. Make slight changes to the structure in a variety of ways to produce “neighbors”.
3. Consider the neighbor with optimal score. Is its score better than the current structure?
 - If “yes”, update the current structure to this neighbor and iterate at step 2.
 - If “no”, return the current structure.

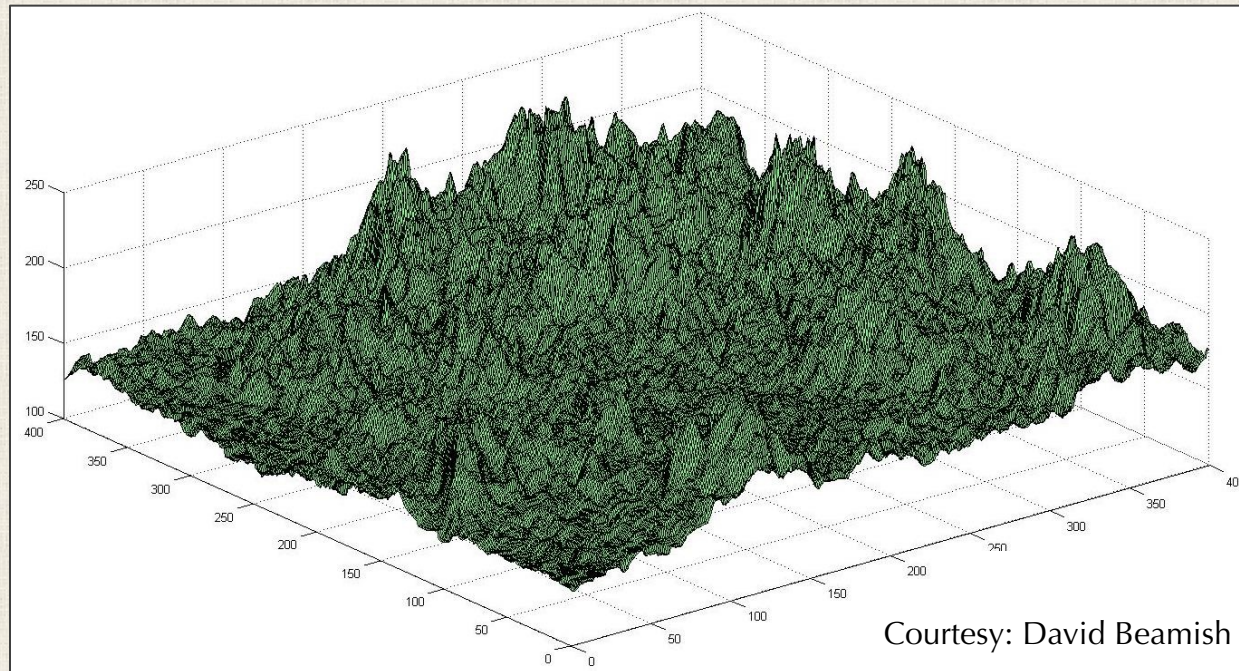


STOP: How could we improve this method?

Improving Local Search

Idea 1: Run algorithm on many different initial values (although search space is huge).

Idea 2: Provide some “jiggle” to allow candidate solutions to “bounce” out of local optima.



Quantifying “Jiggle”

When considering a “neighbor” S' of a candidate protein structure S :

- If $\text{energy}(S') < \text{energy}(S)$, update $S = S'$
- If $\text{energy}(S') > \text{energy}(S)$, then update $S = S'$ with *probability* proportional to $\Delta\text{energy} = \text{energy}(S) - \text{energy}(S')$.

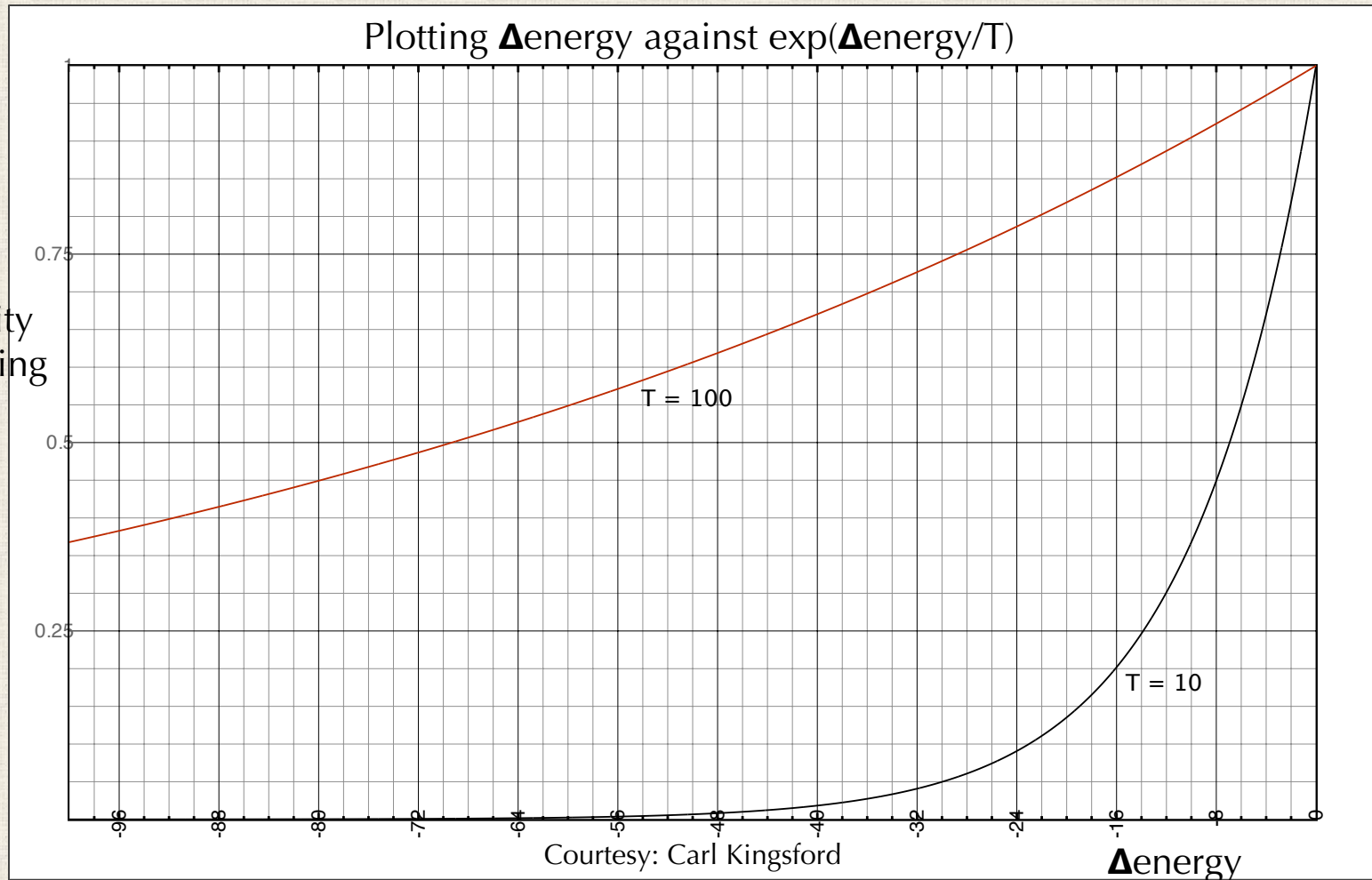
Quantifying “Jiggle”

When considering a “neighbor” S' of a candidate protein structure S :

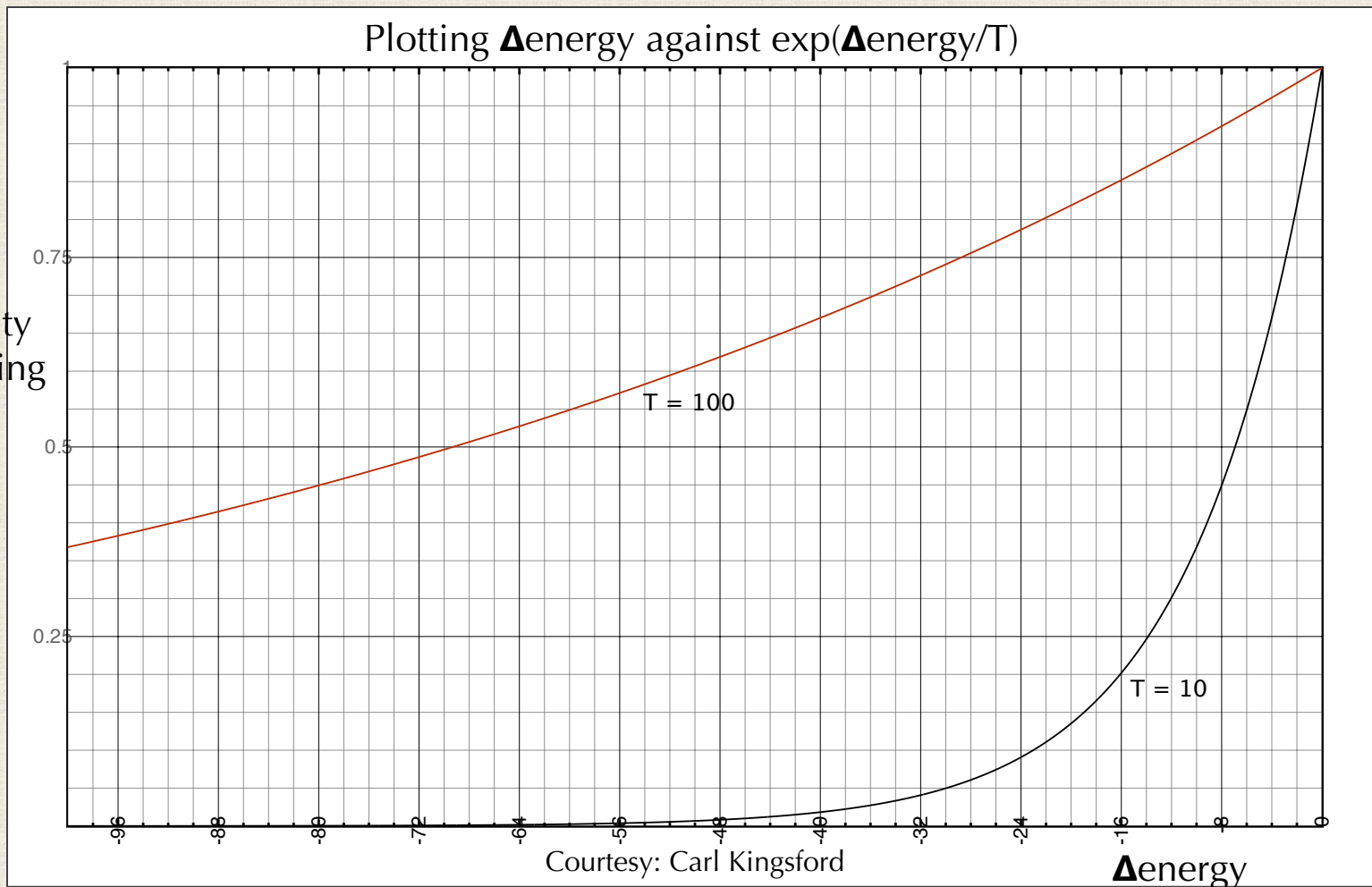
- If $\text{energy}(S') < \text{energy}(S)$, update $S = S'$
- If $\text{energy}(S') > \text{energy}(S)$, then update $S = S'$ with *probability* proportional to $\Delta\text{energy} = \text{energy}(S) - \text{energy}(S')$.

Classic function: $\exp(\Delta\text{energy} / T)$, where T is a “temperature” constant or function. This is called **simulated annealing** because of the analogy of reducing the temperature of a metal slowly.

The “Hotter” the Temperature, the More “Jiggle”



Over time, we lower T , lowering probability of changing structure



The problem with *ab initio* algorithms

Because the search space is so large, and we need to run an algorithm with a lot of initial structures, *ab initio* algorithms still are extremely slow to finish.

The problem with *ab initio* algorithms

Because the search space is so large, and we need to run an algorithm with a lot of initial structures, *ab initio* algorithms still are extremely slow to finish.

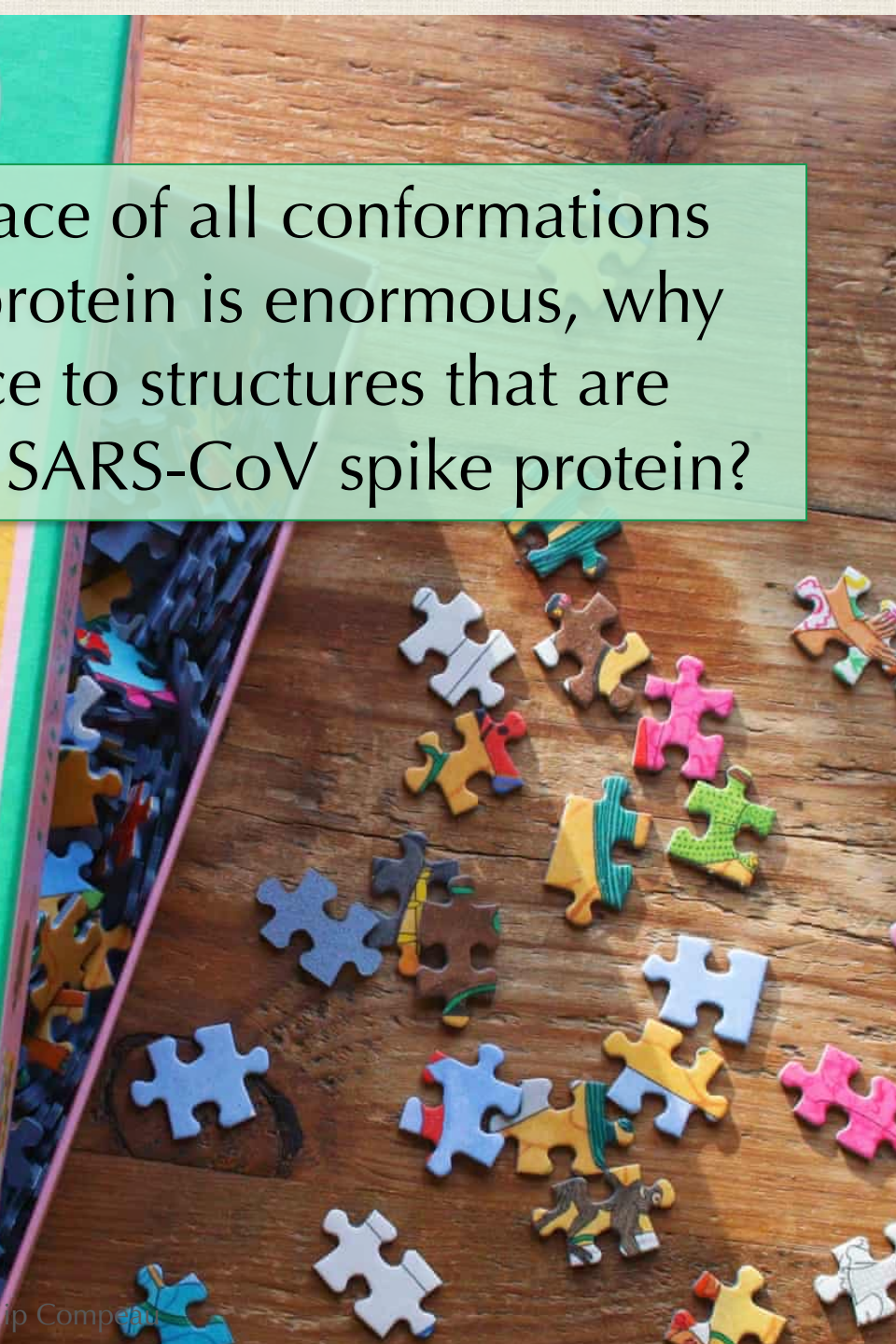
STOP: Say that it's January 2020. Researchers have sequenced and annotated the SARS-CoV-2 genome, but they have not experimentally determined the structure of the spike protein. What might we do?

HOMOLOGY MODELING

DAKAHLO

A jigsaw puzzle by Laura Cooney

Key point: if the search space of all conformations of the SARS-CoV-2 spike protein is enormous, why not restrict the search space to structures that are similar to the shape of the SARS-CoV spike protein?



Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

STOP: If we do not know which template to use before we begin, how could we find a suitable template?

Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

Answer: One natural thing to do would be to search for similar *sequences* for our novel protein in a database using an algorithm like BLAST.

Homology modeling

This idea serves as the foundation of **homology modeling** for protein structure prediction (a.k.a. **comparative modeling**). By using the known protein structure of a homologous protein as a template, we can in theory improve both the accuracy and speed of protein structure prediction.

STOP: Once we have a template, how might we use what we have learned to perform homology modeling?

How does homology modeling work?

One idea is to include an extra “similarity term” in our energy function. The more similar a structure is to the template, the more this similarity term decreases the function we are minimizing.

$$f(S) = \text{energy}(S) - \text{similarity}(S, \text{template})$$

How does homology modeling work?

One idea is to include an extra “similarity term” in our energy function. The more similar a structure is to the template, the more this similarity term decreases the function we are minimizing.

$$f(S) = \text{energy}(S) - \text{similarity}(S, \text{template})$$

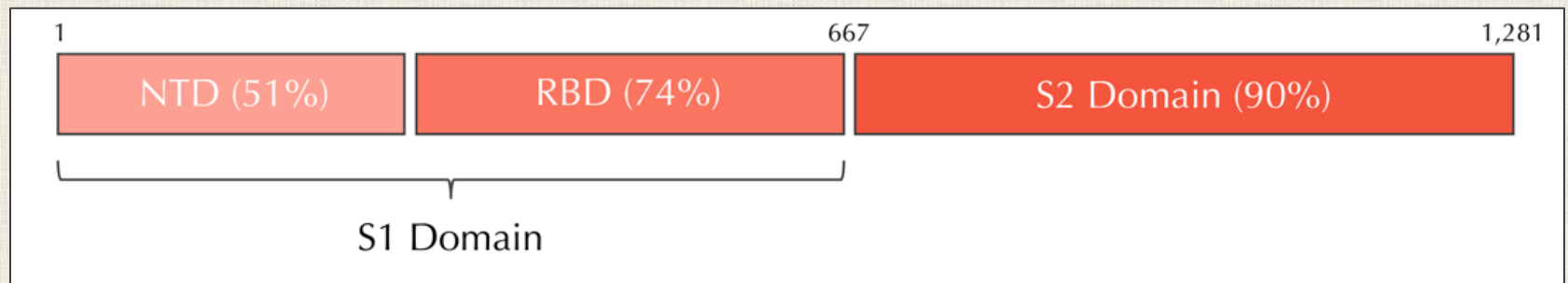
Think of the template protein as “pulling down” nearby structures in the search space.



<https://www.youtube.com/watch?v=cHySqQtB-rk>

How does homology modeling work?

Some algorithms assume that very conserved (similar) regions in two genes correspond to essentially identical structures in the proteins.



How does homology modeling work?

Some algorithms assume that very conserved (similar) regions in two genes correspond to essentially identical structures in the proteins.

We can then use **fragment libraries**, or known protein substructures, to fill in the non-conserved regions and produce a final structure. This approach to homology modeling is called **fragment assembly**.


How does homology modeling work?

Some algorithms assume that very conserved (similar) regions in two genes correspond to essentially identical structures in the proteins.

We can then use **fragment libraries**, or known protein substructures, to fill in the non-conserved regions and produce a final structure. This approach to homology modeling is called **fragment assembly**.

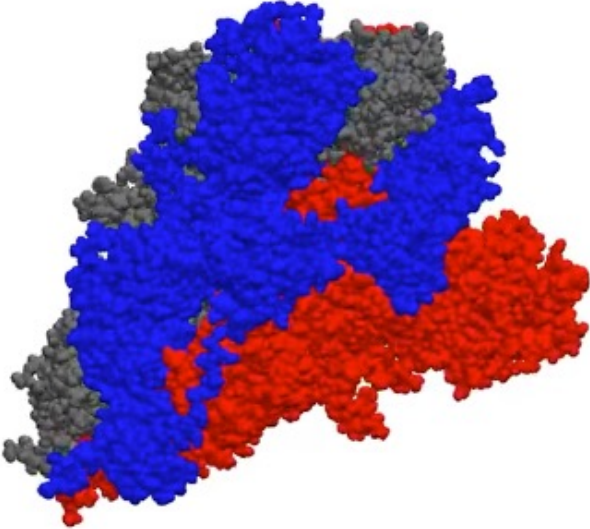
Note: we will use this idea in a SARS-CoV-2 challenge to predict its spike protein structure.

Popular platforms predict structure distributed over many users' computers

 **Greg Bowman** @drGregBowman · Mar 16

As promised, here is our first glimpse of the #COVID19 spike protein (aka the demogorgon) in action, courtesy of @foldingathome . More to come!

<https://twitter.com/drGregBowman/status/1239629911310192640>



0:05 123.7K views

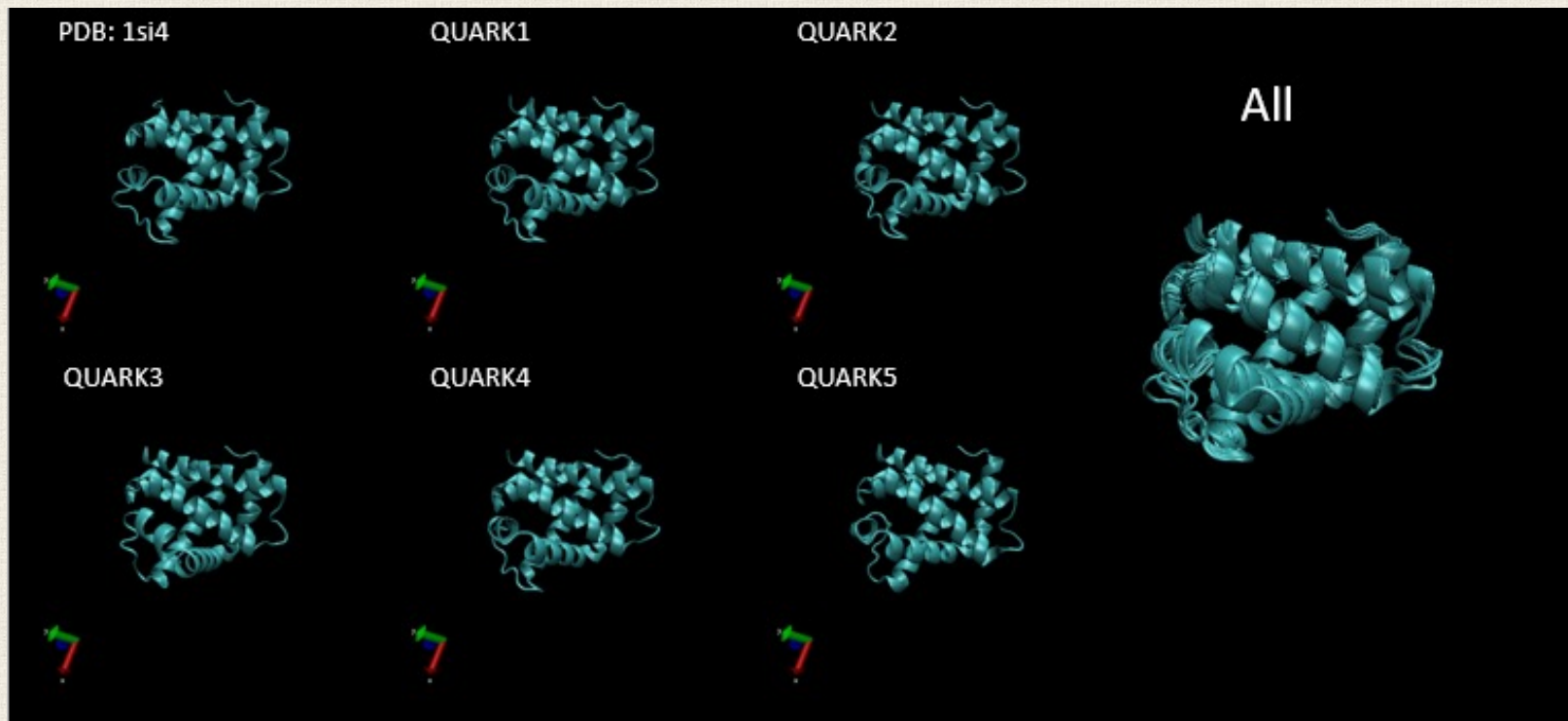
92 1.1K 3K



COMPARING PROTEIN STRUCTURES

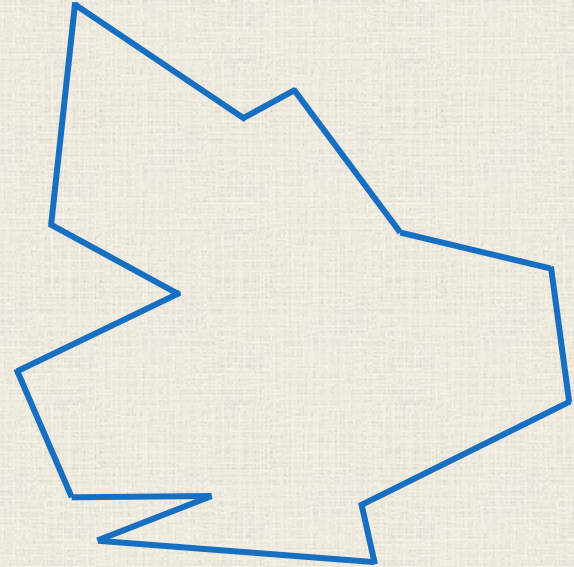
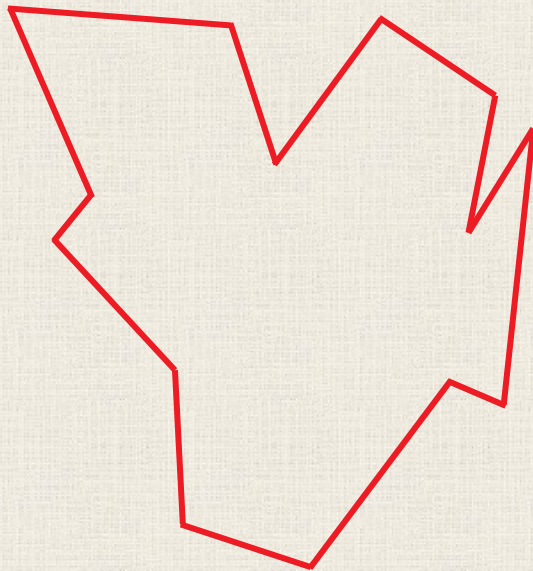
Recall our Second Question

Question 2: How can we compare two similar proteins (e.g., a predicted and experimental structure) quantitatively?



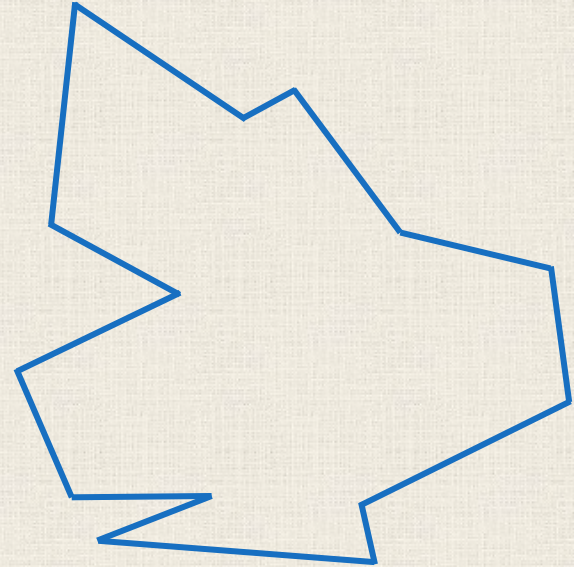
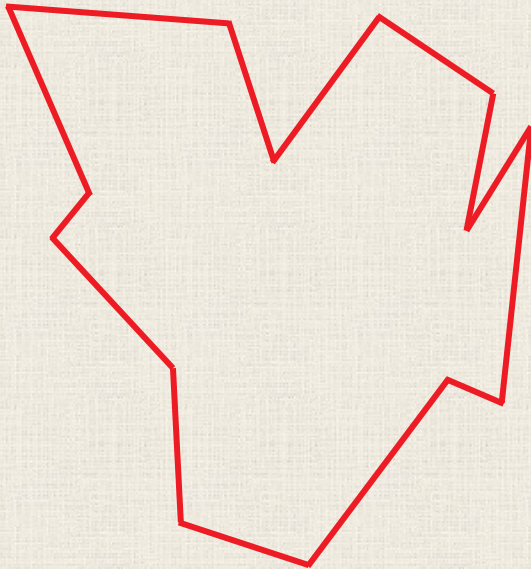
Comparing protein structures is analogous to comparing shapes

Goal: Develop a “distance function $d(S, T)$ that quantifies how different shapes S and T are.



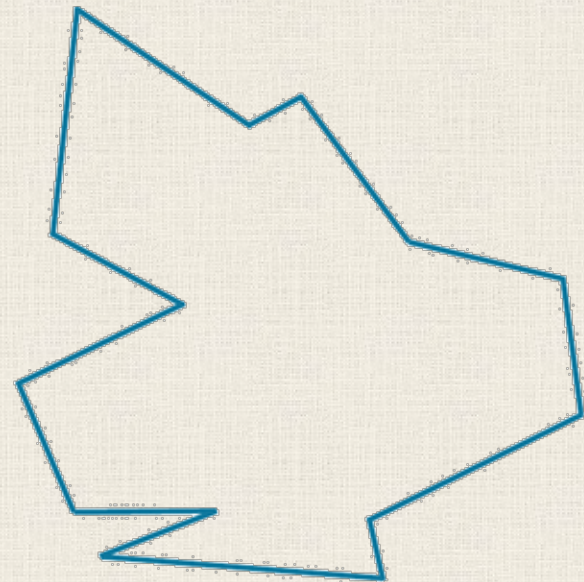
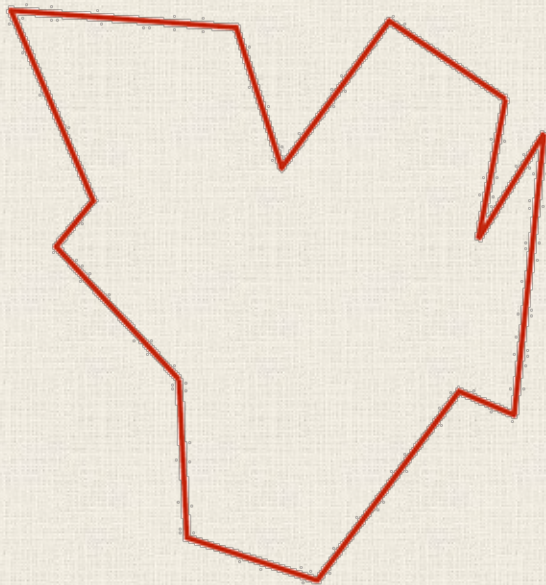
Comparing protein structures is analogous to comparing shapes

STOP: Consider the two shapes in the figure below. How similar are they?



Comparing protein structures is analogous to comparing shapes

Note: The two shapes are in fact the same! We can superimpose/flip/rotate the red shape to see why.



First, align shapes to have same center of mass

Idea: To define $d(S, T)$, first translate/flip/rotate S so that the resulting shape is as similar to T as possible. Then, determine how different the shapes are.

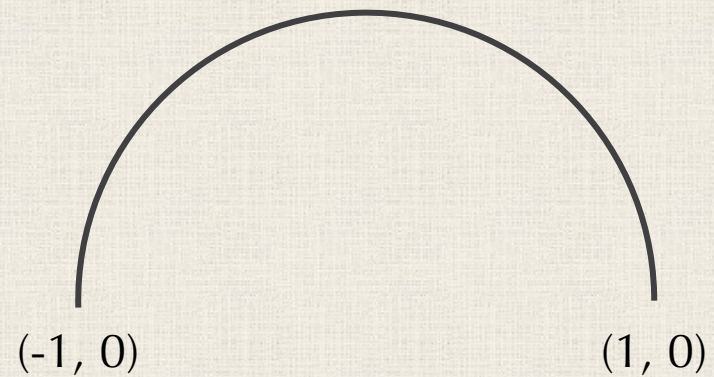
First, align shapes to have same center of mass

Idea: To define $d(S, T)$, first translate/flip/rotate S so that the resulting shape is as similar to T as possible. Then, determine how different the shapes are.

We will first translate S to have the same **centroid** (a.k.a. **center of mass**) as T . The centroid of S is the point (x_S, y_S) such that x_S is the average of x -coordinates on the boundary of S and y_S is the average of y -coordinates on the boundary.

First, align shapes to have same center of mass

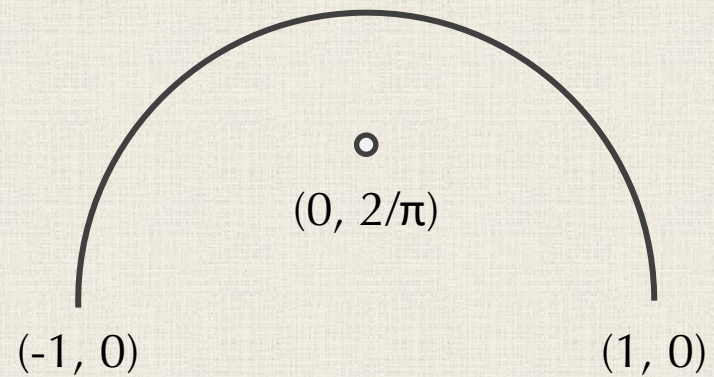
STOP: Let S be the semicircular arc below. What is the centroid of this shape?



First, align shapes to have same center of mass

Answer: The x-coordinate is easy (0), but the y-coordinate is trickier and requires us to integrate over the y-values of the entire semicircle.

$$\begin{aligned} y_S &= \frac{\int_0^\pi \sin \theta}{\pi} \\ &= \frac{-\cos \pi + \cos 0}{\pi} \\ &= \frac{2}{\pi} \end{aligned}$$



Next, rotate and flip S to resemble T as closely as possible

Kabsch algorithm: uses singular value decomposition (matrix algebra) to find flip/rotation of one shape that causes it to be “as similar as possible” to the other shape.

Next, rotate and flip S to resemble T as closely as possible

Kabsch algorithm: uses singular value decomposition (matrix algebra) to find flip/rotation of one shape that causes it to be “as similar as possible” to the other shape.

That is, we must be looking for a rotation/flip minimizing some function between the two shapes. But what function?

Determining Similarity of Aligned Shapes with RMSD

Sample n points along the boundary of S and T , converting S and T into *vectors* $s = (s_1, \dots, s_n)$ and $t = (t_1, \dots, t_n)$.

Determining Similarity of Aligned Shapes with RMSD

Sample n points along the boundary of S and T , converting S and T into vectors $s = (s_1, \dots, s_n)$ and $t = (t_1, \dots, t_n)$.

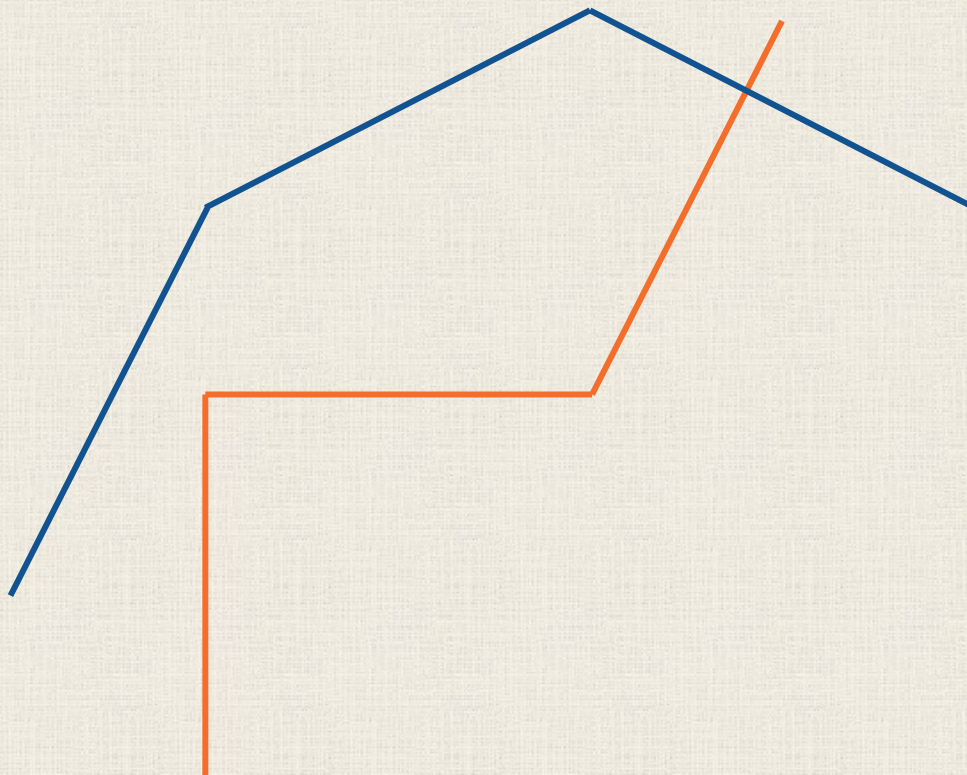
We then compute the **root mean square deviation (RMSD)** between the two shapes,

$$\text{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \dots + d(s_n, t_n)^2)}$$

the square root of the average squared distance between corresponding points in the vectors.

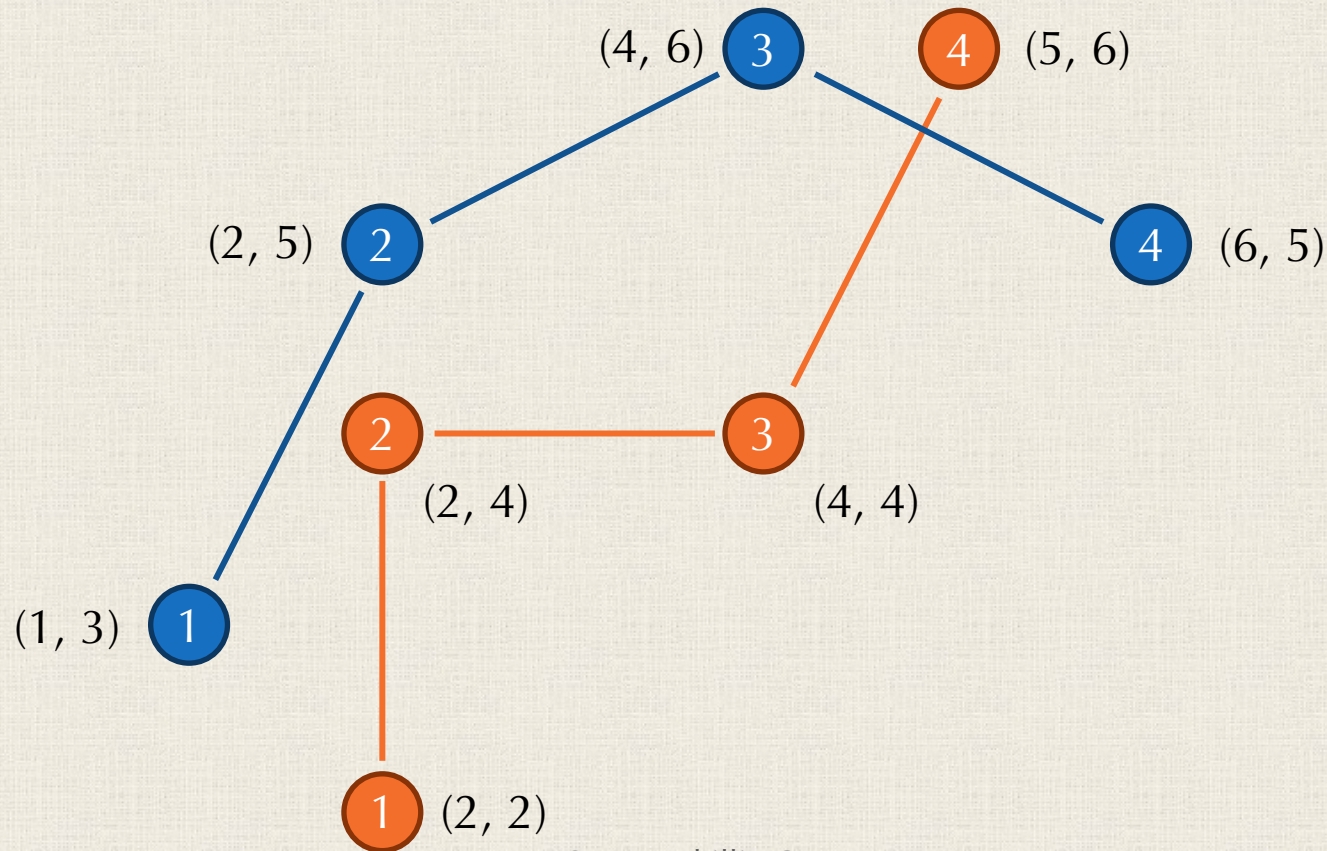
An example of computing RMSD

Consider the two shapes shown below.



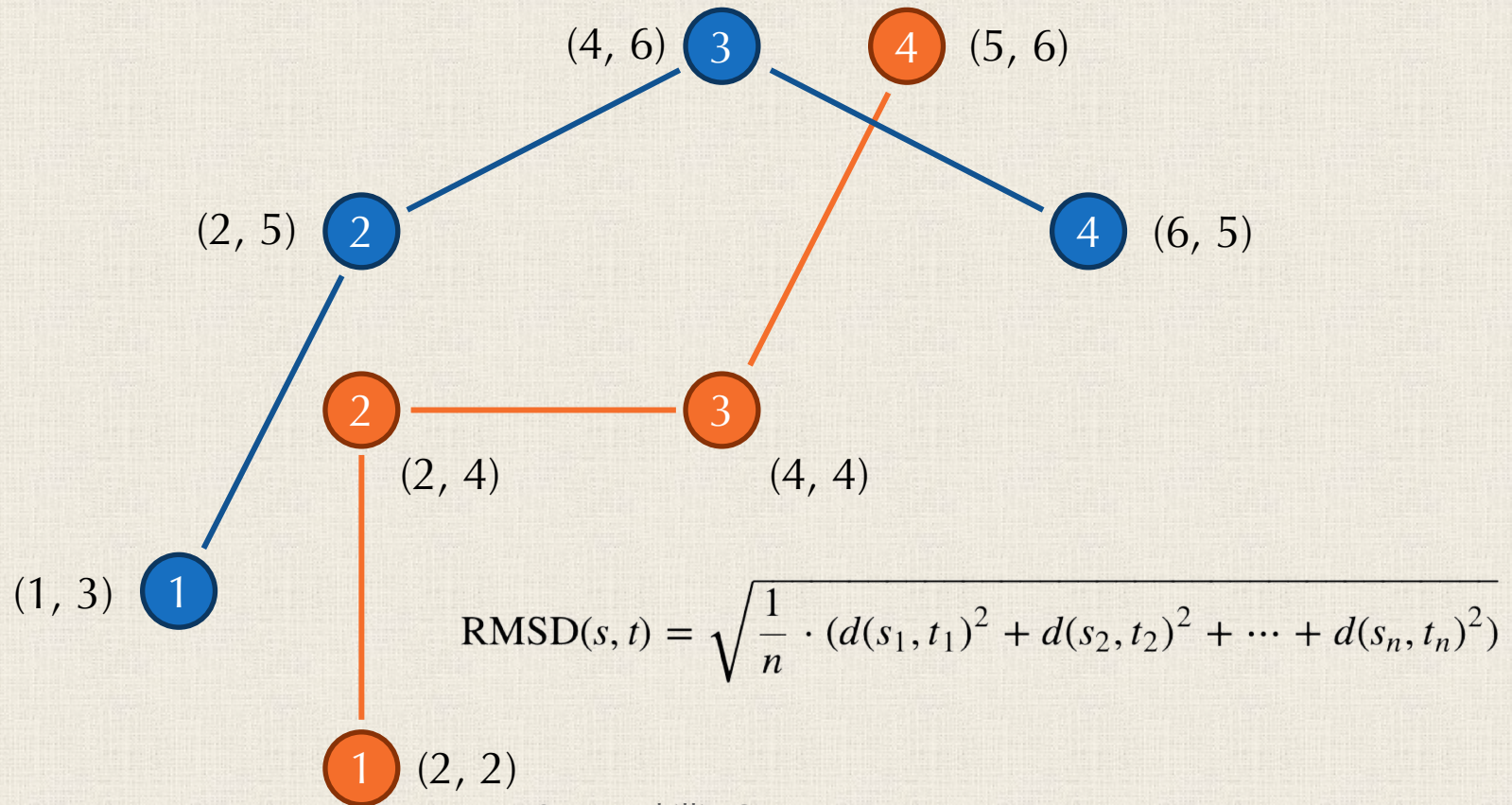
An example of computing RMSD

We vectorize by sampling $n = 4$ points from each.



An example of computing RMSD

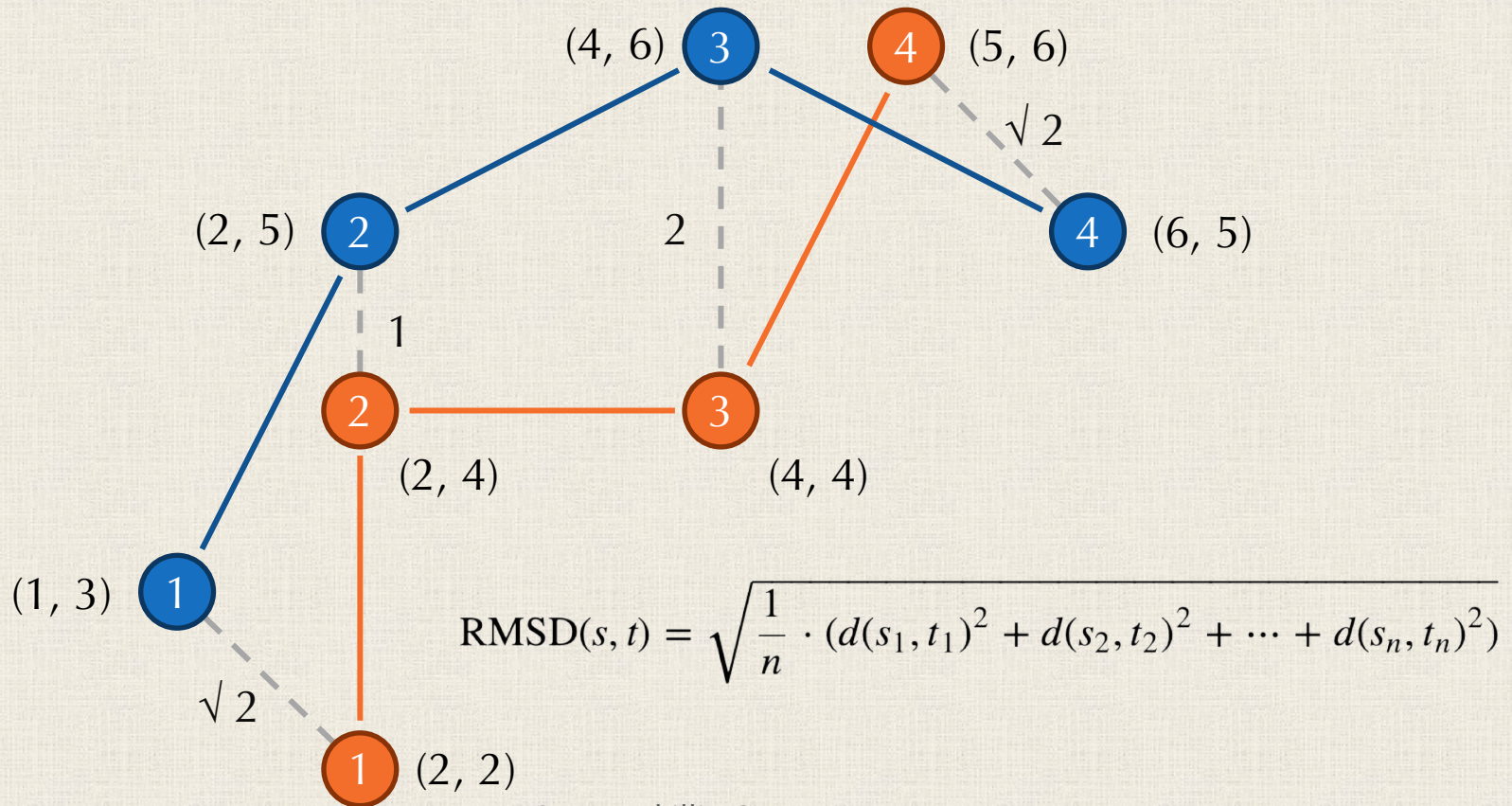
Exercise: Compute the RMSD for this example.



$$\text{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \dots + d(s_n, t_n)^2)}$$

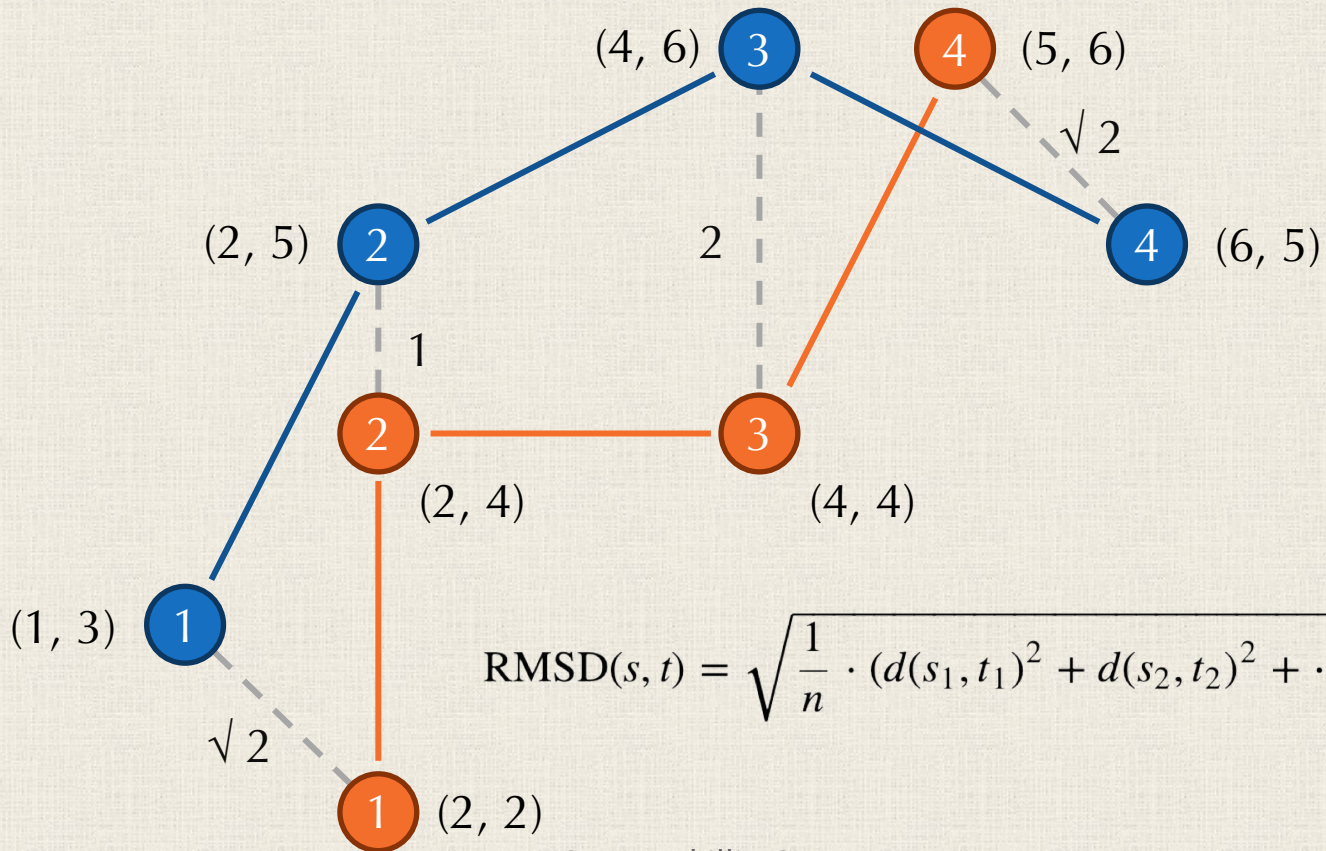
An example of computing RMSD

We first find the distances between corresponding points.



An example of computing RMSD

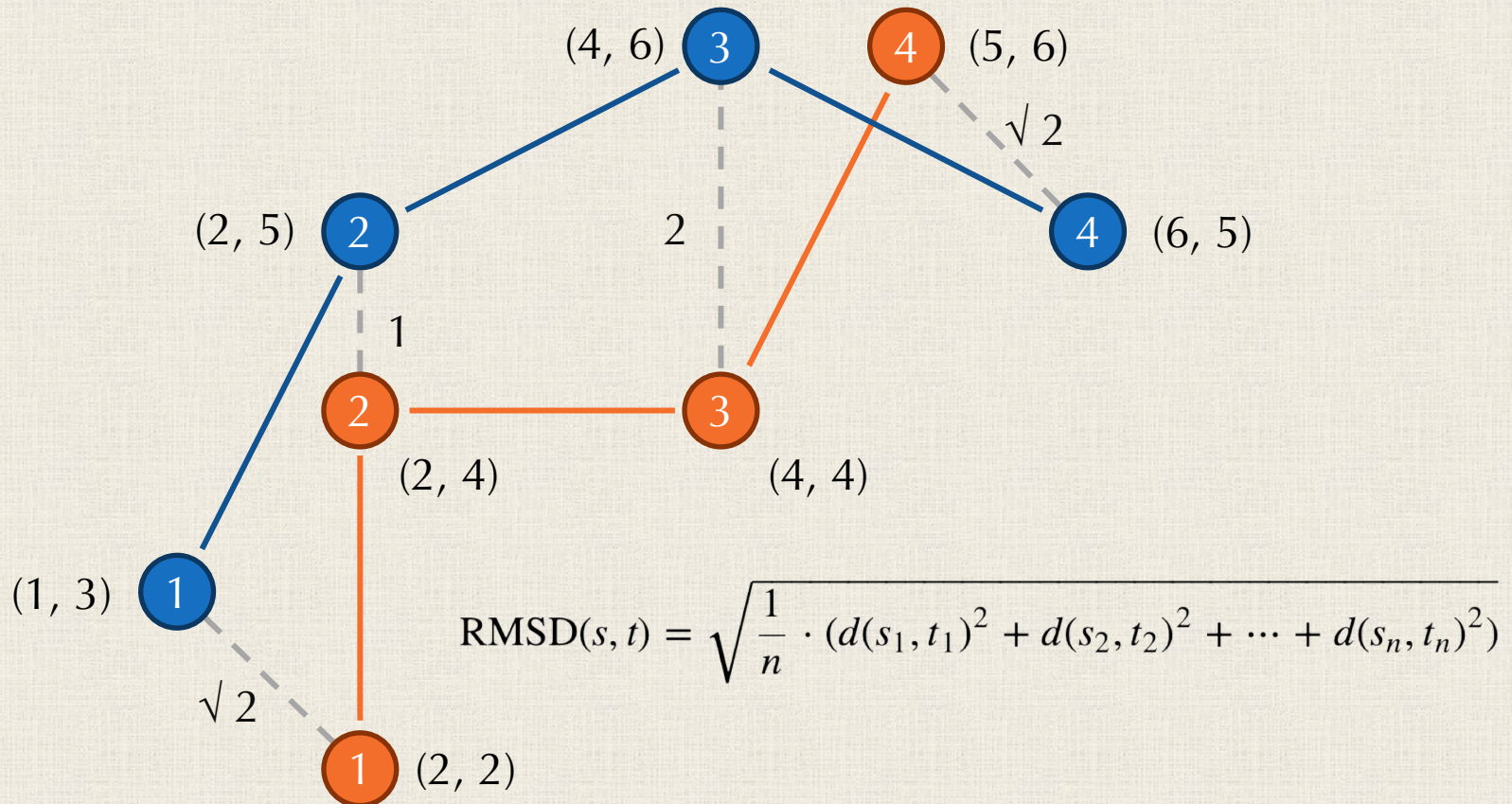
$$\text{RMSD} = \sqrt{\left(\frac{1}{4}\right) \cdot (2 + 1 + 4 + 2)} = \sqrt{\frac{9}{4}} = \frac{3}{2}.$$



$$\text{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \dots + d(s_n, t_n)^2)}$$

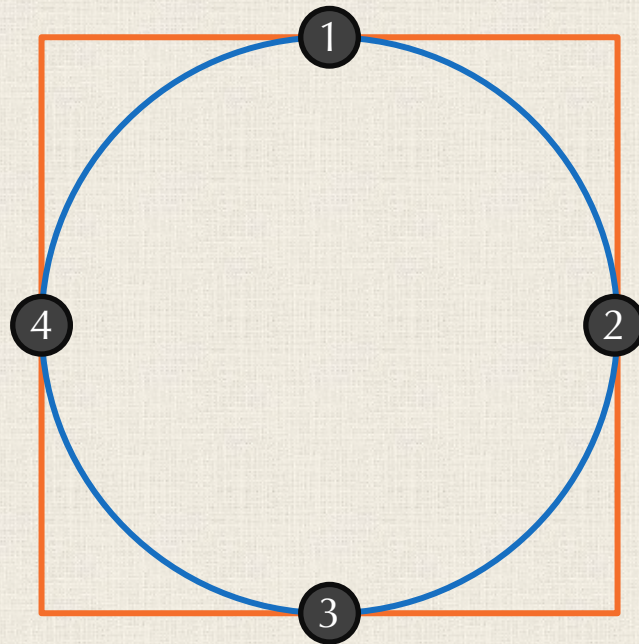
An example of RMSD

STOP: Do you see any issues with using RMSD?



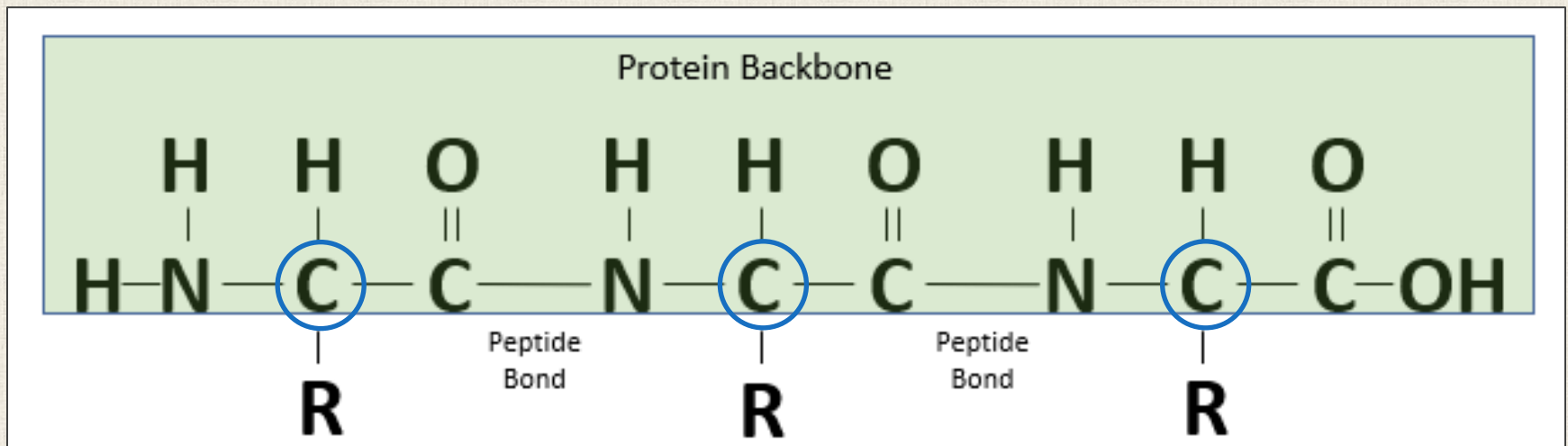
Undersampling can cause issues

Because we didn't sample enough points here, RMSD is zero, but the shapes are not the same.



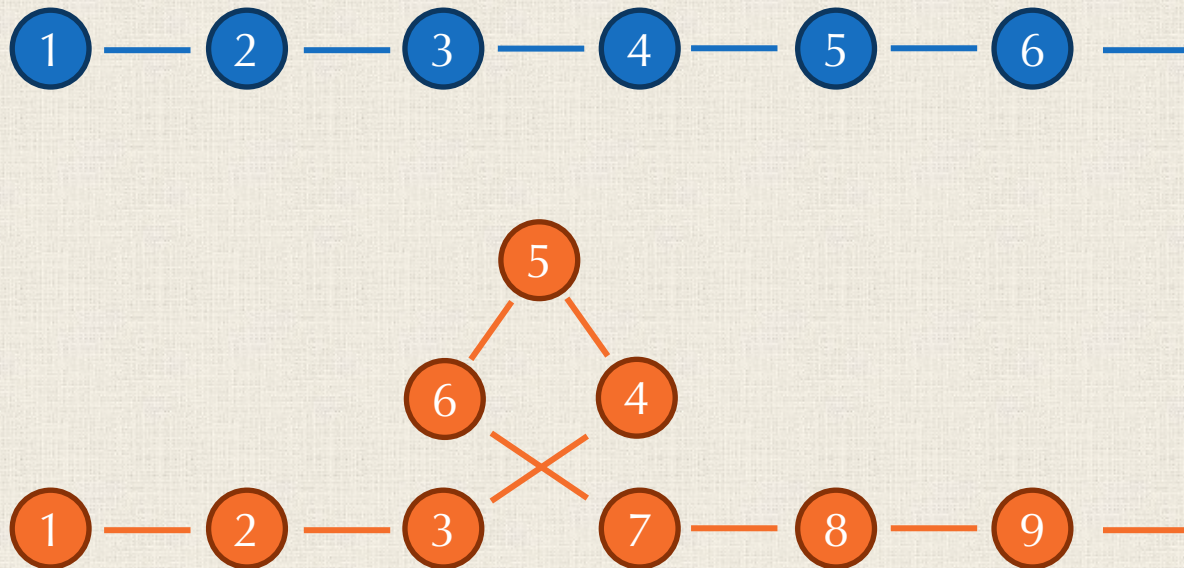
Undersampling can cause issues

In practice, researchers take the “alpha carbon” atom from each amino acid to vectorize a structure and prevent undersampling.



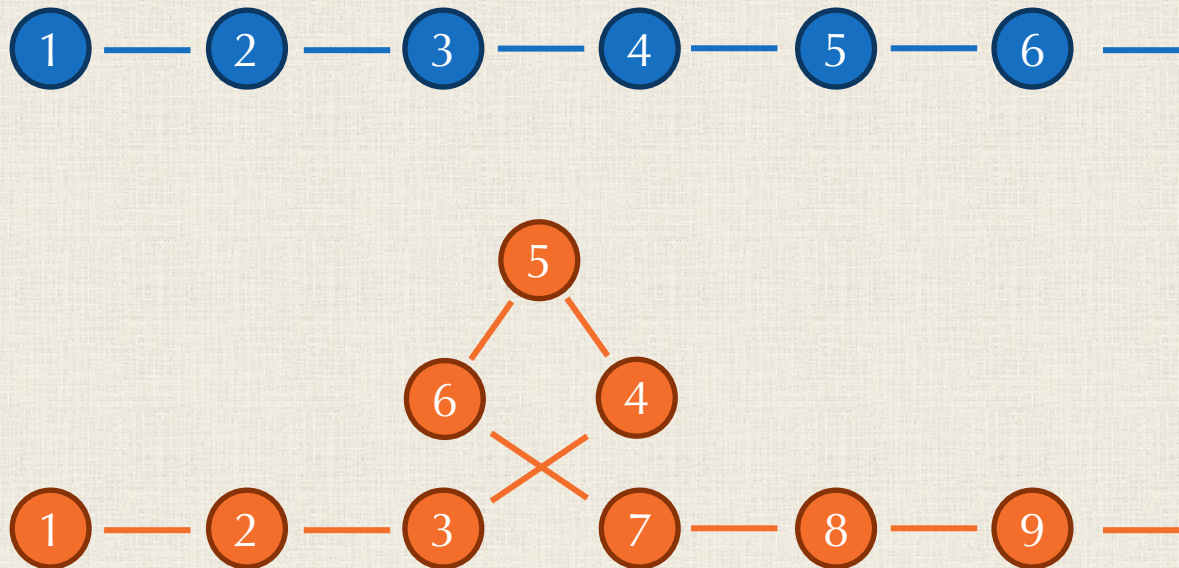
Comparing proteins of differing lengths

The situation below (an inserted substructure) would throw off RMSD for every alpha carbon after #2.



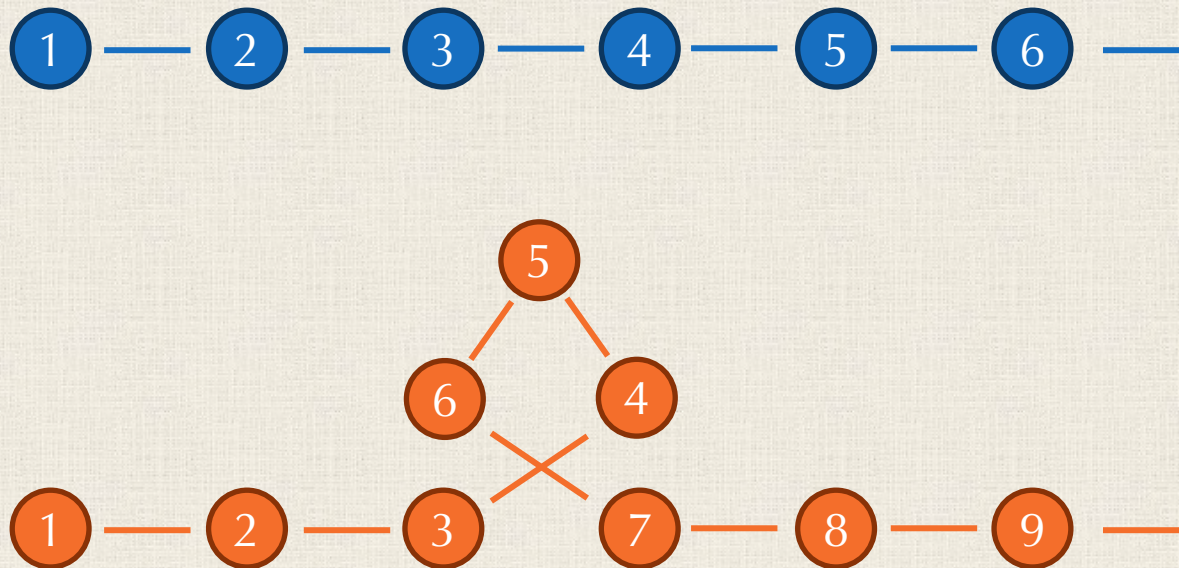
Comparing proteins of differing lengths

STOP: Any ideas on how we could handle situations like this?



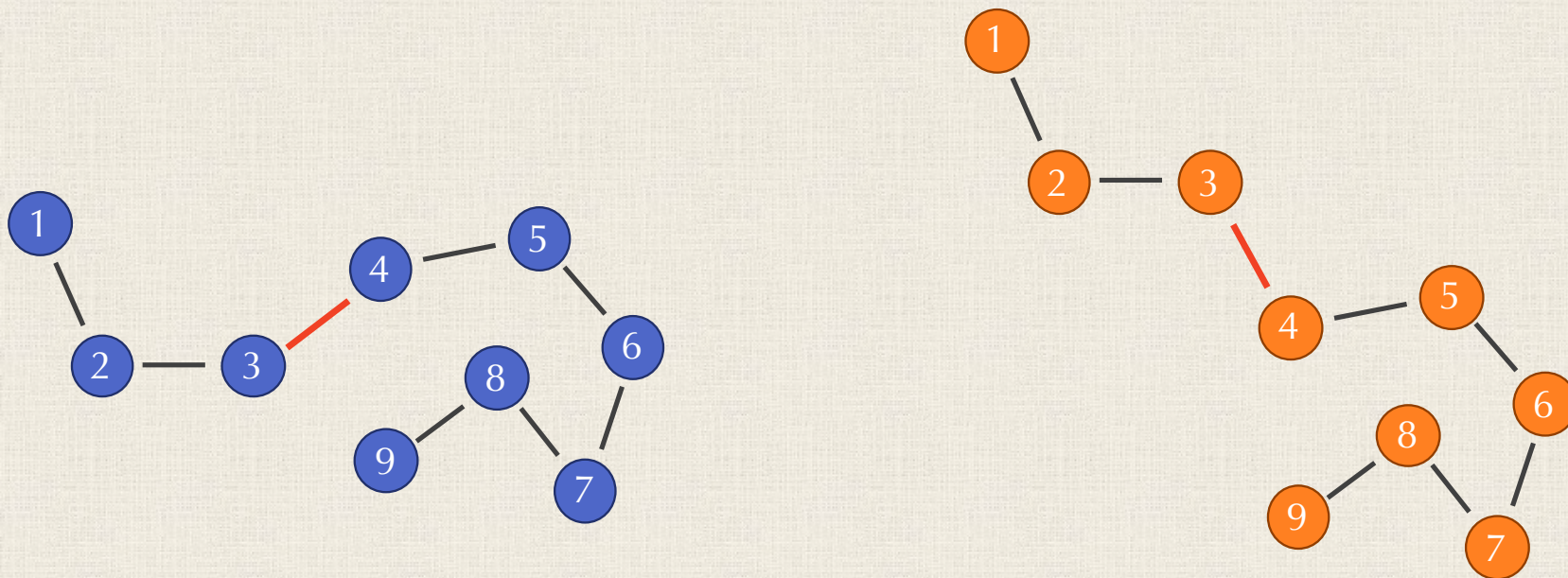
Comparing proteins of differing lengths

Answer: First, we align the protein *sequences*; then, any gap columns will not contribute to RMSD.



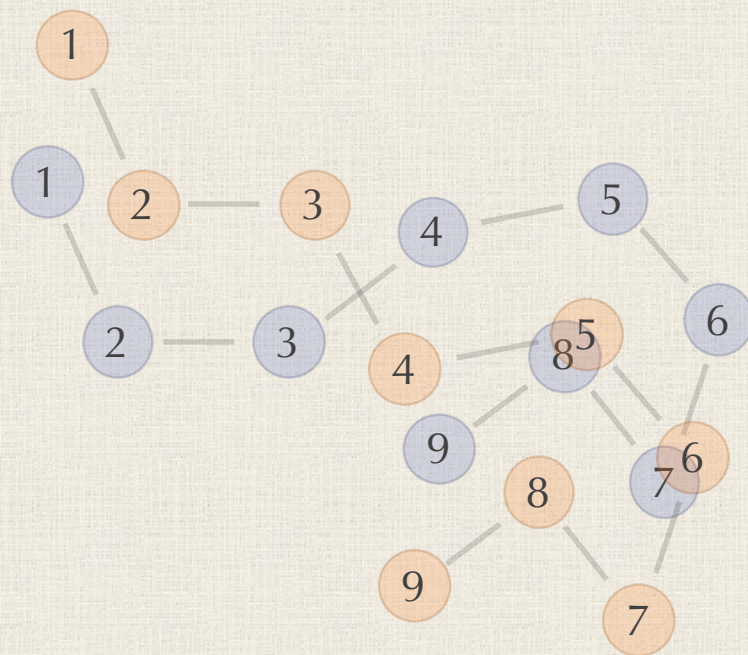
Small protein changes can have a huge impact on RMSD

Here are two protein structures that are *identical* except for changing a single bond angle (red).



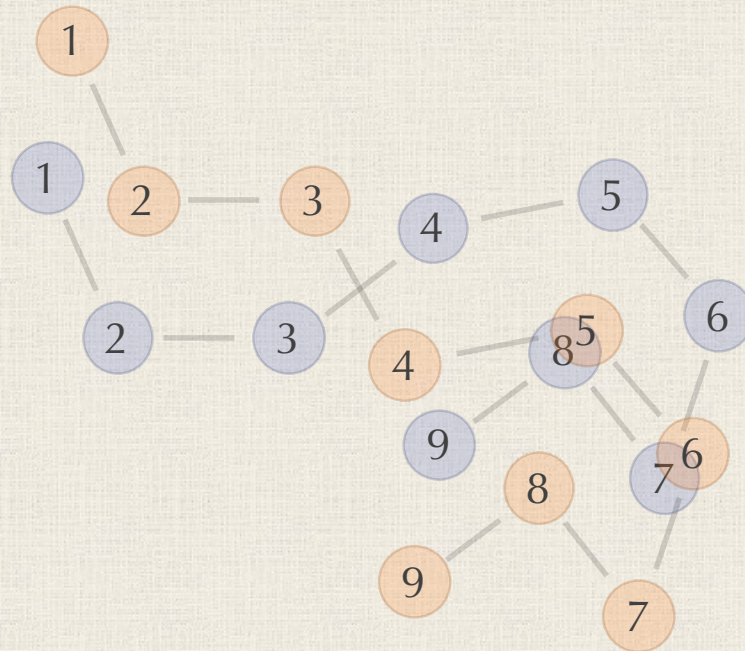
Small protein changes can have a huge impact on RMSD

The Kabsch algorithm will align proteins as shown on the right and miss the similarities.



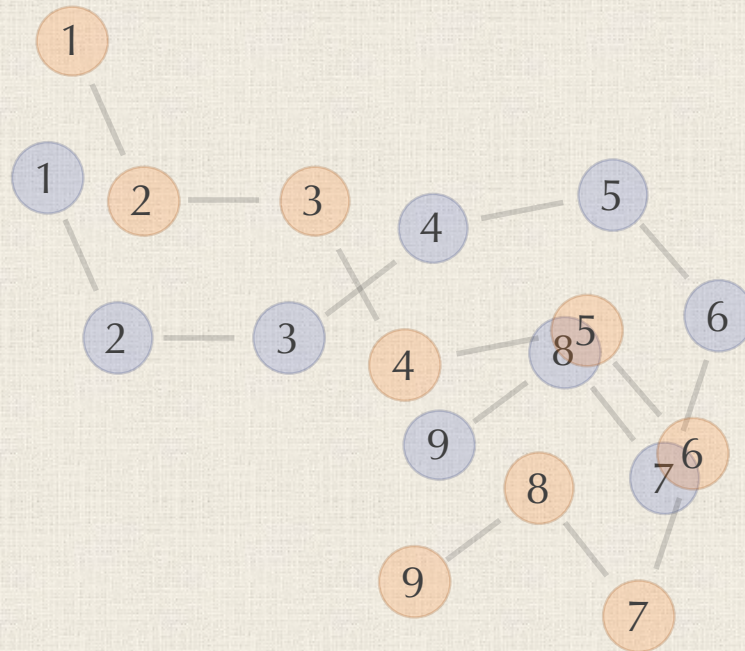
Comparing structures locally

We also haven't discussed how to compare structures *locally*; i.e., at the same position.



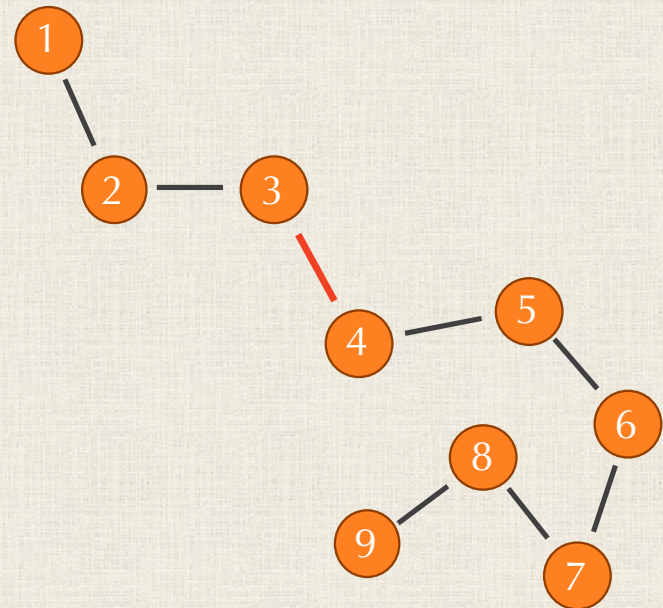
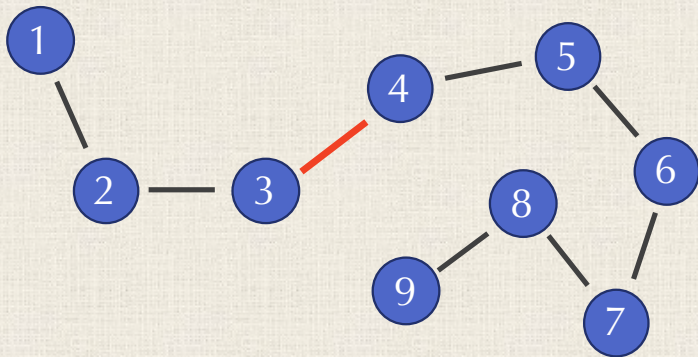
Comparing structures locally

STOP: Why would $d(s_i, t_i)$ be a bad comparison at the i -th alpha carbon? (Hint: look at $i = 6$.)



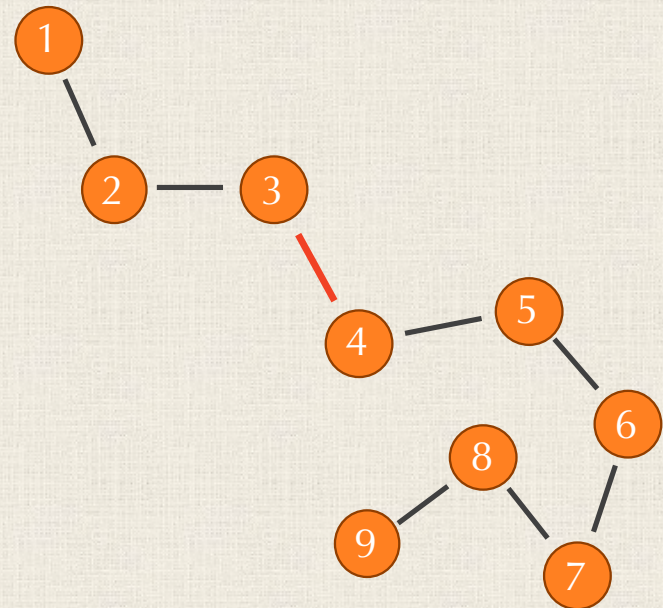
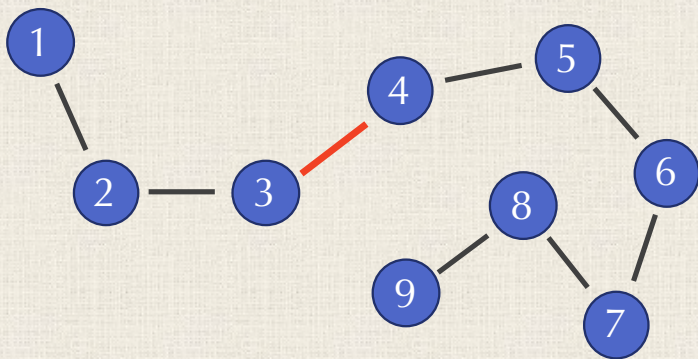
Comparing structures locally

Answer: The proteins aren't really different most spots (positions 1-3, 4-9 are identical substructures).



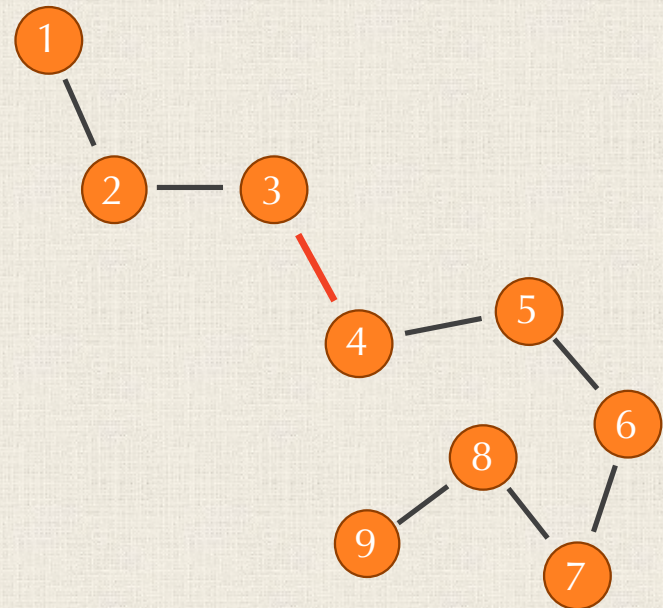
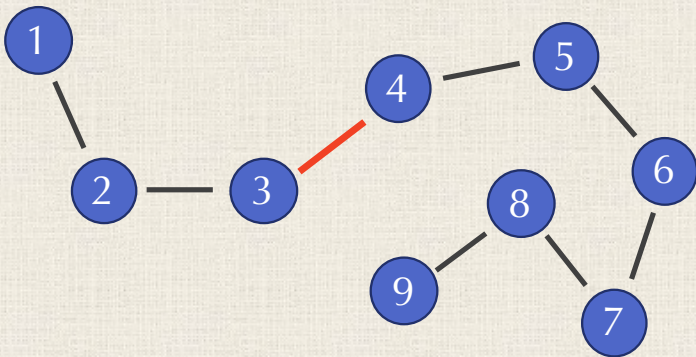
Comparing structures locally

STOP: Do you have any ideas for a better way of comparing structures locally?



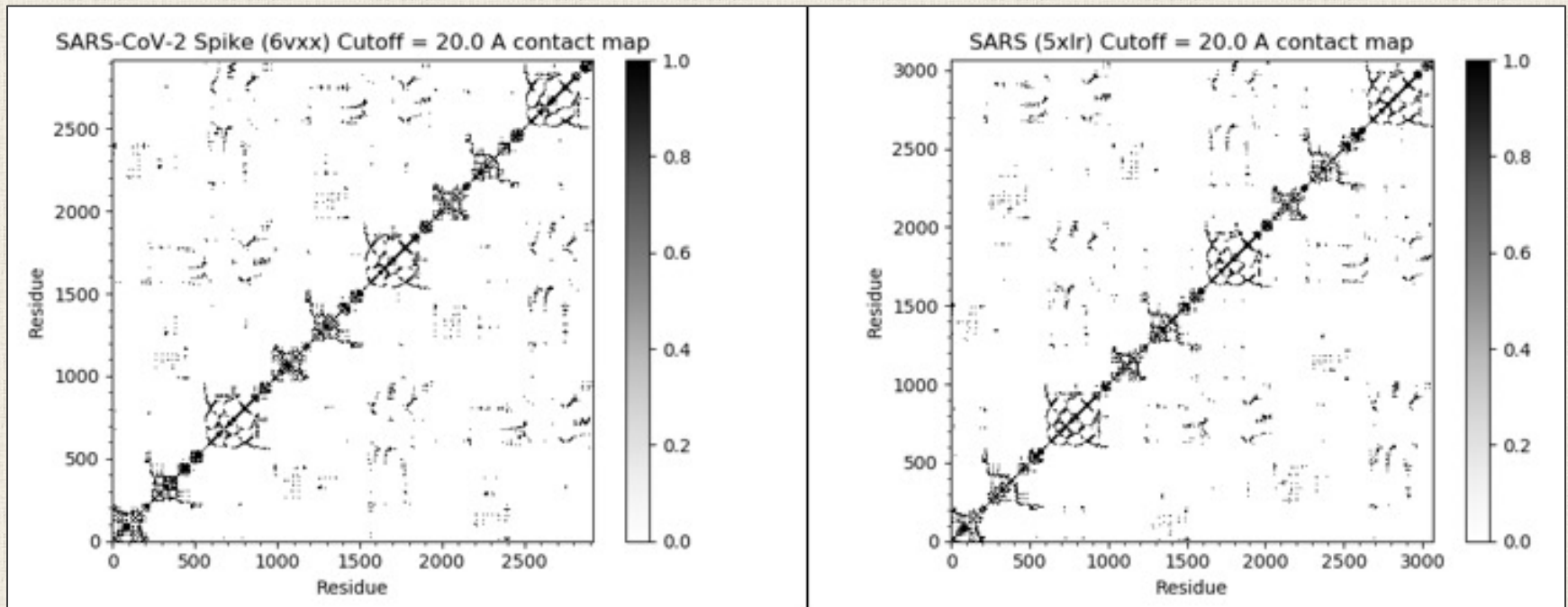
Comparing structures locally

Note: The set of *intraprotein* distances $d(s_6, s_j)$ is similar to the distances $d(t_6, t_j)$.



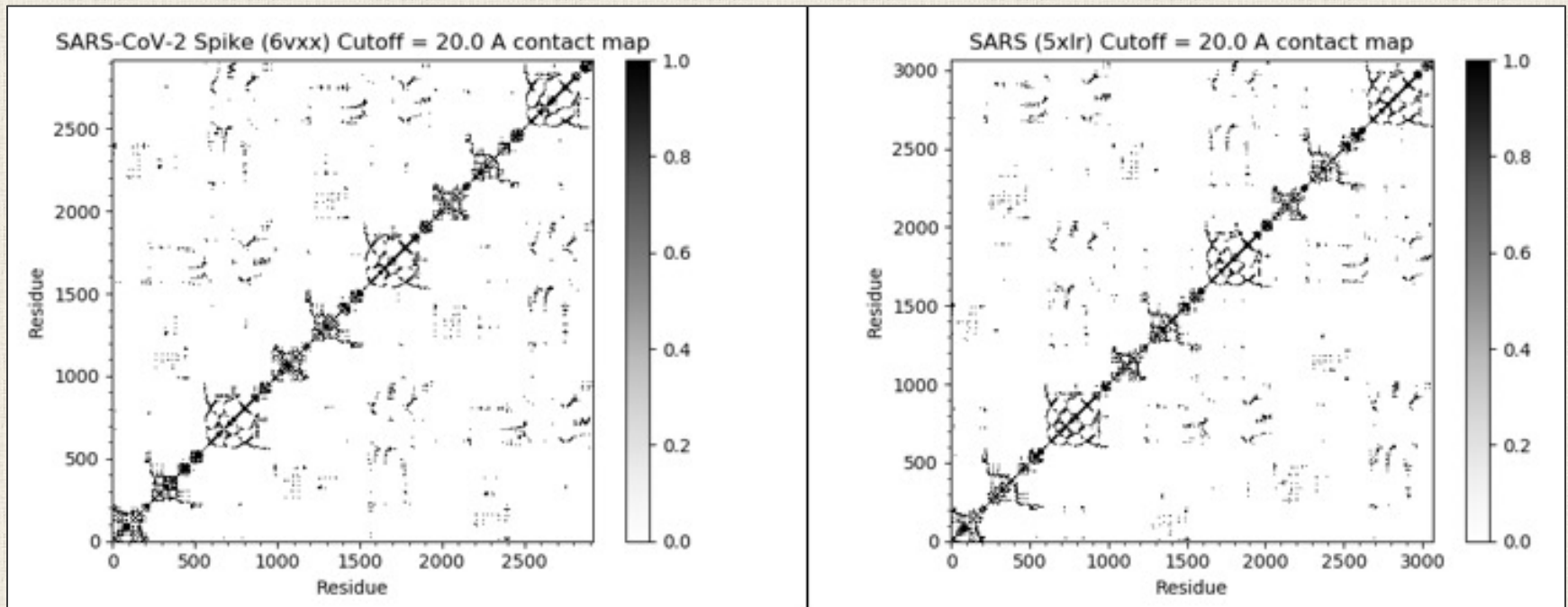
Contact maps help us visualize intraprotein distances

Contact map: for some threshold t , given a structure S , color cell (i, j) black if $d(s_i, s_j) < t$ and white otherwise.



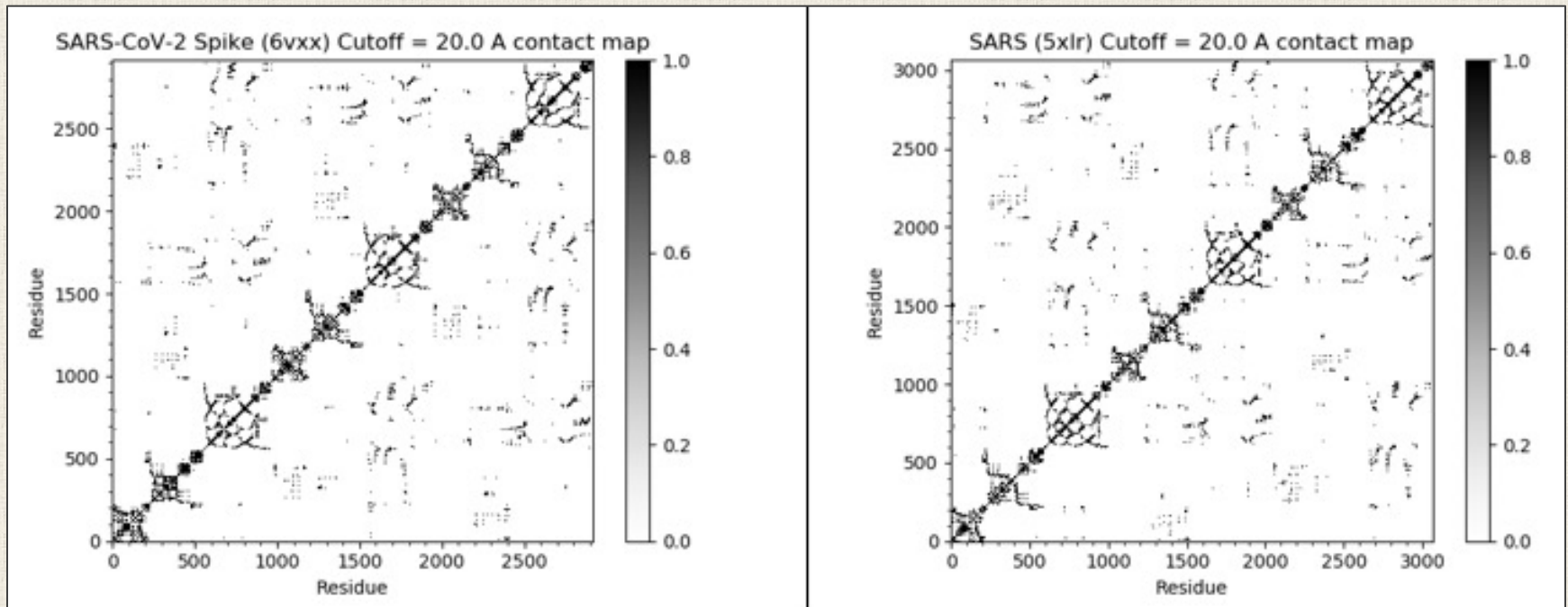
Contact maps help us visualize intraprotein distances

STOP: How might we use a contact map to look for local regions of similarity in protein structures?



Contact maps help us visualize intraprotein distances

Answer: Comparing the i -th row over two maps tells us whether to investigate differences at the i -th amino acid.



Q per residue offers a single value for how much two proteins differ locally

Q per residue (Q_{res}): defined as follows.

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

- N is the number of amino acids in each protein;
- k is equal to 2 when i is at either the start or the end of the protein, and k is equal to 3 otherwise;
- the variance term $\sigma_{i,j}^2$ is equal to $|i - j|^{0.15}$, so that nearby amino acids have more influence.

Q per residue offers a single value for how much two proteins differ locally

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

STOP: What happens to the interior term of the sum if $d(s_i, s_j)$ is comparable to $d(t_i, t_j)$?

Q per residue offers a single value for how much two proteins differ locally

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

STOP: What happens to the interior term of the sum if $d(s_i, s_j)$ is comparable to $d(t_i, t_j)$?

Answer: It heads toward $\exp(0) = 1$.

Q per residue offers a single value for how much two proteins differ locally

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

STOP: What happens to the interior term of the sum if $d(s_i, s_j)$ is very different to $d(t_i, t_j)$?

Q per residue offers a single value for how much two proteins differ locally

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

STOP: What happens to the interior term of the sum if $d(s_i, s_j)$ is very different to $d(t_i, t_j)$?

Answer: It heads toward $\exp(-\infty) = 0$.

Q per residue offers a single value for how much two proteins differ locally

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

STOP: So, what are the possible values of Q_{res} ?

Q per residue offers a single value for how much two proteins differ locally

$$Q_{res}^{(i)} = \frac{1}{N - k} \sum_{\substack{\text{residues} \\ j \neq i-1, i, i+1}} \exp\left[-\frac{[d(s_i, s_j) - d(t_i, t_j)]^2}{2\sigma_{i,j}^2}\right]$$

STOP: So, what are the possible values of Q_{res} ?

Answer: Q_{res} ranges from 0 when proteins are very different at the i -th position, to 1 when proteins are identical at the i -th position.

**PROTEIN STRUCTURE PREDICTION
IS SOLVED! (KINDA?)**

CASP contests establish best structure prediction algorithms

Critical Assessment of protein Structure Prediction (CASP): contest run every two years since 1994 that tests structure prediction algorithms against each other on known (hidden) protein structures.

CASP contests establish best structure prediction algorithms

Critical Assessment of protein Structure Prediction (CASP): contest run every two years since 1994 that tests structure prediction algorithms against each other on known (hidden) protein structures.

CASP14 (2020) was dominated by “AlphaFold”, a deep learning algorithm produced by DeepMind.



Instead of RMSD, CASP scores a predicted structure using a different test

For some threshold t , we first take the percentage of alpha carbon positions for which the distance between corresponding alpha carbons in the two structures is at most t .

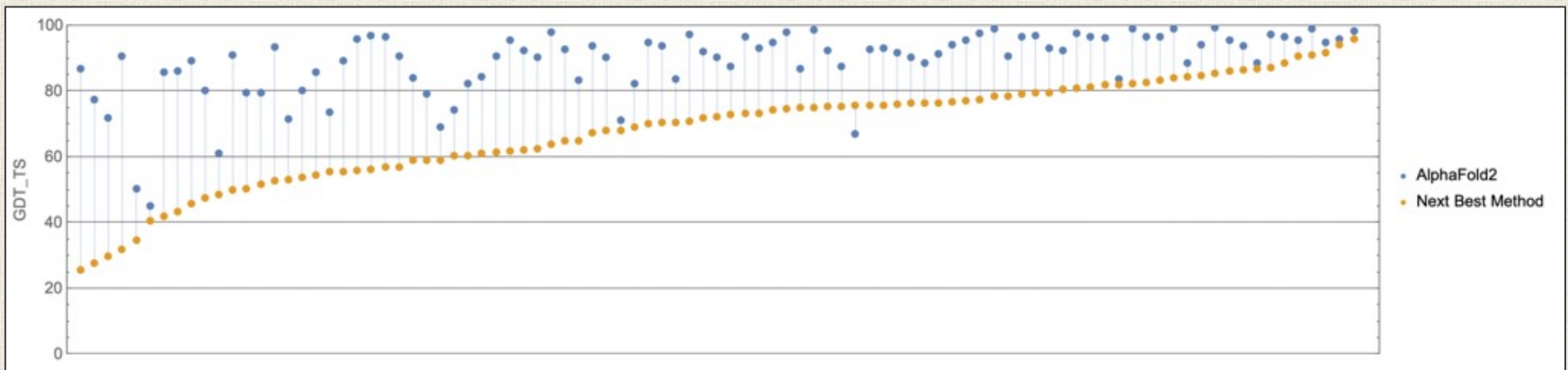
Instead of RMSD, CASP scores a predicted structure using a different test

For some threshold t , we first take the percentage of alpha carbon positions for which the distance between corresponding alpha carbons in the two structures is at most t .

The **global distance test (GDT)** score averages the percentages obtained when t is equal to each of 1, 2, 4, and 8 angstroms. A GDT score of 90% is good, and a score of 95% is excellent (comparable to minor experimental errors).

So, how well did AlphaFold do?

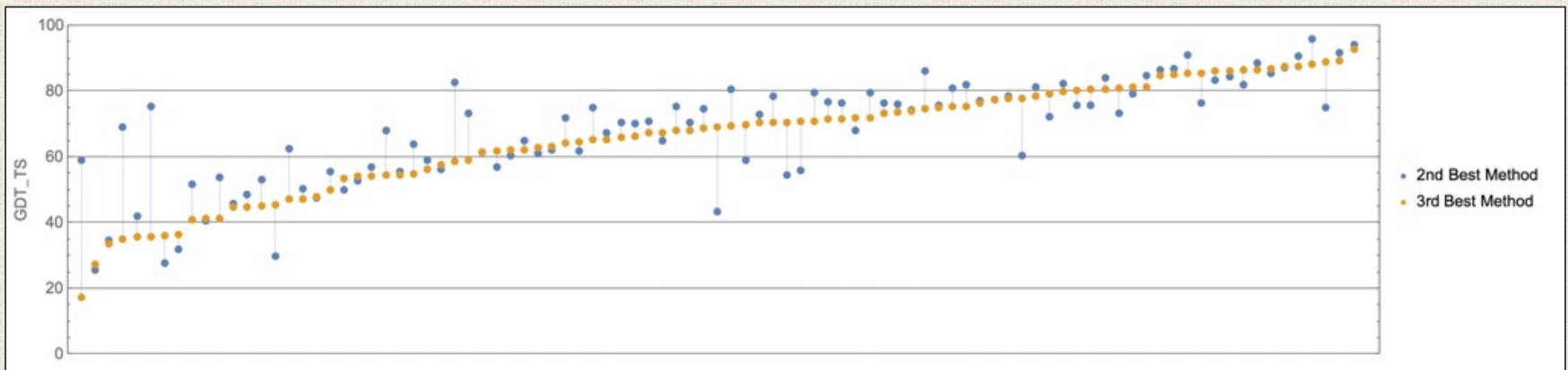
Here's the plot of GDT scores for AlphaFold (blue) and the 2nd place method (orange), produced by same lab that developed Rosetta@Home.



Source: Mohammed AlQuraishi, <https://bit.ly/39Mnym3>.

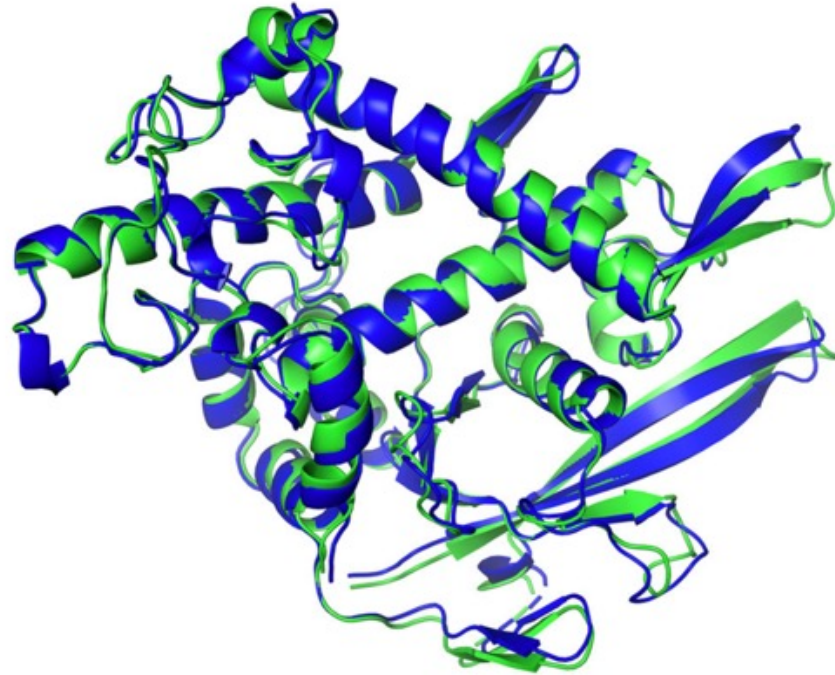
So, how well did AlphaFold do?

To show how decisive the victory is, here is 2nd place vs. the 3rd place method (submitted by the Yang Zhang lab).



Source: Mohammed AlQuraishi, <https://bit.ly/39Mnym3>.

DeepMind received lots of positive press



Structures of a protein that were predicted by artificial intelligence (blue) and experimentally determined (green) match almost perfectly. DEEPMIND

'The game has changed.' AI triumphs at solving protein structures

By **Robert F. Service** | Nov. 30, 2020, 10:30 AM

Science

But some scientists remain skeptical

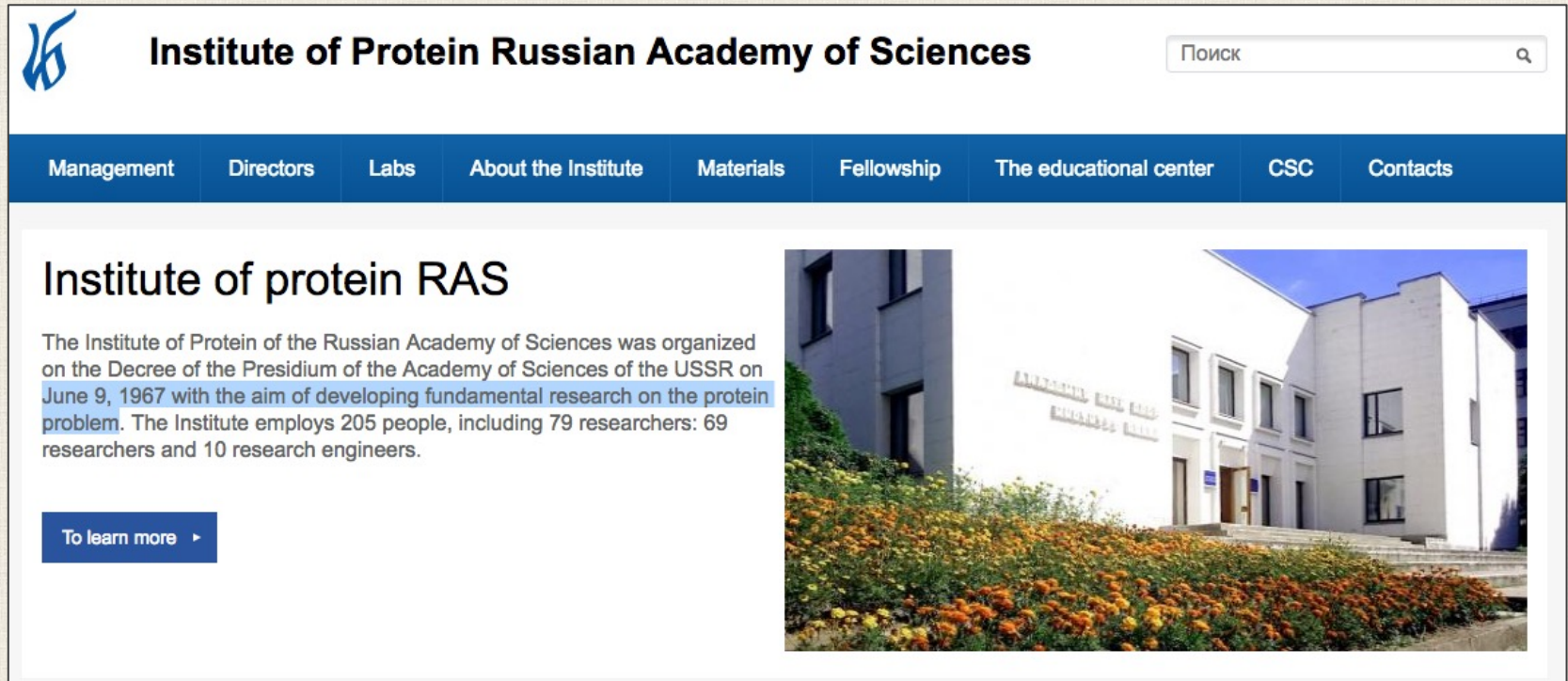
AlphaFold obtained a median RMSD of 1.6, but to be trustworthy for a sensitive application like designing drug targets, it would need an RMSD about 90% lower.

But some scientists remain skeptical

AlphaFold obtained a median RMSD of 1.6, but to be trustworthy for a sensitive application like designing drug targets, it would need an RMSD about 90% lower.

~1/3 of AlphaFold's CASP14 predictions have an RMSD over 2.0, an often-used threshold for whether a predicted structure is reliable. And there is no way of knowing in advance whether AlphaFold will perform well on a given protein, unless we validate the protein's structure, which causes a catch-22.

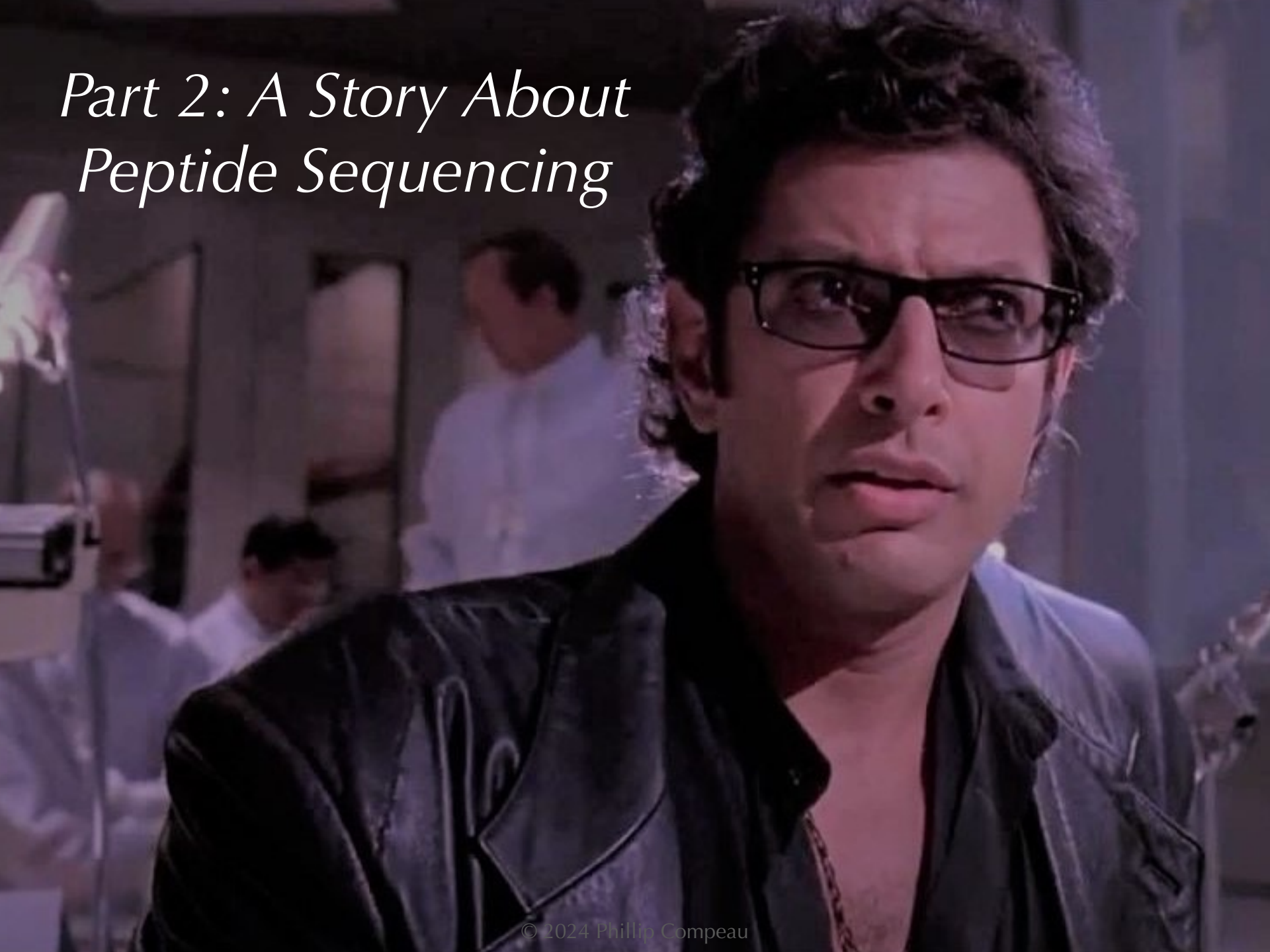
AlphaFold is Still Pretty, Pretty Good



The screenshot shows the website for the Institute of Protein Russian Academy of Sciences. At the top left is the institute's logo, a stylized blue 'IP' monogram. To its right is the text 'Institute of Protein Russian Academy of Sciences'. Further right is a search bar with the Russian word 'Поиск' (Search) and a magnifying glass icon. Below this is a dark blue navigation bar with white text for 'Management', 'Directors', 'Labs', 'About the Institute', 'Materials', 'Fellowship', 'The educational center', 'CSC', and 'Contacts'. The main content area features the title 'Institute of protein RAS' in a large, bold, black font. Below the title is a paragraph of text: 'The Institute of Protein of the Russian Academy of Sciences was organized on the Decree of the Presidium of the Academy of Sciences of the USSR on June 9, 1967 with the aim of developing fundamental research on the protein problem. The Institute employs 205 people, including 79 researchers: 69 researchers and 10 research engineers.' A blue button with white text 'To learn more' and a right-pointing arrow is positioned below the text. To the right of the text is a photograph of a modern, white, multi-story building with large windows and a flat roof. The building is surrounded by a lush garden of orange and yellow flowers in the foreground.

Nevertheless, we may never again see such an *improvement* to the state of the art in a problem that has puzzled biologists for fifty years.

*Part 2: A Story About
Peptide Sequencing*



Let's Hear From Karl Pilkington on the Infinite Monkey Theorem




<https://www.youtube.com/watch?v=FWs0ujLrGI0>

Karl is a wise man

The New York Times

Monkeys-s-s Typing Is-s a Mess-s-s

 Give this article



By The Associated Press

May 10, 2003

1 MIN READ

Give an infinite number of monkeys an infinite number of typewriters, the theory goes, and they will eventually produce the complete works of Shakespeare.

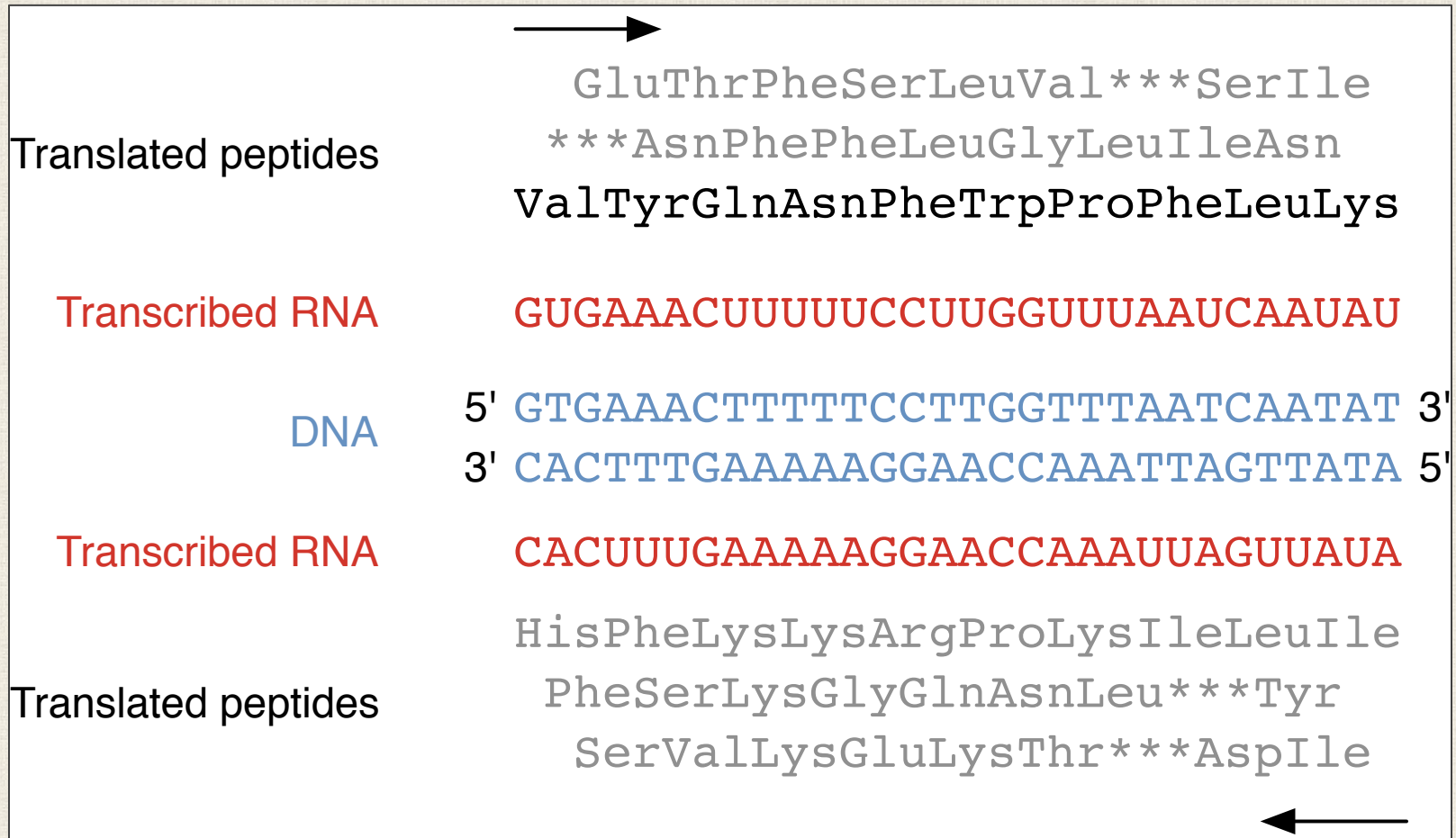
Give six monkeys one computer for a month, and they will make a mess.

Researchers at Plymouth University in England reported this week that monkeys left alone with a computer failed to produce a single word.

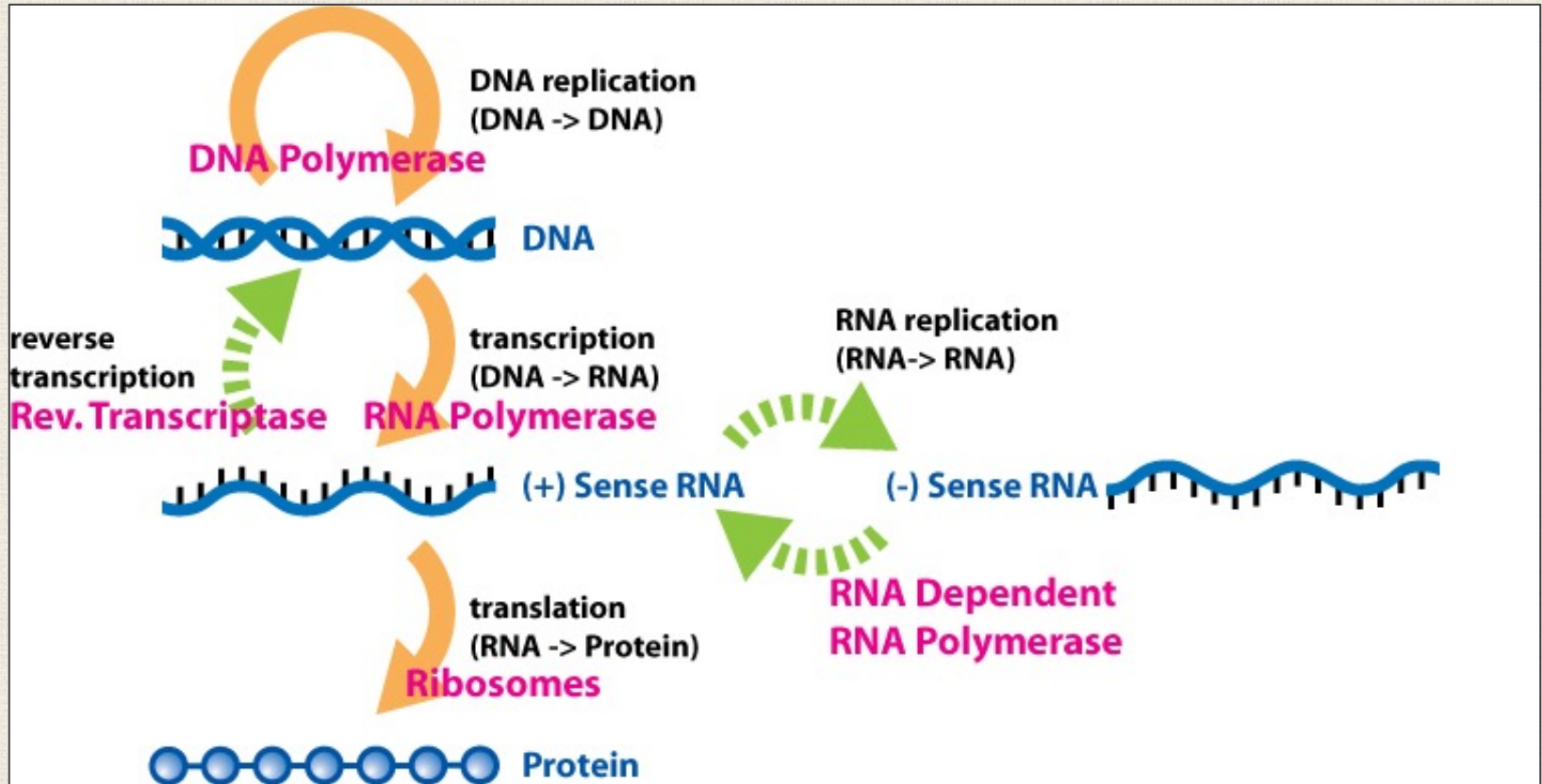
"They pressed a lot of S's," said Mike Phillips, a researcher in the project which was paid for by the Arts Council.



Last Time: We Used RNA as Proxy for Gene Expression

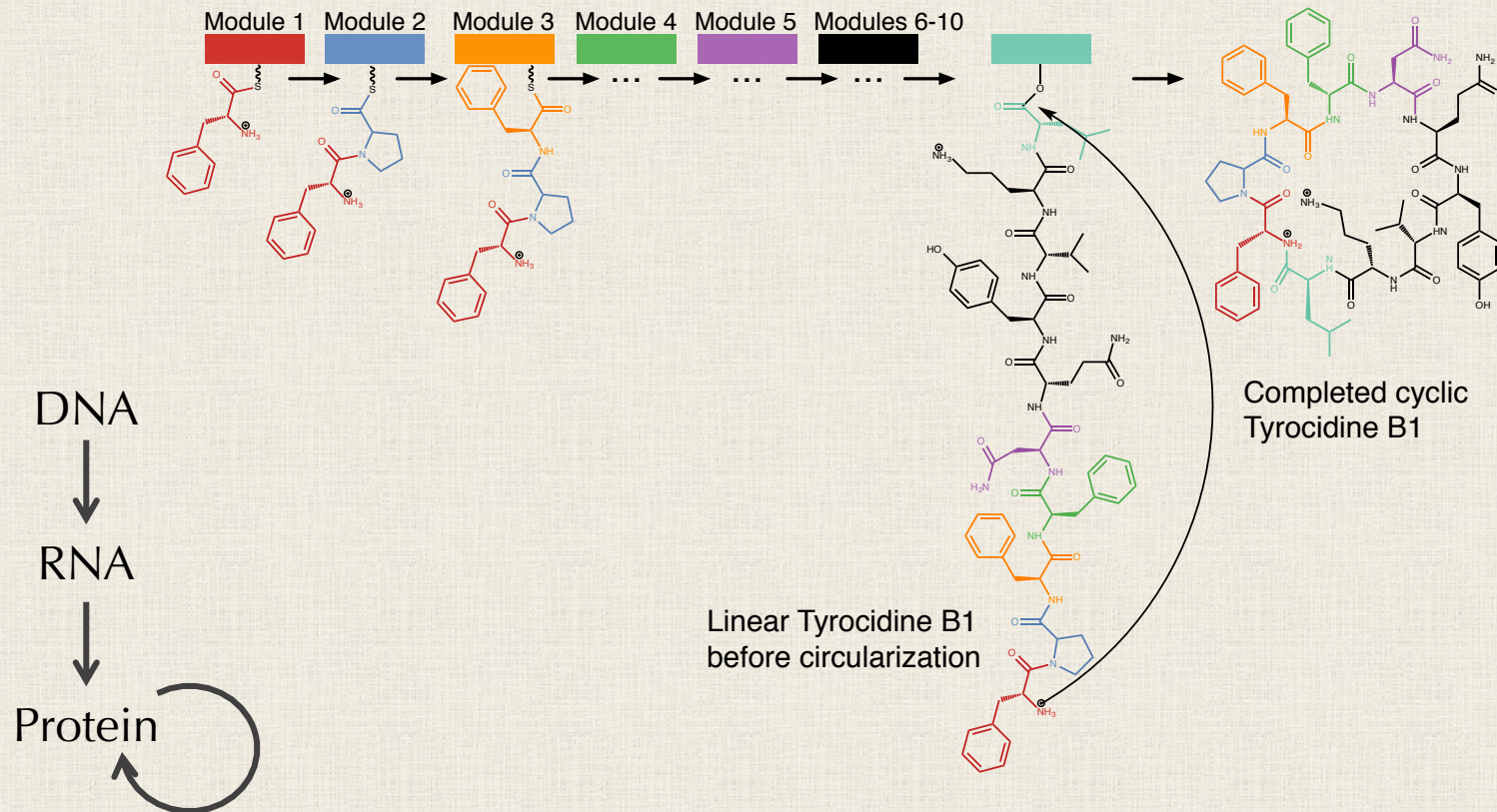


But the Central Dogma Has Exceptions



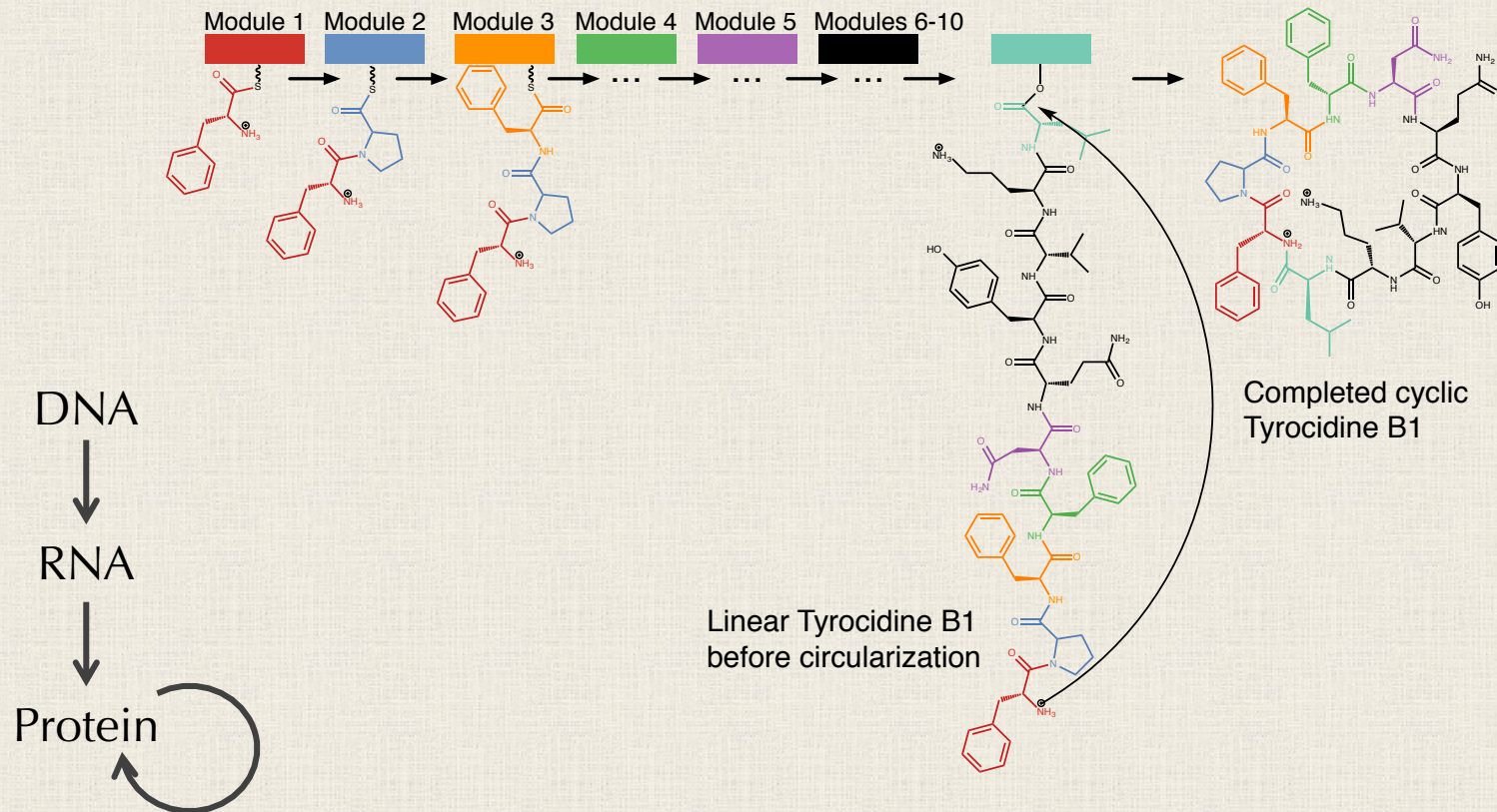
Antibiotic Peptides Can Be Produced Outside the Genetic Code

NRP synthetase: multi-module protein; each module adds single amino acid to peptide.



Antibiotic Peptides Can Be Produced Outside the Genetic Code

So how could we sequence this antibiotic peptide?



Another Application of Peptide Sequencing: Dino Peptides

[Science](#). 2007 Apr 13;316(5822):280-5.

Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry.

[Asara JM](#)¹, [Schweitzer MH](#), [Freimark LM](#), [Phillips M](#), [Cantley LC](#).

Author information

Abstract

Fossilized bones from extinct taxa harbor the potential for obtaining protein or DNA sequences that could reveal evolutionary links to extant species. We used mass spectrometry to obtain protein sequences from bones of a 160,000- to 600,000-year-old extinct mastodon (*Mammuthus americanus*) and a 68-million-year-old dinosaur (*Tyrannosaurus rex*). The presence of *T. rex* sequences indicates that their peptide bonds were remarkably stable. Mass spectrometry can thus be used to determine unique sequences from ancient organisms from peptide fragmentation patterns, a valuable tool to study the evolution and adaptation of ancient taxa from which genomic sequences are unlikely to be obtained.

A Scientific Battle Over Statistics

[Science](#). 2007 Apr 13;316(5822):280-5.

Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry.

[Asara JM](#)¹, [Schweitzer MH](#), [Freimark LM](#), [Phillips M](#), [Cantley LC](#).

[+](#) **Author information**

Abstract

Fossilized bones from extinct taxa harbor the potential for obtaining protein or DNA sequences that could reveal evolutionary links to extant species. We used mass spectrometry to obtain protein sequences from bones of a 160,000- to 600,000-year-old extinct mastodon (*Mammut americanum*) and a 68-million-year-old dinosaur (*Tyrannosaurus rex*). The presence of *T. rex* sequences indicates that their peptide bonds were remarkably stable. Mass spectrometry can thus be used to determine unique sequences from ancient organisms from peptide fragmentation patterns, a valuable tool to study the evolution and adaptation of ancient taxa from which genomic sequences are unlikely to be obtained.

[Science](#). 2008 Aug 22;321(5892):1040; author reply 1040. doi: 10.1126/science.1155006.

Comment on "Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry".

[Pevzner PA](#)¹, [Kim S](#), [Ng J](#).

[+](#) **Author information**

Abstract

Asara et al. (Reports, 13 April 2007, p. 280) reported sequencing of *Tyrannosaurus rex* proteins and used them to establish the evolutionary relationships between birds and dinosaurs. We argue that the reported *T. rex* peptides may represent statistical artifacts and call for complete data release to enable experimental and computational verification of their findings.

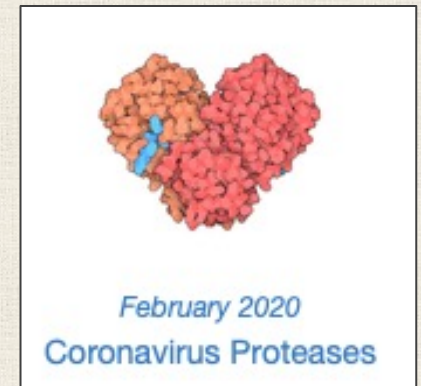
Basics of Mass Spectrometry

Mass spectrometer: a machine that fragments a peptide into two pieces, ionizes the fragments, and then measures the **mass-charge ratio** of fragments.

Basics of Mass Spectrometry

Mass spectrometer: a machine that fragments a peptide into two pieces, ionizes the fragments, and then measures the **mass-charge ratio** of fragments.

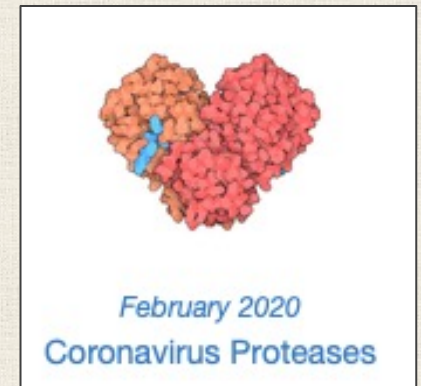
An MS machine can only read short fragments, so we typically first break long proteins into short pieces using other proteins called *proteases*.



Basics of Mass Spectrometry

Mass spectrometer: a machine that fragments a peptide into two pieces, ionizes the fragments, and then measures the **mass-charge ratio** of fragments.

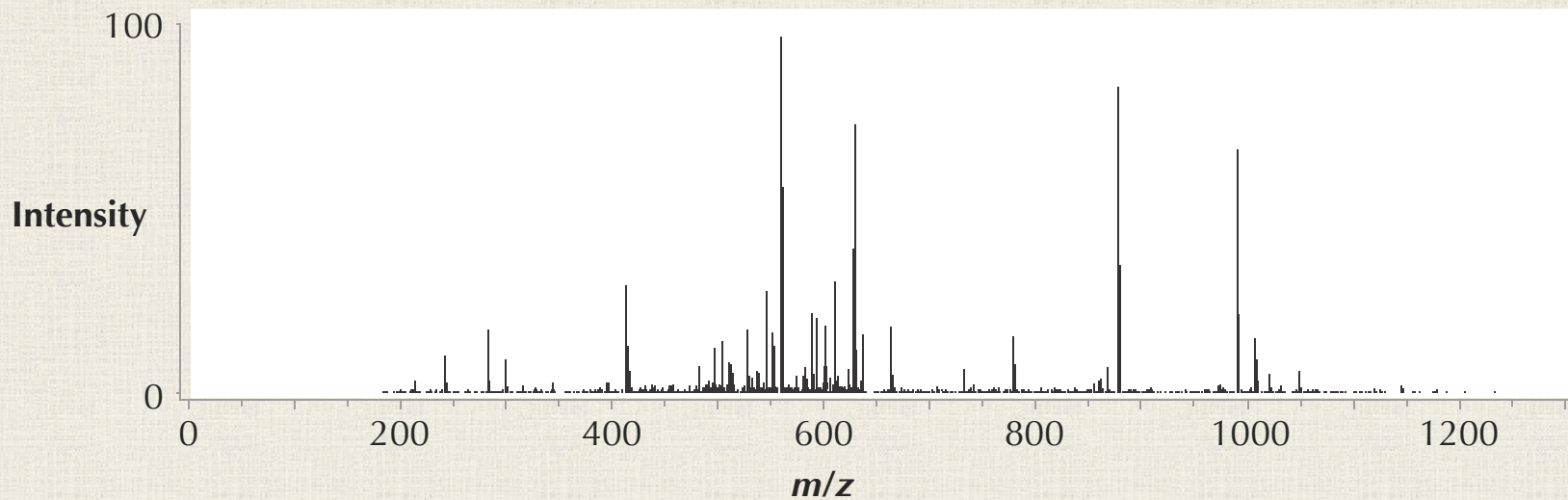
An MS machine can only read short fragments, so we typically first break long proteins into short pieces using other proteins called *proteases*.



Note: the fragmentation process is messy and somewhat unpredictable.

Sample *T. rex* Spectrum

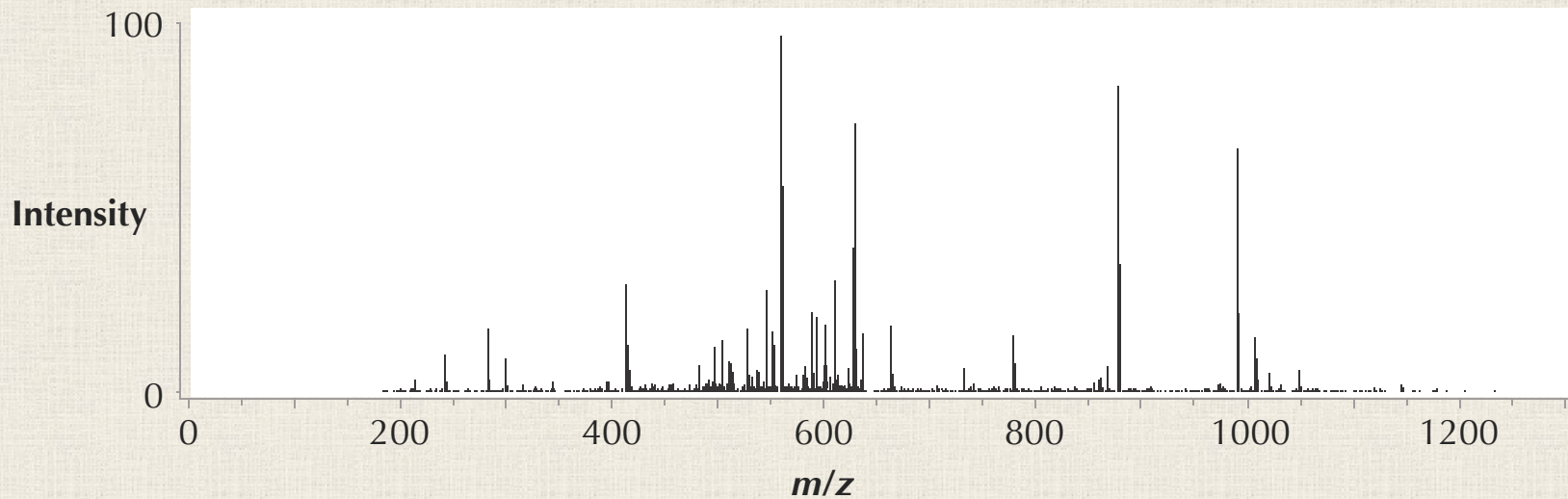
Mass spectrum: range of **intensities** of fragments detected at each mass-charge ratio (denoted m/z) for a given peptide.



Sample *T. rex* Spectrum

Most common charge is $z = +1$, so we can compare all peptide fragment masses against a spectrum using a table of amino acid masses (in Daltons).

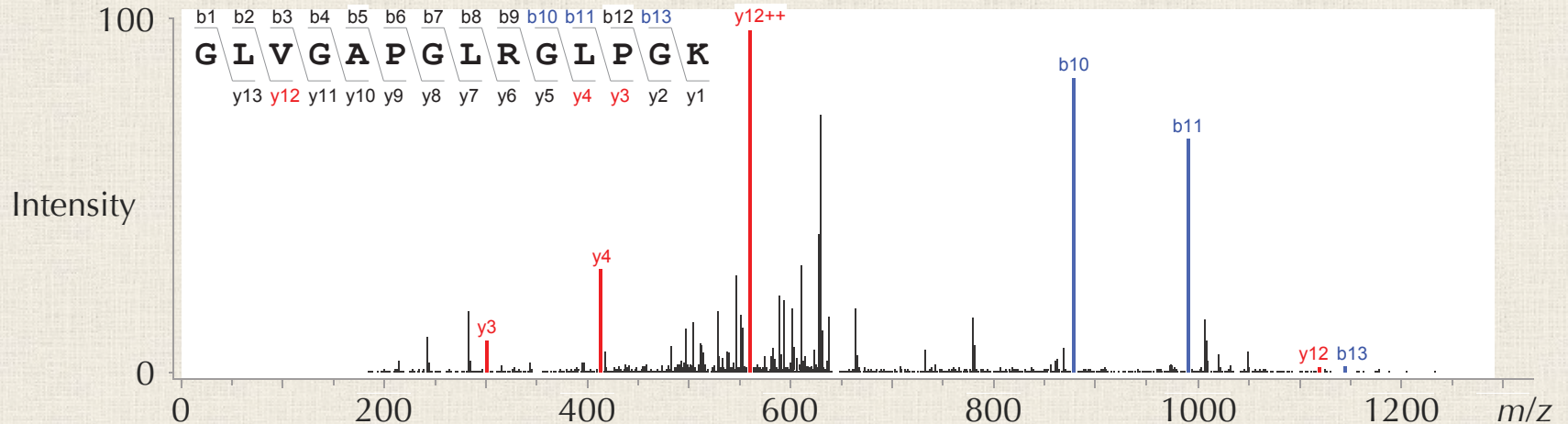
G	A	S	P	V	T	C	I	L	N	D	K	Q	E	M	H	F	R	Y	W
57	71	87	97	99	101	103	113	113	114	115	128	128	129	131	137	147	156	163	186



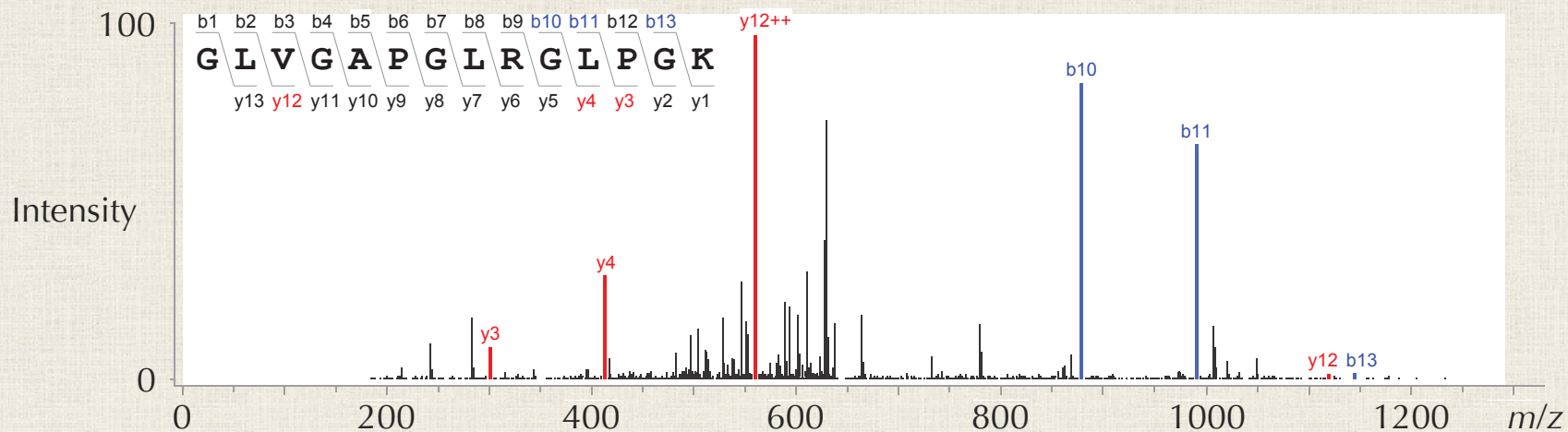
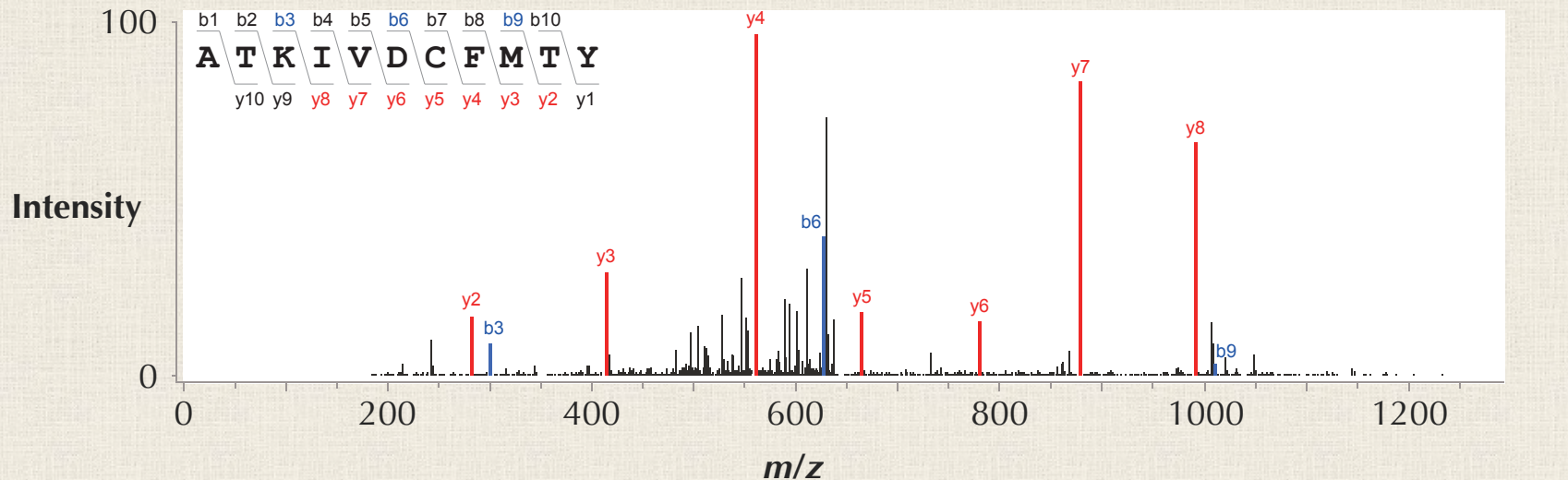
"Annotating" This *T. rex* Spectrum by GLVGAPGLRGLPGK

In this case, y_{12}^{++} means that this peak corresponds to a charge z of +2.

b_i : prefix peptide of length i
 y_i : suffix peptide of length i



How Could we Determine Which Annotation is "Better"?



Toward a Computational Problem

Peptide Sequencing Problem:

- **Input:** A mass spectrum *spectrum* and a peptide-spectrum scoring function *Score()*.
- **Output:** An amino acid string *peptide* that maximizes $Score(peptide, spectrum)$ over all amino acid strings.

An entire area of research is devoted to deriving robust peptide-spectrum scoring functions.

Toward a Computational Problem

Exercise: Count the following two things.

1. The number of possible peptides of length 10.
2. The number of peptides of length 10 in the human proteome (20,000 genes, average length ~400 amino acids).

Toward a Computational Problem

Exercise: Count the following two things.

1. The number of possible peptides of length 10.
2. The number of peptides of length 10 in the human proteome (20,000 genes, average length ~400 amino acids).

Answer:

1. 20 choices at each position, so $20^{10} \sim 10$ trillion.
2. Approx. $20,000 * 400 = 8$ million.

The Problem with Peptide Sequencing

Peptide Sequencing Problem:

- **Input:** A mass spectrum *spectrum* and a peptide-spectrum scoring function *Score()*.
- **Output:** An amino acid string *peptide* that maximizes $Score(peptide, spectrum)$ over all amino acid strings.

The highest-scoring peptide is often not in the *proteome* being considered, missing the biologically correct protein that produced a spectrum.

From Peptide Sequencing to Identification

Peptide Identification Problem:

- **Input:** A mass spectrum *spectrum*, a peptide-spectrum scoring function $Score()$, and a database *proteome* of amino acid strings.
- **Output:** An amino acid string *peptide* that maximizes $Score(peptide, spectrum)$ over all amino acid strings from *proteome*.

Note: a brute force algorithm, which we call **PeptideIdentification()**, is reasonable because the size of *proteome* is manageable.

Peptide Identification Over a Spectrum Database

So, for a family of spectra and a proteome database, we aim to find the collection of peptides scoring at least t against a spectrum for some choice of t .

```
PSMSearch(spectra, proteome,  $t$ )  
  PSMSet  $\leftarrow$  an empty set  
  for every mass spectrum spectrum in spectra  
    peptide  $\leftarrow$  PeptideIdentification(spectrum, proteome)  
    if Score(peptide, spectrum)  $\geq t$   
      PSMSet  $\leftarrow$  append(PSMSet, spectrum)  
  return PSMSet
```

Peptide Identification Over a Spectrum Database

If for some threshold parameter t , we find that the highest-scoring peptide $peptide$ in $proteome$ scores at least t against $spectrum$, then we call $(peptide, spectrum)$ a **peptide-spectrum match (PSM)**.

PSMSearch($spectra, proteome, t$)

$PSMSet \leftarrow$ an empty set

for every mass spectrum $spectrum$ in $spectra$

$peptide \leftarrow$ **PeptideIdentification**($spectrum, proteome$)

if $Score(peptide, spectrum) \geq t$

$PSMSet \leftarrow$ append($PSMSet, spectrum$)

return $PSMSet$

Reported Peptides for *T. rex*

After collecting thousands of spectra, the *T. rex* researchers consulted collagen proteins in the Uniprot database (hundreds of species), along with mutations. (P_{oh} is a hydroxylated version of proline.)

ID	Peptide	Protein
P1	GL V GAPGLRGLPGK	Collagen α1t2
P2	GVVGLP _{oh} GQR	Collagen α1t1
P3	GVQGP _{oh} GPQGPR	Collagen α1t1
P4	GATGAP _{oh} GIAGAP _{oh} GFP _{oh} GAR	Collagen α1t1
P5	GLPGESGAVGPAGPIGSR	Collagen α2t1
P6	GSAGPP _{oh} GATGFP _{oh} GAAGR	Collagen α1t1
P7	GAPGPQGPPSGAP _{oh} GP K	Collagen α1t1

Reported Peptides for *T. rex*

STOP: How can we determine if a single reported PSM is any good?

ID	Peptide	Protein
P1	GL V GAPGLRGLPGK	Collagen α 1t2
P2	GVVGLP _{oh} GQR	Collagen α 1t1
P3	GVQGPP _{oh} GPQGPR	Collagen α 1t1
P4	GATGAP _{oh} GIAGAP _{oh} GFP _{oh} GAR	Collagen α 1t1
P5	GLPGESGAVGPAGPIGSR	Collagen α 2t1
P6	GSAGPP _{oh} GATGFP _{oh} GAAGR	Collagen α 1t1
P7	GAPGPQGPPSGAP _{oh} GP K	Collagen α 1t1

Reported Peptides for *T. rex*

Answer: Rather than ask “Is this peptide above the threshold?”, we ask “What are the odds that a PSM of this quality would occur in a random database?”

ID	Peptide	Protein
P1	GL V GAPGLRGLPGK	Collagen α 1t2
P2	GVVGLP _{oh} GQR	Collagen α 1t1
P3	GVQGP _{oh} GPQGPR	Collagen α 1t1
P4	GATGAP _{oh} GIAGAP _{oh} GFP _{oh} GAR	Collagen α 1t1
P5	GLPGESGAVGPAGPIGSR	Collagen α 2t1
P6	GSAGPP _{oh} GATGFP _{oh} GAAGR	Collagen α 1t1
P7	GAPGPQGPPSGAP _{oh} GP K	Collagen α 1t1

The Monkey and the Typewriter

Exercise: What is the probability that if a monkey typed 11 English letters, that the monkey would type SHAKESPEARE?



The Monkey and the Typewriter

Answer: $1/26^{11}$.



The Monkey and the Typewriter

Exercise: What is the expected number of times that SHAKESPEARE would occur in 20 million randomly generated "words" of length 11?



The Monkey and the Typewriter

Answer: Expected number in one word is the probability of SHAKESPEARE, $1/26^{11}$. Expected number over all words is 20 million $\cdot (1/26^{11}) = 5.45 \cdot 10^{-9}$.



The Monkey and the Typewriter

Answer: Expected number in one word is the probability of SHAKESPEARE, $1/26^{11}$. Expected number over all words is 20 million $\cdot (1/26^{11}) = 5.45 \cdot 10^{-9}$.

This calculation relies on a probabilistic fact called the **linearity of expectation**: the expected value $E(X_1 + X_2 + \dots + X_n)$ is equal to $E(X_1) + E(X_2) + \dots + E(X_n)$ for any collection of random variables X_1, X_2, \dots, X_n .

The Monkey and the Typewriter

Exercise: What is the expected number of occurrences of *all* words from an English dictionary in a randomly generated string of length n ?



The Monkey and the Typewriter

Answer: Expected number of occurrences of a single string $word$ is $(n - |word| + 1) \cdot (1/26^{|word|})$. If n is large, then this is approximately $n \cdot (1/26^{|word|})$.

The Monkey and the Typewriter

Answer: Expected number of occurrences of a single string $word$ is $(n - |word| + 1) \cdot (1/26^{|word|})$. If n is large, then this is approximately $n \cdot (1/26^{|word|})$.

Linearity of expectation yields that the expected number of occurrences of all words is approximately

$$n \cdot \sum_{\text{each string } word \text{ in dictionary}} (1/26^{|word|}).$$

The Monkey and Peptide Identification

Before: “What are the odds of a monkey typing an English word?”

Now: “What are the odds of a PSM with such a good score appearing due to random chance?”

The Monkey and Peptide Identification

Before: “What are the odds of a monkey typing an English word?”

Now: “What are the odds of a PSM with such a good score appearing due to random chance?”

Given a PSM (*peptide, spectrum*) with score s , define its **PSM dictionary** as the set of all peptides scoring at least s against *spectrum*.

The Monkey and Peptide Identification

PSM Dictionary Problem (solvable)

- **Input:** An amino acid string *peptide*, a mass spectrum *spectrum*, and a peptide-spectrum scoring function *Score()*.
- **Output:** The set of all amino acid strings having score at least $Score(\textit{peptide}, \textit{spectrum})$.

Given a PSM (*peptide*, *spectrum*) with score s , define its **PSM dictionary** as the set of all peptides scoring at least s against *spectrum*.

The Monkey and Peptide Identification

We will then compare a given PSM dictionary against a *randomly generated* **decoy proteome** having the same size n as the real protein database – what is the expected number of hits that we find from the PSM dictionary in the decoy?

Given a PSM (*peptide, spectrum*) with score s , define its **PSM dictionary** as the set of all peptides scoring at least s against *spectrum*.

The Monkey and Peptide Identification

We will then compare a given PSM dictionary against a *randomly generated* **decoy proteome** having the same size n as the real protein database – what is the expected number of hits that we find from the PSM dictionary in the decoy?

STOP: If the score of the PSM is good, what does this mean for the expected number of hits against the decoy proteome?

The Monkey and Peptide Identification

We will then compare a given PSM dictionary against a *randomly generated* **decoy proteome** having the same size n as the real protein database – what is the expected number of hits that we find from the PSM dictionary in the decoy?

Answer: It will be very low (hopefully close to zero) because the dictionary will have few strings.

The Monkey and Peptide Identification

Define $E(\textit{Dictionary}, n)$ as the expected number of hits in the PSM dictionary *Dictionary* against a decoy proteome containing n amino acids.

The Monkey and Peptide Identification

Define $E(\textit{Dictionary}, n)$ as the expected number of hits in the PSM dictionary *Dictionary* against a decoy proteome containing n amino acids.

From our previous work with the monkey and the typewriter, we know that

$$E(\textit{Dictionary}, n) \approx n \cdot \sum_{\text{each peptide in dict.}} (1/20^{|peptide|}).$$

We denote the sum as $\textit{Pr}(\textit{Dictionary})$.

The Monkey and Peptide Identification

Note: This assumes a peptide can have up to n hits in a database with n amino acids, but there are $< n$ substrings of length *peptide* in a real database.

From our previous work with the monkey and the typewriter, we know that

$$E(\textit{Dictionary}, n) \approx n \cdot \sum_{\textit{each peptide in dict.}} (1/20^{|\textit{peptide}|}).$$

We denote the sum as $\textit{Pr}(\textit{Dictionary})$.

Running this Analysis on the *T. rex* PSMs

The authors of the *T. rex* peptide paper released the ~31,000 spectra they had found, allowing the following statistical analysis.

ID	Peptide	Protein	Probability
P1	GL V GAPGLRGLPGK	Collagen α 1t2	$1.8 \cdot 10^{-4}$
P2	GVVGLP _{oh} GQR	Collagen α 1t1	$7.6 \cdot 10^{-8}$
P3	GVQGP _{oh} GPQGPR	Collagen α 1t1	$7.9 \cdot 10^{-11}$
P4	GATGAP _{oh} GIAGAP _{oh} GFP _{oh} GAR	Collagen α 1t1	$3.2 \cdot 10^{-12}$
P5	GLPGESGAVGPAGPIGSR	Collagen α 2t1	$9.9 \cdot 10^{-14}$
P6	GSAGPP _{oh} GATGFP _{oh} GAAGR	Collagen α 1t1	$3.2 \cdot 10^{-14}$
P7	GAPGPQGPSGAP _{oh} GP K	Collagen α 1t1	$7.0 \cdot 10^{-16}$

Running this Analysis on the *T. rex* PSMs

Problem 1: When we use expected values, we see that at least two of the hits are very poor.

ID	Peptide	Protein	Probability	$n \cdot \text{Probability}$
P1	GL V GAPGLRGLPGK	Collagen α 1t2	$1.8 \cdot 10^{-4}$	36,000
P2	GVVGLP _{oh} GQR	Collagen α 1t1	$7.6 \cdot 10^{-8}$	16
P3	GVQGP _{oh} GPQGPR	Collagen α 1t1	$7.9 \cdot 10^{-11}$	$1.6 \cdot 10^{-2}$
P4	GATGAP _{oh} GIAGAP _{oh} GFP _{oh} GAR	Collagen α 1t1	$3.2 \cdot 10^{-12}$	$6.4 \cdot 10^{-4}$
P5	GLPGESGAVGPAGPIGSR	Collagen α 2t1	$9.9 \cdot 10^{-14}$	$2.0 \cdot 10^{-5}$
P6	GSAGPP _{oh} GATGFP _{oh} GAAGR	Collagen α 1t1	$3.2 \cdot 10^{-14}$	$6.4 \cdot 10^{-6}$
P7	GAPGPQGPSGAP _{oh} GP K	Collagen α 1t1	$7.0 \cdot 10^{-16}$	$1.4 \cdot 10^{-7}$

Running this Analysis on the *T. rex* PSMs

Problem 2: Other researchers found a more significant PSM that was a match with ostrich hemoglobin (hemoglobin mutates fast and has never been recovered from much younger fossils).

ID	Peptide	Protein	Probability	$n \cdot \text{Probability}$
P1	GL V GAPGLRGLPGK	Collagen α 1t2	$1.8 \cdot 10^{-4}$	36,000
P2	GVVGLP _{oh} GQR	Collagen α 1t1	$7.6 \cdot 10^{-8}$	16
P3	GVQGPP _{oh} GPQGPR	Collagen α 1t1	$7.9 \cdot 10^{-11}$	$1.6 \cdot 10^{-2}$
P4	GATGAP _{oh} GIAGAP _{oh} GFP _{oh} GAR	Collagen α 1t1	$3.2 \cdot 10^{-12}$	$6.4 \cdot 10^{-4}$
P5	GLPGESGAVGPAGPIGSR	Collagen α 2t1	$9.9 \cdot 10^{-14}$	$2.0 \cdot 10^{-5}$
P6	GSAGPP _{oh} GATGFP _{oh} GAAGR	Collagen α 1t1	$3.2 \cdot 10^{-14}$	$6.4 \cdot 10^{-6}$
P7	GAPGPQGPPSGAP _{oh} GP K	Collagen α 1t1	$7.0 \cdot 10^{-16}$	$1.4 \cdot 10^{-7}$
P8	VNVADCGAEALAR	Hemoglobin β	$7.8 \cdot 10^{-17}$	$1.6 \cdot 10^{-8}$

Running this Analysis on the *T. rex* PSMs

Problem 3: For the sake of fairness, we should search spectra against all vertebrate proteins (with up to 1 mismatch). This produces even more baffling results ...

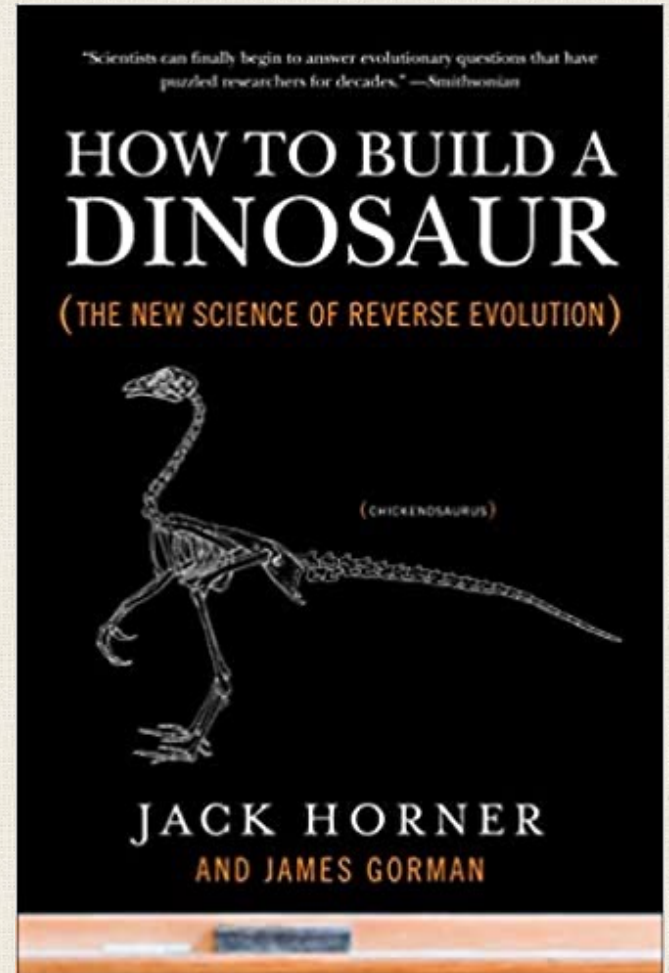
ID	Peptide	Protein	Probability	$n \cdot \text{Probability}$
P9	EDCLSG A KPK	ATG7 (Chicken)	$3.2 \cdot 10^{-12}$	$6.4 \cdot 10^{-4}$
P10	ENAGEDPGLAR	DCD (Human)	$2.7 \cdot 10^{-12}$	$5.4 \cdot 10^{-4}$
P11	E GV DAGAAGDPER	TTL11 (Mouse)	$1.2 \cdot 10^{-12}$	$2.4 \cdot 10^{-4}$
P12	S W I HVALVTGGNK	CBR1 (Human)	$1.2 \cdot 10^{-12}$	$2.4 \cdot 10^{-4}$
P13	SSN V LSGSTLR	MAMD1 (Human)	$5.9 \cdot 10^{-13}$	$1.8 \cdot 10^{-4}$
P14	DEVTPA Y VVVAR	ASPM (Mouse)	$1.9 \cdot 10^{-13}$	$3.8 \cdot 10^{-5}$
P15	R NVADCGAEALAR	HBB (Ostrich)	$3.5 \cdot 10^{-15}$	$7.0 \cdot 10^{-7}$

Running this Analysis on the *T. rex* PSMs

Problem 4: The researchers had worked with ostrich samples beforehand (and ostrich shows up with low probability in both analyses).

ID	Peptide	Protein	Probability	$n \cdot \text{Probability}$
P9	EDCLSG A KPK	ATG7 (Chicken)	$3.2 \cdot 10^{-12}$	$6.4 \cdot 10^{-4}$
P10	ENAGEDPGLAR	DCD (Human)	$2.7 \cdot 10^{-12}$	$5.4 \cdot 10^{-4}$
P11	E GV DAGAAGDPER	TTL11 (Mouse)	$1.2 \cdot 10^{-12}$	$2.4 \cdot 10^{-4}$
P12	S W I HVALVTGGNK	CBR1 (Human)	$1.2 \cdot 10^{-12}$	$2.4 \cdot 10^{-4}$
P13	SSN V LSGSTLR	MAMD1 (Human)	$5.9 \cdot 10^{-13}$	$1.8 \cdot 10^{-4}$
P14	DEVTPA Y VVVAR	ASPM (Mouse)	$1.9 \cdot 10^{-13}$	$3.8 \cdot 10^{-5}$
P15	R NVADCGAEALAR	HBB (Ostrich)	$3.5 \cdot 10^{-15}$	$7.0 \cdot 10^{-7}$

Scientists Are Still Hopeful about Dino Science that Might Not Be Possible



Scientists Are Still Hopeful about Dino Science that Might Not Be Possible



Scientists Are Still Hopeful about Dino Science that Might Not Be Possible



 **zmargotz**
@sissypantz

Can we get an update on this

 **Entrepreneur**  @Entrepreneur · Jun 16, 2015
Scientists Say They Can Recreate Living Dinosaurs Within the Next 5 Years
entm.ag/1R51yQS by @Geoff_Weiss



9:36 AM · Mar 28, 2020 · [Twitter for iPhone](#)

121.9K Retweets 615.1K Likes