Proteins

- 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
- 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
Part 1: Predicting and Analyzing Protein structures
AN INTRODUCTION TO PROTEIN STRUCTURE PREDICTION
Comparing SARS-CoV and SARS-CoV-2

In a SARS-CoV-2 challenge, we aligned the whole SARS-CoV and SARS-CoV-2 genomes.
Comparing SARS-CoV and SARS-CoV-2

In a SARS-CoV-2 challenge, we aligned the whole SARS-CoV and SARS-CoV-2 genomes.

One of the most variable 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
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!

**SARS-CoV**

MFIFLFLLTLSGSDDLRTCCTFFDDVDQAPNYQHTSSMRGVCYYPDEIFRSCTLTLTQLDLPFYSNVTFHTINHTFGNVPVIPFKDGIYFAATEGKNVVR
GWVFGSTMNKSSQSVIIIINNSNVIRACNECLDNPFFAVSKPMGTQHTMDFNADNCTFEYISDASFDLDESEKSNFKHLRESTFVKNKDKLGYV
GYQPIDVRDSLPSNFTLKPIFKLPGLININFRATFAFSP3QDIWTGAASAAAYFVYKPTTFMLKYDENGITTADVCQSNPLALCKCSKSFEIDK
GITYQTSNFRVVRPNFLQFGEVATFKSFYVYAMAVFLPVQFVYFNALNCFVSNCSVAFLSSTFFKCGYVSLATKLDLSFNSVAYDAVFVGGDVRQ
IAPGQTGVIADYNYKLPLDDFMGVCVLAWNTRIDATSTGNYYKYRLHRKHRLPFRERDISNVPSPDGKCPPTPALNCYWPLONYGFTTTTGGYQPR
VVVSFEELNAPATVCGPKSLTDILKIQNCFVSNFNGLTLGVTPLSSKRFQPFQFRGFDSTTDSVRAPKTEILIDSFAFGVGVSTIPIGTMNASEV
AVLQYDNCTVDSTAIHAQDLTPANRIYSTGTNQVFQTAGCGLAGAEHDYDSTEYCDUPIAGACASYHYSSLRRTSQSKSVAYTMSLGDSSAYSNNT
IAIPTNFNSISITTEVMPSAMKTSVDCNMYICGSDTECANLLLQYGFLCTQLNRLSAGIAAEQRDNKREVFAVVKQMKYTPTFLYGGFNSQIIPDPL
KPTKRHSFIDDLNFLKTLADAAGFQYCELGIDNLARLDCALQFNGLTLPLLTDMMATATAYALVSSTAGWTFGAGAALQIFPAFMQAYNFNGI
GVTFQNVLYENQKQIANQFNKAISIQIESLTTTSTALGKLQDVVNFQAQLNTLVKQLSNSFAGASSVNLIDSLRLDKVEAEVQIDRTLITGRQLSFLQTYTV
TQOLIRAIARASIANLATAMKSECVLQNSLLQYGFLCTQANLSAFSTFKCYVSTPKLDNLCFNYV

**SARS-CoV-2**

MFIFLFLLTLSGSDDLRTCCTFFDDVDQAPNYQHTSSMRGVCYYPDEIFRSCTLTLTQLDLPFYSNVTFHTINHTFGNVPVIPFKDGIYFAATEGKNVVR
GWVFGSTMNKSSQSVIIIINNSNVIRACNECLDNPFFAVSKPMGTQHTMDFNADNCTFEYISDASFDLDESEKSNFKHLRESTFVKNKDKLGYV
GYQPIDVRDSLPSNFTLKPIFKLPGLININFRATFAFSP3QDIWTGAASAAAYFVYKPTTFMLKYDENGITTADVCQSNPLALCKCSKSFEIDK
GITYQTSNFRVVRPNFLQFGEVATFKSFYVYAMAVFLPVQFVYFNALNCFVSNCSVAFLSSTFFKCGYVSLATKLDLSFNSVAYDAVFVGGDVRQ
IAPGQTGVIADYNYKLPLDDFMGVCVLAWNTRIDATSTGNYYKYRLHRKHRLPFRERDISNVPSPDGKCPPTPALNCYWPLONYGFTTTTGGYQPR
VVVSFEELNAPATVCGPKSLTDILKIQNCFVSNFNGLTLGVTPLSSKRFQPFQFRGFDSTTDSVRAPKTEILIDSFAFGVGVSTIPIGTMNASEV
AVLQYDNCTVDSTAIHAQDLTPANRIYSTGTNQVFQTAGCGLAGAEHDYDSTEYCDUPIAGACASYHYSSLRRTSQSKSVAYTMSLGDSSAYSNNT
IAIPTNFNSISITTEVMPSAMKTSVDCNMYICGSDTECANLLLQYGFLCTQLNRLSAGIAAEQRDNKREVFAVVKQMKYTPTFLYGGFNSQIIPDPL
KPTKRHSFIDDLNFLKTLADAAGFQYCELGIDNLARLDCALQFNGLTLPLLTDMMATATAYALVSSTAGWTFGAGAALQIFPAFMQAYNFNGI
GVTFQNVLYENQKQIANQFNKAISIQIESLTTTSTALGKLQDVVNFQAQLNTLVKQLSNSFAGASSVNLIDSLRLDKVEAEVQIDRTLITGRQLSFLQTYTV
TQOLIRAIARASIANLATAMKSECVLQNSLLQYGFLCTQANLSAFSTFKCYVSTPKLDNLCFNYV

[https://www.ebi.ac.uk/Tools/psa/emboss_needle/](https://www.ebi.ac.uk/Tools/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.
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
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

© 2021 Phillip Compeau
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?
Институт белка Российской академии наук

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.
...has tried to solve this problem for over 50 years!
Drug discovery often relies on finding drugs that will bond to protein of interest.
We can determine the shape of a protein experimentally.

https://www.youtube.com/watch?v=Qq8DO-4BnlY
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:** barring a major innovation, we will never be able to experimentally determine the structure of all proteins.
What makes proteins structure prediction from sequence hard?

On the one hand, small perturbations in the primary structure of a protein can drastically change the protein’s shape and even render it useless.
What makes proteins structure prediction from sequence hard?

On the one hand, small perturbations in the primary structure of a protein can drastically change the protein’s shape and even render it useless.

On the other, different amino acids can have similar chemical properties, and so some mutations will hardly change the shape of the protein at all.
What makes proteins structure prediction from sequence hard?

Hemoglobin subunit alpha sequences can be very different, but the structures are very similar.

| Sequence Name         | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|-----------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Emu_H-a_A             | V | L | S | A | A | D | K | T | N | T | K | S | V | F | A | K | G | P | H | A | E | E | Y | G | A | C | A | L | L | E | L | F | I | T | Y | P | T | T | K |

Sequence Identities:
- Human to Mako Shark: 43.26%
- Human to Emu: 66.67%
- Mako Shark to Emu: 36.17%
Question 1: What is the 3-dimensional protein corresponding to a string of amino acids?

https://www.cas.org/blog/covid-19-spike-protein
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.
What do we mean by “structure”?

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

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

© 2021 Phillip Compeau
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.

https://www.nature.com/articles/s41401-020-0485-4
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.
An amino acid’s central **alpha carbon** atom is connected to four different molecules:
1. a hydrogen atom (H)
2. a carboxyl group (–COOH)
3. an amino group (-NH$_2$)
4. a **side chain** (denoted “R”), which differs between amino acids.
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 (H₂O) is expelled.
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.
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** ($\phi$) and **psi angle** ($\psi$), 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 $\phi$ and $\psi$ at an amino acid can drastically change a protein’s shape.

Peptide Bonds
Dihedral Angles

Courtesy: Jacob Elmer, https://youtu.be/1usemtlYe_s
Proteins are flexible and can therefore form a huge number of shapes.

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 $n - 1 \phi$ angles and $n - 1 \psi$ 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 \( n - 1 \) \( \phi \) angles and \( n - 1 \) \( \psi \) 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 $n - 1 \phi$ angles and $n - 1 \psi$ 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.
Side chain properties affect protein folding

Amino acids’ side chain variety causes amino acids to have different chemical properties, which can lead to different conformations being more chemically stable than others.
Amino acids’ side chain variety causes amino acids to have different chemical properties, which can lead to different conformations being more chemically stable than others.

For example, nine of the 20 standard amino acids are hydrophobic (a.k.a. non-polar), meaning that their side chains tend to be repelled by water; we tend to find these amino acids sheltered from the environment on the interior of the protein.
Proteins seek the lowest 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 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).
Proteins seek the lowest 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 hypothesis that the structure of a protein can be determined by a protein’s amino acid sequence is called Anfinsen’s dogma (1972 Nobel prize in Chemistry) and is true of most proteins.
PROTEIN STRUCTURE
PREDICTION ALGORITHMS
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.
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.
Answer: This is an optimization problem, and the search space is all conformations of the polypeptide.
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.

Courtesy: David Beamish
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 $\Delta\text{energy} = \text{energy}(S) - \text{energy}(S')$. 
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”

Plotting $\Delta$energy against $\exp(\Delta$energy/$T$)

Probability of changing structure

Courtesy: Carl Kingsford
The problem with \textit{ab initio} algorithms

Because the search space is so large, and we need to run an algorithm with a lot of initial structures, \textit{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?
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.
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?
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.
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.

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

www.nature.com › letters

**Predicting protein structures with a multiplayer online game ...**

by S Cooper - 2010 - Cited by 1388 - Related articles

Aug 5, 2010 - **Protein structure prediction** is one such problem: locating the biologically relevant native conformation of a **protein** is a formidable computational challenge given the very large size of the search space. Here we describe Foldit, a **multiplayer online game** that engages non-scientists in solving hard **prediction** problems.

Scientific articles involving Foldit work:
https://foldit.fandom.com/wiki/Interesting_Articles
Early SARS-CoV-2 Spike Protein Simulations

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

Rommie Amaro
Full length SARS-CoV-2 spike protein with glycans running in MD shows vastly altered accessibility for small molecules and antibodies (aka #glycotime for the #COVID19 #demogorgon) @LCasalino88 (movie) @zied_gaieb @abbydommer @AmaroLab

https://twitter.com/RommieAmaro/status/1241810976866840577
COMPARING PROTEIN STRUCTURES
Recalling our Second Question

Question 2: How can we compare two similar proteins on the level of structure?
Comparing protein structures is analogous to comparing shapes.

**Goal:** Develop a “distance function $d(S, T)$ that quantifies 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.

\[
y_s = \frac{\int_0^\pi \sin \theta \, \pi}{\pi} = \frac{-\cos \pi + \cos 0}{\pi} = \frac{2}{\pi}
\]
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?*
Sample $n$ points along the boundary of $S$ and $T$, converting $S$ and $T$ into vectors $s = (s_1, \ldots, s_n)$ and $t = (t_1, \ldots, t_n)$. 
Determining Similarity of Aligned Shapes with RMSD

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

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

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

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.

\[
\text{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \cdots + d(s_n, t_n)^2)}
\]
An example of computing RMSD

We find distances between corresponding points.

\[
\text{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \cdots + d(s_n, t_n)^2)}
\]
An example of computing RMSD

\[
\text{RMSD} = \sqrt{\left(\frac{1}{4} \cdot (2 + 1 + 4 + 2)\right)} = \sqrt{\frac{9}{4}} = \frac{3}{2}.
\]
An example of RMSD

STOP: Do you see any issues with using RMSD?

RMSD(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1)^2 + d(s_2, t_2)^2 + \cdots + d(s_n, t_n)^2)}
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.
Comparing proteins of differing lengths

**STOP:** Any ideas on how we could handle situations like this?
Answer: First, we align the protein sequences; then, any gap columns will not contribute to RMSD.
Comparing proteins of differing lengths

The situation below (an inserted substructure) would throw off RMSD for every alpha carbon after #2.
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.
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?
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 (Qres):** defined as follows.

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

- \( N \) 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

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

**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.

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

**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

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

**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

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

**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

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

STOP: So, what are the possible values of Qres?
Q per residue offers a single value for how much two proteins differ locally

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

**STOP:** So, what are the possible values of Qres?

**Answer:** Qres 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).

© 2021 Phillip Compeau
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.

So, how well did AlphaFold do?

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

DeepMind received lots of positive press

‘The game has changed.’ AI triumphs at solving protein structures

By Robert F. Service | Nov. 30, 2020, 10:30 AM
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.
But some scientists remain skeptical

Although AlphaFold released some SARS-CoV-2 structures, they have not published their prediction of the SARS-CoV-2 spike protein, or explained the details of their algorithm in a peer-reviewed forum, inviting criticism from open science advocates.
But some scientists remain skeptical

Although AlphaFold released some SARS-CoV-2 structures, they have not published their prediction of the SARS-CoV-2 spike protein, or explained the details of their algorithm in a peer-reviewed forum, inviting criticism from open science advocates.

Finally, AlphaFold does well but is “trained” using a database of known structures, which makes it more likely to correctly predict known structures. But proteins with structures dissimilar to any known structure possess some of the most scientific interest.
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
Last Time: We Used RNA as Proxy for Gene Expression

<table>
<thead>
<tr>
<th>Translated peptides</th>
<th>Transcribed RNA</th>
<th>DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HisPheLysLysArgProLysIleLeuIlePheSerLysGlyGlnAsnLeu<em><strong>TyrSerValLysGluLysThr</strong></em>AspIle</td>
<td>GUGAAACUUUUUUCCUUGGUUUAAUCAAU</td>
<td>5' GTGAAACCTTTTTTCTTGGTTTTAATCAATAT 3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3' CACTTTGAAAAAGGAACCAAATTAGTTATA 5'</td>
</tr>
<tr>
<td></td>
<td>GUGAAACUUUUUUCCUUGGUUUAAUCAAU AU</td>
<td>5'3'</td>
</tr>
</tbody>
</table>
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?

DNA → RNA → Protein

Completed cyclic Tyrocidine B1

Linear Tyrocidine B1 before circularization
Another Application of Peptide Sequencing: Dino Peptides

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

Asara JM, 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.
Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry.

Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC.

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.

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

Pevzner PA, Kim S, Ng J.

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

Most common charge is $z = +1$, so we can compare all peptide fragment masses against a spectrum.

$b_i$: prefix peptide of length $i$

$y_i$: suffix peptide of length $i$
"Annotating" This *T. rex* Spectrum by
GLVGAPGLRLPGK

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”?

© 2021 Phillip Compeau
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.
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. $20,000 \times 400 = 8$ 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.
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:** There are slicker algorithms, but a brute force algorithm will be reasonable here because the size of *proteome* is manageable.
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 $t$.

If for some threshold parameter $t$, we find that the highest-scoring peptide $\text{peptide}$ in $\text{proteome}$ scores at least $t$ against $\text{spectrum}$, then we call $(\text{peptide, spectrum})$ a peptide-spectrum match (PSM).
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 ← an empty set
    for every mass spectrum spectrum in spectra
        peptide ← PeptideIdentification(spectrum, proteome)
        if Score(peptide, spectrum) ≥ t
            PSMSet ← append(PSMSet, spectrum)
    return PSMSet
```
Reported Peptides for *T. rex*

After collecting thousands of spectra, the *T. rex* researchers consulted all collagen proteins in the Uniprot database (hundreds of species), along with mutations of these proteins. They found ...

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>GLVGAPGLRGLPGK</td>
<td>Collagen α1t2</td>
</tr>
<tr>
<td>P2</td>
<td>GVVGLP&lt;sub&gt;oh&lt;/sub&gt;GQR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P3</td>
<td>GVQGPP&lt;sub&gt;oh&lt;/sub&gt;GPQGPR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P4</td>
<td>GATGAP&lt;sub&gt;oh&lt;/sub&gt;GIAGAP&lt;sub&gt;oh&lt;/sub&gt;GFPohGAR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P5</td>
<td>GLPGESGAVGPGAPIGSR</td>
<td>Collagen α2t1</td>
</tr>
<tr>
<td>P6</td>
<td>GSAGPP&lt;sub&gt;oh&lt;/sub&gt;GATGFPohGAAGR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P7</td>
<td>GAPGPQGPGPSGAP&lt;sub&gt;oh&lt;/sub&gt;GPK</td>
<td>Collagen α1t1</td>
</tr>
</tbody>
</table>
Reported Peptides for *T. rex*

**Note:** $P_{\text{oh}}$ is a hydroxylated version of proline.

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>GLV\text{GAPGLRGLPGK}</td>
<td>Collagen $\alpha_{1t2}$</td>
</tr>
<tr>
<td>P2</td>
<td>GVVGLP\text{ohGQR}</td>
<td>Collagen $\alpha_{1t1}$</td>
</tr>
<tr>
<td>P3</td>
<td>GVQGPP\text{ohGPQGPR}</td>
<td>Collagen $\alpha_{1t1}$</td>
</tr>
<tr>
<td>P4</td>
<td>GATGAP\text{ohGIAGAPohGFPohGAR}</td>
<td>Collagen $\alpha_{1t1}$</td>
</tr>
<tr>
<td>P5</td>
<td>GLPGESGAVGPAGPIGR</td>
<td>Collagen $\alpha_{2t1}$</td>
</tr>
<tr>
<td>P6</td>
<td>GSAGPP\text{ohGATGFPohGAAGR}</td>
<td>Collagen $\alpha_{1t1}$</td>
</tr>
<tr>
<td>P7</td>
<td>GAPGPQGPP\text{SGAPohGPK}</td>
<td>Collagen $\alpha_{1t1}$</td>
</tr>
</tbody>
</table>
STOP: How can we determine if a single reported PSM is any good?

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>GLV GAPGLRGLPGK</td>
<td>Collagen α1t2</td>
</tr>
<tr>
<td>P2</td>
<td>GVVGGLP&lt;sub&gt;oh&lt;/sub&gt;GQR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P3</td>
<td>GVQGPP&lt;sub&gt;oh&lt;/sub&gt;GPGPR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P4</td>
<td>GATGAP&lt;sub&gt;oh&lt;/sub&gt;GIAGAP&lt;sub&gt;oh&lt;/sub&gt;GFPohGAR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P5</td>
<td>GLP GESGAVGPAGPIGR</td>
<td>Collagen α2t1</td>
</tr>
<tr>
<td>P6</td>
<td>GSAGPP&lt;sub&gt;oh&lt;/sub&gt;GATGFPohGAAGR</td>
<td>Collagen α1t1</td>
</tr>
<tr>
<td>P7</td>
<td>GAPGPQGPG&lt;sub&gt;oh&lt;/sub&gt;GPK</td>
<td>Collagen α1t1</td>
</tr>
</tbody>
</table>

**FIGURE 11.13** The seven candidate T. rex collagen peptides (P1 - P7) reported by Asara as well as a hemoglobin peptide (P8). The last column shows the probabilities of the PSM dictionaries formed by these peptides. Red symbols indicate mutated amino acids compared to peptides in the UniProt database. The amino acid Poh stands for hydroxyproline, a modified form of proline that is common in collagens.
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: $1/26^{11}$. 
Exercise: What is the expected number of times that SHAKEPEARE would occur in 20 million randomly generated “words” of length 11?
Answer: Probability of a single occurrence is $1/26^{11}$. Expected number of occurrences is therefore 20 million $(1/26^{11}) = 5.45 \times 10^{-9}$. 
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 $word$ is approximately $n \left( 1/26^{\text{word}} \right)$.
Answer: Expected number of occurrences of word is approximately $n \left( \frac{1}{26} \right)^{|word|}$.

Summing over the dictionary gives that the expected number of occurrences of all words is approximately

$$n \sum_{\text{each string word in dictionary}} \left( \frac{1}{26} \right)^{|word|}.$$
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 a given 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 a given 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*. 

© 2021 Phillip Compeau
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*. 
Define $E(Dictionary, n)$ as the expected number of hits in the PSM dictionary $Dictionary$ against a decoy proteome containing $n$ amino acids.
Define $E(Dictionary, n)$ 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

$$E(Dictionary, n) \approx n \ast \sum_{\text{each peptide in dict.}} \frac{1}{20 \mid \text{peptide} \mid}.$$

We denote the sum as $Pr(Dictionary)$. 
STOP: In order to report a high quality PSM, what are we expecting from $\Pr(\text{Dictionary})$ and $E(\text{Dictionary}, n)$?

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

$$E(\text{Dictionary, } n) \approx n \times \sum_{\text{each peptide in dict.}} \left( \frac{1}{20} |\text{peptide}| \right).$$

We denote the sum as $\Pr(\text{Dictionary})$. 
Answer: They should both be close to zero.

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

\[ E(\text{Dictionary, } n) \approx n \times \sum_{\text{each peptide in dict.}} \left( \frac{1}{20} |\text{peptide}| \right). \]

We denote the sum as \( \text{Pr}(\text{Dictionary}) \).
From our previous work with the monkey and the typewriter, we know that

\[ E(Dictionary, n) \approx n \times \sum_{\text{each peptide in dict.}} \left( \frac{1}{20} |\text{peptide}| \right). \]

We denote the sum as \( \Pr(Dictionary) \).
From our previous work with the monkey and the typewriter, we know that

\[ E(Dictionary, n) \approx n \times \sum_{\text{each } peptide \text{ in dict.}} \left( \frac{1}{20} |peptide| \right). \]

We denote the sum as \( Pr(Dictionary) \).
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.

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
<th>Probability</th>
<th>n · Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>GLVGAPGLRGLPGK</td>
<td>Collagen α1t2</td>
<td>1.8 · 10^{-4}</td>
<td>36,000</td>
</tr>
<tr>
<td>P2</td>
<td>GVVGLP_{oh}GQR</td>
<td>Collagen α1t1</td>
<td>7.6 · 10^{-8}</td>
<td>16</td>
</tr>
<tr>
<td>P3</td>
<td>GVQGPP_{oh}GPQGPR</td>
<td>Collagen α1t1</td>
<td>7.9 · 10^{-11}</td>
<td>1.6 · 10^{-2}</td>
</tr>
<tr>
<td>P4</td>
<td>GATGAP_{oh}GIAGAP_{oh}GFP_{oh}GAR</td>
<td>Collagen α1t1</td>
<td>3.2 · 10^{-12}</td>
<td>6.4 · 10^{-4}</td>
</tr>
<tr>
<td>P5</td>
<td>GLPGESGAVGPAGPIGSR</td>
<td>Collagen α2t1</td>
<td>9.9 · 10^{-14}</td>
<td>2.0 · 10^{-5}</td>
</tr>
<tr>
<td>P6</td>
<td>GSAGPP_{oh}GATGFP_{oh}GAAGR</td>
<td>Collagen α1t1</td>
<td>3.2 · 10^{-14}</td>
<td>6.4 · 10^{-6}</td>
</tr>
<tr>
<td>P7</td>
<td>GAPGPQGPSSGP_{oh}GP_{K}</td>
<td>Collagen α1t1</td>
<td>7.0 · 10^{-16}</td>
<td>1.4 · 10^{-7}</td>
</tr>
</tbody>
</table>

© 2021 Phillip Compeau
Running this Analysis on the *T. rex* PSMs

**Problem 1:** 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).

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
<th>Probability</th>
<th>n  · Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>GLVGVAPGLRGLPGK</td>
<td>Collagen α1t2</td>
<td>1.8 · 10^{-4}</td>
<td>36,000</td>
</tr>
<tr>
<td>P2</td>
<td>GVVGLPohGQR</td>
<td>Collagen α1t1</td>
<td>7.6 · 10^{-8}</td>
<td>16</td>
</tr>
<tr>
<td>P3</td>
<td>GVQGPPohGPQGPR</td>
<td>Collagen α1t1</td>
<td>7.9 · 10^{-11}</td>
<td>1.6 · 10^{-2}</td>
</tr>
<tr>
<td>P4</td>
<td>GATGAPohGIAGAPohGPohGAR</td>
<td>Collagen α1t1</td>
<td>3.2 · 10^{-12}</td>
<td>6.4 · 10^{-4}</td>
</tr>
<tr>
<td>P5</td>
<td>GLPGESGAVGPAGPIGSR</td>
<td>Collagen α2t1</td>
<td>9.9 · 10^{-14}</td>
<td>2.0 · 10^{-5}</td>
</tr>
<tr>
<td>P6</td>
<td>GSAGPPohGATGFPohGAAGR</td>
<td>Collagen α1t1</td>
<td>3.2 · 10^{-14}</td>
<td>6.4 · 10^{-6}</td>
</tr>
<tr>
<td>P7</td>
<td>GAPGPQGPGSPohGPohGPK</td>
<td>Collagen α1t1</td>
<td>7.0 · 10^{-16}</td>
<td>1.4 · 10^{-7}</td>
</tr>
<tr>
<td>P8</td>
<td>VNVADCGAEALAR</td>
<td>Hemoglobin β</td>
<td>7.8 · 10^{-17}</td>
<td>1.6 · 10^{-8}</td>
</tr>
</tbody>
</table>

© 2021 Phillip Compeau
Running this Analysis on the *T. rex* PSMs

**Problem 2:** For the sake of fairness, we should search spectra against all vertebrate proteins (with up to 1 mismatch). This produces even more baffling results …

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
<th>Probability</th>
<th>( n \cdot \text{Probability} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>P9</td>
<td>EDCLSGA<strong>K</strong>PK</td>
<td>ATG7 (Chicken)</td>
<td>3.2 ( \cdot ) 10^{-12}</td>
<td>6.4 ( \cdot ) 10^{-4}</td>
</tr>
<tr>
<td>P10</td>
<td>ENAGEDPG<strong>L</strong>AR</td>
<td>DCD (Human)</td>
<td>2.7 ( \cdot ) 10^{-12}</td>
<td>5.4 ( \cdot ) 10^{-4}</td>
</tr>
<tr>
<td>P11</td>
<td>EGVDAGA<strong>A</strong>AGDPER</td>
<td>TTL11 (Mouse)</td>
<td>1.2 ( \cdot ) 10^{-12}</td>
<td>2.4 ( \cdot ) 10^{-4}</td>
</tr>
<tr>
<td>P12</td>
<td><strong>W</strong>IHVALVTGG<strong>N</strong>K</td>
<td>CBR1 (Human)</td>
<td>1.2 ( \cdot ) 10^{-12}</td>
<td>2.4 ( \cdot ) 10^{-4}</td>
</tr>
<tr>
<td>P13</td>
<td>SSN<strong>V</strong>LSG<strong>G</strong>STLR</td>
<td>MAMD1 (Human)</td>
<td>5.9 ( \cdot ) 10^{-13}</td>
<td>1.8 ( \cdot ) 10^{-4}</td>
</tr>
<tr>
<td>P14</td>
<td>DEVTP<strong>Y</strong>YVVVAR</td>
<td>ASPM (Mouse)</td>
<td>1.9 ( \cdot ) 10^{-13}</td>
<td>3.8 ( \cdot ) 10^{-5}</td>
</tr>
<tr>
<td>P15</td>
<td>RNVADCG<strong>A</strong>EAAL<strong>R</strong></td>
<td>HBB (Ostrich)</td>
<td>3.5 ( \cdot ) 10^{-15}</td>
<td>7.0 ( \cdot ) 10^{-7}</td>
</tr>
</tbody>
</table>

---

*FIGURE 11.15* Matching *T. rex* spectra against all vertebrate proteins in the UniProt database (allowing for up to 1 mutation) reveals a diverse set of peptides. Red symbols indicate mutated amino acids. Note the presence of another ostrich hemoglobin peptide (P15), which is slightly heavier (by 57 daltons) than the previously reported hemoglobin peptide in Figure 11.13 (P8). This change in mass may represent either a mutation of **V** into **R** (as shown above) or a modification of an amino acid.

---

**The dinosaur DNA controversy**

As the "*T. rex* peptides" paper continues to age, there is no end in sight to its controversy. Yet it was not the first paper to report the retrieval of genetic material from dinosaurs. In 1994, Scott Woodward announced that he had sequenced DNA from an 80 million year-old dinosaur bone. The most vehement critic of his finding was — believe it or not — Mary Schweitzer, who proved that Woodward had only sequenced contaminated human DNA.

The moral is that although we often present scientific discoveries as clear and incontrovertible, the reality is that some of the interesting avenues of modern science often fall short of this ideal. In a sense, the academic battleground is part of the appeal of becoming a scientist in the first place. But we also cannot help but wonder if we would have a conclusive answer to whether Horner’s fossil really contained dinosaur peptides if it had originally been shared with dozens of independent researchers, who would have undoubtedly unearthed the shocking appearance of hemoglobin in the *T. rex* samples. Fittingly, in their criticism of Woodward’s "dinosaur DNA" paper, Schweitzer wrote, "real advance in [paleontology] will come only when it is demonstrated that those studies can be replicated in independent laboratories."
Problem 3: The researchers had worked with ostrich samples beforehand (and ostrich shows up with low probability in both analyses).

<table>
<thead>
<tr>
<th>ID</th>
<th>Peptide</th>
<th>Protein</th>
<th>Probability</th>
<th>n · Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>P9</td>
<td>EDCLSGAKPK</td>
<td>ATG7 (Chicken)</td>
<td>3.2 · 10^{-12}</td>
<td>6.4 · 10^{-4}</td>
</tr>
<tr>
<td>P10</td>
<td>ENAGEDPGLAR</td>
<td>DCD (Human)</td>
<td>2.7 · 10^{-12}</td>
<td>5.4 · 10^{-4}</td>
</tr>
<tr>
<td>P11</td>
<td>EGVDAGAAGDPER</td>
<td>TTL11 (Mouse)</td>
<td>1.2 · 10^{-12}</td>
<td>2.4 · 10^{-4}</td>
</tr>
<tr>
<td>P12</td>
<td>WIHVALVTGGNK</td>
<td>CBR1 (Human)</td>
<td>1.2 · 10^{-12}</td>
<td>2.4 · 10^{-4}</td>
</tr>
<tr>
<td>P13</td>
<td>SNSVLGSTLR</td>
<td>MAMD1 (Human)</td>
<td>5.9 · 10^{-13}</td>
<td>1.8 · 10^{-4}</td>
</tr>
<tr>
<td>P14</td>
<td>DEVTPAYVVVAR</td>
<td>ASPM (Mouse)</td>
<td>1.9 · 10^{-13}</td>
<td>3.8 · 10^{-5}</td>
</tr>
<tr>
<td>P15</td>
<td>RNVADCGAEALAR</td>
<td>HBB (Ostrich)</td>
<td>3.5 · 10^{-15}</td>
<td>7.0 · 10^{-7}</td>
</tr>
</tbody>
</table>
Scientists Are Still Hopeful about Dino Science that Might Not Be Possible
Scientists Are Still Hopeful about Dino Science that Might Not Be Possible

Can we get an update on this

Scientists Say They Can Recreate Living Dinosaurs Within the Next 5 Years
entm.ag/1R5IyQS by @Geoff_Weiss

9:36 AM · Mar 28, 2020 · Twitter for iPhone
121.9K Retweets 615.1K Likes