Applications in Computational Biology and Explainable AI

CSE 312 Summer 21
Lecture 24

“Real Life”

The usual attitude towards math/theory is: “why should I care?”

This lecture is going to try and motivate an answer, at least from a research perspective

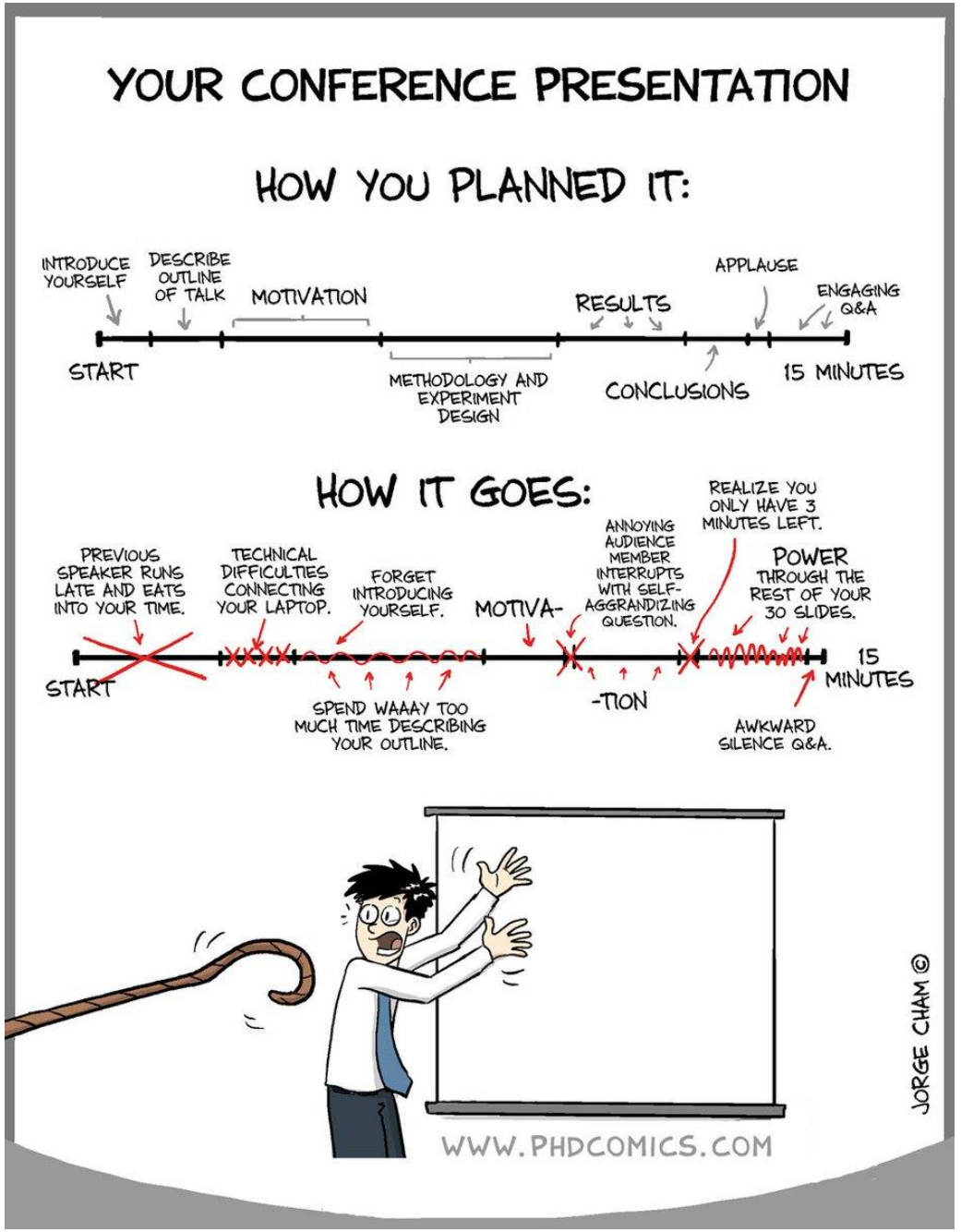
https://www.smbc-comics.com/comic/a-new-method

The first math class.

Caveats

This talk will be *very* high level – I will skip many details

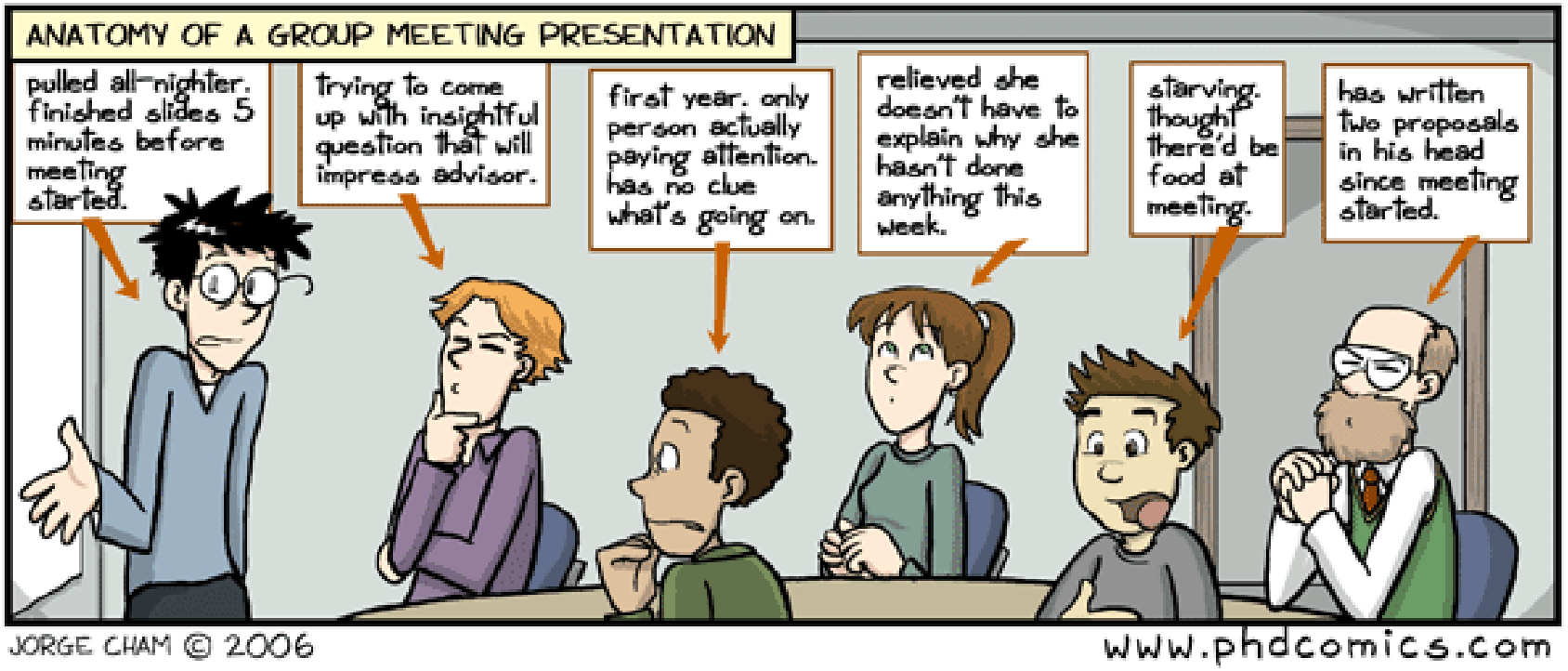
It is mainly an attempt to motivate you to think about how these concepts apply to real life situations

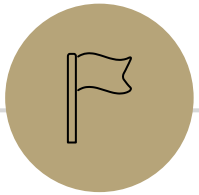


Engagement

This talk will be much more fun if you engage and ask questions

Remember: if you aren't asking questions for me, I will be asking questions to you





An Application in Computational Biology: Protein Sequence Statistics

Karlin, Samuel, and Stephen F. Altschul. "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." Proceedings of the National Academy of Sciences 87.6 (1990): 2264-2268.

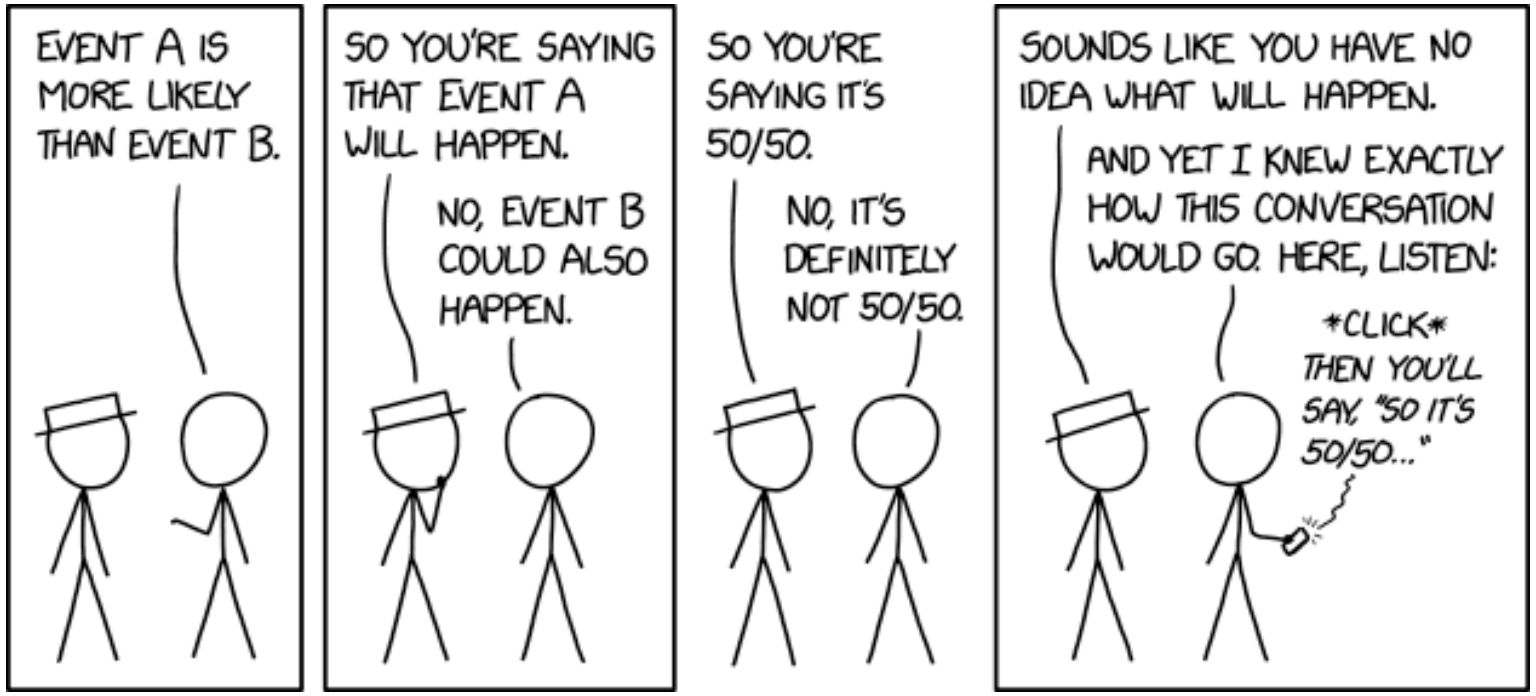
A Coin Flip Problem

Consider the following problem: I flip a coin n times with probability p of heads. What is the expected length of the *longest sequence* of heads?

We will consider this problem shortly, but first, some warm up

Coin Flip Warm Up

I flip a coin n times with probability p of heads. What is...
...the distribution of the number of heads?
...the probability that we see a sequence of only heads?
...the expected number of runs of heads with length k?



Coin Flip Recursion

What is the probability that we see a sequence of k consecutive heads?

Coin Flip Recursion

What is the probability that we see a sequence of k consecutive heads?

f(j, m) be the probability of seeing such a sequence having already observed j consecutive heads with m flips remaining

Coin Flip Recursion

What is the probability that we see a sequence of k consecutive heads?

f(j, m) be the probability of seeing such a sequence having already observed j consecutive heads with m flips remaining

Initial conditions:

f(j, m) = 0 if k – j > m

f(j, m) = 1 if j >= k

Coin Flip Recursion

What is the probability that we see a sequence of k consecutive heads?

f(j, m) be the probability of seeing such a sequence having already observed j consecutive heads with m flips remaining

Initial conditions:

f(j, m) = 0 if k – j > m

f(j, m) = 1 if j >= k

Recursion:

f(j, m) = p * f(j + 1, m – 1) + (1 – p) * f(0, m – 1)

A Coin Flip Problem

Consider the following problem: I flip a coin n times with probability p of heads. What is the expected length of the *longest sequence* of heads?

TTTTTHHTHTTHHHHTHTHHHTTTHHHHHHTHTTT

Before we answer this – how might we go about beginning this problem? What distributions may be involved?

A Coin Flip Problem

Consider the following problem: I flip a coin n times with probability p of heads. What is the expected length of the *longest sequence* of heads?

A very loose answer:
Expected # of a sequence of heads of length k is roughly n(1 – p) * pˆk

There is usually 1 sequence that is the longest, so solve for k:
n(1 – p) * pˆk = 1

k = -log_p(n * (1-p))

A Coin Flip Problem

Consider the following problem: I flip a coin n times with probability p of heads. What is the expected length of the *longest sequence* of heads?

A tighter bound is *much* harder to find

So “Why Do We Care?”

So why am I asking you all these weird questions about sequences of coin flips?

So “Why Do We Care?”

So why am I asking you all these weird questions about sequences of coin flips?

They are a good *model* of a biological phenomenon: the sequences of amino acids that make up proteins in our body

High School Biology Refresher

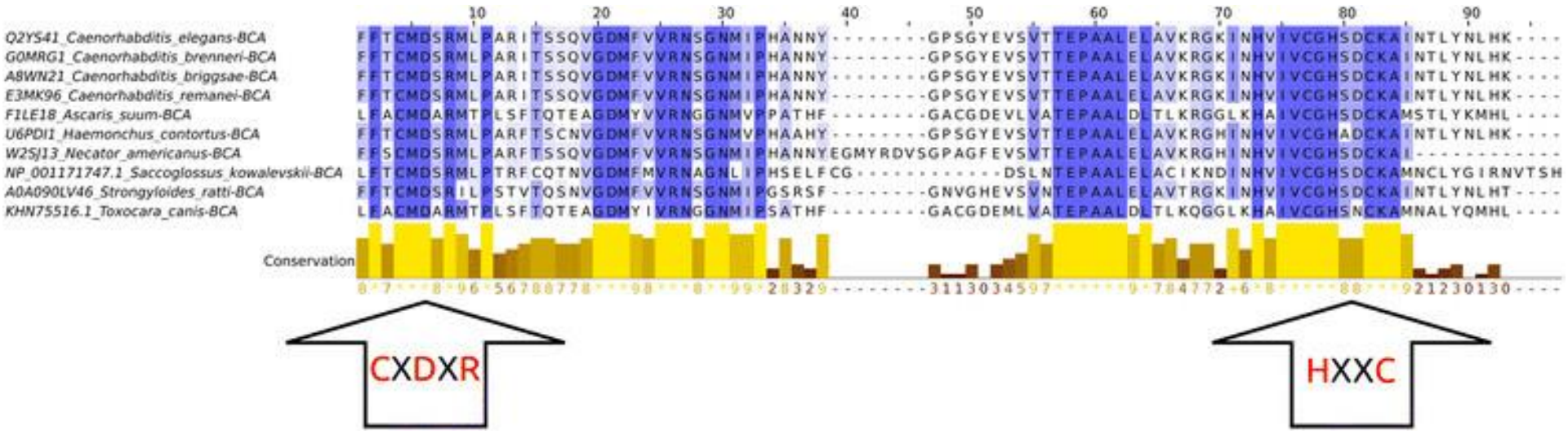
Proteins are a key organic macromolecule – they do almost everything in our bodies, from supporting muscle tissue to carrying oxygen in our blood

As a computer scientist, all you need to know is that proteins are strings from a roughly ~20 amino acid alphabet

The Sequence Distribution

Protein sequences aren't *random* – they have evolved over millions of years to accomplish specific functions using amino acids as a code

Often, we have a particular subsequence that we observe multiple times throughout different sequences. The question becomes: what was the probability of observing that specific subsequence at random?



Amino Acid Modeling

Usually, we define a reference distribution: we assume the ith amino acid occurs with probability p_i independently

We then define a scoring system: s_i is the score of the ith amino acid

We can then ask questions about the probability of seeing subsequences with particular scores. Usually, we are interested in finding the Maximal Scoring Subsequence (MSS)

Amino Acid Modeling

The score of a subsequence is the sum of the scores of each individual amino acid in the subsequence

For example, maybe we care about amino acids with a particular property, like hydrophobicity – so we might give all the hydrophobic amino acids a positive score and all others a negative score

An Example

Suppose my scoring system is:
A = 1
B = 0
C = -1

and my sequence is:
ABBCCABAACACBAAAABACCCAACACCACBBCABCAB

What is my MSS? What is its score?

An Example

Suppose my scoring system is:
A = 1
B = 0
C = –1

and my sequence is:
ABBCCABAACACBAAAABACCCAACACCACBBCABCAB

What is my MSS? What is its score?

An Exercise

Suppose I wanted to find the longest consecutive sequence of ‘A’s – what should my scoring system be?

The Distribution

In general, Karlin and Altschul (1990) proved a formula for the distribution of the MSS:

For a sequence of length n, let $M(n)$ denote the maximal segment score. It can be proved that $M(n)$ is of the order $(\ln n)/\lambda^*$ (24). Subtracting this centering value from $M(n)$, we can ask what is the limiting probability distribution for $\tilde{M}(n) = M(n) - (\ln n)/\lambda^*$.

THEOREM 1. ***The random variable*** $\tilde{M}(n)$ ***(the centered maximal segment score) has the close approximating distribution***

$$Prob\{\tilde{M}(n) > x\} \approx 1 - exp\{-K^* e^{-\lambda^* x}\}. \qquad [2]$$

Math Details

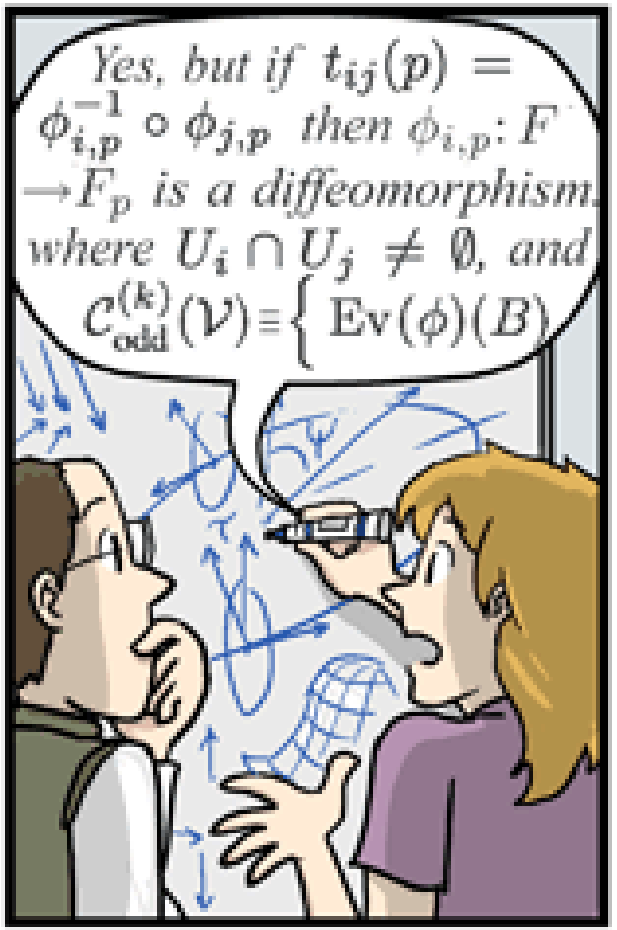
The math behind the proof is *complicated*

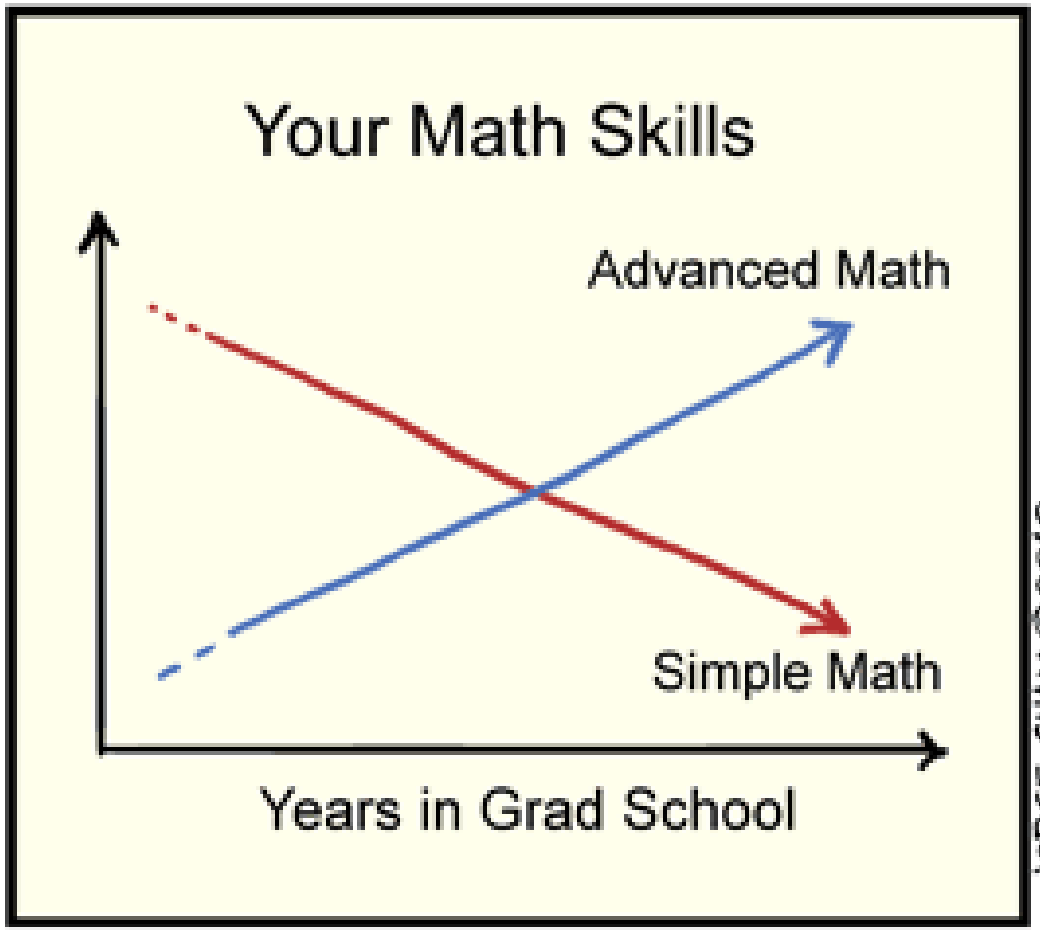
The intuition is not! We assume that in the limit of some very large sequence n, the maximal scoring sequence is a very *rare* event in an infinite process

This is modeled by a poisson distribution, and we ask the question: what is the probability that this distribution exceeds 0?

$$Prob\{\tilde{M}(n) > x\} \approx 1 - exp\{-K*e^{-\lambda*x}\}. \qquad [2]$$

A Comedic Interlude

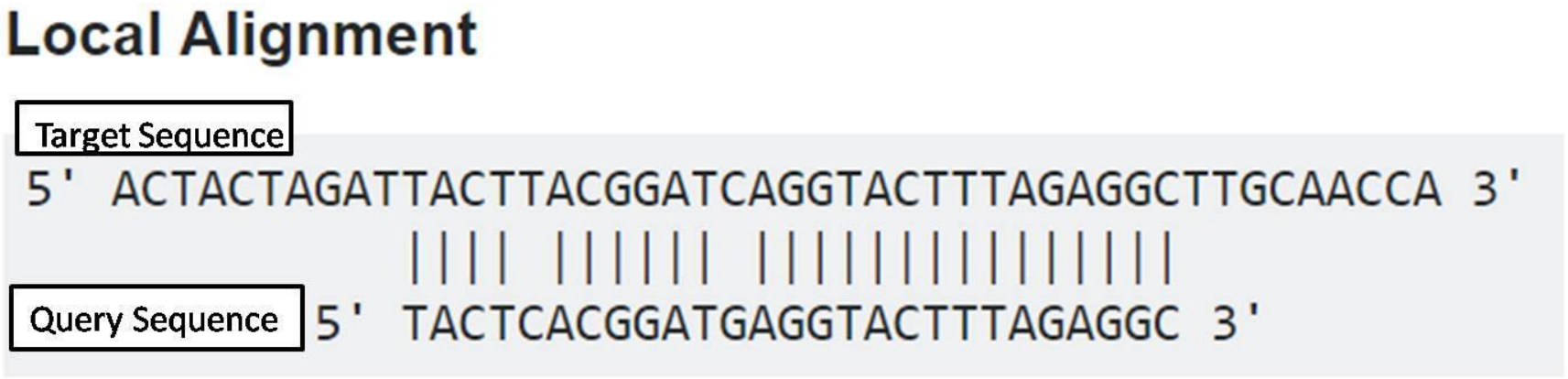




But Wait, There's More!

One of the most important problems in computational biology is called sequence searching: given a sequence and a database of other sequences, which sequences are most "similar" to mine?

Many algorithms for doing this (which you can learn in 527), but suppose an algorithm spits out a similar looking sequence. How do you know if it is a significant match?

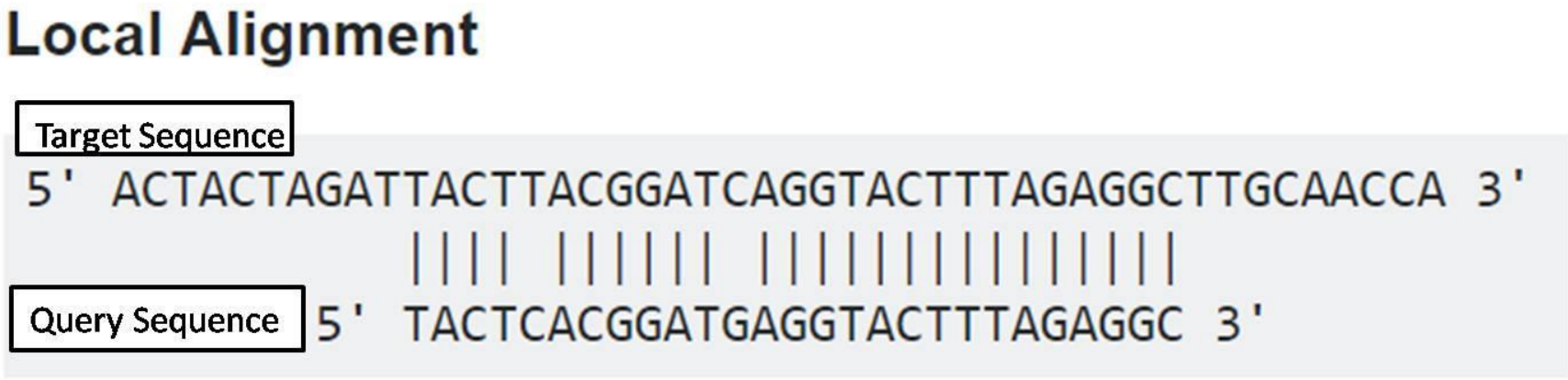


Pairwise Scoring Extensions

We can analyze this model in a very similar way: simply use a pairwise scoring system! Let s_{i,j} be the score of matching i and j, and then ask: what is the probability/expected value/bound of seeing a particular pair of matched subsequences?

Double Trouble

An example: suppose a match is worth 1 point, and a mismatch is worth -1 points. Consider picture at the bottom. What is the associated score? What is the probability of getting such a score at random?



BLAST

This insight – the ability to estimate the probability of high scoring pairwise matches – is the **foundation** of much of modern computational protein biology!

It is a key component of the BLAST algorithm, a tool used by essentially every computational biology researcher.

Basic local alignment search tool

SF Altschul, W Gish, W Miller, EW Myers… - Journal of molecular …, 1990 - Elsevier

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP …

☆ 99 Cited by 94723 Related articles All 92 versions

Citations…

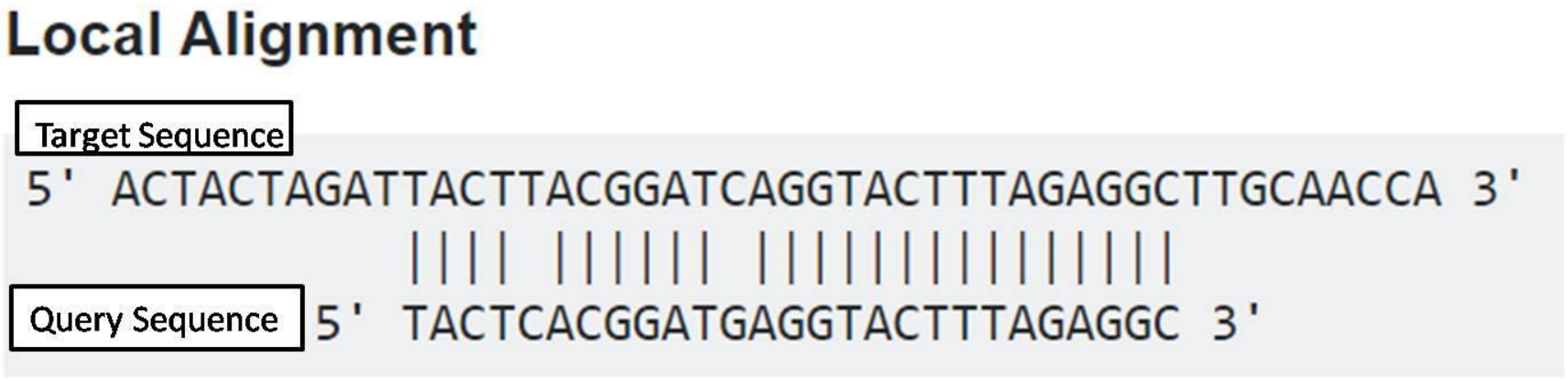
90,000 citations is more than citations than most researchers get over their LIFETIMES…

…so yeah, the paper is pretty important

The Point

…is that you have learned many of the basic tools used to analyze these kinds of problems! All you need to do is apply them creatively.

What other kinds of questions could you ask and answer about the statistics of sequences and sequence matching?



I was cheated…

…where's the AI/research part of the talk??

Most of the AI methods we use go beyond this course. If you are interested, take 545! But a preview…

Mix in a Neural Net

Most research has gone beyond basic sequence statistics. The community is largely interested in training large, deep neural networks to predict things about protein sequences!

Mix in a Neural Net

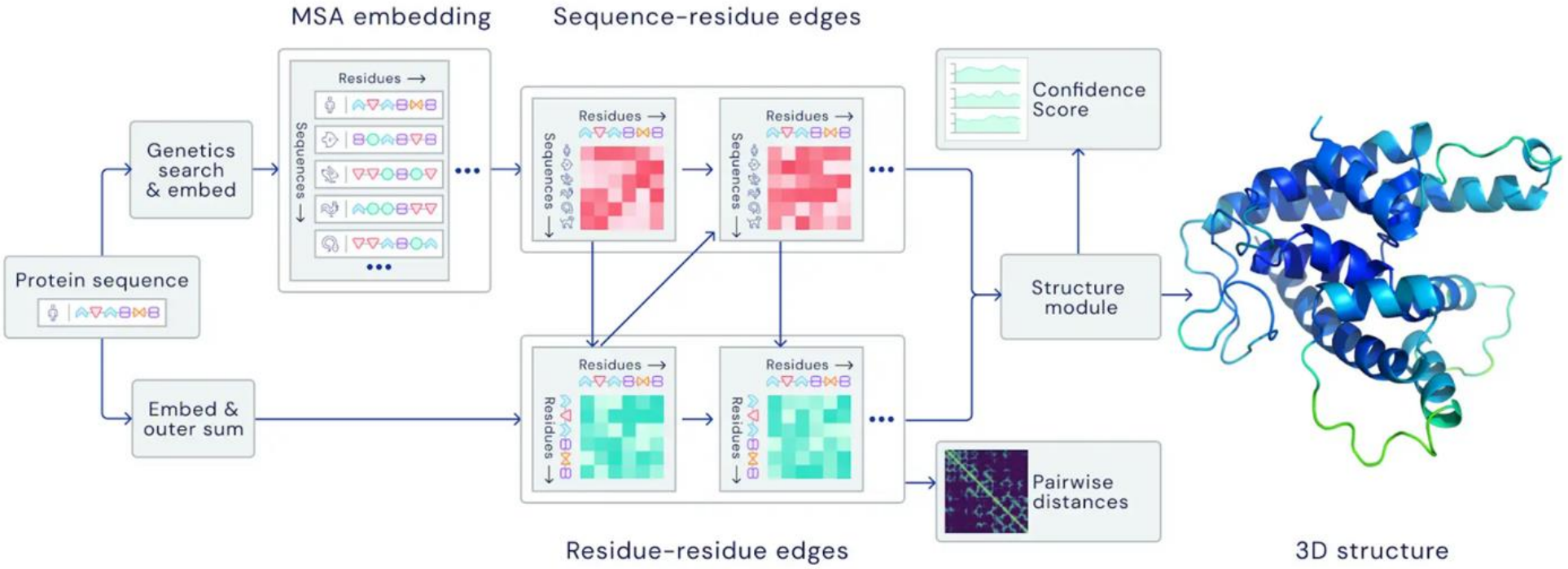
Most research has gone beyond basic sequence statistics. The community is largely interested in training large, deep neural networks to predict things about protein sequences!

Mix in a Neural Net

Most research has gone beyond basic sequence statistics. The community is largely interested in training large, deep neural networks to predict things about protein sequences!

Mix in a Neural Net

For example, the biggest news in the protein world was AlphaFold 2: a system to predict how proteins fold from their individual sequences



Closer to Home…

One of my upcoming research ideas is about using neural networks to *learn* the significance of matches

Instead of using computational/discrete approximations, can we train a model to *predict* the probability of seeing two sequences being matched?

Closer to Home…

For example, certain coin flip sequence problems we've seen in this class have messy, recursive answers. Sometimes they have no closed formulas and we can only approximate them!

Wouldn't it be nice to have a machine learning algorithm to do all the heavy theoretical work for us?

Closer to Home…

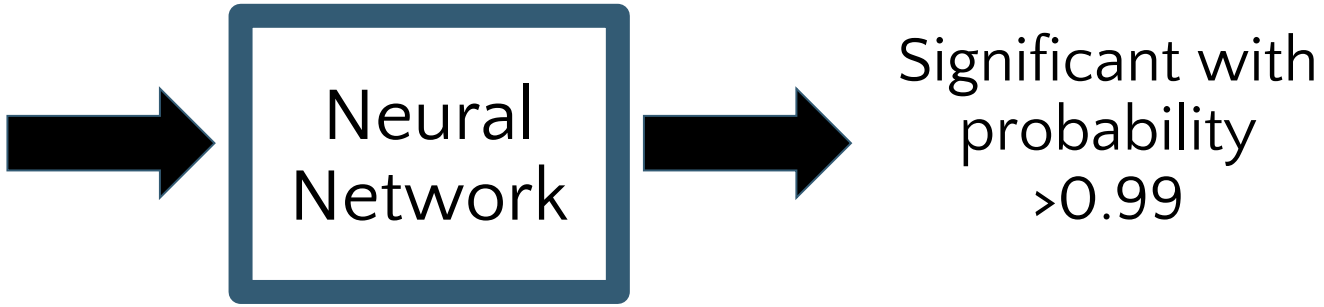
```
MNSFSTSAFGPVAFSLGLLLVLPAAFP-APVPPGEDSKDVAAPHRQPLTS
|...|...|.|||| |||:||...||| :.|..|:.::| ..|:| |:.:
MKFLSARDFHPVAF-LGLMLVTTTAFPTSQVRRGDFTED-TTPNR-PVYT

SERIDKQIRYILDGISALRKETCNKSNMCESSKEALAENNLNLPKMAEKD
:.::...|.::|..|..:|||.||.::.|.::.:|||||||.||::...|
TSQVGGLITHVLWEIVEMRKELCNGNSDCMNNDDALAENNLKLPEIQRND

GCFQSGFNEETCLVKIITGLLEFEVYLEYLQNRF-ESSEEQARAVQMSTK
||:|:|:|:|.||:||.:||||:..||||::|.. ::.:::||.:|..|:
GCYQTGYNQEICLLKISSGLLEYHSYLEYMKNNLKDNKKDKARVLQRDTE

VLIQFLQKKAKNLDAITTPDPTTNASLLTKLQAQNQWLQDMTTHLILRSF
.||...::.|:|..|..|.|.:||.|..||::|.:||:..|...||:|.
TLIHIFNQEVKDLHKIVLPTPISNALLTDKLESQKEWLRTKTIQFILKSL

KEFLQSSLRALRQM
:|||:.:||:.||.
EEFLKVTLRSTRQT
```



Significant with probability >0.99

Remember Your Basics

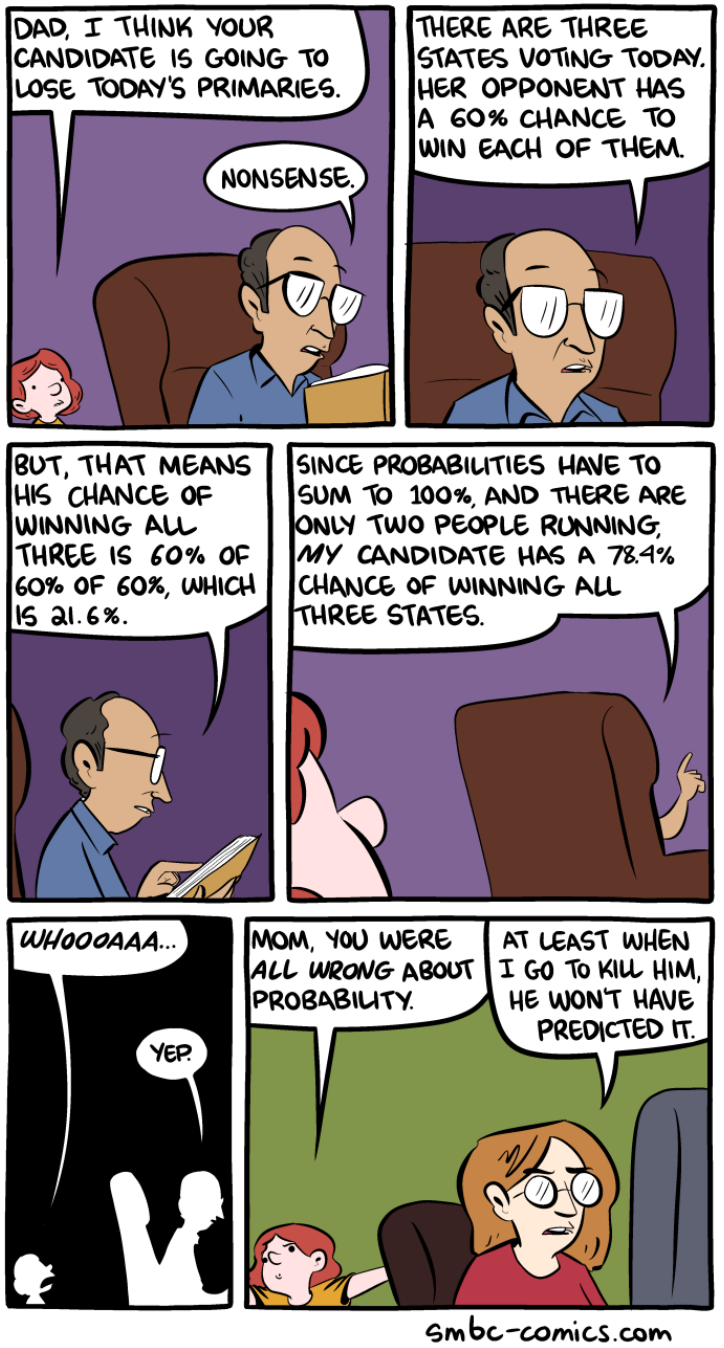
But in order to train such a model, we need DATA about the existing statistics of protein sequences…

…and in order to get DATA, we need to understand the underlying statistics ourselves! So don’t worry, understanding what we’ve learned in class is still very important!

A Recap

Discrete math shows up everywhere in research – especially in computational biology, because our genomes and our proteomes are fundamentally discrete objects!

Understanding the statistics of these sequences is a foundational area of research, and one that has led to many important computational and biological discoveries!



Any Questions?