3000788 Intro to Comp Molec Biol

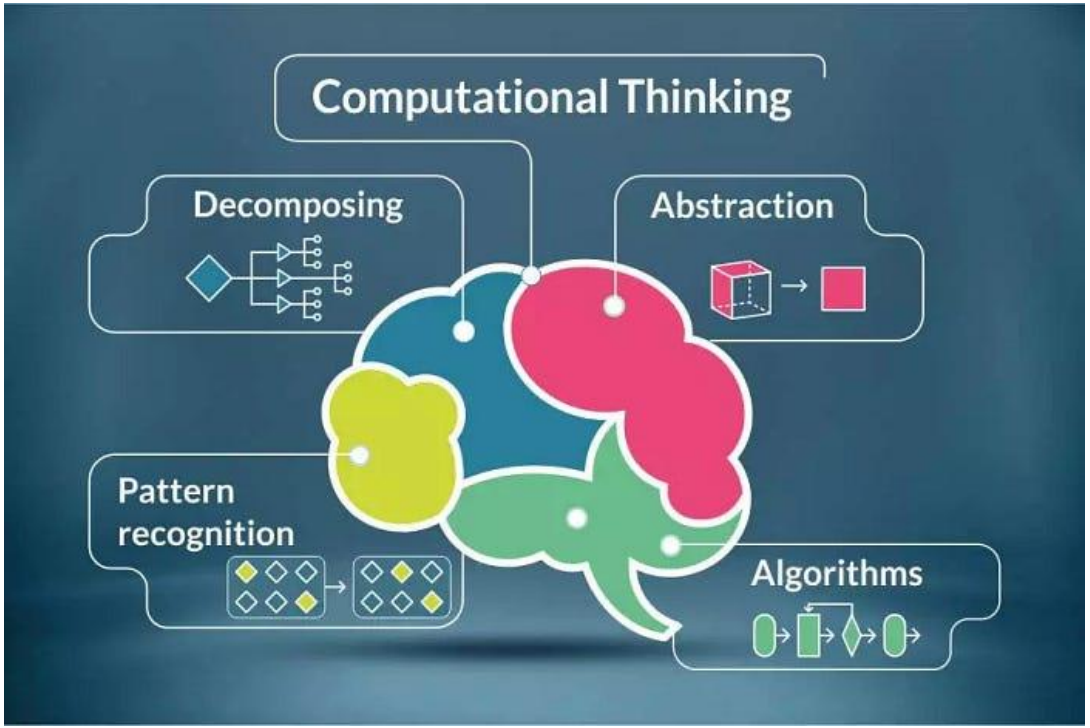
Lecture 3: Computational thinking

August 24, 2023

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Computational thinking



- Transform hypothesis into a computable statement

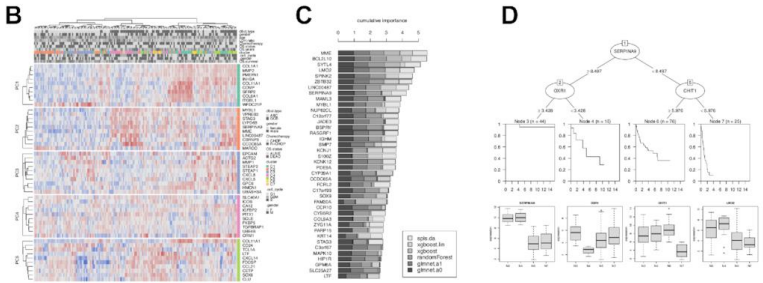
- Genes A and B are co-regulated
- Spearman's correlation coefficient for the expression levels of genes A and B is high

- Transcription factor binds to the promoter of gene A to activate transcription

https://www.nextgurukul.in/thenextworld/

What can you do with computational thinking?

- **Data analysis**
  - Identify calculations that will support or disprove your hypothesis
  - Aware of pros and cons of each calculation
  - Generate effective visualizations that tell the story
- Modeling and simulation
- Algorithm



Akhmedov, M. et al. NAR Genom and Bioinfor, 2(1):lqz019 (2019)

Gene ID	P61_2_C	P62_2_C	P63_2_C	P64_2_C	P68_2_C
ENSG00000000003.14	4.637576	6.183992	5.237635	2.372719	5.665966
ENSG00000000005.5	0	0	0	0	0
ENSG00000000419.12	11.22781	4.813792	2.99782	10.99452	10.7482
ENSG00000000457.13	7.656414	5.082675	7.710682	9.014404	8.488388
ENSG00000000460.16	3.172546	2.245954	5.974815	3.501081	4.162024
ENSG00000000938.12	0	0	0	0.042488	0
ENSG00000000971.15	6.626259	8.19511	5.904925	11.7748	2.050394
ENSG00000001036.13	1.790445	0.76823	3.670635	0.68115	1.894823
ENSG00000001084.11	19.53907	25.08378	11.04872	5.815902	20.23763
ENSG00000001167.14	15.34717	20.00867	17.10001	25.31168	27.41216
ENSG00000001460.17	0.889852	3.090642	0.744581	3.439525	2.417934
ENSG00000001461.16	3.771195	3.12468	1.385353	2.767444	2.973217
ENSG00000001497.16	16.75059	9.662455	15.4965	14.34071	10.62035
ENSG00000001617.11	2.998366	3.712208	3.885852	17.50663	3.019686

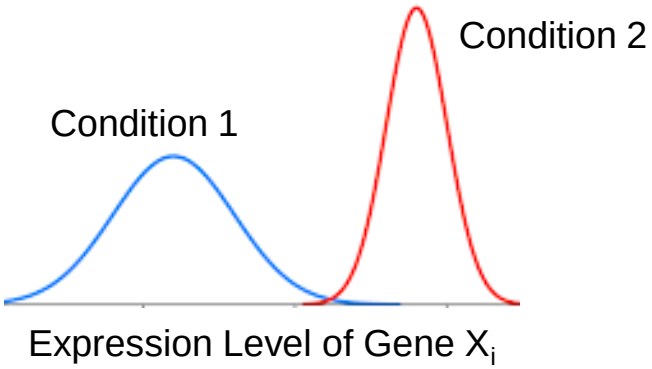
Statistics

- Normal or log-normal
  - Fluorescence signal of 1e5 vs 1e6 → 10-fold increase
  - Cell diameter difference of 1.6 nm vs 1 nm → 0.6 nm increase

- Understand standard tests
  - Test of means
  - Test of variances
  - Test the entire distribution

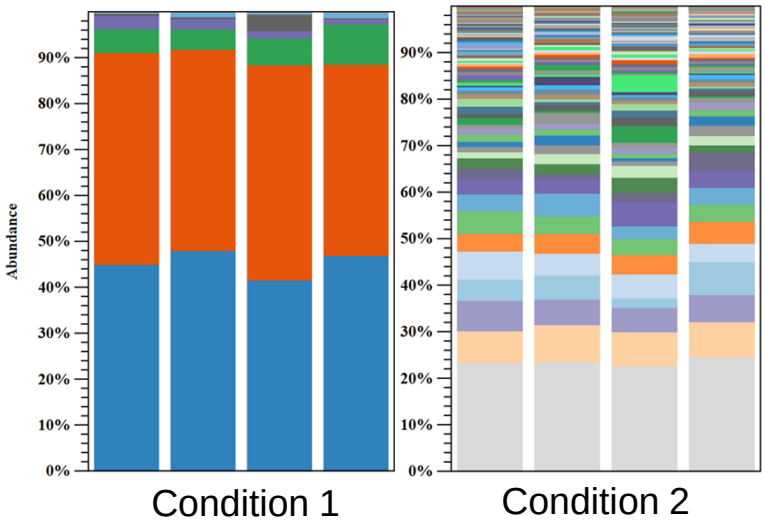
- Permutation test = ability to hand-craft and customize
  - Define appropriate null hypothesis
  - Design test statistics
  - Permute or simulate changes in data

Good data analysis requires mathematics



- Identify significantly differentially expressed genes, X_1, X_2, …, X_{200} with *t*-test
- Want to do more experiments on the “most” differentially expressed genes
  - Can we rank these genes from “more” to “less” differentially expressed?
  - **Hint**: Independent two-sample *t*-test’s statistics $= \frac{\text{Mean}_1 - \text{Mean}_2}{\sqrt{\frac{1}{n}(\text{Variance}_1 + \text{Variance}_2)}}$

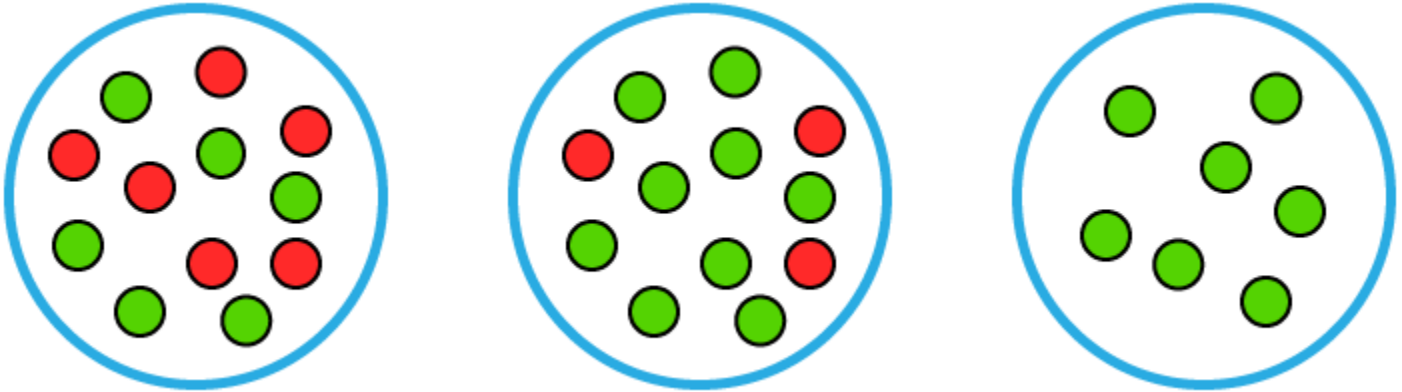
Describing diversity of microbiome



Colors indicate different microbial taxa

- Visually, microbial taxa distribution in Condition 2 is clearly more diverse
- Can we describe this pattern with numbers?
  - Number of taxa?
  - Entropy = $-\text{Frequency}_1 \log_2(\text{Frequency}_1) - \cdots - \text{Frequency}_n \log_2(\text{Frequency}_n)$

Entropy



https://www.javatpoint.com/entropy-in-machine-learning

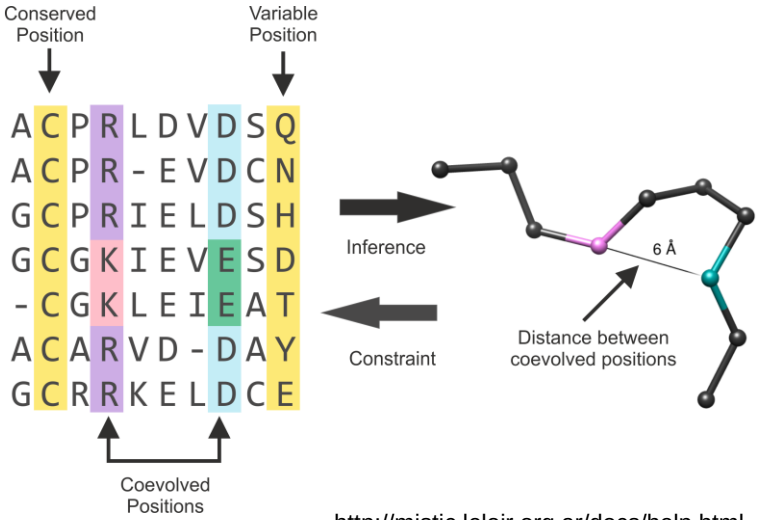
- Entropy = $-\text{Frequency}_1 \log_2(\text{Frequency}_1) - \cdots - \text{Frequency}_n \log_2(\text{Frequency}_n)$
- Left sample: $-\frac{6}{12}\log_2\left(\frac{6}{12}\right) - \frac{6}{12}\log_2\left(\frac{6}{12}\right) = \log_2(2) = 1$
- Middle sample: $-\frac{3}{12}\log_2\left(\frac{3}{12}\right) - \frac{9}{12}\log_2\left(\frac{9}{12}\right)$
- Right sample: $-\frac{0}{12}\log_2\left(\frac{0}{12}\right) - \frac{12}{12}\log_2\left(\frac{12}{12}\right) = 0$

Mutual Information

- MI = difference between P(X, Y) and P(X) P(Y)
  - If X and Y are statistically independent, P(X, Y) = P(X) P(Y), MI = 0

- $\mathrm{MI}(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$

- Difference between distributions
  - Kullback-Leibler (KL) Divergence
  - $D_{KL}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$



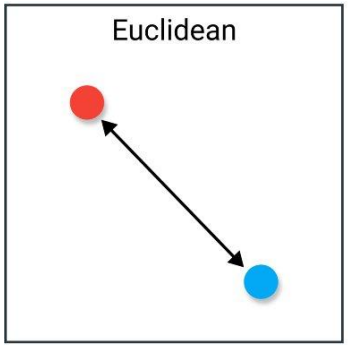
http://mistic.leloir.org.ar/docs/help.html

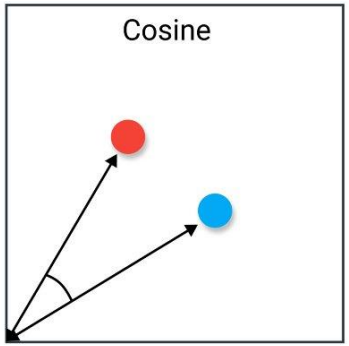
Choosing the right distance (difference)

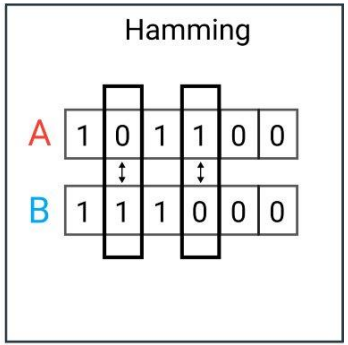
- Difference between patients P61 and P62
- $(4.64-6.18)^2 + (11.23-4.81)^2 + \ldots$
  - Euclidean distance
  - Dominated by highly expressed genes
  - Very sensitive to data scaling
  - Need some **normalization**
- Pearson's versus Spearman's correlation
  - Do you care about magnitude?
- What if there are a lot of zeros?

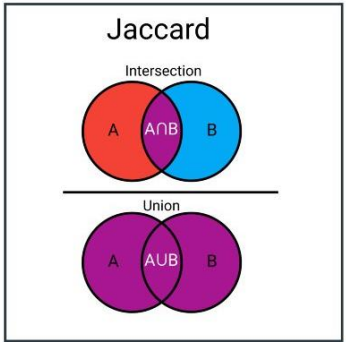
Gene ID	P61_2_C	P62_2_C	P63_2_C	P64_2_C	P68_2_C
ENSG00000000003.14	4.637576	6.183992	5.237635	2.372719	5.665966
ENSG00000000005.5	0	0	0	0	0
ENSG00000000419.12	11.22781	4.813792	2.99782	10.99452	10.7482
ENSG00000000457.13	7.656414	5.082675	7.710682	9.014404	8.488388
ENSG00000000460.16	3.172546	2.245954	5.974815	3.501081	4.162024
ENSG00000000938.12	0	0	0	0.042488	0
ENSG00000000971.15	6.626259	8.19511	5.904925	11.7748	2.050394
ENSG00000001036.13	1.790445	0.76823	3.670635	0.68115	1.894823
ENSG00000001084.11	19.53907	25.08378	11.04872	5.815902	20.23763
ENSG00000001167.14	15.34717	20.00867	17.10001	25.31168	27.41216
ENSG00000001460.17	0.889852	3.090642	0.744581	3.439525	2.417934
ENSG00000001461.16	3.771195	3.12468	1.385353	2.767444	2.973217
ENSG00000001497.16	16.75059	9.662455	15.4965	14.34071	10.62035
ENSG00000001617.11	2.998366	3.712208	3.885852	17.50663	3.019686

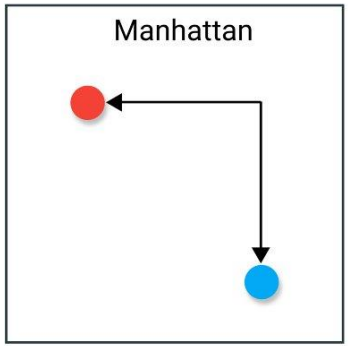
Distance functions







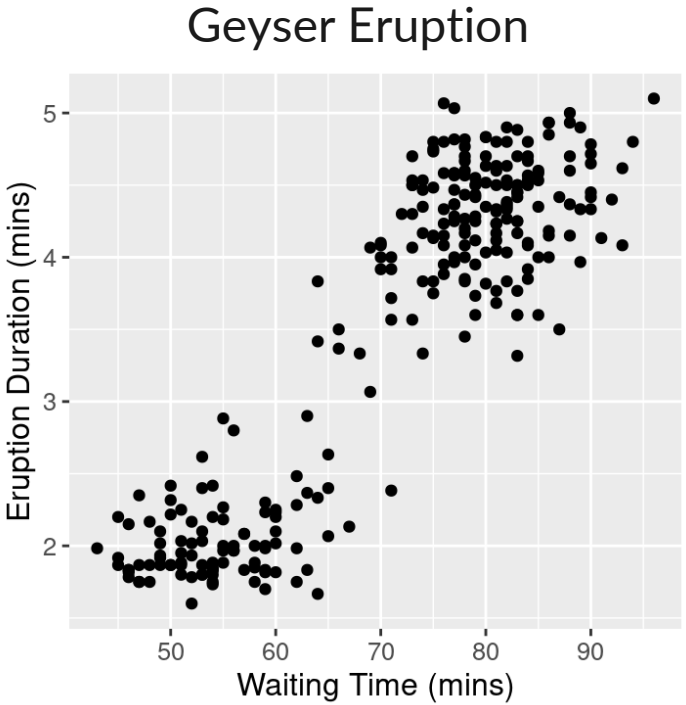




- **Cosine correlation** does not care about **magnitude**
  - $\vec{u} \cdot \vec{v} = |u||v|\cos(\theta)$
  - $\text{CosineCorr}(\vec{u}, \vec{v}) = \cos(\theta)$
- **Hamming distance** ~ number of mutation
- **Jaccard distance** = Intersection / Union

https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Storytelling and interpretation through graph

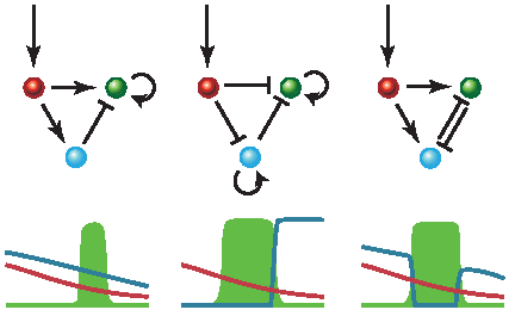


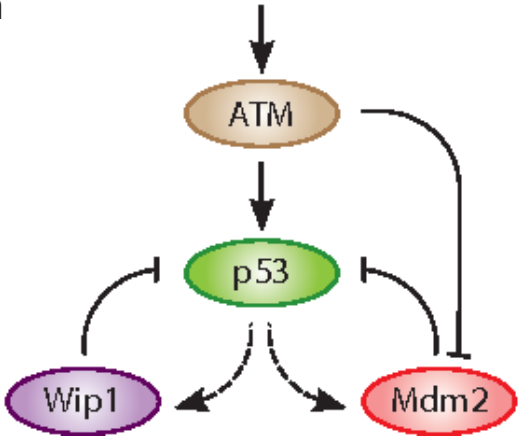
https://datasciencebook.ca/viz.html

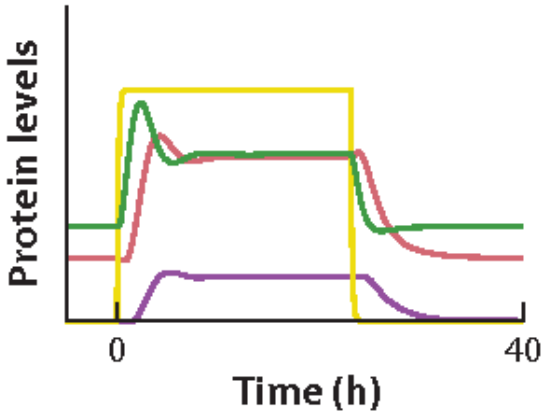
- Eruption time ~ linear function of waiting time
- Two modes of waiting times / eruption times
- The next eruption duration (**Y**) depends on the amount of energy accumulated inside the geyser, which in turn depends on the duration of time since the last eruption (**X**)
- Hence, eruptions after similar waiting times (**X**) tend to have similar durations (**Y**)

What can you do with computational thinking?

- Data analysis
- **Modeling and simulation**
  - Identify mechanisms that underlie the phenomenon or system of interest
  - Develop causal models
  - Study the distribution and behavior of the system
  - Synthesize new data
- Algorithm

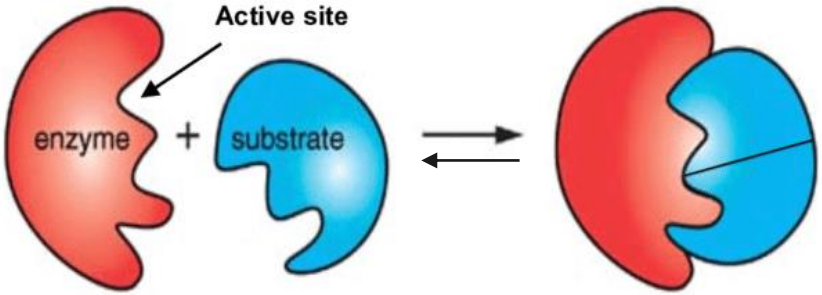






Rue and Garcia-Ojalvo. Annu Rev Biophys 42:605-627 (2013)

Enzyme substrate binding



Graham Hutchings. “Development of new highly active nano gold catalysts for selective oxidation reactions” (2014)

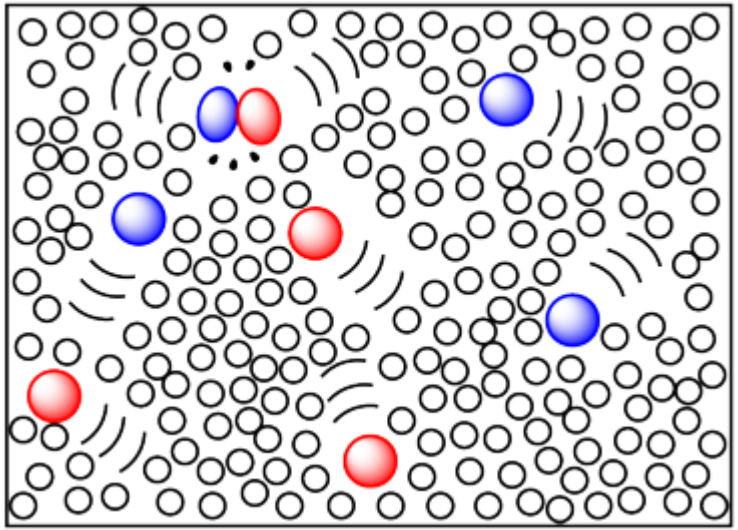
- $K_{dissociation}$ = [E][S] / [E-S] at equilibrium

- Association = P(E meeting S) x $P_{binding}$ = k_1 [E][S]
- Dissociation = Number of E-S molecules x $P_{dissociating}$ = k_2 [E-S]
- At equilibrium, Association = Dissociation
  - k_1 [E][S] = k_2 [E-S] → $K_{dissociation}$ = k_2 / k_1

Mental image for modeling physical system

Rate of E + S → ES =
Rate of E meeting S x
Rate of binding

Rate of E + S → ES =
[E][S] x k



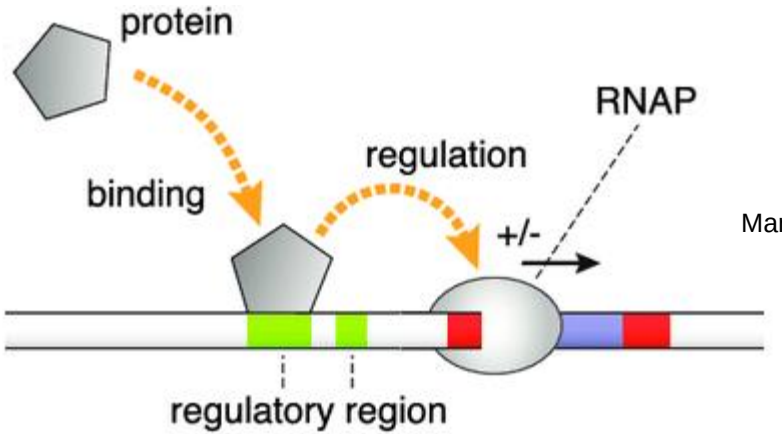
https://employees.csbsju.edu/cschaller/Reactivity/kinetics/rkphase.htm

Rate of E meeting S scales linearly with [E] and [S]

Rate of binding is a constant for E and S

- Molecules must find each other in 3D to bind and interact
- Reaction takes time (and typically energy)

From enzymatic reaction to transcription



Marbach, D. et al. EvoBIO 2007

- $TF + DNA_{inactive} \leftrightarrow TF\text{-}DNA$
- $RNAP + TF\text{-}DNA \rightarrow RNAP + TF\text{-}DNA + RNA$
- $RNA \rightarrow$ degraded RNA

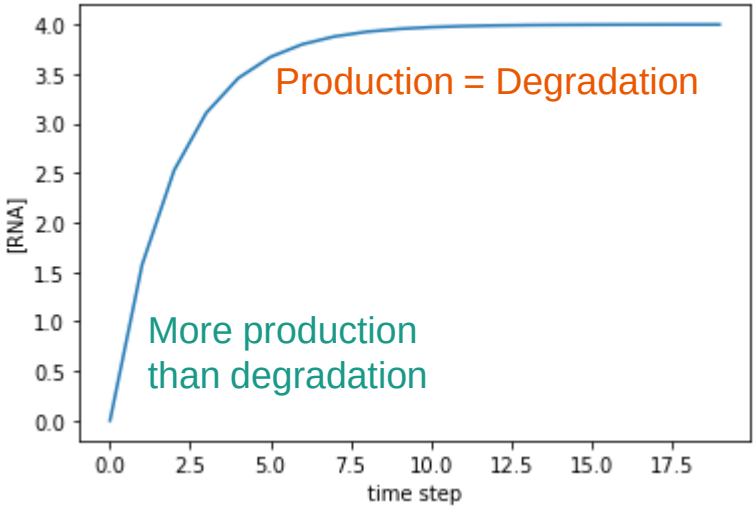
A simple gene expression model

$k_{transcription}$

$$DNA \rightarrow DNA + RNA$$

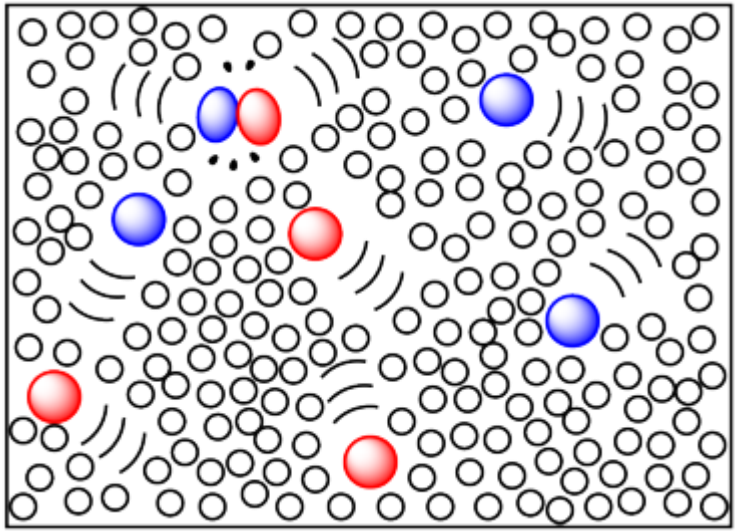
$$RNA \rightarrow \emptyset$$

$k_{degradation}$



- $\frac{d[RNA]}{dt} = k_{transcription} - k_{degradation}[RNA]$
- This is called an **ordinary differential equation**

Multi-gene activation with chromatin folding



https://employees.csbsju.edu/cschaller/Reactivity/kinetics/rkphase.htm

- Genes **A** and **B** are co-regulated by the same transcription factor and must be activated together to respond to signal
- **Scenario 1**: **A** and **B** loci are located at different locations in the nucleus
  - Response = P(TF finds **A**) P(TF finds **B**)
- **Scenario 2**: Chromatin at **A** and **B** loci fold together to form a 3D cluster in nucleus
  - Response = P(TF finds **chromatin cluster**)

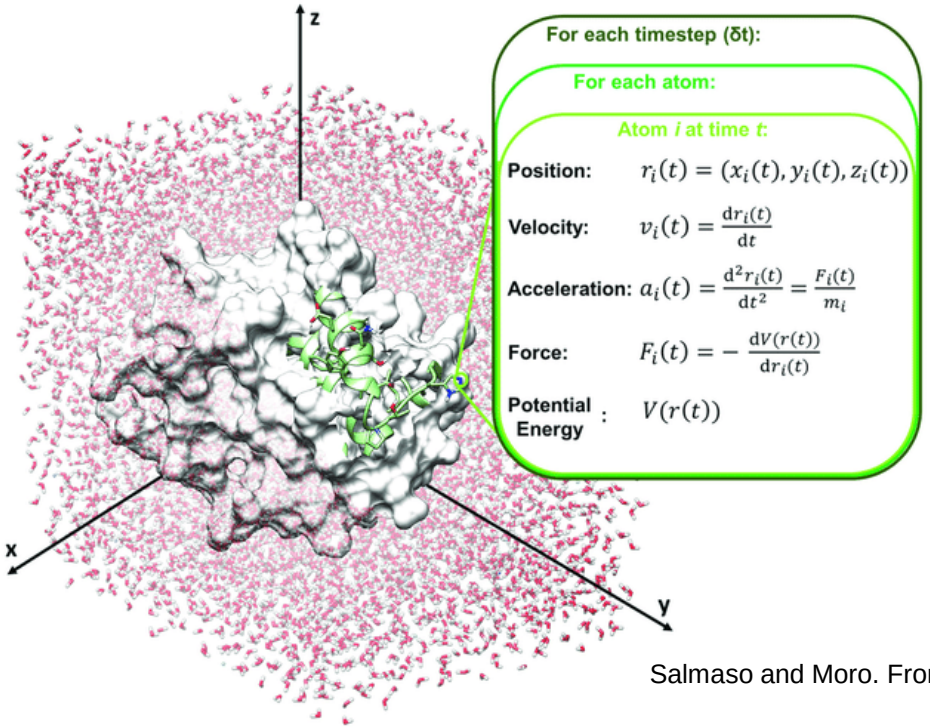
How to simulate a differential equation?

- Define the rate of change in gene expression, $\frac{dx}{dt}$, or $x'(t)$
- Transform from continuous time (calculus) into discrete time (simulation)

- $x_{t+1} = x_t + (x_{t+1} - x_t)$
- $x_{t+1} - x_t$ can be viewed as Δx at time t, or approximately $x'(t)$

- If we start with an initial condition x_0, we can determine x_1, x_2, … by calculating $x'(0)$, $x'(1)$, … and adding them to the current x_i

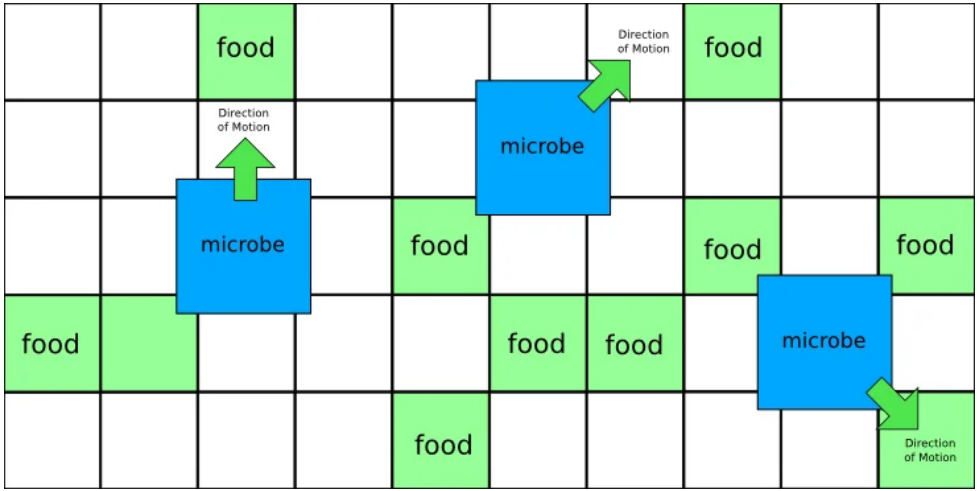
Molecular dynamics simulation



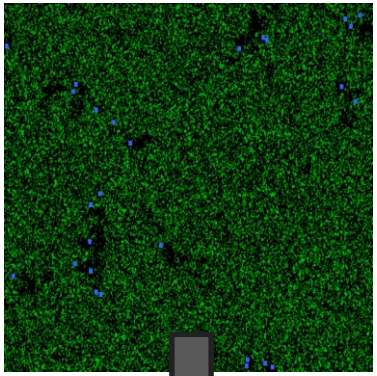
Salmaso and Moro. Frontiers in Pharmacology 9 (2018)

- Applying forces (diffusion, electrostatic, Van der Waals) to the protein, $F = m \times a$

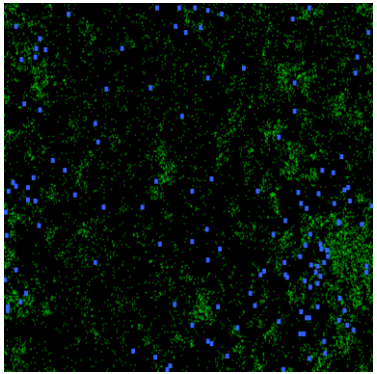
Artificial life



Time = 0



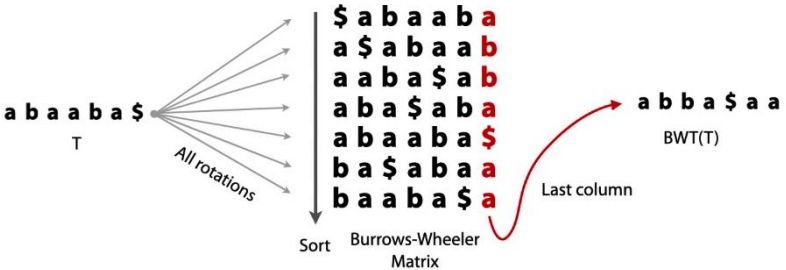
Time = 100



- Model different movement behaviors (probability of moving in a certain direction) with different genes
- See more at **https://beltoforion.de/en/simulated_evolution/**

What can you do with computational thinking?

- Data analysis
- Modeling and simulation
- **Algorithm**
  - Formulating step-by-step instructions
  - Understand the strength and weakness of computer program



GCAT-GCG
G-ATTACA

or

GCA-TGCG
G-ATTACA

Needleman-Wunsch

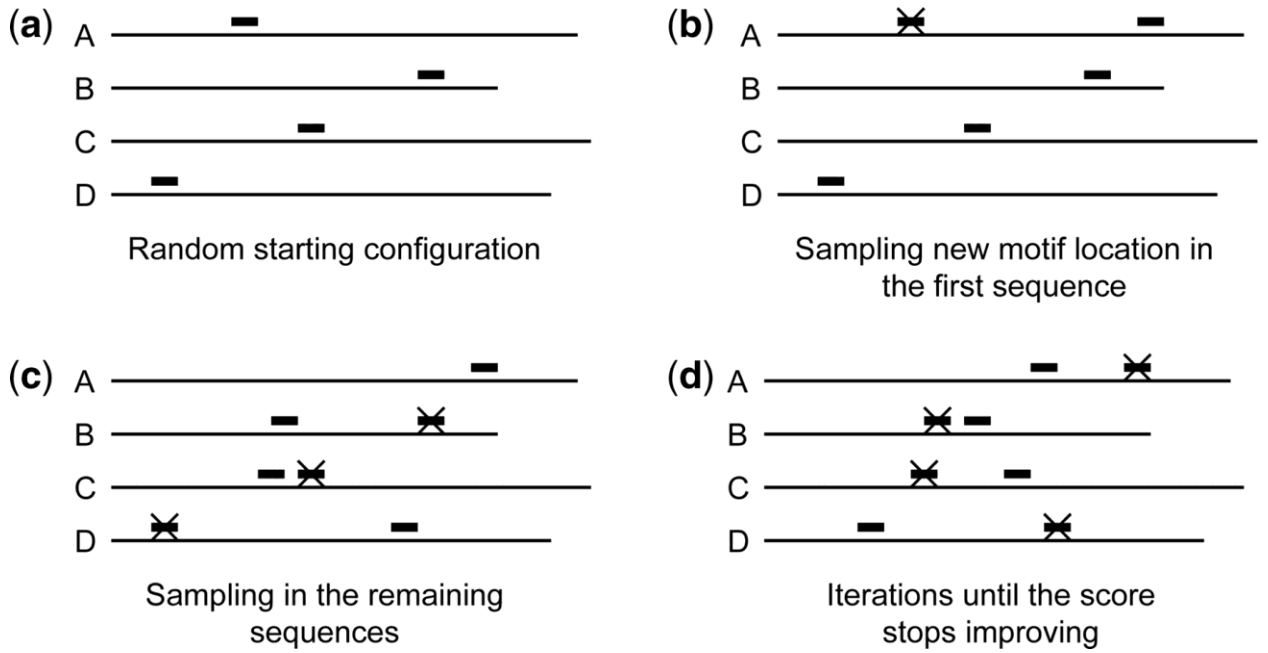
match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	G
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Burrows and Wheeler. A block sorting lossless data compression algorithm (1994)

https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

DNA binding motif discovery



- The key is how to sample and trying multiple random initial conditions

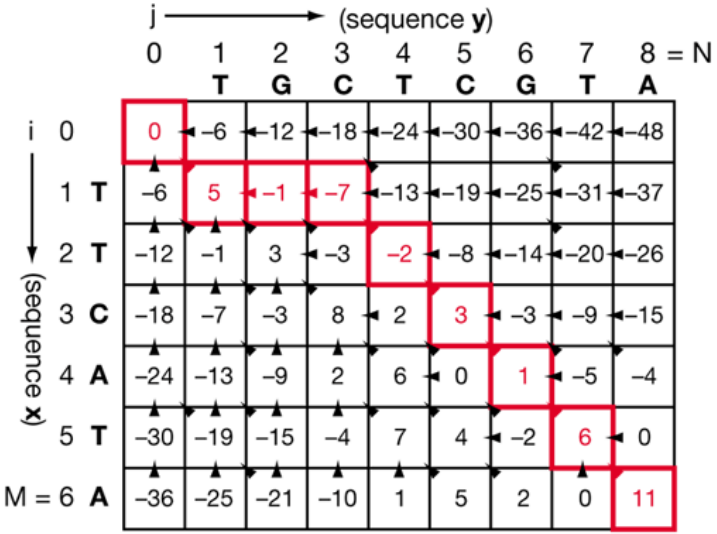
Dynamic programming

No. of rocks	1	2	3	4	5	6	7	8	9	10
Winner										
No. of rocks	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Winner										?

- Dynamic programming build the solution of complex problem using on the solutions of simpler ones
- There is a pile of 20 rocks. Two players take turns by removing 1 or 2 rocks from the pile. Whoever removes the last rock(s) win. Who is the winner?

Dynamic programming for sequence alignment

Dynamic programming matrix:



Optimum alignment scores 11:

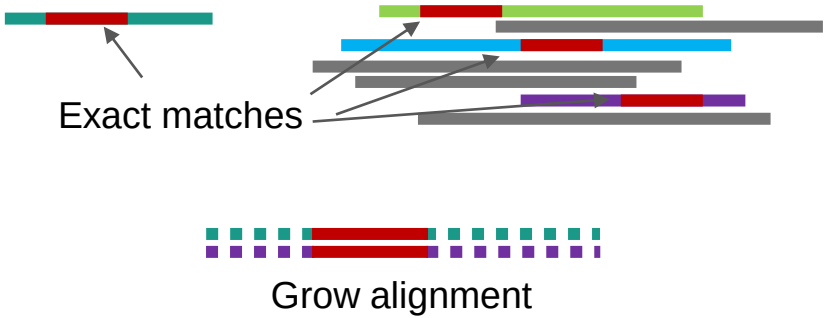
T	–	–	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	–6	–6	+5	+5	–2	+5	+5

- The best alignments for long sequences depend on the best alignments of shorter sequences

- The best alignment for TTCATA vs TGCTCGTA is either
  - T/T + best alignment for TCATA vs GCTCGTA
  - T/– + best alignment for TCATA vs TGCTCGTA
  - –/T + best alignment for TTCATA vs GCTCGTA

Eddy, S.R. Nature Biotechnology 22:909-10 (2004)

BLAST sequence search

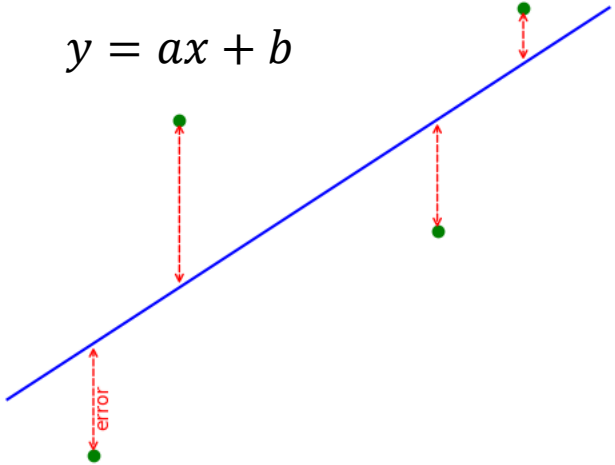


- Searching millions of sequences still take time
- Can we speed up by ignoring obviously unmatchable sequences?

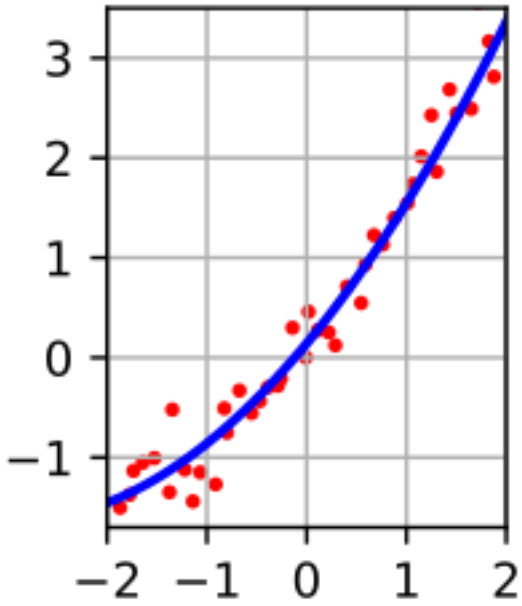
- High similarity implies a run of identical sequences
  - 95% similarity = 5 mismatches in 100 positions = a run of 19 matches!
  - Find these runs of matches first

Curve fitting and optimization

Curve fitting with least square



https://en.wikipedia.org/wiki/Least_squares

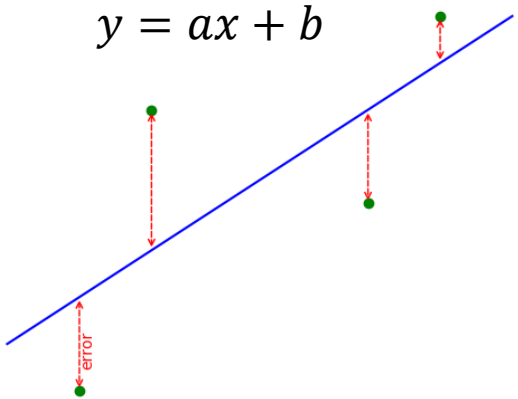


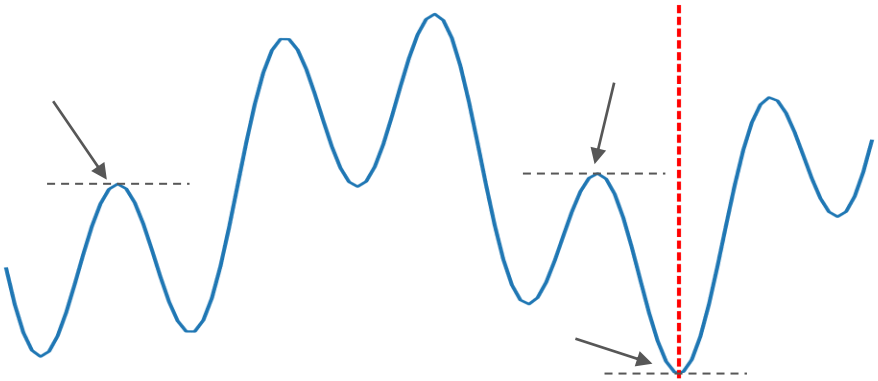
$y = ax^2 + bx + c$

- Finding the **best** *a*, *b*, and *c* that make the curve fit to the observations
- Minimize least square error $\frac{1}{n}\left((\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2\right)$

How can we find the best parameters

- Minimize $\mathrm{L}(a,b) = \frac{1}{n}\left((\hat{y}_1 - y_1)^2 + \cdots + (\hat{y}_n - y_n)^2\right)$
- Randomly search for (a, b)
  - Heuristic guess: $a = \frac{\sum y_i}{\sum x_i}$
- Use calculus
  - At optimal, slopes are zero
  - Solve $\frac{\partial \mathrm{L}(a,b)}{\partial a} = 0$ and $\frac{\partial \mathrm{L}(a,b)}{\partial b} = 0$





Newton-Ralphson method

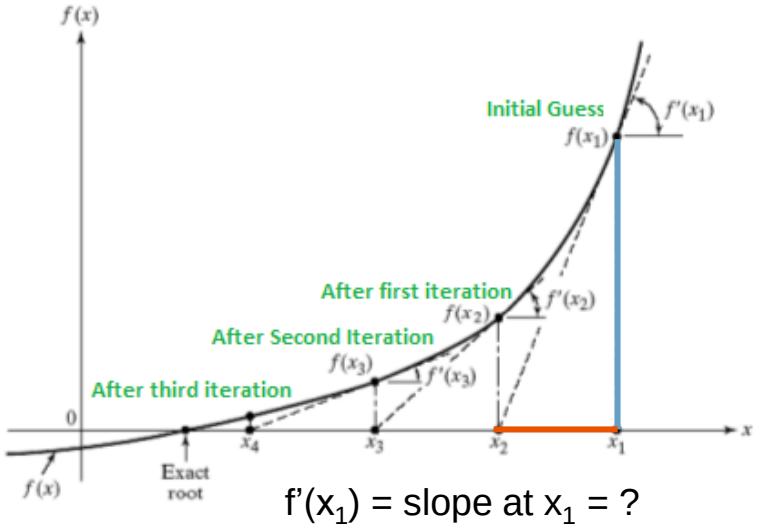
- Want to maximize L(x) by solving

$$f(x) = \frac{dL(x)}{dx} = 0$$

- Start with an initial guess x_1
- Calculate $f(x)$ and $f'(x)$ at x_1
- Define $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$

- Repeat the process for x_3, x_4, …
- Stop when x_i converges to a value

f(x) = the first derivative of the objective L(x)



f'(x_1) = slope at x_1 = ?

https://predictivehacks.com/newton-raphson-method-in-python/

Gradient descent

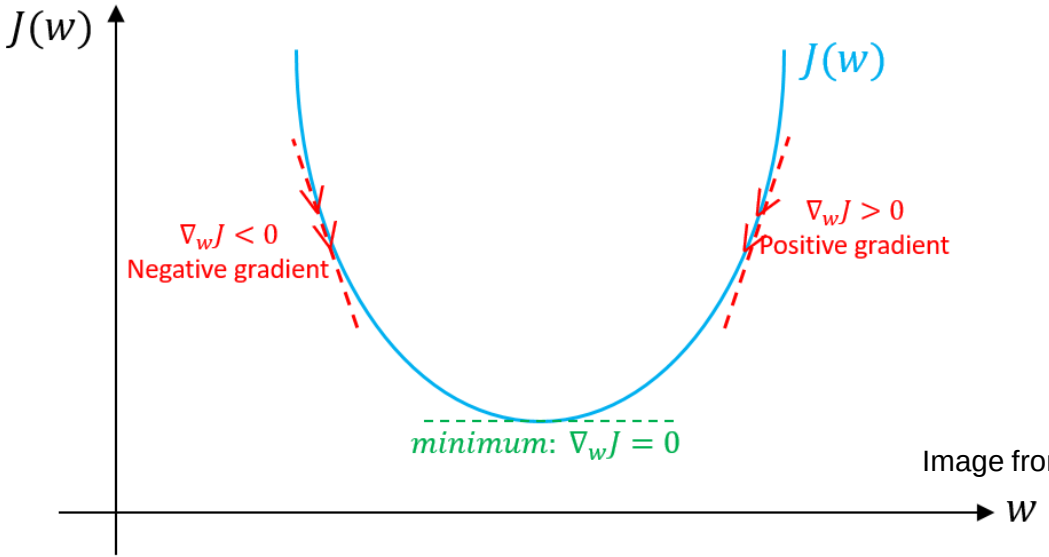
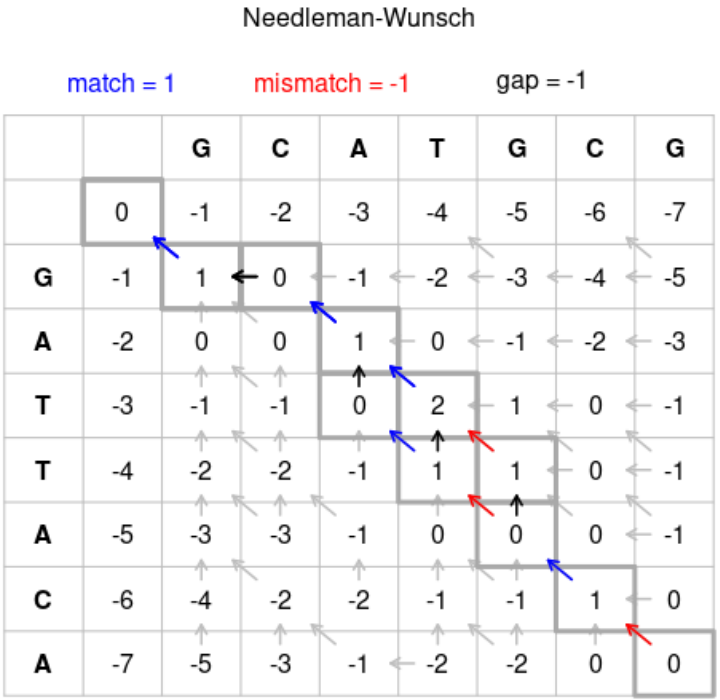


Image from towarddatascience.com

- Gradient is just a generalization of slope in multi-dimension
- Gradient at optimal point = 0
- Gradient points toward the direction of increase in f(x)

(Almost) all algorithms involves optimization

- **Curve fitting** = minimize square error
- **Sequence alignment** = maximize matching between two sequences
- **Protein docking** = minimize energy
- **Motif discovery** = maximize matching between a motif pattern and all sequences

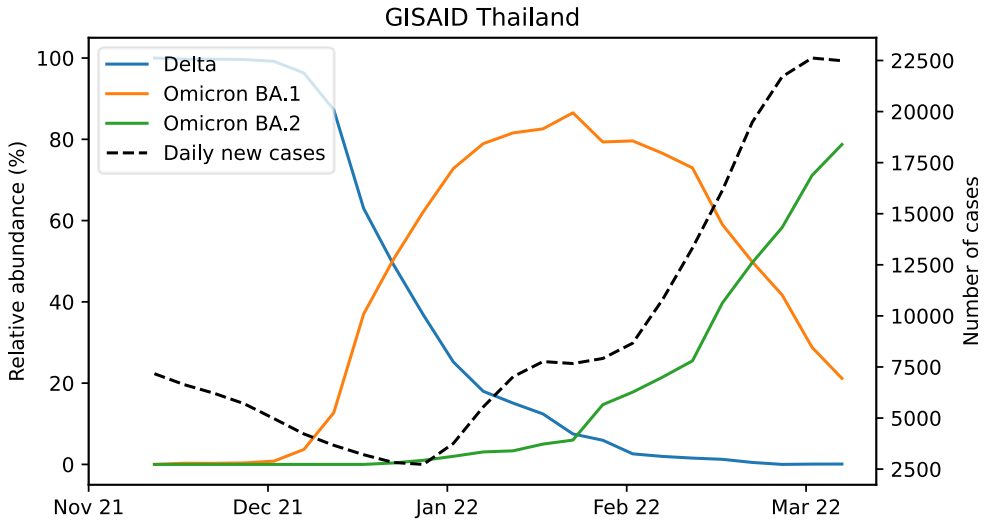


Knapsack problem

https://en.wikipedia.org/wiki/Knapsack_problem

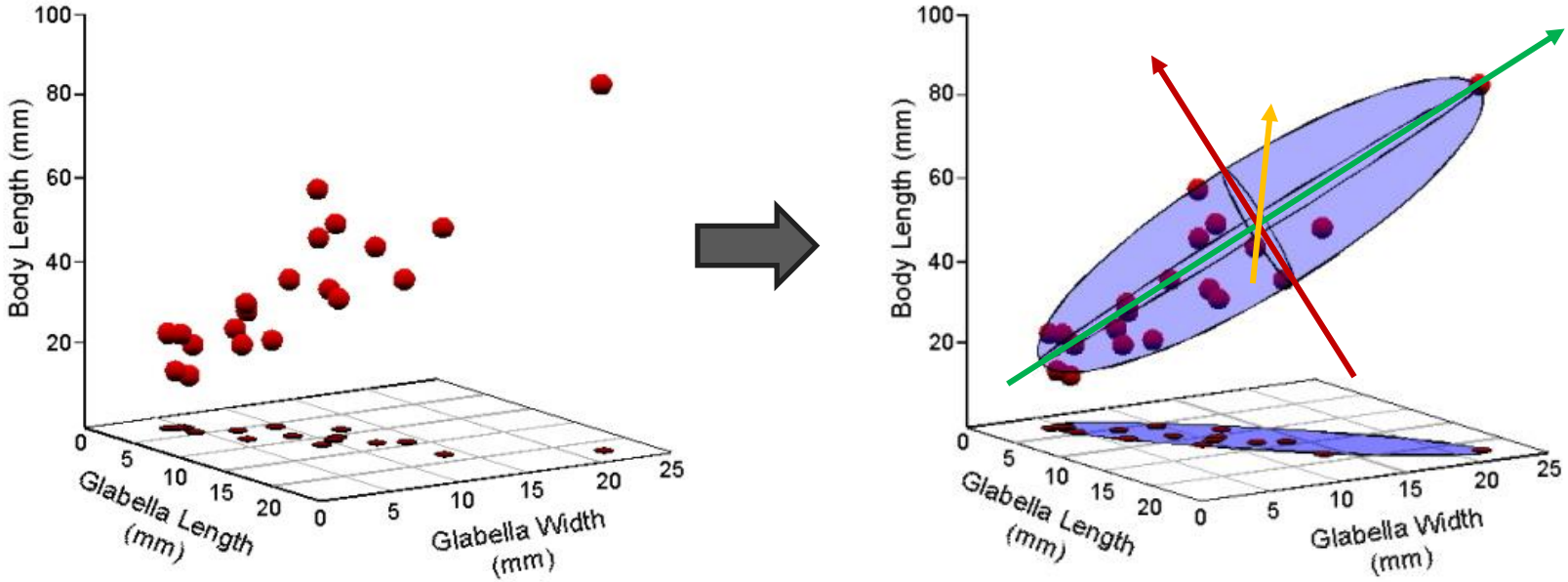
- Limited capacity
- Want to pick as much value as possible while not exceeding the capacity
- General optimization setting
  - **Objective** = Total value
  - **Constraint** = Total weight ≤ W
- Creating a genome from genes with fitness values and maintenance costs

Estimating COVID-19 transmission



- (# new case) = (# past Delta) x (Delta transmission rate) + (# past Omicron) x (Omicron transmission rate)
- Find transmission rates that best fit to the observed number of cases

Principal component analysis (PCA)

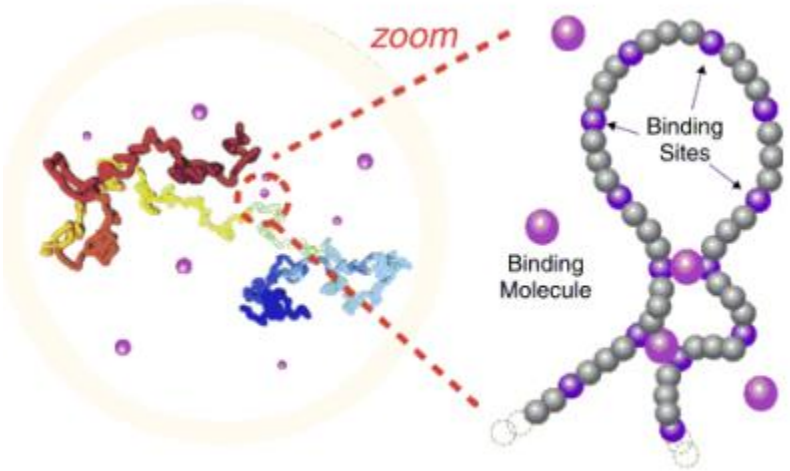


Source: the paleontological association

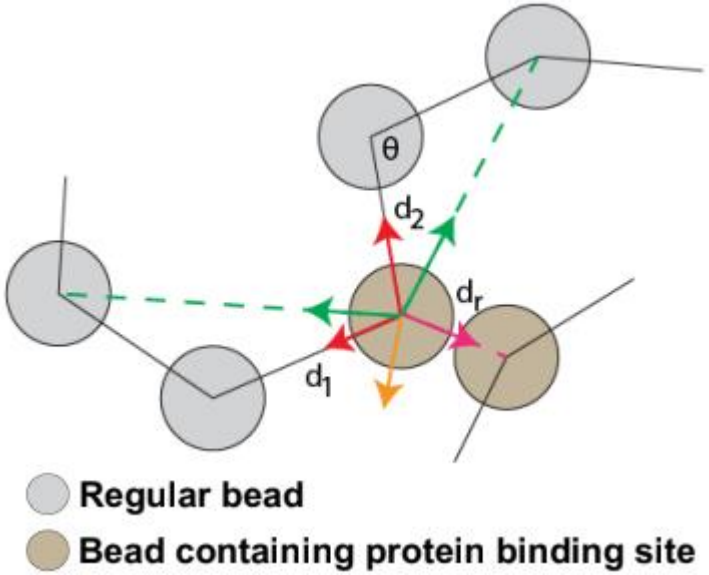
- Find new axes such that the variances are maximized

Simulation

Physical simulation

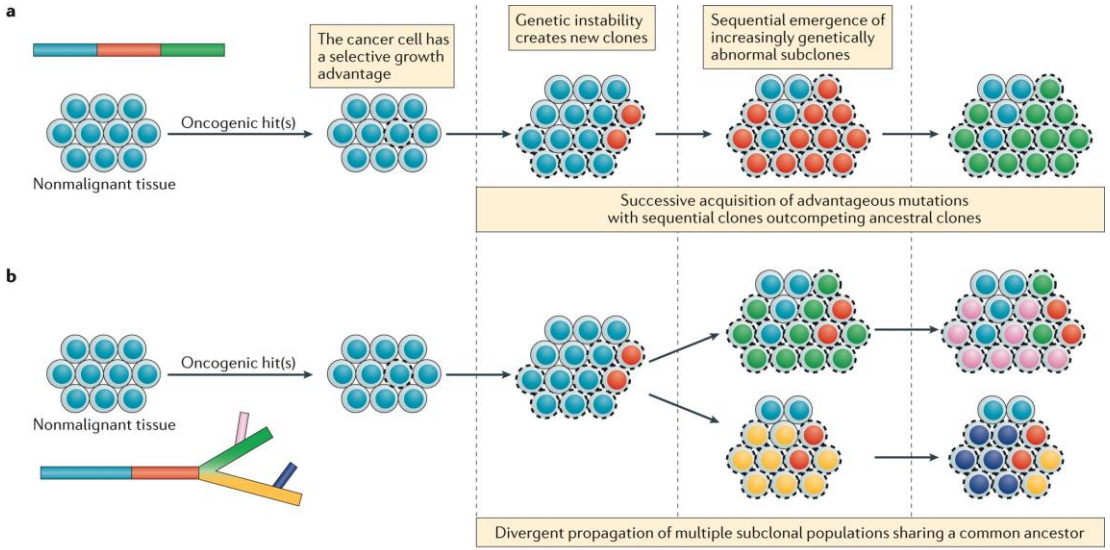


Nicodemi and Pombo. Curr Opin Cell Biol (2014)



- Consider relevant laws of Physics
- Transform forces into motion, $F = m \times a$
- Update **velocity** from **acceleration** → Update **position** from **velocity**

Tumor growth

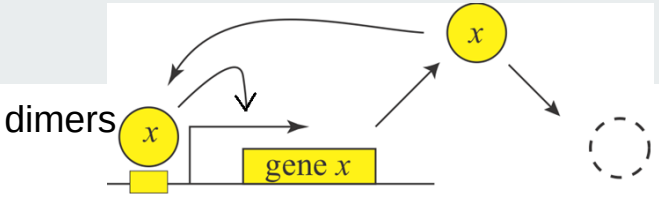


Dagogo-Jack and Shaw, Nat Rev Clin Oncol (2017)

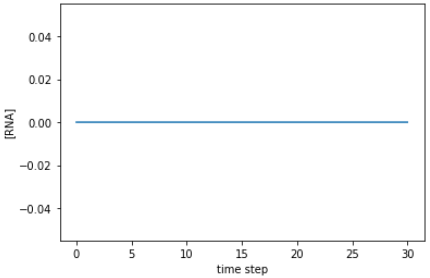
Nature Reviews | Clinical Oncology

- Cell's actions: Gain mutation, Divide, Die
  - What are the parameters influencing these actions?

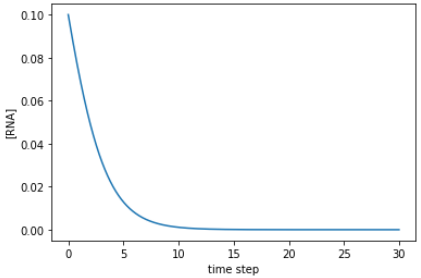
Simulating differential equation



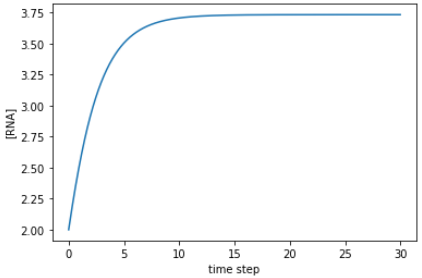
Initial [X] = 0

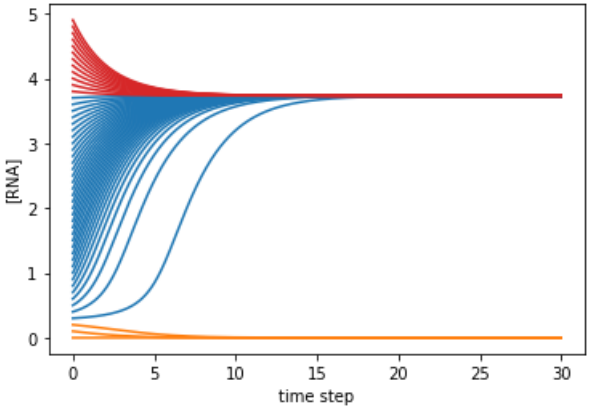


Initial [X] = 0.1



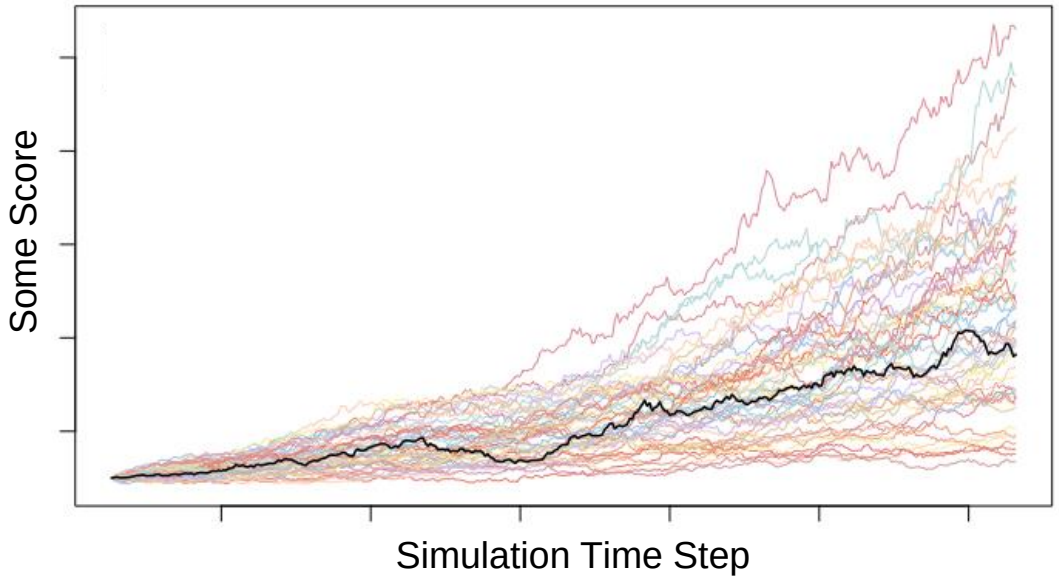
Initial [X] = 2





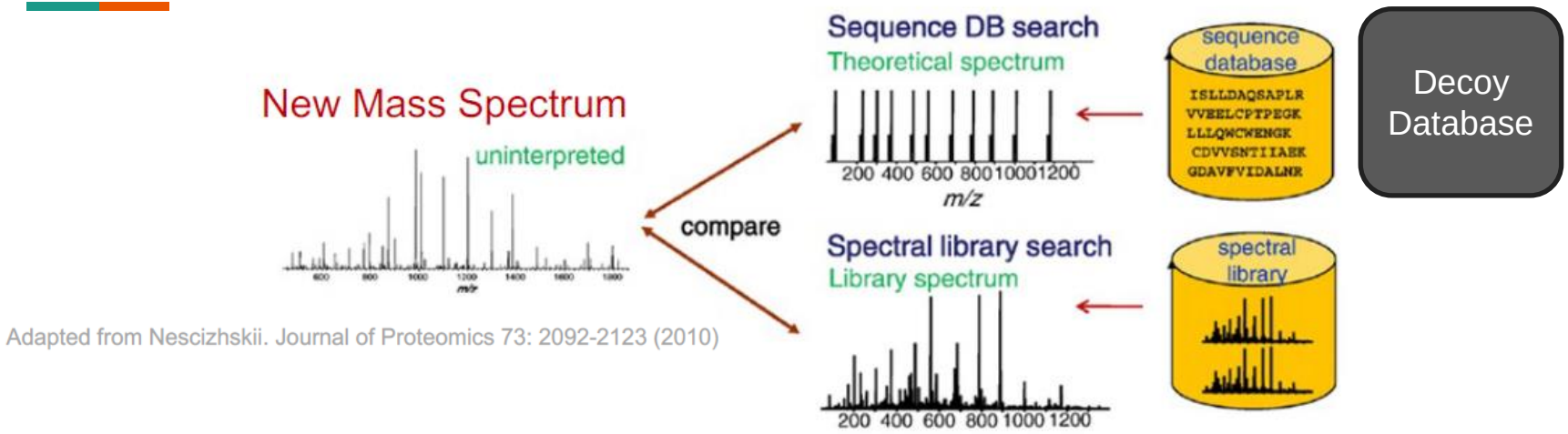
- At equilibrium, $\frac{k_{transcription}(k[X])^2}{1+(k[X])^2} = k_{degradation}[X]$
- There are two solutions, one of which is $[X] = 0$
- Depending on the current $[X]$, the system can converge to either equilibrium
- [X] at equilibrium depends on the constants in the equation

Monte Carlo technique



- Perform repeated sampling to explore broad parameter scenarios
- Provide estimates for the probability of different scenarios and outcomes
- May require millions of repetition (can utilize multiple CPUs)

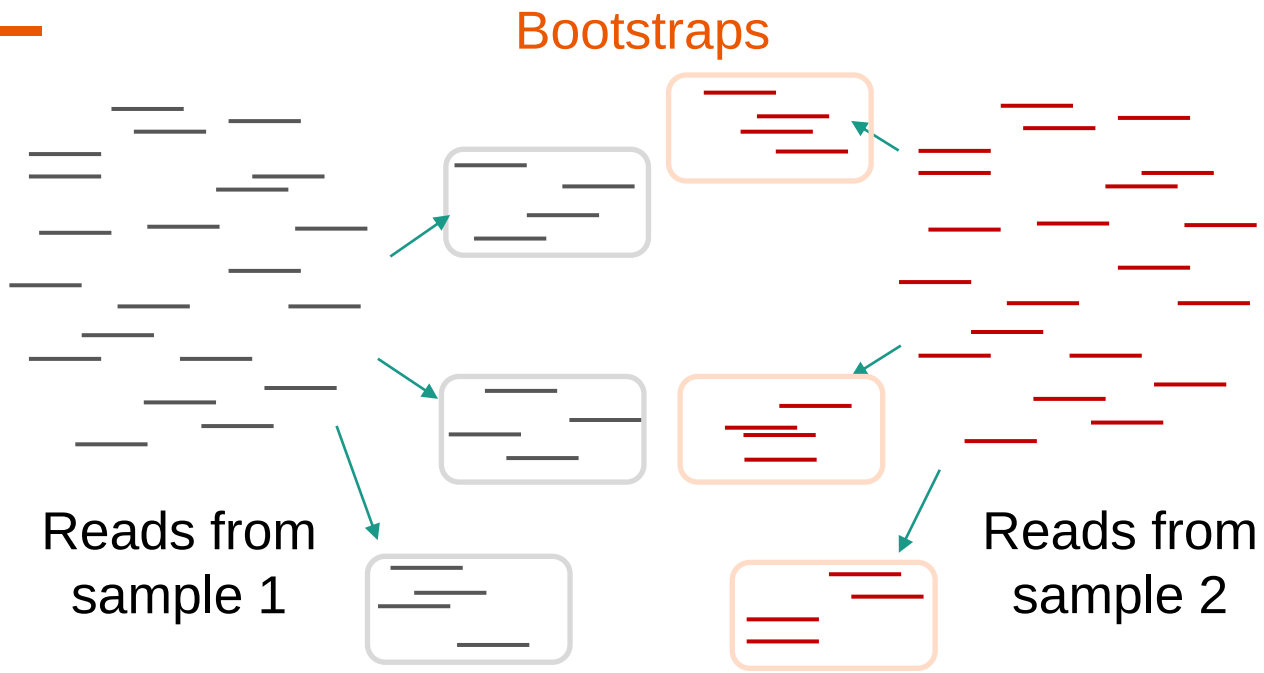
Decoys for database search



Adapted from Nesczihskii. Journal of Proteomics 73: 2092-2123 (2010)

- No statistical model to calculate *p*-value for peptide database search
- Need to estimate the number of positives
  - Generate a decoy database of peptides that shouldn't be in the sample
  - Number of false positives ~ number of hits to the decoy database
  - Sample peptide lengths and amino acid frequencies

Bootstrapping



- Instead of using the whole data to perform one calculation
- Sampling from data to perform multiple calculations → Estimate variance

Summary

- Computational thinking is about **transforming abstract hypothesis into a quantitative statement** that can be tested with data
- **Identify relevant components and mechanisms** that underlie the phenomenon and **develop the mathematical models**
- Use **mathematics to formulate score functions**
- Almost every algorithm boils down to **an optimization problem**

More recommended companion courses

6.0001 I Fall 2016 I Undergraduate

Introduction To Computer Science And Programming In Python

6.0001 Introduction to Computer Science and Programming in Python is intended for students with little or no programming experience. It aims to provide students with an understanding of the role computation can play in solving problems and to help students, regardless of their major, feel justifiably confident of …Show more

6.0002 I Fall 2016 I Undergraduate

Introduction To Computational Thinking And Data Science

6.0002 is the continuation of *6.0001 Introduction to Computer Science and Programming in Python* and is intended for students with little or no programming experience. It aims to provide students with an understanding of the role …Show more

Any question?

- See you on August 24th