Sequence Bioinformatics

Optimal pairwise sequence alignment

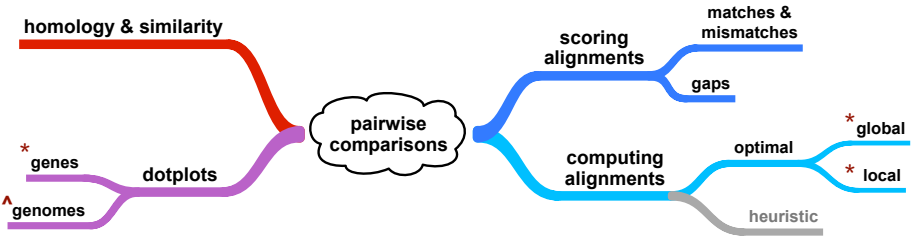
Dec 05, 2023

Stefanie Hartmann

Wintersemester 2023/ 2024

Universität Potsdam

Today's topics



reading

- HB Nicholas, Jr, DW Deerfield, and AJ Ropelewski. Strategies for searching sequence databases. Biotechniques, 28(6):1174–1191, 2000. (Introduction, Search Algorithm, Similarity Matrices, Scoring insertions and deletions)

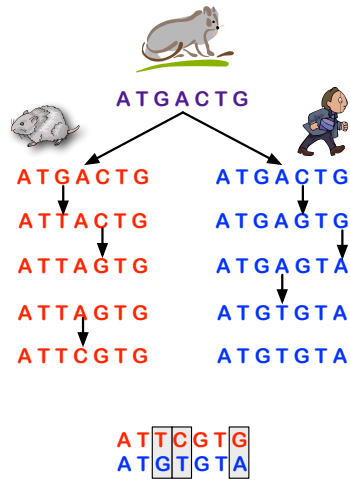
Homology & similarity

homologs

- morphological structures, sequences, domains, alignment positions....
- derived from a common ancestor
- all-or-nothing condition

similarity / identity

- a quantitative measure



Homology & similarity

존나게 중요한 개념이다

homologs, orthologs*, paralogs*, etc.*

- very very very important concepts!
  - clear descriptions, transfer of annotations, evolutionary & functional genomics, …
- evolutionary definition, requires phylogenetic inference, gene & species tree

동일한 조상에서 유래했다는 정의 근데 우리는 타임머신이 없기때문에 직접적으로 확인할 수 없다.

여러 메소드를통해 homology를 판별할 수 잇다.

- cannot be observed or known, just inferred

similarity

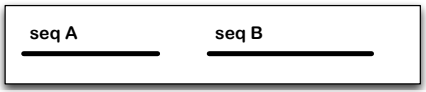
- can be observed & measured & evaluated
- can be used to infer homology, orthology, paralogy?

*will be covered later

Comparative sequence analysis

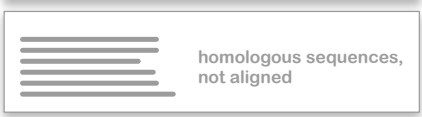
starting with...

두 시퀀스를 비교하는 방식으로 homology를 판별할것임.



similarity / homology?
compute (optimal) alignment

find database sequences that are similar (homologous) to the query sequence

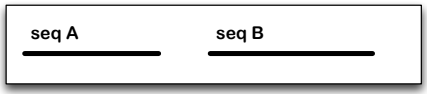


compute a multiple sequence alignment

model the alignment, find additional family members

Pairwise sequence comparison

starting with...



similarity / homology?
compute optimal alignment

즉, functional domain같은 특정 부위의 유사성만 가지나? 아니면 전체적인 유전자가 global하게 ㅇ유사성을 가지나?

graphical overview: dotplot

- genes: (where) are they similar?
  are they locally or globally similar ?
- genomes: syntenic regions? rearrangements?
  유전자 수준뿐만 아니라 아미누산 단위에서도 비슷한가?

alignment

- how to evaluate different alignments?
- how to compute optimal local & global alignments?

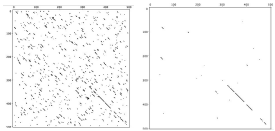
Dotplots

	A	C	T	T	G	C	A	C	T	A
C		X				X		X		
C		X				X		X		
A	X						X			X
C		X				X		X		
T			X	X					X	
G					X					
A	X						X			X
T			X	X					X	
A	X						X			X
A	X						X			X

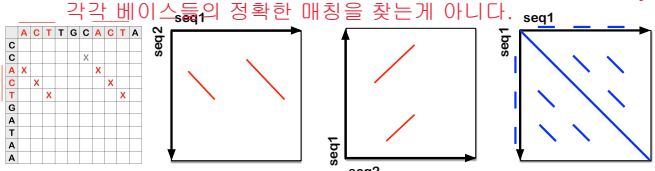
연속적으로 동일한 시퀀스는 더 높은 점수를 가진다.

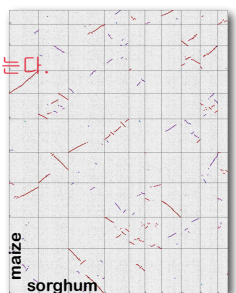
Dotplots

- **individual positions (nt, aa) usually not readable/given**
- **signal**
  - **identity, similarity**
  - **length of consecutive signals**
- **compare two sequences,**
  or compare one sequence against itself 많은 노이즈가 나타난다.



우리는 직접적인 베이스들을 보는게 아니라
이런 매칭 라인 형태로 보게 된다. 이 라인들을 통해 similarity를 찾는다.
각각 베이스들의 정확한 매칭을 찾는게 아니다.





Dotplots in practice

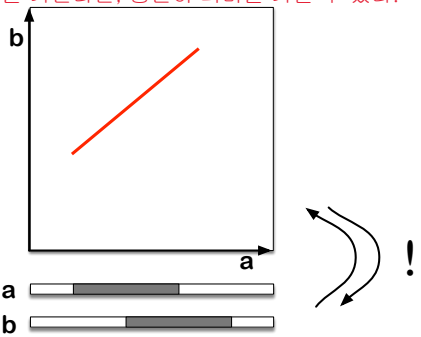
이렇게 매칭된 구역이 실제 gene이 아닐 수 있다...
그러나, non coding region이라도, 정확하게 우리가 예상하는
evolutionaly distance를 가진다면, 충분히 의미를 가질 수 있다.

this week:

- **organellar genomes**
- **MUMmer**

next week:

- **protein sequences**
- **EMBOSS**



위의 라인 매칭에서, 아래의 alignment로 변환 혹은 반대방향으
과정이 가능하도록 연습. `

Pairwise sequence alignment (genes):

similarity가 발견이 된다면, alignment를 시작 할 수 있다.

the comparison & arranging of two sequences by

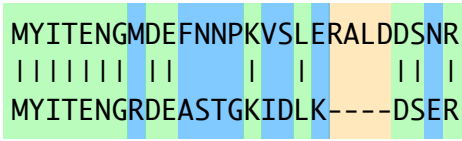
- **searching for pairwise matches and "good mismatches" between their characters**
- **possibly inserting gaps in each sequence**

```
seq1     VQ--PKLQAHAGKVFGMVRDSAAQL
         |   ||  ||  || |   |  | |
seq2     VMGNPKVKAHGKKVLG---DGLAHL
```

여러 갭이나, 포인트 뮤테이션이 있다고 해서 그게 다른게 아니다!!

Scoring an alignment

- matches
- mismatches
- gaps



우리의 alignment가 실제로맞을지, scoring을 통해 점수를 매기고
"better maching"을 찾을 수 있다.

- a **scoring matrix** is used for matches and mismatches
- **gap penalties** are used for opening & extending gaps
- alignment score: sum of scores at each alignment position, $S = \sum s_{ij}$

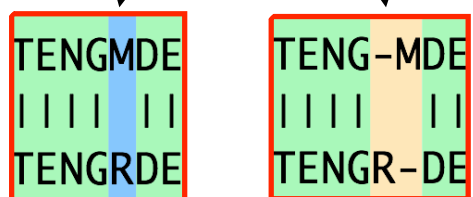
→ **chosen scoring scheme** affects the result!

Scoring an alignment

- matches
- mismatches
- gaps

여러 가능성이 존재한다!

??



내가 어떤 스코어링 알고리즘, 뭘 더 중요하고 뭘더 낫게 생각하는지에 따라, 결과에 크게 영향을 미칠 수 있다.

chosen scoring scheme affects the result!

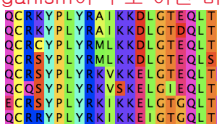
스코어링 점수를 어떻게 설정하고, 스코어링을 어떻게 할 것인가?

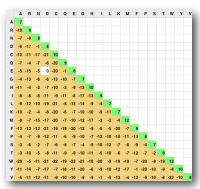
How to obtain a scoring matrix?

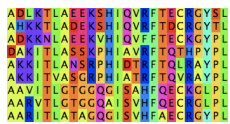
- observe trusted alignments of related proteins
  - which residues are paired?
    (i.e., which substitutions have occurred?)
- different values for sequences of different evolutionary divergence!
  - different scoring matrices!

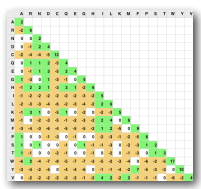
진화적 방향에 따라, 선호되는 변화, 즉 scoring matrix가 변화할 수 있다.

각 mutation이 어떤 영향을 미치는지, 기능에 어떤 영향을 미치는지,
혹은 그 organism이 주로 어떤 미스매치를 선호하는지



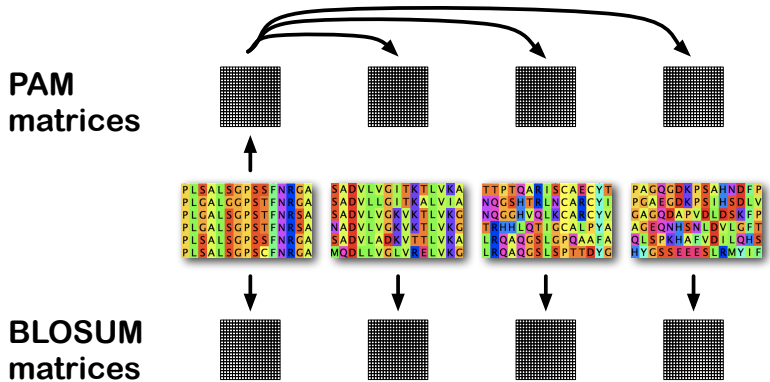






PAM & BLOSUM matrices

이미 스코어링 매트릭스는 완성이 되었따.



PAM matrices

그 정보가 없던 1970년에 마들어졌는데, 아직도 잘 작동한다.

- Point Accepted Mutation, compiled by Margaret Dayhoff and her colleagues in the 1970s
- based on observed amino acid substitutions in families of evolutionarily related proteins
- PAM1 implies 1 substitution per 100 amino acids
- 추론하다
- extrapolation of values for more distantly related proteins:
  $PAM250 = (PAM1)^{250}$
- guidelines:
  - use PAM250 for proteins of 20% identity
    PAM120 40%
    PAM60 60%

BLOSUM matrices

- BLOcks amino acid SUbstitution Matrices
- based on local alignments of divergent sequences
- different BLOSUM matrices are not extrapolated but based on observed alignments
  - BLOSUM50 matrix is derived from alignments of sequences that are 50% identical

- guideline: a BLOSUM matrix index should approximately match the percent identity of the sequences to be aligned
  - BLOSUM50 matrix is best used for sequences that are 50% identical

신기하게도 pam과 blosum은 같은 결과를 낸다.

BLOSUM62

보통 프로그램에서 디폴트임.

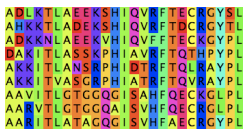
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

가능한 모든 alignment 가능성에 대해 점수를 더한다음, 점수가 가장 나은것을 선택

Substitution matrices

➡ which amino acids occur together in the alignment columns more often than expected by chance?

homologous sequences



$$s(a,b) = log\left(\frac{p_{ab}}{q_a q_b}\right)$$

p_{ab} : observed frequency of residues a and b aligned

q_a, q_b : frequencies of residues a and b

자주 나타나는 pair또한 존재하고, 이런 페어는 일반 M,L따로있는것보다 더 점수가 높을 수도 있다. 아니면 추가점수를 받을 수도 있음/.

```
M:    0.01
L:    0.1
ML:   0.002
```

$$s(M,L) = log(\frac{0.002}{0.01*0.1}) = +1$$

Substitution matrices

proteins

- **all-purpose matrices from & for sequences with a range of evolutionary divergence (PAM, BLOSUM, WAG, …)**
  - **empirical: "counting" or probabilistic approaches**
- **matrices for specialized types of proteins (e.g., transmembrane) or taxonomic lineages**
- **matrices based on the constraints on protein structure (folding, site-site interactions, etc)**

DNA

- **different matrices exist**
- **I'll cover some of them later (phylogenetics!)**

Gap penalties

linear score 모든 갭이 같은 panalties를 가진다

- **each gap position receives independent penalty of *p***

affine gap penalties 갭의 크기에따라, 즉 열림-확장에 따라 다른 패널티를 부여

- **score depends on the length of the contiguous gap**
- **larger gap opening penalty: *d*** 생물학적으로 더 말이 되는 이유: 보통 아무리 큰 갭이라고 하더라도, 한번의 insertion이라는 싱글 이벤트이기 때문.
- **smaller gap extension penalty: *e***

```
CTT--GAT        ATAG-GCT
CTTATGAC        AT-GTGCC
```

p + p	***p + p***
d + e	***d + d***

What is an "optimal" alignment?

뭐가 optimal일지는 나의 관점, 목적에 따라 완전히 다르다!

- **functionally correct** (optimal) alignment
  - aligned residues have the same function
- **structurally correct** (optimal) alignment
  - aligned residues play a similar role / are in corresponding positions in the 3D structure
- **evolutionarily correct** (optimal) alignment
  - aligned residues are homologous, i.e. share a common ancestry
- **algorithmically correct** (optimal) alignment
  - the highest-scoring alignment for a given substitution model and gap penalties

위 세ㄱ개는 보통 우리가 얻을 수 없다.. 우리가 얻을 수 있는 optimal은 오직 algorithmically correct 뿐이다.
alignment는 결국 우리의 substitution matrix에 따라 결정된다. 알고리즘적 최적은 올바른 선택을 통해 이루어 질 수 있고, 이게 실제로 위 세개의 최적과 맞을지는.. 기도메타뿐이다.

Pairwise alignments

GOAL:

- optimal (highest-scoring) pairwise alignment

PROBLEM:

- constructing and scoring all possible alignments and picking the best one is not an option!

number of possible global alignments: $\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{(\pi n)}}$

length of two sequences	number of possible alignments
5	56
10	252
20	15,504
50	658,008
100	2,535,650,040
250	255,244,687,600

SOLUTION:

- dynamic programming

Dynamic programming

J. Mol. Biol. **(1970) 48, 443-453**

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH

Dynamic programming

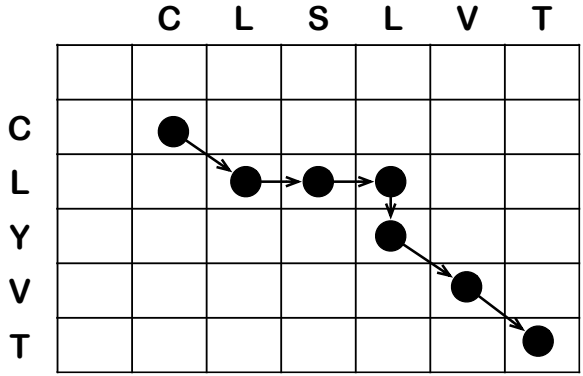
- used for optimization problems
- problems are broken into smaller, nested subproblems
- solutions to subproblems are computed and stored
  - these are used to construct solutions to larger and larger portions of the original problem

- alignment problem:
  - build up the best alignment by using optimal alignments of smaller subsequences

Pairwise alignments

		C	L	S	L	V	T
C							
L							
Y							
V							
T							

Pairwise alignments

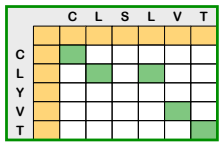


CLSL-VT
CL--YVT

DP: optimal pairwise alignment

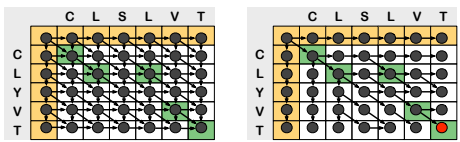
1. initialization of the score matrix

스코어 매트릭스를 만든다.

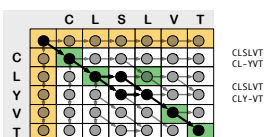


2. matrix fill (calculate alignment score)

각각 next step의 score를 모두 기입한다.



3. traceback and deduction of alignment



도착점부터 시작>?

DP: initialization

match: +1
mismatch: -1
gap: -2

		C	L	S	L	V	T
	0	→ -2	→ -4	→ -6	→ -8	→ -10	→ -12
C	↓ -2	+1	-1	-1	-1	-1	-1
L	↓ -4	-1	+1	-1	+1	-1	-1
Y	↓ -6	-1	-1	-1	-1	-1	-1
V	↓ -8	-1	-1	-1	-1	+1	-1
T	↓ -10	-1	-1	-1	-1	-1	+1

매 스텝마다, 어느 방향으로 움직이는것이 가장 좋을지 선택해서 남겨둔다.
만약 두 스텝이 점수가 같다면, 두개 다 남긴다.

내가 선택한, 가장 좋은 선택지만 기억하고 나머지는 잊어버린다.

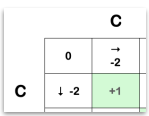
DP: scoring

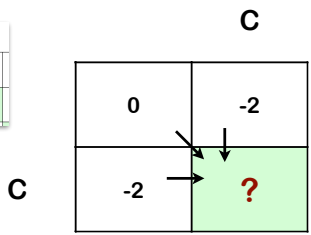
$$S_{i,j} = max \begin{cases} S_{i-1,j-1} + s(x_i, y_i) & x_i, y_j \text{(mis)match} \\ S_{i-1,j} + \delta & x_i \text{ aligned to gap in } y \\ S_{i,j-1} + \delta & y_j \text{ aligned to gap in } x \end{cases}$$

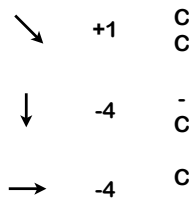
↘ ↓ →

$s(x_i, y_i)$: substitution score for residues ***i, j***
(example: match = 1, mismatch = -1)

δ: **gap penalty (example: gap = -2)**



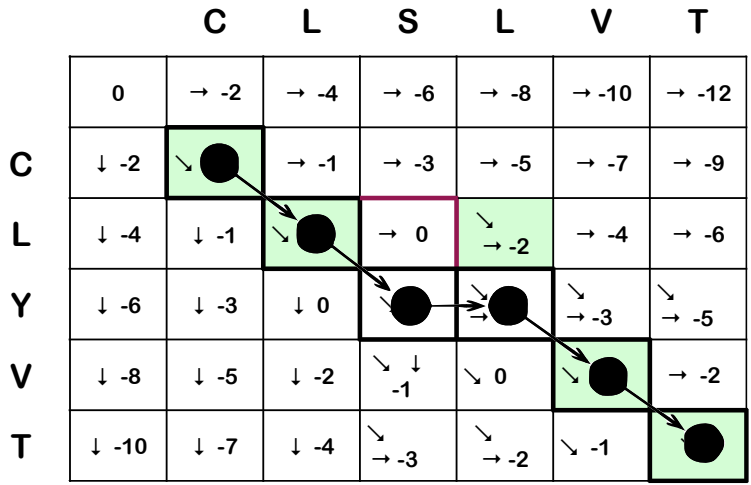




DP: traceback

		C	L	S	L	V	T
	0	→ -2	→ -4	→ -6	→ -8	→ -10	→ -12
C	↓ -2	↘ 1	→ -1	→ -3	→ -5	→ -7	→ -9
L	↓ -4	↓ -1	↘ 2	→ 0	↘ → -2	→ -4	→ -6
Y	↓ -6	↓ -3	↓ 0	↘ 1	↘ → -1	↘ → -3	↘ → -5
V	↓ -8	↓ -5	↓ -2	↘ ↓ -1	↘ 0	↘ 0	→ -2
T	↓ -10	↓ -7	↓ -4	↘ → -3	↘ → -2	↘ -1	↘ 1

DP: traceback



CLSLVT CLSLVT
CLY-VT CL-YVT

Smith-Waterman algorithm

local pairwise alignment

- **cells with negative scores are set to zero**
- **traceback starts at the highest scoring cell**
  stops when 0 is encountered

$$M_{i,j} = max \begin{cases} M_{i-1,j-1} + s(x_i, y_i) & x_i, y_j \text{(mis)match} \\ M_{i-1,j} + \delta & x_i \text{ aligned to gap in } y \\ M_{i,j-1} + \delta & y_j \text{ aligned to gap in } x \\ 0 & \text{if score} \leq 0 \end{cases}$$

Smith-Waterman

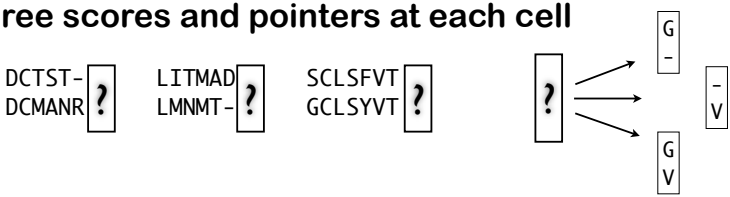
		C	L	S	L	V	T
	0	0	0	0	0	0	0
C	0	↘ 1	0	0	0	0	0
L	0	0	↘ 2	→ 0	1	0	0
Y	0	0	↓ 0	↘ 1	0	0	0
V	0	0	0	0	↘ 0	↘ 1	0
T	0	0	0	0	0	0	↘ 2

Affine gap penalties

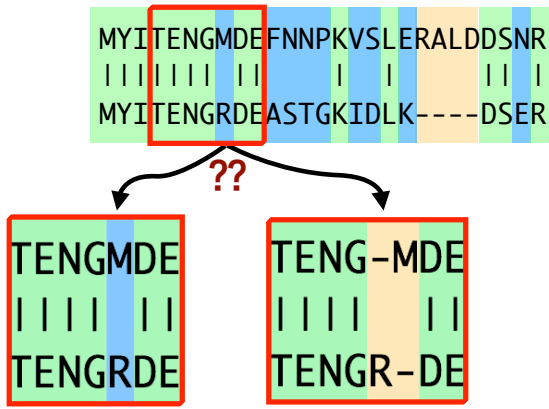
- score depends on the length of the contiguous gap
- gap opening penalty: ***d***
- gap extension penalty: ***e***

$$\gamma(g) = -d - (g-1)e$$

- consequence for dynamic programming implementation: now we have to keep track of three scores and pointers at each cell



Scoring an alignment



→ chosen scoring scheme affects the result!

Review questions / topics

- homology, similarity
- dotplots: generation, interpretation, application
- gap penalties
- substitution matrices: PAM, BLOSUM
  - origin, differences, applications
- dynamic programming
  - principle
  - initialization, scoring, traceback
  - global vs. local
- *optimal* alignments

Next quiz

- today ~2PM to tomorrow 9AM
  - does not cover today's material
  - one page, PDF format
  - please include your name on the page