Needleman-Wunsch global alignment

Mikhail Dozmorov

Spring 2018

Alignment goals

- **Homology** - sequence similarity - helps to infer functions of genes uncharacterized in one organism but known in another
- **Global sequence alighment** (Needleman-Wunch)
- **Gapped local sequence alignment** (Smith-Waterman)

Needleman-Wunsch algorithm

- The problem of finding best possible alignment of two sequences is solved by Saul B. Needleman and Christian D. Wunsch in 1970
- The Needleman-Wunsch algorithm is an example of dynamic programming, a discipline invented by Richard E. Bellman in 1953
- It refers as optimal matching problem or global alignment

Needleman, S. B., and C. D. Wunsch. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." Journal of Molecular Biology 48, no. 3 (March 1970): 443–53. - Online demo: http://experiments.mostafa.io/public/needleman-wunsch/ - Wikipedia: https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

Needleman-Wunsch algorithm

- In its simplest form, assign a value A where the aligned pair consists of the same letters (nucleotides, amino acids)
  - If the letters differ, subtract B - edit penalty.
  - If a gap needs to be made, subtract a gap penalty times the number of gaps

Steps

1. Initialization of the score matrix $D(i, j)$
2. Calculation of scores, such that

$$D(i,j) = max \begin{cases} D(i-1, j-1) + s(i,j) \\ D(i-1, j) + g \\ D(i, j-1) + g \end{cases}$$

Where $s(i, j)$ is the substitution score for entries i and j, and g is the gap penalty

Example of a simple penalty matrix

	A	G	C	T
A	1	-1	-1	-1
G	-1	1	-1	-1
C	-1	-1	1	-1
T	-1	-1	-1	1

In reality, similarity matrices were derived based on evolutionary observations

Example of penalty matrix

Suppose your empirically defined matrix is

```
   A  G  C  T
A 10 -1 -3 -4
G -1  7 -5 -3
C -3 -5  9  0
T -4 -3  0  8
```

with gap penalty= -5

Apply it to the secong sequence:

```
Read: AGACTAGTTAC
Ref:  CGA---GACGT
```

-3+7+10+(3)(-5)+7-4+0-1+0 = 1

PAM250

- The similarity matrix is frequently used to score aligned peptide sequences to determine the similarity of those sequences.
- Derived from comparing aligned sequences of proteins with known homology and determining the "point accepted mutations" (PAM) observed.
- The frequencies of these mutations are in this table as a "log odds-matrix" where: $M_{ij} = 10(log_{10} R_{ij})$, where M_{ij} is the matrix element and R_{ij} is the probability of that substitution as observed in the database, divided by the normalized frequency of occurence for amino acid i.

PAM250

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

http://prowl.rockefeller.edu/aainfo/pam250.htm

BLOSUM- BLOcks Substitution Matrix

- BLOSUM62 - sequences having at least 62% identity are merged together.
- BLOSUM30 - sequences having at least 30% identity are merged together.
- BLOSUM90 - sequences having at least 90% identity are merged together.

Best possible alignment for two sequences

- The N-W algorithm is mathematically proven to find the best alignment of two sequences
- N-W algorithm takes $O(n^2)$ to find best alignment of n letters in two sequences
- Accessing all possible alignments one by one would date $\binom{2n}{n}$, so n^2 is much smaller

Sequence searching and alignment

- **FASTA** - a DNA and protein sequence alignment software package by David J. Lipman and William R. Pearson in 1985.
- **BLAST** (Basic Local Alignment Search Tool) - an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences.
  - Designed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipman
  - Innovation: heuristic database search (speed), followed by optimal alignment (accuracy, statistics)

BLAST

- BLAST is not used for NGS because it is too slow.
- Format for command line version: `blastall -d assemblyfasta -i genefasta -o output.blast -p blastn -e 1e-15`
  - `-i` indicates what is the gene file
  - `-o` indicates what you want the output to be
  - `-p` with ending "n" means nucleotide alignment -e statistical significance of alignments
- Magic-BLAST is an alternative

https://ncbi.github.io/magicblast/

https://ncbiinsights.ncbi.nlm.nih.gov/2016/10/13/introducing-magic-blast/

Significance of the alignment

- For local alignment we want to address how high an alignment score S exceeds a cutoff x.
- If the quality score is within random chance then it probably isn't a good alignment. We use an extreme value (aka Gumbel) distribution with parameter:

$$P(S > x) = 1 - exp(-KMNe^{-\lambda x})$$

- M is the effective length of the query sequence
- N is the effective length of the reference sequence
- K and λ are positive parameters that depend on the score matrix and the composition of the sequences being compared

E-values

- E-values are the number of alignments with scores at least equal to x that would be expected by chance alone. The larger the database the more likely you will have a hit by chance, therefore we must take the size of the database into consideration.
- We can treat E-values as multiple comparison corrected p-values.
  - Low E-value $\sim$ strong match or good hits
  - Commonly used threshold: E-value < 0.05

New Aligners for NGS data: MAQ, BWA, Bowtie, SOAP, Rsubread, etc.

- The main technique for faster alignment: indexing. We make substrings of length k (short integer) and put the substring and its location in a hash table.
- **MAQ** - Mapping and Assemblies with Quality (Li, Ruan, Durbin 2008)
  - Does the alignment but also calls SNPs
  - At the alignment stage it searches for the un-gapped match with the lowest mismatch score, defined as the sum of qualities at mismatching bases
  - Only considers positions that have two or fewer mismatches in the first 28 base pairs (this is the default which can be changed)
- Sequences that fail to reach a mismatch score threshold are searched with Smith-Waterman algorithm that allows for gaps
- Always reports a single alignment, positions aligned equally well to a read are chosen randomly
- Potential problem - multimapped reads will not contribute to variant

BWA, Bowtie

- Bowtie employs Burrows-Wheeler transform (BWT) based on the full-text minute-space (FM) index. Index is built using `bwa index`, `bowtie2-build`

```
acaacg$ → $acaacg →(sort) $acaacg → gc$aaac
          g$acaac          aacg$ac
          cg$acaa          acaacg$
          acg$aca          acg$aca
          aacg$ac          caacg$a
          caacg$a          cg$acaa
          acaacg$          g$acaac
```

- Bowtie has a memory footprint of 1.3 GB for the human genome. Very fast.
- The last first mapping can transform it back to the original sequence.