

Pairwise sequence alignment

Analysis of Biological Sequences 140.638

Conditional probability (Bayes)

$p(A,B) = p(B)p(A|B)$ (Bayes' formula)

$p(A,B) = p(A)p(B|A)$

$p(A)p(B|A) = p(B)p(A|B)$

e.g. A = rolling 2, 4, or 6

B = rolling a 1, 2, 3, or 5

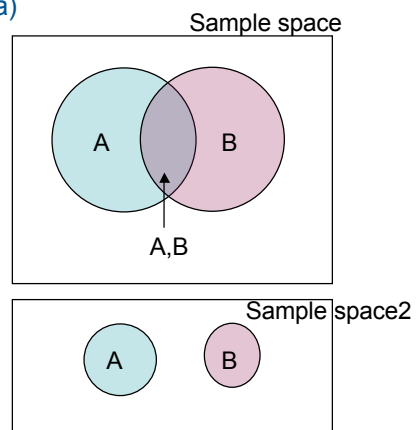
$p(A,B) = 1/6$

$p(A|B) = 1/4$

$p(B|A) = 1/3$

$p(H|D) = p(H)p(D|H)/p(D)$

h=hypothesis, d=data



Goal of sequence alignment

- Examine relationships between sequences
 - Divergence from common ancestor (orthologs or paralogs)
 - Convergent evolution
 - Common motifs
 - Catalytic sites
 - Structurally significant regions
 - Mutational distance
- Sequence similarity != homology, though

Steps in sequence alignment

- Obtain sequences
- Align sequences
- Score alignment
- Is it significant, mathematically? Biologically?

Example alignment

```
Query 2   WTVQPIVLPEKDSWTVN 18
          W V P ++ E + W +N
Sbjct 67  WFVSPHLISENERWRIN 83
```

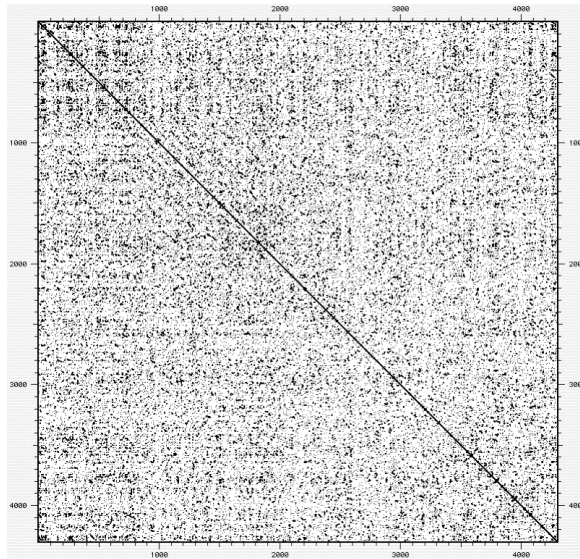
HIV RT vs Aquifex. How do we create this alignment? Is it significant?

Dot matrix analysis

- The simplest, most visual, most intuitive way to create an alignment
- Reference: Gibbs AJ, McIntyre GA. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J. Biochem* 1970 **16**(1):1-11.

Dot matrix alignment

PKD
protein
against
itself

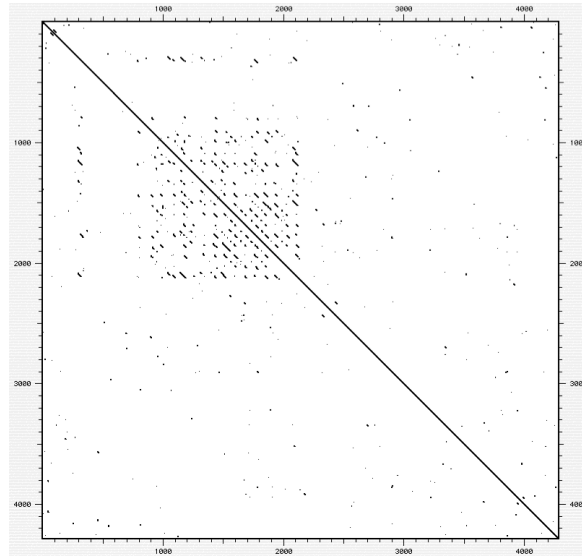


Dot matrix alignment

- Simplicity has its drawbacks:
 - No overall score given
 - Quality of hits judged by eye
 - No good guidelines for penalty/window parameters
 - Few programs provide a way to obtain hits
- Parameters can be modified:
 - For proteins, can use scoring matrices
 - Sliders help interactively determine parameters

Dot matrix alignment

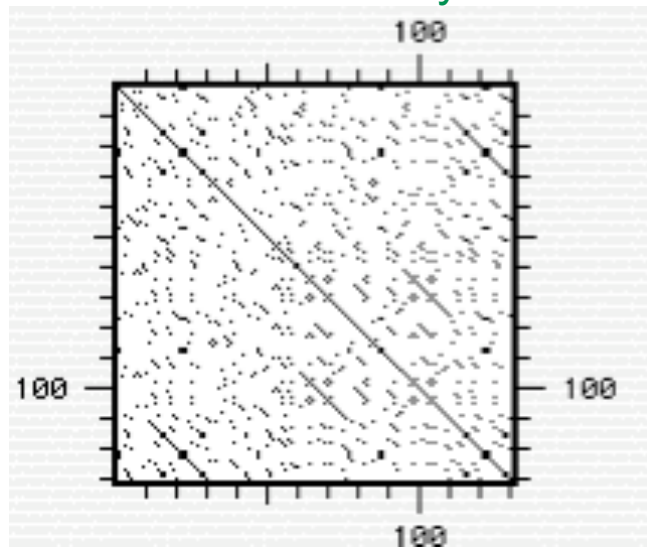
PKD vs
itself with
better
parameters



Dot matrix alignment

- Lots of variations—can align DNA vs DNA, protein vs protein, many scoring schemes
- Sequence repeats and inverse repeats readily apparent
- Can be used to find self-complementary portions of sequences (e.g. RNA) to help predict secondary structure
- Still used today—you will see it even in major papers

Dot matrix: ribozyme



Creating alignments from scratch

Example:

ACCTAGCTAGCCGAT

And

ACCCCTAGGCGAAA

Possible alignment:

ACC--TAGCTAGCCGAT-

ACCCCTAGG----CGAAA

Elements of an alignment

```
ACC--TAGCTAGCCGAT-  
ACCCCTAGG----CGAAA
```

- Matches
- Mismatches
- Gaps/indels

Choosing the best alignment

```
ACC--TAGCTAGCCGAT-  
ACCCCTAG---G-CGAAA
```

```
ACCTAGCTAGCCGAT-  
ACCC--CTAG-CGAAA
```

```
ACC--TAGCTAGCCGAT-  
ACCC----TAGGCGAAA
```

Need Scoring Rules

- For example: score = (#matches) - (#mismatches) - (#gaps)x2

ACC--TAGCTAGCCGAT- Score = 10 - 1 - 7x2 = -5
ACCCCTAG---G-CGAAA

ACCTAGCTAGCCGAT- Score = 10 - 2 - 4x2 = 0
ACCC--CTAG-CGAAA

ACC--TAGCTAGCCGAT- Score = 9 - 1 - 7x2 = -6
ACCC-----TAGGCGAAA

Need Scoring Rules

- For example: score = 3x(#matches) - 4x(#mismatches) - (#gaps)

ACC--TAGCTAGCCGAT- Score = 3x10 - 4x1 - 7 = 19
ACCCCTAG---G-CGAAA

ACCTAGCTAGCCGAT- Score = 3x10 - 4x2 - 4 = 18
ACCC--CTAG-CGAAA

ACC--TAGCTAGCCGAT- Score = 3x9 - 4x1 - 7 = 16
ACCC-----TAGGCGAAA

Dynamic programming

- Goal: find the optimal solution to a complicated problem
- Approach: break the problem down into smaller, tractable problems, and solve those in a way that is optimal for the final solution
- “dynamic” because the program is constantly making decisions and executing bits of code in ways that could not be programmed linearly

Global alignment: Needleman-Wunsch algorithm (Gotoh)

- Dynamic programming: achieve optimal alignment by constructing optimal alignments of smaller subsequences
- Assume that the optimal alignment is known up to a point, and then extend the alignment optimally to create a new optimal alignment
- Recursive algorithm: in programming terms, the function calls itself (example: binary search)

Dynamic programming algorithm: example

- DNA alignment rules: match = 2, mismatch = -1, gap = -2
- Global alignment -> start at the beginning of the sequences and progress to the end
- Score of the alignment = score of the alignment up to the previous character + maximum score of aligning the next two symbols or adding a gap in either sequence.

Dynamic programming algorithm

	-	G	A	T	C
-					
A					
A					
C					

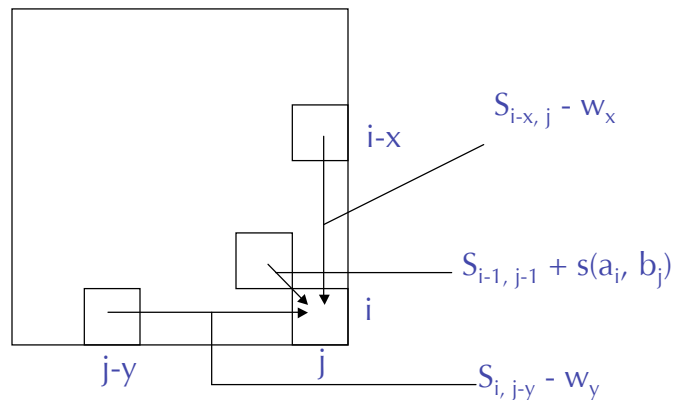
Rules:
m = 2
mm = -1
g = -2

Dynamic programming algorithm (global—Needleman-Wunsch/Gotoh)

$$S_{i,j} = \max \begin{cases} S_{i-1, j-1} + s(a_i, b_j) \\ \max(x \geq 1) (S_{i-x, j} - w_x) \\ \max(y \geq 1) (S_{i, j-y} - w_y) \end{cases}$$

Dynamic programming algorithm (global—Needleman-Wunsch/Gotoh)

$$S_{i,j} = \max \begin{cases} S_{i-1, j-1} + s(a_i, b_j) \\ \max(x \geq 1) (S_{i-x, j} - w_x) \\ \max(y \geq 1) (S_{i, j-y} - w_y) \end{cases}$$



Dynamic programming algorithm (global—Needleman-Wunsch/Gotoh)

Base cases for the recursion:

$$S_{0,0} = 0$$

$$S_{i,0} = S_{i-1,0} + w_x = i * w_x$$

$$S_{0,j} = S_{0,j-1} + w_y = j * w_y$$

Complexity

- Big-O notation
- Space: $O(mn)$
- Time: $O(mn)$
- Filling the matrix: $O(mn)$
- Backtrace: $O(m+n)$

Reference

Needleman and Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins" *J. Mol. Biol.* (1970) **48**:443-453

(available through PubMed)

Local alignment

- A local alignment between two sequences is an alignment with maximum similarity between a substring of sequence a and a substring of sequence b
- Smith and Waterman, "Identification of Common Molecular Subsequences," *J. Mol Biol.* (1981) **147**:195-197 (available through PubMed)

Local alignment

$$S_{i,j} = \max \begin{cases} S_{i-1, j-1} + s(a_i, b_j) \\ \max(x \geq 1) (S_{i-x, j} + w_x) \\ \max(y \geq 1) (S_{i, j-y} + w_y) \\ 0. \end{cases}$$

At every step you have the option of starting the alignment over by setting the score to zero. Otherwise the algorithm is identical to global alignment.

Local alignment

- Rules: match = 2, mismatch = -1, gap = -2

	-	G	A	C	C
-					
A					
A					
C					

Local vs global

- Scoring matrix or match/mismatch scores will determine whether a local alignment is obtained
- Needleman-Wunsch can return a local alignment depending on the weighting of end gaps and other scoring parameters
- Look at alignment: if there are long internal gaps, the alignment is local
- The best way to tell what's going on is to align random or unrelated sequences under the same conditions (next lecture)

Local vs global

	-	A	A	C	T	A
-						
C						
C						
T						
G						

Local vs global

	-	A	A	C	T	A
-						
C						
C						
T						
G						

Useful compilations of alignment programs

http://www.bioinformaticsonline.org/links/ch_03_t_1.html

<http://bioweb.pasteur.fr/seqanal/EMBOSS/>

Sankoff alignment

- No predetermined gap penalties/cell weights
- Uses a deletion/insertion (DI) index
- Effectively decides the number of gaps in advance

Implementation of Sankoff alignment

- Nozaki and Bellgard, Bioinformatics 2005
- Implement, in roughly quadratic space, an algorithm for pairwise alignment that a priori assigns the number of gaps; does not employ gap penalties
- Equivalent to using match=1, mismatch=0 and exploring $n \times m \times (k+1)$ space, where k is the predetermined number of gaps