Sequence Alignment

Bogdan Dumitriu

Department of Computer Science
University of Utrecht

May 2, 2005

Outline

Alignment Significance
Problem formulation

Bram:

What kind of statistical methods can be used to evaluate the significance of an alignment score?

Alignment Significance
Explanation

Obviously, a non-significant alignment:

```
THESEALGRITHMARETR--YINGTFINDTHEBESTWAYTMATCHPTWSEQENCES
TH S++ + ++  +++T   Y    FIND++       YT  +  P +++
THISDESNTMEANTHATTHEYWILLFINDAN-------YTHIN-GPRFND------
```

Alignment Significance

A solution

A common and simple test - a permutation test:

1. Randomly rearrange the order of one or both sequences.
2. Align the permuted sequences.
3. Record the score for this alignment.
4. Repeat steps 1-3 a large number of times.

Then, compare the score with the obtained distribution.

Mutation Independence

Problem formulation

Marjolijn:

Since RNA is a copy of a part of the DNA, why does the independence assumption regarding mutations hold for DNA, but not for RNA?

Mutation Independence

Possible explanation

Messenger RNA (mRNA) undergoes a few processing steps:

- a modified guanine is added at the "front" of the message
- splicing: certain parts of non-coding sequences (introns) are removed
- a sequence of adenine nucleotides are added at the "end" of the message

Also, some errors can occur when copying DNA to RNA.

Relative Entropy
Problem formulation

Jacob:

Why is $\sum_{a,b} q_a q_b \log \frac{q_a q_b}{p_{ab}}$ equal to the relative entropy $H(q^2||p)$ and furthermore, what *is* this entropy and what has it to do with this local alignment algorithm?

Relative Entropy

What is information?

Definition

Information is a decrease in uncertainty.

Alternative definition

Information can be seen as a degree of surprise.

Quantitative approach

Information: $H(p) = \log_2 \frac{1}{p}$ or $H(p) = -\log_2 p$

Relative Entropy

What is information?

Definition

Information is a decrease in uncertainty.

Alternative definition

Information can be seen as a degree of surprise.

Quantitative approach

Information: $H(p) = \log_2 \frac{1}{p}$ or $H(p) = -\log_2 p$

Relative Entropy

What is entropy?

Flat frequency distribution of symbols:

- each symbol has probability $\frac{1}{n}$
- each symbol has holds $\log_2 n$ bits of information

Non-flat frequency distribution of symbols:

- each symbol has probability p_i
- on average, each symbol holds $-\sum_{i}^{n} p_i \log_2 p_i$ bits of information

Definition

Entropy represents the average information per symbol.

Relative Entropy

What is entropy?

Flat frequency distribution of symbols:

- each symbol has probability $\frac{1}{n}$
- each symbol has holds $\log_2 n$ bits of information

Non-flat frequency distribution of symbols:

- each symbol has probability p_i
- on average, each symbol holds $-\sum_{i}^{n} p_i \log_2 p_i$ bits of information

Definition

Entropy represents the average information per symbol.

Relative Entropy

What is relative entropy?

Applying the entropy definition to scoring matrix:

$$
\begin{aligned}
H &= \sum_{i=1}^{20} \sum_{j=1}^{i} q_i q_j \log \frac{q_i q_j}{p_{ij}} \\
&= -\sum_{i=1}^{20} \sum_{j=1}^{i} q_i q_j \log \frac{p_{ij}}{q_i q_j} \\
&= -\sum_{i=1}^{20} \sum_{j=1}^{i} q_i q_j s(i,j) \\
&= -\sum_{a,b} q_a q_b s(a,b)
\end{aligned}
$$

Needleman-Wunsch Algorithm

Problem formulation

Ingmar:

How do you determine which steps backwards to take in the matrix composed using dynamic programming by the Needleman-Wunsch algorithm in order to create the best alignment and why?

Needleman-Wunsch Algorithm

In a nutshell

- $F(i,j)$ - score of the best alignment between $x_1x_2\ldots x_i$ and $y_1y_2\ldots y_j$
- $F(i,j)$ - built recursively based on $F(i-1,j-1)$, $F(i-1,j)$ and $F(i,j-1)$
- $$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_i), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases}$$
- for each $F(i,j)$ we record the choice we make

Needleman-Wunsch Algorithm

How traceback works

- we start at $F(n, m)$
- $F(i, j)$ traces back to one of $(i-1, j-1)$, $(i-1, j)$ or $(i, j-1)$
- depending on the choice, we add x_i / y_i / '_'
- we stop at $F(0, 0)$

Multiple Sequence Alignments

Problem formulation

Adriano:

What are the implications of extending the a pairwise alignment of sequences to multiple sequence alignment?

Multiple Sequence Alignments

Use Needleman-Wunch

Two sequences

Needleman-Wunch complexity for pairs: $O(n^2)$.

Multiple sequences

Needleman-Wunch complexity for multiple: $O(n^m)$, where m is the number of sequences.

Example

With the same 10,000,000,000 operations:

- align 2 sequences of 100,000 nucleotides each.
- align 5 sequences of 100 nucleotides each.

Problem

Needleman-Wunch doesn't scale well.

Multiple Sequence Alignments

Use Needleman-Wunch

Two sequences

Needleman-Wunch complexity for pairs: $O(n^2)$.

Multiple sequences

Needleman-Wunch complexity for multiple: $O(n^m)$, where m is the number of sequences.

Example

With the same 10,000,000,000 operations:

- align 2 sequences of 100,000 nucleotides each.
- align 5 sequences of 100 nucleotides each.

Problem

Needleman-Wunch doesn't scale well.

Multiple Sequence Alignments

Use Needleman-Wunch

Two sequences

Needleman-Wunch complexity for pairs: $O(n^2)$.

Multiple sequences

Needleman-Wunch complexity for multiple: $O(n^m)$, where m is the number of sequences.

Example

With the same 10,000,000,000 operations:

- align 2 sequences of 100,000 nucleotides each.
- align 5 sequences of 100 nucleotides each.

Problem

Needleman-Wunch doesn't scale well.

Multiple Sequence Alignments

Use pairwise algorithm repeatedly

Another approach:

- align sequence 1 with sequence 2 $\Rightarrow$ trial consensus
- align consensus with sequence 3 $\Rightarrow$ new consensus
- carry on until global consensus is achieved

Problem

A different ordering of the sequences yields a different alignment.

Multiple Sequence Alignments

Use pairwise algorithm repeatedly

Another approach:

- align sequence 1 with sequence 2 $\Rightarrow$ trial consensus
- align consensus with sequence 3 $\Rightarrow$ new consensus
- carry on until global consensus is achieved

Problem

A different ordering of the sequences yields a different alignment.

Multiple Sequence Alignments
The CLUSTAL algorithm

Steps of the CLUSTAL algorithm:

1. all pairs are aligned separately, result in distance matrix
2. a guide tree is calculated from the distance matrix
3. the sequences are progressively aligned based on guide tree

Multiple Sequence Alignments

Example of running CLUSTAL (1)

Say we have the set of proteins:

1. Hba_Human: human α-globin
2. Hba_Horse: horse α-globin
3. Hbb_Human: human β-globin
4. Hbb_Horse: horse β-globin
5. Myg_Phyca: sperm whale myoglobin
6. Glb5_Petma: lamprey cyanohaemoglobin
7. Lgb2_Luplu: lupin leghaemoglobin

Multiple Sequence Alignments

Example of running CLUSTAL (2)

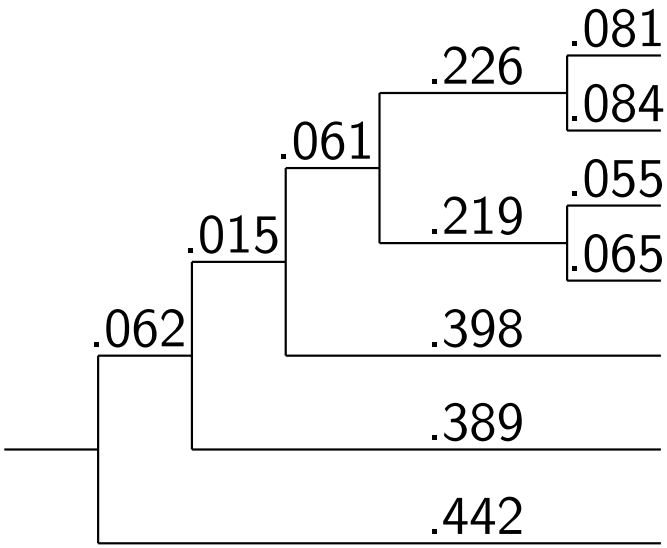
Step 1. The distance matrix / pairwise alignments

Hbb_Human	1	–					
Hbb_Horse	2	.17	–				
Hba_Human	3	.59	.60	–			
Hba_Horse	4	.59	.59	.13	–		
Myg_Phyca	5	.77	.77	.75	.75	–	
Glb5_Petma	6	.81	.82	.73	.74	.80	–
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

Multiple Sequence Alignments

Example of running CLUSTAL (3)

Step 2. The guide tree



Hbb_Human:	0.221
Hbb_Horse:	0.225
Hba_Human:	0.194
Hba_Horse:	0.203
Myg_Phyca:	0.411
Glb5_Petma:	0.398
Lgb2_Luplu:	0.442

Multiple Sequence Alignments
Example of running CLUSTAL (4)

Step 3. Progressive alignment

- order successive alignments based on guide tree
- each step: align a pair of sequences or alignments
- if aligning alignments, use weighed average

Multiple Sequence Alignments

Improved version

CLUSTAL-W: CLUSTAL with some improvements

- gap penalties varies with position/residue
- scoring matrices varied with sequence pairs
- the weight assigned to sequences

More details:

http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=308517&blobtype=pdf

Multiple Sequence Alignments

Improved version

CLUSTAL-W: CLUSTAL with some improvements

- gap penalties varies with position/residue
- scoring matrices varied with sequence pairs
- the weight assigned to sequences

More details:

http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=308517&blobtype=pdf

Two more questions

Marjolijn

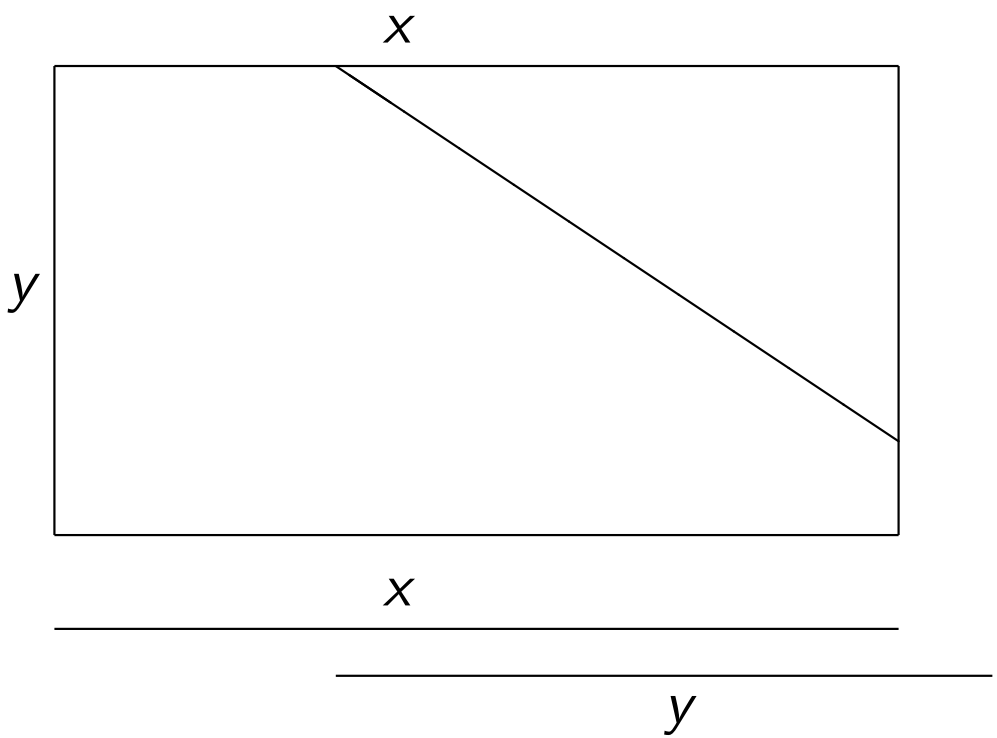
Is it possible that you get the same alignment when using the Needleman-Wunsch algorithm (global) and the Smith-Waterman algorithm (local) to align two different sequences? If it is not, what are the differences in alignment and score?

Lee

Exercise 2.9 gives you a little assignment to calculate the dynamic programming matrix and an optimal alignment for two given DNA sequences. It gives a linear gap penalty of $d = 2$ however. Shouldn't penalties be negative? So I assume this is just a type mistake from the authors?

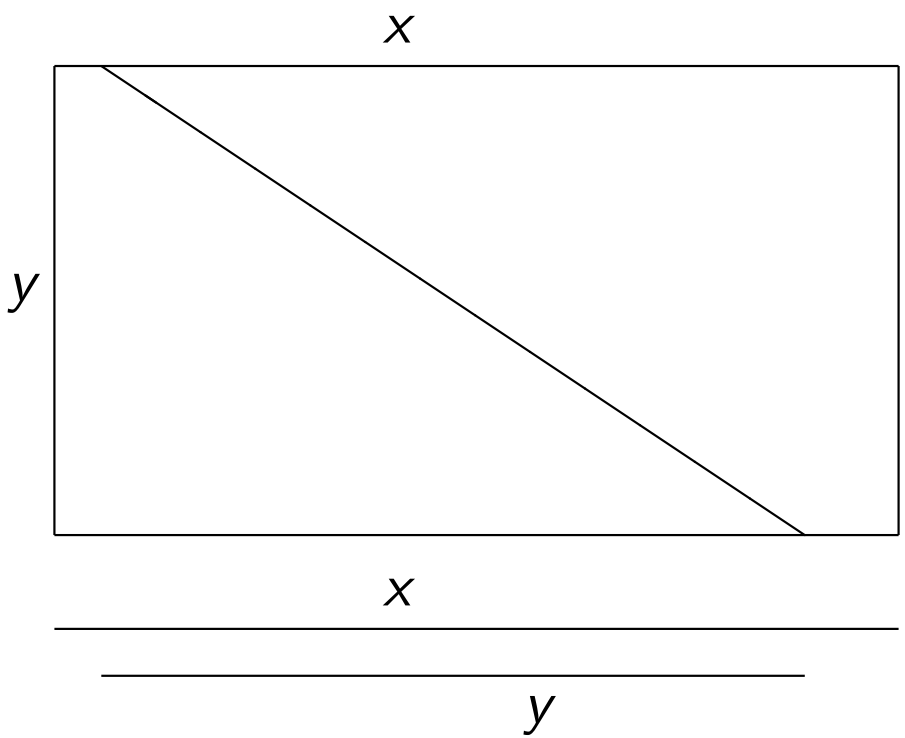
Overlap Matches

... in images (1)



Overlap Matches

... in images (2)



Overlap Matches

Idea

- $F(0,0)$ turns into top or left border
- $F(n,m)$ turns into bottom or right border
- $F(i,0)$ becomes 0, $\forall i$
- $F(0,j)$ becomes 0, $\forall j$
- traceback starts at *max* of bottom/right border