Sequence Analysis Spring 2000 Dr. Richard Friedman
(212)305-6901 (76901) friedman@cuccfa.ccc.columbia.edu 130BB

Lesson 5 - Sequence Comparison and Alignment

Read: GCG Tutorial Chapters 7.1, 7.2 and 9.

Theory

A. Needleman-Wunsch global alignment algorithm (as modified by Gotoh)

Example: Align the sequences GAATTC and GATTA according to the following scoring rules:

match = 2
mismatch = -1
gap = -2
starting score =0

1. Fill the path-graph according to scoring rules:

		G	A	A	T	T	C
	0	-2	-2	-2	-2	-2	-2
G	-2	2	-1	-1	-1	-1	-1
A	-2	-1	2	2	-1	-1	-1
T	-2	-1	-1	-1	2	2	-1
T	-2	-1	-1	-1	2	2	-1
A	-2	-1	2	2	-1	-1	-1

2. Make a path to each cell that maximizes the score for that cell. To do this, start from the upper left hand corner and fill each cell according to the following rule:

$F(i,j)= \max [F(i-1,j-1) + s(x_i, y_j),$
$F(i,j-1) + d),$
$F(i-1,j) + d)]$

$F(i,j)$= the entry in the ith row and jth column of the path-graph.

$s(x_i, y_j)$, = the score for the residues being aligned.

d = gap penalty

$F(i-1,j-1) + s(x_i, y_j)$ represents a diagonal move on the path graph, in which two residues are aligned.

$F(i,j-1) + d$ represents a horizontal move on the path graph, in which a residue in the first sequence is aligned with a gap in the second sequence.

$F(i-1,j) + d$ represents a vertical move on the path graph, in which a residue in the second sequence is aligned with a gap in the first sequence.

		G	A	A	T	T	C
	0	-2	-4	-6	-8	-10	-12
G	-2	2	0	-2	-4	-6	-8
A	-4	0	4	2	0	-2	-4
T	-6	-2	2	3	4	2	0
T	-8	-4	0	1	5	6	4
A	-10	-6	-2	-1	3	4	5

3. The score of the the last cell is the maximum score for the alignment. The path that leads from the first cell to the last cell that gives this score corresponds to the maximum-scoring alignment. To discover this path, trace back a path from the last cell to the first cell, using only transitions that maximize the score.

		G	A	A	T	T	C
	0	-2	-4	-6	-8	-10	-12
G	-2	2	0	-2	-4	-6	-8
A	-4	0	4	2	0	-2	-4
T	-6	-2	2	3	4	2	0
T	-8	-4	0	1	5	6	4
A	-10	-6	-2	-1	3	4	5

Note that there can be more than one path though the path graph, corresoponding to more than one optimal alignment.

4. Trace the path or paths forward to give the optimal alignment:

		G	A	A	T	T	C
	0	-2	-4	-6	-8	-10	-12
G	-2	2	0	-2	-4	-6	-8
A	-4	0	4	2	0	-2	-4
T	-6	-2	2	3	4	2	0
T	-8	-4	0	1	5	6	4
A	-10	-6	-2	-1	3	4	5
		G G	_ A	A A	T T	T T	C A
		G G	A A	_ A	T T	T T	C A

B. Smith-Waterman local alignment algorithm.

1. The first step of the Smith-Waterman algorithm is to generate a path graph identical to the Needleman-Wunsch algorith.

2. In the Smith-Waterman algorithm the minimum allowed final value of a cell is zero- no negative values are allowed. This rule is expressed formally by including 0 as an alternative in the expression for F(i,j):

$$F(i,j) = \max [F(i-1,j-1) + s(x_i, y_j),$$
$$F(i,j-1) + d),$$
$$F(i-1,j) + d),$$
$$0]$$

This rule is equivalent to setting all negative cell entries to zero in the Needleman-Wunsch path graph.

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

3. Find the cell with the highest score. Trace back the path and stop just before a cell with a score of zero is reached. The longest path obtained by this rule represents the optimal alignment.

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

Note that for our example, although there were two optimal global alignment paths, there is only one optimal local alignment path, because one of the local paths is truncated by terminating in a zero. In general, there may be more than one optimal local alignment path and the program will choose one of them arbitrarily. Also, in our case the local alignment terminates on the first residue of both sequences. This need not be the case.

4. Trace the path forward to give the optimal alignment:

		G	A	A	T	T	C
	0	0	0	0	0	0	0
G	0	2	0	0	0	0	0
A	0	0	4	2	0	0	0
T	0	0	2	3	4	2	0
T	0	0	0	1	5	6	4
A	0	0	0	0	3	4	5
		G G	A A	A _	T T	T T	

Summary of commands:

Note: In this document different fonts have different meanings:

Times is used to explain commands.

`Courier is used to indicate commands and command options.`

`Courier italics are used to indicate command parameters, for example, filenames.`

`Courier bold is used to indicate commands that are not displayed.`

`Courier bold italics are used to indicate computer-generated output.`

Helvetica is used to indicate menu items.

`compare` — Prepares a file that summarizes the alignment of two sequences in all registers that can be plotted with dotplot (a *.pnt file).

`tk`	Prepares GCG to send graphics output to the terminal.
`lz`	Prepares GCG to print graphics output on cuccfa's printer.
`dotplot`	Plots a dot-matrix figure based on a *.pnt file generated by compare.
`gap`	Aligns two sequences using the Needleman-Wunsch global alignment algorithm.
`bestfit`	Aligns two protein sequences or two DNA sequences using the Smith-Waterman local alignment algorithm and GCG default parameters.
`bestfit -matr=blosum50.cmp -gap=10 -len=2`	Aligns two protein sequences using the Smith-Waterman local alignment algorithm and parameters that William Pearson found to most likely detect sequence homology of proteins.
`bestfit -matr=blosum50.cmp -gap=10 -len=2 -in1=sequence1 -in2=sequence2 -out=outputfile -D`	Aligns two protein sequences using the Smith-Waterman local alignment algorithm and parameters that William Pearson found to most likely detect sequence homology of proteins, without your having to answer any menu questions.
`bestfit -matr=fastadna.cmp -gap=12 -len=4`	Aligns two sequences using the parameters that William Pearson found to work well (but not necessarily best) in detecting sequence homology of nucleic acids.
`tofasta`	Reads sequence file in GCG format and writes it in Fasta format.
`prss3`	Calculates the statistical significance of a Smith-Waterman alignment. Both files must be in Fasta format. I recommend

	1000 randomizations.
`translate`	Translates a nucleic acid sequence to a protein sequence.
`framealign`	Aligns a protein and a nucleic acid sequence taking translation and frameshifts into account.

PAMX. As X increases, evolutionary distance increases.
BLOSUMY. As Y decreases, evolutionary distance increases.

Probability of occuring by chance in a database search =
PRSS3 probability x number of sequences in database.

Lab :

1. Global Comparison of two Sequences

   A. Make a directory entitled lab5, and go to it.
   B. Copy the file for the beta chain of mouse hemoglobin and for mouse myoglobin into your directory by typing

      ```
      cp /usr2/seq/seqclass/lab5/hbb_myove.pep .
      cp /usr2/seq/seqclass/lab5/myg_mouse.pep .
      ```

   C. Do a Needleman-Wunsch Global alignment of the two sequences by typing:

      ```
      gap
      ```

      and answering the questions that the program asks you. A good output file name would be hbb-myg.gap.

   D. Display the output by typing:

      ```
      cat hbb-myg.gap
      ```

      Are you happy with the alignment?

2. Creating a dotplot representation of the alignment of two sequences:

   A. Create a *.pnt file, which can serve as the input to the dotplot program, by typing:

      ```
      compare
      ```

   B. Initialize the graphics output for Tektronix display by typing

      ```
      tk
      ```

   C. Create a dotplot file of the alignment by executing the program, dotplot.

```
dotplot
```

Does the dotplot show that the sequences are aligned?

D. Set the computer to print output on the computer facility's (excuse me, the Computer and Informatics Resource Center's) laser printer by typing

```
lz
```

E. Rerun the dotplot program. You can pick the output up in Room 130BB after class.

3. Perform a global alignment of the human SRC and CRK proto-oncogene proteins by performing steps analogous to those in part 1 (Their filenames are src_human.pep and crk_human.pep respectively and they are in /usr2/seq/seqclass/lab5). Are you happy with the alignment?

4. Local alignment of two sequences:

Perform a Smith-Waterman local alignment of the sequences in part 3 by typing the following command:

```
bestfit -matr=blosum50.cmp -gap=10 -len=2
```

(where the added options alter the similarity matrix and the gap penalties from the program's default parameters).

Are you happier with this alignment?
Let's quantify "happy".

5. Testing for the statistical significance of the score of the alignment of two proteins:

This is done by shuffling one of the sequences using the PRSS3 program. This program is run by typing

```
prss3
```

Input files for PRSS3 must be in Fasta format.
I recommend 1000 shuffles.

Is the alignment of human SRC and CRK proteins statistically significant?

6. Demonstration of the importance of using a good substitution matrix and good gap penalties in an alignment.

A. Obtain the nucleic acid files spe-ecofrag.seq and aocecofrag.seq (I don't know what they are either, but I thank Dr. Rolf Freter for them) and align them using the bestfit with the nucleic acid default parameters (i.e. just type

```
bestfit
```

and don't change the matrix and gap parameters).

B. Now rerun bestfit with an optimized matrix and set of gap parameters as follows:

```
bestfit -gap=12 -len=4 -matr=fastadna.cmp
```

make sure your output file has a different name than the one in the previous run.

Which alignment looks more realistic to you?
I suggest that you use this second set of parameters in doing an alignment.

7. Translation of a DNA sequence.

A. Obtain the nucleic acid sequence test.seq. Translate it in the first reading frame using the program.

```
translate
```

B. Align the peptide translation of test.seq with the peptide sequence of the human androgen receptor (andr_human.pep) using bestfit with optimized parameters.

Does test.seq seem related to the human androgen receptor?

8. Protein-nucleic acid alignment.

A. Align the human androgen receptor to the test.seq using the program, framealign (Hint: What do you type?).

Does test.seq seem related to the human androgen receptor now?

Why did the alignments with translate and bestfit on the one hand, and framealign, on the other hand, give different results?

B. Devise a way of translating test.seq so that its translation will give a good alignment with the human androgen receptor using bestfit.

Written problems (optiona0:

1. Convince yourself that the matrix representation represents all possible sequence alignments between two sequences.

2. Align the sequences, TCCGGT and TCGGC by hand (i.e. without the computer) using the Needleman-Wunsch-Gotoh and Smith-Waterman algorithms, and the same scoring system used in these notes. This is to be handed in at the beginning of the next class.

Work through the Seqlab tutorial lessons 2&3. Repeat the present exercises using Seqlab.

Web sites:

Links to the following sites:
http://www.ccc.columbia.edu/~friedman/lesson_5.html

On-line GCG Manual
http://www.ccc.columbia.edu/genhelp/

PRSS (an earlier version of PRSS3) is available on the Web at
http://www.med.virginia.edu/~wrp/cshl97/prss.htm

Pairwise compairson programs at the Baylor site
http://dot.imgen.bcm.tmc.edu:9331/seq-search/alignment.html

Bibliography:

1. Biological Sequence Analysis, R. Durbin, S. Eddy, A. Krough, and G. Mitchison, Cambridge, 1998., p. 12 -24.

2. Needleman, S.B. Wunsch, Christian D. A. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. J. Mol. Biol. 48: 443-453, 1970.

3. Smith, T. F., and Waterman, M. S. “Identification of common molecular subsequences”. J. Mol. Biol. 146: 195-197, 1981.

4. Gotoh, O. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705-708.

5. Pearson, W. R. "Effective protein sequence comparison." Methods in Enzymology 266, Computer Methods for Macromolecular Sequence analysis. Doolittle, R. F. ed., pp. 227-258, 1996 Academic Press. San Diego.

6. Maizel, J.V. and Lenk, R.P, “Enhanced graphic matrix analysis of nucleic acid and protein sequences”. PNAS, 78, 7665-7669, 1981.

Relevant entries in the GCG program manual and the Fasta documentation: /usr2/seq/doc/ fasta3.doc