

# Binary Alignments

October 30, 2002

Sequence Analysis & Consulting Service

Seminar Series

Susan Johns

(johns@cgl.ucsf.edu)

First, copy over to your account the necessary data. Enter the following commands. [The first one does the actual copying. The second one moves you into this area.]

```
binary_setup<rtm>
cd binary_align<rtm>
```

As a result of these commands, you are now in a sub-directory called `binary_align` where the necessary files are located. Activate the `gcg` software and set your graphics device.

```
gcg<rtm>
```

Run SETPLOT and select the `TEK4105` option.

How are two sequences compared to one another?

- **alignment approach** zeros in the alike areas
- **graphical approach (dot matrix)** [provides an overall feel for the relationship of one sequence to another learning about all the areas that are alike with possibly of gaining structural or functional insights]

Both of these approaches uses a method of relating the similarity of bases or amino acid residues to one another called a comparison table or scoring matrix.

A comparison table is a probability table for the matching of one sequence character with another. Such tables are also known as scoring matrices. The table assigns a value for the match quality of every possible pair of symbols. When comparing nucleotides, the matrix might contain 1's for matches and 0's (zeros) for mismatches. However, when comparing amino acids, a number could be assigned that is based on chemical similarity or evolutionary distance. This number might be negative if the two residues were very dissimilar.

Consider the following two sequences.

FIGSALKVLGVLPSIVSWVKQA

GIGAILKVLSTGLPALISWIKRKRQE

Just how similar are they?

hint - Similar amino acids in the same position of the sequence.

How would you figure it out?

Now consider these two sequences.

LKCNKLIPIAYKTCPEGKNLCYKMMLASKK  
MVPVKRGCINVCPKNSALVKYVCCSTDRCN

LKCHKAQFPNIETQCKWQTLCFQRDVKPHP  
SSMIVLRGCTSSCGKGAMCCATDLCNGPST  
PST

How would you figure it out?

Most of the time when alignments are needed, dynamic programming algorithms are used to formulate the alignment.

background information:

- creating a matrix
- applying penalties
- revised matrix
- results

Creating a matrix

An example of using dynamic programming.

Using DNA with a unitary match matrix with 1 point for a match and 0 for a mismatch. Create a matrix by placing the sequences along the x and y axis of a plot. Fill in the matrix.

	c	T	A	T	A	t	A	a	g	g
c	1	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	1	1
T	0	1	0	1	0	1	0	0	0	0
A	0	0	1	0	1	0	1	1	0	0
t	0	1	0	1	0	1	0	0	0	0
A	0	0	1	0	1	0	1	1	0	0
a	0	0	1	0	1	0	1	1	0	0
T	0	1	0	1	0	1	0	0	0	0

Applying penalties

A simple gap penalty of subtracting 1 point for every gap inserted, except at the beginning and the end of the sequence will be used.

Points are now added or subtracted based on the best path through the matrix, working diagonally, left to right and top to bottom.

Revised matrix

This results in the following matrix.

	c	T	A	T	A	t	A	a	g	g
c	1	0	0	0	0	0	0	0	0	0
g	0	1	0	0	0	0	0	0	1	1
T	0	1	1	1	0	1	0	0	0	1
A	0	0	2	1	2	0	2	1	0	0
t	0	1	0	3	1	3	1	2	1	0
A	0	0	2	1	4	2	4	3	2	1
a	0	0	1	2	3	4	4	5	3	2
T	0	1	0	2	2	4	4	4	5	4

## Results

Convert the scoring matrix into a trace back path graph by picking the bottom-most, furthest right and highest scoring coordinates and drawing arrows between the boxes to connect them all the way back to the beginning, always choosing the highest scoring trace back route.

There will probably be more than one best path through the matrix. Starting from the top and working down results in two optimum alignments.

```

cTATAtAagg      cTATAtAagg
|  |||||       |  |||||
cg.TAtAaT.     cgT.AtAaT.
  
```

There are two different GCG dynamic programming alignment solutions.

One takes into account the entire sequences to be aligned (global) called **gap**.

[Needleman and Wunsch (1970)]

The other looks for the best local regions alignments between the two sequences being compared (local) called **bestfit**.

[Smith and Waterman (1981)]

```

gap results:
Quality:      302          Length:    111
cyc_euggr x cyc_orysa  January 27, 1999 10:04 ..

 1 .....GDAERGKLFESRAAQCHSAQKGV.NSTGPSLWGVYGRVSGS 41
 1 ASFSEAPPGNPKAGEKIFKTKCAQCHTVDRGAGHKQGNLNLGFRQSGT 50
42 VPGYAYSNAKNAAIIVWEEETLHKFLENPKKYVPGTKMAFAGIKAKKDRQ 91
51 TPGYSYSTANKNMAVIWEENTLYDYLLNPKKYIPGTKMVFGLKKPQERA 100
92 DIIAYMKTLDK 102
101 DLISYLKEATS 111
  
```

```

bestfit results:
Quality:      305          Length:     99
cyc_euggr x cyc_orysa  January 27, 1999 10:14 ..

 1 GDAERGKLFESRAAQCHSAQKGV.NSTGPSLWGVYGRVSGSVPGYAYSN 49
 9 GNPKAGEKIFKTKCAQCHTVDRGAGHKQGNLNLGFRQSGTTPGYSYST 58
50 ANKNAIIVWEEETLHKFLENPKKYVPGTKMAFAGIKAKKDRQDIIAYMK 98
59 ANKNMAVIWEENTLYDYLLNPKKYIPGTKMVFGLKKPQERADLISYLK 107
  
```

There are a number of things that can be changed in order to try to improve an alignment.

- program being used  
gap or bestfit
- penalties values
- comparison table being used  
change actual matching scores in a given table  
change the comparison table being used  
(blosum30 instead of blosum62 for more divergent alignments)  
change the nature of the comparison  
evolution vs structure

While using dot matrix procedures allows the determining of all the areas between two sequences that are alike, using alignment procedures zeros in on the actual areas.

The alignment process uses dynamic programming algorithms to formulate an alignment.

Just because dynamic programming is guaranteed to find an optimal alignment, does not mean

it is the **only** optimal alignment

or

that it is the **right** one

or

even a **biologically relevant** one.

Alignments are bias by the end of the sequence they are started from. Reversing sequences to be aligned and repeating the process might provide additional insight on the nature of the alignment.

To evaluate if an alignment has any meaning or not, use a Z score determination.

Repeat the run that produced your alignment with a -random=100 command switch and then do a Z score calculation to determine its significance.

$$Z \text{ score} = \frac{[(\text{actual score}) - (\text{means of randomized scores})]}{(\text{standard deviation of randomized scores})}$$

### Rough Z score guidelines

Z score	Inference
≤ 3	little, if any evidence for homology
= 6	probably homologous, but may be due to convergent evolution
≥ 9	definitely homologous

AP of: euk\_tata.seq check: 3919 from: 1 to: 10

Eukaryotic promoter TAT box allowed by:  
Philipp Bucher (1990) J. Mol. Biol. 212:563-578  
preferred region: center between -36 and -20

to: prok\_tata.seq check: 2658 from: 1 to: 8

consensus from the standard W. coli  
RNA polymerase promoter "Pribnow" box  
from: 1 to 8

```

////////////////////////////////////
Gap Weight:      0          Average Match: 10.000
Length Weight:   1          Average Mismatch: 0.000

Quality:         59          Length:      10
Ratio:           7.375      Gaps:       2
Percent Similarity: 75.000  Percent Identity: 75.000

Average quality based on 100 randomizations: 51.5 +/- 8.0
  
```

```

Match display thresholds for the alignment(s):
| = IDENTITY
: = 5
. = 1
  
```

euk\_tata.seq x prok\_tata.seq February 2, 1999 13:24 ..

```

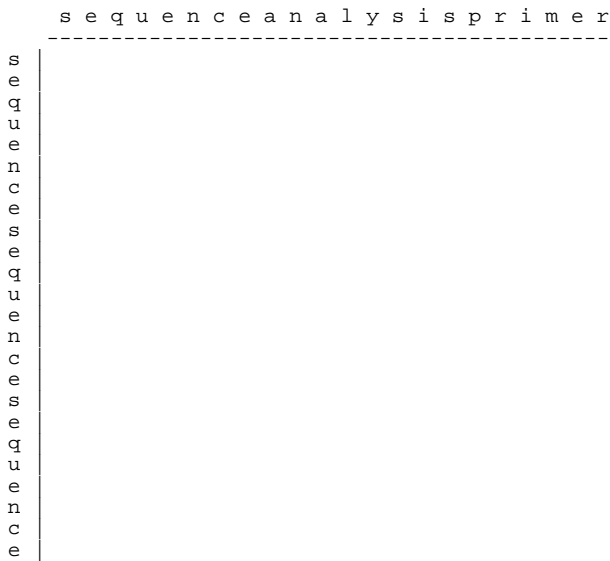
1 cTATAtAagg 10
|  |||||
1 cg.TAtAaT. 8
  
```

$$Z \text{ score} = \frac{59 - (51.5)}{(8.0)} = .9375$$

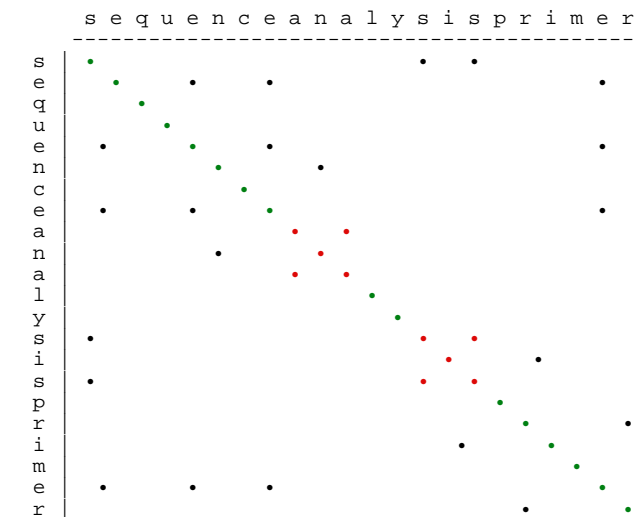
Pairwise alignments are the basis for sequence database searching. Changes are made in the basic approaches to make the procedure faster because of the large number of sequences being looked at.

- wordsearch - compare/dotplot approach [slow]
- ssearch - most sensitive [slow]
- fasta - Pearson and Lipman search
- blast - Altschul et al search

The graphical approach uses the placement of sequences along the x and y axes of a graph to determine relationships.

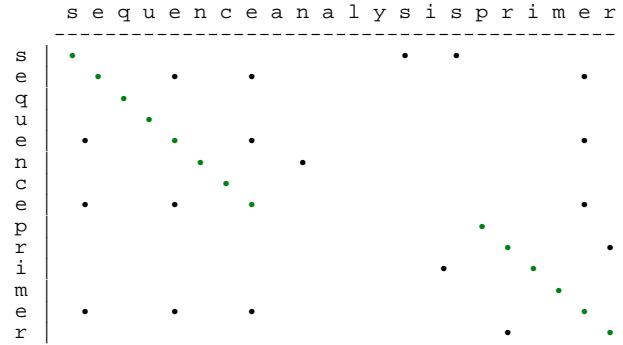


Sequences are written out along the x and y axis. Then a mark (in this case a dot) is placed wherever the two sequences match one another.



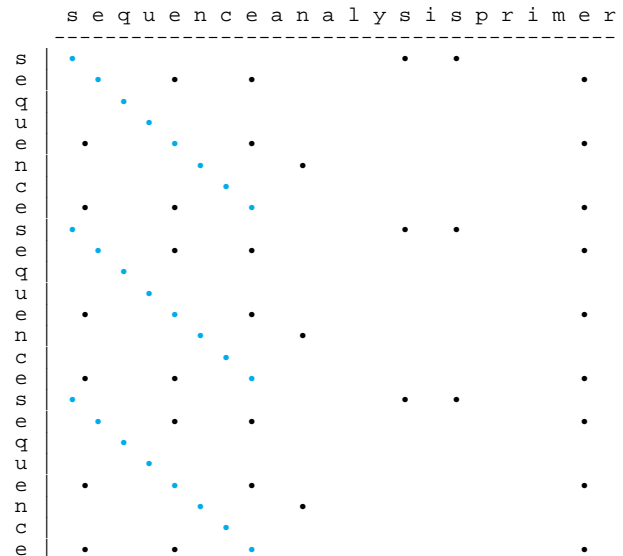
- most obvious features:
- the identity diagonal •
  - two short palindromes (ana and sis) •

Shows the effect of an 'insertion' or a 'deletion'.  
[Since it is impossible to tell if the evolutionary event that caused this was an insertion or a deletion the phenomena is called an 'indel'.]



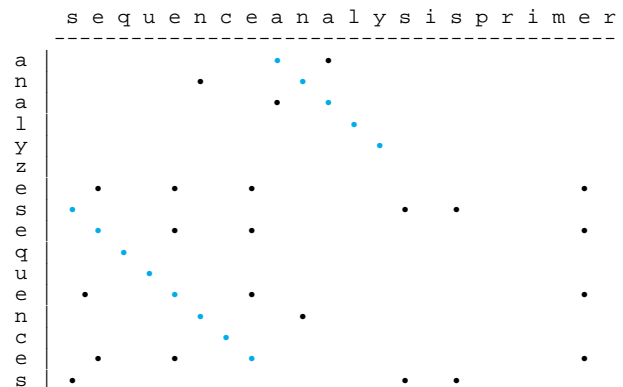
most obvious feature :  
shift in the identity diagonal showing the existence of an indel

Shows the effect of a 'duplication' or a direct repeat.



most obvious feature:  
column of diagonals showing a repeated sequence element

Shows the effect of a 'transposition' or a rearrangement of sequence elements.



most obvious features:  
movement of identity diagonals off the central diagonal  
lack of a diagonal for primer shows its deletion  
ana palindrome is still there

Creating a dotplot is a two step process.

- run **compare** to generate point file sequences to be used  
their starting and ending points  
window size and stringency level  
[comparison table to be used]
- run **dotplot** to display point file point file to be used  
-all to display all the information in a self comparison  
-dot to display dots instead of connecting lines

While using dot matrix procedures allows the determining of all the areas between two sequences that are alike, using alignment procedures zeros in on the actual similar areas.

## Bookmarks for pairwise alignments

### binary alignments

#### align description

#### Sequence Alignment

<http://helix.biology.mcmaster.ca/721/outline2/node37.html>

### align servers

#### Alignment App

<http://genome.cs.mtu.edu/align/align.html>

#### Genestream Resource Center

<http://www2.igh.cnrs.fr/bin/align-guess.cgi>

#### Expasy's SIM Alignment Tool - Protein Sequences )

Canda - <http://ca.expasy.org/tools/sim-prot.html>

China - <http://cn.expasy.org/tools/sim-prot.html>

Korea- <http://kr.expasy.org/tools/sim-prot.html>

Switzerland - <http://www.expasy.org/tools/sim-prot.html>

Taiwan - <http://tw.expasy.org/tools/sim-prot.html>

USA - <http://us.expasy.org/tools/sim-prot.html>

#### Two-sequence alignment server (lalign and prss)

[http://www.isrec.isb-sib.ch/experiment/ALIGN\\_form.html](http://www.isrec.isb-sib.ch/experiment/ALIGN_form.html)

#### BLAST 2 SEQUENCE

<http://www.ncbi.nlm.nih.gov/gorf/bl2.html>

## SACS URLs:

### Bestfit [basic] page

<https://www.sacs.ucsf.edu/secure/cgi-bin/bialign.pl?prog=bestfit>

### Bestfit [advanced] page

<https://www.sacs.ucsf.edu/secure/cgi-bin/bialignadv.pl?prog=bestfit>

### Gap [basic] page

<https://www.sacs.ucsf.edu/secure/cgi-bin/bialign.pl?prog=gap>

### Gap [advanced] page

<https://www.sacs.ucsf.edu/secure/cgi-bin/bialignadv.pl?prog=gap>

### Compare - Dotplot [basic]page

<https://www.sacs.ucsf.edu/secure/cgi-bin/compare.pl>

### Compare - Dotplot [advanced]page

<https://www.sacs.ucsf.edu/secure/cgi-bin/compare.pl>