

Multiple Sequence Alignment

May 15, 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.]

```
multiple_setup<rtm>  
cd multiple<rtm>
```

As a result of these commands, you are now in a sub-directory called multiple where the necessary files are located. Activate the gcg software.

```
gcg<rtm>
```

What is Multiple Sequence Alignment?

MSA is an alignment of 2 or more nucleotide or protein sequences.

Why do MSA?

To determine and investigate relationships between a group of related sequences.

Learn about structural, functional, and motif information present in an evolutionary conserved set of sequences.

Design degenerate primers for a cloning experiment.

Create an input file/format for other Sequence Analysis techniques, e.g. Phylogenic and Profile Analysis.

What are some of the programs available for MSA?

Pileup (GCG), ClustalW, MultAlin, SAM

How do MSA programs differ from one another?

MSA programs use a variety of approaches to solve the problem of aligning multiple sequences. Pileup uses a progressive pairwise approach. ClustalW uses a similar technique, but, puts off the alignment of more dissimilar sequences until later in the process. MultAlin uses a more hierarchical approach to its pairwise clustering. SAM in its alignment mode uses a hidden Markov model approach to the problem.

Every multiple alignment program uses some sort of means to determine the relationship of the sequence characters to one another. In the classical approaches, this is a comparison table or scoring matrix. The default matrix is different each one of them. Most allow the user a selection of comparison tables to choose from to do their alignments. In the HMM approach, various transition probabilities tables are used.

Pileup Overview

Pileup creates a multiple sequence alignment using a progressive pairwise method.

A series of pairwise alignments are carried out between each of the sequences to be aligned.

Clusters of aligned sequences are then formed. Two clusters are aligned by an extension of the pairwise alignment.

Progressive, pairwise alignments are made with increasingly dissimilar sequences and clusters, until all sequences have been included in the final pairwise alignment.

Pileup clusters the sequences by similarity to produce a dendrogram, or a tree-like representation of the sequence relationships.

Directs order of subsequent alignments.

Pileup Details

Because pileup is a GCG program, its input data can be in various forms.

- a wild card set of sequences in the current location
- a list file

list file information

A input file containing a list [UNIX path] of the sequence file names.

sequence list

```
..  
/home/socr/b/johns/seminars/multiple/working/anp4_pseam.swissprot  
/home/socr/b/johns/seminars/multiple/working/anp_limfe.swissprot  
/home/socr/b/johns/seminars/multiple/working/anpa_pseam.swissprot  
/home/socr/b/johns/seminars/multiple/working/anpx_pseam.swissprot  
/home/socr/b/johns/seminars/multiple/working/anpy_pseam.swissprot
```

or

An input file containing a list of database specifiers and accession numbers.

sequence list

```
..  
sp:anp4_pseam  
sp:anp_limfe  
sp:anpa_pseam  
sp:anpx_pseam  
sp:anpy_pseam
```

sequence list

```
..  
sp:anp4_pseam begin:1 end:45  
sp:anp_limfe begin:1 end:60  
sp:anpa_pseam begin:1 end:60  
sp:anpx_pseam begin:1 end:60  
sp:anpy_pseam begin:1 end:60
```

or

A combination of 1 and 2.

ClustalW Overview

In ClustalW similarity scores are calculated by marking fragments of a given length for each sequence and finding all the best matches of this length between two sequences.

Next a dotplot type of analysis takes place, the best diagonals in the dot plot are saved.

The final step is to arrange the diagonals from the dot plot in such a way as to give the highest score.

There are 2 modes of operation, FAST and SLOW.

- The SLOW/Accurate mode is very similar to Pileup.
- The FAST mode uses a different algorithm.

ClustalW Details

ClustalW can accept a number of different input formats [NBRF/PIR, Embl/SwissProt, FASTA, GDE, ClustalW, GCG/MSF, GCG/RSF]. However, all the sequences to be used must be contained in a single input file

MultAlin Overview

MultAlin creates a multiple sequence alignment using a progressive pairwise method with a hierarchical clustering mode designed by the author.

This program provides some interesting comparison tables to use in the alignment process. The output on the web is a colored gif image of the generated alignment. The coloring scheme can be set by the user.

MultAlin Details

The MultAlin documentation states that the input data can be in various formats, however, the program works best in its own and fasta format.

The command version of the program is not easy to use and this program is best run on the web.

While the output format is called msf, it is not a standard GCG msf format.

SAM Overview

SAM creates a multiple sequence alignment using a HMM approach. This is just the first step in a much more complex process of creating a HMM model for biological sequence analysis.

This program is very complex and only the very basics of its alignment procedure will be touched on here.

SAM Details

The SAM documentation states that the input data can be in various formats. It can take any of the formats supported by the program readseq. To keep things simple only the fasta format will be used here.

There is a web page for using SAM at the Pasteur Institute in Paris. It appears to return your results via email since your email address is

requested. However, if you wait and your job is small, the process is interactive in three steps.

Once an alignment has been generated, it needs to be examined to see if it makes sense from a scientific point of view.

Do known functional motifs line up?

Does the alignment at the end of the sequence look as strong as at the beginning?

Does changing end and gap penalties greatly change the produced alignment?

Does the alignment change if a different comparison table or scoring matrix is used?

Could parts of the alignment be improved by hand editing the results?

When a GCG msf alignment is edited by hand the revised alignment needs to be run through the reformat program to correct the parameters that GCG is looking for to insure that it is a proper msf alignment file.

MSA post processing

Raw multiple alignment files contain the desired information, it is just not in a very attractive format. Classically, a highlighter was used to match those columns in an alignment that were conserved. Then, software was developed to do this task. The easiest of these alignment display programs to use is called **prettybox**.

Prettybox is now part of the GCG package and is designed to work with msf formatted files. It will work with any alignment that can be converted over into this format.

At times even a prettybox output image is not dramatic enough for a given circumstance. When color is needed try the colorization tool **color-pb**. This allows the changing of the black, dark gray and pale gray boxes into colored boxes.

Color is not the only means that can be used to highlight important features. At times it is the differences that are important. Such a case is attempting to find the few differences in a highly conserved set of sequences. Here the **pretty** program can be used with command switches to make these regions highly visible.

```
pretty -mat=identpep.cmp -dif -cons t2p.msf{*} -def
```

Things to Remember

This method (pairwise progressive) does not create the best possible alignment. Gaps introduced early influence the rest of the process.

Look carefully at your data. You may be able to improve MSAs by hand.

Use what you know about the sequences (functional, structural, domain info) in a critical examination of the output.

Because of the nature of the algorithm, the order of the sequence input matters.

A MSA program will align ANY sequence you give it. These programs do not "tell" you if your sequence is related by doing the alignment.

URLs used in this seminar:

SACS pileup pages:

<https://www.sacs.ucsf.edu/secure/cgi-bin/pileup.pl> - limit of 10 sequences
<https://www.sacs.ucsf.edu/secure/cgi-bin/pileup2.pl> - uses list file
<https://www.sacs.ucsf.edu/secure/cgi-bin/pileup3.pl> - uses fasta collection file

SACS ClustalW page:

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

SACS MultAlin page:

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

Pasteur SAM page

<http://bioweb.pasteur.fr/seqanal/motif/sam-uk.html>

SACS visualization seminar:

<http://www.sacs.ucsf.edu/Training/viewer/7-28-99-viewer-intro.html>

ClustalW distribution page:

<http://bigfoot.eecs.umich.edu/pub/NetBSD/packages/pkgsrc/biology/clustalw/README.html>

Public multiple alignment sites:

clustalw:

<http://www.ebi.ac.uk/clustalw/> (EBI)
<http://www.ch.embnet.org/software/ClustalW.html> (EMBnet)
<http://stl.wustl.edu/msa/> (Washington Univ. St. Louis - one of the options)

multalin:

<http://www.toulouse.inra.fr/multalin.html>