Tutorials and Comparative Analysis

On,

Multiple Sequence Alignment Tools

Submitted to,

Dr. Deendayal Dinakarpandian

In
Introduction to Bioinformatics (CS 566)

By,

Anusha, Krishna, Mridula
**Tool 1: Multalin – Multiple Sequence Alignment**

**About Multalin:**

Multalin creates a multiple sequence alignment from a group of related sequences using progressive pairwise alignments. The method used is described in "Multiple sequence alignment with hierarchical clustering", F. Corpet, 1988, Nucl. Acids Res. 16 10881-10890. This uses dynamic programming. MultiAlin is based on hierarchical clustering algorithm.

**CLUSTERING ALGORITHM:**

1) MultAlin begins by computing similarity scores for every possible pair of sequences using a fast algorithm.

2) These scores are used to create a hierarchical clustering represented by a dendrogram.

3) A consensus sequence and pairwise scores are computed.

4) To achieve step 3, the scores of all the pairwise alignments included in the multiple one are computed and they can be used to do step 1 again;

If the clustering order is different, a new multiple alignment can be done following this new clustering. i.e., steps 2 and 3. This process can be iterated until the clustering order remains unchanged by iteration.

**User Manual for Multalin:**

**Step1:** Navigate to the URL - [http://prodes.toulouse.inra.fr/multalin/multalin.html](http://prodes.toulouse.inra.fr/multalin/multalin.html)

**Step2:** On the MultAlin home page you will see a large rectangle. Paste (as in cut and paste) your sequences in the rectangle. Instead of pasting your sequences, you can give the name of your sequences file, or select it with the Browse button.

**Step3:** The next step is to set the parameters. Use the pop up menus or type in text or numbers where required. When you are ready click on the "submit data" button (you can use either the buttons at top or at bottom of the page).

**Step4:** The result will be sent back to your internet browser in the form of a GIF image (default), a plain text or a coloured html page.
Figure: Alignment
**Pros & Cons of Multalin:**

**Pros:**

1. It has 7 scoring matrices and gives a choice for user-defined matrix.
2. The output can be seen in a text page, html page, gif image, with color indications, alignment and tree description.
3. It gives an option for users to specify their values of gap create and gap extend rather than selecting a pre-defined value from drop-down.

**Cons:**
1. It dosen’t allow for the PAM matrices.
2. It forces the usage of cost matrices, there are minimal options for selecting the matrices.
3. It also forces the gap open penalty and gap extension penalties.
4. This works better for short sequences not for longer sequences.
Tool 2: ClustalW – Multiple Sequence Alignment

About ClustalW:

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

What ClustalW uses:

- Global alignment algorithm
- Pair-wise Progressive alignment.
- Considers the sequence redundancy.

User Manual for ClustalW

Step 1: Simply navigate to site: [http://www.ebi.ac.uk/clustalw/](http://www.ebi.ac.uk/clustalw/)

Step 2: From here, we can either use the tool hosted online or download and install it for local use.

Step 3: Either upload your protein sequences or paste the protein sequences in the given text box and hit “Run” button to view the aligned output with default settings

Step 4: Use the various output files to analyse the output
Sample Outputs

Regions with high conservation amongst the five sample input sequences of Keratin protein

Phylogenetic analysis by means of two trees

- Cladogram: Cladogram is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length, thus cladograms show common ancestry, but do not indicate the amount of evolutionary "time" separating taxa.

- Phylogram: Phylogram is a branching diagram (tree) assumed to be an estimate of a phylogeny, branch lengths are proportional to the amount of inferred evolutionary change.
**Pros:**

1. ClustalW has a very simple and sweet interface, with various options that can be tinkered with.
2. It shows alignments with an option of full or fast, so depending on the trade off between fastness and correctness we can choose the option we need.
3. Most commonly used algorithm hence a wide acceptance.
4. A self explanatory Help and FAQ documentation

**Cons:**

1. Does not provide option for various BLOSUM and PAM matrices.
2. Can be extended to have more additional functionalities.
3. Does not allow us to specify Gap open or Gap extend penalties
**Tool 3: Geneious 2.0.10– Multiple Sequence Alignment**

**About the tool**

This is the latest global alignment tool developed by Biomatters Development team. Geneious is based on BioJava which is an open source project dedicating to provide a Java framework for processing biological data. **Geneious** is a self-organizing, automatically-updating library of genomic and genetic data; that provides a fully integrated, visually advanced toolset for: Sequence alignment and phylogenetics, Sequence analysis, BLAST Protein structure viewing, NCBI, EMBL, Pubmed auto-find & more API for creating your own plugins. It includes an advanced and integrated suite of tools which is designed to support a wide variety of data formats. Geneious provides a huge list of databases like the Entrez Genome Database, PubMed database, Popset database etc that can be searched. It is based on a time-efficient method and a heuristic algorithm called the progressive alignment.

**Progressive alignment algorithm**

The most popular and time-efficient method of multiple sequence alignment is progressive pairwise alignment. The idea is very simple. At each step, a pairwise alignment is performed.

In the first step, two sequences are selected and aligned. The pairwise alignment is added to the mix and the two sequences are removed. In subsequent steps, one of three things can happen:

- Another pair of sequences is aligned
- A sequence is aligned with one of the intermediate alignments
- A pair of intermediate alignments is aligned

This process is repeated until a single alignment containing all of the sequences remains.

**User Manual**

1. Click the url [www.geneious.com](http://www.geneious.com)

2. Select Download Geneious Pro 2.0.10 which is in the lower bottom at the right

3. After downloading, open the tool and enter the sequences. You can do this in two ways. You can either enter this by creating new sequences which is useful if you want to store the sequences within the tool locally or you can import the file directly.

4. This is how the tool shows all the sequences:
5. We can now align the sequences by clicking the Alignment option. This is the screen we get for choosing different alignment options.
6. The tools lets us define the gap open penalty and the gap extension penalty. The default values are 12.0 for gap open penalty and 3.0 for gap extension penalty. The sequence alignment type can be selected which is either Geneious Alignment or ClustalW. Cost matrix can be selected and Refinement iterations can be set from 1 to any number of iterations. The default value used for Cost matrix is Blosum 62
7. The new aligned sequence can be viewed by opening the sequence.
8. There is an option for generating the tree by enabling the build tree option. This is a screen shot for the tree obtained:

Pros:
1. It has a friendly UI which has a detailed tutorial and is much easier to navigate than the existing applications
2. This tool allows for refining an alignment after it is done where it removes the sequence one at a time and then re-aligns the removed sequences to a profile of the remaining sequences.
3. It has a unique feature of build tree via alignment which actually speeds up the whole process of building a tree
4. This also has an extra feature of performing alignment using CLUSTALW which is the most widely used algorithm
5. New plugins can be written and incorporated with relative ease.
6. It provides numerous alignment options and also provides a folder structure where you can store your own local data in the tool thus avoiding numerous imports everytime you open the application.
7. A direct connection to databases
Cons:

1. It uses a probabilistic way of aligning the sequences and not an optimal algorithm as the tool does not analyze the sequences structurally and evolutionarily.
2. This tool does not let us define the k-tuple (word-size).
3. This tool does not specify the time taken to perform an alignment. This might be useful for analyzing the complexities.
SEQUENCES USED

We performed a search on NCBI for Keratin protein sequences and we traced out 5 sequences which are mostly similar ones. The following are the sequences in FASTA format:

> homo sapien
vtlartdlem qieglkeela ylrknheeem lalrgqtggd vnvemdaapg vdlrsilnem rdqyeqmaek
nrrdaetwlf skteelnevk asnselqqs rsveltelrv lqgleielqqs qlstkaslen sleetkryc mqlsqiqgli
gsveeqlaql rcemeqqsse yqilidvktr leheiatyr llxgedahls sqqasqgqsys sreffssss sssrqrptil
keqsssssfq qgss

> mus musculus
matsrqrfts ssmkgsccgi gggssrmssv laggscrapsc tcggmsvss tfssggyvcejq ggggyggsfss
ssfegggsf sgfgdfgfqgg fggflqggl ggqdgdfllv gsekvtnqnl ndrlatyldk vraleenard
levkirdwq qrqqteikdy spyktiedl ksklaaptq naqflqidn araladfrt kyenelflfrq svegdinglr
kvldeltls adlemqienl reaelflkkn heemamlrg qtggdvnem daapgvlslr lihnmrdqye
qmaeknrvd fawfrknte lnkevasnsdf liqnsnrs eelrrfqqgle ileqsqlmsmk aslensleet
kgrycmqlssq iqqgissvee qlaqlrceme qqqseynill dkvrkrrqi atyrllldge nhissqshss gqssqsyss
refvssssr prsilkeqgs tsfsqssqs srd

> Rattus
matsrqrfts ssmkksccgi gggssrmssv laggscrapsc tyggmsvss tfssgaacsq ggggyggsfss
ssfegggsf sgfgdfgfqgg fggflqggl ggqdgdfllv lgvsekvtm qnlndrtyld kdvllaremos ndslevkird
wyqrqrpqtei kydpffrii edlqskivra kqenaaqsvld inlraandd frtkydnets lrqlvesd inlrrstdl
msrademqi esrelayl kkhneemla lrrqggsdvn vemdaapvgd lsrilnsmrd qyeqmaekn
rdveaqflsqkq teelnqvas nheliqcrs eveslrffqgle iieqsqlsmk ssmkanfls eetkrycveq sqiqqlgs
leqqlavrlc emeqqssqyen illdkvtrk qeiatyrrrl dgenvhssss qhssqgsqys gevffssssrr prsilkeqgs
tsfsqssqsq sgy

> Bos Taurus
mttitrhhss gskgsgxxla gggssrscrvs lsgllsgsrll slaqggsqgsl glueyscscs yggsgggygsg fggydggflvg
gckatmtnl drlasyldkv raleantel elkirdwqq qapppadys syftktiedl nkhltavdn
anllqisnrdn lralldftrtk feteqarlrvs veadinglr vldeltlara ldemqienk eelayrknh eemkalrgq
vggeinemd aapvgdlsri lmenmdqyek maenrrldae dwffskteel nrevtnsel vqsgkeseise
lrlqlqalei elqsqlssmak smeslalarte nrycsvqssq gqglgsveeq qlaqlrcemeq qnyqeykilld tkvtrkqen
qyrlllgege ahtlqytktc pqntqrqvrti vevqdvqgrvi srsqyqhts h

> Canis lupus familiaris
maatttsirq fstgsvkql egppggsfpm ssyrggcacr aspllgvsc gtmstvssrf ssalggggyggy
yteslggggg sqgggsfsgf gaqgggsgf ssadlgggs ekemqmnld rlasylkavr aleenanadle
vkhihwyykk qgpgpavdys fttktelml kilaaidnaq slvqinmar ladafftke etelnmsv eadtnlvrr
ldeltlara laimqelske eleyakknhe eemailrqgv gggdsvemda apgvlslr nmermdqykekm
aeenkkedaf wffskteel nrevtanqal rsrseitel rrvqnliee lqsqslmska legsaltea rgyaqlaqrlq
glsisieeqq gelrdmerq nqeyqvlldv ktlrrqueiat yrrillega hlatqysssl isqpteatv ttrqvrti
evqdvckvss vqvhrth
Comparitive Analysis:

In this section we compared all the above discussed tools by playing with the various parameters passed to perform the sequence alignment. Below are the observed results.

We used the following 5 sequences of Human, Mouse, Rat, Buffalo, Dog to study these tools.

CLUSTALW

1. Impact of Gap penalties

<table>
<thead>
<tr>
<th>Gap Open</th>
<th>Gap Ext</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>10</td>
<td>14452</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>15522</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
<td>15908</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>16329</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>16658</td>
</tr>
</tbody>
</table>

Plot of above data
**Interpretation:** As we decrease the penalty parameters our alignment score increases, the reasoning for this is that when we increase the penalties we are penalizing each Gap creation and hence the number of matches or alignments made would be less and hence a less alignment score.

2. **Alignment Scores by varying the number of sequences for ClustalW**

<table>
<thead>
<tr>
<th>No. of sequences</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2450</td>
</tr>
<tr>
<td>3</td>
<td>5922</td>
</tr>
<tr>
<td>4</td>
<td>11249</td>
</tr>
<tr>
<td>5</td>
<td>15522</td>
</tr>
</tbody>
</table>

**Interpretation:** As the number of sequences increases the complexity to solve the problem also increases. From the above plot we observe that as we go on increasing the
number of sequences, the alignment score also increases, this observation agrees to the fact that by increasing the number of sequences we also increase the probability of finding many matched or aligned pairs amongst sequences

**MULTIALIN**

1. Observations made with MultiAlin

<table>
<thead>
<tr>
<th>Gap open penalty</th>
<th>Gap Extension Penalty</th>
<th>MSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10</td>
<td>476</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>477</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>479</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>479</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>492</td>
</tr>
</tbody>
</table>

![Variation of MSF with Gap Open penalty](image)
Inferences: As we decrease the penalty parameters the msf score increases

GENEIOUS

Experiment

We try to perform the global alignment alignment of 3 sequences, 4 sequences and 5 sequences using Blosum 62 as the cost matrix and 12 as gap open penalty, 3 as gap extension penalty and compare the similarities.

The first set of 3 sequences are keratin sequences of Mouse, rat and buffalo.
The next set of 4 sequences are keratin sequences of dog, mouse, rat and buffalo.
The 5 sequences are keratin sequences of mouse, rat, buffalo, dog and human.

These were our observations:

<table>
<thead>
<tr>
<th>No of Sequences</th>
<th>% of Similarity</th>
<th>Gap open penalty</th>
<th>Gap extn penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>72</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>58</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>
Inference made:
1. As the number of sequences increases, the similarity score gradually decreased.
2. Mouse, rat, buffalo had the best alignment. The human sequence had a worst match with the other four

**Impact of Cost matrix:**

The Geneious tool has a wide range of cost matrix selection. So we choose to analyze the similarity changes when the cost matrix changes for all the 5 sequences. These were our observations (Blosum):

<table>
<thead>
<tr>
<th>Cost Matrix</th>
<th>Gap open penalty</th>
<th>Gap extension penalty</th>
<th>Similarity %</th>
<th>Identical Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blosum 45</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 50</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 55</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 60</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 65</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 70</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 75</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 80</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>151</td>
</tr>
<tr>
<td>Blosum 85</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>Blosum 90</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
</tbody>
</table>
For the PAM

<table>
<thead>
<tr>
<th>Cost Matrix</th>
<th>Gap open penalty</th>
<th>Gap extension penalty</th>
<th>Similarity %</th>
<th>Identical Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 100</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 110</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 120</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 130</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 140</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 150</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 160</td>
<td>12</td>
<td>3</td>
<td>57</td>
<td>150</td>
</tr>
<tr>
<td>PAM 170</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>151</td>
</tr>
<tr>
<td>PAM 180</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 190</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 200</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 210</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 220</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 230</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 240</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
<tr>
<td>PAM 250</td>
<td>12</td>
<td>3</td>
<td>58</td>
<td>150</td>
</tr>
</tbody>
</table>

**Inferences**

1. There was no change in the similarity % but there was one case where it decreased and then again increased.