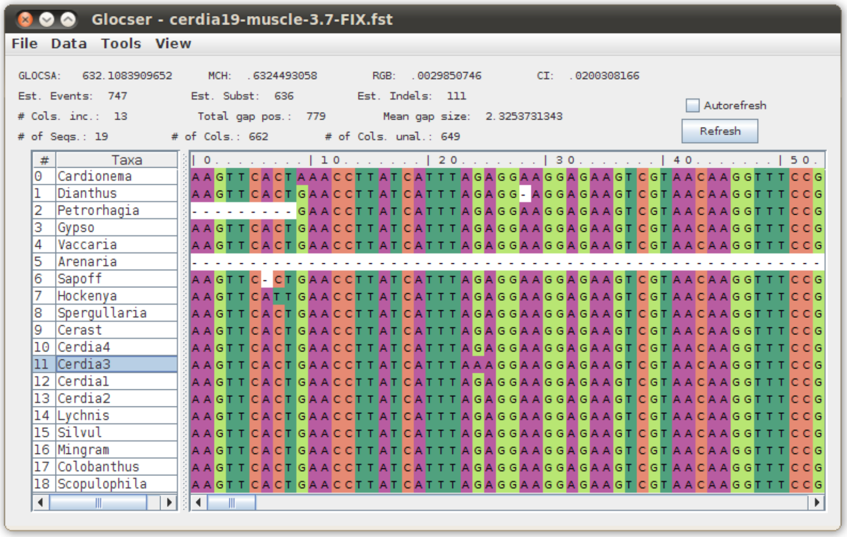
Glocser v1.0 Manual

Edgar D. Arenas-Díaz

March 21, 2013

1 Introduction

Glocser is a software tool intended to help in the process of aligning a set of DNA sequences. It reads FASTA and HENNIG86 alignments files; incorporates indicators and charts about the alignment quality including the Glocsa evaluation criterion. It allows to perform manual alignments and it's able to call an external alignment tool.



```
GLOCSA:  632.1083909652    MCH:  .6324493058       RGB:  .0029850746      CI:  .0200308166
Est. Events:  747          Est. Subst:  636        Est. Indels:  111
# Cols. inc.:  13          Total gap pos.:  779    Mean gap size:  2.3253731343
# of Seqs.: 19             # of Cols.: 662         # of Cols. unal.: 649

#  Taxa          |0.........|10........|20........|30........|40........|50.
0  Cardionema    AAGTTCACTAAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
1  Dianthus      AAGTTCACTGAACCTTATCATTTAGAGG-AGGAGAAGTCGTAACAAGGTTTCCG
2  Petrorhagia   ---------GAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
3  Gypso         AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
4  Vaccaria      AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
5  Arenaria      ------------------------------------------------------
6  Sapoff        AAGTTC-CTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
7  Hockenya      AAGTTCATTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
8  Spergullaria  AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
9  Cerast        AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
10 Cerdia4       AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
11 Cerdia3       AAGTTCACTGAACCTTATCATTTAAAGGAAGGAGAAGTCGTAACAAGGTTTCCG
12 Cerdia1       AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
13 Cerdia2       AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
14 Lychnis       AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
15 Silvul        AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
16 Mingram       AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
17 Colobanthus   AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
18 Scopulophila  AAGTTCACTGAACCTTATCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCG
```

Figure 1: Glocser with an alignment

2 Running Glocser

Glocser is programmed using the Java Software Development Kit v1.6 , and it should run everywhere the Java Runtime Enviroment v1.6 (or newer) runs. As with most Java programs a jar file (`glocser.jar`) is to be executed in order to run Glocser. If you need to open large alignment files, a couple of scripts are provided (`glocser.bat` for Windows and `glocser.sh` for GNU/Linux or Mac OS X) than give to the Java Virtual Machine the parameters needed to use more memory.

If you intend to use an external alignment tool from Glocser, you need to have the path to the executable and give it to Glocser in the frame where it is to be called. Alternatively you can put a copy of the exectuable in the `ext-tools` folder, which is the default location, and just edit the name of the executable in the frame where it is to be called.

3 File Managment

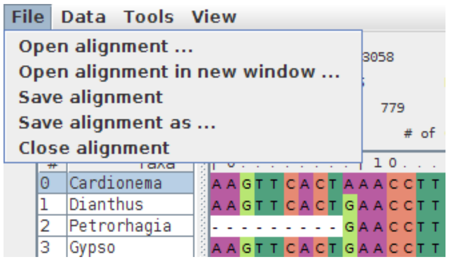


Figure 2: File menu

3.1 Opening files

Only FASTA and HENNIG86 DNA alignment files are currently supported, ther are opened through the `File->Open alignment ...` menu item. The item `File->Open alignment in new window ...` opens the alingment in a new window.

3.2 Saving files

Alignments can only be saved in FASTA format. With the `File->Save alignment` menu item, the open alignment is saved with the same file name and location. Using `File->Save alignment as ...` a new location and file name can be set.

3.3 Closing files

An alignment file can be closed explicitly through the `File->Close alignment` menu item. When an alignemnt is already opened and another one is opened, the previous one is closed. When an opened alignment is saved with a differnt name, the alignment with the previous name is closed.

4 Sequence reordering

Sequences can be reordered in an open alignment using the pop-up menu that appears when right clicking on the sequences' names (figure 3). In this menu are menu items to move the selected sequence (if any) to the top, to the bottom, to te position in which the pop-up menu was invoked. There are also an item to sort the sequences alphabetically (useful when comparing alignments) and an item to restore the order that was present when the alignment was opened.

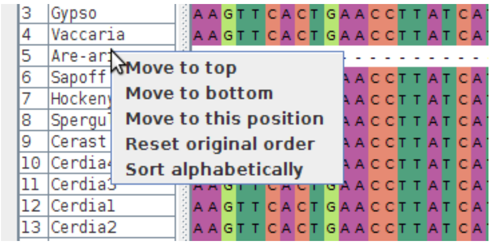


Figure 3: Sequences' name contex menu

5 Manual Alignment

Glocser allows manual adjustments to alignments of multiple sequences. These can be done through some menu items in the `Data` menu (figure 4), or with the keyboard and mouse.

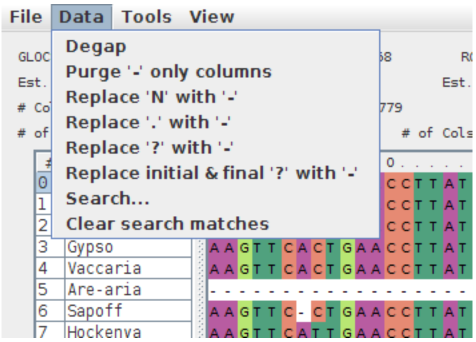


Figure 4: Data menu

5.1 Degap

With the `Data->Degap` menu item every occurrence of '-' (gap) is removed from the alignment, making it *unaligned*.

5.2 Purge '-' only columns

Every column in the alignment that has only gap codifications '-' is removed with the `Purge `-' only columns` menu item.

5.3 Replace 'N' with '-'

Each occurence of 'N' is replaced with a standard gap codifitacion '-'.

5.4 Replace '.' with '-'

Each occurence of '.' is replaced with a standard gap codifitacion '-'.

5.5 Replace '?' with '-'

Each occurence of '?' is replaced with a standard gap codifitacion '-'.

5.6 Replace initial & final '?' with '-'

Occurrences of '?' heading and tailng a sequence are replaced with a standard gap codifitacion '-'.

5.7 Inserting and deleting gaps using keyboard and mouse

5.7.1 Inserting gaps

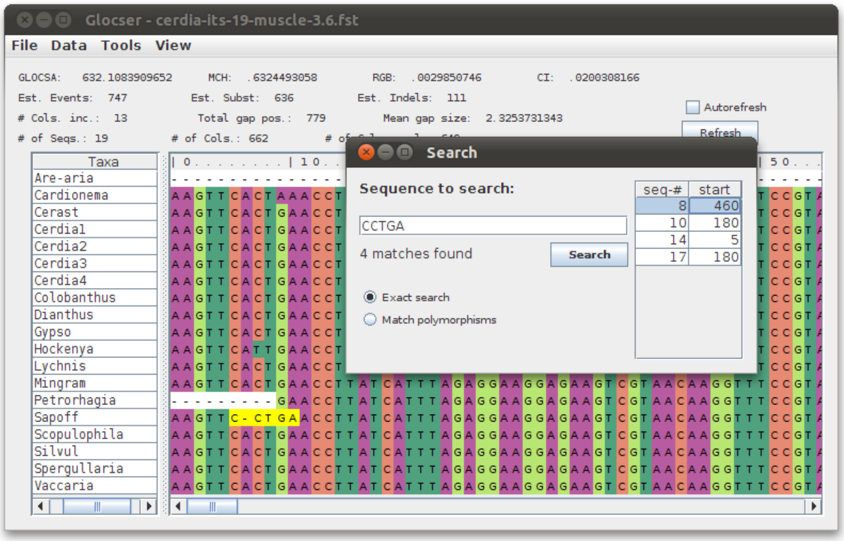
Selecting a range of positions in the alignment and pressing the `INSERT` or `SPACE` keys introduces new '-' symbols *before* the first column in the selected range of positions. Pressing `TAB` inserts 3 '-' symbols.

5.7.2 Deleting gaps

Pressing `BACKSPACE` deletes one '-' symbol before the first column of the selected range (if there is one in every sequence in the selected region). Selecting a range of positions with '-' and pressing the `DELETE` key suppresses one column of the selected '-' (If not every sequence in the selection has a '-' in the first column, nothing is deleted).

5.7.3 Moving bases within gaps

Selecting a range of positions which has bases in it and it is between other contiguous '-', and then pressing `Shift + INSERT` or `Shift + SPACE` moves the selected range one position to the right (if there are '-' next to the right of the selection in every sequence). Pressing `Shift + BACKSPACE` has the analogous effect but to the left.



GLOCSA: 632.1083909652 MCH: .6324493058 RGB: .0029850746 CI: .0200308166
Est. Events: 747 Est. Subst: 636 Est. Indels: 111
Cols. inc.: 13 Total gap pos.: 779 Mean gap size: 2.3253731343
# of Seqs.: 19 # of Cols.: 662

Sequence to search: CCTGA

4 matches found

● Exact search
○ Match polymorphisms

seq-#	start
8	460
10	180
14	5
17	180

Figure 5: Search in Glocser

6 Search of motifs

In Glocser motifs can be searched using the `Data-> Search...` menu item. The search dialog permits two different behavoirs: *Exact search* and *Match polymorphisms*.

Using *Exact search* will strictly match the characters in the search dialog, with no regard to polymorphisms, considering them as independent characters.

Match polymorphisms, on the other hand, considers polymorphisms as any of their corresponding base possibilities.

When the search is performed, in the right side of the search dialog the occurrences are listed, identified by the sequence number in which it is and its start position in it. Also, they are highlighted in yellow in the alignment panel. Selecting an occurence in the search dialog will also select it in the alignment panel.

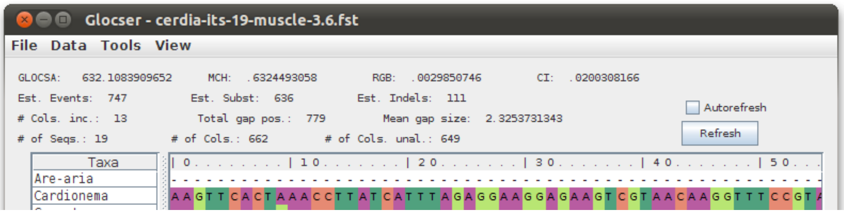
7 GLOCSA and other indicators

7.1 In the main frame

In the main Glocser frame, several indicators about the alignment are shown, and can be autoupdated when the alignment is modified or refreshed manualy.

The indicators shown are the following:

- GLOCSA. The GLOCSA score for the alignment, higher values indicate



```
GLOCSA:   632.1083909652    MCH:   .6324493058       RGB:   .0029850746       CI:   .0200308166
Est. Events:  747           Est. Subst:  636         Est. Indels:  111
# Cols. inc.:  13            Total gap pos.:  779       Mean gap size:  2.3253731343
# of Seqs.: 19            # of Cols.: 662         # of Cols. unal.: 649

    Taxa      | 0 . . . . . . . . | 1 0 . . . . . . . | 2 0 . . . . . . . | 3 0 . . . . . . . | 4 0 . . . . . . . | 5 0 . . .
Are-aria      - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Cardionema    A A G T T C A C T A A A C C T T A T C A T T T A G A G G A A G G A G A A G T C G T A A C A A G G T T T C C G T A
```

Figure 6: GLOCSA in the main frame

a higher quality alignments. It is composed of three indiviual criteria: Mean Column Homogeneity (MCH), Reciprocal of Gap Blocks (RGB) and Columns Increment (CI). See Appendix A.

- MCH. Mean Column Homogeneity, a criterion that evaluates how homogenious are the columns of the alignment on average. A greater homogeneity indicates a better alignment. See section A.1.

- RGB. Reciprocal of Gap Blocks, the reciprocal of the number of gap blocks (contigous gap codifications “-”, not including trailing and initial gaps). See section A.2.

- CI. Columns Increment, the ratio of the increment in the number of columns after alignment. See section A.3.

- Estimated Events. The estimated number of synapomorphies in the alignment, considering substitutions and indels.

- Estimated Substitutions. The estimated number of subsitution in the alignment.

- Estimated Indels. The estimated number of indel in the alignment.

- Number of columns incremented. How many columns are added to the alignment after aligning it.

- Total gap positions. The total number of gap codifications in an alignment (not including trailing and initial gaps).

- Mean gap size. The mean size of the gap blocks (contigous gap codifications “-”, not including trailing and initial gaps).

- Number of sequences. How many sequences are in the alignment.

- Number of columns. Present number of columns in the alignment.

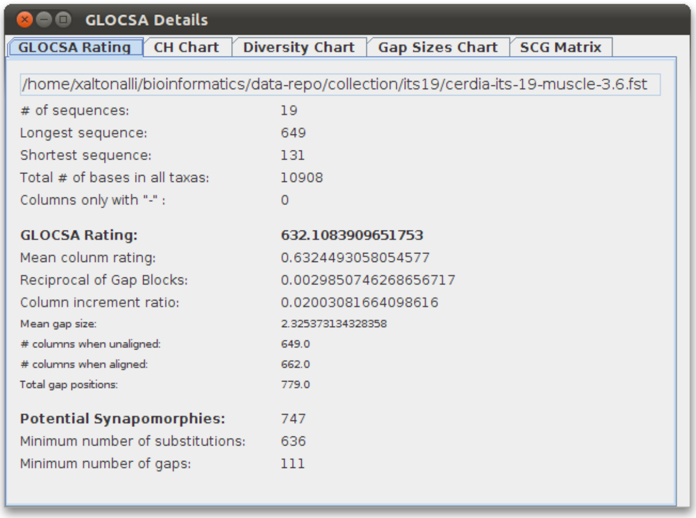
- Number of columns when unligned. How many columns the alignment has if no gaps are in it.

7.2 GLOCSA Details

In the `View->GLOCSA Details` menu item, GLOCSA and other indicators are available, along with three related charts about the alignment and the simple gap coding matrix.

- CH Chart. A Column Homogeneity Histogram for the whole alignment.
- Divesity Chart. A pie chart of the diversiy of the composition of the columns in the alignment.
- Gap Sizes Chart. An histogram of the gap block sizes in the alignment.
- SGC Matrix. The simple gap coding matrix can be genetared and saves to a file in this tab.

Copy, save, and zoom capabilities are available for the charts in the contextual menu (right mouse click).



/home/xaltonalli/bioinformatics/data-repo/collection/its19/cerdia-its-19-muscle-3.6.fst

# of sequences:	19
Longest sequence:	649
Shortest sequence:	131
Total # of bases in all taxas:	10908
Columns only with "-" :	0
GLOCSA Rating:	**632.1083909651753**
Mean colunm rating:	0.6324493058054577
Reciprocal of Gap Blocks:	0.0029850746268656717
Column increment ratio:	0.02003081664098616
Mean gap size:	2.325373134328358
# columns when unaligned:	649.0
# columns when aligned:	662.0
Total gap positions:	779.0
Potential Synapomorphies:	747
Minimum number of substitutions:	636
Minimum number of gaps:	111

Figure 7: GLOCSA Details

8 External Alignment Tool Calling

Within Glocser an external alignment tool can be called to align the opened set of sequences. This can be done with the `Align-> Align with ext. tool`. In the *Align with ext. tool* dialog, the first text field indicates the path to the *external tool* executable, and the second the options passed to it. Both can be edited to suit a particular installation or desired parameters.

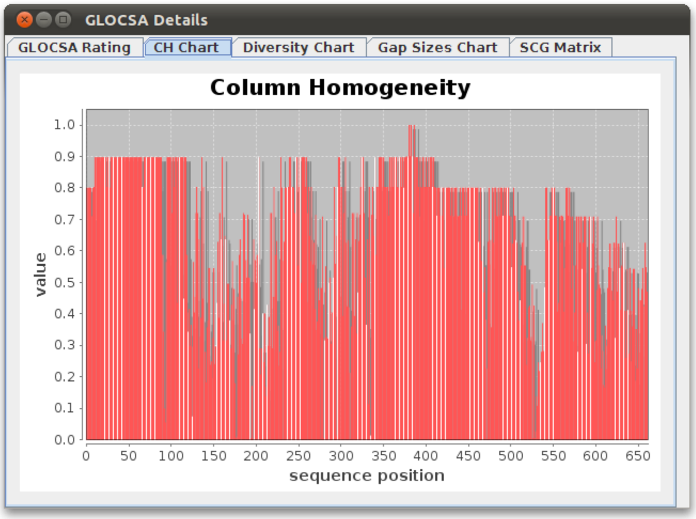


Figure 8: Column Homogeneity Chart

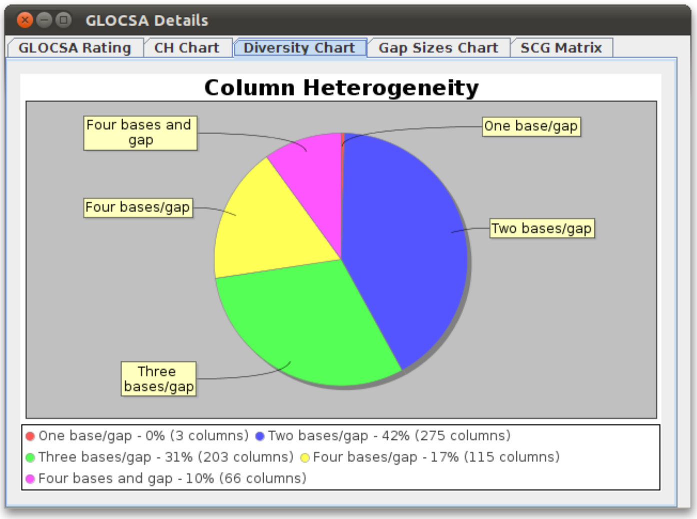
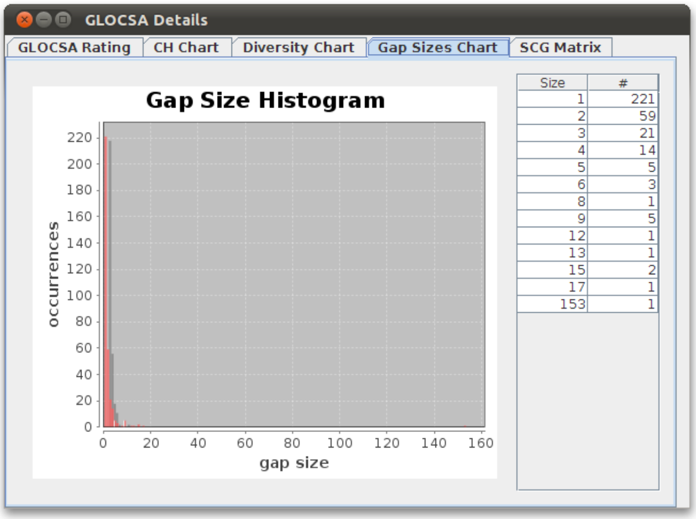
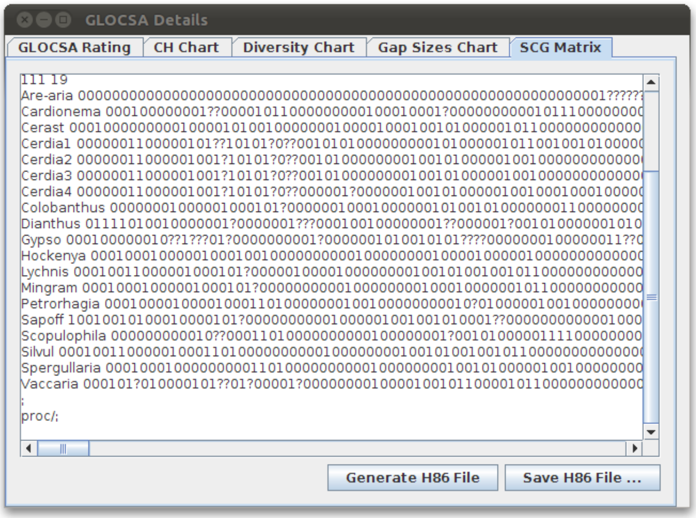

GLOCSA Details
GLOCSA Rating
CH Chart
Diversity Chart
Gap Sizes Chart
SCG Matrix
Column Heterogeneity
Four bases and gap
One base/gap
Four bases/gap
Two bases/gap
Three bases/gap
● One base/gap - 0% (3 columns) ● Two bases/gap - 42% (275 columns)
● Three bases/gap - 31% (203 columns) ● Four bases/gap - 17% (115 columns)
● Four bases and gap - 10% (66 columns)

Figure 9: Diversity Chart



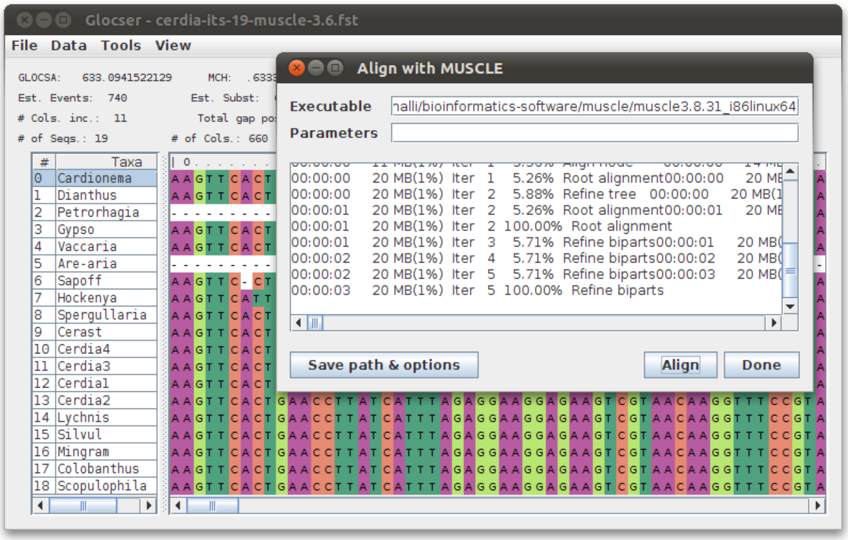
Size	#
1	221
2	59
3	21
4	14
5	5
6	3
8	1
9	5
12	1
13	1
15	2
17	1
153	1

Figure 10: Gap Sizes Chart



```
111 19
Are-aria 000000000000000000000000000000000000000000000000000000000001?????
Cardionema 0001000000001??000010110000000001000100001?0000000000101110000000
Cerast 0001000000000100001010010000000010000100010010100000101100000000000
Cerdia1 00000011000001 01??10101?0??001010100000000010100000101100100101000000
Cerdia2 000000110000010 01?10101?0??00101000000000100101000001001000000000000
Cerdia3 000000110000010 01?10101?0??00101000000000100101000001001000000000000
Cerdia4 000000110000010 01?10101?0??000001?0000001001010000010010001000100000
Colobanthus 0000000100000100010 1?000000100010000000101001010000000110000000
Dianthus 01111010010000001?0000001???0001001000000001??000001?00101000000101 0
Gypso 000100000010??1???01?000000001?00000010100101 01???00000001000000 11??0
Hockenya 0001000100000100010010000000000010000000010000100000100000000000 00
Lychnis 000100110000010 00101?0000010000100000000100101001001011000000000000
Mingram 00010001000001000101?000000000010000000001000100000010110000000000
Petrorhagia 000100001000010001101000000010010000000000 10?010000010010000000 0
Sapoff 100100101000100000101?0000000000100000100100101000 1??00000000000001000
Scopulophila 000000000010??0001101000000000001000000001?00101000001111000000 00
Silvul 0001001100000100011010000000000010000000010010100100101100000000000000
Spergullaria 000100010000000001101000000000001000000000100101000001001000000000
Vaccaria 000101?0100000101??01?00001?0000000010000100101100001011000000000000
;
proc/;
```

Figure 11: Simple Gap Coding Matrix



```
GLOCSA:   633.0941522129     MCH:  .6333
Est. Events:  740            Est. Subst:
# Cols. inc.:  11            Total gap po
# of Seqs.: 19               # of Cols.: 660
```

#	Taxa
0	Cardionema
1	Dianthus
2	Petrorhagia
3	Gypso
4	Vaccaria
5	Are-aria
6	Sapoff
7	Hockenya
8	Spergullaria
9	Cerast
10	Cerdia4
11	Cerdia3
12	Cerdia1
13	Cerdia2
14	Lychnis
15	Silvul
16	Mingram
17	Colobanthus
18	Scopulophila

Executable: nalli/bioinformatics-software/muscle/muscle3.8.31_i86linux64

Parameters:

```
00:00:00   20 MB(1%)  Iter  1   5.26%  Root alignment00:00:00    20 ME
00:00:00   20 MB(1%)  Iter  2   5.88%  Refine tree   00:00:00   20 MB(1
00:00:01   20 MB(1%)  Iter  2   5.26%  Root alignment00:00:01    20 ME
00:00:01   20 MB(1%)  Iter  2 100.00%  Root alignment
00:00:01   20 MB(1%)  Iter  3   5.71%  Refine biparts00:00:01   20 MB(
00:00:02   20 MB(1%)  Iter  4   5.71%  Refine biparts00:00:02   20 MB(
00:00:02   20 MB(1%)  Iter  5   5.71%  Refine biparts00:00:03   20 MB(
00:00:03   20 MB(1%)  Iter  5 100.00%  Refine biparts
```

Figure 12: Align with Muscle

In the text area below, the output of the external alignment tool is shown when the tool completes its execution.

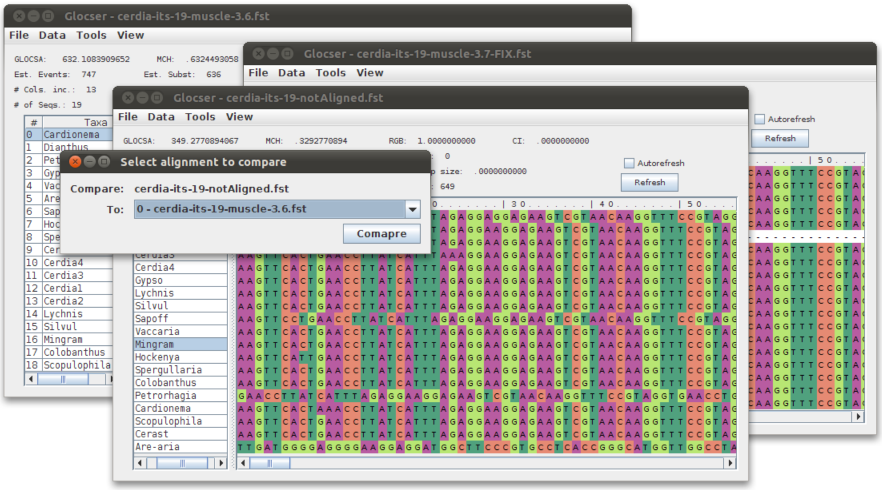
In theory any alignment tool that can read the input from `stdin` and write the output to `stdout` can be called from Glocser, only the path and options would need to be changed for that purpose. But only *MUSCLE v3.6 - 3.8* were tested.

9 Compare alignments

Two alignments can be compared after they have been opened in Glocser in different windows, using the menu item `Tools-> Compare alignments...` in any alignment frame. If more than two alignments are opened, a frame will open in which the second alignment for the comparison is to be chosen (the first alignment being the one in which the menu item was selected). If only two alignments are open, they are compared without any further questions.

The alignments to be compared must be have the same sequences and in the same order. If they are in a different order, they can be sorted alphabetically before comparing them.

The Alignment Comparison frame shows a table with every difference between the two alignments, one in each row. A difference is considered as a gap of different length preceding a specific base in a sequence of the alignment. Thus, they are identified in a single row with a sequence number and a base index, followed by the position and size of the preceding gap in both alignments.



Glocser - cerdia-its-19-muscle-3.6.fst
Glocser - cerdia-its-19-muscle-3.7-FIX.fst
Glocser - cerdia-its-19-notAligned.fst
File Data Tools View
Autorefresh
Refresh

GLOCSA: 632.1083909652 MCH: .6324493058
Est. Events: 747 Est. Subst: 636
Cols. inc.: 13
# of Seqs.: 19

GLOCSA: 349.2770894067 MCH: .3292770894 RGB: 1.0000000000 CI: .0000000000

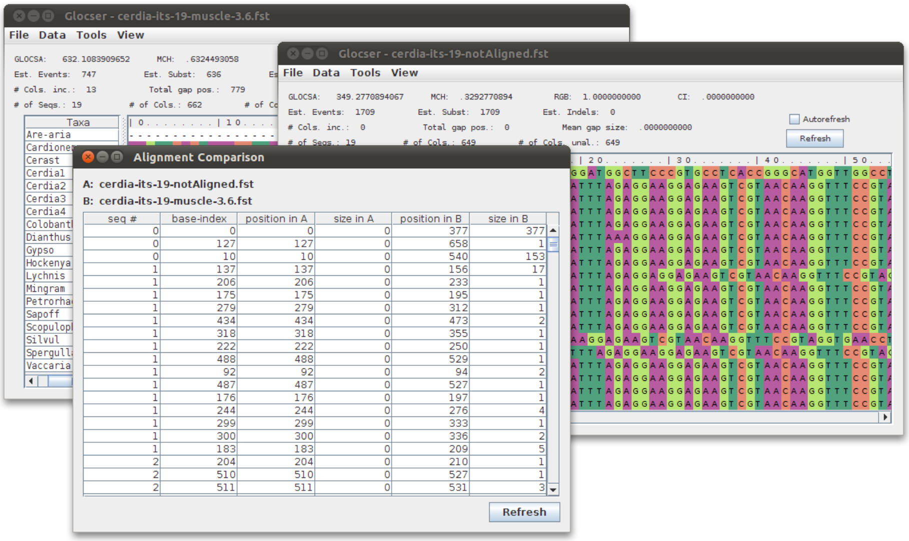
Select alignment to compare

Compare: cerdia-its-19-notAligned.fst

To: 0 - cerdia-its-19-muscle-3.6.fst

Comapre

Figure 13: Select alignment to compare



Glocser - cerdia-its-19-muscle-3.6.fst
Glocser - cerdia-its-19-notAligned.fst
File Data Tools View
Autorefresh
Refresh

GLOCSA: 632.1083909652 MCH: .6324493058
Est. Events: 747 Est. Subst: 636
Cols. inc.: 13 Total gap pos.: 779
# of Seqs.: 19 # of Cols.: 662

GLOCSA: 349.2770894067 MCH: .3292770894 RGB: 1.0000000000 CI: .0000000000
Est. Events: 1709 Est. Subst: 1709 Est. Indels: 0
Cols. inc.: 0 Total gap pos.: 0 Mean gap size: .0000000000
# of Seqs.: 19 # of Cols.: 649 # of Cols. unal.: 649

Alignment Comparison

A: cerdia-its-19-notAligned.fst

B: cerdia-its-19-muscle-3.6.fst

seq #	base-index	position in A	size in A	position in B	size in B
0	0	0	0	377	377
0	127	127	0	658	1
0	10	10	0	540	153
1	137	137	0	156	17
1	206	206	0	233	1
1	175	175	0	195	1
1	279	279	0	312	1
1	434	434	0	473	2
1	318	318	0	355	1
1	222	222	0	250	1
1	488	488	0	529	1
1	92	92	0	94	2
1	487	487	0	527	1
1	176	176	0	197	1
1	244	244	0	276	4
1	299	299	0	333	1
1	300	300	0	336	2
1	183	183	0	209	5
2	204	204	0	210	1
2	510	510	0	527	1
2	511	511	0	531	3

Refresh

Figure 14: Comparison of two alignments

A GLOCSA - A new objective function

The Global Criterion for Sequence Alignment (GLOCSA) is a new proposed function to assess the quality of multiple sequence alignments of DNA. It has been build from the ground up with simplicity and a global approach in mind. By global it is understood that it rates the alignment as a whole, i.e. all sequences considered simultaneously, not taking pairs of sequences to score their corresponding alignment. It also takes into account the gaps, seeking to favor parsimony.

GLOCSA is composed of three individual criteria, *Mean Column Homogeneity (MCH)*, *Reciprocal of Gap Blocks (RGB)* and *Columns Increment (CI)*. These are combined in a polynomial with a set of corresponding weights (w_{mch}, w_{rgb} and w_{ci}). These weights are set by default to the values shown in table 1. This default values were determined empirically, adjusting them to assign better scores to better alignments using a set of artificial examples and some real-world alignments.

$$GLOCSA = w_{mch}MCH + w_{rgb}RGB + w_{ci}CI \tag{1}$$

Table 1: GLOCSA Weights

w_{mch}	$= 1000$
w_{rgb}	$= 20$
w_{ci}	$= -20$

The main problem faced when scoring alignments is that the exact evolutionary history of the involved sequences is never known. Theories can be stated about which alignment reflects the more plausible or probable evolutionary history (which is what produces the differences in the sequences), but certainty cannot be guaranteed.

Compared to the other schemes of sequence alignment evaluation rating them on a pair basis, such as *weighted sum of pairs*, GLOCSA has the advantage of rating the whole alignment at a time (with the *Mean Column Homogeneity* criterion). It also has the advantage of considering parsimony, favoring more concentrated *gaps* (with *Reciprocal of Gap Blocks*) and smaller alignment matrices (with *Columns Increment*).

At the moment it is intended to rate only multiple sequences of DNA composed of the standard IUB/IUPAC codifications for nucleic acids, shown in table 2.

To score an alignment of multiple sequences, a matrix with C columns and S lines is considered, where C is the maximum number of positions in a sequence, and S is the number of sequences in the alignment. Initially, to perfectly fit all the sequences in the matrix, gap positions are appended ("$-$") at the end of the shorter sequences.

Table 2: Nucleic Acid Codifications Supported

A	Adenosine
C	Cytosine
G	Guanine
T	Thymine
R	G A (puRine)
Y	T C (pYrimidine)
K	G T (Ketone)
M	A C (aMino group)
S	G C (Strong interaction)
W	A T (Weak interaction)
B	G T C (not A) (B comes after A)
D	G A T (not C) (D comes after C)
H	A C T (not G) (H comes after G)
V	G C A (not T, not U) (V comes after U)
N	A G C T (aNy)
-	gap
?	any base or gap

A.1 Mean Column Homogeneity

In the alignment matrix each position is represented in a column, and the column homogeneity has the purpose of rating the grade of diversity in the elements of a given position, scoring higher the more homogeneous columns.

The basic idea is that the occurrences of each of the four bases in a column are counted. A,C,G and T are counted with a weight of 1.0 while polymorphisms are counted as an equal fraction of a unit for each base they represent (e.g. A counts 1.0 for A while R is either G or A, so it counts 0.50 for G and 0.50 for A). Gaps are also counted, with a unit for each. Using these counts the column homogeneity for each column is computed.

The count of bases and gaps are computed in $wc_{jt} \quad \forall \quad 0 \leq t \leq 4$, where t is the index for a base or gap which is being counted and j is the column. These weighted counts are the result of adding up to wc_{jt} the corresponding weight (shown in table 3) for the codification of each sequence in the column. This can be expressed as,

$$wc_{jt} = \sum_{i} T_w \left(t, am(i,j)\right) \tag{2}$$

where $am(i,j)$ is a function that retrieves the codification in the sequence i at column j of the alignment, and the function $T_w(t, P_c)$ looks up the weight associated with the base t and the codification P_c (in this case P_c is given by $am(i,j)$) in table 3.

After counting, the column homogeneity of a given column is computed using the following formula,

Table 3: Base count weights matrix

t		A	C	G	T	R	Y	K	M	S	W	B	D	H	V	N	$-$
0	A	1	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	0
1	C	0	1	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	0
2	G	0	0	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$	$\frac{1}{4}$	0
3	T	0	0	0	1	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{4}$	0
4	$-$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

$$CH_j = \frac{\sum_{t=0}^{3} \left(wc_{jt}\right)^2}{\left(\sum_{t=0}^{4} wc_{jt}\right)^2} \tag{3}$$

It is to be noted that in the numerator of the fraction only the four bases are considered (A,C,G and T indexed by 0, 1, 2 and 3), and in the denominator the gap ($-$, indexed by 4) is considered along with the bases. This is considered in order to penalize the insertion of gaps, assuming that as the gaps are not counted in the numerator but they are counted in the denominator, the column homogeneity value decreases when there are more gaps.

In the case that a position in a sequence has a ? codification, that position for that sequence is discarded (as it was not observed) for the computing of that column homogeneity value. This is because a ? implies that in that position the sequence has no information.

An special consideration is taken when all the elements in a column are gap codifications ($-$), in that case the column homogeneity is given a value of zero, to penalize the existence of such columns.

When the column homogeneity value for all the columns has been computed, the mean value is obtained and that is the *Mean Column Homogeneity.*

This criterion gives higher scores to more homogeneous columns, penalizing diversity of bases in a column (as shown in the examples of table 4).

A.2 Reciprocal of Gap Blocks

The gap codifications ("$-$") which are contiguous are grouped into blocks (not counting initial and trailing gaps); and the reciprocal of the number of gap blocks is calculated, as shown in the next equation.

$$RGB = \frac{1}{GB} \tag{4}$$

where GB is the number of gap blocks in the alignment. If there are no gap blocks, the Reciprocal of Gap Blocks criterion is given a value of 1.0.

This criterion serves the purpose of rewarding the alignments where the gap codifications are located in a more concentrated manner, i.e. where there are fewer larger blocks of gap codifications rather than more blocks of smaller

Table 4: Column Homogeneity evaluation examples

	column												
	0	1	2	3	4	5	6	7	8	9	10	11	12
seq0	A	A	A	A	A	A	A	A	A	A	–	A	A
seq1	A	A	A	A	A	A	A	A	A	A	–	–	G
seq2	A	A	A	A	A	A	A	A	A	G	–	–	–
seq3	A	A	A	A	A	A	A	A	A	G	–	–	–
seq4	A	A	A	A	A	A	A	A	G	T	–	–	–
seq5	A	A	A	A	A	A	A	A	G	T	–	–	–
seq6	A	A	A	A	A	A	A	G	T	T	–	–	–
seq7	A	A	A	A	A	A	G	G	T	C	–	–	–
seq8	A	A	–	A	G	G	T	T	C	C	–	–	–
seq9	A	–	–	G	G	T	C	T	C	C	–	–	–
CH	1.00	0.81	0.64	0.82	0.68	0.66	0.52	0.44	0.28	0.26	0.00	0.01	0.02

length. Fewer blocks imply less evolutionary events to be explained, and a more parsimonious alignment.

In tables 5, 6 and 7 three alignments of an hypothetical set of sequences are shown. The three alignments have the same number of ”−“, but the example in table 5 has them in 3 blocks, the example in table 6 in 2 blocks and finally the example in table 7 in just 1 block, a difference which is noticeable in the reciprocal gap blocks criterion, and thus favoring the alignment which implies less evolutionary events (parsimony).

A.3 Columns Increment

Inserting gaps to align a set of sequences is common and the number of columns increases. *Columns Increment* is the ratio of this augmentation, defined by

$$CI = \frac{C}{C_0} - 1 \tag{5}$$

where C is the number of columns after aligning, and C_0 the number of columns before aligning, which is equivalent to the number of nucleotides of the longest sequence.

An example of a hypothetical set of sequences for which two different alignments are shown in tables 8 and 9 is given. Each alignment has a different value for the *Columns Increment* criterion. A smaller alignment is preferred because a smaller matrix probably implies less evolutionary events (parsimony).

Table 5: Alignment to exemplify the *Reciprocal of Gap Blocks* criterion. $RGB = 0.3\bar{3}$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
seq0	A	A	A	A	G	G	C	A	T	C	A	T	C	A	T	C	A	G	G	A	A	A	A
seq1	A	A	A	A	G	G	-	-	-	C	-	-	-	A	-	-	-	G	G	A	A	A	A

Table 6: Alignment to exemplify the *Reciprocal of Gap Blocks* criterion. $RGB = 0.50$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
seq0	A	A	A	A	G	G	C	A	T	C	A	T	C	A	T	C	A	G	G	A	A	A	A
seq1	A	A	A	A	G	G	-	-	-	-	-	-	C	A	-	-	-	G	G	A	A	A	A

Table 7: Alignment to exemplify the *Reciprocal of Gap Blocks* criterion. $RGB = 1.0$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
seq0	A	A	A	A	G	G	C	A	T	C	A	T	C	A	T	C	A	G	G	A	A	A	A
seq1	A	A	A	A	G	G	C	-	-	-	-	-	-	-	-	-	A	G	G	A	A	A	A

Table 8: Alignment to exemplify the *Columns Increment* criterion. In this case, the number of columns remain the same after aligning. $CI = 0$

	0	1	2	3	4	5	6	7	8
seq0	A	T	C	A	T	C	A	T	C
seq1	A	T	C	A	T	C	A	T	C
seq2	A	T	C	A	T	C	A	T	C

Table 9: Alignment to exemplify the *Columns Increment* criterion. Here, the number of columns increased to 6 after aligning. $CI = 0.6\bar{6}$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
seq0	A	T	C	A	T	C	-	-	-	A	T	C	-	-	-
seq1	A	T	C	-	-	-	A	T	C	A	T	C	-	-	-
seq2	A	T	C	-	-	-	-	-	-	A	T	C	A	T	C