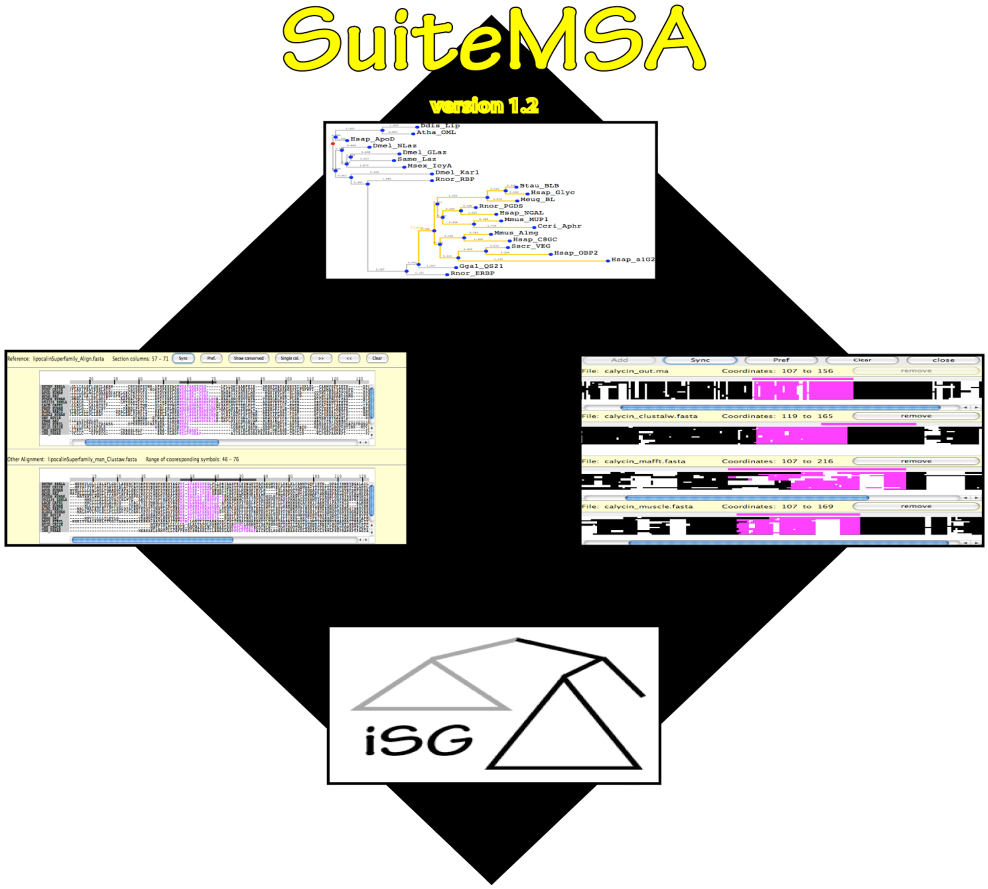


SuiteMSA User's Manual

SuiteMSA-1.2.06 release date: 6-14-2011
User's Manual last updated: 6-14-2011

Developed by Catherine L. Anderson

anderson@cse.unl.edu

Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, Nebraska

Table of Contents

+

1. What is SuiteMSA?

SuiteMSA is a java-based application that provides unique multiple sequence alignment (MSA) viewers. Users can directly compare multiple MSAs and evaluate where the MSAs agree (are consistent) or disagree (are inconsistent). **SuiteMSA** also provides a graphical user interface (GUI) for a sequence simulator, **iSGv2.1** (Strope *et al.* 2009), and allows visual inspections of MSAs and phylogenies. The MSA visualization tools (viewers) can be used with any MSA in FASTA format. These viewers are independent of the simulation program and do not require the installation of **iSGv2.1**. Once communication is set up between **SuiteMSA** and **iSGv2.1**, parameters can be configured and simulations can be launched from **SuiteMSA**. A simulation log tracks all simulations performed recording the information including parameters used, date and time stamps, and communications with **iSGv2.1**. Once the simulation is done, the true MSA and phylogeny with indel events mapped can be displayed.

Grant support. Development of **iSGv2** and **SuiteMSA** has been supported by:
NSF Assembly of Tree of Life (AToL) grant 0732863.

2. How to cite SuiteMSA

If you use **SuiteMSA** in any of your research or use any graphics obtained from **SuiteMSA** in your publication, please cite **SuiteMSA**. If you also use results simulated by **iSGv2.1**, please cite both of **SuiteMSA** and **iSGv2.1**.

For **SuiteMSA**:
Anderson, C. L., Strope, C. L., and Moriyama, E. N. (2011) SuiteMSA: Visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC Bioinformatics* **12**:184.

For **iSGv2.1**:
Strope, C. L., Abel, K., Scott, S. D., and Moriyama, E. N. (2009) Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol Biol Evol* **26**: 2581-2593.

3. System and software requirements

3.1 Operating System (OS)

SuiteMSA is written in Java. The programs (both **SuiteMSA** and **iSGv2.1**) have been tested on OS X 10.5.8 and 10.6.6 on Intel-processor Macintosh as well as on Linux. This release of **SuiteMSA** does not run on Windows.

3.2 Java Runtime Environment (JRE)

If you have been diligently updating your system software, your OS should already have the most recent compatible JRE. We have not conducted extensive backward compatibility testing for **SuiteMSA**. If you experience any problem with **SuiteMSA** and if an older version of JRE is suspected, you can download the most recent compatible JRE from

http://www.apple.com/support/ for Macintosh

http://www.java.com/en/download/index.jsp for Linux.

4. Installing SuiteMSA and related programs

4.1 Downloading programs

All programs are available from **SuiteMSA** website:
http://bioinfolab.unl.edu/~canderson/SuiteMSA/

i) For Mac OS X 10.5 or higher, download **"SuiteMSA_Package_forMacOSX.dmg"**. Double-click on the dmg file, and drag **"SuiteMSA_Package"** folder to your computer.

ii) For other OS, you need to download each package separately.

- Download **"SuiteMSA_Package.zip"**, and double-click the zip file to uncompress.
- Download **"indel-seq-gen-2.1.03.tar.gz"** or **"indel-seq-gen-2.1.03.zip"** file. Depending on the OS, you can double-click the file to uncompress or on Linux-like OS, type the following two commands:

```
gunzip indel-seq-gen-2.1.03.tar.gz
tar -xvf indel-seq-gen-2.1.03.tar
```

- To use **ClustalW2** with **SuiteMSA** GUI, you need to download **ClustalW2 v2.1** from **ClustalW2** website (http://www.clustal.org/download/current/). Depending on the OS, follow their instructions. Note that you need to download **ClustalW**, not **ClustalX**.
- To use **Muscle** with **SuiteMSA** GUI, you need to download a muscle package from **Muscle** website (http://www.drive5.com/muscle/downloads.html) depending on the OS, follow their instructions.

4.2 Installing programs

If you downloaded the **"SuiteMSA_Package_forMacOSX.dmg"**, you can skip the following steps.

4.2.1 Installing SuiteMSA

Nothing needs to be done for **SuiteMSA**. The executable is **"SuiteMSA-1.2.jar"** located in **"SuiteMSA-1.2"** folder. Do not move or delete the executable and other files in this folder. You can create an alias of the jar file and put the alias wherever you like for your convenience.

4.2.2 Installing iSGv2.1

To set up **iSGv2.1** in Linux-like environment (use Terminal in MacOS X), change the directory to **"indel-seq-gen-2.1.03"** directory and type the following two commands:

```
./configure
make
```

This should create **"indel-seq-gen"** executable file in **"src"** directory. Try to run it by typing:

```
      ./indel-seq-gen
```

You should see the message:

```
      You must specify the substitution model (-m).
```

If you see any error message, consult the iSG website for trouble-shooting (http://bioinfolab.unl.edu/~cstrope/iSG/).

NOTE: If `./configure` does not succeed (you see error messages and indel-seq-gen executable cannot be generated), it is possible that you have file/directory names with irregular characters (', ", *etc.*). Please check the directory path and try to remove any irregular characters.

4.2.3 Installing ClustalW2

Follow the instruction provided by **ClustalW2**. If you downloaded the binary (executable), nothing needs to be done. To confirm if **ClustalW2** is properly installed in Linux-like environment, change the directory to **"clustalw-2.1-macosx"** (or similarly named) directory and type the following command:

```
      ./clustalw2
```

You should see the **ClustalW** menu. If you see any error message, consult the Clustal website for trouble-shooting (http://www.clustal.org/).

4.2.3 Installing Muscle

Follow the instruction provided by **Muscle**. If you downloaded the binary (executable), nothing needs to be done. To confirm if **Muscle** is properly installed in Linux-like environment, change the directory to where **Muscle** is installed and type the name of the executable, for example:

```
      ./muscle3.8.31_i86darwin32
```

You should see the **Muscle** help. If you see any error message, consult the **Muscle** website (http://www.drive5.com/muscle/).

5. Start using SuiteMSA

5.1 Input files

Three types of input files are used with **SuiteMSA**.

i) *MSA Viewer*, *MSA Comparator*, and *Pixel Plot* require alignments in FASTA format. An example alignment in FASTA format is shown below.

```
>Dmel_Karl
YQTRRRSGPSNRCPKVGAIKNFDLERMMGCWHVVQYYASTELP---------EYACMRSH
FSFSKEDQHITMNFSYIFAEDPL--RKLVGNITWMIPKFQEPGHWQHTEDIYEG-----I
YNTYVLDTDYDTWGVMHCAEKKK--QPRYLSALLLSRKTSLADNEISFLRGKLPQD-IDT
SF-MFNIGQESCDNLMESSRDDPLAYV
>Meug_BL
------------VENIRSKNDLGVEKFVGSWYLREAAKT-------------MFSIPLFD
MDIKEVNLTPEGNLELVLLEKA-----DRCVEKKLLLKKTQKPTEFEIYISSES----AS
YTFSVMETDYDSYFFCLYNISDR----EKMACAHYVRRIE-ENKGMNEFKKILRTLAMPY
TV-IEVRTRDMCHV-------------
```

Sequence names including a space character:
Although using ' ' (space) in sequence names is allowed in **SuiteMSA**, many programs including **ClustalW** ignore any characters after the first space character in the sequence name. For example, "Taxon 1" is recognized as "Taxon" in these programs. In order to avoid such unintentional name changes, avoid using the space character in sequence names.

Irregular characters in sequences:
SuiteMSA allows the use of hyphens (-) and question-marks (?) in aligned sequences. Any other irregular characters will be shown as a solid black box.

ii) *Phylogeny Viewer* requires a phylogeny in Newick format. An example of a Newick format tree is shown below.

```
(((Ddis_Lip:1.0093059,Atha_OML:0.8969969):1.3570963,Hsap_ApoD:
0.3447207):0.0852,((Dmel_NLaz:0.8008651,((Dmel_GLaz:1.4075732,
Same_Laz:1.2573009):0.1178207,Msex_IcyA:1.6703565):0.1439733):
0.1856264,(Dmel_Karl:2.3475366,(Rnor_RBP:1.8656970, ((((((Btau_BLB:
0.3160180,Hsap_Glyc:0.6381421):0.5326599,Meug_BL:0.8755637):
1.4551225,((Rnor_PGDS:0.4357870,Hsap_NGAL:1.0253870):0.5292846,
(Mmus_MUP1:0.7890084,Ccri_Aphr:1.7487024):0.9120180):0.1651272):
0.1007552,((Mmus_A1mg:0.7808197,Hsap_C8GC:1.3278706):0.7052975,
((Sscr_VEG:0.6334171,Hsap_OBP2:1.8798683):0.8045697,Hsap_a1G2:
4.3324329):0.5309832):0.1734662):0.4519230,Ggal_QS21:1.0870960):
0.3539863,Rnor_ERBP:1.1812067):1.1271613):0.4814631) :0.4631459):
0.0852);
```

iii) *iSG Simulation* requires several input support files. See the section 9 for these files.

5.2 Sample files

The following sample files are included in the **"sample"** folder of the **SuiteMSA** distribution package:

lipocalinSuperfamily_seqs.fasta (unaligned sequences in FASTA format)
lipocalinSuperfamily_alignment.fasta (manually aligned sequences in FASTA format)
lipocalinSuperfamilyPhylogeny.tree (a phylogeny in Newick format)
lipocalinSuperfamily.tree (a guide tree file used for iSGv2 simulation)
lipocalinSuperfamily.spec (a lineage specification file used for iSGv2 simulation)
lipocalinSuperfamily_template.maroot (a MSA root file used for iSGv2 simulation)
lipocalin.freq (an amino acid frequency file used for iSGv2 simulation)
lipocalin.idlen (an indel length-distribution file used for iSGv2 simulation)
lipocalinSuperfamily_SecondaryStructureS.fasta (secondary structure prediction for each sequence in alignment).

MSAs generated by **ClustalW2** (Larkin *et al.* 2007), **Muscle** (Edgar 2004), and **MAFFT** (Katoh and Toh 2008) (in FASTA format) are also available in this folder. More sample files are available in the **"additionalSamples"** folder.

See Anderson *et al.* (2011a and 2011b) for more description on these sample files and the examples of using these samples with SuiteMSA.

5.3 Running SuiteMSA

Double-click on the executable file, **SuiteMSA-1.2.jar**, in the **"SuiteMSA-1.2"** folder. It brings up the **SuiteMSA** main window listing the six tools available as shown in **Figure 5.1**. Click on a button to start any tool. See individual manual section for detailed description of each tool.

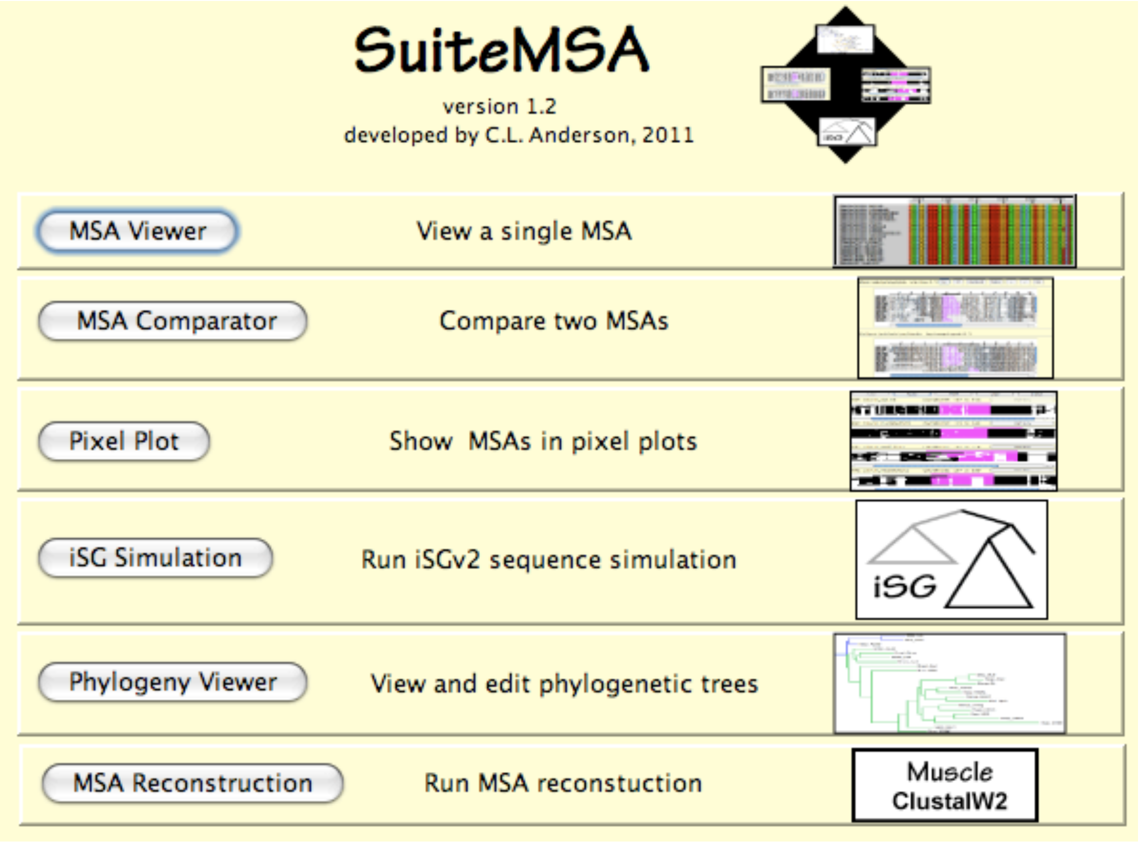
5.4 How to increase the memory available for SuiteMSA

The default size of the memory available for a java application is fairly small (about 32mb). Dealing with big alignments uses a large amount of memory. If you experience slowness or difficulty in using *MSA Viewer* or *MSA Comparator*, you need to increase the memory allocated to **SuiteMSA**.

On Linux or Mac OS X (using Terminal), use the following command line to start the program:

```
    java -Xms350m -Xmx350m -jar SuiteMSA-1.2.jar
```

Replace **"SuiteMSA-1.2.jar"** with the correct name of the SuiteMSA executable file. This increases the **SuiteMSA**'s memory (heap) size to 350mb. How much memory you can allocate to the program depends on how much free memory is available on your computer.



SuiteMSA

version 1.2
developed by C.L. Anderson, 2011

View a single MSA

Compare two MSAs

Show MSAs in pixel plots

Run iSGv2 sequence simulation

View and edit phylogenetic trees

Run MSA reconstuction

Figure 5.1. The SuiteMSA main window.

6. *MSA Viewer*

6.1 Viewing a single MSA: *MSA Viewer*

The *MSA Viewer* allows the user to view any single MSA in FASTA format. This viewer is available from the **SuiteMSA** main screen as well as from the **Alignments** menu of the **ClustalW2** and **Muscle** GUIs. After selecting an MSA file, the *MSA Viewer* window as shown in **Figure 6.1** will be displayed.

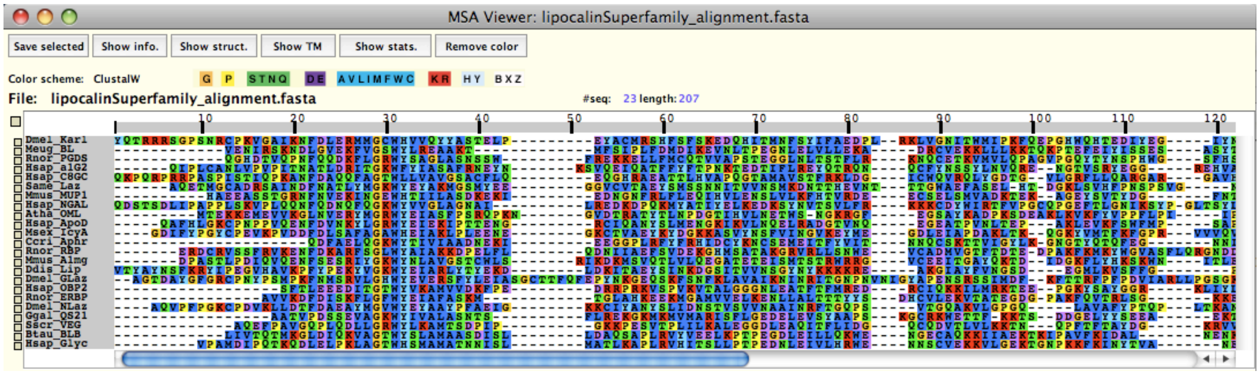
The viewer recognizes if the alignment is for nucleotide or amino acid sequences and chooses the appropriate color-coding automatically. This is done based on the composition of the letters included in the sequences. If A, T, U, G, and C compose more than 50% of the sequences, they are considered to be nucleotide sequences. Letters allowed for ambiguity codes are as follows:

- Ambiguous nucleotide letters allowed: N, R, Y, K, M, S, W, B, V H, D and X
- Ambiguous amino acid letters allowed: B, X, and Z.

Note that the viewer shows all letters in the sequences in upper cases even if the input file uses lower cases in the sequences. Note also that, as described in the section 5.1, irregular characters other than hyphens (-) and question-marks (?) are represented as black squares.

6.1.1 Color schemes used for nucleotide and amino acid letters

The *MSA Viewer* offers several color schemes for the nucleotide and amino acid letters within the MSA. The name of the color scheme along with the color key in current use in the display is shown immediately above the file name of the MSA. To change the color scheme, choose **Set color scheme ...** from the menu **Display options**. This brings up a color choice window (shown in **Figure 6.2**). The choices available depend on the type of sequences in the MSA. For amino acids, the color-coding is mainly based on their grouping of physico-chemical properties. The choices are Clustalw, Taylor, Zappo, Helix

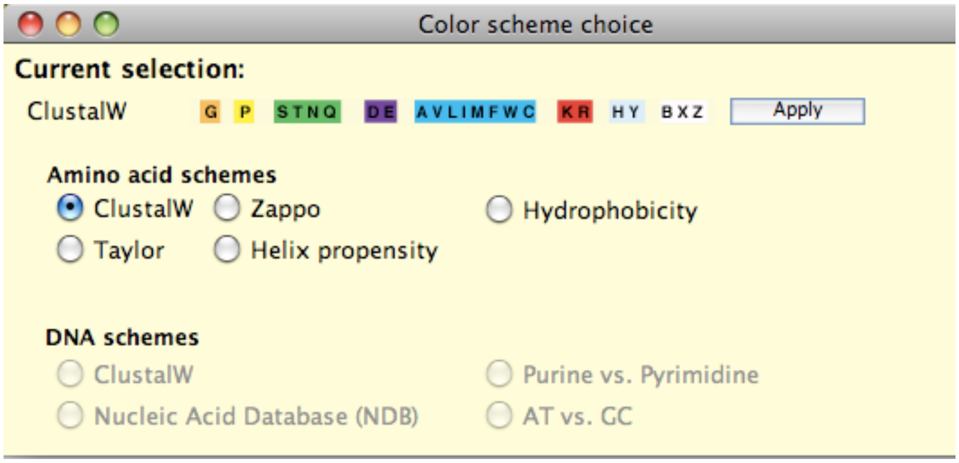


Color scheme: ClustalW

File: lipocalinSuperfamily_alignment.fasta

#seq: 23 length: 207

Figure 6.1: The *MSA Viewer*. The alignment for the lipocalin superfamily is displayed in this example. The sample file (lipocalinSuperfamily_alignment.fasta) is included in the **"sample"** folder.



Color scheme choice

Current selection:

ClustalW G P STNQ DE AVLIMFWC KR HY BXZ

Amino acid schemes

● ClustalW ○ Zappo ○ Hydrophobicity

○ Taylor ○ Helix propensity

DNA schemes

○ ClustalW ○ Purine vs. Pyrimidine

○ Nucleic Acid Database (NDB) ○ AT vs. GC

Figure 6.2: The color-scheme selection window. The available color-scheme options and the colors used for each letter are shown.

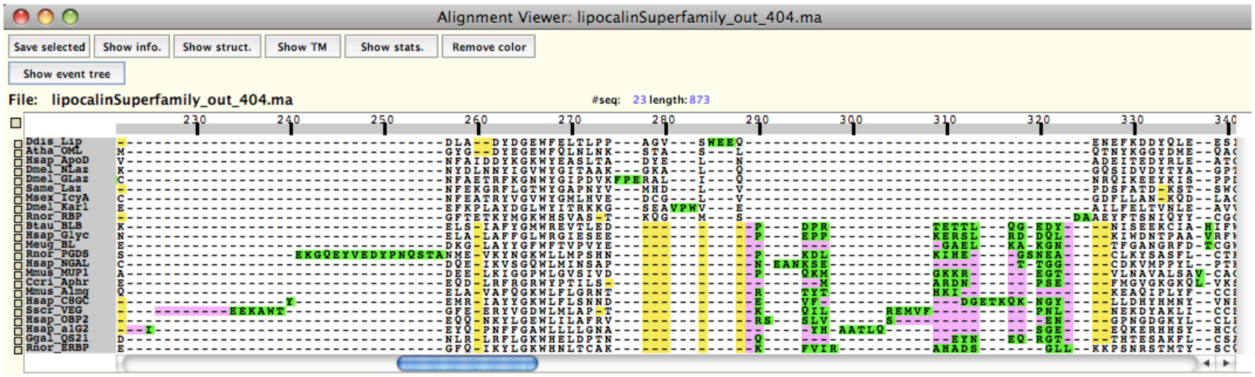
propensity (Proctor *et al.* 2010) and Hydrophobicity (Kyte and Doolittle, 1982). For nucleic acids, the choices include ClustalW, Nucleic Acid Database, as well as Purine *vs.* Pyrimidine and AT *vs*. GC. Letters used for ambiguity codes are colored light gray regardless of the color-coding scheme.

To show the MSA in black and white without colors, click on the **Remove color** button. To restore the color scheme last used, click on the **Restore color** button.

6.2 Viewing a simulated MSA: *Alignment Viewer*

A variation of the *MSA Viewer*, called the *Alignment Viewer*, is available in the **Alignments** menu in the *iSG Simulation* (the *iSG Simulation* tool is explained in the section 9). The *Alignment Viewer* works very similar to the *MSA Viewer*. The difference is if the MSA is from a simulated data set created by **iSGv2.1** and the event trace file (xxx.trace) is present, the color-coding in the display represents the type of indel events that have taken place. As shown in **Figure 6.3**, insertions are colored in green and deletions are colored in yellow. In sites where an insertion occurred followed by a deletion the resulting gap is colored in pink. Clicking on a gap symbol (-) brings up the event label. Note that the current version of **iSGv2.1** (2.1.03) does not track substitution events.

Use the **Show event tree** button (unique to the *Alignment Viewer*) to open the guide tree used in the simulation with the *Phylogeny Viewer*. It displays the guide tree with the indel events mapped. For more on the *Phylogeny Viewer*, see the section 10.



File: lipocalinSuperfamily_out_404.ma #seq: 23 length: 873

```
                   230       240       250       260       270       280       290       300       310       320       330       340
Ddis_Lip   ------------------------------------DLA--DYDGEWFELTLPP---AGV---SWEEQ-------------------------------------ENEFKDDYQLE--ES
Atha_OML   M----------------------------------GYG--DYEGEWFQLNLNK---STA---S---L-------------------------------------QTNYKGGYDME--QA
Hsap_ApoD  V----------------------------------NFAIDDYKGKWYEASLTA---DYE---L---N-------------------------------------ADEITEDYRLE--AT
Dmel_NLaz  K----------------------------------NYDLNNYIGVWYGITAAK---GKA---L---Q-------------------------------------GQSIDVDYTYA--GP
Dmel_GLaz  C----------------------------------NFAETRFKGNWYGIPDVKFPERAL---I---Q-------------------------------------NRQIKEEYKIS--PP
Same_Laz   -----------------------------------NFEKGRFLGTWYGAPNYV---MHD---L---V-------------------------------------PDSFATD-KST--SW
Msex_IcyA  C----------------------------------NFEATRYVGVWYGMLHVE---DCG---L---V-------------------------------------GDFLLAN-KQD--LA
Dmel_Karl  E----------------------------------EFKPLAYDGLWYITRKKG---SEAVPWV---E-------------------------------------AILFELTVNLE--AV
Rnor_RBP   -----------------------------------GFTETKYMGKWHSVAS-T---KQG---M---S-----------------------------------DAAEYFTSNIQYY--CG
Btau_BLB   K----------------------------------ELS-IAFYGMWREVTLED--------------P----DPR-----------TETTL---QG-EDY-----NISEEKCIA-HIF
Hsap_Glyc  N----------------------------------ELA-LAFFGLWRGIESEE--------------P----EPP-----------KERSL---RD-DQL-----KIWDNTPAA-VRF
Meug_BL    E----------------------------------DKG-LAYYGFWFTVPVYE-------------------------------------GAEL---KA-KGN-----TFGANGRFD-TCG
Rnor_PGDS  S------------------EKGQEYVEDYPNQSTANME-VKYNGKWLLMPSHN--------------P----KDL-----------KIHE-----GSNEA-----CLKYSASFL--CT
Hsap_NGAL  C----------------------------------DQE-IKVSGQWLMINSAP--------------N-EANKSE--------------------T-TGG-----CDKVMPPYL--PT
Mmus_MUP1  A----------------------------------DEE-LKIGGPWLGVSIVD--------------P----QKM-----------GKKR-------EGT-----VLNAVALSAV-CA
Ccri_Aphr  E----------------------------------EQD-LRFRGRWYPTILS-----------------------M-----------ARDN-------PSE-----FMGVGKGKQL-VK
Mmus_Almg  Q----------------------------------ELA-VAFQGKWLFLGRNT--------------R----TYT-----------HKI----------------KEAQIPLYF--CC
Hsap_C8GC  -----------------Y----------------EMR-IAYYGKWLFLSNND--------------E----VF--------------DGETKQK-NGY-----LLDHYHMNY--VN
Sscr_VEG   -----------EEKAWT-----------------GFE-ERYVGDWLMLAP-T--------------K----QIL------REMVF-----------PNL-----NEKDYAKLI--CC
Hsap_OBP2  -----------------------------------EQQ-NKYLGEWLILAFRV--------------RS---SLV------S----------------EN-----GPNGDGKYL--CL
Hsap_a1G2  ---I-------------------------------EYQ-PNFFGAWLLLLGNA---------------------YH-AATLQ----------------SGE-----EQKERHHSY--HC
Ggal_QS21  D----------------------------------NLR-LRFLGKWHELDPTN--------------Q--------------------EYN---EQ-RGT-----THTESAKFL--CS
Rnor_ERBP  E----------------------------------GFQ-IKYLGKWHNLTCAK--------------K----FVIR----------AHADS-------GLL--KKPSNRSTMTY--SC
```

Figure 6.3: The *Alignment Viewer*. A true alignment based on the lipocalin superfamily simulation is displayed in this example. The insertions and deletions are colored in green and yellow, respectively. The pink gaps indicate sites where an insertion occurred followed by a deletion in the same site. The sample output files used (lipocalinSuperfamily_out_404.ma and lipocalinSuperfamily_out_404.trace) are included in the **"iSG_sample_output"** folder.

6.3 Single alignment statistics

The *MSA Viewer* provides statistics to help characterize an individual alignment. These are described in this section.

6.3.1 Information content

The information content is an indication of the conservation of each alignment column. As shown in **Figure 6.4**, it is displayed in a bar chart beneath the MSA. Each bar shows the information content of the alignment column. The completely conserved alignment columns are indicated with full-height bars (the maximum bar height depends on the type of sequences: maximum value is 2.0 for DNA and 4.32 for amino acids). In **Figure 6.4**, positions 29 and 31 are completely conserved and show the maximum information content. To display the information content bar chart, click on the **Show info.** button. To remove it, click on the **Remove info.** button**.**

The information content is calculated based on the Shannon entropy (Schneider and Stephens, 1990). Refer to Anderson *et al.* (2011a) for the details of the method.

6.3.2 Alignment Statistics

Various statistics calculated for the alignment are displayed by clicking on the button **Show stats.** They are shown at the top of the *MSA Viewer* window (**Figure 6.4**). To remove the display, click on the button **Remove stats.**. The statistics calculated are as follows:

- **% gaps**. The percent of sites that contain gaps. It is calculated by: (number of gap symbols)/(number of sequences x alignment length).
- **% conserved.** The number of columns fully conserved divided by the total number of columns.

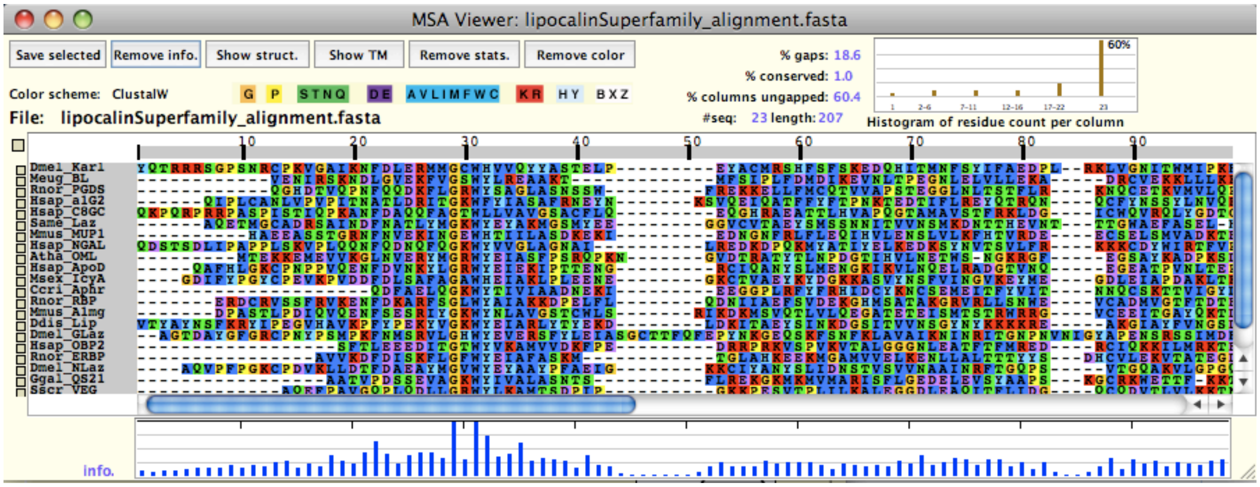


Figure 6.4: The *MSA Viewer* with information content and statistics. The alignment for the lipocalin superfamily is displayed in this example along with the information bar chart and the statistics. The fully conserved columns in positions 29 and 31 show the maximum information content. The sample file (lipocalinSuperfamily_alignment.fasta) is included in the **"sample"** folder.

- **% columns ungapped.** The number of ungapped columns divided by the total number of columns.
- **Histogram of residue count per column.** This shows the frequency distribution of gapped and ungapped (no-gap) columns within the MSA. While the sixth, right-most, column shows the frequency of the alignment positions that have no gap, the first, left-most, column shows the frequency of the alignment positions that are occupied by all gaps but one non-gap character. Four columns in-between show the frequencies of alignment positions <24% (excluding single-character sites), 25-50%, 51-75%, and >76% (excluding 100%) filled with non-gap characters.
- **# seq.** The number of sequences in the MSA (always displayed).
- **Length.** The number of columns in the MSA (always displayed).

6.4 Secondary structure and transmembrane prediction display

SuiteMSA allows for the display of both secondary structure and transmembrane predictions to help evaluate how well these structures are aligned in the MSA. The input files for both these displays must be in FASTA format. Note that in the current version of SuiteMSA, white space characters (*e.g.*, space, tab) are not allowed within the structural sequence.

6.4.1 Secondary structure display

Secondary structure information for both RNA and protein sequences can be displayed. Symbols used for secondary structures are as follows:

i) Secondary structures for proteins. Alpha-helices are represented by H (or h) and in green color. Beta-strands are represented by E (or e) and in brown color. Random coils, loops, and turns are represented by C (or c), T (or t), and S (or s) in cream, tan, and pink, respectively. An example FASTA format of a protein secondary structure is shown below.

```
>HV1J3
CEEEEEECCCCCCCCCCEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEECCCCCCC
CCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCEEEECCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCEEEECCCCCCHHHHHHHCEECCCEEECCCCCCCCCCCCCCCCEEEECCCCCHHC
CCCCCCCCCCCCCCCCCCCCCEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCEEEEECCCCCCCCCEEEEECCCCEEEECCCCCCCCCCCCCCCCCCCEEECCCCCCH
```

ii) Secondary structures for RNA sequences. Nested parenthesis format is used. Base pairs are represented by (), <>, [], or {}. Unpaired regions are indicated with '.', ',', and ';'. An example FASTA format of an RNA secondary structure is shown below.

```
>Toxoplasma_gondii
...(((((...[[[.))))).(.((.(((((((.(((((((((..(..(((.(((...((
.....(((.....(.(.............))....)))))........(((.......((
((..((..((.....((((......................................(((
(....(((((((.....))))))).....)))).......((((.(.((((((.(((((.
........))))))))))...))))))....((((((........(((............
```

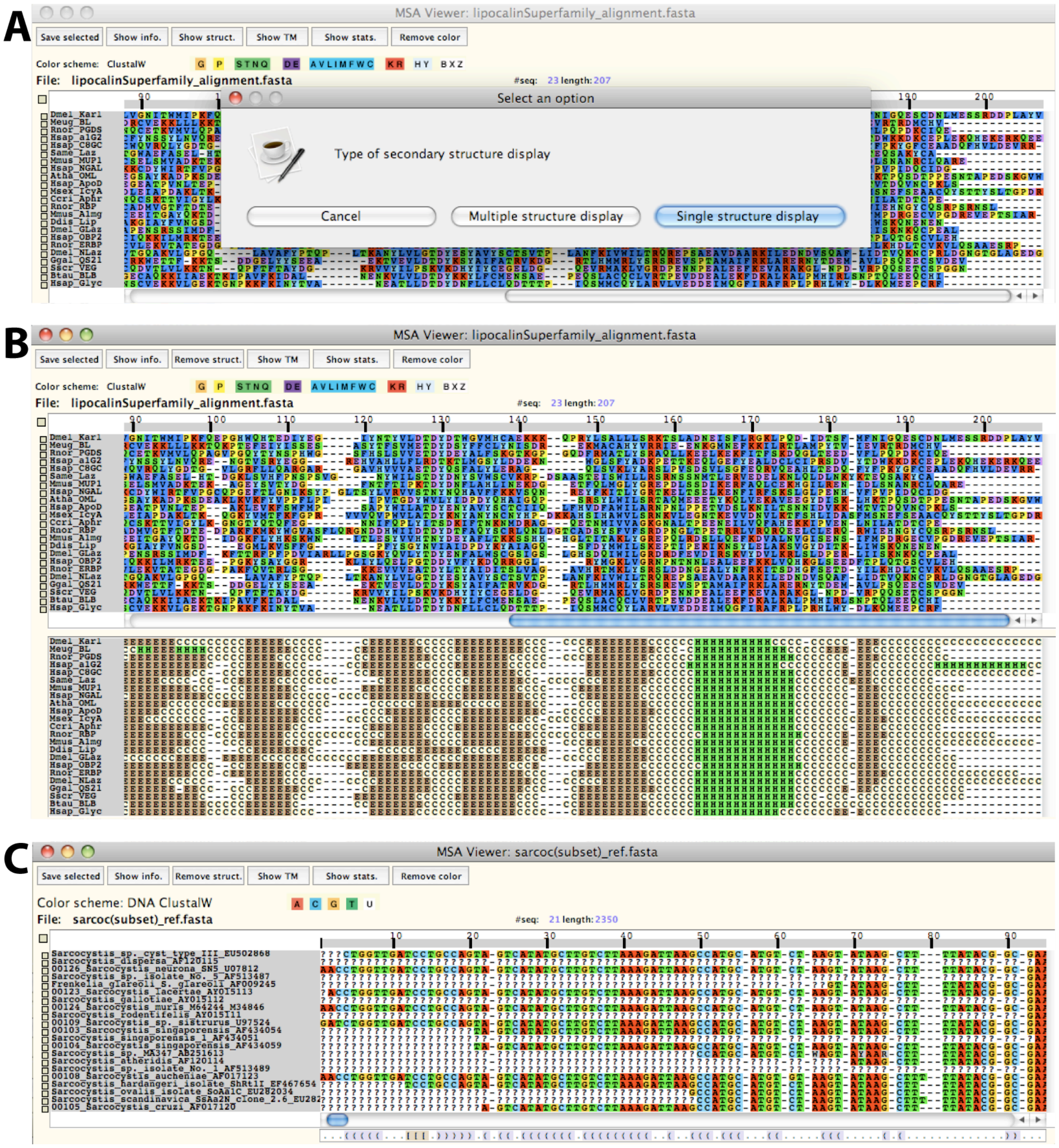
To display a secondary structure, click on the **Show struct.** button. To remove the structure display, click on the **Remove struct**. button.

Two types of secondary structure display options are available (shown in **Figure 6.5A)**:

i) Multiple structure display. This is shown in **Figure 6.5B**. Use this option when each sequence has its own prediction data. The *MSA Viewer* configures the secondary structure display aligned along with the MSA displayed. The structure file has the following requirements:

- The sequence IDs used in the structure file must be identical to those used in the sequence file. This includes cases (upper or lower). Its order is not critical as the structure sequences will be sorted to the same order as the MSA they represent.
- There must be one structure sequence for each sequence within the MSA.
- The structure sequences should not include any gap or irregular symbols.
- Each structure sequence must have the same number of characters as the corresponding sequence (excluding gap symbols).

ii) Single structure display. **Figure 6.5C** shows this display. Use this option when one secondary structure sequence is used as a representative (*e.g.,* consensus) for the whole alignment. Such a representative structure sequence could contain gaps, and the length must be the same as that of the alignment. If multiple structure data are included in the input file, the first structure data will be used as the representative.



Color scheme: ClustalW

File: lipocalinSuperfamily_alignment.fasta #seq: 23 length: 207

Type of secondary structure display

Color scheme: DNA ClustalW

File: sarcoc(subset)_ref.fasta #seq: 21 length: 2350

Figure 6.5: The *MSA Viewer* with its secondary structure display. A) The popup window asking for the type of display: multiple or single structure. B) The lipocalin superfamily MSA is displayed along with their secondary structures predicted by PSIPRED (McGuffin *et al.* 2000). The sample files, lipocalinSuperfamily_alignment.fasta and lipocalinSuperfamily_SecondaryStructureS.fasta, are included in the **"sample"** folder. C) The 18S rDNA MSA is displayed with a single representative secondary structure at the bottom. The sample files (sarcoc(subset)_ref.fasta and sarcoc_toxo_SecondaryStructure.fasta) are available in the **"additionalSample/rDNA"** folder.

6.4.2 Transmembrane structure display

The *MSA Viewer* allows for the display of transmembrane (TM) structural information aligned with the corresponding MSA (see **Figure 6.6**). Letters and symbols used to identify amino acids located in various structural regions are as follows. Amino acids located in TM regions are represented by X, x, H, or h. Amino acids located in extracellular loop regions are represented by O, o, or -, and those located in internal (cytoplasmic) loop regions are represented by I (capital), i, or +. An example FASTA format of a TM protein structure is shown below. Please note that the character “-“ from the TM sequence is changed to an equal sign (=) in the display to distinguish it from the gaps of the alignment.

```
>GP01_09_ptafr_human
----------------OOOOOOXXXXXXXXXXXXXIIIIII+++++++++++IIIIIXXXXXXXXXXXX
XXOOOOOO-----------OOOOOOXXXXXXXXXXXXXIIIIII+++++++++++++++++++IIIIII
XXXXXXXXXXXXOOOOOO-----------------------------OOOOOOXXXXXXXXXXXXXIII
III++++++++++++++++++++IIIIIIXXXXXXXXXXXXXOOOOOO------------------OOO
OOXXXXXXXXXXIIIII+++++++++++++++++++++++++++++++++++++++++++++++++
```

To open the TM structure display, click on the **Show TM** button. To remove the display, click on the **Remove TM** button. The TM structure display is available only in the

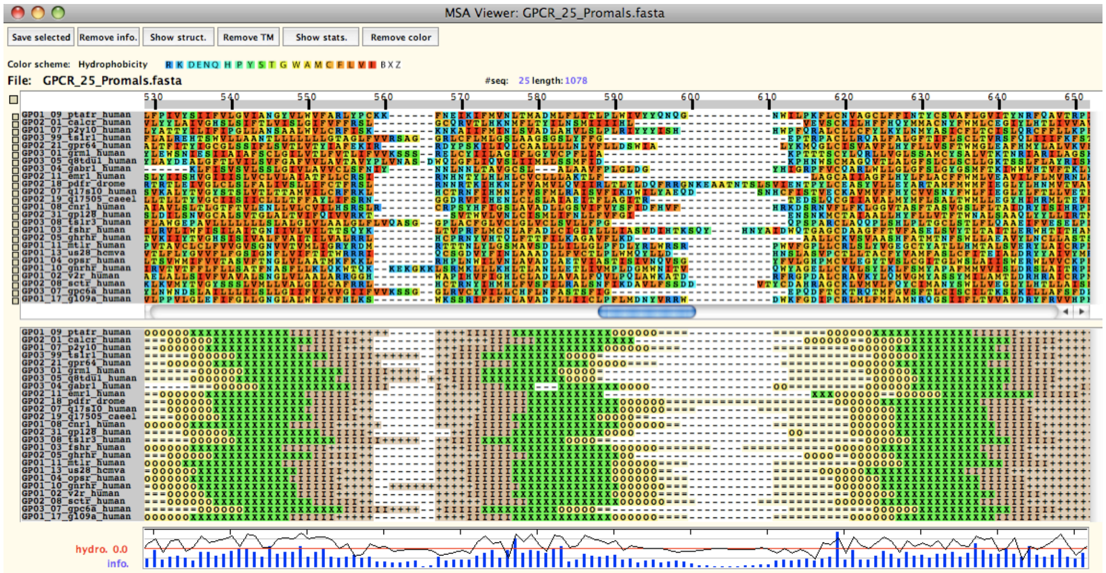


Figure 6.6: The *MSA Viewer* with transmembrane structure prediction. The MSA of G-protein coupled receptors produced by PROMALS (Pei and Grishin, 2007) is displayed along with their transmembrane structures predicted by MEMSAT3 (Jones, 2007). The transmembrane regions are colored green, the inner cap residues are dark brown, the inner loop residues are tan, while the outer cap residues are yellow and the outer loop residues are cream. Note that the color scheme “Hydrophobicity” is used for the MSA, and the average hydrophobicity plot is overlaid on the information content bar chart at the bottom. The sample files (GPCR_25_Promals.fasta and GPCR_25_TransMembranePrediction.fasta) are available in the **"additionalSamples/GPCR"** folder.

multiple structure display mode. See the previous section for the requirements for the input file.

6.4.3 Average hydrophobicity plot

If the color scheme "Hydrophobicity" is selected for a protein MSA, the average hydrophobicity for each alignment position can be plotted. As shown in **Figure 6.6**, the plot is overlaid on the information content bar chart. Use the **Show info**. button to display it. The hydrophobicity index for each amino acid used was obtained from (Kyte and Doolittle, 1982)

6.5 Other utilities available in the *MSA Viewer*

The *MSA Viewer* provides several utilities for convenient sequence manipulation. These include extracting a subset of sequences from an MSA and sorting sequences within an MSA.

6.5.1 Extracting a subset of sequences from an MSA

A selection box is located to the left of each sequence in the MSA displayed (see for example, **Figure 6.6**). To extract a subset of the sequences in the MSA, select the sequences by clicking on each of the corresponding boxes. For convenience, to select/deselect all sequences, click on the larger box at the top. Once sequences have been selected, click on the **Save Selected** button. This will bring up a file saving dialog window to save the selected sequences (in aligned FASTA format). After the file has been saved, you will have the option to display the subset of the alignment in a new *MSA Viewer* window. The original *MSA Viewer* window including the entire alignment will still be accessible.

6.5.2 Sorting sequences within an MSA

The order of sequences within an alignment can be different depending on the method used. In order to compare two different alignments using two *MSA Viewer* windows, for example, having the sequences in the same order is useful. This can be achieved by using the Sort MSA utility using the following steps:

1. Find the alignment where sequences are in the desired order.
2. Open this alignment using the *MSA Viewer*. We call this alignment MSA1.
3. Choose the **Sort MSA ...** item from the **Data** menu. This will open a file selection window. Select the alignment file you wish to sort its sequences to the same order as those in MSA1. Note that the two MSAs must have the same set of sequences (the same number of the sequences with exactly the same names).
4. Enter a new file name for the file the sorted alignment will be saved in.
5. This new file now contains the MSA where the sequences are ordered according to the order found in MSA1.

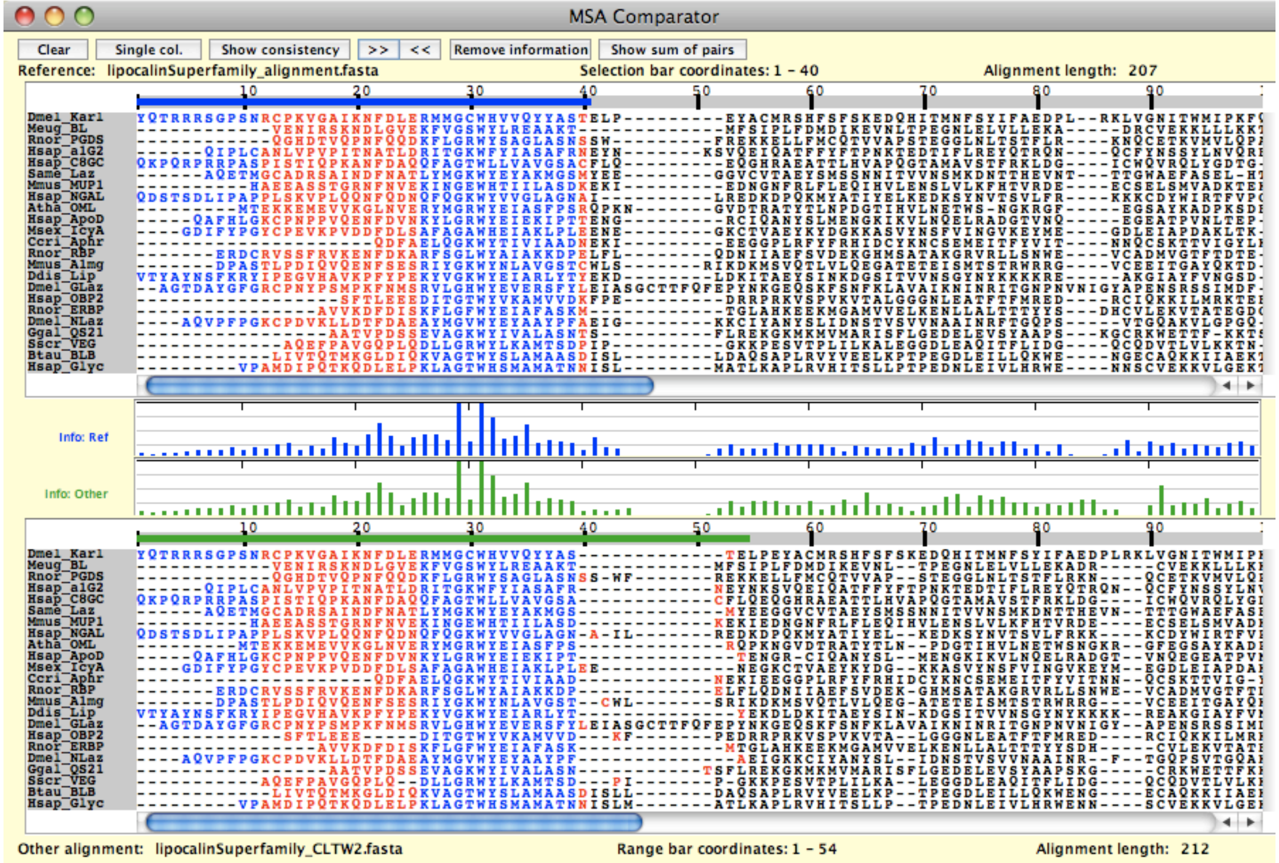
7. *MSA Comparator*

7.1 Comparative viewing of two MSAs: *MSA Comparator*

The *MSA Comparator* allows the user to compare two MSAs containing the same sequences (**Figure 7.1**). The *MSA Comparator* is available from the **SuiteMSA** main screen as well as from the **Alignments** menu in the *iSG Simulation,* the *ClustalW2* GUI, and the *Muscle* GUI.

In order to view two MSAs using the *MSA Comparator*, there are a few requirements:

- Sequence names (including upper or lower cases) must be identical between the two MSAs (including non-alphanumeric characters such as '_' or '-').



Reference: lipocalinSuperfamily_alignment.fasta

Selection bar coordinates: 1 - 40

Alignment length: 207

Other alignment: lipocalinSuperfamily_CLTW2.fasta

Range bar coordinates: 1 - 54

Alignment length: 212

Figure 7.1: The *MSA Comparator*. This example uses the alignment of the lipocalin superfamily proteins (lipocalinSuperfamliy_ailgnment.fasta) as the reference alignment (the top alignment), and compares it with the alignment reconstructed by **ClustalW2** (the second alignment). The blue selection bar is set to cover 40 alignment positions of the reference alignment. The green range bar above the second alignment shows the column range that covers the sequence positions selected in the reference alignment. The highlighted areas under the selection and range bars in the alignments show the sequence positions either in agreement (consistent) in blue, or else in red. The maximum height on the information content graphs is shown in the completely conserved columns in positions 29 and 31 of both alignments.

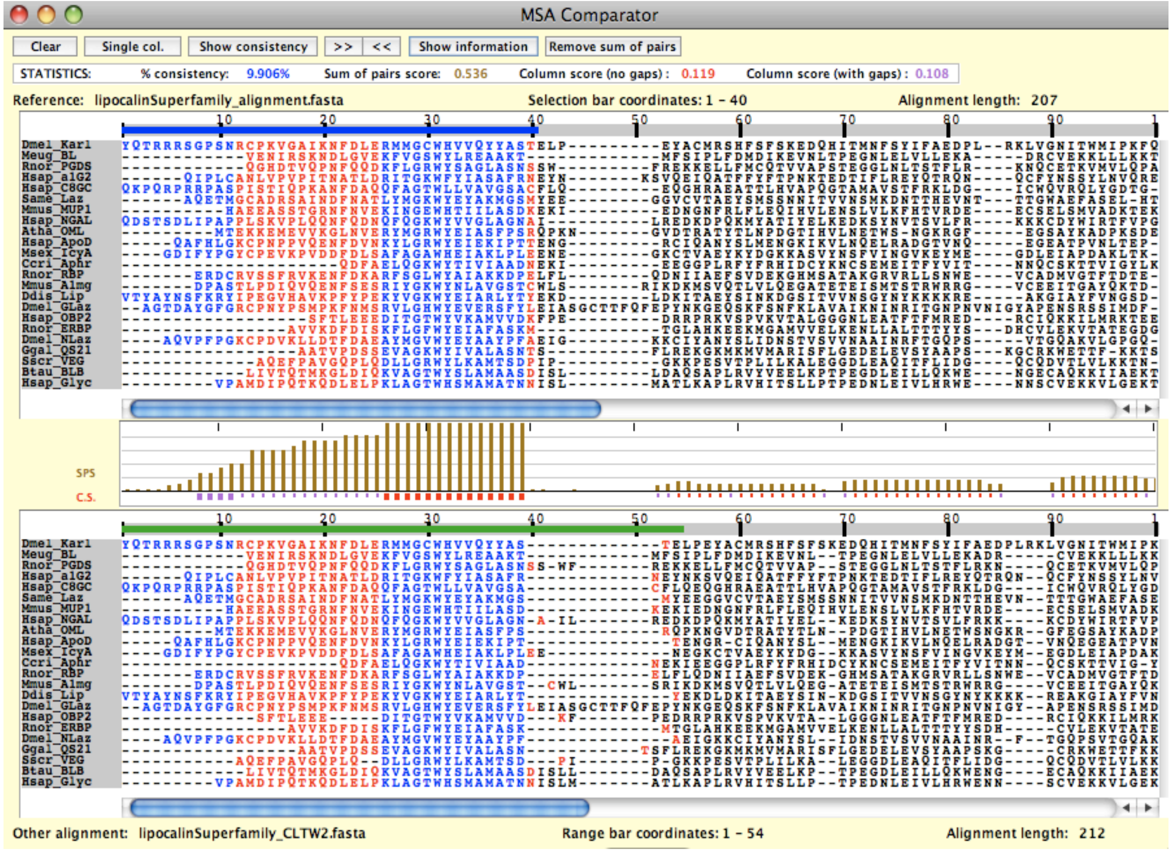
- Sequences in both MSAs must be identical. They may differ only in the location of the gaps.
- Only aligned sequences can be compared. All sequences in each MSA must contain the same number of sites. White-space characters (space, tab, line breaks) will be ignored and not counted for the alignment length.

The first MSA selected is used as the "reference" MSA, and the second MSA selected is compared against it. The order of the sequences included in the reference MSA will be used to display both MSAs. If the order of the sequences in the second MSA is different from the reference MSA, the sequences will be sorted to match the order of the reference MSA.

The blue selection bar above the reference alignment indicates the region selected for comparison. The green range bar above the second alignment shows the column range that covers the sequence positions selected in the reference alignment. The highlighted areas under the selection and range bars in the alignments show the sequence positions either in agreement (consistent) in blue, or else in red.

The following display functions are available with the *MSA Comparator*:

- ***Show/Clear*** *button.* Clicking this button will toggle between two states:
  - **Show** will display the consistency between the two alignments. The area below the blue selection bar above the reference alignment is compared with the corresponding positions in the second alignment (the area under the green range bar). Columns that are consistent (identical) between the two alignments are colored blue. Columns that are inconsistent (different) are colored red.
  - ***Clear*** will clear the consistency color-coding.
- ***Single col.*** *button.* Clicking this button will change the size of the selection bar to a single-column length. See the later description of the **Display options** menu on how to change the selection bar size.
- ***Show consistency*** *button.* Clicking this button will display all consistent columns in the entire alignment in blue highlighting.
- ***<<*** *and* ***>>*** *buttons.* Clicking these buttons will shift the selection bar above the reference alignment to the left or right, respectively. See the later description of the **Display options** menu on how to change the stepping size.
- ***Show information*** button. Clicking this button will display the information content bar chart for each alignment (shown between the MSAs in **Figure 7.1**). Clicking the button again will remove the display. The height of each bar represents the information content of each column of the alignment. The maximum bar height is achieved when the alignment column is fully conserved. The maximum value depends on the type of sequences in the alignment: 2.0 for nucleotide and 4.32 for amino acid sequences. Scrolling of each information content display is synchronized with that of corresponding MSA. For more information on how the information content is calculated, see Anderson *et al*. (2011a).



STATISTICS: % consistency: 9.906% Sum of pairs score: 0.536 Column score (no gaps) : 0.119 Column score (with gaps) : 0.108

Reference: lipocalinSuperfamily_alignment.fasta Selection bar coordinates: 1 - 40 Alignment length: 207

Other alignment: lipocalinSuperfamily_CLTW2.fasta Range bar coordinates: 1 - 54 Alignment length: 212

Figure 7.2: The *MSA Comparator* showing SPS display and statistics panel. For the same two example alignments shown in **Figure 7.1**, the column-wise SPS display (between the MSAs) and alignment statistics (above the reference MSA) are shown. The maximum SPS value is shown for the ungapped and consistent columns in positions 26 to 39. The large red squares under the SPS bar chart for the same positions indicate that these positions are used in calculation of both the ungapped and gapped column scores.

- ***Show sum of pairs*** button. Clicking this button will show the column-wise 'sum of pairs score' (SPS) bar chart between the two MSAs and the statistics panel above the alignment file name (shown in **Figure 7.2**). The column-wise SPS is calculated for the second (bottom) MSA compared against the reference MSA. The SPS bar chart is aligned to the bottom MSA. In the statistics panel, the following four scores are given.
    - **% consistency**. This is the percent of columns in the second MSA that are 100% identical to the columns in the reference MSA.
    - **Sum of pairs score**. This is the sum of pair score (SPS) for the whole alignment. This is calculated from the column-wise SPS's summed up and normalized.
    - **Column score (no gaps).** This is the column score calculated excluding all columns that have gaps. It is the total number of ungapped columns in the bottom alignment that have the maximum SPS divided by the total number of ungapped columns. Ungapped columns (columns that have no gaps) are

marked by a red square beneath the SPS bar chart (indicated by C.S.). If the column contains the maximum SPS, this square is large, else it is small. Only the columns marked with the large red squares are used in the calculation.

- **Column score (with gaps).** This is the column score calculated including also columns that have gaps. It is the total number of all columns in the bottom alignment that both match the column in the top (reference) alignment identically and contain at least 20% non-gap characters (maximum of 80% gaps). To distinguish from the ungapped columns (marked with red squares), gapped columns (columns that have gaps) with greater then 20% non-gap characters are marked by a purple square beneath the SPS bar chart (indicated by C.S.). If the column of the bottom alignment matches the corresponding column in the reference alignment, this square is large, else it is small. Only the columns marked with the large purple squares are used in the calculation of the gapped column score.

For more details on how these scores are calculated, see Anderson *et al.* (2011a).

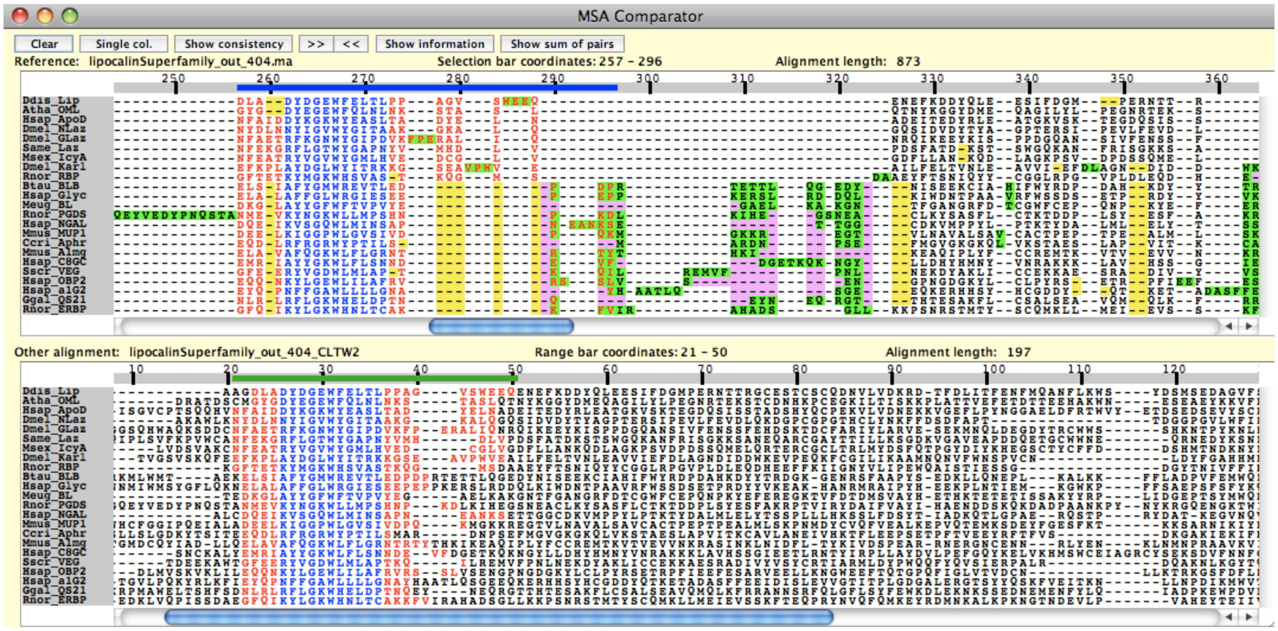
Selection bar can be also moved to a position directly by clicking on anywhere in the gray scale bar above the reference alignment.

The **Display options** menu allows the user to change display parameters:

- **Set step size…** allows the user to change the number of columns that the selection bar will move when the navigation buttons (**<<** or **>>**) are clicked. The default value is 1.
- **Set selection bar width…** allows the user to change the number of columns that the selection bar will cover. The default value is 25.

7.2 Comparing with simulated true MSAs

When the *MSA Comparator* is used with simulated true MSAs generated by **iSGv2.1** and the event trace file (xxx.trace) is present, the insertions and deletions are colored in green and yellow, respectively, as shown in **Figure 7.3**. In sites where an insertion occurred followed by a deletion the resulting gap is colored in pink. Clicking on a gap symbol (-) brings up the event label.



Reference: lipocalinSuperfamily_out_404.ma Selection bar coordinates: 257 - 296 Alignment length: 873

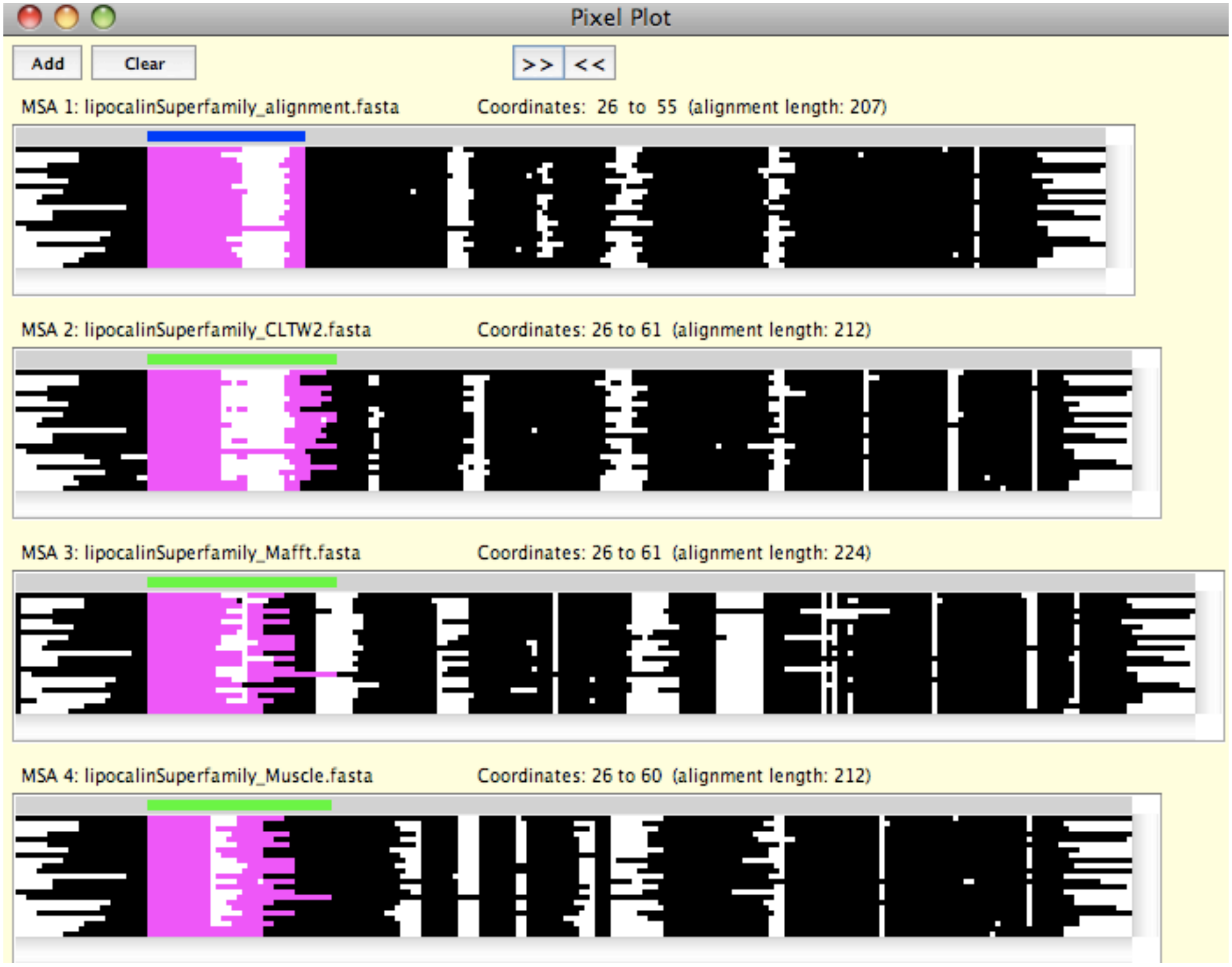
Other alignment: lipocalinSuperfamily_out_404_CLTW2 Range bar coordinates: 21 - 50 Alignment length: 197

Figure 7.3: The *MSA Comparator* showing indel color-mapping. In this example, the true alignment provided by **iSGv2.1** (simulating 23 lipocalin-superfamily protein sequences) was used as the reference and compared against the MSAs reconstructed by **ClustalW2 v2.1**. Both sample files (lipocalinSuperfamily_out_404.ma and lipocalinSuperfamily_out_404_CLTW2) are included in the **"iSG_sample_output"** folder.

8. *Pixel Plot*

The *Pixel Plot* allows the user to view multiple MSAs (**Figure 8.1**). It allows for the inspection of extremely large MSAs with long and/or many sequences. It depicts the general shape (gap distribution) of the compared alignments in pixel patterns. The *Pixel Plot* is available from the **SuiteMSA** main screen as well as from the **Alignments** menu in the *iSG Simulation*, the *ClustalW2* GUI, and the *Muscle* GUI.

The requirements in MSAs for the *Pixel Plot* is the same as the *MSA Comparator*:



MSA 1: lipocalinSuperfamily_alignment.fasta Coordinates: 26 to 55 (alignment length: 207)

MSA 2: lipocalinSuperfamily_CLTW2.fasta Coordinates: 26 to 61 (alignment length: 212)

MSA 3: lipocalinSuperfamily_Mafft.fasta Coordinates: 26 to 61 (alignment length: 224)

MSA 4: lipocalinSuperfamily_Muscle.fasta Coordinates: 26 to 60 (alignment length: 212)

Figure 8.1: The *Pixel Plot*. In this example, the alignment of the lipocalin superfamily proteins (MSA 1; lipocalinSuperfamliy_ailgnment.fasta) is compared with three MSAs generated by **ClustalW2 v2.1** (MSA 2), **MAFFT v6.843** (MSA 3), and **Muscle v3.8.13** (MSA 4). The selection bar is shown in blue at the top of the reference alignment. The green range bar above each of the reconstructed alignments shows the column range that covers the sequence positions selected in the reference alignment. The highlighted areas under the selection and range bars in the alignments show the positions of the residues selected in the reference alignment.

- Sequence names (including upper or lower cases) must be identical among the MSAs (including non-alphanumeric characters such as '_' or '-').
- Sequences in all MSAs must be identical. They may differ only in the location of the gaps.
- Only aligned sequences can be compared. All sequences in each MSA must contain the same number of sites. White-space characters (space, tab, line breaks) will be ignored and not counted for the alignment length.

The first alignment selected (MSA 1) is considered as the 'reference' alignment. All other alignments are compared against the reference. Once the reference alignment (MSA 1) is displayed, the **Add** button can be used to add additional alignments. The order of the sequences included in the reference MSA will be used to display all other MSAs. If the order of the sequences in the added MSA is different from the reference MSA (MSA 1), the sequences will be sorted to match the order of the reference MSA.

The blue selection bar above the reference alignment (MSA 1) indicates the region selected for comparison. The green range bars above the other MSAs show the column range that covers the sequence positions selected in the reference alignment. The magenta-highlighted pixels under the selection and range bars show the corresponding sequence positions.

Following display functions are available with the *Pixel Plot*:

- ***Add** button.* Use this button to add another MSA.
- ***Show/Clear** button.* Clicking this button will toggle between two states:
  - **Show** will display the location of the characters selected in the reference alignment by magenta highlighting.
  - **Clear** will clear the highlighting under the selection and range bars**.**
- ***<<** and **>>** buttons.* Clicking these buttons will shift the selection bar above the reference alignment (MSA 1) in the direction indicated. See the description of the **Display options** menu on how to change the stepping size.

Selection bar can be also moved to a position directly by clicking on anywhere in the gray scale bar above the reference alignment.

The **Display options** menu allows the user to change display parameters:

- **Set step size…** allows the user to change the number of columns that the selection bar will move when the navigation buttons (**<<** or **>>**) are clicked. The default value is 1.
- **Set selection bar width…** allows the user to change the number of columns that the selection bar will cover. The default value is 25.

9. *iSG Simulation*

The *iSG Simulation* button in the **SuiteMSA** main screen is used to start the **iSGv2.1** sequence simulation.

In this section, we will go through how to set up and run **iSG** simulation using the sample files included in the **SuiteMSA** package. For more details about options available in **iSGv2.1**, refer to the manual distributed with **iSGv2.1**, which is also available from: http://bioinfolab.unl.edu/~cstrope/iSG/

9.1 Setting up file/folder locations

Before starting the simulation setup, you need to set the locations of the various input, output, log, and executable files. Select the **Set folder locations...** from the **File** menu at the top of *iSG Simulation* screen. The **Set iSG v2.1 folder locations** window will appear as shown in **Figure 9.1**

Four folders need to be located. You can either choose these folders by using the **Browse...** button or typing the complete path in each field by yourself. For your convenience, if you use the **Browse...** button to choose the input folder location at the top, the output and log file folder locations will be auto-filled. You can edit these paths if

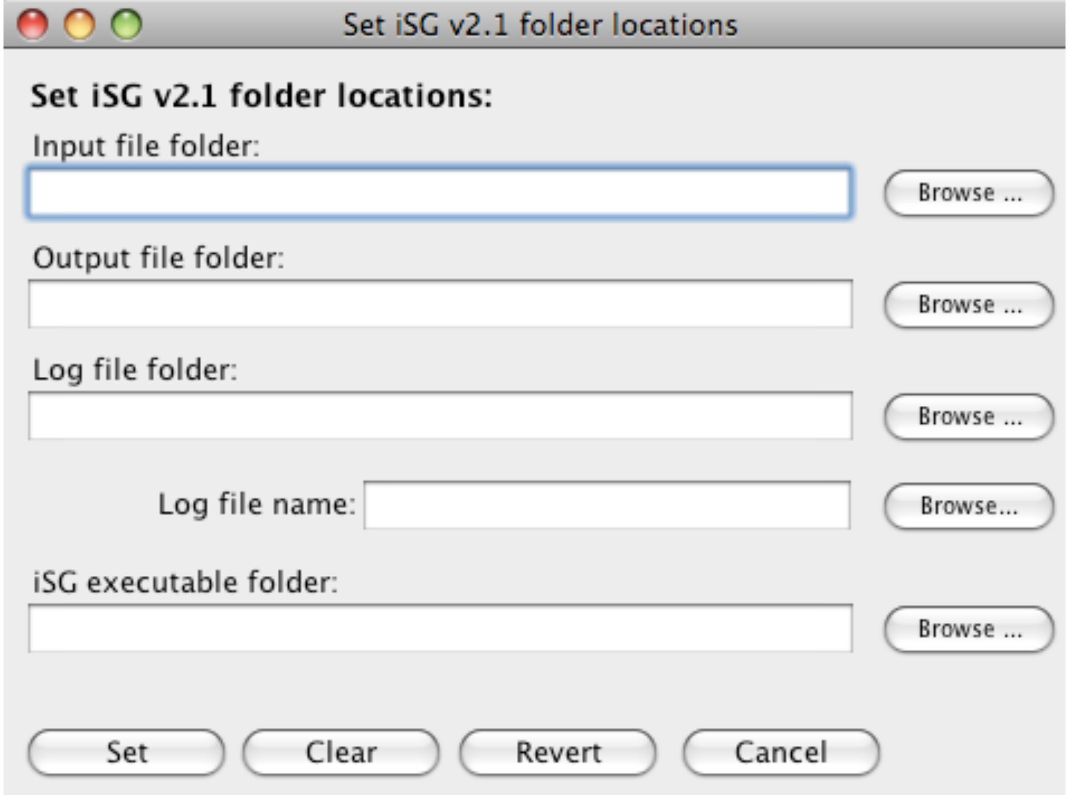


Figure 9.1: The "Set iSGv2.1 folder locations" window.

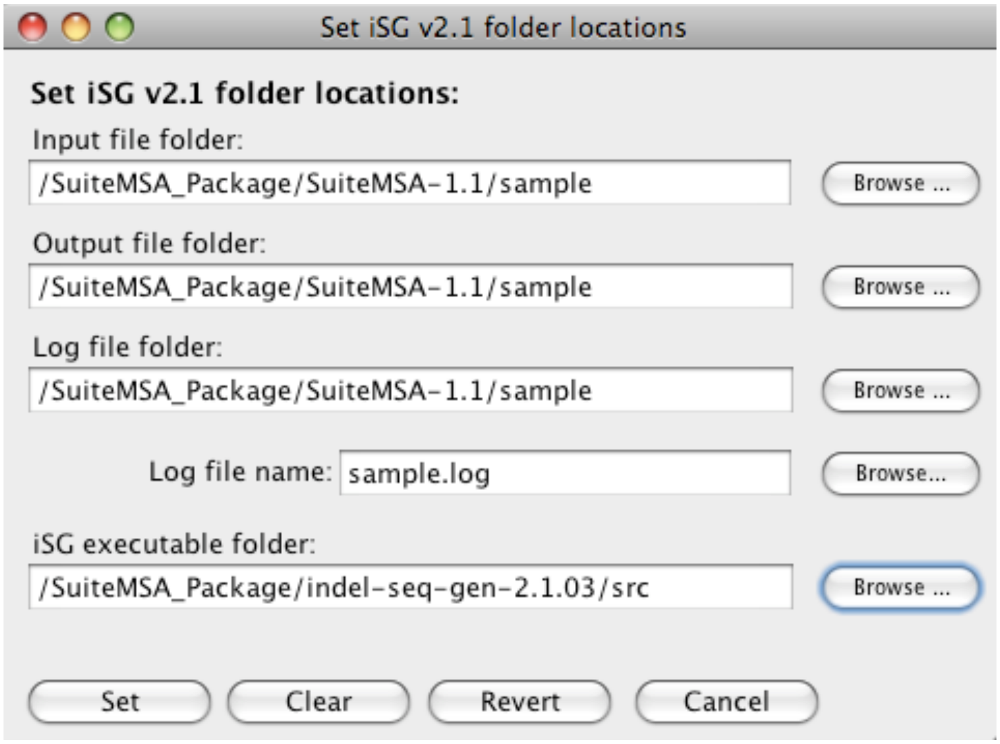
necessary. Make sure that these folders exist in the specified locations. When you push the **Set** button, if any of the folders does not exist, you will see an error message until you create these folders. The required folders and file are as follows:

i) **Input file folder.** This is where all support files for a specific run of the **iSGv2.1** simulation must be stored. The guide tree file is the minimum requirement. Other support files include: frequency file, indel length distribution file, root sequence file, lineage (spec) file, *etc*. Any files specified in the command-line options, in the guide tree, or in the lineage specification file should be found in this input folder. These files are explained in the later sections. For more details, refer to the **iSGv2.1** manual.

ii) **Output file folder.** This is where all files generated by **iSGv2.1** are placed.

iii) **Log file folder.** This is where the log file generated by **SuiteMSA** is stored. This log tracks each simulation run complete with the command-line options used, date and time stamps, and any error messages received from **iSGv2.1**. This log will be valuable in helping you solve any issue you might have in setting up **iSGv2.1** simulations as well as in providing a record of simulation parameters. The log from each run is appended to the same log file.

iv) **Log file name**. This is the actual name of the log file. While the folder needs to be present, the log file will be created if the file does not exist already. The content of



Set iSG v2.1 folder locations:

Input file folder: /SuiteMSA_Package/SuiteMSA-1.1/sample

Output file folder: /SuiteMSA_Package/SuiteMSA-1.1/sample

Log file folder: /SuiteMSA_Package/SuiteMSA-1.1/sample

Log file name: sample.log

iSG executable folder: /SuiteMSA_Package/indel-seq-gen-2.1.03/src

Figure 9.2: The "Set iSGv2.1 folder locations" window with folder and file names filled in.

the log file can be cleared using the **Clear log file** in the **File** menu.

v) **iSG executable folder.** This is where the binary executable file for **iSGv2.1**, indel-seq-gen, has been stored.

Figure 9.2 shows how the set-up window looks like after all folder locations and the log file name have been set to the folders included in the SuiteMSA_Package. Once all folders and file have been selected, click on the **Set** button.

The locations of these folders and file are saved and recalled next time you run **SuiteMSA**. If you do not have to change the folder/file information, simply dismiss the window by clicking on either **Set** or **Cancel** button.

The **Revert** button can be used to restore the file/folder locations to those saved last time. Note that once the **Set** button is used, the new locations will be saved and the older settings will be over-written.

9.2 Running iSGv2.1 simulation: an overview

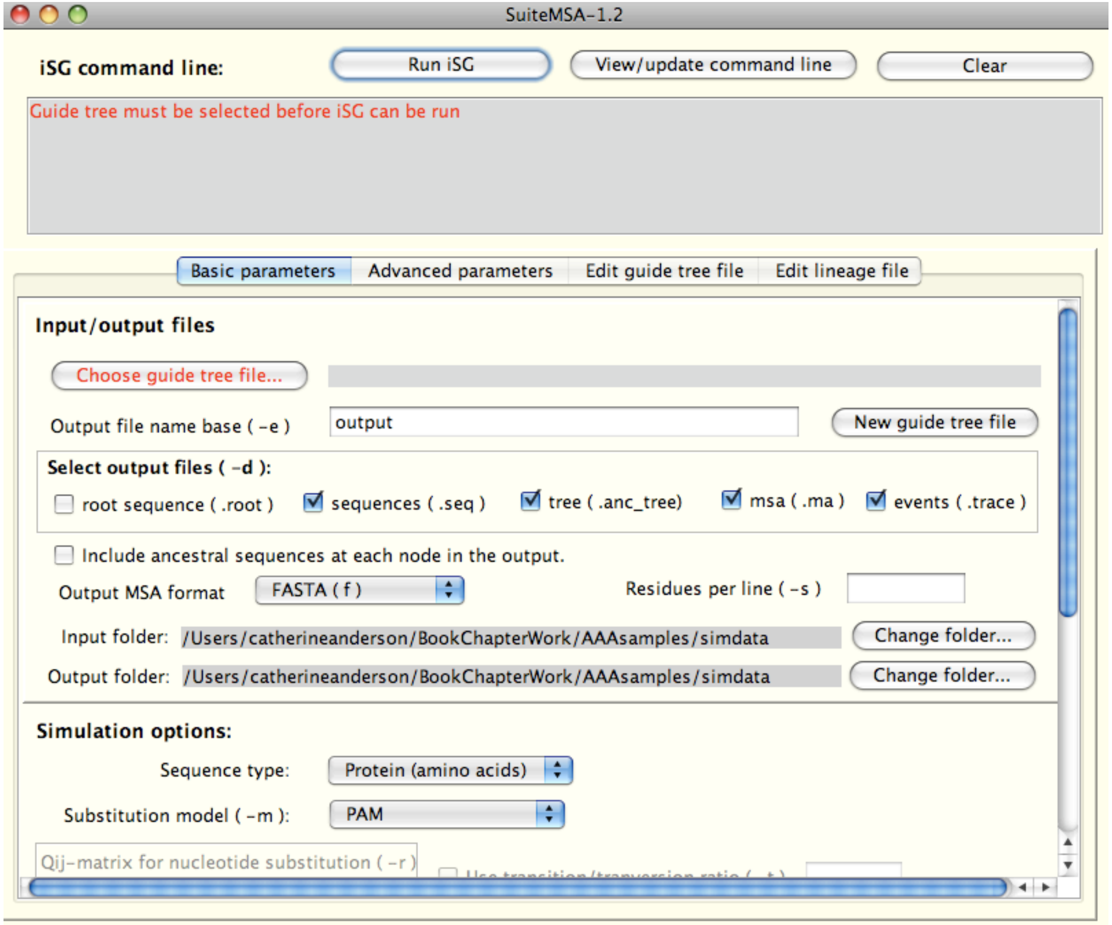
All options used with **iSGv2.1** can be set using **SuiteMSA** GUI. **SuiteMSA** provides basic error-checking for the input support files making it easier for a new user to successfully run **iSGv2.1**. Note, however, that the current version of **SuiteMSA** does not perform consistency-checking for information files included in guide tree and lineage files.

Setting up and running a simulation is done through the main window as shown in **Figure 9.3**. The **Run iSG** button starts the simulation. The gray area below **"iSG command line:"** shows the iSG command line used for the simulation.

As shown in **Figure 9.3**, when the main window is first opened, the command-line field displays a message in red indicating that a guide tree file (the minimum requirement) must be selected before **iSGv2.1** can be run. As soon as a guide tree file is selected, the iSG command-line field will display the full command as shown in **Figure 9.4**. As you make changes in the options for a simulation, use the **View/update command line** button to update the command-line display. Examining the command line will help you learn the proper syntax to run **iSGv2.1** directly in the future.

When you click on the **Run iSG** button, the command line is generated afresh, displayed, and used to run **iSGv2.1**. Therefore, it is not necessary to "update" the command line before running **iSGv2.1**. In order to check the iSG command line with all options, however, it is useful to view the complete command line before starting the simulation.

Note that the background of the iSG command line field is light gray. Whenever a text area is light gray, it indicates that the text area is read only, *i.e.*, not editable.



iSG command line:

Guide tree must be selected before iSG can be run

Input/output files

Output file name base (-e) output

Select output files (-d):

☐ root sequence (.root) ☑ sequences (.seq) ☑ tree (.anc_tree) ☑ msa (.ma) ☑ events (.trace)

☐ Include ancestral sequences at each node in the output.

Output MSA format: FASTA (f) Residues per line (-s)

Input folder: /Users/catherineanderson/BookChapterWork/AAAsamples/simdata

Output folder: /Users/catherineanderson/BookChapterWork/AAAsamples/simdata

Simulation options:

Sequence type: Protein (amino acids)

Substitution model (-m): PAM

Qij-matrix for nucleotide substitution (-r)

☐ Use transition/transversion ratio (-t)

Figure 9.3: The main *iSG Simulation* window when it is first opened. Note the text in red in the **"iSG command line"** field. It indicates that a guide tree must always be specified before running an **iSGv2.1** simulation.

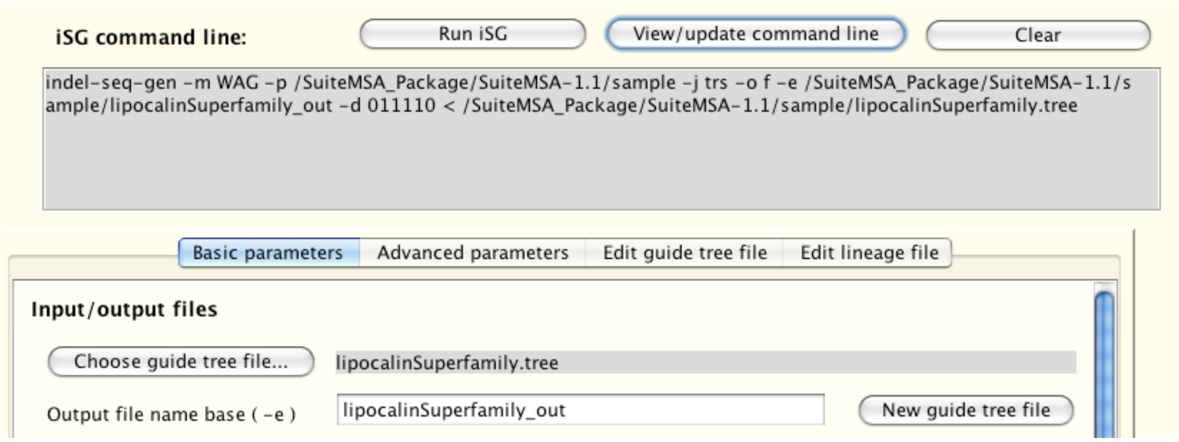
The **Clear** button at the top clears all option settings including the guide tree file.

All options for **iSGv2.1** can be selected in the lower section of the main window, which is grouped into four tabbed pages (see **Figure 9.3**). Each of these pages is described in detail in the following sections.

To help you learn and understand the **iSGv2.1** syntax, the command-line flags for all options used in **iSGv2.1** are shown next to the option labels (*e.g.*, '-e' for the "Output file name base" as shown in **Figure 9.3**). Furthermore, when you hover your cursor over any option label, it brings up a tool tip (a help tag), which shows a short explanation of the option.

9.3 Setting up iSGv2.1 simulation

This section explains the options available in the four option panels. The screen shots show the settings used for the lipocalin superfamily sequence simulation wherever possible. The complete list of options used for this simulation is also given at the end of



iSG command line:

```
indel-seq-gen -m WAG -p /SuiteMSA_Package/SuiteMSA-1.1/sample -j trs -o f -e /SuiteMSA_Package/SuiteMSA-1.1/s
ample/lipocalinSuperfamily_out -d 011110 < /SuiteMSA_Package/SuiteMSA-1.1/sample/lipocalinSuperfamily.tree
```

Input/output files

Choose guide tree file... lipocalinSuperfamily.tree

Output file name base (-e) lipocalinSuperfamily_out

Figure 9.4: The top section of the main window showing the iSG command line. It shows that a guide tree file, lipocalinSuperfamily.tree, has been selected. The output file name base has been also changed to reflect the guide tree file name.

this section. All support files required for this simulation are included in the **"sample"** folder within the **"SuiteMSA_Package"** under the **"SuiteMSA-1.2"** folder.

9.3.1 Basic parameters

The **"Basic parameters"** page presents the most basic options for the **iSGv2.1** simulation. For more detailed information on each option, refer to the iSGv2 User Manual.

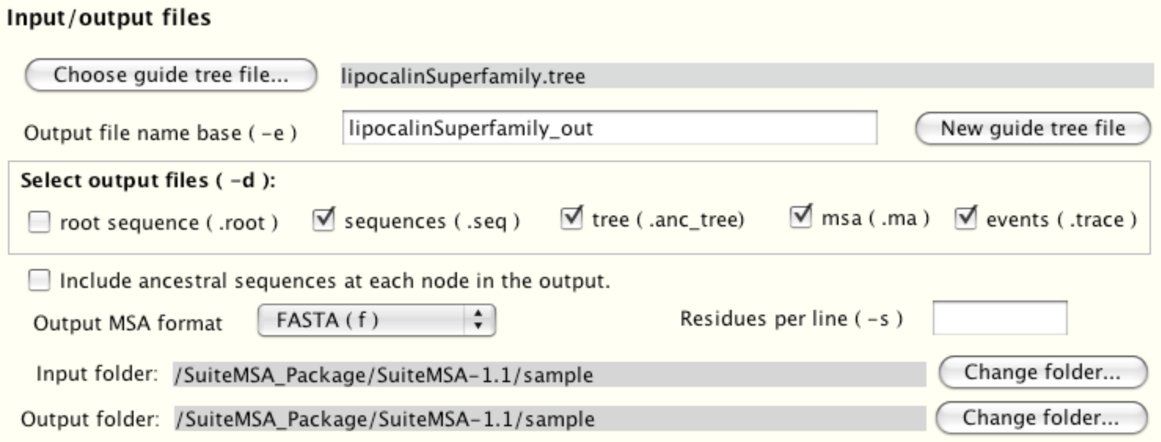
a. Input/output files

Figure 9.5 shows the **"Input/output files"** section, which deals with the organization of the files used with and generated by **iSGv2.1**.

i) **Choose guide tree file...** Use this button to select the iSG guide tree. The file should exist in the designated **"Input folder"**. This file should contain the guide tree in Newick format as well as information pertaining to the evolutionary model (substitution model, indel probability, *etc.*) for each partition (subsequence). The content of this file can be edited using the **"Edit guide tree file"** page. As a matter of good practice, select the guide tree file before setting any other parameters. Until a guide tree file is selected, the iSG command line cannot be displayed with the **"View/update command line"** button (shown in **Figure 9.4**).

ii) **Output file name base.** An output file-name base is automatically generated based on the guide tree file name. If the guide tree file name has an extension (*e.g.*, .tree), the extension part is removed and ' _out' is appended. If the file name has no extension, '_out' will be simply appended. This name base is used for all output files generated by **iSGv2.1**.

In the example shown in **Figure 9.5**, the guide tree file is "lipocalinSuperfamily.tree". The output files will be named as "lipocalinSuperfamily_out.ma",



Input/output files

Choose guide tree file... lipocalinSuperfamily.tree

Output file name base (-e) lipocalinSuperfamily_out

New guide tree file

Select output files (-d):

☐ root sequence (.root) ☑ sequences (.seq) ☑ tree (.anc_tree) ☑ msa (.ma) ☑ events (.trace)

☐ Include ancestral sequences at each node in the output.

Output MSA format FASTA (f) Residues per line (-s)

Input folder: /SuiteMSA_Package/SuiteMSA-1.1/sample Change folder...

Output folder: /SuiteMSA_Package/SuiteMSA-1.1/sample Change folder...

Figure 9.5: The Input/output files section of the "Basic parameter" page. In this example, the guide tree file, lipocalinSuperfamily.tree, is chosen. The **"Output file name base"** is auto-filled based on the input file name.

"lipocalinSuperfamily_out.trace”, "lipocalinSuperfamily_out.anc_tree", *etc*. The output file-name base can be changed by typing directly in the text box labeled "Output file name base”.

iii) **New guide tree file.** This button will allow you to create a new guide tree file from scratch. (This function is currently not available.)

iv) **Select output files. iSGv2.1** generates five output files. Four of these files are selected by default. If you do not need any of these files, you can suppress saving the files. It saves disk space when running a large number of simulations. The output files are as follows:

- ***root sequence*** *(xxxxx.root).* This file contains the root sequence of the simulation.
- ***sequences*** *(xxxxx.seq).* This file contains the simulated sequence products in FASTA format.
- ***tree*** *(xxxxx.anc_tree).* This file contains the final tree information. This file is needed to display the event-mapping tree.
- ***msa*** *(xxxxx.ma).* This file contains the true MSA of the simulated sequences. This file is needed to display the true alignment with the events tracked.
- ***events*** *(xxxxx.trace).* This file contains indel event records from the simulation. This file is needed when displaying either the event-mapping tree or the true alignment.

v) **Include ancestral sequences at each node in the output.** By selecting this option, the ancestral sequence generated at each node (including the root) during the simulation can be saved in the output file. Files affected by this option are: the MSA file (xxxxx.ma), the trace file (xxxxx.trace), and the sequence file (xxxxx.seq).

vi) **Output MSA format.** Choose from FASTA, Phylip, or NEXUS format. While the default format for iSGv2.1 is Phylip, the default used with **SuiteMSA** is FASTA. The alignment must be in FASTA format in order to be displayed in the **SuiteMSA** viewers.

vii) **Input/Output folder.** The input and output folders you chose in the **"Set folder locations..."** (see the section 9.1) are shown in the gray text area. You can change these folder locations by using each **Change folder…** button.

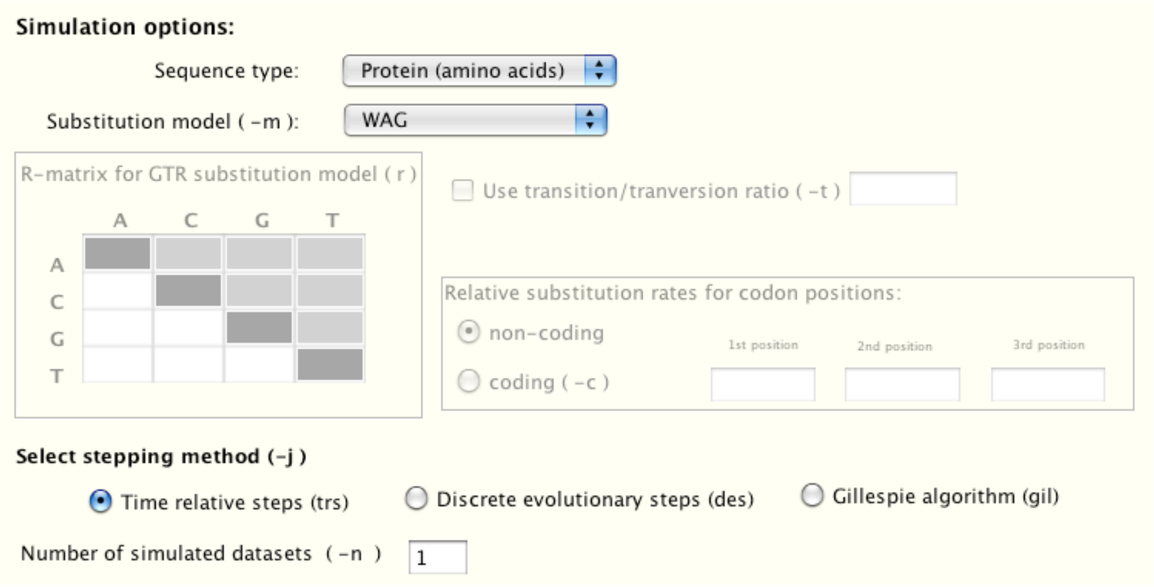
b. Simulation options

This section (shown in **Figure 9.6**) deals with the parameters used to control the actual simulation.

i) **Sequence type.** There are two types: Protein (amino acid) or DNA (nucleotide) sequences. This selection affects the choices of other options (*e.g.,* substitution model). Note that no consistency-checking is done with sequence type and the content of the support files, such as frequency files or lineage specification file. Check the appropriate content of these files depending on the sequence type.

ii) **Substitution model.** The available substitution models depend on the sequence type. The default protein substitution model is PAM, and HKY85 for nucleotide sequences.

The general time reversible (GTR) model available for nucleotide simulations requires the entry of the "R-matrix", which will become active when the GTR model is chosen. For convenience, the rate fields are auto-filled with a uniform rate. The user can change these values as needed. It is mandatory that all six fields be filled. On the other hand, it is not necessary for these rate values to sum to 1.0 as **iSGv2.1** normalizes them before running the simulation.



Simulation options:

Sequence type: Protein (amino acids)

Substitution model (-m): WAG

R-matrix for GTR substitution model (r)

	A	C	G	T
A				
C				
G				
T				

□ Use transition/tranversion ratio (-t)

Relative substitution rates for codon positions:

● non-coding

○ coding (-c)

1st position 2nd position 3rd position

Select stepping method (-j)

● Time relative steps (trs) ○ Discrete evolutionary steps (des) ○ Gillespie algorithm (gil)

Number of simulated datasets (-n) 1

Figure 9.6: Simulation options on the "Basic parameters" page. In this example, WAG is chosen for the amino acid substitution model.

iii) **Transition/transversion ratio.** This option is available only if the nucleotide sequence type is chosen. When first selected, for convenience, the field is auto-filled with the default ratio of 1.0, which can be changed to any positive number by the user. If this option is selected, the simulation will not run unless the field is filled in, or the option unchecked.

iv) **Relative substitution rates for codon positions.** This option is also available only if the nucleotide sequence type is chosen. It specifies the relative rates among the three codon positions. Selecting the **"non-coding"** option applies a uniform substitution rate among the three positions. Selecting the **"coding"** option requires that you enter the relative substitution rate for each of the three codon positions. If the **"coding"** option is selected, the simulation will not run until all three rates are entered or the option unchecked.

v) **Select stepping method.** There are three simulation-stepping methods to choose from: Time relative steps (trs), Discrete evolutionary steps (des), and the Gillespie algorithm (gil). For more detailed information on these options, see the iSGv2 User Manual.

vi) **Number of simulated datasets to generate.** The default number of datasets is one. To run the simulation more than once, enter the desired number in the field. This number is mandatory, and without a number, you will see an error message.

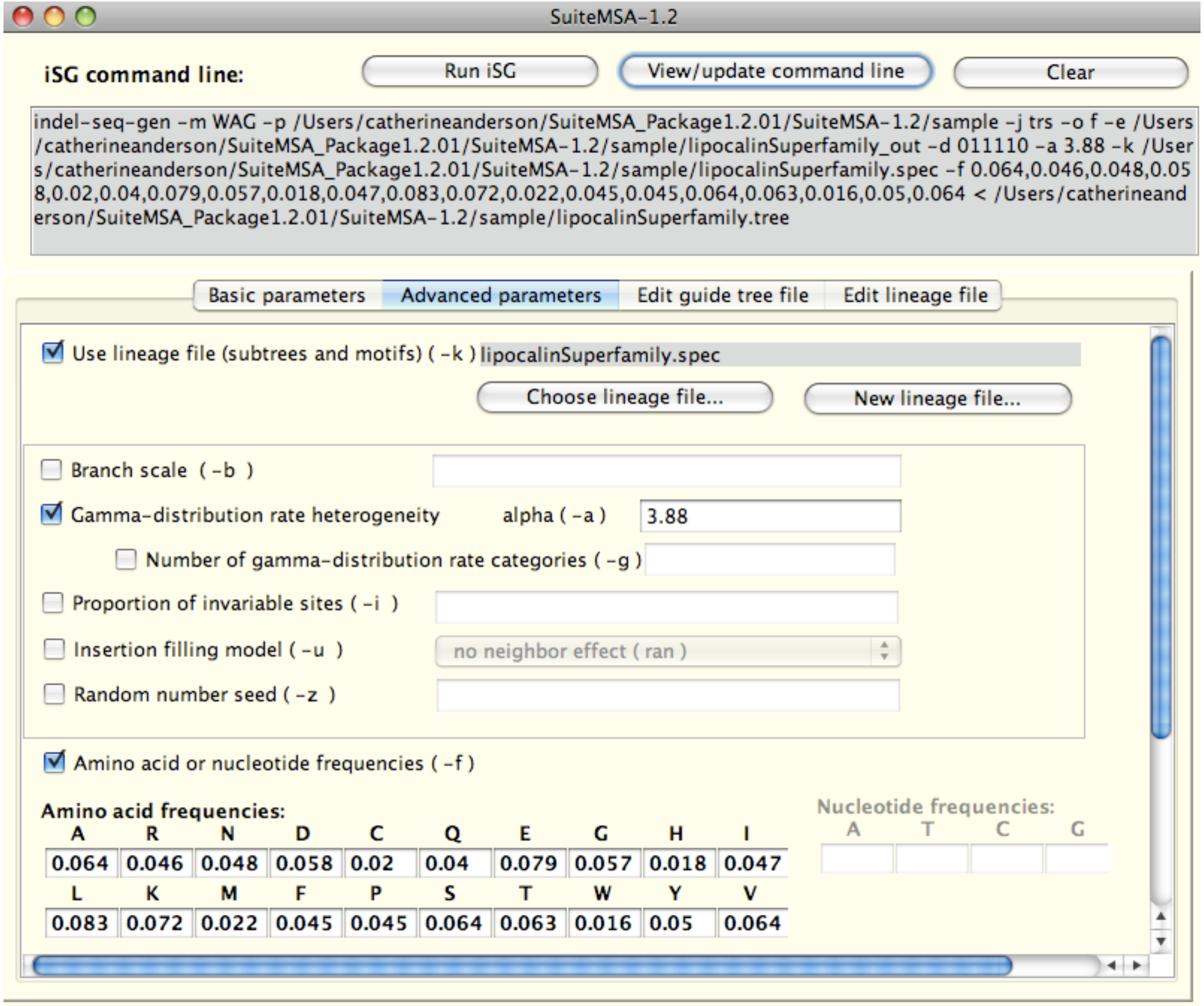
9.3.2 Advanced parameters

The **"Advanced parameters"** page presents the more advanced options that controls simulation by **iSGv2.1**. **Figure 9.7** shows this page.

i) **Use lineage file.** Subtree and motif specifications are included in the lineage file. See the iSGv2 User Manual for the details of the lineage file format. After the file has been selected, its content can be viewed and edited using the **"Edit lineage file”** page. Note that consistency error-checking between the lineage file and other support files (*e.g.*, guide tree file) is not performed by **SuiteMSA** currently. Make sure that all input files contain consistent contents.

ii) **Branch scale.** This option provides a scaling factor for the guide tree branch lengths. When this option is checked, the field is auto-filled with the default value of 1.0. The simulation will not run unless this field is filled out or the option unchecked.

iii) **Gamma-distribution rate heterogeneity.** Selecting this option allows for the use of gamma-distribution rate heterogeneity in the simulation. When this option is checked, the alpha (or shape) parameter is auto-filled with the default value of 1.0. Change the value as needed. The simulation will not run unless this field is filled out or the option unchecked.



iSG command line:

```
indel-seq-gen -m WAG -p /Users/catherineanderson/SuiteMSA_Package1.2.01/SuiteMSA-1.2/sample -j trs -o f -e /Users
/catherineanderson/SuiteMSA_Package1.2.01/SuiteMSA-1.2/sample/lipocalinSuperfamily_out -d 011110 -a 3.88 -k /User
s/catherineanderson/SuiteMSA_Package1.2.01/SuiteMSA-1.2/sample/lipocalinSuperfamily.spec -f 0.064,0.046,0.048,0.05
8,0.02,0.04,0.079,0.057,0.018,0.047,0.083,0.072,0.022,0.045,0.045,0.064,0.063,0.016,0.05,0.064 < /Users/catherineand
erson/SuiteMSA_Package1.2.01/SuiteMSA-1.2/sample/lipocalinSuperfamily.tree
```

☑ Use lineage file (subtrees and motifs) (-k) lipocalinSuperfamily.spec

☐ Branch scale (-b)

☑ Gamma-distribution rate heterogeneity alpha (-a) 3.88

☐ Number of gamma-distribution rate categories (-g)

☐ Proportion of invariable sites (-i)

☐ Insertion filling model (-u) no neighbor effect (ran)

☐ Random number seed (-z)

☑ Amino acid or nucleotide frequencies (-f)

Amino acid frequencies:

A	R	N	D	C	Q	E	G	H	I
0.064	0.046	0.048	0.058	0.02	0.04	0.079	0.057	0.018	0.047
L	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
0.083	0.072	0.022	0.045	0.045	0.064	0.063	0.016	0.05	0.064

Nucleotide frequencies:

A	T	C	G

Figure 9.7: The "Advanced parameters" page. In this example, the lineage file, lipocalinSuperfamily.spec, is chosen. Gamma-distribution for rate heterogeneity is chosen and the alpha value of 3.88 is typed in. Amino acid frequencies are also given. See the section 9.3 for an alternative way of specifying the amino acid (or nucleotide) frequencies.

With the **"Gamma-distribution rate heterogeneity"** option checked, the **"Number of gamma-distribution rate categories"** can be specified. If this option is checked, the accompanying field must be filled with an integer between 2 and 32. For convenience the field is auto-filled with 32.

iv) **Proportion of invariable sites.** The proportion value needs to be between 0.0 and 1.0 (exclusive). For convenience, the field is auto-filled with 0.5. The simulation will not run unless this field is filled out or the option unchecked.

v) **Insertion filling model.** This option is available only for protein simulation. Three models are provided for amino acid insertions based on neighbor effects: random, original (Xia and Xie 2002), and swiss-prot. For more information, refer to the iSGv2 User Manual.

vi) **Random number seed.** A random number seed can be provided in order to make a specific simulation run exactly reproducible. As with all other options, the simulation will not run unless this field is filled out or the option unchecked.

vii) **Amino acid or nucleotide frequencies.** This option allows the user to specify either amino acid or nucleotide frequencies. If this option is not checked, the default frequencies used by **iSGv2.1** are based on the substitution model. When the option is first checked, the fields are auto-filled with the uniform frequency (0.05 for each amino acid or 0.25 for each nucleotide). The values entered do not need to sum to 1.00 since **iSGv2.1** normalizes the values before running the simulation. All frequency fields must be filled out or the simulation will not run.

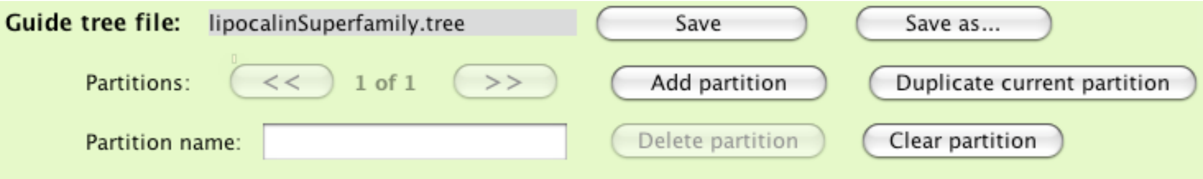
9.3.3 Edit guide tree file

The guide tree file used in **iSGv2.1** has a strict format. The **"Edit guide tree file"** page facilitates building and learning the proper format of the file. In the following sections, options are explained in five groups.

a. File components

The top section of the page (**Figure 9.8**) deals with the file and partition structure. The guide tree file name shown in the gray area is the file selected in the **"Basic parameters"** page. In the **"Edit guide tree file"** page, the guide tree file can be edited and the changes can be saved using the **Save** button. By using the **Save as...** button the guide tree information can be also saved in a new file.

Note that after changing the guide tree file contents through this page, the file must be saved before the changes take effect for the simulation run. Saving the file also triggers an error-checking routine, and the file cannot be saved until all errors are corrected. Note that when saving the guide tree file, **SuiteMSA** always saves the partition options in the following order: subsequence options, the partition name, root-sequence options, indel options, and finally the guide tree in Newick format. While **SuiteMSA** can read a guide tree file with partition options in any order, it always saves the options in this order.



Guide tree file: lipocalinSuperfamily.tree

Partitions: 1 of 1

Partition name:

Figure 9.8: The top section of the "Edit guide tree file" page. In this example, the partition information from the guide tree file lipocalinSuperfamily.tree is shown. There is only one partition in this file, so the navigation buttons are not enabled.

You can navigate through the partitions by using **<<** and **>>** buttons. The current partition number is displayed in the format of "*n* of *m*" (1 of 1 in **Figure 9.8**), where *n* is the number of the current partition and *m* is the total number of partitions in the guide tree file. Navigating away from the current partition also triggers the error-checking routine. It checks if all necessary fields are filled out with the correct data type and, where needed, within the correct range.

The **Add partition** button will insert a new partition immediately after the current partition. This new partition inherits the topology information from the previous partition. All other options are cleared except that the random root-sequence option with 200 characters is chosen in default.

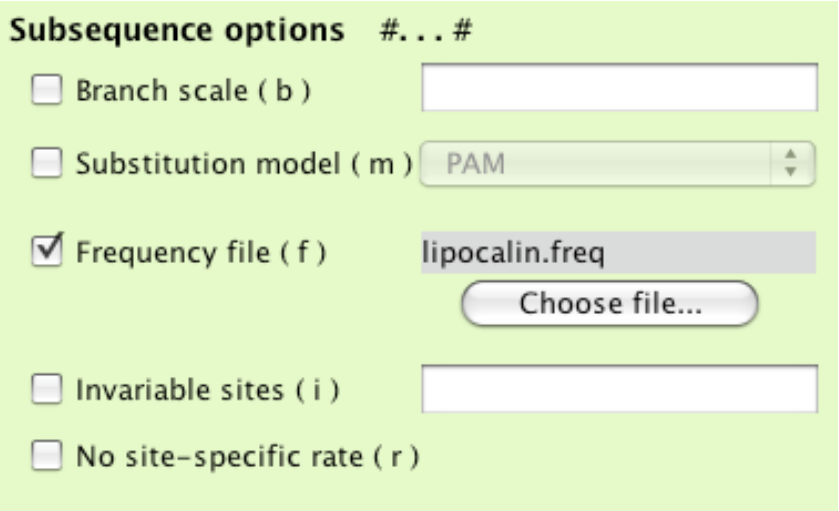
A partition can also be added by using the **Duplicate current partition** button. This will create a new partition by duplicating the current partition. The second copy of the partition is named by attaching '_copy' to the original partition name. This duplicated partition becomes the current partition and is displayed.

The current partition may be deleted by using the **Delete partition** button. The **Clear partition** button will clear all text fields of the current partition and uncheck all options.

Giving the **"Partition name"** is optional. When more than one partition is used, it would be useful to use meaningful partition names to help the user better track which partition information is being displayed.

b. Subsequence options

Figure 9.9 shows the **"Subsequence options"** section. This section deals with those options that are enclosed between the two pound signs (#...#) within the guide tree file. These options override the global parameter set on the command line. The letters



Subsequence options #. . . #

☐ Branch scale (b)

☐ Substitution model (m) PAM

☑ Frequency file (f) lipocalin.freq

☐ Invariable sites (i)

☐ No site-specific rate (r)

Figure 9.9: The "Subsequence options" section of the “Edit guide tree file” page. In this example, the frequency file, lipocalin.freq, is used. This is specified in the selected guide tree file, lipocalinSuperfamily.tree.

displayed within the parentheses after the option name is the letter used within the guide tree file. For more information on the format of these options, see the iSGv2 User Manual.

i) **Branch scale.** This is the scaling factor that will be applied to all branch lengths of the guide tree during the simulation.

ii) **Substitution model.** The selected substitution model will be applied to the partition in place of the model specified by in the **"Basic parameters"** page.

iii) **Frequency file.** The frequencies specified in the selected frequency file will be applied to the partition during the simulation. A frequency file for nucleotides should contain four comma-delimited values, while one for the amino acids should contain 20 comma-delimited values.

Note that the frequency information can be also provided in the **"Advanced parameters"** page as shown in **Figure 9.7**. When there is only one partition, either of these two methods (using the option in the **"Advanced parameters"** page or using a frequency file specified in the guide tree file as shown in **Figure 9.9**) can be used. Note, however, that if there are more than one partition, the guide tree file options are specific to each partition.

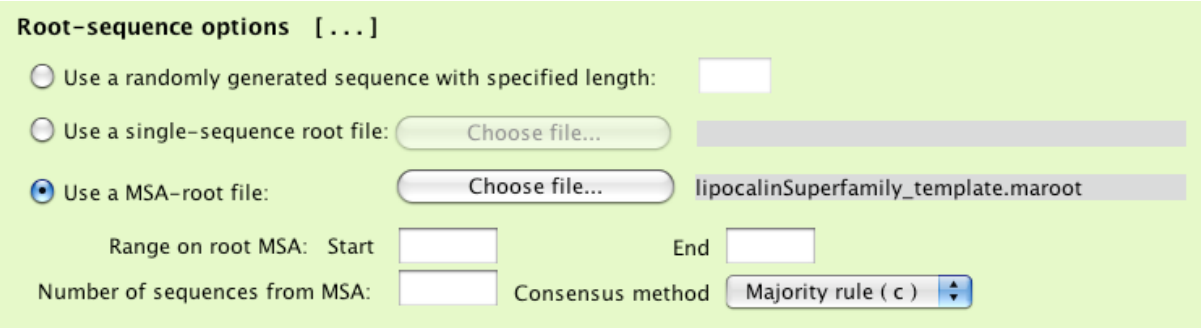
iv) **Invariable sites.** The proportion of invariable sites can be given for each partition.

v) **No site-specific rate.** When this option is selected, the site-specific rate settings including gamma-distribution heterogeneity and codon-position specific rates are removed.

c. Root-sequence options

Figure 9.10 shows the **"Root-sequence options"** section of the **"Edit guide tree file"**. This section deals with the options for generating the root sequence (those options that are enclosed between the square brackets in the guide tree file). Available root-sequence options are as follows:

- ***Use a randomly generated sequence with specified length.*** This option



Root-sequence options [. . .]

○ Use a randomly generated sequence with specified length:

○ Use a single-sequence root file:

● Use a MSA-root file: lipocalinSuperfamily_template.maroot

Range on root MSA: Start End

Number of sequences from MSA: Consensus method: Majority rule (c)

Figure 9.10: The root-sequence section of the "Edit guide tree file" page. In this example, the MSA-root root file, lipocalinSuperfamily_template.maroot, is chosen.

generates a random sequence of the given length as the root sequence.

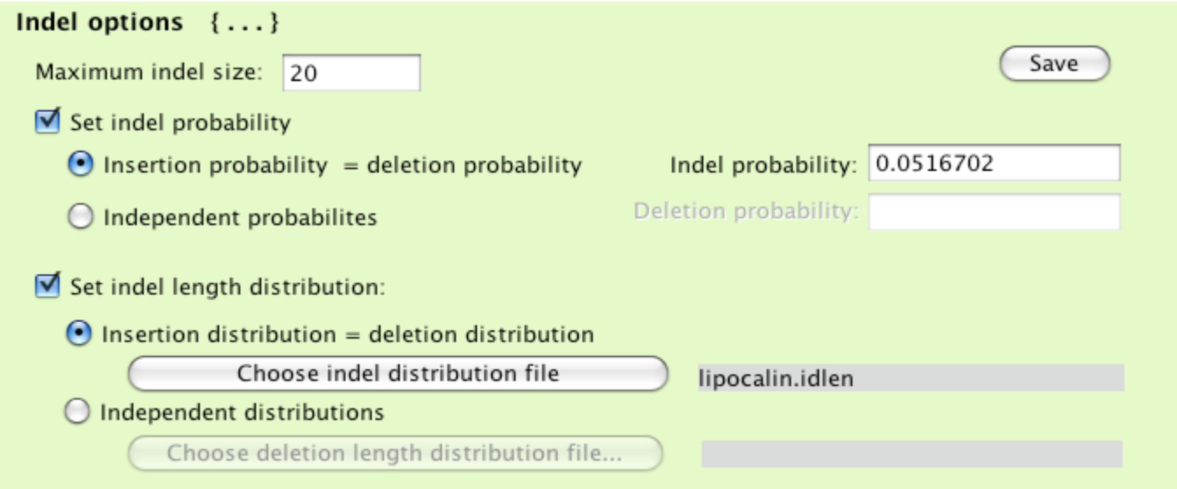
- ***Use a single-sequence root file.*** This option uses the root sequence included in the given file. The format for this file is given in iSGv2 Users manual. If motifs are specified in the lineage files (chosen in the **"Advance parameters"** page), the root sequence file must have the corresponding motif markers.
- ***Use an MSA-root file.*** This option requires a file that contains a multiple sequence alignment. If the **Start** and **End** fields for the **Range on root MSA** are left blank, the whole length of the alignment is used. Otherwise, enter the start and stop position of the subsection you wish used. If the **Number of sequences from MSA** field is left blank, all sequences within the alignment are used. Or enter the number of sequences to be used. Finally select the **Consensus method** from the pop-up menu.

  Note: If the MSA file contains a motif marker line, a lineage file including the motif information needs to be chosen in the **"Advanced parameters"** page. Having the marker present in the MSA root file and not selecting the lineage file with motif information will generate a warning from **iSGv2.1**. For more information on these options see iSGv2 User Manual.

d. Indel options

As shown in **Figure 9.11**, the **"Indel options"** section of the **"Edit guide tree file"** page specifies the parameters controlling the indel occurrence for the partition (these options are enclosed between the curly braces within the guide tree file). The indel options include their maximum size, probabilities, and length distribution. To avoid including indels in your simulation, uncheck these options and leave the maximum indel size field blank or type in '0'.

i) **Maximum indel size.** Enter an integer to specify the maximum length of insertions


Indel options {...}

Maximum indel size: 20

☑ Set indel probability

● Insertion probability = deletion probability

Indel probability: 0.0516702

○ Independent probabilites

Deletion probability:

☑ Set indel length distribution:

● Insertion distribution = deletion distribution

lipocalin.idlen

○ Independent distributions

Save
Choose indel distribution file
Choose deletion length distribution file...

Figure 9.11: The "Indel options" section of the "Edit guide tree file" page. In this example, the maximum indel size is set to 20. The indel probability is set to 0.0516702 (for both insertions and deletions). The indel length distribution is provided in the file lipocalin.idlen.

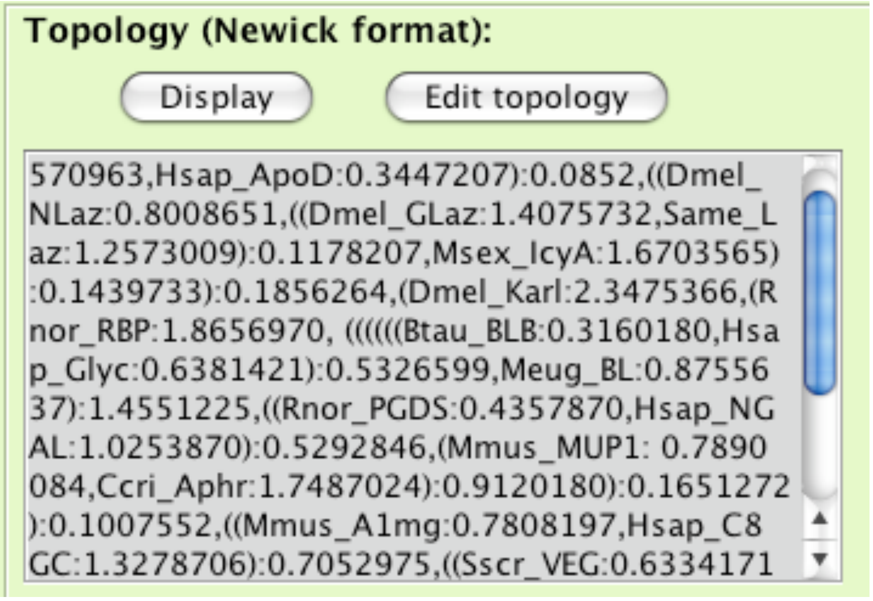
or deletions. If the maximum indel size is larger than 0 and no other indel options are checked, indel probability and length distributions as given by Chang and Benner (2004) are used (see iSGv2 User Manual for details). If either of the following two options is used, the **"Maximum indel size"** must be entered.

ii) **Set indel probability.** If this option is not selected, the indel probability given by Chang and Benner (2004) is used (see iSGv2 User Manual for details). When this option is selected, you have the choice of using the same probability for both insertions and deletions (select **"Insertion probability = deletion probability"**) or having them different probabilities (select **"Independent probabilities"**). If **"Independent probabilities"** is chosen, both **"Insertion probability"** and **"Deletion probability"** must be entered.

iii) **Set indel length distribution from file.** If this option is not selected, the indel length distribution given by Chang and Benner (2004) is used. Otherwise, the length distribution can be given in a file that contains a comma-delimited set of values of at least as many values as the maximum indel size entered. Again you have the option of using the same distribution file for both insertion and deletion (**"Insertion distribution = deletion distribution"**) or using separate files (**"Independent distributions"**).

e. Topology

The **"Topology"** section displays the guide tree in Newick format (**Figure 9.12**). The tree topology in Newick format can be edited by clicking on the **Edit topology** button. Since the current release of **SuiteMSA** does not perform any format-checking for the tree topology, it is for the user to ensure that the topology adheres to the correct format.



```
570963,Hsap_ApoD:0.3447207):0.0852,((Dmel_
NLaz:0.8008651,((Dmel_GLaz:1.4075732,Same_L
az:1.2573009):0.1178207,Msex_IcyA:1.6703565)
:0.1439733):0.1856264,(Dmel_Karl:2.3475366,(R
nor_RBP:1.8656970, ((((((Btau_BLB:0.3160180,Hsa
p_Glyc:0.6381421):0.5326599,Meug_BL:0.87556
37):1.4551225,((Rnor_PGDS:0.4357870,Hsap_NG
AL:1.0253870):0.5292846,(Mmus_MUP1: 0.7890
084,Ccri_Aphr:1.7487024):0.9120180):0.1651272
):0.1007552,((Mmus_A1mg:0.7808197,Hsap_C8
GC:1.3278706):0.7052975,((Sscr_VEG:0.6334171
```

Figure 9.12: The "Topology" area of the "Edit guide tree file" page. The tree displayed is the 23-taxon tree specified in the file guide tree file, lipocalinSuperfamily.tree.

To edit the tree graphically, click on the **Display** button. It will open an editable *Phylogeny Viewer*. The section 10 describes this tool in detail.

9.3.4 Edit lineage file

The **"Edit lineage file"** page deals with the subtree (clade or lineage) and motif options included in the linage specification file. For detailed information on these options, see the iSGv2 User Manual. The **"Edit lineage file"** page allows you to display and change the options for each subtree. Note that at least one subtree or one motif should exist before you can make any changes to these options. The options are divided into five groups and described below.

a. The lineage file content

The top section of the **"Edit lineage file"** page (**Figure 9.13**) shows the overview of the lineage file content. The information displayed is:

- ***Lineage file.*** The lineage file selected in the **"Advanced parameters"** page is shown.
- ***Subtrees loaded.*** This field shows the number of subtrees in the lineage file. The number of the subtrees is updated as a subtree is added or deleted.
- ***Motifs loaded.*** This field shows the number of motifs in the lineage file. This field is similarly updated.

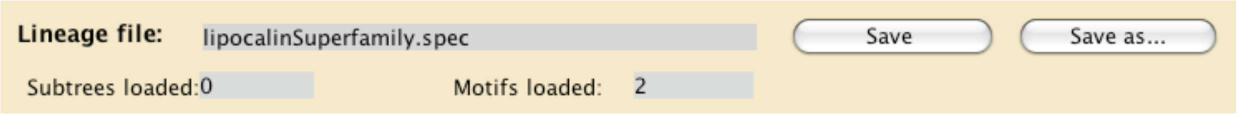
Note that any changes made in subtree and motif options using the **"Edit lineage file"** page will not take effect until the changes are saved in the lineage file. Use **Save** or **Save as…** button to save any changes to the lineage file.

b. Subtree navigation

This section deals with the options for an individual subtree (**Figure 9.14**).

i) **<<** and **>>.** Use these navigation arrows to scroll through the subtrees. Information related to the current subtree is displayed. Error checking is done prior to shifting to the previous or next subtree. Until all errors (*e.g.*, missing values) are corrected, the current subtree cannot be changed.

ii) **Add subtree.** Clicking on this button inserts a subtree at the end of the subtree list. The subtree ID field is auto-filled with a temporary name, which should be changed. The default name is **"new subtree *n*"** where *n* is the number of the subtrees in the file.

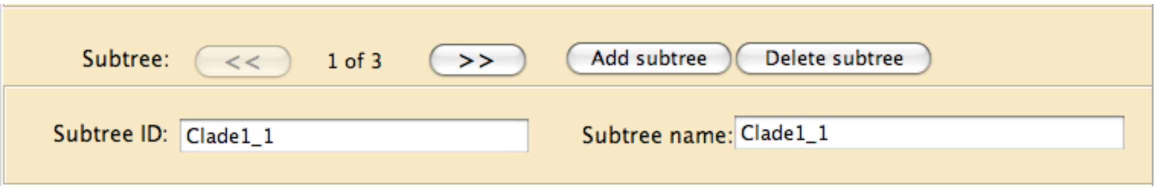


Lineage file: lipocalinSuperfamily.spec

Subtrees loaded: 0

Motifs loaded: 2

Figure 9.13: The top section of the "Edit lineage file" page showing the number of subtrees and motifs. The lipocalinSuperfamily.spec file contains no subtree (clade) specification, so the number of subtrees is 0.



Subtree ID: Clade1_1

Subtree name: Clade1_1

Figure 9.14: The subtree navigation section of the "Edit lineage file" page. For this example, the lineage specification file, exon_intron.spec (found in the **"indel-seq-gen-2.1.03/data"** folder), is used. This lineage specification file contains 3 subtrees (or clades).

iii) **Delete subtree.** Clicking on this button deletes the subtree.

iv) **Subtree ID.** This is mandatory and each ID should be unique within the file. All clades should be identified in the Newick-format guide tree with their IDs. Since no consistency-checking is done, it is for the user to ensure the subtree IDs and their consistency between the guide tree file and the lineage file. Note that a special subtree ID 'root' is used to specify the entire guide tree. This 'root' subtree does not have to be specified in the guide tree.

v) **Subtree name.** This is optional. It can be used to supply a description of the subtree as needed by the user.

Note that a new subtree cannot be saved if no subtree-specific setting is created using either or both of **"Use subtree-specific subsequence options"** or **"Use subtree-specific indel options"** (described next). If neither is checked, an error message will be displayed if you try to save the file or navigate away from the invalid subtree.

c. Subtree subsequence options

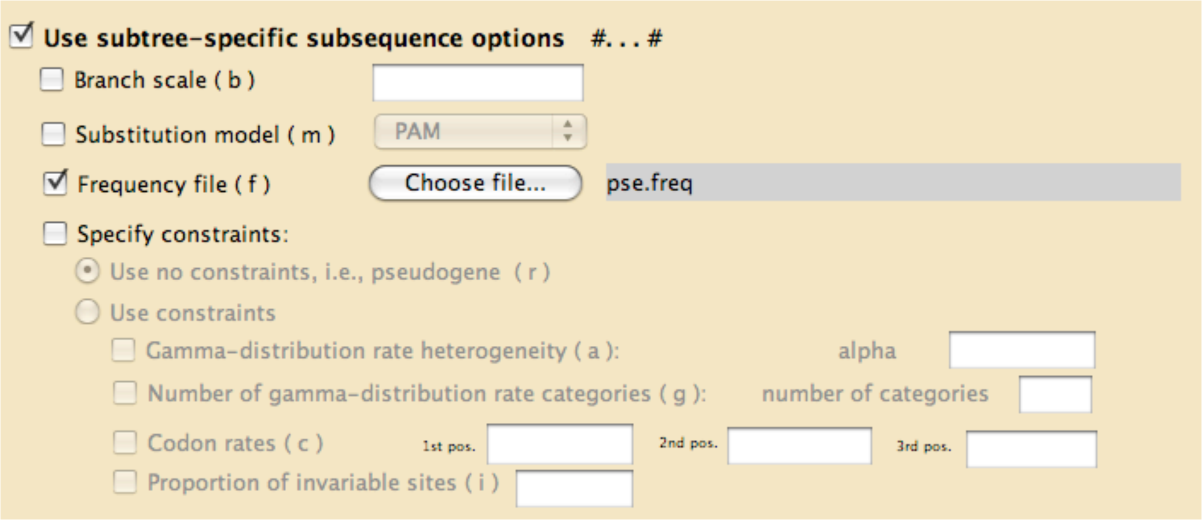
This section deals with overriding the parameter set specified in the **"Basic parameters"** and the **"Advanced parameters"** pages as well as in the partition options set in the **"Edit guide tree file"** page. See the iSGv2 User Manual for details regarding the precedence of these options.

Check the box for the **Use subtree-specific subsequence options** to make changes in the subsequence options (**Figure 9.15**). If this box is not checked, the subsequence settings given in the guide tree file as shown on the **"Edit guide tree file"** page will be used for the subtree. The subtree-specific subsequence options are as follows:

i) **Branch scale.** Check this box to change the scaling factor of the branch length (in a decimal number) for the subtree.

ii) **Substitution model.** Check this box to change the substitution model for the subtree.

iii) **Frequency file.** Check this box to change the amino acid or nucleotide frequencies for the subtree. A frequency file needs to be given.



☑ **Use subtree-specific subsequence options** #...#

□ Branch scale (b)

□ Substitution model (m) PAM

☑ Frequency file (f) pse.freq

□ Specify constraints:

● Use no constraints, i.e., pseudogene (r)

○ Use constraints

□ Gamma-distribution rate heterogeneity (a): alpha

□ Number of gamma-distribution rate categories (g): number of categories

□ Codon rates (c) 1st pos. 2nd pos. 3rd pos.

□ Proportion of invariable sites (i)

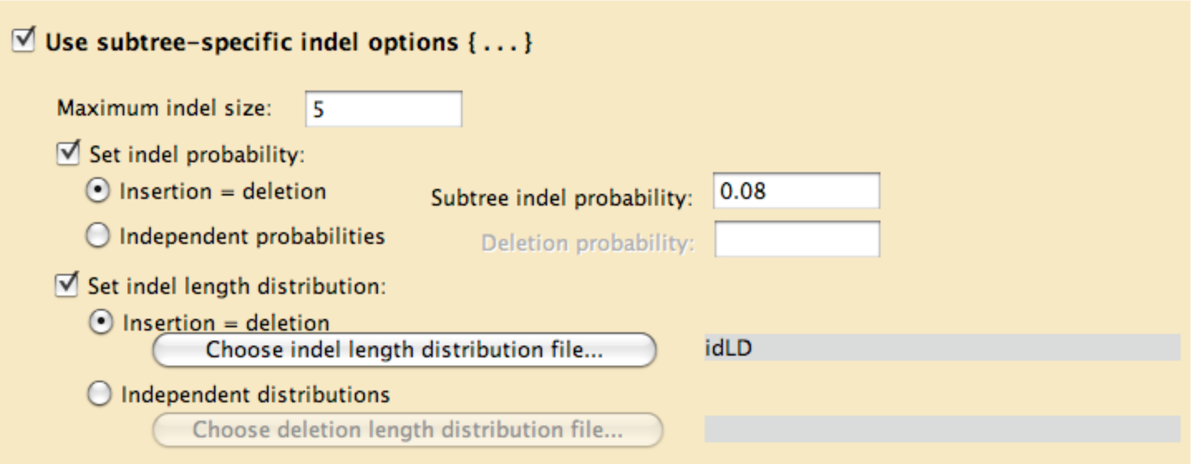
Figure 9.15: The subtree-specific subsequence options in the "Edit lineage file" page. For this example, the lineage specification file exon_intron.spec (found in the **"indel-seq-gen-2.1.03/data"** folder) is used. The frequency file, pse.freq, is applied to this subtree.

iv) **Specify constraints.** Check this box to change how to deal with constraints.

- ***Use no constraints.*** Selecting this option means that a uniform rate is applied to all positions by removing the use of gamma-distribution rate heterogeneity and different rates among codon positions. It effectively simulates a neutral evolution as in a pseudogene for the subtree.
- ***Use constraints.*** Subtree-specific constraints can be applied using this option.
    - ***Gamma-distribution rate heterogeneity.*** Check this box to use the rate heterogeneity and/or change the alpha (shape) parameter for the gamma distribution (in a decimal number) for the subtree.
    - ***Number of gamma-distribution rate categories.*** Check this box to use the discrete gamma-distribution rate categories or change the number of the categories (in an integer from 2 to 32). This option is available only if the gamma-distribution rate option above or the same gamma-distribution rate option in the **"Advanced parameters"** page is selected.
    - ***Codon rates.*** Check this box to use or change the codon position-specific rates (in three decimal values) for the subtree.
    - ***Proportion of invariable sites.*** Check this box to use or change the proportion of invariable sites (in a decimal value) for the subtree.

d. Subtree-specific indel options

This section can be used to change the indel parameters for the subtree.


Choose indel length distribution file...
Choose deletion length distribution file...

☑ Use subtree-specific indel options { . . . }

Maximum indel size: 5

☑ Set indel probability:

● Insertion = deletion Subtree indel probability: 0.08

○ Independent probabilities Deletion probability:

☑ Set indel length distribution:

● Insertion = deletion

idLD

○ Independent distributions

Figure 9.16: The subtree-specific indel options in the "Edit lineage file" page. For this example, the lineage specification file exon_intron.spec (found in the **"indel-seq-gen-2.1.03/data"** folder) is used. For the first subtree, the maximum indel length is set to 5, the indel probability is set to be 0.08 for both insertions and deletions, and the length distribution file idLD is used for both insertions and deletions.

Check the box for the **Use subtree-specific indel options** to make changes in the indel options (**Figure 9.16**). If this box is not checked, the same indel settings from the guide tree file as shown on the **"Edit guide tree file"** page will be used for the subtree.

i) **Maximum indel size.** It changes the maximum length of insertions or deletions for the subtree. To suppress indel occurrence for the subtree, type in 0. The maximum indel size needs to be specified in an integer (0 or larger). You cannot keep this field empty.

ii) **Set indel probability.** If this option is not selected, the indel probability given by Chang and Benner (2004) is used (see the iSGv2 User Manual for more information). If this option is selected, you have the choice of using the same probability for both insertions and deletions (select **"Insertion = deletion"**) or having them different probabilities (select **"Independent probabilities"**). If **"Independent probabilities"** is chosen, both of **"Insertion probability"** and **"Deletion probability"** must be given.

iii) **Set indel length distribution.** If this option is not selected, the indel length distribution given by Chang and Benner (2004) is used. Otherwise, a length distribution can be given in a file that contains a comma-delimited set of values with at least as many values as the maximum indel size specified. Again you have the option of using the same distribution file for both insertion and deletion (**"Insertion = deletion"**) or using separate files (**"Independent distributions"**). See the iSGv2 User Manual for more information.

e. Motifs

Figure 9.17 shows the motif section, which allows you to navigate through the motif information provided in the lineage specification file.

The **"motif marker"** (in **Figure 9.17**, 'b' is used for this motif) must agree with the marker in the MSA or single-sequence root file specified in the guide tree file. Conversely, any marker used in either the MSA or single-sequence root file must be defined in the motif section of the lineage file (see the iSGv2 User Manual for more details).

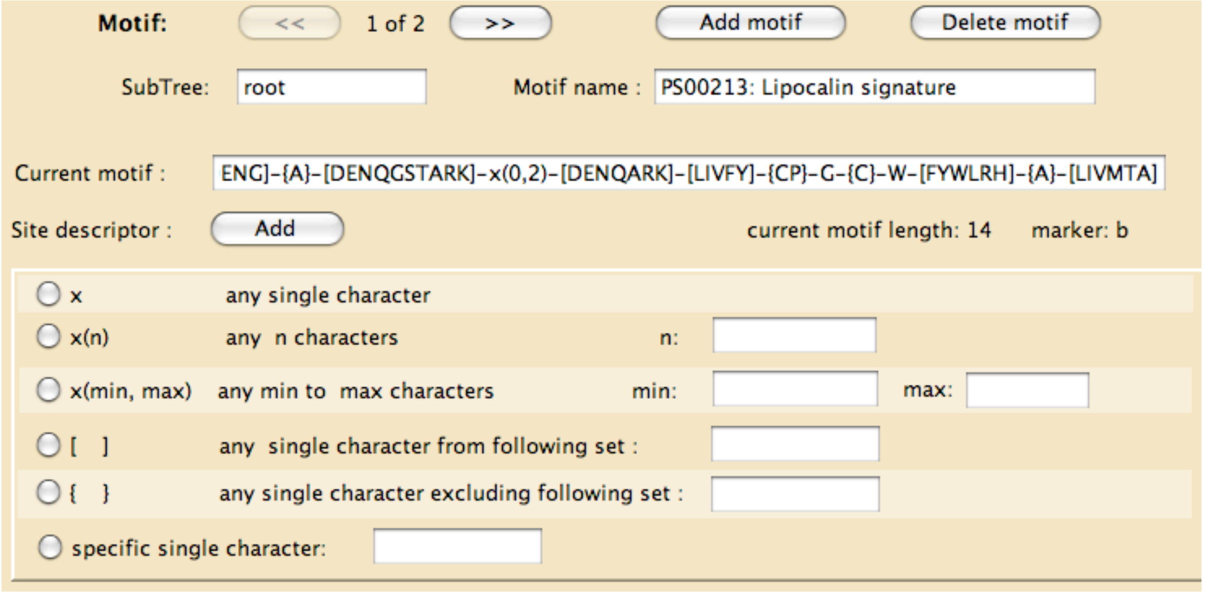
Note that **SuiteMSA** generates a motif marker (a single alphabet) when a new motif is added. The user needs to ensure that the MSA or single-sequence root file uses corresponding motif markers correctly.

i) **<< and >>.** Use these motif navigation arrows to scroll through the motifs specified in the lineage specification file. It displays options used for the current motif. Moving from the current motif to the previous or next motif triggers error checking for proper field values. Any errors must be corrected before moving from the current motif to another motif.

ii) **Add motif.** Clicking on this button inserts a motif at the end of the list.

iii) **Delete motif.** Clicking on this button deletes the current motif.

iv) **Subtree ID.** A motif can be specified for a subtree using its ID (Subtree ID). Use the special subtree ID "root" if the motif applies to the entire tree (see the iSGv2 User



SubTree: root

Motif name : PS00213: Lipocalin signature

Current motif : ENG]-{A}-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-{A}-[LIVMTA]

current motif length: 14 marker: b

○ x any single character

○ x(n) any n characters n:

○ x(min, max) any min to max characters min: max:

○ [ ] any single character from following set :

○ { } any single character excluding following set :

○ specific single character:

Figure 9.17: The motif section of the "Edit Lineage File" page. This section allows editing, adding, or deleting of motifs. In this example, the PROSITE motif contained in the lineage specification file lipocalinSuperfamily.spec is shown.

Manual for more information).

v) **Motif name.** This is an optional field that can be used to describe the motif.

vi) **Motif pattern.** The motif is expressed with the regular expression. The motif pattern can be directly typed in or choose a motif site descriptor and add it by using the **Add** button. The following motif site descriptors are available:

- ***any single character.*** Use this for a site that can take any character. This will add '-x' to the end of the current pattern.
- ***any* n *characters.*** Use this for a site that is made up of any *n* characters. For instance x(3) describes a site made up of any 3 characters.
- ***any* min *to* max *characters.*** Use this for a site that is made up of any *min* to *max* characters. For instance x(3,6) describes a site made up of any 3 to 6 characters.
- ***any single character from the following set***. Use this for a site that is made up of any character given in the set. For instance [WKY] would allow any character from a W, a K, or a Y.
- ***any single character excluding the following set***. Use this for a site that is made up of any character not in the set given. For instance {WKY} would allow any character except a W, a K, or a Y.
- ***specific single character.*** Use this for a site that takes only a single character.

vii) **Current motif length.** It displays the maximum length of the current motif described by the regular expression.

viII) **Motif marker.** It displays the motif marker assigned to the current motif.

9.3.5 Options used for the lipocalin protein superfamily simulation

Followings are the options set in each option panel for the lipocalin protein superfamily simulation. All support files are included in the **"sample"** folder within the **"SuiteMSA_Package"** under the **"SuiteMSA-1.2"** folder. The model alignment (included in lipocalinSuperfamily_alignment.fasta as well as in lipocalinSuperfamily_template.maroot) and phylogeny (included in lipocalinSuperfmaily.tree) are based on those given in Sánchez *et al.* (2003). See Anderson *et al.* (2011b) for more details about how specific values are chosen.

i) **Basic parameters.**

- Guide tree file: lipocalinSuperfamiliy.tree
- Substitution model: WAG

ii) **Advanced parameters.**

- Lineage file: lipocalinSuperfamiliy.spec
- Gamma-distribution rate heterogeneity: alpha = 3.88

- Amino acid frequencies*: A: 0.064, R: 0.046, N: 0.048, D: 0.058, C: 0.02, Q: 0.04, E: 0.079, G: 0.057, H:,0.018, I: 0.047, L: 0.083, K: 0.072, M: 0.022, F: 0.045, P: 0.045, S: 0.064, T: 0.063, W: 0.016, Y: 0.05, V: 0.064

iii) **Edit guide tree file.**

- Frequency file*: lipocalin.freq
- MSA root file: lipocalinSuperfamiliy_template.maroot
- Maximum indel size: 20
- Indel probability: 0.0516702 (insertions = deletions)
- Indel length distribution file: lipocalin.idlen (insertions = deletions)

iv) **Edit lineage file.**

The motif pattern is provided in the lineage file (lipocalinSuperfamily.spec).

*Only one of these settings is needed.

9.4 Running the simulation

After all simulation parameters are entered, click on the **Run iSG** button at the top of the *iSG Simulation* screen (shown in **Figure 9.7**). The simulation will run and when it is completed, the iSGv2 summary window will appear as shown in **Figure 9.18**.

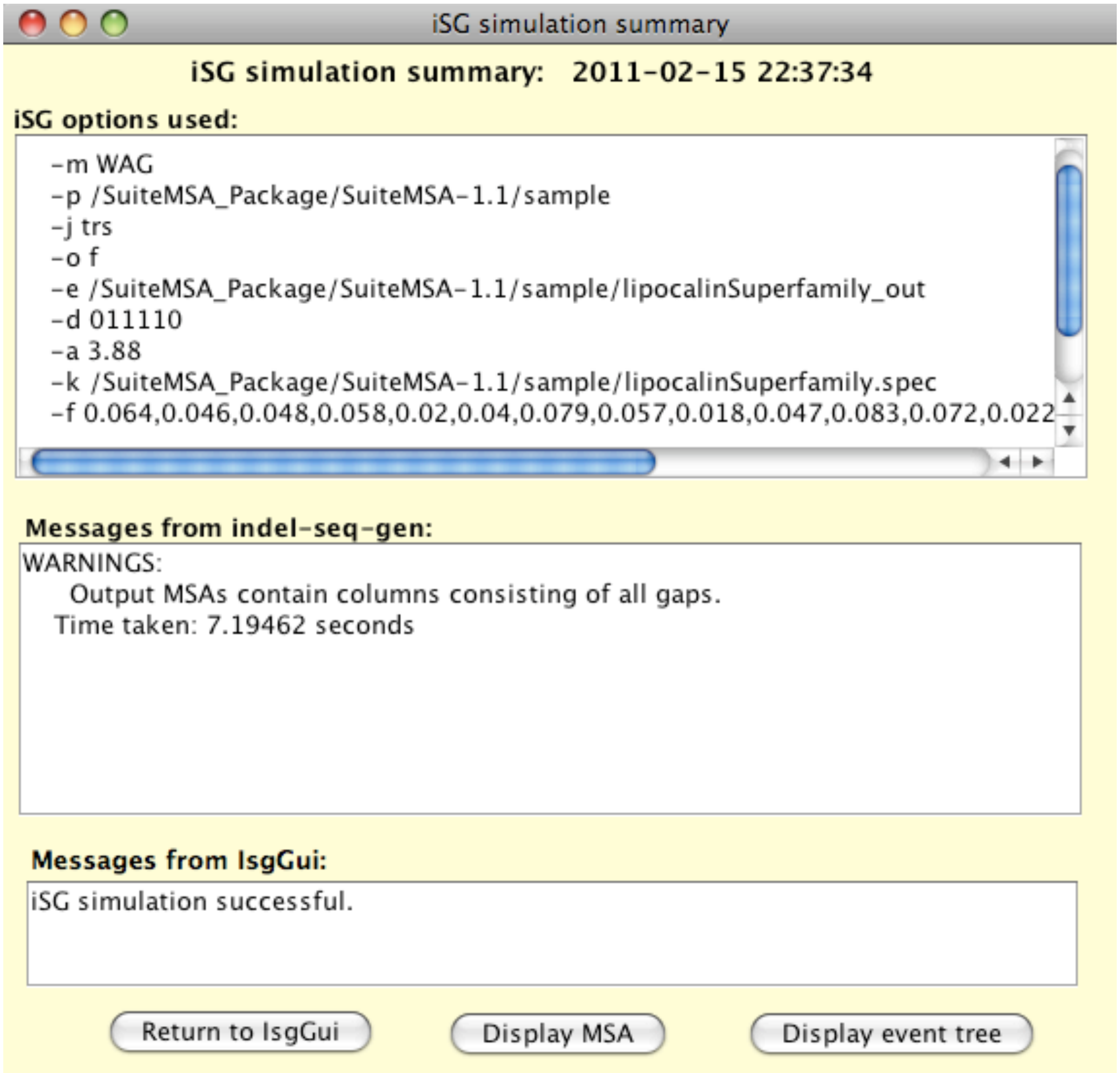
9.4.1 iSG simulation summary window

When the simulation run is complete, the **"iSG simulation summary"** window will appear as shown in **Figure 9.18**. This window contains the information on the options and guide tree file used, the time used for simulation, and any messages from **iSGv2.1** and **SuiteMSA**.

The **"iSG simulation summary"** window has three buttons. The **Return to SuiteMSA** button will bring you back to the main *iSG Simulation* window. The **Display MSA** and **Display event tree** buttons will bring up viewers for the output MSA and simulation guide tree with indel events tracked. See the section 9.5 for these viewers.

Note that if the simulation is done with the **"Discrete evolutionary steps (des)"** option (the **"Stepping method"** option is found in the "**Basic parameters"** page), clicking on the **"Display event tree"** button will produce an error message. The event-tracking on phylogeny is available only if the **"Time relative steps (trs)"** or **"Gillespie algorithm (gil)"** is used for simulation. With any simulation method, however, events can be tracked on the true MSA using the **"Display MSA"** button.

If the display buttons are disabled, it indicates that the simulation did not successfully complete for some reason. Check the error messages from **iSGv2.1** and **SuiteMSA** in the summary window. Note also that if the MSA output format other than FASTA is chosen, the **Display MSA** button will be disabled.



iSG simulation summary: 2011-02-15 22:37:34

iSG options used:

```
-m WAG
-p /SuiteMSA_Package/SuiteMSA-1.1/sample
-j trs
-o f
-e /SuiteMSA_Package/SuiteMSA-1.1/sample/lipocalinSuperfamily_out
-d 011110
-a 3.88
-k /SuiteMSA_Package/SuiteMSA-1.1/sample/lipocalinSuperfamily.spec
-f 0.064,0.046,0.048,0.058,0.02,0.04,0.079,0.057,0.018,0.047,0.083,0.072,0.022
```

Messages from indel-seq-gen:

```
WARNINGS:
    Output MSAs contain columns consisting of all gaps.
  Time taken: 7.19462 seconds
```

Messages from IsgGui:

```
iSG simulation successful.
```

Figure 9.18: The iSG simulation summary window. The "iSG options used" shows the parameters used for the lipocalin protein superfamily simulation.

9.4.2 Notes on a large simulation and how to kill the process

Note that while **iSGv2.1** is running, **SuiteMSA** stops responding to any of mouse clicking. You need to wait until **iSGv2.1** stops running before you can resume working on **SuiteMSA**. For a complex large simulation this could take awhile. The lipocalin superfamily simulation used as an example will require between 5 to 15 seconds depending on the system.

In order to monitor the activity of **iSGv2.1**, use a CPU monitoring program such as 'top' or 'ps' (on Linux-like system). It shows the CPU and memory usage of **iSGv2.1** under the name of 'indel-seq-gen'. To kill the indel-seq-gen process on Linux-like system, make note of the PID (process ID) of indel-seq-gen and use the following Linux command:

```
     kill pid
```

or

```
     kill -9 pid
```

where the *pid* should be replaced with the actual PID for indel-seq-gen. This will kill indel-seq-gen and release **SuiteMSA**, although no error message will appear from indel-seq-gen. On Macintosh, you can also use **"Activity Monitor"** application to monitor and quit the process.

9.5 Viewing the simulation results

9.5.1 Viewing simulation MSAs with event tracking

To view the true MSA produced by the simulation, click on the **Display MSA** button on the **"iSG simulation summary"** screen or *Alignment Viewer* from the **Alignments** menu of the *iSG Simulation* main screen. Using the **Display MSA** button from the summary window will bring up directly the MSA from the last simulation run. All indel events can be traced on the true MSA of simulated sequences. **Figure 6.2** shows the MSA from the lipocalin superfamily simulation. Clicking on a gap symbol (-) color-coded for an insertion or deletion event brings up the event label. Note that the current version of **iSGv2.1** (2.1.03) tracks only insertion and deletion events, but not substitution events. For details on *Alignment Viewer*, see the section 6.

The *MSA Comparator* and *Pixel Plot* are also available in the **Alignments** menu of the *iSG Simulator*. As shown in **Figure 7.2**, the *MSA Comparator* can show indel events color-coded if the MSA is produced by **iSGv2.1** and the event trace file (xxx.trace file) is present.

For details on the *MSA Comparator* and *Pixel Plot*, see the sections 7 and 8, respectively.

9.5.2 Viewing phylogenies with event tracking: *Phylogeny Viewer with events*

Note that, as mentioned before, the *Phylogeny Viewer with events* cannot be used if the simulation is done with the **"Discrete evolutionary steps (des)"** option. If the simulation is done with the **"Time relative steps (trs)"** or **"Gillespie algorithm (gil)"** option, indel events can be tracked on the simulation guide tree using the phylogeny viewer.

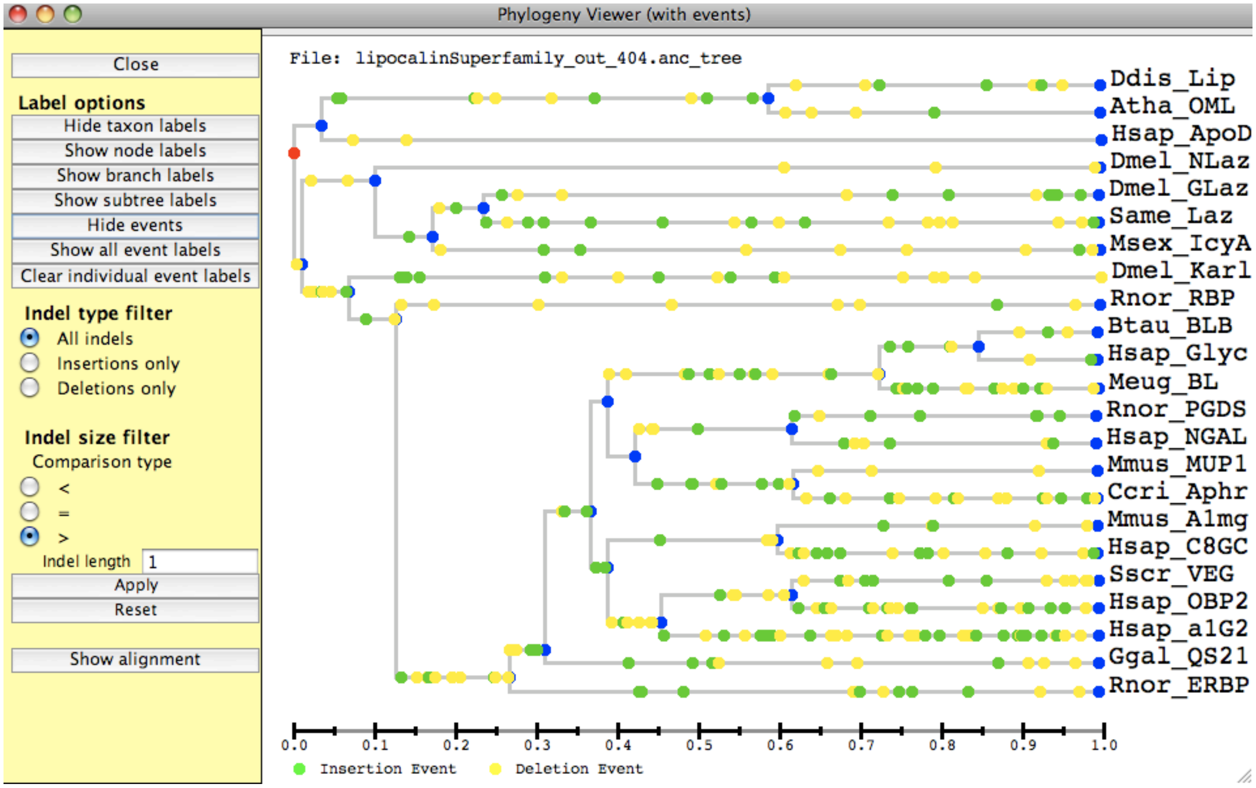
The phylogeny viewers are available by clicking on the **Display event tree** button on the **"iSG simulation summary"** screen or from the **Trees** menu of the *iSG Simulation* main screen. In the **Trees** menu, the first viewer, **Phylogeny Viewer (guide tree, Newick tree)...** can be used to view any trees in Newick format. It does not show indel event mapping. For details on this *Phylogeny Viewer* see the section 10.

For tree files with events associated (generated by **iSGv2.1** simulation), the second viewer, **Phylogeny Viewer (with events)...** can be used to map indel events on the simulation guide tree. The iSG output tree file has the name like 'xxx.anc_tree'. In addition, the event trace file associated to the tree file (xxx.trace) is needed. **Figure 9.19** shows a phylogeny with indel events mapped from the lipocalin superfamily simulation.

The *Phylogeny Viewer with events* has the following display functions.

i) **Label options.**

- ***Hide/Show taxon labels.*** It hides and shows the taxon labels on the tree.
- ***Show/Hide node labels.*** It shows and hides the node labels on the tree.
- ***Show/Hide branch labels.*** It shows and hides the branch length on the tree.
- ***Show/Hide subtree labels.*** It shows and hides the subtree labels on the tree if they are available.
- ***Hide/Show events.*** It hides and shows the indel events on the tree. Insertions and deletions are shown with green and yellow dots, respectively.
- ***Show/Hide all event labels.*** It shows and hides the indel event labels on the tree. This option is available only when events are visible.



File: lipocalinSuperfamily_out_404.anc_tree

Indel type filter
● All indels
○ Insertions only
○ Deletions only

Indel size filter
Comparison type
○ <
○ =
● >
Indel length 1

Figure 9.19: The *Phylogeny Viewer with events*. It shows the guide tree with the indel events generated during the simulation of the lipocalin protein superfamily. The sample output files used, lipocalinSuperfamily_out_404.anc_tree and lipocalinSuperfamily_out_404.trace, are included in the **"iSG_sample_output"** folder.

- ***Clear individual labels.*** It clears the individual indel event labels on the tree that were activated by clicking on the individual event. This option is available only when events are visible.

ii) **Display filter.** Indel events can be filtered using different conditions for display.

- ***Indel type filter***
  - ***All indels.*** No filter is applied and all indel events are shown.
  - ***Insertions only.*** Only insertion events (green dots) are displayed.
  - ***Deletions only.*** Only deletion events (yellow dots) are displayed.
- ***Indel size filter***

  < It displays only the events that are shorter than the length given in the **"indel length"** field.

  = It displays only the events whose lengths are equal to that given in the **"indel length"** field.

  > It displays only the events that are longer than the length given in the **"indel length"** field.

Use the **Apply** button to apply the size filter after the length is changed.

Use the **Reset** button to clear the display filter and show all events regardless of size or type.

iii) **Show alignment**. This button displays the alignment corresponding to the event tree.

iv) **Close.** This button will close the *Phylogeny Viewer*.

10. *Phylogeny Viewer*

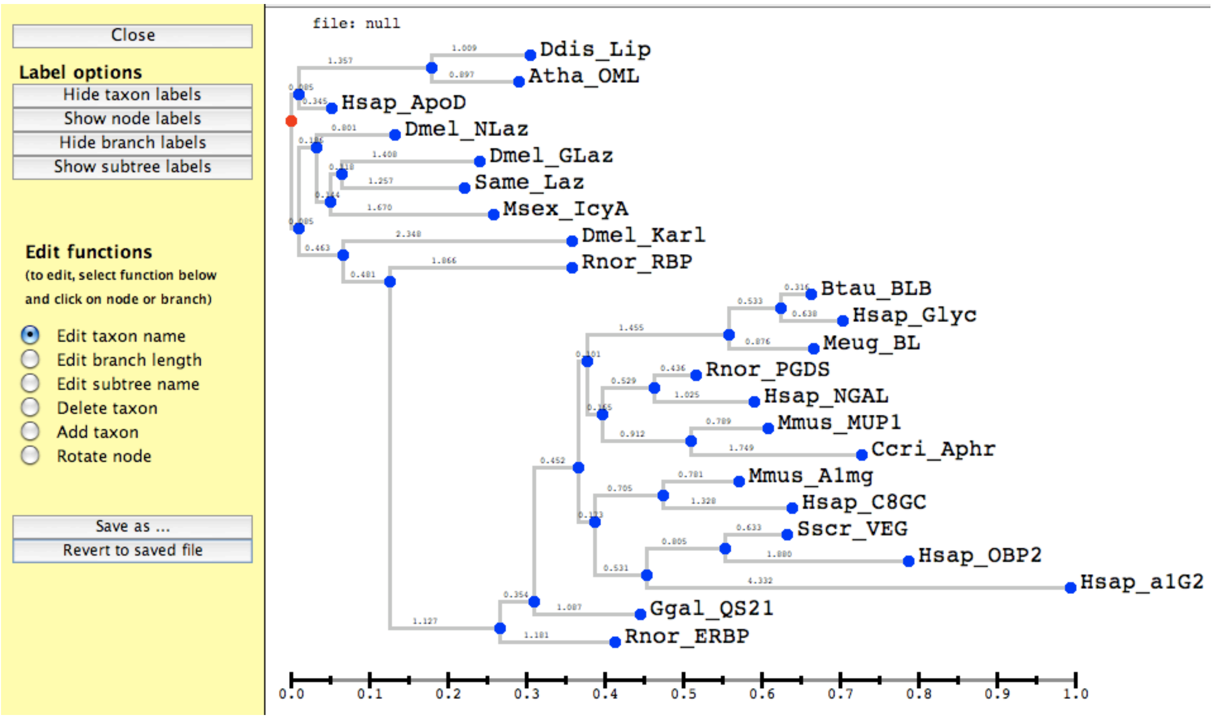
The *Phylogeny Viewer* allows the user to display and edit any trees in Newick format as well as in **iSGv2.1** guide tree format. This viewer is available from the **SuiteMSA** main screen as well as from the **Trees** menu of the *iSG Simulation*.

The *Phylogeny Viewer with events* is also available in *iSG Simulation* **Trees** menu. This viewer is described in the section **9.5.2 Viewing phylogenies with event tracking**.

10.1 Viewing phylogeny

Figure 10.1 shows the *Phylogeny Viewer* displaying the guide tree used for the lipocalin superfamily simulation. The following label-display options are available:

- ***Hide/Show taxon labels*** *button.* Clicking this button will toggle the hiding or displaying of the taxon names.
- ***Show/Hide node labels*** *button.* Clicking this button will toggle the displaying or hiding of the internal node labels.
- ***Hide/Show branch labels*** *button.* Clicking this button will toggle the hiding or displaying of the branch lengths.
- ***Show/Hide subtree labels*** *button.* Clicking this button will toggle the displaying or hiding of the subtree (or clade) labels.



● Edit taxon name
○ Edit branch length
○ Edit subtree name
○ Delete taxon
○ Add taxon
○ Rotate node

Figure 10.1: The *Phylogeny Viewer*. It can be used to display and edit phylogenetic trees in Newick format. In this example, the guide tree used for the lipocalin superfamily simulation is displayed. The guide tree file, lipocalinSuperfamily.tree, is available in the **"sample"** folder.

10.2 Editing phylogeny

The *Phylogeny Viewer* also allows graphical editing of a phylogenetic tree. First choose one editing function from the radio buttons in the side bar. Next choose the taxon (terminal node), internal node, or branch you want to edit. Each editing function is explained below:

- ***Edit taxon name.*** It allows the user to choose a taxon (terminal) node. The selected taxon node will be displayed in green. Enter a new taxon label in the pop-up window.
- ***Edit branch length.*** It allows the user to choose a branch. The selected branch will be displayed in red. Enter a new length in the pop-up window.
- ***Edit subtree name.*** It allows the user to choose a subtree (clade). This is done by clicking on the internal branch leading to the subtree. The selected branch will be displayed in red. Enter a new subtree name in the pop-up window.
- ***Delete taxon.*** It allows the user to delete a taxon (terminal) node. The selected node will be displayed in green. It will delete the selected taxon node as well as the internal node directly above (ancestral). **Figure 10.2A** illustrates how the sibling node of the deleted node will be attached directly to its grandparent node. The terminal branch (Y) is extended by joining it with the internal branch (X) as shown in the figure. Note that internal nodes cannot be deleted.

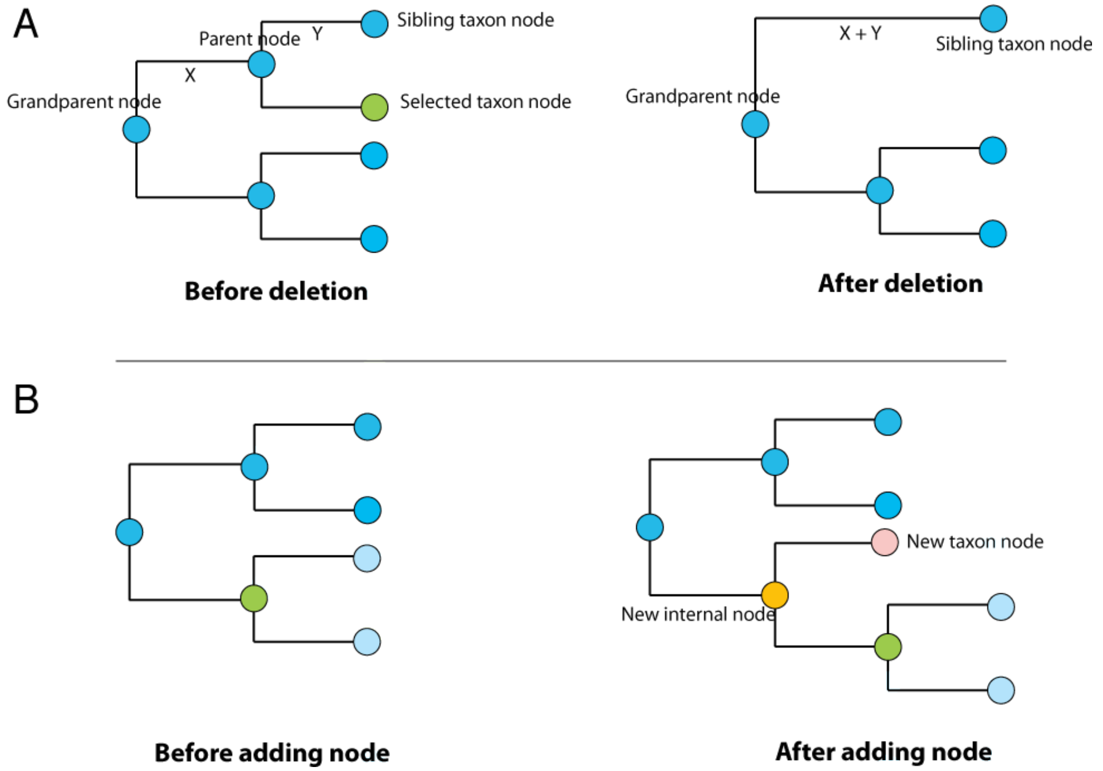


Figure 10.2: Deleting and adding a node to a phylogeny. Deleting a terminal (taxon) node will delete also the parent node (**A**). Adding a node results in the addition of a new internal node (**B**).

- ***Add taxon.*** It allows the user to add either taxon (terminal) or internal node. Choose a node where a new node should be added. The selected node will be displayed in green. After entering the name of the new node in the pop-up window, a new node as well as a new internal and terminal branch will be added as shown in **Figure 10.2B**.
- ***Rotate node.*** It allows the user to rotate a subtree (clade) at the selected internal node. This function has no effect on terminal (taxon) nodes.

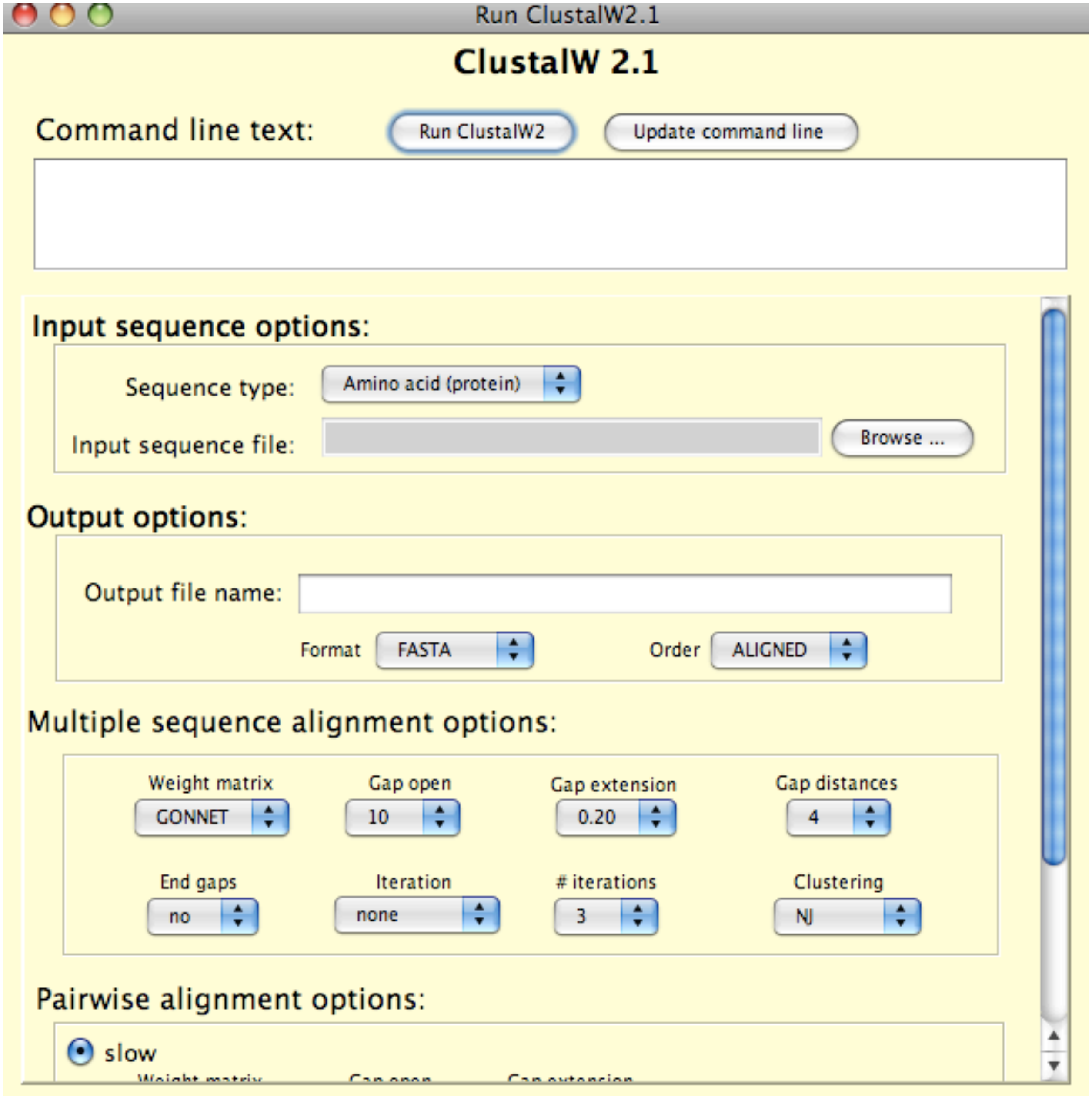
Once the phylogeny is edited, click on the **Save as…** button to save the new tree in a file.

The **Revert to saved file** button will restore the display to the tree last saved in the file.

11. *ClustalW2* GUI

ClustalW2/ClustalX2 is a multiple alignment program developed by Larkin *et al.* (2007). **SuiteMSA** provides a simple GUI similar to **ClustalX2**. The *ClustalW2* GUI is available from the main **SuiteMSA** screen using the **MSA Reconstruction** option.

First you must set the location of the **ClustalW2** executable by selecting the **ClustalW2 location** from the **Setup** menu. Navigate to the folder containing the **ClustalW2**



Command line text:

Input sequence options:

Sequence type: Amino acid (protein)

Input sequence file:

Output options:

Output file name:

Format FASTA

Order ALIGNED

Multiple sequence alignment options:

Weight matrix GONNET

Gap open 10

Gap extension 0.20

Gap distances 4

End gaps no

Iteration none

iterations 3

Clustering NJ

Pairwise alignment options:

● slow

Weight matrix Gap open Gap extension

Figure 11.1: Main screen for *ClustalW2* GUI.

executable and select this folder.

The main *ClustalW2* GUI screen is shown in **Figure 11.1**. Note that more options are available in **ClustalW2**. Only the most commonly used options are included in this interface.

11.1 Input sequence options

i) **Sequence type.** Choose from "Amino acid (protein)" or "Nucleotide (DNA)". The default sequence type is "Amino acid (protein)".

ii) **Input sequence file.** Choose a file containing the sequences to be aligned. When selected, the file name will appear in the **Input sequence file** field.

11.2 Output options

i) **Output file name.** The suggested output file name will be auto-filled into the **Output file name** field, which is based on the input file name with the suffix '_CLTW' attached. This name may be changed simply by overwriting or editing it. The output file will be saved in the same directory where the input file is located.

ii) **Format.** The default output alignment format when running **ClustalW2** from this GUI is FASTA. This is to view the alignments using the various viewers available in **SuiteMSA**. However, the user may select any output format from: CLUSTAL, GCG (MSF), GDE, PHYLIP, PIR, NEXUS, or FASTA.

iii) **Order.** The order of the sequences in the alignment can be chosen from "Aligned" (ClustalW2 alignment order) or "Input" (the input sequence order). The default is the "Aligned" order.

11.3 Multiple sequence alignment options

- **Weight matrix:** choices are GONNET series (default), BLOSUM series, PAM series, and ID (identity matrix) for protein sequences, and IUB (default) and CLUSTALW for DNA sequences.
- **Gap open:** the gap opening penalty ranges from 1 to 100 (default: 10).
- **Gap extension:** the gap extension penalty ranges from 0.05 and 10.0 (default: 0.2).
- **Gap distances:** the gap distance penalized ranges from 0 to 10 (default: 4).
- **End gaps:** the end gap separation penalty can be turned on (yes) or off (no) (default: no).
- **Iteration:** iteration scheme can be chosen from none (no iteration: default), tree (iterations are done at each step of the progressive alignment), or alignment (iterations are done only on the final alignment).
- **# iterations:** the iteration number can be chosen from 1 to 10 (default: 3).
- **Clustering:** clustering method can be chosen from NJ (default) or UPGMA.

11.4 Pairwise alignment options

There are two pairwise alignment modes: fast and slow ("slow" is the default). Each mode has different options available. For detailed description of each option, refer to the ClustalW2 documentation.

11.4.1 Slow mode

The slow/accurate mode is based on the dynamic programming algorithm.

- **Weight matrix:** choices are GONNET (default), BLOSUM, PAM, and ID (identity matrix) for protein sequences, and IUB (default) and CLUSTALW for DNA sequences.
- **Gap open:** the gap opening penalty ranges from 1 to 10 (default: 10).
- **Gap extension:** the gap extension penalty ranges from 0.05 to 10.0 (default: 0.1).

11.4.2 Fast mode

The fast/approximate mode is based on the Wilbur and Lipman's algorithm.

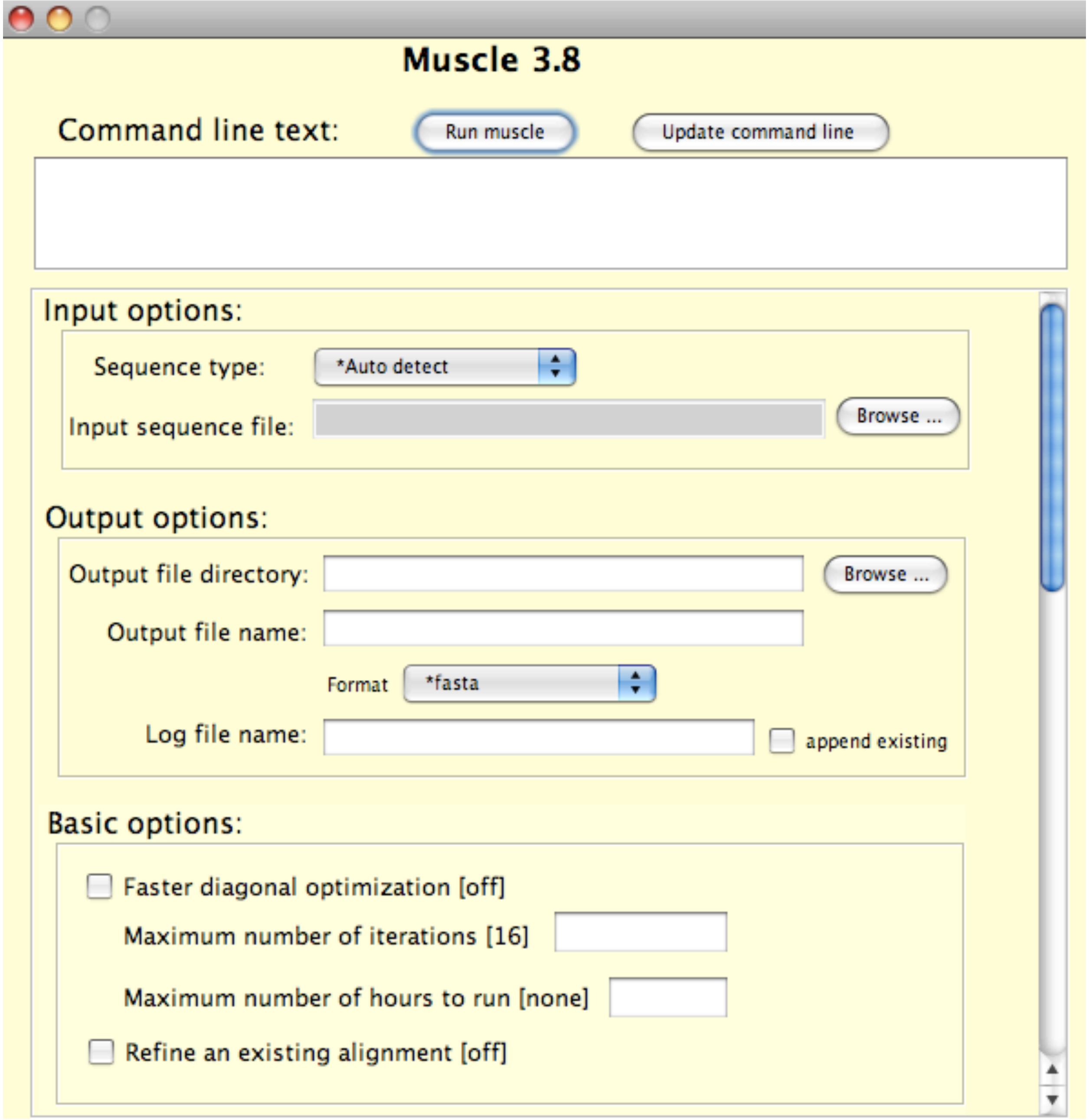
- **KTUP (word size):** the k-tuple size ranges from 1 to 5 (default: 1).
- **Window size:** choose from 0 to 10 (default: 5).
- **Gap penalty:** gap penalty ranges from 1 to 500 (default: 3).
- **Top diag:** number of top diagonals ranges from 1 to 10 (default: 5).
- **Score type:** choose from percent (default) or absolute.

11.5 Alignment viewers

All three of the alignment viewers (*MSA Viewer*, *MSA Comparator*, and *Pixel Plot*) are available from the **Alignments** menu. See the sections 6, 7, and 8 for detailed information on these viewers.

12. *Muscle* GUI

Muscle is a multiple alignment program developed by Edgar (2004). **SuiteMSA** provides a simple GUI for **Muscle v3.8**. The *Muscle* GUI is available from the main **SuiteMSA** screen using the **MSA Reconstruction** option.



Input options:

Sequence type: *Auto detect

Input sequence file:

Output options:

Output file directory:

Output file name:

Format *fasta

Log file name: □ append existing

Basic options:

□ Faster diagonal optimization [off]

Maximum number of iterations [16]

Maximum number of hours to run [none]

□ Refine an existing alignment [off]

Figure 12.1: Main screen for *Muscle* GUI.

First you must choose the **Muscle** executable file by selecting the **Muscle executable** from the **Setup** menu. Navigate to the folder containing the **Muscle** executable file and select the executable.

The main *Muscle* GUI screen is shown in **Figure 12.1**. Note that more options are available in **Muscle**. Only the most commonly used options are included in this interface. For each option, the default value is shown in [ ] or marked with '*' for the drop-down list. For more information on each option, see the MUSCLE User Guide available from http://www.drive5.com/muscle/.

12.1 Input options

i) **Sequence type.** Choose from "Auto detect", "Amino acid (protein)" or "Nucleotide (DNA)". The default sequence type is "Auto detect".

ii) **Input sequence file.** Choose a file containing the sequences to be aligned. When selected, the file name will appear in the **Input sequence file** field.

12.2 Output options

i) **Output file directory.** The suggested output directory name will be auto-filled into the **Output file directory** field, which is based on the input file directory. This directory may be changed simply by overwriting, editing, or browsing for another directory

ii) **Output file name.** The suggested output file name will be auto-filled into the **Output file name** field, which is based on the input file name with the suffix '_Muscle' attached. This name may be changed simply by overwriting or editing it.

iii) **Format.** The default output alignment format when running **Muscle** from this GUI is FASTA. This is to view the alignments using the various viewers available in **SuiteMSA**. However, the user may select any output format from: clw (clustalw), clwstrict (a strict clustalw format including the header), msf (GCG format), html (with color coding for conserved sites), phylip sequential, or phylip interleaved.

iv) **Log file name.** If you wish to maintain a log file, type the name of the log file in the text field provided. If you wish to append the log to the existing file, type this file name and click the **"append existing"** check box. The log file will be saved to the output file directory selected above.

12.3 Basic options

i) **Faster diagonal optimization.** Click the check box to use the faster optimization using the k-mer extension method. This is off in default.

ii) **Maximum number of iterations.** Enter the maximum number of iterations in an integer. The default is 16.

iii) **Maximum number of hours to run.** Default is no time limit. Use a floating-point number of hours.

iv) **Refine an existing alignment.** This option refines the already aligned input file. It skips the first two iterations and begins the tree-dependent refinement. The default setting is off.

12.4 Scoring

Profile Score can be selected from the following options.

i) For a protein alignment:

- **le:** log-expectation profile score using the VTML240 matrix (default)
- **sp:** sum-of-pairs protein profile score using the PAM200 matrix
- **sv:** sum-of-pairs protein profile score using the VTML240 matrix

ii) For a nucleotide alignment:

- **spn:** sum-of-pairs nucleotide profile score based on BLASTZ. This is the default and only option for nucleotide alignments.

12.5 Iteration related options

There are three major steps in Muscle, and they are called as "iterations". In the first iteration, a guide tree is built based on pairwise distances calculated using k-mer counting, In the second iteration, pairwise distances are estimated using Kimura's method based on the multiple alignment. The third and later iterations are called as "Tree-dependent refinement".

i) **1st and 2nd iterations.**

- **Clustering method.** The guide tree can be built using the following methods:
  - upgmb (default)
  - upgma,
  - neighbor-joining
- **Distance (first iteration only).** K-mer counting is used to estimate distances for the first iteration. The choices are:
  - kmer6_6 (default for protein sequences)
  - kmer20_3
  - kmer20_4
  - kbit20_3
  - kmer4_6 (default for nucleotide sequences)
- **Tree rooting method.** The guide tree is rooted using one of the following methods:
  - pseudo: the last node created is used as the pseudo root (default).
  - midlongestspan: locating the root at the mid-point of the longest span.

  - minavgleafdist: locating the root to minimize the average branch weight.
- **Sequence weighting scheme.** Sequences are weighted to correct for the effects of biased sampling from a group of similar sequences. Available methods are:
  - clustalw: the method used in ClustalW (default).
  - henikoff: from Henikoff and Henikoff (1994).
  - henikoffpg: a variant of the Henikoff & Henikoff method used in PSI-BLAST.
  - threeway: from Gotoh (1995).
  - gsc: from Gerstein *et al.* (1994).
  - none: no sequence weighting is used.

ii) **Tree dependent refinement.** These are the options for the third and later iterations.

- **Clustering method.** The guide tree can be built using the following methods:
  - upgmb (default)
  - upgma,
  - neighbor-joining
- **Distance (iteration 2 and later refinement).** Distance estimation is done based on multiple alignments. The choices are:
  - pctid_kimura (default)
  - pctid_log
- **Tree rooting method.** The guide tree is rooted using one of the following methods:
  - pseudo: the last node created is used as the pseudo root (default).
  - midlongestspan: locating the root at the mid-point of the longest span.
  - minavgleafdist: locating the root to minimize the average branch weight.
- **Sequence weighting scheme.** Sequences are weighted to correct for the effects of biased sampling from a group of similar sequences. Available methods are:
  - clustalw: the method used in ClustalW (default).
  - henikoff: from Henikoff and Henikoff (1994).
  - henikoffpg: a variant of the Henikoff & Henikoff method used in PSI-BLAST.
  - threeway: from Gotoh (1995).
  - gsc: from Gerstein *et al.* (1994).
  - none: no sequence weighting is used.
- **Use anchor optimization.** Anchor optimization is used in default to speed up the refinement. Uncheck it to turn it off.
- **Minimum anchor spacing.** Enter an integer value to set the minimum spacing between anchor columns. (default: 32)
- **Minimum column score for an anchor.** Enter a decimal value to set the minimum score for a column to be an anchor. Default is dependent on the profile score function.

12.6 Other options

i) **Maximum distance between diagonals.** Enter an integer value to set the maximum distance between two diagonals to be merged. (default: 1)

ii) **Minimum length of diagonals.** Enter an integer value to set the minimum length of a diagonal. (default: is 24).

iii) **Diagonal end positions deleted.** Enter an integer value to set the number of positions at diagonal ends to be discarded. (default: is 5).

iv) **Hydrophobic window size.** Enter an integer value to set the window size to find a hydrophobic region. (default: 5)

v) **Gap penalty multiplier for hydrophobic regions.** Enter a decimal value to set the multiplier for gap open/close penalties in hydrophobic regions. (default: 1.2)

vi) **Use guide tree.** Check this box to use a given tree as the guide tree, and choose the file that includes the guide tree in Newick format. (default: off)

13. References

Anderson CL, Strope CL, Moriyama EN: **Assessing multiple sequence alignments using visual tools.** ***In Bioinformatics/ Book 1 (ed. Mahdavi, MA)*** 2011a, InTech, in press.

Anderson CL, Strope CL, Moriyama EN: **SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation**. *BMC Bioinformatics* 2011b, **12**:184.

Chang MS, Benner SA: **Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments.** *J Mol Biol* 2004, **341**:617-631.

Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004*,* **32**:1792-1797*.*

Gerstein M, Tsai J, Levitt M: **The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra**. *J Mol Biol* 1995, **249**:955-966.

Gotoh O: **A weighting system and algorithm for aligning many phylogenetically related sequences**. *Comput Appl Biosci* 1995*,* **11**:543-551.

Henikoff S, Henikoff JG: **Position-based sequence weights**. *J Mol Biol* 1994, **243**: 574-578.

Jones DT: **Improving the accuracy of transmembrane protein topology prediction using evolutionary information**. *Bioinformatics* 2007, **33**:538-544.

Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program***.* *Brief Bioinform* 2008, **9**:286-298.

Kyte J, Doolittle R: **A simple method for displaying the hydropathic character of a protein**. *J Mol Biol* 1982, **157**:105-132.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**:2947-2948.

McGuffin LJ, Bryson K, Jones DT**: The PSIPRED protein structure prediction server.** *Bioinformatics* 2000 **16**:404-405.

Pei J, Grishin NV: **PROMALS: towards accurate multiple sequence alignments of distantly related proteins.** *Bioinformatics* 2007, **23:**802-808**.**

Procter JB, Thompson JD, Letunic I, Creevey C, Jossinet F & Barton JG: **Visualization of multiple alignments, phylogenies and gene family evolution***. Mol Biol Evol* 2010, **7**:S16 - S25.

Sánchez D, Ganfornina MD, Gutiérrez G, Marín A: **Exon-intron structure and evolution of the lipocalin gene family**. *Mol Biol Evol* 2003, **20**:775-783.

Schneider TD, Stephen RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**:6097-6100.

Strope CL, Abel K, Scott SD, Moriyama EN: **Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0***. Mol Biol Evol* 2009, **26**:2581-2593.

Xia X, Xie Z: **Protein structure, neighbor effect, and a new index of amino acid dissimilarities.** *Mol Biol Evol* 2002, **19**:58–67.