indel-Seq-Gen v2.0.3 Manual

Cory L. Strope*, Computer Science and Engineering,
Kevin Abel, Computer Science and Engineering,
Stephen D. Scott, Computer Science and Engineering,
Etsuko N. Moriyama, School of Biological Sciences and the Center for Plant Science Innovation

University of Nebraska – Lincoln

March 16, 2010

Contents

*Corresponding Author, University of Nebraska, email: corystrope@gmail.com

1 Acknowledgments

We would like to thank the authors of Seq-Gen, Drs. Andrew Rambaut and Nick Grassly, which provided the starting point for the development of iSG. As such, when citing indel-Seq-Gen, please note that Seq-Gen [4] should also be cited. The bibtex citation is:

```
@ARTICLE{Rambaut97,
        AUTHOR = "A. Rambaut and N.C. Grassly",
        TITLE = "{Seq-Gen}: an application for the {Monte Carlo} simulation of
{DNA} sequence evolution along phylogenetic trees",
        JOURNAL = "{CABIOS}",
        VOLUME = 13,
        NUMBER = 3,
        YEAR = 1997,
        PAGES = "235-238"       }
```

We are also grateful for the NSF AToL grant 0732863 and Department of Education grant P200A040150 for funding this work.

2 Introduction

indel-Seq-Gen is a tool to simulate the evolutionary events of highly diverged DNA (coding and non-coding) and protein sequences. Long-term evolution often include dynamic changes such as insertions and deletions (indels), but some subsequences, such as domains and motifs, retain their original sequences and structures better than less functionally important regions. indel-Seq-Gen allows for the simulation of many different evolutionary patterns over different regions of a sequence, in the end outputting the "true" multiple alignment of the sequences. indel-Seq-Gen also has multiple unique features, including input a multiple alignment for the root sequence[1], placing functional constraints on sequence sites, tracing mutations along the simulation and noting their positions in the true multiple alignment, changing evolution parameters and site conservation between subtrees (i.e., lineages), and more. In addition iSGv2.0.3 fixes a fundamental flaw in representing in representing insertions and deletions through the introduction of evolving along the guide tree in discrete steps. Finally, iSGv2.0.3 also introduces a novel method of representing conservation patterns with respect to insertions versus deletions versus substitutions. These features can be used in many evolutionary studies, such as to test the accuracy of multiple alignment methods, phylogenetic methods, evolutionary hypotheses, ancestral sequence reconstruction methods, and superfamily classification methods.

3 Getting Started

indel-Seq-Gen v2.0.3 (iSGv2.0.3) is freely downloadable at `http://bioinfolab.unl.edu/~cstrope/iSG/`. iSGv2.0.3 has been tested MacOS X (versions 10.4.7 and Leopard). You will need `g++` compiler if the provided executables do not run on your system.

1. Download the `indel-seq-gen-2.0.3.tar.gz` source archive, open the archive by typing the two commands:

   `gunzip indel-seq-gen-2.0.3.tar.gz`, and then

   `tar -xvf indel-seq-gen-2.0.3.tar`

2. This creates a directory `indel-seq-gen-2.0.3` with all of the files. Go to this directory `cd indel-seq-gen-2.0.3`. Make the executables of indel-seq-gen by typing the command "`./configure`", followed by the command "`make`". For further installation instructions, refer to the file `INSTALL`.

3. After making the executables, type the command "`cp src/indel-seq-gen data/`". This copies the executable into the `data` directory.

4. Go to the `data` directory using the command "`cd data/`".

5. In the `data` directory, there are sample files that can be used to run indel-seq-gen. The following commands will execute indel-seq-gen using various capabilities. Cut-and-paste (or type) these examples to run the simulations. After each run, you can examine the output of these runs by opening the filenames that begin with the name following the `--outfile` or `-e` flags:

```
./indel-seq-gen --matrix HKY --outfile DNA_out < simple_nuc.tree
./indel-seq-gen --matrix HKY --codon_rates 0.2,0.05,0.75 --outfile DNA_out < simple_nuc.tree
./indel-seq-gen -m HKY -e mid_noncoding --num_runs 5 < mid_nuc_noncoding.tree
./indel-seq-gen -m HKY -e mid_noncoding -n 5 -c 2,1,8 --invar 0.02 < mid_nuc_coding.tree
./indel-seq-gen -m HKY -c 2,1,8 -e exon_intron --step_type trs < exon_intron.tree
./indel-seq-gen -m HKY -c 2,1,8 --lineage exon_intron_lineage.spec < exon_intron_lineage.tree
./indel-seq-gen -m JTT -k lipocalin.spec -w a  --alpha 1.3 < lipocalin.tree
./indel-seq-gen -m JTT -k lipocalin_ma.spec < lipocalin_ma.tree
```

   For further understanding of these examples, refer to Section 8.

[1] This can be used to create different but related ancestral sequences in each run

4 Overview

In order to simulate heterogeneous sequences and lineage- and site-specific iSGv2.0.3 uses *enviroments*. The environment of simulation is comprised of substitution parameters (Θ; four components: $\theta^m \rightarrow$ substitution matrix , $\theta^f \rightarrow$ character frequencies, $\theta^r \rightarrow$ site rates, and $\theta^i \rightarrow$ percent invariable sites) and indel parameters (Λ; three components: $\lambda^p \rightarrow$ probability of an indel occurring as the number of indels per subsitution, $\lambda^l \rightarrow$ length distribution, and $\lambda^m \rightarrow$ maximum length). iSGv2.0.3 implements four environments:

1. *Global environment*: Sets the default substitution parameters for the simulation run (indel parameters cannot be specified globally).
2. *Partition environment*: Defines the partition-specific simulation parameters.
3. *Subtree environment*: Defines the lineage-specific simulation parameters.
4. *Motif environment*: Defines site-specific simulation parameters.

Besides the motif environment, these environments define the same sets of parameters. For this reason, iSGv2.0.3 implements precedence for each variable, where the precedence, from lowest to highest, is global, partition, subtree. If any components of the substitution or indel parameter ar not changed in a higher precedence environment, that component inherits the settings from the next lower precedence environment. The motif environment does not conflict with the substitution and indel settings of the other environments, as it works specifically on sites. Precedence and settings can be found in Table 1, and an example simulation run using the precedence rules is shown in Figure 1.

Table 1: *Listing of the precedences of environments and the sections in hich environments are explained.*

		Precedence		
Environment	Set in	Θ[1]	Λ[1]	Section
Global	Command line	Low	N/A	5.1
Partition	Guide tree file	Middle	Low	5.2
Subtree	Specification file	High	High	5.3
Motif	Specification file	N/A	N/A	5.3

[1] $\Theta = \{\theta^m, \theta^f, \theta^r, \theta^i\}$ are the substitution parameters, as defined in Section 4.

[2] $\Lambda = \{\lambda^p, \lambda^l, \lambda^m\}$ are the indel parameters, as defined in Section 4.

Table 2 gives the precedence overriding options for each of the environments.

Table 2: *Environment options and the options that override them for the different input files to iSGv2.0.3 .*

		Lowest ⟵ Precedence ⟶ Highest		
		Global[1]	Subtree[2]	Lineage[3]
Θ:	θ^f	`-f (--frequencies) <float_list>`	`#<filename>#`	`#f<filename>#`
	θ^m	`-m (--matrix) <matrix>`	`#m<matrix>#`	`#m<matrix>#`
	θ^r	`-g (--num_gamma_cats) <int>`	`#r#`	`#r#` OR `#g<int>#`
		`-a (--alpha) <float>`	`#r#`	`#r#` OR `#a<float>#`
		`-c (--codon_rates) <float_list>`	`#r#`	`#r#` OR `#c<float_list>#`
	θ^i	`-i (--invariable) <float>`	`#r#` OR `#i<float>#`	`#r#` OR `#i<float>#`
Λ:[4]	λ^m	N/A	{ * , - , - }	{ * , - , - }
	λ^p	N/A	{ - , * , - }	{ - , * , - }
	λ^l	N/A	{ - , - , * }	{ - , - , * }

[1] Global options are set at the command line. For the complete list of global options, see Table 3.

[2] Subtree options are set in the guide tree file, explained in Section 5.2. For a complete list of options for the guide tree file, see Table 4.

[3] Lineage options are set in the specification file, explained in Section 5.3. For a complete list of options for the specification file, see Table 4 for indel options (Λ) and Table 5 for substitution options (Θ).

[4] Setting indel parameters is the same for both the subtree file and the specification file. The '*' indicates where the option is set. For formatting options, see Table 4.

5 iSGv2.0.3 Parameters

Creating realistic sequences requires many parameters. Because of this, iSGv2.0.3 has many options. This section begins by describing both for the global simulation run and for subsequences, called *partitions*.

5.1 Global Options

To run indel-Seq-Gen from the command prompt, type the following line:

```
indel-seq-gen -m <matrix> [-options] < guide_tree_file > outfile
```

Notes:

- This assumes that `indel-seq-gen` is in your `PATH` and data files locations are specified relative to your present working directory,
- The `guide_tree_file` is described in Section 5.2
- The "> `outfile`" portion of the command specifies that all non-error messages generated by `indel-Seq-Gen` will be saved in `outfile`. Otherwise, messages will be displayed on the screen.

(A)

```
[:exon(,,)]"Partition_1"{9,0.031/0.01,codonLD}
(
 (Taxon1:0.3,Taxon2:0.14
 )Clade1_1:0.5,
 (Taxon3:0.34, Taxon4:0.5
 ):0.12
);
[:intron]"Partition_2"#r,b0.5#{9,0.1,idLD}
(
 (Taxon1:0.3,Taxon2:0.14
 )Clade1_2:0.5,
 (Taxon3:0.34, Taxon4:0.5
 ):0.12
);
```

(B)

```
LINEAGES =
{
   Clade1_1:#fpse.freq,r#{5,0.08,idLD};
   Clade1_2:#p#{8,0.08,idLD};
   Taxon3(2):#a 0.7#{0,0};
}
```

(C)

```
          00000000000000000000000000000000000000000
Taxon1    TTACTT---TTTCCTAACCGG------CCCGAGCTAATG
Taxon2    TTAGTTTTTAATCCCAACCTG------CCCCATCTGATA
Taxon3    CGAAAT------CTTACCGGAGGTAATACGCGGCGCATC
Taxon4    CTGCAT------TCAACCTGT------TAGCATCGCATC
Rootseq:  CTACAT    TATCCTAACCGG      CCCCAGCGCATC
```

(D)

```
30
33000000000000100000000000033
GUTTCAGGTAAAATGCANNGACTTAGRYAG
```

(E)

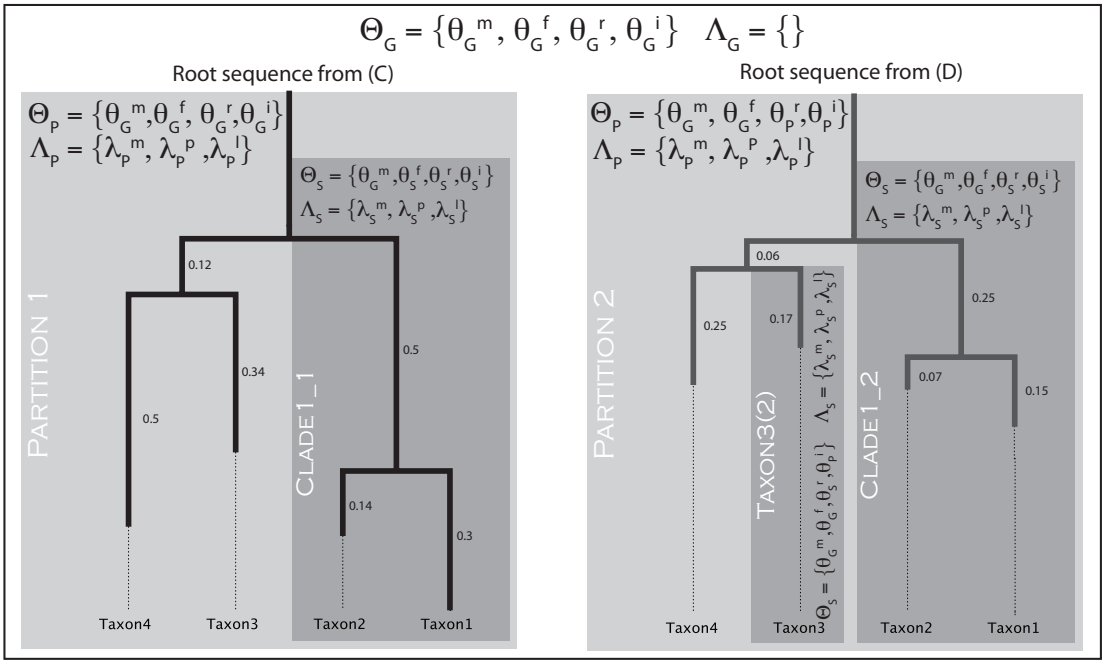

Figure 1: *Depiction of simulation parameters and environments using the exon_intron_lineage example above, "`./indel-seq-gen -m HKY -c 2,1,8 --lineage exon_intron_lineage.spec < exon_intron_lineage.tree`". (A) The `exon_intron_lineage.tree` file specifies two partitions, "Partition 1" and "Partition 2". (B) The `exon_intron_lineage.spec` file specifies environment changes for the subtrees in (E). (C) The file `exon`, a multiple alignment input for partition 1. The final root sequence is shown below the alignment in italics. (D) The file `intron`, a root sequence input for partition 2. (E) The simulation run, showing the substitution (Θ_X) and indel (Λ_X) settings, where X specifies which environment the settings come from: $G \rightarrow$ global, $P \rightarrow$ partition, and $S \rightarrow$ subtree.*

`indel-Seq-Gen` has many options. We suggest that you run some of the examples shown in Section 8. All necessary files are included in the provided `indel-seq-gen-2.0.3/data` directory. If you get any error message and the program does not run, please make sure if you followed the steps explained in the previous section. If you still cannot run the program, please contact us. The contact information is found in the first page.

Table 3: *Global options (entered at the command line) and their effects for the indel-Seq-Gen run. Subsequence options (described in the next Section) will override the global options if there are conflicts. For input type {list},* **do not** *use spaces to separate list items.*

Option	Long Option	Type	Effect	EV[a]
-a	--alpha	{*float*}	Shape (alpha) for gamma rate heterogeneity. Used only for DNA sequences.	θ^r
-b	--branch_scale	{*float*}	Scaling factor for all branch lengths in simulation guide tree [default = 1.0].	
-c	--codon_rates	{*list*}	#1, #2, #3 = rates for codon position heterogeneity. Example: -c 0.15,0.05,0.8. Numbers are not required to sum to 1, as iSGv2.0.3 will normalize them. Used only for DNA sequences.	θ^r
-d	--select_outputs	{*0—1*}*[6]*	Six binary numbers used to select which files to output. The file selection order is {root, seq, ma, tree, trace, verb}, i.e., to select only the ".ma" file, use option -d 001000 on the command line. *Must be used in conjunction with -e option.*	
-e	--outfile	{*string*}	Base filename for output files: <filename>.root, <filename>.seq, <filename>.ma, and <filename>.trace. These files will be created to hold the root sequences, sequence files, multiple alignments, and traced events output from the run, respectively.	
-f	--frequencies	{*list*}	Frequencies for individual amino acids (20, ARNDCQEGHILKMFPSTWYV) or nucleotides (4, ACGT), separated by commas, with **no spaces** between. Use "e" to for equal frequencies (0.05 or 0.25 for each state, respectively) [default = use frequencies based on the substitution matrix]. Example: -f 0.2,0.2,0.3,0.3 for 20% A and 20% C.	θ^f

Continued on next page...

Table 3: (continued)

Option	Long Option	Type	Effect	EV[a]
-g	--num_gamma_cats	{*int*}	Number of categories for the discrete gamma-distribution rate heterogeneity. Must be between 2 and 32 [default = none]. Higher numbers of discrete categories severely affect the speed of iSGv2.	θ^r
-h	--help		Output the usage instructions	
-i	--invar	{*float*}	Proportion of invariable sites [default = none].	θ^i
-k	--lineage	{*filename* }	Subtree specification file name. See Section 5.3.	
-j	--step_type	{*des, trs, gil*}	des = discrete evolutionary steps [default], trs = time relative steps, gil = Gillespie algorithm used for indel creation (fast). These are described below, in Section 5.1.1.	
-m	--matrix	{*string*}	Substitution matrix: HKY, F84, GTR for nucleotides, PAM, JTT, MTREV, CPREV, GENERAL for amino acids.	θ^m
-n	--num_runs	{*int*}	The number of datasets to simulate for each tree [default = 1].	
-o	--outfile_format	{*char*}	Output format: either Phylip (p), NEXUS (n), or FASTA (f) [default = Phylip].	
-q	--quiet		Quiet mode: only the root sequence, resultant sequences, and multiple alignment are printed.	
-r	--rel_rates	{*list*}	Six comma-separated numbers specifying general rate matrix, in the following order: A to C, A to G, A to T, C to G, C to T and G to T. DNA only.	θ^{m} [b]
-s	--option_width	{*int*}	The number of residues per line on the multiple alignment output [default = 60].	
-t	--tstv	{*float*}	Transition-transversion ratio.	

Continued on next page...

Table 3: (continued)

Option	Long Option	Type	Effect	EV[a]
-u	--indel_fill	{*string*}	Indel fill model, based on neighbor effects (preferring certain amino acids based on the neighboring amino acids): xia = original from [6], built on the E. coli K-12 proteins, sp = swiss-prot, ran = no neighbor effect [default = ran].	
-w	--write_anc		Write ancestral sequences.	
-z	--rng_seed	{*int*}	Manually set the random number seed.	
-1	--proportion_motif	{*float*}	The proportion of a random root sequence to be set as PROSITE regular expression motifs (as of PROSITE v20.60, 1038 patterns).	

[a] EV refers to the environment variable of which the global option is categorized. If left unspecified, the option does not refer to any of the environment variables. Environment variables are explained in Section 4.

[b] Relative rates apply to the GTR matrix, not site rates.

5.1.1 Discrete steps and Trace file

iSGv2.0.3 has two discrete-stepping methods and one non-discrete stepping method. Discrete stepping methods are implemented because (i) There is currently no continuous model for indel evolution, and thus most current simulation methods simulate with a flawed indel representation, and (ii) discrete steps allow for concrete tracking of events during the simulation (refer to [5]). However, such methods are slow, and impractical for very large simulation runs. Thus, iSGv2.0.3 also includes a new non-discrete stepping method that also fixes the flaw in indel representations. These methods are described below.

Discrete Stepping Methods:

- The Discrete Evolutionary Steps (DES) model simply calculates ϵ, and simulates sequences along the guide tree in ϵ step sizes.
- Event occurrence times can be related along the simulation guide tree with the *time relative steps*, or TRS, representation. iSGv2.0.3 scales the branch lengths of the guide trees and branch-specific ϵ's so that all mutation events that occur in simulation run are ordered with respect to the relative time of the run (i.e., the simulation starts at relative time point 0.0, and all taxa end at time point 1.0).

Non-Discrete Steps: The DES and TRS methods of simulation are computationally intensive, and can take a very long time for large datasets. For this reason, `indel-Seq-Gen` also allows the user to choose an indel creation procedure that utilizes the Gillespie algorithm (GIL), as described by Cartwright [1]. Using the GIL of indel creation greatly speeds up the the simulation time, and the output of such runs is comparable to the TRS method (i.e., the indel events are still traced). **WARNING: If you choose to change the amino acid/nucleotide frequencies along lineages, DO NOT use this option. Use either DES or TRS.** The DES and TRS options will immediately begin shifting the amino acid/nucleotide distributions, while the GIL option will shift the distributions gradually.

Event Tracking

A benefit introduced by the above methods is the ability to track indel events by the following properties:

1. ID
2. Relative time of occurrence (from 0.0 (root of guide tree) to 1.0 (tips of guide tree); for the TRS model),
3. Set of taxa affected by change (i.e., subtree of occurrence),
4. Event type (insertion/deletion),
5. Indel length, and
6. Locations (columns) in the output MSA that reflect the event.

The time of occurrence for events is listed in different ways for each model: DES reports events by sequence partition (specified in the `guide_tree_file` in the order they occur. TRS and GIL reports events by the relative time they occur in the guide tree.

5.2 Partition Options

Specifications for each partition are given in the `guide_tree_file`. Specifications for each subsequence are separated by ';'. Refer to the Examples section (Section 8) for some sample tree files.

Partition options allow the user to create subsequence-specific effects by overriding the global parameters for specific subsequences. Partition options are shown in Table 4.

	Partition-specific Parameters																	
	Root Sequence			Name				Substitution			Indel			Guide Tree				
Partition 1	[	Root Sequence 1	]	“	Label 1	”	#	Substitution 1	#	{	Indel 1	}	(	Tree 1	);			
⋮	[	⋮	]	“	⋮	”	#	⋮	#	{	⋮	}	(	⋮	);			
Partition n	[	Root Sequence n	]	“	Label n	”	#	Substitution n	#	{	Indel n	}	(	Tree n	);			

Figure 2: *A template of the guide tree file for a sequence with* n *partitions/subsequences. Parameters in small-caps are required, others are optional. Parameter values are listed in Table 4.*

Table 4: *Partition options.*

Options	Command Styles	EV[a]	Description
rootseq	[:<root_sequence_file>] or [:<root_sequence_file>, #] or [:<mult_align_file> $(1, 2, 3)$] or [:<mult_align_file> $(1, 2, 3)$, #] or [length] or [length, #]		Six formats are allowed: The first two formats specify the root sequence file, in which a single root sequence along with the quaternary invariable array and any motifs/lineages are specified. The third and fourth format specify that the root sequence will be created from a file containing an input the multiple alignment file, with the quaternary invariable array and motifs/lineages specified. Sub-options ‘1’, ‘2’, and ‘3’ specify the range of the multiple alignment to use, the number of sequences to select from the alignment, and the method of creating the consensus root sequence, respectively (for further details, see Section 6). The last two formats take as input a numeric value (not preceded by ‘:’), and generate a random sequence of the given length to use as the root sequence. When using the options with ‘#’, relative rates will be assigned to each partition. Using the ‘#’ option should be done only when the trees for each partition are the same.
label	“Partition Label”		This option, when present, will label the boundaries of the subsequence in the multiple alignment in the `*.verb` file with the name given.

Continued on next page...

Table 4: (continued)

Options	Command Styles	EV[a]	Description
subseq	#i<% invariable>, b<branch scale>, f<AA frequency file>, m<substitution matrix>, r#	θ^i θ^f θ^m θ^r [b]	Option 'b' modifies and options 'i','f','m','r' replaces the global options given at the command line, respectively. Option 'b' will modify the default branch length scale by the value given. Option 'r' is a flag that specifies that no site-specific rates are used for this partition (i.e., codon, gamma, and discrete gamma rates are reset to uniform rates). This is helpful for specifying introns in coding region sequences, as shown in 8.3. Multiple options should be separated by commas.
indel	Format: {$\#_1$, $\#_2$, <*file*$_1$>/<*file*$_2$>} $\#_1$: Max indel size $\#_2$: Indel probability distribution If $\#_2 = 0$: Use Chang & Benner If $\#_2 > 0$: P(ins)=P(del)=$\#_2$ If $\#_2/\#_3$: P(ins)=$\#_2$, P(del)=$\#_3$ <*file*$_1$>/<*file*$_2$>	 λ^m λ^p λ^l	These options specify the different indel models and parameters. Only the first two parameters are required. The last parameter, <file$_1$>/<file$_2$ >, can be used to specify the two file names, which provide the user defined insertion length distribution (file$_1$) and deletion length distribution (file$_2$). If no distribution file is provided, the distributions given by Chang and Benner (2004) will be used. If only one distribution file is given, it will be used for both insertions and deletions.
Guide tree	Modified Newick Format		The rooted evolutionary tree for the partition. Trees must have the same number of taxa, with each taxon named the same in all trees for all partitions. The branching pattern as well as branch lengths, however, may vary. Branch lengths are assumed to be the expected number of substitutions per site. **Taxon names cannot begin with a number**. Subtrees can be named inside of the tree, as shown in Figure 3.

[a] EV refers to the environment variable of which the global option is categorized. If left unspecified, the option does not refer to any of the environment variables. Environment variables are explained in Section 4.

[b] This option is used to specify an intron from an exon. This assumes that the global option '`-c`' (i.e., '`--codon_rates`') is set. This flag sets the codon rates back to uniform rates.

5.2.1 Specifying Subtrees in Newick Format

To specify subtrees for lineages and motifs, the Newick Tree Format must be clade-labelled, as shown in Figure 3. For multi-partition simulation runs, note that internal node clade labels can be named arbitrarily, thus for a particular subtree, each partition can have a unique lineage-specific parameterization (e.g., Clade1_1 for partition 1, Clade1_2, for partition 2, ...). However, since taxon labels *must* be consistant between partitions, unique subtree parameters cannot be made (e.g., specifying `Taxon1` is ambiguous if more than one partition is used). To specify partition-specific taxon subsequence parameters, refer to Figure 3 for the method of specifying subtrees for taxon subtrees.

```
(
 (
  (Taxon1:0.1,Taxon2:0.1
  )Clade2:0.1,Taxon3:0.1
 )Clade1:0.1,
 (
  (Taxon4:0.1,Taxon5:0.1
  )Clade4:0.1,Taxon6:0.1
 )Clade3:0.1
);
```

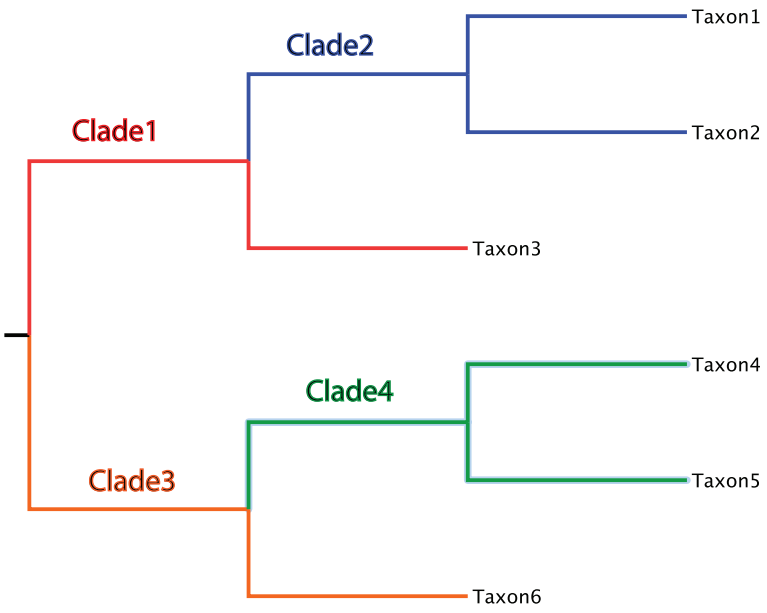

Figure 3: *The subtree-labelled Newick format tree (left) and the resulting tree, which is colored to show the branches affected by the subtree label.*

Node Type	Specification		
Internal Node	<clade_list>		: options;
Taxon	<Taxon_name>		: options;
Taxon	<Taxon_name>	(partition_list)	: options;

Figure 4: *The format of subtree specification in iSGv2.0.3 .*

5.3 Lineage and Motif Options

Lineages and motifs are input to iSGv2.0.3 through the use of the global option "`-k`" or "`--lineage`" accompanied by the name of the lineage/motif file. This file consists of two parts:

1. LINEAGES, which change the conditions under which a subtree evolves, and
2. MOTIFS, which impose site-specific functional constraints on specific sequence positions for a specific subtree.

The basic outline of this file is:

```
LINEAGES =
{
    subtree_list: "lineage_name"#subseq#{indel};
    .
    .
    .
}

MOTIFS =
{
    subtree:
        MARKER=<marker>;
        NAME=<motif_description>;
        PATTERN=<modified_PROSITE_motif>;
    .
    .
    .
}
```

Lineages and motifs both work on subtrees (internal nodes named in input guide tree). To make a lineage or motif effective from the root of the guide tree, name `subtree_list` in LINEAGES or `subtree` in MOTIFS as `subtree:` to `root:`, respectively.

5.3.1 LINEAGES

Lineages allow for a specific subtree to evolve under different conditions as specified for partitions in the tree file. Subsequence options (##) are the only set of options affected by these restrictions, shown in Table 5. Labelling ("") and indel ({}) options are no different than in partitions. *Note: In order to remove indels from a lineage, include the option {0,0} in the lineage.* iSGv2.0.3 prohibits unrealistic changes, *e.g.* when changing site rates (gamma rates (γ), discrete gamma rates, codon rates, and uniform rates), iSGv2.0.3 only allows a subset of ancestor→descendant category changes: $\gamma \rightarrow$\{γ, uniform\}, discrete $\gamma \rightarrow$\{discrete γ, uniform\}, codon→\{codon, uniform\}, and uniform→\{γ, discrete γ, uniform\}.

Table 5: *Novel functions and restrictions imposed for lineage-specific subsequence options (##).*

Option:	Restrictions	Function	EV[a]
Site rates:			
a	Partition must be either gamma, discrete gamma, or uniform rates.	Change α parameter for gamma site rates.	θ^r
c	Partition must be codon rates.	Change codon position frequencies.	θ^r
g	Partition must be discrete gamma or uniform rates.	Change the number of discrete categories.	θ^r
Other:			
b	Cannot change branch scale.	None.	
f	None.	Change character frequencies.	θ^f
i	Only affects random input sequence (i.e., [100])	Change percentage invariable.	θ^i
m	Must be a matrix used for specifiec character set (nucleotides vs. amino acid), cannot take value GENERAL or GTR.	Change substitution matrix.	θ^m
r	None.	Removes *all* constraints (e.g., invariable sites, motifs, site rates) from a lineage, effectively making all changes in the lineage occur using uniform rates.	θ^r, θ^i

[a] EV refers to the environment variable of which the global option is categorized. If left unspecified, the option does not refer to any of the environment variables. Environment variables are explained in Section 4.

5.3.2 MOTIFS

Two steps are required to specify a motif: (1) Marking the root sequence in the *rootseq* file (as in Table 4, formats 1–4), and (2) Specifying constraints in the lineage file.

Motif types: iSGv2.0.3 allows two formats for motifs: PROSITE-like motif format and sequence template motif format. PROSITE-like motif format specifies any length of motif that follows the PROSITE regular expression pattern. Sequence template motif format is required to cover the entire sequence, and are used to place minimum and maximum length parameters on subsequences.

Specifying sites on root sequence: Marking sites is done in the root sequence input file, below the input root sequence or input multiple alignment. To begin marking the sequence, place a '*' corresponding to the first position in the root sequence. This *must* be present for motifs to be correctly parsed. All other positions in the input root sequence must have a corresponding '*' or motif label (anything but '*'). Motif specification characters also must be contiguous, i.e., after specifying the positions that are in a motif, the rest of the motif specification line must be '*'s. As a regular expression, the PROSITE-like motif specification is: $*^+x^+**$, where x is the character for the motif specification. For template-like motifs, the regular expression is $*[0-9]^{n}$[2], where n is the length of the root sequence input. There is no limit to the number of motifs that can be specified on a root sequence. Figure 5 shows some errors that can be made in specifying motifs.

Specifying motif constraints: After the motif-participating sites are labelled in the root sequence, specify the conditions that apply to each site in the lineage file. The fields for specifying motifs are shown in

[2]For multiple sequence alignment input, the regular expression is $*[[0-9],.]^n$, see Section 6.2 for details.

Motif specification	Error	Fix
`aaaaa***************`	Motif specification must begin with '*'	
`*aaaaaa******aaaaa**`	Motif specification must be contiguous	split motif into 2 motifs, specifying each area, or make entire motif contiguous.
`*******aaaaaabbbbb**`	Can only use one motif marker per motif specification	split motif into 2 motifs, specifying each area, or make entire motif contiguous.

Figure 5: *Erroneous motif specifications for a root sequence of length 20, with the assumption that the motif is labelled by 'a'.*

In root sequence specification file

```
50
00000000000000000000000000000000000000000000000000
ASPISTIQAATVPDSS--EVAGKWYIVALASNTSFLREKGKMKMVMARIL
**************bbbbbbbbbbbbbb**********************
```

In motif specification file

<subtree$_i$ >:

```
    MARKER=b;
    NAME=PS00213: Lipocalin signature, ALL;
    PATTERN=[DENG]-{A}-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-{A}-[LIVMTA];
```

Figure 6: *An example of the specifying the lipocalin motif in the single root sequence input. The pattern in the motif specification file corresponds to the marked sites in the last line of the root sequence file. The sites corresponding to* `x(0,2)` *in this instance is specified as 0 sites, with the '-' placeholder being used to fill the maximum possible number of sites (2) that could occupy the "variable-length site".*

Table 6.

Note that when marking motifs, the marked positions must be the same size as the PATTERN specified. In the case of variable positions, iSGv2.0.3 requires that the maximum size of the motif (i.e., n in $\mathtt{x}(m,n)$) are represented in the root sequence. To make this possible, you may need to place '-' in the root sequence or root MSA, as shown in Figure 6. If a character is placed in the variable motif instead of '-', iSGv2.0.3 will assume that this position is part of the variable motif in the root sequence.

6 Root Sequence Input

The root sequence is specified by the square brackets (`[ ]`) in the guide tree file. There are three parameters, if used, for root sequence input:

1. Specifying the length of root sequence. indel-Seq-Gen will randomly generate a root sequence of the specified length. For example `[40]` calls for a root sequence of length 40 characterss.
2. Root sequence input.
3. Multiple alignment input.

Table 6: *List of the options for specifying motifs. See Figure 6 for an example of a motif specification.*

Field	Description	Motif type
MARKER	A single character, value can be anything except for the character '*' (see Section 6 for specifying sites for motif). This is the ID of the motif as specified in the root sequence input.	Prosite-like, template (0-9 only).
NAME	Not required. This is the description of the motif.	Prosite-like, template
PATTERN	Regular expression patterns, as below.	
`[list]`	A list of acceptable amino acids or nucleotides for a site, site can take value of `list`.	Prosite-like
{`list`}	A list of unacceptable amino acids or nucleotides for a site, site can take any value *not* in `list`.	Prosite-like
`char`	Invariable site, takes value of `char`.	Prosite-like
x	Site can take any amino acid or nucleotide value.	Prosite-like, template
x(n)	Next n sites can take any amino acid or nucleotide value.	Prosite-like, template
x(min, max)	Variable length motif sequence. Cannot start a motif with a variable length motif position. This specifies that length of the motif positions can vary between m to n sites (variance caused by indels). Variable sites cannot overlap between different motifs.	Prosite-like, template

Each of these options is explained in the following subsection. Make sure that the input file is in the same directory when using a root sequence input file (options 2 and 3).

6.1 Root Sequence Input

This option specifies that the user has a root sequence in a file for a partition.

In the tree file, the root sequence input is specified as:

```
[:<rootseq_file>]
```

The format of the file `<rootseq_file>` is:

```
<length of sequence>
<invariable array>
<sequence>
<motif_spec>
```

Figure 7 gives an example of a root sequence for conserving a Thioredoxin-fold protein sequence motif.

```
20
00003223100000000000
LARDCVLCSTWVTIALACLK
```

```
20
00000000100000000000
LARDCVLCSTWVTIALACLK
****aaaa************
```

In motif specification file

```
root:
    MARKER=a;
    NAME=Trx-like;
    PATTERN=C-[GATPLVE]-[PHYWSTA]-C;
```

Figure 7: *Thioredoxin-fold (Trx-fold) proteins have a characteristic "CXXC" motif that is conserved for all proteins in the family. (Left): This root sequence input requires the CXXC motif to remain constant for all sequences created through the use of the invariable array (described later), listed above the root sequence. The serine in position 9 will also be held invariable, though insertions are allowed to occur between itself and the previous cysteine. Finally, the length of the root sequence is given by the first line. Note that this sequence is not truly a Trx-fold sequence, but an example to show the usage of the invariable array. (Right): Motif specification of the Thioredoxin active site. Motif specifications preserve both the length dependence and character subsets. Note also that the invariable array is no longer used to preserve the motif; it is recommended that when using motif specification, all non-motif positions in invariable array should be set to 0.*

6.2 Multiple Alignment Input

This option specifies that the user has a multiple alignment of their sequence set, and wants a root sequence created from the multiple alignment. In the guide tree file, the root sequence input is specified as:

```
[:<ma_file>(1,2,3)]
```

Where options 1, 2, and 3 stand for:

1. The range of the multiple alignment to use, where the format is *beginning*`:`*end*. An input of `21:67` specifies the range from the 21st to the 67th spot (inclusive) of the input multiple alignment. Default for this option is the entire range of the multiple alignment.

2. The number of sequences to choose from the multiple alignment. indel-Seq-Gen randomly with replacement selects the specified number of sequences from the multiple alignment. Default for this option is to use all sequences.

3. Method for collapsing the multiple alignment into a root sequence, either random `'r'` or consensus `'c'`. Consensus is a majority-rule method, using a coin flip to break ties. Random uses a weighted coin toss based on the character composition at the site to choose the representative character, except for invariable positions, which will be chosen by consensus. For an example of the weighted coin toss, look at the first column in Figure 8 in which all sequences emit an amino acid, column 6. In this column, there are 2 T's, 1 V, and 1 C. A weighted coin toss on this column will be a T 50% of the time, a V 25% of the time, and a C 25% of the time. The default for this option is consensus.

When specifying the multiple alignment in the tree file, a blank field specifies that the default entry for the field is desired. For example, `[:input_MA(,,r)]` indicates that the entire range and all sequences from `input_MA` will be used, but that the character that will represent the column will be chosen by a weighted coin toss based on the characters that appear in that column. Note that the size of the root sequence can fluctuate between simulation runs. For example, using the option `[:input_MA(,1,)]` will randomly choose a single sequence to create root sequence, thus the sequence that indel-Seq-Gen chooses will be the length of the root sequence for that subsequence.

Format:

```
<invariable array>
<sequence 1>
<sequence 2>
 .      .
 .      .
 .      .
<sequence n>
<motifs>
```

Figure 8 is an example of an input multiple alignment of the SET-C region of the SET-domain family. The '-' is the only character that will be accepted for the gap character. Irregular characters ("YRUN" in nucleotides, "BZJX" in amino acids) cause each character for which they stand for to be counted as the 1 over the number of characters they represent (e.g., 'B' increments both 'N' and 'D' by 0.5). All '.'s in the template *must* also be '.'s in all motifs coinciding with the same site. The reverse, however, is fine, as shown in this example, in the position just before the GxxL motif.

6.2.1 Multiple alignment root sequence input for template motifs

Note that the multiple alignment root sequence input will have variable length, depending on the number of columns in the multiple alignment that are inferred to be gap columns ($> 50\%$ '-' in a column). When specifying template motifs, all sites, except the first, must belong to a position in the template. For this reason, when specifying a template motif, iSGv2.0.3 also allows the character '.' on a position in the motif specification. The '.' character specifies that this site should be excluded from the template sites. For example, in Figure 8, 6 positions of the multiple sequence alignment are excluded from the template specification.

6.3 Invariable Array

The invariable array in indel-Seq-Gen is a quaternary array that specifies how a region is allowed to evolve, and is specified in the root sequence input. Table 7 shows the effect of each numerical entry in the invariable array, and Figure 9 shows an example of how the algorithm finds possible positions for indels based on the invariable array.

Input MA, using invariable array

```
       00000000000000000000000000000000100322300010000
Taxon1 -----TKEDF-----TSDQTNPIGQDSATRLILKGEELTCNYKLFD
Taxon2 RNSCNVVPKIVQVNGDFRIRFTALRD-----IK-GEELFFNYGENF
Taxon3 -----TGVSNNQFGG---YDFVALGD-----IE-GEELTWBYETTE
Taxon4 -----CELVQLTEFS---LGVVAICN-----IE-GEELSFDYAWEA
```

Input MA, using motif specification

```
       00000000000000000000000000000000000000000010000
Taxon1 -----TKEDF-----TSDQTNPIGQDSATRLILKGEELTCNYKLFD
Taxon2 RNSCNVVPKIVQVNGDFRIRFTALRD-----IK-GEELFFNYGENF
Taxon3 -----TGVSNNQFGG---YDFVALGD-----IE-GEELTWBYETTE
Taxon4 -----CELVQLTEFS---LGVVAICN-----IE-GEELSFDYAWEA
       ******************************bb.bbbb********
       *....00000000000000000000..00000000000000000
```

In motif specification

```
<subtree>:
    MARKER=b;
    PATTERN=I-x-G-x(2)-L;
<subtree' >:
    MARKER=0;
    NAME=Multiple alignment template example;
    PATTERN=x(10,22)-x(2,6)-x(5)-x(3,8);
```

Figure 8: **input_MA***: (Top) An example of a 4 sequence multiple alignment from the SET-domain family of sequences using the invariable array. The GXXL motif is conserved in both the invariable array, along with the tyrosine and isoleucine. (middle, bottom) The motif specification input and root sequence file. The motif will be conserved in the same manner as the invariable array on the top. However, specifying this subsequence as a motif allows the user to conserve the motif on a subtree, rather than over the entire guide tree as is done with the invariable array.*

Table 7: *Representation of positions in the invariable array, and effects on the sequence. Invariable sites (1 and 3) block both deletion and substitution of the corresponding position in the sequence array. No-indel sites (2 and 3) block deletion of the sequence array, but also block insertion events from occurring between consecutive no-indel positions (between 2-2, 2-3, 3-2, and 3-3) in the invariable array.*

#	Effects
0	None
1	Invariable
2	No-indel
3	Invariable + No-indel

```
Accepting Positions:

  Invariable Array:       0 0 1 2 3 0
  Insertion (any size):  1 1 1 1 0 1 1

  Invariable Array:       0 0 1 2 3 0
  Deletion (size 1):      1 1 0 0 0 1
  Deletion (size 2):      1 0 0 0 0 0
  Deletion (size 3):      0 0 0 0 0 0
```

Figure 9: *The invariable array and accepting positions for insertion and deletion events. For insertions, the accepting positions (denoted by '1' and '0' above) are located in between consecutive positions in the invariable array, while deletion accepting positions correspond exactly those in the invariable array. In the accepting positions, the site with '1' is allowed to have indels. In the rare case that an accepting position cannot be found (as in the size 3 deletion example above), indel-Seq-Gen will output an error, but continue the simulation run.*

7 Other Input Files

7.1 Character Frequencies `#f`*freq_file*`#`

Protein and nucleotide subsequences often evolve under different functional constraints, causing them to display different character frequencies. For this, a file containing 20 amino acid frequencies (order: ARND-CQEGHILKMFPSTWYV) or 4 nucleotide frequencies (order: ACGT) separated by commas can be entered for each subsequence with the option: `#f<freq_file>#`. Values from the input file are read and normalized to create a distribution, where the number of values is specified by the maximum indel size of the subsequence. For an example, see the file `aaf.freq` included with all iSG `.tar.gz` archive downloads.

7.2 Indel Probabilities {`#`$_1$`,#`$_2$`,`*indel_file(s)*}

Insertion and deletion frequencies can be provided for each subsequence. The format of the file is shown in Figure 10. This example is for the indel probabilities of sizes 1–10 of the Zipfian distribution described in Chang and Benner [2]. indel-Seq-Gen will read in the number of values corresponding to the `max_indel_size` ($\#_1$) specified for the subsequence (for Figure 10, the maximum indel size of a subsequence using this file can be up to size 10), and then normalize the values to create a distribution. This means that the frequencies can be given in absolute numbers or in any scale (as shown in Figure 10). If the maximum indel size is greater than the number of indel positions in the length distribution file, indel-Seq-Gen will output a message that it is unable to read the input distribution file.

```
2628,743.8,355.5,210.5,140.2,100.6,76,59.6,48.1,39.7
```

Figure 10: *An example length distribution input file. This has the frequencies of indels with lengths from 1 to 10 amino acids, taken from the first 10 values of the Zipfian distribution. The number of values in this file should not be smaller than the given max indel size.*

```
[999]((Taxon1:0.3,Taxon2:0.14):0.5,(Taxon3:0.34, Taxon4:0.5):0.12);
```

Figure 11: *A simple example of a nucleotide simulation partition file. This simulates a 999 nucleotide sequence with no indels, along a guide tree of 4 taxa.*

mid_nuc_coding.tree
```
[999]{9,0.1/0.3,codonLD}((Taxon1:0.3,Taxon2:0.14):0.5,(Taxon3:0.34, Taxon4:0.5):0.12);
```

codonLD
```
0,0,3222,0,0,233, 0,0, 0.23
```

mid_nuc_noncoding.tree
```
[999]{9,0.1/0.3,idLD}((Taxon1:0.3,Taxon2:0.14):0.5,(Taxon3:0.34, Taxon4:0.5):0.12);
```

idLD
```
2628,743.8,355.5,210.5,140.2,100.6,76,59.6,48.1,39.7
```

Figure 12: *The specification files for a simulation run with indels, and the associated files, codonLD and idLD.*

8 Examples

8.1 Basic coding and non-coding

Despite the complexity that exists in iSGv2, the main goal of the simulation method is to be an all-purpose sequence simulator. Therefore, included in the distribution is a file called "`simple_nuc.tree`", shown in Figure 11.

Given this example, assume the usage of the Hasegawa, Kishino, and Yano [3] substitution matrix. The following two commands will create non-coding and coding simulation runs, respectively:

```
./indel-seq-gen --matrix HKY --outfile DNA_out < simple_nuc.tree
./indel-seq-gen --matrix HKY --codon_rates 0.2,0.05,0.75 --outfile DNA_out < simple_nuc.tree
```

This will also create five outfiles: DNA_out.ma, DNA_out.seq, DNA_out.root, DNA_out.trace, and DNA_out.verb. In the first examples, the sites are mutated uniformly as is non-coding DNA. In the second example, the third coding mutates 75% of the time, while the first and second codn mutate 20% and 5% of the time, respectively, mimicking codon position rates.

8.2 Including indels

A little more complex example including indels is shown in Figure 13. Notice that we now need to two trees to represent coding or non-coding sequences, since indel sizes that are not a multiple of three would cause a nonsense mutation.

The commands:

```
./indel-seq-gen -m HKY -e mid_noncoding --num_runs 5 < mid_nuc_noncoding.tree
./indel-seq-gen -m HKY -e mid_noncoding -n 5 -c 2,1,8 --invar 0.02 < mid_nuc_coding.tree
```

exon_intron.tree

```
[:exon(,,)]{9,0.031/0.01,codonLD}((Taxon1:0.3,Taxon2:0.14):0.5,(Taxon3:0.34, Taxon4:0.5):0.12);
[:intron]#r,b1.2#{9,0.1,idLD}((Taxon1:0.3,Taxon2:0.14):0.5,(Taxon3:0.34, Taxon4:0.5):0.12);
```

exon

```
          000000000000000000000000000000000000000000
Taxon1    TTACTTT---TTCCTAACCGG---C---CCGAGCT---AATG
Taxon2    TTAGTTTTTAATCCCAACCTG---C---CCCATCT---GATA
Taxon3    CGAAATC------TTACCGGAGGTAATACGCGGCGC---ATC
Taxon4    CTGCATT------CAACCTGT---T---AGCATCGC---ATC
```

intron

```
30
330000000000001000000000000033
GUTTCAGGTAAAATGCANNGACTTAGRYAG
```

Figure 13: *The specification files for a simulation run of one exon partition and one intron partition with indels, and the associated root sequence files, exon and intron.*

In this example, the previously introduced options `--matrix`, `--outfile`, and `--codon_rates` have been replaced by their short options. The `--num_runs (-n)` option has also been included, so that five runs will be simulated. The multiple alignments of this example is very gappy, since in the indel options, the maximum indel size is 9 nucleotides, and an insertion occurs once for every 10 substitutions, a deletion once every 3.3 substitutions. Finally, the option `--invar` sets 2 percent of the sites to be invariable. Codon rate input also does not need to add up to 1, as iSGv2.0.3 will normalize the values.

8.3 Coding and non-coding together

iSGv2.0.3 has the unique ability to simulate exons and introns in a single run. This adds yet more complexity, as you can see by the necessary files, as listed in Figure 12. Files that remain the same as previous examples are excluded from this figure.

The command:

```
./indel-seq-gen -m HKY -c 2,1,8 -e exon_intron --step_type trs < exon_intron.tree
```

There are many interesting things to note:

1. Codon rates are input at the command line. The subsequence option #r# in the intron sequence resets the rates to uniform rates in the intron. The additional subsequence options #b1.2# multiplies all branch lengths in the tree by a factor of 1.2.
2. The file `intron` conserves the nuclear-like spliceosomal sites (GU at beginning, G in the middle, and AG at the end).
3. The option `--step_type` sets the simulation run to be in time relative steps. The output file `exon_intron.trace` will contain where each indel event occurred and the relative time of the event, as explained in Section 5.1.1.

8.4 Lineage Introduction

Lineages are one of the unique abilities of iSGv2, paricularly the ability to cause a lineage to evolve under no specifications, which we call "pseudogene" evolution (the `p` option in subsequence options `##`). To add

exon_intron_lineage.tree

```
[:exon(,,)]{9,0.031/0.01,codonLD}
(
 (Taxon1:0.3,Taxon2:0.14
 )Clade1_1:0.5,
 (Taxon3:0.34, Taxon4:0.5
 ):0.12
);
[:intron]#r,b0.5#{9,0.1,idLD}
(
 (Taxon1:0.3,Taxon2:0.14
 )Clade1_2:0.5,
 (Taxon3:0.34, Taxon4:0.5
 ):0.12
);
```

exon_intron_lineage.spec

```
LINEAGES =
{
    Clade1_1:#fpse.freq,p#{5,0.08,idLD};
    Clade1_2:#p#{8,0.08,idLD};
    Taxon3(1):#a 0.7#{0,0};
}
```

pse.freq

```
10,23,15,6
```

Figure 14: *The specification of subtrees in exon_intron_lineage.tree and the corresponding specifications in exon_intron_lineage.spec. Subtrees on internal nodes can be uniquely named, thus the convention Clade_<partition_number>. Taxa cannot be uniquely specified between partitions, since they are required to be the same for all partitions. Thus, the taxon lineage (Taxon3) contains the list of partitions it is active in in parentheses. In this case, the lineage specifications for Taxon3 are active only for the first partition.*

lineages to the previous example, we need to specify clades in the partition file, create the lineage file and specify this file on the command line. Figure 14 shows the file changes, and Figure 15 shows the output of the command:

```
./indel-seq-gen -m HKY --lineage exon_intron_lineage.spec --alpha 1.3 < exon_intron_lineage.tree
```

No output files are given, thus the output is printed to the screen. Note that the file `outfile.verb` will still be created.

Some finer points:

1. The global parameter `--alpha` specifies that, unless otherwise specified in clades, all branches in all trees will evolve with site rates, where the gamma parameter $\alpha = 1.3$. However, Clade1_1 and Clade1_2 both become pseudogene lineages, which releases site rates from them. Taxon3 in the first tree also does not evolve with this gamma, since the lineage file specifies it to evolve with $\alpha = 0.7$ (`#a 0.7#`).

2. Taxon3, partition 1, evolves without indels ({0,0}).

```
>Dataset_0__partition_1
[0,D,,1100,4,16:17:18:19]
[1,I,,1100,1,23]
[2,D,,1100,4,30:31:32:33]
[3,D,,1000,1,15]
[4,I,,1000,2,21:22]
[5,I,,0011,3,12:13:14]
>Dataset_0__partition_2
[6,D,,1000,2,59:60]
[7,I,,1000,4,39:40:41:42]
[8,D,,1000,6,61:62:63:64:65:66]
[9,D,,0001,1,62]
CTACATTATCCTAACCGGCCCCAGCGCAATCGTTTCAGGTAAAATGCACAGACTTAGACAG
 4 61
Taxon1    CTACAGCAGCCACCGACCACCCGCCGCCGTTGTCAGTTGGTCTGCAGCCAG
Taxon2    CTACAGCAGCCTAGGCAAGAACCCGTGTCAGTTGAACTGCAAAGACTCAGCCAG
Taxon3    ATACTGTCTCCCAATACCCAGCGCTATAGCAAATGTATAAGGTAAAATGGACAGACTTTGACAG
Taxon4    TTATCTTAGGACAGTACCCAGACCCCGGGCAATCGTGTGAACTAGAACGCAGAGCATAGCCAG
 4 61
Taxon1    CTACAGCAGCC--------ACCGACCACC----CGCCGCCGTTGTCAGTTGGTCTGCA--
Taxon2    CTACAGCAGCC---T----A--GGCAAGA----ACCCG----TGTCAGTTGAACTGCAAA
Taxon3    ATACTGTCTCCCAATACCCA---GCGCTATAGCAAATG----TATAAGGTAAAATGGACA
Taxon4    TTATCTTAGGACAGTACCCA---GACCCCGGGCAATCG----TGTGAACTAGAACGCAGA

Taxon1    ------GCCAG
Taxon2    GACTCAGCCAG
Taxon3    GACTTTGACAG
Taxon4    G-CATAGCCAG
```

Figure 15: *The output for a lineage simulation, consisting of event tracing, root sequence, sequences, and multiple alignment. Taxon1 and Taxon2 are "pseudogenes".*

8.5 Prosite-like motif

To create Prosite-like motifs, a root sequence input is necessary. In this example, the lipocalin root sequence and motif specification in Figure 6 is used, along with `lipocalin.tree`, shown in Figure 8.5.

```
indel-seq-gen -m JTT -k lipocalin.spec -w a < lipocalin.tree
```

lipocalin.tree

```
[:lipocalin]
{5,0.1,idLD}
(((Taxon1:0.4,Taxon2:0.14):0.5,Taxon3:0.34):0.1, Taxon4:0.5);
```

Figure 16 shows the output of this command.

```
>Dataset_0__partition_1
[0,D,,1100,1,10]
[1,D,,1100,2,9:11]
[2,I,,1000,5,44:45:46:47:48]
[3,D,,1000,5,50:51:52:53:54]
[4,D,,1000,2,8:12]
[5,D,,1000,2,35:36]
[6,D,,1000,1,42]
[7,D,,1000,2,39:40]
[8,I,,0100,2,18:19]
[9,D,,0001,1,42]
[10,D,,0001,1,52]
[11,D,,0001,2,55:56]
[12,I,,0001,1,4]
ASPISTIQAATVPDSSEVAGKWYIVALASNTSFLREKGKMKMVMARIL
 7 48
5         ASPISTIQAATVPDSSEVAGKWYIVALASNTSFLREKGKMKMVMARIL
6         DSPIDTIWAAVVPDSSDIAGRWYLMALVSDTSFLREKAKLKMVVAGVL
7         DRPVDMSVVGDSSAVDGTWLFMQYVTEVSFVRQRLKFKMLVSTKL
Taxon1    EGPLDTAVEEEQISGNWLGMRSTYHVVSLFNQGILQEV
Taxon2    ERSLEMTVVGDSSGDAIDGVWLFLQYINEVGFLRQRLKFAALGTIKL
Taxon3    DGPIDTIQTNIVLDSSDIAGRWYVMDLIGDAMFRRTMKGLKNVLSGPL
Taxon4    PFPNLATKLIGVQPDQDEIIGQWYELSHHSKSAIFGDTSMKLVTK

 7 48
5         ASP-ISTIQAATVPDSS--EVAGKWYIVALASNTSFLREKGKM-----KMVMARIL
6         DSP-IDTIWAAVVPDSS--DIAGRWYLMALVSDTSFLREKAKL-----KMVVAGVL
7         DRP-VDMS---VVGDSS--AVDGTWLFMQYVTEVSFVRQRLKF-----KMLVSTKL
Taxon1    EGP-LDT-----AVEEE--QISGNWLGMRSTYHV--VS--L-FNQGILQ-----EV
Taxon2    ERS-LEMT---VVGDSSGDAIDGVWLFLQYINEVGFLRQRLKF-----AALGTIKL
Taxon3    DGP-IDTIQTNIVLDSS--DIAGRWYVMDLIGDAMFRRTMKGL-----KNVLSGPL
Taxon4    PFPNLATKLIGVQPDQD--EIIGQWYELSHHSKSAIFGDTS-M-----KLV-TK--
```

Figure 16: *The output for a lipocalin simulation, consisting of event tracing, root sequence, sequences, and multiple alignment.*

8.6 Template motif

To set up a run with a template, we will using the true multiple alignment sequences output in Figure 16 as the input root sequence. The resultant file and template specifications are given in Figure 8.6, and we use `lipocalin_ma.tree`, a tree similar in Figure 8.5, where `[:lipocalin]` is replaced by `[:lipocalin_ma(,,)]`.

`lipocalin_ma`

```
          0000000000000000000000000000000000000000000000000000000000
Taxon1    EGP-LDT-----AVEEE--QISGNWLGMRSTYHV--VS--L-FNQGILQ-----EV
Taxon2    ERS-LEMT---VVGDSSGDAIDGVWLFLQYINEVGFLRQRLKF-----AALGTIKL
Taxon3    DGP-IDTIQTNIVLDSS--DIAGRWYVMDLIGDAMFRRTMKGL-----KNVLSGPL
Taxon4    PFPNLATKLIGVQPDQD--EIIGQWYELSHHSKSAIFGDTS-M-----KLV-TK--
PS00213   **************bbbbbbbbbbbbbb**************************
lipo_ma   *77777777777777777777777777777777777777777.....77777777
```

`lipocalin_ma.spec`

```
MOTIFS =
{
        root:
                MARKER=b;
                NAME=PS00213: Lipocalin signature;
              PATTERN=[DENG]-{A}-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-{A}-[LIVMTA];
        root:
                MARKER=7;
                NAME=lipo_ma: Lipocalin partial template;
                PATTERN=x(5,20)-x(10,30);
}
```

Note that in this example, there is a very tight restriction on the number of insertions, while deletions are more acceptable. Figure 17 shows these restrictions based on the input root sequence. Note that the sites marked by '.' are removed consideration when building the root sequence.

```
template:  ---- x(5,20), 17 --- -------- x(10,30), 30 --------
positions: 12345678911111111112 12345678911111111122222222223
                    01234567890          01234567890123456789O
              *             **
Taxon1     EGP-LDT-----AVEEE--QI SGNWLGMRSTYHV--VS--L-FQ-----EV
Taxon2     ERS-LEMT---VVGDSSGDAI DGVWLFLQYINEVGFLRQRLKFAALGTIKL
Taxon3     DGP-IDTIQTNIVLDSS--DI AGRWYVMDLIGDAMFRRTMKGLKNVLSGPL
Taxon4     PFPNLATKLIGVQPDQD--EI IGQWYELSHHSKSAIFGDTS-MKLV-TK--
lipo_ma    *7777777777777777777 77777777777777777777777777777
```

Figure 17: *The input multiple alignment* `lipocalin_ma` *from Figure 8.6 as iSGv2.0.3 reads it in. Above the alignment is the placement of the template sites as iSGv2.0.3 places them;* `template:` *Gives the template spec (e.g.,* `x(5,20)`*) and the current number of sites residing in the template as iSGv2.0.3 reads the root sequence,* `positions:` *the position in the template. Numbers with a '*' below them are positions that are not included in the root sequence built. A space has been placed between templates to make it easier to read.*

To run this example, simply type:

```
./indel-seq-gen -m JTT -k lipocalin_ma.spec < lipocalin_ma.tree
```

Note that this example incorporates both a motif and template, and that they do not need to coincide. This is often the case for real protein evolution, such as the lipocalins in this example, where the lipocalin signature begins in a coil region and ends in a beta strand. A good use of the template for this type of protein sequences is to make each $x(min, max)$ match with the coil regions and beta strands, as is done in the published work [5].

References

[1] R.A. Cartwright. DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*, 21:iii31–iii38, 2005.

[2] Mike S. S. Chang and S.A. Benner. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *Journal of Molecular Biology*, 341:617–631, 2004.

[3] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:672–677, 1985.

[4] A. Rambaut and N.C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13:235–238, 1997.

[5] C.L. Strope, K. Abel, S.D. Scott, and E.N. Moriyama. Biological sequence simulation for complex evolutionary hypotheses with indel-Seq-Gen version 2. *Molecular Biology and Evolution*, 26:2581–2593, 2009.

[6] X. Xia and Z. Xie. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. *Molecular Biology and Evolution*, 19:58–67, 2002.