SPIMAP: Species Informed Maximum A Posteriori Gene Tree Reconstruction

Documentation for the SPIMAP software package

April 8, 2011

Author: Matthew D. Rasmussen (rasmus@mit.edu)
Software website: http://compbio.mit.edu/spimap

citation:
Matthew D. Rasmussen and Manolis Kellis. *A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction*. Molecular Biology and Evolution. 2010. doi: 10.1093/molbev/msq189

Contents

```
>KLLA0C08239g
ATGAGTCTCAAACGTGTAGTTGTCACTGGTCTTGGGGCCTACACGCCCCTTGGTTCTACAGTTTCAAAGTCTTGGGCAGG
TTTGCTT---GCTGCTAAGCAATCACTAATACCCTTAGATGCTTTCTACAACAGAGAA---GACTTTGCAAAAGTGAAAA
AGTTGGTCCCACTAGATACAGCAGTGAGTAGGTTACAT------------------------------------------
>ADL072C
---ATGCATCCCCGAGTGGTCGTGACCGGCATTGGGTGCTATACTCCTCTGGGGCCGTCGCTAGCCCAGTCTTGGAAGGA
GCTGTTG---CGAGGGACGAGCGGCCTTGTCAGGCTGCAAGATCTGGCAGAGTACGAGGGCGATTACAAACCACTGTCGA
GGCTTATATCCGGTGATCTTCGAGTCGGGAAAGTTGGATTTGAG------------------------------------
>kwal_5828
---ATGACTTCCAGAGTCGTTGTTACTGGGCTTGGTGCTATCACTCCACTTGGGAGGACTGTTTCCGAGTCATGGAGAGC
TTTATTG---GCAGGCAAGTCCGGAATTCGTCCCATTCGCGATCTTCCC------------AATGCTAAAAGCTACGAAG
GACACTGTCCTGCATCTGTTGCCGTTGCAGACATTCCTGATTTC------------GATCCA------------------
```

Figure 1: **Example *.align file.** Three gene DNA sequences are given, each with 240 sites.

1 File formats

1.1 Sequence alignment format (*.align)

SPIMAP uses the FASTA file format (`http://en.wikipedia.org/wiki/FASTA_format`) for sequences alignments. The file extension is not important and many different extensions are in common use (*.fa, *.mfa, *.fasta, *.align).

Each line starting with ">" is used to indicate a gene name (Figure 1). The entire line after the ">" is used as the gene name. The aligned sequence of is then given on the following lines. It may be wrapped to any number of columns (no not wrapped at all). Gaps in the alignment are represented with the "-" character.

At this time (version 1.1), SPIMAP can only use DNA sequences. The sequence can be in both upper case and lower case (SPIMAP ignores case) and degeneracy codes can be used ("NnRrYyWwSsKkMmBbDdHhVv"), however at this time SPIMAP treats all degeneracies as completely missing data ("N").

1.2 Species tree file format (*.stree)

Species trees should be specified using the Newick file format. See `http://en.wikipedia.org/wiki/Newick_format` for details. Beyond the newick format, SPIMAP has only a few additional requirements. First, the species names given in the species tree should match those given in the SMAP file (Section 1.3). Second, the branch lengths of the species tree should be expressed in units of time (Figure 2). Any unit of time can be used (e.g. millions of years, generations, relative units, etc). The only requirement is that the duplication and loss rates are also expressed in compatible units. Therefore, if branch lengths are in *millions of years*, the duplication rate (specified by `spimap`'s "-D" option) should be in units of duplications/gene/*million years*.

```
(((((((scer:7.061760,
       spar:7.061760
      )n7:4.999680,
      smik:12.061440
     )n6:5.970600,
     sbay:18.032040
    )n5:52.682400,
    cgla:70.714260
   )n4:7.220700,
   scas:77.934960
  )n3:23.181480,
  (
   (
    agos:78.553260,
    klac:78.553260
   )n9:10.434960,
   kwal:88.988220
  )n8:12.128400
 )n2:78.883560,
 (
  (
   (
    calb:41.275620,
    ctro:41.275980
   )n12:29.632860,
   (
    cpar:52.323120,
    lelo:52.323120
   )n13:18.585720
  )n11:31.149540,
  (
   (
    cgui:75.615840,
    dhan:75.615840
   )n15:14.006880,
   clus:89.622720
  )n14:12.435660
 )n10:77.941620
)n1;
```

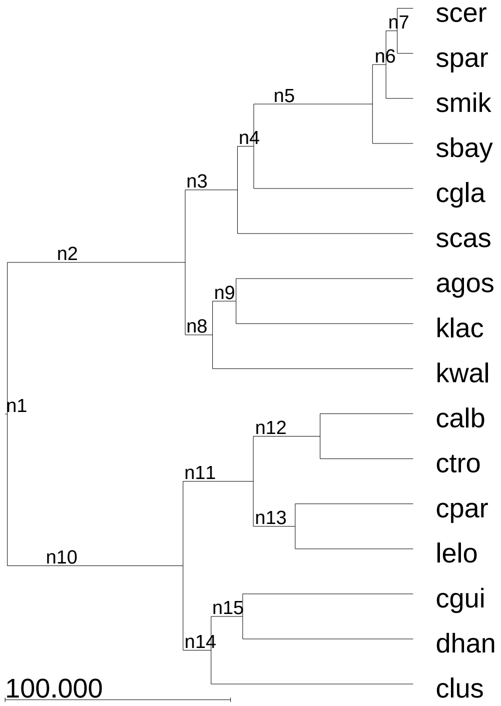

Figure 2: **Example *.stree file and corresponding tree.** This file (left) specifies the species tree (right) using the newick file format. Branch lengths should be expressed in units of time (e.g. millions of years). Ancestral nodes can also optionally be named (the names “n1”, “n2”, etc are used in this example).

Naming ancestral nodes. SPIMAP also supports naming ancestral nodes in the species tree using the newick format. For example, the parental node of `human` and `chimp` can be named `primate` using the following syntax:

```
((human:5,chimp:5)primate:70,mouse:75)mammal;
```

If ancestral nodes are named, they will be used in the output of the reconciliation mapping (Section 1.4).

1.3 Gene to species name mapping file format (*.smap)

SPIMAP uses a special file format (`*.smap`) to specify which genes belong to which species. Each line contains two tab-delimited fields:

1. pattern matching a gene ID
2. species ID

Only 3 types of gene ID patterns are supported. The pattern can either be an exact matching string, a prefix (denoted "`text*`"), or a suffix (denoted "`*text`"). The "*" is the only special wildcard character.

```
A*        agos
orf19*    calb
CDUG_*    cdub
CAGL*     cgla
IPF_*     cgla
CGUG_*    cgui
sbay_*    sbay
scas_*    scas
smik_*    smik
spar_*    spar
SCP*      spom
YALI*     ylip
Y*        scer
Q*        scer
```

Figure 3: **Example *.smap file.** This file specifies how to map gene names to their corresponding species. The first column indicates a gene name pattern (in this case a prefix) and the second column specifies a species name. Note: this example only gives a partial list of the species in Figure 2.

The species ID should be the same as the ones used in the species tree files. All patterns and IDs are case-sensitive.

When mapping a gene name to a species name all exact matches are processed first. If no exact match is found, the patterns are then processed in the same order as they appear in the file until a match is found. For example in the SMAP file given in Figure 4, the gene ID "`YALI123`" should match the species "`ylip`", instead of "`yeast`", because the pattern "`YALI*`" occurs before "`Y*`".

1.4 Reconciliation file format (*.recon)

When SPIMAP's "-r" option is used, the reconciliation found for the gene tree and species is saved to a file "`OUTPUT_PREFIX.recon`". The reconciliation file format is tab-delimited, where each line has three fields:

1. gene node ID.
2. species node ID.
3. event (one of the following: "gene", "spec", "dup")

Each line specifies the mapping of one node in the gene tree (field 1) to one node or branch in the species tree (field 2). Branches are indicate using the node ID directly below it (i.e. the younger of the two incident nodes). The lines can be given in any order.

If the gene node is a leaf, it will map to a leaf in the species tree and the event field will contain the event "gene". All internal nodes of the gene tree are marked either as speciations (event "spec") or duplications (event "dup"). Specaition nodes map directly to the indicated species node, and duplication nodes map to the indication species branch. The time of the duplication along the species branch is not indicated in this file format nor is it inferred by SPIMAP.

```
KLLA0C08239g   klac   gene
ADL072C        agos   gene
kwal_5828      kwal   gene
CAGL0J02970g   cgla   gene
scas_g715.48   scas   gene
smik_6662      smik   gene
sbay_7039      sbay   gene
smik_6659      smik   gene
sbay_7037      sbay   gene
YER061C        scer   gene
spar_6281      spar   gene
n10            n5     spec
n9             n7     spec
n8             n6     spec
n7             n5     spec
n6             n5     dup
n5             n3     spec
n4             n3     spec
n3             n9     spec
n2             n8     spec
n1             n2     spec
```

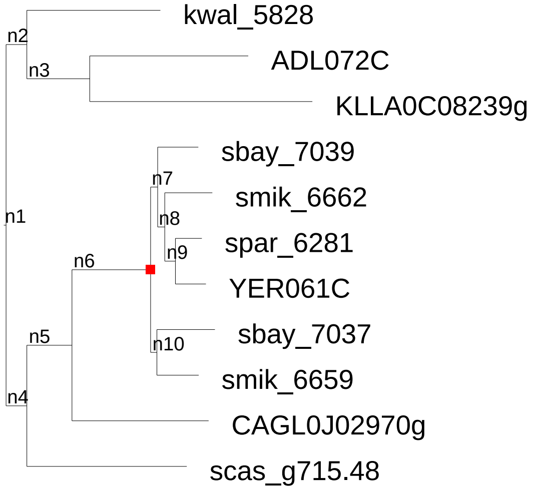

Figure 4: **Example *.recon file.** The reconciliation file format (left) specifies how all the nodes in a gene tree (right) map to the nodes and branches in the species tree (see Figure 2). Notice that gene node “n6” (red dot) represents a duplication event along species branch “n5” (shown Figure 2). The gene tree and species tree have their own name space (“n5” in the gene tree is not the same as “n5” in the species tree).

If gene IDs are not given to the ancestral nodes of a gene tree or species tree, SPIMAP will by default name them with “nXXX” where XXX is the preorder traversal of the internal nodes.

1.5 SPIMAP model parameters file format (*.params)

SPIMAP has several parameters for its substitution rates model. These parameters are learned by the `spimap-train-rates` program, which saves the parameters in a custom `*.params` file format. The `spimap` program reads these parameters using the “-p” option. Most uses of SPIMAP do not require understanding the contents of a `*.params` file.

The `*.params` file format is tab-delimited and each line is processed one at a time.

If the first field of a line is the word “baserate”, then the remaining two fields are interpreted as floating point values α_G and β_G, which are the two parameters, shape and scale, of the inverse-gamma distributed gene-specific rate.

If the first field of the line does not match “baserate”, then the first field indicates a species tree branch and the remaining two fields are interpreted as floating point values α_i and β_i, which are the two parameters, shape and scale, of the gamma distributed species-specific rate. Each branch is indicated by its more recent

```
baserate  6.98457288742   5.98457288742
1         3.28887700831   394.209221588
2         4.64684152603   551.109741211
3         1.13027572632   164.191940308
4         0.610769152641  75.0393371582
5         7.14405012131   927.631103516
6         2.96983885765   238.195861816
7         5.63683271408   632.264831543
8         0.974860072136  94.9837493896
9         0.856632292271  78.6899032593
10        4.64683914185   544.528686523
11        1.92581880093   271.891052246
12        3.84569692612   624.703308105
13        3.14617466927   335.446655273
14        0.699178874493  84.1814575195
15        0.746283352375  137.345901489
scer      8.42576217651   763.305847168
ctro      6.70220327377   999.845153809
scas      9.14448356628   1253.45031738
agos      8.84074497223   801.648925781
sbay      6.95680332184   1048.7590332
kwal      14.3321857452   1962.9083252
dhan      15.7483224869   2699.00878906
smik      10.2562847137   1143.78076172
cgla      9.81903266907   1015.43951416
spar      5.80616807938   799.18963623
calb      8.38038921356   1233.68322754
lelo      9.40990924835   973.772583008
cpar      9.43262672424   1184.28100586
klac      6.6709280014    767.418823242
clus      8.37989234924   881.762878418
cgui      11.9692239761   1187.47314453
```

Figure 5: **Example *.params file.** The `*.params` file contains the parameters for SPIMAP's substitution rate model.

node. Ancestral nodes are indicated by an integer, where are assigned in pretraversal order.