DeCoSTAR: Ancestral genome or gene organizations using reconciled phylogenies - Supplementary Materials

February 20, 2017

Scaffolding mode

DeCoSTAR can be used to infer some extant adjacencies, typically to account for badly assembled genomes. In this case, the cost to put a new adjacency between two leaves (of the same species) is defined by the formulas:

$c_1(a,b) = -T * log(p)$
$c_0(a,b) = -T * log(1-p)$

Where: $p = F_{adj} * BP$
$T = \frac{Break}{SPI * log(\frac{1-BP}{BP})}$)

And: $BP = \frac{\#ctg - \#chr}{2 * \#ctg * (\#ctg - 1)}$

$\#ctg$: number of contigs

$\#chr$: expected number of chromosomes

F_{adj} is defined as follow. We make the assumption that a genome's organisation is linear: a gene can have at most two neighbors.

If the extremity of the adjacency a (respectively b) already has two neighbors (ie. is in the middle of a contig), then this adjacency is not possible: $F_{adj} = 0$ (leading to $c_1(a,b) = \infty$)

If both extremities of the adjacency already have one neighbor, $F_{adj} = 1$.

If the extremity of the adjacency a (respectively b) already has one neighbor and the extremity b (resp. a) has no neighbors, then $F_{adj} = 2$, to account for the fact that a could be either one of the two neighbors of b.

If both extremities of the adjacency have no neighbors, $F_{adj} = 4$, to account for the different senses in which the two genes a and b could be linked together.

In the special case where DeCoSTAR is used with oriented adjacencies, a and b aren't genes, but extremities of genes (ie. a gene start or stop). We consider that gene extremity is always linked to the other extremity of the same gene. As a consequence, F_{adj} is always computed considering that a and b have one more neighbor than before.

SPI, or Scaffolding Propagation Index, is a parameter that accounts for the ditribution of poorly assembled genomes along the species tree. More precisely, it is the size of the clade c where

DeCoSTAR can still infer new adjacencies even though the adjacency has no extant homologues in c (ie. the only extant adjacency that support the new one is on the outgroup of this clade).

Data-sets

18 Anopheles genomes

The species tree is taken from (Fontaine *et al.*, 2015) , and constructed from a concatenate of genes on the X chromosome. Gene families have been obtained from (Neafsey *et al.*, 2015). There are 17 780 gene families in the database, and we discarded gene families with only 1 gene, families containing a gene included in another gene, and a few families which did not pass an alignment quality filter (no resulting site after GBlock). This resulted in 14 940 families. Gene trees were first inferred by RaxML with Muscle alignments as input, then corrected with ProfileNJ, keeping only 100% bootstrap support branches. Adjacencies were then computed as the set of consecutive genes on the same scaffold. DeCoSTAR was used with scaffolding mode and Boltzmann sampling with a temperature of 0.05.

References

Anselmetti, Y., *et al.* (2015). Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics*, **16**(Suppl 10), S11.

Bérard, S., *et al.* (2012). Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics (Oxford, England)*, **28**(18), i382–i388.

Chauve, C., Ponty, Y., and Zanetti, J. P. P. (2015). Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. *Lecture Notes in Computer Science (Advances in Bioinformatics and Computational Biology)*, **8826 LNBI**, 49–56.

Fontaine, M., *et al.* (2015). introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**.

Neafsey, D. E., *et al.* (2015). Highly evolvable malaria vectors: The genomes of 16 anopheles mosquitoes. *Science*, **347**(6217), 1258522.

Patterson, M., Szöllősi, G., Daubin, V., and Tannier, E. (2013). Lateral gene transfer, rearrangement, reconciliation. *BMC Bioinformatics*, **14**(Suppl 15), S4.

Szöllosi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, **62**(3), 386–397.

DeCoSTAR's Documentation

DeCoSTAR: Detection of Co-evolution version 1, 24/11/2016

Given a species tree S, a set of gene family (unrooted) tree distributions G, a set of extant adjacencies A and a set of costs for adjacencies events (adjacency gain and adjacency breakage), DeCoSTAR can compute:

- reconciled gene trees R from the gene trees in G, such that these forms a most parsimonious reconciliation between S and G according to the TERA algorithm described in [1].

- an history of the given adjacencies in A along the reconciled gene trees R such that this history minimizes an adjacency gains and breakages cost(with respect to their relative costs); according to the models described in [2][3].

The history of adjacencies comes in the form of one or several adjacency trees (phylogenetic trees in which nodes represents adjacencies between the nodes of gene trees). We refer to such an history as an adjacency forest. Note that DeCoSTAR can, but usually does not output adjacency forest but instead provide lists of inferred adjacencies in ancestral species.

Rather than an adjacency forest minimizing the number of adjacency gains and breakages, DeCoSTAR can sample adjacency forests in such a manner that adjacency forests with a lower adjacency cost have a higher chance to be sampled. This is done according to the algorithm described in [4], extended to include transfers.

It is also possible to use DeCoSTAR to infer adjacencies in extant species as described in [5] by using the scaffolding mode.

There are three required parameters. The first is the species tree file (`species.file`), which contains a tree in newick format. The second is a file containing the names (one per line) of files containing gene tree distributions with one distribution per gene family (`gene.distribution.file`). Even if each gene tree distribution only contains one tree (for instance if you used a maximum likelihood approach to get that tree), DeCoSTAR ask for one file per gene family. The third is a file containing the extant adjacencies (`adjacencies.file`). There must be one adjacency per line. An adjacency correspond to a couple of leaves names separated by a space.

Description of the software parameters

The parameters that are boolean (ie. whose default is true or false) should be specified using 0 or 1 (for, respectively, false and true). Parameters must be given using the format `<name>=<value>` (note the absence of `'-'` before the option name as well as the absence of space after the `'='`. For instance, a typical command line might look like:

```
DeCoSTAR species.file=my.sp.tree.txt gene.distribution.file=my/distribs.txt adjacencies.file=adjs.t
```

For a non dated (but with transfer) analysis where the cost of a single adjacency gain is 3 (the other events will have their default costs).

Input parameters:

- `species.file` required. Species tree file (newick). NB: default behavior wants it to be ultrametric if transfers are used (use `dated.species.tree=0` to circumvent)
- `parameter.file` no default value. A file with input parameters (one per line)
- `gene.distribution.file` required. Gene distribution files file (one file name per line)
- `adjacencies.file` required. Adjacencies file (one adjacency per line; leaf names separated by a space)
- `dated.species.tree` default : true. The species tree is ultrametric and dates will be used to subdivide the trees in time slices.

- `char.sep` default : `'_'`. Character separating gene names in gene tree files. One character only. Caveat : I you decide to use `char.sep='|'`, be aware that the character `'|'` is also used as a separator in the newick output of reconciled trees (if you use `write.newick=1`).
- `ale` default : false. Gene tree distribution are ALE files
- `already.reconciled` default : false. Gene tree distribution are reconciled gene trees in recPhyloXML format. Will skip the reconciliation phase
- `rooted` default : false. Specify that the root of the given gene trees must be kept. This option turns off amalgamation when switched on.

Reconciliation parameters:

In its default form, DeCoSTAR performs the reconciliation of gene families using the TERA algorithm [1]. As such, it include some of the algorithm's options.

The options that concern reconciliation are:

- `with.transfer` default: true. Allows transfers in the reconciliation and adjacency histories reconstruction.
- `dupli.cost` default: 2. cost of a single gene duplication
- `HGT.cost` default: 3. cost of a single Horizontal Gene Transfer
- `loss.cost` default: 1. cost of a single gene loss
- `try.all.amalgamation` default: true. try all possible amalgamation when reconciling gene trees. Otherwise only the best tree (ie. The most frequent) of the distribution is used.
- `Topology.weight` default: 1. In the case of amalgamation, this is the weight associated to the topology part of the score guiding the reconciliation

Not all options are included. In the case were some specific, non-included, options are required, it is recommended to perform the reconciliation independently using ecceTERA (or any other reconciliation software) and then to directly give the reconciliation to DeCoSTAR (using `already.reconciled=1`).

Basic adjacency history parameters:

- `Again.cost` default: 2. Cost of a single adjacency gain
- `Abreak.cost` default: 1. Cost of a single adjacency breakage

Adjacency history sampling parameters:

- `use.boltzmann` default: false. Use Boltzmann sampling for the adjacencies history computation.
- `boltzmann.temperature` default: 0.1. Temperature to use in the Boltzmann sampling (if used)
- `nb.sample` default: 1. Number of samples to get from the adjacency matrix. NB: it can be used together with `use.boltzmann` or not.

Adjacency history assembly related parameters (scaffolding mode [5]):

- `scaffolding.mode` default : false. Use scaffolding algorithm to improve extant genomes scaffolding/assembly.
- `chromosome.file` no default value. A file containing the number of chromosome in each each species (one species per line, each line comprised name of the species followed by the number of chromosome, separated by a tabulation)
- `adjacency.score.log.base` default :10000. Used in the case where the adjacency file also contains a score between 0 and 1. Base of the logarithm applied to this score.
- `scaffold.includes.scored.adjs` default : false. Used in the case where the adjacency file also contains a score between 0 and 1 AND `scaffolding.mode` is true. If true, include the adjacencies with a score ¡ 1 in the computation of the number of contigs.

Advanced adjacency history parameters:

- `C1.Advantage` default: 0.5. Between 0 and 1. Probability to choose C1 (presence of adjacency) over C0 (absence of adjacency) in case of a score tie at the root of an equivalence class
- `all.pair.equivalence.class` default: false. Compute adjacency histories for all pair of gene families (even if they share no adjacencies).
- `bounded.TS` default: false. Use bounded time slices in adjacency history computations (only if the species tree is dated)
- `always.Again` default: true. Always put an adjacency Gain at the top of an equivalence class tree
- `absence.penalty` default: -1. If set to -1 (the default), nothing changes. Otherwise, specify the cost of having an adjacency at a pair of leaves which are not part of the list of adjacencies given at initialization
- `substract.reco.to.adj` default: false. If set to 1, the weighted cost of a reconciliation event will be used to favor co-event in the adjacency matrix computation. Unavailable for Boltzmann computation.
- `Reconciliation.weight` default: 1. Weight of the reconciliation events when `substract.reco.to.adj=1`
- `Adjacency.weight` default: 1. Weight of the adjacency events when `substract.reco.to.adj=1`

Output parameters:

- `verbose` default: 1. Show progress and timing.

  0: nothing is reported short of error.
  1: basic report (default).
  2: various information about reconciliation, adjacency matrix and backtracking
  3: maximal amount of information

- `write.newick` default: false. Use newick format rather than phyloXML-like format for ouputs.
- `hide.losses.newick` default: false. If true, losses and the branches leading to them will be removed from newick formatted output.
- `write.adjacencies` default: true. Write the adjacencies inferred in ancestral species in a file.
- `write.genes` default: false. Write the genes inferred in ancestral and extant species.
- `output.dir` default: none. Directory to print files in.
- `output.prefix` default: none. A prefix to prepend to all output files.
- `write.adjacency.trees` default: false. Write the inferred adjacency trees.

Input formats:

The default parameters assume an ultrametric dated (i.e., with branch lengths), binary species tree. An undated species tree is input using `dated.species.tree=0`.

The gene trees are expected to be unrooted and in a newick format (unless the rooted option is used). Leaves names should be composed of the name the the species in which the leaf is and the gene name, linked by a separator (by default, this separator is `'_'` and can be changed using `char.sep`). Rather than gene trees, the user may supply ale files instead (specified with `ale=1`), which are files that sum up a gene distribution in the form of conditional clades probabilities. Such files can be obtained with the ale [6] software. If the gene family ave already been reconciled (for instance if you don't want to use the same method as DeCoSTAR), they can be provided instead of the gene tree distributions (with `already.reconciled=1`) . Reconciled trees should be provided using the recPhyloXML format (see http://phylariane.univ-lyon1.fr/recxml/ for a description of the format).

The adjacencies given in the file (`adjacencies.file`) may present two additional fields describing the orientation of the genes forming the adjacencies. These orientations are specified using the `'+'` and `'-'` character (respectively for a sense and anti-sense gene). This will cause DeCoSTAR to treat the extremities of a gene as two different entities when it comes to adjacencies (but not reconciliation : the two extremities of a same gene have the same history).

Additionally, the adjacencies given in the file (adjacencies.file) may have a third field that should be a number between 0 and 1. This number will be used as a score denoting the confidence that the adjacency really exists (1 meaning that the adjacency is certain; 0 that the adjacency is not possible) that DeCoSTAR algorithm will take into account, allowing the possibility to create adjacency histories without this adjacency. This is an advanced functionality, and it is linked to the option (`adjacency.score.log.base`) which determine the base of the logarithm that is used to go from this 0-to-1 score to a parsimony cost.

If both orientation and score are specified for an adjacency, they should come in that order: orientation, then score; such that a valid line could look like:

```
g1 g2 + + 0.9
```

Rather than being all provided in the command lines, arguments can be given in a file specified with the parameter.file argument. In that file, parameters can be given using the format

<name>=<value>. Any parameters given on the command line will take precedence and the duplicated parameter will be ignored.

Output formats:

If the output.dir option has been used, all file will be written in the specified directory. Otherwise they are written in the current directory.

If write.adjacencies is set to true (it is by default), DeCoSTAR will output the a file containing the adjacencies inferred at ancestral speciation nodes and leaves such that each line represent an adjacency. The fields of these lines are separated by spaces and correspond to, in order:

- the species the adjacency is in
- the gene forming the first extremity if the adjacency
- the gene forming the second extremity if the adjacency
- the orientation of the first extremity if the adjacency (as described in the input format section)
- the orientation of the second extremity if the adjacency (as described in the input format section)
- the eventual score given to that adjacency at input (NB: ancestral adjacency have an input score of 0)
- the frequency of observation of the adjacency (ie. how many time the adjacency was observed across all sample divided by the number of samples)

If write.genes is set to true (it is set to false by default), DeCoSTAR will output a file describing all extant and ancestral genes. Each line correspond to one gene and begins with the code of the species the gene is in followed by the gene name followed by the list of the gene's extant descendants (all separated by spaces).

DeCoSTAR will also output the species tree and will create a reconciliations.suffix file. If write.adjacencies.trees is set to true (it is set to false by default), then an adjacencyTrees.suffix file will also be written. Here suffix is either 'xml' or 'newick' depending on the chosen output format. These files contain respectively the reconciled tree of each gene family and the adjacency trees computed from the extant adjacencies.

By default, all trees are written in XML format. The species tree follows a classical phyloXML format. The reconciled gene tree are in the recPhyloXML format (see http://phylariane.univ-lyon1.fr/recxml/ for a detailed description of the format).

Adjacency trees follow a format close to the recPhyloXML one, adapted to include adjacency related events. As each node represent an adjacency, its given name is actually the name (or the id, if they have no name) of the two genes it links, separated by '-'. Each clade has a <eventsAdj> tag that contains an ordered list of event the adjacency has undergone. Each event has a coevent property. If the coevent property is set at \1", then this indicate that the event spanned both end of the adjacency at the same time. If it is set at \0", then it means that the event only spanned one end of the adjacency. There is an additional event when compared to recPhyloXML: the adjBreak event which marks an adjacency breakage. Furthermore, the different adjacency equivalence class families (ie. a group of adjacencies linking gene from the same couple of gene families) are grouped

together under the `<EquivalenceClassFamily>` tag which specifies which gene families are linked. If several samples were done, then an additional `<sample>` tag is present.

If the option `write.newick` has been activated, reconciled gene trees and adjacency trees will be written in a newick format where reconciliation or adjacency information will be written in place of the bootstrap.

This information consists in the name of the gene (or its id in the case of an internal node), or the name of the two genes it links for adjacency trees, followed by the event associated with the node, the species it is in and the time slice it occurs at (if applicable). These fours fields are separated by a `'|'` character. For reconciliation trees, events may be:

- `Extant` : for leaves
- `Spe` : speciation
- `Loss` : gene Loss
- `Dup` : gene duplication
- `SpeOut` : speciation to an extinct/unsampled lineage (otherwise called SpeciationOut)
- `Reception` : transfer reception
- `Null` : no event (to account for time slices)
- `BifOut` : bifurcation in an extinct/unsampled lineage (otherwise called BifurcationOut)

For adjacency trees, the following events are added:

- any reconciliation event might have the prefix `\co-"` marking the fact that both extremities of the adjacency underwent the same event at the same time (for instance, a `co-duplication` means that the two adjacent genes were duplicated together)
- `Breakage` : adjacency breakage

The trees linking different gene families are separated by an information line beginning by `'>'` and specifying which families are linked and, if necessary, the sample the trees belong to.

NB: in this model of reconciliation, lateral gene transfer are modeled as a process where a gene first undergoes a speciation to an extinct or unsampled lineage of the species tree (otherwise called `SpeciationOut`) where it evolves for a certain time before being transferred from this unsampled lineage to a sampled lineage (ie. a branch) of the species tree (transfer reception). See [1] or [5] for a more detailed view of this process when applied to reconciliation inference.

NB2: adjacency trees do not explicitly contain any adjacency gains because any adjacency gains actually gives rise to a new adjacency tree. In other words, there is an implicit adjacency gain at the root of every adjacency tree.

REFERENCES

1. Celine Scornavacca, Edwin Jacox, and Gergely Szöllősi. Joint Amalgamation of Most Parsimonious Reconciled Gene Trees. Bioinformatics (2014): btu728.

2. Sèverine Bérard, Coralie Gallien, Bastien Boussau, Gergely J. Szöllősi, Vincent Daubin and Eric Tannier. Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics (Oxford, England) Vol. 28, No. 18 (2012) p. i382-i388

3. Murray Patterson, Gergely Szöllősi, Vincent Daubin, Eric Tannier. Lateral gene transfer, rearrangement, reconciliation. BMC Bioinformatics Vol. 14, Suppl. 15 (2013) S4

4. Cedric Chauve, Yann Ponty and João Paulo Pereira Zanetti. Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. BMC Bioinformatics Vol. 16 Suppl. 19 (2015) S6

5. Yoann Anselmetti, Vincent Berry, Cédric Chauve, Annie Chateau, Eric Tannier and Sèverine Bérard. Ancestral gene synteny reconstruction improves extant species scaffolding. BMC genomics Vol. 16, Suppl. 10 (2015) S11

6. Gergely Szöllősi, Wojciech Rosikiewicz, Bastien Boussau, Eric Tannier and Vincent Daubin. Efficient Exploration of the Space of Reconciled Gene Trees. Systematic biology Vol. 62, No. 6 (2013) pp. 901-912