

PhySIC_IST: cleaning source trees to infer more informative supertrees.

CELINE SCORNAVACCA,
Vincent Berry,
Vincent Ranwez et Emmanuel J.P. Douzery

LIRMM, UMR CNRS 5506
ISEM, UMR CNRS 5554
University of Montpellier II

June 16, 2008



Reconstruction of phylogenies

- INPUT: (different) source datasets

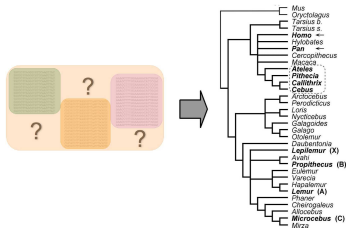
```

AAAGCTTGGAA  AAGCTTGGAA  16,19,20,24,25,26,28,29,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100
AAAGCTTGGAA  CC  CB  BB  BC  BB  CB  BB  BG  DD  BB  18,18,20,18,24
AAAGCTTGGAA  CB  CB  BB  BC  BB  CB  BB  EG  DD  BB  18,18,20,15,24
AAAGCTTGGAA  CB  CB  BB  BC  BB  CB  BB  EI  DD  BC  18,19,20,16,25
AAACCTTGGAA  CB  CB  BB  BB  BB  CB  BB  EG  DD  BB  18,19,20,18,25
AAACCTTGGAA  CC  CB  BB  BB  BB  BB  BK  DD  BB  18,20,21,14,18
AAACCTTGGAA  CC  CB  BB  BB  BB  BB  BK  DG  BB  18,20,20,12,19
AAACCTTGGAA  CC  CB  BB  BB  BB  CB  EK  DD  BB  18,19,20,12,20
AAACCTTGGAA  DB  CE  BB  BB  BB  CB  EG  DD  BB  18,19,20,12,19
AAACCTTGGAA  DB  CE  BB  BB  BB  CB  EG  DD  BF  CACGC  25,23,19
AAACCTTGGAA  CB  CB  BB  BB  BB  CB  EG  DD  BF  CACGC  26,24,20
AAACCTTGGAA  CB  CB  BB  BB  BB  FB  CB  EL  DD  BF  CACGC  27,28,19
20,21,18,22,23,20,19  CB  BB  BB  BB  FB  CB  EL  DD  BF  CACGC  27,28,31
20,19,16,22,20,20,19  CB  BB  BB  BB  FB  CB  EL  DI  BF  ACACGC  19,23,20
20,19,16,22,20,20,19,20  CB  BB  BB  BB  FB  CB  EL  DI  BF  ACACGC  19,23,20
20,19,16,22,20,20,19,20,22  CCGCGGCCCTTAGGGTTTCAACTACACGC  19,23,20
20,19,19,21,20,20,19,20,22  CCGCGGCCCTTAGGGTTTCAACTACACGC  20,23,22
20,19,19,21,20,20,19,20,22  CCGCGGCCCTTAGGGTTTCAACTACACGC  20,22,19
21,19,19,21,20,20,20,19,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  20,22,19
21,19,18,24,20,20,19,19,18  CCGCGGCCCTTAGGGTTTCAACTACACGC  19,22,19
21,19,18,24,20,20,19,19,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  19,19,19
21,20,18,24,20,20,18,20,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  19,18,20
21,20,18,24,20,20,16,22,20  CCGCGGCCCTTAGGGTTTCAACTACACGC  19,19,20
21,20,21,24,20,20,16,22,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  20,19,20
21,20,21,24,20,20,15,22,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  21,20,18,24,20,20,16,22,20,20
21,20,21,24,20,20,16,22,19,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  21,20,21,24,20,20,16,22,19,19
21,20,21,24,20,20,15,22,19,19  CCGCGGCCCTTAGGGTTTCAACTACACGC  21,20,21,24,20,20,15,22,19,19
  
```


Reconstruction of phylogenies for multiple datasets

Two main approaches

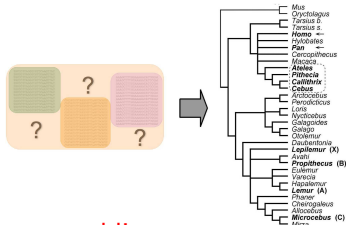
- Supermatrix approach: **assembling datasets**



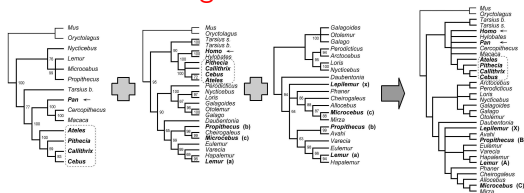
Reconstruction of phylogenies for multiple datasets

Two main approaches

- Supermatrix approach: **assembling datasets**



- Supertree approach: **assembling trees**



Interest of supertrees

Supertrees are useful for:

- Combining heterogeneous data
- Obtaining a phylogeny using several genes:
 - ▶ Avoids having to deal with too much missing data
 - ▶ Evolutionary models adapted for each gene sequence
- Pointing out problematic areas of the phylogeny
 - ▶ agreement and disagreement among input trees.
 - ▶ measuring taxon overlap

Supertree methods

VOTE vs VETO methods

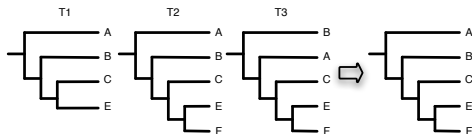
Supertree methods can be classified into two categories, depending on the way they deal with incongruent data:

Supertree methods

VOTE vs VETO methods

Supertree methods can be classified into two categories, depending on the way they deal with incongruent data:

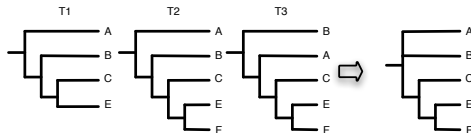
- **Vote** methods resolve conflicts, opting for the resolution that maximizes their optimization criteria.
- **worrying feature**: this approach can lead to propose clades contradicting all source trees.



Supertree methods

VOTE vs VETO methods

- Veto methods do not allow the resulting supertree to contain clades that a source tree would vote against.
 - ▶ pruning some taxa:
OR
 - ▶ proposing multifurcations
- worrying feature: this approach can lead to propose unresolved supertrees.



PhySIC

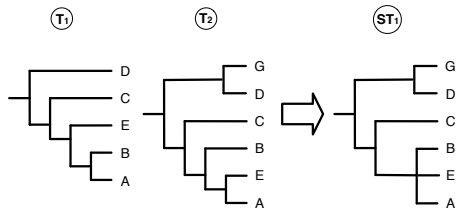
A VETO method with desirable properties

- The resulting supertree does not contain relationships contradicting the source trees (**non-contradiction** property, denoted by **PC**);

PhySIC

A VETO method with desirable properties

- The resulting supertree does not contain relationships contradicting the source trees (**non-contradiction** property, denoted by **PC**);



PhySIC

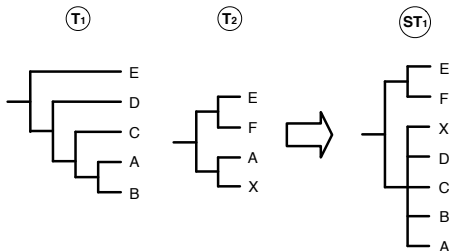
A VETO method with desirable properties

- The resulting supertree only contains relationships that are present in a source tree or collectively induced by several source trees (**induction** property, denoted by **PI**).

PhySIC

A VETO method with desirable properties

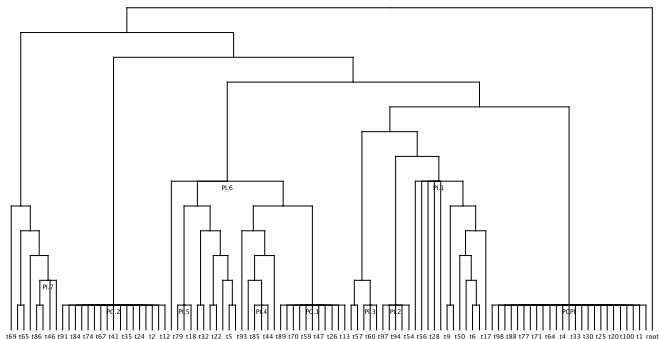
- The resulting supertree only contains relationships that are present in a source tree or collectively induced by several source trees (**induction** property, denoted by **PI**).



PhySIC

A VETO method with UNdesirable proprieties

- BUT**, when \mathcal{T} contains numerous contradictions or small overlap, the supertrees built with *PhySIC* can be highly unresolved.



An improved version of *PhySIC*

- To cope with this, we propose a second method that:

An improved version of *PhySIC*

- To cope with this, we propose a second method that:
 - ▶ maximizes the information contained in the produced supertree;

An improved version of *PhySIC*

- To cope with this, we propose a second method that:
 - ▶ maximizes the information contained in the produced supertree;
 - ▶ returns a supertree T that still respects PC and PI by:

An improved version of *PhySIC*

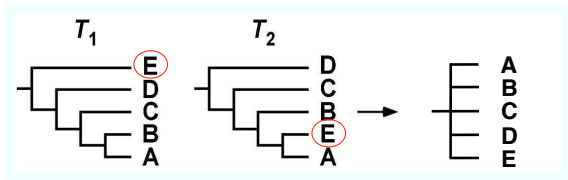
- To cope with this, we propose a second method that:
 - ▶ maximizes the information contained in the produced supertree;
 - ▶ returns a supertree T that still respects PC and PI by:
 - ★ allowing multifurcations;

An improved version of *PhySIC*

- To cope with this, we propose a second method that:
 - ▶ maximizes the information contained in the produced supertree;
 - ▶ returns a supertree T that still respects PC and PI by:
 - ★ allowing multifurcations;
 - AND
 - ★ pruning rogue taxa:

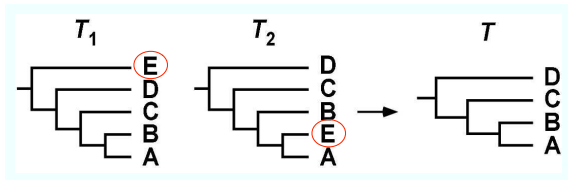
An improved version of *PhySIC*

- To cope with this, we propose a second method that:
 - ▶ maximizes the information contained in the produced supertree;
 - ▶ returns a supertree T that still respects PC and PI by:
 - ★ allowing multifurcations;
 - AND
 - ★ pruning rogue taxa:



An improved version of *PhySIC*

- To cope with this, we propose a second method that:
 - ▶ maximizes the information contained in the produced supertree;
 - ▶ returns a supertree T that still respects PC and PI by:
 - ★ allowing multifurcations;
 - AND
 - ★ pruning rogue taxa:



PhySIC_IST

Outline of *PhySIC_IST*

- *PhySIC_IST* (*PHYlogenetic Signal with Induction and non-Contradiction Inserting a Subset of Taxa*) is an algorithm that operates successive insertions of taxa on a backbone tree.

PhySIC_IST

Outline of *PhySIC_IST*

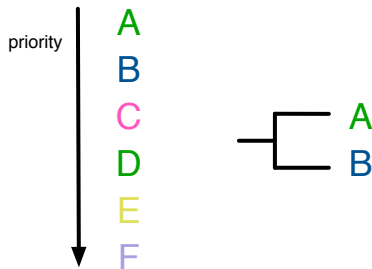
- *PhySIC_IST* (*PHYlogenetic Signal with Induction and non-Contradiction Inserting a Subset of Taxa*) is an algorithm that operates successive insertions of taxa on a backbone tree.

the order of the insertions has to be chosen carefully!

PhySIC_IST

Outline of *PhySIC_IST*

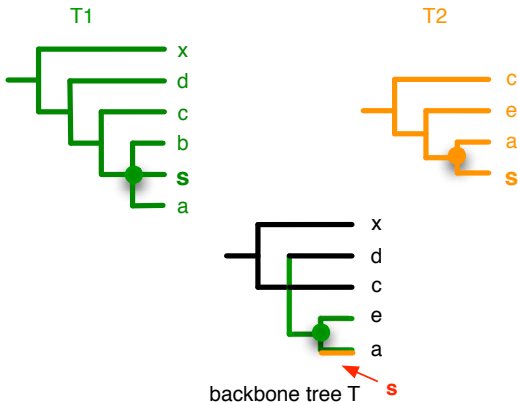
- We order taxa in decreasing priority order
- The first taxa to be inserted are those present in as much source trees as possible and involved in as few contradictions as possible
- We build the backbone tree



PhySIC_IST

Supports

- Within which region of the backbone tree can a taxon s be inserted without contradicting T_1 and T_2 ?



PhySIC_IST

Outline of *PhySIC_IST*

- one best supported position (PI) and all trees agree (PC)

PhySIC_IST

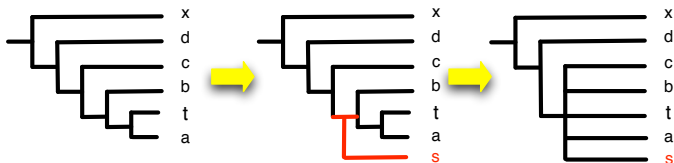
Outline of *PhySIC_IST*

- one best supported position (PI) and all trees agree (PC)
- more than one best supported position and/not all trees agree (PI and PC???)

PhySIC_IST

Outline of PhySIC_IST

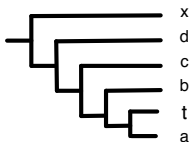
- one best supported position (PI) and all trees agree (PC)
- more than one best supported position and/not all trees agree (PI and PC???)



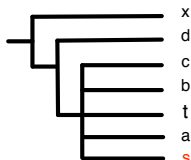
PhySIC_IST

Outline of PhySIC_IST

- one best supported position (PI) and all trees agree (PC)
- more than one best supported position and/not all trees agree (PI and PC???)



VS



PhySIC_IST

CIC criterion

- We need to evaluate the amount of information of a tree.

PhySIC_IST

CIC criterion

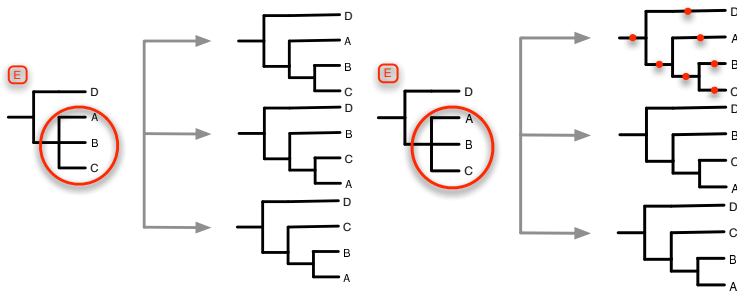
- We need to evaluate the amount of information of a tree.
- We use a variant of the CIC criterion (Thorley, Wilkinson, Charleston 1998) that also takes into account missing taxa and we define it as:

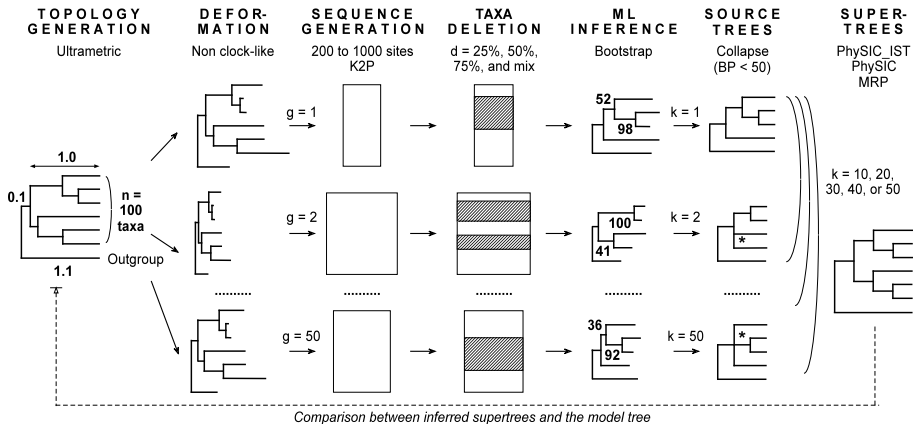
PhySIC_IST

CIC criterion

- We need to evaluate the amount of information of a tree.
- We use a variant of the CIC criterion (Thorley, Wilkinson, Charleston 1998) that also takes into account missing taxa and we define it as:

$$CIC(T, n) = -\lg \frac{\text{number of permitted binary trees with } n \text{ taxa}}{\text{number of possible binary trees with } n \text{ taxa}}$$

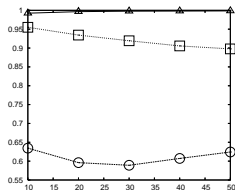




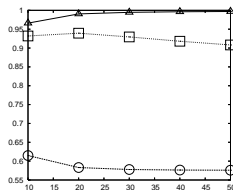
Large-scale simulations

Average *CIC* values

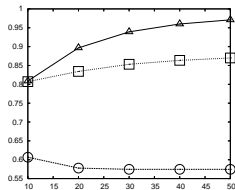
MRP \triangle , *PhySIC* \circ , *PhySIC_IST* \square



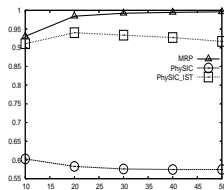
$d = 25\%$



$d = 50\%$



$d = 75\%$

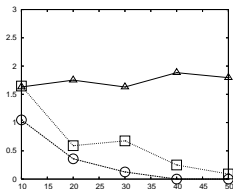


mixed d

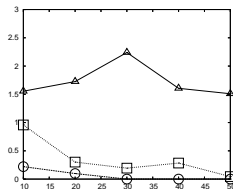
Large-scale simulations

Average percentage of type I error

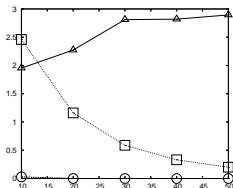
MRP \triangle , *PhySIC* \circ , *PhySIC_IST* \square



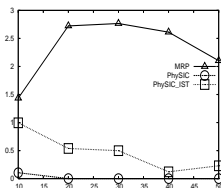
$d = 25\%$



$d = 50\%$



$d = 75\%$



mixed d

The improvement of *PhySIC_IST* on *PhySIC*

The improvement of *PhySIC_IST* on *PhySIC* is a consequence of three fundamental differences between them:

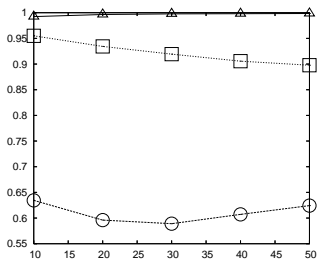
- the new version operates successive insertions of taxa on a backbone and is not based on a revised version of the Build algorithm (unlike *PhySIC*)
- the two methods do not have the same optimization criterion
 - ▶ *PhySIC* \Rightarrow nb of triplets
 - ▶ *PhySIC_IST* \Rightarrow *CIC*
- *PhySIC_IST* can propose non-plenary supertrees

- 1 Introduction
 - Combining data for phylogenetic inferences
 - Vote methods
 - Veto methods
- 2 VETO methods with desirable proprieties
 - Physic
 - *PhySIC_IST*
- 3 STC preprocess

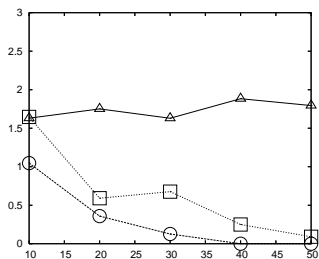
Limits of veto methods

- As the amount of available information continues to increase, the number of conflicts between source trees increases

MRP \triangle , *PhySIC* \circ , *PhySIC-IST* \square

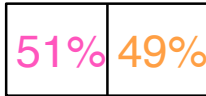


informativeness

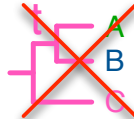
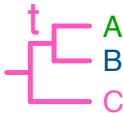


inaccuracy

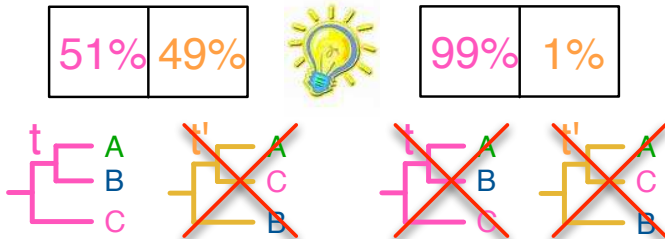
Vote VS veto methods?



Vote VS veto methods?



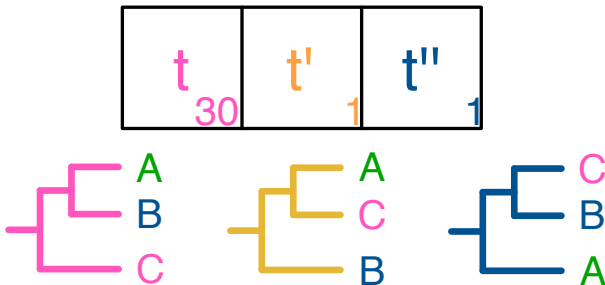
Vote VS veto methods?



- **IDEA:** flexible liberal(voting) preprocessing of the input trees before a veto approach.

Source Tree Correction (STC) preprocess

We want to drop the statistically less supported alternative(s), if any exists.



STC preprocess

- After that, the STC preprocess modifies the source trees (*PhySIC_IST*), forcing them not to contain the dropped resolutions.

STC preprocess

- After that, the STC preprocess modifies the source trees (*PhySIC_IST*), forcing them not to contain the dropped resolutions.
- Each modified tree may contain either new multifurcations, or lack some of its former taxa.

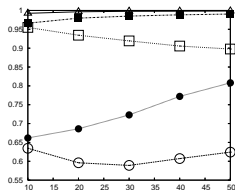
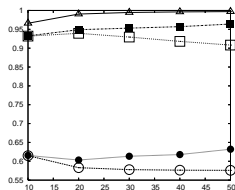
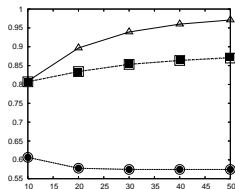
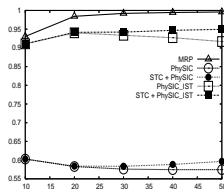
STC preprocess

- After that, the STC preprocess modifies the source trees (*PhySIC_IST*), forcing them not to contain the dropped resolutions.
- Each modified tree may contain either new multifurcations, or lack some of its former taxa.
- A threshold α is chosen by the user.

VOTE

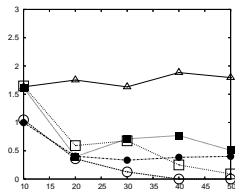
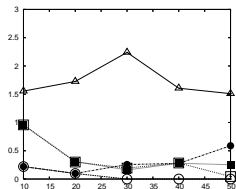
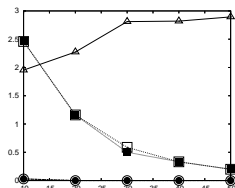
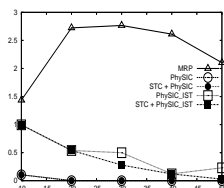
VETO



Large-scale simulations ($\alpha = 0.05$)Average *CIC* valuesMRP \triangle , *PhySIC* \circ , *PhySIC_IST* \square , STC+ *PhySIC* \bullet and STC+ *PhySIC_IST* \blacksquare  $d = 25\%$  $d = 50\%$  $d = 75\%$ mixed d

Large-scale simulations ($\alpha = 0.05$)

Average percentage of type I error

MRP \triangle , *PhySIC* \circ , *PhySIC_IST* \square , STC+ *PhySIC* \bullet and STC+ *PhySIC_IST* \blacksquare  $d = 25\%$  $d = 50\%$  $d = 75\%$ mixed d

Conclusions

- *PhySIC_IST*: new version of *PhySIC*
 - ▶ more informative but still reliable supertrees
- STC: a statistical preprocess of the source trees to detect and correct artifactual positions of taxa
- This approach has the advantage of separating the liberal resolution of conflicts in the data from the assemblage of the supertree.
 - ▶ feedback of the source trees
- Test STC+ *PhySIC_IST* on biological datasets

Conclusions

- http://www.atgc-montpellier.fr/physic_ist/

The screenshot shows the PhysIC_IST server website. At the top, there is a header with the ATGC logo (A, T, G, C) and the ISEM logo. Below the header, the text reads "LIRMM Montpellier bioinformatics platform ISEM". The main content area is titled "PhysIC_IST server: healing source trees to infer healthy supertrees." and includes the following information:

- Scornavacca C., Berry V., Douzery E.J.P., Ranwez V. Submitted to BMC Bioinformatics.
- Please cite THIS paper if you use PhysIC_IST.

The "PhysIC_IST online execution" section contains the following fields and options:

- Source tree file: File Example file
- Backbone tree file (optional):
- Outgroup file (optional):
- File format: UNIX Windows Mac
- Bootstrap threshold for source clade selection:
- Correction threshold used by STC:
- Your name:

A left sidebar contains a navigation menu with the following items: Home, Organization, Citations & Statistics, Online programs, PhysIC, PhysIC_IST, Downloads, Online execution, Papers & contacts, User's guide, Binaries, Databases, and Datasets.

Thanks

- Olivier Gascuel and Vincent Lefort
- Céline Brochier, Vincent Daubin and Frédéric Delsuc