

Notes I



Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

ReVa



APOGeT: Automated Phylogeny over Geological Time-scales

GODIN-DUBOIS Kevin & CUSSAT-BLANC Sylvain
& DUTHEN Yves

August 2, 2019

Taxonomy

Biological taxonomy:

- Manual clustering from population
- Increasing reliance on genomic data
- Further confirmations through fossil record

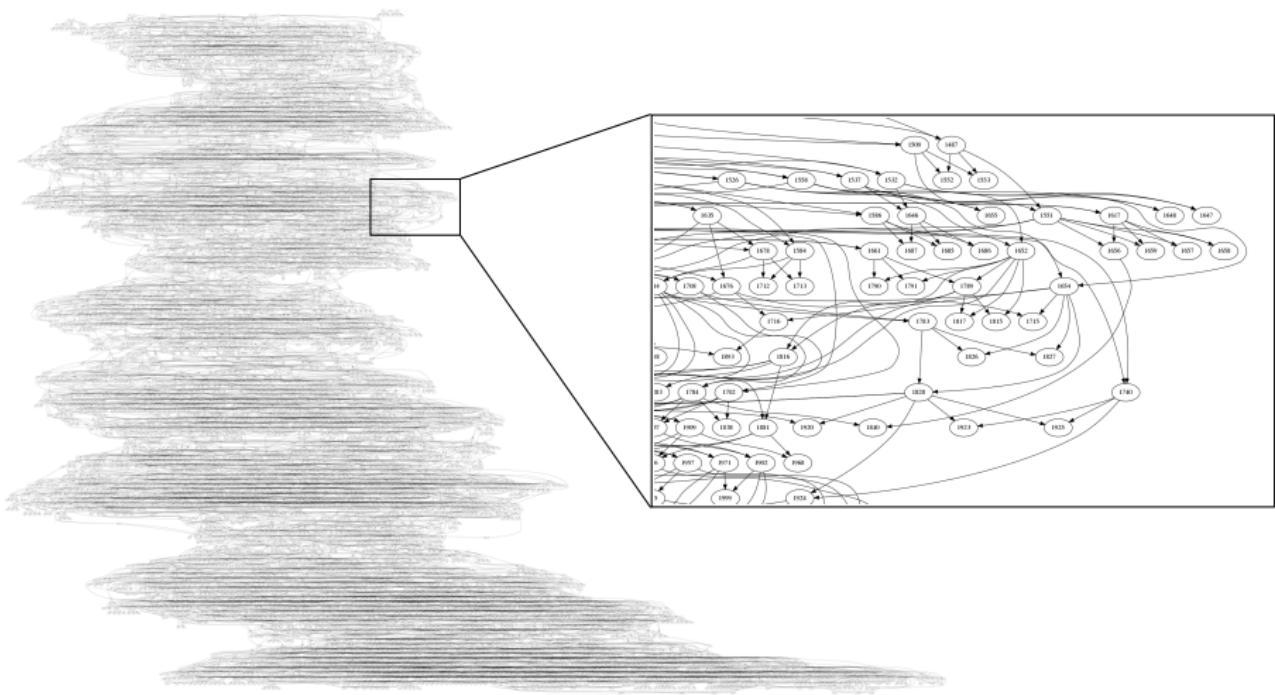
Taxonomy

Biological taxonomy:

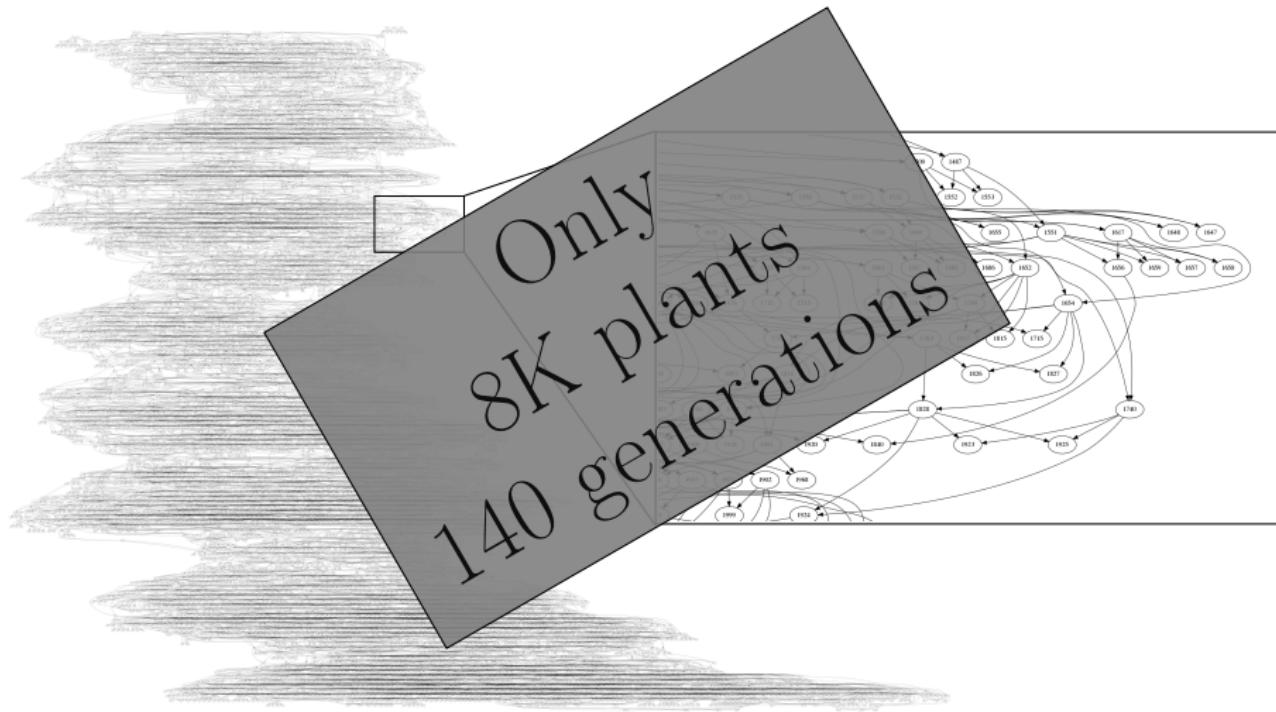
- Manual clustering from population
- Increasing reliance on genomic data
- Further confirmations through fossil record

No such limitations in ALife!

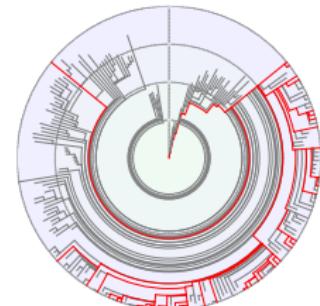
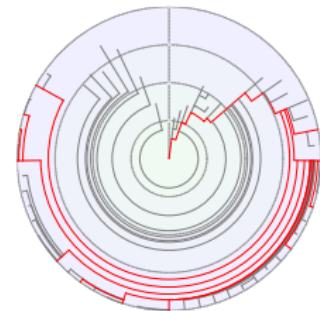
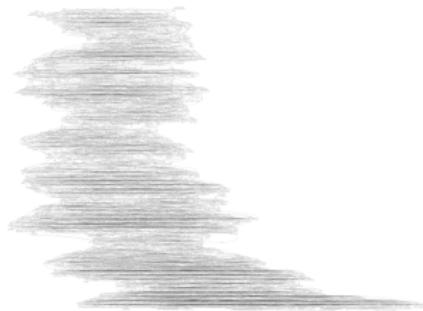
Genealogy?



Genealogy?



Genealogy?



Biological species

Used definition:

“... group of potentially interbreeding natural population reproductively isolated from other such groups.” Singh (2012)

Words of warning

From the genealogy's point of view:

- No species 'birth'
- No true tree of life

Therefore all classification is **arbitrary**

Forewords
○○○○○

APOGeT
●○○○○○○○○

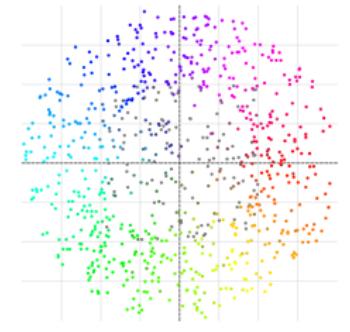
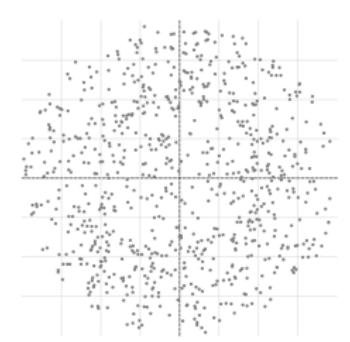
Parameters
○○○○○

Limitations
○○○○○

- Forewords
- APOGeT
 - ▶ Centroids vs Representatives
 - ▶ R-Set concept
 - ▶ Species affectation
 - ▶ Hybridism
 - ▶ Visualisation
- Parameters
- Limitations

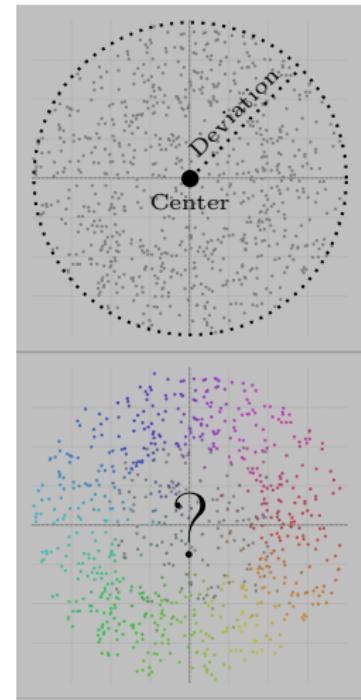
Centroids vs Representatives

- Smallest genome:
 - Position ($\in [0, 1]^2$)
 - Color in ($\in [0, 1]^3$)
- How to represent these species?



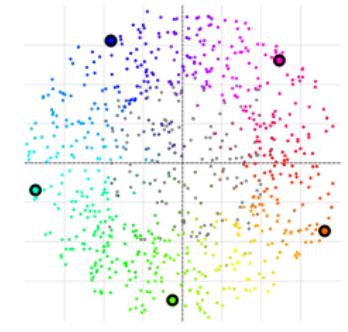
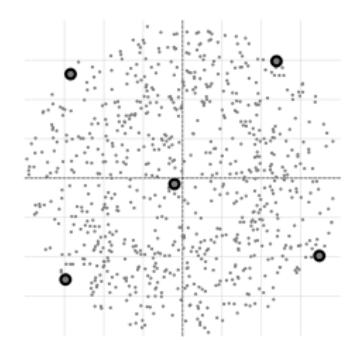
Centroids vs Representatives

- Smallest genome:
 - Position ($\in [0, 1]^2$)
 - Color in ($\in [0, 1]^3$)
- How to represent these species?
 - Centroids?



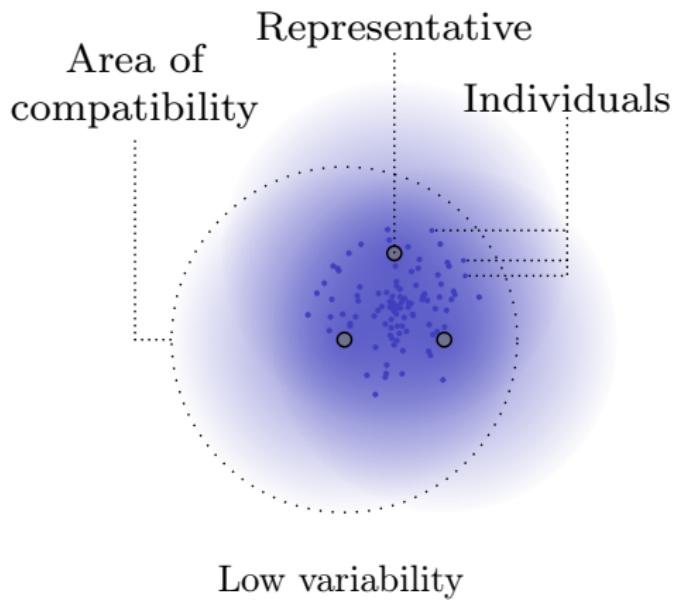
Centroids vs Representatives

- Smallest genome:
 - Position ($\in [0, 1]^2$)
 - Color in ($\in [0, 1]^3$)
- How to represent these species?
 - Centroid
 - Representatives



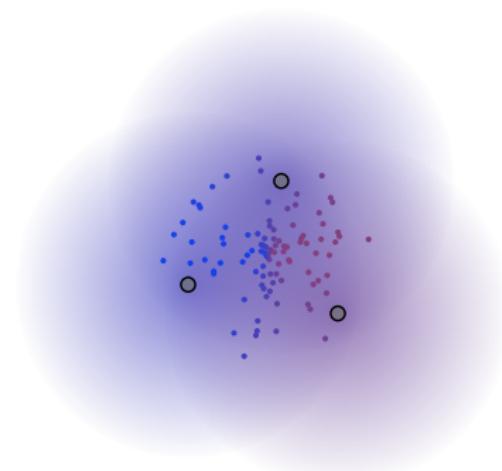
R-Set concept

Informally



R-Set concept

Informally



Diversity increasing

Forewords
○○○○○

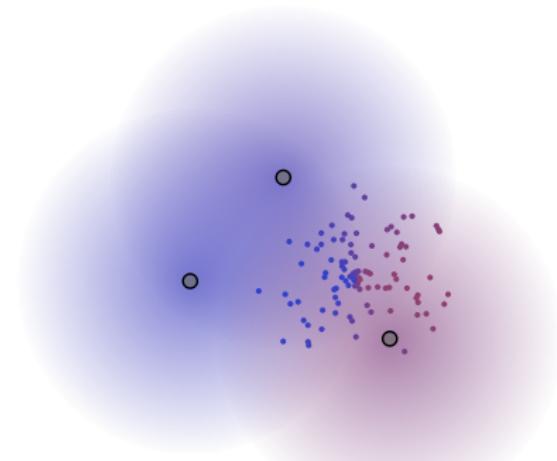
APOGeT
○○●○○○○○○

Parameters
○○○○○

Limitations
○○○○○

R-Set concept

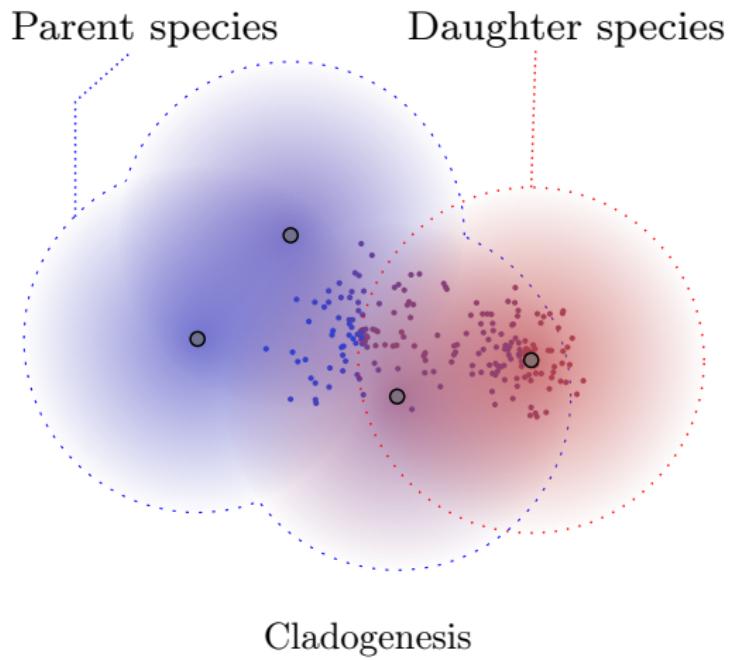
Informally



Low intra-compatibility

R-Set concept

Informally



R-Set concept

Formally

g : genome

R : R-Set

$r^i, r^j \in R$: representatives

$$C(r_i, g) = - \min_{r^j \neq r^i} dist(r^i, r^j) + \min_{r^j \neq r^i} dist(g, r^j)$$

$$C(R, g) = \max_{r^i \in R} C(r_i, g)$$

$C(R, g) \leq 0$, g is not a better representative
 > 0 , g replaces $r_i = \operatorname{argmax} C(R, g)$

Species affection

g : genome

S_0 : species of g 's parents

S_1, \dots, S_n : subspecies of S_0

$$d = dist(g, r_S)$$

$$xcompat(g, r_S) = \min(g.compat(d), r_S.compat(d))$$

$$match(g, S) = \frac{1}{|R_S|} \sum_{r_S} xcompat(g, r_S)$$

Species affection

$$\text{match}(g, S) = \frac{1}{|R_S|} \sum_{r_S} xcompat(g, r_S)$$

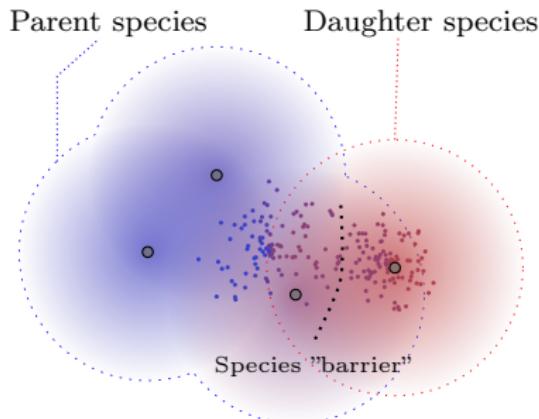
$\text{match}(g, S_0) \geq T$, insert in S

$\exists \underset{i \in [1, n]}{\text{match}}(g, S_i) \geq T$, insert in $\underset{i \in [1, n]}{\text{argmax}} \text{match}(g, S_i)$

otherwise, insert as new subspecies

Hybridism

- Potential subject of inquiry
- Guaranteed by-product of the algorithm
Child on other side of the species “barrier”



Hybridism

Changes to the algorithm

$S_0 \neq S_1$: species of g 's parents

$S_i^0, i \in [1, n]$: subspecies of S_0

$S_i^1, i \in [1, m]$: subspecies of S_1

$$S = \operatorname{argmax}_{S_0, S_1} \operatorname{match}(g, S_i)$$

$$\max_{S_0, S_1} \operatorname{match}(g, S_i) \geq T, \text{ insert in } S$$

$$\exists_{i \in [1, n+m]} \operatorname{match}(g, S_i) \geq T, \text{ insert in } \operatorname{argmax}_{i \in [1, n+m]} \operatorname{match}(g, S_i)$$

otherwise, insert as new subspecies of S

Hybridism

Major contributor

- Contribution pool $C_{S_i} = \{(S_j, c_j) \dots\}$ with S_j , species id and c_j contribution count
- Preserves graph information
- Parent species defined as the major contributor:

$$(S_j, c_j) \in C_i / \forall (S_k, c_k) \in C_i, c_k \leq c_j \\ \text{with } S_i \neq S_j \neq S_k$$

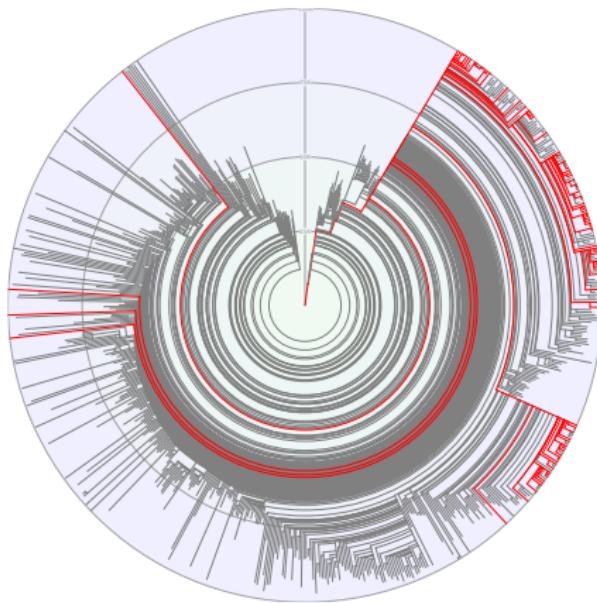
Forewords
○○○○○

APOGeT
○○○○○○○●

Parameters
○○○○○

Limitations
○○○○○

Visualisation



Full tree

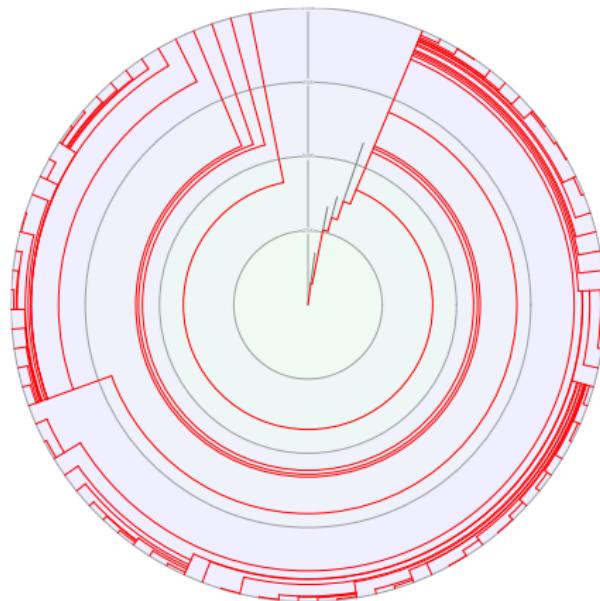
Forewords
○○○○○

APOGeT
○○○○○○○●

Parameters
○○○○○

Limitations
○○○○○

Visualisation



Survivors only

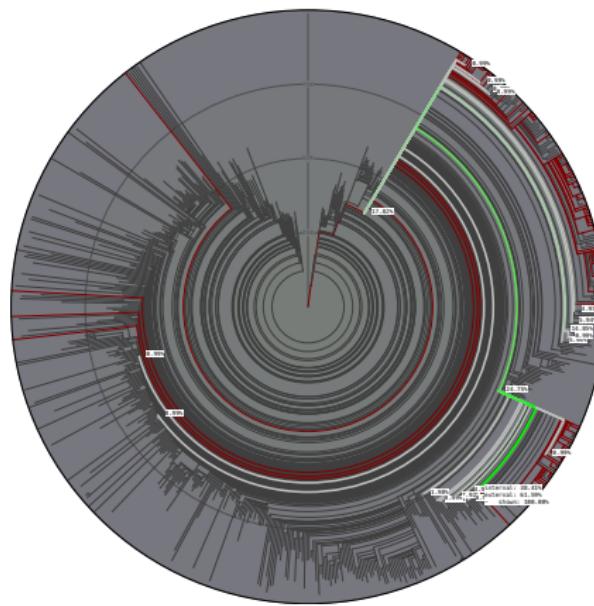
Forewords
○○○○○

APOGeT
○○○○○○○●

Parameters
○○○○○

Limitations
○○○○○

Visualisation



Hybridism

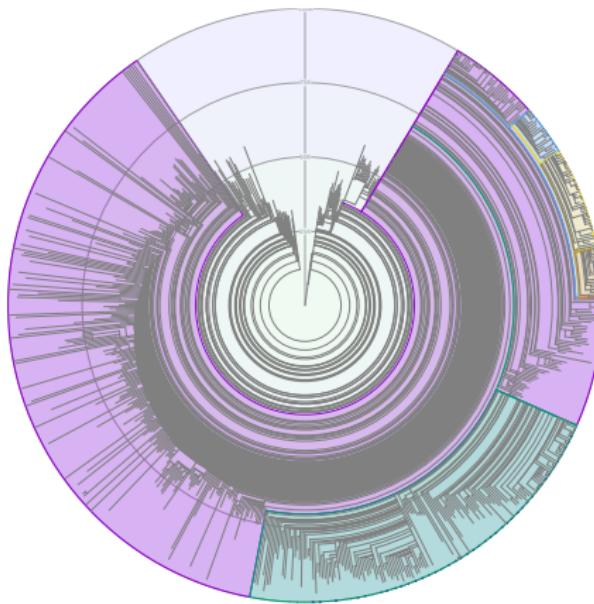
Forewords
○○○○○

APOGeT
○○○○○○○●

Parameters
○○○○○

Limitations
○○○○○

Visualisation



Taxonomic levels

Forewords
ooooo

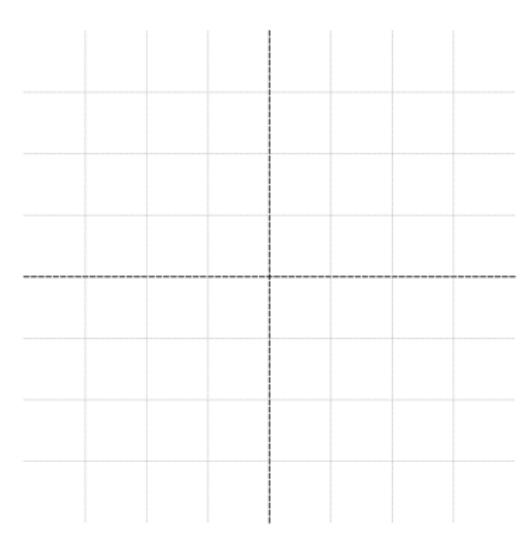
APOGeT
oooooooooo

Parameters
●○○○○

Limitations
oooo

- Forewords
- APOGeT
- Parameters
 - ▶ Belonging threshold
 - ▶ R-Set size
 - ▶ Compatibility function
- Limitations

Parameters



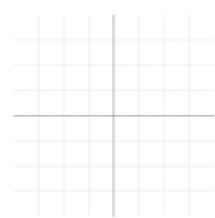
Waiting for input

- $T = .25$, minimal matching score threshold
- $K = 5$, rset size
- $C = BOC^1$, compatibility function

¹as defined in Godin-Dubois et al. (2019)

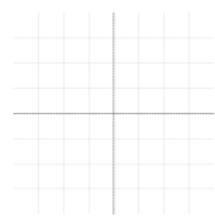
Belonging threshold

$T = .125$



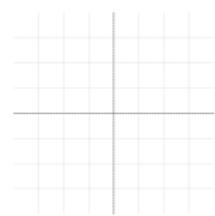
Waiting for input

$T = .25$



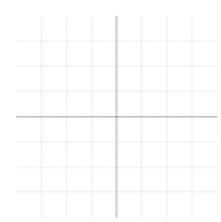
Waiting for input

$T = .5$

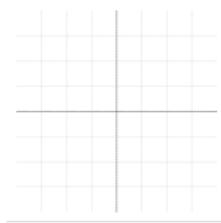


Waiting for input

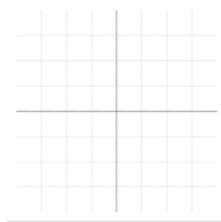
R-Set size

 $K = 2$ 

Waiting for input

 $K = 5$ 

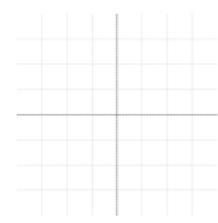
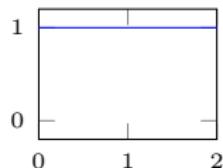
Waiting for input

 $K = 10$ 

Waiting for input

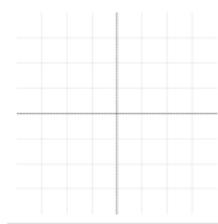
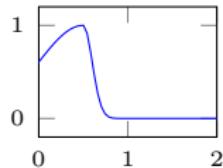
Compatibility function

Constant



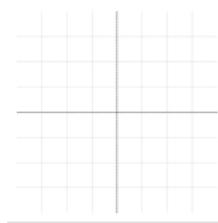
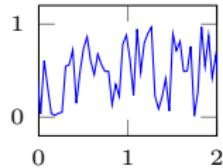
Waiting for input

BOC



Waiting for input

Random



Waiting for input

Forewords

○○○○○

APOGeT

○○○○○○○○○○

Parameters

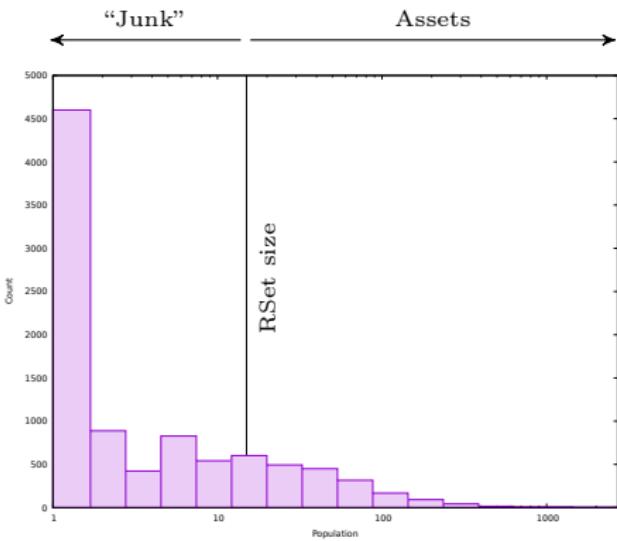
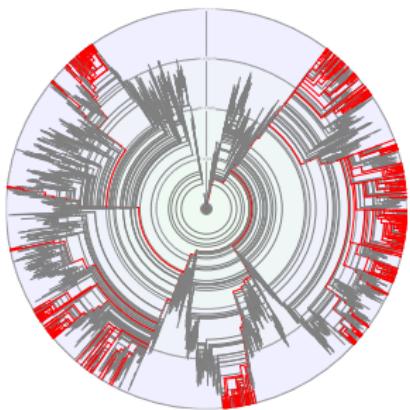
○○○○○

Limitations

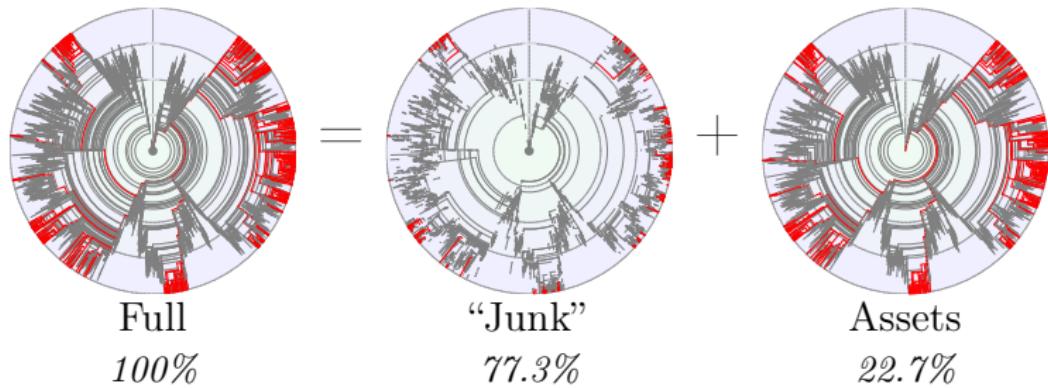
●○○○

Limitations

Bloating



Bloating



Forewords
○○○○○

APOGeT
○○○○○○○○○○

Parameters
○○○○○

Limitations
○○●○

MiSC

- Parameter calibration
- Compatibility function design
- Horizontal gene transfer

Dependencies

- C++17 & Qt ≥ 5
- Tools repository (*configuration, introspection, ...*)

<https://github.com/kgd-al>

- json (*serialization*)

<https://github.com/nlohmann/json>

- cxxopts (*command line arguments parsing*)

<https://github.com/jarro2783/cxxopts>

Summary

- Clustering of genomes stream
- Applicable across a broad range of substrates
- Built-in visualisation tools
- Handles both sexual and asexual genealogy

Source



Videos



References

○

Annexes

- References

References



Kevin Godin-Dubois, Sylvain Cussat-Blanc, and Yves Duthen. “Speciation under changing environments”. In: *The 2019 Conference on Artificial Life*. Vol. 31. Cambridge, MA: MIT Press, 2019, pp. 349–356.



B N Singh. “Concepts of species and modes of speciation”. In: *Current Science* 103.7 (2012).