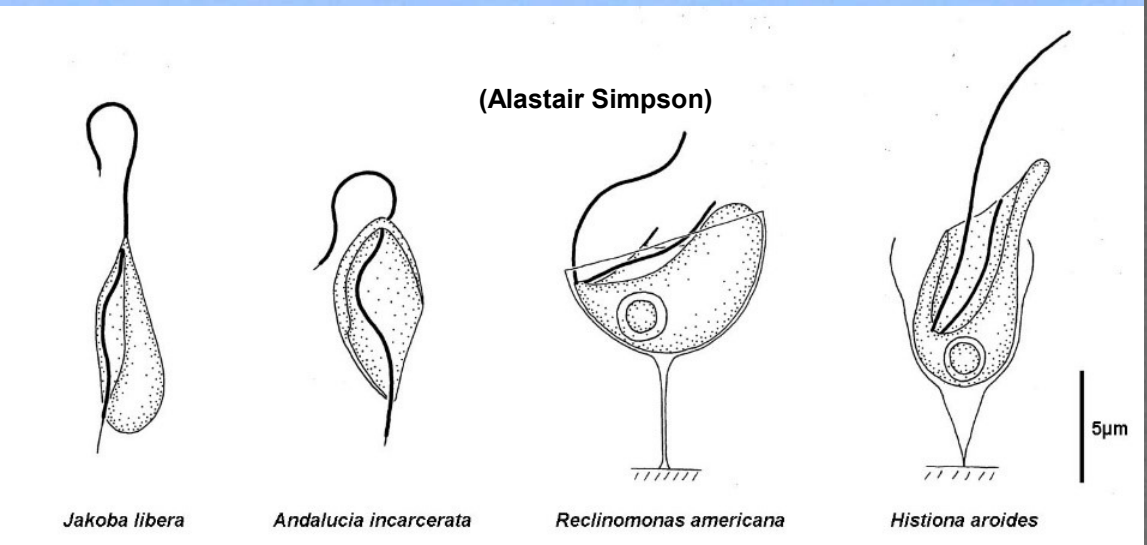
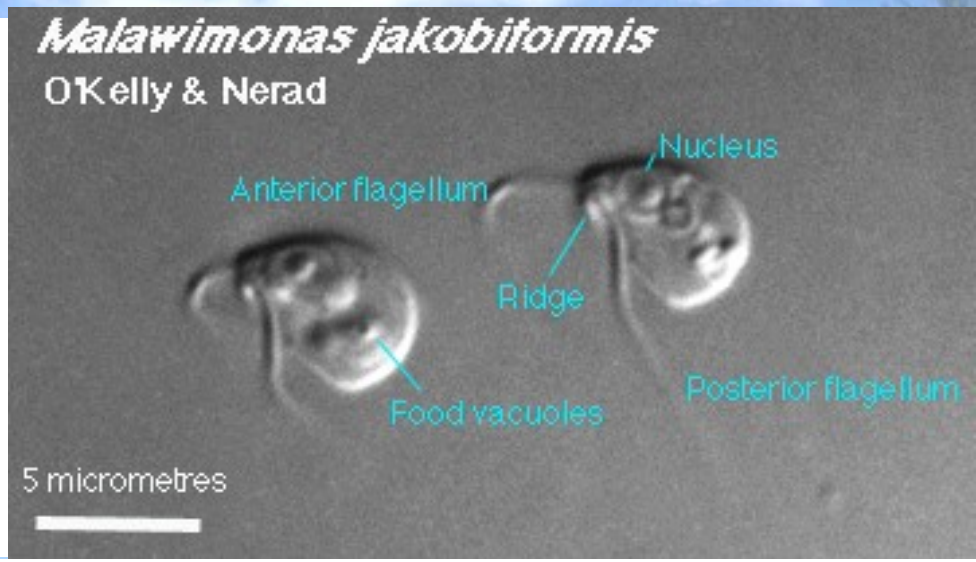
Aligning and modeling a few thousand unknown ncRNAs in 'primitive' unicellular eukaryotes

S. Prince, M. Sarrasin, G. Burger and B.F. Lang

Department of Biochemistry, Robert-Cedergren Centre of Bioinformatics and Genomics, Université de Montréal, Montréal, QC, H3C 3J7, Canada





Topics of my presentation

- **Why jakobids and malawimonads?**
- **Why does RFAM provide little help in ncRNA finding – an evolutionary approach to updating CMs**
- **How to build quality CMs from a few thousand single ncRNAs sequences**

Species selection:

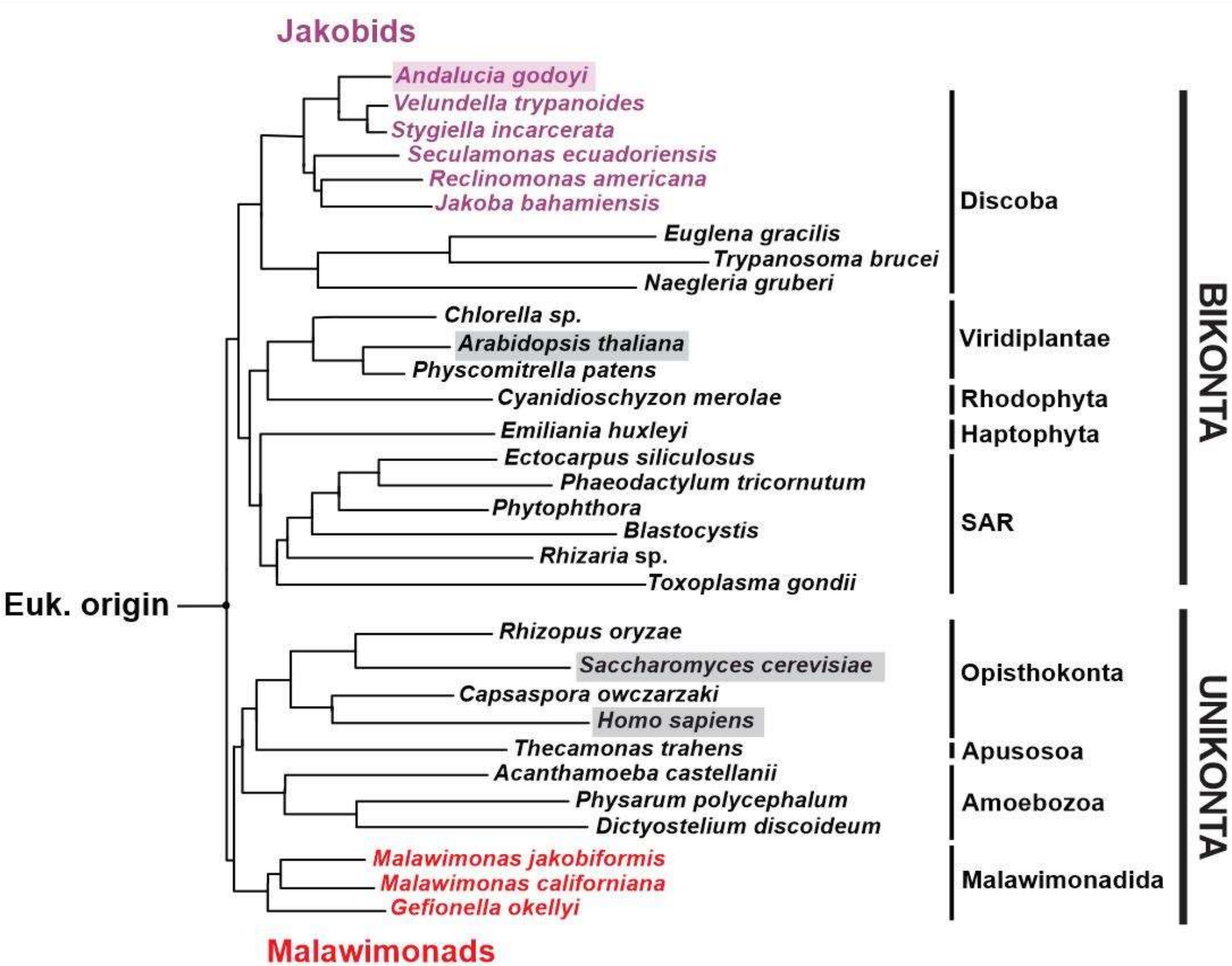
Jakobids:

Andalucia godoyi
Jakoba bahamiensis
Jakoba libera
Reclinomonas americana
Seculamonas
Stygiella incarcerata
Velundella trypanoides

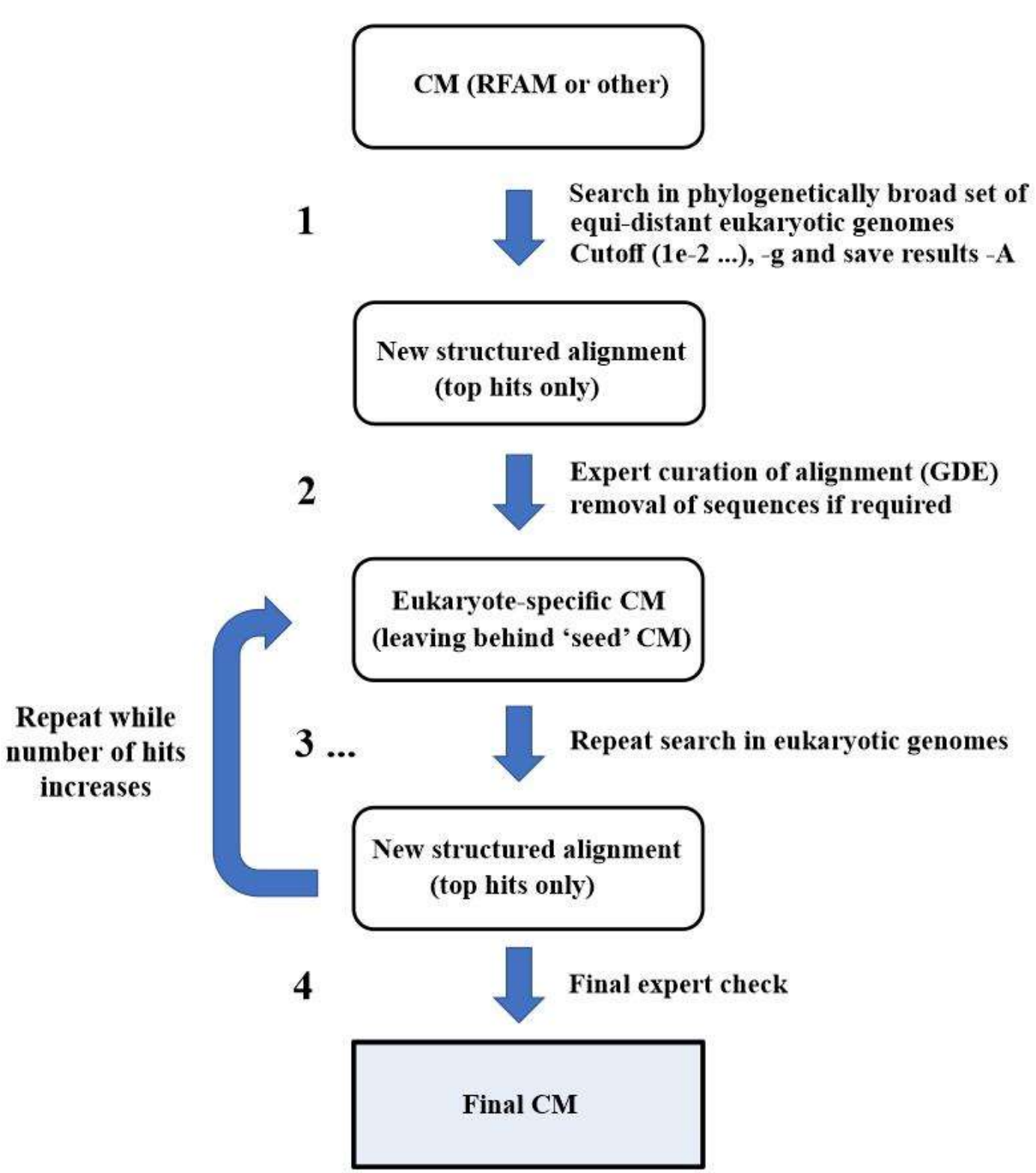
Malawimonads:

Malawimonas californiana
Malawimonas jakobiformis
Malawimonas sp.

All require live bacteria as food source – contamination issues


Jakobids
Andalucia godoyi
Velundella trypanoides
Stygiella incarcerata
Seculamonas ecuadoriensis
Reclinomonas americana
Jakoba bahamiensis
Euglena gracilis
Trypanosoma brucei
Naegleria gruberi
Chlorella sp.
Arabidopsis thaliana
Physcomitrella patens
Cyanidioschyzon merolae
Emiliania huxleyi
Ectocarpus siliculosus
Phaeodactylum tricornutum
Phytophthora
Blastocystis
Rhizaria sp.
Toxoplasma gondii
Euk. origin
Rhizopus oryzae
Saccharomyces cerevisiae
Capsaspora owczarzaki
Homo sapiens
Thecamonas trahens
Acanthamoeba castellanii
Physarum polycephalum
Dictyostelium discoideum
Malawimonas jakobiformis
Malawimonas californiana
Gefionella okellyi
Malawimonads
Discoba
Viridiplantae
Rhodophyta
Haptophyta
SAR
Opisthokonta
Apusosoa
Amoebozoa
Malawimonadida
BIKONTA
UNIKONTA

RFAM CMs do not find much in these protist genomes as (i) seed alignments are often poor, with almost as many alignment gaps as aligned positions, and (ii) they are biased towards animals, fungi and plants. What can be done about it?



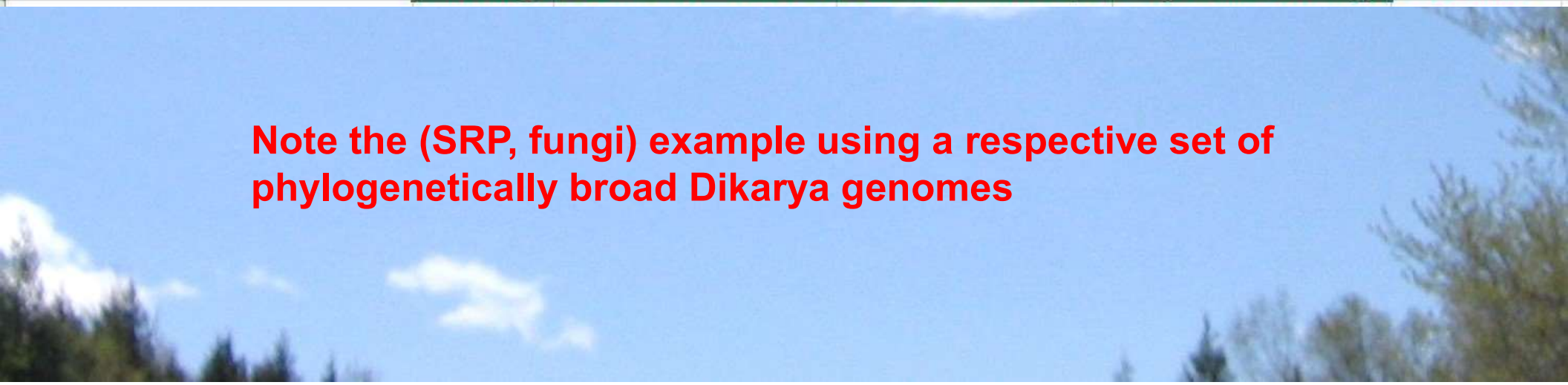
Evolutionary approach to updating CMs

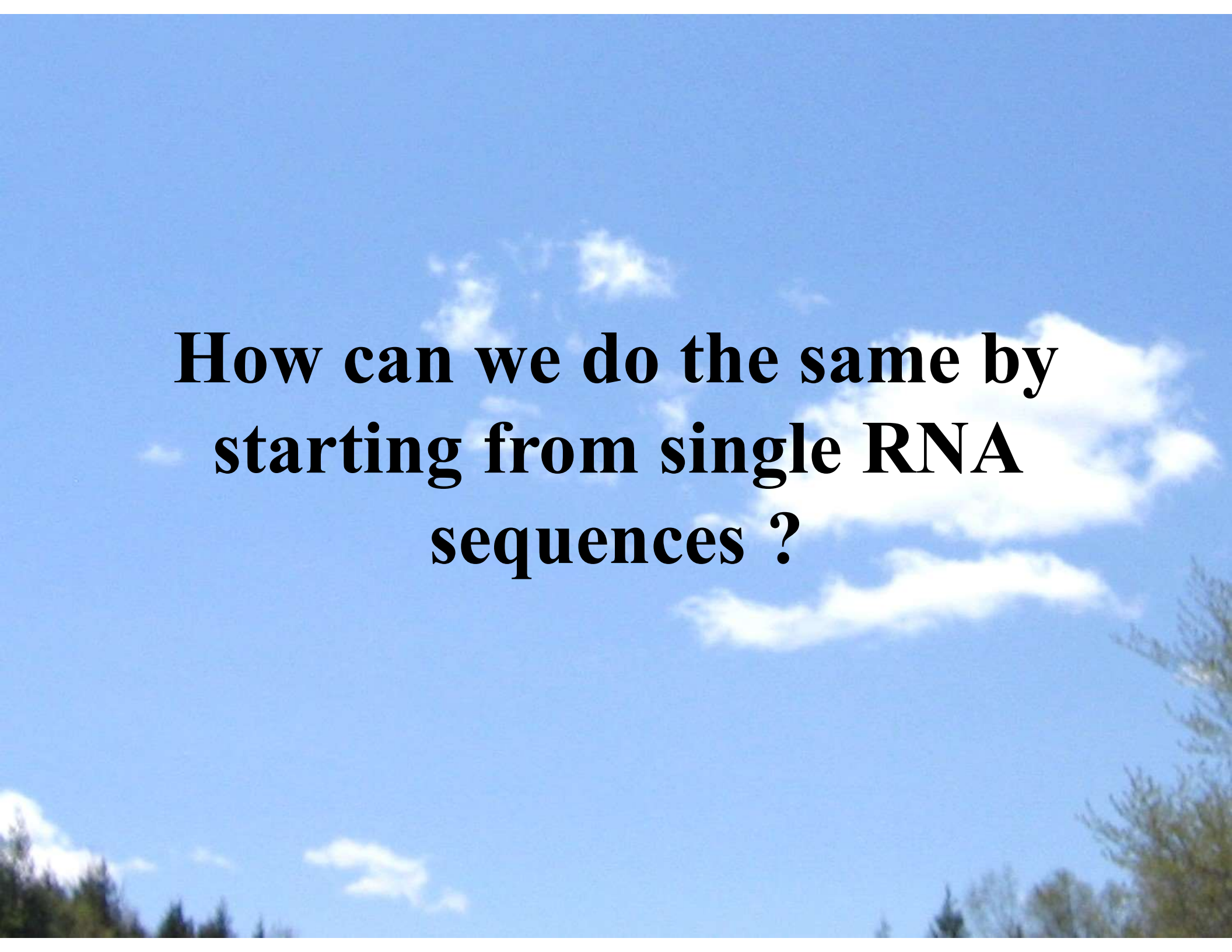
Will they be as good as the original CMs?

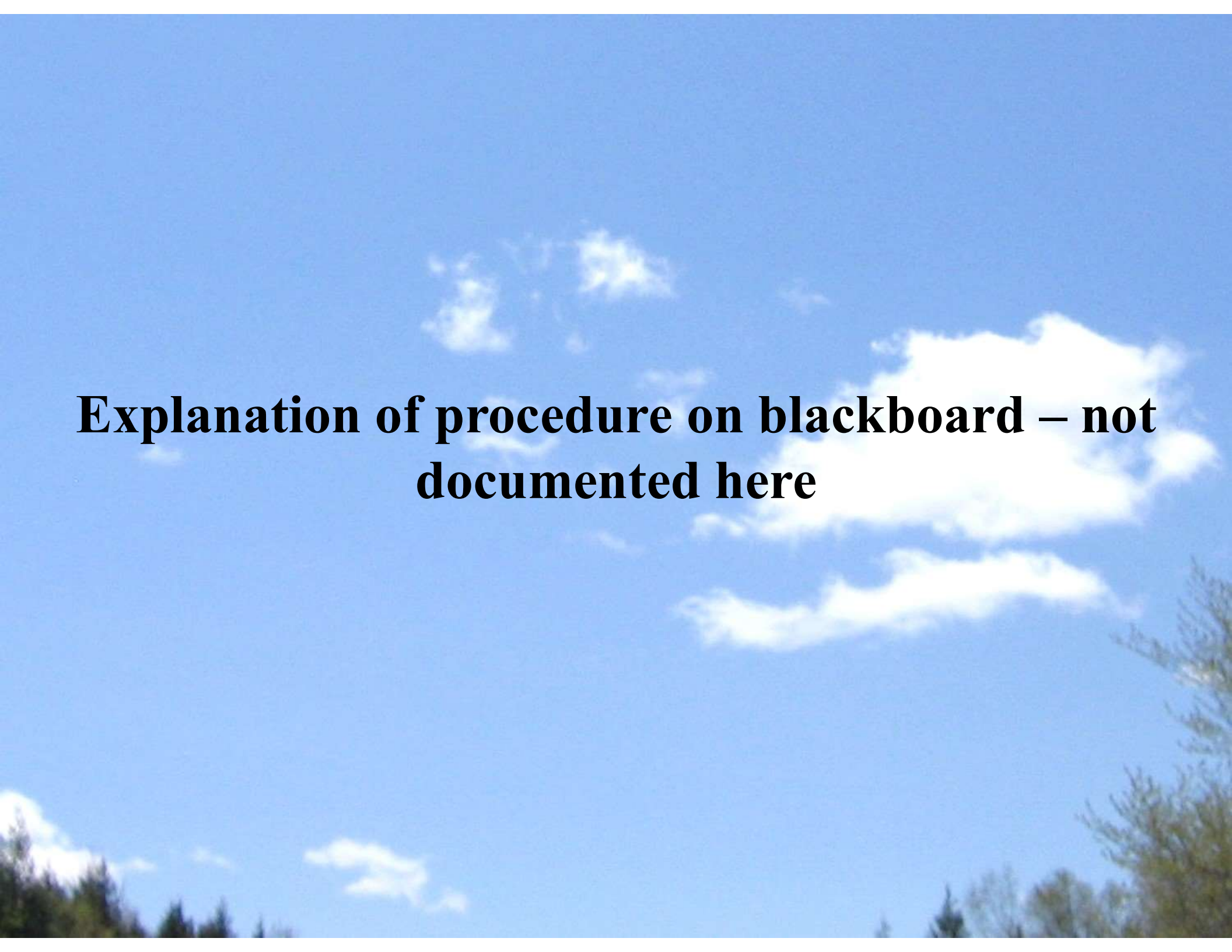
Comparison of CM model performance - RFAM *vs* Cmit

Collections of eukaryotic, phylogenetically broad genomes used (NC records, + 12 of our protist nuclear genomes)

RNA	RFAM #	RFAM models	Genomes	Cmiter models	# iterations
RNase P, nuclear	RF00009	60 (71) 1.6e-39	Eukaryota (107)	62 (74) 1.7e-49	2
MRP	RF00030	80 (89) 3.9e-35	Eukaryota (107)	86 (96) 2.3e-40	4
SRP, protist	RF01856	80 (220) 1.2e-30	Eukaryota (107)	84 (232) 4e-39	3
SRP, fungi	RF01502	115 (117) 6.1e-43	Dikarya (177)	155 (202) 2.7e-55	8
SRP, plant	RF01855	68 (208) 6.2e-72	Eukaryota (107)	85 (244) 5.6e-53	3
U1	RF00003	76 (295) 1.9e-42	Eukaryota (107)	97 (816) 5.5e-55	3
U2	RF00004	98 (744) 4.6e-36	Eukaryota (107)	98 (835) 2.8e-50	2
U4	RF00015	86 (186) 1.1e-22	Eukaryota (107)	89 (201) 2.7e-35	3
U5	RF00020	70 (235) 2.3e-14	Eukaryota (107)	74 (247) 1.1e-17	5
U6	RF00026	98 (503) 1.2e-25	Eukaryota (107)	97 (499) 1.8e-27	2
U4atac	RF00618	6 (6) 2.2e-07	Eukaryota (107)	23 (48) 9.2e-24	2
U6atac	RF00619	28 (35) 8.5e-24	Eukaryota (107)	25 (39) 6.9e-42	3
U11	RF00548	25 (28) 2e-14	Eukaryota (107)	28 (31) 1e-33	3
U12	RF00007	22 (25) 9.3e-17	Eukaryota (107)	25 (29) 6.8e-29	2
	Meaning of numbers: total # of hits in distinct species (total # of hits) best E-value				




How can we do the same by starting from single RNA sequences ?


Explanation of procedure on blackboard – not documented here

Results of an iterative search with a presumed single mito *rnpB* sequence in Dipodascaceae yeasts

```
Query:       new  [CLEN=420]
Hit scores:
 rank     E-value  score  bias  sequence        start    end   mdl trunc   gc  description
 ----   --------- ------ -----  -------------- ------ ------   --- ----- ----  -----------
  (1) !     2e-45  384.4  15.2  Geot.cand.5768  23220  23544 +  cm    no 0.32  Geotrichum.candidum.CLIB918 Eukaryota;Fungi;Dikarya;Asco
  (2) !   6.7e-27  226.3  83.3  Magn.capi.4095  10338  10091 -  cm    no 0.15  Magnusiomyces.capitatus.NRRLY-17686 Eukaryota;Fungi;Dika
  (3) !   5.2e-16  133.3   0.3  Magn.magn.0043  10876  10702 -  cm    no 0.36  Magnusiomyces.magnusii.CBS234.85 Eukaryota;Fungi;Dikarya
  (4) !   1.6e-15  129.1   0.1  Sapr.suav.3801  10120   9947 -  cm    no 0.37  Saprochaete.suaveolens.NRRLY-17571 Eukaryota;Fungi;Dikar
  (5) !   2.4e-12  102.1  94.7  Magn.inge.4093   8314   8126 -  cm    no 0.08  Magnusiomyces.ingens.NRRLY-17630 Eukaryota;Fungi;Dikarya
  (6) !   2.4e-12  102.1  94.7  Sapr.inge.6489   9487   9299 -  cm    no 0.08  Saprochaete.ingens.NRRLY-7930 Eukaryota;Fungi;Dikarya;As
  (7) !   4.1e-12  100.1   0.0  Magn.tetr.4094  17518  17683 +  cm    no 0.42  Magnusiomyces.tetraspermus.NRRLY-7288 Eukaryota;Fungi;Di
  (8) !   8.6e-10   80.2   0.0  Sapr.fung.3800  18389  18230 -  cm    no 0.41  Saprochaete.fungicola.CBS625.85 Eukaryota;Fungi;Dikarya;
```

Results of an iterative ERPIN search based on the previous CM alignment

```
***************************************************
Condensed alignment for magnusio.epn-16.fullalignment
***************************************************
There are 16 matches

Species name                   E-value   # of nt     Start..Stop      Str  #  Structure
------------------------------ --------- -------- ------------------- -- --- ------------------
                                                                           [[[[[[[[[[[    [[[   [[[[[                                                                                                                   ]]]]] ]]] ]]]]]]]]]]]
Geot.cand.5768               |  2.96e-20    29008    23220..23544     FW   1 ATTAATATAAT GTA AAG T CTAAT aaa...(22)...agg  TACTATATAGAAAT aag..(151)...taa  ATTAAT aat...(68)...tta  TTAGTGAAATAA tt                ATACTAAA ATTAG CTT ATTATATTAAT
Magn.capi.4095               |  1.74e-19    43486    10091..10338     RC   1 ATTAATATAAT GAA AAG T CTAAA taa...(24)...agg  TACTTAATAGAAAT taa...(80)...ata  ATAAAT cct...(32)...tta  AAAGTGAAATAA ata...(30)...tta ATACAAAA ATTAG CTT ATTATATTAAT
Magn.inge.4093               |  9.09e-20    37684     8126..8314      RC   1 ATTAATATTAT GAA AAG T CTAAT tat...(17)...tgg  AACTTTATAGAAAT ata...(45)...aaa  TCCAAT att...(35)...tat  ATAGTGAAATTA tattaatata        ATACAAAA ATTAG CTT ATAATATTAAT
Magn.magn.0043               |  4.23e-22    42757    10702..10876     RC   1 GTCAATAATAT GTA AAG T CATGT aat...(28)...tgg  CACTAAGGAGATAA gaa...(33)...aag  ACCAAT cca...(20)...aac  TTAGTGAGATGA tatgagaacata      ACACAAGA ACTAG CTT ACATTATTGAT
Magn.tetr.4094               |  3.04e-20    44469    17518..17683     FW   1 GTCAATAATAA GTA AAG T CATGT aat...(18)...agg  CACTGAGGAGAGAA gac...(34)...gaa  TCTAAT cct...(20)...agc  TTAGTGAGATGA gatctgcggatg      GCACGGGA ACTAG CTT ACATTATTGAT
Sapr.fung.3800               |  1.39e-14    33027    18231..18388     RC   1 TGCAATAGTAT GTA AAG T CATGG aatcggcagtgtctgg  AACTCGATAGAAGA tac...(30)...taa  AGGAAT cca...(18)...gca  CGAGTGTGATCA catgtgaacatc      GAACAAAA ACTAG CTT ATATTATTGGA
Sapr.suav.3801               |  8.32e-24    49739     9947..10120     RC   1 GTCAATAATAT GTA AAG T CATGT aat...(28)...tgg  CACTAAGGAGAGAA aag...(32)...aag  ACCAAT cca...(20)...aac  TTAGTGAGATGA tatgagaacata      ACACAAGA ACTAG CTT ACATTATTGAT
Magn.oveten                  |  9.96e-20    45300    30427..30593     RC   1 ATTAATATTAT GAA AAG T CTAAT aat...(17)...tgg  AACTTAATAGAAAT aaa...(40)...aaa  TCCAAT ata...(22)...gga  AAAGTGAAATAA aaatta            AAACAAAA ATTAG CTT ATAATATTAAT
Magn.starmeriCBS780          |  5.17e-19    38526     9021..9191      RC   1 ATTAATATTAT GAA AAG T CTAAA tat...(23)...agg  AACTAAATAGAAAT taa...(36)...aaa  TTTAAT cct...(24)...aat  ATAGTGAAATTA tatatt            ATACTAAA ATTAG CTT ATAATATTAAA
Sapr.gigasCBS126             |  4.23e-22    48823    11063..11237     RC   1 GTCAATAATAT GTA AAG T CATGT aat...(28)...tgg  CACTAAGGAGATAA gaa...(33)...aag  ACCAAT cca...(20)...aac  TTAGTGAGATGA tatgagaacata      ACACAAGA ACTAG CTT ACATTATTGAT
Sapr.saccharoCBS_252-91      |  4.47e-12    30915     7176..7339      RC   1 TACCAATATTG GGA AAG T CTAAG aac...(18)...agg  CACTGAGGAGAAAA aag...(35)...caa  GAAAAC cct...(17)...agt  GAAGTGAAAAAA aatgtgaacata      ACACAGGA ACTAG CTT ACAATATTGAT
Dipod.albidusCBS766.85       |  2.50e-17    58286    39101..39344     FW   1 TTTAATATAAT AGA AAG T CTAAA tat...(17)...agg  AACTAAAAAGAAAG agt...(83)...ata  ATAAAT att...(17)...att  TTAGTGAAATCA ata...(45)...att GTATCGAA ATTAG CTT ATTATATTAAT
Dipo.geniculatus             |  8.53e-18    79898    69904..70077     FW   1 ATTAATATAAT AGA AAG T CTAAA tat...(18)...agg  AACTAAAAAGAAAG aga...(32)...aaa  AATAAT cct...(27)...ata  CTAGTGAAATCA ataaaaataattatt  GTATCGAA ATTAG CTT ATTATATTAAT
Sapr.psychrCBS765.8          |  5.74e-19    40642    30369..30542     FW   1 ATAAATATAAT GTA AAG T CTAAA tat...(19)...agg  AACTATATAGAAAA taa...(56)...att  ATTAAT ata...(17)...ata  ATAGTGAAATAA                   AAACTAAA ATTAG CTT ACTATATTTAT
Gala.candidusGEOT13          |  2.96e-20    29011    23223..23547     FW   1 ATTAATATAAT GTA AAG T CTAAT aaa...(22)...agg  TACTATATAGAAAT aag..(151)...taa  ATTAAT aat...(68)...tta  TTAGTGAAATAA tt                ATACTAAA ATTAG CTT ATTATATTAAT
Gala.reessii                 |  1.65e-20    36080     9846..10144     RC   1 ATTAATATAAT GTA AAG T CTAAT gag...(32)...agg  TACTATATAGAAAT aaa..(161)...gct  ATAAAT ttt...(22)...tca  TTAGTGAAATAA tt                ATACTAAA ATTAG CTT ATTATATTAAT
```

More homologs are found with ERPIN, with hits in the e-14 to e-20 range and for all available genomes (Infernal has issues with the extreme AT content …).

Other examples are inference of new group I intron classes from single unidentified potential group I introns

➔ To be applied to potential ncRNA in *Andalucia*

Conclusions

- **Improving CMs from public sources**
- **Iterative procedure to go from single RNA sequences to CMs**
- **Finding previously unknown ncRNAs (mito *rnpB*, group I introns …)**
- **To be extended and streamlined for thousands of potential ncRNAs**

Thanks to collaborators of this project …

Romain Derelle (Barcelona)
Alastair Simpson (Halifax, Canada)
Marek Elias (Czech Republic)
Andrew Roger (Halifax, Canada)
Mike W. Gray (Halifax, Canada)
David Morse (Montreal, Canada)
Eric Nawrocki

and financing from

NSERC CRSNG

Université de Montréal

Fonds de recherche Nature et technologies Québec