Marius Audenis, Gabriel Chiche, Leana Ortolani
Supervisor: Carlos Pulido Quetglas

HMM based classification for conserved protein domains (SNARE)

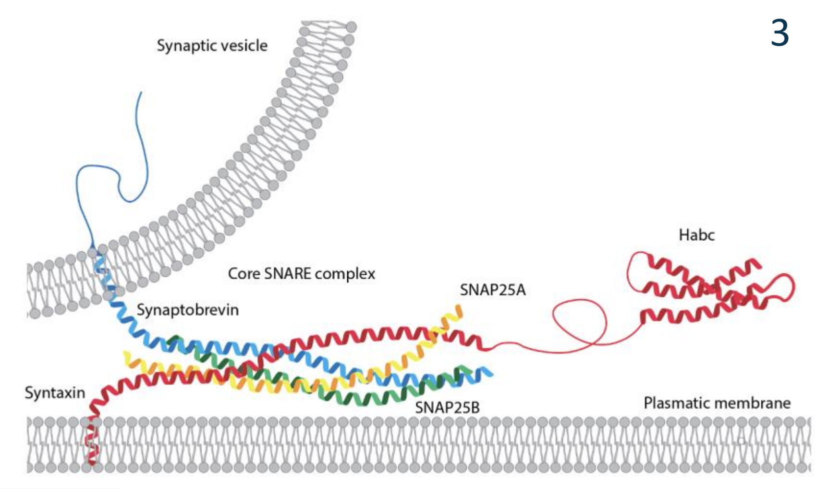
01

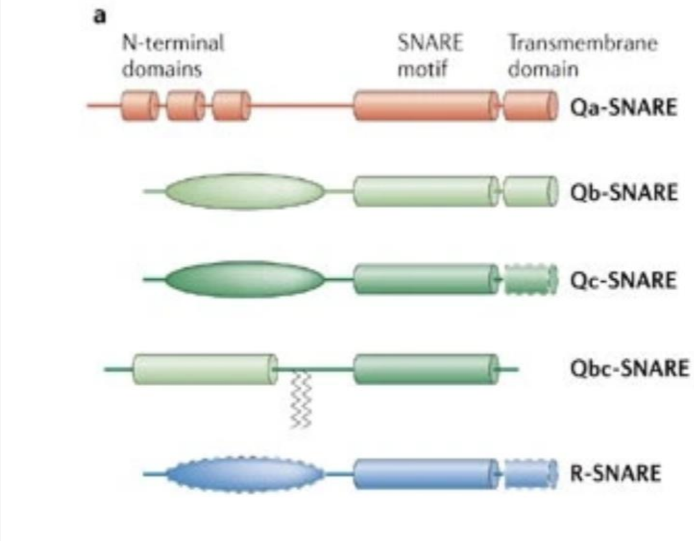
Background

SNAREs

- Role : membrane fusion
- Conserved domain : SNARE motif
- Classification : Qa, Qb, Qc, R, SNAP

- **Goal** : automatise the classification with a model
- Classification can be used to hypothesize important informations about a protein (function, location, etc…)

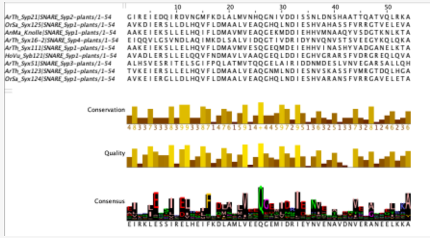




Classification – broad picture

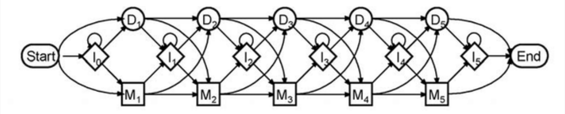
Sequences

MSA

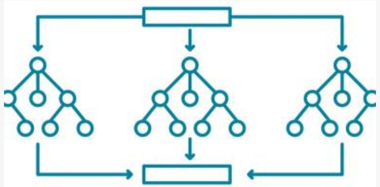


```
                                         10        20        30        40        50
ArTh_Syp21|SNARE_Syp2-plants/1-54     GIREIEDQIRDVNGMFKDLALMVNHQGNIVDDISSNLDNSHAATTQATVQLRKA
OrSa_Syx125|SNARE_Syp1-plants/1-54    AVKDIERSLLDLHQVFLDMAALVEAQGHQLNDIESHVAHASSFVRRGTVELEVA
AnMa_Knolle|SNARE_Syp1-plants/1-54    AAKEIEKSLLELHQIFLDMAVMVEAQGEKMDDIEHHVMNAAQYVSDGTKNLKTA
ArTh_Syx16-2|SNARE_Syp4-plants/1-54   EIQQVLGSVNDLAQIMKDLSALVIDQGTIVDRIDYNVQNVSTSVEEGYKQLQKA
ArTh_Syx111|SNARE_Syp1-plants/1-54    AAKEIEKSLLELHQVFLDMAVMVESQGEQMDEIEHHVINASHYVADGANELKTA
HoVu_Syb121|SNARE_Syp1-plants/1-54    AVADLERSLLELQQVFNDMAVLVAAQGEQLDDIEGHVGRARSFVDRGREQLQVA
ArTh_Syx51|SNARE_Syp3-plants/1-54     ALHSVESRITELSGIFPQLATMVTQQGELAIRIDDNMDESLVNVEGARSALLQH
ArTh_Syx123|SNARE_Syp1-plants/1-54    TVKEIERSLLELHQVFLDMAALVEAQGNMLNDIESNVSKASSFVMRGTDQLHGA
OrSa_Syx124|SNARE_Syp1-plants/1-54    AVKEIERGLLDLHQVFLDMAALVEAQGHQLNDIESHVARANSFVRRGAVELETA

                        Conservation
                                      48337333839933871476159144459729513632513373281246236
                             Quality

                           Consensus
                                      EIRKLESSIRELHEIFKDLAMLVEEQGEMIDRIEYNVENAVDNVERANEELKKA
```

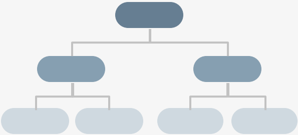
HMM profiles



Scores and E-values



Random forest



Classification

Goal : input = sequence | output = classification of the sequence (group & subgroup)

02

MSA

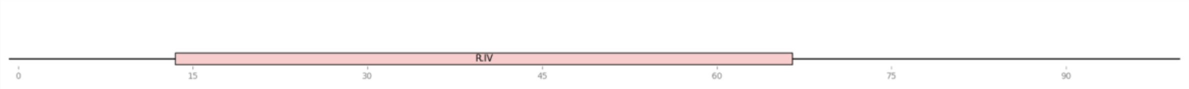
Multiple Sequence Alignment

Data Collection

- Tracey database : SNAREs protein collection
- Motif sequences and full sequences
- Sequences from specific taxa :
  - Archaeplastida --> Viridiplantae
  - Opisthokonta --> Metazoa and Fungi

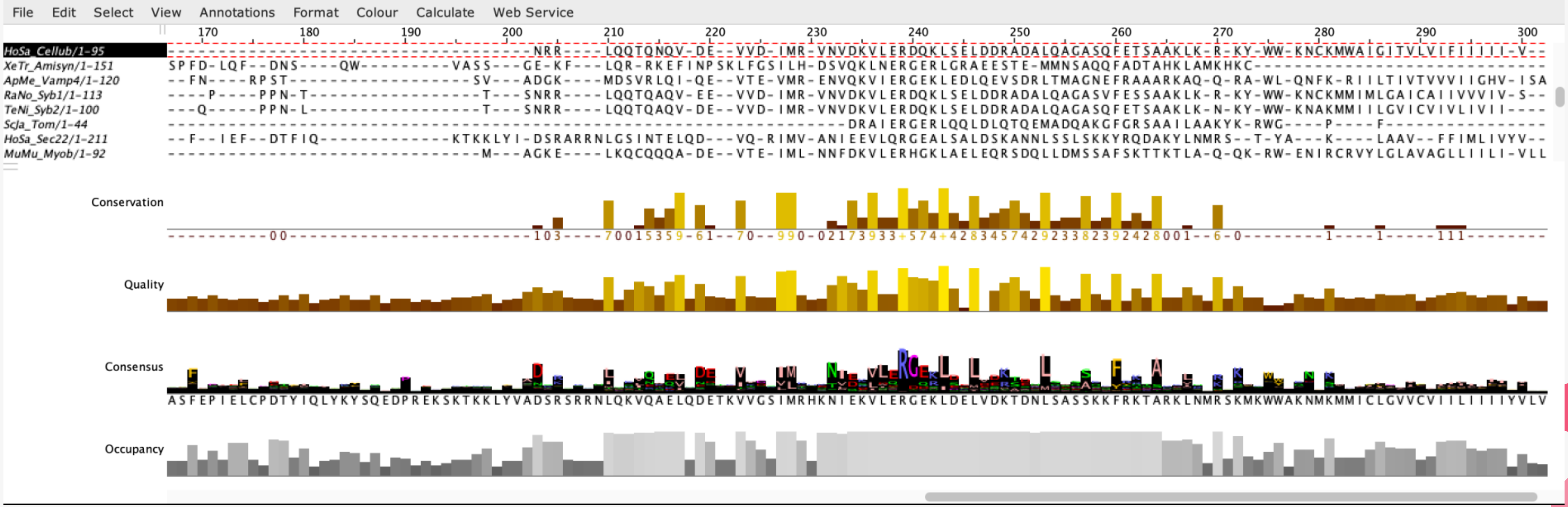
Alignment

HoSa_Cellub



Sequence with motif:

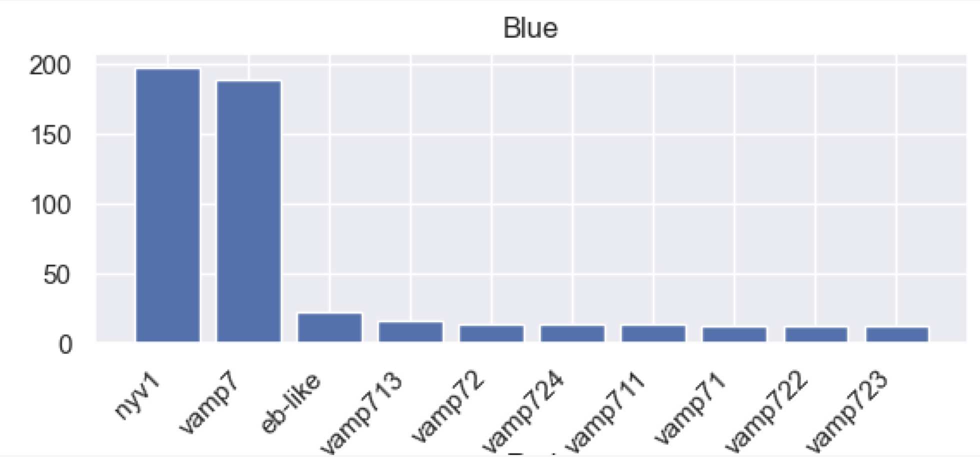
MSTGPTAATGSNRRLQQTQNQVDEVVDIMRVNVDKVLERDQKLSELDDRADALQAGASQFETSAAKLKRKYWWKNCKMWAIGITVLVIFIIIIIVWVVSS

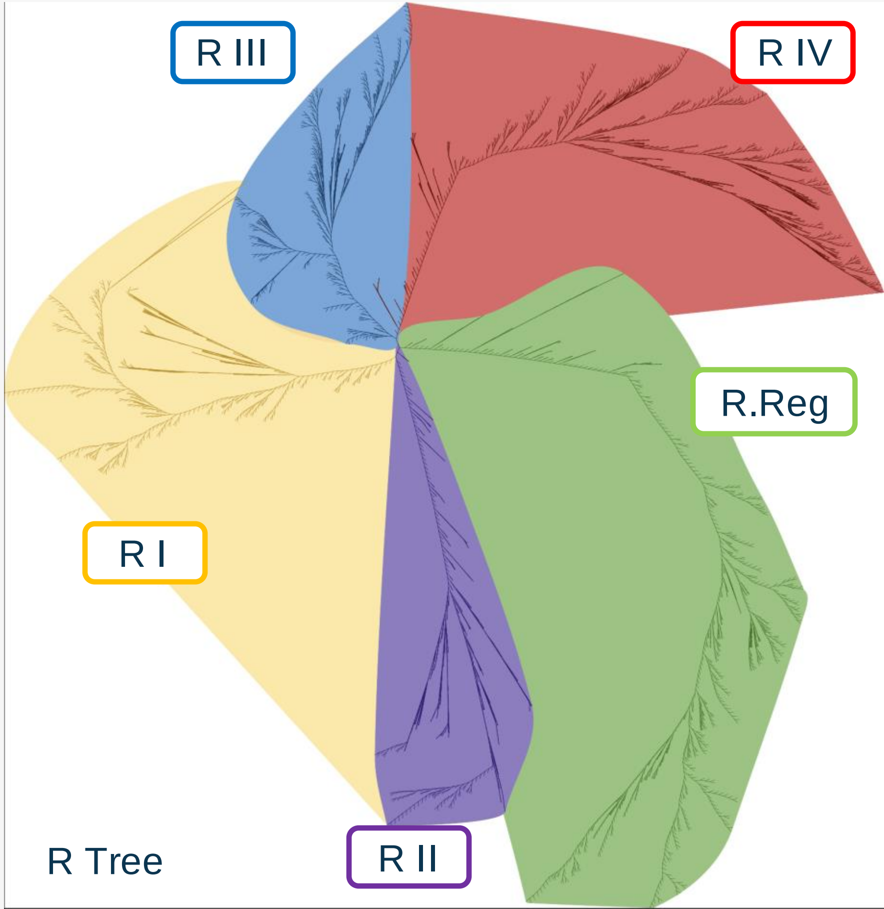


```
HoSa_Cellub/1-95   ------------------------------------NRR----LQQTQNQV-DE--VVD-IMR-VNVDKVLERDQKLSELDDRADALQAGASQFETSAAKLK-R-KY-WW-KNCKMWAIGITVLVIFIIIII-V--
XeTr_Amisyn/1-151  SPFD-LQF--DNS----QW---------VASS---GE-KF---LQR-RKEFINPSKLFGSILH-DSVQKLNERGERLGRAEESTE-MMNSAQQFADTAHKLAMKHKC-----------------------------
ApMe_Vamp4/1-120   --FN----RPST------------------SV---ADGK----MDSVRLQI-QE--VTE-VMR-ENVQKVIERGEKLEDLQEVSDRLTMAGNEFRAAARKAQ-Q-RA-WL-QNFK-RIILTIVTVVVIIGHV-ISA
RaNo_Syb1/1-113    ----P----PPN-T-----------------T---SNRR----LQQTQAQV-EE--VVD-IMR-VNVDKVLERDQKLSELDDRADALQAGASVFESSAAKLK-R-KY-WW-KNCKMMIMLGAICAIIVVVIV-S--
TeNi_Syb2/1-100    ---Q-----PPN-L-----------------T---SNRR----LQQTQAQV-DE--VVD-IMR-VNVDKVLERDQKLSELDDRADALQAGASQFETSAAKLK-N-KY-WW-KNAKMMIILGVICVIVLIVII----
ScJa_Tom/1-44      -------------------------------------------------------------------DRAIERGERLQQLDLQTQEMADQAKGFGRSAAILAAKYK-RWG---P----F-----------------
HoSa_Sec22/1-211   --F--IEF--DTFIQ-------------KTKKLYI-DSRARRNLGSINTELQD---VQ-RIMV-ANIEEVLQRGEALSALDSKANNLSSLSKKYRQDAKYLNMRS--T-YA---K----LAAV--FFIMLIVYV--
MuMu_Myob/1-92     -------------------------------M---AGKE----LKQCQQQA-DE--VTE-IML-NNFDKVLERHGKLAELEQRSDQLLDMSSAFSKTTKTLA-Q-QK-RW-ENIRCRVYLGLAVAGLLIILI-VLL

Conservation       ----------00------------------------103----70015359-61--70--990-02173933+574+428345742923382392428001--6-0--------1----1-----111--------

Consensus          ASFEPIELCPDTYIQLYKYSQEDPREKSKTKKLYVADSRSRRNLQKVQAELQDETKVVGSIMRHKNIEKVLERGEKLDELVDKTDNLSASSKKFRKTARKLNMRSKMKWWAKNMKMMICLGVVCVIILIIIIYVLV
```

Identification of sub-groups

- Full sequences alignment to build the trees
- Average distance, visualisation with ITOL
- Labelling of sub-groups
- MSA for each of the sub-groups





Sequences distribution across taxa

	V	M	F
Qa	471	3081	1481
Qb	316	1042	972
Qc	335	947	1327
R	448	2294	1416
SN	60	786	283

03

HMM

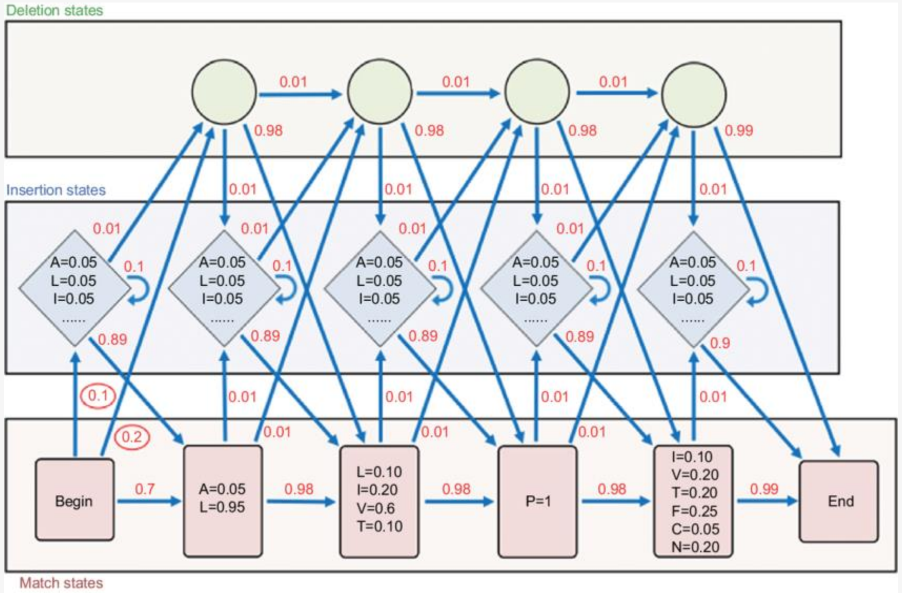
Hidden Markov Models, Profiles and search

What are HMM profiles?

- Based on multiple sequences alignements
- Probabilistic models used to represent a family of sequences

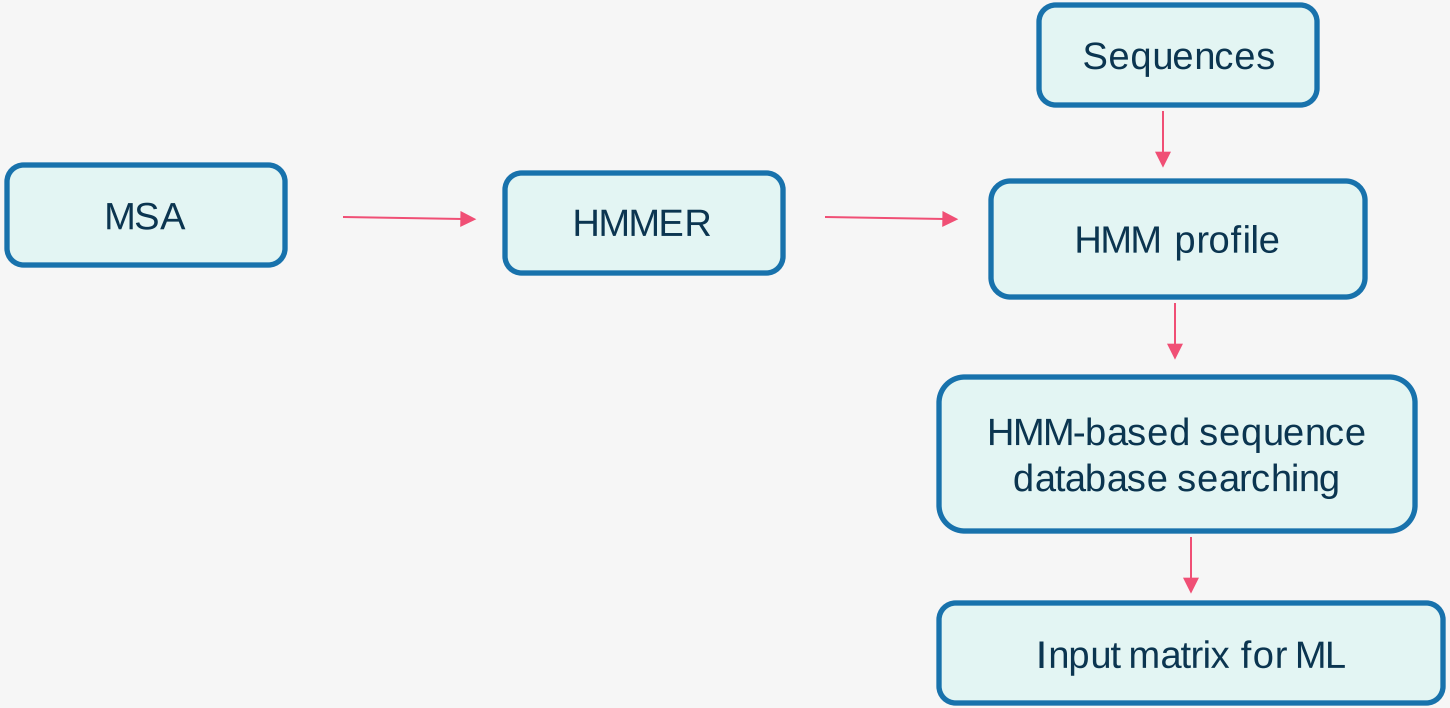
- They capture conserved and variable regions, as well as insertions and deletions

- Allow to determine how likely it is for a sequence to belong a specific group

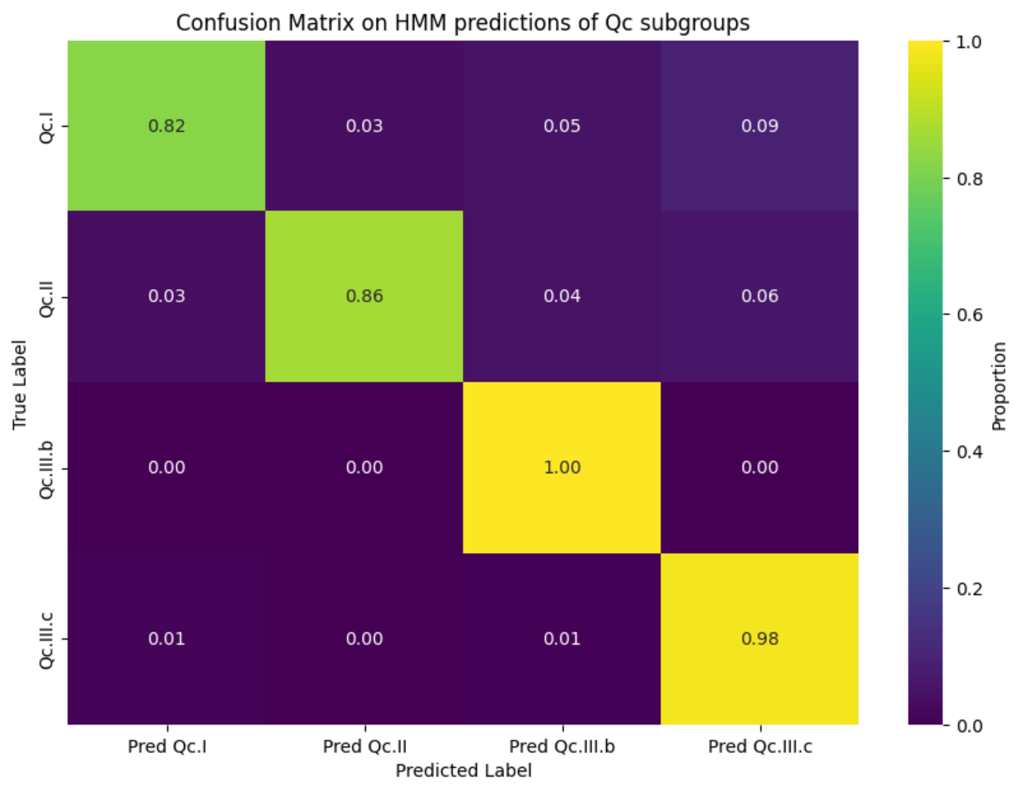


Arthur Gruber ResearchGate

Building and using the HMM profiles



HMM profiles performance



Accuracy = 0.92

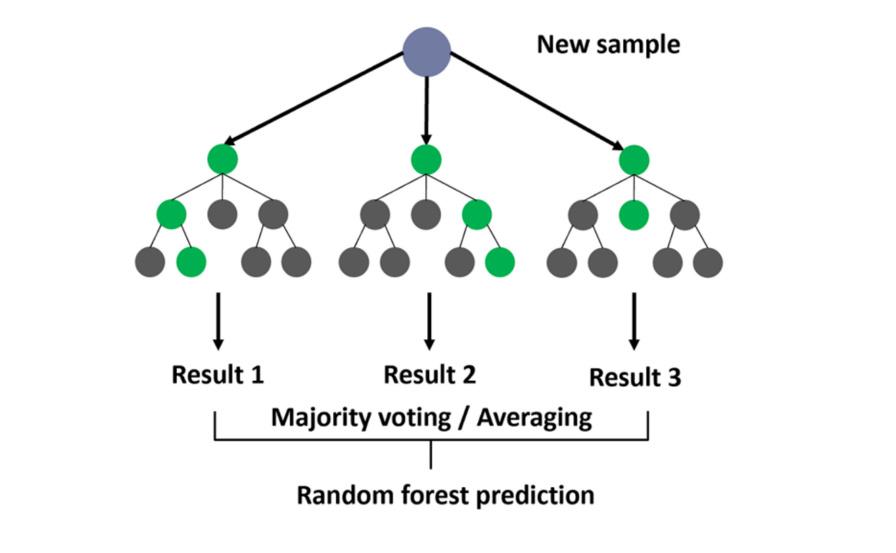
04

Machine learning

Random forests

Machine learning – Random forests

- Commonly used model for classification
- Ensemble learning technique
- Creation of multiple decision trees
- Aggregation of results
- High accuracy and reduced overfitting



Dr Roi Yeoshua, Medium

Machine learning – input

Sequence	Score/length Qa	EvalQa	Score/length Qb	Eval Qb	Score/length Qc	Eval Qc	Score/length R	Eval R	Label
DiOr_Syx1a	0.86254	2.2e-79	NaN	NaN	0.076	6.6e-06	NaN	NaN	Qa
TrVi_Bos1	NaN	NaN	0.47964	8.2e-40	NaN	NaN	NaN	NaN	Qb
ChMy_Syx6	0.38248	9.4e-15	0.13983	0.0011	1.1024	6.2e-41	NaN	NaN	Qc
…	…	…	…	…	…	…	…	…	…

- No hits from the HMM profile do not return anything
- Need to set an appropriate value for the missing data for the model to work
- Different replacement values were tested, from 0.001 to 0.95
- Specific to the score
- Specific to the E-value

Machine learning – input

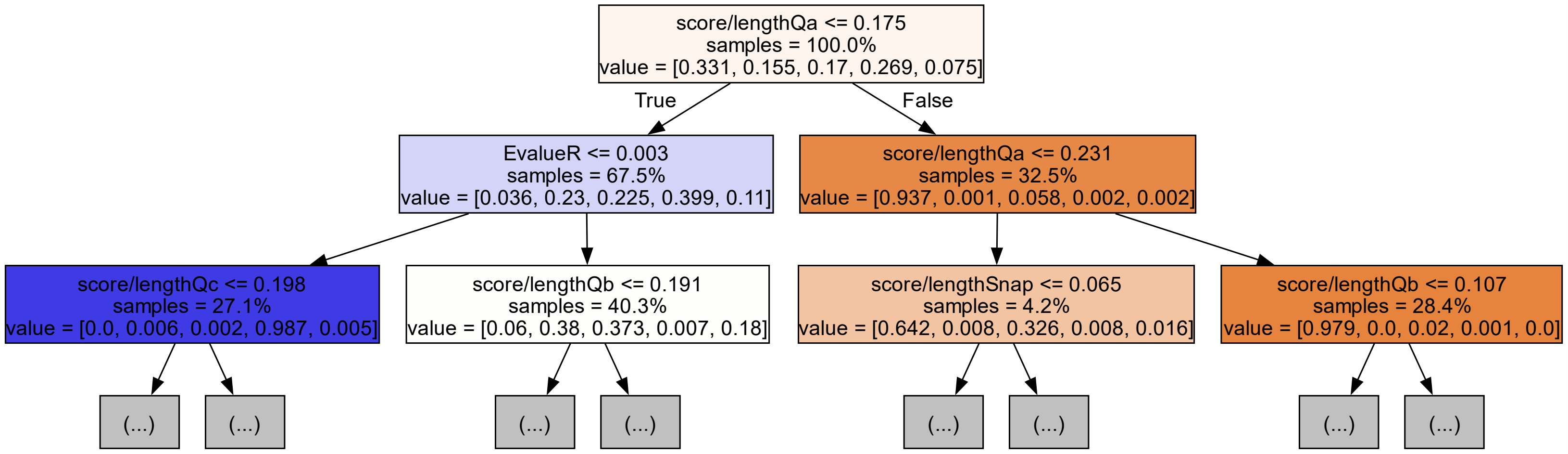
Sequence	Score/length Qa	EvalQa	Score/length Qb	Eval Qb	Score/length Qc	Eval Qc	Score/length R	Eval R	Label
DiOr_Syx1a	0.86254	2.2e-79	0.01	0.9	0.076	6.6e-06	0.01	0.9	Qa
TrVi_Bos1	0.01	0.9	0.47964	8.2e-40	0.01	0.9	0.01	0.9	Qb
ChMy_Syx6	0.38248	9.4e-15	0.13983	0.0011	1.1024	6.2e-41	0.01	0.9	Qc
…	…	…	…	…	…	…	…	…	…

- No hits from the HMM profile do not return anything
- Need to set an appropriate value for the missing data for the model to work
- Best working values :

  Low replacement value for the score : 0.01

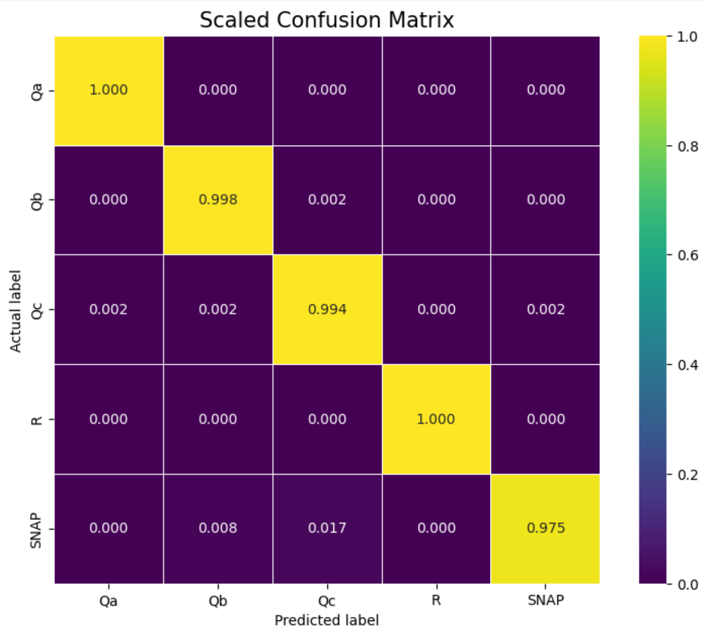
  High replacement value for the E-values : 0.9

Machine Learning – Example tree


score/lengthQa <= 0.175
samples = 100.0%
value = [0.331, 0.155, 0.17, 0.269, 0.075]
True
False
EvalueR <= 0.003
samples = 67.5%
value = [0.036, 0.23, 0.225, 0.399, 0.11]
score/lengthQa <= 0.231
samples = 32.5%
value = [0.937, 0.001, 0.058, 0.002, 0.002]
score/lengthQc <= 0.198
samples = 27.1%
value = [0.0, 0.006, 0.002, 0.987, 0.005]
score/lengthQb <= 0.191
samples = 40.3%
value = [0.06, 0.38, 0.373, 0.007, 0.18]
score/lengthSnap <= 0.065
samples = 4.2%
value = [0.642, 0.008, 0.326, 0.008, 0.016]
score/lengthQb <= 0.107
samples = 28.4%
value = [0.979, 0.0, 0.02, 0.001, 0.0]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Machine learning – Main group results

- Confusion matrix



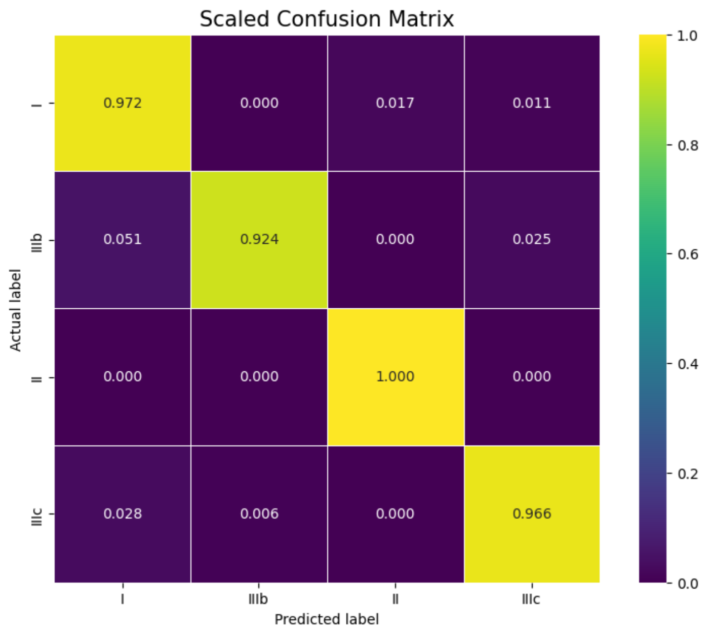
Actual label \ Predicted label	Qa	Qb	Qc	R	SNAP
Qa	1.000	0.000	0.000	0.000	0.000
Qb	0.000	0.998	0.002	0.000	0.000
Qc	0.002	0.002	0.994	0.000	0.002
R	0.000	0.000	0.000	1.000	0.000
SNAP	0.000	0.008	0.017	0.000	0.975

- Evaluation metrics:
- Accuracy: Correctly classified instances out of the total instances
- Precision: Ratio of TPs to the sum of TPs and FPs
- Recall: Ratio of TPs to the sum of TPs and FNs
- F1 score: Harmonic mean of precision and recall

- Accuracy: 0.9996
- Precision: 0.9996
- Recall: 0.9996
- F1: 0.9996

Machine learning – Qc subgroups results

- Confusion matrix

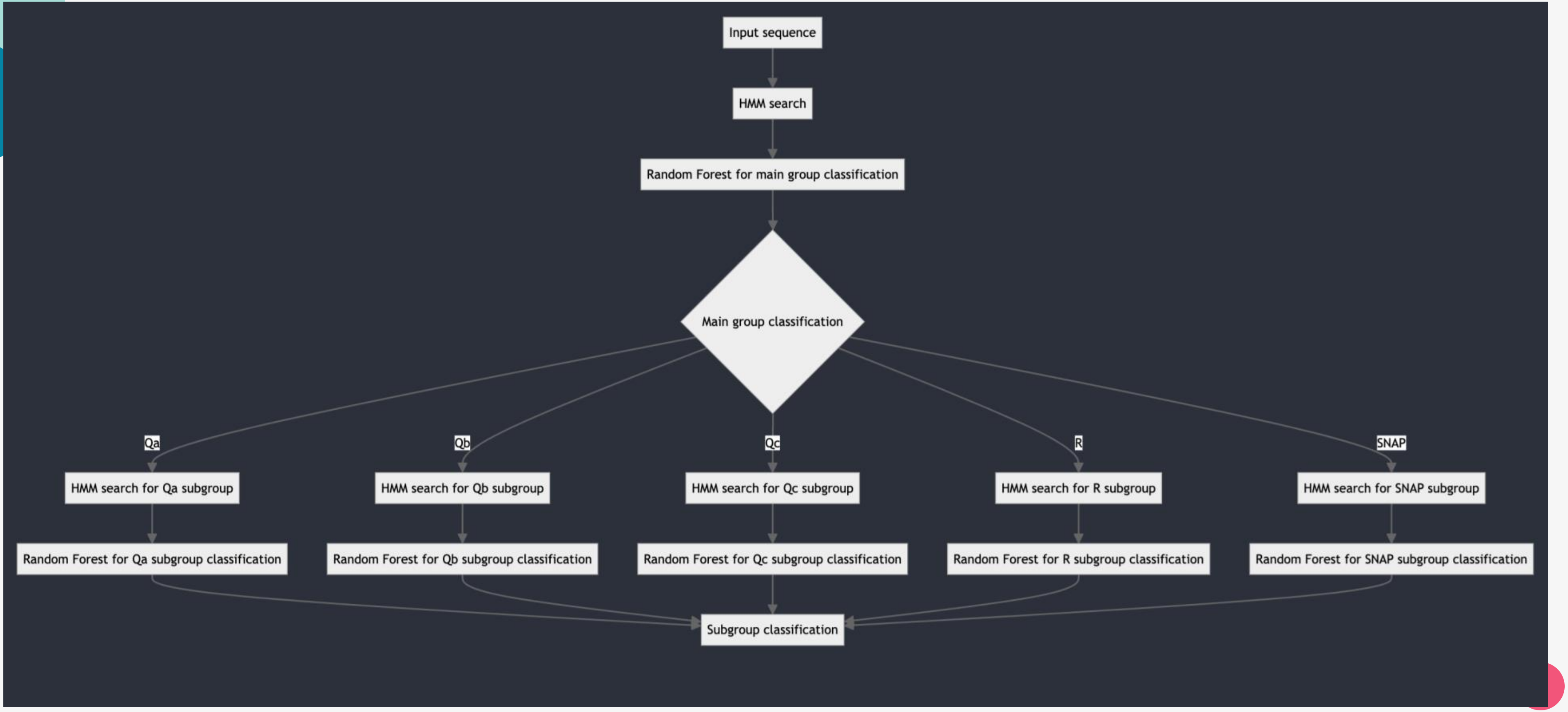


Scaled Confusion Matrix

Actual label \ Predicted label	I	IIIb	II	IIIc
I	0.972	0.000	0.017	0.011
IIIb	0.051	0.924	0.000	0.025
II	0.000	0.000	1.000	0.000
IIIc	0.028	0.006	0.000	0.966

- Evaluation metrics:
- Accuracy: Correctly classified instances out of the total instances
- Precision: Ratio of TPs to the sum of TPs and FPs
- Recall: Ratio of TPs to the sum of TPs and FNs
- F1 score: Harmonic mean of precision and recall

- Accuracy: 0.967
- Precision: 0.967
- Recall: 0.967
- F1: 0.967



Input sequence

HMM search

Random Forest for main group classification

Main group classification

Qa

Qb

Qc

R

SNAP

HMM search for Qa subgroup

HMM search for Qb subgroup

HMM search for Qc subgroup

HMM search for R subgroup

HMM search for SNAP subgroup

Random Forest for Qa subgroup classification

Random Forest for Qb subgroup classification

Random Forest for Qc subgroup classification

Random Forest for R subgroup classification

Random Forest for SNAP subgroup classification

Subgroup classification

05

What’s next ?

Future things to implement

Challenges

- Using IqTree for better classification
- Using a tree to classify the main groups
- Investigating other ML models to upgrade the HMMs performance

Feedback

- It was really interesting to work practically
- Projects had a real biological meaning and were based on real data
- More time for the presentation would be appreciated

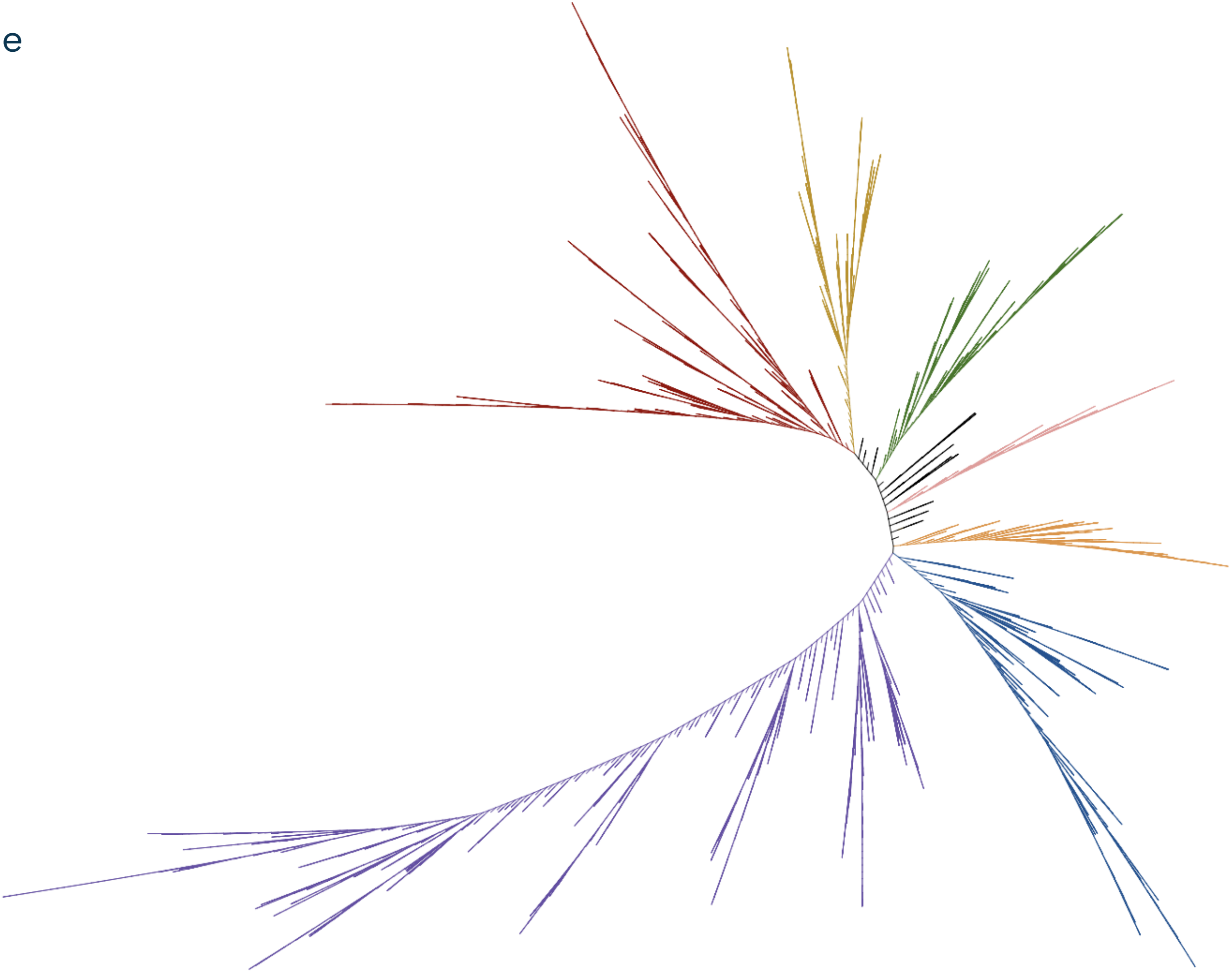
Thanks!

Do you have any questions?

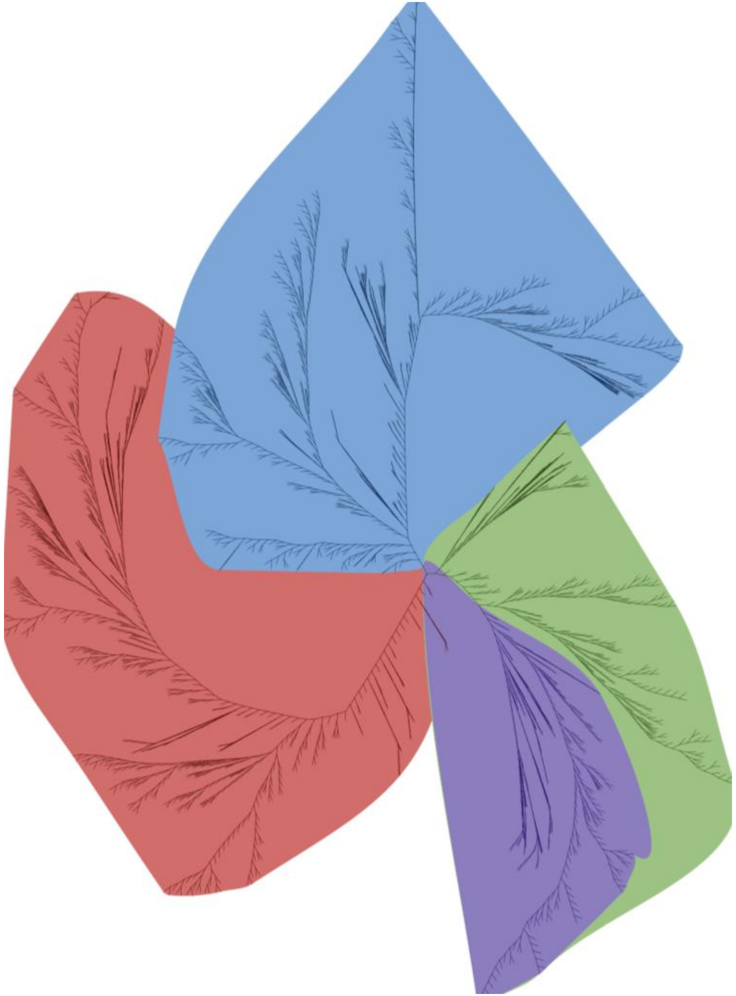
Gabriel Chiche, Marius Audenis, Leana Ortolani
Supervisor Carlos Pulido Quetglas

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Qa Tree



Qc Tree



SNAP Tree