I Meeting of Systematics, Biogeography and Evolution (SIBE)

@ManoloFPerez

sites.google.com/site/manolofperez

Using deep learning for unifying genomic data and traits in species delimitation

MF Perez; I Sanmartín; BC Faircloth; LAC Bertollo; MB Cioffi

REAL JARDÍN BOTÁNICO

Louisiana State University

Introduction

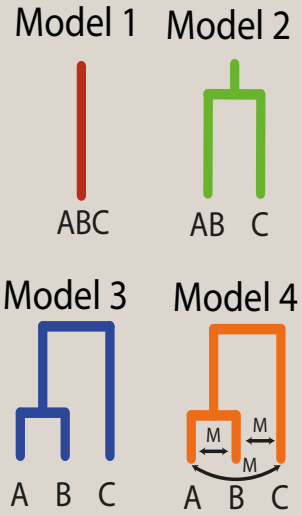
The presence of distinct **species concepts** leads to different strategies to **identify species boundaries** (de Queiroz 2007). It is important to adopt a **multidisciplinary approach**, by **integrating** different **sources of evidence** (Carstens et al. 2013).

Most approaches consist in analyzing **genetic and phenotypical/geographical** information **separately,** followed by **visual/qualitative comparison**. Methods that actually **integrate** different data are **limited** to up to a **few hundreds of loci** and **simple models** of evolution (Solís-Lemus et al. 2015).

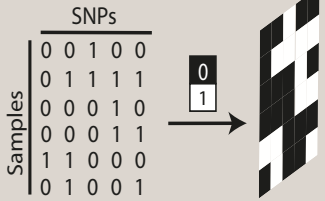
We present a method based on **simulated data and machine learning**, that **combines** both **genomic and trait** information in a unified framework.

Methods

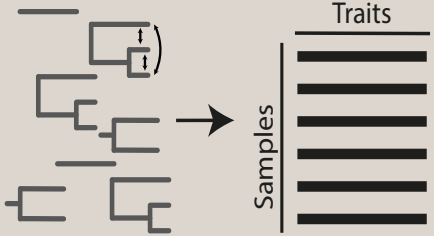
Simulate SNPs & trees with ms (Hudson 2002)



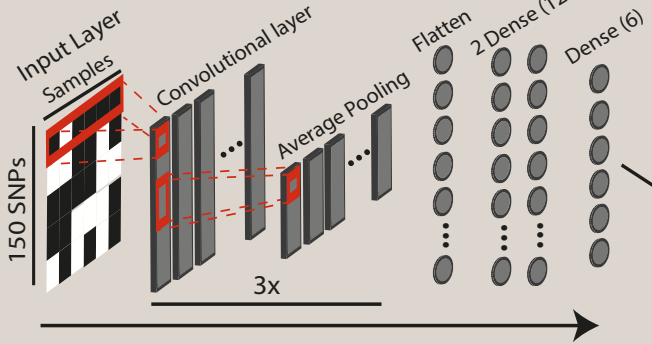
Transform SNPs to image



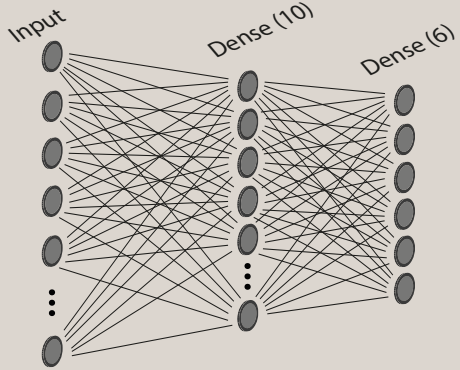
Simulate traits (discrete, BM and OU) from trees



Train CNN with SNP data



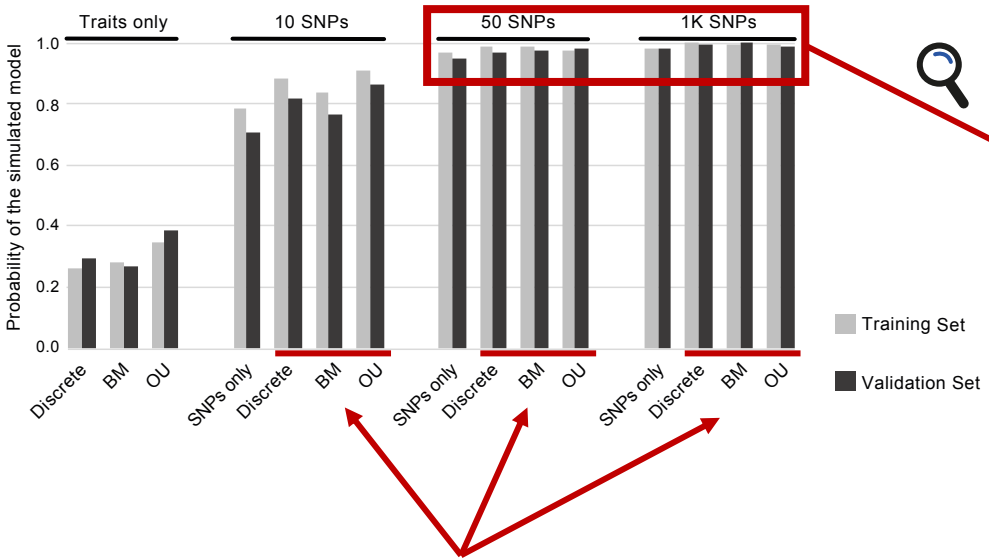
Train MLP with trait data



Combine the Dense layers from both networks

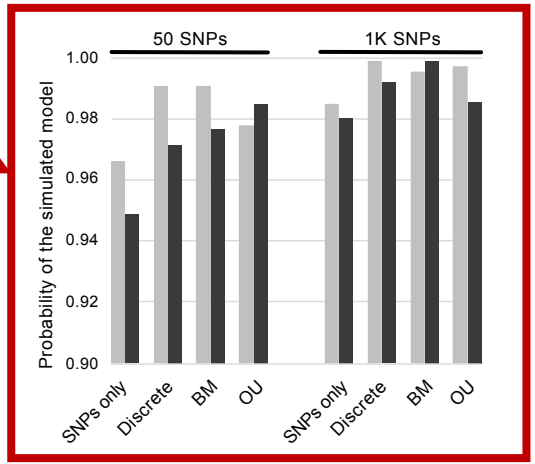
Results

Probabilities are low when using only traits.



Increased nº of SNPs also led to a higher probability of recovering the right model.

- little improvement with > 50 SNPs.

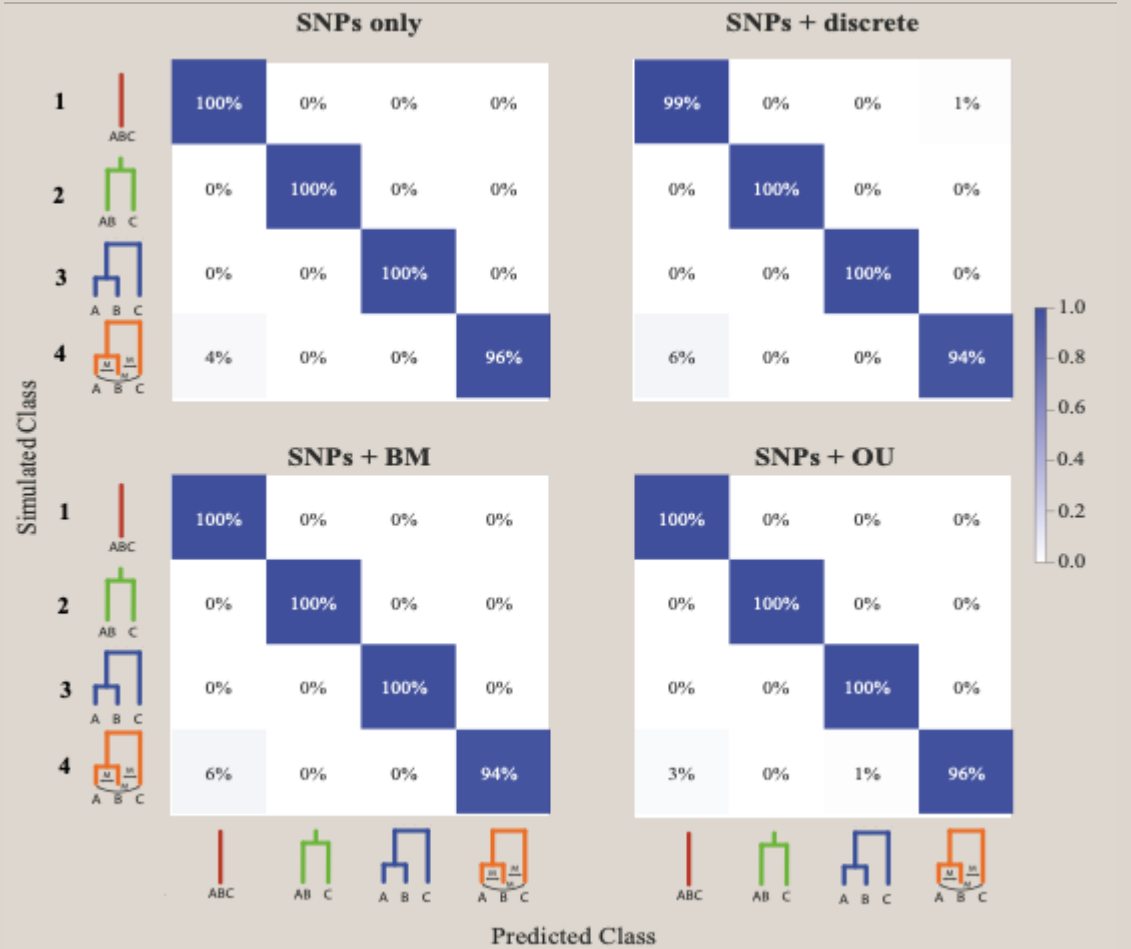


Using both genomic and trait data recovered slightly better results than using only SNPs.

Conclusions

The **accuracy** of our approach was **very high** (confusion matrix with the **test set**). **Lowest accuracy** between model 4 (**migration**) with model 1 (**one species**).

Incorporating traits resulted in **similar accuracy** to using only SNPs.

Traits incorporate information **complementary** to genomic data.

The method is **flexible**, allowing **complex scenarios** and the use of both types of data.