Eukaryote Genome annotation at the Plant Systems Biology department

Stephane Rombauts[1] , Lieven Sterck[1] , Francis Dierick[1] , Steven Robbens[1] , Jan Wuyts[1] , Thomas Schiex[3] , Pierre Rouzé[2] , Yves Van de Peer[1]

1 Bioinformatics & Evolutionary Biology, Plant Systems Biology, UGent-VIB Technologiepark 927, Gent 9052 Belgium
2 INRA-associated to Bioinformatics & Evolutionary Genomics, Plant Systems Biology, UGent-VIB Technologiepark 927 Gent 9052 Belgium
3 INRA, Département de Biométrie et Intelligence Artificielle Chemin de Borde Rouge, BP 27 Castanet-Tolosan 31326 Cedex, France

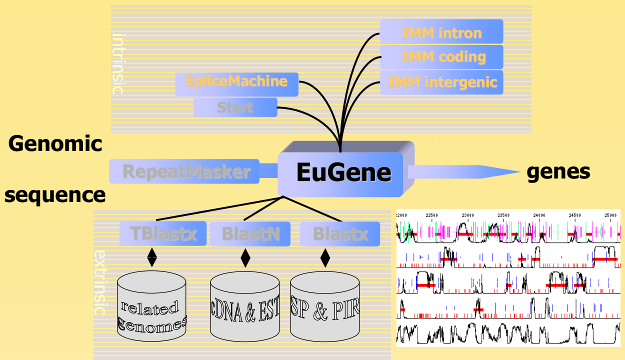
[1] E-mail: {strom,yvpee}@psb.ugent.be URL: http://bioinformatics.psb.ugent.be/

Introduction

Genome annotation is one of the main research topics of our group, and we have been able to demonstrate the strength of our genome annotation platform in collaborative efforts to predict genes on a wide variety of genomes. Our involvement in this broad diversity of eukaryotic genomes will give us insights in the genome structures and their evolution, and enable us to perform complex comparative analyses to better understand the biology within those genomes.

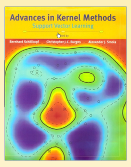
Technology

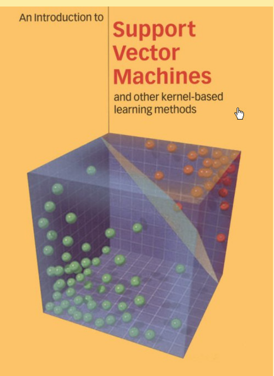
EuGène



Each site is represented as a feature vector that is a point in an n-dimensional space.

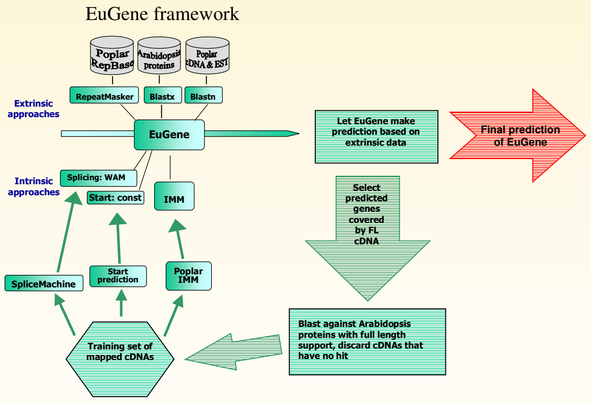
We employ a large margin hyperplane induction method (Suport Vector Machine) to build a decision function in this space.




An Introduction to Support Vector Machines and other kernel-based learning methods

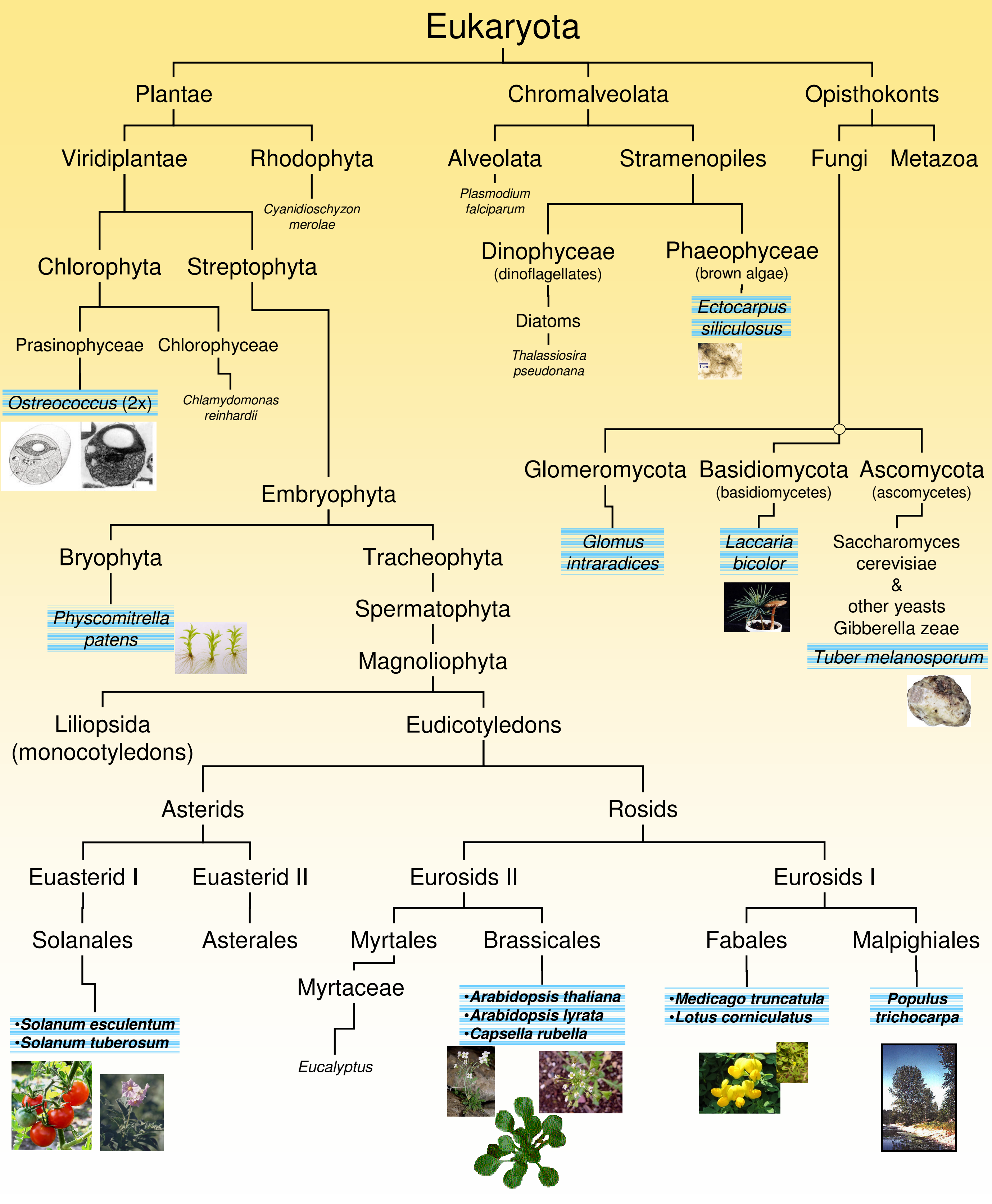
Constructing datasets

case of Poplar


EuGene framework
Poplar RepBase
Arabidopsis proteins
Poplar cDNA & EST
RepeatMasker
Blastx
Blastn
Extrinsic approaches
EuGene
Let EuGene make prediction based on extrinsic data
Final prediction of EuGene
Intrinsic approaches
Splicing: WAM
Start: const
IMM
Select predicted genes covered by FL cDNA
SpliceMachine
Start prediction
Poplar IMM
Training set of mapped cDNAs
Blast against Arabidopsis proteins with full length support, discard cDNAs that have no hit

Advantages

- Speed: within a few days we could construct a new dataset to train for the poplar genome
- We can use 2 ESTs that have a to small or no overlap to be considered as a full length mRNA



Conclusion & Perspectives

Strengths of EuGene

- exploits probabilistic models like Markov models for discriminating coding from non coding sequences
- integrates information from several signal (splice site, traduction start...) prediction software, propriety or 3rd party software
- Exploits the wealth of existing sequences (EST, mRNA, 5'/3' EST couples, proteins, genomic homologous sequences)...
- Based on all the available information, EuGene will output a prediction of maximal score i.e., maximally consistent with the provided information.
- integrates each source of information through small independent software components, called "plugins".
- There exists currently more than 25 plugins, but if needed EuGene's users have the ability to write new ones

Weaknesses of EuGene

- No ability to predict alternative splicing (yet)
- Complex to train, using genetic algorithms
  - Each individual component
  - Each component in the frame of EuGene
  - Evaluate the weights and penalties for the extrinsic data

References

1: *Schiex T, Moisan A, and Rouzé P. (2001)* EuGène: An Eucaryotic Gene Finder that combines several sources of evidence. Computational Biology, Eds. O. Gascuel and M-F. Sagot, LNCS 2066, pp. 111-125, 2001

This work is supported by the European Commision (QLRI-CT-2001-00006)

2: Tuskan et al. The genome of western black cottonwood, *Populus trichocarpa* (Torr. & Gray ex Brayshaw) (submitted)
3: Derelle et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils unique genome heterogeneity (submitted)