The Big Data Challenge in Bioinformatics

Applications in Molecular Plant Breeding

Dr. Elizabeth Tapia

CIFASIS - Centro Int. Franco Argentino de Ciencias de la Información y de Sistemas

CONICET

Jornadas de Definición Estratégica en Big Data , 26-30 Agosto 2013, Fundación Sadosky, Buenos Aires, Argentina

Agenda

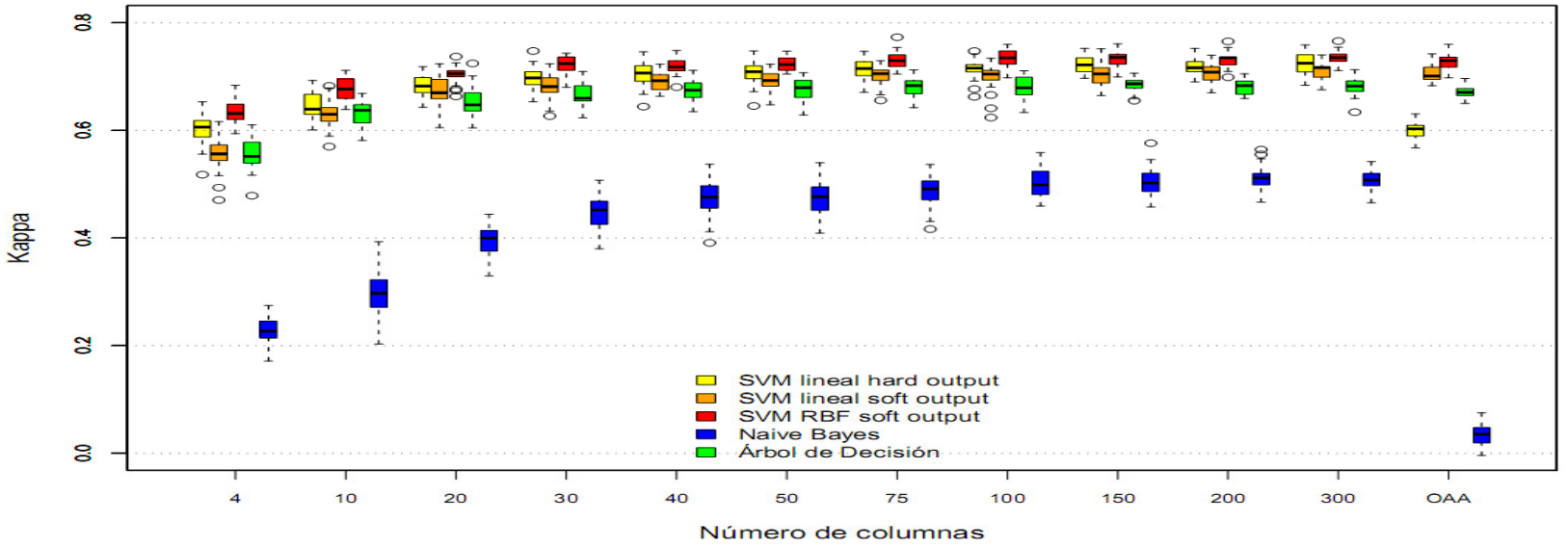
* The Big Data Opportunity
  * Molecular Plant Breeding
  * Annotation of non-model organisms
* Big Data Services
* Conclusions

Molecular Plant Breeding
Low Density Marker Data (from 1998 to 2010)

data	#Ins. [a]	#Atrib.	#Clase	Distribución de Clases
Liu data	197	188	10	71(CCB), 13 (CCB-B14), 11 (CCB-B73) 9 (CCB-SSS), 28 (NCB), 29 (Tuxpeño) 8 (CCB-103), 17 (South), 8 (TuxpeñoCaribe)
Xia data	73	164	8	22 (Pop21), 17(Pop43), 7 (Pop62), 5 (Pool24) 5 (Pop28), 5 (pool26), 5 (Pop26), 7 (Pop24)
Morales data	26	42	4	4(GH1), 8(GH2),6 (GH3), 8(GH4)

Adapted from Ornella L. and Tapia E. **Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data.** Computers and Electronics in Agriculture, 74 (2010), pp. 250–257

Heterotic Prediction in Maize
Low Density Molecular Marker Data

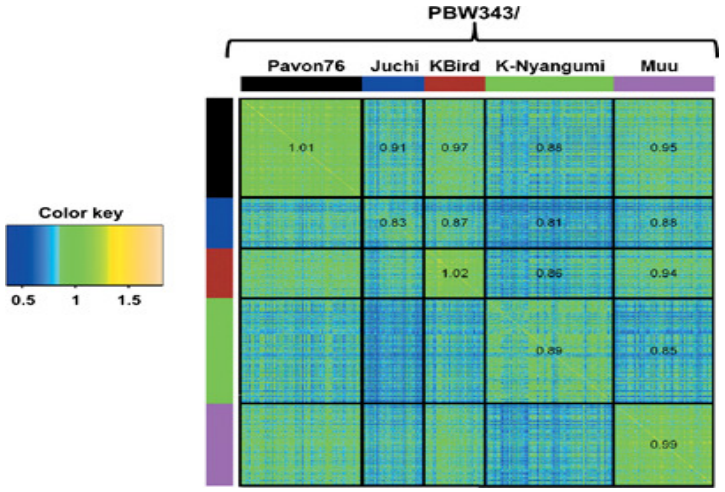


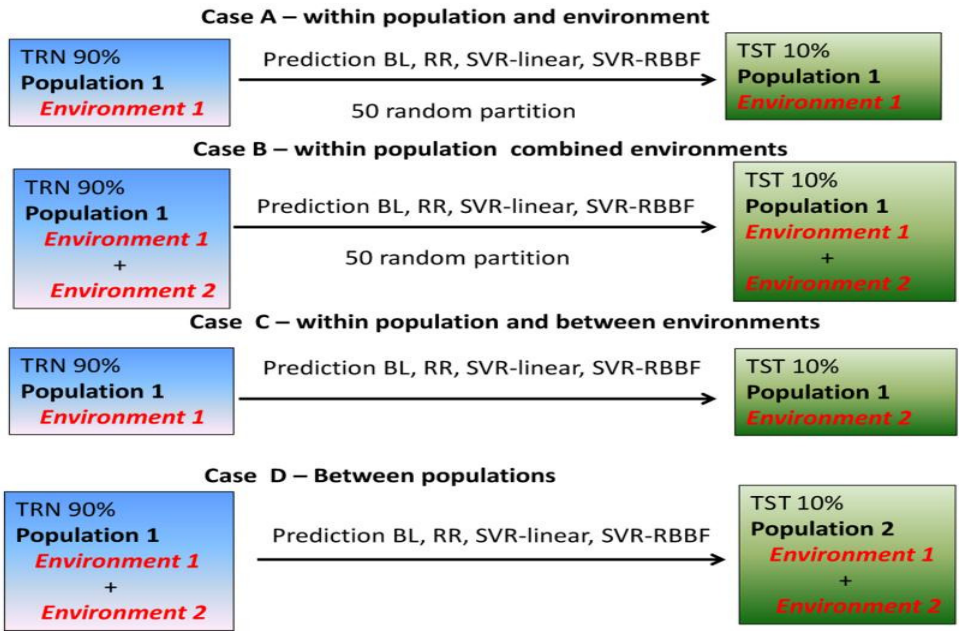
Liu Data. Supervised Prediction of Heterotic Groups with Binary Reductions.

Kappa index for 30x10 Fold CV and Majority Voting.

Wheat Rust Resistance Median Marker Data – From 2011

Heat map of the genomic relationship matrix G of five wheat populations characterized with 1400 DART markers





Ornella L, Singh S, Pérez P, Burgueño J, Singh R, Tapia E *et al.* (2012). Genomic prediction of genetic values for resistance to wheat rusts. *The Plant Genome* 5: 136–148.

Predicting Resistance to Wheat Rusts

Table 3. Pair-wise correlations between observed and predicted stem rust values of two models, Bayesian LASSO and the GBLUP, trained in one population and evaluated in the other population for five populations (adapted from Ornella *et al.*, 2012)

Previous table | Figures and tables index | Next table

	Training[a]					
	PBW343/Juchi	***PBW343/Kingbird***	***PBW343/K-Nyangumi***	***PBW343/Muu***	***PBW343/Pavon76***	
Testing						
PBW343/Juchi	—	0.48	0.14	0.28	0.31	Bayes LASSO
PBW343/Kingbird	0.53	—	0.29	0.25	0.54	
PBW343/K-Nyangumi	0.14	0.30	—	0.28	0.28	
PBW343/Muu	0.18	0.30	0.33	—	0.29	
PBW343/Pavon76	0.37	0.51	0.22	0.33	—	
	GBLUP					

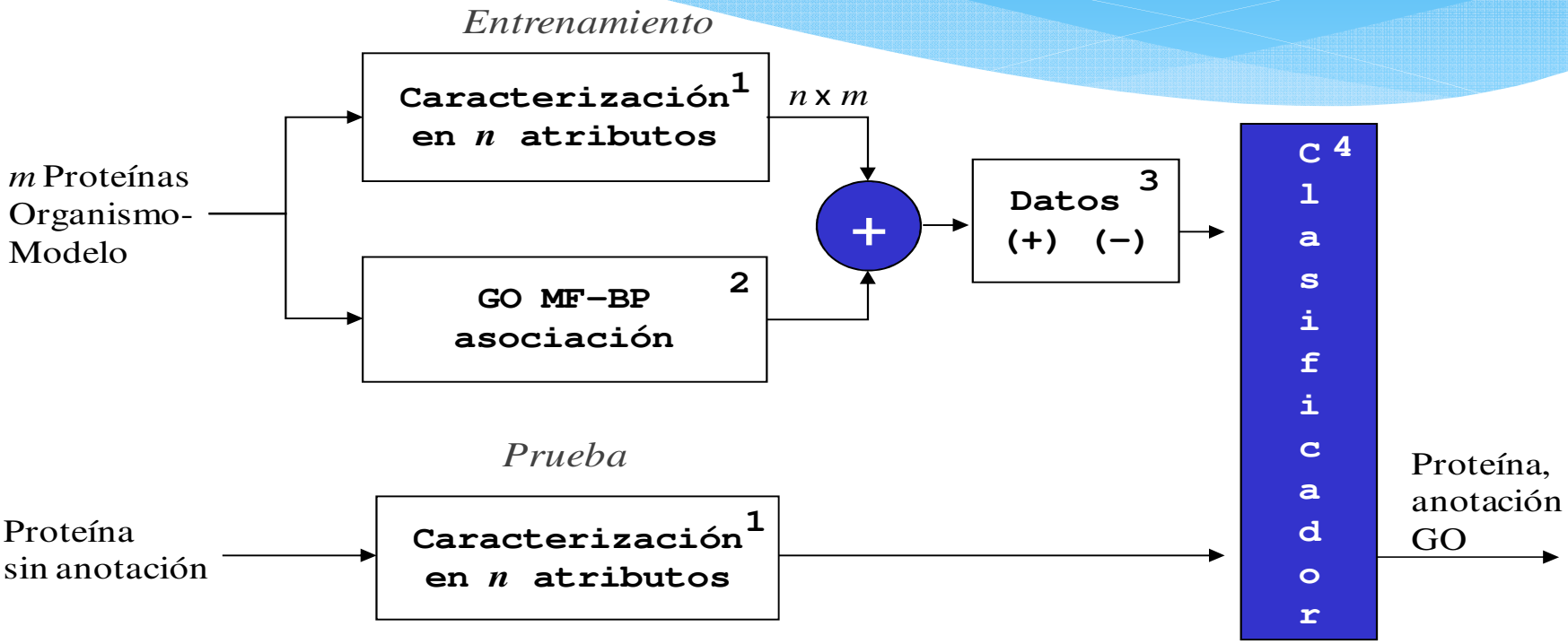
There are five related populations: PBW343/Juchi, PBW343/Kingbird, PBW343/K-Nyangumi, PBW343/Muu and PBW343/Pavon76.

[a] The triangle on the upper-right shows the prediction ability (correlation) of Bayes LASSO, with the rows indicating the training population (that is, PBW343/Juchi) and the columns the testing population (that is, PBW343/Kingbird, 0.48); the triangle on the lower-left gives the prediction ability of GBLUP, with the columns indicating the training population (that is, PBW343/Juchi) and the rows the testing population (that is, PBW343/Kingbird, 0.53).

Previous table | Figures and tables index | Next table

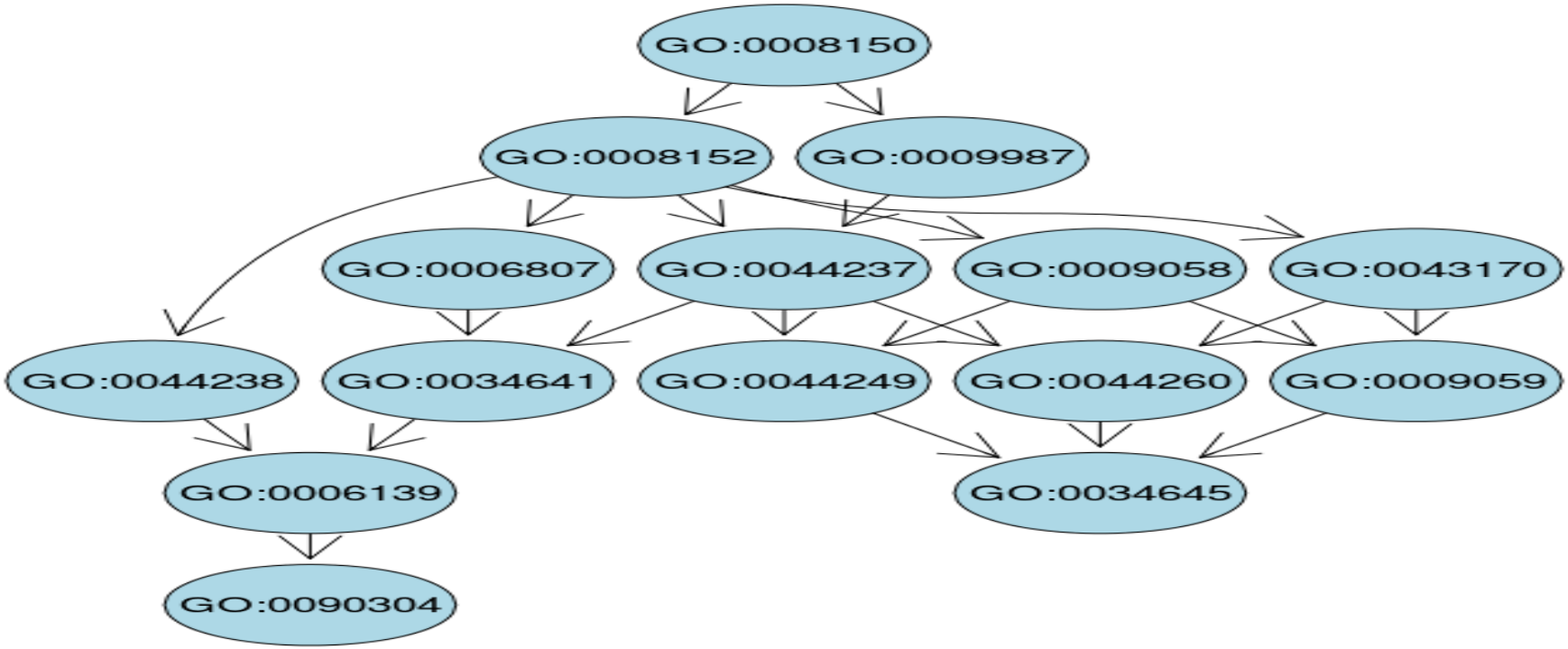
Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L *et al (2013)*. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* advance online publication 10 April 2013

Non Model Plants
Electronic Annotation of Protein Sequences

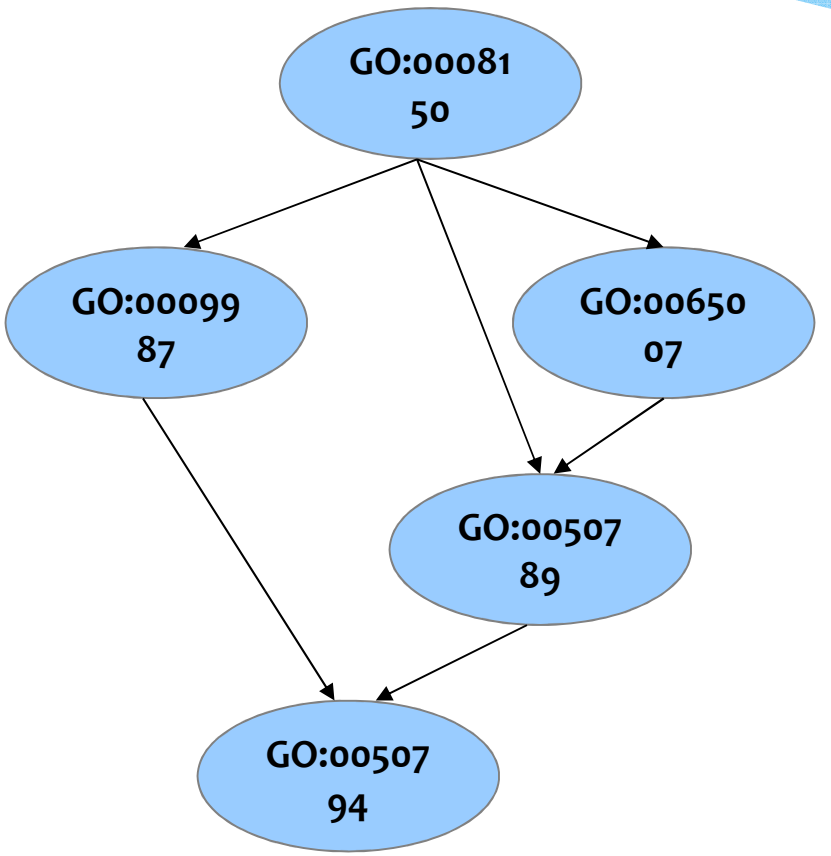


Fernández P et al. (2010) Sunflower Functional Genome Database, a curated unigene database to support functional diversity studies en sunflower, ISCB Latin America, Uruguay.

Computational Model
GO BP Annotations



Predicting GO BP Terms



at2g37740:
Description
zinc finger (C2H2 type) family protein
Family
C2H2
3D structure (top 5)
1NJQ
From: Plant Transcription Factor Database

Valentini, G. and Cesa-Bianchi, N. (2008) HCGene: a software tool to support the hierarchical classification of genes Bioinformatics. 24(5):729-31.

Big Data Services Main Demands

* Confidentiality
* Integrity
* Availability

DNAnexus

At a Glance

What they wanted to do
- Build a secure mirrored version of the NCBI SRA repository
- Develop a web-based user-friendly interface
- Allow for scalability and future data growth

What they did
- Used Google Cloud Storage to host 350 terabytes of DNA sequencing data
- Focused resources on developing the user interface, since Google handles security, scalability and other issues

What they accomplished
- Mirrored a comprehensive cloud-based DNA archive using minimal internal resources
- Created a platform that has received praise from researchers on how easy it is to use
- Help expedite research by allowing scientists to sort through and pinpoint specific DNA sequencing data more easily

Conclusions

* Big Data in Bioinformatics
  * Boosted by NGS and related technologies
* Big Data requires multidisciplinar collaboration
  * To design secure and friendly services for sequence processing and analysis
* Big Data services will boost Molecular Biology research