

Predictive clustering relates gene annotations to phenotype properties extracted from images

Dragi Kocev, Bernard Ženko, Petra Paul, Coenraad Kuijl, Jacques Neefjes, and Sašo Džeroski

Jožef Stefan Institute, Dept. of Knowledge Technologies, Ljubljana, Slovenia {dragi.kocev, bernard.zenko, saso.dzeroski}@ijs.si
Division of Cell Biology and Centre for Biomedical Genetics, NKI, Netherlands {p.paul, c.kuijl, j.neefjes}@nki.nl

Introduction

- Grouping genes with similar phenotypes upon siRNA mediated downregulation
- Phenotypes are described by features extracted from images using some free general-purpose or custom-made software (e.g., CellProfiler)
- siRNA screen designed to study MHC Class II antigen presentation
 - A major regulatory process in the immune system
 - Controls most aspects of the adaptive immune response
 - Strongly linked to almost all autoimmune diseases

Gene id	GO000166	GO000267	...	GO001883	Feature_1	Feature_2	...	Feature_t
9625	1	0		1	2.991	-2.296		0.68
21	1	1		1	-0.013	-1.048		0.767
84519	0	0		0	-0.843	0.155		1.396
...								
23621	0	1	...	0	-0.531	0.843	...	2.304

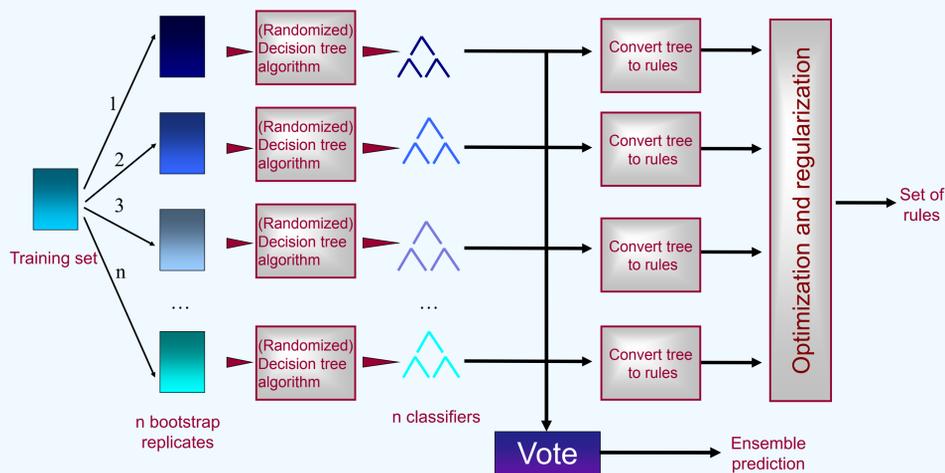
Methodology

Predictive clustering trees

- Induced by standard TDIDT algorithm
- Able to make a prediction for given structured output
- Heuristic score: minimization of intra-cluster variance
- Definition of a distance and prototype function for a given output
- Medoid is taken as a prototype in each leaf
- Cosine similarity as distance measure

$$\text{CosineSim}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Ensembles of PCTs and Rule ensembles

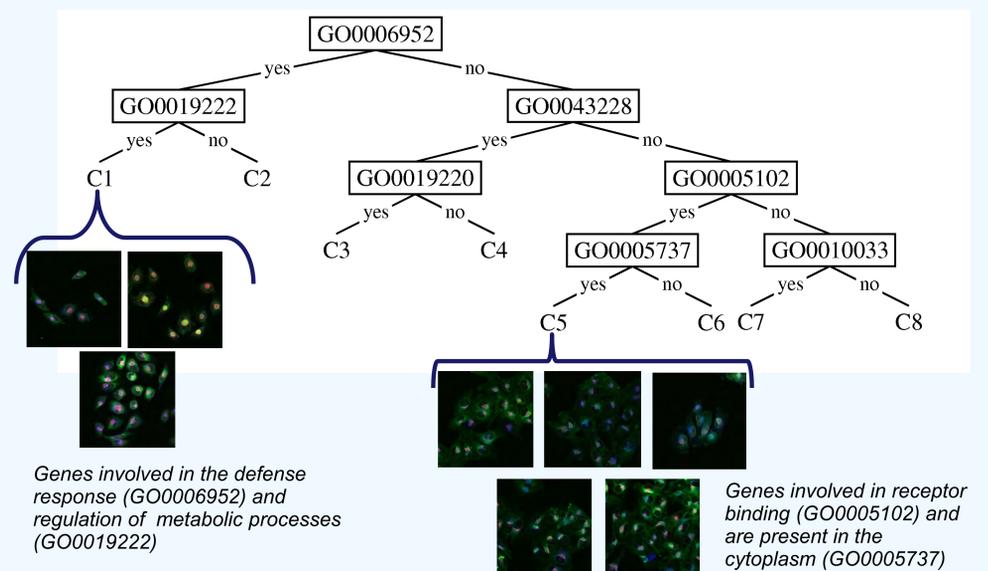


- Ensembles are able to lift the predictive performance of a single predictive clustering tree
- Random forests are efficient to learn
- Ensembles are not interpretable
 - conversion to fitted rule ensembles
- Each tree from the ensembles is converted to a set of rules
- Using some optimization techniques, select the rule set with best performance
- Easy to interpret by domain experts

Data description

- siRNA screen performed on 269 genes, from which 20 were hypothetical
- Each gene is described by:
 - its annotation with terms from the Gene Ontology
 - resulting phenotypes (images from confocal microscopy)
- Only the GO terms that are used to annotate at least 1 gene from the ones analysed (in total - 334)
- CellProfiler for extracting features from the images: in total 700 features, from which 13 most relevant to the study are used

Results



	C1	C2	C3	C4	C5	C6	C7	C8
Mean Cells Intensity StdIntensity GFP	-1.66	1.03	-0.77	-0.93	-1.31	-1.66	0.48	-1.6
Mean Cells Texture AngularSecondMoment GFP 50	-1.41	-1.48	0.54	0.22	0.88	0.48	-0.75	-0.45
Mean Cytoplasm Intensity IntegratedIntensityE GFP	-1.24	0.25	-0.23	0.54	-1	0.2	0.07	2.34
Mean Cytoplasm Texture InfoMeas1 GFP 50	1.29	1.86	1.05	-0.41	1.47	1.66	0.46	0.44
Mean Means ClassII per Cells AreaShape Eccentricity	1.73	0.24	-3.26	0.22	-3.04	0.25	2.07	0.76
Mean Means ClassII per Cells Texture Entropy GFP 3	-5.48	3.84	-1.93	-3.06	0.85	-5.11	-0.89	-7.8
Mean Cells Children EE Count	-1.47	0.08	-1.38	1.9	2.63	7.77	-1.7	3.47
Mean Means EE per Cells AreaShape Perimeter	2.26	-2.24	-2.48	-1.96	-1.32	0.28	0.17	-3.37
Mean Nuclei AreaShape Solidity	1.03	-0.69	-1.63	1.14	0.21	-1.18	-0.66	-0.67
Mean Means Golgi per Cells Intensity IntegratedIntensity RFP	-0.5	0.65	-1.69	-0.93	-1.55	-1.2	-0.42	-0.34
Mean Means Golgi per Cells Intensity IntegratedIntensityE RFP	-1.02	1.49	-1.61	-0.6	-1.43	-0.8	-0.48	0.79
Mean Means Golgi per Cells RadialIntensityDist FracAID RFP 2	0.87	-0.9	-2.09	-0.34	-2.95	-3.82	0.83	-1.26
Mean Means Golgi per Cells RadialIntensityDist FracAID RFP 4	-0.98	1.67	-0.17	0.49	1.12	1.76	-0.92	2.08
Size of Cluster	3	8	3	33	5	4	7	186

IF GO0006139 = 1 AND GO0065007 = 1 THEN

Genes involved in regulation (GO0065007) and in particular cellular nucleobase, nucleoside, nucleotide and nucleic acid metabolic processes (GO0006139)

Conclusions

- Application of the predictive clustering paradigm for analysis of phenotype images from siRNA screen for MHC Class II antigen presentation
- The clusters and their descriptions are obtained in a single step
- Identified and described groups of genes which yield similar phenotypes upon siRNA mediated downregulation