Logic Programming for Big Data in Computational Biology

Nicos Angelopoulos

Wellcome Sanger Institute
Hinxton, Cambridge

nicos.angelopoulos@sanger.ac.uk

18.9.18

overview

- knowledge for Bayesian machine learning over model structure
- applied knowledge representation for biological data analytics

Bayesian inference of model structure (Bims)

A Bayesian machine learning system that can model prior knowledge by means of a probabilistic logic programming.

Nonmeclature

- DLPs = Distributional logic programs
- Bims = Bayesian inference of model structure

Timeline

- Theory (York, 2000-5)
- Applications (Edinburgh, 2006-8, IAH 2009, NKI 2013)
- Bims library and theory paper 2015-2017

Bims Overview

- syntax of DLPs
- a succinct classification tree prior program
- Bayesian learning of model structure
- learning classification and regression trees
- Bayesian learning of Bayesian networks
- the bims library

DLPs- description

We extend LP's clausal syntax with probabilistic guards that associate a resolution step using a particular clause with a probability whose value is computed on-the-fly.

The intuition is that this value can be used as the probability with which the clause is selected for resolution.

Thus in addition to the logical relation, a clause defines over the objects that appear as arguments in its head, it also defines a probability distribution over aspects of this relation.

DLPs example

$$\begin{aligned} &\mathit{member}(H,[H|\_T]).\\ &\mathit{member}(El,[\_H|T]) \ :- \\ &\qquad \mathit{member}(El,T). \end{aligned} \tag{C_1}$$

$$L :: \mathit{length}(\mathit{List}, L) \sim El :: \mathit{umember}(El, \mathit{List}) \tag{G_1}$$

$$\frac{1}{L} :: L :: \mathit{umember}(El,[El|\mathit{Tail}]). \tag{C_2}$$

$$\begin{aligned} 1-\frac{1}{L} :: L :: \mathit{umember}(El,[H|\mathit{Tail}]) \ :- \\ \mathit{umember}(El,\mathit{Tail}). \end{aligned} \tag{C_3}$$

DLPs probabilistic goals

$$\frac{1}{L} :: L :: umember(El, [El|Tail]). \tag{C_4}$$

$$\begin{aligned} 1 - \frac{1}{L} :: L :: umember(El, [H|Tail]) \ :- \\ K \ \text{ is } \ L - 1, \\ K :: umember(El, Tail). \end{aligned} \tag{C_5}$$

DLPs query

$$?-umember(X,[a,b,c]).$$

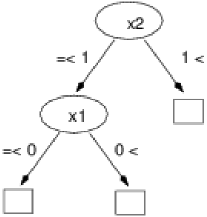
$$X = a \quad (1/3 \text{ of the times } = 1/3);$$

$$X = b \quad (1/3 \text{ of the times } = 2/3 * 1/2);$$

$$X = c \quad (1/3 \text{ of the times } = 2/3 * 1/2 * 1).$$

simple tree prior

?- cart(ζ, ξ, A, M).



M=nd(x2,1,nd(x1,0,lf,lf),lf)

$$
\begin{array}{lrl}
(C_0) & & cart(\zeta, \xi, M, Cart) :- \\
 & & \quad \psi_0 \text{ is } \zeta, \\
 & \psi_0: & \quad split(0, \zeta, \xi, M, Cart).
\end{array}
$$

$$
\begin{array}{lrl}
(C_1) & \psi_D: & split(D, \zeta, \xi, M_B, nd(F, Val, L, R)) :- \\
 & & \quad \psi_{D+1} \text{ is } \zeta * (1+D)^{-\xi}, \\
 & & \quad D_1 \text{ is } D+1, \\
 & & \quad r\_select(F, Val, M_B, L_B, R_B), \\
 & \psi_{D+1}: & \quad split(D_1, \zeta, \xi, L_B, L), \\
 & \psi_{D+1}: & \quad split(D_1, \zeta, \xi, R_B, R).
\end{array}
$$

$$
(C_2)\ 1 - \psi_D: \quad split(D, \zeta, \xi, M_B, lf).
$$

Bims theory

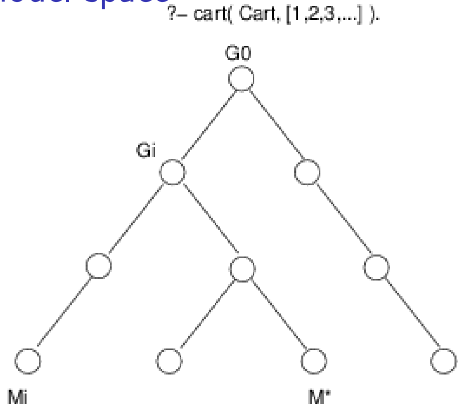
Bayes' Theorem

$$p(M|D) = \frac{p(D|M)p(M)}{\sum_M p(D|M)p(M)}$$

Metropolis-Hastings

$$\alpha(M_i, M_*) = min\left\{\frac{q(M_*, M_i)P(D|M_*)P(M_*)}{q(M_i, M_*)P(D|M_i)P(M_i)}, 1\right\}$$

DLP defined model space



From M_i identify G_i then sample forward to $M_\star$.
$q(M_i, M_\star)$ is the probability of proposing $M_\star$ when M_i is the current model.

Pyruvate kinase interactors

objective

improve chances of discovering binding molecules based on examples from screened chemical libraries.

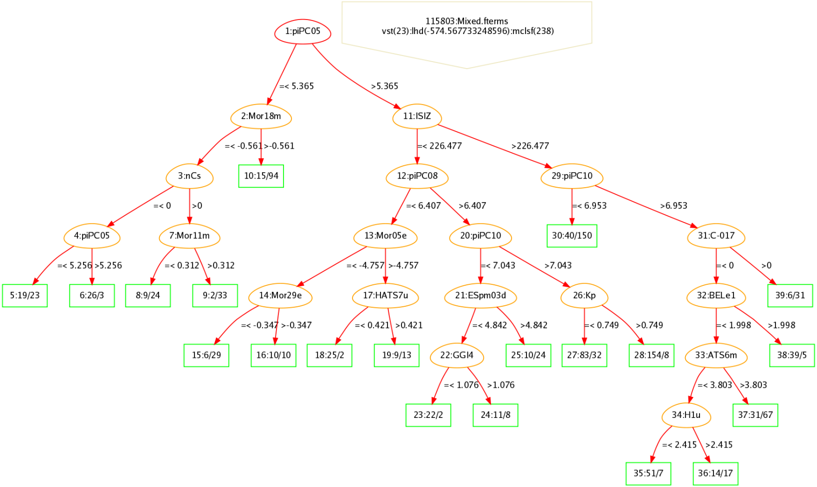
pyruvate kinase affinity data

582 Active and 582 Inactive. Dragon software produces 1500 property descriptors for each molecule, about 1100 were used.

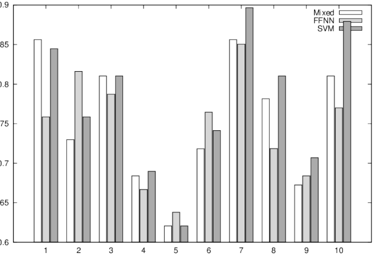
ten-fold cross-validation

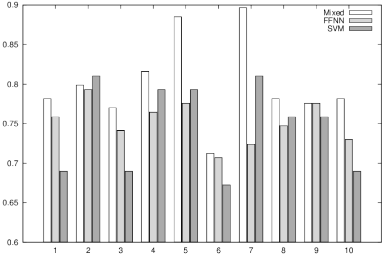
Compared to Feed Forward Neural Networks and Support Vector Machines by splitting the data into ten train/test segments.

best likelihood model



ten-fold validation

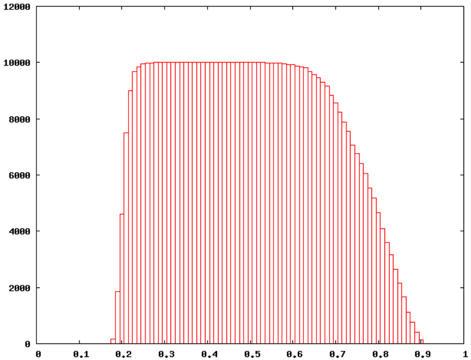




$$Sensitivity = \frac{T^{+}}{T^{+} + F^{-}}$$

$$Specificity = \frac{T^{-}}{T^{-} + F^{+}}$$

molecules of Eduliss according to BCarts



Bims: Bayesian inference of model structure

Released in 2016 as an easily installable SWI-Prolog library
Includes (IJAR paper in 2017)

- priors and likelihoods for: CARTs and Bayesian networks
- hooks for user defined models

Probabilistic logic programming

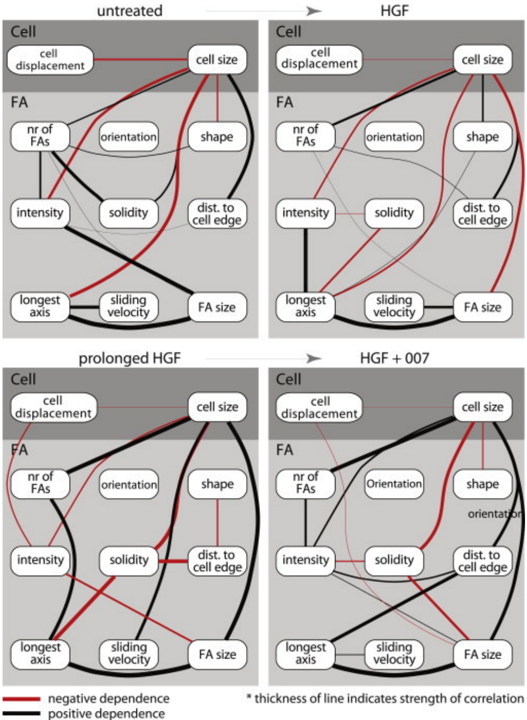
- thesis: probabilistic finite domains
- PLP workshop and IJAR associated issues (5th edition)

knowledge-based computation biology

- graphical models
  (focal adhesion dynamics, NKI, 2011-3)
- proteomics functional analysis
  (TKSilac,KSR1,ATG9A, Imperial, 2014-5)
- mutational profiling
  (14MG, Sanger, 2016-8)

Graphical models of FAD

Graphical models (aka Bayesian networks) can provide a network view of dependencies among variables, capturing much richer information than pairwise correlations.
In this project, microscopy based variables characterising focal adhesion in time are connected for a number of conditions in the HGF pathway.


untreated
HGF
Cell
cell displacement
cell size
FA
nr of FAs
orientation
shape
intensity
solidity
dist. to cell edge
longest axis
sliding velocity
FA size
Cell
cell displacement
cell size
FA
nr of FAs
orientation
shape
intensity
solidity
dist. to cell edge
longest axis
sliding velocity
FA size
prolonged HGF
HGF + 007
Cell
cell displacement
cell size
FA
nr of FAs
orientation
shape
intensity
solidity
dist. to cell edge
longest axis
sliding velocity
FA size
Cell
cell displacement
cell size
FA
nr of FAs
Orientation
shape
orientation
intensity
solidity
dist. to cell edge
longest axis
sliding velocity
FA size
negative dependence
positive dependence
* thickness of line indicates strength of correlation

tkSilac: tyrosine kinase screen

- MCF7 cell line
- 33 SILAC runs
- 65/66 expressed tyrosine kinases

- 4739 quantified in some experiment
- 1000 quantified in 60 or more TK KO

Figure 2

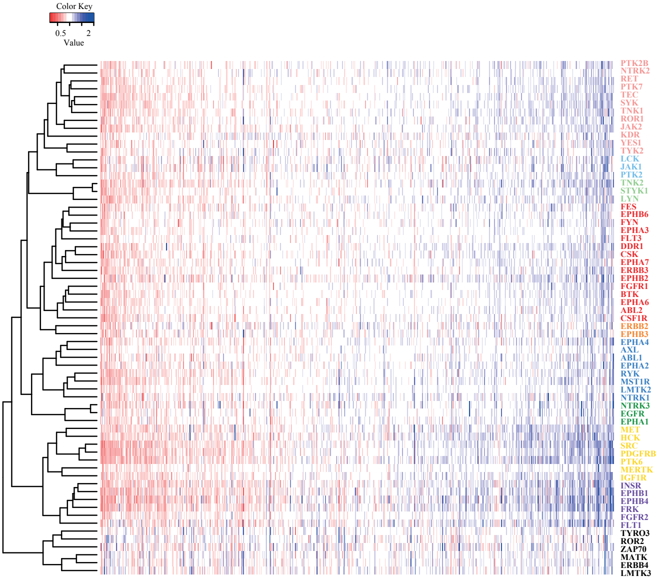
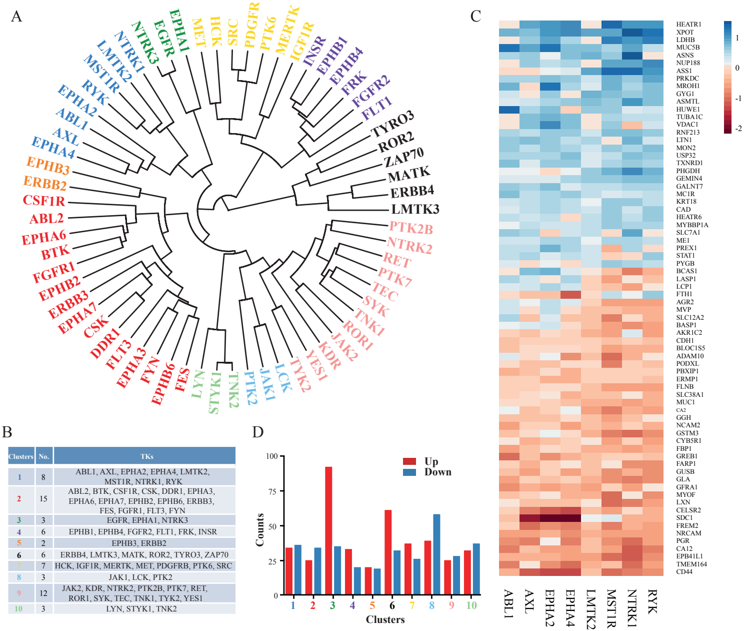


Fig. 2. Heatmap of quantified proteins after TK silencing. The overall pattern of regulation is shown in the heatmap of quantified values. After normalized to siControl, values of fold changes are all above 0, with value 1 showing that the expression levels of the specific protein are not altered after silencing TKs. For each knockdown (rows)

Figure 4



Clusters	No.	TKs
1	8	ABL1, AXL, EPHA2, EPHA4, LMTK2, MST1R, NTRK1, RYK
2	15	ABL2, BTK, CSF1R, CSK, DDR1, EPHA3, EPHA6, EPHA7, EPHB2, EPHB6, ERBB3, FES, FGFR1, FLT3, FYN
3	3	EGFR, EPHA1, NTRK3
4	6	EPHB1, EPHB4, FGFR2, FLT1, FRK, INSR
5	2	EPHB3, ERBB2
6	6	ERBB4, LMTK3, MATK, ROR2, TYRO3, ZAP70
7	7	HCK, IGF1R, MERTK, MET, PDGFRB, PTK6, SRC
8	3	JAK1, LCK, PTK2
9	12	JAK2, KDR, NTRK2, PTK2B, PTK7, RET, ROR1, SYK, TEC, TNK1, TYK2, YES1
10	3	LYN, STYK1, TNK2

Figure 5

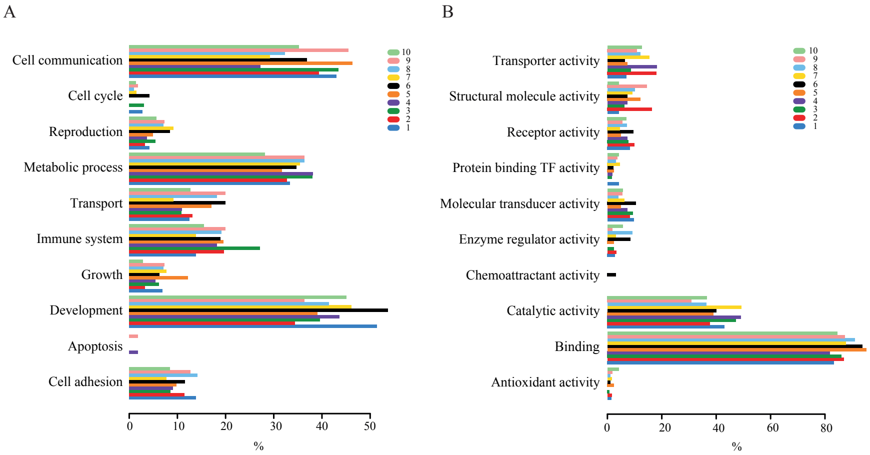
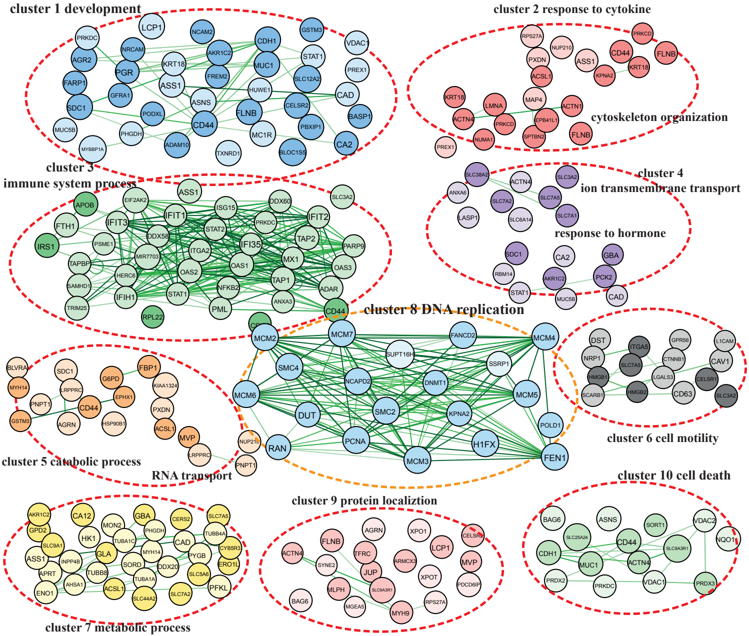
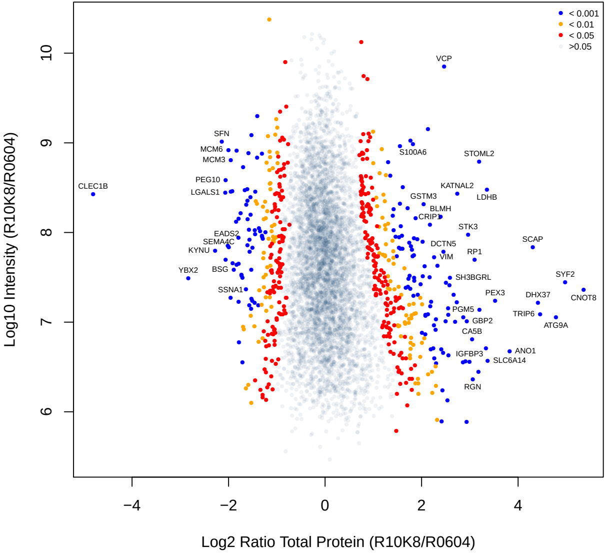


Fig. 5. Characterization of a functional portrait for each cluster. A, A functional profile of top GO biologic processes that the up- and downregulated proteins belong to is presented. x-axis shows the percentage of hits in each cluster that belong to a GO biologic process term. The color coding and the number for each cluster are indicated as above. B, A functional profile of top GO molecular functions that the up- and downregulated proteins belong to is presented. x-axis shows the percentage of hits in each cluster that belong to a GO molecular function term.

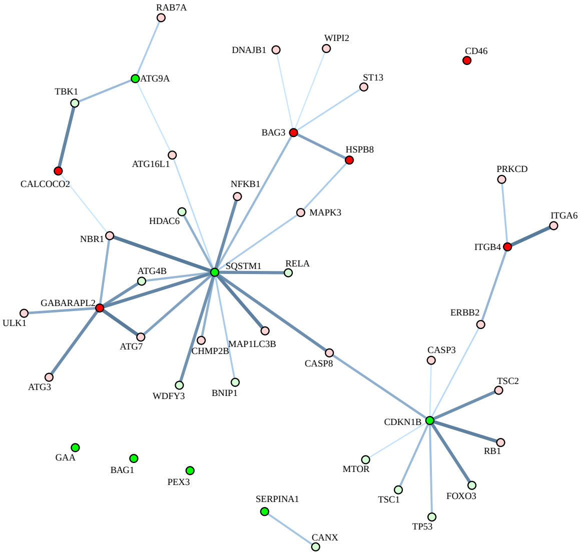
Figure 6


cluster 1 development
PRKDC
LCP1
NCAM2
GSTM3
VDAC1
NRCAM
CDH1
AGR2
AKR1C2
STAT1
KRT18
PGR
MUC1
PREX1
FARP1
FREM2
SLC12A2
ASS1
HUWE1
CAD
GFRA1
ASNS
SDC1
PODXL
CELSR2
FLNB
BASP1
CD44
MUC5B
PHGDH
PBXIP1
MC1R
MYBBP1A
ADAM10
TXNRD1
CA2
BLOC1S5
cluster 2 response to cytokine
RPS27A
PRKCD
NUP210
CD44
FLNB
PXDN
ASS1
KRT18
ACSL1
KPNA2
KRT18
MAP4
LMNA
ACTN1
ACTN4
PRKCD
EPB41L1
cytoskeleton organization
FLNB
SPTBN2
NUMA1
PREX1
cluster 3
immune system process
APOB
EIF2AK2
ASS1
DDX60
SLC3A2
FTH1
IFIT3
IFIT1
ISG15
IFIT2
PRKDC
STAT2
DDX58
TAP2
IRS1
PSME1
ITGA2
IFI35
PARP9
MIR7703
MX1
TAPBP
OAS2
OAS1
OAS3
HERC6
TAP1
SAMHD1
NFKB2
ADAR
IFIH1
STAT1
TRIM25
ANXA3
PML
RPL22
CD44
cluster 4
ion transmembrane transport
SLC38A2
ACTN4
SLC3A2
ANXA6
SLC7A2
SLC7A5
LASP1
SLC6A14
SLC7A1
response to hormone
SDC1
CA2
GBA
RBM14
AKR1C2
PCK2
STAT1
MUC5B
CAD
cluster 8 DNA replication
MCM2
MCM7
FANCD2
MCM4
SUPT16H
SSRP1
SMC4
NCAPD2
DNMT1
MCM6
MCM5
DUT
SMC2
KPNA2
POLD1
RAN
PCNA
H1FX
MCM3
FEN1
DST
ITGA5
GPR56
L1CAM
NRP1
CTNNB1
CAV1
SLC7A5
HMGB1
LGALS3
CELSR1
SCARB1
HMGB2
CD63
SLC3A2
cluster 6 cell motility
BLVRA
SDC1
FBP1
MYH14
G6PD
KIAA1324
LRPPRC
PNPT1
CD44
EPHX1
PXDN
GSTM3
AGRN
HSP90B1
ACSL1
MVP
NUP210
LRPPRC
PNPT1
cluster 5 catabolic process
RNA transport
cluster 10 cell death
cluster 9 protein localiztion
AKR1C2
CA12
GBA
CERS2
SLC7A5
GPD2
MON2
PHGDH
TUBB4A
HK1
TUBA1C
CAD
SLC9A1
MYH14
CYB5R3
ASS1
INPP4B
GLA
PYGB
SORD
ERO1L
APRT
TUBB8
DDX20
SLC5A6
AHSA1
TUBA1A
ENO1
ACSL1
SLC44A2
SLC7A2
PFKL
cluster 7 metabolic process
AGRN
XPO1
CELSR2
FLNB
LCP1
ACTN4
TFRC
ARMCX3
SYNE2
MVP
JUP
XPOT
MLPH
SLC9A3R1
PDCD6IP
BAG6
MGEA5
MYH9
RPS27A
BAG6
ASNS
SORT1
VDAC2
NQO1
SLC25A24
CD44
CDH1
SLC9A3R1
MUC1
ACTN4
PRDX2
PRKDC
VDAC1
PRDX3

volcano plot (BT474HR H/M)

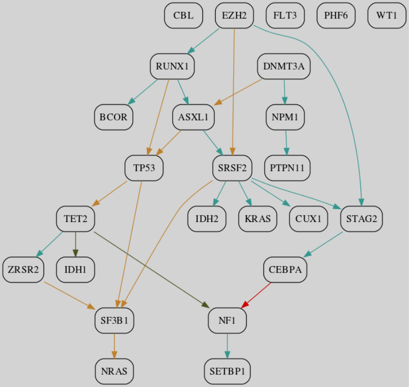


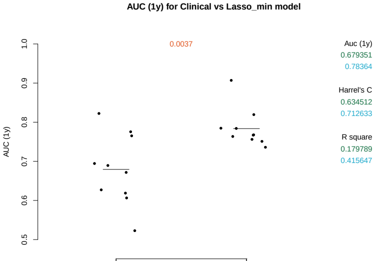
Autophagy



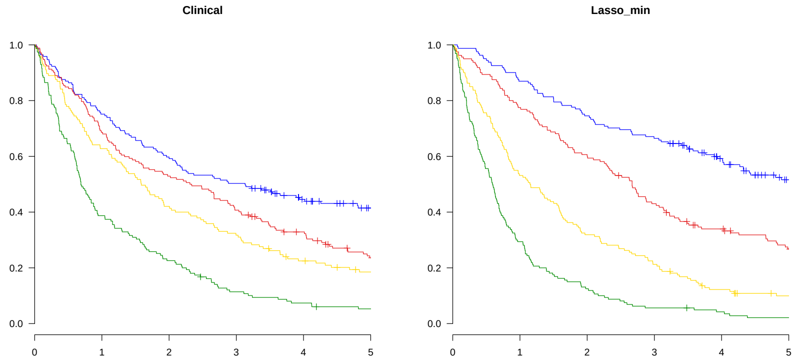
RAB7A
DNAJB1
WIPI2
CD46
ATG9A
TBK1
ST13
BAG3
HSPB8
ATG16L1
CALCOCO2
PRKCD
NFKB1
MAPK3
ITGA6
HDAC6
NBR1
ITGB4
SQSTM1
RELA
ATG4B
GABARAPL2
ULK1
ERBB2
ATG7
MAP1LC3B
CHMP2B
CASP3
CASP8
ATG3
WDFY3
BNIP1
TSC2
CDKN1B
RB1
GAA
BAG1
MTOR
PEX3
FOXO3
TSC1
SERPINA1
TP53
CANX

Myelodysplastic syndrom, NGS somatic mutations profiling

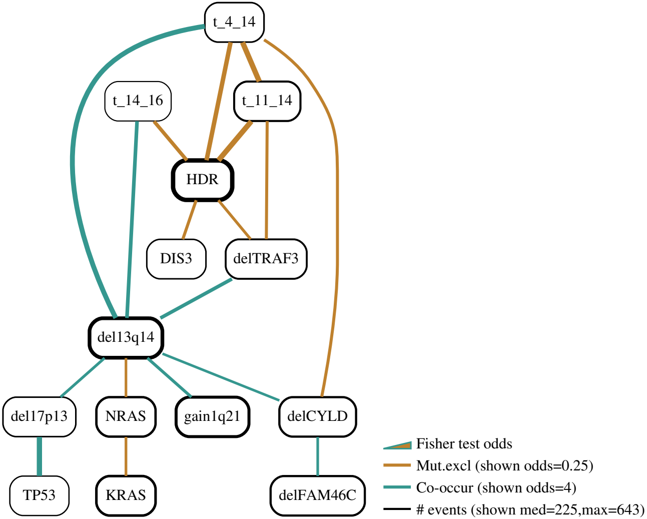




5 year: Clinical vs Lasso (Optimal)



myeloma structural variations



logic programming for (biological) data analytics

Positives

- interpreted
- memory management
- clean and high level
- probabilistic ML & reasoning (Prism,Bims,Pepl)
- intuitive database integration (db_facts,bio_db)
- multi-threaded and web-capable
- talking to other systems (R:Real,ODBC,proSQLite)
- (largely) OS independence

Negatives

- graphics
- SWI-Prolog, at core a one-person project
- code sharing in toddler stage (but showing promise)
- in-browser interaction with other technologies

KR bottom line

(probabilistic) logic programming and Bayesian networks are powerful tools for

explainable, accountable, open and **shareable** AI & ML

KR bottom line

(probabilistic) logic programming and Bayesian networks are powerful tools for

explainable, accountable, open and **shareable** AI & ML

symbolic AI education, can be a central player in

contributing tangibly to the current AI resurgence, while
managing expectations of modern AI
see media coverage of Facebook/Cambridge-Analytica
& Uber/Tesla driveless accidents

KR bottom line

(probabilistic) logic programming and Bayesian networks are powerful tools for

explainable, accountable, open and **shareable** AI & ML

symbolic AI education, can be a central player in

contributing tangibly to the current AI resurgence, while
managing expectations of modern AI
see media coverage of Facebook/Cambridge-Analytica
& Uber/Tesla driveless accidents

biology presents a unique application area, where

unprecedented volumes of data are generated
knowledge is a crucial concept, currently being shaped
transferable to other big data areas