knowledge AI in bio-data analytics

Nicos Angelopoulos

Cardiff University

https://stoics.org.uk/~nicos
nicos@stoics.org.uk

22.03.29

background

1989-92 - Keele University, UK
BSc in Computer Science and Statistics

1992-93 - Imperial College, London
MSc in (Classical) AI

1996-01 - City University, London
PhD (Thesis: Probabilistic Finite Domains)

2002-08 - York University (+Edinburgh)
Bayesian Inference of Model Structure (Bims)

2009-14 - NKI + Imperial
AI for big-data cancer analytics

2015-18 - Wellcome Sanger Institute
Bayesian Networks for genomic cancer cohorts

overview

- Stream 1. prior knowledge for Bayesian machine learning
- Stream 2. applied knowledge representation for biological big data analytics
- Stream 3. Bayesian networks for cancer and biological datasets

Stream 1. (York) Bayesian inference of model structure

A probabilistic programming framework for Bayesian machine learning of structured statistical models, such as classification trees and graphical models (Bayesian networks).

Allows the encoding of prior information in the form of a probabilistic logic program.

Nomenclature

- 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 theory - Bayesian machine learning

Bayes' Theorem

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

Metropolis-Hastings

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

Bims

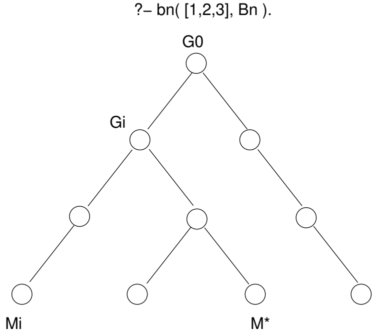
Bims utilises a probabilistic logic programming language to express detailed prior information

$$p(M)$$

Bims uses the implicitly defined space to find new proposed models and thus does away with calculating

$$q(M_*, M_i)$$

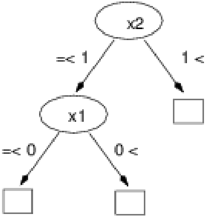
DLP defined model space



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.

simple tree prior

?- cart(ζ, ξ, A, M).



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

$$
\begin{array}{lrl}
(C_0) & & cart(\zeta,\xi,M,Cart) :- \\
 & & \psi_0 \text{ is } \zeta, \\
 & \psi_0: & split(0,\zeta,\xi,M,Cart). \\
 & & \\
(C_1) & \psi_D: & split(D,\zeta,\xi,M_B,nd(F,Val,L,R)) :- \\
 & & \psi_{D+1} \text{ is } \zeta * (1+D)^{-\xi}, \\
 & & D_1 \text{ is } D+1, \\
 & & r\_select(F,Val,M_B,L_B,R_B), \\
 & \psi_{D+1}: & split(D_1,\zeta,\xi,L_B,L), \\
 & \psi_{D+1}: & split(D_1,\zeta,\xi,R_B,R). \\
 & & \\
(C_2) & 1-\psi_D: & split(D,\zeta,\xi,M_B,lf).
\end{array}
$$

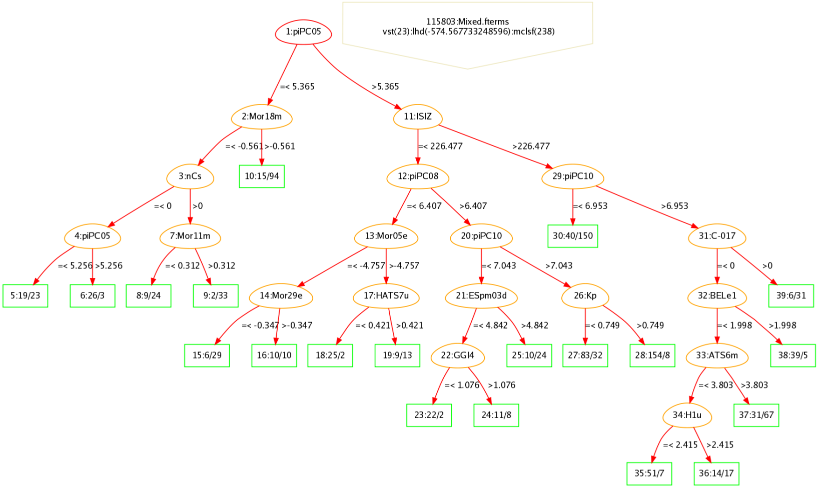
(Edinburgh) Pyruvate kinase interactors

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

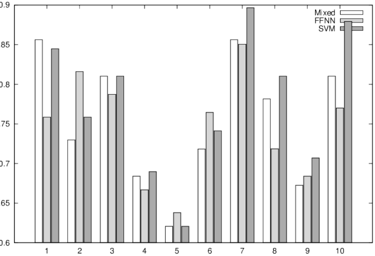
pyruvate kinase affinity data
582 Active and 582 Inactive. Dragon software produces 1500 property descriptors for each molecule, about 1100 were used.

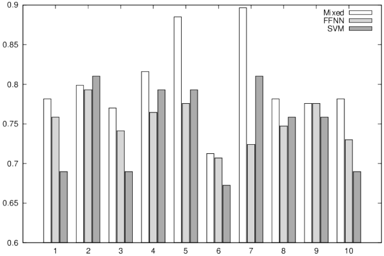
ten-fold cross-validation
Compared to Feed Forward Neural Networks and Support Vector Machines by splitting the data into ten train/test segments.

best likelihood model



ten-fold validation





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

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

Bims: Bayesian inference of model structure

Early publications:
UAI ’01, ICML 2005, IJCAI 05, (journal) AMAI 2008.

More recent developments: in 2016 as an easy to install SWI-Prolog library
(IJAR paper in 2017, (SJR: Q1))

Includes

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

Stream 2. (Imperial) Knowledge-based data analytics

tkSilac: tyrosine kinase screen

- MCF7 cell line
- 33 SILAC runs
- 65/66 expressed tyrosine kinases
- 4739 proteins quantified in some experiment
- 1000 proteins quantified in 60 or more TK KO

Molecular and Cellular Proteomics (MCP) 2015

Tk screen- input matrix

Figure 2

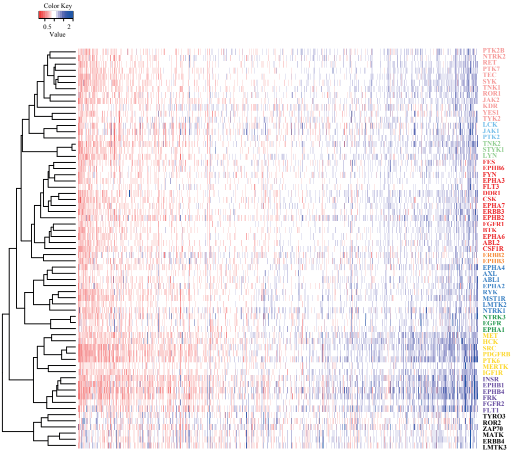
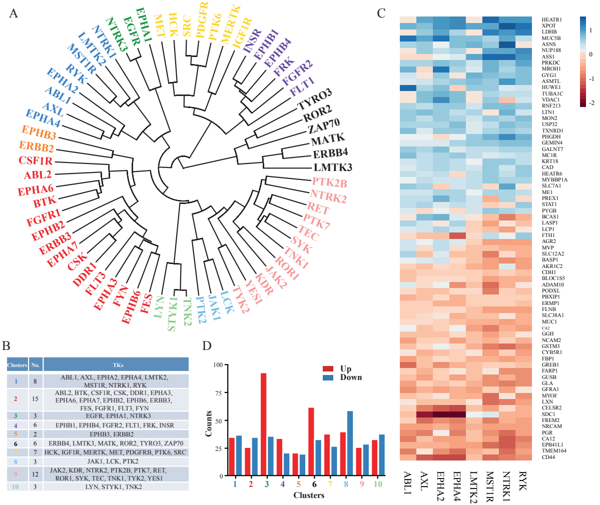


Fig. 2. Heatmap of quantified proteins after TK silencing. The overall pattern of regulation is shown in the heatmap of quantified values. After normalized to siControl, values of fold changes are all above 0, with value 1 showing that the expression levels of the specific protein are not altered after silencing TKs. For each knockdown (rows) the quantified value for an identified protein is plotted in red for down regulated proteins (below 1), white for non-differential and non-identified and blue for up-regulated proteins (above 1). The row labels indicate the knock out experiment and the colors correspond to the clusters described below.

Tk screen- clusters

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

Fig. 4. Hierarchical clustering of the 65 TKs expressed in MCF7 cells. A, Hierarchical clustering of the 65 TKs was performed using R's hclust function. The complete linkage method which aims to identify similar clusters based on overall cluster measure was used. 10 distinctive clusters were obtained and the complete dendrogram is

Tk screen- Gene Ontology

Figure 5

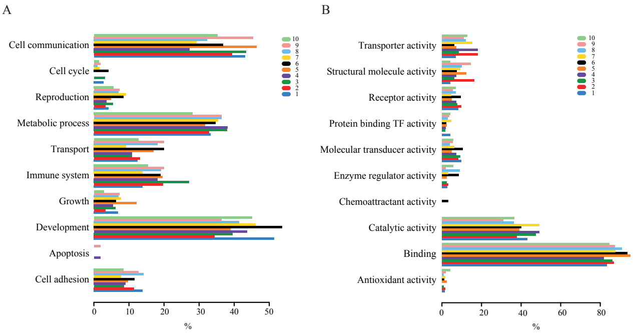


Fig. 5. Characterization of a functional portrait for each cluster. A, A functional profile of top GO biologic processes that the up- and downregulated proteins belong to is presented. x-axis shows the percentage of hits in each cluster that belong to a GO biologic process term. The color coding and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

Tk screen- GO terms + STING edges

Figure 6

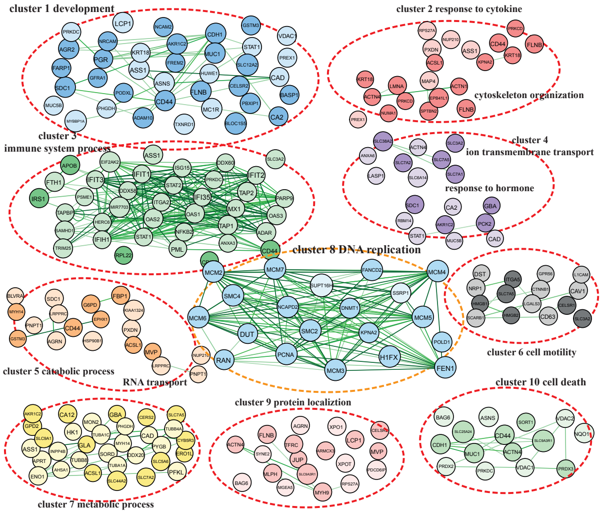
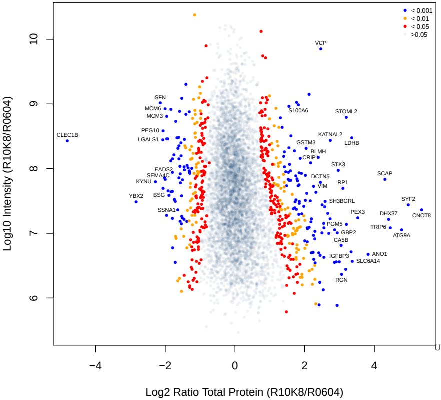
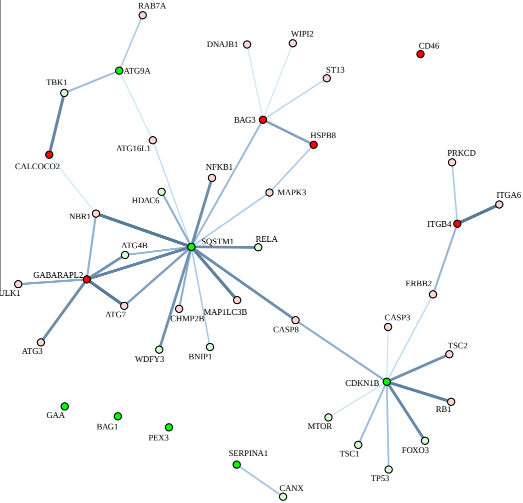


Fig. 6. Representatives of defined functional networks in each classified TK cluster. The functional networks were generated using GO analysis combined with the STRING platform. Proteins in lighter color are up-regulated, whereas brighter color indicates down-regulation. Arrows show the interactions between connected proteins. Rep-

herceptin resistance (BT474HR) — ATG9A / autophagy





proteomics data analytics (Imperial)

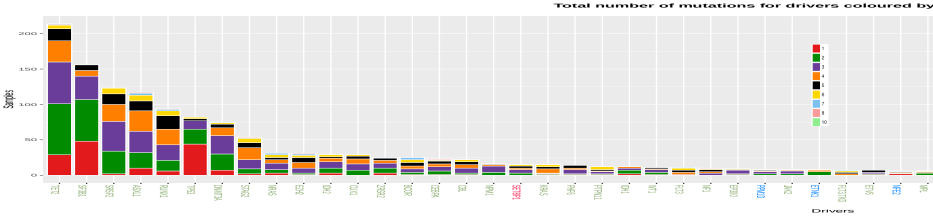
tyrosine kinase screen
Molecular and Cellular Proteomics (MCP) 2015

KSR1:
Breast Cancer Res. and Treat., 2015

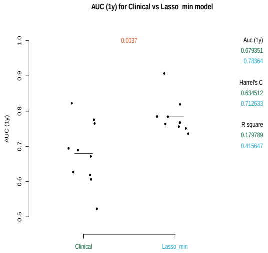
ATG9A:
Oncotarget 2016

Prolog libraries:
Real (> 550), proSQLite (> 700), bio_db, bio_analytics

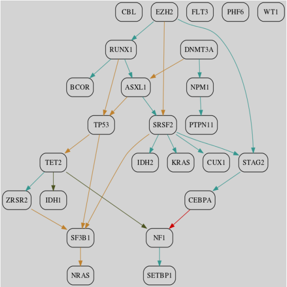
Stream 3: (Sanger) Bayesian networks in cancer genomics

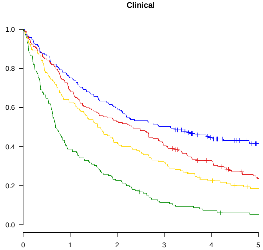


Sanger- survival models

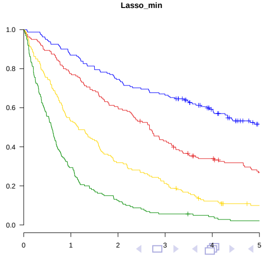


AUC (1y) for Clinical vs Lasso_min model



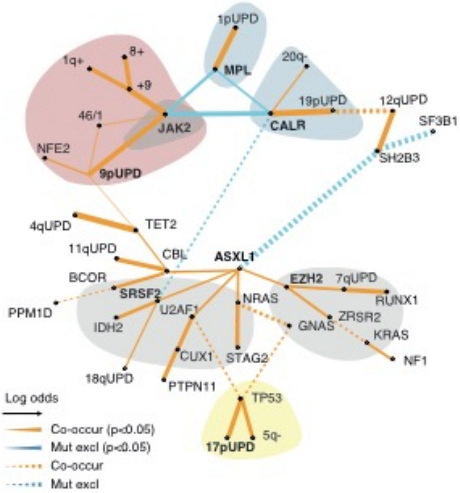


Clinical



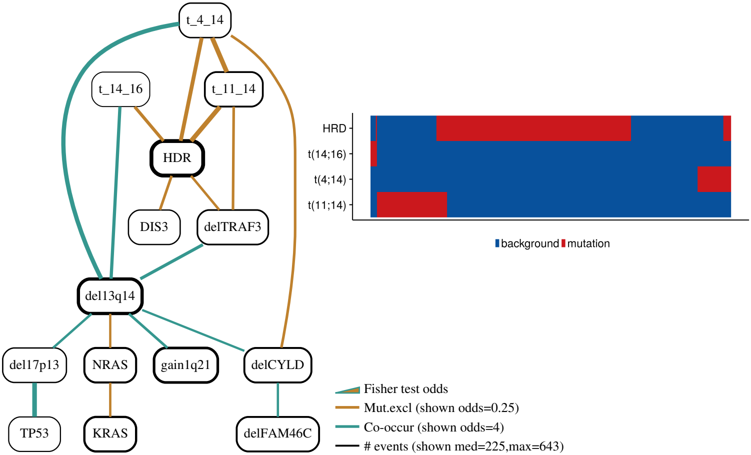
Lasso_min

MPN: myeloproliferative neoplasms



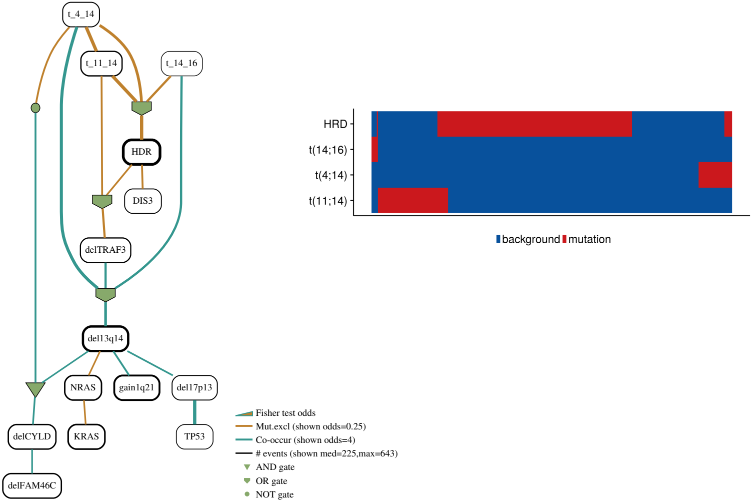
New England Journal of Medicine, October 2018

myeloma structural variations



Nature Communications, August 2019

myeloma gated structural variations



BNs in cancer genomics

- MPN published in New England J. of Medicine, Oct, 2018
- multiple myeloma: in Nature Communications (3rd author), Aug, 2019
- colorectal: January 2020
  (with Dutch collaborators - J. of Clin. Oncology)
- 1st author methods paper:
  accepted late February in Communications Biology

Renal carcinoma, Bayesian estimate

Cell

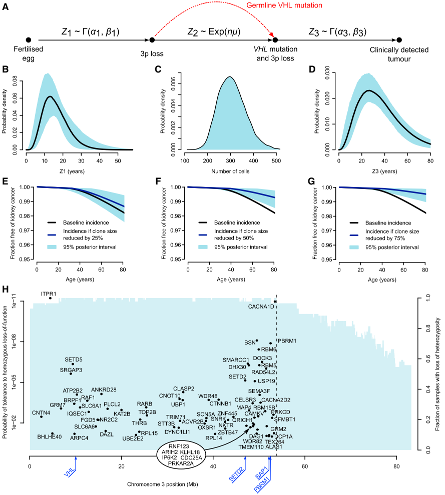


Figure 7. Mathematical Modeling of Clear Cell Renal Cell Carcinoma Evolution

(A) Schematic depicting how the age of incidence of renal cell carcinoma may be modeled as the sum of waiting times; Z_1 representing the time to 3p loss, Z_2 representing the time to *VHL* inactivation, and Z_3 representing the time from bi-allelic loss of *VHL* to clinically detected tumor. Z_1 and Z_3 are modeled by gamma distributions and Z_2 by an exponential distribution of the product of *n*, the number of cells with 3p loss and μ, the calculated *VHL* mutational rate.

(B–D) The posterior distribution of the waiting times for Z_1 (B), the number of cells with 3p loss (C), and the waiting time for Z_3 (D) with 95% posterior intervals.

collaborators

Worked done in collaboration with colleageus in medicine/biology

- Dr Fransesco Maura (myeloma, Sloan Kettering, New York)
- Dr Peter Campbell (hemato-oncologist, Sanger)
- Dr Jyoti Nangalia (MPN, Cambridge/Sanger)
- Dr Georgios Giamas (kinase signalling, Sussex)
- Dr David MacIntyre (prenatal metabolomics, Imperial)

computer science

- Dr James Cussens (Bayesian networks, Bristol University)
- Dr Jan Wielemaker (SWI-Prolog, Amsterdam)

themes and leadership

Research themes

- AI models of disease evolution and signalling
- machine learning with priors
- knowledge based big data bio analytics

Leadership

- translational data science: from lab to clinic
- precision medicine
- computational biology
- AI, knowledge representation
- machine learning
- probabilistic logic programming