

knowledge based computational cancer biology

Nicos Angelopoulos

The Pirbright Institute
head of computational biology

<https://www.pirbright.ac.uk/users/dr-nicos-angelopoulos>
nicos.angelopoulos@pirbright.ac.uk

23.09.07

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
- 2019- - Essex + Cardiff + Pirbright
Indepent research in computational biology

prehistoric AI - Prolog

BSc - Keele University

Project supervisor: Paul Singleton

MSc - Imperial College

Project supervisor: Robert, "Bob", Kowalski

Prolog = Programming in Logic.

A practical programming language that was originally developed for classical AI reasoning based on logic (theorem proving).

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 4. computational biology at Pirbright

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

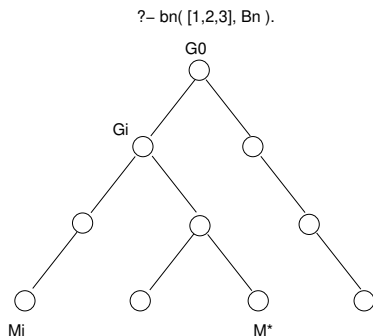
Bims utilises a probabilistic logic programming language to express detailed prior information

$$p(M)$$

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

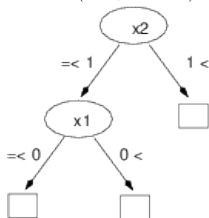


From M_i identify G_i then sample forward to M_{\star} .

$q(M_i, M_{\star})$ is the probability of proposing M_{\star} when M_i is the current model.

simple tree prior

?- $\text{cart}(\zeta, \xi, A, M)$.



$M = \text{nd}(x2, 1, \text{nd}(x1, 0, \text{lf}, \text{lf}), \text{lf})$

(C₀) $\text{cart}(\zeta, \xi, M, \text{Cart}) : -$

ψ_0 is ζ ,

ψ_0 : $\text{split}(0, \zeta, \xi, M, \text{Cart})$.

(C₁) ψ_D : $\text{split}(D, \zeta, \xi, M_B, \text{nd}(F, \text{Val}, L, R)) : -$

ψ_{D+1} is $\zeta * (1 + D)^{-\xi}$,

D_1 is $D + 1$,

$r_select(F, \text{Val}, M_B, L_B, R_B)$,

ψ_{D+1} : $\text{split}(D_1, \zeta, \xi, L_B, L)$,

ψ_{D+1} : $\text{split}(D_1, \zeta, \xi, R_B, R)$.

(C₂) $1 - \psi_D$: $\text{split}(D, \zeta, \xi, M_B, \text{lf})$.

(Edinburgh) Pyruvate kinase interactors

objective

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

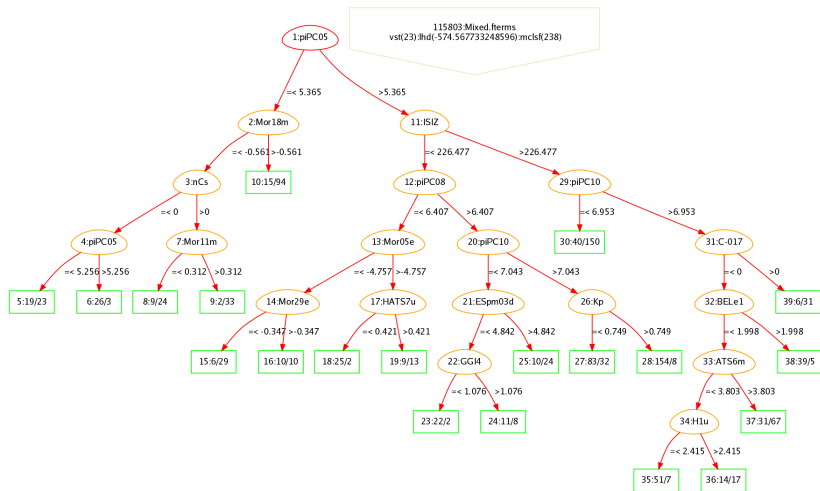
pyruvate kinase affinity data

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

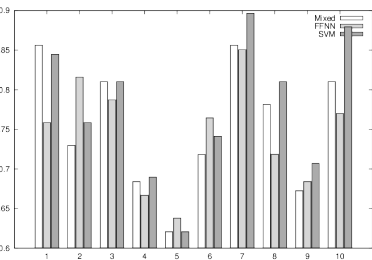
ten-fold cross-validation

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

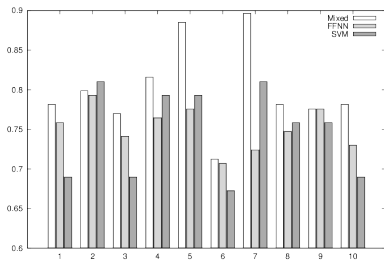
best likelihood model



ten-fold validation



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



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

Bims: Bayesian inference of model structure

Early publications:

UAI '01, ICML 2005, IJCAI 05, (journal) AMAI 2008.

Later stages: in 2016 as an easy to install SWI-Prolog library
(IJAR paper in 2017)

Includes

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

Bims: Lessons

Working with colleagues with complementary skills can lead to good quality work.

Combining logic (programming) and probability is a vexed question with a long history. *Bims* is an applied tool that provides a therotically sound way for Bayesian machine learning.

Biomedicine is an extremely fitting general application area, as there is a background truth (nature) and a huge (growing) body of prior knowledge expressed in biological databases (Gene Ontology, Reactome and many more)

Ability to include prior knowledge to machine learning and statistical inference when analysing the data from specific biological experiments, can lead to better results and better explainable AI models.

Stream 2. (Knowledge-based data analytics)

Postdoc/research positions

- ▶ NKI: Netherlands Cancer Institute
- ▶ Imperial College: tyrosine kinases (LMTK3)

Stream 2. (Imperial) Knowledge-based proteomics

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

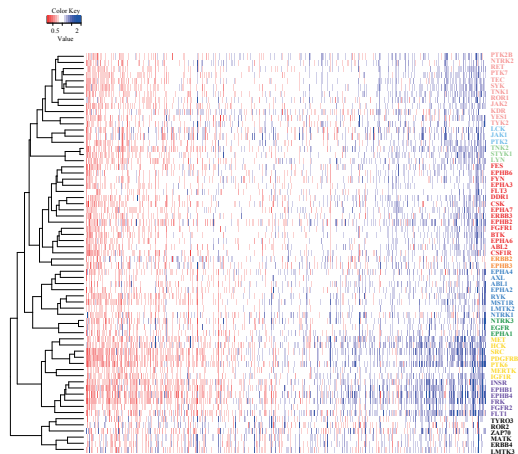


Fig. 2. Heatmap of quantified proteins after TK silencing. The overall pattern of regulation is shown in the heatmap of quantified values. After normalized to siControl, values of fold changes are all above 0, with value 1 showing that the expression levels of the specific protein are not altered after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

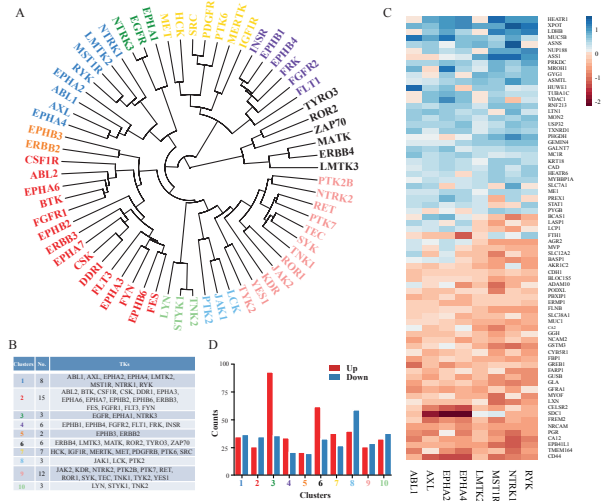


Fig. 4. Hierarchical clustering of the 65 TKs expressed in MCF7 cells. A, Hierarchical clustering of the 65 TKs was performed using R's hclust function. The complete linkage method which aims to identify similar clusters based on overall cluster measure was used. 10 distinctive clusters were obtained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

Tk screen- Gene Ontology

Figure 5

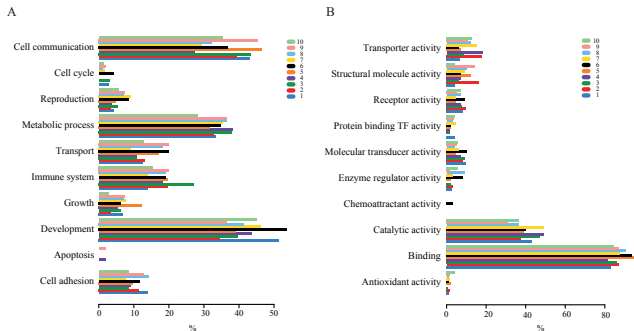


Fig. 5. Characterization of a functional portrait for each cluster. A, A functional profile of top GO biologic processes that the up- and downregulated proteins belong to is presented. x-axis shows the percentage of hits in each cluster that belong to a GO biologic process term. The color coding and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

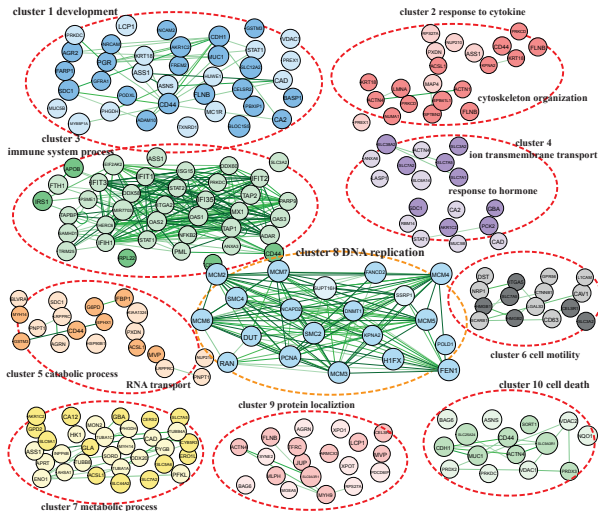
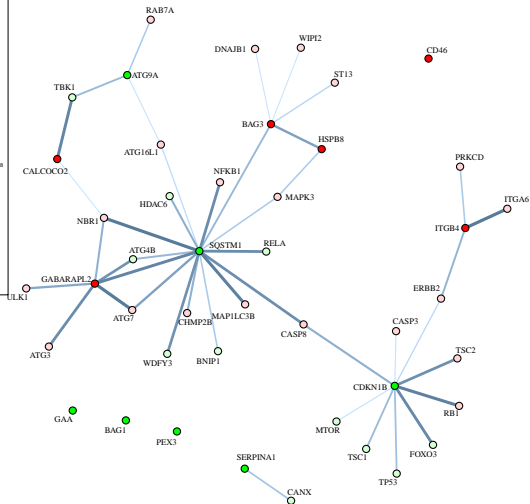
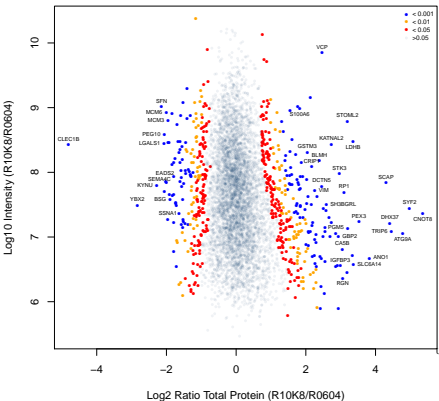


Fig. 6. Representatives of defined functional networks in each classified TK cluster. The functional networks were generated using GO analysis combined with the STRING platform. Proteins in lighter color are up-regulated, whereas brighter color indicates down-regulation. Arrows show the interactions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

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

Knowledge cancer analytics: lessons

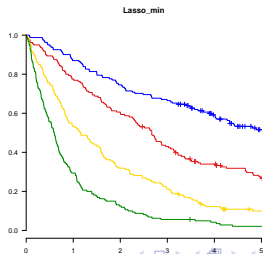
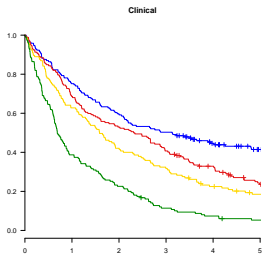
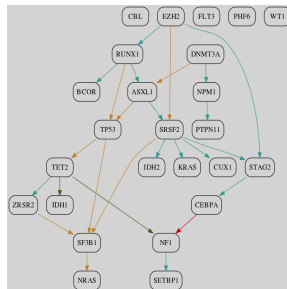
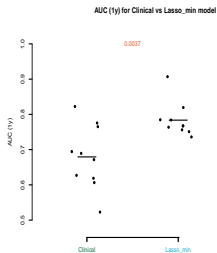
It is always a tightrope walk between how much time to what quality a piece of bio data analysis tool is written.

Once a robust pipeline is in place it can lead to a number good publications.

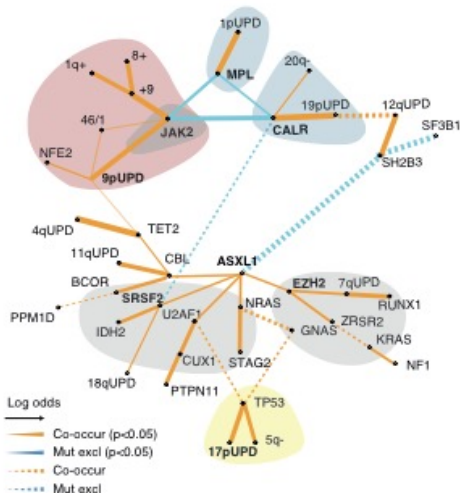
Using niche tool chains is only viable if you first invest in interface libraries.

Using niche tools that are "cleaner" and at "higher level" can allow small teams to be more productive.

Sanger- survival models

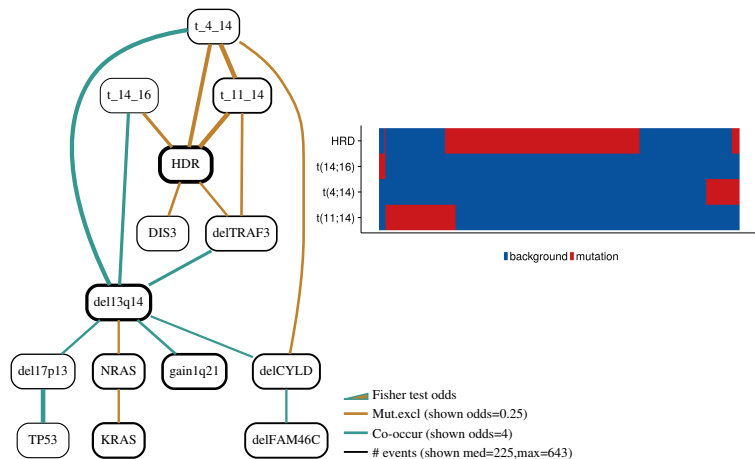


MPN: myeloproliferative neoplasms



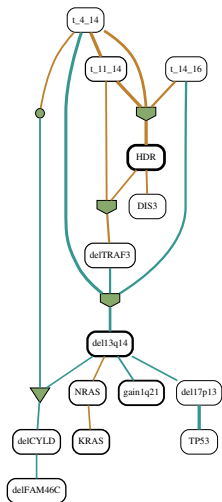
New England Journal of Medicine, October 2018

myeloma structural variations

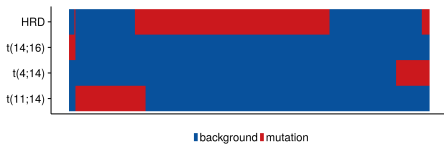


Nature Communications, August 2019

myeloma gated structural variations



- Fisher test odds
- Mut.excl (shown odds=0.25)
- Co-occur (shown odds=4)
- # events (shown med=225,max=643)
- AND gate
- OR gate
- NOT gate



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:
Communications Biology (2022)

Renal carcinoma, Bayesian estimate

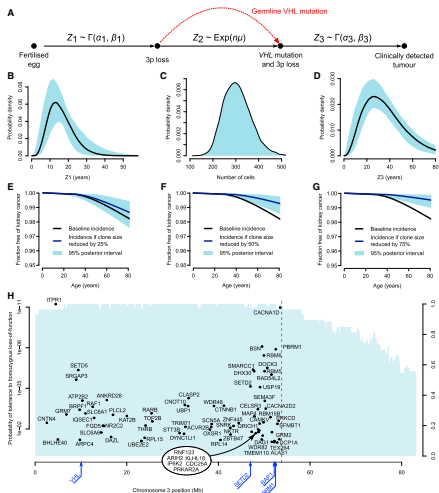


Figure 7. Mathematical Modeling of Clear Cell Renal Cell Carcinoma Evolution

(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.

Knowledge in cancer genomics: lessons

BNs can hit a sweet spot between complexity and explainability in many biomedicine settings.

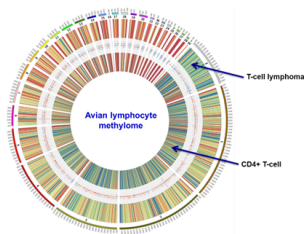
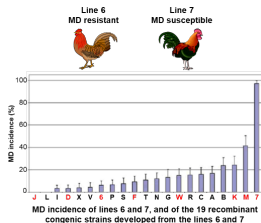
They work well as Figure 1 when there are less than 50 variables and for hypothesis generation.

Convincing medicine researchers to use any model other than a *p-value* can be a lengthy business.

Having in-house analytics people can accelerate and deepen interactions.

(Be mindful of the culture gap.)

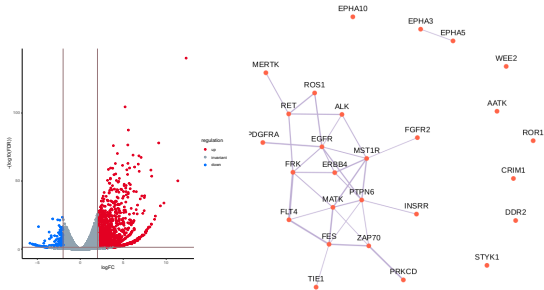
Pirbright Avian Oncogenic Viruses - genetic



Marek's disease virus.

Pac bio sequencing of lines 6 (resistant) and 7 (susceptible) and 6 recombinants. Particularly, in combination with other modalities such as methylation.

Avian Oncogenic Viruses - hypoxia



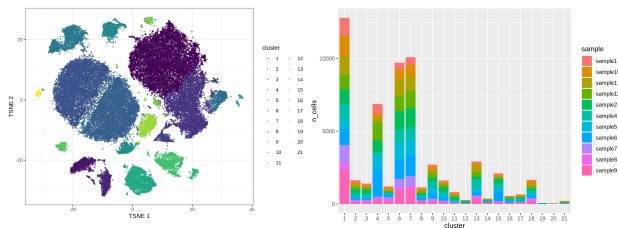
peptidyl-tyrosine
phosphorylation

Analyses of hypoxia and CRISPR screen experiments.

Statistical, significance, visualisation, gene ontology and protein-protein interactions.

Chicken genomic information blueprint of incorporating genomes relevant to Pirbright.

scRNA Corona viruses



Single cell RNA provide rich data...

pig viral BAL samples,

to then extend to more bespoke investigations,
including comparatives to human and mouse
in collaboration with Babraham (Dr Arianne Richard).

focus areas

- ▶ knowledge based analytics
- ▶ computational biology
- ▶ AI, knowledge representation
- ▶ machine learning
- ▶ probabilistic logic programming

In collaboration

- ▶▶ incorporate knowledge in analyses of experimental data
- ▶▶ interface with high quality external databases
- ▶▶ use and develop state-of-the-art algorithms
- ▶▶ well engineered robust, re-deployable pipelines
- ▶▶ large scale multi-omics projects

collaborators

medicine/biology

- ▶ Dr Francesco Maura (myeloma, Miami, USA)
- ▶ Dr Peter Campbell (hemato-oncologist, Sanger)
- ▶ Dr Jyoti Nangalia (MPN, Cambridge/Sanger)
- ▶ Dr Georgios Giamas (kinase signalling, Sussex)
- ▶ Prof Tassos Karadimitris (haematologist-Imperial College)

computer science

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

Thank you