Algebra and tensors give interpretable groups for crosstalk mechanisms in breast cancer

Mariano Beguerisse Díaz

Mathematical Institute
University of Oxford

June 12, 2018

Acknowledgements

Collaborators:

Anna Seigal (UC Berkeley)

Heather Harrington (Oxford)

Mario Niepel (Harvard)

Birgit Schoeberl (Merrimack)

Funding:

UNIVERSITY OF OXFORD

Emirates

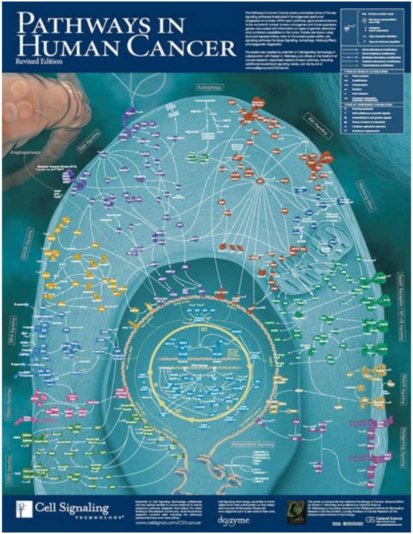
Pre-print: `arXiv:1612.08116`

Biological motivation

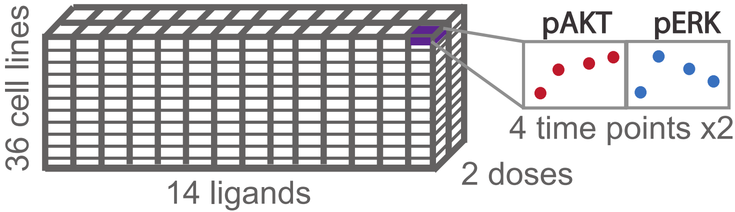
Chemotherapy is a blunt tool that kills indiscriminately all rapidly dividing cells.

Cancer physiology is complex.

Need for focused therapies to target cellular decision making of cancer cells.



Tensor data



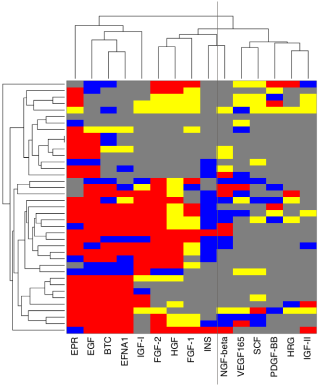
Five dimensional tensor containing results of 36×14 experiments.

The challenge is to **determine the signalling mechanisms** at play in these data.

Cluster experiments with similar responses.

Can be difficult to interpret mechanistically.

Need to impose constraints to **facilitate interpretation.**

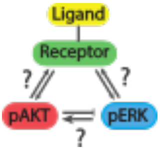


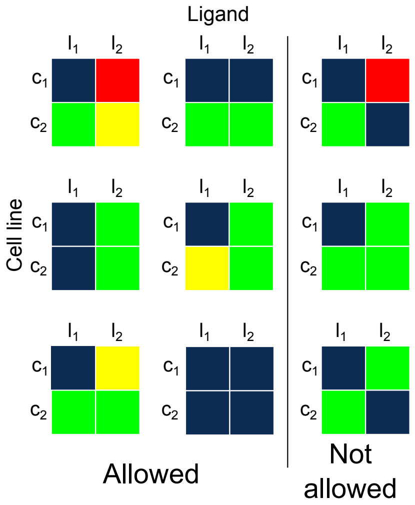
Rectangular clusters

Constrain clusters' shape.

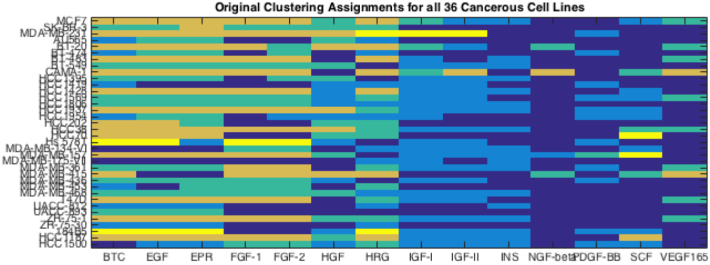
Rectangle-shaped clusters: single explanatory mechanism.

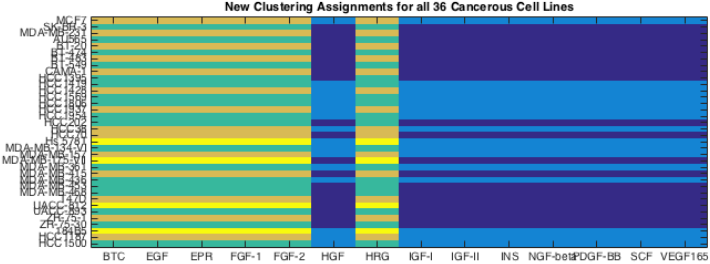
Find an ODE model for each cluster.



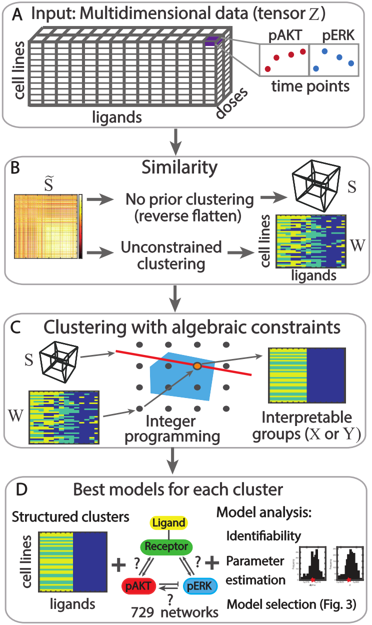


Rectangular clusters





Overview of method



Multi-indexed data $\mathbf{Z}$: In this example $\mathbf{Z} \in \mathbb{R}^{36\times 14\times 2\times 3\times 2}$.

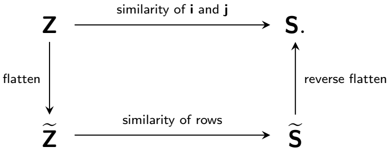
Flattened tensor: $\widetilde{\mathbf{Z}}$. In this example $\widetilde{\mathbf{Z}} \in \mathbb{R}^{504\times 12}$.

Similarity matrix: $\widetilde{\mathbf{S}}$ between the rows of $\widetilde{\mathbf{Z}}$. Here $\widetilde{\mathbf{S}} \in \mathbb{R}^{504\times 504}$.

Similarity tensor: The similarity of the data indexed by $\mathbf{i} = (i_1, i_2)$ and $\mathbf{j} = (j_1, j_2)$:

$$s_{\mathbf{i},\mathbf{j}} = \mathrm{sim}\left(\mathbf{Z}(i_1, i_2, :, \ldots, :), \mathbf{Z}(j_1, j_2, :, \ldots, :)\right) \in \mathbb{R}.$$

We summarize these relationships in the following diagram:



Where **i** and **j** are the multi-indices of experiments (i.e., cell-type/ligand combinations).

Structured clustering

Given $\mathbf{S}$ we cluster the experiments indexed by $\mathbf{i} = (i_1, i_2)$, $\mathbf{j} = (j_1, j_2)$, where $i_1, j_1 \in \{1, \ldots, 36\}$ and $i_2, j_2 \in \{1, \ldots, 14\}$.

Partition is encoded in a $(36 \times 14) \times (36 \times 14)$ tensor $\mathbf{X}$ with entries

$$x_{\mathbf{ij}} = \begin{cases} 0 & \text{if } \mathbf{i} \text{ and } \mathbf{j} \text{ belong to the same cluster,} \\ 1 & \text{otherwise,} \end{cases}$$

that are a coarse approximation of the "distance" between $\mathbf{i}$ and $\mathbf{j}$. A valid assignment must fulfil

$$\begin{aligned} \text{Reflexivity:} \quad & x_{\mathbf{ii}} = 0, \\ \text{Symmetry:} \quad & x_{\mathbf{ij}} = x_{\mathbf{ji}}, \\ \text{Transitivity:} \quad & 0 \leq -x_{\mathbf{ik}} + x_{\mathbf{ij}} + x_{\mathbf{jk}} \leq 2. \end{aligned}$$

Structured clustering

The $(36 \times 14) \times m$ tensor $\mathbf{Y}$ has entries

$$y_{\mathbf{i}k} = \begin{cases} 1 & \text{if the data indexed by } \mathbf{i} \text{ belongs to cluster } k, \\ 0 & \text{otherwise.} \end{cases}$$

We require that

$$\sum_{k=1}^{m} y_{\mathbf{i}k} = 1,$$

to ensure that each data item has been assigned to exactly one cluster.

The tensors $\mathbf{X}$ and $\mathbf{Y}$ are related by equation:

$$1 - x_{\mathbf{ij}} = \sum_{k=1}^{m} y_{\mathbf{i}k} y_{\mathbf{j}k}.$$

Two implementations

Need to classify experiments $\mathbf{i}$ into rectangular clusters.

Two ways to do this:

- Starting from scratch (i.e., no previous clustering information).
- Starting from a pre-existing, non-rectangular clustering of experiments.

Two implementations

Starting from scratch:

$$\begin{aligned} \max_{\mathbf{X}} \quad & \langle \mathbf{S}, (\mathbf{1} - \mathbf{X}) \rangle + \lambda \langle \mathbf{1}, \mathbf{X} \rangle, \\ \text{subject to} \quad & b_l \leq \mathbf{V} \cdot \mathrm{vec}(\mathbf{X}) \leq b_u, \end{aligned}$$

where $\mathbf{V}$ encodes the rectangular constraints:

$$\begin{aligned} & x_{i_1 i_2 j_1 j_2} = x_{i_1 j_2 j_1 i_2}, \\ & 0 \leq x_{i_1 i_2 j_1 j_2} - x_{i_1 i_2 j_1 i_2} \leq 1, \\ & 0 \leq x_{i_1 i_2 j_1 j_2} - x_{i_1 i_2 i_1 j_2} \leq 1. \end{aligned}$$

From pre-existing clustering $\widetilde{\mathbf{Y}}$:

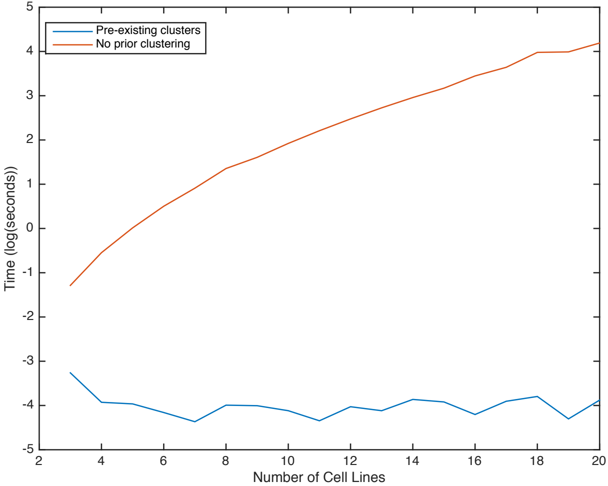
$$\max_{\mathbf{Y}} \quad \langle \widetilde{\mathbf{Y}}, \mathbf{Y} \rangle,$$

subject to

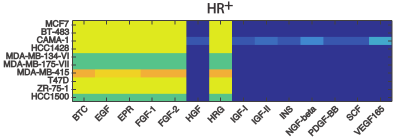
$$\begin{aligned} & \sum_{r=1}^{m} y_{ijr} = 1, \\ & -1 \leq y_{ikr} + y_{jlr} - y_{ilr} \leq 1. \end{aligned}$$

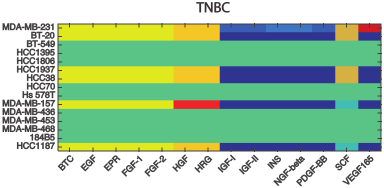
Both are integer programs that we optimise with a branch and cut algorithm.

Performance

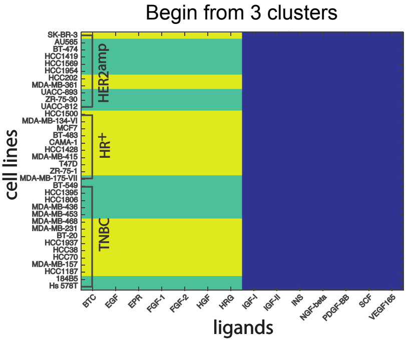


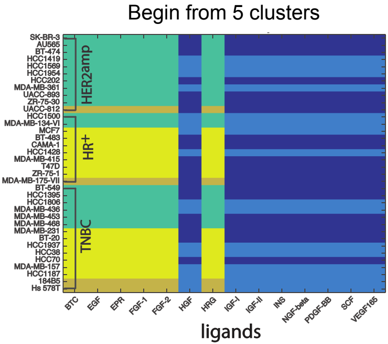
Test on HR^+ cells and Triple Negative Breast Cancer (TNBC) only.





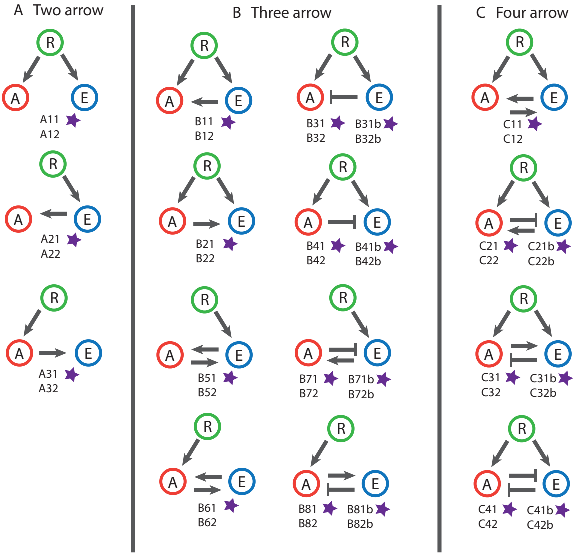
Test on all cells based starting on initial non-rectangular partitions into 3 and 5 clusters.





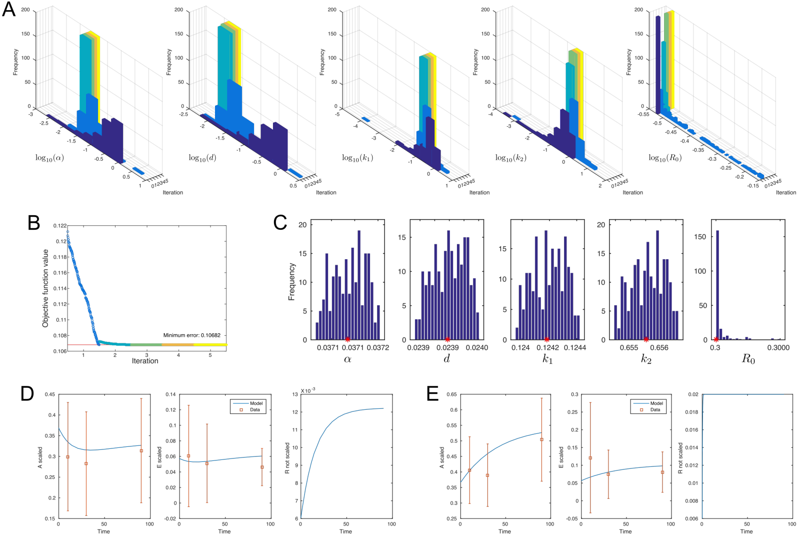
Results

Systematic search for models



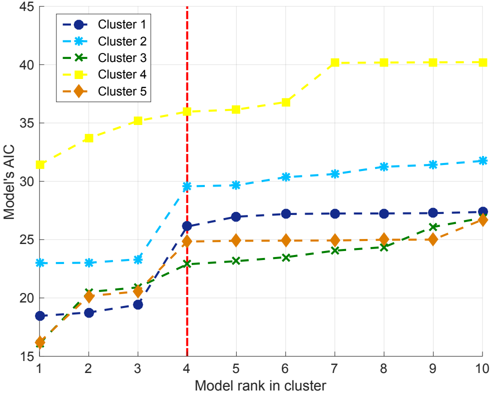
Results

Systematic search for models

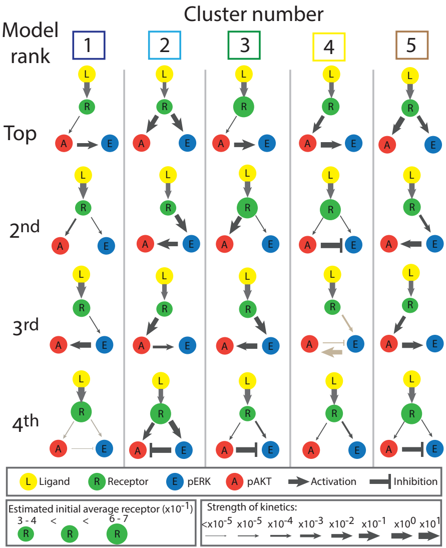


Results

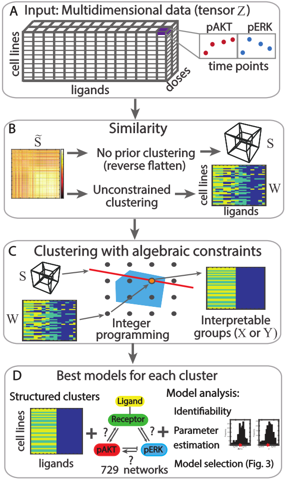
Ranking models for each cluster



Ranking models for each cluster



Recap



arXiv:1612.08116

Method for clustering multi-indexed data.

Encode interpretatibility constraints as algebraic constraints in integer program.

Clustering from scratch or find nearest compliant clustering to initial guess.

36 cell lines with 14 ligands into 5 clusters with ranking of mechanistic hypotheses.

Thank you!