Exploiting weak modularity in cancer progression to infer large Mutual Hazard Networks

Simon Pfahler[1], Leon Ernstberger[1], Peter Georg[1], Andreas Lösch[2], Rudolf Schill[3], Lars Grasedyck[4], Rainer Spang[2], Tilo Wettig[1]

[1] Department of Physics, University of Regensburg, [2] Faculty of Informatics and Data Science, University of Regensburg,

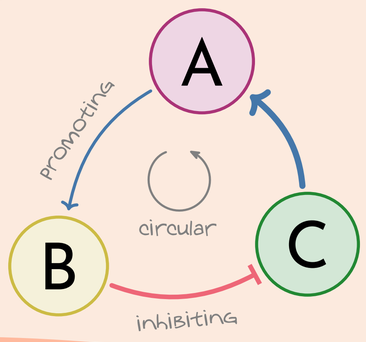
[3] Department of Biosystems and Engineering, ETH Zürich, [4] Institute for Geometry and Applied Mathematics, RWTH Aachen

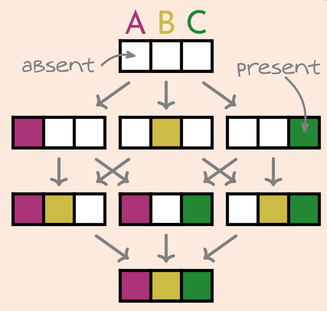
online version

Studienstiftung des deutschen Volkes

Summary

- In cancer progression, the rate of occurence of genetic events depends on the state of a tumor
- Mutual Hazard Networks infer promoting and inhibiting effects between genetic events from patient data
- Splitting the events in a dataset into clusters allows us to infer approximate MHNs for hundreds of events, overcoming a major runtime limitation of MHN
- Investigating the obtained clusters can give valuable input into the underlying biology, e.g. the role that different events play in cancer progression





Mutual Hazard Networks [1]

- Cancer progresses by accumulating genetic events
- This progression can be modeled as a Markov chain with transition rates

$$\mathbf{Q}_{x^{+i},x} = \Theta_{ii} \prod_{x_j=1}^{d} \Theta_{ij}$$

(Θ_{ii}: base rate of event i; Θ_{ij}: influence of event j on event i; $\mathbf{Q}_{x^{+i},x}$: rate to acquire event i)

- Patient data of observed tumors define a probability distribution $\mathbf{p}_{\mathcal{D}}$

→ Parameters Θ can be infered by comparing to the time-marginalized probability distribution

$$\mathbf{p}_{\Theta} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{p}_0$$

($\mathbf{p}_0$: initial distribution contains only healthy patients)

via the log-likelihood (LL)

- Exact calculation of $\mathbf{p}_{\Theta}$ is limited to under 25 active events per patient due to runtime behavior

Clustering

- Cancer progression is widely assumed to be weakly modular [2]

→ Perform calculations on smaller clusters and combine results in the end

- To estimate Θ_{ij}, we need a cluster with < 25 events, containing both i and j
- We use hierarchical clustering [3] to obtain possibly overlapping clusters

Θ matrix

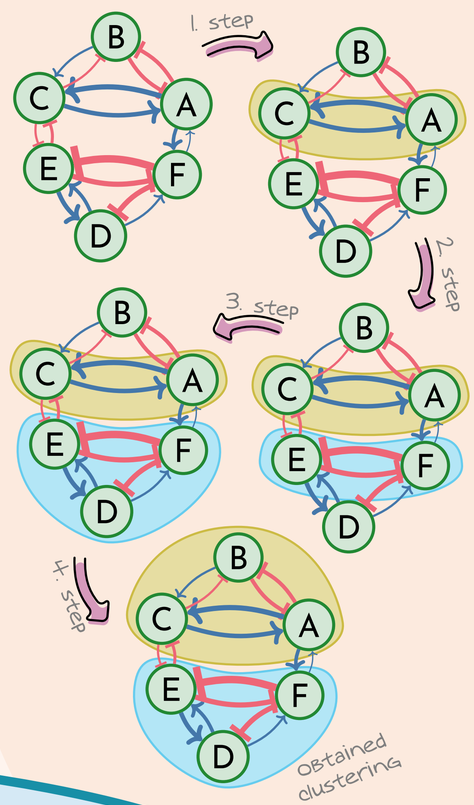
	A	B	C	D	E	F
A	2	.3	4	1	1	1.5
B	.3	1	.5	1	1	1
C	5	2	1	1	.4	1
D	1	1	1	1.5	4	.2
E	1	1	.5	3	2	.1
F	3	1	1	2	.2	1

$\max(|\log \Theta_{ij}|, |\log \Theta_{ji}|)$

Distance matrix

	A	B	C	D	E	F
A	∞	.8	.6	∞	∞	.9
B	.8	∞	1.4	∞	∞	∞
C	.6	1.4	∞	∞	1.1	∞
D	∞	∞	∞	∞	.7	.6
E	∞	∞	1.1	.7	∞	.4
F	.9	∞	∞	.6	.4	∞

Clustering algorithm



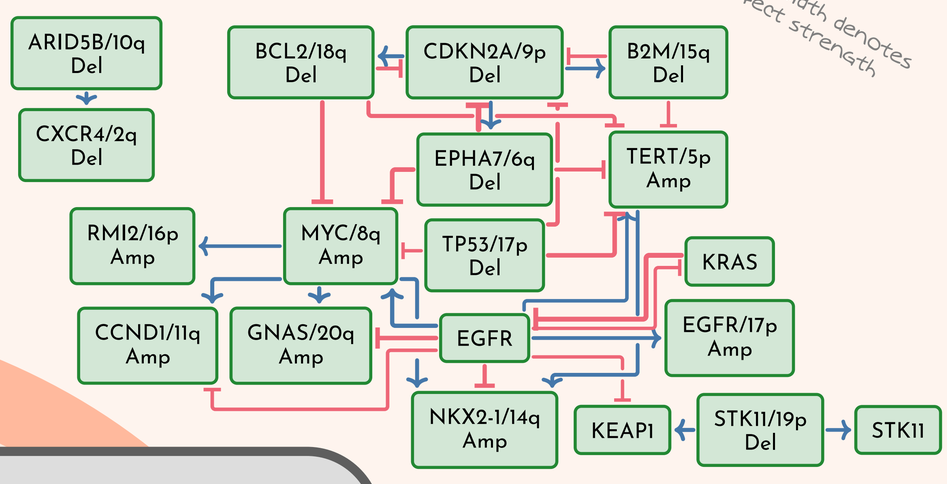
Learning Process

- To infer MHNs from patient data, we can utilize the structure found by our clustering:
  1. Start at independence model, i.e. $\Theta_{i \neq j} = 1$
  2. For every parameter Θ_{ij}: Get gradients of the LL score by considering a cluster containing events i and j
  3. Get new parameters Θ through one optimizer step
- The clusters used to calculate gradients adapt throughout the optimization process to fit the data

Biological results

MSK-CHORD [4] data of 5907 LUADs, trained on 125 genetic events

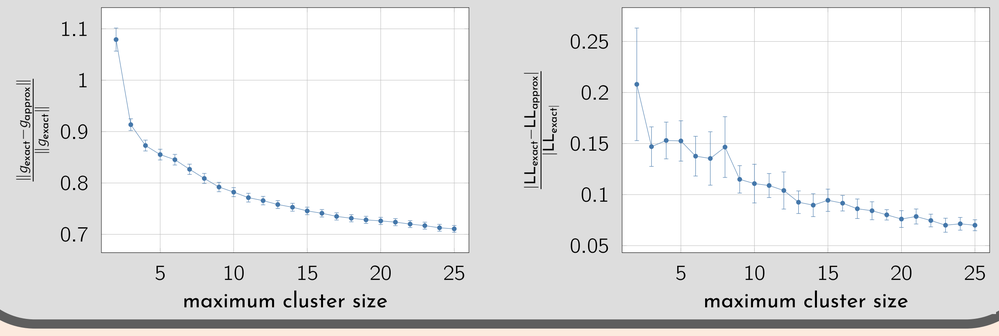
strongest 30 connections shown



Validation

To validate our method using artificial datasets and MHNs with 80 events, we check:

1. Gradient approximation accuracy
2. Accuracy of learned MHNs



Next steps

- Investigate choice of distance measure analytically and define it to minimize $\frac{\|g_{exact} - g_{approx}\|}{\|g_{exact}\|}$
- Obtain an approximation of the score along with the gradient approximation
- Consider different clustering strategies

→ Spectral clustering is of particular interest, as first tests showed promising results on graphs obtained from MHNs

- Investigate biological interpretability of clusters further

References

1. Schill, R. *et al.* in *Research in Computational Molecular Biology* (2024).
2. Iranzo, J., Gruenhagen, G., Calle-Espinosa, J. & Koonin, E. V. *Cell Reports* **40** (2022).
3. Rokach, L. & Maimon, O. in *Data Mining and Knowledge Discovery Handbook* (2005).
4. Jee, J. *et al. Nature* **636** (2024).