Supplementary Material to

Modelling Cancer Progression using Mutual Hazard Networks

Rudolf Schill[1], Stefan Solbrig[2], Tilo Wettig[2], and Rainer Spang[1]

[1]Department of Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg
[2]Department of Physics, University of Regensburg

Contents

1 Mutual Hazard Networks as stochastic approximations of Conjunctive Bayesian Networks

The functional form in eq. 2 of the main manuscript can be seen as a stochastic approximation of conjunctive dependencies. We would like to illustrate this approximation in an easy example: Consider the CBN in Fig. S1A which defines the rate of event 4 by the function

$$f_4^{\mathrm{CBN}}(\mathbf{x}) = \begin{cases} 2.50 & \text{if } x_1 = 1 \text{ and } x_2 = 1 \text{ and } x_3 = 1 \\ 0 & \text{otherwise} \end{cases}$$

We can approximate this function by the MHN in Fig. S2B. For illustration we set $\Theta_{41} = \Theta_{42} = \Theta_{43} = \sqrt[3]{2.5/\Theta_{44}}$, such that the rate of event 4 depends equally on events 1, 2 and 3, and evaluates again to 2.50 after all of them have occured:

$$f_4^{\mathrm{MHN}}(\mathbf{x}) = \Theta_{44}(\Theta_{41}^{x_1+x_2+x_3})$$

Figure S1C shows that f_4^{MHN} approximates f_4^{CBN} in the limit of $\Theta_{44} \to 0$ (although only CBNs allow strictly zero rates).

In a more complex situation, where the rate of an event does not depend equally strong on its parent events, we see a similar behaviour. For the CBN most transitions between genotypes are forbidden, while an approximating MHN assigns the highest rates to the same transitions but also allows transitions with smaller rates from states where only some parent events have occured (Fig. S2).

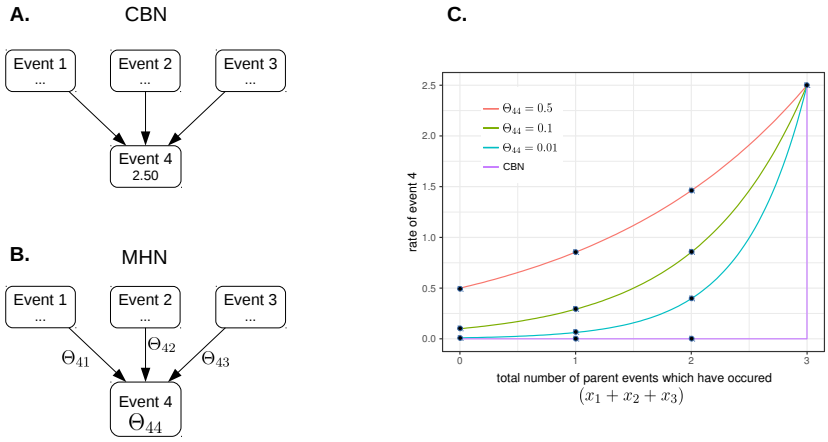

Figure S1: An example for a CBN (A) which can be approximated by an MHN (B). (C) shows the rate of event 4 as a function of the number of occured parent events. The purple curve corresponds to CBN where event 4 can only occur if all three parent events have occured. The other three curves to MHNs where event 4 can occur spontaneously with different rates Θ_{44} but becomes more and more likely the more parent events have occured. With Θ_{44} becoming smaller, the curves approximate the CBN.

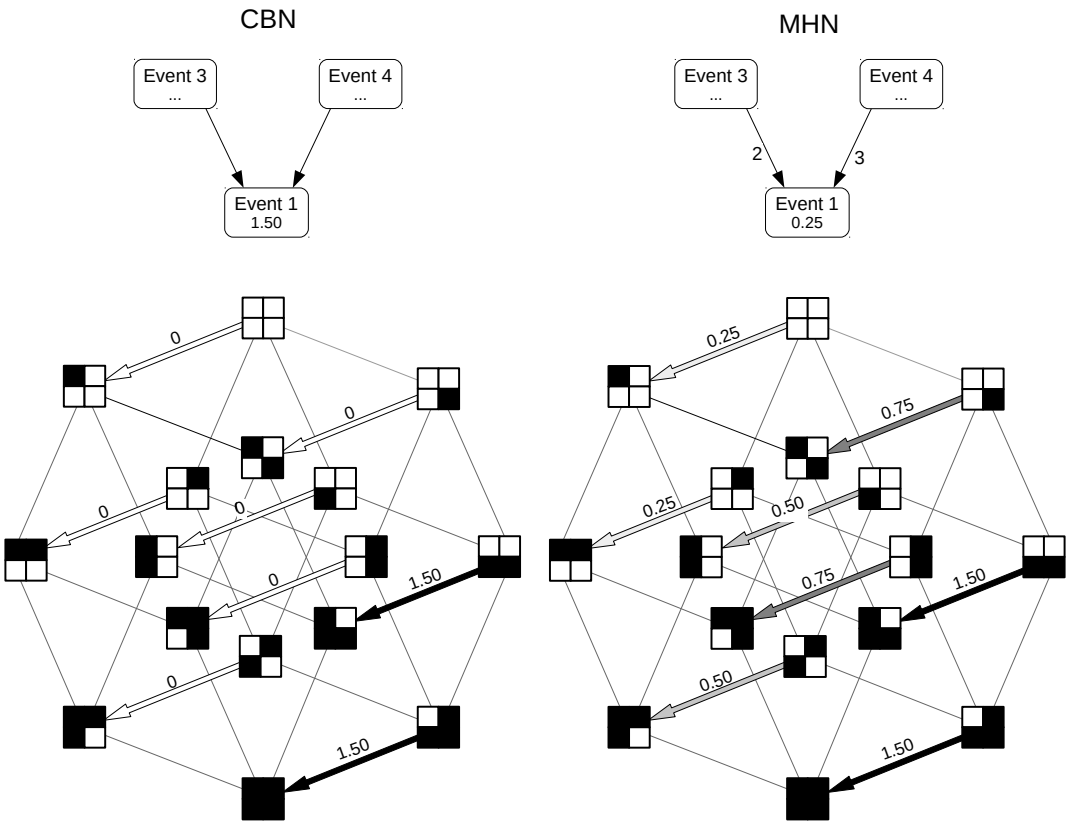

Figure S2: A CBN (left) and its approximating MHN (right) defining the Markov processes shown below them. For illustration we only show transitions that introduce event 1 to the genotype.

2 Runtime of MHN in simulation experiments

We tested the performance of our implementation in simulation experiments. MHN was written in R, and its performance-critical parts were implemented in C (using the R package `inline`) to avoid unnecessary memory-copy operations. We made explicit calls to BLAS routines and compiled R to use the Intel MKL library for vectorized and threaded matrix and vector operations. Fig. S3 shows the runtime of a single gradient step for random and dense Θ on a Dell OptiPlex 9020 workstation with 8GB RAM and an Intel® Core™ i5-4590 CPU. The runtime was about 1 minute for $n = 20$ and scaled exponentially with n as expected.

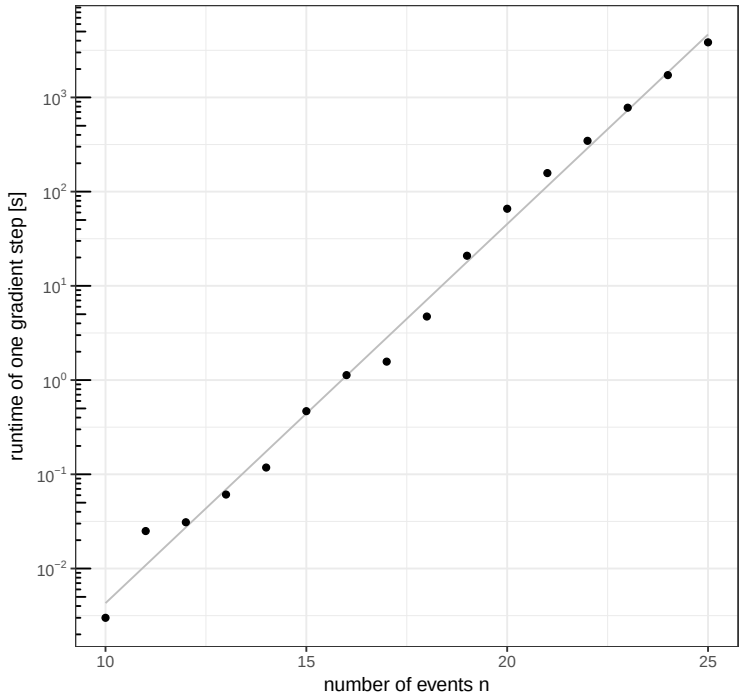

Figure S3: Runtime of a single gradient step for random and dense Θ.

3 Illustration of the Kronecker product in eq. 9

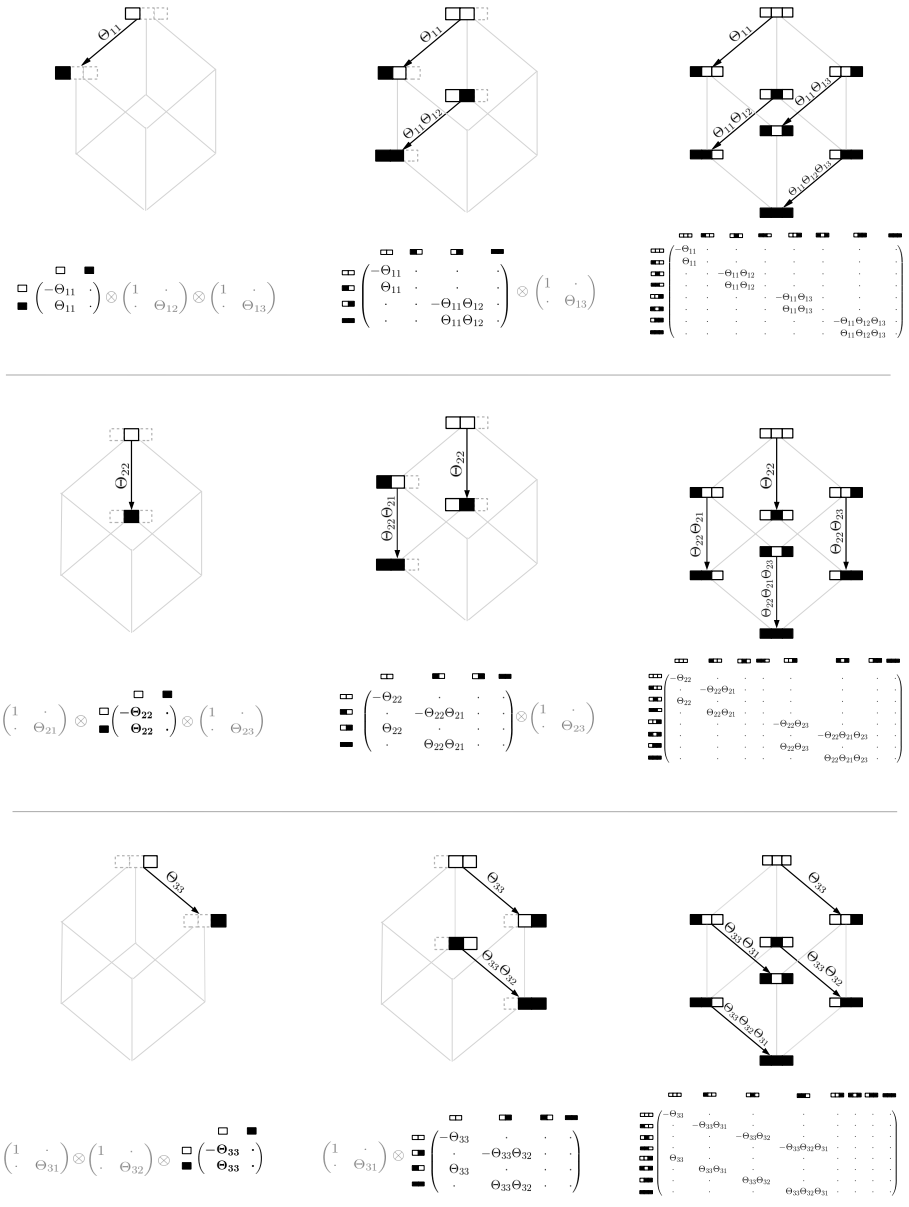

Figure S4: Illustration of Q_Θ represented as a sum of Kronecker products for $n = 3$ in equation 9. The i-th row corresponds to the i-th term in the sum and contains all transitions that introduce event i to the genome. A row is read from left to right and shows how the Kronecker product successively describes all possible transition rates that can arise due to multiplicative interactions with other events. The first highlighted Kronecker factor describes the two possible states of event i and a transition with base rate Θ_{ii}. Each subsequent Kronecker factor that is multiplied from the left or from the right appends the two states of the corresponding event j to all previously modelled states. This doubles the number of modelled states, where one half lacks the event j and retains their previous transition rates, while the other half has j present, which modulates their transition rates by the factor Θ_{ij}.

4 Comparison of Mutual Hazard Networks to Conjunctive Bayesian Networks

The following table that is also part of the main manuscript compares MHNs and CBNs on three cytogenetic cancer datasets (Baudis and Cleary, 2001) which were previously used by Gerstung *et al.* (2009).

Table 1: MHNs compare favourably to CBNs on three datasets in terms of the log-likelihood scores per tumour, averaged over 5 folds in cross-validation. They also compare favourably in terms of the AIC which penalizes the number of parameters in a model and is weighted by the sample size. While MHNs have n^2 continuous parameters, CBNs have n continuous parameters and a discrete graph structure that is hard to quantify in terms of degrees of freedom, hence we ignore the latter and bound the AIC of CBNs from below.

	cross-validated	in-sample		AIC	
dataset	MHN	CBN	MHN	CBN	MHN
Breast cancer	-5.63	-5.73	-5.54	$\geq$9373	9152
Colorectal cancer	-5.64	-5.79	-5.41	$\geq$6612	6288
Renal cell carcinoma	-5.02	-5.13	-4.81	$\geq$2587	2559

Below are plots showing (A) the raw data, (B) the learned CBNs, and (C) the corresponding MHNs. Note that the MHNs show mutual dependencies as well as inhibiting edges, features that a CBN does not have. Our observation that MHNs compare favorably in terms of both cv-log-likelihood and AIC can be interpreted as evidence that such mutual dependencies between progression events exist in all three types of cancer. The models show striking overlap as well as marked differences.

Breast cancer: Both models have the edges +8q → -8p → -11q in common (green). The tumour initiating role of +8q may arise from the oncogene MYC on chromosome arm 8q. The two models also agree in the edge +20q → -13q (blue), where the oncogene AURKA and the tumor suppressor BRCA2 are located.

However, MHN identifies mutual exclusivity between +17q and -16q (red) which the CBN cannot. Interestingly, gains at 17q, the locus of the oncogene ERBB2, are associated with a poor prognosis (Buerger *et al.*, 1999) while losses at 16q are associated with a good prognosis (Roylance *et al.*, 1999). Moreover, in the CBN the event +1q predisposes cancers to a subsequent 16q loss. The MHN model agrees that the two events are related but interprets their interplay differently: Here +17q inhibits both +1q and -16q and their association can be explained away by the absence of +17q. Hence the MHN does not see a driver event in +1q which facilitates a subsequent 16q loss and thus a favorable course of progression.

Colorectal cancer: Both models agree that +20p is an initiating event and promotes -18q (red), -4q (green), +7q (orange), and +13q (purple), likely due to genetic instability caused by the oncogene AURKA (Bischoff, 1998). They also agree in the edges +20q → +8q → -8p. CBN further identifies a subsequent major event in the loss of 18q which is the locus of the tumour suppressor SMAD4 and appears to trigger -1p, -15q, -17p and -8p. In contrast, MHN identifies -8p as an intermediate rather than terminal event which promotes both -18q and -15q, thereby explaining away their

positive association. Compared to the CBN this interpretation avoids assigning the exceptionally large rate of 18.2 to the event -8p, which would mean that -8p occurs immediately after its parent.

Renal cell carcinoma: Both models identify -3p as an initiating event which is the locus of VHL, a tumour suppressor which regulates the hypoxia response pathway (Cowey and Rathmell, 2009) and plays a known initiating role in RCC (Gnarra *et al.*, 1994). The models also agree that -3p promotes -6p. While both models find that -4q is related to -4p, -6q, -13q and +17q, these edges point away from -4q in the CBN and point towards -4q in the MHN. Similarly, CBN finds that +17q promotes +17p, while MHN finds that +17p promotes +17q.

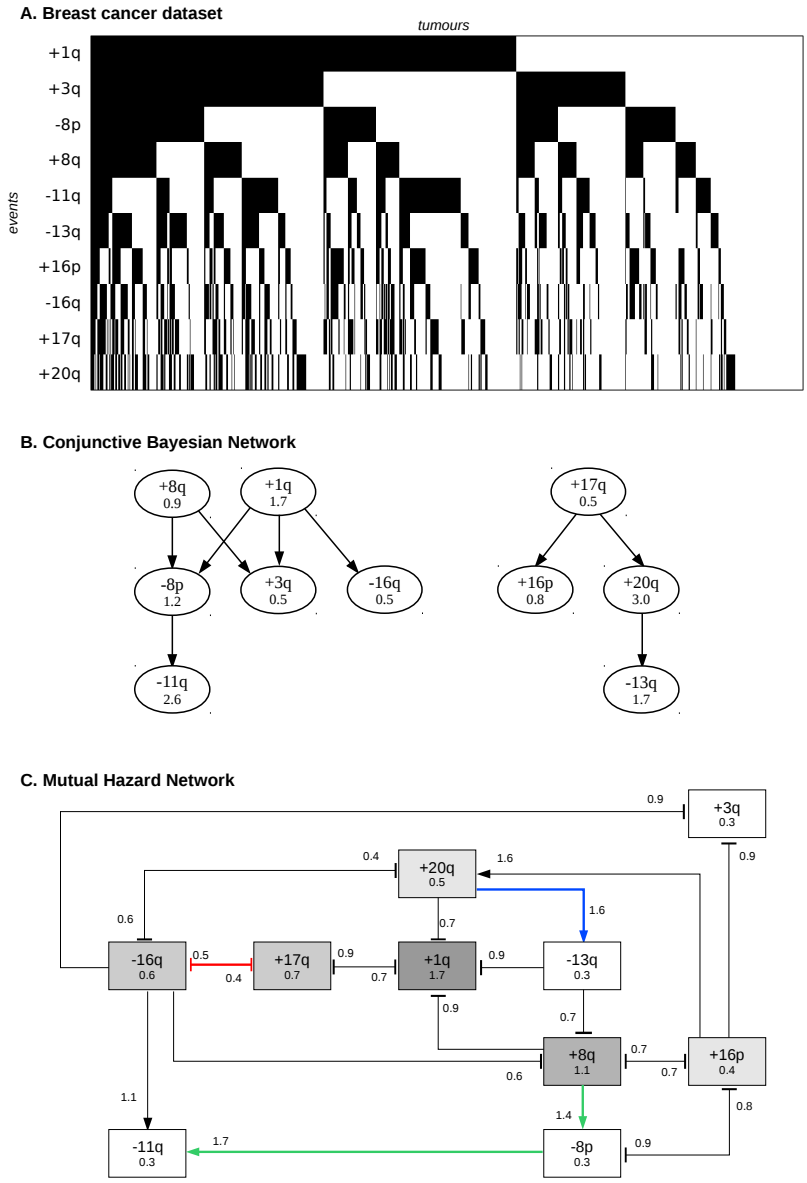

Figure S5: Breast cancer: (A) shows the raw data, where rows are copy number alterations and are sorted by frequency, while columns are event tumours whose 0/1-patterns are sorted lexicographically. (B) shows the CBN estimated from this data, where edges denote that all parent alterations must have occured before the child alteration can. Afterwards this alteration happens with the rate annotated in the corresponding node. (C) shows the MHN estimated from this data, where alterations initially happen with rate annotated in the corresponding node. Once an alteration has occured, it multiplies the rate of other events by the factor annotated on the edges.

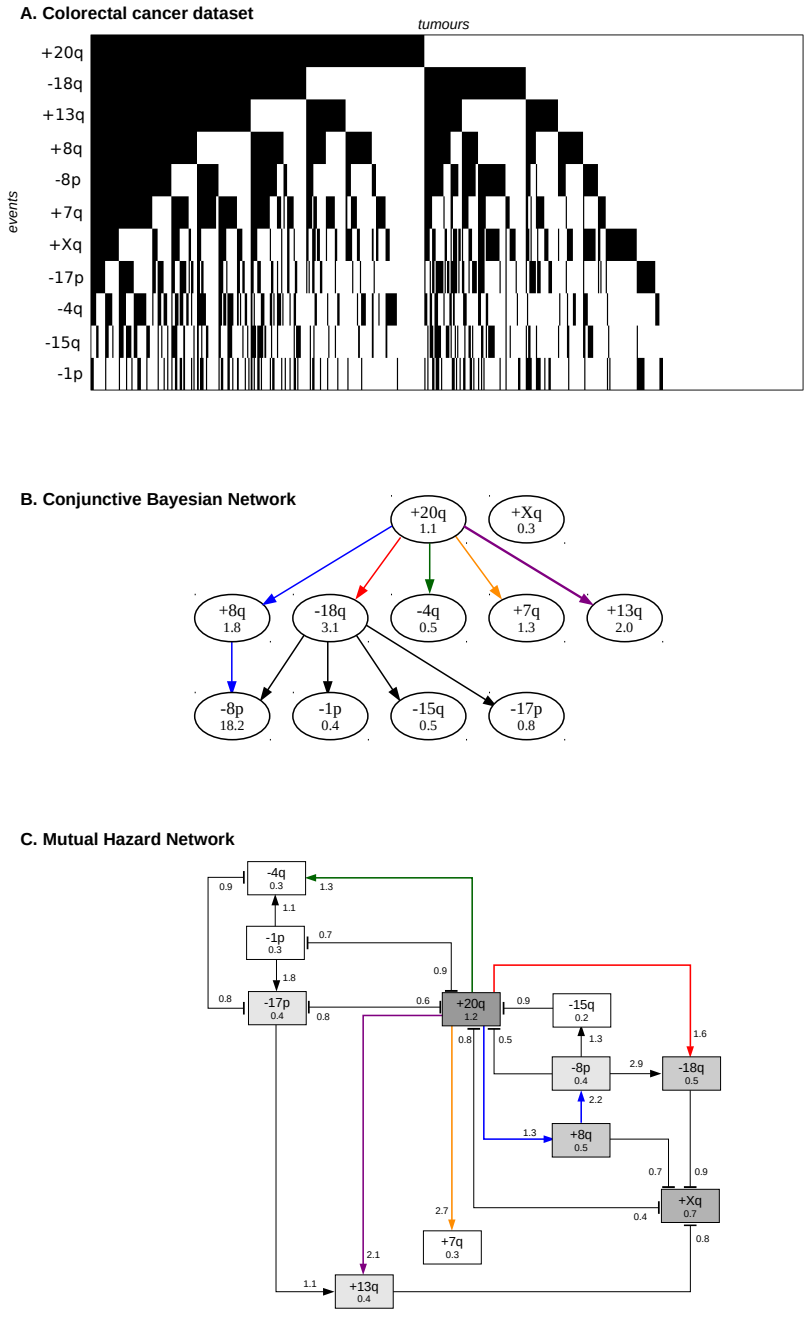

Figure S6: Colorectal cancer: (A) shows the raw data, where rows are copy number alterations and are sorted by frequency, while columns are event tumours whose 0/1-patterns are sorted lexicographically. (B) shows the CBN estimated from this data, where edges denote that all parent alterations must have occured before the child alteration can. Afterwards this alteration happens with the rate annotated in the corresponding node. (C) shows the MHN estimated from this data, where alterations initially happen with rate annotated in the corresponding node. Once an alteration has occured, it multiplies the rate of other events by the factor annotated on the edges.

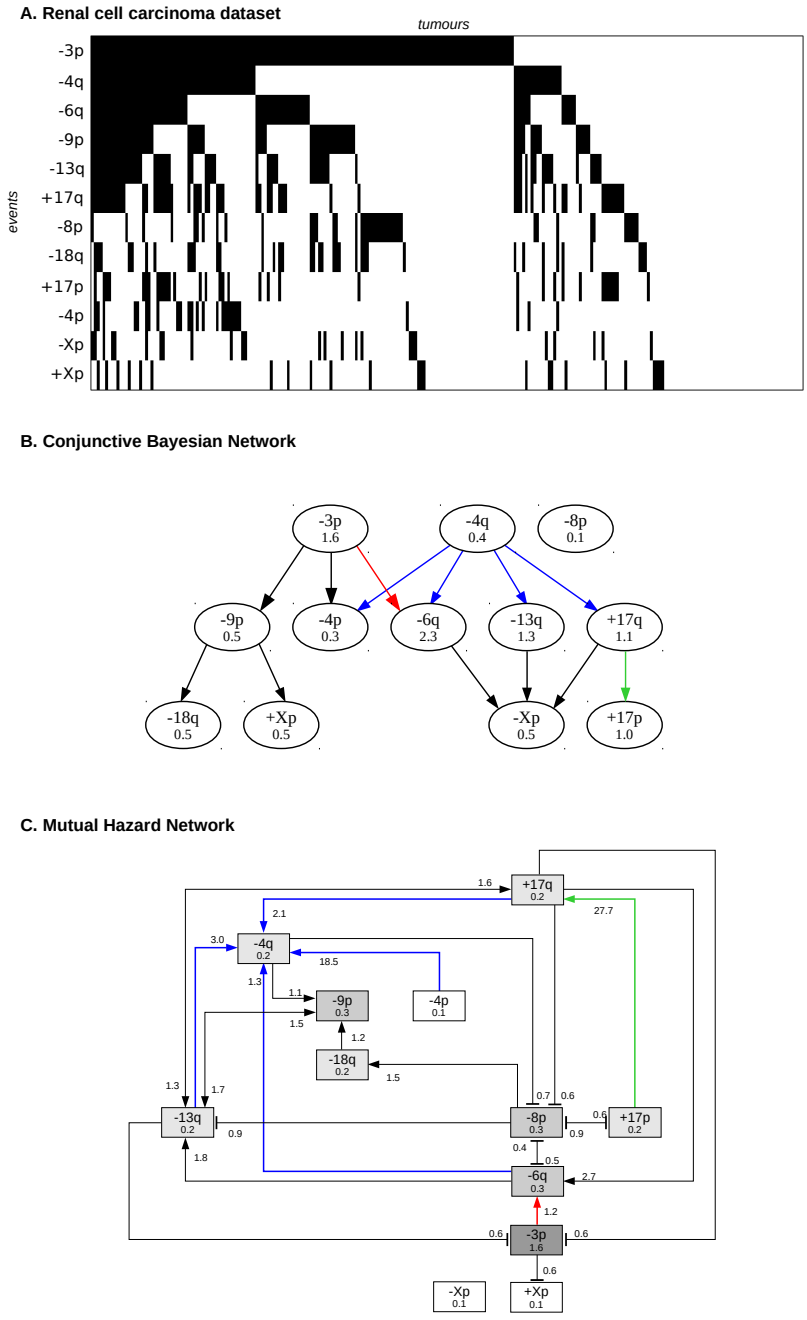

Figure S7: Renal cell carcinoma: (A) shows the raw data, where rows are copy number alterations and are sorted by frequency, while columns are event tumours whose 0/1-patterns are sorted lexicographically. (B) shows the CBN estimated from this data, where edges denote that all parent alterations must have occured before the child alteration can. Afterwards this alteration happens with the rate annotated in the corresponding node. (C) shows the MHN estimated from this data, where alterations initially happen with rate annotated in the corresponding node. Once an alteration has occured, it multiplies the rate of other events by the factor annotated on the edges.

References

Baudis, M. and Cleary, M. L. (2001). Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, **17 12**, 1228–9.

Bischoff, J. R. (1998). A homologue of Drosophila aurora kinase is oncogenic and amplified in human colorectal cancers. *The EMBO Journal*, **17**(11), 3052–3065.

Buerger, H., Otterbach, F., Simon, R., Schfer, K.-L., Poremba, C., Diallo, R., Brinkschmidt, C., Dockhorn-Dworniczak, B., and Boecker, W. (1999). Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes. *The Journal of Pathology*, **189**(4), 521–526.

Cowey, C. L. and Rathmell, W. K. (2009). VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy. *Curr Oncol Rep*, **11**(2), 94–101.

Gerstung, M., Baudis, M., Moch, H., *et al.* (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25**(21), 2809–2815.

Gnarra, J., Tory, K., Weng, Y., Schmidt, L., Wei, M., Li, H., Latif, F., Liu, S., Chen, F., Duh, F.-M., Lubensky, I., Duan, D., Florence, C., Pozzatti, R., Walther, M. M., Bander, N., Grossman, H., Brauch, H., Pomer, S., Brooks, J., Isaacs, W., Lerman, M., Zbar, B., and Linehan, W. (1994). Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nature Genetics*, **7**(1), 85–90.

Roylance, R., Gorman, P., Harris, W., Liebmann, R., Barnes, D., Hanby, A., and Sheer, D. (1999). Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. *Cancer Res.*, **59**(7), 1433–1436.