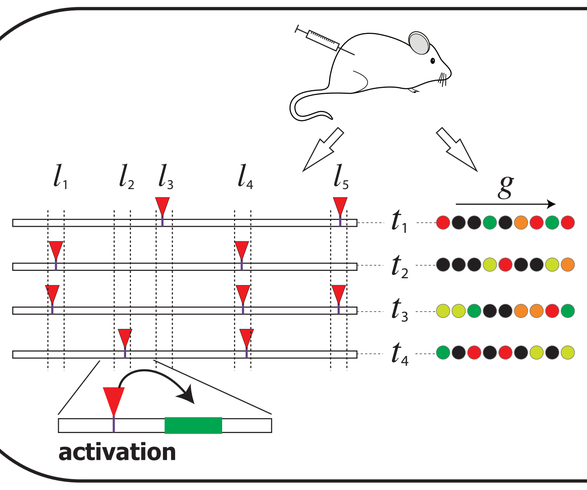
Inference of logic networks
from insertion and expression data

Jeroen de Ridder[1,2,*], Marcel Reinders[1] and Lodewyk Wessels[1,2]

*J.deRidder@TUDelft.nl

[1]Information & Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands. [2]Division of Molecular Biology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands

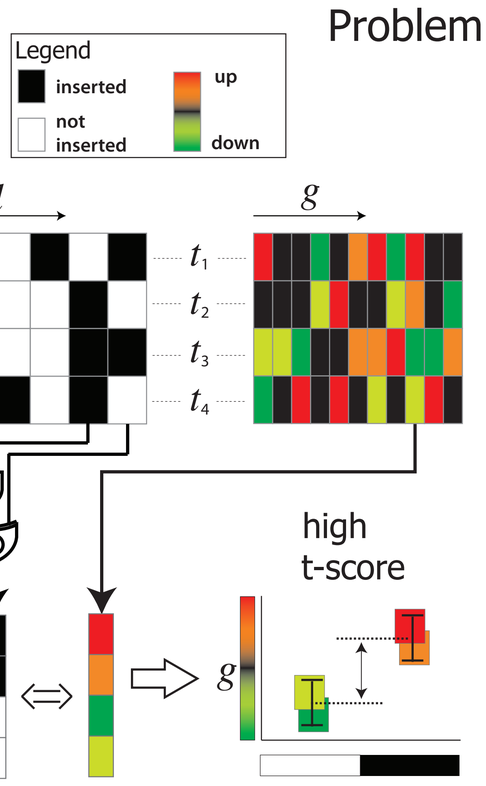


Abstract

In this study, 43 tumors, that were induced by retroviral insertional mutagenesis, are profiled, resulting in a dataset for which both the initiating events (the viral integration sites) as well as the consequent expression profiles are available.

We infer associations between insertion loci and gene expression profiles, while explicitly incorporating simple boolean logic, modelling multiple and parallel oncogenic pathways. We show that this results in the discovery of interesting causal associations between virally inserted loci and differentially expressed genes in tumorigenesis.

Problem description



Virally induced tumors are harvested and expression profiles as well as insertion profiles are determined

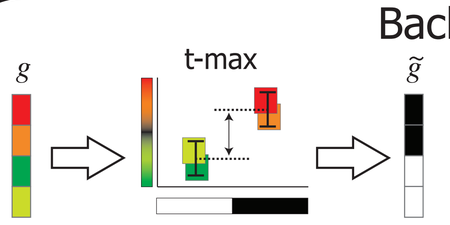
Associations between individual insertion loci and gene expression profiles may not be identifiable. This can be explained by the fact that often multiple viral triggers are required for tumorigenesis. Alternatively, parallel pathways lead to multiple possibilities in which viral triggers can disrupt healthy cell proliferation.

Therefore, we infer association between small logic networks consisting of AND and xor functions.

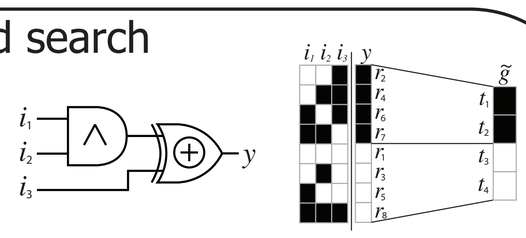
Association is measured by means of the t-score.

Since the data consists of 112 insertion loci in 43 tumors, and expression measurements for all genes, a challenging search task results. Therefore, we propose a backward branch-and-bound search procedure.

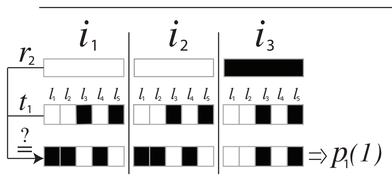
Backward search



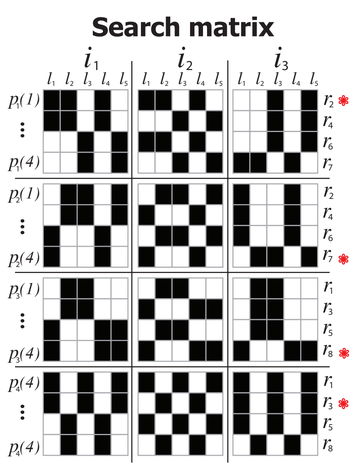
The optimal binary representation of the continuous expression vector is determined by maximizing the t-score.



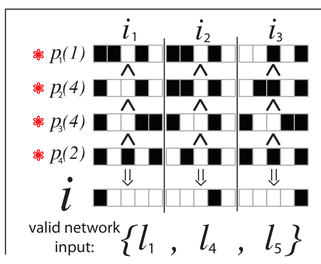
Logic network under investigation and its truth table. The rows in the truth table with output '1' (black) apply to the tumors for which g̃ equals '1', and vice versa.



For each row in the truth table it can be determined which loci need to be selected as input to the logic network in order to reach the required output for a certain tumor. For t_1, row r_2 requires the inputs i_1 and i_2 to be either l_1, l_2 and l_4 input i_3 to be l_3 or l_5.



The search matrix, that results from the this procedure, contains, for each tumor separately, all possible input combinations for which the logic network reaches the required output specified in g̃. To obtain an input combination valid for more than one tumor, the AND across one row per tumor is taken. For instance, the AND across the starred rows, gives an input combination valid for all four tumors (see below).



A valid combination of inputs is obtained when, for all inputs, the AND across a selection of rows in the search matrix results in at least one locus that is selected. For the starred rows, this is true, since: i_1=l_1, i_2=l_4, and i_3=l_5.

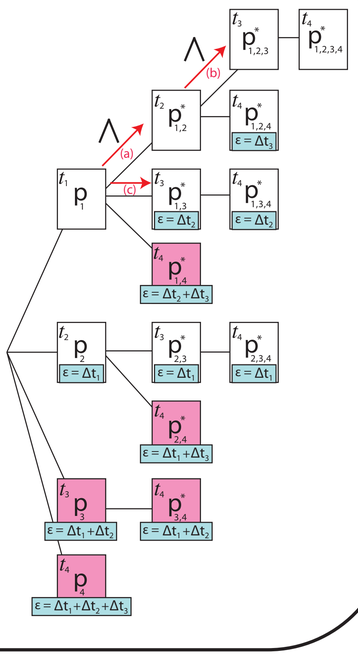
Branch-and-bound

It is desirable to allow some mismatch between g̃ (the binary version of the gene expression) and the output of the logic network. Therefore we introduce Δt, that, for each tumor, measures the effect of a mismatch on the total t-score.

For this search problem a branch-and-bound search strategy is employed.

p_n contains the rows from the search matrix corresponding to tumor n. In step (a) the AND of all possible combinations of rows in p_1 and p_2 is taken. This result is pruned, removing invalid, and duplicate results (indicated by $p^*_{1,2}$). In step (b) the same is done for all possible combinations of rows from $p^*_{1,2}$ and p3. In case no valid solutions result (e.g. $p^*_{1,2}$ is empty), the search continues directly to p_3 (step (c)).

Whole branches in the search tree can be abandoned, by requiring the error to be below a certain tolerance level. (Exemplified by the branches shaded in red).

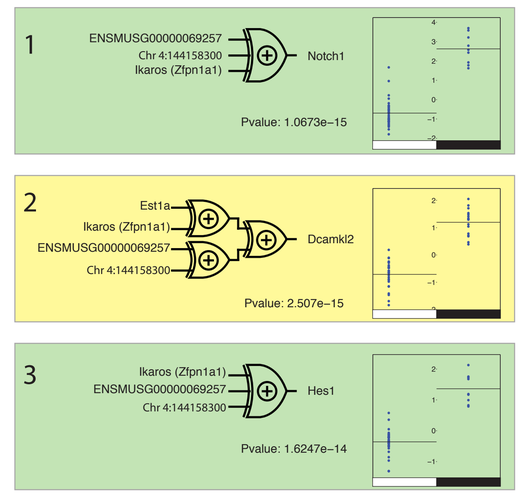


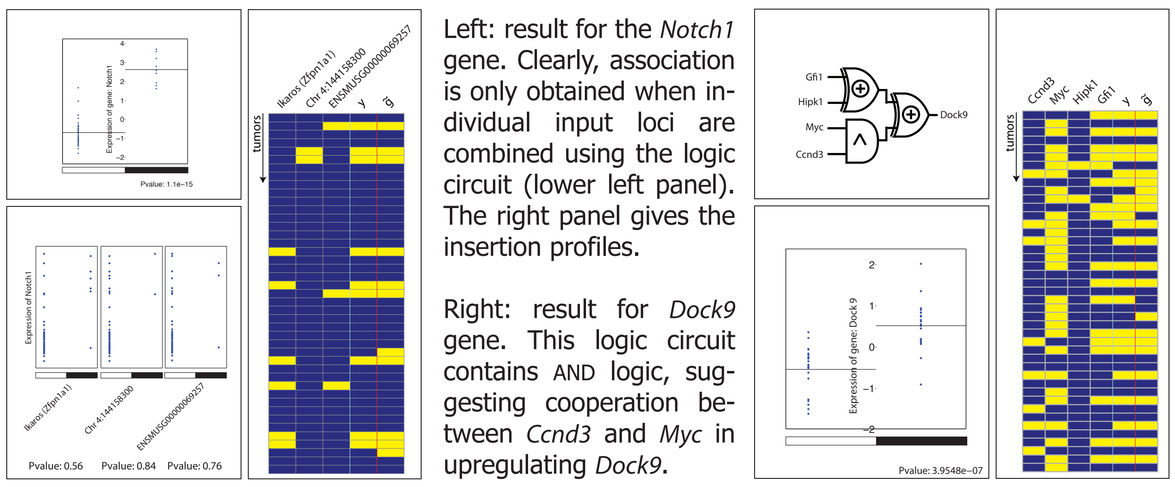
Conclusions

We evaluated 17 candidate logic networks for the 1000 genes with highest variance across the tumors. This resulted in 55 genes for which the expression profiles associated with a logic network of the insertion loci with p-value < 10^{-6}.

On the right the top 3 networks are given. For all three networks it holds that a hit in any of the input loci is sufficient to upregulate the expression of the target gene.

In conclusion, by integrally analyzing the combination of insertion data and expression data from the same tumor, it is possible to gain insight in the mechanisms in which viral triggers act on their downstream targets, and, as a result, discover interesting genes for further study.





Left: result for the *Notch1* gene. Clearly, association is only obtained when individual input loci are combined using the logic circuit (lower left panel). The right panel gives the insertion profiles.

Right: result for *Dock9* gene. This logic circuit contains AND logic, suggesting cooperation between *Ccnd3* and *Myc* in upregulating *Dock9*.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

nbic Netherlands Bioinformatics Centre

TU Delft
Delft University of Technology