Understanding Myelome Multiple mechanisms by automatic reasoning on an integrated model of transcriptomic data and large-scale signaling pathways

Bertrand Miannay[1,2], Florence Magrangeas[2], Olivier Roux[1], Stephane Minvielle[2], and Carito Guziolowski[1]

[1] École Centrale de Nantes, IRCCyN UMR CNRS 6597, Nantes, France.

[2] Centre de recherche en cancérologie Nantes-Angers, CRCNA, UMR 892 INSERM- 6299 CNRS, Nantes, France

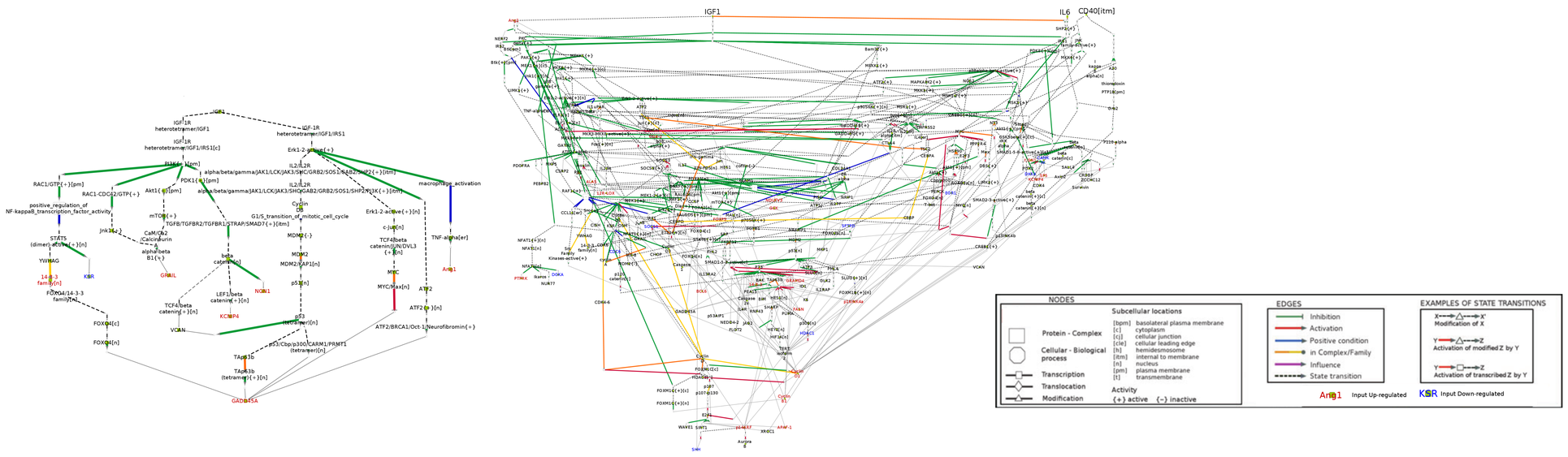

FIGURE 1 : Example of graphs extracted from PID

Biological context

Multiple Myeloma (MM) is an incurable haematological malignancy cancer ; representing 1% of cancer, and 2% of the associated mortality. This disease is particularly aggressive and the current treatments can't ensure the survival of patients.

Our aim is to better understand mechanisms of dysregulation phenomena by comparing cancer cells expression profiles with the regulation network of a normal cell.

In this work, we studied gene expression data from 611 individuals (9 healthy donors, 602 MM) and their consistency with a large-scale causal network of signaling and transcriptional events[1]

Constraint based model : Confrontation between the data and the regulatory network [2]

- Data : over-expressed (+), under-expressed (-), invariant (0)
- Regulatory network : a simple and signed directed graph {+,-}
- Experimental data : for each node, three possible colors : +, -, 0
- Consistency rule : each node variation (+ or -) has to be explained by, at least, one of its predecessor variation

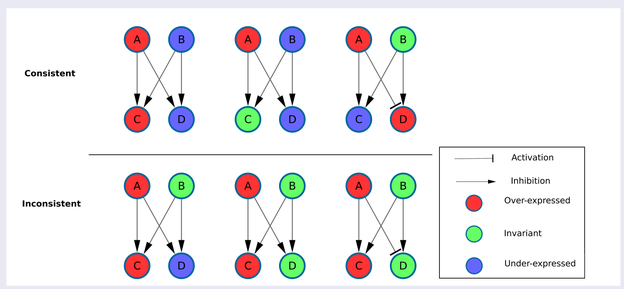

FIGURE 2 : Examples of consistency and inconsistency between the graph and the data

- If network is inconsistent with experimenta data, then correction by adding Minimal Correction Sets (MCOS) [3]

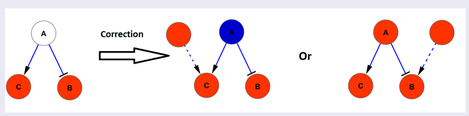

FIGURE 3 : Example of MCOS and predictions with two colorations

Methods and tools

- Data discretization
  - Discretization with a two fold relative to healthy donors for variant genes (over/under-expressed)
  - Discretization with a 0.02 fold for invariant genes
- Graph
  - Source : Pathway Interaction Database (PID-NCI) [4]
  - Subgraph extraction,downstream events from three proteins (IL6, IGF1 et CD40) to all variants genes by the shortests paths
  - Compaction of the subgraph to reduce the order while maintaining dependencies
- Prediction of nodes' signs
  - Implementation : Constraint logic programming (Answer Set Programming) : Iggy [a]
  - Exhaustive exploration of all graph coloring (associate a sign to a node in the graph)
  - Determination of the MCOS to restore the consistency
  - Intersection of all consistent graphs coloring
  - Prediction of the signs of each node in the graph with the observed data (signed genes) : +, -, 0, Change {+,-}, not+ {0,-} or not- {+,0}
- Predictions' validation :
  - 50% of measured genes {+,-,0} used to predict the other half for each individual
  - Comparison between measured and predicted data
  - Computation of the prediction's precision : $Precision = \frac{|True\ prediction|}{|all\ predictions|}$
  - Comparison with precisions from permuted data (randomized data)

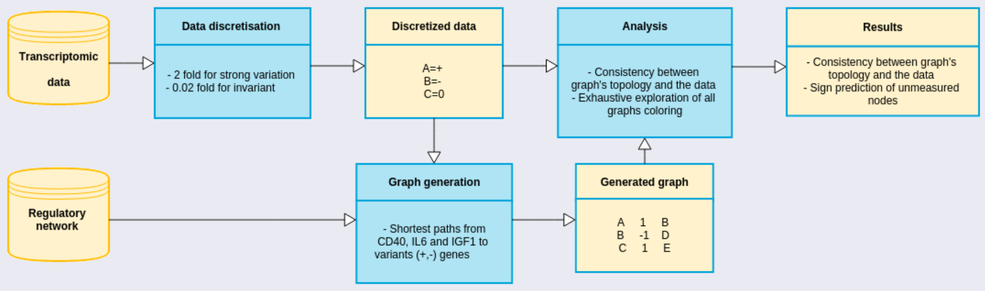

FIGURE 4 : Method Pipeline

a. Thiele, S. (2014). iggy-0.4 : a tool for consistency based analysis of influence graphs and observed systems behavior

Results & Discussions

- Discretization
  - Between 48-214 measured genes for each individual in PID
  - Identification of 17404 variant genes associated to 51596 proteins

Measured genes	before PID mapping		after PID mapping	
	healthy (9)	MM (602)	healthy (9)	MM (602)
signed "+"	12 %	27 %	13 %	31 %
signed "-"	8 %	44 %	6 %	49 %
signed "0 "	80 %	29 %	81 %	20 %
total	59507	5867413	482	60532

TABLE 1 : Proportion and number of signs for healthy and MM

- Graph
  - First generated graph with 2031 nodes and 2414 edges
  - Compacted graph with 538 nodes for 703 edges
  - It contains 476 genes of 634 in PID
  - It contains known proteins in signaling pathways : NFκB, MAP-kinase, AKT3 [5]
  - only 3 roots (IL6, IGF1 et CD40) are fixed, independently of patients → Interpatient heterogeneity is reduced
- Prediction of nodes' signs
  - Decision tree with predictions separates healthy and MM with 4 nodes : FOXA1 (in nucleus), CSF1R, VGFR1 and phosphorylated RB1-TFDP1 complex (Figure 5, left)
- Predictions' validation :
  - Precisions with real data are better than 87 to 100 % of randomized data precisions (Figure 5, right)
  - *Precisions' predictions*$_{real\ data}$ > *Precisions' predictions*$_{randomized\ data}$: *pvalue* < 10^{-57}

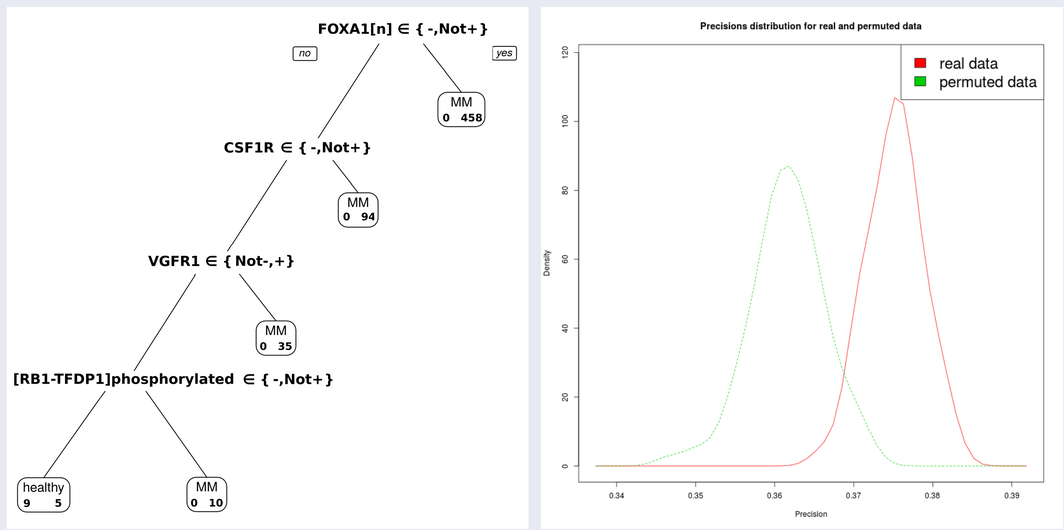

FIGURE 5 : left : Decision tree with predictions
right : Precisions distribution with real and permuted data

Conclusions

- Graph :
  - PID seems appropriate for our analysis, but it is not updated since 2012 and contains only a small portion of measured data
  - Our approach allows to connect important signaling transcription paths to variants genes [5]
- Predictions
  - Able to deduce node's sign from partial informations
  - Currently, our method is unable to identify most informative nodes (key nodes) : those that, if perturbed, will generate different predictions profiles
- Biological conclusions
  - We are able to deduce conditions of genes, proteins and complexes from transcriptomic informations
  - Identification of specific conditions of proteins for phenotypes
- Further work
  - Add more regulatory knowledge (other databases) and improve graph generation by using other paths search
  - Use classification approach to identify subtypes of MM [6]
  - Identify key nodes correlated to patient's prognosis : subset of signature nodes which are characteristic of a cancerous state or a resistance phenotype [7]

References

[1] C. Guziolowski, A. Kittas, F. Dittmann, and N. Grabe, "Automatic generation of causal networks linking growth factor stimuli to functional cell state changes.," *The FEBS journal*, vol. 279, pp. 3462–74, Sept. 2012.

[2] C. Guziolowski, A. Bourdé, F. Moreews, and A. Siegel, "BioQuali Cytoscape plugin : analysing the global consistency of regulatory networks.," *BMC genomics*, vol. 10, p. 244, Jan. 2009.

[3] I. N. Melas, R. Samaga, L. G. Alexopoulos, and S. Klamt, "Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs.," *PLoS computational biology*, vol. 9, p. e1003204, Jan. 2013.

[4] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID : the Pathway Interaction Database.," *Nucleic acids research*, vol. 37, pp. D674–9, Jan. 2009.

[5] B. Klein, "Positioning NK-kappaB in multiple myeloma.," *Blood*, vol. 115, pp. 3422–4, Apr. 2010.

[6] H. Avet-Loiseau, C. Li, F. Magrangeas, W. Gouraud, C. Charbonnel, J.-L. Harousseau, M. Attal, G. Marit, C. Mathiot, T. Facon, P. Moreau, K. C. Anderson, L. Campion, N. C. Munshi, and S. Minvielle, "Prognostic significance of copy-number alterations in multiple myeloma.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 27, pp. 4585–90, Oct. 2009.

[7] R. J. Gillies, D. Verduzco, and R. A. Gatenby, "Evolutionary dynamics of carcinogenesis and why targeted therapy does not work.," *Nature reviews. Cancer*, vol. 12, pp. 487–93, July 2012.