UNIVERSITEIT GENT

FACULTY OF **ENGINEERING AND ARCHITECTURE**

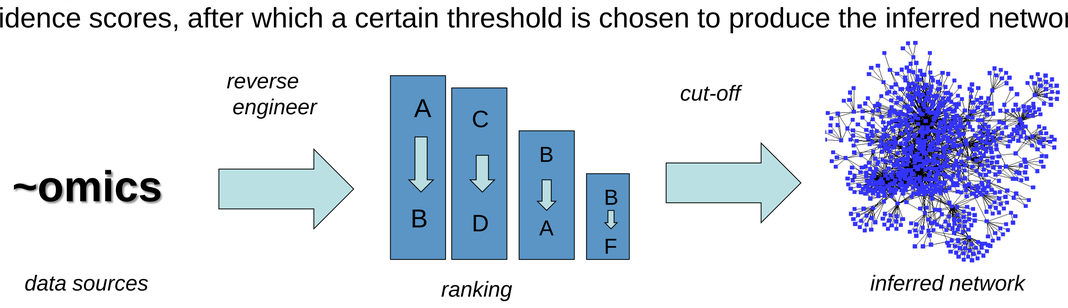
A re-ranking algorithm for gene regulatory network predictions using graphlets and graph-invariant properties

Joeri Ruyssinck, Tom Dhaene and Yvan Saeys
joeri.ruyssinck@intec.ugent.be

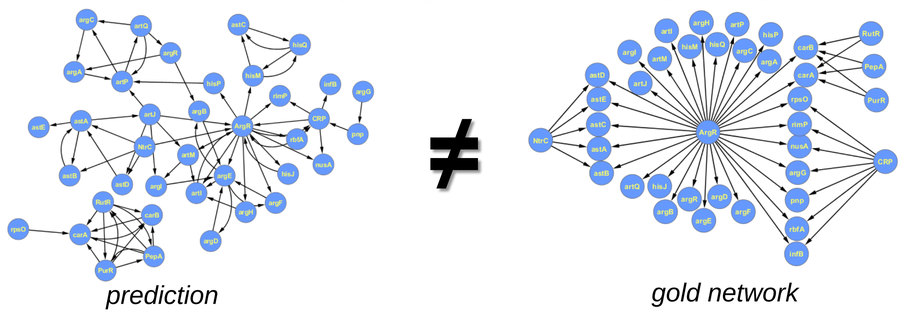
Problem statement

A vast amount of algorithms have been proposed that try to deduce the topology of large gene regulatory networks from high throughput data.

These algorithms typically produce a ranking of links between genes with associated confidence scores, after which a certain threshold is chosen to produce the inferred network.



However, the structure of the predicted network does not resemble the typical structure of a gene regulatory network, as most algorithms only take into account connections found in the data and do not include known typical graph properties of gene regulatory networks.



prediction

- Mesh topology
- Direct effects vs. indirect effects
- Cause-effect ambiguity

gold network

- Scale free topology
- Network motifs
- Modular structure

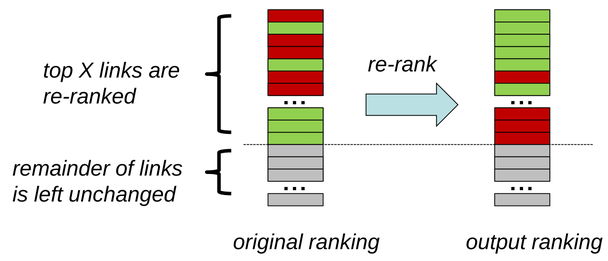
Research goal: include the known graph-invariant properties of gene regulatory networks as a prior knowledge in the inference process.

Rather than developing a new network inference algorithm which specifically takes into account several network structure properties, we opt to design a post processing algorithm which is applicable to any ranking of regulatory links.

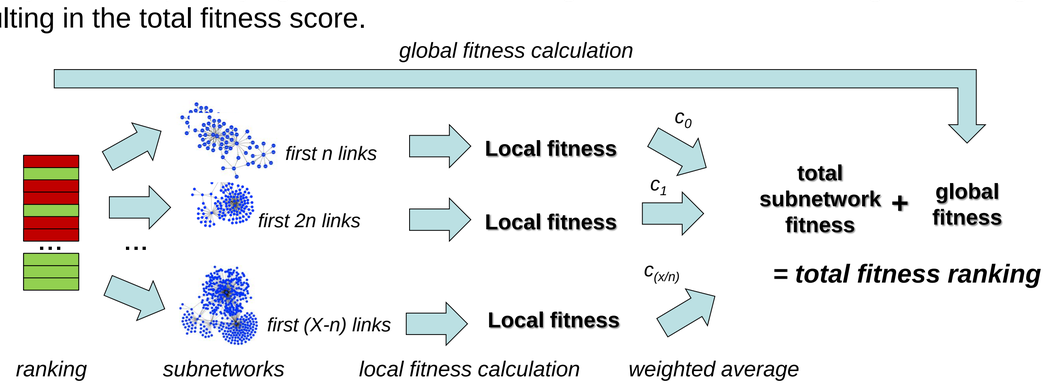
As such, our algorithm is not dependent on the data sources used to infer the network and the user can continue to use his or her preferred algorithm.

Re-ranking approach

The proposed re-ranking algorithm, named *Netter,* works as follows. The top *X* links of the ranking are extracted and will be re-ranked based on topological properties in a simulated annealing framework. The goal is to move the true positive links to the top of the ranking.



The fitness of a ranking can be calculated by first creating subnetworks of increasing size consisting of the first *n, 2n, 3n, ..., X* links. For each of these subnetworks, a local fitness score is calculated depending on several penalty functions based on graph-invariant properties. These local scores are then aggregated using a weighted average, in which the fitness score of smaller subnetwork is considered more important than those of a larger subnetwork. Finally this score is combined together with a global penalty function quantifying how much the ranking is different from the original ranking, resulting in the total fitness score.

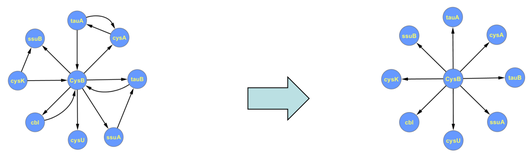


Starting from the original ranking, a new optimised ranking in the simulated annealing framework is generated by randomly moving links up or down and adjusting the subnetworks accordingly. Based on the new and old fitness score, the new ranking is either accepted or reverted and this process is repeated until the a certain amount of iterations have passed. The optimised network is stored and the entire optimising process is repeated several times.

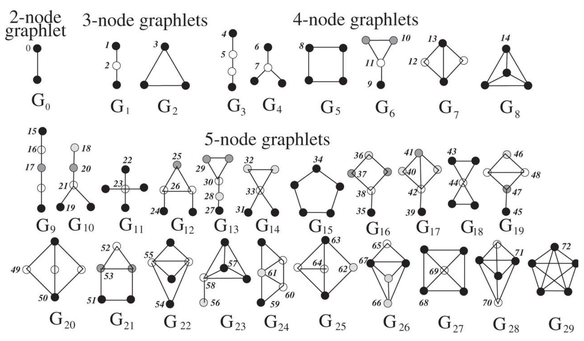
The final output ranking is created by taking the average position of each link in the optimised networks.

Structure properties

We defined several structure penalty functions that try to quantify how much a given network resembles the typical structure of a gene regulatory network. Of course such a penalty function must find balance between being informative and being a valid assumption for a wide array of gene regulatory networks. Simple penalties were developed such as discouraging the amount of bidirectional links or the amount of nodes with outgoing links.



However, more advanced penalties make use of *graphlets*. Graphlets are small connected non-isomorphic induced subgraphs of a larger network. By counting the relative occurrences of these graphlets, one can steer the prediction towards more realistic networks structures. As for example it is clear that a G_{29} graphlet should occur much less than a G_{11} graphlet.



Because of the small differences between subsequent networks in our optimization framework it was infeasible to count graphlet occurrences from scratch after each iteration. Therefore we have developed software which efficiently keeps track of the graphlet occurrences in the network using incremental updates.

Preliminary tests

First exploratory tests have been conducted using several gene regulatory network inference methods such as CLR, GENIE3 and NIMEFI which focus on inferring large networks from microarray data. Results shown were conducted on the DREAM4 InSilico Size 100 Multifactorial benchmark dataset and on randomly generated networks created using GeneNetWeaver.

We show the AUPR value of the ranking of the first 750 links compared to the new ranking obtained by applying Netter. Please note that the AUPR value is dependant on the amount of true links contained in the original prediction, as such AUPR values can not be compared between different algorithms.

	CLR				GENIE3				NIMEFI		
	Original	New	Diff		Original	New	Diff		Original	New	Diff
Dream4-1	0.34	0.23	-0.11	Dream4-1	0.40	0.26	-0.14	Dream4-1	0.35	0.19	-0.15
Dream4-2	0.26	0.42	0.16	Dream4-2	0.34	0.38	0.04	Dream4-2	0.33	0.43	0.10
Dream4-3	0.36	0.41	0.05	Dream4-3	0.49	0.51	0.01	Dream4-3	0.46	0.45	0.00
Dream4-4	0.38	0.54	0.16	Dream4-4	0.45	0.50	0.05	Dream4-4	0.40	0.46	0.06
Dream4-5	0.37	0.58	0.20	Dream4-5	0.42	0.47	0.05	Dream4-5	0.44	0.53	0.10
GNW-1	0.28	0.61	0.33	GNW-1	0.37	0.68	0.31	GNW-1	0.31	0.62	0.32
GNW-2	0.32	0.69	0.37	GNW-2	0.44	0.74	0.30	GNW-2	0.37	0.68	0.31
GNW-3	0.25	0.27	0.03	GNW-3	0.29	0.37	0.08	GNW-3	0.28	0.32	0.04
GNW-4	0.29	0.41	0.11	GNW-4	0.31	0.35	0.04	GNW-4	0.35	0.28	-0.07
GNW-5	0.20	0.24	0.04	GNW-5	0.27	0.33	0.05	GNW-5	0.28	0.23	-0.04
Average	0.30	0.44	0.13	Average	0.38	0.46	0.08	Average	0.36	0.42	0.07

Results show that in most cases our algorithm, Netter, indeed succeeds in significantly improving the top of the ranking of these algorithms. However, further tests and the development of new penalty functions are needed to validate our findings and increase the effectiveness of our algorithm.

iMinds CONNECT.INNOVATE.CREATE
INTEC IBCN
Internet Based Communication Networks and Services, IBCN • www.ibcn.intec.ugent.be
Department of Information Technology • Gaston Crommenlaan 8 bus 201, 9050 Gent, Belgium
VIB Department of Molecular Biomedical Research, Ghent University, Belgium
FWO VLAANDEREN