SFFS-MR: a floating search strategy for GRNs inference

Fabrício M. Lopes – UTFPR/Brazil
David C. Martins-Jr – UFABC/Brazil
Junior Barrera – USP/Brazil
Roberto M. Cesar-Jr – USP/Brazil

Summary

- GRN inference
- GRN inference by feature selection
- Intrinsically Multivariate Prediction
- SFFS with multiple roots (SFFS-MR)
- Experimental results
- Conclusion and future work

GRN Inference

GRN inference

- Important problem in Bioinformatics field
- Systems Biology: study of live organisms viewed as integrated and interacting networks of genes, proteins and biochemical reactions
- Emergent field of study since the advent of high-throughput technologies for extraction of gene expressions (mRNA abundances or transcripts)
  - DNA Microarrays, SAGE, RNA-Seq
  - Allow the analysis of thousands of transcripts simultaneously
- **Problem:** large number of variables (thousands) and small number of samples (dozens)

GRN inference

- The inference of GRNs from temporal expression data is a great challenge
- Several approaches for modeling and identification of GRNs
  - Boolean Networks
  - Differential equations
  - Bayesian Networks, etc...
- Probabilistic Boolean Networks (PBN) is the stochastic version of Boolean Networks, being suitable in situations with limited data samples
  - our focus in this work

GRN Inference by feature selection

GRN inference by feature selection

- Feature selection: commonly used approach to infer GRNs
- Composed by two main parts
  - search algorithm
  - criterion function
- For each gene considered as target, the best predictor set with respect to the target is chosen according to a search algorithm guided by a criterion function

GRN inference by feature selection

- Commonly used criterion functions
  - Correlation
    - Considers only 1-to-1 relationships
    - Suitable to identify co-regulation between genes, functional modules and clusters
    - Ignores multivariate (N-to-1) relationships
  - Bayesian error based
    - Non-linear Coefficient of Determination (CoD)
    - Capture N-to-1 relationships
    - Based on the maximum conditional probability of the target given the considered subset of predictors
  - Information theory based
    - Entropy, mutual information
    - Used to infer 1-to-1 and N-to-1 relationships
    - Based on the conditional probability distribution as a whole

GRN inference by feature selection

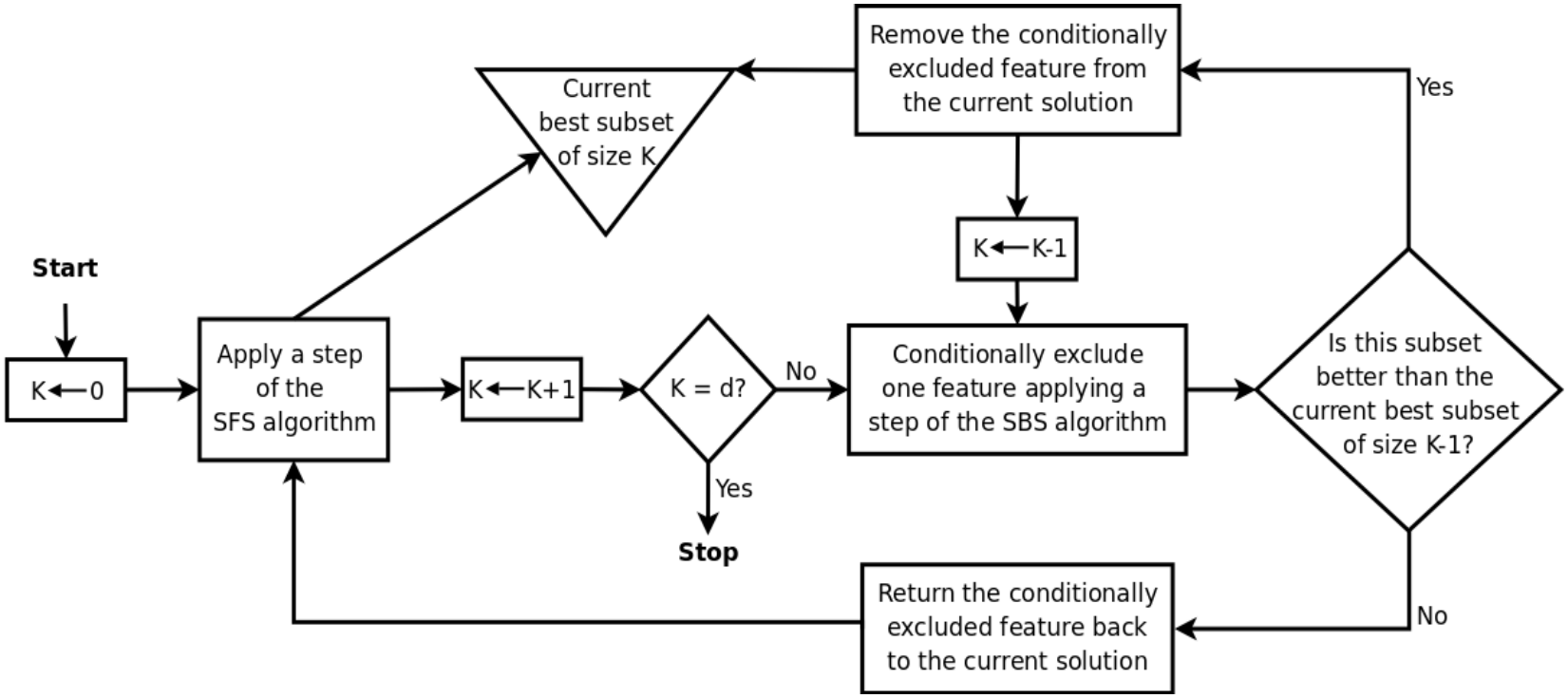
- Optimal search algorithms such as exhaustive or branch-and-bound searches return the best feature subspace
  - **Large computational cost**
  - **Unfeasible for problems with thousands of variables (such as GRN inference)**
- Sub-optimal algorithms such as Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS)
  - **Optimal solution not guaranteed**
  - Some of them present **good cost-benefit** in terms of computational cost and quality of the solution (e.g. SFFS)

GRN inference by feature selection

- SFS: genuinely greedy algorithm
  - Feature subset S starts empty
  - While a stop criterion is not satisfied
    - Include the feature X_i in the partial solution S such that $S \cup X_i$ is the best of all $S \cup X_j$ (according to a given criterion function)
  - Return S
- **Nesting effect:** inserted features are never discarded
  - It is very common to insert features that do not make part of the optimum solution

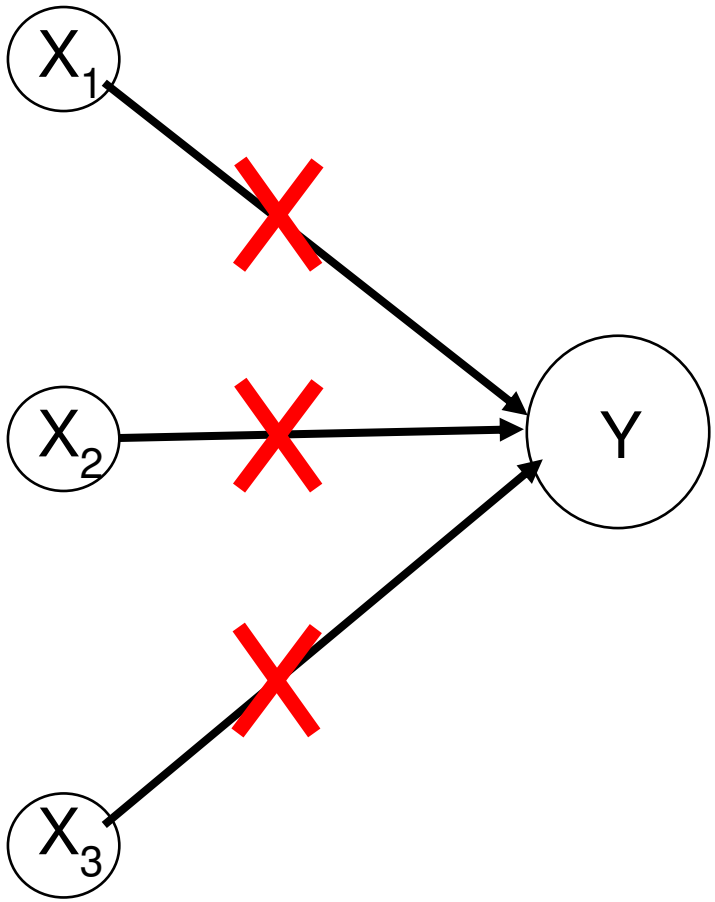
GRN inference by feature selection

- SFFS: designed to alleviate the nesting effect
  - It can add or remove one feature at each step
  - Good cost-benefit, but they do not avoid the nesting effect completely



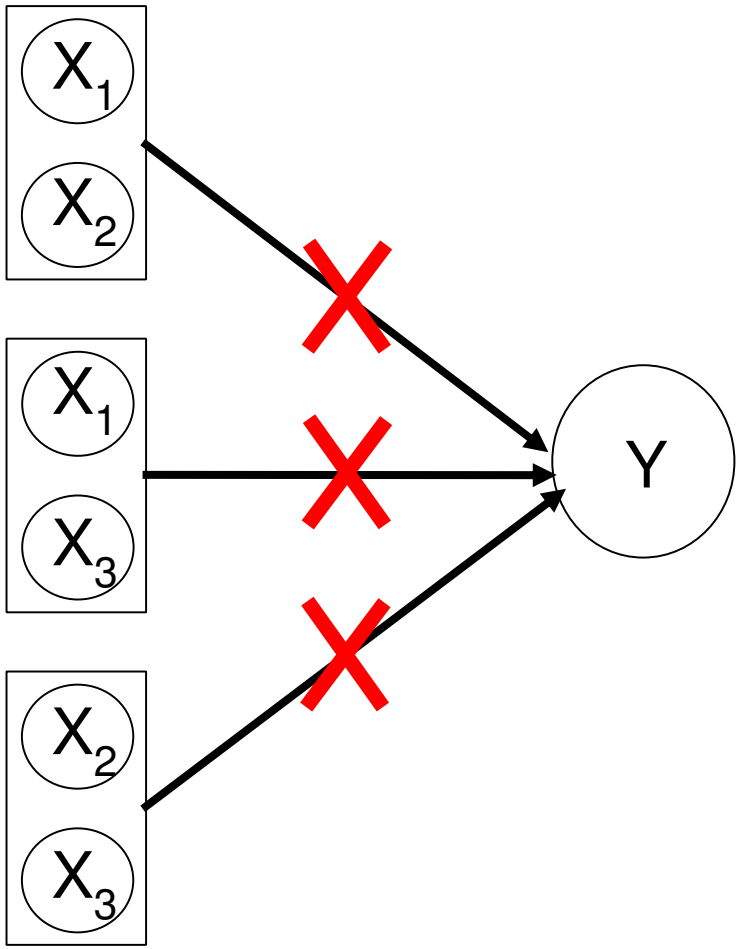
Intrinsically Multivariate Prediction

Intrinsically multivariate prediction



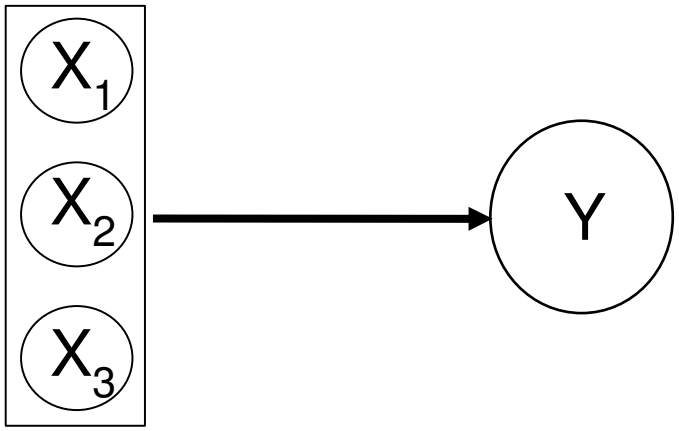
Each predictor is individually bad

Intrinsically multivariate prediction



All predictor pairs are bad

Intrinsically multivariate prediction



The triple is an excellent predictor of the target

- Y is intrinsically multivariate predicted by X1, X2, X3 ({X1, X2, X3, Y} form an IMP set)

Intrinsically multivariate prediction

- Formally, a set of features **X** is intrinsically multivariate predictive (IMP) for the target feature Y with respect to λ and δ, if:

$$\max_{\mathbf{Z} \subsetneq \mathbf{X}} \mathcal{F}_Y(\mathbf{Z}) \leq \lambda \wedge \mathcal{F}_Y(\mathbf{X}) \geq \delta$$

where *F* is a criterion function that varies from 0 to 1

  - (0 – absence of prediction; 1 – full prediction)

- The IMP score is a measure of how much IMP is a predictor set with its corresponding target. It is defined by:

$$I_Y(\mathbf{X}) = \mathcal{F}_Y(\mathbf{X}) - \max_{\mathbf{Z} \subsetneq \mathbf{X}} \mathcal{F}_Y(\mathbf{Z})$$

Intrinsically multivariate prediction

- The concept of IMP is related to the nesting effect
  - Two bad features may originate a very good pair for prediction of a given target (several heuristics would discard such features and never include them)
    - **Large IMP score**
  - Two good features may be redundant as a pair for prediction of a given target (e.g. they may be highly correlated)
    - **Small IMP score**

SFFS with Multiple Roots
SFFS-MR

SFFS with Multiple Roots (SFFS-MR)

- Why not consider an algorithm that applies a small number of SFFS executions considering good and bad initial features (roots)?
- **SFFS with Multiple Roots (SFFS-MR)**
- SFFS-MR differs from SFFS because of the exploration of multiple roots
- The B best individual features and W worst features are chosen to compose the root set. SFFS is applied for each root as initial subset of cardinality 1
- At the end, there will be B+W solutions. The best of these solutions will be returned

SFFS with Multiple Roots (SFFS-MR)

- Computational complexity
  - If B and W are small constant values compared to the total number of features, its asymptotical computational cost is not worse than SFFS

Experimental Results

Experimental Results

- Synthetic networks
- Artificial Genetic Networks (AGN) generated by considering uniformly-random Erdös-Rényi (ER) topology (random graphs)
- Probabilistic Boolean Networks (PBN) applied to generate the network dynamics
  - Temporal expression profiles
- Comparison of the results using SFS, SFFS and SFFS-MR

Experimental Results - Parameters

- Average input degree (number of predictors per node) varied from 1 to 5
- Number of observed instants of time varied from 5, 10, 15, 20 to 100 in steps of 20
- For each gene g_i of the network, its value is given by a randomly selected function from 3 possible Boolean functions $\{f_1^{(i)}, f_2^{(i)}, f_3^{(i)}\}$
  - The probabilities of each function be selected are given by $c_1^{(i)} = 0.95$, $c_2^{(i)} = 0.025$, $c_3^{(i)} = 0.025$ (quasi-deterministic setting)
- Experimental results were obtained from 50 simulations for each signal size and average input degree

Experimental Results – Algorithms settings

- Criterion function: based on mutual information (the same criterion for SFS, SFFS and SFFS-MR)
- SFFS-MR settings
  - Roots: 1 best individual feature and 5 worst individual features (6 roots in total)

Experimental Results - Assessment

- Similarity measures adopted to compare the AGN and the inferred networks
  - Positive Predictive Value (PPV)
  - Sensitivity (or recall)

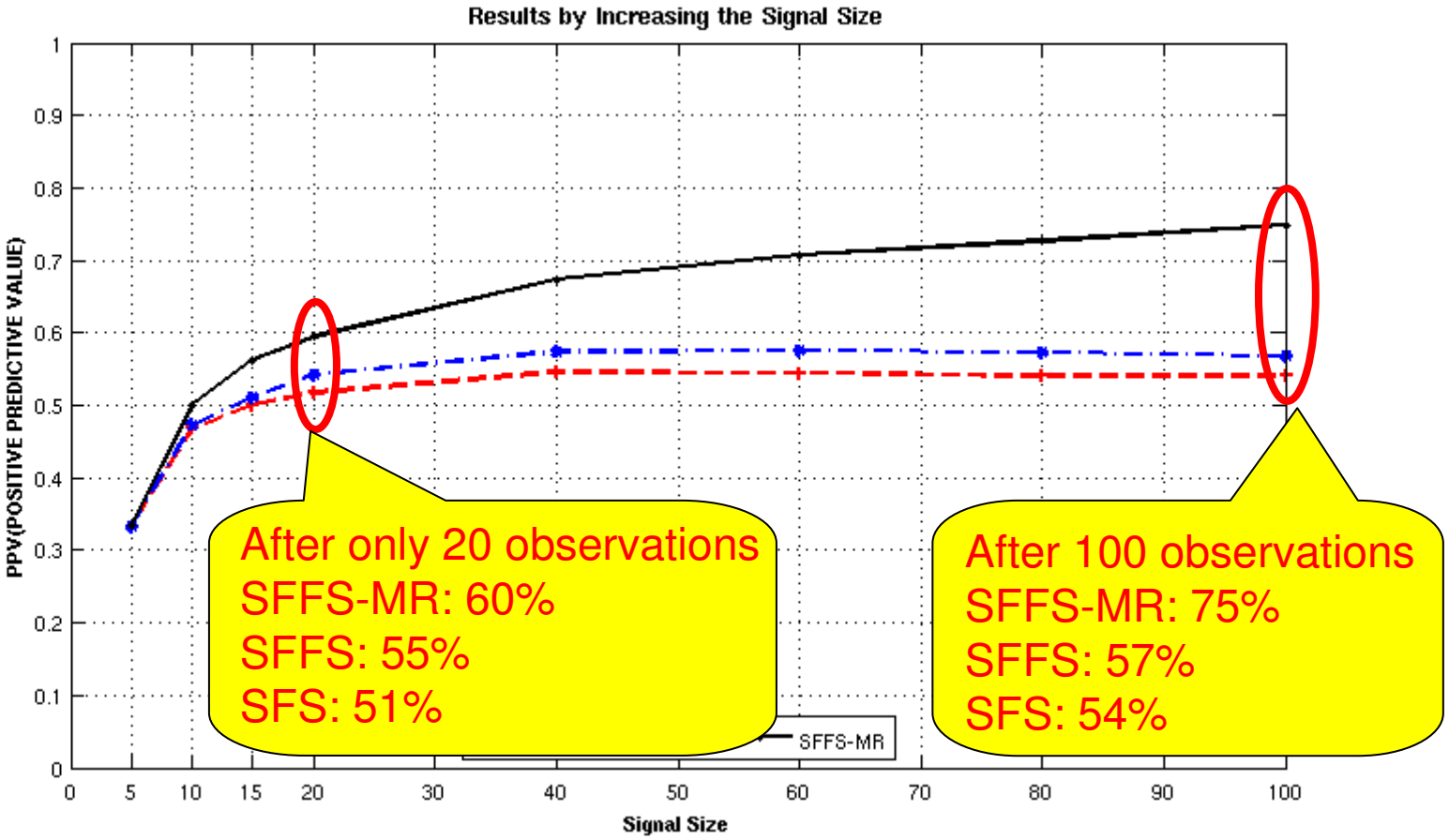
$$PPV = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Similarity} = \sqrt{PPV \times \text{Sensitivity}}$$

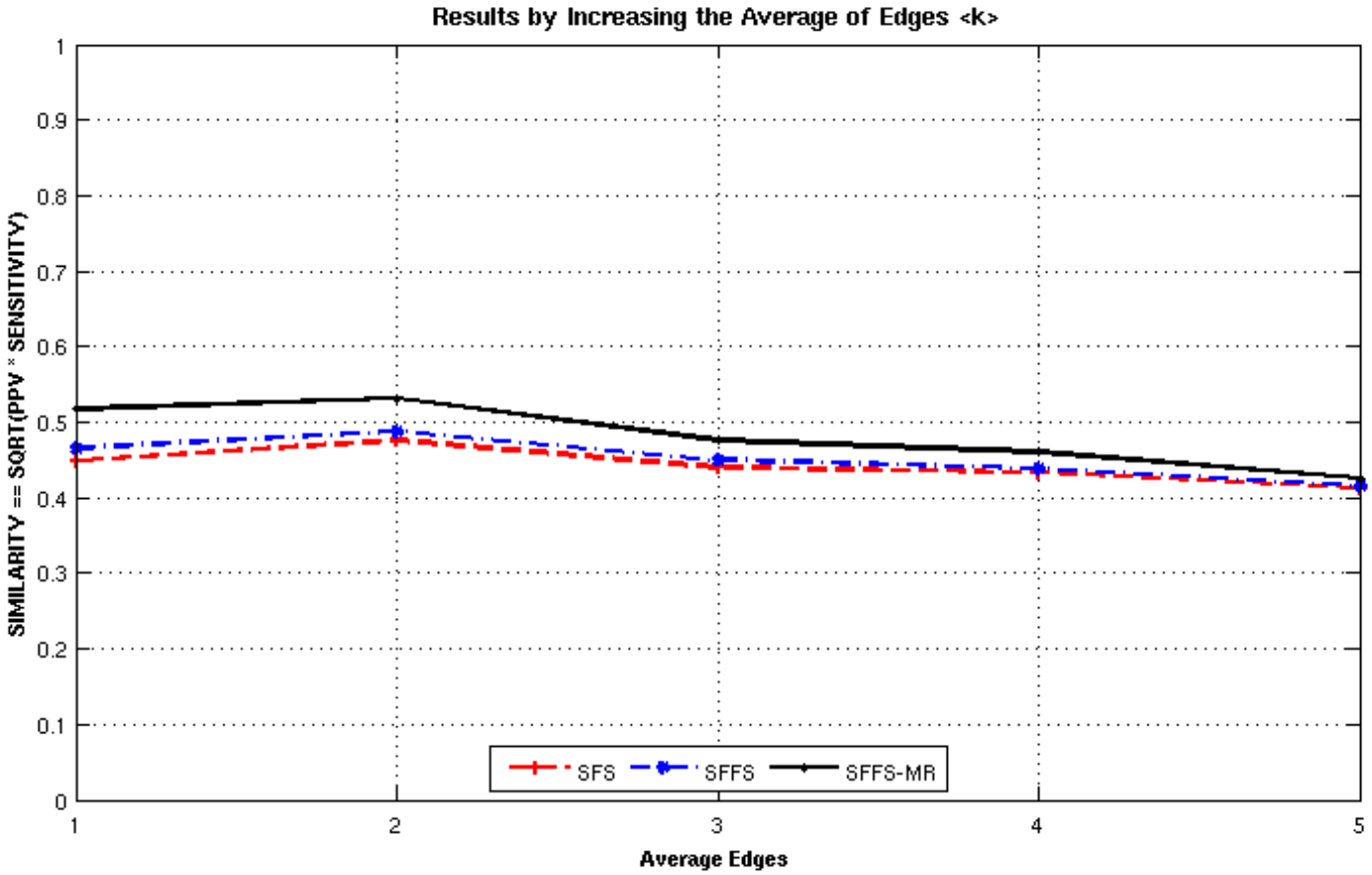
Experimental Results – 1st experiment

- PPV vs Signal Size



Experimental Results – 2nd experiment

- Similarity vs Average number of edges



Conclusion

Conclusion

- New feature selection strategy that applies SFFS for multiple initial features
- Assumption: some genes in biological organisms presents intrinsically multivariate prediction (IMP)
- The new method exploits the IMP property by exploring good and bad individual features
- Although the search space traversed by the proposed method is a little wider than SFFS, it does not worse its asymptotical computational cost

Conclusion

- Experimental results show that the SFFS-MR provides better inference accuracy (PPV) than SFS and SFFS when considering both small (15-20 time-points) and large (100 time-points) signal sizes
  - 60% of accuracy after only 20 observations from a state-space of size 2^{20}
  - Robustness in terms of the increasing average input degree

Future works

- Evaluation in large-scale networks
- Comparison of SFFS-MR with other network inference methods based on feature selection
- Application of this technique to infer GRNs from real data (e.g. microarray, SAGE, RNA-Seq)
- Development of a better strategy to choose the initial features (roots)
  - Features that are not correlated with each other may be interesting roots with potential to achieve better results than just considering best and worst individual features

Future works

- Inclusion of topology information such as small-world (WS) and scale-free (BA) in order to guide the search process for the correct topology inference of these networks
  - **Work in progress**

References

- Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., Guthke, R.: Gene regulatory network inference: Data integration in dynamic models - A review. Biosystems 96(1), 86-103 (2009)
- Lopes, F.M., Martins-Jr, D.C., Cesar-Jr, R.M.: Feature selection environment for genomic applications. BMC Bioinformatics 9(1), 451 (October 2008)
- Lopes, F.M., Cesar-Jr, R.M., da Fontoura Costa, L.: AGN simulation and validation model. In: Bazzan, A.L.C., Craven, M., Martins, N.F. (eds.) Advances in Bioinformatics and Computational Biology, Third Brazilian Symposium on Bioinformatics. LNBI, vol. 5167, pp. 169-173. Springer (August 2008)
- Martins-Jr, D.C., Braga-Neto, U., Hashimoto, R.F., Dougherty, E.R., Bittner, M.L.: Intrinsically multivariate predictive genes. IEEE Journal of Selected Topics in Signal Processing 2(3), 424-439 (June 2008)
- Pudil, P., Novovicova, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters 15(11), 1119-1125 (November 1994)
- Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W.: Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18(2), 261-274 (2002)

Authors

Fabrício M. Lopes
Assistant Professor (UTFPR)
PhD student (IME-USP)

David C. Martins-Jr
Professor (UFABC)

Junior Barrera
Full Professor (FFCLRP-USP)

Roberto M. Cesar-Jr
Full Professor (IME-USP)

Acknowledgements

- FAPESP, CNPq and CAPES for financial support