Learning parsimonious ensembles for biomedical prediction problems and DREAM challenges

Ana Stanescu and Gaurav Pandey*

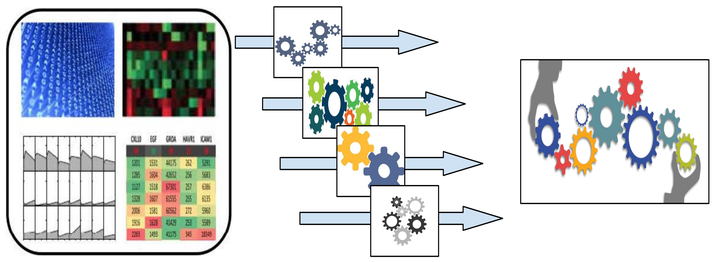
Icahn School of Medicine at Mount Sinai, New York, NY

gaurav.pandey@mssm.edu *http://research.mssm.edu/gpandey/*

National Institute of General Medical Sciences

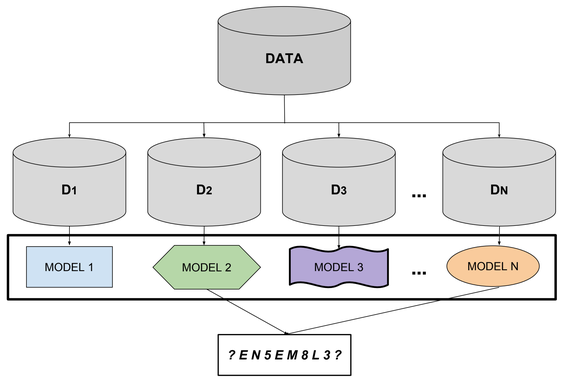
HETEROGENEOUS ENSEMBLES FOR DREAM CHALLENGES:

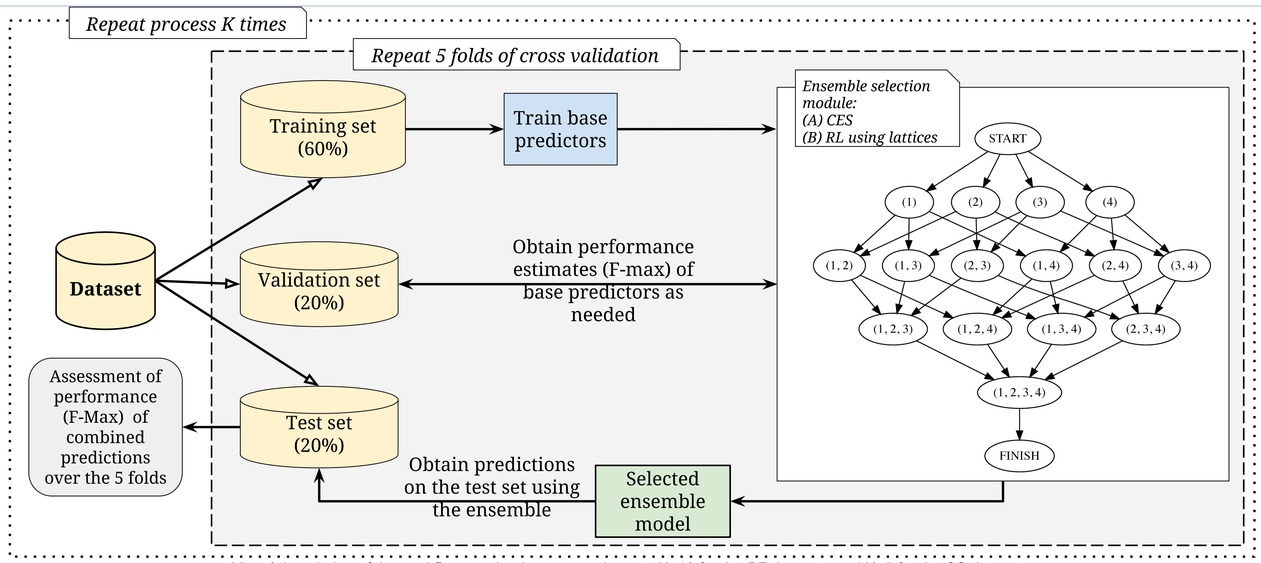
- DREAM challenges are a great mechanism for identifying the most effective solution(s) for challenging biomedical problems.
- Can we improve the (predictive) ability of DREAM challenges by considering the contributions of non-winning submissions/ models also?



- **GOAL**: build **heterogeneous ensembles**
- **Parsimony** of such an ensemble can be of even greater value for DREAM challenges due to enhanced interpretability
- Ensemble Selection(ES)/Pruning is a potential approach for this, but popular algorithms like Caruana's ES (CES) are ad-hoc (sub-optimal) and non-exhaustive

- **IDEA**: A novel ensemble selection approach based on reinforcement learning (RL), which provides a systematic way of exhaustively exploring the many possible combinations of base predictors that can be selected into an ensemble



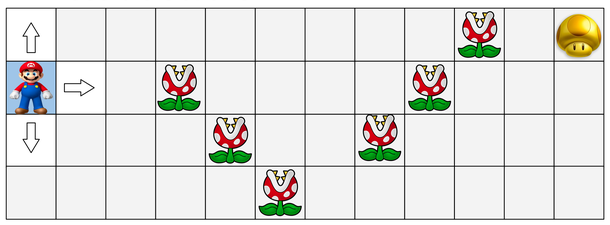


Visual description of the workflow used to in our experiments. K=10 for the PF datasets and K=5 for the SS datasets.

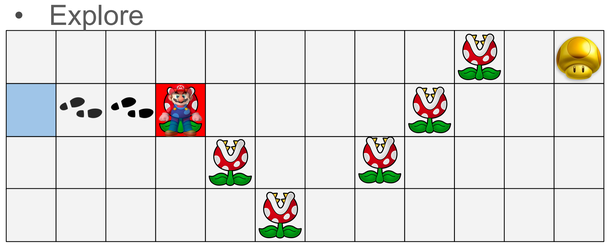
MATERIALS and METHODS:

Reinforcement Learning (RL)

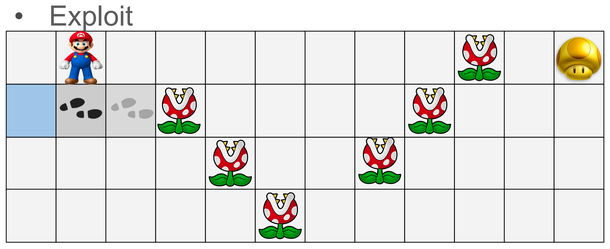
- Possible actions



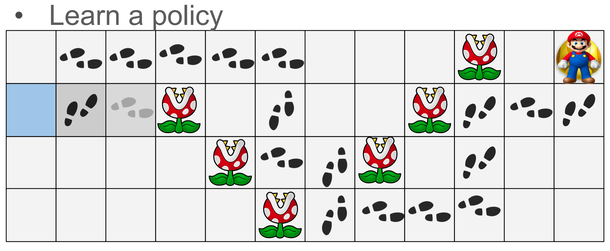
- Explore

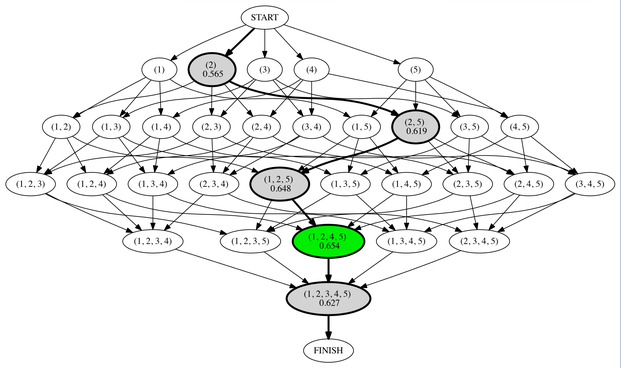


- Exploit



- Learn a policy





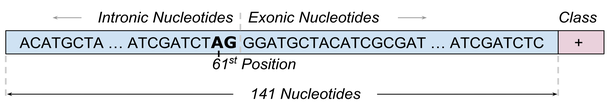
RL Strategies for ES

- **RL_greedy**
  - Reward is given by ensemble performance
- **RL_pessimistic**
  - Reset to start as soon as performance drops
- **RL_backtrack**
  - Go back one position when performance drops

Q-Learning

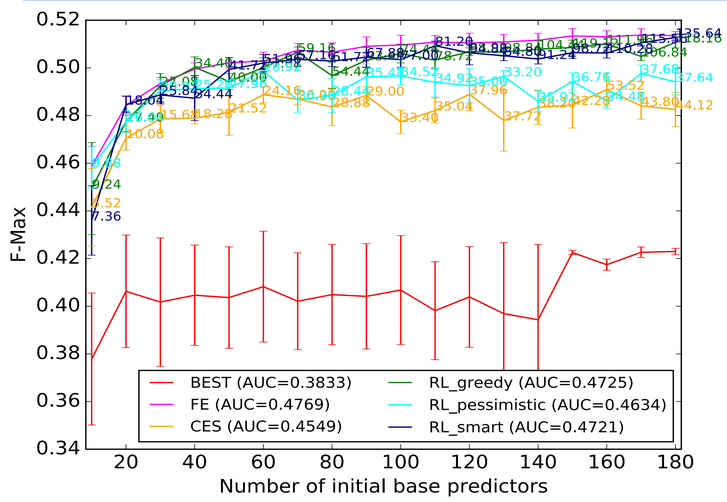
- Algorithm to find an optimal action-selection policy.
- Proven to converge to an optimal solution (i.e. find an optimal action-selection policy) under certain constraints.

Target problem: Splice Site Prediction



Problem	*C. elegans*	*D. melanogaster*	*P. pacificus*	*C. remanei*	*A. thaliana*
#Features	141	141	141	141	141
#Positives	1,598	997	1,596	1,600	1,600
#Negatives	158,150	99,003	156,326	157,542	158,377
Total	159,748	100,000	157,922	159,142	159,977

RESULTS:

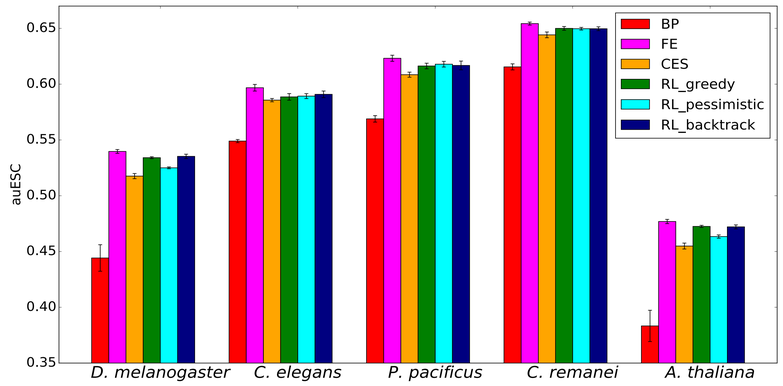


- RL able to better capture predictive performance close to the full ensembles with a much smaller number of base predictors.
  - More capable of achieving this balance than CES, especially for larger datasets.
- The downstream performance or sizes of the RL selected ensembles is not sensitive to RL parameters (*e.g.*, exploitation/exploration probability), showing **robustness to parameters** as compared to other, more ad-hoc ES methods.

A. thaliana	auESC	size_ratio@60	size_ratio@120	size_ratio@180	perf_ratio@60	per_ratio@120	perf_ratio@180
BP	0.3833	0.0167	0.0083	0.0056	0.8118	0.7912	0.8237
FE	0.4769	1	1	1	1	1	1
CES	0.4549	0.4	0.31	0.24	0.9710	0.9577	0.9379
RL_greedy	0.4725	0.5	0.5	0.51	0.9946	0.9927	0.9945
RL_pessimistic	0.4634	0.48	0.29	0.21	0.9919	0.9649	0.9623
RL_backtrack	0.4721	0.87	0.79	0.75	0.9983	0.9919	0.9985

- **Our approach can help extract more useful knowledge from DREAM challenges by constructing predictive and parsimonious ensembles of the submissions.**
  - Will be applied in the DREAM Respiratory Viral challenge
- **Implementation available:**

https://github.com/GauravPandeyLab/lens



Acknowledgements:

This work was partially supported by NIH grant # R01-GM114434 and an IBM faculty award to GP. We thank the Icahn Institute for Genomics and Multiscale Biology and the Minerva supercomputing team for their financial and technical support. We also thank Om P. Pandey and Gustavo Stolovitzky for their technical advice.

REFERENCES:

- **A. Stanescu and G. Pandey, *Learning parsimonious ensembles for unbalanced computational genomics problems*. In: Pacific Symposium on Biocomputing, PSB 2017 (In press.)**
- S. Whalen, O. P. Pandey and G. Pandey, *Predicting protein function and other biomedical characteristics*. Methods 93(15): 92-102, 2016.
- R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, *Ensemble selection from libraries of models*. ICML 2004.
- R. Caruana, A. Munson and A. Niculescu-Mizil, *Getting the Most Out of Ensemble Selection*. ICDM 2006.
- CJCH Watkins and P. Dayan, *Q-Learning*. Machine Learning 8(3-4), 1992.
- G. Schweikert, G. Rätsch, C. Widmer, and B. Schölkopf, *An empirical analysis of domain adaptation algorithms for genomic sequence analysis*. NIPS 2009.