**INTRODUCTION**

Expression2Kinases (X2K) is a computational pipeline that takes as input lists of differentially expressed genes, it then performs enrichment analysis to prioritize transcription factors that most likely regulate the observed changes in expression (Chen et al., 2012). Next, known protein-protein interactions are used to connect the identified transcription factors to form a subnetwork. Finally, kinases enrichment analysis is performed to prioritize protein kinases known to phosphorylate substrates within the subnetwork of transcription factors and intermediate proteins. X2K is available as a desktop, web, and command line tools. The X2K pipeline is made of these three components:

[1] *Transcription factor (TF) enrichment*: ChIP Enrichment Analysis (ChEA) (Lachmann et al., 2010) was used to rank transcription factors based on TF targets that are enriched in the input gene list.

[2] *Protein-protein Interaction (PPI) network construction:* Genes2Networks (G2N) (Berger, Posner, & Ma’ayan, 2007) was used to generate protein-protein interactions (PPI) subnetworks given a list of TFs.

[3] *Kinase-enrichment*: Kinase Enrichment Analysis (KEA) (Lachmann & Ma’ayan, 2009) was used to rank protein kinases that are known to phosphorylate the proteins within the subnetwork identified by G2N.

Each step in the X2K pipeline has a number of modifiable parameters. However there are far too many unique combinations of parameters to manually identify the optimal parameters settings to accurately predict kinase targets. We therefore developed a Genetic Algorithm (GA) to search the fitness landscape ( in terms of parameter combinations) in order to efficiently identify an optimal set of parameters for the X2K pipeline.

**MATERIALS & METHODS**

*GA Description*

The GA operates by coding all changeable parameters in each step of X2K as a binary string. It then produces a population of binary strings, runs the X2K pipeline each string’s corresponding parameter combination to assess fitness (in terms of accuracy of predicting kinases), crossover the fittest individuals to create a new population, introduces random mutations into that population to avoid the GA getting stuck at a local optimum, and repeating until population fitness stabilizes around an optimum (Figure X).

The following GA settings were used:

* 20 generations
* 100 individuals/generation
* 4 crossover points
* Random mutation rate of 0.01
* Breed the top 10 individuals
* Retain the top 10 individual in the next population

*X2K Pipeline Parameters:*

* **[1] TF-enrichment** 
  + ***TF Sort***: Three mutually-exclusive sort methods could be used to sort significantly enriched TFs, including, 1) *P-value*: sort by raw p-value from Fisher’s exact test, 2) *Rank*: the rank by p-value minus the standard deviation, 3) *Odds ratio*: log(p-value) x (observed overlap/the expected overlap = Odds Ratio), 4) *Combined Score*: log(p-value) x z-score.
  + ***TF Databases***: 8 non-mutually exclusive TF databases were available, including ChEA 2016, ARCHS4 human, ARCHS4 mouse, ENCODE 2016, huMAP, BioGRID, JASPAR-TRANSFAC, and CREEDS ()
  + ***TF Species:*** Three mutually-exclusive options to include different subsets of the TF Databases by species, including 1) human, 2) mouse, 3) both.
  + ***Top TFs***: Three mutually exclusive options for selection the number of top-enriched TFs to be used in the PPI-construction step (5, 10 or 20).
* **[2] PPI-construction:**
  + ***PPI Databases***: Eight non-mutually exclusive PPI databases were available, including BIND, BIOCARTA, BIOGRID, DIP, FIGEYS, HPRD, INNATEDB, INTACT, KEGG, MINT, MIPS, MURPHY, PDZBASE, PPID, PREDICEDPPI, SNAVI, STELZL, VIDAL, and HUMAP (REFERENCES).
  + ***PPI Path Length***: Two mutually exclusive options for the number of nodes from the TF-input by which to extend the PPI network (1 or 2).
* **[3] Kinase-enrichment:** 
  + ***Kinase Sort***: Three mutually-exclusive sort methods could be used to sort significantly enriched TFs (*same options as TF Sort*).
  + ***Kinase Data Types***: Two different types of data are available in KEA; kinase-protein interactions (KP) and phosphorylation reactions (P).
  + ***Top Kinases***: The maximum number of kinases predicted by the kinase-enrichment was held constant at 20 to standardize the fitness measurements.
* *X2K Validation:*
  + Databases:
    - Run 1: GEO gene perturbation data
    - Run 2: LINCS L1000 drug perturbation data combined with KINOMEscan
  + Fitness scores of each parameter combination are defined as the % of experiments in which the known perturbed kinase was recovered, weighted by the perturbed kinase’s ranking in the significance-ordered output.

**RESULTS**

* Run 1
  + Annealed at gen n
  + Optimal parameters
  + Overfitting tests
  + ANOVA parameter results
* Run 2
  + Annealed at gen n
  + Optimal parameters
  + Overfitting tests
  + ANOVA parameter results

**DISCUSSION**

* Consistency between runs.
* Optimal parameters overall. Insights as to why these are optimal (type, quality of database, previous comparisons)

**CONCLUSIONS**

* 1)
* 2)
* 3)

FIGURES

   