An Integrative Multi-Network and Multi-Classifier Approach to Predict Genetic Interactions

Gaura, Bin, Aaron, et. al.

By: Xiangxue Wang
Background

**Genetic Interaction:** mutations in two genes produce a phenotype that is surprising in light of each mutation's individual effects

**Importance:** Genetic Interactions capture functional redundancy, and thus are important for:
1) predicting function;
2) and dissecting protein complex into functional pathways;
3) and exploring the mechanistic underpinnings of common human diseases.

   e.g. The synthetic sickness and lethality are the most studied types of genetic Interactions in yeast.
Background

**Motivation:** only a small proportion of gene pairs have been tested for genetic interactions, even in yeast. Because of large number of possible combinations of gene pairs.

**Proposal:**

**Goal:** expanding the set of known synthetic lethal (SL) interactions

**Approach:** an integrative, multi-network approach for predicting genetic interactions has been devised

e.g. synthetic sickness and lethality (SSL) arises when a combination of mutations in two or more genes leads to cell death, whereas a mutation in only one of these genes does not, and by itself is said to be viable.
Previous Work

The multiple network decision tree (MNDT)
----most comprehensive work to predict SSL interactions precisely

1. MNDT extracted both SSL-dependent features and SLL-independent features to train a decision tree-based classifier.

2. Given two networks characterizing relationships between gens, a 2-hop feature for two genes A and B is used represent whether there is a 2-step path between A and B through a third gene C in different network

Ad: 2-hop feature derived from the overlay of the multiple networks are the most effective differentiating SSL and non-SSL interactions, in particular when one of the networks was known SSL network.

Dis: 1)the use of only SSL-independent features led to a low accuracy
   2) SSL-independent features are unavailable for many gene pairs.

Overall, MNDT is an extensive and effective organism-specific approach in the literature and has been the basis for validating new algorithm
Multi-Network and Multi-Classifer (MNMC) Framework

Main parts:

1. Defining a large number of features that are independent of known SL interactions for characterizing the relationships between pairs of genes.

2. Using these features, a non-parametric multi-network and multi-classifier system for predicting SL interactions has been developed.

Ad: more appropriate for settings where very few gene pairs have been tested, including higher organism where large-scale interactions are not available.

Application:
Using above approach to predict the genetic interactions between the known transcription factors (TFs) and uncovered a number of novel SL interactions between TFs, which were supported by knowledge.
Main Parts:

Feature Extraction:

Goal: identifying a set of features that capture information about the relationships between genes.

Results: 152 SL-independent features are extracted:
   1) 62 by capturing the likelihood of genes being directly related
   2) 90 by overlaying pairs of networks (individual features)

The value of the feature obtained by overlaying two such networks N1 and N2 is computed as \[ \max[w(g_1, c), w(c, g_2)] \], where \((g_1, c) \in N1 \) and \((c, g_2) \in N2 \) or vice versa.

One feature overlay with another to generate overlay feature.
Feature Evaluation:

Goal: evaluating the ability of differentiating SL and non-SL classes
Approach: Kolmogorov-Smirnov (KS) for distribution difference

D-statistic from KS for discriminative power of features.

Most discriminative SL-independent features used in MNMC prediction approach
Under-Sampling
Goal: getting a balanced data classes for training.
Approach: under-sampling of the majority (negative) class s.t. the number
is equal to the positive. (92.9% vs 7.1%)

Classifier Training
The balanced combination of two sets is used to train a non-parametric
multi-classifier system that enabled the simultaneous use of multiple
classification, such as SVM, NN, decision tree.

Validation
Testing each of the individual classifiers and the ensemble (MNMC) on
SGD-SL dataset.
Receiver operating characteristic (ROC) curves plotted based on 10-fold
Cross-validation.
Precision-Recall curves plotted as a score threshold of 0.2.

Precision = #True SL prediction / #Complete set of SL predictions
Recall = #True SL prediction / #Complete set of Known SL examples
a. ROC of 6 individual classifiers and the ensemble using SL-independent features.
b. ROC for the ensemble using data with all features (AUC = 0.837) and SL-independent features (AUC = 0.741)
c. Precision-Recall using all features and SL-independent Features.
Results compare on different dataset and with MNDT

(A) ROC (AUC of MNMC.all = 0.819 and MNMC.slif = 0.851) and (B) Precision-recall for classification of the 337 least connected SL and 199 corresponding non-SL interactions using our SL-independent and SL-dependent features;

(C) ROC (AUC, MNDT.all = 0.862, MNMC.all = 0.897, MNDT.slif = 0.598 and MNMC.slif = 0.805) and (D) Precision-recall for classification on Wong et al.'s SSL dataset [8] using MNMC and MNDT based on on either all the features (all) or the SL-independent features (slif);

(E) ROC (AUC, MNMC.all = 0.616 and MNMC.slif = 0.633) and (F) Precision-recall curves for classification on an independent test set constructed from the SGD interaction database using MNMC.all and MNMC.slif.
Using MNMC to Study Functional Redundancy in Transcriptional Factors

467 predicted and 8 experimentally verified interactions between 106 transcription factors from SGD (highlighted in red)

Similar networks obtained, and is well supported by Literature.

In summary, MNMC works in predicting interaction network very well

Prediction of a global TF SL interaction network
Methodology Highlights

Feature extraction:
1) semantic similarity measure of two classes in one gene ontology

\[ linsim(c_1,c_2) = \frac{2 \times \left[ \log p_{ms}(c_1,c_2) \right]}{\log p(c_1) + \log p(c_2)} \]

\( c_i, i \) classes; \( p(c) \), probability of a protein being annotated with class \( c \);
\( p_{ms}(c_1,c_2) = \min_{c \in S(c_1,c_2)} p(c) \), \( S(c_1,c_2) \) is the set of common ancestor of \( c_1, c_2 \)

2) functional similarity measure of two genes

\[ FunctionalSim(A,B) = \frac{2 \times \sum_{(a_i,b_i) \in p} linsim(a_i,b_i)}{|A| + |B|} \]

\( A \), set of annotations of two genes in the entire ontology, namely group A

Results: the similarity measure takes the specificity and relative positioning of the annotations into account more robustly than a simple count of common Functional annotations.

Finally, created three features for functional similarity of each gene pair, each corresponding to one GO ontologies, Biological Process, Molecular Function and Cellular Component
Methodology Highlights

Multiple Classification System:
1) Noise-AND function based Combination Strategy

\[ p(x) = \prod_{i=1}^{N} p_i(x) - \prod_{i=1}^{N} (1 - p_i(x)), \]

\( x, \) a given gene pair;
\( p_i(x), \) represents the probability that \( x \) is predicted as SL by classifier \( i \)

Thus, this score simply computes a difference between the products of the probabilities of the example belong to the SL and non-SL classes from each of the classifiers, and the higher this score, the more likely there is SL interaction.
Thank you!