NASH gene prioritization using novel feature engineering techniques

2/11/2021

Nikki Taylor, M.S. Candidate in Biomedical Informatics

Process for predicting novel disease genes

1. Data Organization

- String PPI
- Biological module gene sets
- Disease genes

2. Feature engineering

- Node2vec PPI embeddings
- Gene-module similarity scores
- Feature selection: ANOVA F-value

3. Model selection

- Linear SVC
- Resampling
- 10-fold cross validation
- Validation with NLP literature mined data, experimental data

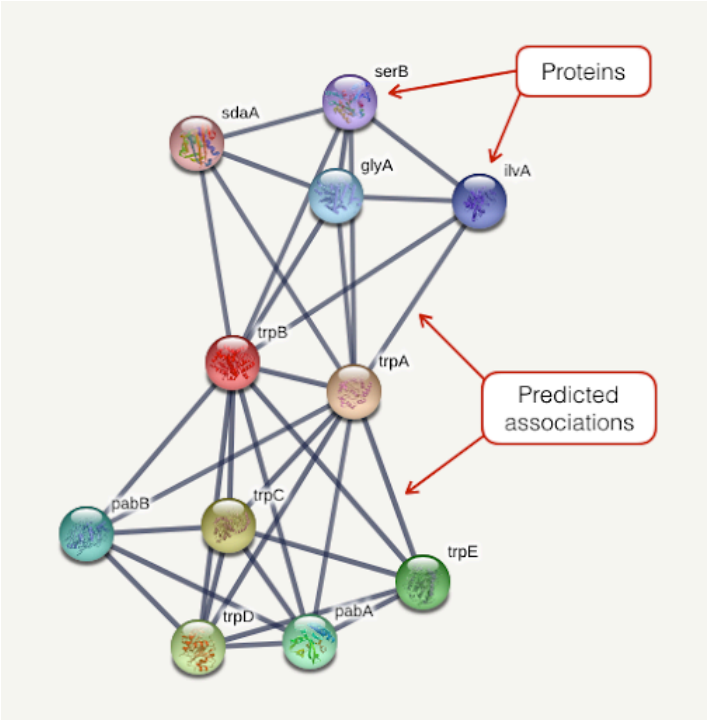
4. Gene prioritization

- Predict probabilities
- Experimental validation
- Drug target prioritization

1. Data Organization

Data Sources: String PPI

- Includes protein-protein interactions derived from experimental data, computational prediction methods, and public text collections
- We extracted 728K high confidence connections



Data Sources: Biological Modules

Type of data	Description	Number of gene sets
Immune response	Proteomics data from ImmProt – modules derived from 6982 proteins enriched in immune cells	47
Hallmark signaling pathways	Molecular Signatures Database (MSigDB) hallmark gene set collection	50
Metabolic subsystems	Genome-scale metabolic pathways derived from Human Metabolic Reaction Database (Human GEM)	137
Cytokines and cytokine receptors	Our cytokine network	3

Data Sources: Known NASH genes

Gold standard genes:

- 70 genes from DisGeNET curated list
  - Derived from UNIPROT, CGI, ClinGen, Genomics England, CTD (human subset), PsyGeNET, and Orphanet.

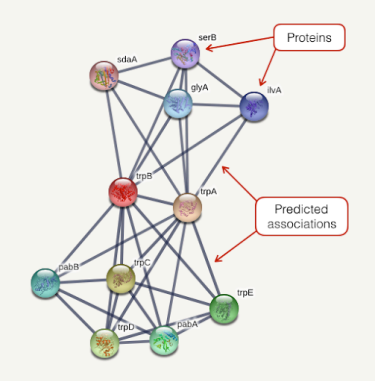
Genes for validation:

- 308 genes from DisGeNET ”Befree” NLP literature mined list
  - Text mining of medline abstracts
- 118 experimentally derived genes from the Svensson Lab
  - Human mappable genes from set of 200 identified by scRNA sequencing of liver cells from diet-induced NASH mice

2. Feature engineering

Generating annotated gene embeddings

String PPI



node2vec

Gene embeddings

Gene	1	2	…	64
Gene 1	.5	.2	…	.6
Gene 2	.01	.45	…	-.25
Gene 3	.73	7.2	…	.9
…				
Gene 14,704	.05	12	…	-6

14,704 genes x 64 dimensions

Annotate with biological modules

Annotated gene embeddings

Mod	Gene	1	2	…	64
Mod 1	Gene 1	.5	.2	…	.6
Mod 1	Gene 2	.01	.45	…	-.25
Mod 7	Gene 3	.73	7.2	…	.9
…	…				
Mod 27	Gene 14,704	.05	12	…	-6

14,704 genes x 64 dimensions

Embeddings:

- low dimensional vectors to represent genes that preserve relationships in the graph
- nodes that are similar in the graph are similar in the embedding space (by cosine distance)

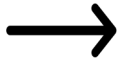
Creating module vectors

Annotated gene embeddings

Module	Gene	1	2	…	64
Mod 1	Gene 1	.5	.2	…	.6
Mod 1	Gene 2	.01	.45	…	-.25
Mod 7	Gene 3	.73	7.2	…	.9
…	…				
Mod 27	Gene 14,704	.05	12	…	-6

14,704 genes x 64 dimensions

Sum embeddings by module



Module vectors

Module	1	2	…	64
Mod 1	10	.8	…	3
Mod 2	.9	4	…	8
…				
Mod 237	.11	3	…	-9

237 modules x 64 dimensions

Why sum embedding vectors?

- To represent all embeddings in a module as one vector
- Cosine similarity is based on angles between vectors and is not impacted by scaling, so averaging is not necessary

Calculating gene-module scores

Pairwise cosine similarity

Gene embeddings

Gene	1	2	…	64
Gene 1	.5	.2	…	.6
Gene 2	.01	.45	…	-.25
Gene 3	.73	7.2	…	.9
…				
Gene 14,704	.05	12	…	-6

14,704 genes x 64 dimensions

Module vectors

Module	1	2	…	64
Mod 1	10	.8	…	3
Mod 2	.9	4	…	8
…				
Mod 237	.11	3	…	-9

237 modules x 64 dimensions

Gene-module scores

	Mod 1	Mod 2	…	Mod 237
Gene 1	.5	.2	…	.6
Gene 2	.01	.45	…	-.25
Gene 3	.73	7.2	…	.9
…				
Gene 14,704	.05	12	…	-6

14,707 genes x 237 modules

Feature selection: top 64 features

Gene-module scores

	Mod 1	Mod 2	…	Mod 237
Gene 1	.5	.2	…	.6
Gene 2	.01	.45	…	-.25
Gene 3	.73	7.2	…	.9
…				
Gene 14,704	.05	12	…	-6

14,707 genes x 237 modules

Select top 64 modules by ANOVA F-score

Gene-module scores

	Mod 1	Mod 2	…	Mod 64
Gene 1	.5	.2	…	.5
Gene 2	.01	.45	…	-.3
Gene 3	.73	7.2	…	.96
…				
Gene 14,704	.05	12	…	5

14,707 genes x 64 modules

Feature engineering

1. Use node2vec to generate 14,707 gene embeddings from String PPI
2. Annotate genes with biological module involvement
3. Sum gene embeddings by module
4. Calculate gene-module cosine similarities to use as features (237 scores for each gene)
5. Select 64 best modules as features using ANOVA F-score
   - During cross validation, feature selection is based on training set

3. Model selection

Model specifics and benchmarking

Model:

- Linear SVC
- Positive NASH genes = set of 70 from curated list
- Negative genes = all other genes not in positive or validation sets

Benchmarking:

- 10 fold cross validation
- Sampling method: SMOTE oversampling vs undersampling
- Feature type: embeddings vs. module scores
- Feature selection
- Validation using 118 experimental genes and 408 NLP genes, 200 randomly selected negative genes

Benchmarking results

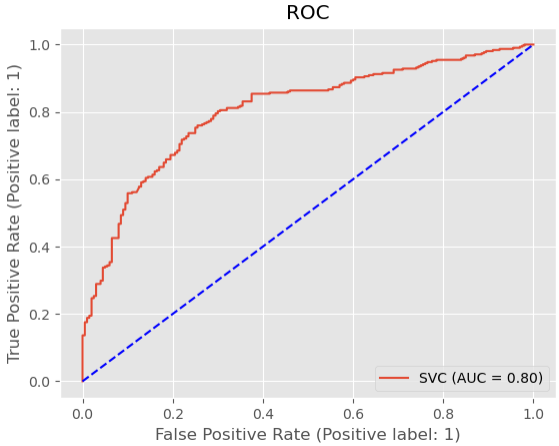
Features	Sampling Method	Test set positives	CV 1	CV 2	CV 3	CV 4	CV 5	CV 6	CV 7	CV 8	CV 9	CV 10	All 70 genes
Gene embeddings	Oversample	Curated (70 total, 14 in test set)	0.87	0.77	0.75	0.93	0.96	0.92	0.84	0.88	0.92	0.88	
		Befree (308)	0.69	0.7	0.7	0.72	0.7	0.71	0.72	0.72	0.69	0.71	0.73
		Svensson genes (118)	0.72	0.72	0.72	0.74	0.7	0.73	0.74	0.73	0.69	0.69	0.74
	Undersample	Curated (70 total, 14 in test set)	0.84	0.75	0.81	0.9	0.91	0.9	0.85	0.75	0.86	0.8	
		Befree (308)	0.74	0.74	0.75	0.76	0.76	0.7	0.75	0.75	0.7	0.75	0.73
		Svensson genes (118)	0.79	0.75	0.75	0.78	0.78	0.69	0.81	0.71	0.73	0.7	0.77
Module scores	Oversample	Curated (70 total, 14 in test set)	0.92	0.89	0.92	0.9	0.8	0.88	0.82	0.92	0.91	0.92	
		Befree (308)	0.78	0.78	0.77	0.77	0.79	0.77	0.78	0.76	0.78	0.77	0.77
		Svensson genes (118)	0.78	0.79	0.78	0.77	0.77	0.76	0.73	0.76	0.78	0.76	0.77
	Undersample	Curated (70 total, 14 in test set)	0.94	0.9	0.9	0.88	0.88	0.89	0.87	0.91	0.88	0.89	
		Befree (308)	0.8	0.79	0.79	0.77	0.79	0.78	0.79	0.77	0.79	0.76	0.77
		Svensson genes (118)	0.79	0.8	0.79	0.76	0.79	0.77	0.79	0.76	0.78	0.75	0.79
Feature selected module scores	Oversample	Curated (70 total, 14 in test set)	0.9	0.83	0.86	0.89	0.95	0.94	0.91	0.89	0.91	0.9	
		Befree (308)	0.78	0.78	0.79	0.78	0.78	0.79	0.8	0.81	0.78	0.78	0.79
		Svensson genes (118)	0.8	0.78	0.79	0.78	0.79	0.79	0.79	0.8	0.77	0.76	0.8
	Undersample	Curated (70 total, 14 in test set)	0.89	0.85	0.87	0.89	0.9	0.93	0.92	0.89	0.9	0.88	
		Befree (308)	0.79	0.79	0.8	0.79	0.78	0.8	0.8	0.79	0.78	0.77	0.80
		Svensson genes (118)	0.79	0.77	0.79	0.78	0.78	0.76	0.79	0.8	0.77	0.71	0.82

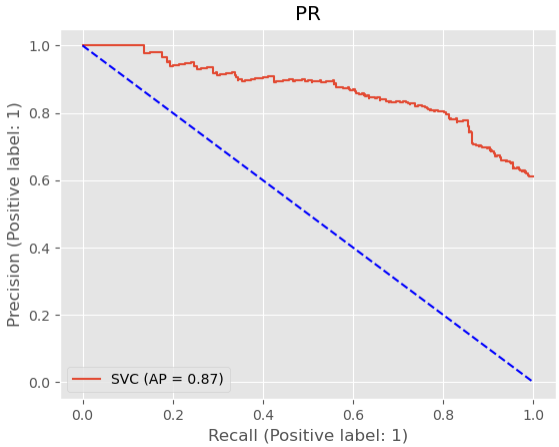
Validation set:

Final model

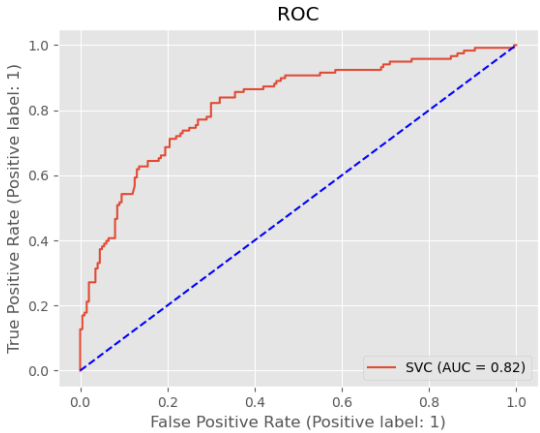
- Linear SVC
- Feature selection
- Undersampling

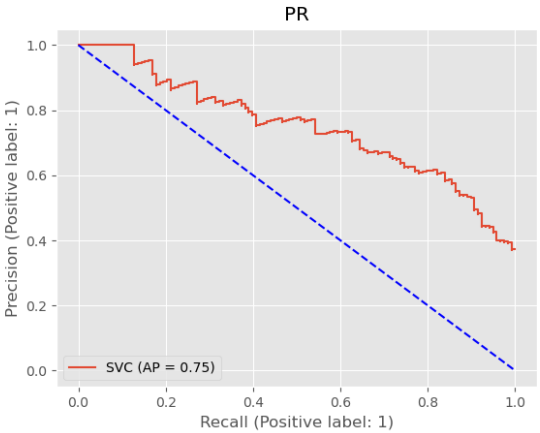
Befree genes





Svensson genes





4. Gene prioritization

Top features

Module	P-value
heme_degradation	2.33E-36
TNFA_SIGNALING_VIA_NFKB	2.89E-24
carnitine_shuttle_cytosolic	3.77E-24
IL6_JAK_STAT3_SIGNALING	4.86E-23
eicosanoid_metabolism	3.57E-21
carnitine_shuttle_endoplasmic_reticular	6.16E-21
ros_detoxification	1.76E-20
essential cytokines	3.22E-20
beta_oxidation_of_di_unsaturated_fatty_acids_n_6_peroxisomal	1.13E-18
arachidonic_acid_metabolism	2.25E-18
omega_3_fatty_acid_metabolism	9.28E-18
APOPTOSIS	1.17E-17
ME 44	2.22E-17
carnitine_shuttle_peroxisomal	2.32E-17
beta_oxidation_of_unsaturated_fatty_acids_n_9_peroxisomal	2.99E-17
beta_oxidation_of_even_chain_fatty_acids_peroxisomal	2.99E-17
beta_oxidation_of_unsaturated_fatty_acids_n_7_peroxisomal	2.99E-17
ME 22	1.17E-16
protein_assembly	2.16E-16

Top 20 false positive genes (unknown)

Gene	Score
HMOX1	0.99
PTGS2	0.98
NOS2	0.98
FABP1	0.98
FABP9	0.97
IDO1	0.97
STAT6	0.97
POR	0.97
BLVRA	0.97
FMO5	0.97
IL12B	0.97
ENSG00000264813	0.97
LTC4S	0.97
BBOX1	0.96
HSD17B4	0.96
CYP4A11	0.96
SLC27A5	0.96
FOS	0.96
LIPC	0.96
GPR29	0.96

- Scores are produced using Platt scaling (logistic transformation of classifier scores)

Drug target scores

Drug	Average score	Notes
OBETICHOLIC ACID	0.92	Achieved primary endpoint in phase 3 trial
MGL-3196	0.9	In phase 3
SELONSERTIB	0.89	Halted in phase 3 (2016)– lack of efficacy
ARAMCHOL	0.89	In phase 2b
BMS-986036 (Pegbelfermin)	0.88	In phase 2
ELAFIBRANOR	0.84	Halted in phase 3 (2020) – lack of efficacy
VOLIXIBAT	0.82	Terminated phase 2 (2019) – lack of efficacy
GR-MD-02	0.78	Beginning phase 2b/3
CENICRIVIROC	0.46	In phase 3
EMRICASAN	0.19	Halted in phase 2b (2019) – lack of efficacy
SEMAGLUTIDE	0.06	Completed phase 2
LIRAGLUTIDE	0.06	Completed phase 2