CLASSIFYING SUBSTRATES OF TRANSPORTERS BY USING NEIGHBORING GENES

Tran Vu Ha

Supervisor: Prof. Volkhard Helms

Advisor: Ahmad Barghash

CONTENTS

1. INTRODUCTION

1.1. The Problem

1.2. Operon

1.3. Aim of the project

THE PROBLEM

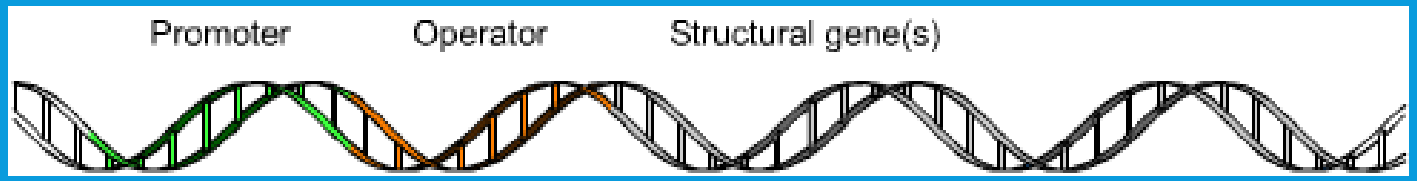
- **So far the problem of classifying substrates transported by membrane transporters is unsolved**

- Methods for prediction these substrates:
  - Nadine Schaadt: Classifying substrate specificities of membrane transporters from Arabidopsis thaliana based on amino acid composition (AAC).
  - Ahmad Barghash uses sequence similarity and motif.

  - **We propose a method for this problem by using information of neighboring genes**

OPERON

- Jacob, F.; Perrin, D.; Sanchez, C.; Monod, J. (Feb 1960). "Operon: a group of genes with the expression coordinated by an operator
- Wikipedia: **operon** is a functioning unit of genomic DNA containing a cluster of genes under the control of a single regulatory signal or promoter.

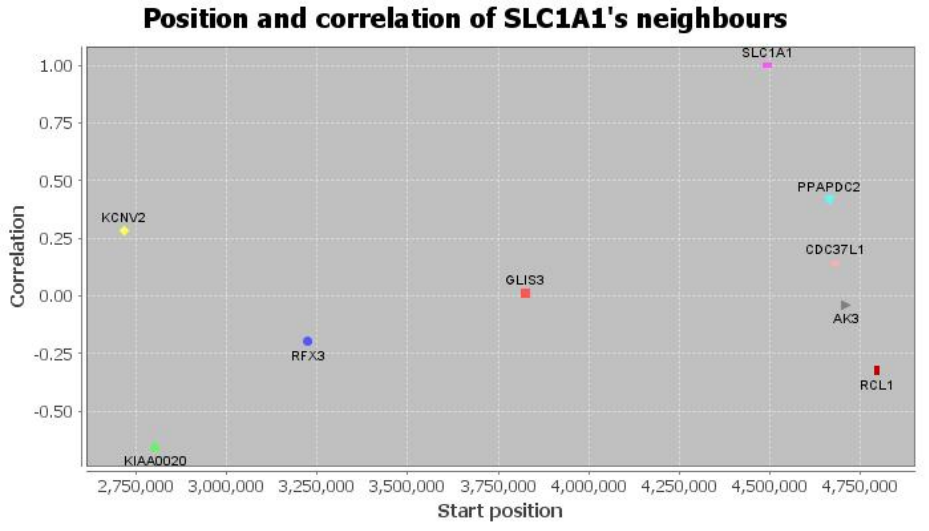


- The genes contained in the operon are either expressed together or not at all
- Operons exist in prokaryotes, and also in eukaryotes

Image source: http://en.wikipedia.org/wiki/File:Operon_1.png

AIM OF THE PROJECT

- Find out relations between gene and its neighbors
- These relations then can be used to classify substrate of transporter proteins
- For example:
  - Find relation between genes and their expression correlations



2. RELATED WORKS

2.1. Data sets and data sources

2.2. Features & Tools

DATASETS AND SOURCES

- Amino acid, metal and sugar transporters of human, E.coli and Saccharomyces Cerevisiae.
- Neighboring genes from UCSC and EcoCyc
- Expression data from TCGA (human) and GEO (E.coli and yeast)

FEATURES & TOOLS

- Neighboring genes
- Gene co-expression
- Gene Ontology
- Support Vector Machines

NEIGHBORING GENES

- For E. coli, we downloaded data file from EcoCyc
- For Human, saccharomyces cerevisiae … we can find neighboring genes from UCSC's MySQL server
- Download data table and save it to local computer

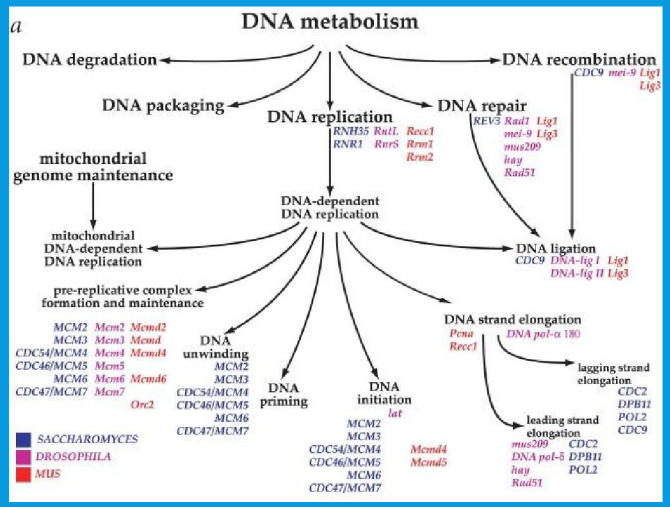
GENE CO-EXPRESSION

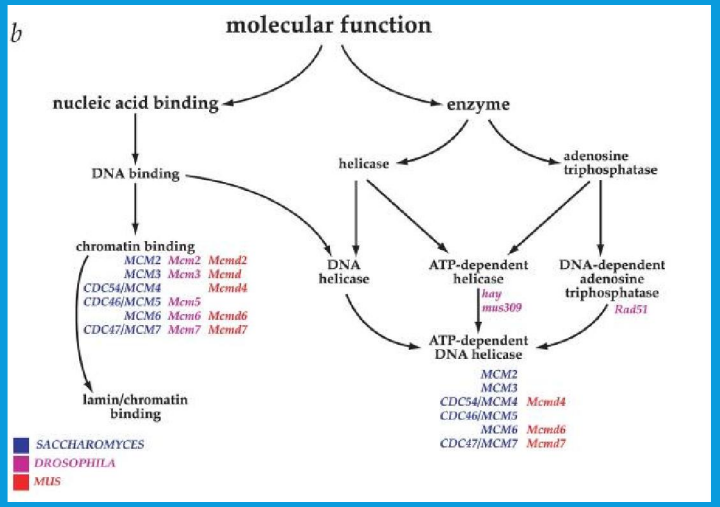
- Genes are co-expressed if they have the same expression levels.
- In this project we used Pearson correlation to measure co-expression
- Expression data from TCGA (human colon adenocarcinoma) and GEO (e.coli (GDS2768) and yeast (GDS91))

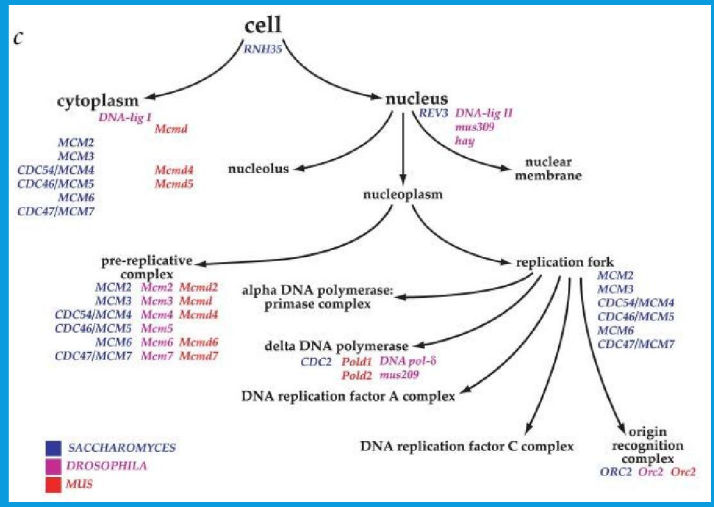
Central Gene	Neighboring Gene	Correlation	Start position	End position
LYP1	BNI1	0.331440142	129521	135383
LYP1	SEC2	-0.29621985	126804	129084
LYP1	GOR1	-0.170978515	121117	122170
LYP1	LYP1	1	138549	140385
LYP1	PIK1	-0.493177603	140877	144078
LYP1	PDR17	0.280377717	145562	146615
LYP1	YIF1	0.21630743	146895	147840

GENE ONTOLOGY

- Gene Ontology
  - Three categories of GO: Biological Process, Molecular Function, Cell Component







- Tools for Gene Ontology:
  - Uniprot (http://www.uniprot.org) to get GO term ID from Uniprot Accession
  - QuickGO (http://www.ebi.ac.uk/QuickGO/) to get other information of a GO term ID

Image source: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/figure/F1/

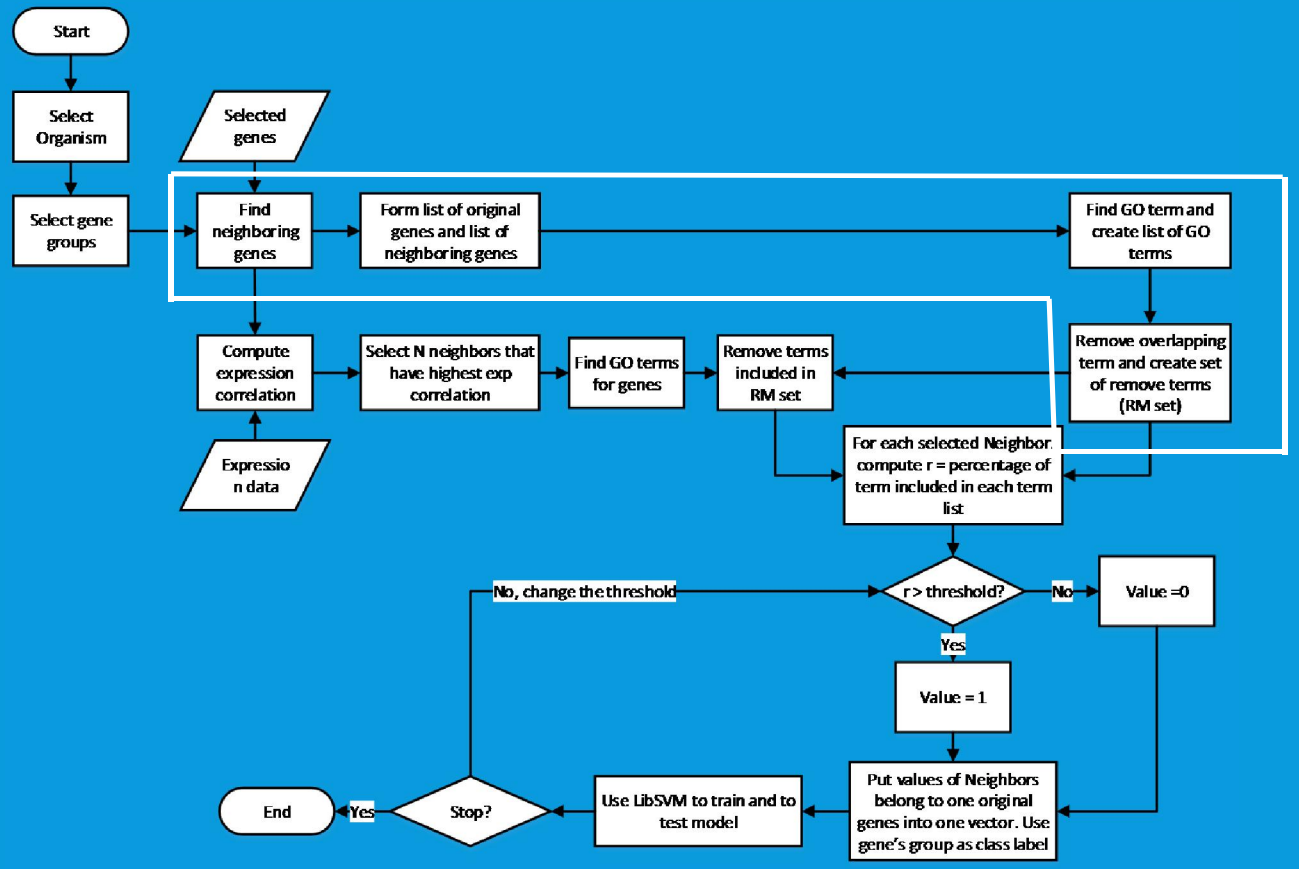
SUPPORT VECTOR MACHINES (SVM)

- Support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.
- LIBSVM
  - A Library for Support Vector Machines by Chih-Chung Chang and Chih-Jen Lin
  - The goal of LIBSVM is to help users from other fields to easily use SVM as a tool

3.RESULT

- Method
- Testing result

WORKFLOW

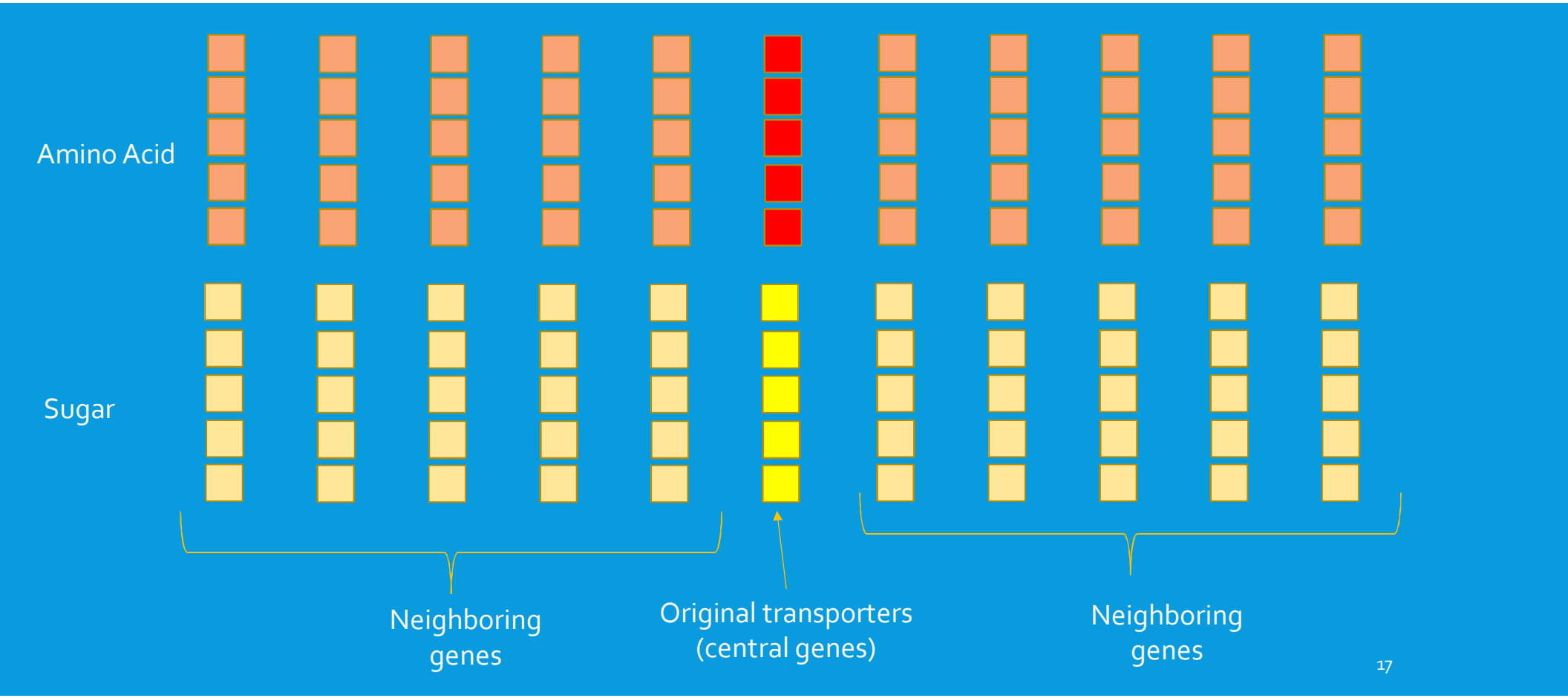

Start
Select Organism
Select gene groups
Selected genes
Find neighboring genes
Form list of original genes and list of neighboring genes
Find GO term and create list of GO terms
Compute expression correlation
Select N neighbors that have highest exp correlation
Find GO terms for genes
Remove terms included in RM set
Remove overlapping term and create set of remove terms (RM set)
Expressio n data
For each selected Neighbor compute r = percentage of term included in each term list
r > threshold?
No
Value =0
No, change the threshold
Yes
Value = 1
Put values of Neighbors belong to one original genes into one vector. Use gene's group as class label
Use LibSVM to train and to test model
Stop?
Yes
End

SELECTED GENE GROUPS

Amino Acid Transporter genes

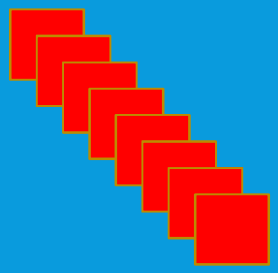
Sugar Transporter genes

FIND NEIGHBORING GENES



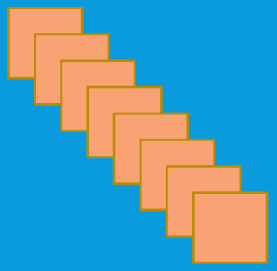
FORM THE LISTS

Amino Acid transporter genes

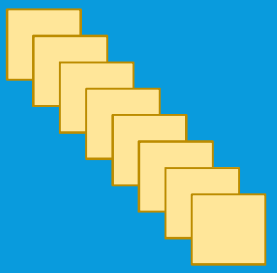


Sugar transporter genes

Neighbors of Amino Acid transporter genes

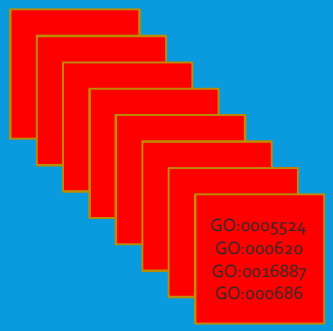


Neighbors of Sugar transporter genes

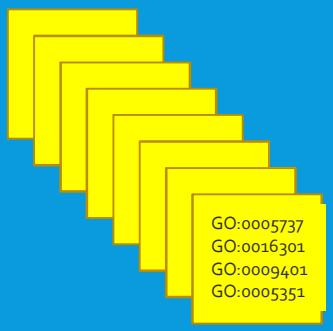


FIND GO TERM FOR ALL GENES

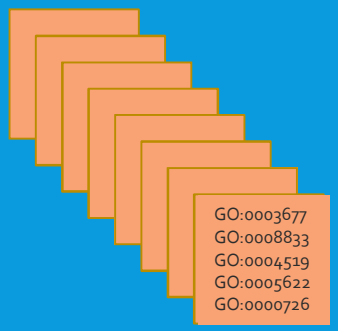
All GO terms of Amino Acid transporter genes



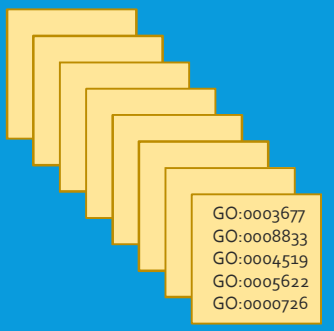
All GO terms of Sugar transporter genes



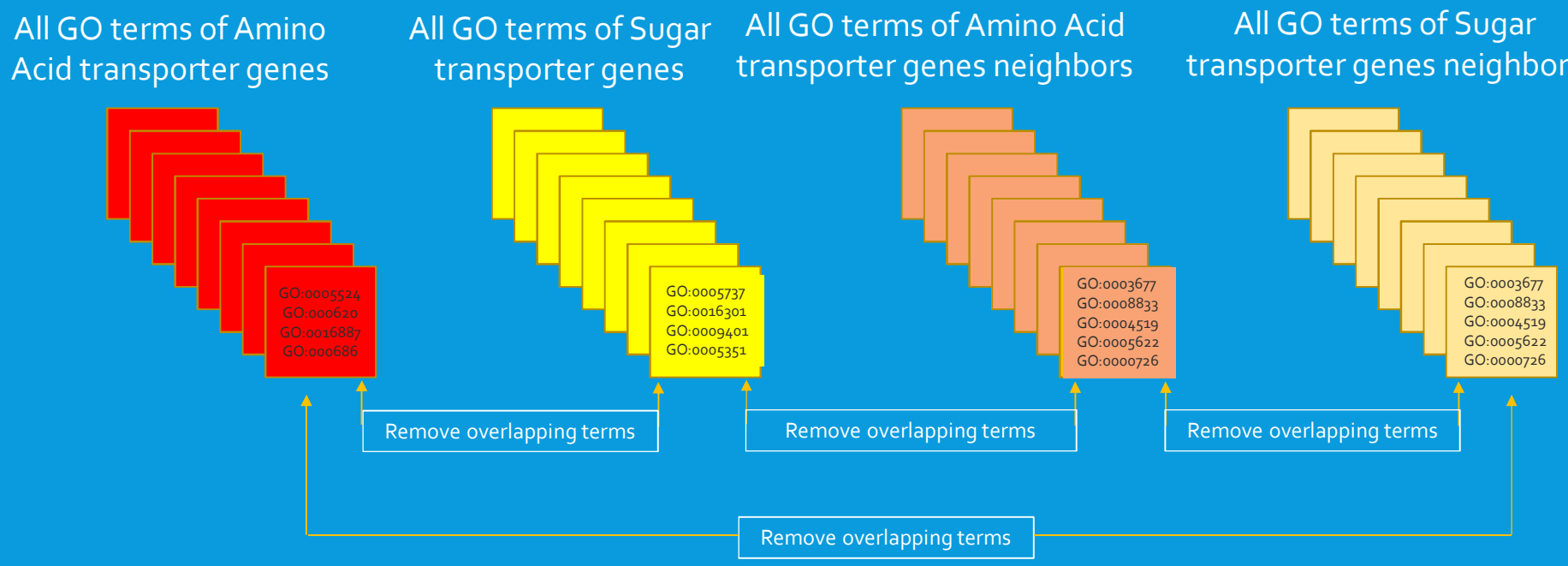
All GO terms of Amino Acid transporter genes neighbors



All GO terms of Sugar transporter genes neighbors

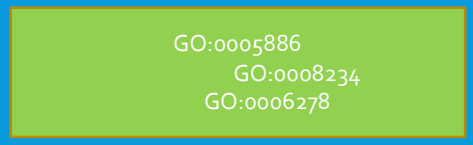


REMOVE OVERLAPPING TERMS

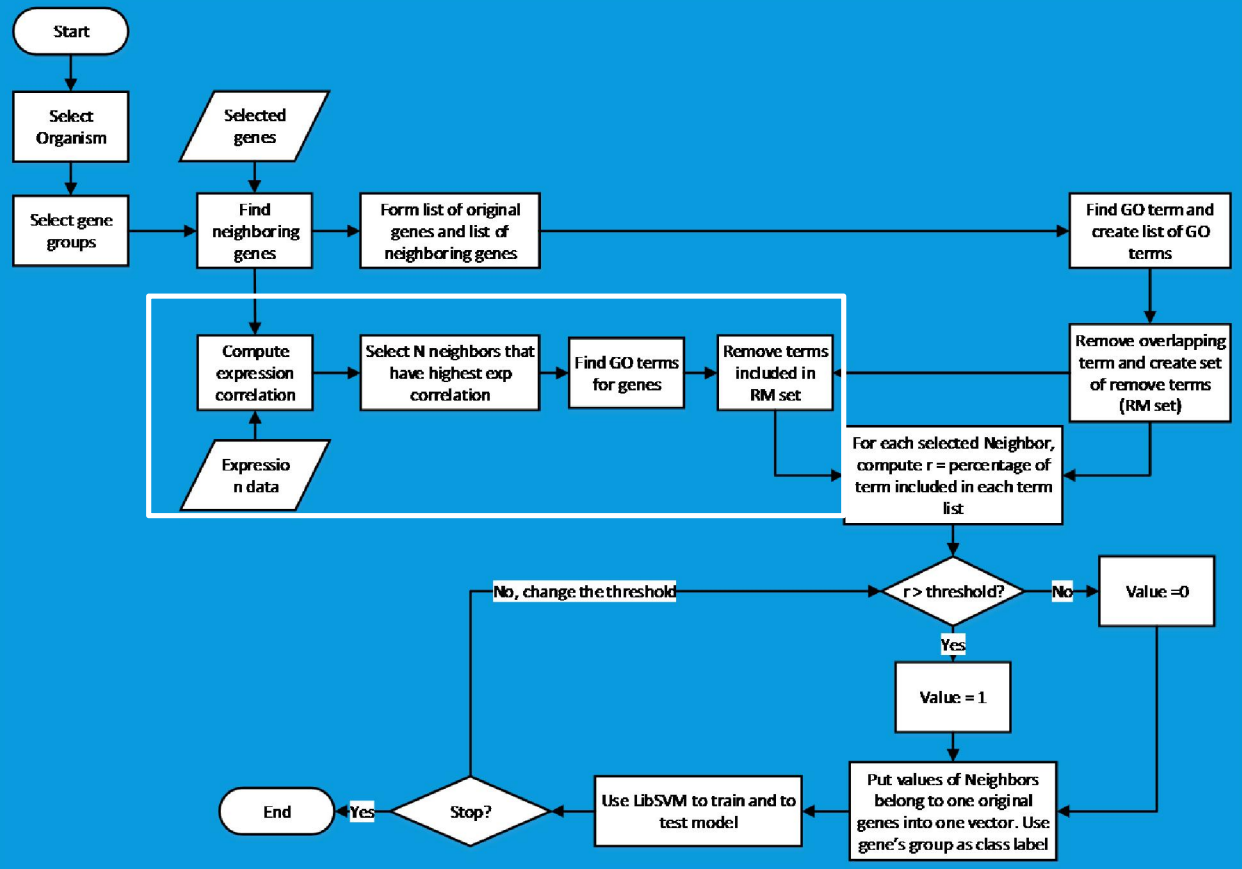


COLLECT ALL REMOVED TERMS

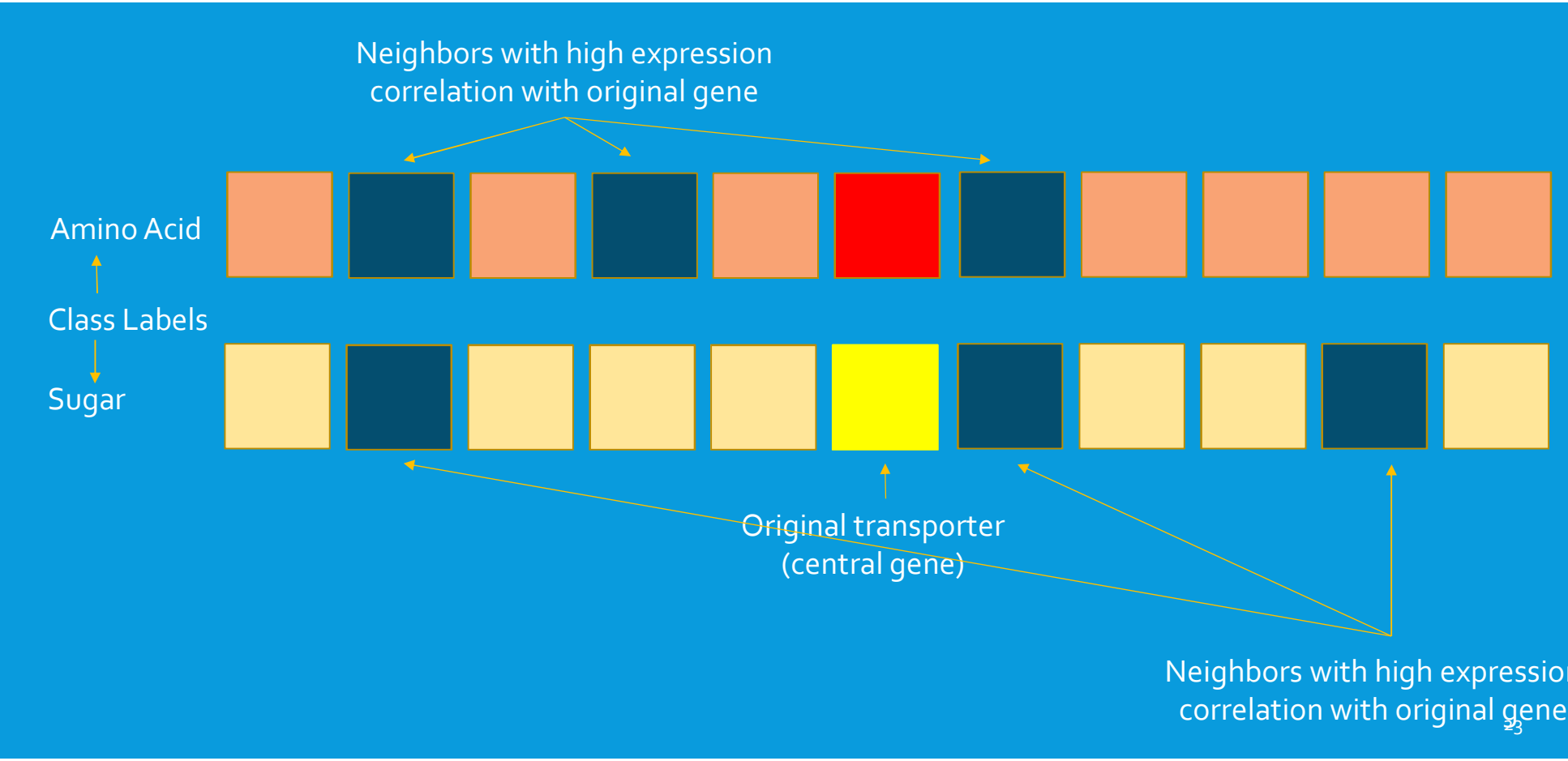
All removed terms
(RM set)



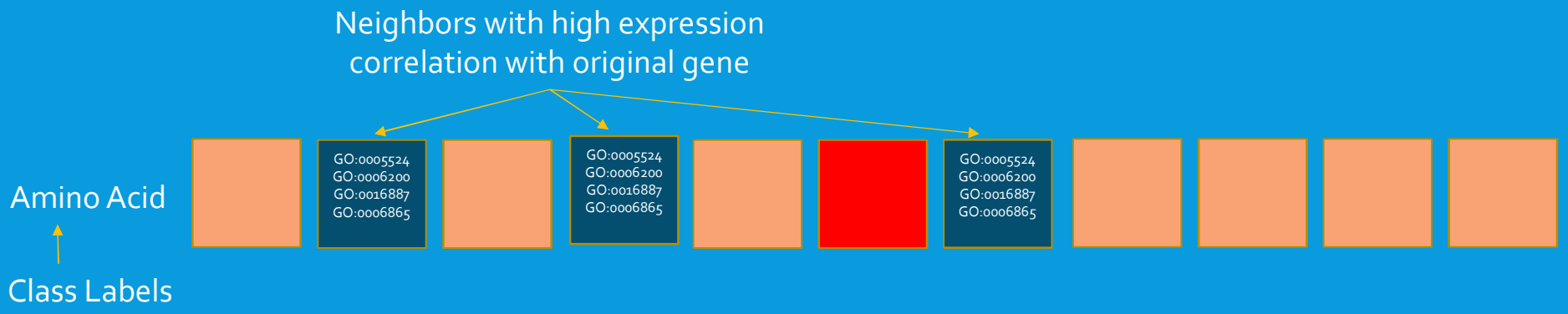
WORKFLOW


Start
Select Organism
Selected genes
Select gene groups
Find neighboring genes
Form list of original genes and list of neighboring genes
Find GO term and create list of GO terms
Compute expression correlation
Select N neighbors that have highest exp correlation
Find GO terms for genes
Remove terms included in RM set
Remove overlapping term and create set of remove terms (RM set)
Expressio n data
For each selected Neighbor, compute r = percentage of term included in each term list
No, change the threshold
r > threshold?
No
Value =0
Yes
Value = 1
Put values of Neighbors belong to one original genes into one vector. Use gene's group as class label
Use LibSVM to train and to test model
Stop?
Yes
End

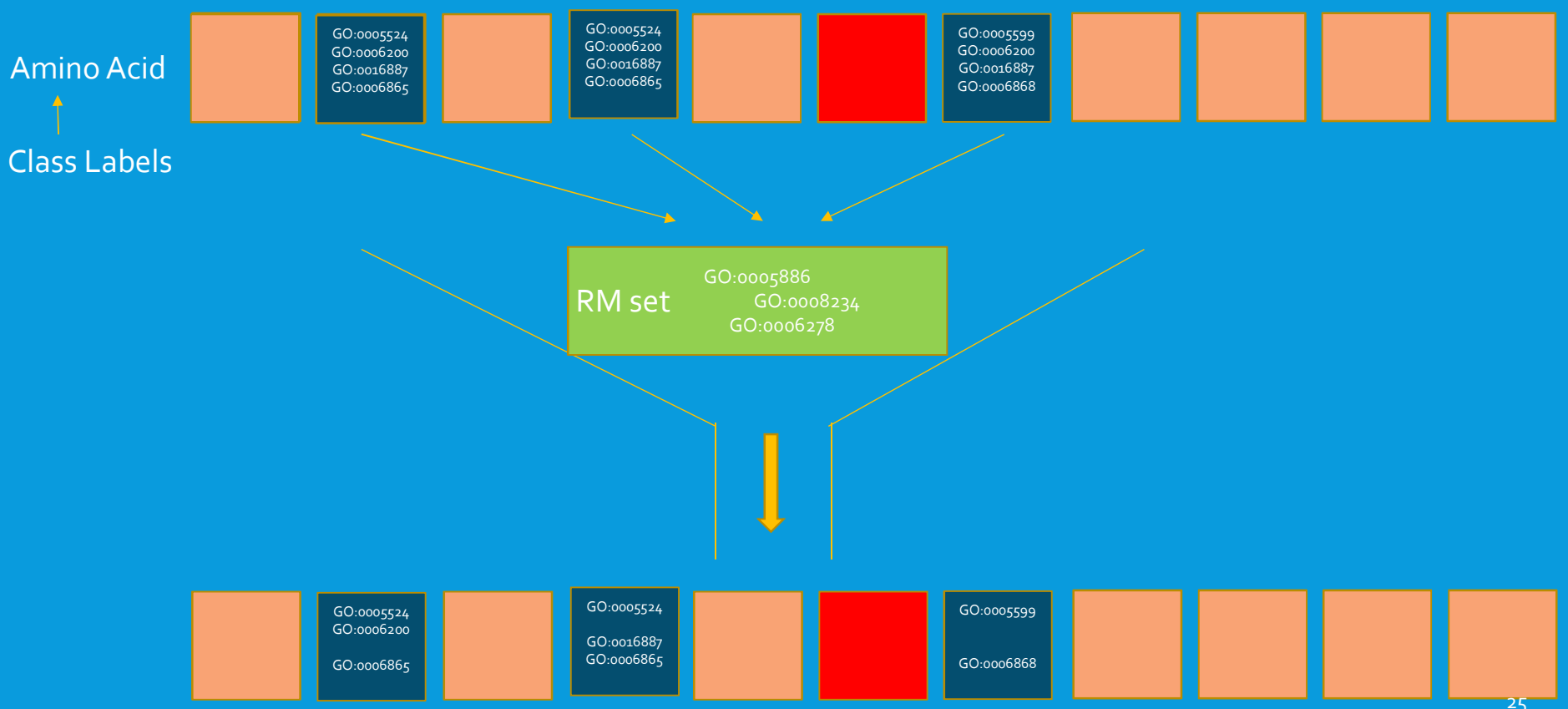
FIND HIGHLY CO-EXPRESSED NEIGHBORS



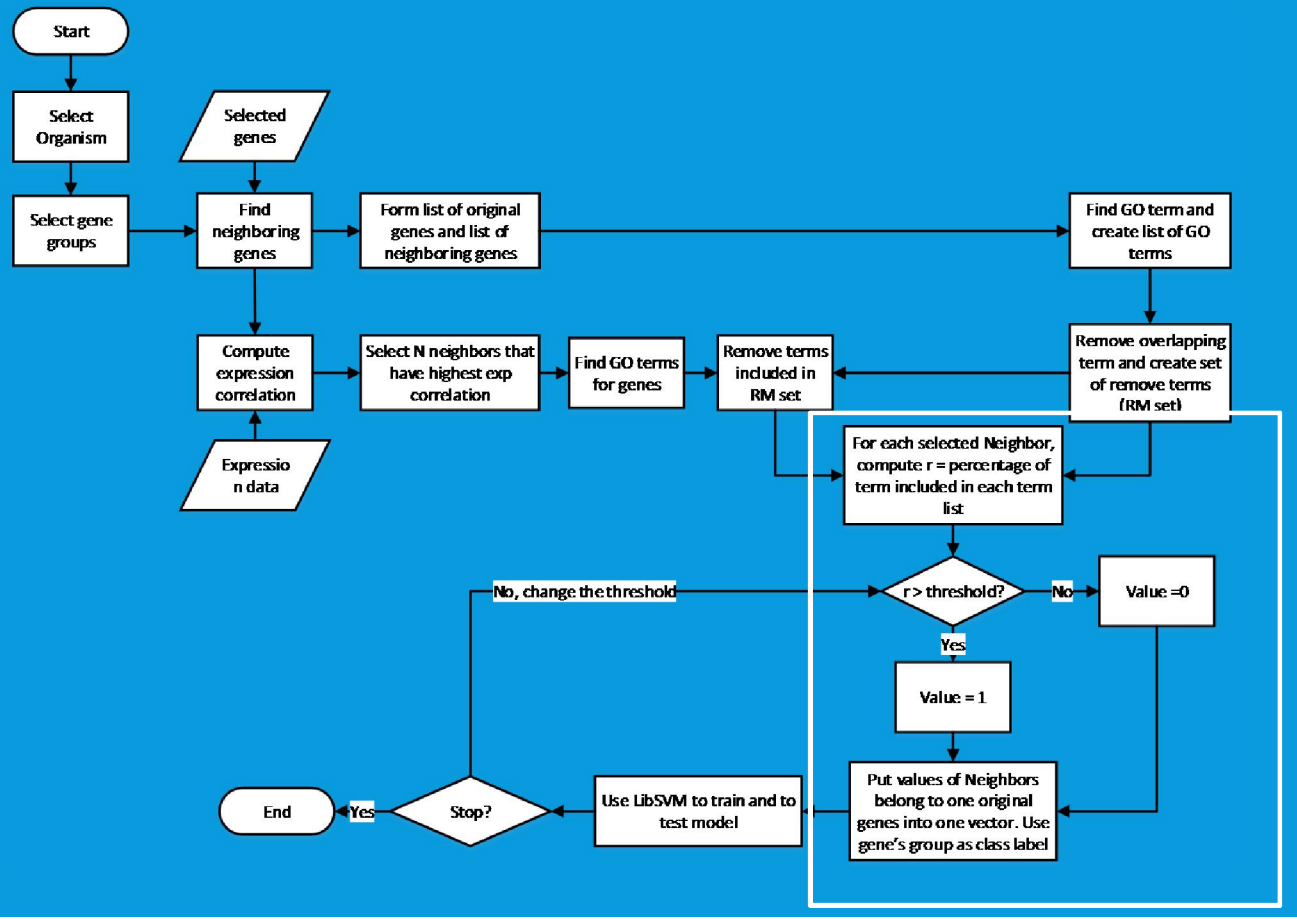
FIND GO TERMS FOR SELECTED NEIGHBORS



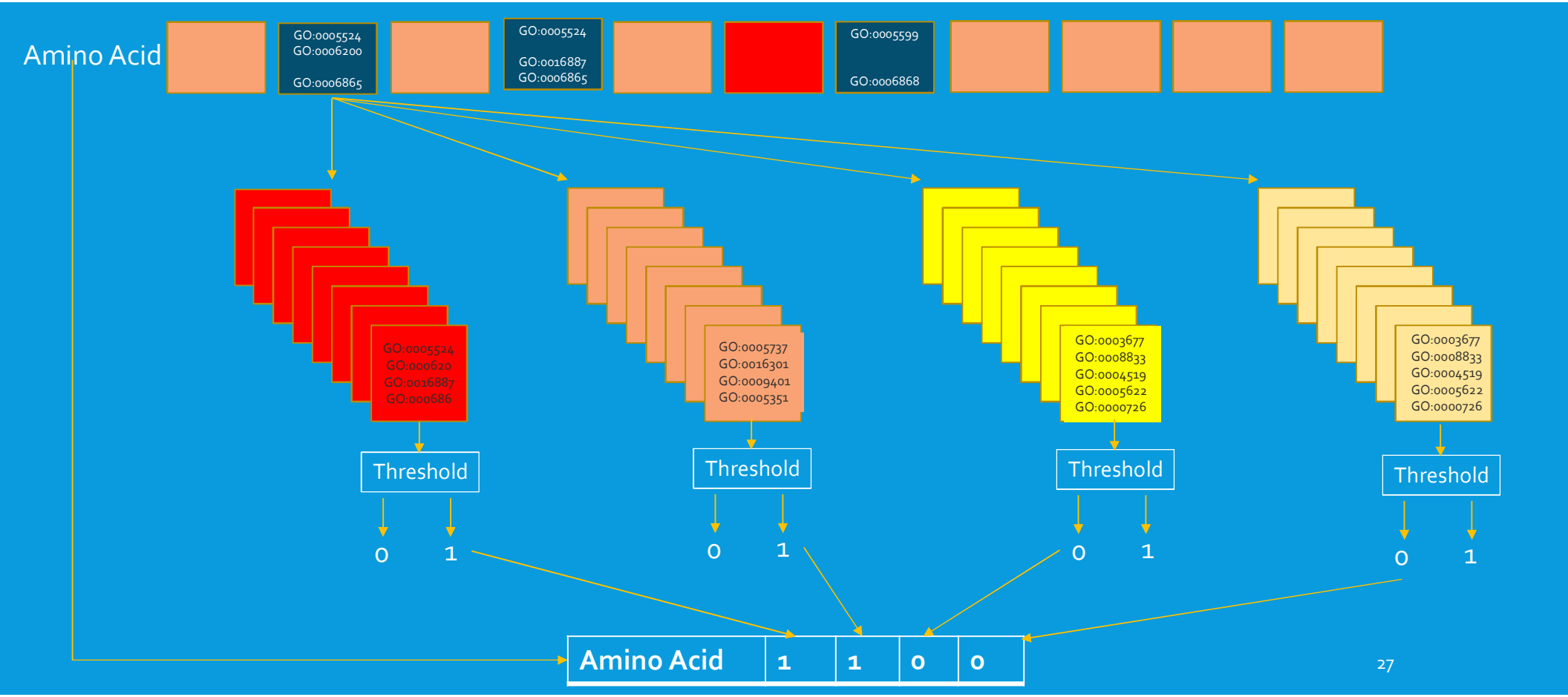
APPLY FILTER FOR SELECTED NEIGHBORS



WORKFLOW


Start
Select Organism
Selected genes
Select gene groups
Find neighboring genes
Form list of original genes and list of neighboring genes
Find GO term and create list of GO terms
Compute expression correlation
Select N neighbors that have highest exp correlation
Find GO terms for genes
Remove terms included in RM set
Remove overlapping term and create set of remove terms (RM set)
Expressio n data
For each selected Neighbor, compute r = percentage of term included in each term list
r > threshold?
No
Value =0
No, change the threshold
Yes
Value = 1
Put values of Neighbors belong to one original genes into one vector. Use gene's group as class label
Use LibSVM to train and to test model
Stop?
Yes
End

CREATE DATA FOR SVM



DATA FOR SVM

Class	Neighbor 1				Neighbor 2				Neighbor 3			
	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4
Amino acid	0	1	0	0	0	1	0	0	0	1	0	0
Amino acid	0	0	0	0	0	1	0	0	0	1	0	0
Amino acid	0	1	0	0	0	1	0	0	0	1	0	0
Amino acid	0	1	0	0	0	1	0	0	0	1	0	0
Sugar	0	0	0	1	0	0	0	1	0	0	0	0
Sugar	0	0	0	1	0	0	0	1	0	0	0	0
Sugar	0	0	1	1	0	0	0	0	0	0	0	0
Sugar	0	0	0	0	0	0	1	1	0	0	0	0

List 1 : All GO terms of AA

List 2 : All GO terms of AA's neighbors

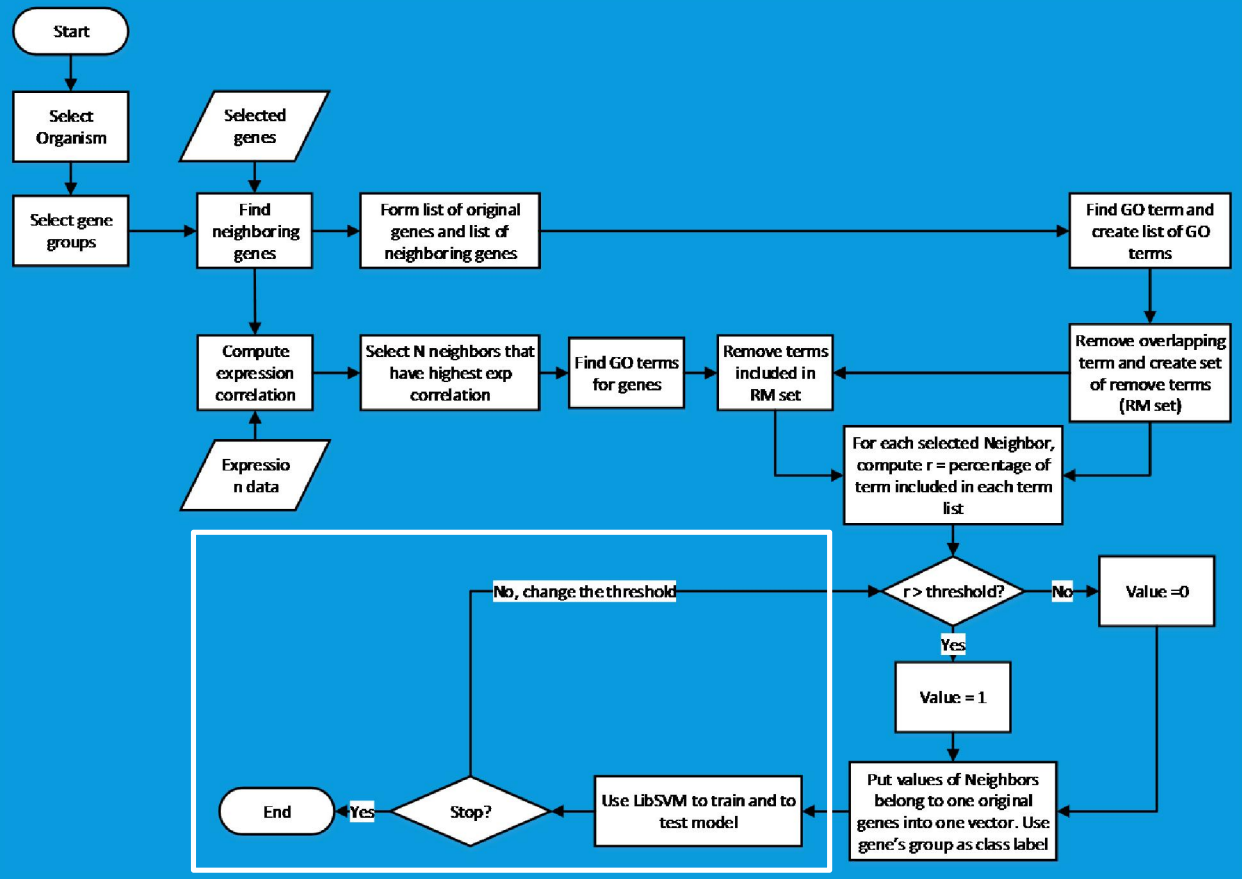
List 3 : All GO terms of Sugar

List 4 : All GO terms of Sugar neighbors

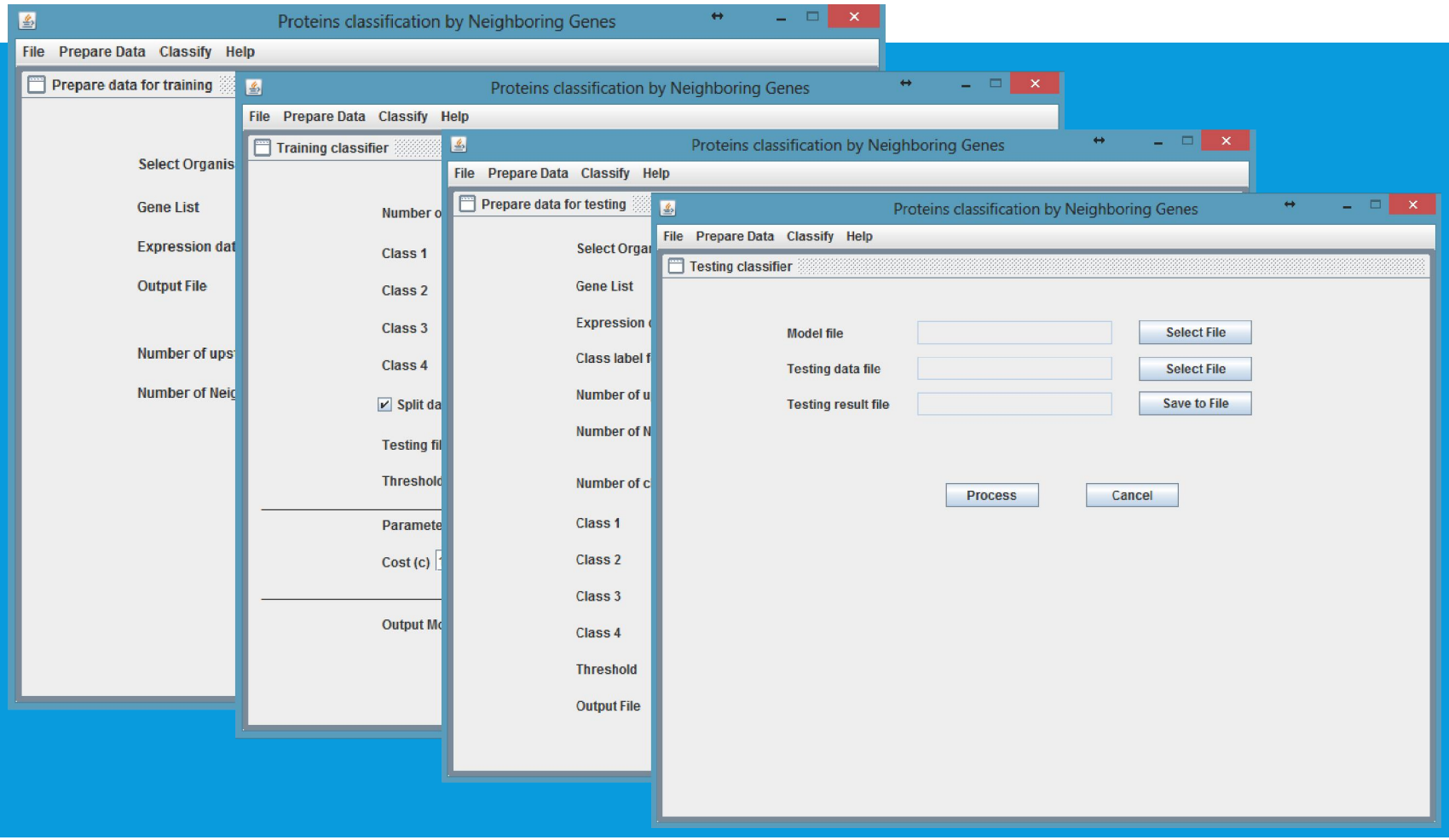
LIBSVM INPUT

```
1    1:0 2:1 3:0 4:0 5:0 6:0 7:0 8:0 9:0 10:0    11:0    12:0
1    1:0 2:1 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:1    11:0    12:0
1    1:0 2:0 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:0    11:0    12:0
1    1:0 2:1 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:1    11:0    12:0
1    1:0 2:0 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:1    11:0    12:0
1    1:0 2:1 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:1    11:0    12:0
1    1:0 2:1 3:0 4:0 5:0 6:1 7:0 8:0 9:0 10:1    11:0    12:0
2    1:0 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0 10:0    11:0    12:0
2    1:0 2:0 3:1 4:0 5:0 6:0 7:0 8:1 9:0 10:0    11:0    12:0
2    1:0 2:0 3:0 4:1 5:0 6:0 7:0 8:0 9:0 10:0    11:0    12:1
2    1:0 2:0 3:0 4:1 5:0 6:0 7:1 8:1 9:0 10:0    11:0    12:0
2    1:0 2:0 3:0 4:1 5:0 6:0 7:0 8:1 9:0 10:0    11:0    12:1
2    1:0 2:0 3:1 4:1 5:0 6:0 7:0 8:0 9:0 10:0    11:0    12:1
2    1:0 2:0 3:0 4:1 5:0 6:0 7:0 8:0 9:0 10:0    11:0    12:1
```

WORKFLOW


Start
Select Organism
Selected genes
Select gene groups
Find neighboring genes
Form list of original genes and list of neighboring genes
Find GO term and create list of GO terms
Compute expression correlation
Select N neighbors that have highest exp correlation
Find GO terms for genes
Remove terms included in RM set
Remove overlapping term and create set of remove terms (RM set)
Expressio n data
For each selected Neighbor, compute r = percentage of term included in each term list
No, change the threshold
r > threshold?
No
Value =0
Yes
Value = 1
End
Yes
Stop?
Use LibSVM to train and to test model
Put values of Neighbors belong to one original genes into one vector. Use gene's group as class label

THE TOOL


Proteins classification by Neighboring Genes
File Prepare Data Classify Help
Prepare data for training
Select Organis
Gene List
Expression dat
Output File
Number of ups
Number of Neig
Proteins classification by Neighboring Genes
File Prepare Data Classify Help
Training classifier
Number o
Class 1
Class 2
Class 3
Class 4
Split da
Testing fil
Threshold
Paramete
Cost (c)
Output Mo
Proteins classification by Neighboring Genes
File Prepare Data Classify Help
Prepare data for testing
Select Orga
Gene List
Expression
Class label f
Number of u
Number of N
Number of c
Class 1
Class 2
Class 3
Class 4
Threshold
Output File
Proteins classification by Neighboring Genes
File Prepare Data Classify Help
Testing classifier
Model file
Select File
Testing data file
Select File
Testing result file
Save to File
Process
Cancel

TESTING RESULT

Features	E.Coli	S.Cerevisiae
Gene group:	Amino acid + sugar transporter genes	Amino acid + sugar transporter genes
Training set	27 (14 AA + 13 Sugar)	17 (10 AA + 7 Sugar)
Test set	23 (12 AA + 11 Sugar)	14 (9 AA + 5 Sugar)
Threshold for r	0.8	0.8
Accuracy	87% (20/23)	100% (14/14)

TESTING RESULT

Features	E.Coli	S.Cerevisiae
Gene group:	Amino acid-, sugar- and metal transporter genes	Amino acid-, sugar- and metal transporter genes
Training set	31(14 AA + 6Metal+13 Sugar)	25 (10 AA + 8Metal + 7 Sugar)
Test set	27 (12 AA + 4 Metal+1 1Sugar)	21 (9 AA + 7 Metal + 5 Sugar)
Threshold for r	0.8	0.5
Accuracy	74% (20/27)	90.47% (19/21)

TESTING RESULT

- Human testing

Features	Model built from E.Coli	Model built from S.Cerevisiae
Gene group:	Amino acid + sugar transporter genes	Amino acid + sugar transporter genes
Training set	50 (26 AA + 24 Sugar)	31 (19 AA + 12 Sugar)
Test set (human)	15 (9 AA + 6 Sugar)	15 (9 AA + 6 Sugar)
Threshold for r	0.1	0.1
Accuracy	40% (6/15)	66.66% (10/15)

4. OUTLOOK

- Test our method with other gene groups and other organisms
- Extend the tool to work with more organisms
- Take into account the distance between genes, gene orientation and gene duplication
- Classification with AdaBoost, Artificial neural network…
- Develop an web-based tool

5.REFERENCES

- Schaadt NS, Christoph J, Helms V: **Classifying substrate specificities of membrane transporters from Arabidopsis thaliana**. *Journal of chemical information and modeling* 2010, **50**(10):1899-1905.
- Jacob F, Perrin D, Sanchez C, Monod J: **[Operon: a group of genes with the expression coordinated by an operator]**. *Comptes rendus hebdomadaires des seances de l'Academie des sciences* 1960, **250**:1727-1729.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nature genetics* 2000, **25**(1):25-29.
- Chih-Chung Chang and Chih-Jen Lin: **LIBSVM : a library for support vector machines**. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011

5.REFERENCES

- http://cancergenome.nih.gov/
- http://en.wikipedia.org/wiki/Operon
- http://www.genome.jp/
- http://www.ncbi.nlm.nih.gov/gds
- http://www.uniprot.org/

- Pauli Miettinen's lecture note

THANK YOU FOR YOUR ATTENTION!!!