A Predictive Model To Estimate The Level Of Recombinant Protein Overexpression In *Escherichia Coli*

By Narges Habibi

Supervisor: Assoc. Prof. Dr Siti Zaiton Mohd Hashim (Faculty of Computing, UTM)

Co-supervisor: Prof. Mohd Razip Samian (School of Biological Sciences, USM)

Outlines

- Introduction
- Literature Review
- Methodology
- Results & Discussion
- Conclusion

Introduction

- Recombinant Protein Overexpression
- Problem Background
- Problem Statement
- Research Goal
- Research Objectives
- Research Scope
- Research Significance

Recombinant Protein Overexpression

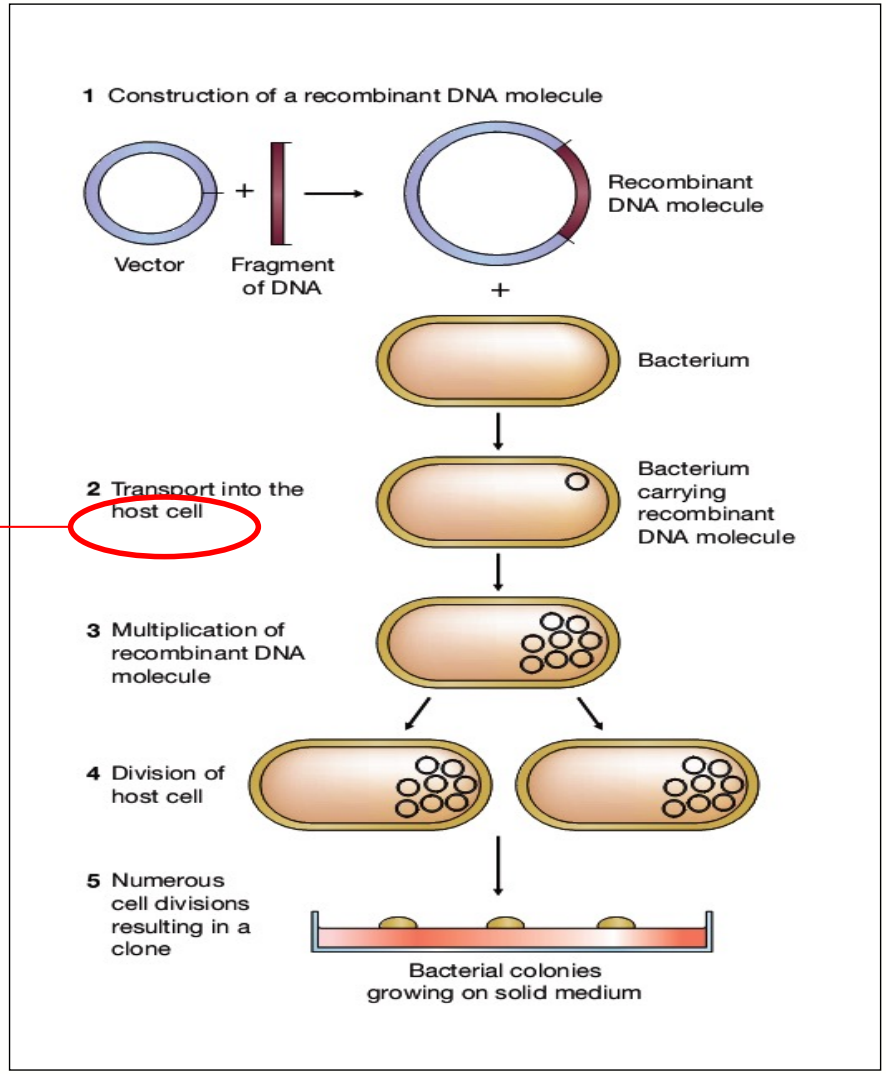
- Recombinant protein overexpression is a significant biotechnological process as it allows researchers to produce a specific protein in desired quantity.

- *Escherichia coli* (*E. coli*) bacteria is the major expression host used for recombinant protein expression. Approximately 30% of all of the recombinant pharmaceutical proteins are produced in *E. coli* (Kucharova, 2012).

Recombinant Protein Overexpression (Cont.)

- Steps of recombinant

Protein overexpression:

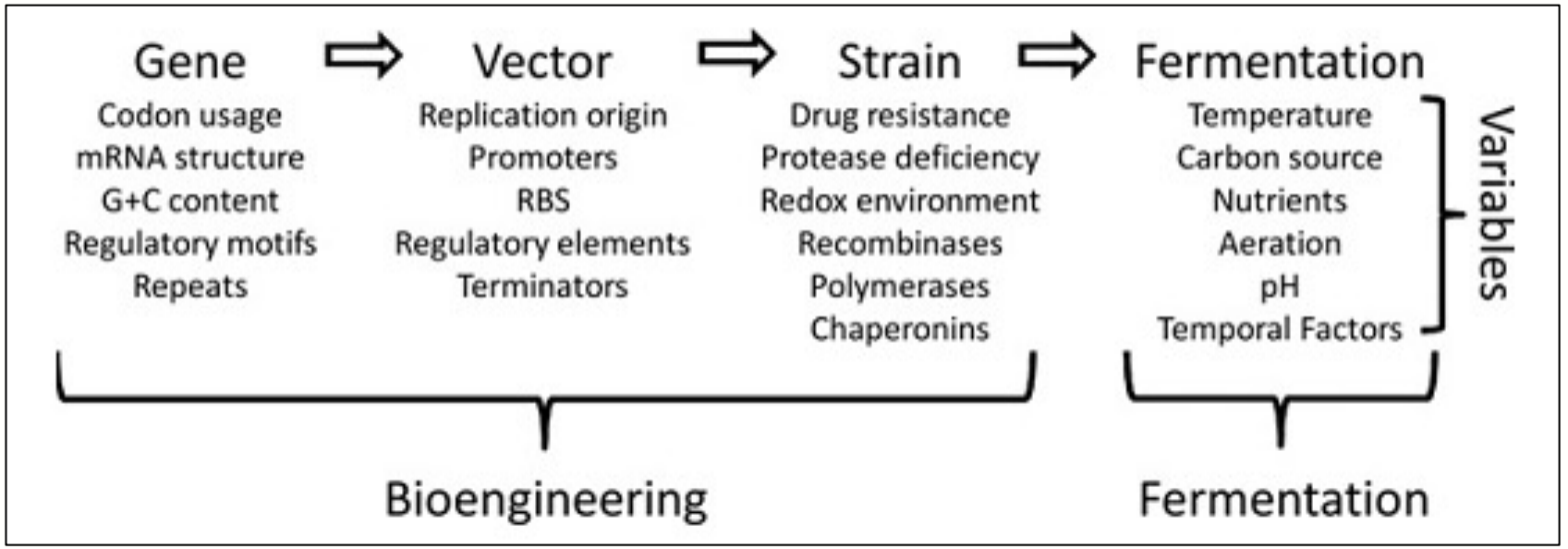


Example: E. coli

(Brown, 2010)

Recombinant Protein Overexpression (Cont.)

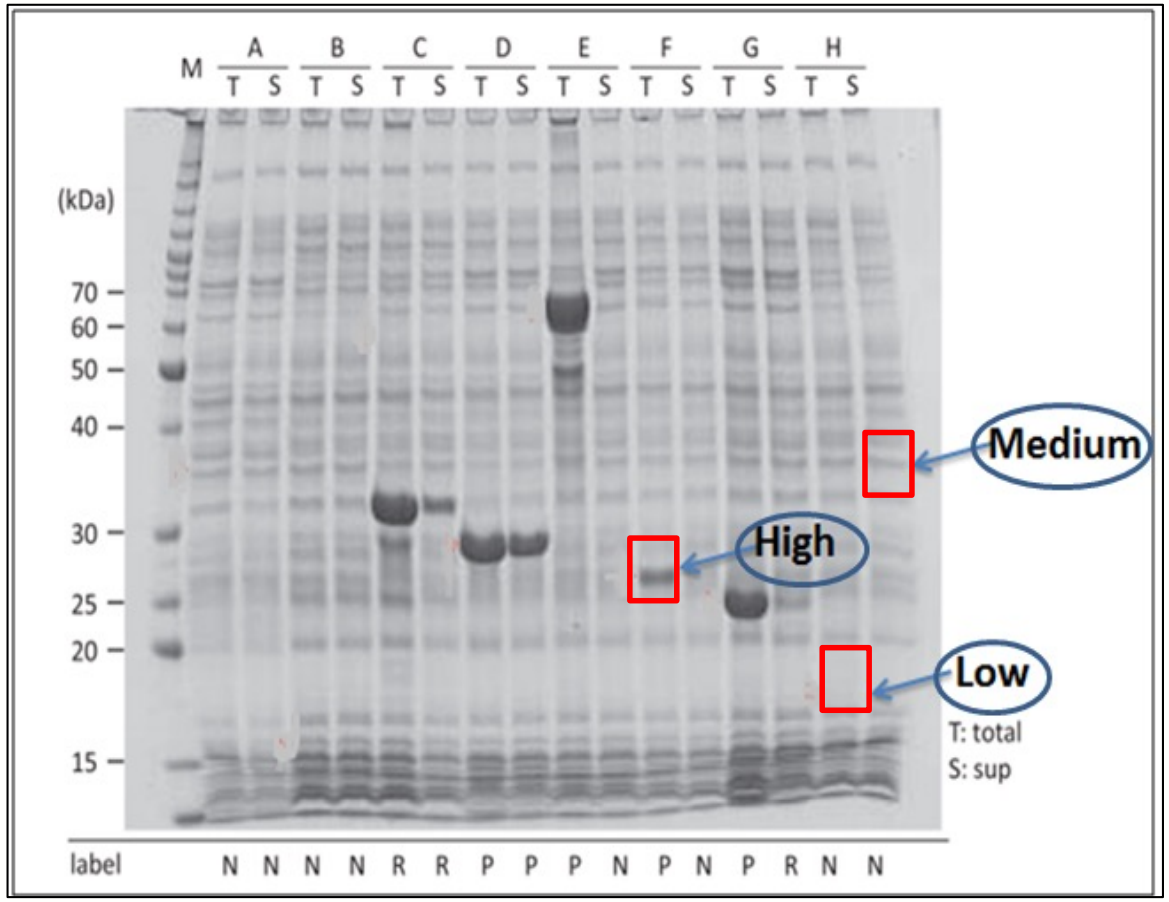
- Influential factors on RPO:



(Gustafsson et al., 2012)

Recombinant Protein Overexpression (Cont.)

- The level of RPO:



(Hirose et al., 2011)

Problem Background

- Being able to conduct the theoretical prediction of protein overexpression level will aid developing the large-scale proteomics studies (Chan et al., 2010, Hirose et al., 2011, 2013, Chang et al., 2013).

- There are some rules to anticipate the overexpression level of a recombinant protein before the actual experiment, but due to the existence of several parameters and their complicated relationships, which are partially known, it is not possible to know the expected level of overexpression in advance.

- To the best of our knowledge, no previous study has addressed the problem of recombinant protein overexpression level prediction.

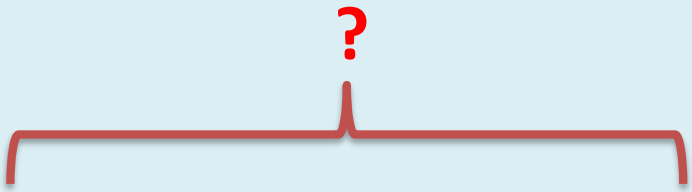
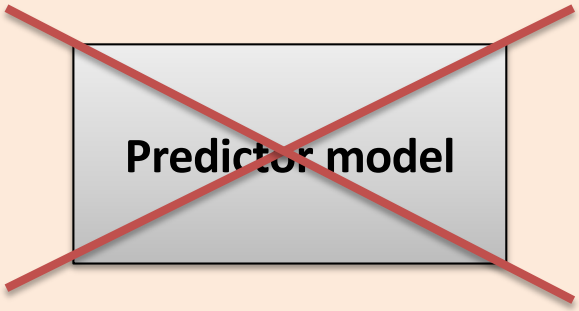
Problem Statement

- Predicting the recombinant protein overexpression level prior to perform the real laboratory experiments, is hardly mentioned in the literature. The ability to predict the level would lead to reduced labour, time and cost.

Research Goal

- The goal of this research is to **predict** the **overexpression level** of a recombinant protein in *E. coli*, based on its **sequence**, as well as the expression **vector** and expression **host** employed, using a **machine learning-based model**.

Research Objectives

#	Problem	Objective
1		**Obj1:** To construct a new dataset of recombinant protein overexpression in *E. coli*, containing the gene sequence, vector, host and the expression level.
2	 {Feature_1, Feature_2, …, Feature_N}	**Obj2:** To identify the influential features on recombinant protein overexpression level in *E. coli*, using the constructed dataset.
3		**Obj3:** To develop a predictive model using the identified features to estimate the overexpression level of a recombinant protein in *E. coli*, based on machine learning techniques.

Research Scope

Issue	Scope	Reason
Expression host	**E. coli**	The most common host used (Kucharova, 2012).
Overexpression result	**{Low, Medium, High}**	To simplify the modelling problem.
Dataset	**In-house developed dataset: "EcoliOverExpressionDB"**	No available dataset with all the required fields (features) for this research (i.e. gene sequence, vector, host, expression level).
Predictive model development	**Machine learning approach**	The most common approaches in the related researches.
Model evaluation	**K-fold cross validation + Common statistical metrics for classification (e.g. F-score)**	The most common method and metrics in the related researches.

Research Significance

- The findings of this research can be useful to estimate the recombinant overexpression level in *E. coli* before doing real laboratory experiment. It helps a biologist to decide whether to perform a specific experiment, and hence to decrease the involved time, effort and expense.

Literature Review

- Classification Concepts
  - Features
  - Feature Selection
  - Multi-class classification
  - Class Imbalance learning
  - Ensemble Learning
  - Random Forest
- Comparative Study
- Challenges of This Research

Classification Concepts

- Features:
  - A feature is an individual measurable heuristic property of a phenomenon being observed;
  - The set of features of a given data instance is often grouped into a feature vector.

- Feature selection:
  - Reducing the feature space dimensionality .

- Multi-class classification:
  - A classification task with more than two classes/categories;
  - Harder, relative to the two-class situation.

Classification Concepts (Cont.)

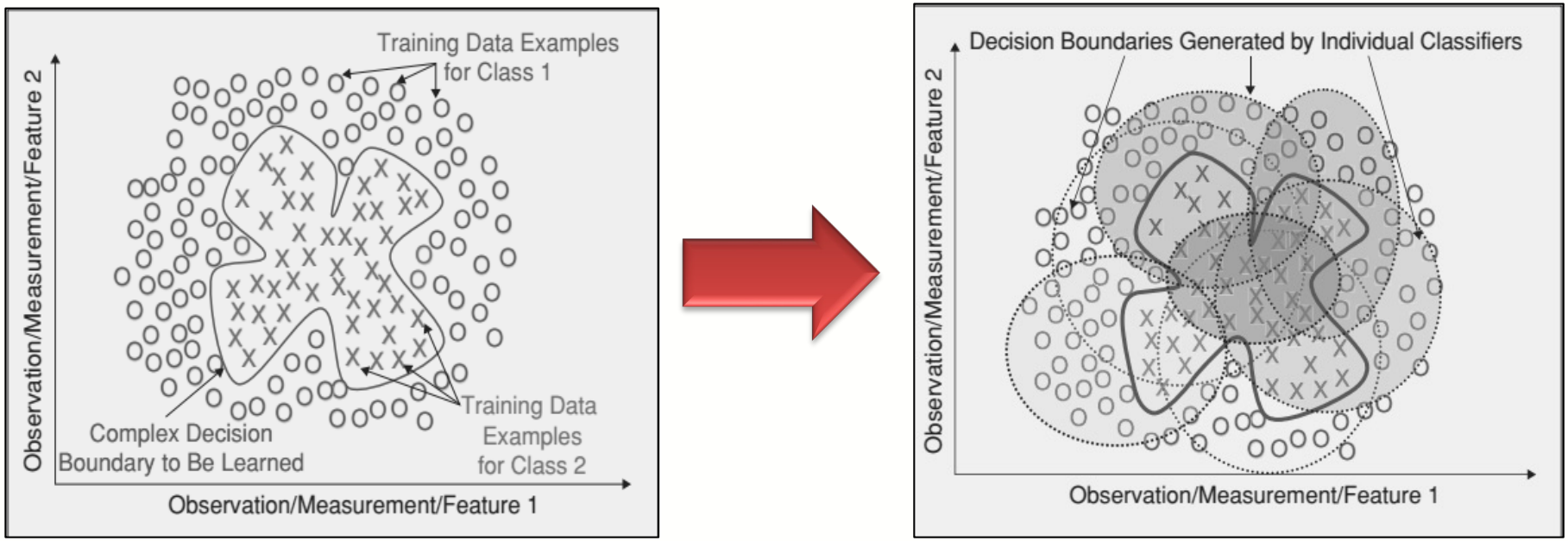
- Class Imbalance learning:
  - A classification problem, where some classes are highly underrepresented compared to other classes.
  - The skewed distribution causes many machine learning algorithms perform less effective, particularly in predicting minority class.

Classification Concepts (Cont.)

- Class Imbalance learning (Cont.):
  - The learning objective: "obtaining a classifier with high accuracy for the minority class without critically endangering the accuracy of the majority class".
  - Numerous solutions have been proposed at the data and algorithm levels to deal with class imbalance (e.g. ensemble learning).

Classification Concepts (Cont.)

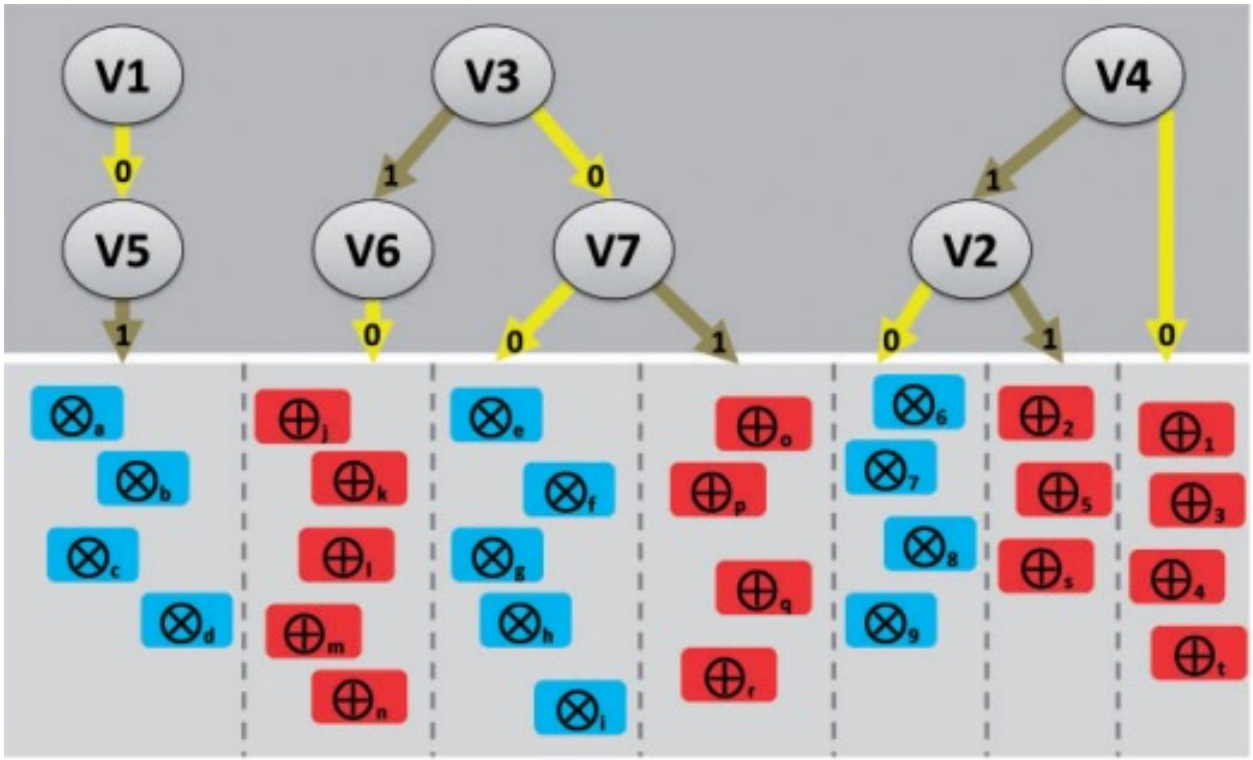
- Ensemble learning:



(a) Complex decision boundary (b) Ensemble of classifiers spanning the decision space (Polikar, 2006)

Classification Concepts (Cont.)

- Random forest:



Classification Concepts (Cont.)

- Random forest (Cont.)
  - Benefits from a computational viewpoint:
    - It can deal with both regression and (multiclass) classification;
    - It is relatively fast to train and to use (make prediction);
    - It depends on very few tuning parameters;
    - It has a built in estimate of generalization error;
    - It can be employed directly for high-dimensional problems;
    - It can simply be implemented in parallel.

Classification Concepts (Cont.)

- Random forest (Cont.):
  - Benefits from a statistical viewpoint:
    - Measures of variable importance;
    - Differential class weighting;
    - Missing value imputation;
    - Visualization;
    - Outlier detection;
    - Unsupervised learning.

Comparative Study

- The most similar previous studies to our research:
  - Recombinant protein overexpression prediction to estimate whether the given gene sequence will be expressed or no;
  - It is a two-class prediction problem.

Comparative Study

#	Reference	Dataset(s)	Feature Selection Method(s)	Modeling Technique(s)	Performance
1	(Hirose and Noguchi, 2013)	HGPD *E. coli* Size: 7768 Expressed: 4102 Non-expressed: 3666	Filter: Student's t-test	Two techniques: 1. Support vector machine 2. Sequence pattern-based method	Accuracy=82% AUC=87%
2	(van den Berg et al., 2012)	Van den berg Homologous: Size: 345 Expressed: 178 Non-expressed: 167 Heterologous: Size: 991 Expressed: 163 Non-expressed: 828	Filter: Student's t-test	Support vector machine	Accuracy=N/A AUC=80%

Comparative Study (Cont.)

#	Reference	Dataset(s)	Feature Selection Method(s)	Modeling Technique(s)	Performance
3	(Hirose et al., 2011)	HGPD *E. coli* Size: 13180 Expressed: 7744 Non-expressed: 5436	Filer: Student's t-test	Random forest	Accuracy=78% AUC=N/A
4	(Chan et al., 2010)	121 genes from different species were expressed in 6 different vectors. Size: 726 Soluble: 231 Insoluble: 236 Non-expressed: 259	Feature selection package in LIBSVM: filter (F-score) + wrapper (SVM)	Support vector machine	Accuracy=83% AUC=89%

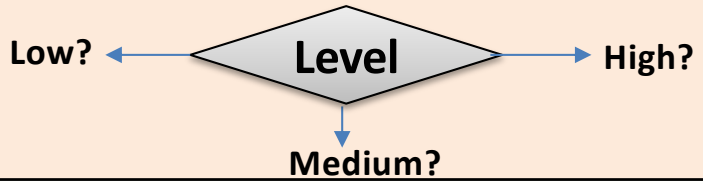
Comparative Study (Cont.)

#	Reference	Dataset(s)	Feature Selection Method(s)	Modeling Technique(s)	Performance
5	(van den Berg et al., 2010)	Homologous proteins from *A. niger* CBS 513.88 with a signal sequence predicted by SignalP. Size: 638 Expressed: 268 Non-expressed: 370	Two techniques: 1. Filer: Student's t-test 2. Wrapper: backward	Linear discriminant	Accuracy=0.84% AUC=N/A
6	(Luan et al., 2004)	ORFs of *C. elegans* with one expression vector and one Escherichia coli strain. Size: 10167 Expressed: 4854 Non-expressed: 5313	Filter: linear correlation coefficient (LCC)	-	N/A

Comparative Study (Cont.)

#	Reference	Dataset(s)	Feature Selection Method(s)	Modeling Technique(s)	Performance
7	(Goh et al., 2004)	TargetDB Size: 27000	Embedded: random forest	Decision tree	Accuracy=76% AUC=N/A
8	(Christendat et al., 2000)	SPINE	-	Decision tree	N/A

Challenges of This Research

#	Challenge	Solution
1	? { Feature_1, Feature_2, …, Feature_N }	Finding a large set of potential features.
2	{ large feature set }	Using feature extraction and selection techniques.
3	Small dataset 	Using small sample size techniques, e.g. ensemble learning.
4	Imbalance dataset 	Using appropriate data-level and algorithm-level techniques, e.g. ensemble learning.
5	Multi-class classification problem 	Finding an appropriate classifier by intensive comparison.

Methodology

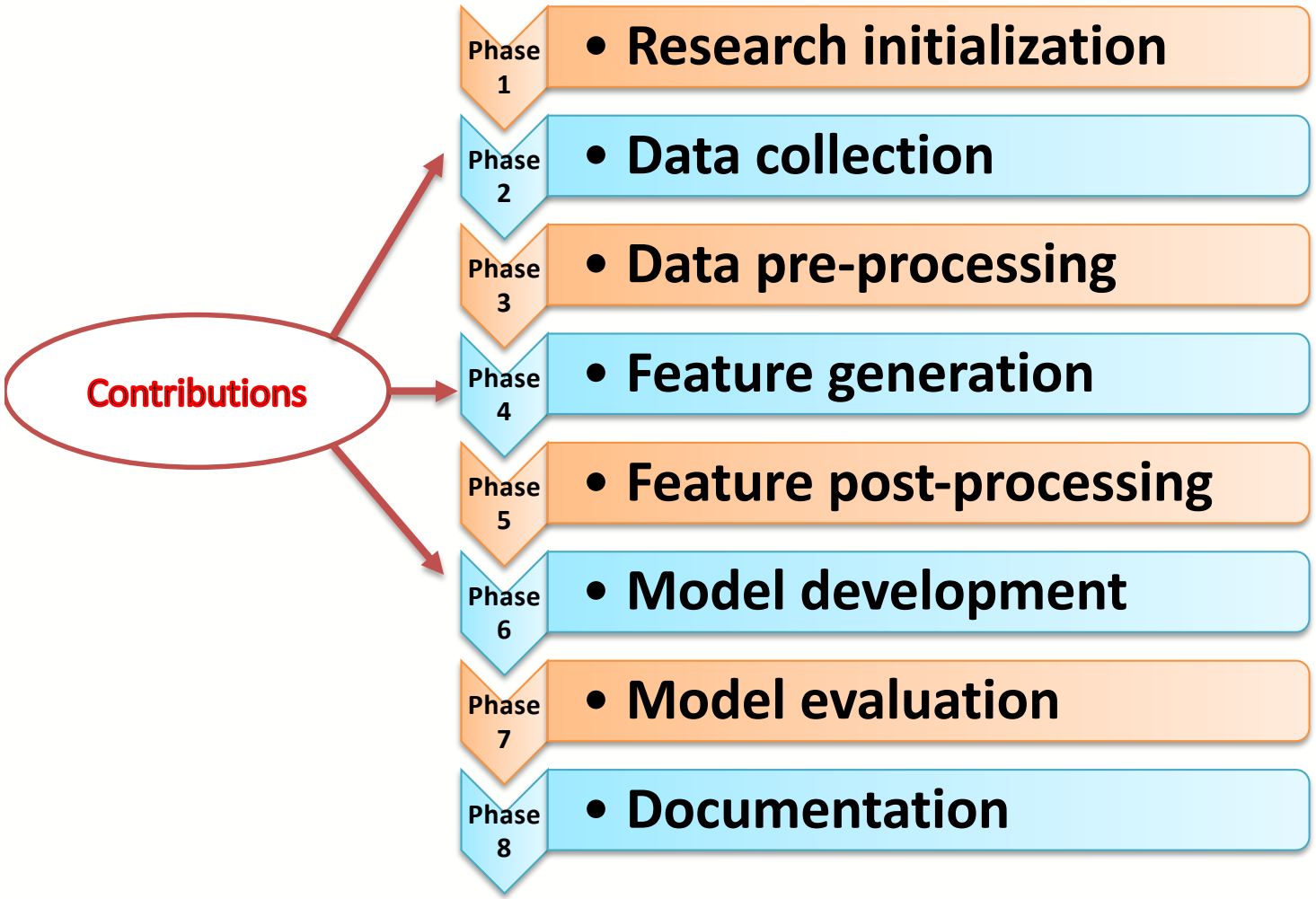
- Problem Formulation
- Research Framework
- Database Establishment
- Feature Set Generation
- Predictive Model Development
- Predictive Model Evaluation

Problem Formulation

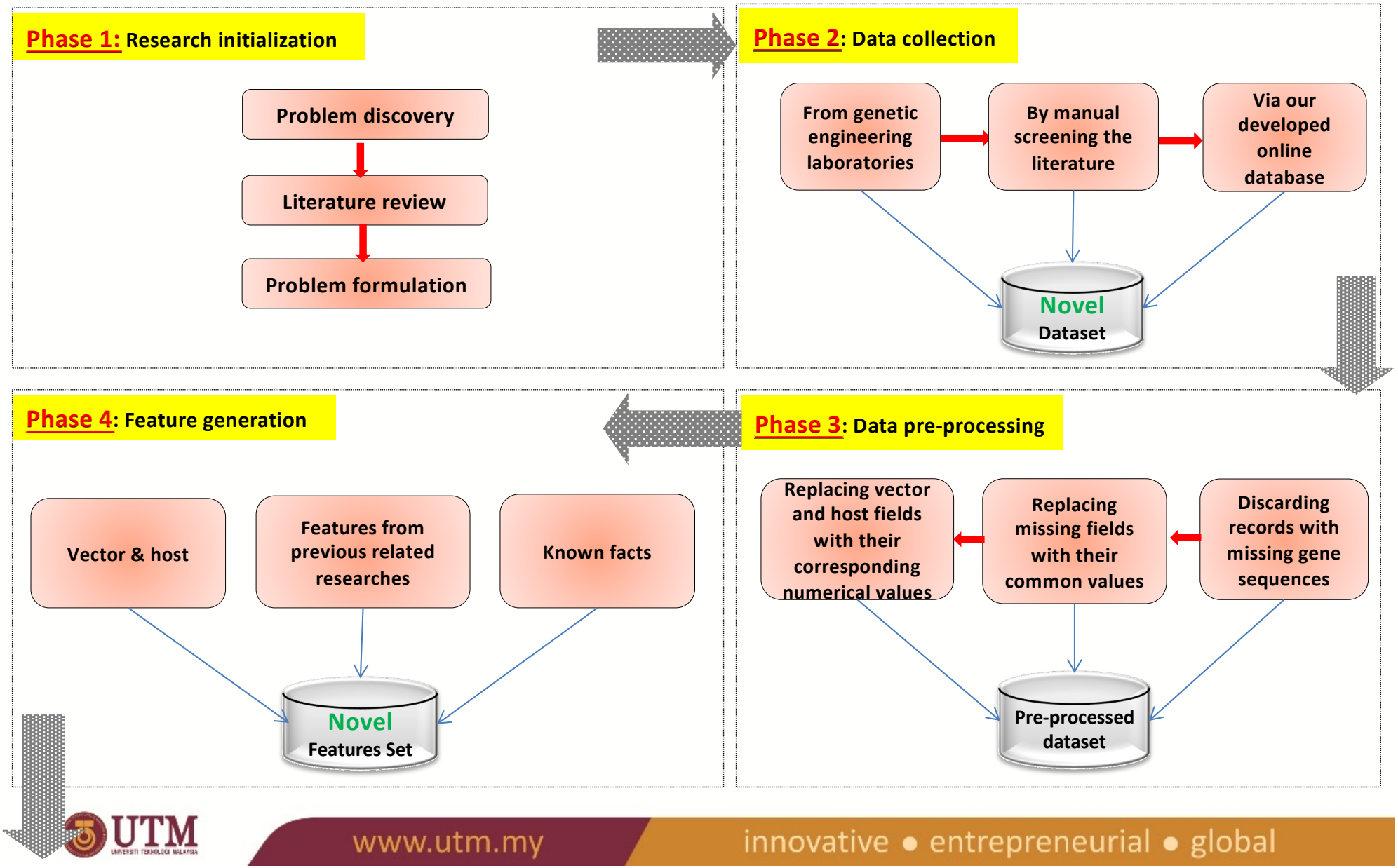
- The problem of this research can be formulated as follows:

"Given a recombinant gene sequence "*S*", vector "*V*" and *E. coli* host strain "*H*", predict the overexpression level of "*S*", carried by "*V*", in "*H*".

Research Framework



Research Framework (Cont.)



Research Framework (Cont.)

Phase 5: Feature post-processing

- Normalizing the features to have mean=0 and standard deviation=1
- Scaling the features by liner transformation
- Post-processed Features Set

Phase 6: Model development

- Evalauting several classifiers and feature representation methods
- Determining the best classifier and feature representation method
- Tuning model parameters
- Novel predictor model

Phase 7: Model evaluation

- Computational evaluation

Phase 8: Documentation

- Paper writing on the developed database
- Paper writing on the reviewed literature
- Paper writing on the proposed model
- Thesis writing

Database Establishment

Data collection form UTM & USM genetic engineering laboratories

Data extraction from the literature

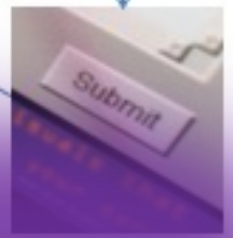
Initial Dataset

Establishing an online database system

Inviting the researchers to submit their records online

http://

EcoliOverExpressionDB



Validation by an biotechnology expert

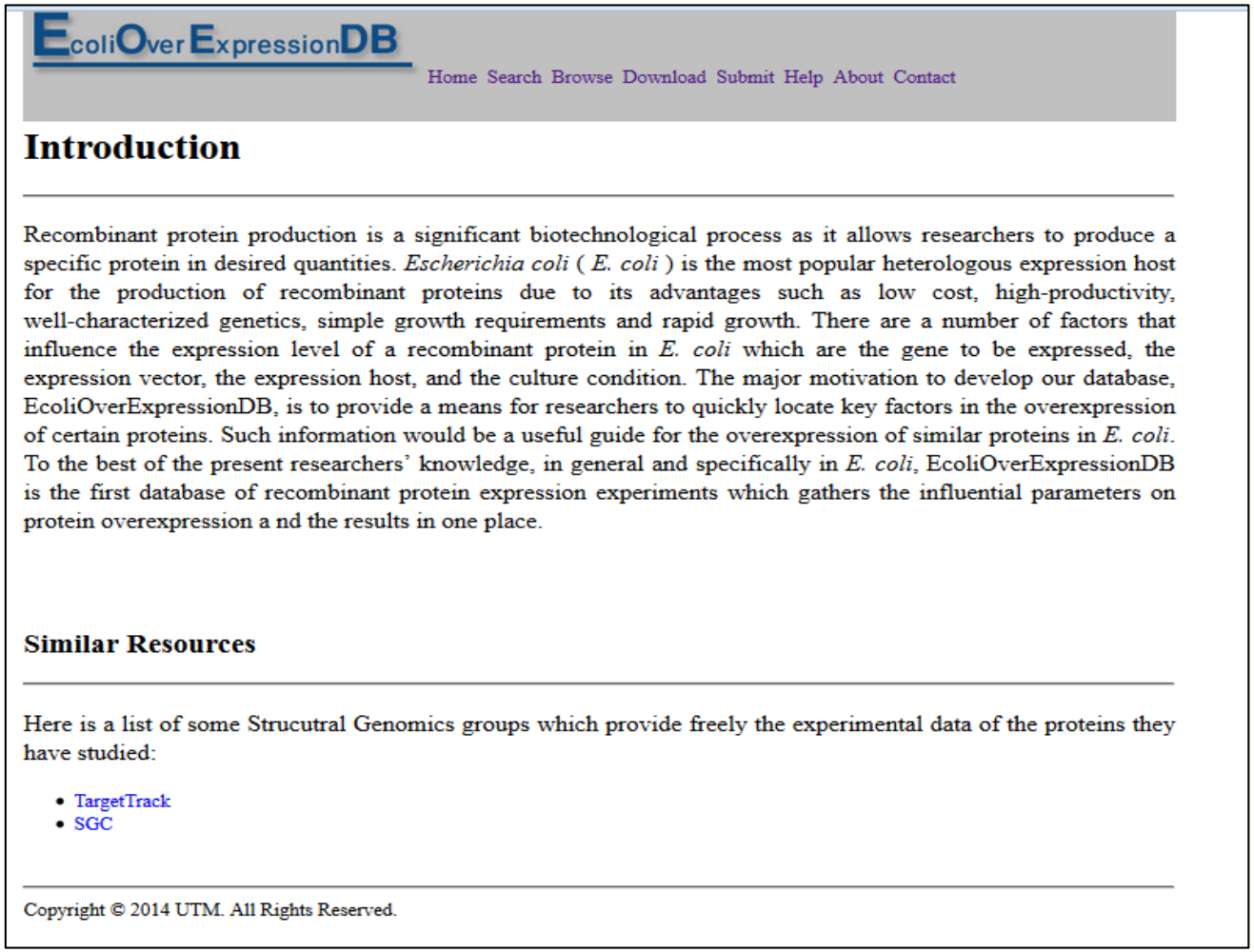
Validation by PEP journal through publishing a paper

Database Establishment (Cont.)

#	Name	Example
1	**Database ID**	EC00023
2	**Gene name**	VHb-His6 (vitreoscilla Hemoglobin)
3	**Gene ID (e.g. GenBank ID)**	AF292694.1
4	**Gene sequence**	ATGTTA…CACCAC
5	**Vector name**	pET-22b(+)
6	***E. coli* strain name**	BL21(DE3)
7	**Inducer concentration (e.g. Arabinose, Autoinduction, IPTG)**	0.8 mM IPTG
8	**Temperature after induction (C)**	30
9	**Expression level (Low, Medium or High)**	Medium
10	**Expression yield (mg/l)**	82.7
11	**Reference paper**	Protein Expr Purif, 2012. 86(1): p. 21-6.
12	**Note**	Before purification

Database Establishment (Cont.)

- Home page: http://birg4.fbb.utm.my:8080/EcoliOverExpressionDB/



EcoliOverExpressionDB

Home Search Browse Download Submit Help About Contact

Introduction

Recombinant protein production is a significant biotechnological process as it allows researchers to produce a specific protein in desired quantities. *Escherichia coli* (*E. coli*) is the most popular heterologous expression host for the production of recombinant proteins due to its advantages such as low cost, high-productivity, well-characterized genetics, simple growth requirements and rapid growth. There are a number of factors that influence the expression level of a recombinant protein in *E. coli* which are the gene to be expressed, the expression vector, the expression host, and the culture condition. The major motivation to develop our database, EcoliOverExpressionDB, is to provide a means for researchers to quickly locate key factors in the overexpression of certain proteins. Such information would be a useful guide for the overexpression of similar proteins in *E. coli*. To the best of the present researchers' knowledge, in general and specifically in *E. coli*, EcoliOverExpressionDB is the first database of recombinant protein expression experiments which gathers the influential parameters on protein overexpression a nd the results in one place.

Similar Resources

Here is a list of some Strucutral Genomics groups which provide freely the experimental data of the proteins they have studied:

- TargetTrack
- SGC

Database Establishment (Cont.)

- Browse page:

EcoliOverExpressionDB

Home Search Browse Download Submit Help About Contact

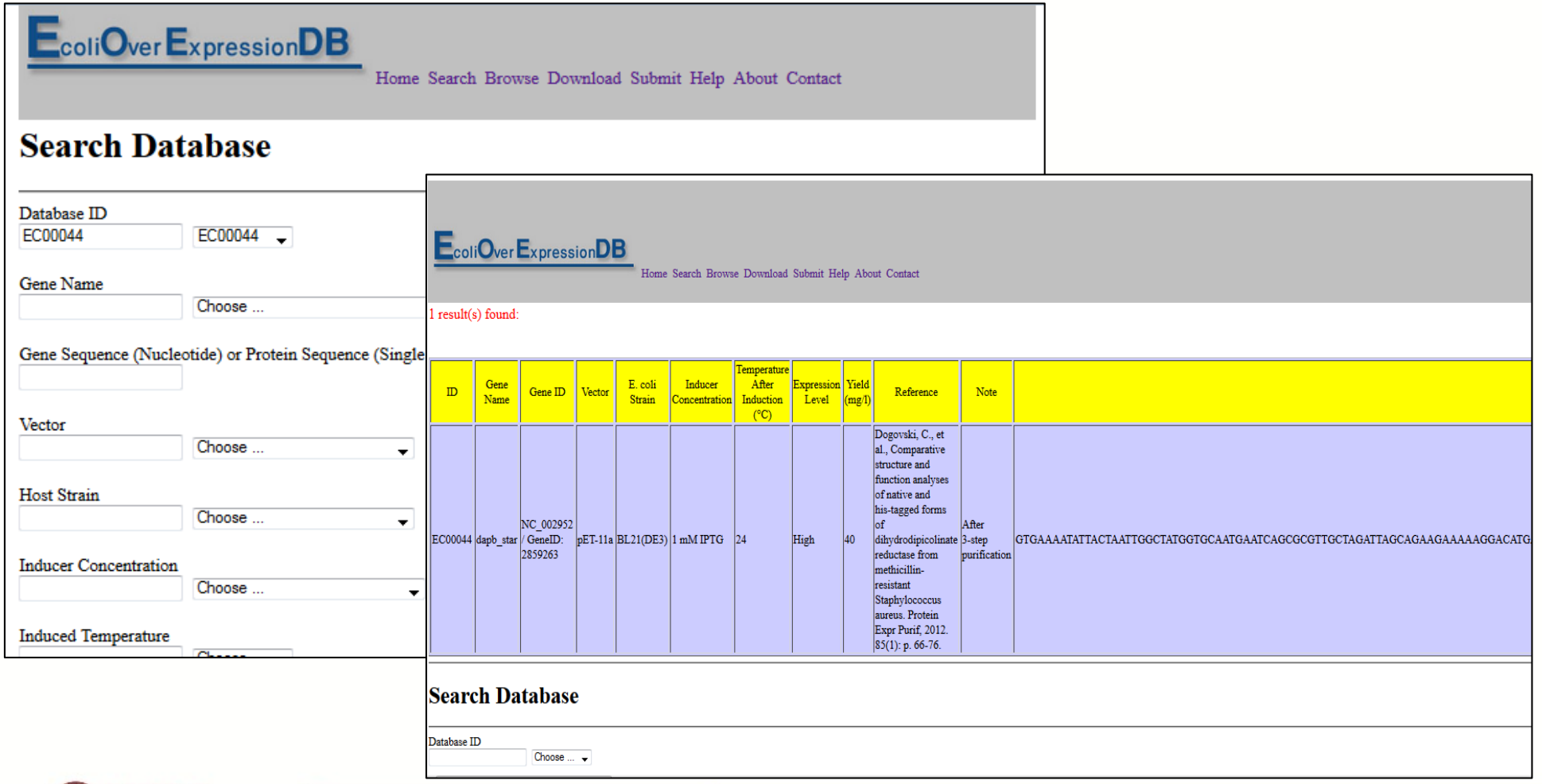
Browse

305 result(s) found:

ID	Gene Name	Gene ID	Vector	E. coli Strain	Inducer Concentration	Temperature After Induction (°C)	Expression Level	Yield (mg/l)	Reference	Note
EC00001	phaC	gi|302785999	pET-32 Xa/LIC	BL21 (DE3)	0.04 mM IPTG	15	High	?	Prof Razip's Lab- by Adrian Chek	-
EC00002	PhaC	gi|302785999	pCold DNA	BL21 (DE3)	0.1 mM IPTG	15	High	?	Prof Razip's Lab- by Adrian Chek	-
EC00003	phaC of Pseudomonas sp USM-4-55	?	pUC18	XL1- Blue	?	37	Low	?	Prof Razip's Lab- by Aida Baharuddin	-
EC00004	hGH	?	pKT52	JM107	1 mM IPTG	24	Low	?	Prof Razip's Lab-by Prof Razip	human growth hormone (with signal petide)
EC00005	hGH	?	pKT52	MM294	1 mM IPTG	24	Low	?	Prof Razip's Lab-by Prof Razip	human growth hormone (with signal petide)
EC00006	MBP-sNgb	BT059199	pETM41	BL21 (DE3)	1 mM IPTG	15	?	0.0001	Bjorlykke, G.A., et al., Cloning, expression and purification of Atlantic salmon (Salmo salar,	cleared lysate

Database Establishment (Cont.)

- Search page:



Database Establishment (Cont.)

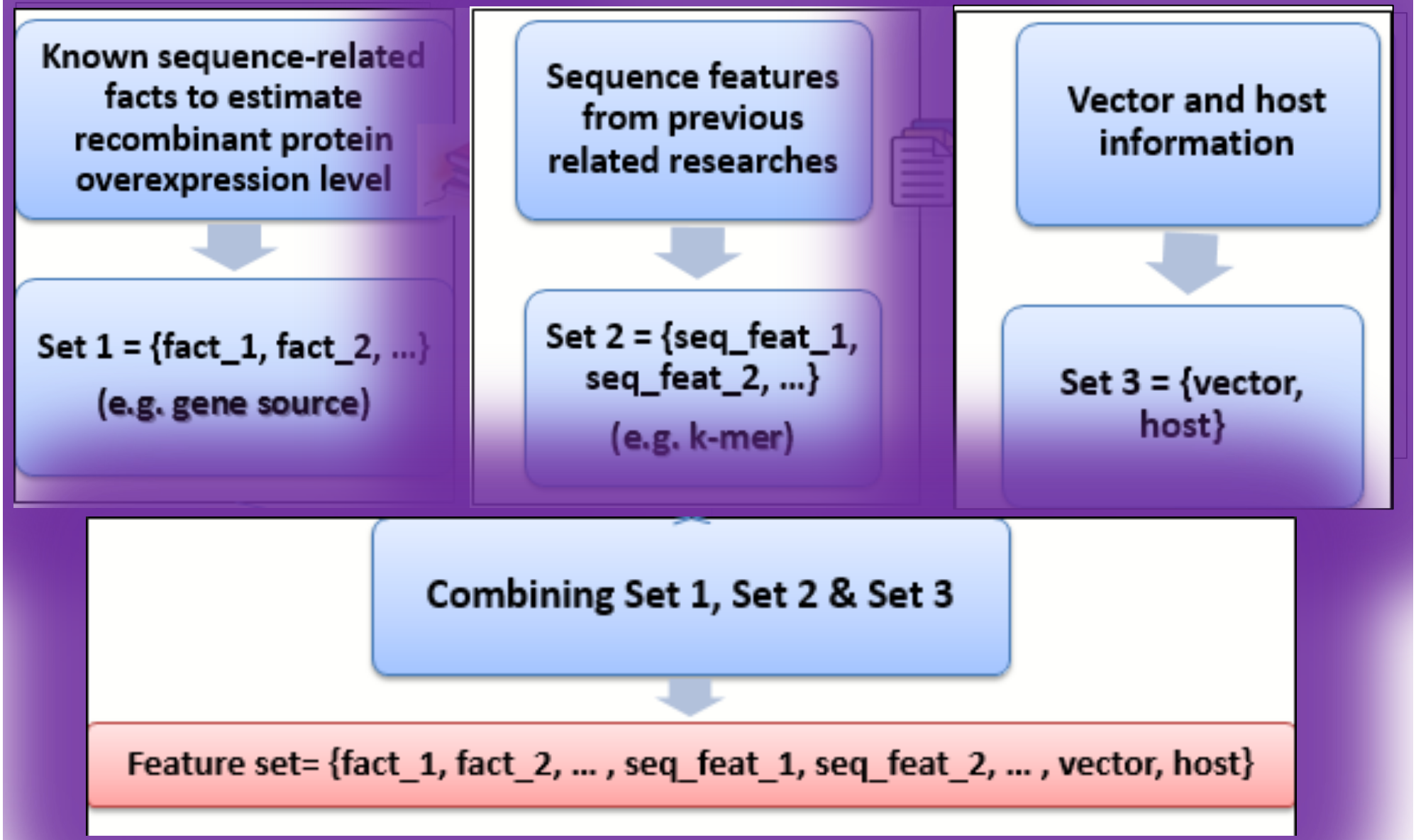
- Final dataset specification (after pre-processing):

Class	Number of instances	Proportion
High	52	46%
Medium	48	42.5%
Low	13	11.5%
Total records	113	

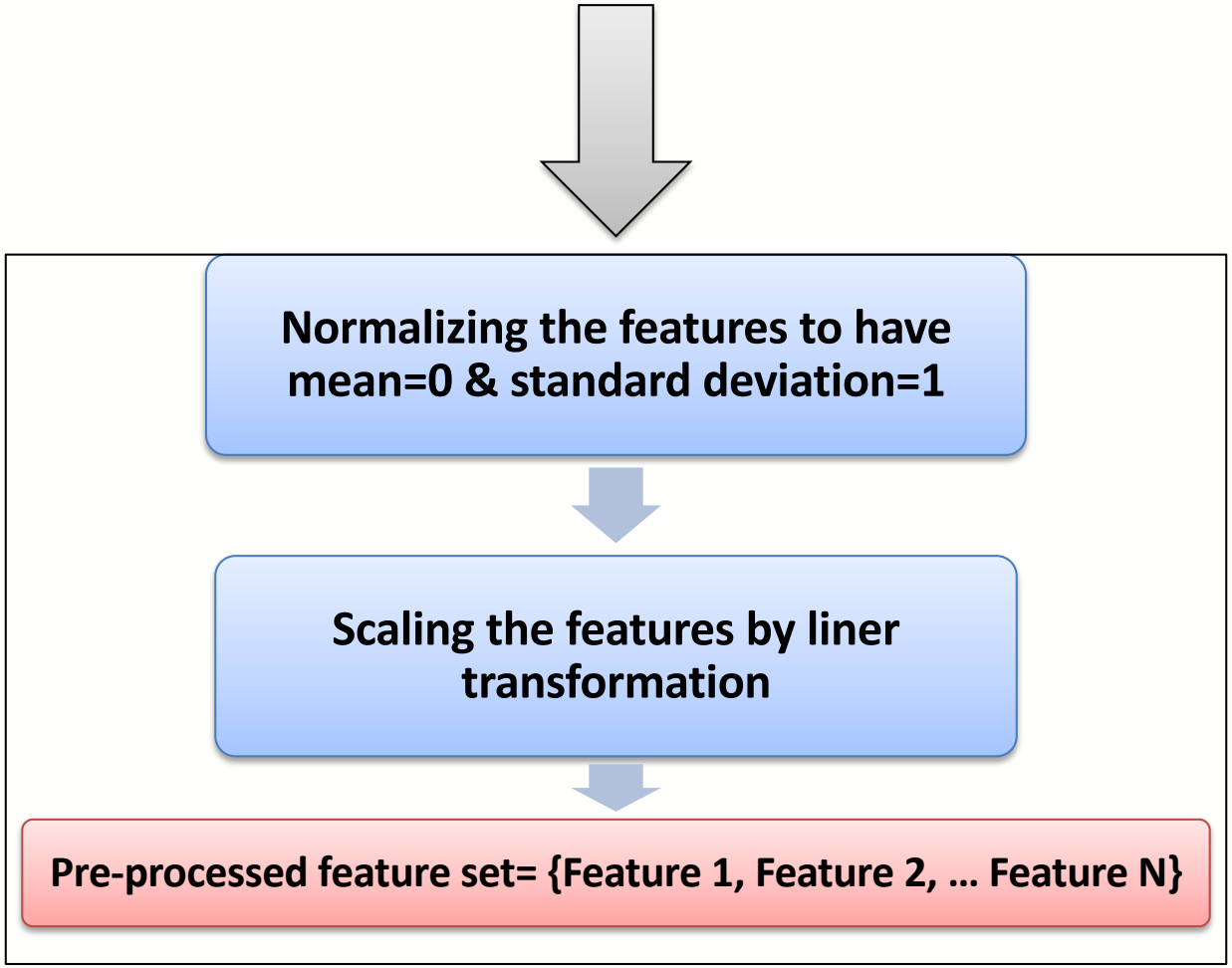
Imbalance

CHALLENGE!

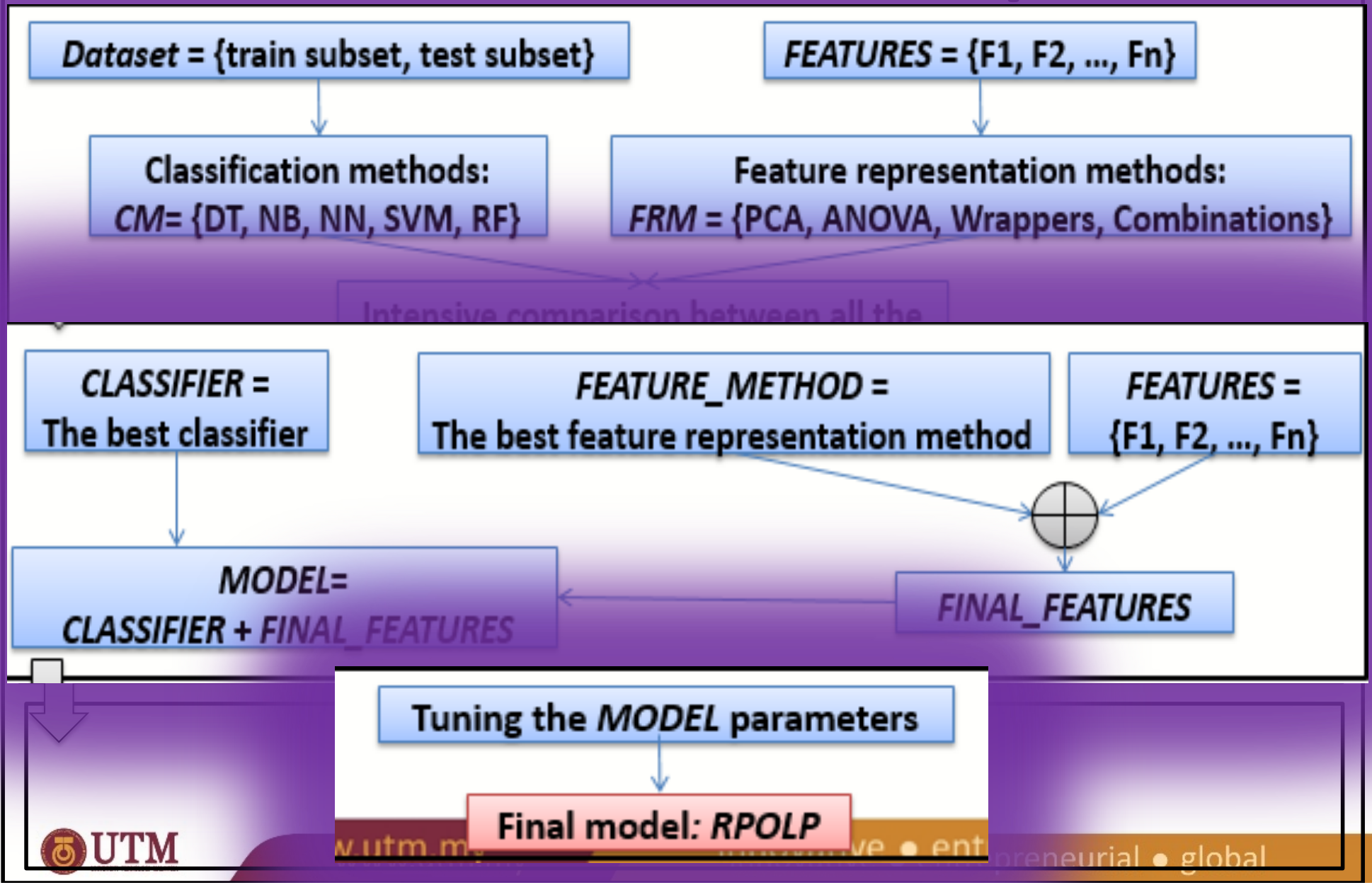
Feature Set Generation



Feature Set Generation (Cont.)



Predictive Model Development



Predictive Model Development (Cont.)

- Feature extraction and selection methods employed:

#	Method	Type
1	PCA	Feature Extraction
2	One-way ANOVA	Feature Selection: Filter
3	Forward	Feature Selection: Wrapper
4	Backward	
5	Random	
6	Floating	
7	Random forest	
8	ANOVA_Forward	Combined Method
9	ANOVA_RF	
10	Forward_RF	
11	RF_Forward	

Predictive Model Evaluation

#	Name	Abbr.	Formula	Description
1	**Accuracy**	ACC	(TP+TN) / (TP+TN+FP+FN)	The number of correctly classified instances divided by the total number of instances (Smialowski et al., 2012).
2	**Area under ROC curve**	AUC	-	AUC measures the discriminating ability of the model and it takes values between 0.5 for random drawing and 1.0 for perfect classifier (Smialowski et al., 2012).
3	**Error**	ERR	1-ACC	-
4	**F-score**	FS	2×Precision×Recall / (Precision+Recall)	The harmonic mean of recall and precision (Hirose and Noguchi, 2013).

Predictive Model Evaluation (Cont.)

#	Name	Abbr.	Formula	Description
5	**Gain**	GAIN	Precision / Proportion of the given class in the full data set.	Gain is an important performance measure that quantifies how much better the decision is in comparison with random drawing of instances (Smialowski et al., 2012).
6	**Matthew's Correlation Coefficient**	MCC	$\frac{(TP \times TN - FP \times FN)}{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$	It indicates the correlation between the classifier assignments and the actual class in the two-class case and is a good measure of classifier performance even when classes are unbalanced (Smialowski et al., 2012). The MCC ranges between −1 and 1, and a large positive value indicates a better prediction (Hirose and Noguchi, 2013).

Predictive Model Evaluation (Cont.)

#	Name	Abbr.	Formula	Description
7	**Precision (Selectivity)**	PRC	TP/(TP+FP) Or TN/(TN+FN)	The ratio of the number of correctly classified positive or negative instances to the number of all instances classified as positive or negative, for positive and negative class respectively (Smialowski et al., 2012).
8	**ROC Curve**	ROC	Plotting the "FP-rate" against the "TP- rate", while the probability is increased from 0 to 1.0 with 0.01 increments.	The receiver-operator characteristic curve, showing the trade-off between the ratio of false positives and false negatives in testing a classifier (de Ridder, de Ridder and Reinders, 2013). A larger area value indicates a more robust prediction method (Hirose and Noguchi, 2013).

Predictive Model Evaluation (Cont.)

#	Name	Abbr.	Formula	Description
9	**Recall (Sensitivity) (True positive rate) (TP- rate)**	REC	TP/(TP+FN)	The ratio of the number of correctly classified positive instances to the number of all instances from the positive class (Smialowski et al., 2012).
10	**Specificity (True negative rate) (TN-rate)**	SPC	TN/(TN+FP)	The ratio of the number of correctly classified negative instances to the sum of all negative instances (Smialowski et al., 2012).
11	**G-mean**	G-mean	sqrt (recall_1 * recall_2 * …)	Geometric mean of all recall values of all classes (Wang et al., 2012).

Predictive Model Evaluation (Cont.)

- Training and Testing Protocol:
  - K-fold cross validation.
  - k=10; Reasons:
    1. Theoretical reasons (not too small, not too large);
    2. The most common value in the related researches;
    3. The trial and error approach.

Results & Discussion

- Results
  - Group 1 Experiments
  - Group 2 Experiments
  - Prediction Performance
- Discussion

Results

- There are two groups of experiments in this research:

#	Group	Description
1	Selecting feature representation & classifier for RPOLP	To determine the best representation of features and then choosing an appropriate classifier by comparative analysis.
2	Adjusting parameters of RPOLP	To tune two classifier's parameters: number of trees & number of features.

Results (Cont.)

- Experiments basics:

Aspect	Description	Reason to use
Classifier	• DT • NB • NN • SVM • RF	The most common classifiers used in the related literature.
Performance comparison	Based on the achieved G-mean, AUC and F-Score, in order.	The main metrics used in class imbalance problems (Wang and Yao, 2012).

Results (Cont.)

- Group 1 experiments
  - Based on the obtained results, random forest is able to achieve superior performance compared to the other four tested classifiers.
  - One-way ANOVA is the best method to filter out the input features.

Results (Cont.)

- Group 1 experiments (Cont.)

Classifier: Random Forest										
Metric=> / Features	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
Original	0.69	0.31	0.66	0.56	0.75	3.84	0.38	0.64	0.77	0.28
PCA	0.71	0.29	0.70	0.61	0.75	3.87	0.40	0.60	0.73	0.41
ANOVA	**0.78**	**0.22**	**0.75**	**0.69**	**0.82**	4.01	**0.54**	**0.69**	**0.81**	**0.50**
Forward	0.73	0.27	0.69	0.64	0.77	3.77	0.46	0.64	0.76	0.45
Backward	0.68	0.32	0.63	0.55	0.74	3.61	0.38	0.61	0.69	0.28
Random	0.71	0.29	0.68	0.60	0.76	3.71	0.42	0.63	0.75	0.38
Floating	0.74	0.26	0.69	0.63	0.78	3.35	0.45	0.61	0.78	0.43
RF	0.75	0.25	**0.75**	0.62	0.79	3.74	0.51	0.66	0.76	0.43
ANOVA=>Forward	0.72	0.28	0.72	0.59	0.77	**4.55**	0.48	0.68	0.71	0.33
ANOVA=>RF	0.77	0.23	0.70	0.62	0.80	3.53	0.45	0.63	0.80	0.34
RF=>Forward	0.76	0.24	0.70	0.67	0.80	3.50	0.48	0.65	0.78	0.47

Results (Cont.)

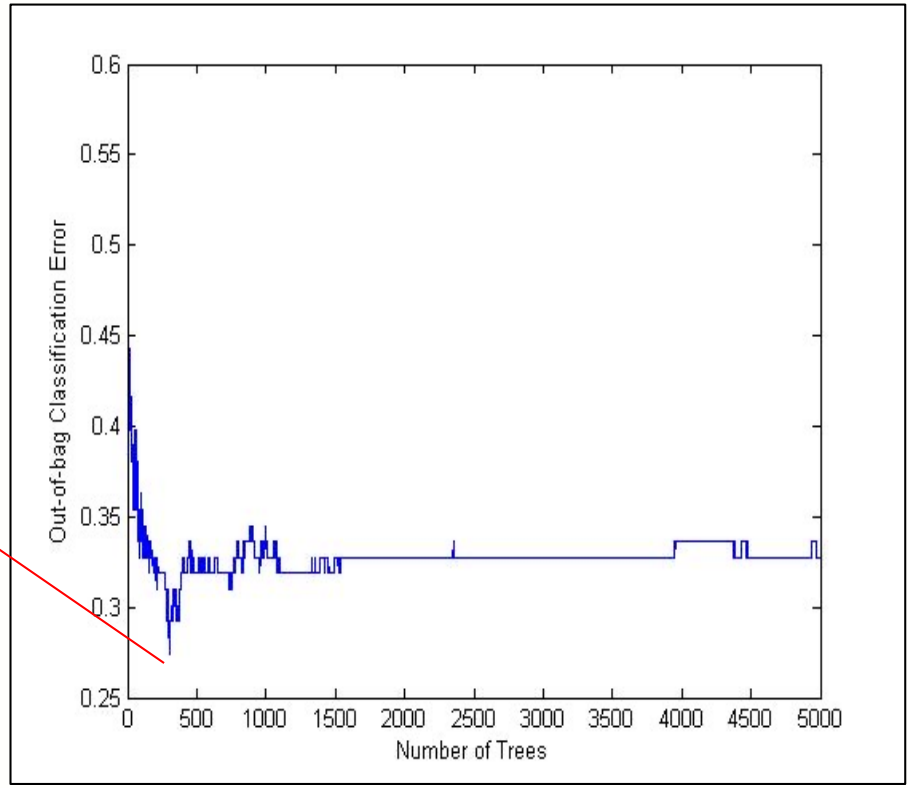
- Group 2 experiments

  - Number of trees:
    - To tune the number of trees, 10 learning curves of out-of-bag classification error versus number of trees, are computed (Maximum number of trees=5000).
    - Then the obtained values for the best number of trees in 10 experiments are averaged (=233).
    - The results are shown in the next slide.

Results (Cont.)

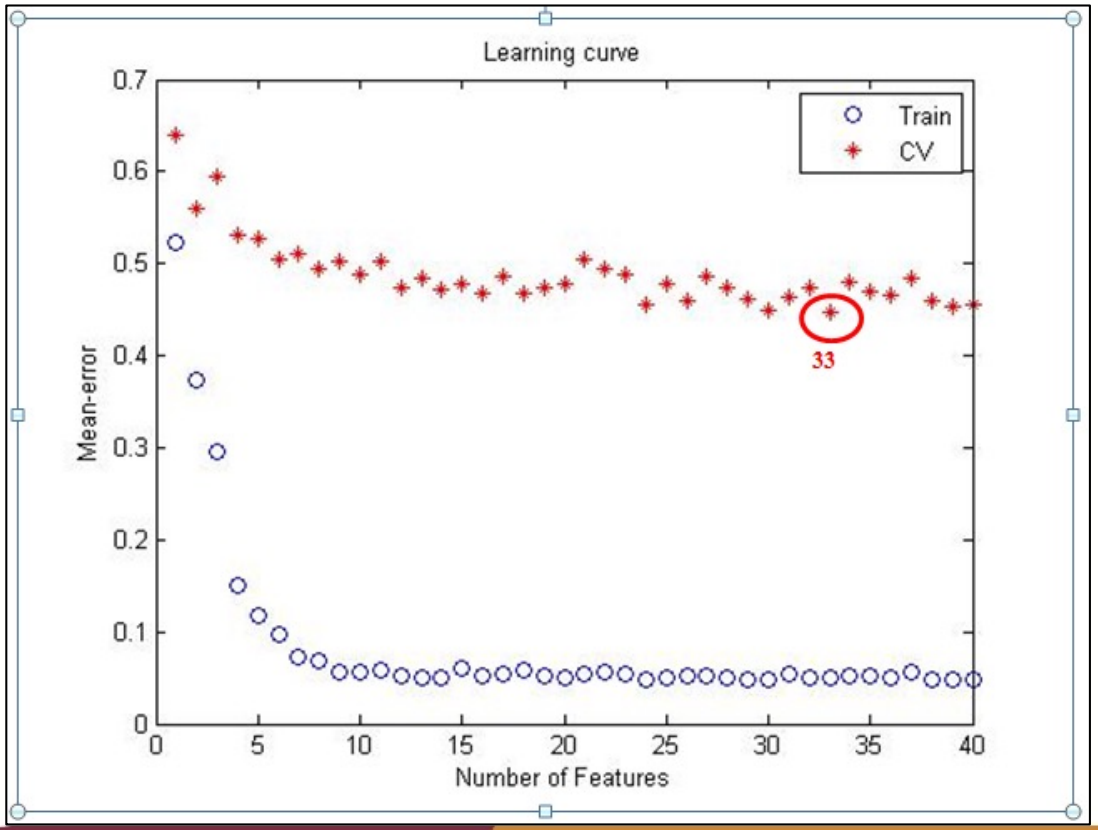
- Group 2 experiments (Cont.)

Experiment	Minimum Error	Number of Trees (which produces the minimum error)
1	0.3274	93
2	0.3097	131
3	0.3097	498
4	0.2920	161
5	**0.2743**	307
6	0.2920	293
7	0.3009	60
8	0.3009	69
9	0.3186	515
10	0.3186	210
Average number of trees ~ **233**		



Results (Cont.)

- Group 2 experiments (Cont.)
  - Number of features: Learning curve of RF mean-error versus number of features.



Results (Cont.)

- Prediction performance
  - Comparing 10-fold and leave-one-out cross validation methods
    - **=> If more training data is available, the model is able to gain higher performance.**

Classifier: RF

Feature Representation: one-way ANOVA

Model Parameters: Number of trees=233; Number of features=sqrt(40) ~ 6

Metric=> / Evaluation Method	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
10-fold	0.81	0.19	0.82	0.71	0.84	4.27	0.64	0.76	0.83	0.52
Leave-one-out	**0.91**	**0.09**	**0.93**	**0.88**	**0.93**	**3.08**	**0.89**	**0.93**	NaN	**0.74**

Results (Cont.)

- Confidence intervals of the measurements (95%):

Measurement	Value
Accuracy	0.78 +- 0.07
Error	0.22 +- 0.07
Precision	0.77 +- 0.09
Recall	0.68 +- 0.12
Specificity	0.81 +- 0.05
Mathew Correlation Coefficient	0.53 +- 0.16
Gain	4.09 +- 1.19
F-score	0.67 +- 0.12
Geometric Mean	0.53 +- 0.15
Area Under Curve	0.82 +- 0.07

Discussion

Aspect	Discussion

RPOLP model	• Feature selection method: One-way ANOVA
	• Size of feature set: 40
	• Classifier: Random forest:
	  ▪ Number of feature=233
	  ▪ Number of features=6 (=square root of 40)
	• Evaluation: 10-fold cross validation
	• Results: G-mean=0.52, AUC=0.83, F-score=0.76

Discussion (Cont.)

Aspect	Discussion
Vector & host features	• Host: No relationship between host and overexpression level is found. • Vector: Just two feature selection methods (Random and RF) selected vector feature. • Probable reason: Small available dataset. • Further investigation: Representing the vector and host in the dataset with their features (e.g. RBS for vector), instead of their names, makes finding some relationships between them and overexpression level more probable.

Discussion (Cont.)

Aspect	Discussion
Significant features	• Repeat features (for amino acids, and chemical and physical groups) in both normal and terminal regions. • Serine amino acid. • Chemical and physical group of RKH in both normal and N-terminal features. • Proportion of disordered regions. • Terminal features: ▪ Gene 3-mers at N-term and C-term features. ▪ Frequencies of some amino acids (H, S, W) in N-term. ▪ Repeats of some amino acids in N-term and C-term.

Discussion (Cont.)

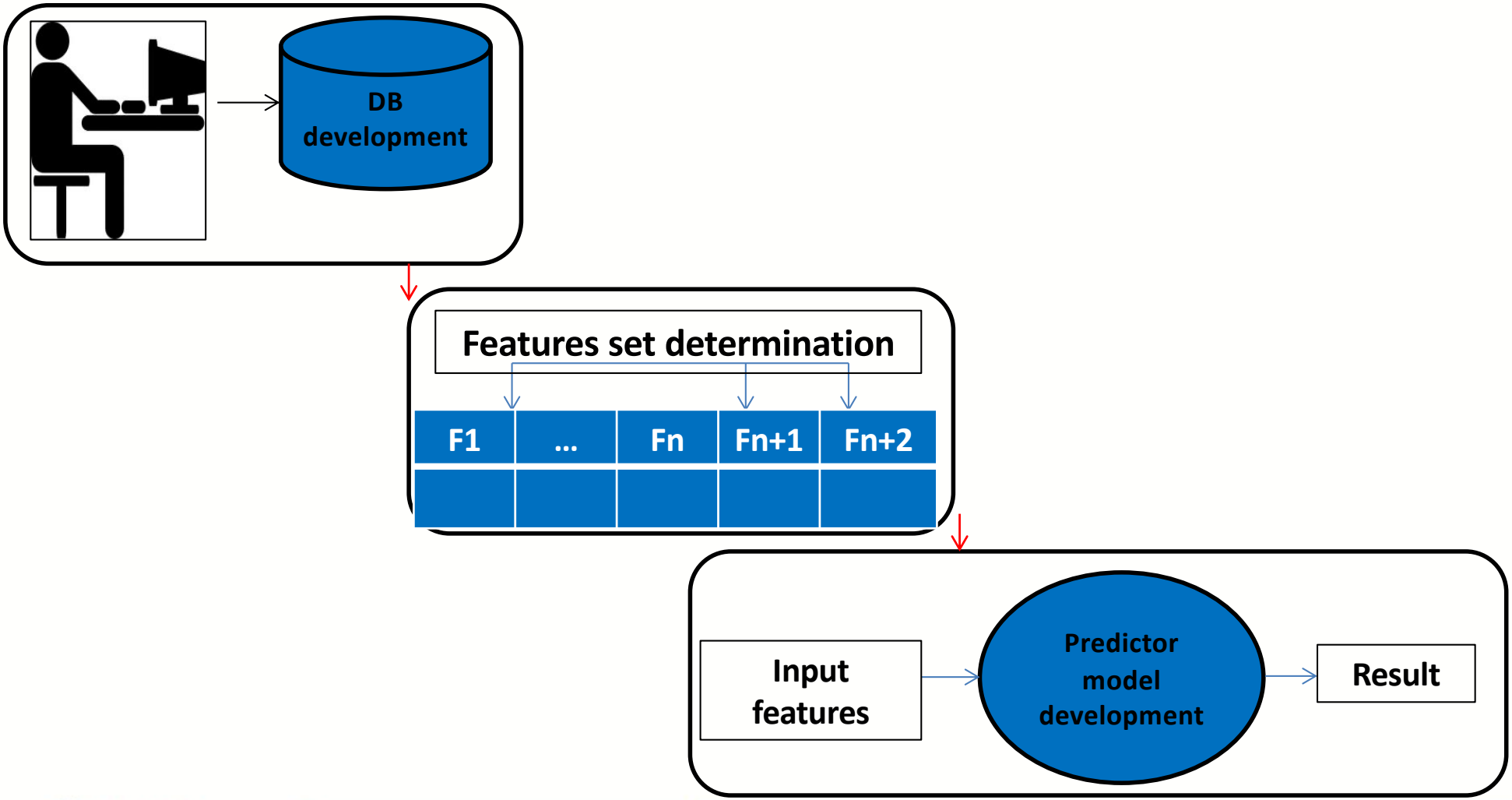
Aspect	Discussion
Random forest	• It is concluded that random forest is very less sensitive to the following issues, compared to the other tested classifiers: ▪ Feature space dimension: Random forest builds each tree with a subset of features. ▪ Imbalance data: Each tree of random forest is built using a random sub-sample of data. Randomness increases the chance of building a tree with a more balanced subset. ▪ Small size data: Random forest creates several trees with overlapping random subsets of data.

Conclusion

- Conclusion
- Thesis Contributions
- Future Works
- Publications

Conclusion

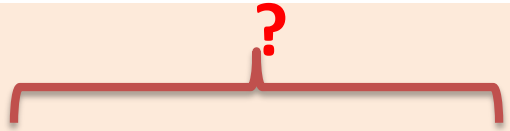
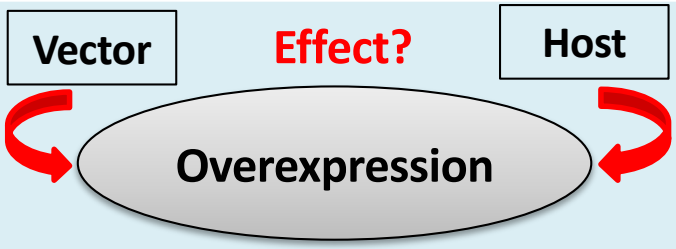
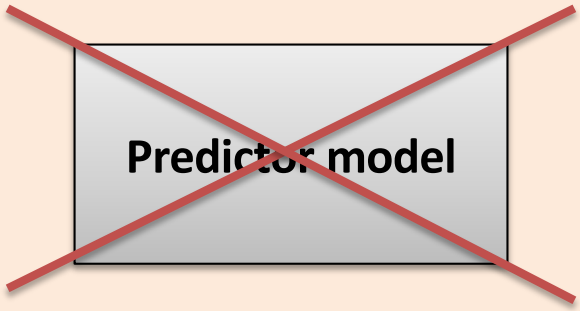
Recombinant Protein Overexpression Level Predictor



Thesis Contributions

Objective	Challenge	#	Contribution
1, 2 & 3	 Recombinant protein overexpression level prediction	1	A model (RPOLP) to predict recombinant protein overexpression level is proposed.
1	 Overexpression dataset	2	A novel dataset is constructed.
		3	A web-based database (EcoliOverexpressionDB) is established to facilitate public access to the constructed dataset .

Thesis Contributions (Cont.)

Objective	Challenge	#	Contribution
2		4	A set of significant features are discovered through feature extraction and selection techniques.
		5	Vector and host are considered and investigated in addition to the gene sequence.
3		6	A new predictive model (RPOLP) to estimate recombinant protein overexpression level is developed which is able to handle properly this multi-class classification problem with the small imbalance collected dataset.

Future Works

Aspect	Future Work
Data	1. Extending the constructed dataset by incorporating more data fields and data records.
	2. Applying techniques of missing value estimation for the vector, host and overexpression level fields.
Features	3. Discovering and incorporating more relevant features to the overexpression level.
	4. Investigating in more depth the effect of the vector and host on the overexpression level by incorporating their features (e.g. “RBS” for vector and “drug resistance” for host).
	5. Evaluating the effect of fermentation condition (e.g. temperature) on the overexpression level.

Future Works (Cont.)

Aspect	Future Work
Modeling	6. Detecting outlier data samples and removing them before model building.
	7. Employing other feature extraction and selection methods.
	8. Dealing with the imbalance dataset problem more specifically.
	9. Investigating other types of classifiers.
	10. Applying extensions of random forest.
	11. Evaluating other types of ensemble learning techniques.

Publications

Journals

1. **Narjeskhatoon Habibi**, Alireza Norouzi, Siti Z Mohd Hashim, Mohd Shahir Shamsir, Razip Samian, "*Prediction of Recombinant Protein Overexpression in Escherichia coli Using Machine Learning Approach",* Computers in Biology and medicine, Elsevier, 66 (2015), pp. 330-336 (**ISI Journal, IF=1.24)**
2. **Narjeskhatoon Habibi**, Siti Z. Mohd Hashim, Alireza Norouzi, Mohammed Razip Samian, *"Review of Machine Learning Methods to Predict the Solubility of Overexpressed Recombinant Protein",* BMC Bioinformatics, BioMed Central, 15:134 (2014), doi:10.1186/1471-2105-15-134 (**ISI Journal, IF=3.02)**
3. **Narjeskhatoon Habibi**, Mohd Razip Samian, Siti Zaiton Mohd Hashim, *"EcoliOverExpressionDB: A Database of Recombinant Protein Overexpression in E. coli",* Protein Expression and Purification, Elsevier, 95 (2014), pp. 92–95. (**ISI Journal, IF=1.4**)
4. **Narjeskhatoon Habibi**, Siti Zaiton Mohd Hashim, Cesar A. Rodriguez, Razip Samian, *"A Review of CADS, Languages and Data Models for Synthetic Biology"*, Jurnal Teknologi, Penerbit UTM Press, 63:1 (2013), pp. 87–96. (**Scopus Journal**)

Proceedings

1. **Narjeskhatoon Habibi**, Siti Zaiton Mohd Hashim, Razip Samian, Cesar A. Rodriguez, Alireza Norouzi, *"Intelligent Host Cell Selection for Synthetic Biology Applications"*, SB6.0: The Sixth International Meeting on Synthetic Biology, 9-11 July 2013, Imperial College London, London, Published online (http://sb6.biobricks.org/poster/intelligent-host-cell-selection-synthetic-biology/).
2. **Narjeskhatoon Habibi**, Siti Zaiton Mohd Hashim, Cesar A. Rodriguez, Mohd Saberi Bin Mohamad, Safaai Bin Deris, *"The Emerging Field of Synthetic Biology: a Review"*, 2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012), 12-14 June 2012, Kuala Lumpur, Malaysia, Vol 1, pp. 160-164. (**IEEE Conference**)

Acknowledgment





And all the people who have helped me during 23 years of study! Thank you ☺

Thanks for Your Attention

Q&A

Extra Slides

Feature Set Generation (Cont.)

Group	#	Name	Size	How to compute (MATLAB/Web Service)
Cassette	1	Vector	1	-
	2	Host	1	-
Gene	3	Gene 1-mer	4	*nmercount*
	4	Gene 2-mer	16	*nmercount*
	5	Gene 3-mer	64	*nmercount*
	6	Gene length	1	-
	7	GC-content	1	*Oligoprop*
	8	CAI	1	CAI (http://emboss.bioinformatics.nl/cgi-bin/emboss/cai) Codon Usage Table: *Eecoli.cut*

Feature Set Generation (Cont.)

Group	#	Name	Size	How to compute (MATLAB/Web Service)
Protein-Sequence	9	Size of polypeptide	1	-
	10	Frequencies of mono-peptide	22	-
	11	Frequencies of chemical groups	8	-
	12	Frequencies of Physical groups	5	-
	13	Repeat of amino acids	22	-
	14	Size of polypeptide	8	-
	15	Frequencies of mono-peptide	5	-

Feature Set Generation (Cont.)

Group	#	Name	Size	How to compute (MATLAB/Web serv.)
Protein-Structure	16	Frequencies of single amino acids in surface area	22	RVPnet (Ahmad, Gromiha and Sarai, 2003)
	17	Frequencies of chemical groups in surface area	8	RVPnet (Ahmad, Gromiha and Sarai, 2003)
	18	Frequencies of physical groups in surface area	5	RVPnet (Ahmad, Gromiha and Sarai, 2003)
	19	Number of transmembrane regions	1	TMHMM (Krogh et al., 2001)
	20	Disordered regions: number	1	POODLE-L (Hirose et al., 2007)

Feature Set Generation (Cont.)

Group	#	Name	Size	How to compute (MATLAB/Web Serv.)
Protein-Structure	21	Disordered regions: proportion	1	POODLE-L (Hirose et al., 2007)
Terminal-Protein/ Gene Sequence	22	Gene N-term: 3-mer	64	*nmercount*
	23	Gene C-term: 3-mer	64	*nmercount*
	24	N-term: frequencies of mono-peptides	22	-
	25	C-term: frequencies of mono-peptides	22	-
	26	N-term: frequencies of chemical groups	8	-
	27	C-term: frequencies of chemical groups	8	-
	28	N-term: frequencies of physical groups	5	-

Feature Set Generation (Cont.)

Group	#	Name	Size	How to compute (MATLAB/Web Serv.)
Terminal-Protein/ Gene Sequence	29	C-term: frequencies of physical groups	5	-
	30	N-term: repeat of amino acids	22	-
	31	C-term: repeat of amino acids	22	-
	32	N-term: repeat of chemical groups	8	-
	33	C-term: repeat of chemical groups	8	-
	34	N-term: repeat of physical groups	5	-
	35	C-term: repeat of physical groups	5	-

Predictive Model Evaluation (Cont.)

- **Extension of Binary Performance Measures for Multi-class Situation:**

  - For all the measures (except AUC & G-mean):
    1. In multi-class classification with *N* classes, for an individual class C_i (*i=1, 2 ,.., N*), the measures are calculated from the counts for C_i.
    2. The overall classification performance is assessed in this study by **macro-averaging**; a measure is the average of the same measures calculated for $C_1;...;C_N$.

  - For G-mean: *G-mean = sqrt (Recall_positive* Recall_negative)*

  - For AUC: Averaging the AUCs of all the one-vs-all classifiers.

Results

- Classifiers' structures & specification:

Classifier	Specification
DT	It splits the available DATA into training and validation parts (80% & 20%). Then it builds the tree (without pruning) on the training part and estimates error using the validation subset. Tree with minimal validation set error is returned.
NB	It uses 20 bins and histograms are corrected to avoid zeroes for in-range value. Out-of-range values are handled in 'strict' mode. The number of histogram bins is estimate automatically.
NN	Feed-forward neural network with one hidden layer and 10 neurons (*units*) in the hidden layer. *'noscale'* option is used.
SVM	Polynomial kernel is used and polynomial degree is selected by grid-search.*'noscale'* option is used.
RF	20 tree classifiers (*COUNT/trees*) are built, using 20% of features (*dim*) selected randomly at each tree stage.

Results (Cont.)

- Original feature set:

Feature representation method: Original Features Number of features=466										
Metric=> **Classifier**	**ACC**	**ERR**	**PRE**	**REC**	**SPC**	**GAIN**	**MCC**	**F-Score**	**AUC**	**G-mean**
DT	0.62	0.38	0.39	0.41	0.70	1.13	0.11	0.42	NaN	0.09
NB	0.62	0.38	NaN	0.33	0.67	NaN	NaN	NaN	NaN	0.00
NN	0.62	0.38	NaN	0.33	0.67	NaN	NaN	NaN	NaN	0.00
SVM	0.62	0.38	NaN	0.33	0.67	NaN	NaN	NaN	NaN	0.00
RF	**0.69**	**0.31**	**0.66**	**0.56**	**0.75**	**3.84**	**0.38**	**0.64**	**0.77**	**0.28**

Results (Cont.)

- PCA:

Feature representation method: PCA Dimensions of the feature space transform=[113*446 => 113*112]										
Metric=> **Classifier**	**ACC**	**ERR**	**PRE**	**REC**	**SPC**	**GAIN**	**MCC**	**F-Score**	**AUC**	**G-mean**
DT	0.67	0.33	0.45	0.46	0.73	1.56	0.21	0.50	NaN	0.10
NB	**0.72**	**0.28**	**0.84**	0.59	**0.76**	4.29	0.50	0.58	NaN	0.31
NN	0.71	0.29	0.60	0.58	**0.76**	2.65	0.34	0.57	0.70	0.36
SVM	0.57	0.43	0.66	0.48	0.73	2.01	0.35	0.37	NaN	0.14
RF	0.71	0.29	0.70	**0.61**	0.75	**3.87**	**0.40**	**0.60**	**0.73**	**0.41**

Results (Cont.)

- ANOVA:

Feature representation method: One-way ANOVA

Number of features=40

Features=[100, 112, 129, 132, 141, 150, 153, 160, 176, 182, 185, 198, 203, 216, 223, 227, 250, 251, 261, 277, 282, 286, 292, 293, 303, 305, 311, 318, 333, 342, 368, 400, 403, , 404, 408, 411, 427, 435, 446, 461]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.63	0.37	0.56	0.41	0.71	2.34	0.30	0.50	NaN	0.13
NB	0.67	0.33	0.60	0.51	0.72	3.09	0.33	0.56	0.67	0.26
NN	0.74	0.26	0.64	0.53	0.78	2.91	0.47	0.78	0.72	0.14
SVM	0.67	0.33	0.65	0.48	0.73	3.65	0.36	0.75	NaN	0.05
RF	**0.78**	**0.22**	**0.75**	**0.69**	**0.82**	**4.01**	**0.54**	**0.69**	**0.81**	**0.50**

Results (Cont.)

- Wrapper, Forward:

Feature representation method: Forward

Number of features=13

Features=[100,253,197,146,131,455,179,174,437,119,464,158,196]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.62	0.38	0.42	0.45	0.71	1.56	0.14	0.46	NaN	0.10
NB	0.66	0.34	0.59	0.47	0.71	3.01	0.27	0.51	0.62	0.20
NN	0.60	0.40	0.36	0.34	0.67	1.21	-0.01	NaN	NaN	0.00
SVM	0.65	0.35	NaN	0.37	0.68	NaN	NaN	NaN	NaN	0.00
RF	**0.73**	**0.27**	**0.69**	**0.64**	**0.77**	**3.77**	**0.46**	**0.64**	**0.76**	**0.45**

Results (Cont.)

- Wrapper, Backward:

Feature representation method: Backward

Number of features=17

Features=[437, 442, 447, 448, 450, 451, 452, 453, 454, 455, 458, 459,461, 462, 463, 464, 466]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.64	0.36	0.43	0.36	0.69	1.00	0.20	NaN	NaN	0.00
NB	0.61	0.39	0.43	0.43	0.68	1.83	0.11	0.38	NaN	0.11
NN	0.64	0.36	NaN	0.37	0.69	NaN	NaN	NaN	0.53	0.00
SVM	0.55	0.45	0.57	0.46	0.72	1.71	0.26	0.38	NaN	0.17
RF	**0.68**	**0.32**	**0.63**	**0.55**	**0.74**	**3.61**	**0.38**	**0.61**	**0.69**	**0.28**

Results (Cont.)

- Wrapper, Random:

Feature representation method: Random

Number of features=8

Features=[1, 4, 102, 111, 114, 128, 222, 426]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.63	0.37	0.43	0.38	0.70	1.14	0.09	0.42	NaN	0.03
NB	0.60	0.40	0.39	0.42	0.70	1.28	0.10	0.39	NaN	0.10
NN	0.63	0.37	0.50	0.37	0.69	2.02	0.21	0.42	0.54	0.02
SVM	0.55	0.45	0.29	0.43	0.67	0.95	-0.01	0.41	NaN	0.03
RF	**0.71**	**0.29**	**0.68**	**0.60**	**0.76**	**3.71**	**0.42**	**0.63**	**0.75**	**0.38**

Results (Cont.)

- Wrapper, Floating:

Feature representation method: Floating

Number of features=23

Features=[88, 134, 184, 207, 223, 236, 247, 256, 274, 278, 310, 312,329, 332, 365, 376, 421, 430, 449, 179, 4, 245, 6]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.58	0.42	0.52	0.37	0.66	1.90	0.17	0.41	NaN	0.11
NB	0.66	0.34	0.64	0.54	0.73	3.23	0.36	0.55	0.69	0.34
NN	0.62	0.38	0.51	0.40	0.69	3.55	0.11	0.39	0.59	0.02
SVM	Not converged									
RF	**0.74**	**0.26**	**0.69**	**0.63**	**0.78**	**3.35**	**0.45**	**0.61**	**0.78**	**0.43**

Results (Cont.)

- Wrapper, Random forest:

Feature representation method: Random forest

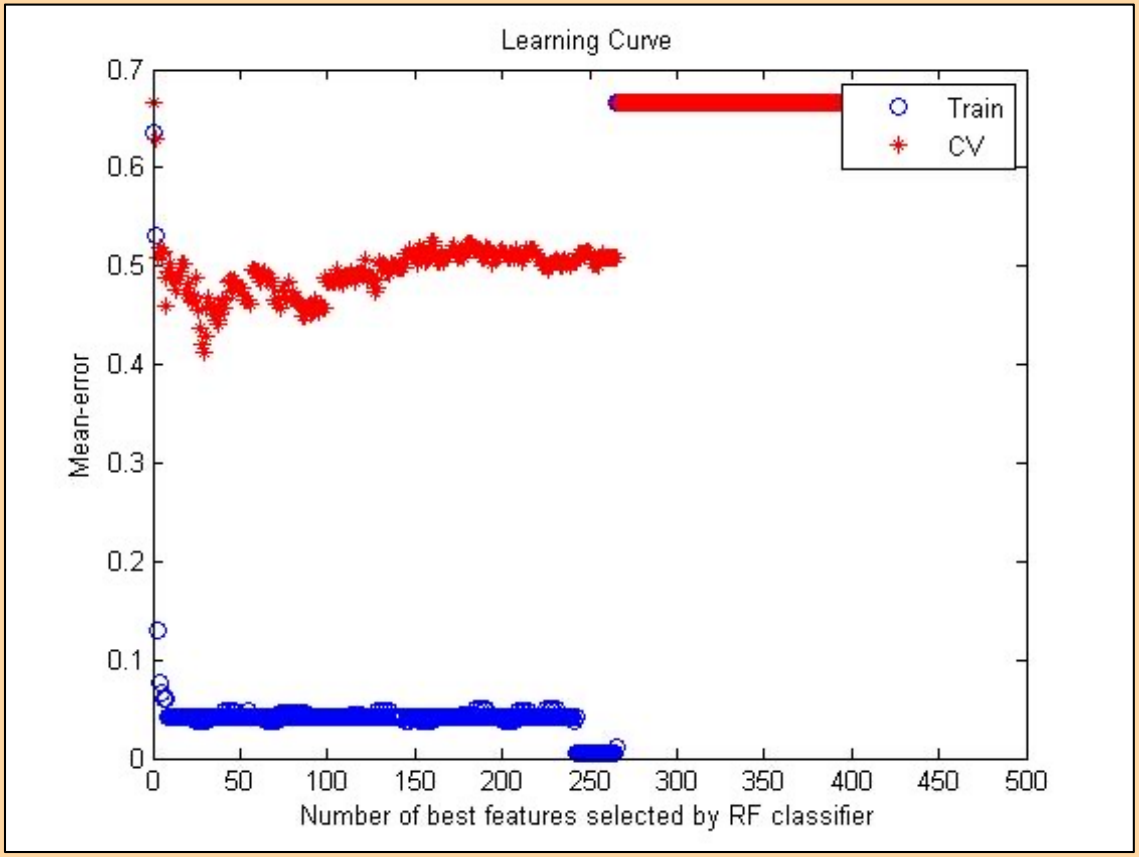
Number of features=29

Features=[293, 277, 1, 160, 13, 198, 184, 40, 42, 193, 357, 177, 333, 157, 360, 341, 375, 82, 292, 112, 381, 461, 197, 101, 384, 41, 318, 283, 89]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.70	0.30	0.59	0.51	0.75	2.40	0.42	0.59	NaN	0.25
NB	0.63	0.37	0.50	0.43	0.70	2.32	0.20	0.60	NaN	0.18
NN	0.70	0.30	NaN	0.44	0.75	NaN	NaN	NaN	0.65	0.00
SVM	0.60	0.40	0.58	0.48	0.72	1.94	0.30	0.42	NaN	0.11
RF	**0.75**	**0.25**	**0.75**	**0.62**	**0.79**	**3.74**	**0.51**	**0.66**	**0.76**	**0.43**

Results (Cont.)

- Wrapper, Random Forest (Cont.)
  - learning curve for finding the optimal number of best features:



Results (Cont.)

- **Combinations of Feature Space Dimensionality Reduction Methods**

  - Four combinations of the methods which seem to be more promising (ANOVA, Forward and RF) are examined:
    - ANOVA_Forward
    - ANOVA_RF
    - Forward_RF
    - RF_Forward

  - RF_Forward outperforms the other combined techniques.

Results (Cont.)

- ANOVA_Forward :

Feature representation method: ANOVA_Forward

Number of features=11

Features=[10, 19, 38, 35, 5, 2, 6, 26, 17, 28, 9]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.64	0.36	0.57	0.46	0.71	3.32	0.25	0.57	NaN	0.17
NB	0.63	0.37	0.49	0.48	0.69	2.34	0.16	0.48	0.61	0.29
NN	0.70	0.30	0.59	0.51	0.75	3.30	0.33	0.74	**0.72**	0.18
SVM	0.63	0.37	0.53	0.47	0.70	1.84	0.29	0.45	NaN	0.08
RF	**0.72**	**0.28**	**0.72**	**0.59**	**0.77**	**4.55**	**0.48**	**0.68**	0.71	**0.33**

Results (Cont.)

- ANOVA_RF:

Feature representation method: ANOVA_RF

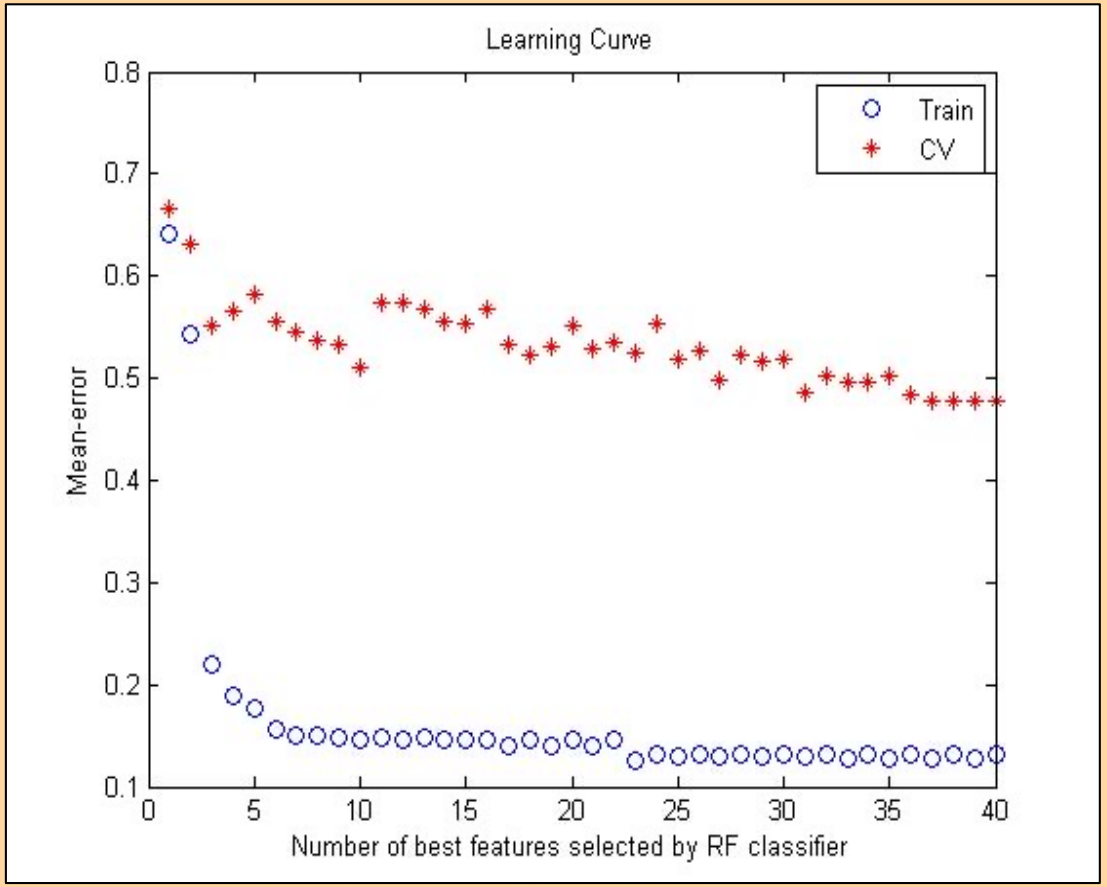
Number of features=37

Features=[24, 20, 12, 25, 19, 38, 34, 32, 13, 26, 2, 10, 16, 8, 7, 21, 15, 36, 39, 30, 5, 17, 11, 27, 28, 1, 37, 18, 31, 9, 29, 3, 22, 6, 14, 35, 23]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.66	0.34	0.53	0.41	0.71	2.04	0.25	0.66	NaN	0.10
NB	0.64	0.36	0.58	0.49	0.70	3.08	0.29	0.53	NaN	0.22
NN	0.71	0.29	0.72	0.46	0.74	2.89	**0.66**	**0.73**	NaN	0.08
SVM	0.66	0.34	**0.81**	0.55	0.75	**4.26**	0.57	0.65	NaN	0.23
RF	**0.77**	**0.23**	0.70	**0.62**	**0.80**	3.53	0.45	0.63	**0.80**	**0.34**

Results (Cont.)

- ANOVA_RF (Cont.)
  - Choosing the number of top features ranked by RF for ANOVA_RF method:



Results (Cont.)

- Forward_RF:

Feature representation method: Forward_RF

Number of features=13

Features: Forward=RF=[100, 253, 197, 146, 131, 455, 179, 174, 437, 119, 464, 158, 196]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.62	0.38	0.42	0.45	0.71	1.56	0.14	0.46	NaN	0.10
NB	0.66	0.34	0.59	0.47	0.71	3.01	0.27	0.51	0.62	0.20
NN	0.60	0.40	0.36	0.34	0.67	1.21	-0.01	NaN	NaN	0.00
SVM	0.65	0.35	NaN	0.37	0.68	NaN	NaN	NaN	NaN	0.00
RF	**0.73**	**0.27**	**0.69**	**0.64**	**0.77**	**3.77**	**0.46**	**0.64**	**0.76**	**0.45**

Results (Cont.)

- RF_Forward:

Feature representation method: RF_Forward

Number of features=16

Features:

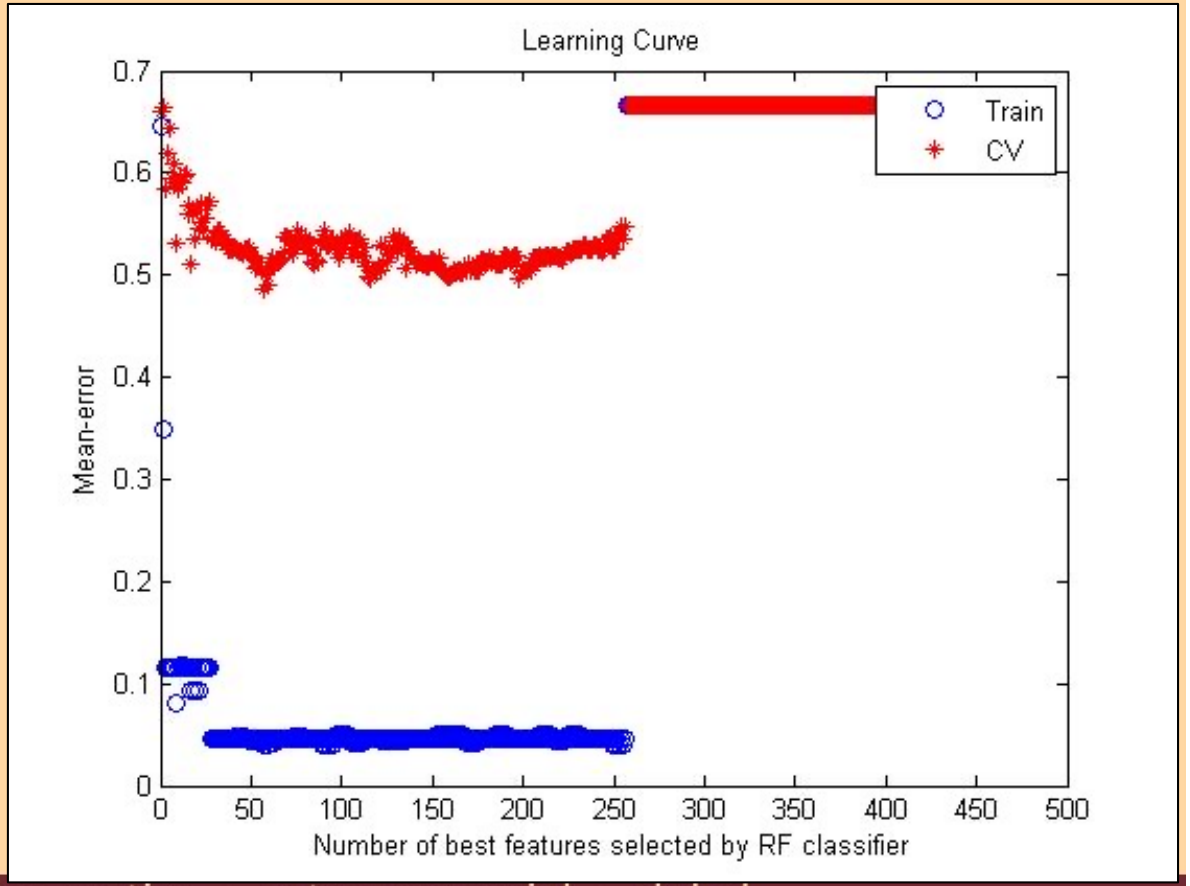
RF=[282, 52, 116, 103, 40, 43, 46, 114, 177, 106, 337, 54, 397, 361, 199, 387, 324, 219, 231, 430, 191, 383, 373, 129, 382, 425, 58, 1, 277, 185, 47, 217, 311, 198, 16, 334, 187, 48, 65, 381, 360, 238, 165,384,70, 194, 105, 153, 72, 303, 190, 263, 8, 302, 200, 79, 385]

Forward=[28, 24, 10, 56, 3, 25, 8, 6, 45, 19, 11, 41, 12, 32, 17,16]

Metric=> Classifier	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
DT	0.69	0.31	0.64	0.53	0.75	2.64	0.42	0.56	NaN	0.28
NB	0.62	0.38	0.49	0.46	0.69	2.24	0.19	0.54	0.62	0.22
NN	0.63	0.37	**0.75**	0.44	0.71	**4.40**	0.45	0.58	NaN	0.09
SVM	0.60	0.40	0.47	0.41	0.69	1.99	0.14	0.39	NaN	0.05
RF	**0.76**	**0.24**	0.70	**0.67**	**0.80**	3.50	**0.48**	**0.65**	**0.78**	**0.47**

Results (Cont.)

- RF_Forward (Cont.)
  - Choosing the number of top features ranked by RF for RF_Forward:



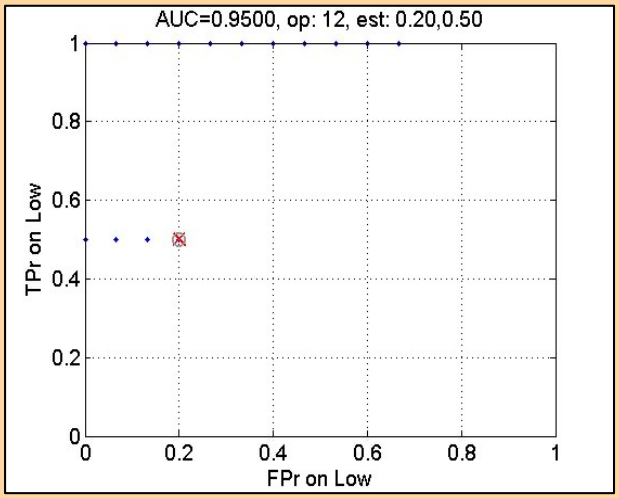
Results (Cont.)

- The performance of the Default classifier using 10-fold and leave-one-out cross validations:

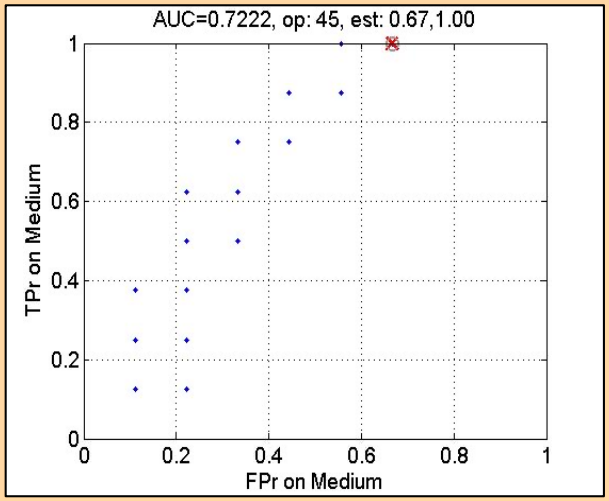
Classifier: Default										
Metric=> / Evaluation Method	ACC	ERR	PRE	REC	SPC	GAIN	MCC	F-Score	AUC	G-mean
10-fold	0.55	0.45	NaN	0.33	0.67	NaN	NaN	NaN	NaN	0.00
leave-one-out	0.48	0.52	NaN	0.33	0.67	NaN	NaN	NaN	NaN	0.00

Results (Cont.)

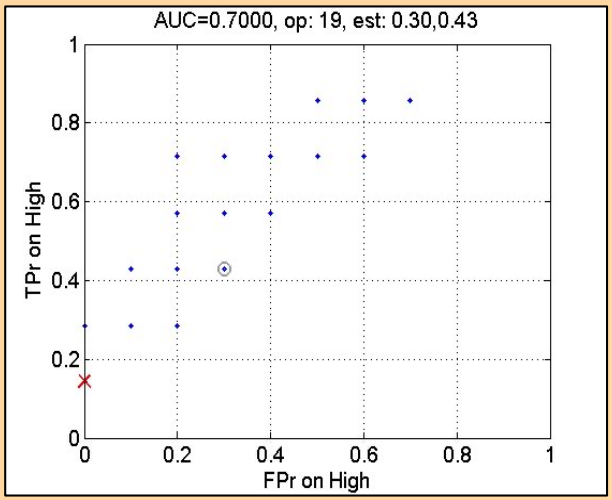
- Prediction performance (Cont.)
  - ROC for the 3 classes Low, Medium and High, obtained for row 1 of the previous table , in the last fold of cross validation:



Low



Medium



High

Discussion (Cont.)

Aspect	Discussion
Large feature space	• All the examined classifiers, except random forest, performed very poor on the original feature set because of the curse of dimensionality problem. • As random forest builds each tree with a subset of data records and a subset of features, the curse of dimensionality problem is avoided to a great extent.
Imbalance data	• The imbalance data obviously affects classifiers performances in a negative manner and makes the prediction task more difficult (compared to a balance dataset). • The structure of random forest, as an ensemble learner, helps to tackle this problem. It is because each tree of random forest is built using a random sub-sample of data. Randomness increases the chance of building a tree with a more balanced subset.

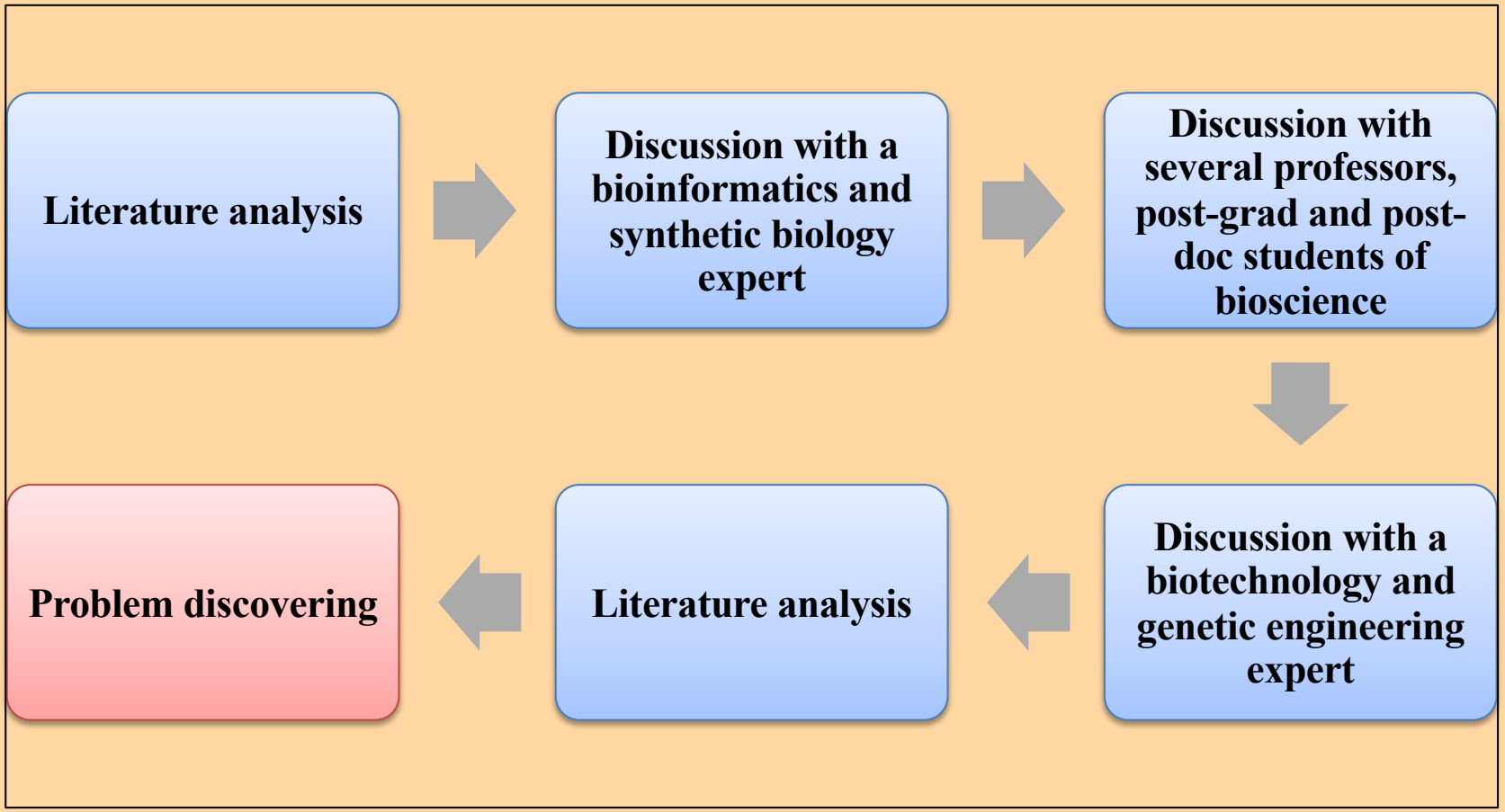
Discussion (Cont.)

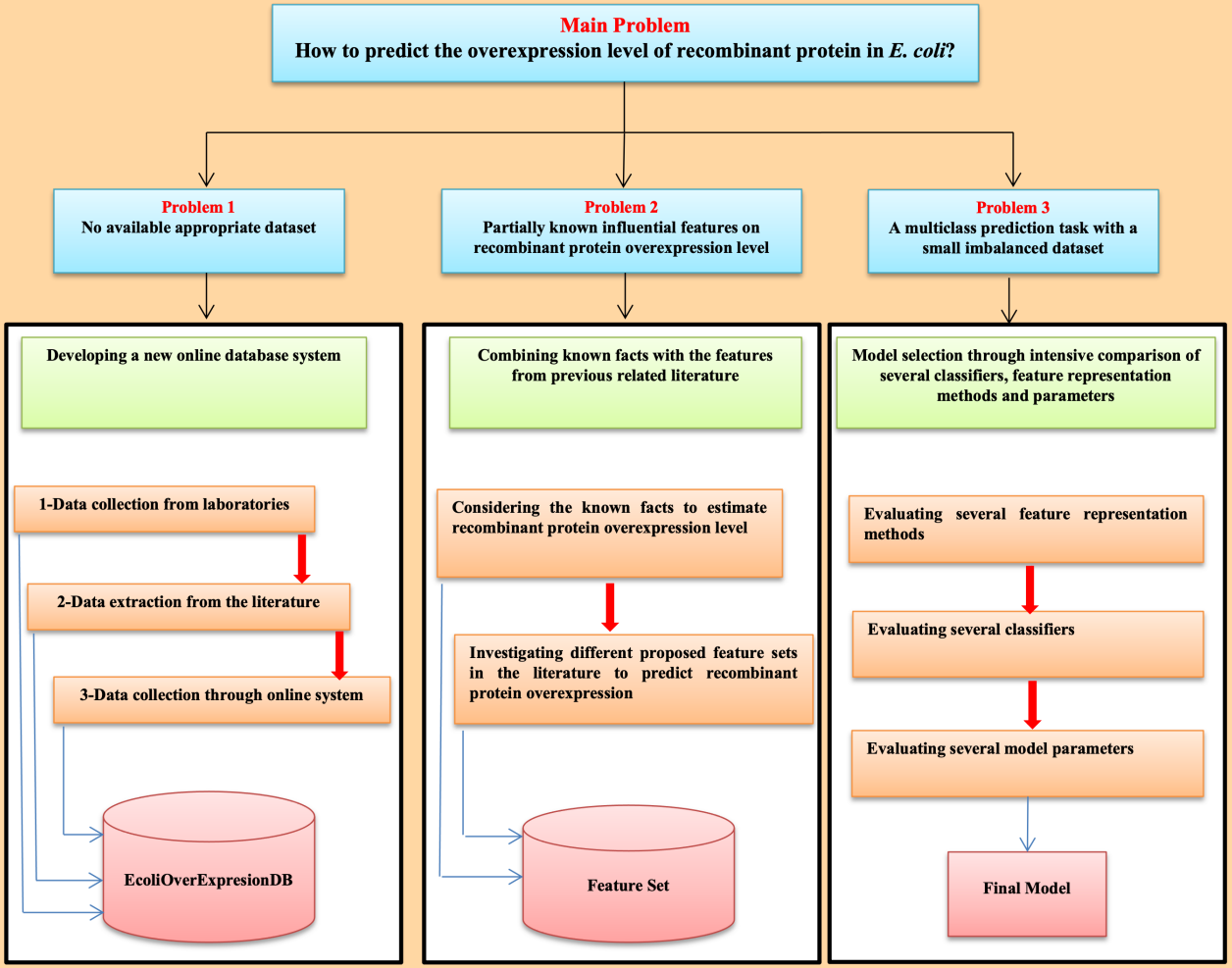
Aspect	Discussion
Small dataset	• Because random forest creates several trees with overlapping random subsets of data, the problem is diminished to some extent.
Random forest classifier	• Totally, it is concluded that random forest is very less sensitive to the feature space dimension. • Also it shows less sensitivity to the imbalance and small data, compared with the other tested classifiers.

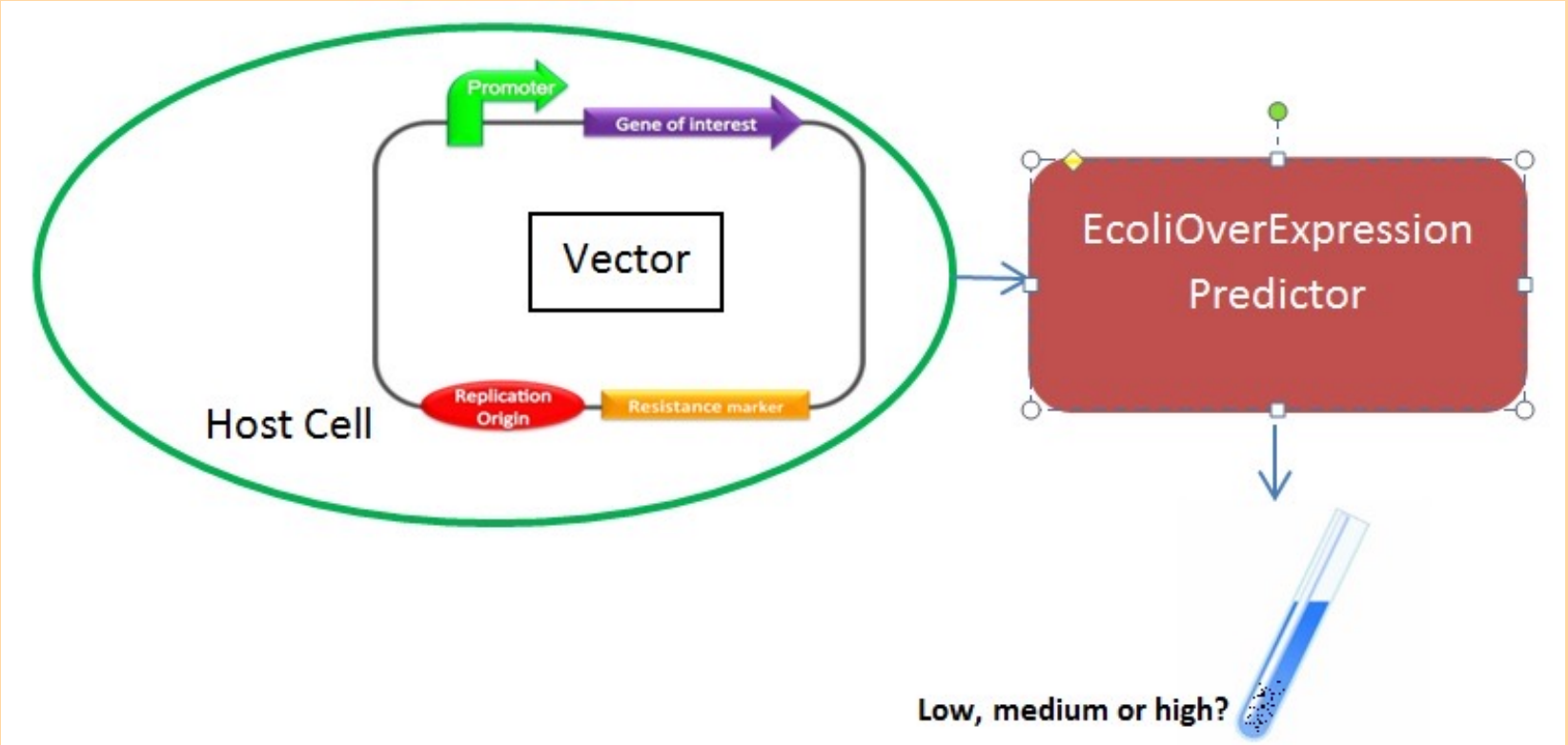
Number of Trees

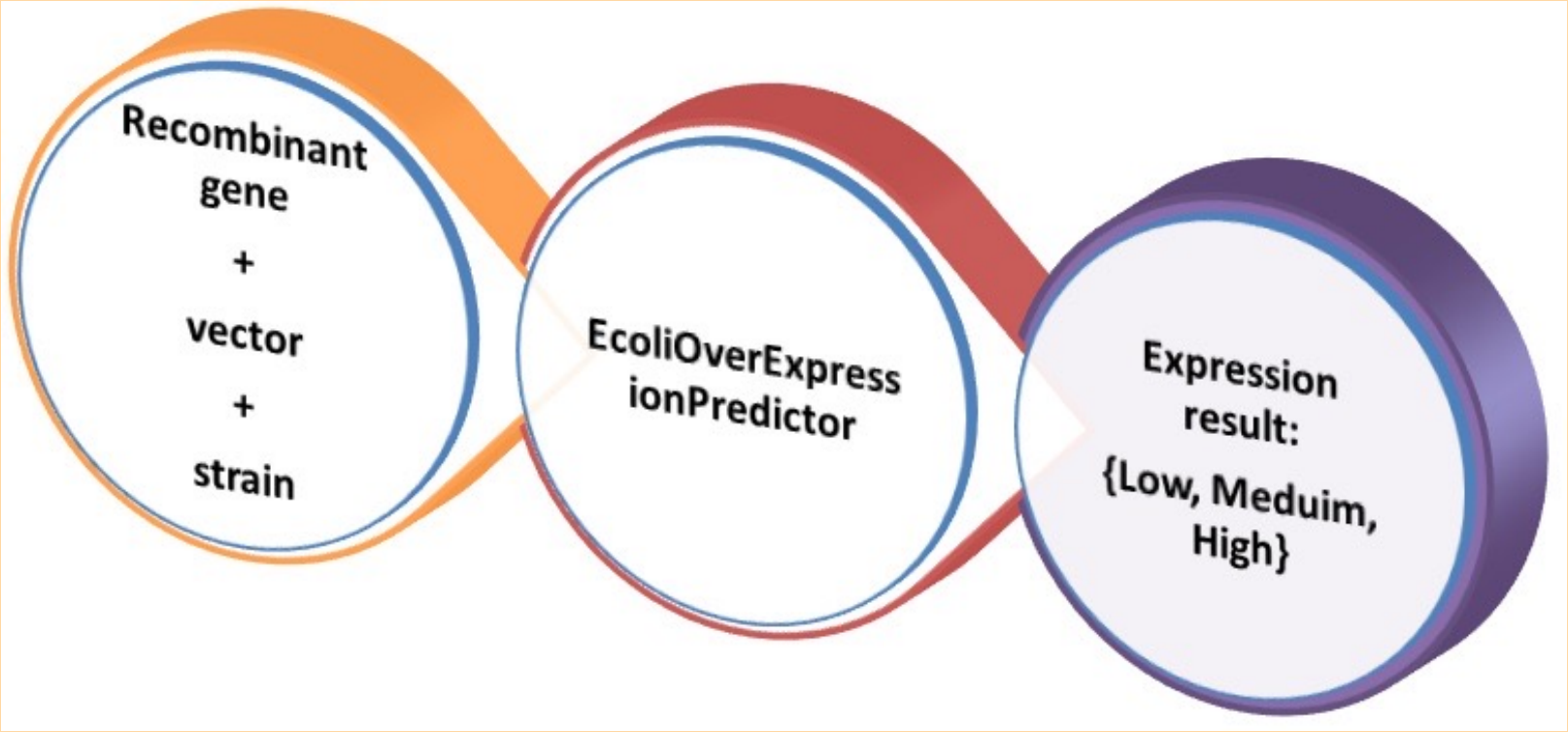
Experiment	Minimum Error	Number of Trees (which produces the minimum error)
1	0.3274	93
2	0.3097	131
3	0.3097	498
4	0.2920	161
5	**0.2743**	307
6	0.2920	293
7	0.3009	60
8	0.3009	69
9	0.3186	515
10	0.3186	210
Average number of trees ~ **233**		

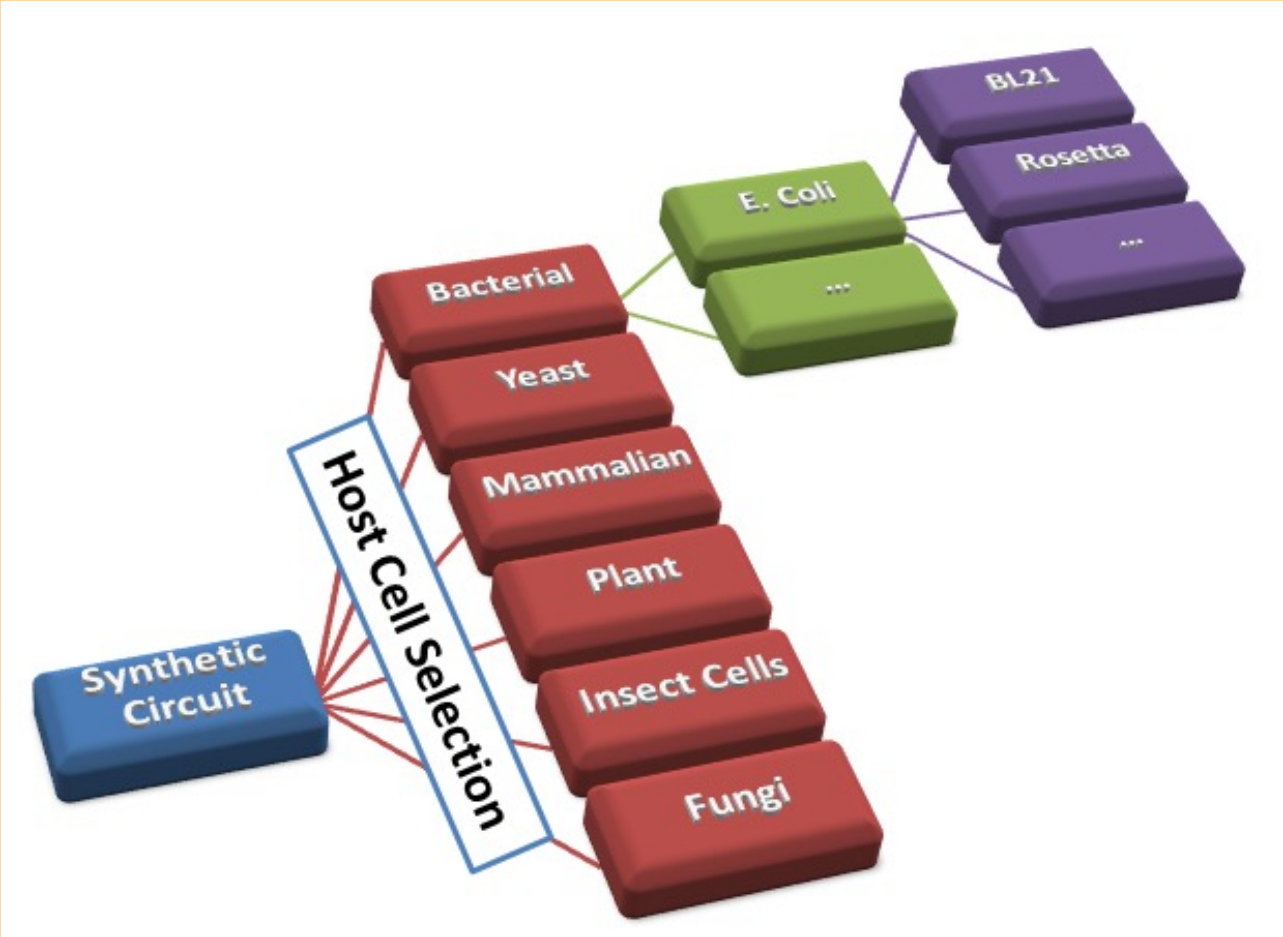
Problem Discovery Procedure




Main Problem
How to predict the overexpression level of recombinant protein in E. coli?
Problem 1
No available appropriate dataset
Problem 2
Partially known influential features on recombinant protein overexpression level
Problem 3
A multiclass prediction task with a small imbalanced dataset
Developing a new online database system
1-Data collection from laboratories
2-Data extraction from the literature
3-Data collection through online system
EcoliOverExpresionDB
Combining known facts with the features from previous related literature
Considering the known facts to estimate recombinant protein overexpression level
Investigating different proposed feature sets in the literature to predict recombinant protein overexpression
Feature Set
Model selection through intensive comparison of several classifiers, feature representation methods and parameters
Evaluating several feature representation methods
Evaluating several classifiers
Evaluating several model parameters
Final Model


Promoter
Gene of interest
Vector
Replication Origin
Resistance marker
Host Cell
EcoliOverExpression Predictor
Low, medium or high?


Recombinant gene
+
vector
+
strain
EcoliOverExpressionPredictor
Expression result:
{Low, Meduim, High}


Synthetic Circuit
Host Cell Selection
Bacterial
Yeast
Mammalian
Plant
Insect Cells
Fungi
E. Coli
...
BL21
Rosetta
...