DISSERTATION INFORMATION

Title: BINNING OF METAGENOMIC SEQUENCES BASING ON CLASSIFICATION AND CLUSTERING TECHNIQUES.

Major: **Computer Science.**

Major code: **62.48.01.01.**

PhD student: **Le Van Vinh.**

Advisors: **1. Assoc. Prof. Dr. Tran Van Lang**

2. Assoc. Prof. Dr. Tran Van Hoai

Institution: **Ho Chi Minh City University of Technology, Vietnam National University – Ho Chi Minh City.**

1. Objectives

This dissertation aims to propose efficient binning approaches for metagenomic reads which are able to work well with short reads, and deal with the limited availability of reference databases.

2. Contributions of the dissertation

- **Contribution 1:** This dissertation proposes methods to enhance the quality of the genome abundance-based binning of metagenomic sequences. There are three major contributions of this study to the problems: (1) using a reduced statistical model which requires small costs to find maximum likelihood estimates of its parameters; (2) applying a model selection method to detect the number of clusters in datasets automatically, which could improve the classification quality; (3) proposing a variable-length *l*-mer counting method in order to boost the quality of abundance-based binning approaches in case of error-free sequencing sequences.
- **Contribution 2:** The dissertation proposes an idea of selecting a representative of a group of reads belonging to the same genomes using the sequence overlapping information between reads. The representative shows that it still statistically

contains similarity-based and composition-based features. Thus, it is able to preserve features of the original read group and can be applied to reduce computational costs while still keeping the quality of binning approaches

- **Contribution 3:** A novel unsupervised method is proposed to classify metagenomic reads using the feature of *l*-mer frequency and the sequence overlapping information between reads. The proposed approach, called BiMeta, uses the idea of the selection of group representatives to reduce computational costs as well as achieve good classification quality. BiMeta consists of two main phases. In the first phase, reads are grouped by utilizing the information of sequence overlapping. The second phase merges the groups basing on the feature of *l*-mer frequency extracted from their representative.
- **Contribution 4:** A semi-supervised method, called SeMeta, is proposed to classify and label reads. SeMeta also applies the idea of selecting representative of a read group, but it utilizes the ability of preserving the similarity-based feature to separate reads. The approach consists of two major steps. After clustering reads using an improvement of BiMeta, it assigns each cluster to the best suitable taxon basing on the similarity between reads in a representative of the cluster and reference databases. Besides, an efficient filtering technique is also proposed to reduce noises (ambiguous hits) in results of the similarity search, which aims to produce better classification quality.

Advisors	PhD Student
Assoc. Prof. Dr. Tran Van Lang	Le Van Vinh
Assoc. Prof. Dr. Tran Van Hoai	