

TSINGHUA UNIVERSITY

REFERENCE FOR UNDERGRADUATE

A Short Material for *Introduction  
to Computational Biology*

*Songpeng Zu*

supervised by  
Dr. Shao Li

March 24, 2016



# Preface

This material is specific for an undergraduate class called *Introduction to Computational Biology*, which is taught by Dr. Shao Li in Tsinghua University. In this class, we mainly cover:

- Basic knowledge about the central law from DNA to protein.
- Basic diverse omics knowledge and analysis.
- Network pharmacology and modern traditional Chinese medicine.

We also have several exercise classes focus on computational methods on bioinformatics, such as

- Sequence alignment, like Smith-Waterman algorithm, multiple sequence alignment.
- Unsupervised machine learning methods, such as principle component analysis, clustering (hierarchical clustering, K-means) for gene expression data, phenotype-genotype data or other related omics data.
- Supervised machine learning methods, such as linear regression, logistic regression, LASSO for predicting drug responses based on genomic data and so on.

In other words, this class involves both basic biological knowledge and basic machine learning tools for bioinformatics.

Since bioinformatics is an extremely hot field, and lots of papers are being published to change our traditional ways to see, to analyze the biological data. It's quite important and effective way for students to take active part in our class, especially read recently published high-quality articles and make the presentations in class. Reading published articles can make students quickly understand lots of concepts, current progress, and how we do research on bioinformatics.

The difficulty lies on the limited time for our class (usually one 90-minute class for one week, and in total 16 weeks for one semester). In the past, we have five classes

as the exercise classes to talk about the frequently used computational methods. And students are hard to take active part in our classes. In this year, we try a new way. Around 24 students were classified as three groups, and each group focuses on one particular topics in bioinformatics, such as sequence analysis, genome-wide association study and network pharmacology. Each group will present one recently published article for 20 minutes in one exercise class. And we use the left 30 minutes to talk about the computational methods. Students need to pay more time to teach themselves these methods to complete the homework, and also prepare for the paper presentations.

It is indeed an interesting try, and we find students are quite interested in the class now. The shortcut is that we need a specific materials for our exercise class since we cover several parts of machine learning, statistics, and optimization in the limited time.

So we decide to write the specific material for our exercise class, and hope it would be helpful for students to teach themselves.