訂正確認報告書

訂正承認日	2016年1月25日	訂正申請日	2015年12月7日
題名	Study on the Predictions of Gene Function and Protein Structure Using Multi-SVM and Hybrid EDA		
著者氏名	Benhui CHEN		
報告者氏名	古月　敬之	確認者氏名	吉江　修

本論文は、学位規則第23条第1項に照らし、学位の取消には該当しないが、訂正を要する箇所が認められたため、これに対して著者によりなされた訂正について確認した結果を下表の通り報告する。

Gene, DNA, RNAに関する記述

訂正前：

３ページ２行目から３ページ最後まで
Genes are the units of heredity. 中略… which proteins are being produced.

訂正後：

３ページ２行目から３ページ最後まで
Genes are defined as the DNA segments which carry the heredity information. This information can produce an organism and decide the organism’s characteristics. Gene-finding programs are implemented for finding where the genes locate in a DNA sequence. If a proper stretch of DNA (such as length, start position and end position in DNA sequence, etc.) is identified in gene-finding programs. This stretch of DNA can be signed as an Open Reading Frame (ORF).

DNA

DNA is a long chain molecule which is composed of a backbone of phosphate groups and alternate sugars. A base is utilized for attaching to each sugar. And the base sequences along the backbone produces the code. There are four categories of bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). According to a view of computer science, DNA can be considered as a long letter string with four letters A, G, T and C.

Encoding and replicating the information required to produce proteins are the main functions of DNA. All the approximately twenty amino acids that produce proteins are coded by different combinations of four DNA bases. A codon is defined as each triple of DNA. Three of the triples are utilized for encoding “stop” codons, which indicate the cellular reading code mechanism where to stop.

DNA is double stranded structure which two long chain of molecules entwined together in the well-known double helix. The complementary base pairing exists in two strands. Each A in one strand is paired with a T in the other and each C with a G.

RNA

RNA is utilized for translating DNA to proteins. RNA is a nucleic acid which its structure very likes to DNA but only single stranded. There are four bases of RNA: A, G, C and U (Thymine is replaced by Uracil). RNA undertakes several important roles in a cell. Taking a copy of one of the strands of DNA is the primary role of RNA. This messenger RNA can undertake splicing to remove non-coding regions (introns) of a gene. At last, the base sequences of RNA are translated into amino acids for producing the proteins.

訂正理由と内容・訂正を認めた理由：

引用に関して訂正を要する箇所が認められたため、該当部分の記述を改めた。本部分は研究背景で、周知の専有名詞に関する記述であり、訂正することは本旨に影響を与えないことから、本訂正は妥当と判断する。

Protein and Protein Folding と HP モデルに関する記述

訂正前：

５ページ節 1.3.1 下の行から８ページの下から３行目まで
Proteins are extremely important molecules 中略… by one amino acid and the walk cannot intersect itself.

訂正後：

５ページ節 1.3.1 下の行から８ページ下から９行目まで
Investigating the proteins’ roles in cell is very important to understand the operation of the whole cell. Biology knowledge indicate that proteins are extremely important molecules. Proteins participate almost all the works in the cell, such as immunity, transportation, muscle structure, metabolism, hormones, repair, respiration and control of genes.

As we all know, a protein molecule is a string of amino acids that are connected by peptide bonds. There are approximately twenty different types of amino acids in protein. For computational convenience, a protein always is represented by a character string in computational algorithms.

The protein conformation means how the long chain of amino acids folds in 3-D space. Many research works have been implemented on protein structure determination because of it is associated with the functions of the protein. Investigating protein conformation is a key to understanding its interaction with RNA, DNA and enzyme. The protein conformation information can indicate essential knowledge for protein engineering and drug design.

A protein can be described in terms of its four main structural descriptions as follows. 1) **Primary Structure** refers to the linear sequence of protein amino acid units. These sequences act as instruction for protein, and are results of the human genome project. 2) **Secondary Structure** presents a protein by substructures or regular patterns in the polypeptide backbone. These substructures or regular patterns are defined as one of three motifs (alpha-helix, random coil and beta-sheet). 3) **Tertiary Structure** describes the 3-D conformation of protein molecule. A protein cannot realize functional until its tertiary structure is formed. The spatial relationships between the protein secondary substructures are presented in the tertiary structure. 4) **Quaternary**

Structure describes the molecular relationships and structures when a group of protein molecules are assembled for forming larger molecules. The protein three-dimensional tertiary structure is defined and folded by the protein primary form. Some environmental conditions (such as the specific chaperon or helper molecules and the correct pH condition) are required in the protein self-folding procedure. Some means of thermal or chemical kinetics can be utilized to unfold a protein. When the agent is withdrawn, some proteins also can refold into their tertiary structures. A protein self-folding process can be completed in a few microseconds. Some large protein molecules may spent minutes or hours for self-folding. Protein structure prediction is a research area that focuses on accurately predicting the protein tertiary structures based on their primary structures. In the conventional structure elucidation biological experiment for protein tertiary structure, Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography are utilized to identify the positions of atoms within the molecular environment.

However, the conventional biological method is a slow and expensive process. As a result, only a few proteins’ structures are experimentally determined while the number of known protein sequences exceeds millions. Therefore, utilizing computer programs to predict the protein conformations from protein sequences becomes a key means for uncovering proteins’ 3-D structures and functions [4, 48, 9].

Under specific conditions, the protein sequence folds into a unique native 3-D structure. Each possible protein fold has an associated energy. According to the thermodynamic hypothesis, a native protein conformation corresponds to the one which the free energy fulfils the global minimum. Based on this hypothesis, many methods that search for the protein native structure define an approximation of the protein energy and use optimization algorithms that look for the protein fold that minimizes this energy. These approaches mainly differ in the type of energy approximation employed and in the characteristics of the protein modeling. Due to the complexity of PSP task, some simplified models (such as Dill’s HP-lattice [56] model) are always used as tools to investigate the properties of protein conformation.

1.3.2 Hydrophobic Polar (HP) Model

According to the assumption of thermodynamical approaches to the protein tertiary structure prediction, a protein folding procedure searches the global minimum of Gibbs free energy. The simplest models of this assumption further simplifies that the interactions between hydrophobic amino acids are the major contributions to the free energy of a protein’s native folding [31]. Biology knowledge indicate that the hydrophobic amino acids in a protein sequence tend to push together, on the contrary, the hydrophilic amino acids are inclined to push on the molecule outside [30]. This well-studied simplified model for protein folding is called as hydrophobic polar (HP) model. [56].

Commonly studied proteins consists of approximately 20 different amino acids (as showed in Tab. 1.2). Only two properties of amino acid information in the sequence, whether the amino acid belongs to hydrophobic

(H) or belongs to polar (P) (also known as hydrophilic), are considered in HP model. Table 1.2 shows the assignment of residues as hydrophobic or polar [31]. The string of amino acids are folded on a 2-D or 3-D lattice for seeking conformations which the number of topologically adjacent H − H neighbors achieves the global maximum.

訂正理由と内容・訂正を認めた理由：

引用に関して訂正を要する箇所が認められたため、該当部分の記述を改めた。本部分は研究背景で、周知の専有名詞・術語に関する記述であり、訂正することは本旨に影響を与えないことから、本訂正は妥当と判断する。

Benchmark datasetsに関する記述

訂正前：

42ページ節3.4.1下の行から43ページ6行目まで
Nowadays, the number of protein sequences being　中略… is associated with a label that identifies the functional family of the sequence (if known).

訂正後：

42ページ節3.4.1下の行から43ページ6行目まで
Nowadays, many protein sequences are stored in central protein databases from labs all over the world. And the number of sequences is constantly increasing. Only a fraction of these protein sequences has been experimentally analyzed for detecting their structure and their functions in the corresponding organism. The main reason is that experimental determination of protein structure is time-consuming and labor-intensive. Therefore, the automatic computer program tools that can classify new proteins to their corresponding structural families are important and imperative. With the contribution of modern data analysis techniques, such as machine learning and knowledge discovery, the issue has been approached computationally, thus providing fast and more flexible solutions.

The proposed improved multi-label classification method is used to solve two benchmark biology functional prediction problems of Yeast functional genomics [34] and Genbase motif-based protein classification [32].

Yeast functional genomics data. The yeast Saccharomyces cerevisiae is one of the best-studied organisms. The yeast functional genomics dataset from Ref. [34] is studied in our experiments. Each gene is described by the phylogenetic profile and concatenation of microarray expression data. And each gene is associated with a set of functional labels whose maximum size is very large (more than 190). The whole set of functional classes is structured by hierarchies up to four levels deep. In order to make it simplified, the dataset is preprocessed by Elisseeff and Weston. Only the known structure of the functional classes are utilized in dataset. In this chapter, the same data set as used in [34] is adopted. In this data set, only functional classes in the top hierarchy are considered.

Genbase motif-based protein classification data. This problem is studied in [32]. The protein chain can be mapped into a proper motif sequence representation according to attributes. A very important issue in the data mining process is the efficient choice of attributes. The motif sequence representation supports the efficient function of data-driven algorithms, which represent samples as classified part of a fixed set of attributes. In this problem, protein chains are represented utilizing a proper motif sequence vocabulary. Suppose there are N motifs in the vocabulary. Given a protein sequence typically contains a few motifs in the vocabulary. A protein sequence is encoded as an N bit binary vector. The i-th bit is encoded as 1 if the corresponding motif is present in the sequence; otherwise the corresponding bit is encoded as 0. And each N-bit binary vector is associated with a set of functional label (if known).

訂正理由と内容・訂正を認めた理由：

引用に関して訂正を要する箇所が認められたため、該当部分の記述を改めた。本部分は周知のテスト用 Benchmark Datasets に関する記述であり、訂正することは本旨に影響を与えないことから、本訂正は妥当と判断する。

TPR 一貫性アンサンブル法に関する記述

訂正前：

51ページ節4.2.3下の行から51ページの下から10行目まで
TPR consistency ensemble method. 中略… according to the true path rule.

訂正後：

51ページ節4.2.3下の行から51ページの下から10行目まで
TPR consistency ensemble method has proposed in Ref. [90] for solving the HMC problem. The True Path Rule can guarantee the uniformity of gene function annotations in FunCat taxonomies: “If the child term describes the gene product, then all its parent terms must also apply to that gene product”. That is to say, if a gene is labeled by a specific functional class, then it should be labeled by all its “parent” classes, and also should be labeled by all its ancestor classes in a recursive way.
The ensemble method is implemented by a two-way asymmetric information flow that traverses the graph-structured classes. The positive predictions for a node influence its ancestor nodes by a recursive way. And the negative predictions influence its offspring nodes.

For a given example x and a class node ci, a classifier in TPR consistency ensemble should obey the rules as follows: {

$$di = 1 \Rightarrow dpar(i) = 1$$

$$di = 0 \Rightarrow dchild(i) = 0 \qquad (4.2.1)$$

The TPR method realizes an ensemble that respects the “true path rule” by putting together the classification results predicted at each node by local “base” classifiers. Positive predictions of local classifiers are propagated from bottom to top across the graph in a recursive way. They effect the predictions of their ancestor nodes by traversing the graph towards higher level nodes. Negative predictions for a given node are propagated to their descendant nodes for preserving the consistency of the hierarchy according to the true path rule.

訂正理由と内容・訂正を認めた理由：

引用に関して訂正を要する箇所が認められたため、該当部分の記述を改めた。本部分は既存法に関する記述であり、訂正することは本旨に影響を与えないことから、本訂正は妥当と判断する。

EDA進化型アルゴリズムに関する記述

訂正前：

69ページ節5.2.2下の行から69ページの下から6行目まで
In EDAs[67], there are neither crossover nor mutation operators. 中略… of direction encoding on an EAs performance.

訂正後：

69ページ節5.2.2下の行から70ページ2行目まで
There are neither mutation nor crossover operators in EDAs [55]. A probabilistic model is used to capture the probability distribution from a database that includes the selected individuals from previous generation. And this probabilistic model is utilized to sample the new population for EDA. Therefore, through the joint probability distribution of the selected individuals at each generation, the interrelations between the different variables that represent the individuals are explicitly expressed by the probabilistic model. EDAs suppose that it is feasible to obtain a probabilistic model for the promising areas of search space. And this model can be utilized to guide the search for the optimum solutions. A condensed representation of the features shared by the selected individuals is used to build the probabilistic graphical model in EDAs. By the probabilistic graphical model, different patterns of interactions between subsets of the problem variables can be captured efficiently, and this knowledge can be utilized to sample new solutions conveniently. In the algorithm of protein folding optimum, one of the important problems is how to present a specific conformation. To embed a hydrophobic pattern $S \in \{H, P\}+$ into a lattice, there are three methods of Cartesian Coordinate, Internal Coordinate and Distance Matrix [52] as follows.

1) Cartesian Coordinate representation. The residue position is represented by its Cartesian Coordinate. The positions of residues are independently.
2) Internal Coordinate representation. The residue position is decided by its predecessor residues in the sequence. Two types of internal coordinate are often used in lattice model. The one is relative direction representation where the residue directions depend on the direction of the previous move. The other is absolute direction representation where the residue directions depend on the axes defined

by the lattice.

3) Distance Matrix representation. A distance matrix is introduced to represent the residue positions. Given a residue, its location can be computed by the distance matrix.

In Ref. [52], the authors implemented a detailed comparative research on absolute and relative directions utilizing evolutionary algorithms. According to the experimental results, relative directions always have better performances than absolute directions based on square and cubic lattice. On the contrary, absolute directions outperform relative directions based on triangular lattices. Evaluation the effectiveness of direction encoding on an evolutionary algorithm is uncertain in general. But the internal coordinates with relative directions can be suggested based on the experimental evidence.

訂正理由と内容・訂正を認めた理由：

引用に関して訂正を要する箇所が認められたため、該当部分の記述を改めた。本部分は既存のアルゴリズムに関する記述であり、訂正することは本旨に影響を与えないことから、本訂正は妥当と判断する。

Appendix: 既存法に関する記述

訂正前：

109ページから130ページまで
Appendix A, B and C

訂正後：

Appendix A, B and C削除

訂正理由と内容・訂正を認めた理由：

引用に関して訂正を要する箇所が認められたため、該当部分を削除した。本部分は既存法に関する詳細記述であり、本文にこれらの手法の引用が既にあるので、削除することは本旨に影響を与えないことから、本訂正は妥当と判断する。