Studying Bioinformatics Based on Word Distributions in Proteins

055762F Kenta MOTOMURA
Supervisor : Morikazu NAKAMURA, Joji M. OTAKI

1 Introduction

In biology, studying molecular similarities are very important for understanding life (Figure 1).

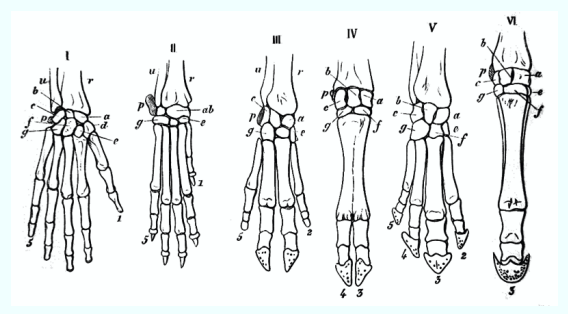

Figure 1: Similarities on old type biology [2]

There is already a very good method to study molecular similarity called "Sequence Alignment". This traditional method is very good for studying similarities of molecules but has some weak points. For example, if you attempt to calculate similarity of very large number of sequences, the calculation time increases significantly.

Therefore my research has three main objectives.

- To propose more efficient methods of studying similarities between molecules based on word distribution.
- To implement algorithms for solving biology problems with the new method.
- To verify effectiveness and usefulness.

2 Word Distributions in Proteins

We define a *word* as a short sequence (Figure 2). Proteins are made up of amino acids, so a word in a protein is a short amino acid sequence.

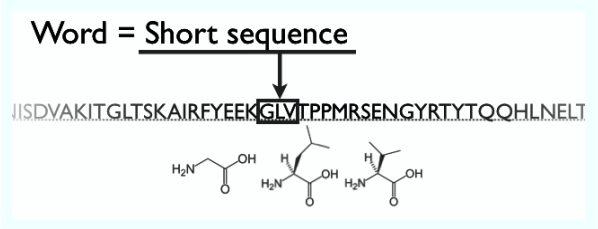

Figure 2: A word (a short sequence) in a protein

We define *word distribution* as distribution of words in a group of sequences. To make distribution data, we extract all words of a specified length from all the sequences in a database and calculate the availability[1] for each word. Then distribution for each word is obtained. Figure 3 shows a map of the word distribution where the horizontal axis means kind of words and vertical axis means availabilities of words.

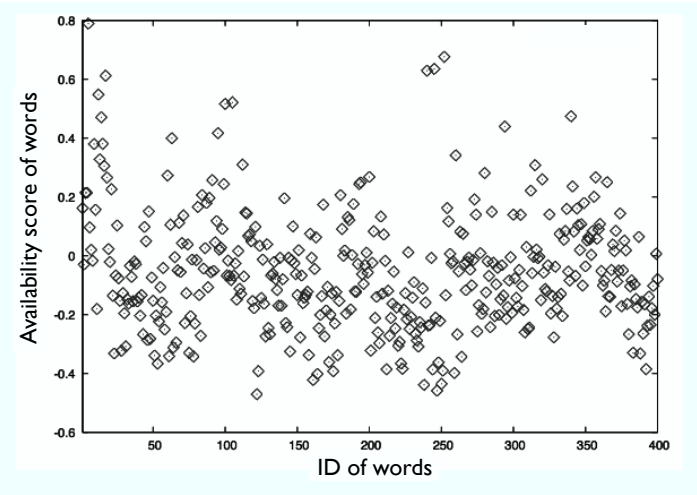

Figure 3: A distribution map of words in a protein DB

3 Similarity based on Word Distribution

With the traditional method, we just compare some sequences to acquire similarity among the sequences. On the other hand, with our new method, we compare word distribution to acquire similarity among groups of sequences (Figure 4). That is, we can obtain easily similarity between species but not between just sequences.

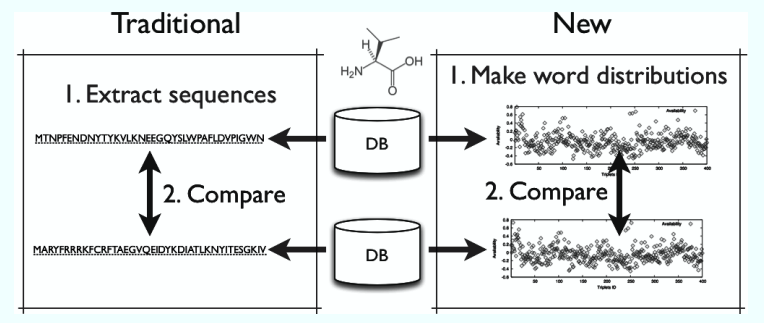

Figure 4: Calculation of similarities

The traditional method is good for comparing sequences, and the new method is good at comparing groups of sequences.

If you try to calculate similarities between groups with many sequences using the traditional alignment method, you have to compare many pair of sequences. You may have to make over 10 trillion comparisons. The new method needs only several comparisons of distributions, that is far more beneficial time-wise.

4 Case study

4.1 A phylogenetic tree

A phylogenetic tree is a tree showing the evolutionary relationships among species, DNA or other biological entities. We made a phylogenetic tree based on our approach, word distribution, to check its usefulness.

4.2 Procedure

We made a phylogenetic tree with our new method by the following procedure.

1. Download protein databases from websites such as NCBI in the U.S.

2. Make word distribution data for each database.

3. Compare word distributions and make distance data from the similarities.

4. Make a phylogenetic tree from the distance data.

If we can construct a biologically-accurate phylogenetic tree within a short calculation time, we can say the new method is efficient. We do not discuss the biological accuracy of the obtained tree since it can be performed by biologists. Our research project is a collaboration work with a biological laboratory.

4.3 Results

We produced a phylogenetic tree including 110 species with a comprehensive comparison of sequences (Figure 5). We compared species, that is, 110 groups of sequences, and made the tree in a very short period of time.The calculation time of our method is about 500 seconds.

By using the new method, you can make trees with comprehensive comparisons of sequences in a very short time and that are not subject to humans subjectivity.

Suppose we need 0.05 seconds per one pairwise alignment. Let a_i be the number of sequences for species i. The comprehensive pairwise alignment requires $\sum_{i=1}^{k-1}\sum_{j=2}^{k} a_i a_j$ time execution. In the experiment, we calculated similarity for 110 species, these are several thousand to several ten thousand sequences for a species, for example, 59679, 24780, 12252, 7603. Therefore, we need more than 63 billion seconds.

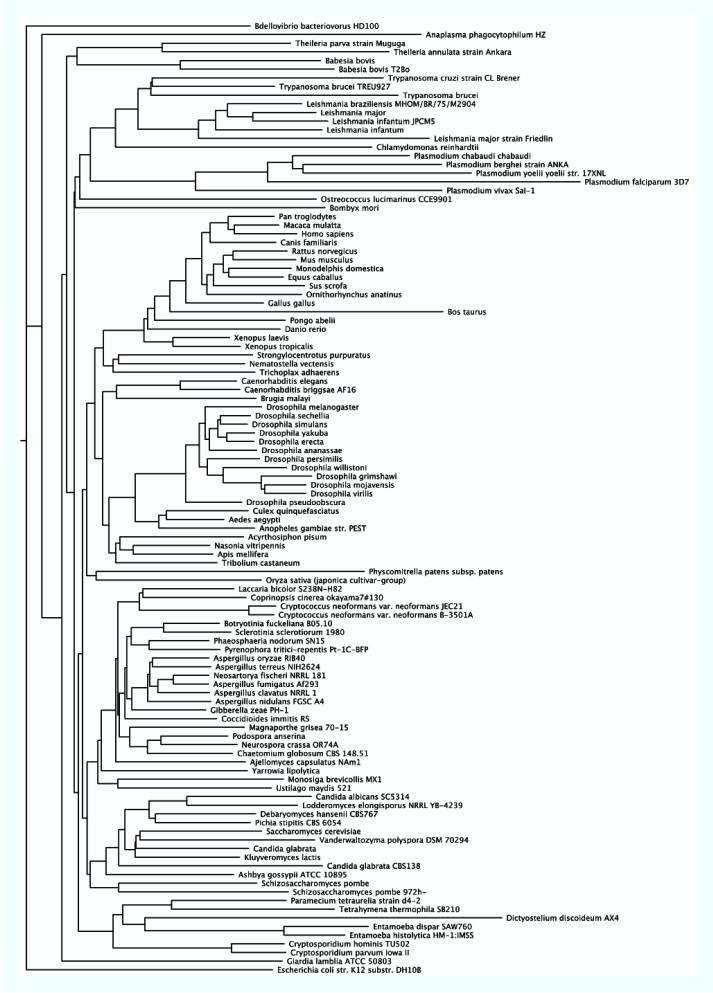

Figure 5: A phylogenetic tree created with our new method

5 Conclusion

In this research, we proposed a new method to calculate molecular similarities. The most important concept of the method is using word distributions. And we showed the usefulness by making a phylogenetic tree with proposed method. In that experiment, we made a biologically-accurate tree in a very short time.

In terms of future research, we plan on clarifying the potential of the method to solve other biology problems.

References

[1] Joji M. Otaki, Tomonori Gotoh and Haruhiko Yamamoto. Potential implications of availability of short amino acid sequences in proteins: An old and new approach to protein decoding and design, Biotechnology Annual Review 14: 109-141 (2008).

[2] Gegenbaur, Carl Grundzüge der vergleichenden Anatomie. 2. umgearb. Auflage. Mit 319 Holzschnitten. Leipzig, Verl. von Wilhelm Engelmann: 692 (1870).