Adapting Normalized Google Similarity in Protein Sequence Comparison

Lee Jun Choi and Nur'Aini Abdul Rashid
School of Computer Science
University Sains Malaysia, Penang
leejunchoi@gmail.com;nuraini@cs.usm.my

Abstract

Biological sequence comparison faced various challenges. Although dynamic programming based solution claimed to be the optimal solution for the comparison process, the computation limitation and some fundamental challenges still make it inefficient for mass sequence comparison. Statistical method explores the statistics of sequences by the frequency of the words in the sequence; it provides a comparison solution without loss of statistical information, and also caters some of the fundamental problem in sequence comparison. Normalized Google Distance is a way of finding semantic similarity in web pages, with significant related characteristics; in this research, we propose an algorithm that will integrate Normalized Google Similarity into protein sequence comparison.

1. Introduction

Protein Comparison had been one of the crucial process in protein analysis. The similarity of sequences is an important variable for biological operation such as protein clustering and functionality study. Various approaches had been explored to obtain similarity between sequences, the most common approaches are through sequence alignment, which two or more sequences are lined up together and amino acid or DNA that similar to each other are aligned together. Dynamic Programming based protein alignment is the optimal solution for sequence comparison with the high cost computational load.

Although optimal solution had been identified for biological sequence comparisons, there still exists some limitation in the process, especially with alignment based comparison. The major issue of current stage of biological sequence comparison is viewing the biological data as linear based instead of 3D structured. And in alignment approaches, we also faced the issues of well-documented long-range interactions of sequences are ignored and the variety of alignment because first aligning reference homologous sequence to derive a score for the alignment of individual units. Besides the biological concept, alignment approaches also have problem with high cost of computer load and computational load and not catering the sequence divergences and difficulty of accessing the statistical relevant of the resulting score.

Statistical based sequence comparison approach was introduced to cater some of the protein comparison challenges like the memory consumption, the loss of statistical relevant of the sequence compared and the issue of relying on the first aligning reference homologous sequence to derive a score in alignment approach; this approach is a non-alignment based sequence comparison approach.

Google Distance is a semantic interrelatedness measurement derived from the number of hits in the Google search engine for a given set of keywords. The Google Distance value trends to be very small if the keywords were closely related, while the value will be larger if the keywords are less related.

In Normalized Google Distance between term X and term Y,

$$NGD(X,Y) = \frac{MAX\{\log f(X), \log f(Y)\} - \log f(X,Y)}{\log M - MIN\{\log f(X), \log f(Y)\}}, \quad (1)$$

where M is the total number of web pages searched by Google; *f*(X) and *f*(Y) are the number of hits for terms X and Y respectively; and *f*(X,Y) is the number of web pages on which both X and Y occur.

The Normalized Google Distance is infinite if the term X and Y never occurs together, and the Normalized Google Distance will be equal to zero if the terms always occur together.

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

2. Related Work

There are various statistical based sequence comparisons where most of the method uses k-word or k-tuple frequency as the foundation of the comparison. The simplest similarity measurement is by using the Euclidean distance of the k-word frequency which introduced in 1986 [2]. The method was extended by Pevzner and Torney's group by applying filtration techniques to deduct several of the characteristic measures for search optimization [12] and weight of individual K-tuples were add in the Euclidean distance measurement to maximize the variance of reference sequences with regard to random sequences [4]. The Euclidean distance is further explored in biological sequence comparison with the Standard Euclidean Distance and the Mahalanobis distance which consider the variances of k-words in the calculation [20,21].

Besides Euclidean distances, measurement based on information theoretic measures that uses relative entropy, which known as the Kullback-Leibler distance [21] is also been used for sequence similarity. Finchant and Gautier introduced the correlation coefficient structure [6] while Petrilli implemented the measurement by calculating the linear correlation coefficient (LCC) between two sequences in the process of classifying proteins based on di-peptide frequencies [13]. The angle metrics that investigate the distance between sequences with the angle between the k-tuple count vectors is also been studied and applied [7,8].

Google Distance (NGD) had been implemented in hierarchical clustering, machine learning and also language translation [15]; it also had been apply in the study of extracting meaning from world-wide-web [1].

3. Contribution

The main study in this paper is to investigate if Normalized Google Similarity, which derived from Normalized Google Distances, can be used to measure the similarity between two protein sequences by using k-word frequencies. Different k-word frequency will be compared to obtain the impact of different k-word distribution against the Normalized Google Similarity in comparing protein sequences.

4. Method

This study is to evaluate the significant of Normalized Google Similarity for protein sequence similarity by studying the correlation coefficient between similarities obtained from Normalized Google Similarity with the Smith Waterman Score from FASTA result. We also run an accuracy test for the similarity search using Normalized Google Similarity with FASTA search results as the benchmark. The method is shown in Figure 1.

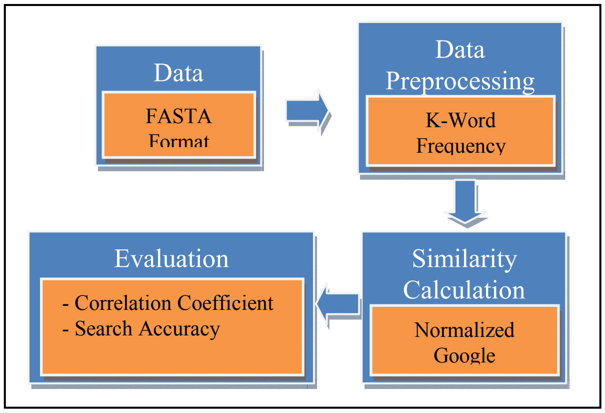

Figure 1: Methods of study and evaluation

4.1. Data

Two types of data were needed for this study, first is the result of FASTA search on particular protein sequence, which will be used for correlation coefficient study and also benchmark data for accuracy, and the second data is linear sequence of all the protein returned in FASTA search result, these sequences will be used to calculate the sequence similarity using the selected similarity indexes or calculation modules.

Several FASTA search services are available online in different sites that involves in biological research. Among the sites was the site European Bioinformatics Institute (EBI), which is a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL). The result of the FASTA search will be obtained from EBI online FASTA protein search service. The service can be access through the URL http://www.ebi.ac.uk/fasta33/. Each FASTA search will returns the first fifty protein IDs that have the highest similarity in the protein database.

Protein sequences are needed as input for the similarity calculation using the selected similarity indexes or calculation modules. Since the correlation study and accuracy test will based on the FASTA result, all the fifty protein sequences in the search result for each FASTA search will be used as input for the similarity calculation, with the FASTA search results as the comparison subject. Since there are two sets of FASTA results are generated for the study, there will

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

be one hundred protein sequences that will be used in the similarity study.

A full sequence for each of the protein ID returned by FASTA search will be retrieved from SWISS-PROT protein database.

4.2. Data Preprocessing

The protein sequences obtained from SWISS-PROT protein database are in FASTA format, which consist the protein ID, protein name, protein description and the biological sequence. A process to extract the different information in the protein sequence will be done, and the result will be stored in a relational database.

K-words/K-tuples frequency takes the occurrences of the fixed length words as the statistical characteristic of a sequence. For that, a sequence is treated as a long linear sentence or a text document, where the occurrences of words with specific length are calculated. The result is the number of occurrences for each possible word. In the process of calculating the word occurrences, words with overlapping features are ignored, only none overlapping word are calculated. (see Figure 2)

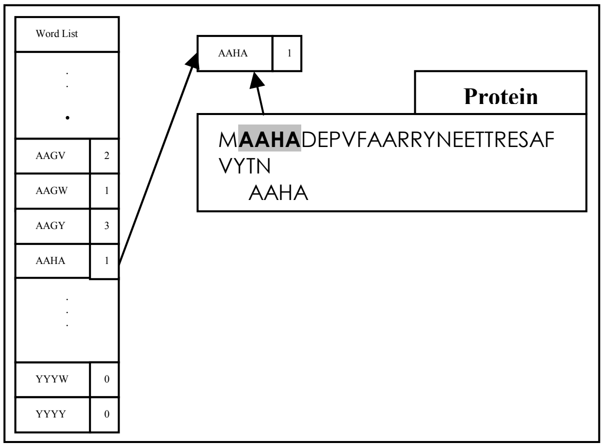

Figure 2: Example of Words Frequency Calculation

Since the study will also compare the impact of different k-words frequency distribution on the Normalized Google Similarity, four different k-words frequency will be studied. The four k-words frequencies differ in the size of words used in the k-words frequency or the K value of the frequency. The four k-words frequencies that will be used are uni-character (k = 1), bi-characters (k = 2), tri-characters (K = 3) and quad-character (k = 4).

4.3. Normalized Google Similarity Calculation

Google Distance is a measure of semantic interrelatedness derived from the number of hits returned by the Google search engine for a given set of keywords. Therefore it is estimated that it also able to locate the interrelatedness of any two protein sequences.

In the effort to compare any two protein sequences, the sequences are treated as two different web pages and the each words frequency represents terms found in each webpage, and the Normalized Google Distance of two sequences can be obtain, with

$$NGD = \frac{MAX(\sum W_X, \sum W_Y) - \sum MIN(W_X, W_Y)}{(\sum W_X + \sum W_Y) - MIN(\sum W_X, \sum W_Y)}, \qquad (2)$$

where, NGD is the Normalized Google Distance between sequence X and sequence Y, $\sum W_X$ is number of words or N-Gram partition in the first sequence X, $\sum W_Y$ is the number of words or N-Gram partition in the second sequence Y, $MAX(\sum W_X, \sum W_Y)$ is the maximum values between the number of words or N-Gram partition in sequence X and Y, $MIN(\sum W_X, \sum W_Y)$ is the minimum number of words or N-Gram partition in sequence X and Y, And $\sum MIN(W_X, W_Y)$ is the sum of minimum value between number of each different words or N-Gram partition in sequence X and Y, this also represent the number of matches between sequence X and sequence Y.

The Normalized Google Distance (NGD) for two sequences represents the dissimilarity between the two sequences. The maximum values for NGD is 1.0, which means two sequences are totally not similar to each other, and the minimum values for NGD is 0.0 values, which means both of the sequences are closely related. Therefore, the similarity of the two sequences can be obtain by

$$NGS = 1 - NGD. \qquad (3)$$

Where NGS is the Normalized Google Similarity and NGD is the Normalized Google Distance.

4.4. Evaluation

To evaluate if the Normalized Google Similarity is capable of represents the significant of comparing two sequences, three evaluations: Correlation Coefficient, Accuracy test will be conducted with the result obtained by using FASTA results as benchmark.

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

4.4.1. Correlation Coefficient. To study the correlation coefficient between the Normalized Google similarity and the Smith Waterman score obtained from FASTA search, scatterplots chart and Spearman's rank correlation coefficient will be used to study the correlation coefficient of between the similarities obtained from FASTA and the different approaches.

For each of the approaches, a scatterplots chart between the similarities obtained from FASTA and a particular approach is produced. The chart will help in observing the correlation between the FASTA and a particular approach.

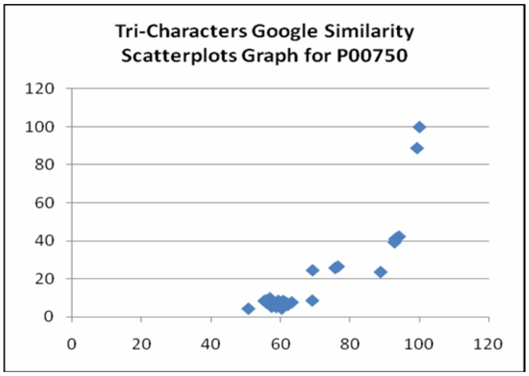

Figure 3. Sample of scatterplots graph for Tri-Characters Google Similarity based on Words Frequency

Besides the scatterplots chart, a correlation coefficient values also been calculated using the Spearman's Rank Correlation Coefficient, where if the similarity in FASTA result is represent by X, and the similarity of a particular approach is represented by Y, then

$$r_{XY} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (4)$$

where r_{xy} is the Spearman's rank correlation coefficient between the compared values , ∑ is the summation sign, d_i^2 is the difference between each rank of corresponding values of x and y, and n the number of pairs of values.

The correlation coefficient (r_{xy}) will reflect the degree of relationship between the FASTA similarity and the similarity for a particular approach. If the correlation coefficient presented a strong relationship between the FASTA similarity and the similarity for that approach, it would mean that the particular approach can be use to predict the similarity between two proteins.

4.4.2. Accuracy Test. Correlation coefficient tests studies the possibility of any existing relationship between Normalized Google Similarity and Smith Waterman score. An accuracy test examines the practicality of using Normalized Google Similarity in protein similarity based on the FASTA search result. The test is done by implementing a simple similarity search using both the selected approach on a partial SWISS-PROT protein database that consists of all the sample protein obtained from FASTA search and proteins from the similar families from the result of the FASTA search. The result is then compared with FASTA result. In this accuracy test, precision and recall of the particular combined approaches against FASTA result will be measured, where

$$\text{precision} = \frac{(\text{FASTA result} \cap \text{Approach Result})}{\text{Approach Result}}, \qquad (5)$$

And

$$\text{recall} = \frac{(\text{FASTA result} \cap \text{Approach Result})}{\text{FASTA Result}}, \qquad (6)$$

From the precision and recall values, the weighted harmonic mean of precision and recall is calculated for each of the approaches tested, where

$$F_1 = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{(\text{Precision} + \text{Recall})}, \qquad (7)$$

F_1 value is the evenly weighted F-measure for precision and recall values.

5. Expected Outcome

All different approaches of calculating protein similarity using different size of words length will go through the Correlation Coefficient evaluation. The result of this evaluation will be a list of the different k-tuple frequency and its correlation coefficient with the FASTA search result.

The correlation coefficient will shows how strong is the connection between the selected approaches against FASTA search result. The schema for the correlation coefficient against the relatedness is shows in Table 1.

Table 1 Correlation Coefficient and Relatedness

Correlation Coefficient	Relationship
0.8 and 1.0	Very Strong
0.6 and 0.8	Strong
0.4 and 0.6	Moderate
0.2 and 0.4	Weak
0.0 and 0.2	Very Weak

Similar with the Correlation Coefficient evaluation, all word size of k-word frequencies of Normalized Google Similarity will be evaluated for the accuracy of its result. The result of this evaluation will be each of the different word size of k-words frequencies with its F-measures. The higher the F-measures will mean a better accuracy for Normalized Google Similarity.

5. References

[1] A. Evangelista and B. Kjos-Hanssen, Google Distance Between Words, *Frontiers in Undergraduate Research*, University of Connecticut, 2006.

[2] B.E. Blaisdell, A measure of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. J.Mol.Evol., 1986, 29,526-573.

[3] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press, 1997.

[4] D.C. Torney, C. Burks, D. Davison, and K.M. Sirotkin, Computation of d2: a measure of sequence dissimilarity. In George,I. and Bell,T.G.M. (eds), *Computers and DNA : the proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, held December 12 to 16, 1988 in Santa Fe, New Mexico*. Addison-Wesley, Redwood City, CA, 1990,pp. 109–125.

[5] FASTA, Wikipedia, accessed 6 January 2008, http://en.wikipedia.org/wiki/FASTA

[6] G. Finchant, and C. Gautier, Statistical methods for predicting protein coding regions in nucleic acid sequences.*Comput.Appl.Biosci.,* 1987, 3,287-295.

[7] G.W. Stuart, K. Moffet, and S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, *Bioinformatics,* 2002, 18, 100-108.

[8] G.W. Stuart, K. Moffet, and J.J. Leader, A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol.Biol.Evol.*, 2002, 19,554-562.

[9] L. Brillouin, Science and Information Theory, Mineola, N.Y.: Dover, [1956, 1962] 2004. ISBN 0-486-43918-6.

[10] L. Wang and T. Jiang, On the complexity of multiple sequence alignment, J Comput Biol 1:337–348, 1994.

[11] O. Gotoh, An improved algorithm for matching biological sequences, J.Mol.Biol.. 1982, 162. 705-708.

[12] P.A. Pevzner, Statistical distance between texts and filtration methods in sequence comparison, *Comput. Appl. Biosci.*, 1992, 8,121-127 .

[13] P. Petrilli, Classification of protein sequences by their dipeptide composition, *Comput. Appl. Biosci.*, 1993, **9**, 205–209.

[14] R. Shamir, Course Notes for Algorithms for Molecular Biology, Tel Aviv University School of Computer Science, 2001.

[15] R.L. Cilibrasi, P.M.B. Vitanyi, The Google Similarity Distance, IEEE Trans. Knowledge and Data Engineering, 19:3(2007), 370-383.

[16] S. Henikoff and J.G. Henikoff, Amino acid substitution matrices from protein blocks. Proc. Natl Acad. Sci. USA, 1992, 89, 10915-10919.

[17] S. Vinga and J. Almeida, Alignment-free sequence comparison – a review, Bioinformatics, Vol. 19, 2002, pp 513-523.

[18] S.B. Needleman and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J.Mol.Biol.*, 1970, 48, 443-453.

[19] T.F. Smith and M.S. Waterman, Identification of common molecular subsequences, J.Mol.Biol., (1981), 147, 195-197.

[20] T.J. Wu, Y.C. Hsieh, and L.A. Li, A measures of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words.*Biometrics,* 1997, 53, 1431-1439.

[21] T.J. Wu, Y.C. Hsieh, and L.A. Li, Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition, *Biometrics*, 2001, **57**, 441–443.

978-1-4244-2328-6/08/$25.00 © 2008 IEEE