The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Data Compression-Based Approaches to Analysis of Biological Networks

Tatsuya Akutsu[1,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University
Kyoto 611-0011, Japan

Abstract Data compression-based methods have been effectively applied to analysis of biological sequences and protein structures. We have been extending this approach for analysis of biological networks. We are also developing grammar-based compression algorithms for structured data. In this extended abstract, we briefly review these approaches.

Keywords Data Compression; Protein Structure; Metabolic Networks; L-System; Context-Free Grammar; Graph Grammar

1 Introduction

Data compression is one of widely used technologies in computer science. It is not only useful for compression but also for useful for measuring the similarity between two objects because concatenated data of similar objects can be well compressed. Therefore, data compression has been applied to comparison of large sequence data [7, 8]. This approach has also been applied to comparison of protein structures [3, 6]. One of the advantages of use of data compression for comparison of these kinds of data is that data compression methods are usually very efficient and thus can be applied to large-scale data. Since it is still difficult to compare large biological networks, it is reasonable to try to apply data compression methods to comparison of biological networks. We have recently examined such an approach [4].

On the other hand, the human genome consists of around 3 billion base pairs whereas the number of cells in the human body is estimated to be 60 trillion. Therefore, it is considered that information on the human body consisting of 60 trillion cells is compressed into 3 billion base pairs. Deciphering this data compression mechanism is one of major goals of systems biology. We began to tackle this problem using *grammar-based compression*, where grammar-based compression is to find a small grammar that uniquely generates a given data [2]. Based on this concept, we are trying to develop a method to infer a small size *L-system* (Lindenmayer system) that can generate a given structural data, where an L-system is a grammatical system for modeling the growth processes of plant development and the morphology of various organisms [10]. As a first step towards this objective, we recently developed an approximation algorithm for computing the smallest tree grammar that generates a given tree structured data.

*E-mail: takutsu@kuicr.kyoto-u.ac.jp

In this extended abstract, we briefly review basic ideas of data compression-based approaches for comparison of biological sequences and protein structures. Then, we review our proposed method for comparison of biological networks. Finally, we review our proposed algorithm for approximating the smallest tree grammar.

2 Comparison of DNA Sequences and Protein Structures

Li et al. proposed the *universal similarity metric* (USM) [8], based on their earlier work [7]. USM is based on the concept of *Kolmogorov complexity*. Kolmogorov complexity $K(x)$ for an object x is defined to be the length of the shortest program P for a universal Turing machine which outputs x. The conditional Kolmogorov complexity of x given y is defined to be the length of the shortest program P which outputs x when y is given. Then, USM between two objects x and y is defined by

$$USM(x,y) = \frac{\max(K(x|y^*), K(y|x^*))}{\max(K(x), K(y))},$$

where x^*, y^* denote the shortest programs for generating x and x, respectively.

Since it is impossible to compute the exact Kolmogorov complexity, Li et al. proposed the *normalized compression distance* (NCD) measure [8] defined by

$$NCD(x,y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))},$$

where xy denotes the concatenation of x and y, and $C(z)$ denotes the compressed size of z.

For text data (i.e., sequence data), a number of data compression algorithms have been proposed. Thus, by combining text compression algorithms with the USM/NCD measure, we can measure the similarity between DNA (or protein) sequences. Li et al. applied a text compression algorithm, which is specialized for DNA sequences, to measuring similarity between genome data [7, 8].

This data compression-based approach was also applied to comparison of protein structures. Krasnogor and Pelta developed such a method using a *contact map* [6]. A contact map is a binary matrix representing the distance between every two residues of a three-dimensional protein structure, in which an element of the matrix is 1 if the distance between the corresponding C_α atoms is less than a predetermined threshold. In their method, each protein structure is first transformed into a contact map and then into a sequence by raster scanning of the contact map. Finally, the standard data compression algorithms are applied using USM. Though a contact map was transformed into sequence data in their method, we developed a method which directly applies image compression algorithms to contact maps [3].

3 Comparison of Metabolic Networks

It is reasonable to try to apply the data compression-based approach to comparison of biological networks because it is computationally hard to compare large-scale graph structures (due to NP-hardness of the subgraph isomorphism problem). However, it is difficult to apply existing graph compression algorithms [9] because it is not guaranteed that the same compression result can always be obtained for identical graphs (i.e., the

compression size may depend on the ordering of vertices). Therefore, we developed a novel graph compression algorithm in which identical graphs are always compressed into the same data [4]. We applied this compression algorithm using USM/NCD to comparison of metabolic networks. The results of computational experiments suggest that the proposed method is useful for comparison of large-scale network data [4].

4 Approximation of Smallest Tree Grammar

Grammar-based compression is one of well-studied methodologies for compression of text data [2]. Grammar-based compression seeks for the minimum size grammar that uniquely generates a given string, where a *context-free grammar* (CFG) is usually employed as a grammar and the size of a grammar is measured by the total number of letters appearing in right hand side of production rules in a grammar. For example, the following string:

acgtacgtacgtacgtacgtacgtacgt

is compressed into the following minimum size context-free grammar:

$$S \to AA,\ \ A \to BB,\ \ B \to CC,\ \ C \to \text{acgt}.$$

Grammar-based compression is useful not only for data compression but also for extraction of patterns in a given string. Since grammar-based compression is NP-hard, approximation algorithms have been developed [2].

It is reasonable to try to extend grammar-based compression for analysis of graph structured data. In particular, it is reasonable to try to infer L-systems from given complex morphological data because L-systems have been extensively studied for modeling developmental processes of various organisms. For that purpose, since an L-system is considered as a kind of graph grammar, we should develop grammar-based compression algorithms for graph grammars. However, to our knowledge, there had been no algorithm with a guaranteed approximation ratio even for grammar-based compression of tree structured data. Therefore, we recently defined an elementary ordered tree grammar (EOTG) by extending CFG for ordered trees and developed a grammar-based compression algorithm (named TREE-BISECTION) for EOTG [1] by extending the BISECTION algorithm for string data [2]. Though the approximation ratio of TREE-BISECTION is $O(n^{5/6})$ and thus is not good, we hope that it stimulates further theoretical and practical studies and leads to discovery of patterns in generation of complex biological systems. We also extended the BISECTION algorithm for grammar-based compression of rectangular image data [5], which may be useful for comparison of protein structures.

References

[1] Akutsu, T.: A bisection algorithm for grammar-based compression of ordered trees, *Information Processing Letters*, 110:815-820, 2010.

[2] Charikar, M., Lehman, E., Liu, D., Panigrahy, R., Prabhakaran, M., Sahai, A., Shelat, A.: The smallest grammar problem, *IEEE Transactions on Information Theory*, 51:2554-2576, 2005.

[3] Hayashida, M., Akutsu, T.: Image compression-based approach to measuring the similarity of protein structures, *Proc. 6th Asia-Pacific Bioinformatics Conference*, 221-230, 2008.

[4] Hayashida, M., Akutsu, T.: Comparing biological networks via graph compression, *BMC Systems Biology* (special issue for OSB 2009), in press.

[5] Hayashida, M., Ruan, P., Akutsu, T.: A quadsection algorithm for grammar-based image compression, *Post Proc. 2nd International Conference on Advanced Science and Technology*, in press.

[6] Krasnogor, N., Pelta, D. A.: Measuring the similarity of protein structures by means of the universal similarity metric, *Bioinformatics*, 20:1015-1021, 2004.

[7] Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P. E., Zhang, H.: An information-based sequence distance and its application to whole mitochondrial genome phylogen, *Bioinformatics*, 17:149-154, 2001.

[8] Li, M., Chen, X., Li, X., Ma, B., Vitányi, P. M. B.: The similarity metric, *IEEE Transactions on Information Theory*, 50:3250-3264, 2004.

[9] Peshkin, L.: Structure induction by lossless graph compression, *Proc. 2007 Data Compression Conference*, 53-62, 2007.

[10] Prusinkiewicz, P., Lindenmayer, A.: *The Algorithmic Beauty of Plants*, Springer-Verlag. 1990.