Application to Gene Cluster Analysis of Inductive Inference of Languages over Patterns with Conceptual Hierarchy

Yukako Tohsato
yukako@ics.es.osaka-u.ac.jp

Hideo Matsuda
matsuda@ics.es.osaka-u.ac.jp

Akihiro Hashimoto
hasimoto@ics.es.osaka-u.ac.jp

Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

1 Introduction

The MINL problem is a problem that finds a minimum and reduced set of patterns explaining a given set of positive example strings. By restricting the number of patterns to be a fixed constant in advance, a polynomial time algorithm that solves this problem is known [1]. This algorithm is applicable to determining gene clusters based on functional classification if genes having the same function are expressed with the same character. However, since gene function is typically classified hierarchically, the above algorithm can only be applied on a single level of the classification hierarchy. In this paper, we extend the MINL problem to cover hierarchical classifications, and propose a novel polynomial time algorithm utilizing entropy to solve the extended problem. The effectiveness of our method is demonstrated by applying the method to a gene cluster analysis on 9 complete genomes.

2 Method

If there is a word which has a half-order relation like Fig. 1 and two sentences like Fig. 2 are given, we can consider finding the pattern using the partially ordered structure of words. When there is a portion common to the given positive examples, it can be represented by a pattern; otherwise by the variable "*". We introduce an evaluation function to express a family of sentences as follows:

$$I(P) = (-\log_2 \frac{k}{n}) \cdot (\sum_{p \in P} \frac{n_s}{n} I(p))$$

$I(p)$ is the sum of information entropies of words and variables composing a pattern p. The entropy $I(w)$ of a word w is the logarithm of the occurrence probability of word w in the positive examples. The entropy $I(*)$ of variable "$*$" is 0. We introduce an extended alignment algorithm (see Fig. 3), that finds a pattern that explains two positive examples. The right matrix in Fig. 3 stores the nearest common ancestor between the words. "b" and "c" are patterns between "b_1" and "b_2", and between "c_1" and "c_2", respectively. The left matrix in Fig. 3 expresses the trace of the optimum alignment between two sequences "ab_1c_1" and "b_1c_2", such as "$*bc$". The recurrence equation is:

$$M_{i,j} = \max \{M_{i-1,j-1} + I(\{w_{1i}, w_{2j}\}), M_{i-1,j}, M_{i,j-1}\}.$$

When more than two sequences are given as the positive examples, we employ a greedy algorithm for aligning those sequences (see Fig. 4).

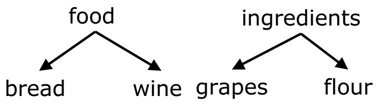

Figure 1: Conceptual hierarchy

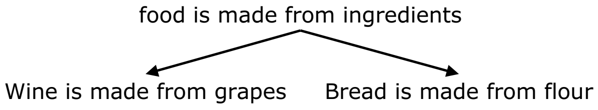

Figure 2: A common pattern is found from two sequences

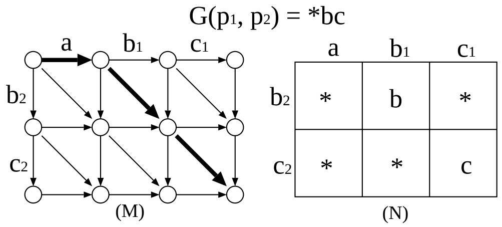

Figure 3: Our extended alignment algorithm

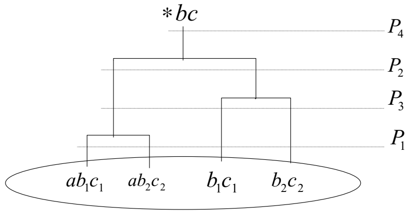

Figure 4: Process of multiple alignments

3 Application to the gene cluster analysis

We apply this algorithm to the gene cluster analysis. In the experiment, we use the data of gene clusters that encode enzymes participating in the tryptophan synthesis pathway in different organisms (see Fig. 5) [2,3]. We consider the EC number expresses a conceptual hierarchy, such as “4.1.1.20” consists of four classes: “4”, “4.1”, “4.1.1”, and “4.1.1.20”, and these classes can be used as patterns in our extended alignment algorithm. The result of this experiment is as shown in Fig. 7. The pattern we obtained is as shown in Fig. 6.

References

[1] Arimura, H. and Shinohara, T. and Otsuki, S., “Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data”, *In* Proc. the 11th STACS, *LNCS 775, Springer-Verlag*, pp.649–660, 1994.

[2] Bono, H. and Goto, S. and Fujibuchi, W. and Ogata, H. and Kanehisa, M., “Systematic Prediction of Orthologous Units of Genes in the Complete Genome”, Genome Informatics 1998, pp.32–40, 1998.

[3] Dandekar, T. and Snel, B. and Huynen, M. and Bork, P., “Conservation of gene order: a fingerprint of proteins that physically interact”, Trends in Biochemical Sciences, vol. 23, 9, pp.324–328, 1998.

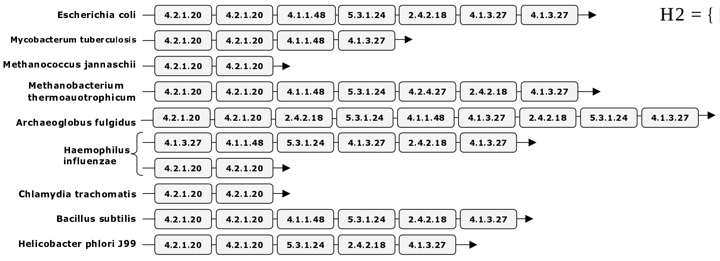
Escherichia coli 4.2.1.20 4.2.1.20 4.1.1.48 5.3.1.24 2.4.2.18 4.1.3.27 4.1.3.27
Mycobacterum tuberculosis 4.2.1.20 4.2.1.20 4.1.1.48 4.1.3.27
Methanococcus jannaschii 4.2.1.20 4.2.1.20
Methanobacterium thermoauotrophicum 4.2.1.20 4.2.1.20 4.1.1.48 5.3.1.24 4.2.4.27 2.4.2.18 4.1.3.27
Archaeoglobus fulgidus 4.2.1.20 4.2.1.20 2.4.2.18 5.3.1.24 4.1.1.48 4.1.3.27 2.4.2.18 5.3.1.24 4.1.3.27
Haemophilus influenzae
4.1.3.27 4.1.1.48 5.3.1.24 4.1.3.27 2.4.2.18 4.1.3.27
4.2.1.20 4.2.1.20
Chlamydia trachomatis 4.2.1.20 4.2.1.20
Bacillus subtilis 4.2.1.20 4.2.1.20 4.1.1.48 5.3.1.24 2.4.2.18 4.1.3.27
Helicobacter phlori J99 4.2.1.20 4.2.1.20 5.3.1.24 2.4.2.18 4.1.3.27

Figure 5: Gene cluster data

H1 = { [4.2.1.20] [4.2.1.20] * ,
* [4] * [4.1.1.48] * [4.1.3.27] [2.4.2.18] * [4.1.3.27] }
H2 = { [4.2.1.20] [4.2.1.20],
* [4.1.1.48] [5.3.1.24] [4.1.3.27] [2.4.2.18] [4.1.3.27],
[4.2.1.20] [4.1.1.20] * [4.1.1.48] * [2.4.2.18] [4.1.3.27] }

Figure 6: Obtained patterns

	H1	H2
Accuracy	4/4	3/4
Entropy	12.47	5.98
Execution time(second)	2.1	1.7

Figure 7: Performance results