Finding Conserved Amino Acid Sequences among Prokaryotic Proteomes

Naohisa Goto ngoto@gen-info.osaka-u.ac.jp
Ken Kurokawa ken@gen-info.osaka-u.ac.jp
Teruo Yasunaga yasunaga@gen-info.osaka-u.ac.jp
Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan

Keywords: complete genome, proteome, conserved sequence, protein synthesis

1 Introduction

Since the first complete genome sequence of *Haemophilus influenzae* determined in 1995, the genomes of more than 80 organisms have been completely sequenced. Complete genome sequence also reveals complete set of protein sequences (proteome) of the organism. At the time when numbers of complete proteomes from a wide variety of species become available, one of the most interesting questions is: what is the most conserved amino acid sequence among species? As highly conserved sequences among many species are expected to play essential roles in living organisms, it is important to answer the question. However, we had many difficulties to answer the above simple question, because of the large number of protein sequences and the long total length of the sequences.

We previously reported that the longest contiguous conserved DNA sequence among 31 prokaryotic complete genomes exists in 16S ribosomal RNA gene [6], by using a software "CONSERV" [4]. CONSERV is a tool for finding contiguous conserved sequences among biological sequences by using suffix tree technique. One of the important features of CONSERV is that it can detect conserved sequences in realistic time with sufficient memory even if given sequences are very long. By using CONSERV, we can search complete proteomes as well as complete genomes to find conserved sequences.

We have improved CONSERV to support proteome analysis, and applied it to find longest contiguous amino acid sequences conserved among prokaryotes.

2 Method

The complete sets of protein sequences were retrieved from NCBI genome database [1]. There were 86 prokaryotic genome sequences (70 bacteria, 16 archaea) which are completely sequenced and annotated. We used all protein sequences of the 86 completely sequenced organisms. There were 235,366 proteins in the 86 organisms, and the total length of these proteins were 72,201,488 amino acids.

We used CONSERV to find contiguous conserved amino acid sequences among all of the 86 organisms. The current version of CONSERV was modified to facilitate analyzing proteomes. Though CONSERV can only detect exact matches, its running speed is very fast and it can efficiently detect conserved sequences among large numbers of proteomes at a time. The high speed and high efficiency are achieved by using suffix tree and Ukkonen's suffix tree construction algorithm [5].

Table 1: Conserved sequences among 86 prokaryotes.

Length	Sequence	Proteins	Positions*
8	GHVDHGKT	EF-Tu, IF-2	19(EF-Tu), 398(IF-2)
8	DTPGHVDF	EF-G, LepA	88(EF-G), 77(LepA)
7	GAGKSTL	ABC transporter	36(BtuD), 43(AraG)**

Sequences shorter than 7 aa are not shown.

Common sequences which are part of longer common sequences are not shown.

∗ Homologous positions on *Escherichia coli* proteins are shown.

∗∗ Only representative proteins are shown.

3 Results and Discussion

Detected conserved amino acid sequence longer than or equal to 7 amino acids among all of the 86 prokaryotes are shown in Table 1. The longest contiguous sequence conserved among the 86 organisms were GHVDHGKT and DTPGHVDF, both were 8-aa in length.

The sequence GHVDHGKT was found on elongation factor Tu (EF-Tu), initiation factor 2, and some other types of GTP-binding proteins. The sequence contains a motif "$GX_4GK(S/T)$", which is known as G-1 region of GTPase superfamily [2].

The sequence DTPGHVDF was found on elongation factor G(EF-G) and LepA. The sequence contains a GTP-binding motif "DXXG" [2].

Elongation factors and Initiation factors are involved in protein synthesis at ribosomes. The function of LepA is unknown, but supposed to regulate ribosome function [3]. The fact that longest conserved sequences are found on proteins involved in protein synthesis suggests that there may be tight constraints at the regions of these conserved sequences. This also suggests that any type of mutations occurred at the regions might be lethal.

It is not well known why the above sequences are extremely conserved at the specific regions of the particular proteins listed above. Further analysis may be required to determine the factors of the highly conserved sequences.

References

[1] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L., GenBank, *Nucleic Acids Res.*, 30:17–20, 2002.

[2] Bourne, H.R., Sanders, D.A., and McCormick, F., The GTPase superfamily: Conserved structure and molecular mechanism, *Nature*, 349:117–127, 1991.

[3] Caldon, C.E., Yoong, P., and March P.E., Evolution of a molecular switch: Universal bacterial GTPases regulate ribosome function, *Mol. Mocrobiol.*, 41:289–297, 2001.

[4] Goto, N., Kurokawa, K., and Yasunaga T., CONSERV: A tool for finding exact matching conserved sequences in biological sequences, *Genome Informatics*, 11:307–308, 2000.

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

[6] Kurokawa, K., Goto, N., and Yasunaga T., Searching the most conserved sequence in bacterial whole genomes, *Genome Informatics*, 11:309–310, 2000.