scienceWATCH.com
TRACKING TRENDS & PERFORMANCE IN BASIC RESEARCH

2009 : October 2009 - Fast Breaking Papers : Kazutaka Katoh on Developing Practical Tools that Support Biological Research

FAST BREAKING PAPERS - 2009

October 2009

📄 PDF

Kazutaka Katoh talks with *ScienceWatch.com* and answers a few questions about this month's Fast Breaking Paper in the field of Computer Science.

SW: Why do you think your paper is highly cited?

Sequence alignment algorithms have been a classical topic in bioinformatics since Saul B. Needleman and Christian D. Wünsch's paper, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J Mol Biol* 48(3):443-53, 1970. A sequence alignment provides a basis of various molecular biological analyses and experiments, such as phylogenetic tree inference, structure prediction, etc.

Recent advancements in biological research and experimental technology have generated quantitatively and/or qualitatively new types of data, which have given a new impetus to studies of alignment algorithms. For example, genome projects and the subsequent development of new-generation sequencers have yielded enormous amounts of sequence data. Such a situation urgently requires an improvement in the scalability of sequence alignment algorithms.

The introduction of residue-to-residue interaction into sequence alignment has been accelerated by the identification of a new "RNA continent" from genomic data. The accumulation of protein structural data, associated with the progress of structural proteomics, has catalyzed the integrated use of sequence and structure information to improve the alignment accuracy of distantly related proteins. As a result, alignment algorithms have become a hot research topic.

SW: Does it describe a new discovery, methodology, or synthesis of knowledge?

This paper describes the latest version of a sequence alignment program, MAFFT. The detailed methodologies and algorithms were described in individual papers (Katoh *et al.*, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.* 30[14]:3059-66, July 15, 2002; Katoh *et*

al., "MAFFT version 5: improvement in accuracy of multiple sequence alignment," *Nucleic Acids Res.* 33[2]:511-8, Jan 20, 2005; Katoh and Toh, "PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences" *Bioinformatics* 23[3]:372-4, FEB 1, 2007; Katoh and Toh, "Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework," *BMC Bioinformatics* 9:212, Apr 25, 2008).

SW: Would you summarize the significance of your paper in layman's terms?

The genes of organisms are regarded as strings consisting of four alphabet letters (nucleotides). Proteins, which are the products of genes, are linear polymers of 20 types of amino acids. Therefore, proteins are again regarded as strings consisting of 20 alphabet letters. Different genes derived from a common ancestral gene are usually similar to each other but not identical. The differences in the strings are generated by the substitutions, insertions, and deletions of "letters" during the course of evolution.

Comparisons of such strings provide various types of information. For example, from a set of similar proteins, we can identify conserved parts, which are interpreted as the important sites that have been conserved during evolution to maintain the protein's functions and/or structure. For such a string comparison, we have to infer the correspondence of the "letters" among the strings based on their common ancestry, while considering substitutions and insertions/deletions. This process is called alignment.

We are developing an alignment tool, MAFFT. Its target is basically protein and DNA sequences, but we are expanding the target into RNA and protein structures, because structural information greatly improves the alignment accuracy.

We have also improved the scalability, such that it can process the large numbers of sequences yielded by recent experimental technologies. Current biological problems require further improvements in the alignment accuracy, speed, and scalability.

SW: How did you become involved in this research, and were there any problems along the way?

I initiated the development of a sequence alignment tool around 1996, when I was a graduate student in Takashi Miyata's lab at Kyoto University. The aim was to conduct a large-scale molecular evolutionary analysis, in which sequence alignments were critically important.

One potential problem was that the lab's and my research interests were mainly in molecular evolution, rather than in computer science. However, I received helpful suggestions from lab members. Kazuharu Misawa and Kei-ichi Kuma raised important ideas on algorithmic points, and the other members gave me suggestions from users' viewpoints.

Such an environment was favorable for the incubation of a computational tool that meets the needs of biologists. Now I am receiving a lot of helpful feedback from around the world, and trying to make the software useful for a broader range of biologists and for informaticians who study related research problems.

SW: Where do you see your research leading in the future?

Our basic intention is to develop practical tools that support actual biological research. As mentioned earlier, demands on alignment algorithms are extensively growing. First, the progress of sequencing technology has resulted in large amounts of sequence data, which require a high-throughput alignment method. Second, structural proteomics projects are now yielding structural information for many proteins, which enables analyses of distantly related proteins. Third, the discovery of non-coding RNAs has shed light on the importance of RNA secondary structures.

We are working on such issues by developing new techniques and by using external structure analysis tools, such as ASH, SCARNA, and LaRA. According to the progress of biological research and technology in the future, we will try to provide basic tools for various comparative analyses of sequences, structures, and other types of data.

SW: Do you foresee any social or political implications for your research?

We do not expect that our research will have a direct effect on social or political problems. However, we hope that our software will contribute to progress in medical and industrial research, as accurate and large-scale alignments are needed in various areas of life science.

Kazutaka Katoh, Ph.D.

"According to the progress of biological research and technology in the future, we will try to provide basic tools for various comparative analyses of sequences, structures and other types of data."

Associate Professor
Division of Bioinformatics
Digital Medicine Initiative
Kyushu University, Fukuoka, Japan
Web

KEYWORDS: SECONDARY STRUCTURE PREDICTION; PROTEIN SEQUENCES; RNA SEQUENCES; STRUCTURAL ALIGNMENT; PHYLOGENETIC TREES; IMPROVED ALGORITHM; PARTITION-FUNCTION; TERTIARY STRUCTURE; NONCODING RNAS; GAP PENALTIES.

PDF

back to top

2009 : October 2009 - Fast Breaking Papers : Kazutaka Katoh on Developing Practical Tools that Support Biological Research