APPLICATION NOTE Genomics & Informatics Vol. 7(3) 178-180, September 2009

COCAW: A Genome-wide Pattern Search System for Designing Microbial Probes

Seunghee Ryu[1], Kiejung Park[2,3,¶], Dohoon Lee[1] and Cheol-Min Kim[4]*

[1]Visual and Biomedical Computing Lab., Pusan National University, Busan 609-735, Korea, [2]PNU SuperComputing Center, Pusan National University, Busan 609-735, Korea, [3]Information Technology Institute, SmallSoft Co. Ltd., Daejeon 305-343, Korea, [4]School of Medicine, Pusan National University, Busan 609-735, Korea

Abstract

A few bioinformatics tools have been used to find out conserved regions as probes. We have developed a system based on a heuristic method with web interfaces to find out conserved regions against microbial genomes. The system runs in real time by using relative entropy in limited narrow regions and detecting similar regions between pair regions with local alignment. The system could be useful to find out conserved regions as genome-wide scale.

Availability: The features of this system is introduced at http://164.125.37.216/Projects/COCAW/

Keywords: genome project, genome-wide search, primer design, probe design

Introduction

Many bioinformatics tools have been used for biomedical applications, especially for genome-based researches. A lot of biomedical researches have focused on finding out genetic characteristics related with diseases. Specifically conserved regions among strains, species or genus can be interpreted as diverse meanings depending on cases. Specific regions among some microbes can be used for diagnostic purpose, especially when using microarray chips.

Homology search tools such as FASTA (Pearson, 1990), BLAST (Altschul, 1998), ClustalW (Thompson, 2002), BLAT (Kent, 2002) are generally used to find similar regions. While they can be used to find out specific regions among a given genomes, additional processing should be done to manipulate the homology search results to extract conserved regions. Since the most popular Primer3 was released, a few programs have been developed for primer design and probe design. Some programs such as PCR Suite (van Baren, 2004) and BatchPrimer3 (You, 2004) provide user-friendly Primer3 interfaces including web-based environment. Some programs have been developed for designing very specific primers (Moro, 2009). iCODEHOP (Boyce, 2009) uses multialignment result from ClustalW. Genomemasker (Andreson, 2006) and UniPrime (Boutros, 2009) treat genome sequences for primer design.

We have developed COCAW (Comparative Observer for Conserved Areas among Whole genomes), which has web interfaces for the input of multiple genomes and the output of graphical display of detected conserved regions.

Overview of the System

Fig.1 shows the general algorithm of COCAW. Common k-mer patterns are detected in genomes as seed patterns for finding out the candidates of conserved pair regions. To select comparison regions instead of com-

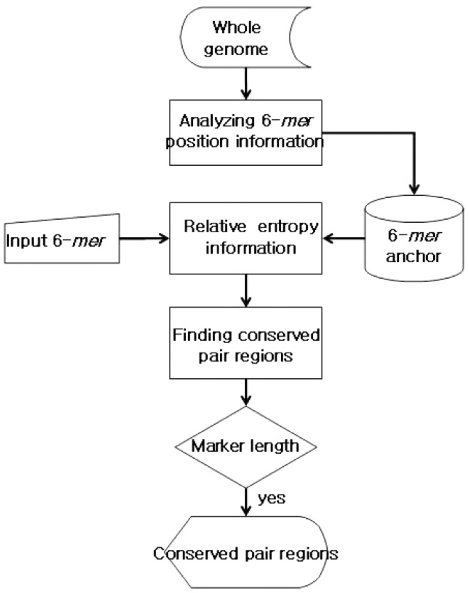

Fig. 1. The general algorithm of COCAW.

¶Co-first author.
*Corresponding author: E-mail kimcm@pusan.ac.kr, Tel +82-51-510-2173, Fax +82-51-510-2174
Accepted 30 August 2009

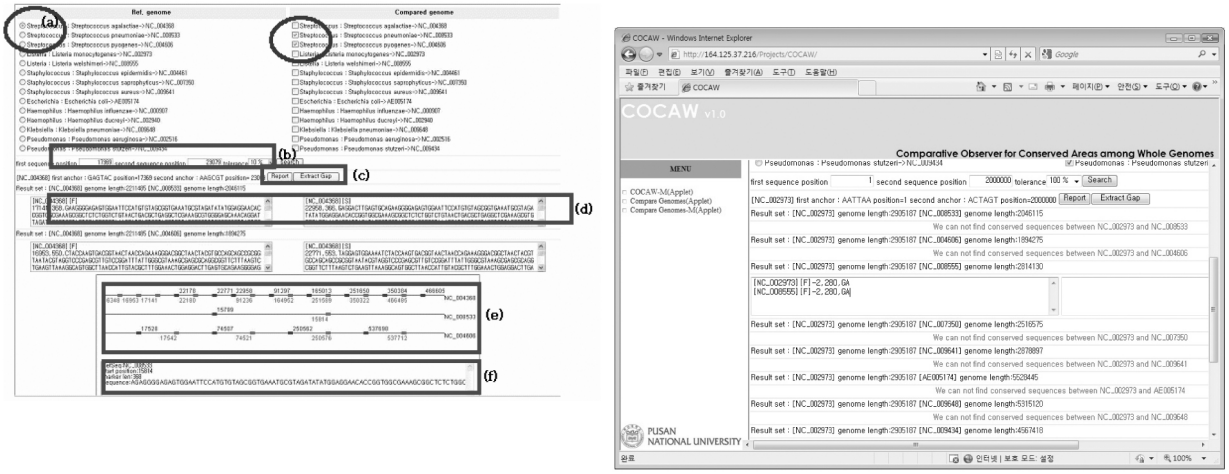

Fig. 2. Main interface of COCAW.

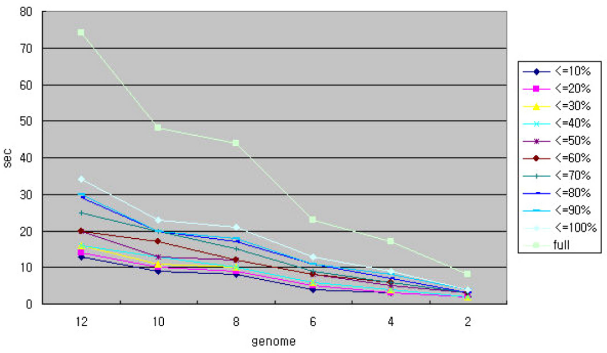

Fig. 3. Running time analysis of COCAW.

paring whole genomes, the information entropy for each k-mer is calculated against whole genomes. After selecting the regions to be compared based on the information entropy, local alignment is calculated to detect final conserved pair regions.

Fig. 2 shows the main web interface of COCAW. Inputs are (a)∼(d) including a reference genome, compared genomes, the comparison range of the reference genome, and tolerance value of the information entropy comparison. And the detected conserved regions are displayed at (e)∼(f). The right shot shows an example of the summary of the searched result, where one genome has a similar region with the reference genome and no similar regions are found between the reference genome and other genomes.

Fig. 3 shows the running time analysis of COCAW. X-axis indicates the number of genomes and Y-axis indicates the running time. And multiple plots for error cutoffs of information entropy are displayed. While the running time increases rapidly by the number of compared genomes, it is still very practical for conserved region analysis of species and genus.

Discussion

COCAW is a tool for scanning multiple genomes for conserved regions and it could be useful for genome-wide detection of the patterns as probes among a group of genomes. Currently it shows the detected conserved regions with a little graphical interface, and it will be upgraded for more practical and informational interface especially for multiple genome processing.

COCAW finds out conserved regions among multiple genomes. As a group-specific pattern is an exclusively conserved region, it is not only conserved in the group but also should not be in other genomes. So for detecting very strict group specific patterns, more complicated and time-consuming processes are additionally required. More practical algorithms should be developed for massive number of genomes with practical user interfaces for diverse applications.

References

Andreson, R., Reppo, E., Kaplinski, L., and Remm, M. (2006). GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinform.* 7, 172.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389-3402.

Boutros, R., Stokes, N., Bekaert, M., and Teeling, E.C. (2009). UniPrime2: a web service providing easier Universal Primer design. *Nucl. Acids Res.* 37, W209-213.

Boyce, R., Chilana, P., and Rose, T.M. (2009). iCODEHOP: a new interactive program for designing COnsensus-

DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. *Nucl. Acids Res*. 37, W222-228.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res*. 12, 656-664.

Moro, C.V., Crouzet, O., Rasconi, S., Thouvenot, A., Coffe, G., Batisson, I., and Bohatier, J. (2009). New design strategy for development of specific primer sets for PCR-based detection of Chlorophyceae and Bacillariophyceae in environmental samples. *Appl. Environ. Microbiol*. 75, 5729-5733.

Pearson, W.R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*. 183, 63-98.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinform. Chapter* 2, Unit 2.3.

van Baren, M.J., and Heutink, P. (2004). The PCR suite. *Bioinformatics*. 20, 591-593.

You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J., and Anderson, O.D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinform*. 9, 253.