S5-3

Development of Bioinformatic Tools for Microbial Phylogeny, Epidemiology and Ecology

Jongsik Chun

School of Biological Sciences, Seoul National University

In this presentation, I will introduce bioinformatic tools for microbial phylogeny, epidemiology and ecology.

JPHYDIT program

Molecular systematics, the major growing field of bioinformatics, has been focused as a key procedure of molecular biology, genomics and molecular phylogenetics itself. With the need of managing the large amount of data efficiently, a large variety of computational approaches and software packages have been developed and those software packages were as powerful as the computation power growing at a high rate. jPHYDIT is a Java-based software package for molecular systematics and has major distinctive features as the followings.

i) 'Integrated Environment for Molecular Systematics'. jPHYDIT was aimed at providing 'Integrated Environment for Molecular systematics' that consists of sequence of sub-tasks ranging from gathering homologous sequences to building phylogenetic tree. ii) 'Database Connectivity. In addition to capability of handling a huge amount of sequence entry data, the database module gave access to multi users so that information could be shared readily. jPHYDIT adopts MySQL as DBMS (database management system) and JDBC package to interface between jPHYDIT and the DBMS. iii) 'Semi-Automated Pairwise Alignment'. Because the molecular sequence data is a prime-determining factor for molecular phylogeny, the importance of pairwise alignment cannot be more emphasized, so most of computational pairwise alignment methods need manual adjusting process to promote the accuracy of the alignment and this process is very laborious. jPHYDIT relieves this laboriousness by using pre-aligned multiple sequences as a 'Template' sequence from the database. This semi-automated pairwise alignment could reduce the number of manual adjusting operation. iv) 'Secondary Structure'. Using 16s ribosomal DNA as a molecular chronometer, 'Secondary Structure' was introduced in alignment process to achieve more reliable phylogeny. Many pairwise alignment algorithms such as Smith-Waterman's and Gotoh's were modified from string compare algorithms

outputting the biologically meaningless alignment. 'Intra strand pairing information' was introduced to make the result of pairwise alignment have the biological meaning. Additionally, intra strand pairing information was used to screen out the sequencing errors.

v) 'Platform Independence'. Being written in JAVA of which outstanding feature is 'Platform Independence', jPHYDIT could be installed and executed on any host operation system such as Linux.
vi) 'Graphical User Interface'. jPHYDIT provides the users with convenient and intuitive graphical user interface. Model-View data structure was hired so that data can be displayed and modified via both Align-Window and Tag-Window. Figure 1)

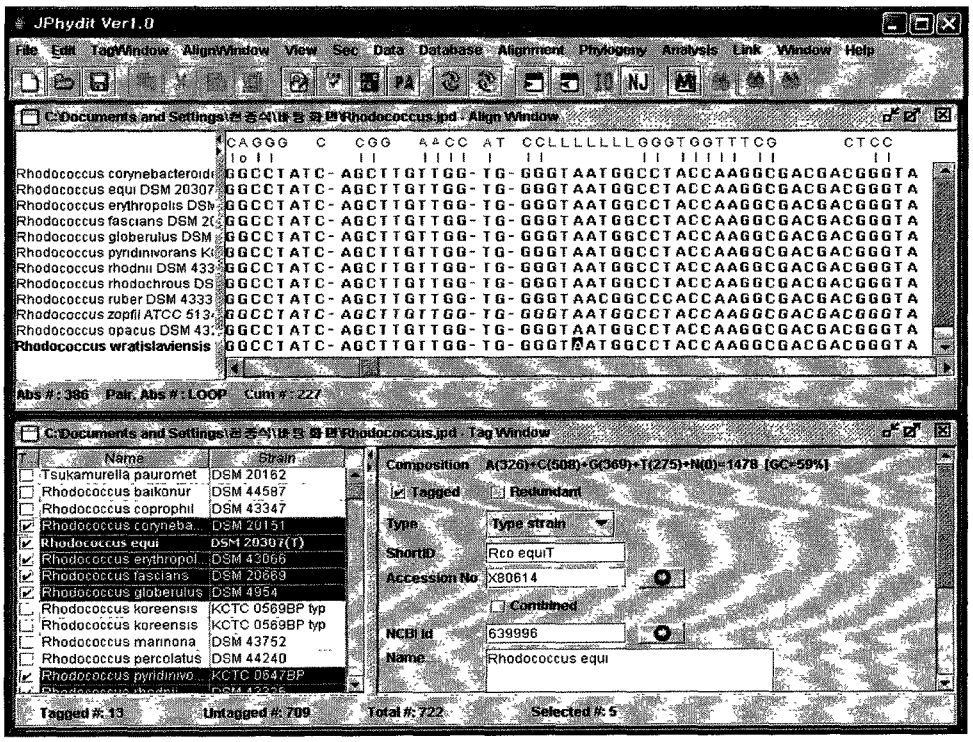

Figure 1. The graphical user interface of jPHYDIT. A sequence (bottom sequence in the upper half. Align-Window) can be imported from Genbank and added to the existing multiple alignment using a semi-automated manner. The nucleotide sequence editor displays intra-strand base paring (top of alignment), which can be used in manual adjustment. L indicates the loop structure. The window for editing general information of a sequence is shown in the lower half (Tag-Window).

As an integrated environment for molecular systematics, jPHYDIT contains essential functional modules such as treeing module, parsing module and editing module etc. and will involve various methods for each module and protein level analysis module in future according to modularity strategy.

MATCHUP program

DNA-based detection needs universal primers and group-specific probe set.

Highly conserved regions in multiple sequences are good candidates of universal primers. Specific primer set also can be used for detection purpose. So, We must search conserved region or specific(unique) region for these works. Finding universal or specific primer region from multiple sequences is not easy work and it takes much laborious time and it has been shown that a good software for that purpose does not exist currentyl from our survey. So, We aimed for developing new tool to design universal primer and internal probes with convenience and exactness for biological researcher and Matchup program was developed.

The major contributions of this program gives us following effects.

First, we tackled the problem of designing universal primers which can amplify multiple products from only one primer pair. All target DNA sequences should be multiply aligned in advance. From those aligned sequences, the most conserved regions should be found and they are used to design universal primer pairs. Groupspecific universal primer or semi-universal primers can be designed to amplify only partial products from all relevant sequences. The procedure to classify all sequences into their groups, finding semi-universal primer, and checking their uniqueness was implemented. This process was developed and named as UPMA (Universal Primer by Multiple Alignment) algorithm.

Second, to overcome the drawback of multiple alignment, new algorithm in which multiple alignment is unnecessary was developed to find universal primer. This method was based on suffix tree and multiple common substrings. From general suffix tree of all target DNA sequences, all multiple common substring could be produced with the number of matching sequnces. These common strings could be extended to left or right direction looking for regions of relatively low degeneracy. If some low-degeneracy regions were found, those regions could be checked for universal or semin-universal primers.

Third, Group or sequence specific primer pairs could be designed from multiple sequences. Their specificity was checked for each sequence and the distictness of their product size were available by manual checks or genetic algorithm based optimization. To minimize the number of primer pairs used for their PCR based separation, specific primers can be searched with its forward or reverse primer fixed. If one primer was fixed, then the other pairing primers were searched with product size constraints satisfied. So, AFLP(Amplification Fragment Length Polymorphis) experiment can be supplied and designed using this program.

Fourth, the design of group specific probe set is possible. Microarray based detection procedure can be performed easily if there exists some unique probe set which hybridize against only their target sequences but does not hybridize non-target sequences. When highly close target sequences are used for experiment, it is more likely that the unique probe set for them cannot be founded. Excluding those sequenes without unique probe set will make the experiment feasible but decoding range will be decreased. In this case, non-unique probes can be alternative choice. All candidates of probes for oligo array were searched. Each candidate was checked for its specificity about how it could hybridize to all

target or non-target sequences. Next, to minimize the number of probe set, optimization was done to exclude the redundancy of probe set. This method was based on integer linear programming.

Fifth, Previous tasks to design oligonucleotides and supplementary works could be executed with graphical user interface and platform dependent framework. It is also possible to execute most of the functions which were supplied at primer3 program since newly developed software incorporated primer3 as its basis. Further, multiple target sequences could be handled more easily for oligonucleotdie design. Design outputs also could be displayed and validated in graphical interface.

Finally, DNA sequences are too large to be searched by traditional algorithms. All target sequences for DNA-based detection should be fully and recurrently scanned for its matching or not matching information. Specially, when whole genome sequences were used for that purpose, large memory usage and high time complexity were essential. Suffix tree data structure was chosen to be used as search engine since it has linear construction and searching time. It was used for designing universal primers, checking the speicificy of prbe set, and various match process so that it is possible to supply feasible solution at the genome-level design of oligonucleotides.

Considering the similarity of target sequences, different methods can be used to design specific primers or probes and the process can be easily executed under graphic user interface and it has been proven Matchup program gives out exact results.