Study on the Measure of Relaxed Sequence Similarity (*Oligostickiness*): Demonstration of Effectiveness in Experimental and Computational Applications

緩和型配列類似性測度「オリゴスティッキネス」の研究―実験的・計算科学的応用への有効性の証明

２００９年 ３月

埼玉大学大学院理工学研究科　(博士後期課程)
理工学専攻（主指導教員　西垣　功一）

SHAMIM AHMED

Summary of the thesis:

Genes have been used for identification and classification of organisms. For this purpose, those of cytochrome C, hemoglobin/myoglobin, ribosomal RNA and others have been adopted and analyzed extensively. However, to identify or compare a species at the gene level, one has to select the gene used for this purpose carefully. Because some genes may be less substituted during the evolution and thus may be insufficient to discriminate species while some may be too much altered in their sequence even at a different evolution speed. Among such genes, ribosomal RNAs (16S/18S rRNA) have been most widely used for this purpose. The superiority of this approach is that it is based on the popular and well-established sequencing technology and can provide the determinate result of nucleotide sequence, which can be further computer-analyzed and can fuel the activity of bioinformatics. Obviously, such an approach may be applicable for classifying a limited number of genome-sequenced species but not for most of the other species with their genomes left not sequenced. Even in future, a huge number of organisms will be left not sequenced due to the logistic reason (too high cost).

In this context a whole-genome-based, non-sequencing method for identification/classification of species was invented and termed as genome profiling [Nishigaki *et. al.*, 2000]. Technologically, genome profiling (GP) is based on a temperature gradient gel electrophoresis (TGGE) analysis of random PCR products [Nishigaki *et al.*, 2000]. For the sake of data refinement, a computer-aided technology was developed by employing species identification dots (spiddos), which correspond to structural transition points of DNAs, and pattern similarity score (*PaSS*) [Naimuddin *et al.*, 2000]. *PaSS* had been shown to be usable for quantitatively measuring the closeness (or distance) between genomes and Thus GP can be used for analyses such as species identification and clustering [Naimuddin *et al.*, 2000].

In this paper, it is demonstrated for the first time that the GP method, when applied to the classification of insects (31 species), can provide a much better correspondence with the established classical taxonomical tree (phenotype-based one) than the conventional 18S rRNA approach does, supporting the common belief that the

whole-genome-based approach must be a better way of classification or clustering than a single gene-based one.

In Genome profiling, there are some steps that can influence determination of the *PaSS* value: for example, random PCR may not select a DNA fragment containing mutations, and the degree of displacement of spiddos caused by a point mutation depends on the type of mutation such as A to G or A to T substitution [Myers *et al.*, 1985, Salimullah *et al.*, 2006]. In this sense, with all such mutations the robustness of GP in clustering of organisms is another interest of this study. Robustness is defined as the degree to which a system works correctly in the presence of fluctuations. Therefore, we measured the "robustness of Genome profiling (GP)", by building an *in-silico* model experiment. The effect of *spiddos-shift*, which is caused by mutations on the score of *PaSS* and the consequent clustering tree was analyzed, expressing the result with a score, altered (0), not altered (1). It was found that, if the experimental fluctuation was confined within less than 3% (usually less than 1% was found experimentally) in PaSS value, then we could get an unchanged clustering result with a more than 99% confidence.

The human genome project and its followers have been producing a huge number of DNA sequence data owing to the accelerating advance in the sequencing technology. Accordingly, the microbiome analysis (metagenome analysis) is becoming more and more popular, resulting in the generation of large quantities of unassigned bacteria and viruses. Up to now, viruses have been assigned to its host by way of microscopic observations, which has been an only possible method for this purpose for a long time.

In this paper, we first introduced a novel method to assign the host-parasite relationship of bacteria and viruses without microscopic observations but with computer analysis of genome sequences. For this aim, we have developed "oligostickiness analysis" which is a kind of oligonucleotide hybridization analysis, i.e., how strongly and where does a particular oligonucleotide (probe) bind to a genome sequence. Based on this analysis, we have further developed a tool to measure the similarity of genome sequences, a parameter SOSS (Set of *Oligostickiness* Similarity Score). This parameter enabled us to assign phages, especially lysogenic ones (i.e. viruses, the genome sequences of which are integrated into the host genome), to its

host bacterium. According to SOSS values, lysogenic phages (such as lambda phage against *E. coli*) have distinctively higher similarity to their hosts than do non-lysogenic (excretive or virulent) ones (such as fd and T4 against *E. coli*). We also investigated the relationship in codon usage frequency and G+C content of genomes between these phages and hosts as a means of interpreting the phenomenon revealed by SOSS analysis. The analysis supported the hypothesis that higher SOSS values result from longer cohabitation in the same environment which must cause common biased mutation. Thus, lysogenic phages which remain in contact with host for a longer period resemble the host and thus being assignable of host-parasite relationship. This finding has great potential for assigning unassigned phages to their hosts, the numbers of which have increased from recent metagenome analyses.

We believe that one of the most important concepts presented in our study is that a relaxed sequence similarity analysis, *oligostickiness*, can extract a large amount of information from genomes as well as the information unavailable by the conventional strict sequence similarity analyses such as repeat sequence analysis. As has already been well-studied, those sequences (genes) which had the same ancestral sequence (gene) are descended with multiple mutations and present as homologs and paralogs. Therefore, the degree of similarity is an important concept in the study of genome sequences. In this sense, Random PCR, a significant step in the GP method, in which various portions of a template DNA can be amplified using one or more primers (probes) by allowing relaxed (mutational or mismatched) forms of primer-binding. Similarly, another method, *oligostickiness*, which is a kind of measure of binding affinity of oligonucleotide, based on the calculation of the free energy (ΔG) of all of the possible hybridization structures formed between the template and the probe (oligonucleotide) at each position along a genome sequence, which allows the counting up all of the possible structures including a lot of mismatch-containing hybridization as long as they have a certain level of stability in terms of ΔG. Therefore, the *oligostickiness* or genome profiling (GP) analysis is statistical and found robust against mutations. This nature is beneficial for identifying and classifying highly diverged (i.e., relaxed) genome sequences. This is why we call *oligostickiness* analysis a measure of relaxed sequence similarity.