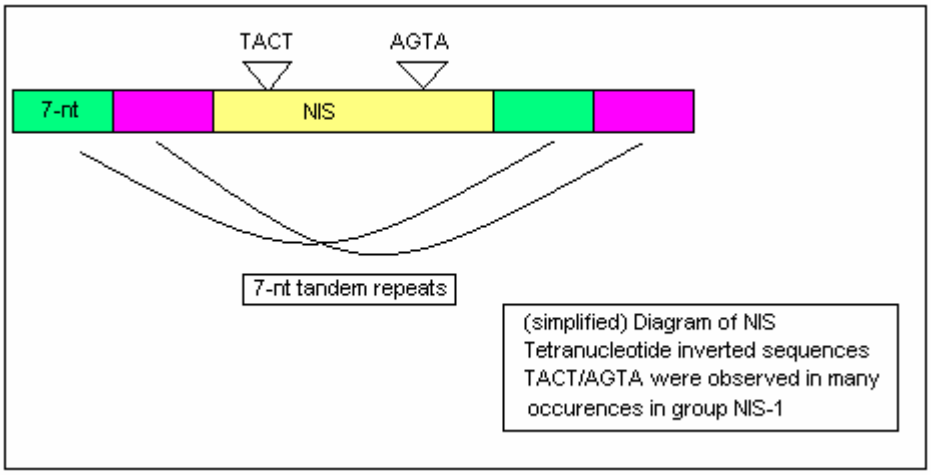
BBSI Research Proposal – Summer 2005
Michiko Kato

Computational Survey of NIS (Nostoc Iterated Sequence) in Cyanobacterial Genomes

Introduction

Both eukaryotic and prokaryotic genomes are known to have repetitive DNA sequences. However, it was thought that repeats were rare in prokaryotic genomes due to their compact genomes and minimal non-coding regions[1]. The advancement in computational genome analysis has permitted a more thorough examination of DNA sequences. As a result, it is now clear that repeat sequences are widespread in prokaryotic genomes. Short interspersed repeats are widely distributed within and among different bacterial species. One such sequence, REP (Repetitive Extragenic Palindrome), was first identified in *E.coli*. REP sequence is characterized as a 38-nt palindromic sequence capable of forming stem-loop or cruciform [2, 3]. Another dispersed sequence, ERIC (Enterobacterial Repetitive Intergenic Consensus), is preferentially located in non-coding transcribed regions and consists of conserved inverted repeats[3]. The ubiquitous presence of these patterned sequences suggests a function that selects for their persistence and/or a rapid means of propagation. Several functional roles of these dispersed sequences have been hypothesized but no role is fully understood.

NIS (Nostoc Iterated Sequence) is a novel type of dispersed repeats found in *Nostoc* genome. During the examination of heptameric tandem repeats in $tRNA^{leu}$ introns by Costa el al., it was discovered that a few of the strains contain a short, non-repetitive sequence within the repeat regions[4]. Currently, NIS is known to be 24-, 42-, 45-, or 48-nt sequence and some have tetranucleotide inverted repeats within the sequence. Six different sequence patterns have been categorized, and the distribution of each type in the genome varies.



So what is the significance of the presence of NIS? The existence of these sequences throughout *Nostoc* genome does not seem to be coincidental. The questions we

hope to be able to answer are their origins, how they spread themselves in the genome, how common they are in prokaryotic genomes, and what their functions are. To start this long quest, the first thing to do is to characterize all of the NIS sequences and compare them in other related cyanobacterial genomes. *Nostoc punctiforme* ATCC29133 has phylogenetic relationship with *Anabaena* PCC 7120 and its close relative *Anabaena variabilis* ATCC29413. Comparative examination of these genomes would provide clues to the evolutionary origin of NIS. My research entails the computational survey of the NIS sequences in these three cyanobacterial genomes. The trend in the occurrence and distribution of all NIS groups will be examined.

Materials and Methods

BioLingua will be the primary tool for this research[5]. It is a web-based bioinformatics tool that facilitates programming and integrated analysis of biological data. Unlike traditional BLAST (Basic Local Alignment Search Tool), BioLingua allows searches on the basis of sequence identity, near identity, and pattern. This allows more thorough search of the interesting sequences. The returned results will be examined by hand to see the patterns and significance.

Possible Results and Implications

By the end of the program, I will be able to complete and compare all NIS sequences in the three cyanobacterial genomes. I hope to run a search of NIS in other bacterial genomes such as *E.coli* if the time allows. The outcome of this project should demonstrate the relationships of each NIS group in these genomes, and whether NIS has ancient origin. Also, the results should establish testable hypothesis to investigate their transfer mechanisms and potential functions.

Reference

1. Rocha E.P.C, Danchin A, and Viari A (1999). Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiology. 150: 725-733
2. Stern MJ, Ferro-Luzzi Ames G et al. (1984). Repetitive Extragenic Palindromic Sequences: A major component of the bacterial genomes. Cell. 37: 1015-1026
3. Lupski JR and Weinstock GM (1992). Short, interspersed repetitive DNA sequences in prokaryotic genomes. Journal of bacteriology. 174: 4525-4529.
4. Costa JL, Paulsrud P, and Lindblad P (2002). The cyanobacterial $tRNA^{leu}$ (UAA) intron: Evolutionary paterns in a genetic marker. Mol. Biol. Evol. 19: 850-857.
5. Massar JP, Travers M, Elhai J, and Shrager J (2005). BioLingua: a programmable knowledge environment for biologists. Bioinformatics. 21: 199-207.