A Novel Method for RNA Sequence Data Analysis

Taishin KIN
Computational Biology Research Center
e-mail: kin-taishin@aist.go.jp

AIST Today Vol. 3, No. 4 (2003) 11

We proposed a novel method to deliver kernels for RNA sequence data using stochastic context free grammar (SCFG)[1]. Our previous work was to deliver kernels for general biological sequences using hidden Markov model (HMM)[2]. RNA sequences can not be dealt with HMM because they involve remote base interactions which consequently form stem-loop structures. The stem-loop structure thermally stabilizes secondary structures of RNA, which is essential in terms of evolutionary conservation. SCFG is more powerful stochastic language model than HMM which allows dealing with the stem-loop structures (Fig1). We call our novel kernel *Marginalized Kernel over SCFG*. The kernel shows good performances in several demonstrations. Fig2 shows a result of kernel PCA for three-class human tRNAs.

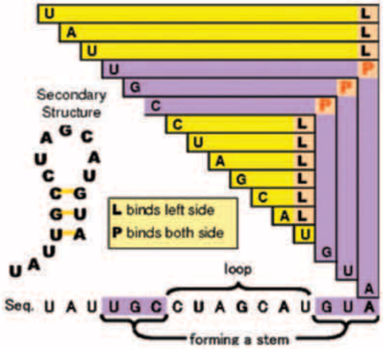

Fig.1 Binding structural information labels to an RNA sequence

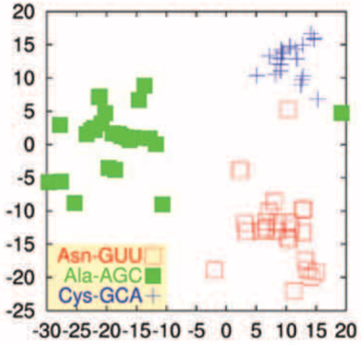

Fig.2 Kernel PCA for three-class human transfer RNA