Title	生物配列解析のためのカーネル設計
Author(s)	金，大真
Citation	
Issue Date	2003-12
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/805
Rights	
Description	Supervisor:佐藤　賢二，知識科学研究科，博士

Designing Kernels for Biological Sequence Data Analysis

Taishin KIN
School of Knowledge Science,
Japan Advanced Institute of Science and Technology

Dec., 2003

Abstract

In order to uncover the nature of life, it is very important to understand the nature of biological sequences such as DNAs, RNAs or proteins. Our research focuses on a fundamental issue of modeling and comparison of biological sequence data in general. The objectives of this research are to propose a general framework where modeling and comparison of biological sequence data can be organized and to develop efficient methods to extract features of biological sequence data.

Firstly, we proposed *the Self-Identification Learning* (SIL) for a system based on hidden Markov models to predict protein coding regions. SIL is a learning algorithm that does not require training data. By making use of its prediction results for its training data in the next iteration, it trains itself through iterative feedback loop of learning and prediction. Whereas existing genefinding systems are not useful when there are insufficient training data of a target organism because a high quality training dataset is indispensable to perform accurate predictions, SIL allows to perform genefinding in such a case.

Many of genefinding systems uses a discriminative property known as *dicodon usage measure* (DUM), one of the best measures that distinguishes protein coding regions and non-coding regions. It has been widely believed that the biological meanings of DUM can be decomposed into several biological properties while such belief is not backed by any objective examination. We found that a portion of dicodon usage parameters suffices to yield reasonable performance for prediction of coding regions. Hinted by this indication of redundancy of DUM, we devised six dicodon approximators based on some combinations of biological properties because a good dicodon approximator will lead to understanding biological meanings of dicodon and to a good basement for designing a better prediction measure. We carried out performance comparisons among DUM and the six approximators by using 17 microbial plus 6 eukaryotic genomic sequence data. However, no approximators could match performance of DUM. Thus dicodon usage cannot be interpreted with the approximators we devised. This result revokes conventional belief on the biological meanings of DUM.

Use of DUM is limited to discriminating coding and non-coding regions. Therefore we started to find a general method to extract features of sequences by thinking two essential issues of biological sequence data analysis i.e. "what is the feature of biological sequence?" and "how we can utilize such features for analysis?" We proposed a novel method to extract features of biological sequence data, *the marginalized kernel*, that is a general framework to design similarity measures among biological sequences. There are two properties dealt with this framework: feature representation and similarity quantification. We use latent variable models (e.g. hidden Markov models) for the feature representations, which allows to bind a hidden variable to a certain biological feature. The hidden variables can be estimated with regular algorithms. The highlight of our method is that we use all probable estimations which allow us to incorporate implicit feature representations. We use kernel functions for similarity quantification so that we can exploit kernel methods such as support vector machines for discriminant analysis. In order to evaluate validity of our method, we performed computational experiments to classify *gyrB* protein sequence data. The experiments show that our method successfully classified most of the proteins.

We developed a novel method: *the marginalized kernel for RNA sequence data*, which defines similarities between RNA sequences by utilizing stochastic context free grammar (SCFG). With our method, powerful multivariate analysis tools such as support vector machines, kernel PCA and etc. become available to RNA sequence data analysis. We demonstrated performance of our method with clustering experiments by using kernel PCA and supervised classification experiments by using support vector machines. The experiments show promising results.

Key Words: biological sequence, di-codon measure, kernel method, marginalized kernel, SCFG