# Marginalized kernels for biological sequences

Koji Tsuda, Taishin Kin and Kiyoshi Asai

AIST, 2-41-6 Aomi Koto-ku, Tokyo, Japan

Presented by Shihai Zhao

May 8, 2014

# Overview

1. Introduction
    - kernel functions
    - hidden Markov model
2. Methods
    - new kernel
    - connections to the Fisher kernel
3. Results and Conclusion

# kernel functions

In kernel methods such as S.V.M., a kernel function should be determined a priori.

## Supervised learning

Objective function is clear. Kernels are designed to optimize the function.

## Unsupervised learning

The choice of kernel is subjective. It is determined to reflect the user's notion of similarity.

# kernel functions for sequences

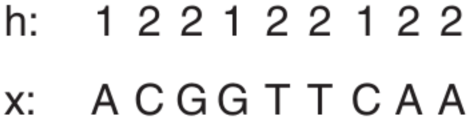
## Texts

Count features, which represent the number of each symbol contained in a sequence

## Biological sequences

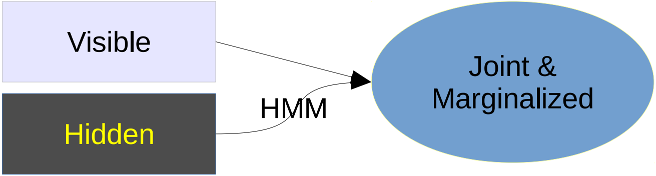
Count does not work 'out of the box' primary due to frequent context change

A DNA sequence with hidden context information. Suppose the hidden variable ('h') indicates coding/noncoding regions.



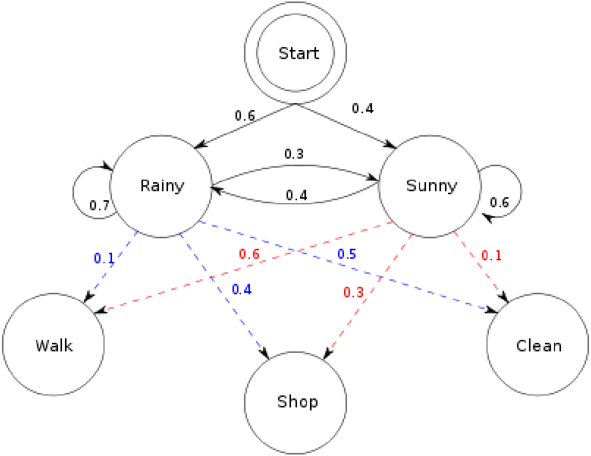
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| h: | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| x: | A | C | G | G | T | T | C | A | A |

# New way to design a kernel



# HMM

A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states.

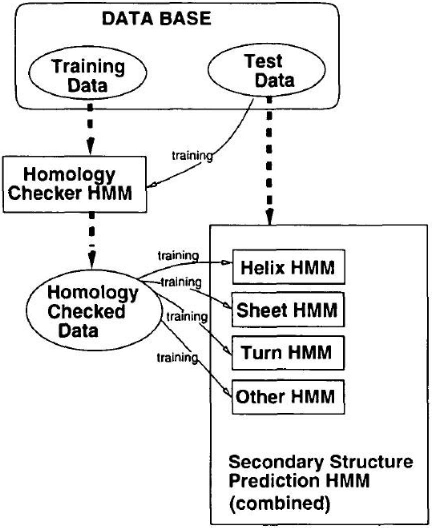


# Example of HMM

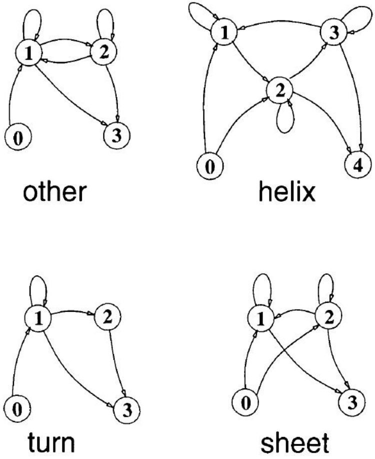
A limited number of sequences whose structures are known. We want to train the four HMMs of secondary structures to make the prediction

- Helix
- Sheet
- Turn
- Other

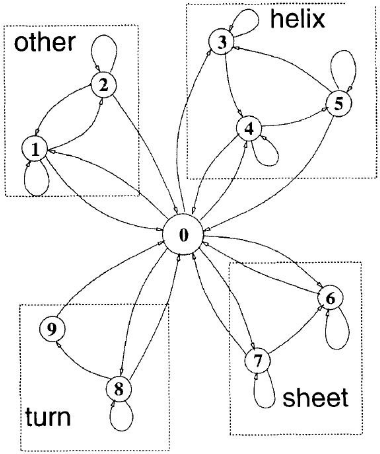
# block diagram



HMMs of secondary structures



Combined HMM for prediction



# marginalized kernel

$x, x \in X, h, h' \in H$, where $H$ is a finite set. $z = (x, h), z' = (x', h')$

$$K(x, x') = \sum_{h \in H} \sum_{h' \in H} p(h|x)p(h'|x')K_z(z, z')$$

$p(x|x)$ has to be estimated from the data. When the cardinality of $H$ is too large, the calculation can be intractable.

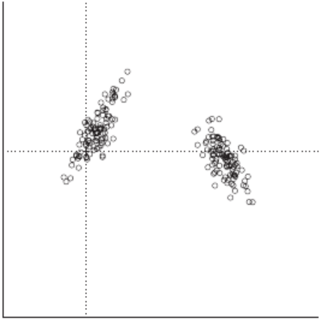
# marginalized kernel from Gaussian mixture

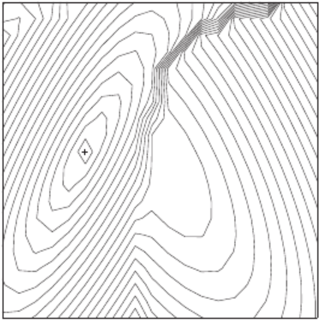
$$K(x, x') = \sum_{h \in H} p(h|x)p(h'|x')x^T A_h x'$$

where $A_h$ is the inverse of covariance matrix.

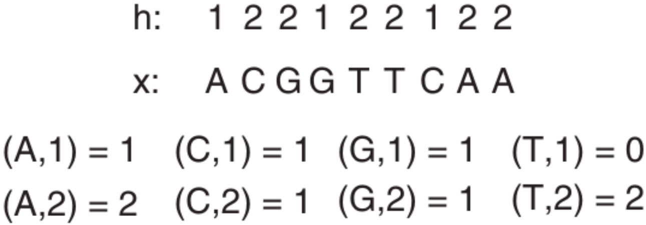
Distance in feature space

$$D(x, x') = \sqrt{K(x, x) + K(x', x') - 2K(x, x')}$$





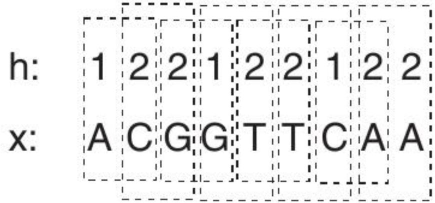
# marginalized count kernel



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| h: | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 |
| x: | A | C | G | G | T | T | C | A | A |

| | | | |
|---|---|---|---|
| (A,1) = 1 | (C,1) = 1 | (G,1) = 1 | (T,1) = 0 |
| (A,2) = 2 | (C,2) = 1 | (G,2) = 1 | (T,2) = 2 |

# second-order marginalized count kernel



# Definition of the Fisher kernel

Assume a probabilistic model $p(x|\theta)$ is defined on $X$, where $\theta$ is a parameter vector. Let $\hat{\theta}$ denote parameter values which are obtained by some learning algorithm. Then the Fisher kernel between two objects is defined as

$$K_f(x, x') = s(x, \hat{\theta})^T Z^{-1}(\hat{\theta}) s(x', \hat{\theta})$$

where $s$ is the Fisher score

$$s(x, \hat{\theta}) := \nabla_\theta \log p(x|\hat{\theta})$$

and $Z$ is the Fisher information matrix

$$Z(\hat{\theta}) = \sum_{x \in X} p(x|\hat{\theta}) s(x, \hat{\theta}) s(x, \hat{\theta})^T$$

# Fisher kernel from latent variable models

the Fisher score is described as

$$\begin{aligned}\nabla_\theta \log p(x|\hat{\theta}) &= \frac{\sum_{h\in H} \nabla_\theta p(x,h|\hat{\theta})}{p(x|\hat{\theta})} \\ &= \sum_{h\in H} \frac{p(x,h|\hat{\theta})}{p(x|\hat{\theta})} \frac{\nabla_\theta p(x,h|\hat{\theta})}{p(x,h|\hat{\theta})} \\ &= \sum_{h\in H} p(h|x,\hat{\theta}) \nabla_\theta p(x,h|\hat{\theta})\end{aligned}$$

The Fisher kernel is described as a marginalized kernel

$$\begin{aligned} K_f(x,x') &= \nabla_\theta p(x|\hat{\theta})^T Z(\hat{\theta})^{-1} \nabla_\theta p(x'|\hat{\theta}) \\ &= \sum_{h\in H} \sum_{h'\in H} p(h|x,\hat{\theta}) p(h'|x',\hat{\theta}) K_z(z,z') \end{aligned}$$

where the joint kernel is $K_z(z,z') = \nabla_\theta p(x,h|\hat{\theta})^T Z(\hat{\theta})^{-1} \nabla_\theta p(x',h'|\hat{\theta})$

# experiment settings

84 amino acid sequences from 5 genera in Actinobacteria

The number of sequences in each genus is listed as 9,32,15,14,14

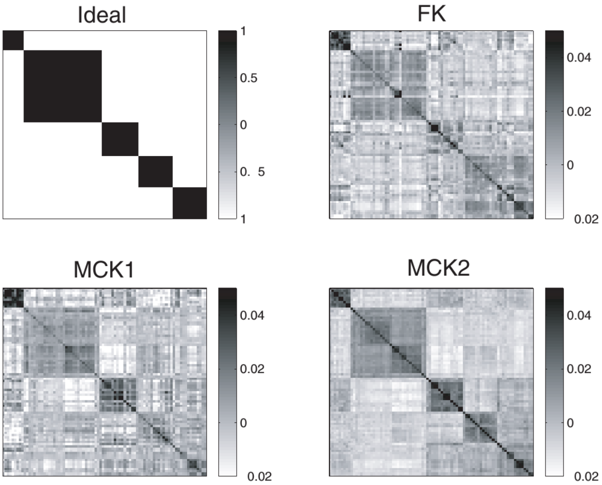
Pairwise identity is 62%-99%

BLAST scores cannot directly be converted to kernels

Two kinds of experiments–clustering and supervised classification are performed on the following kernels:

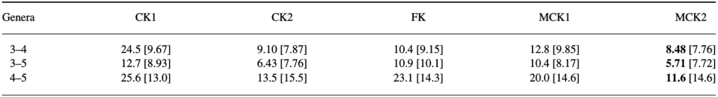
- CK1: Count kernel
- CK2: Second-order count kernel
- FK: Fisher kernel
- MCK1: Marginalized count kernel
- MCK2: Second-order marginalized count kernel

# clustering result



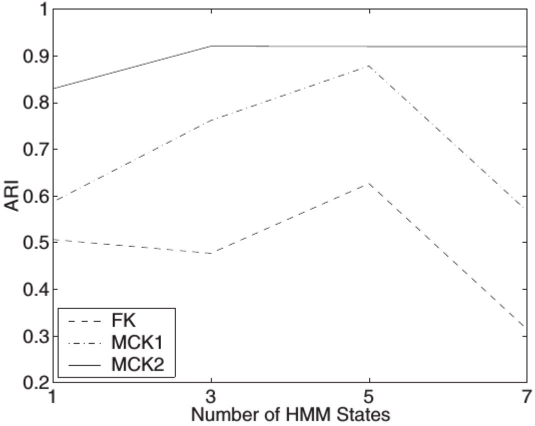
# classification result

Genera 1 and 2 are not used because they can be seperated easily by all kernels. We do one vs one for the rest three.



| Genera | CK1 | CK2 | FK | MCK1 | MCK2 |
|---|---|---|---|---|---|
| 3–4 | 24.5 [9.67] | 9.10 [7.87] | 10.4 [9.15] | 12.8 [9.85] | **8.48** [7.76] |
| 3–5 | 12.7 [8.93] | 6.43 [7.76] | 10.9 [10.1] | 10.4 [8.17] | **5.71** [7.72] |
| 4–5 | 25.6 [13.0] | 13.5 [15.5] | 23.1 [14.3] | 20.0 [14.6] | **11.6** [14.6] |

# effect of HMM states



# conclusion

- Fisher kernel is a special case of MCK.
- second-order kernels perform better than first-order kernels
- number of HMM states' effect