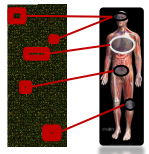
GaKCo

A Fast Gapped k-mer String Kernel using Counting

Ritambhara Singh, Arshdeep Sekhon, Kamran Kowsari, Jack Lanchantin,
Beilun Wang, and Yanjun Qi

Challenge of data explosion

Molecular signatures of tumor / blood sample

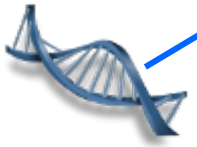


Signs & Symptoms

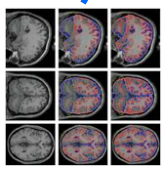
Patient Medical History & Demographics

Public Health Data

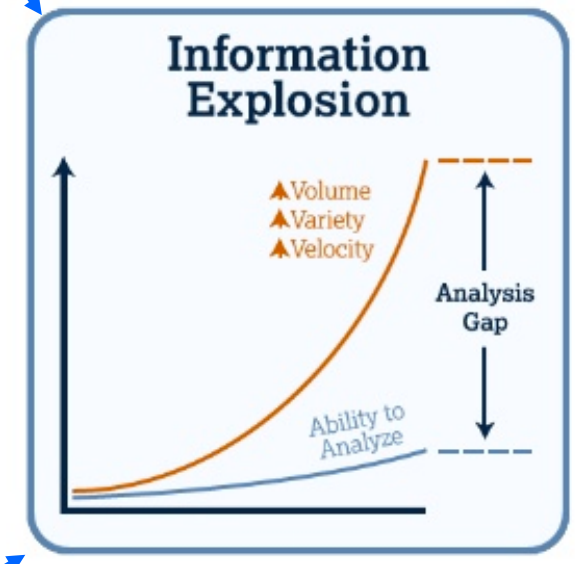
Genetic Data



Mobile medical sensor data



Medical Images



Traditional Approaches

Data-Driven App

Machine Learning

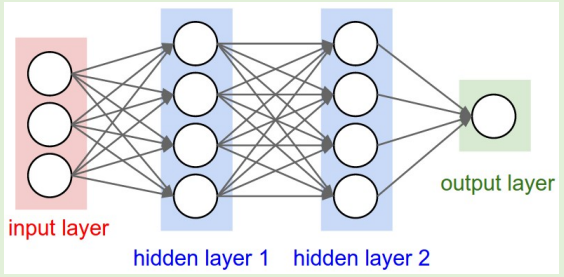
String Classification

X

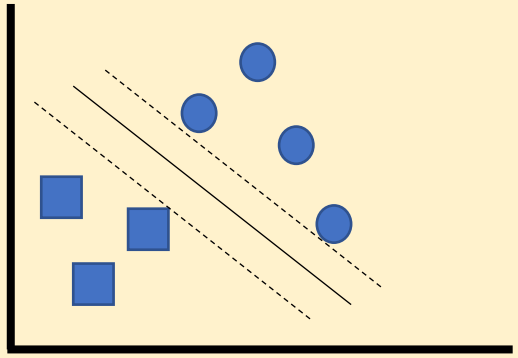
This Food is not good.

NO
(-1)

State-of-the-art Machine Learning Models

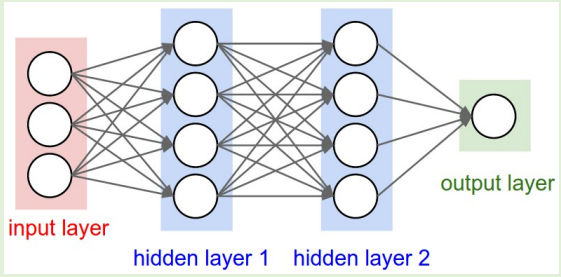


Deep Neural Network (DNN)

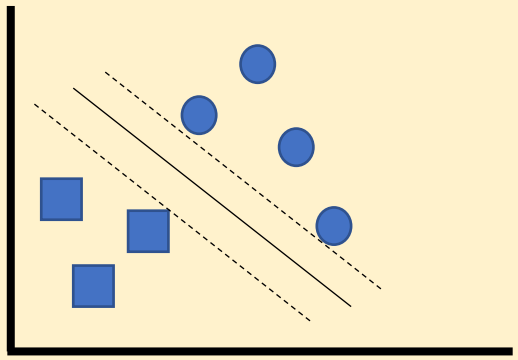


Support Vector Machine (SVM)

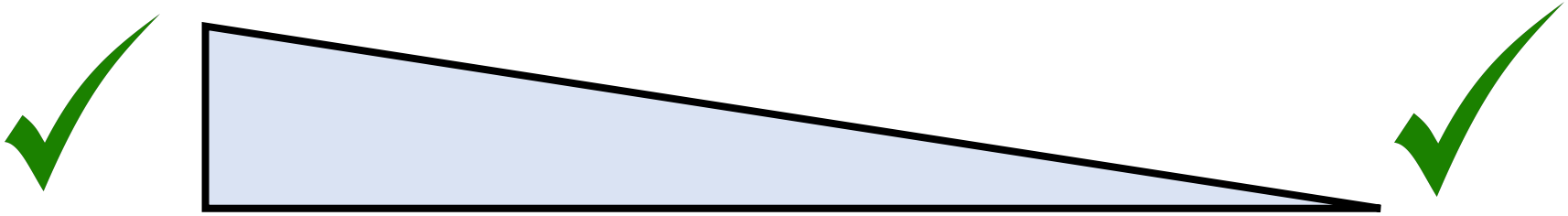
State-of-the-art Machine Learning Models



Deep Neural Network (DNN)



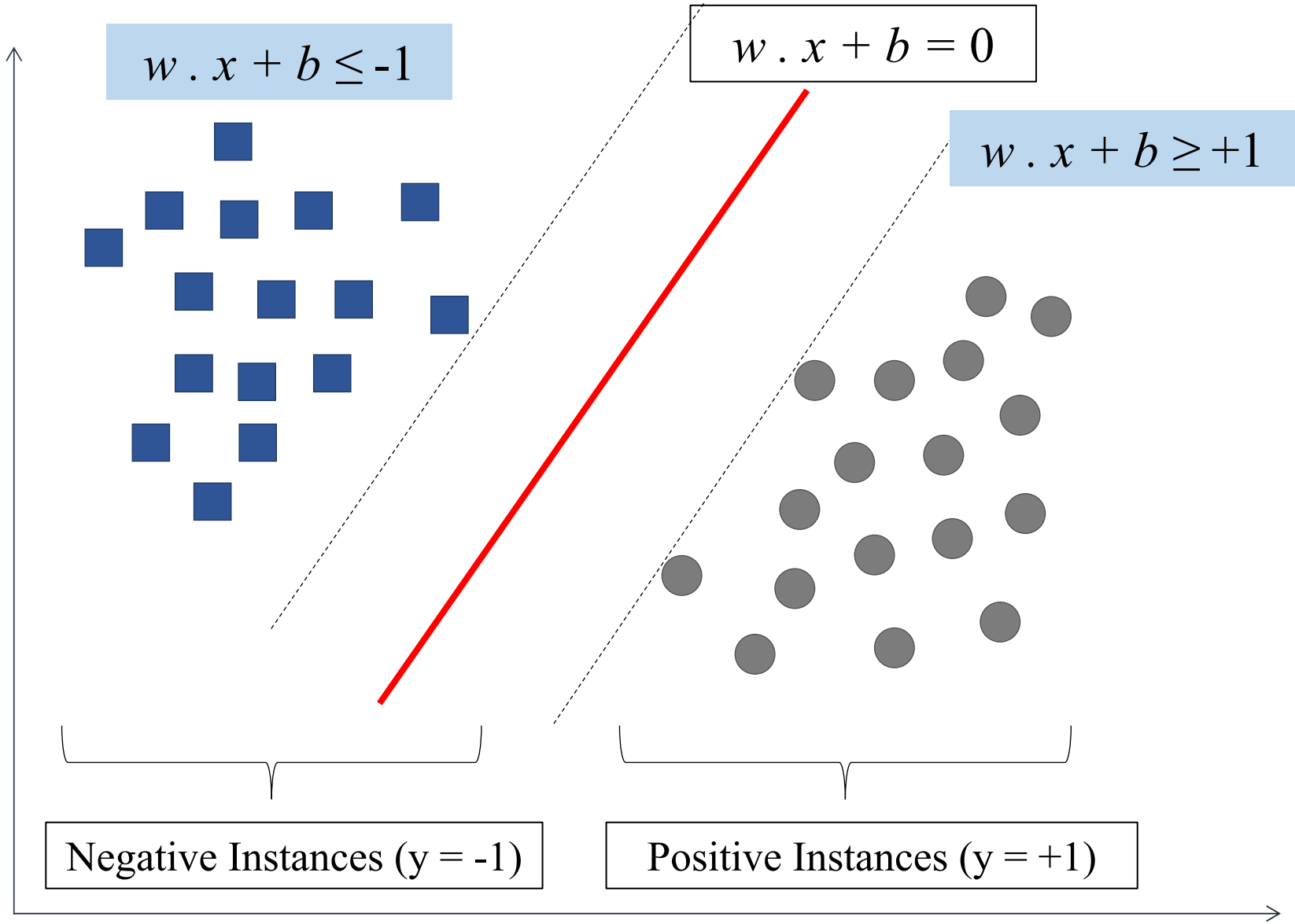
Support Vector Machine (SVM)



Number of Training Samples

Support Vector Machine (SVM)

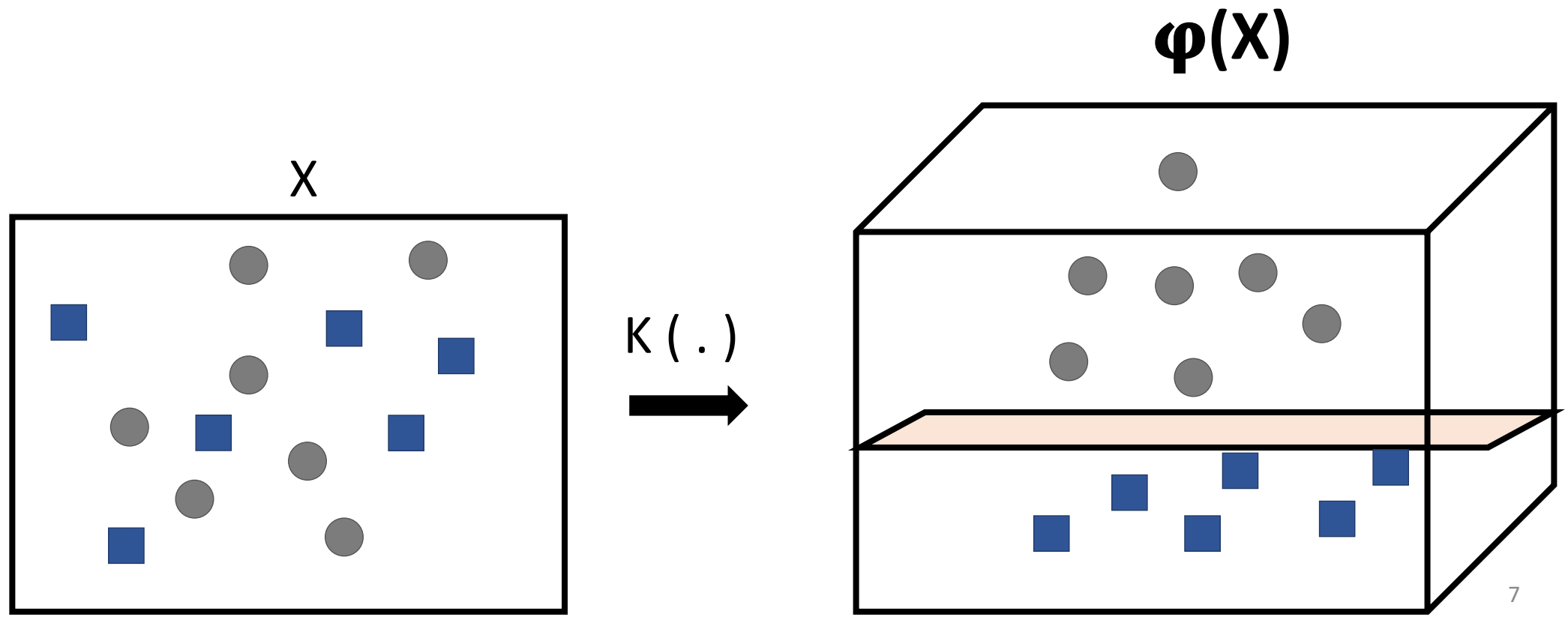
$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$



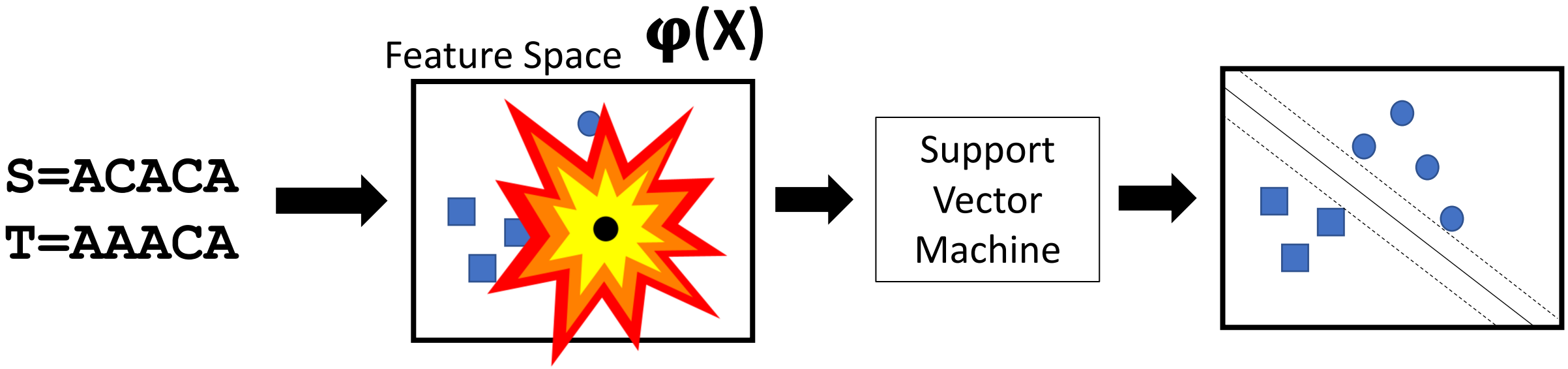
Kernel

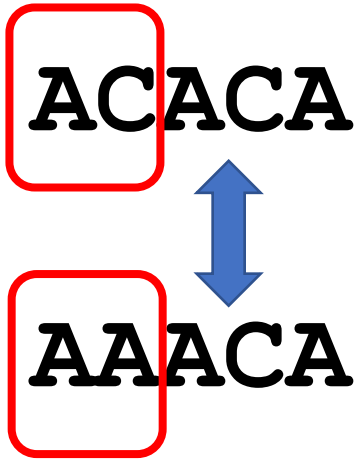
X

$$f(x) = w \cdot \Phi(x) + b.$$



Challenge : SK-SVM methods are slow

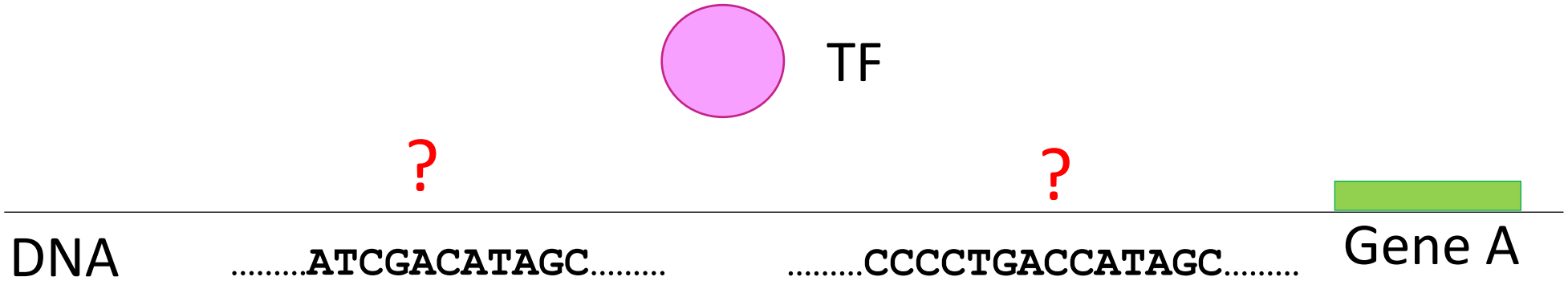




Solution: GaKCo

- Faster than state-of-the-art string kernel
- Independent of dictionary size
- Parallelizable

Does TF bind to this sequence?



ATCGAATCCG ✓

CGCTGAATCG ✗

ATCGCTATCG ✓

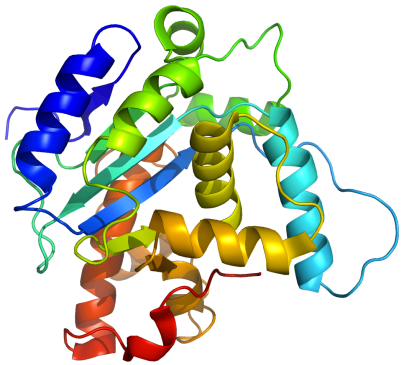
ATCCCGCTCG ✗

$\sum=4$

Minimum Number of Training Samples

1000

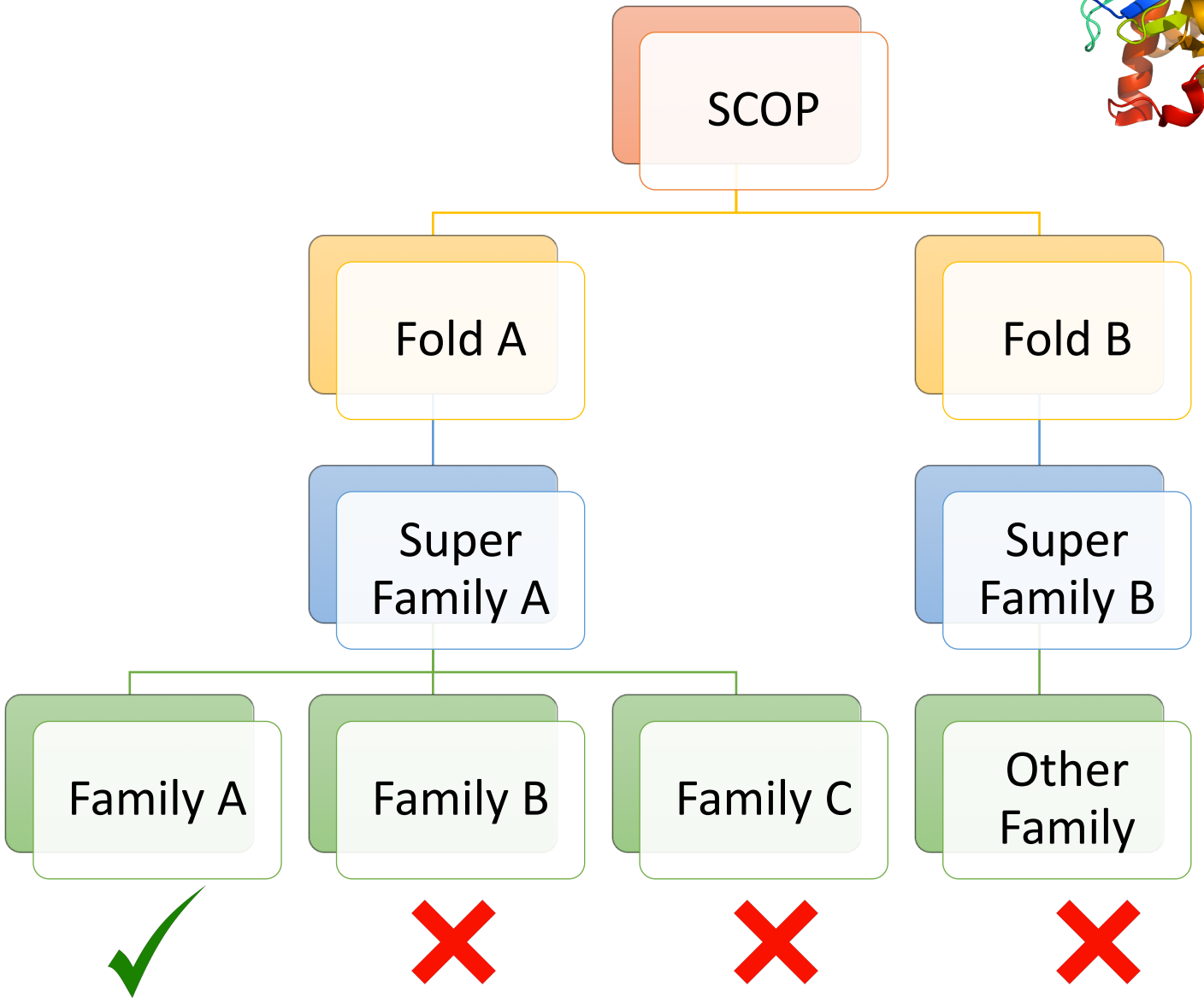
What family does the protein belong to?



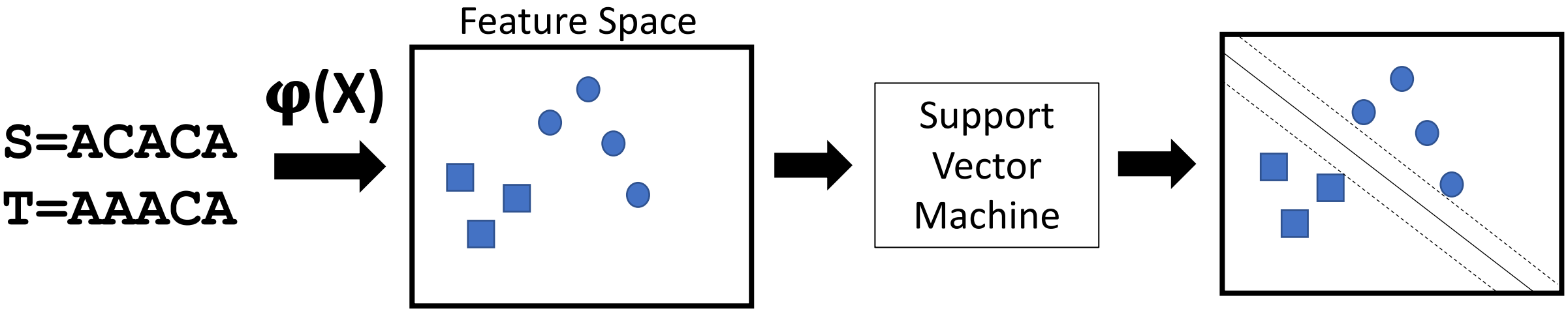
VQGGHACCAKKQQQ

∑=20

Minimum Number of Training Samples
500



String Kernel + SVM Framework



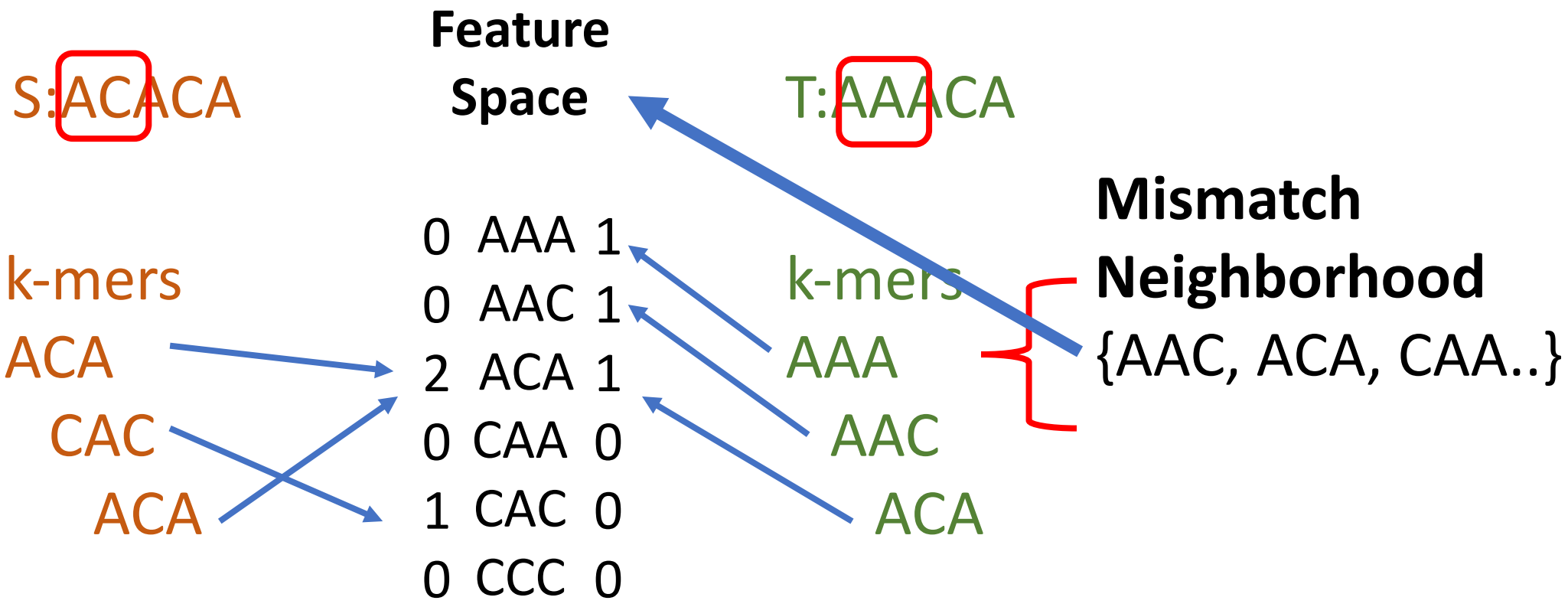
String Kernel

Mismatch Kernel

∑=2 {A,C}

k=3

m=1



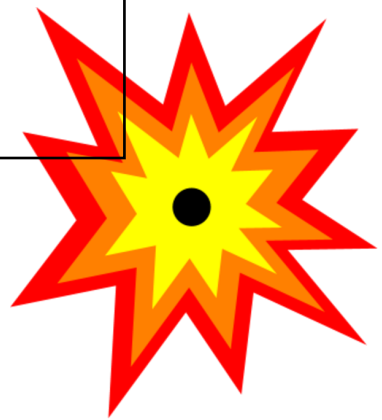
Bag of k-mers

Computational Challenge

∑=20

k=10

Feature Space $\sum^k$



$(20)^{10} = 10^{13}$

Sparse!

AAA	1
AAC	1
ACA	0
CAA	0
CAC	0
CCC	0

Related Work: Gapped k-mer Kernel

∑=2 {A,C}

g=3

k=2

S:ACACA

g-mers	k-mers
ACA	{AC,CA,…}
CAC	
ACA	

T:AAACA

g-mers	k-mers
AAA	{AA,AA,…}
AAC	
ACA	

Gaps=g-k=1

Feature Space $g(|S| + |T|) \ll \sum^k$

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. "Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features".

Related Work: Gapped k-mer Kernel

(1) $$K(x, x') = \sum_{\gamma \in \Theta_g} c_x(\gamma) \cdot c_{x'}(\gamma)$$

(2) $$K(x, x') = \sum_{i=0}^{l_1} \sum_{j=0}^{l_2} h_{gk}(g_i^x, g_j^{x'})$$

(3) $$K(x, x') = \sum_{m=0}^{g-k} N_m(x, x') h_m$$

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. "Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features".

Related Work: Gapped k-mer Kernel

gkm-SVM

∑=2 {A,C}

Dependence on term $(\sum)^m$

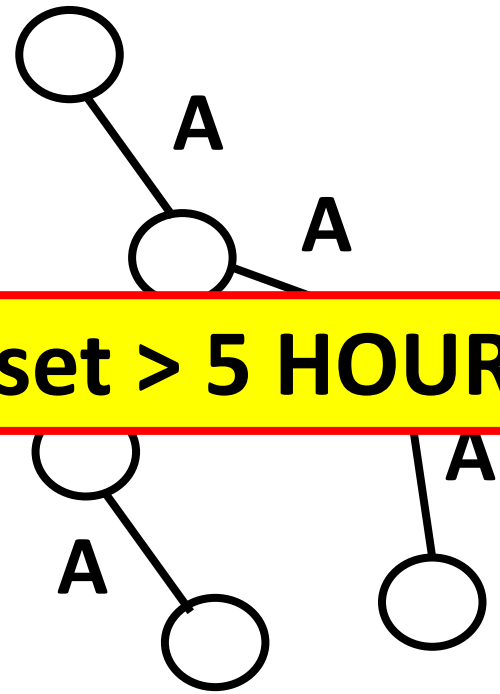
g=3

S:ACACA

T:AAACA

Time taken to compute Kernel for Protein dataset > 5 HOURS

Trie



ACA

CAC

ACA

Mismatch Profile

Coefficient

AAA

AAC

ACA

$$K(S,T) = \sum_{m=0}^{g-k} h_m N_m$$

m=0 $N_m = 2$

m=1 $N_m = 3$

Gaps=g-k=1

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. "Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features".

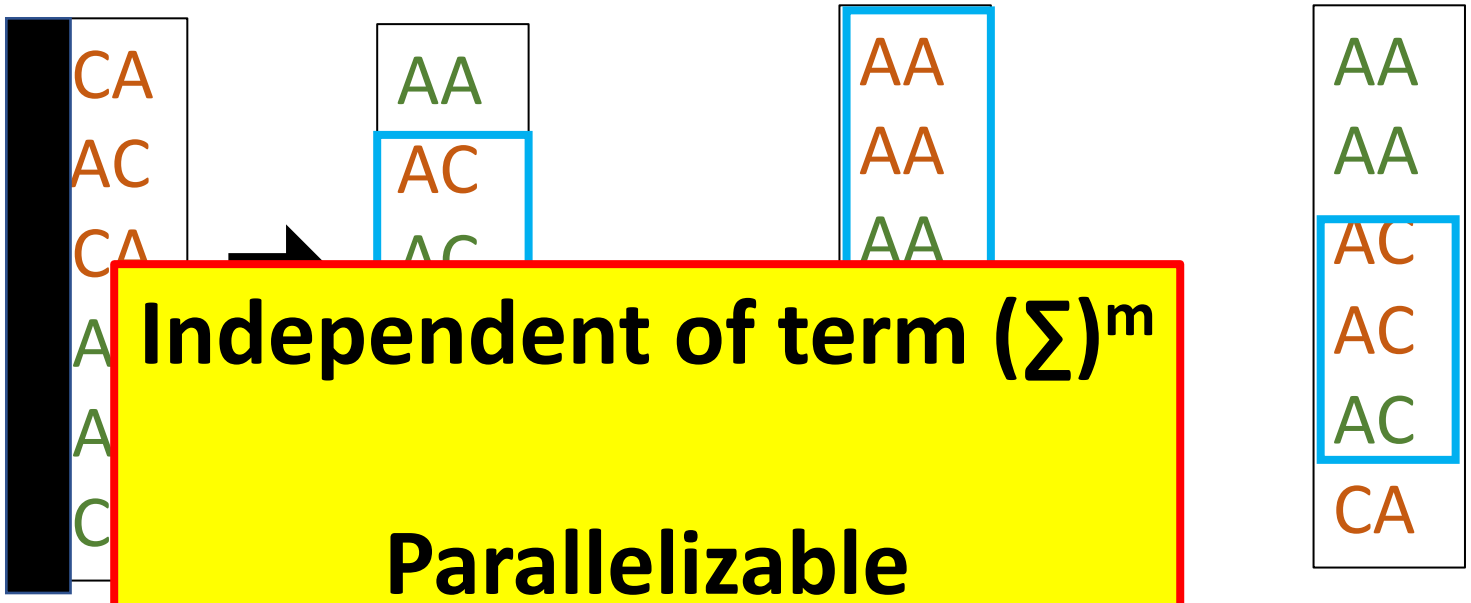
GaKCo : Gapped k-mer Kernel using Counting

S:ACACA

T:AAACA

m=1

g=3



Independent of term $(\sum)^m$

Parallelizable

C_m (S,T)=1+2=3

C_m (S,T)=4

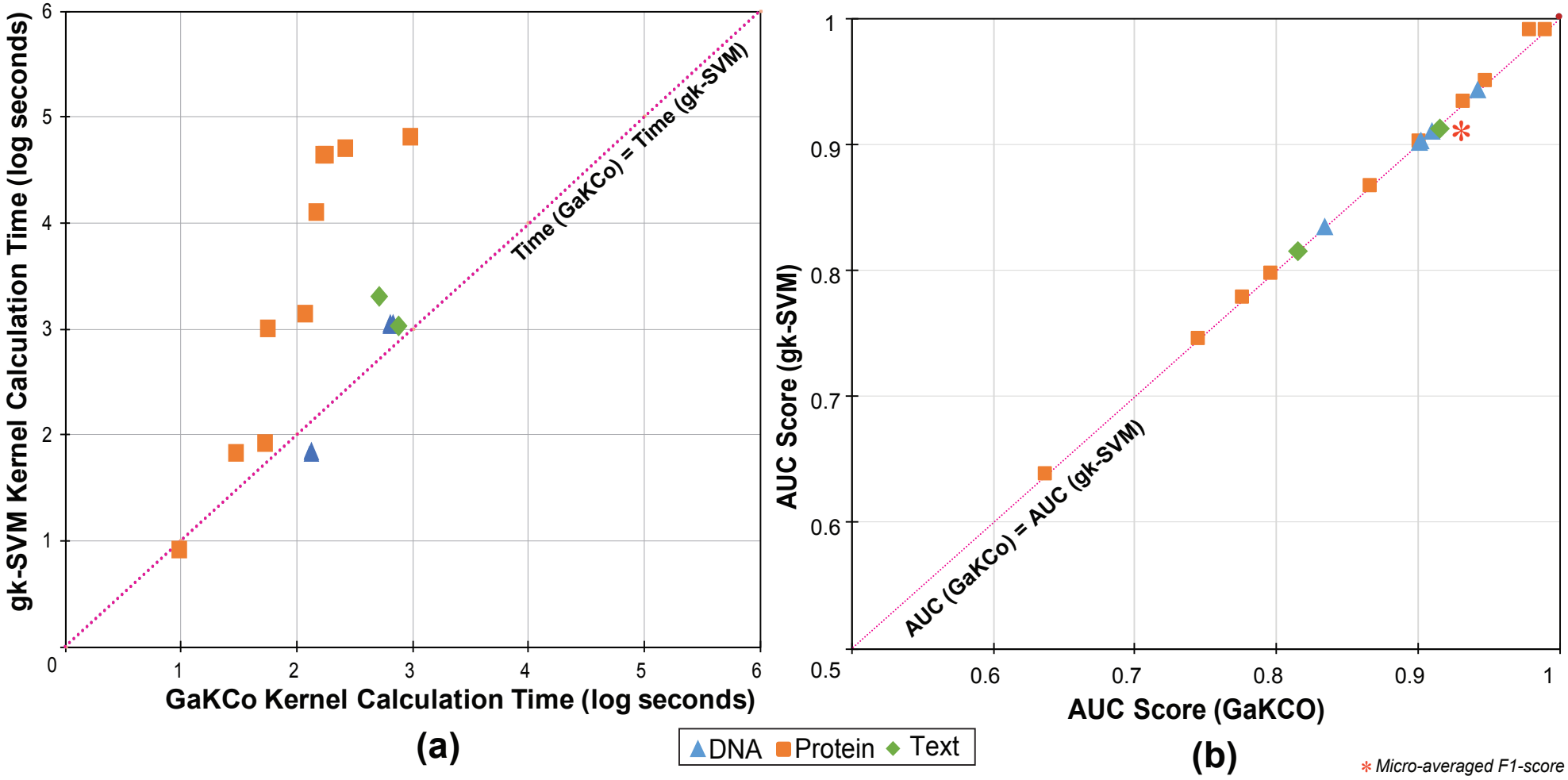
C_m (S,T)=2

Total C_m (S,T)=3+4+2=9

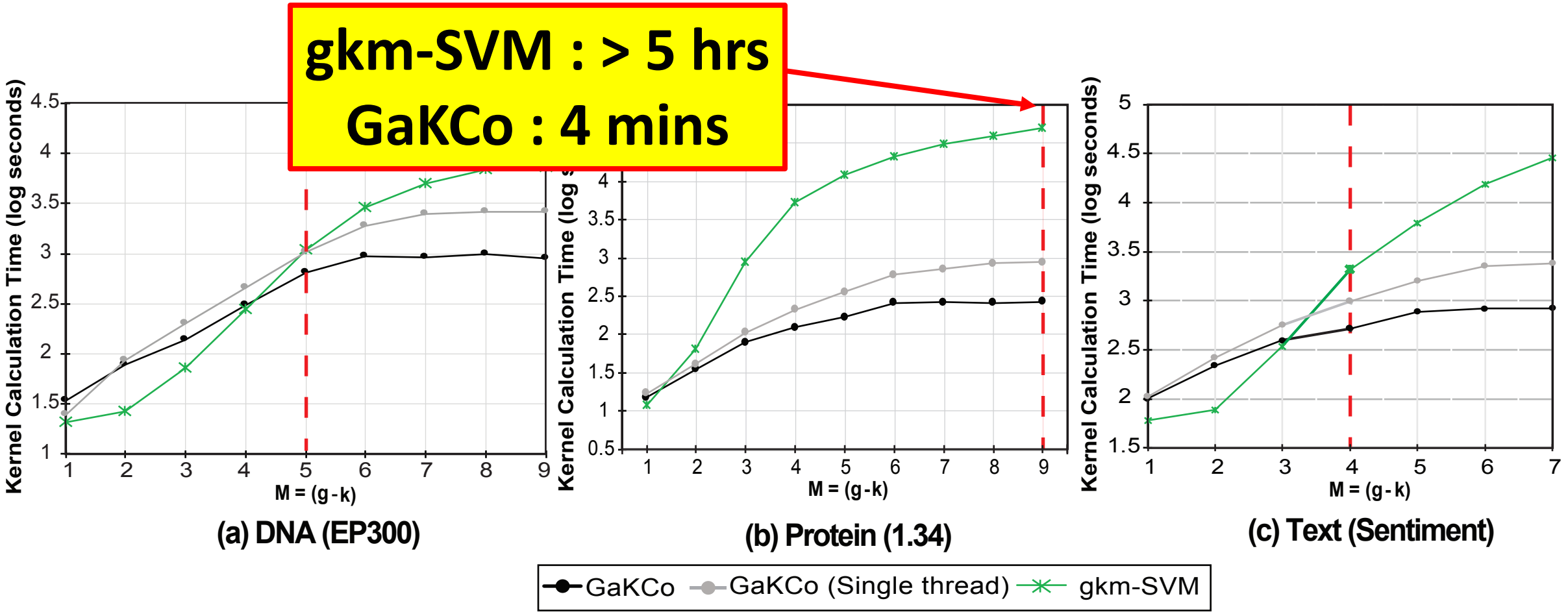
Over Counting = $\binom{3}{1} N_{m=0}$ = 3x 2 = 6

$N_{m=1}$ (S,T)= 9 – 6 =3

GakCo is Faster!

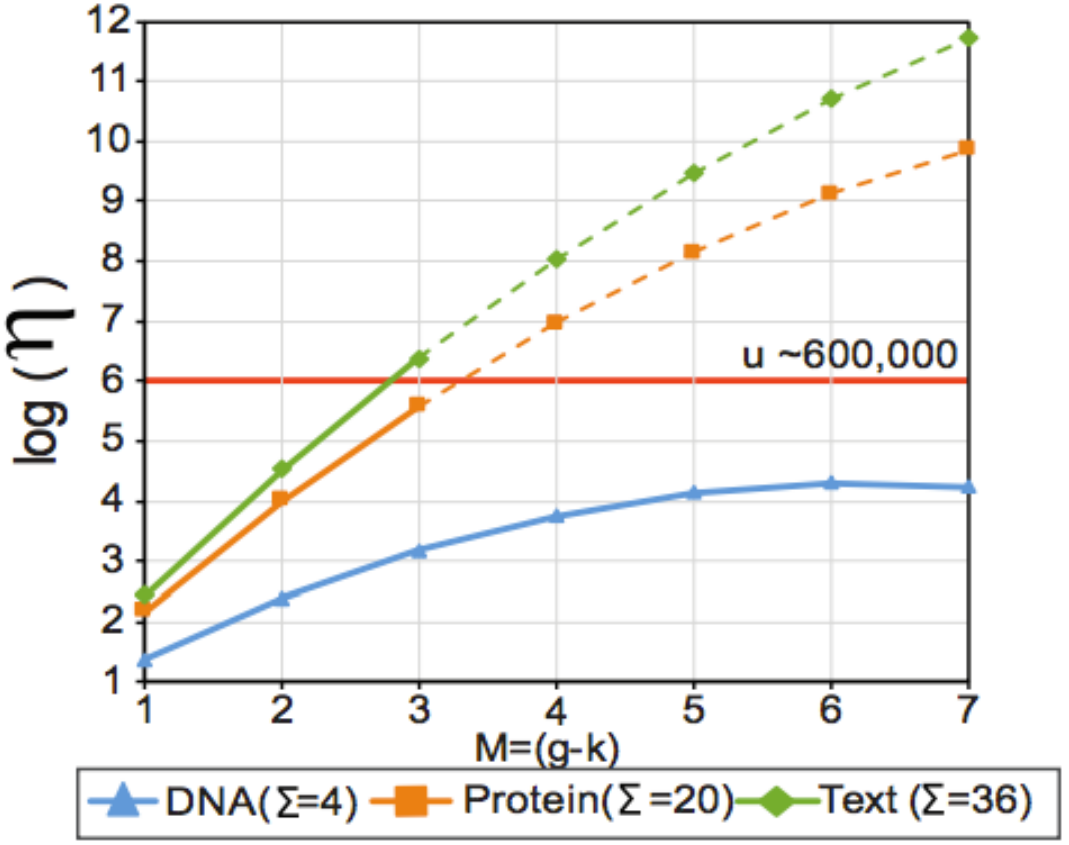


Scales well with increasing ∑ and m



Detailed Theoretical Analysis (in paper)

	GaKCo	gkm-SVM
Pre-processing	$c_{gk}gNl$	$gNl + \eta ug$
Kernel updates	$c_{gk}zN^2$	ηuN^2



Summary: GaKCo

- Fast – (1) reduced gapped k-mer feature space and (2) counting based method
- Scalable – allows larger dictionary size and gap size
- Parallelizable – naturally parallelizable implementation
- Code and datasets available at: https://github.com/QData/GaKCo-SVM
- ArXiv: https://arxiv.org/abs/1704.07468


Thank you