Character based String Kernels for Bio-Entity Relation Detection

Ritambhara Singh and Yanjun Qi

ACL BioNLP Workshop 2016

August 12, 2015

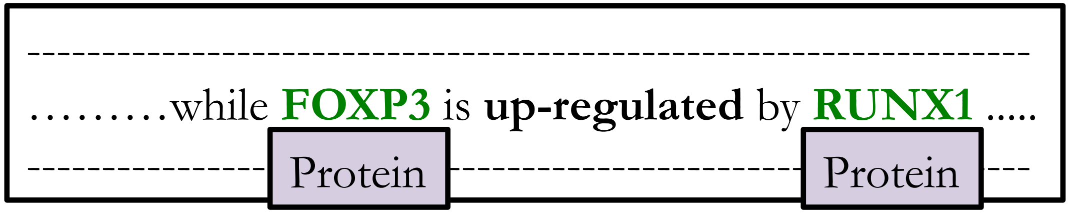
Computer Science at the University of Virginia

University of Virginia Engineering

Introduction

Extraction of bio-entity relations:

- Protein-protein Interaction (PPI)

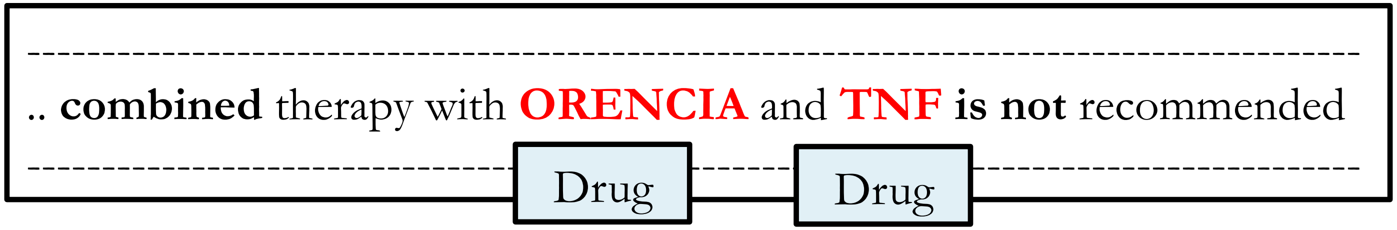


………while **FOXP3** is **up-regulated** by **RUNX1** …..

Protein

Protein

- Drug-drug Interaction (DDI)



.. **combined** therapy with **ORENCIA** and **TNF** **is not** recommended

Drug

Drug

Motivation

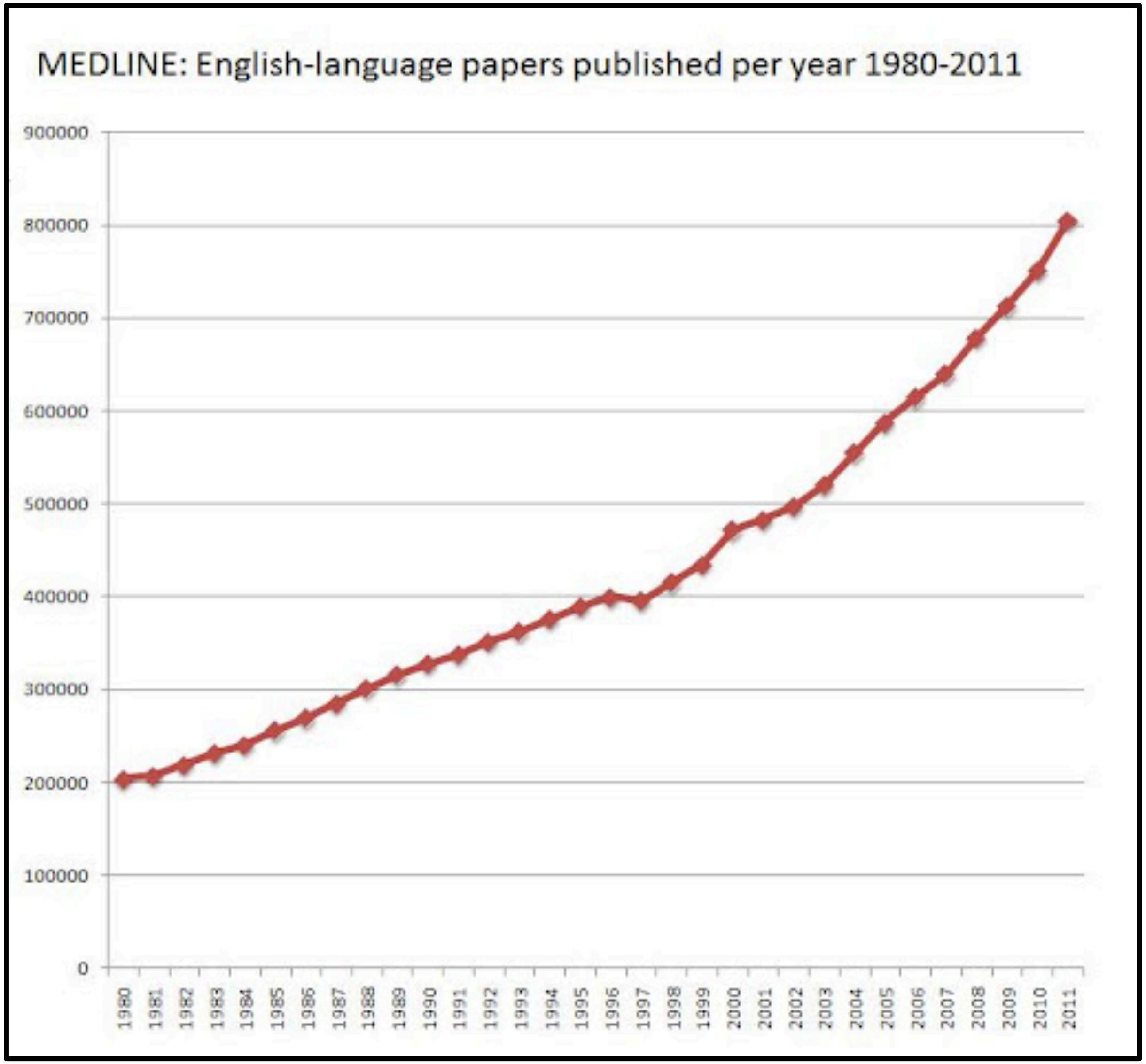
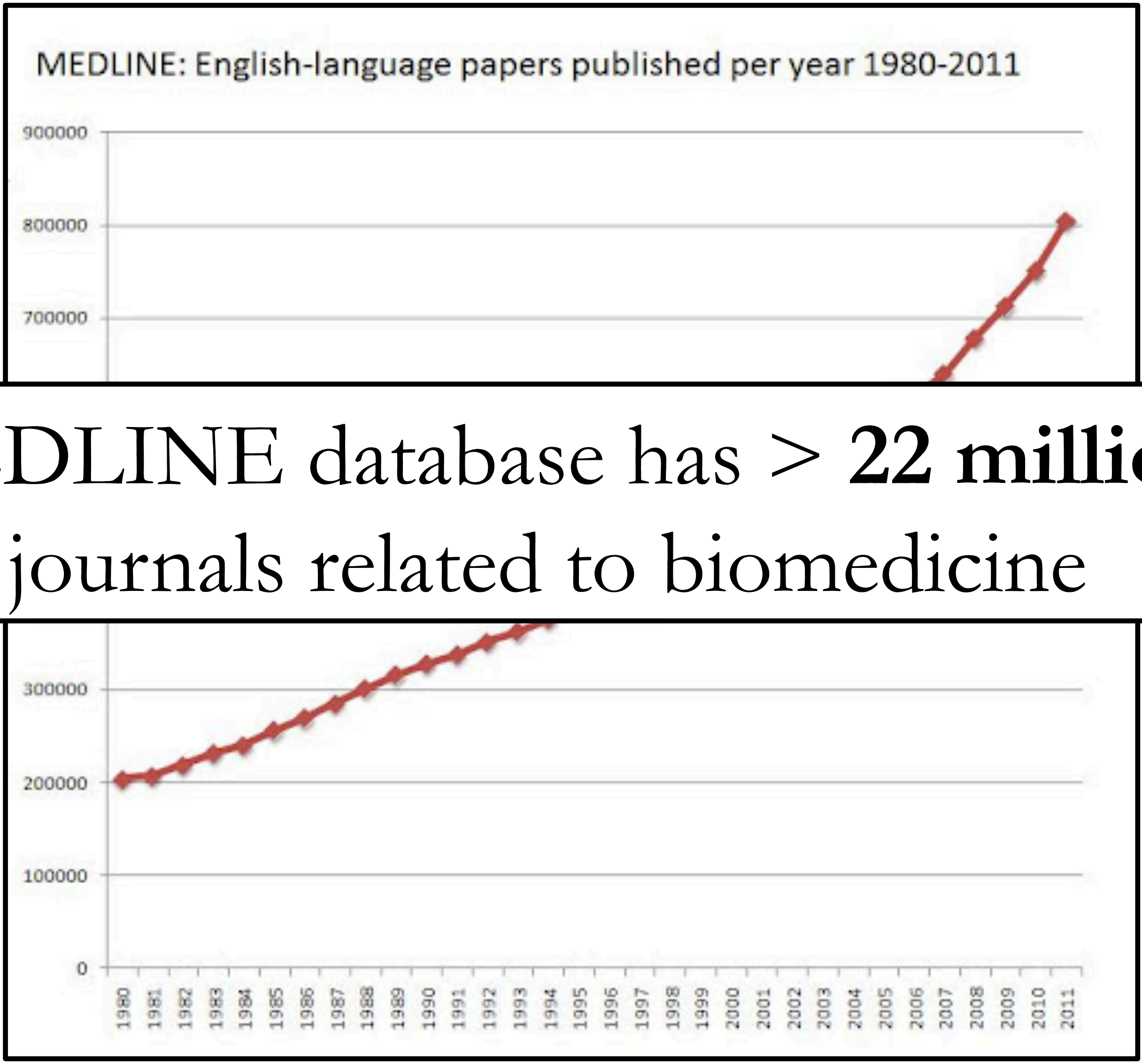


Image Courtesy: "Science: Growing Too Fast?" (Discovery Magazine Blog)

Motivation



MEDLINE database has > **22 million** journals related to biomedicine

Image Courtesy: "Science: Growing Too Fast?" (Discovery Magazine Blog)

Outline

- Background
- Approach
  - Overview
  - String Kernels
- Experimental Setup
- Results

Outline

- **Background**
- Approach
  - Overview
  - String Kernels
- Experimental Setup
- Results

Background

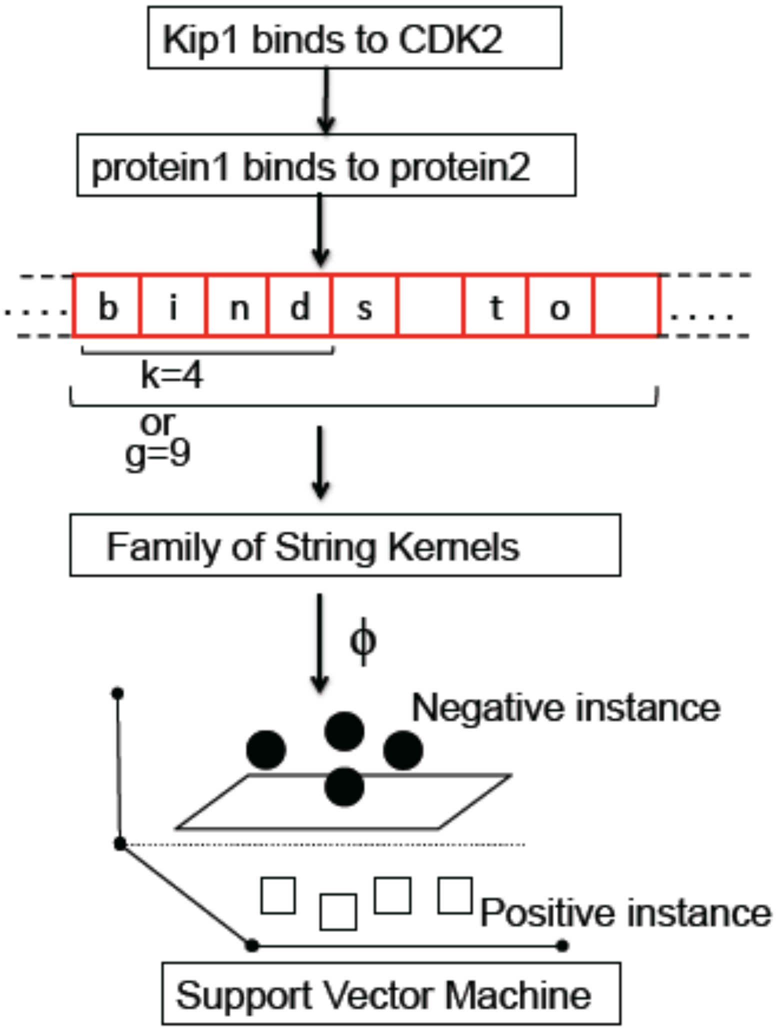
- Convolutional kernels :
  - Shallow Linguistic Kernels (SL)
  - Constituent parse tree-based kernels: subtree (ST), partial tree (PT) etc.
  - Dependency parse tree-based kernels: k-band shortest paths (kBSPS), all-path graph kernel (APG)

- **State-of-the-art: SL, APG, kBSPS**

Outline

- Background
- **Approach**
  - **Overview**
  - **String Kernels**
- Experimental Setup
- Results

Overview



(1) Original Instance

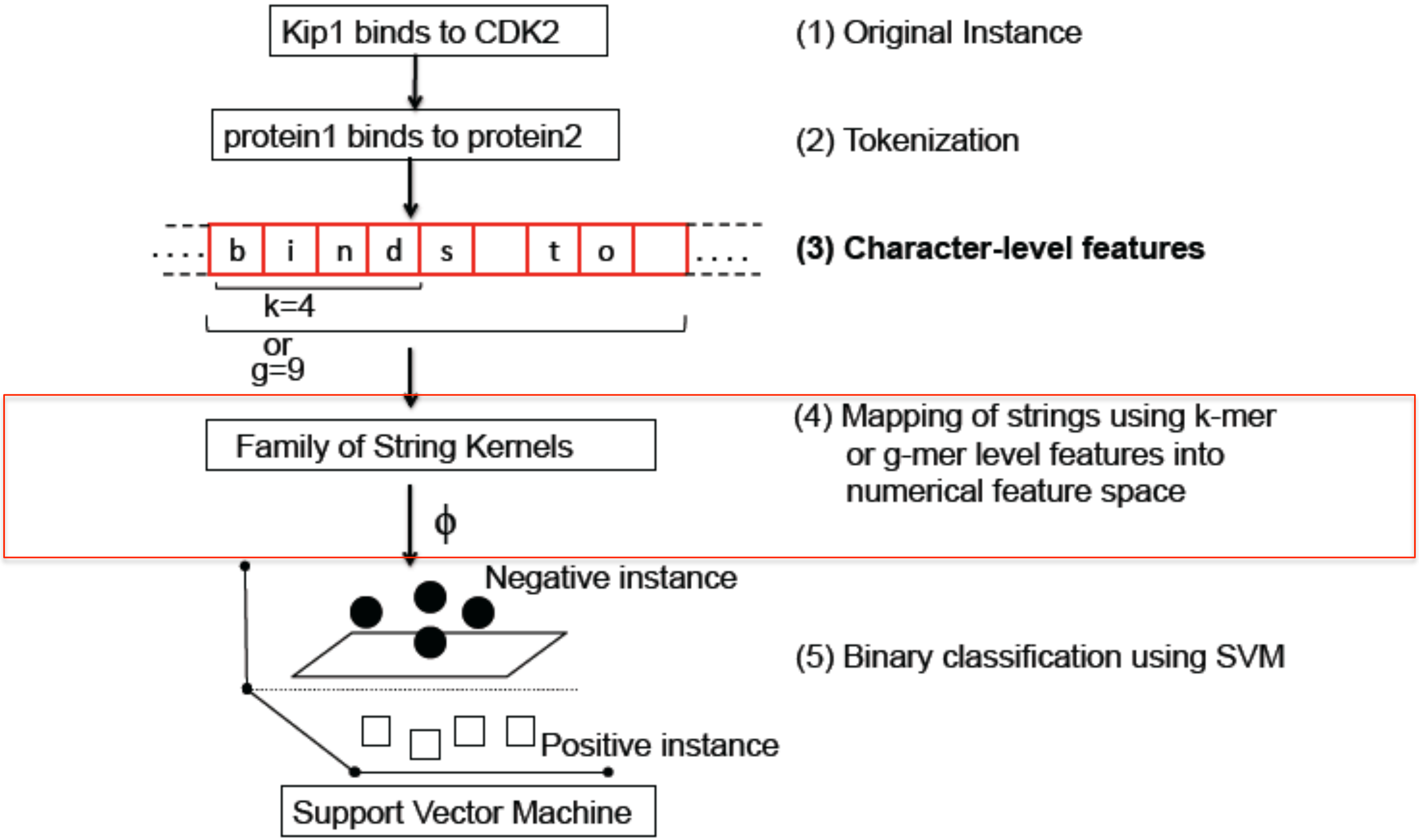
(2) Tokenization

(3) Character-level features

(4) Mapping of strings using k-mer or g-mer level features into numerical feature space

(5) Binary classification using SVM

Overview



Kip1 binds to CDK2

(1) Original Instance

protein1 binds to protein2

(2) Tokenization

b	i	n	d	s		t	o	

(3) Character-level features

k=4

or

g=9

Family of String Kernels

ϕ

(4) Mapping of strings using k-mer or g-mer level features into numerical feature space

Negative instance

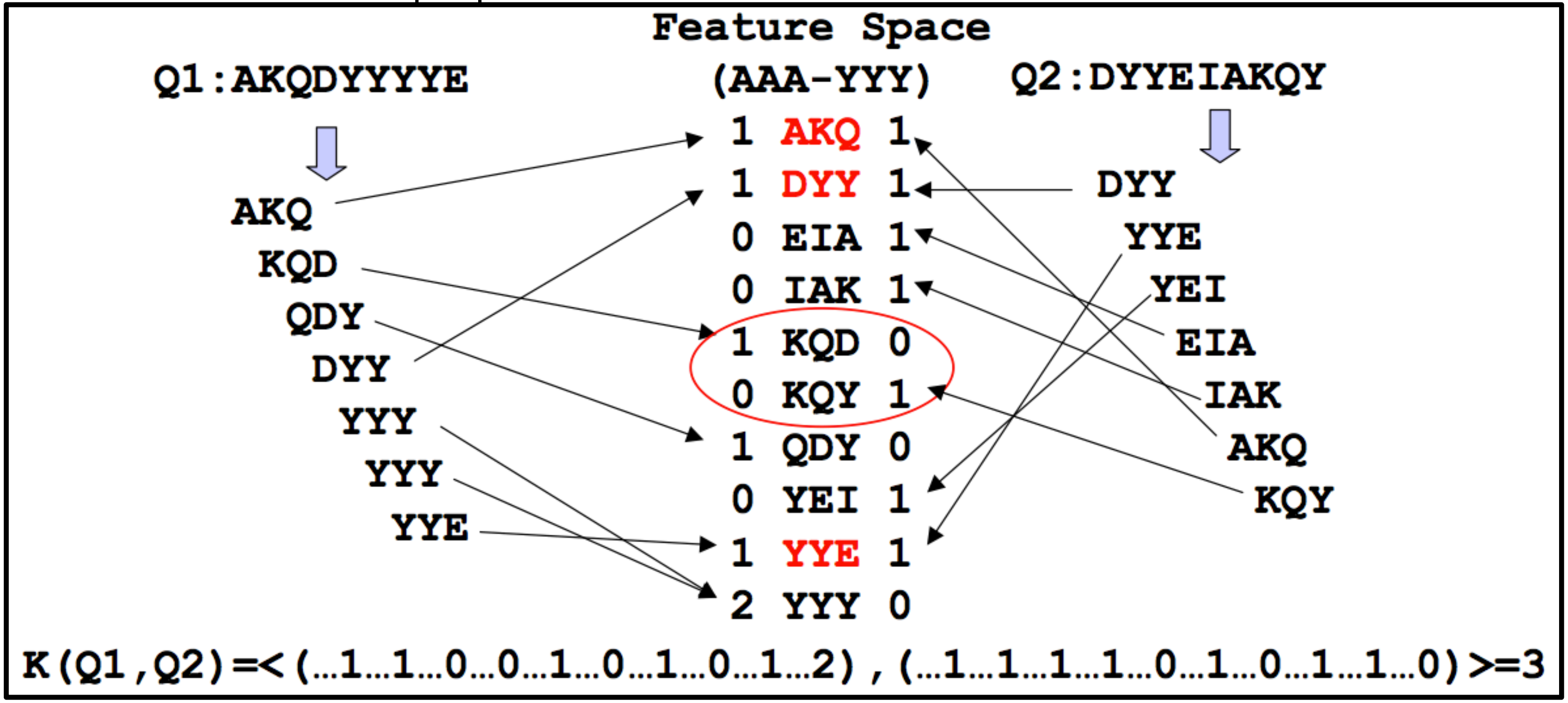
Positive instance

(5) Binary classification using SVM

Support Vector Machine

Spectrum Kernel (SK)

Feature map indexed by all k-length subsequences ("k-mers") from alphabet Σ of amino acids, |Σ|=20



Feature Space

Q1:AKQDYYYYE

AKQ
KQD
QDY
DYY
YYY
YYY
YYE

	(AAA-YYY)	
1	AKQ	1
1	DYY	1
0	EIA	1
0	IAK	1
1	KQD	0
0	KQY	1
1	QDY	0
0	YEI	1
1	YYE	1
2	YYY	0

Q2:DYYEIAKQY

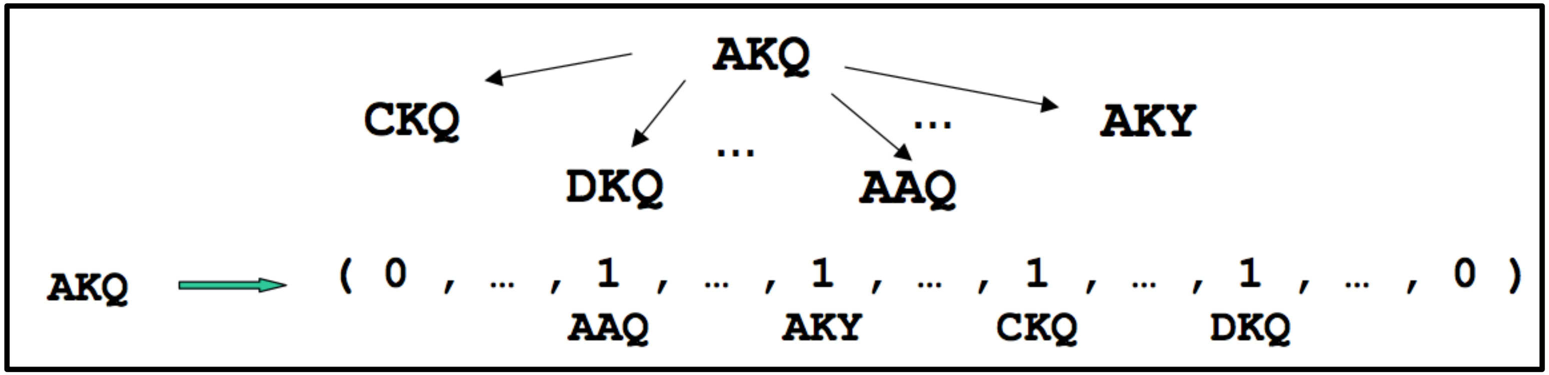
DYY
YYE
YEI
EIA
IAK
AKQ
KQY

K(Q1,Q2)=<(…1…1…0…0…1…0…1…0…1…2),(…1…1…1…1…0…1…0…1…1…0)>=3

Leslie, Eskin and Noble, PSB 2002

Mismatch Kernel (MK)

For k-mer **s**, the mismatch neighborhood $N_{(k,m)}(s)$ is the set of all k-mers *t* within *m* mismatches from **s**.



Leslie, Eskin, Weston and Noble, NIPS 2002

Wildcard Kernel (WK)

Dictionary is augmented with a wild character ‘ * ’

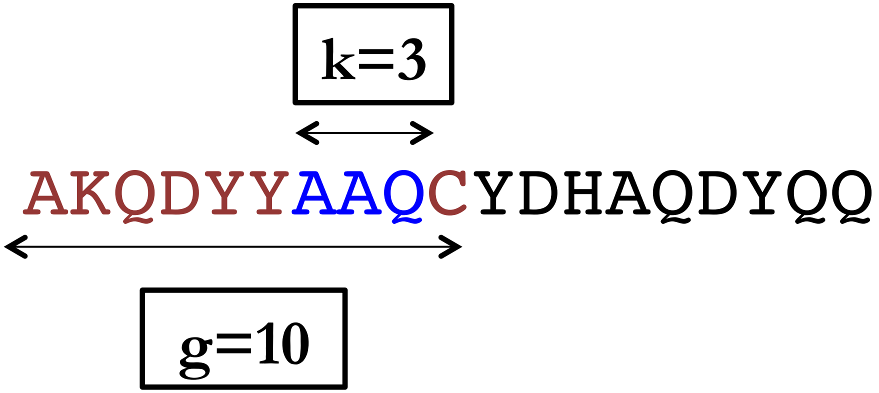
$$\Sigma \cup \{\star\}$$

Feature Space

```
AAA
AA*
A*A
AA*
AAQ
...
...
```

Leslie and Kuang, JMLR 2004

Gapped k-mer based Kernel (GK)



k=3

AKQDYYAAQCYDHAQDYQQ

g=10

number of gaps
(d)=10-3=7

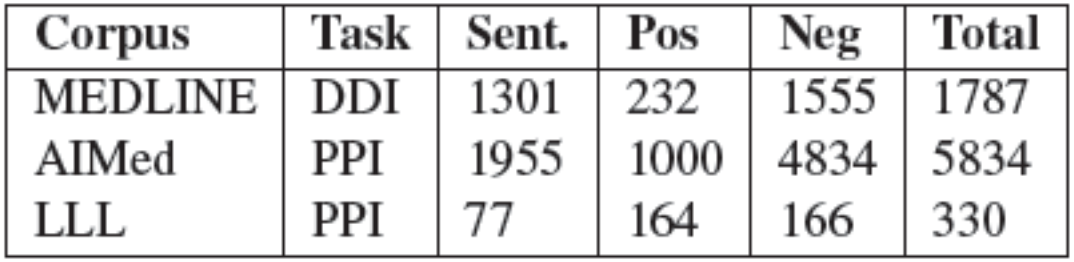
Ghandhi, Mohammed-Noori Ghareghani, Lee, Garraway and Beer,
Bioinformatics 2016

Outline

- Background
- Approach
  - Overview
  - String Kernels
- **Experimental Setup**
- Results

Experimental Setup

- **Dataset:**



Corpus	Task	Sent.	Pos	Neg	Total
MEDLINE	DDI	1301	232	1555	1787
AIMed	PPI	1955	1000	4834	5834
LLL	PPI	77	164	166	330

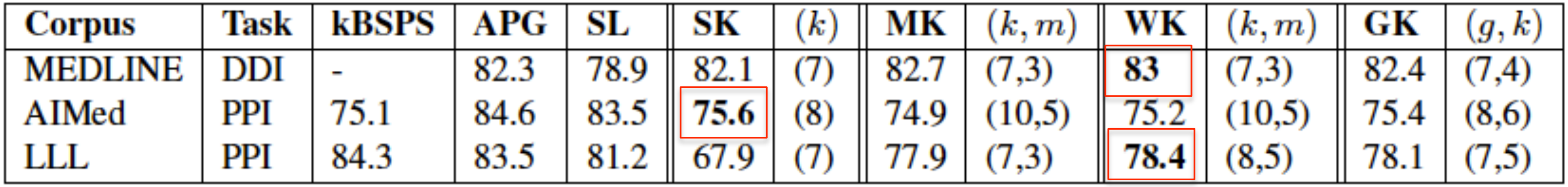
- **Baselines:** SL, APG and kBSPS
- **Parameters:**
  - SK : `k={6,7,..,10}`
  - MK, WK : `k={6,7,..,10};m={1,2,..,k-1}`
  - GK : `g={6,7,..,10};k={1,2,..,g-1}`
- **Evaluation Metric:** AUC Score

Outline

- Background
- Approach
  - Overview
  - String Kernels
- Experimental Setup
- **Results**

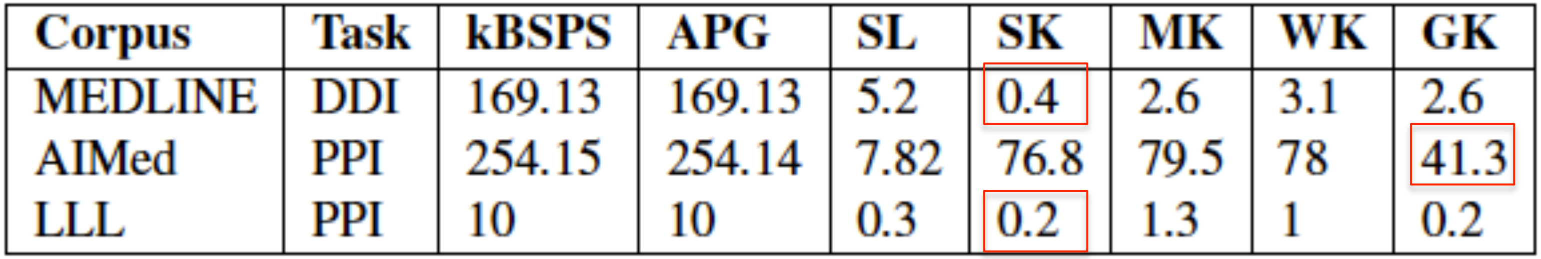
Results

- **Performance:**



Corpus	Task	kBSPS	APG	SL	SK	(k)	MK	(k, m)	WK	(k, m)	GK	(g, k)
MEDLINE	DDI	-	82.3	78.9	82.1	(7)	82.7	(7,3)	**83**	(7,3)	82.4	(7,4)
AIMed	PPI	75.1	84.6	83.5	**75.6**	(8)	74.9	(10,5)	75.2	(10,5)	75.4	(8,6)
LLL	PPI	84.3	83.5	81.2	67.9	(7)	77.9	(7,3)	**78.4**	(8,5)	78.1	(7,5)

- **Time:**



Corpus	Task	kBSPS	APG	SL	SK	MK	WK	GK
MEDLINE	DDI	169.13	169.13	5.2	0.4	2.6	3.1	2.6
AIMed	PPI	254.15	254.14	7.82	76.8	79.5	78	41.3
LLL	PPI	10	10	0.3	0.2	1.3	1	0.2

Conclusion

- Simple and novel character-based representation
- Implement family of string kernels
- Fast and flexible for any bio-NLP dataset
- Complimentary to existing state-of-the-art methods

Thank You