

# Predicting Kinase Functional Sites Using Hierarchical Stochastic Language Modeling

Minghua Deng  
School of Mathematical Sciences,  
Center for Theoretical Biology  
Peking University, Beijing 100871

# Outline

- Motivation
- Methods
- Results
- Summary and conclusion

# Motivation

- Motif finding problems
  - DNA motifs (transcription factor binding sites)
  - Protein motifs
- Motif Modeling
  - Position Specific Scoring Matrix (PSSM)
  - Does not take (insert/delete) into consideration
  - Site dependency.
- Motif cooperation

# Motif Finding Methods

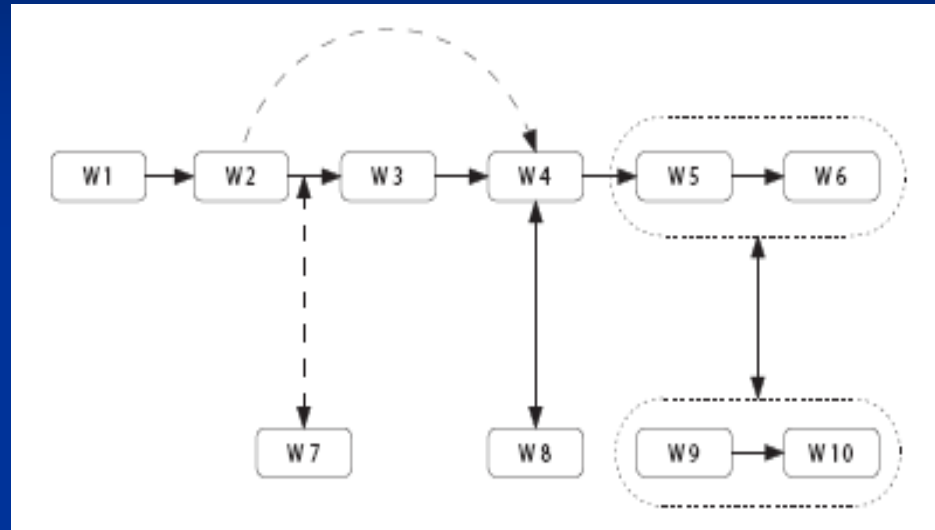
- Methods based on PSSM, using statistical approaches such as EM, Gibbs sampler. Ex: MEME, Biospector, AlignACE.
- Method based on word-counting: finding the over-represented short sequences. Ex: YMF
- Our method is to combine these two paradigms.

# Main Problems

- Input data
  - Kinase sequences: Swiss-Prot (11,115 sequences)
  - Kinase Families: E.C. number in ENZYME database (81 families)
  - We also use PDB, PROSITE and Pfam databases for validation.
- Objective:
  - Finding the functional sites for each kinase family.
  - Kinase discrimination based on the extracted sites.

# Stochastic Language Model

- $\{E; S; R; P; T\}$ 
  - $E$  is a set of symbols
  - $S$  is the starting state
  - $R$  is the set of grammar rules
  - $P$  is the probabilities of different rules in  $R$
  - $T$  is the maximum variations allowed from  $S$ .



# Hierarchical Stochastic Language Model (HSL)

- Keyword model
- Sentence model

# Keyword Model

- Model

- $E$ : 20 amino acids as symbols
- $R$ : insertion/deletion/mutation
- $P$ : alignment with BLOSUM50 matrix and gap penalty  $\delta$
- $T^w$ : Fisher rule to discriminate keywords and the background

$$\frac{\mu_{i1}\sigma_{i2} + \mu_{i2}\sigma_{i1}}{\sigma_{i2} + \sigma_{i1}}$$

# Training Keyword Model

- Constructing the starting state (key)
  - over-represented 4-mers.
    - The top 30 most frequent 4-mers.
    - Chi-square test with Yates correction
  - Greedy algorithm is used to concatenate the 4-mers.
- Wilcoxon rank sum test on the extended word.

# Greedy Concatenating

ABCD

BCDE

CDEF

ABED

BEDD

ABCDE

~~ABEDD~~  
~~ABCDEF~~

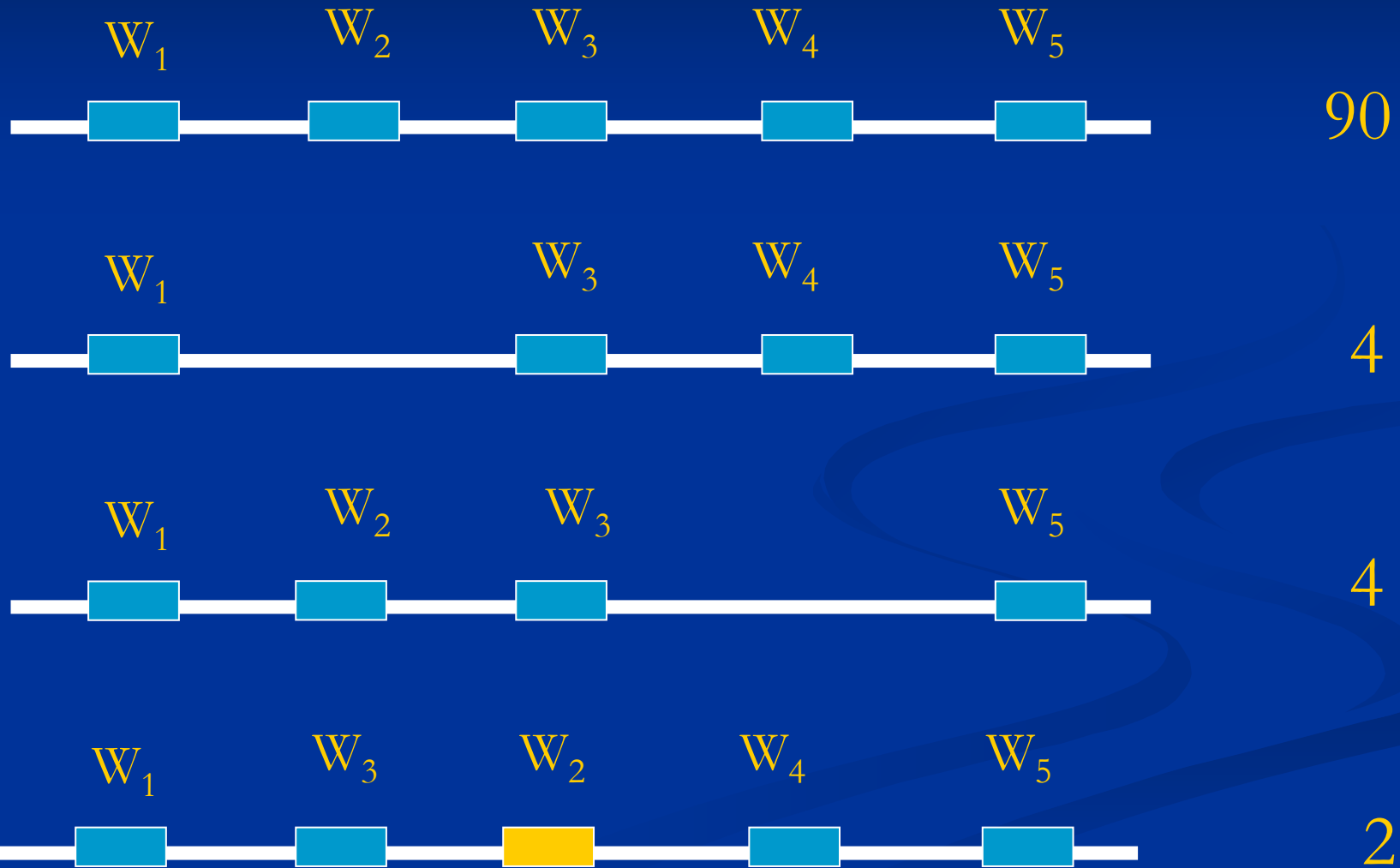
# Sentence Model

- Sentence model is to explore the inter-relationship among the keywords.
- Impliedly, we assume that the order of the keywords are conserved in most of the training sequences. Such a conserved order constitute the starting sentence.
- In some training sequences, some keywords may not present, they are “deleted” from the starting sentence.

# Sentence Model

- Model elements
  - E: Keywords
  - S: Start sentence, ordered keywords.
  - R: deletion.
  - P: The probabilities of deletion of keywords.
  - $T^s$ : Maximum variation allowed.

# Example



# Training Sentence Model

- Given the sequence, its score under the HSL model contains two parts
  - Sum of the matched keyword score
  - Keywords present probability

$$\sum_{i=1}^n \left( 3 \log_2(1 - p_i) + S_i \right) \times m_i$$

Where  $n$  is the number of keywords,  $m_i$  represent whether  $i$ th keyword present in the sequence or not,  $p_i$  is the deletion probability of  $i$ th keyword, as  $S_i$  is the keyword matching score. Here 3 is a scaling number same as BLOSUM 50.

# Training Sentence Model

- All keywords identified in the keyword model are the elements of the sentence model.
- The starting sentence is defined as the most common order of keywords located in the positive training sequences.
- Align keywords back to sequences in the positive training data using DP algorithm.
- An iterative procedure is used to training deletion probability of each keywords.

# Iterative Procedure

- Maximization:  $P(\text{SEQ}, \text{Match}, P^s)$  with missing data “Match”.
- Initialization: set all deletion probabilities as 0.
- Finding the best match at current estimation of  $P^s$ .
- Updating  $P^s$  by counting.
- Repeating until convergence.

# Find the Best Match

0 3 3 0 1 0 0 0 2 0 T=3.5

0 0 0 7 03 10 56 28 90 17 T=5.3

3 2 10 05 9 13 17 30 16 20 T=5.6

0 4 6 6 0 4 0 15 4 0 T=3.6

Maximization:

$$\sum_{i=1}^n \left( 3 \log_2(1 - p_i) + S_i \right) \times m_i$$

# Sub-families

- Our model preferred to the families evolved from same ancestor---homologues.
- It's possible that one kinase family contains sequences evolved from different ancestors---analogues.
- Divide the family into two sub-families
  - More than 10 sequences.
  - More than 5% sequences.

# Results

## ■ PROSITE

- HSL: Recall/Precision=83.5%/23.0%
- MEME: Recall/Precision=37.3%/29.1%.

## ■ PDB

- HSL: Recall/Precision=66.1%/79.9%.
- MEME: Recall/Precision=44.7%/74.7%.

## ■ Pfam

- HSL: 90% keywords covered by Pfam.
- MEME: 77.2% keywords covered by Pfam.

# Results

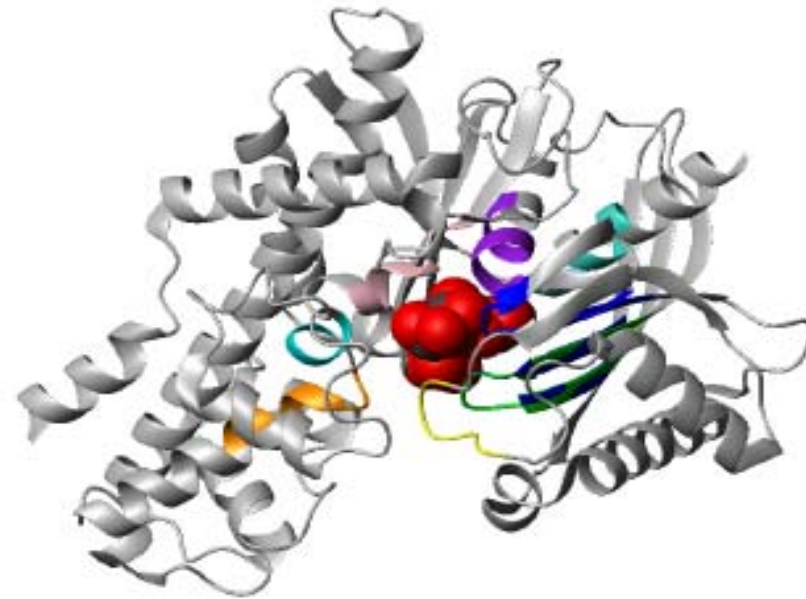
**Table 1.** Comparing our keywords with the results with PDB structures and PROSITE patterns.

Source	Hexokinase (EC 2.7.1.1)			
HSL model	1aLDLGGT <del>NFRV</del> - <i>LGFTFSFP</i> -	<i>WTKGF</i> -VNDTVGT-	iNMEWG-	YEKM-SGMYlgei-DGSG-GAa1
GLC (1cza)	<b>DLGGT</b> -	<i>TFSFP</i> c-lit <b>WTKGF</b> -VNDTVGT-livgtgsn-NMEWGafgd-kqr <b>YEKM</b> iSGMY		
G6P (1cza)	<b>LDLGGT</b> NFRV1- <i>FTFSF</i> -	<b>KGF</b> -VNDTVGT <del>mt</del> -livgtgsn-vdgtlykl-		lse <b>DGSG</b> kGA
PROSITE	[Livm]- <i>G-F</i> -[Tn]- <i>F-S</i> -[Fy]-P-x(5)-[livm]-[dnst]-x(3)-[livm]-x(2)- <i>W-T-K</i> -x-[LF]			

Comparing our keywords with the results from PDB structures and PROSITE patterns. HSL model: the matching keywords of the HSL model on the sequence; GLC: the contacting regions for GLC in PDB:1cza; G6P: the contacting regions for G6P in PDB:1cza; PROSITE: the pattern found in PROSITE. The bold capitals in the table are those residues where our keywords match exactly with the putative binding sites in PDB. The italic capitals are matching residues with pattern of PROSITE.

Mutant monomer of recombinant human hexokinase of type 1 (PDB code: 1cza) from hexokinase family. It has two similar domains which complex with glucoses (GLC) and glucose-6-phosphates (G6p).

# Results



**Fig. 3.** The ribbon structure of 1 domain in hexokinase from human brain (PDB code: 1cza) complex with GLC and G6P. GLC and G6P are shown in space-filling model. The keywords of the HSL model are shown as color ribbons. These ribbons are close to the ligands and form a pocket to bind with them. Residue Asp657(N) of the purple ribbon, "VNDTVGT", is the putative catalytic base and the conformation of green ribbon "LDLGGTNFRV" is identical with that observed in the G6P/GLC complexes of the wild-type enzyme (Aleshin *et al.*, 2000). This drawing was prepared with the program MOLMOL. c



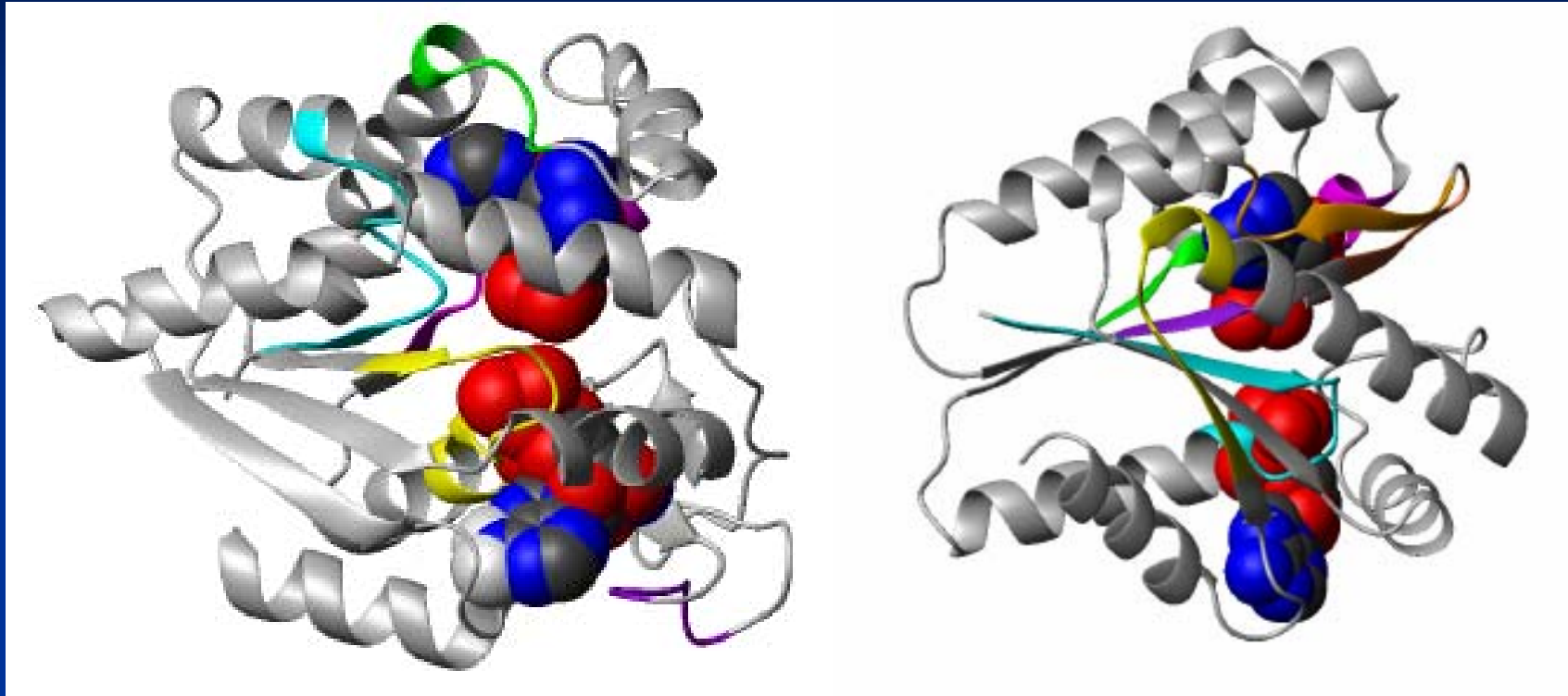
# Results

**Table 3.** Comparing our keywords with the results with PDB structures and PROSITE patterns.

Source	The first sub-family of E.C. 2.7.4.3.				
HSL model	<b>LGAPGAGKGTQA-</b>	<b>QISTGDMLR-</b>	<b>LVTDE-</b>	<i>gFL</i> <b>LDGFPRTI-</b>	<b>FNPPK</b>
2eck:AMP	<b>1LGAPGAGKGTQA-</b>	<b>QISTGDMLR-</b>	gkqakdimdagk	<b>LVTDE</b> lvialv-	<b>LDGFPRTI</b> pga-rkddqeetvrkrlvey
2eck:ADP	<b>1LGAPGAGKGTQA</b> qf-	<b>QISTGDMLR-</b>		<b>DGFPR-</b>	grvyhvk <b>FNPP</b> -dgtkpvaev
PROSITE	[livmFywca]-[Livmfyw] (2)- <b>D-G</b> -[Fyi]- <b>P-R</b> -x(3)-[nq]				
Source	The second sub-family of E.C. 2.7.4.3.				
HSL model	kig <b>IVTGIPGVGK-</b>	<b>INYG-</b>	<b>RDEMR-</b>	<b>IDTH-</b>	<b>IRTP-</b> gy <b>LPGLP-</b> vlagstvkv
Inks:AMP	<b>IVTGIPGVGK</b> stvl-	<b>INYG</b> dfm-	d <b>RDEMR</b> kl-qkklq-	<b>IDTH</b> av <b>IRTP-</b>	<b>LPGLP-</b> rnr
Inks:ADP	<b>VTGIPGVGK</b> stv-			<b>DTH-</b> srqkrdttr-	vivnvegdps
PROSITE	[livmfywca]-[livmfyw] (2)-d-g-[fyi]-p-r-x(3)-[nq] this pattern can not be found in the sequences of this sub-family.				

Comparing our keywords with the results from PDB structures and PROSITE patterns. HSL model: the matching keywords of the HSL model on the sequence; AMP: the contacting regions for AMP in chain A of PDB:2eck and chain F of PDB: Inks; ADP: the contacting regions for ADP in chain A of PDB:2eca and chain F of PDB: Inks; PROSITE: the pattern found in PROSITE. The bold capitals in the table are those residues where our keywords match exactly with the putative binding sites in PDB. The italic capitals are matching residues with pattern of PROSITE.

# Results



The ribbon structure of two adenylate kinases. The left one shows a domain structure of E.Coli, the right one shows the domain structure of adenylate kinase from a trimeric archaeal, both of them with bound AMP and ADP. The color ribbons are our predicted keywords.

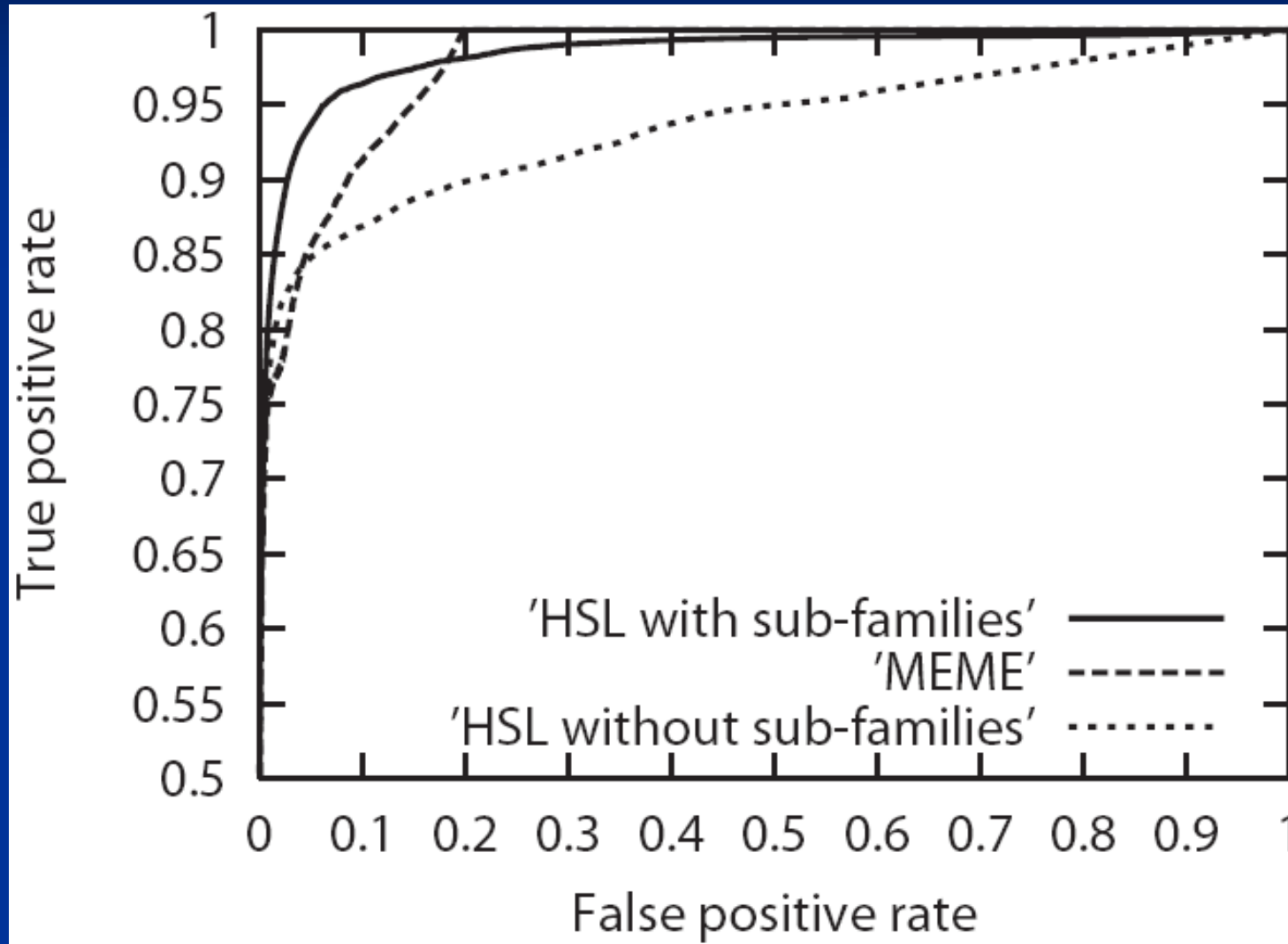
# Results

**Table 2.** Cross-validation results

Method	<i>SN</i>	<i>SP</i>
HSL model with sub-families	94.7%	94.0%
HSL model without sub-families	83.1%	96.9%
MEME with same <i>SN</i> as HSL model	94.5%	85.8%
MEME with same <i>SP</i> as HSL model	86.8%	93.7%

the average *SN* and *SP* of HSL model with sub-families, HSL model without sub-families and MEME.

# Results



# Summary

- Hierarchical stochastic language model (HSL).
- HSL can capture the most important signals.
- HSL takes relationship between keywords into consideration.
- Our approach can detect sub-families automatically.
- HSL outperforms MEME in kinase family classification.

# Acknowledgements

- Huan Yu, PhD student (PKU)
  - Guojun Pei, Graduate student (PKU)
  - Minping Qian (PKU)
  - Luhua Lai (PKU)
  - Fengzhu Sun (USC)
- 
- NSFC, Ministry of Science and technology of People's Republic of China

Thank You!