

# Predicting Protein Interdomain Linkers using Hidden Markov Models

Christine Elsik, Kyoungghwa Bae,  
Bani Mallick

Texas A&M University

# Outline

- Objective and Motivation
- Evolutionary and Structural Domains
- Dataset
- Linker Index
- Model I: Conventional Hidden Markov Model
- Model II: Nonstationary Hidden Markov Model

# Objective and Approach

- Identify domain boundaries using sequence alone for proteins without known homology
- Approach is to identify linker regions based on amino acid propensity (linker index)

# Our Motivation

- Improve automated clustering algorithms to group proteins with similar domain architecture - for protein family studies and annotation
- Existing algorithms do not effectively cluster full-length multidomain proteins
- Algorithms that combine clustering and domain dissection often generate fragmented domains

# Evolutionary and Structural Domains

- Structural Domain - independently folding unit
  - ◆ Domain prediction methods often use atomic coordinates from experimentally determined 3D structures or predicted structure
- Evolutionary Domain - based on sequence homology
  - ◆ Domain prediction uses regions of conservation in sequence alignments

# Evolutionary and structural domains are not always equivalent

- Evolutionary domains can merge structural domains if two structural domains are always found together and there is insufficient phylogenetic representation to detect divergence in linker region
- Structural domain can be split into two evolutionary domains if there is a long divergent non-linker loop region or if sequence database contains domain fragments

# Pfam: Evolutionary Domains

- PfamA
- Based on profile Hidden Markov Models built on multiple sequence alignments
- Generated using Swissprot/TrEMBL
- Requires sufficient number of alignable homologs in database
- There is an attempt to reconcile domain boundaries with structure for domains with PDB structures

# Our Dataset

- Select Swissprot/TrEMBL proteins that can be entirely classified by PfamA domains, with the exception of linkers (4-20 residues) and short tails ( $\leq 20$  residues)
- Why 20 residue length cutoff?
  - ◆ Longer segments may be unknown domains
- Remove proteins with transmembrane regions and proteins annotated as fragments
- 11,968 sequences with at least one linker (14,339 linkers with 28,726 corresponding domain regions)

# Removing Redundancy

- First we grouped 11,968 proteins into homeomorphic families (identical domain organization)
- Then all-by-all FASTA comparison of 11,968 proteins
- Single Linkage clustering using E-value  $\leq 10^{-6}$  and  $\geq 80\%$  alignment
- Due to transitive nature of single linkage clustering, some clusters had  $> 1$  domain organization
- Selected one sequence from each domain organization within each cluster
- 802 proteins with at least one linker (993 linkers with 1988 corresponding domain regions)

# Linker Index

Linker index,  $y_l$ , represents the preference of amino acid  $l$  for linker regions relative to domain regions

$$y_l = -\ln\left(\frac{f_l^{linker}}{f_l^{domain}}\right)$$

Where  $f_l$  is the relative frequency of amino acid  $l$  in the linker or domain region.

$y_l$  will be negative if the relative frequency is greater in the linker than domain regions.

## Amino Acid Frequencies and Linker Index for AA with significant difference between linker and domain

Amino Acid	Frequency in Linker*	Frequency in Domain	<i>y/l</i>
Pro	6.63 (6.07)	4.30	-0.4188
Lys	6.97 (5.72)	5.81	-0.2134
Ser	7.20 (5.55)	6.13	-0.1629
Glu	7.97 (6.89)	6.60	-0.1794
Asp	6.32 (5.28)	5.60	-0.1278
Gln	3.90 (4.05)	3.33	-0.1051
Val	6.24 (6.64)	7.34	0.1782
Tyr	2.46 (3.47)	3.38	0.2500
Leu	7.51 (9.60)	9.54	0.2523
Phe	2.74 (4.34)	4.03	0.3561
Ile	4.73 (5.13)	6.37	0.2758
Trp	0.81 (1.24)	1.32	0.3836
Cys	0.89 (1.24)	1.50	0.5724

\*() = Frequency in George and Heringa (2002, Protein Eng 15:871-879) structural linker dataset

# Differences in frequencies between our dataset and structural dataset

- Possible reasons
  - ◆ Our dataset was based on evolutionary domains instead of structural domains?
  - ◆ Linkers in our dataset were artificially truncated?

# Smoothed Linker Index

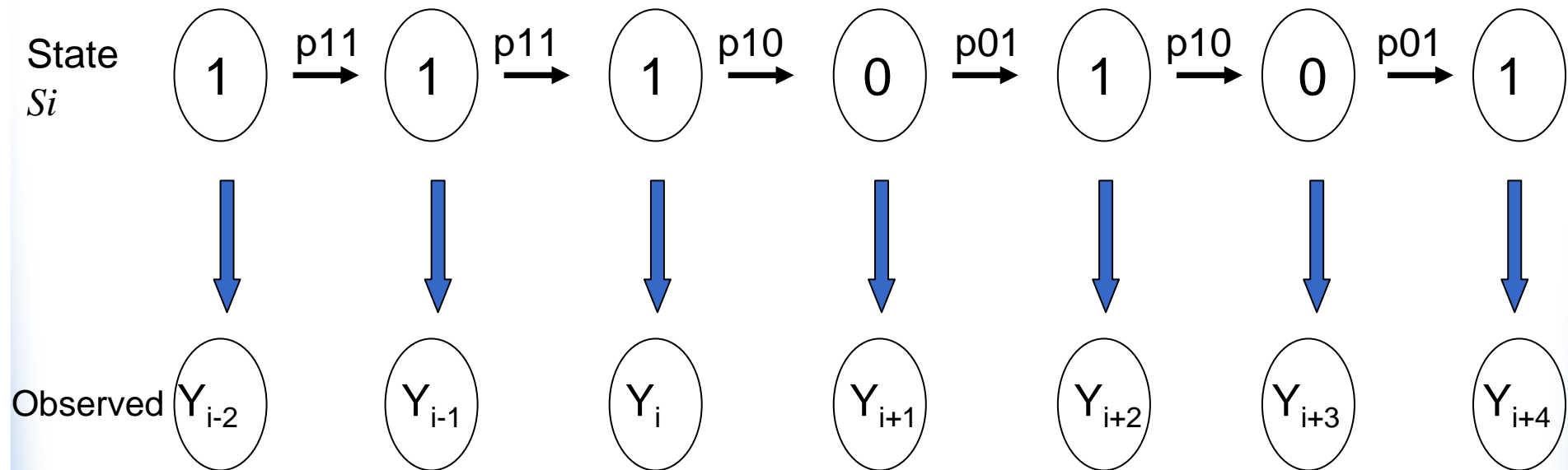
- $y_i$
- Sliding window of length 9
- Take average  $y_l$  within each window and assign this averaged value,  $y_i$ , to the center residue of the window
- Window size 9 gave the maximum difference between linker and domain regions of the sizes tested (3-20)

# Model I

- Conventional Hidden Markov Model
- Protein sequences are assumed to be produced by a HMM and to consist of regions that are homogeneous within a region and differ between regions.
- Each region may be classified into one of two states (linker and non-linker).
- We wish to estimate the hidden states given the observed protein sequence.
- Our sequence data is continuous (string of linker index values) instead of categorical (string of amino acids).
- We must identify linker index values that discriminate linker and non-linker regions.

# Representation of HMM

Transition probabilities



For our model:

States are linker (0) or non-linker (1).

Observations are smoothed linker index ( $y_i$ ).

# Estimation of Parameters

- Our objective is to infer:
  - ◆ the hidden state (linker or non-linker)
  - ◆ Parameters of the model, which include mean  $y_i$  for linkers, mean  $y_i$  for non-linkers and variance in  $y_i$
  - ◆ Parameters of transition probabilities ( $p_{00}$ ,  $p_{11}$ )

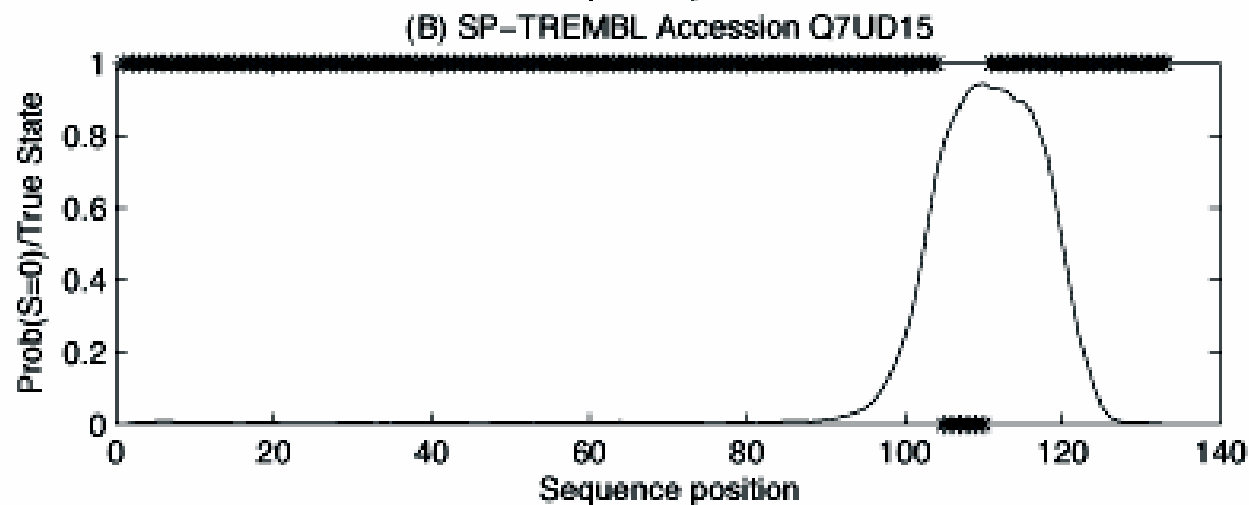
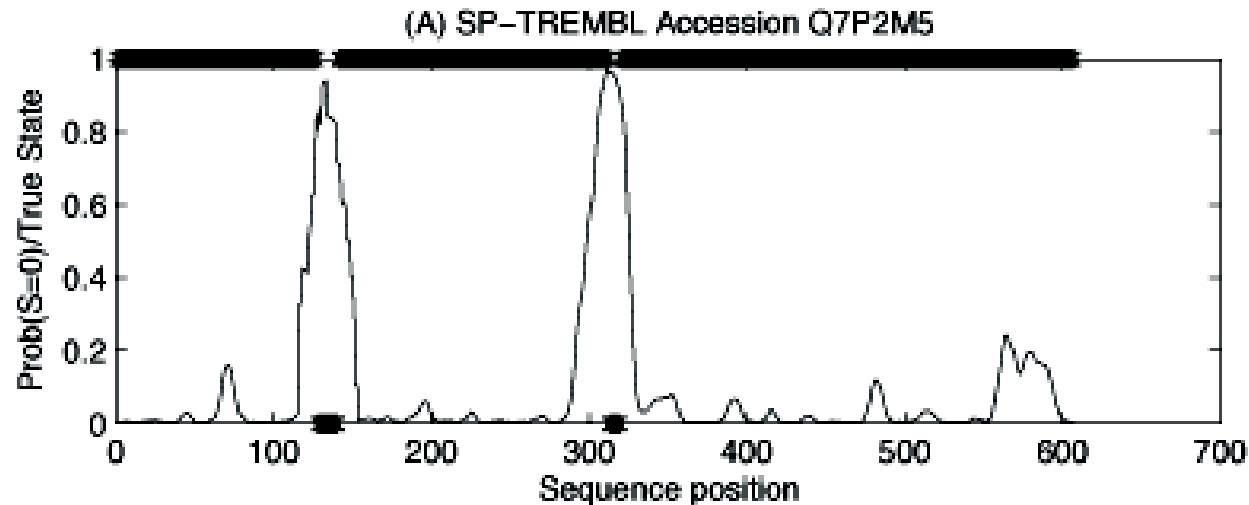
## Estimation of Parameters (cont.)

- Bayesian approach to infer the parameters
- The challenge is determining the posterior distribution of each parameter
- We simulate unknown parameters using a Markov Chain Monte Carlo method, Gibbs Sampling

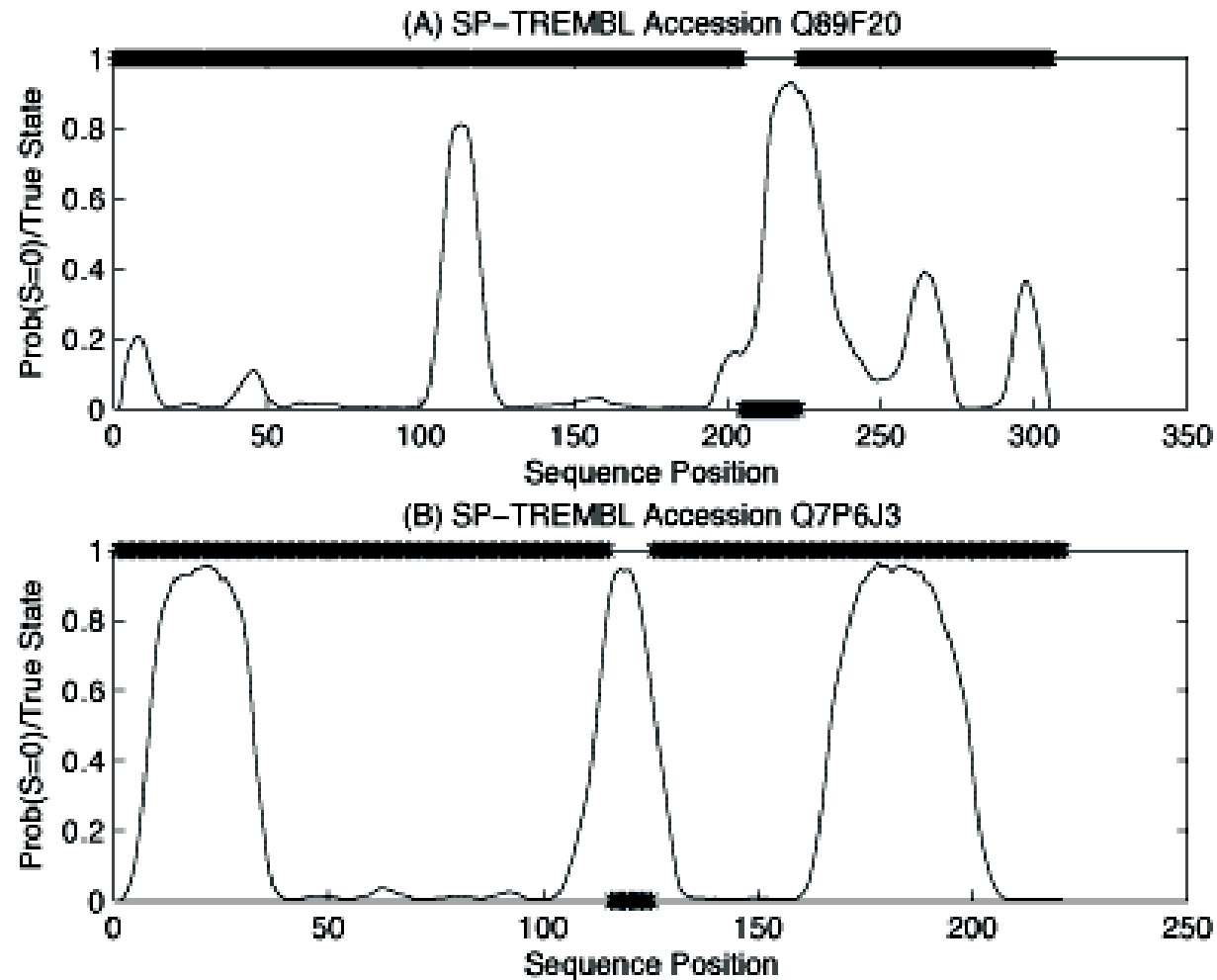
# Predicting the State of a Residue

- The MCMC simulated sample values (posterior distribution) allow us to determine the posterior expectation  $E$ :
- Expected probability that the state  $S_i$  is  $k$  (linker or non-linker) given  $y_i$  and  $S_{i-1}$
- We predict the state of a residue using a classification variable (CV)
  - ♦ CV is 1 if  $E \leq x$
  - ♦ CV is 0 if  $E > x$
  - ♦  $x$  = selected cutoff

# Model I Example - “Well behaved” proteins



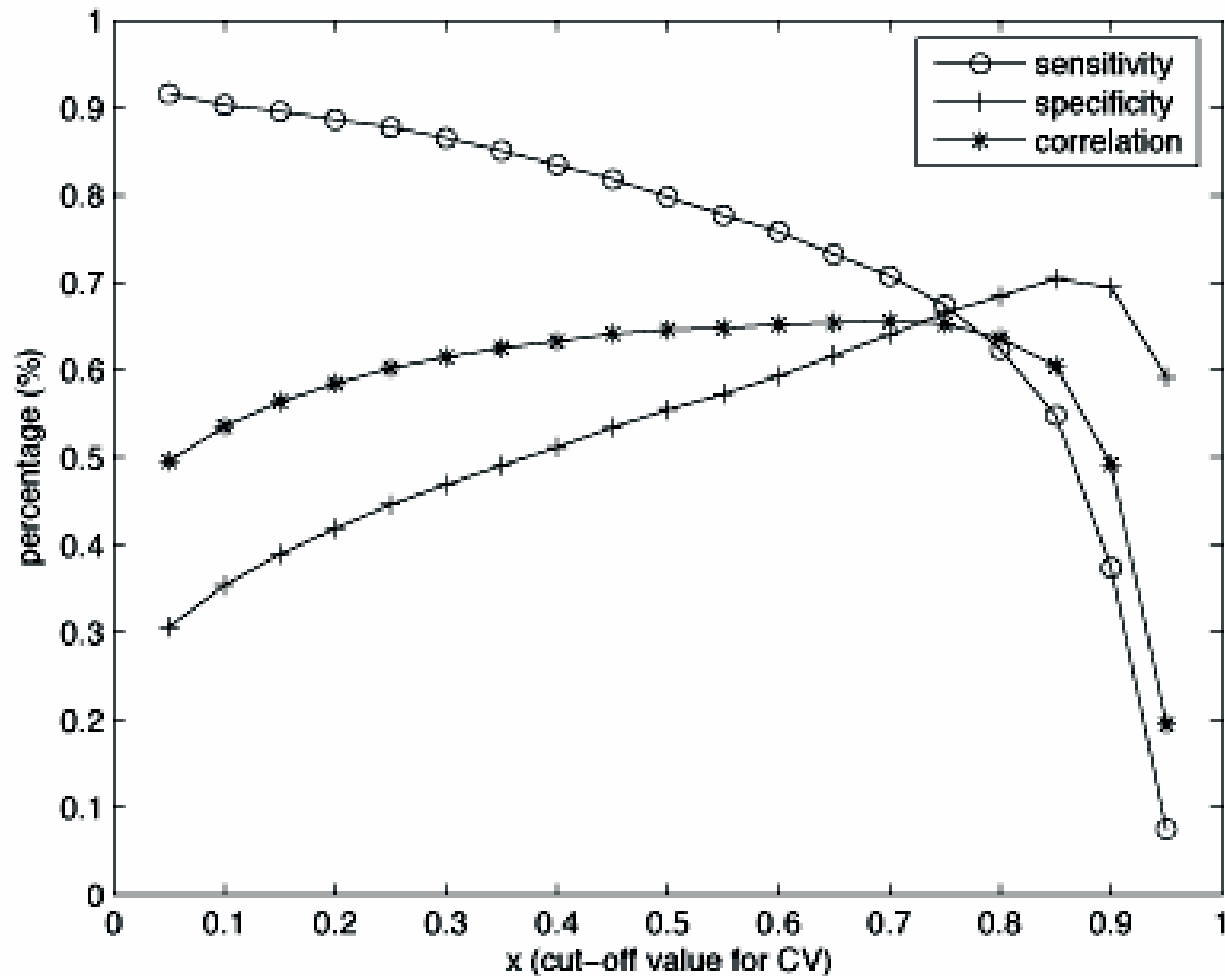
# Model I Example - Overpredictions



# 5-Fold Cross Validation

- Randomly divided dataset into training and test datasets with a 4:1 ratio
- Trained model with 642 sequences and tested with 160 sequences
- Repeated five times
- Sensitivity and specificity were computed on a per residue basis

# Sensitivity, Specificity, Matthews Correlation Coefficient



# Sn and Sp - Residue vs. Region

- On a per residue basis (CV cutoff probability = .75):
  - ◆ Sn and Sp = 67%,
  - ◆ Matthews correlation coefficient was 65%
- On a per linker region basis:
  - ◆ Criteria: length >4 residues with probability > .5 and maximum probability > .8
  - ◆ Sn = 63%, Sp = 93%

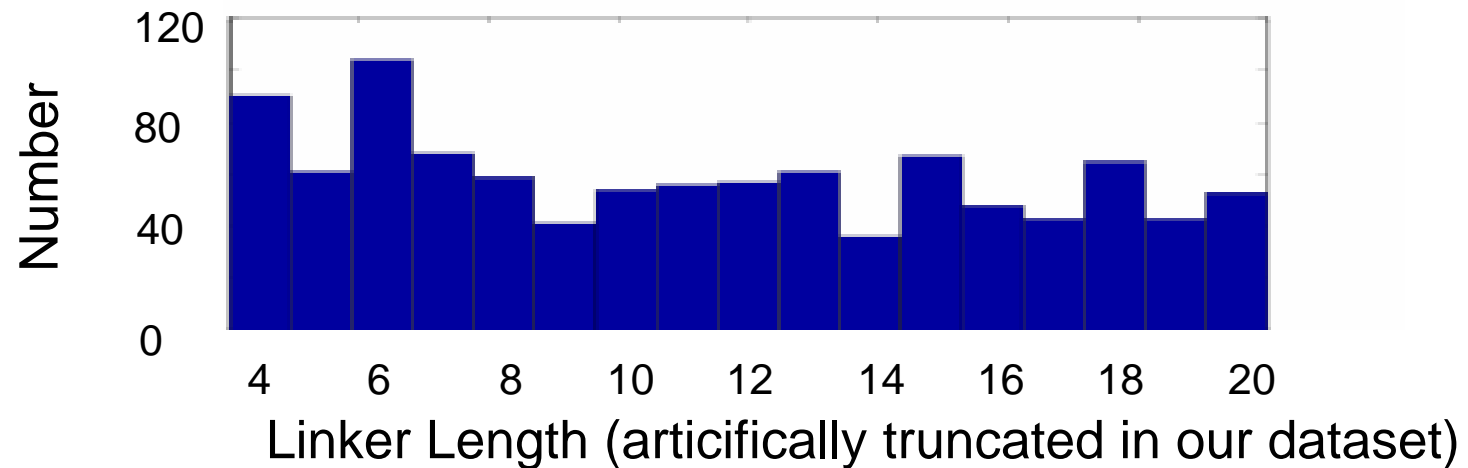
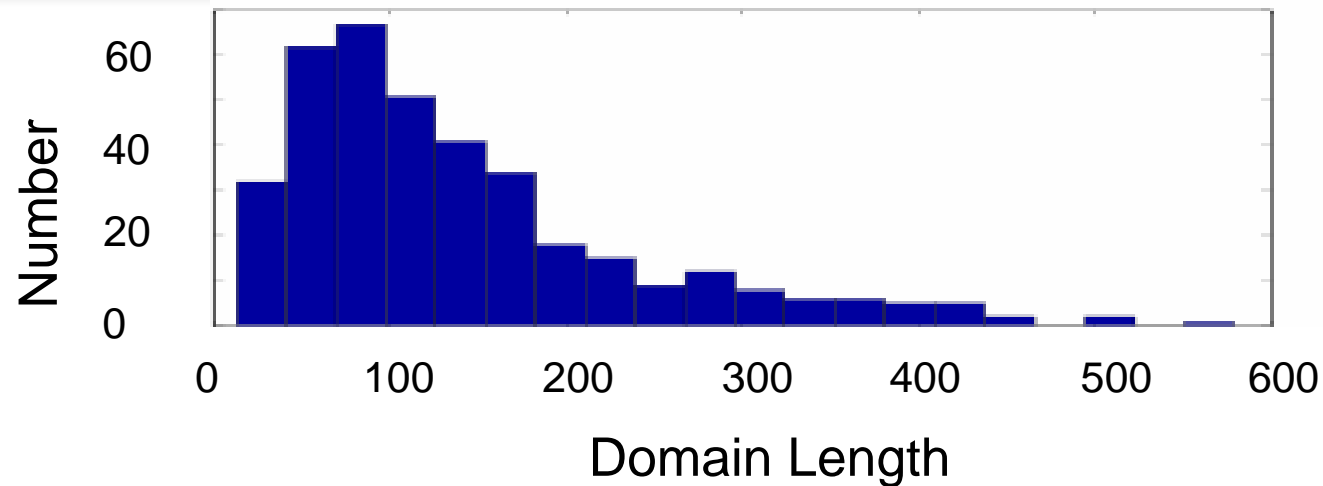
# Implications of Evolutionary Domain Dataset

- A notorious problem in structural linker prediction is distinguishing linkers from intradomain loops.
- Our method appears to perform well in this regard.
- However, the relatively low number of false positives may be due to the evolutionary domain dataset
  - ◆ non-conserved intradomain loops can cause single domains to be annotated as multiple domains in PfamA

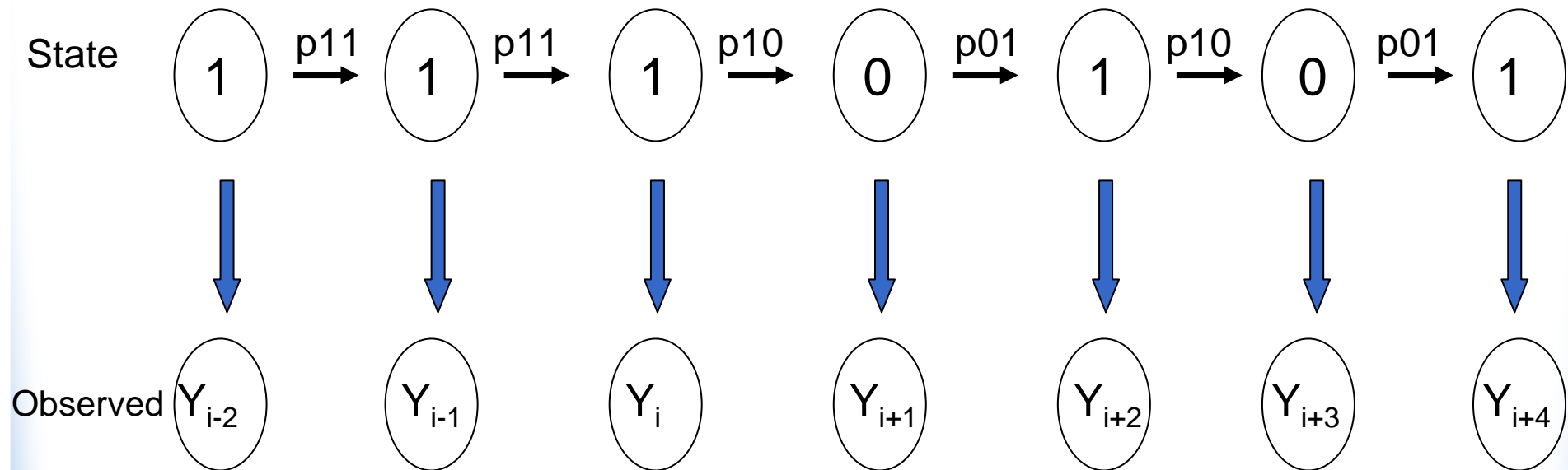
# Model II

- To take advantage of domain length distribution - two possibilities:
- Variable Duration HMM (VDHMM)
- Nonstationary HMM (NSHMM) - our approach

# Length Distributions



# Review - Representation of Simple HMM

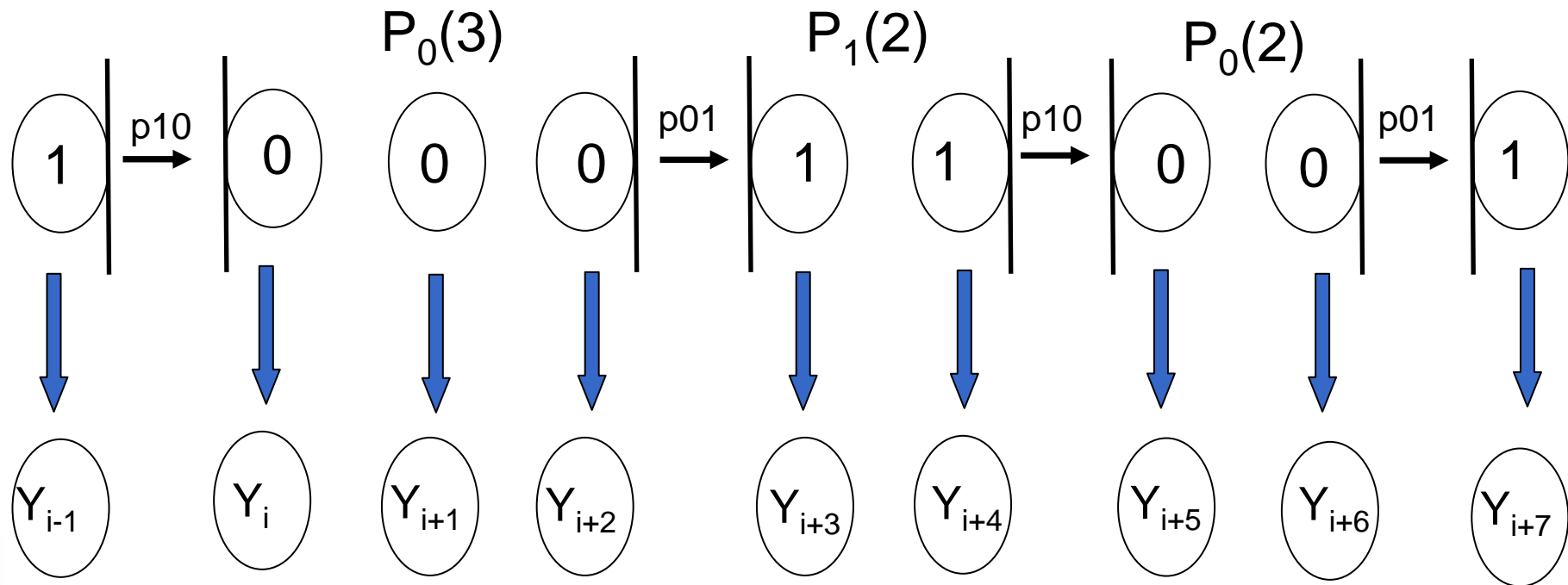


For our model:

States are linker (0) or non-linker (1).

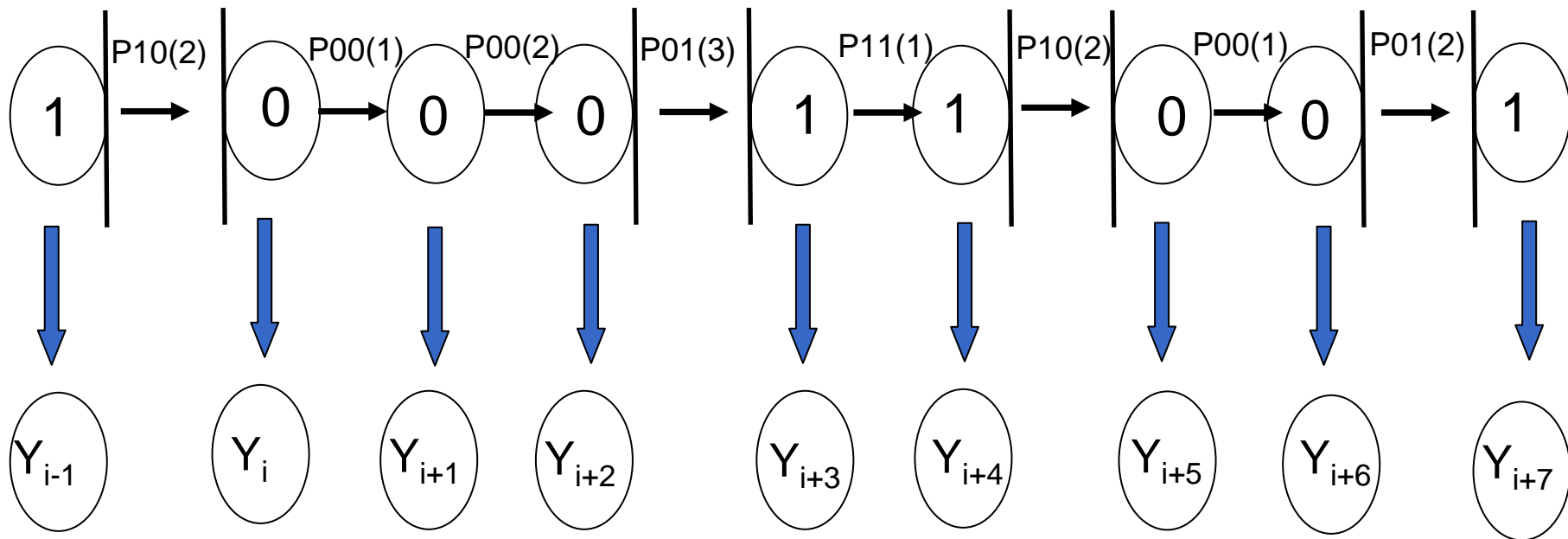
Observations are smoothed linker index ( $y_i$ ).

# Representation of VDHMM



The duration of each state is also modeled.  
Also known Generalized HMM or Segmental HMM.

# Representation of NSHMM

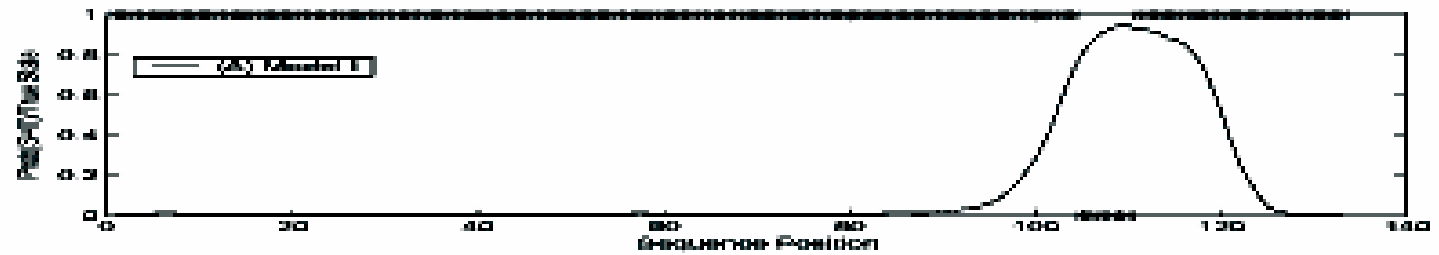


The state transition probabilities  $p_{ij}$  are modeled state duration probabilities  $p_{ij}(d)$ , where  $d$  is state duration. If  $d=1$  this becomes a conventional HMM. If  $d$  is constant for each  $p_{ij}$  where  $i \neq j$ , this becomes a VDHMM.

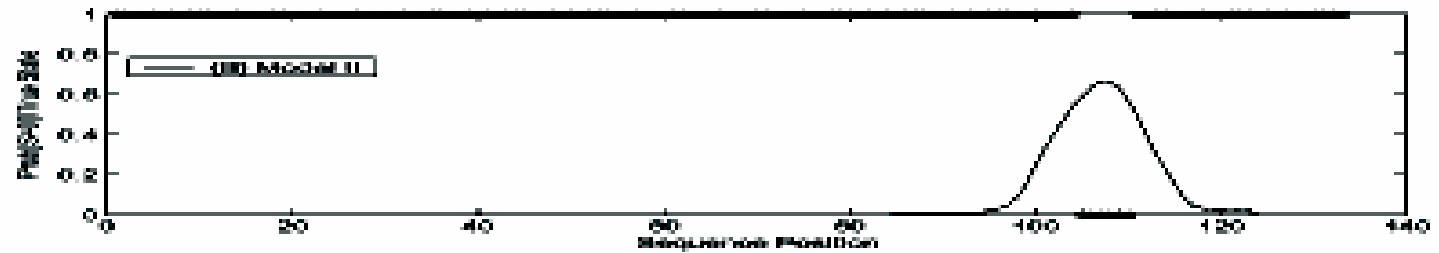
# Model I vs. Model II

Q7UD15

Model I



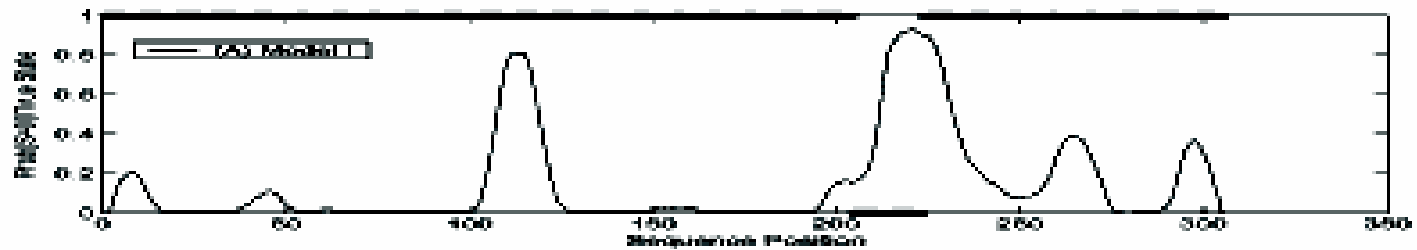
Model II



# Model I vs. Model II

Q89F20

Model I



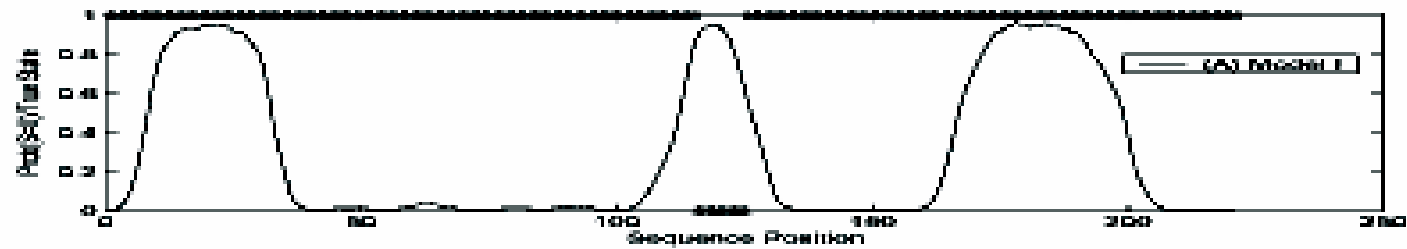
Model II



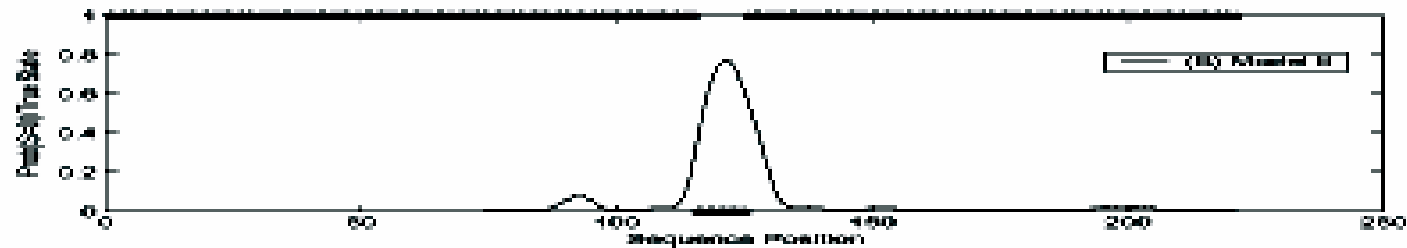
# Model I vs. Model II

Q7P6J3

Model I



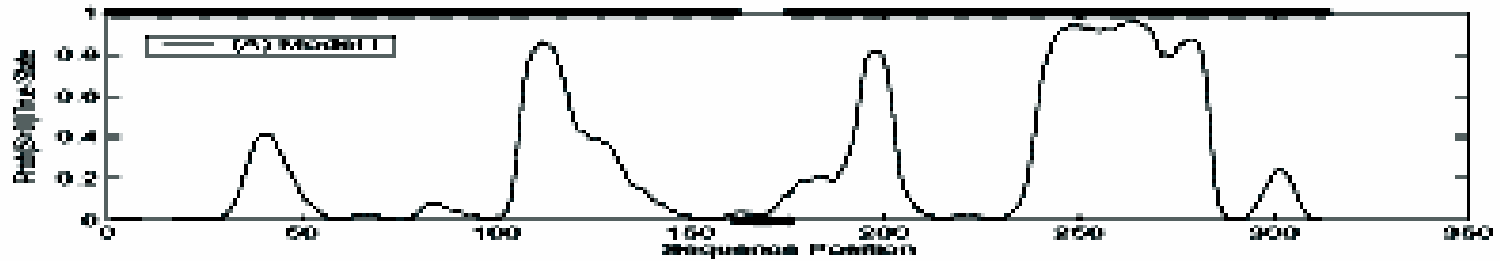
Model II



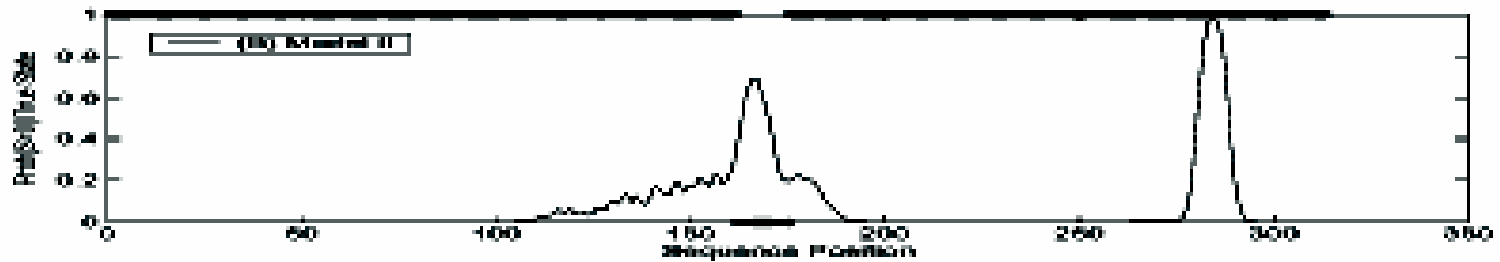
# Model I vs. Model II

Q81JL7

Model I



Model II



# Model I vs. Model II

- At CV cutoff .5:  
Residues based evaluation

	Sn	Sp	Matthews C
Model I	81.4	59.6	67.7
Model II	77.3	71.6	69.6

# Model I vs. Model II

Region based evaluation

	Sn	Sp	
Model I	63	93	
Model II	62	97	

# Summary

- Our methods perform well at defining evolutionary domains.
- We predict linker regions as well as their boundaries because each residue is associated with a probability.
- A NSHMM which models domain length is more sensitive than a conventional HMM.

# Current and Future

- Closer examination of window length effect
- Test using structural dataset
- Development into software package with parallelized version

# Acknowledgements

- Kyoungghwa Bae
- Bani Mallick - Department of Statistics, TAMU
- Funding: Texas Agricultural Experiment Station, Texas A&M University