Statistical Machine Learning Methods for Biomedical Informatics

II. Hidden Markov Model for Biological Sequences

Jianlin Cheng, PhD

William and Nancy Thompson Missouri Distinguished Professor

Department of Electrical Engineering & Computer Science

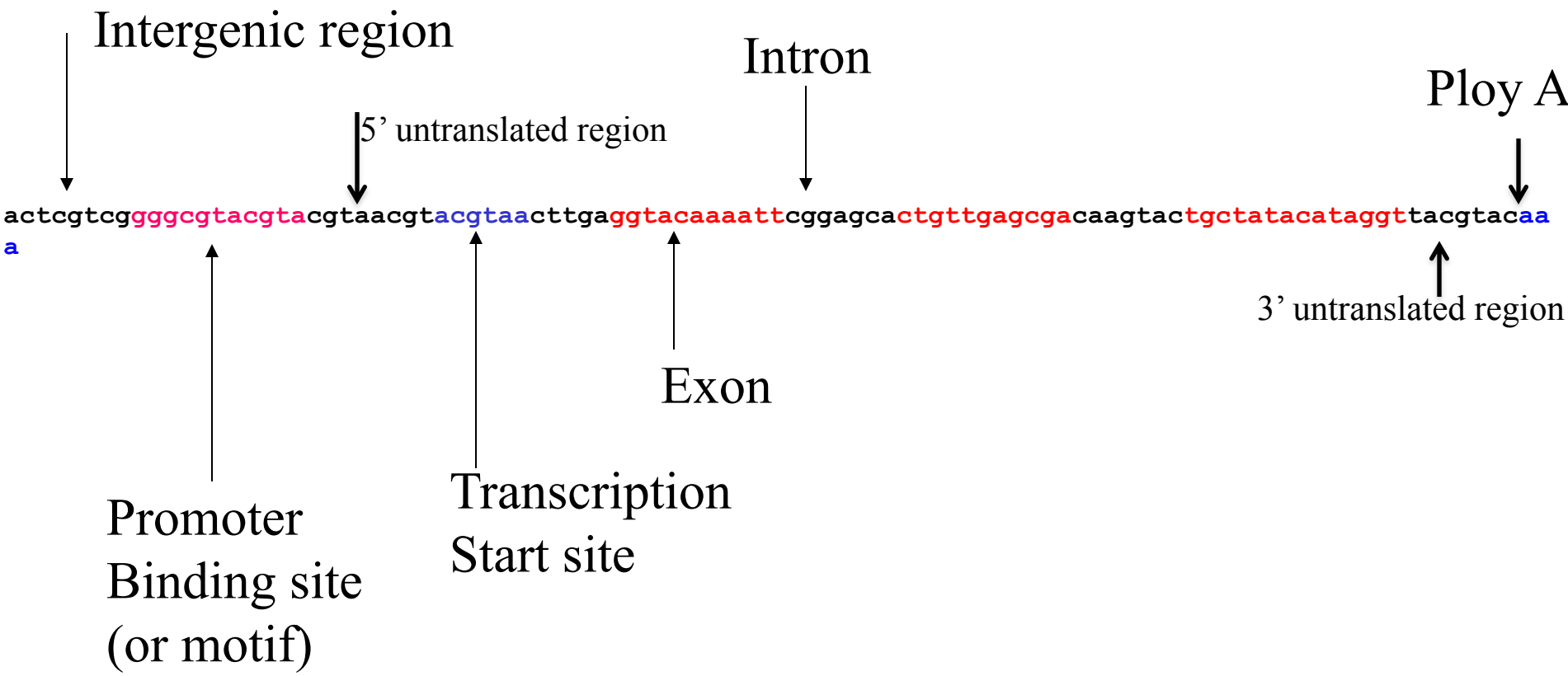
University of Missouri

Application of HMM in Biological Sequence Analysis

- Gene prediction
- **Protein sequence modeling (learning, profile)**
- **Protein sequence alignment (decoding)**
- **Protein database search (scoring, e.g. fold recognition)**
- Protein structure prediction
- …

Genome

Motif and Gene Structure

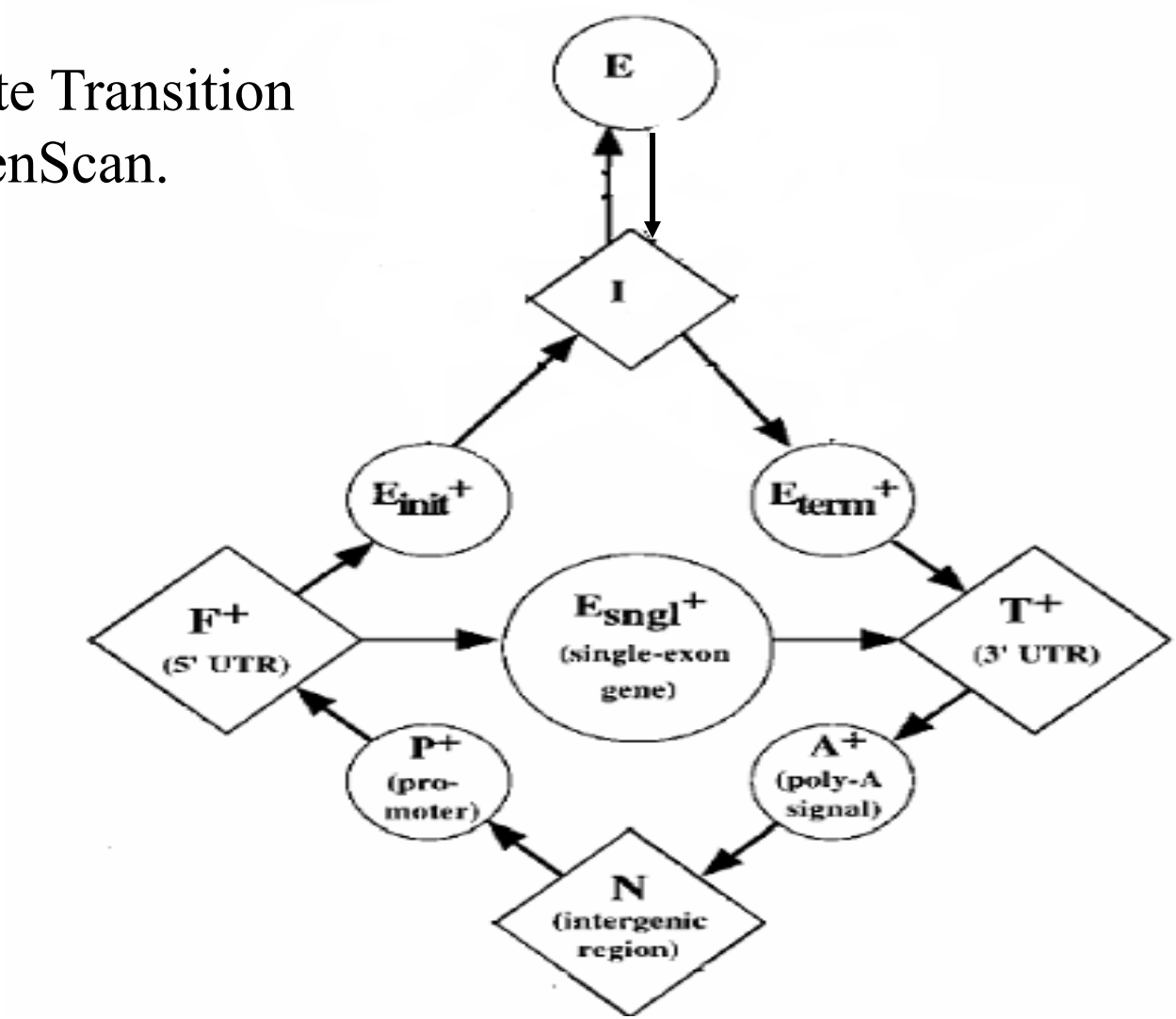


HMM has been used for modeling binding site and gene structure prediction.

GENSCAN

(genes.mit.edu/GENSCAN.html)

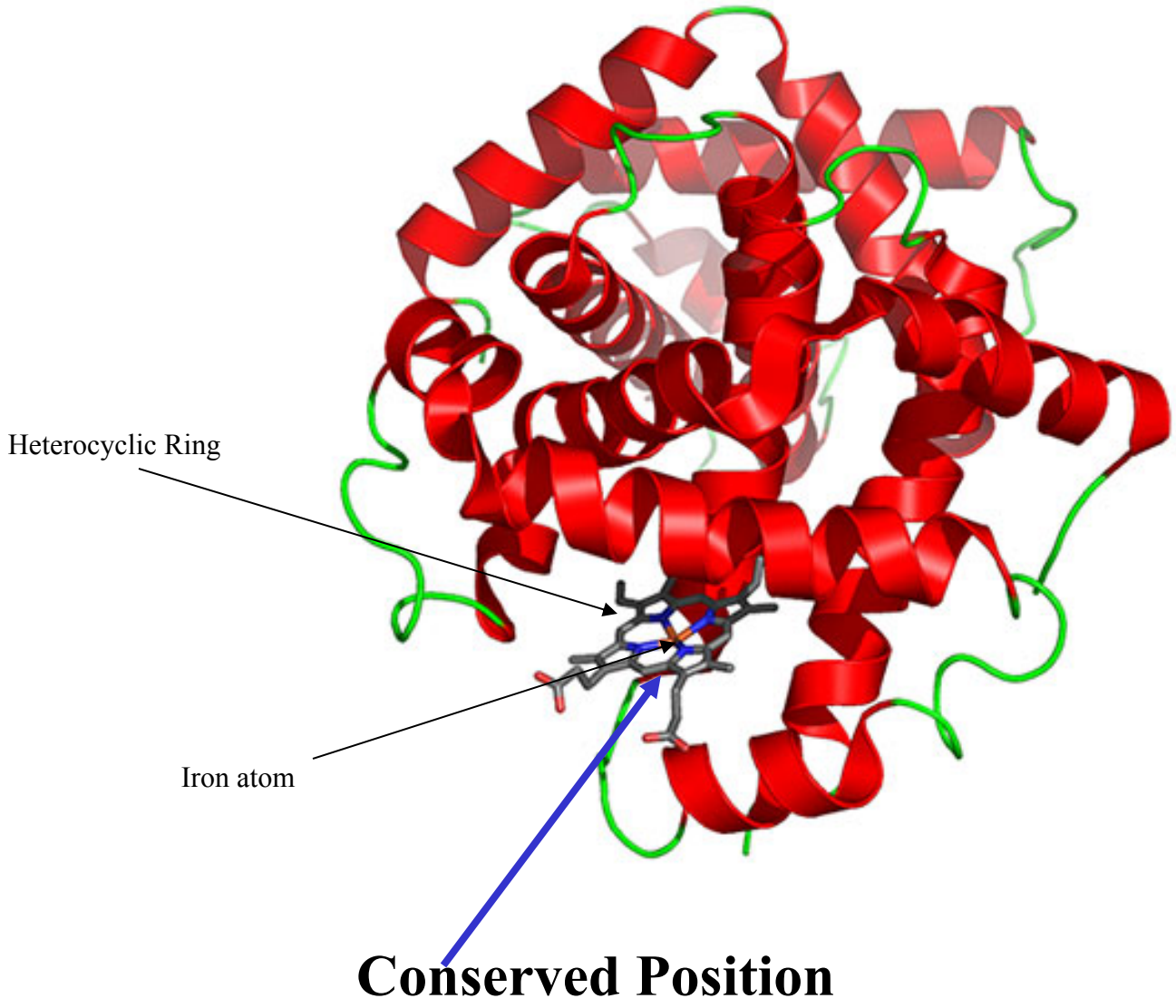
Simplified State Transition Diagram of GenScan.



Model Protein Family (Profile HMM)

- Create a statistical model (HMM) for a group of related protein sequences (e.g protein family)
- Identify core (conserved) elements of homologous sequences
- Positional evolutionary information (e.g. insertion and deletion)

Example: Hemoglobin Transports Oxygen



Why do We Build a Profile (Model)?

- Understand the conservation (core function and structure elements) and variation
- Sequence generation
- Multiple sequence alignments
- Profile-sequence alignment (more sensitive than sequence-sequence alignment)
- Family / fold recognition
- Profile-profile alignment

Protein Family

```
seq1 VRRNNMGMPLIESSSYHDALFTLGYAGDRISQMLGMRLLAQGRLSEMAGADALDV
seq2 NIYIDSNGIAHIYANNLHDLFLAEGYYEASQRLFEIELFGLAMGNLSSWVGAKALSS
seq3 SAETYRDAWGIPHLRADTPHELARAQGTARDRAWQLEVERHRAQGTSASFLGPEALSW
seq4 DRLGVVTIDAANQLDAMRALGYAQERYFEMDLMRRAPAGELSELFGAKAVDL
```

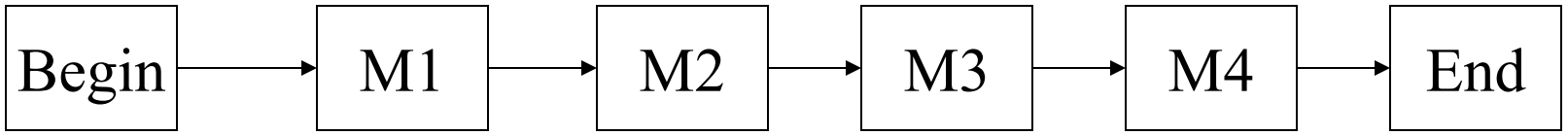
```
seq1 ---VRRNNMGMPLIESSSYHDALFTLGY--AGDRISQMLGMRLLAQGRLSEMAGADALDV
seq2 --NIYIDSNGIAHIYANNLHDLFLAEGYYEASQRLFEIELFG-LAMGNLSSWVGAKALSS
seq3 SAETYRDAWGIPHLRADTPHELARAQGT--ARDRAWQLEVERHRAQGTSASFLGPEALSW
seq4 -----DRLGVVTIDAANQLDAMRALGY--AQERYFEMDLMRRAPAGELSELFGAKAVDL
```

Imagine these sequences evolve from a single ancestral sequence and undergo evolutionary mutations. How to use a HMM to model?

Key to Build a HMM is to Set Up States

- Think about the positions of the ancestral sequence is undergoing mutation events to generate new sequences in difference species. A position can be modeled by a **dice**.
- **Match** (match or mutate): the position is kept with or without variations / mutations.
- **Delete**: the position is deleted
- **Insert**: amino acids are inserted between two positions.

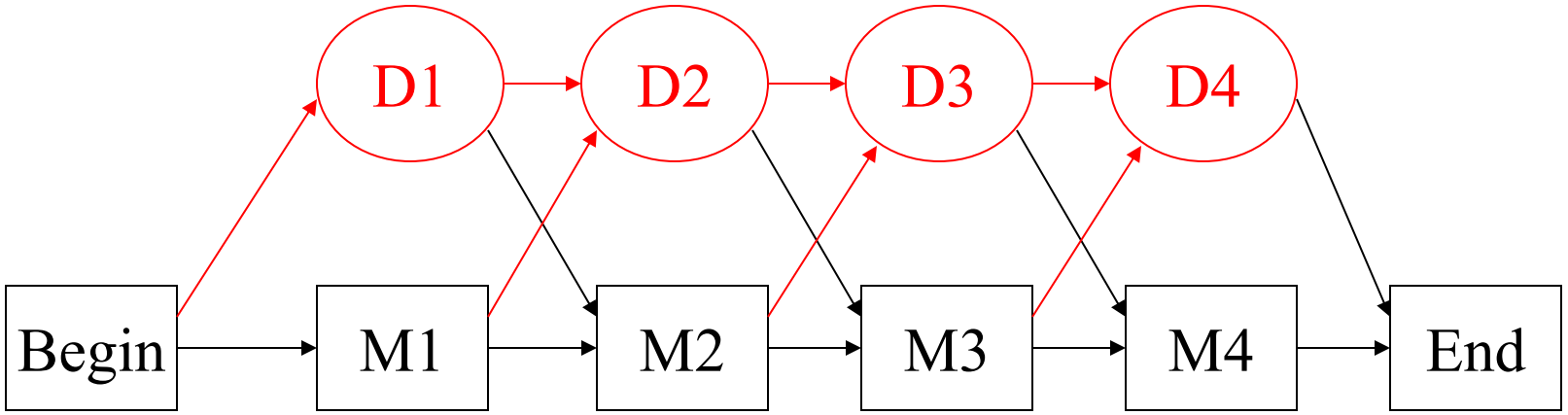
Hidden Markov Model



Each match state has an emission distribution of 20 amino acids; one match state for a position.

Krogh et al. 1994, Baldi et al. 1994

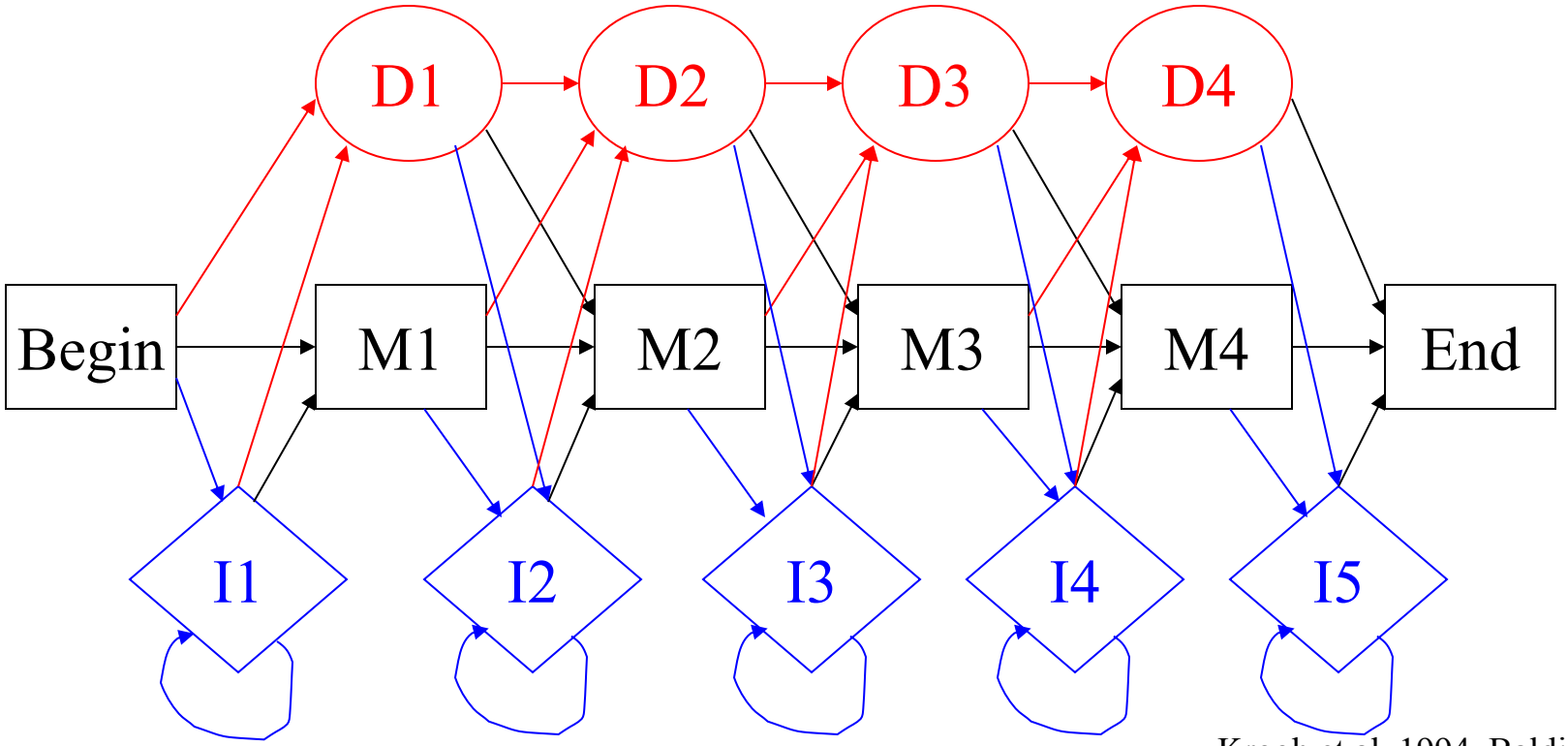
Hidden Markov Model



Each match state has an emission distribution of 20 amino acids.
Deletion state is a mute state (emitting a dummy)

Krogh et al. 1994, Baldi et al. 1994

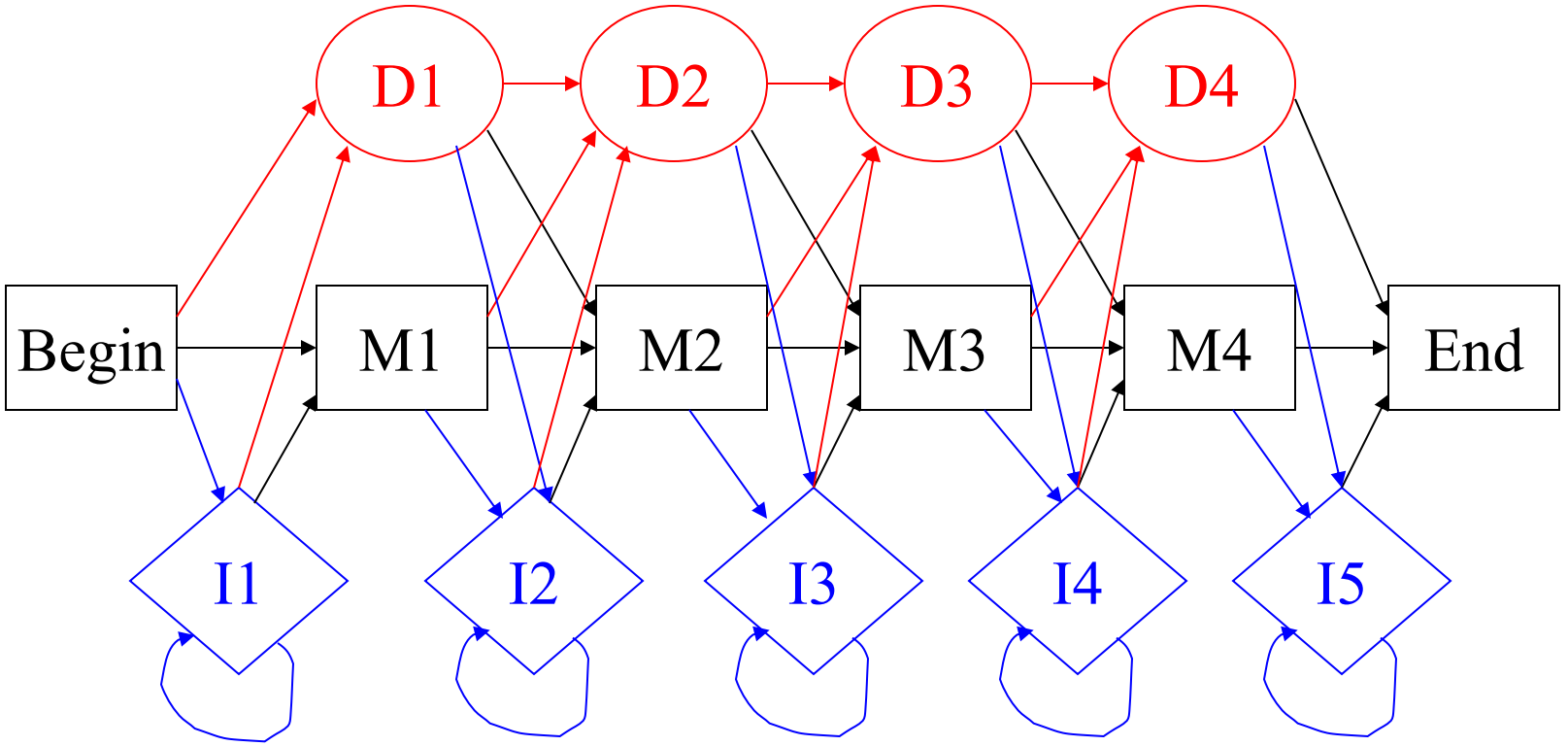
Hidden Markov Model



Krogh et al. 1994, Baldi et al. 1994

Each match state has an emission distribution of 20 amino acids.
Each insertion state has an emission distribution of 20 amino acids.
Variants of architecture exist. (see Eddy, bioinformatics, 1997)

Hidden Markov Model

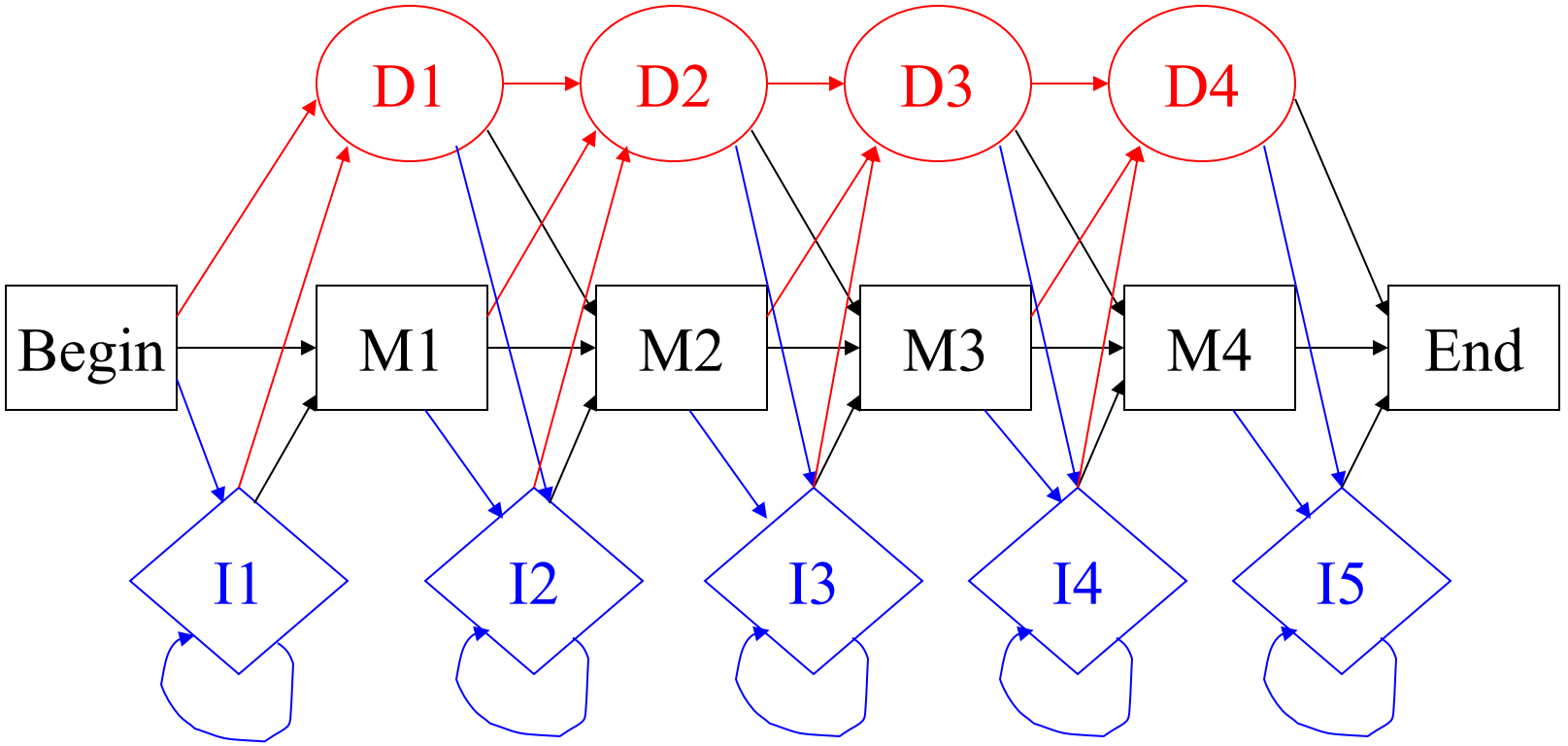


How many states? (M positions: length of model)

M (match) + M (deletion) + (M+1) (insertion) + 2 = 3M + 3

Krogh et al. 1994, Baldi et al. 1994

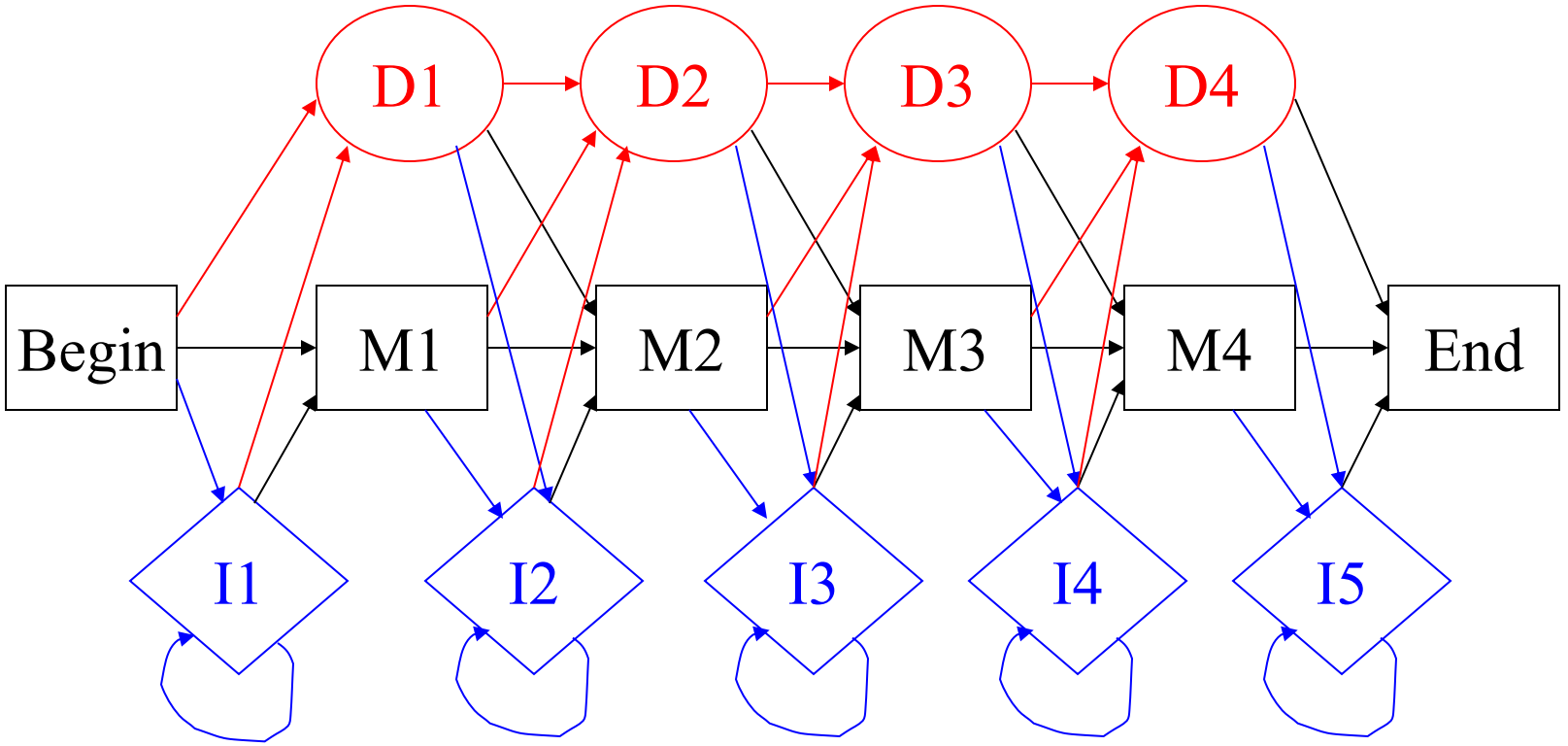
Hidden Markov Model



How many transitions? (M positions: length of model)
Deletion: 3M – 1, Match: 3M – 1, Insertion: 3(M+1) – 1, B/E: 3
Total = 9M +3.

Krogh et al. 1994, Baldi et al. 1994

Hidden Markov Model



How many emissions? (M positions: length of model)
M * 20 (match) + (M+1)*20 (insertion) = 40M + 20

Krogh et al. 1994, Baldi et al. 1994

Initialization of HMM

- How to decide model length (the number of match states)?
- How to initialize transition probabilities?
- How to initialize emission probabilities?

How to Decide Model Length?

- **Learn**: Use a range of model length (centered at the average sequence length). If transition probability from a match (M_i) state to a delete state (D_{i+1}) > 0.5, remove the M_{i+1}. If transition probability from a match (M_i) state to an insertion state (I_{i+1}) > 0.5, add a match state.
- **Get from multiple alignment**: assign a match state to any column with <50% gaps.

How to Initialize Parameters?

- Uniform initialization of transition probabilities is ok in most cases.
- Uniform initialization of emission probability of insert state is ok in many cases.
- Uniform initialization of emission probability of match state is **bad**. (lead to bad local minima)
- Using amino acid distribution to initialize the emission probabilities is better. (need regularization / smoothing to avoid zero)

Initialize from Multiple Alignments

```
seq1 ---VRRNNMGMPLIESSSYHDALFTLGY--AGDRISQMLGMRLLAQGRLSEMAGADALDV
seq2 --NIYIDSNGIAHIYANNLHDLFLAEGYYEASQRLFEIELFG-LAMGNLSSWVGAKALSS
seq3 SAETYRDAWGIPHLRADTPHELARAQGT--ARDRAWQLEVERHRAQGTSASFLGPEALSW
seq4 -----DRLGVVTIDAANQLDAMRALGY--AQERYFEMDLMRRAPAGELSELFGAKAVDL
```

First, assign match / main states, delete states, insert states from MSA

Get the path of each sequence

Count the amino acid frequencies emitted from match or insert states, which are converted into probabilities for each state (need smoothing/ regularization / pseudo-count).

Count the number of state transitions and use them to initialize transition probabilities.

Estimate Parameters (Learning)

- We want to find a set of parameters to maximize the probability of the observed sequences in the family: maximum likelihood: P(sequences | model) = P(sequence 1| model) * … * P(sequence n | model).
- Baum-Welch's algorithm (or EM algorithm) (see my previous lectures about HMM theory)

Demo of HMMER (learning)

DOWNLOAD DOCUMENTATION SEARCH PUBLICATIONS BLOG

HMMER: biosequence analysis using profile hidden Markov models

Get the latest version

v3.1b2

Download (MacOSX / Intel)

Alternative Download Options

HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query *sequences*, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with **phmmer**, or do an iterative search with **jackhmmer**.

HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST.

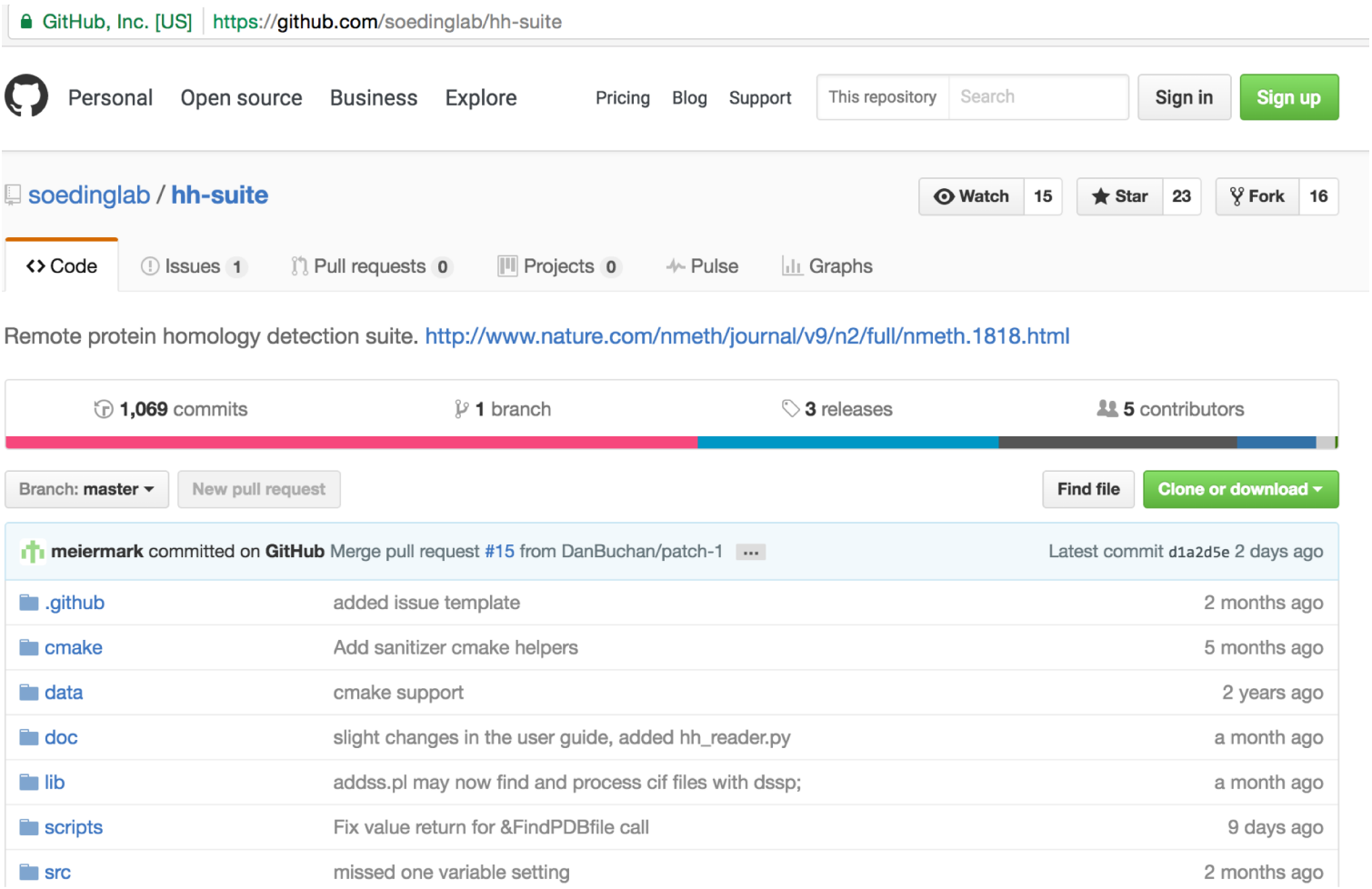
HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via new search servers at the European Bioinformatics Institute.

http://hmmer.org/

Demo

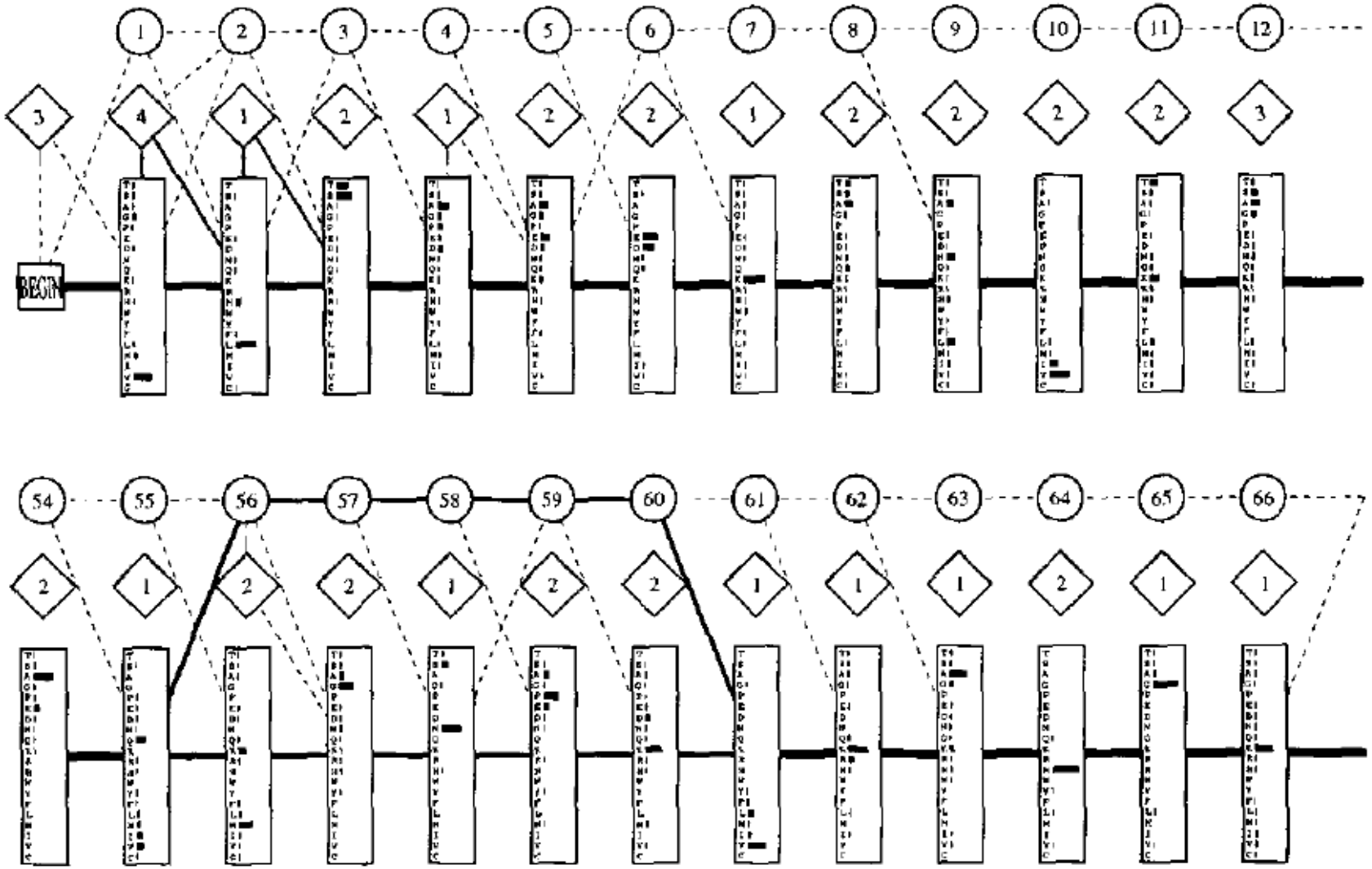
- /Users/jianlincheng/Desktop/work_2018/Teaching_2018/machine_learning_bioinfo_2018/hmmer-3.0/demo
- **Build a HMM**

HHSuite



Remmert M, Biegert A, Hauser A, Söding J (2011). "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment.". Nat. Methods. 9 (2): 173–175

Visualization of Features and Structure in HMM



Myoglobin protein family. How to interpret it?

Krogh et al. 94

Demo of HMMEditor

J. Dai and J. Cheng. **HMMEditor: a visual editing tool for profile hidden Markov model**. *BMC Genomics*, 2007.

Support HMM models built by HMMer 2.0 - 3.0.

Paper: http://www.biomedcentral.com/qc/1471-2164/9/S1/S8

Demo

- /Users/jianlincheng/Desktop/work_2018/Teaching_2018/machine_learning_bioinfo_2018/HMMVE_1.2

- Transition probabilities and emission probabilities

Protein Family Profile HMM Databases

- **Pfam database** (protein family database) (https://pfam.xfam.org)
- Pfam 31.0 contains a total of 16712 families

What Can We Do With the HMM?

- **Recognition and classification. Widely used for database search**: does a new sequence belong to the family? (database search)
- **Idea**: The sequences belonging to the family (or generated from HMM) should receive higher probability than the sequence not belong to the family (unrelated sequences).

Two Ways to Search

- Build a HMM for each family in the database. Search a query sequence against the database of HMMs. (Pfam)
- Build a HMM for a query family, and search HMM against of a database of sequences

Compute P(Sequence | HMM)

- Forward algorithm to compute P(sequence | model)
- We work on: -log(P(sequence|M): distance from the sequence to the model. (negative log likelihood score)
- Unfortunately, -log(P(sequence|M) is length dependent. So what can we do?

Normalize the Score into Z-score

- Search the profile against a large database such as Swiss-Prot
- Plot –log(P(seuqence|model), NULL scores, against sequence length.

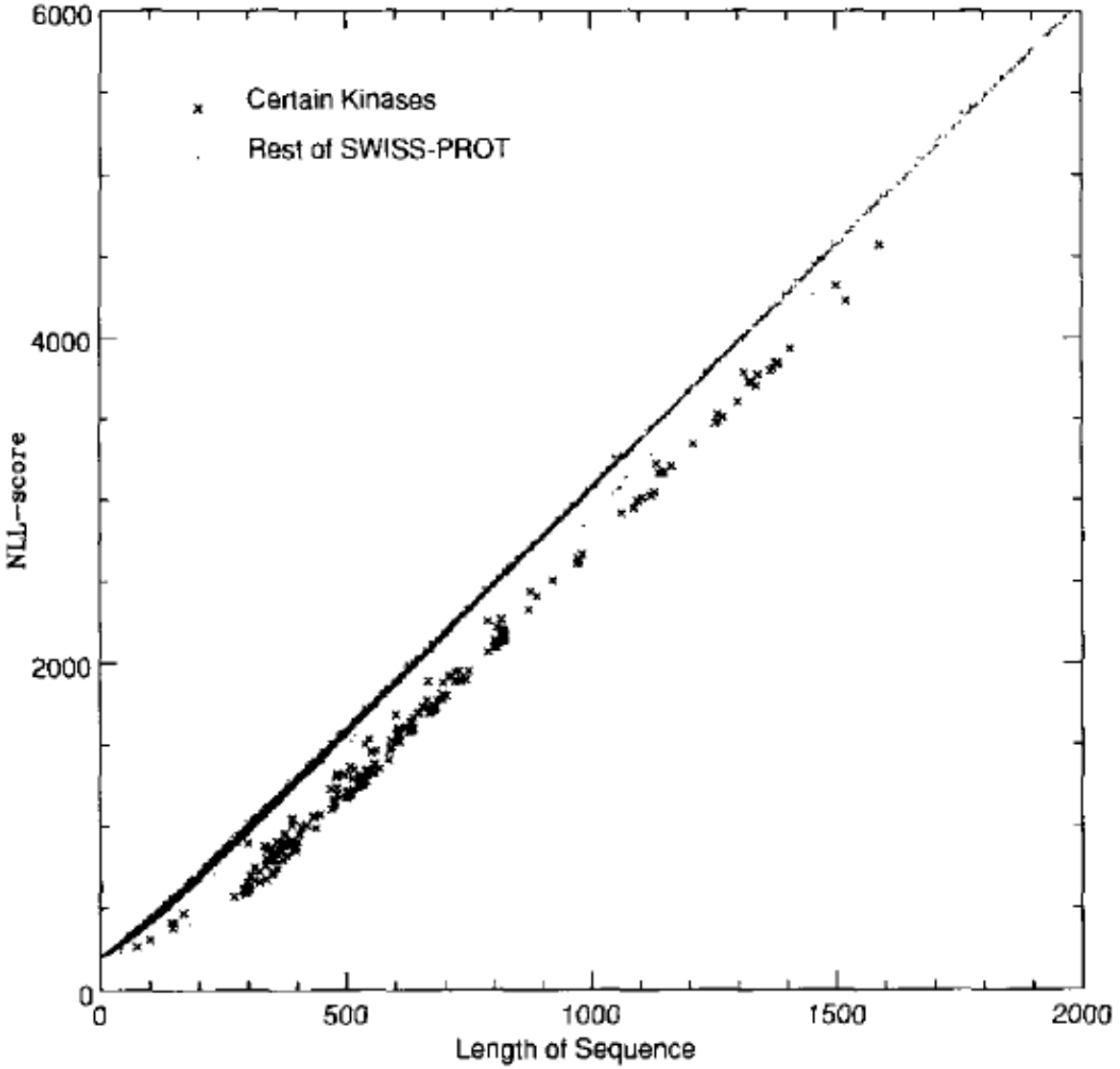


Figure 9. Scatter plot of NLL-score versus length for sequences in SWISS-PROT using the Kinase HMM.

NLL score is linear to sequence length.

NLL scores of the same family is lower than un-related sequences

We need normalization.

Krogh et al. 94

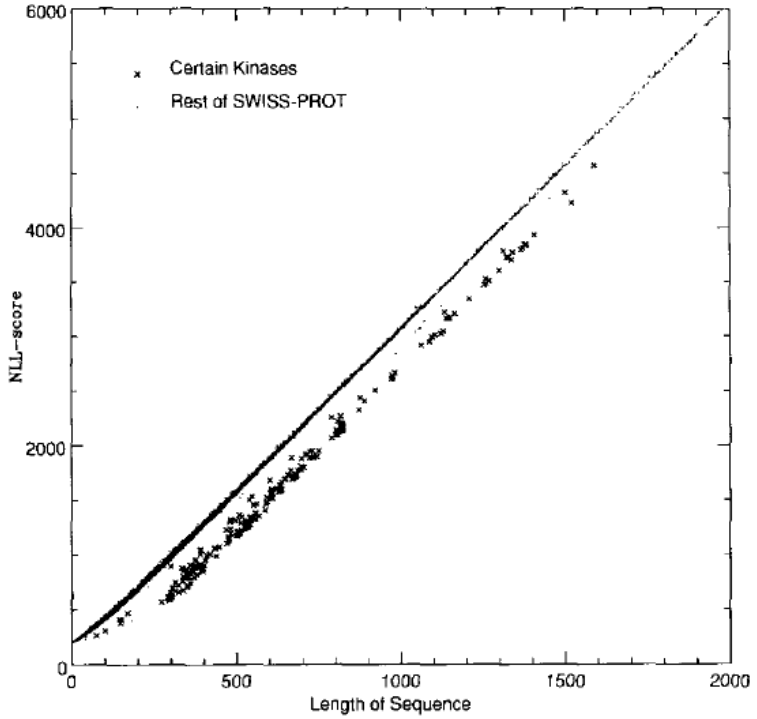


Figure 9. Scatter plot of NLL-score versus length for sequences in SWISS-PROT using the Kinase HMM.

NULL model of unrelated sequences:

Length	Mean (u)	Std (σ)
100	500	5
101	550	6

Compute Z-score: $|s - u| / \sigma$

$Z > 4$: the sequence is very different from unrelated sequence.
(for non-database search, a randomization can work.)

Extreme Value Distribution
(Karlin and Altschul)

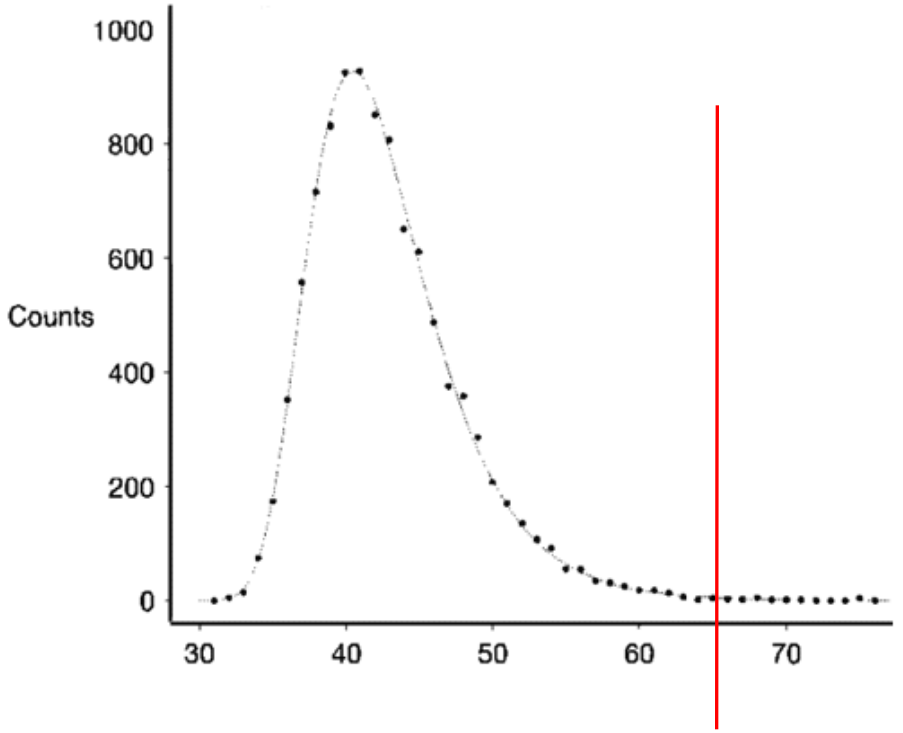
http://www.people.virginia.edu/~wrp/cshl02/Altschul/Altschul-3.html

Log-odds score = log(P(seq|λ) / P(seq |null))

$$P(S \geq x) = 1 - \exp(-Kmn\, e^{-\lambda x})$$

P-value

E-value



K and lamda are statistical parameters. m,n model and sequence length.

HMMer search demo

/Users/jianlincheng/Desktop/work_2018/Teaching_2018/machine_learning_bioinfo_2018/hmmer-3.0/demo

Insight I: Evaluating a Sequence Against a Profile HMM is Fold Recognition Process or Function Prediction

- Check if a sequence is in the same family (or superfamily) as the protein family (superfamily) used to build the profile HMM.
- If they are in the same family, they will share the similar protein structure (fold), possibly protein function.
- The known structure can be used to model the structure of the proteins without known structure.
- The known function can be used to predict function.

Insight II: Evaluating a Sequence Against a HMM is Sequence-Profile Alignment

- Align a query sequence against a HMM of the target sequence to get the most likely path (Viterbi algorithm) (or vice versa)
- Match the path of the query sequence with the path of the target sequence, we get their alignment.
- Represented work: HMMER.

Pairwise Alignment via HMM

Seq 1: A T G R K E
Path : M_1 I_1 I_1 M_2 D_3 M_4 I_4

Seq 2: V C K E R P
Path : M_1 I_1 M_2 M_3 M_4 I_4

Path: M_1 M_2 M_3 M_4
Seq 1: A T G R - K E
Seq 2: V C - K E R P

HMM for Multiple Sequence Alignment

- Build a HMM for a group of sequences
- Align each sequence against HMM using Viterbi algorithm to find the most likely path. (dynamic programming)
- Match the main/match states of these paths together.
- Add gaps for delete states
- For insertion between two positions, use the longest insertion of a sequence as template. Add gaps to other sequence if necessary. (see Krogh's paper)

Demo for Multiple Sequence Alignment Using HMMEditor (based on hmmer2 models)

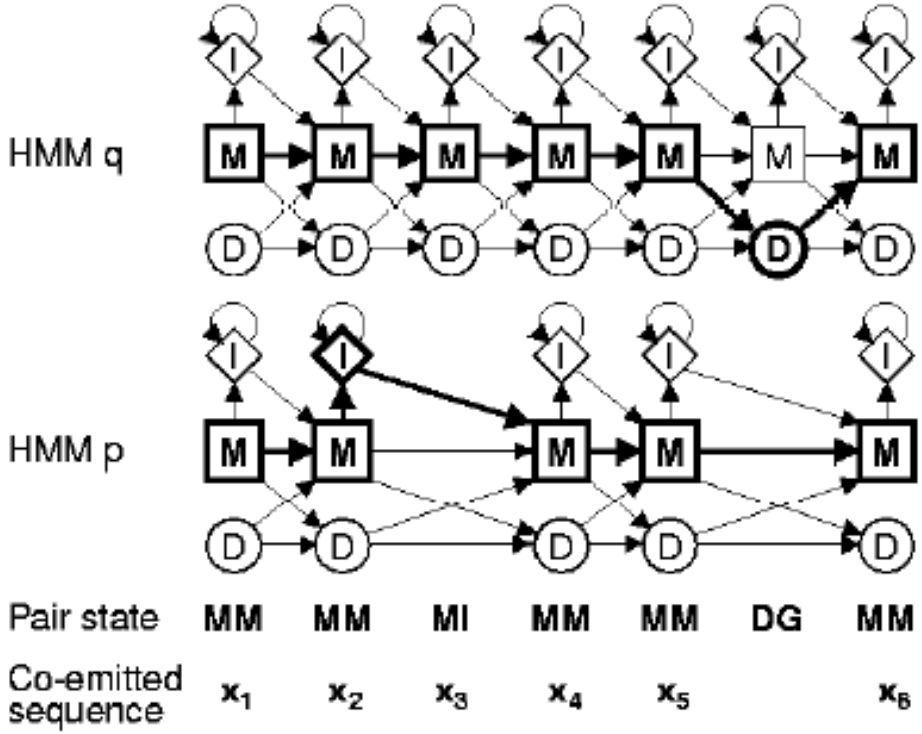
How About Evaluating the Similarity between HMMs?

- Can we evaluate the similarity of two HMMs?

- Can we align two profile HMMs? (profile-profile alignment). Compare HMM with HMM.

J. Soeding, Bioinformatics, 2005

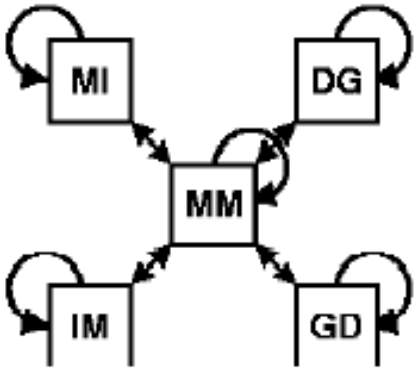
HMM-HMM Comparison
Profile-Profile Alignment



Generalized Sequence (of States) 1

Generalized Sequence (of States) 2

Allowed pair state transitions



Goal: find a sequence of paired states with Maximum sum of log odds scores

J. Soeding, Bioinformatics, 2005

Sequence Weighting

- **Henikoff-Henikoff**: sum of position-based weight. For each position, each type of amino acid is assigned weight 1. The weight is of each amino acid is 1 / frequency of the amino acid at the position. The weight of a sequence is the sum of positional weights. (gap is not counted. A position with more than 50% gaps may be removed from counting.) An easy, useful algorithm
- **Tree algorithm**: construct a phylogenetic tree. Start from root, weight 1 flows down. At any branch, the weight is cut to half.

Null Model

- Background null model (log-odds)
- Reverse null model (SAM)
- Sum of log-odds score of local alignment obeys extreme-value distribution (same as **PSI-BLAST**): good for estimating the significance of sequence-HMM match.
- Sum of log-odds score of global alignment is length dependent.

HMM Software and Code

- HMMER: http://hmmer.org/
- SAM: http://www.cse.ucsc.edu/research/combio/sam.html
- HHSearch: http://toolkit.tuebingen.mpg.de
- MUSCLE: http://www.drive5.com/muscle/
- HHSuite: https://github.com/soedinglab/hh-suite
- HHblits

Remmert M, Biegert A, Hauser A, Söding J (2011). "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment.". Nat. Methods. 9 (2): 173–175

Project 1: Multiple Sequence Alignment Using HMM

- Dataset: BaliBASE: *http://lbgi.fr/balibase/*
- Generate multiple alignment using Clustalw (http://www.ch.embnet.org/software/ClustalW.html) for a family of sequences
- Design and implement your HMM (you may refer to open source HMM code)
- Construct a HMM for a family of sequences (initialization, number of states)
- Estimate the parameters of HMM using the sequences
- Analyze the emission probability of states (visualization)
- Analyze the transition probability between states (visualization)
- Compute the probability of each sequence
- Generate multiple sequence alignments and compare it with the initial multiple alignment

Reference: A. Krogh et al, JMB, 1994 and open source HMM.
J. D. Thompson, F. Plewniak and O. Poch. Bioinformatics, 1999