Protein structure prediction with statistical alignment

Bayesian Phylogeny Workshop
25-29 June, 2008, Renyi Institute

Estimating the Rate of Evolution of the Rate of Molecular Evolution

Jeffrey L. Thorne, Hirohisa Kishino,† and Ian S. Painter**

Mol. Biol. Evol. 15(12):1647–1657. 1998

BMC Bioinformatics

IMPACT FACTOR 3.62

Methodology article

Open Access

How reliably can we predict the reliability of protein structure predictions?

Istvan Miklos ✉, **Adam Novak** ✉, **Balazs Dombai** ✉ and **Jotun Hein** ✉

BMC Bioinformatics 2008, **9**:137 doi:10.1186/1471-2105-9-137

Published: 3 March 2008

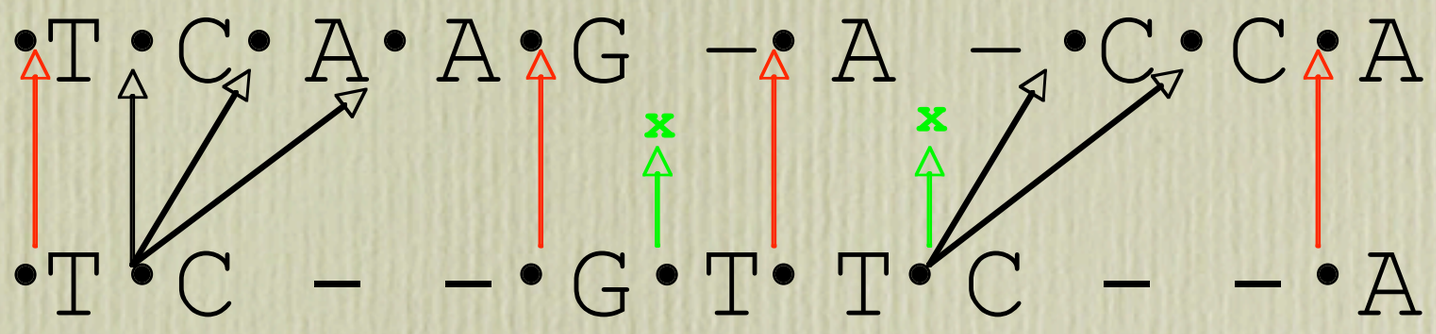
Overview

- Stochastic sequence alignment
- Pair and multiple HMMs
- Markov chain Monte Carlo
- Homstrad database
- Results

A TKF91 model

- Explicit model for insertions and deletions
- Time-continuous Markov model (birth-death model)
- Transition probabilities can be calculated analytically
- Likelihood calculation by dynamic programming
- Maximum Likelihood estimation of parameters

A TKF91 model



The probability of an alignment is the product of probabilities of several patterns

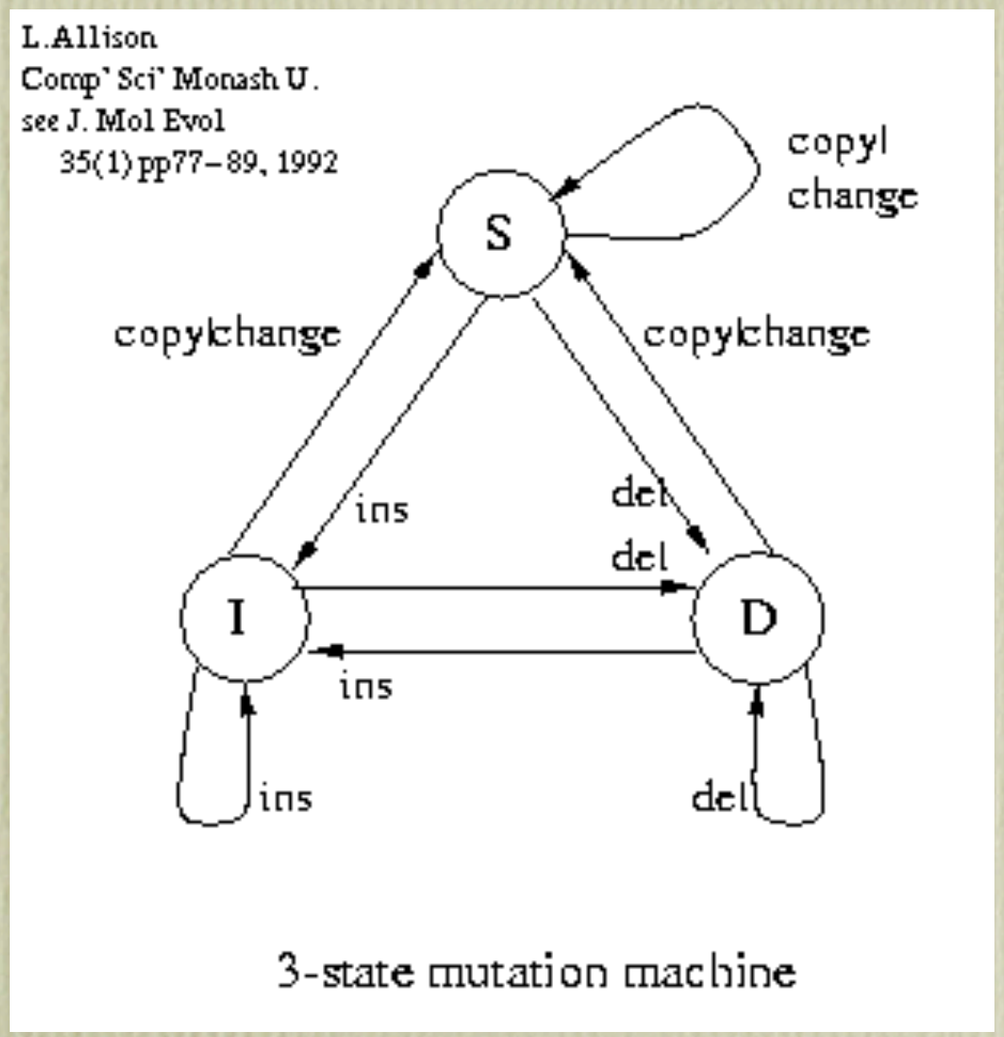
```
# - - -
# # # #
      k
```

```
# - - - -
- # # # #
        k
```

```
* - - - -
* # # # #
        k
```

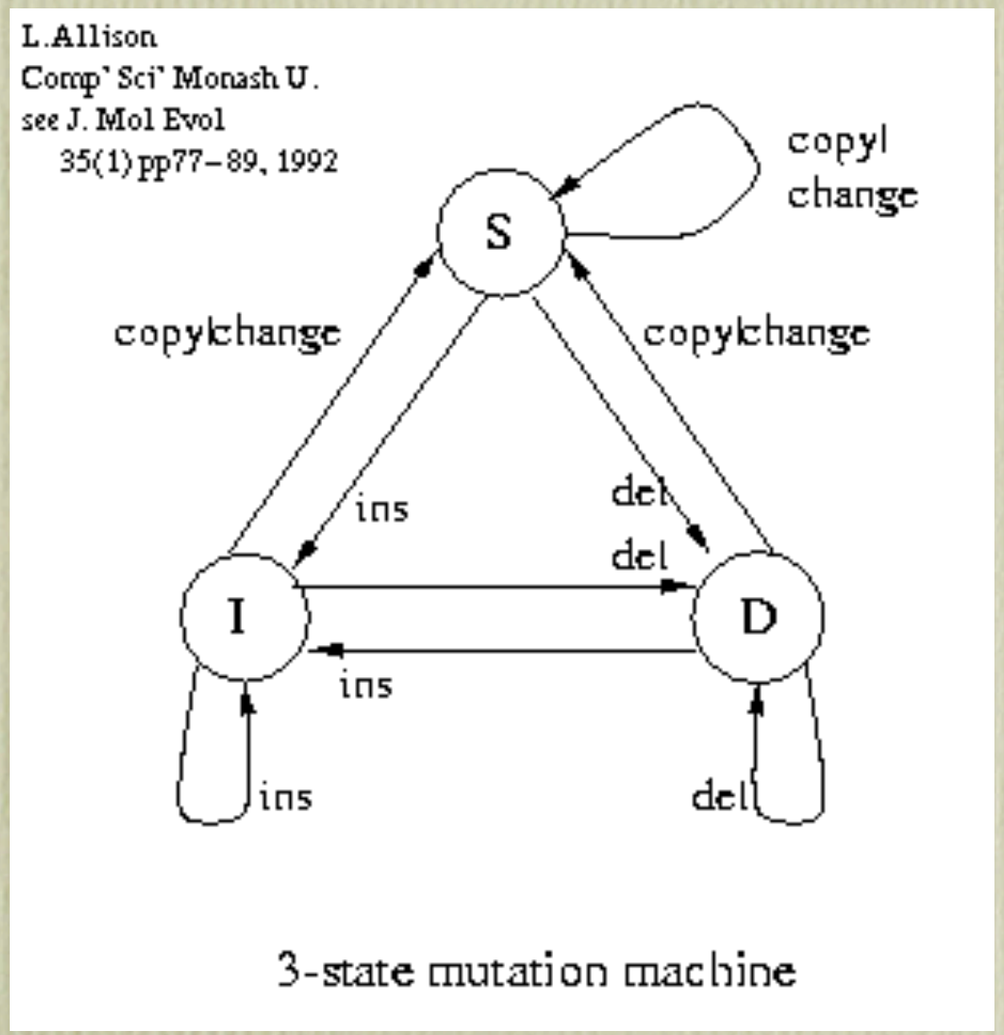
This gives the possbility for a Forward and Viterbi-like algorithm to calculate the observation probabilities and the most likely alignment.

Pair-Hidden Markov Models



3-state mutation machine

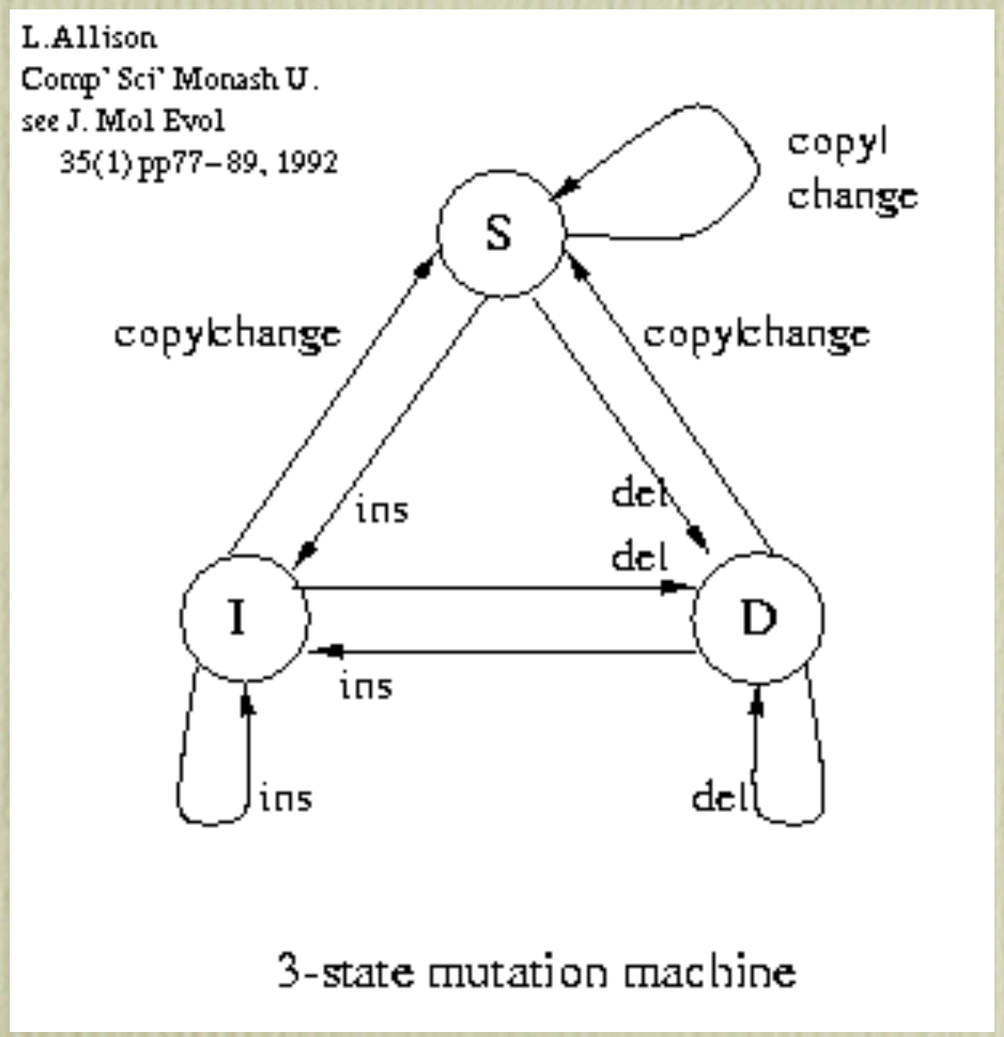
Pair-Hidden Markov Models



3-state mutation machine

Emits characters into two sequences.

Pair-Hidden Markov Models



3-state mutation machine

Emits characters into two sequences.

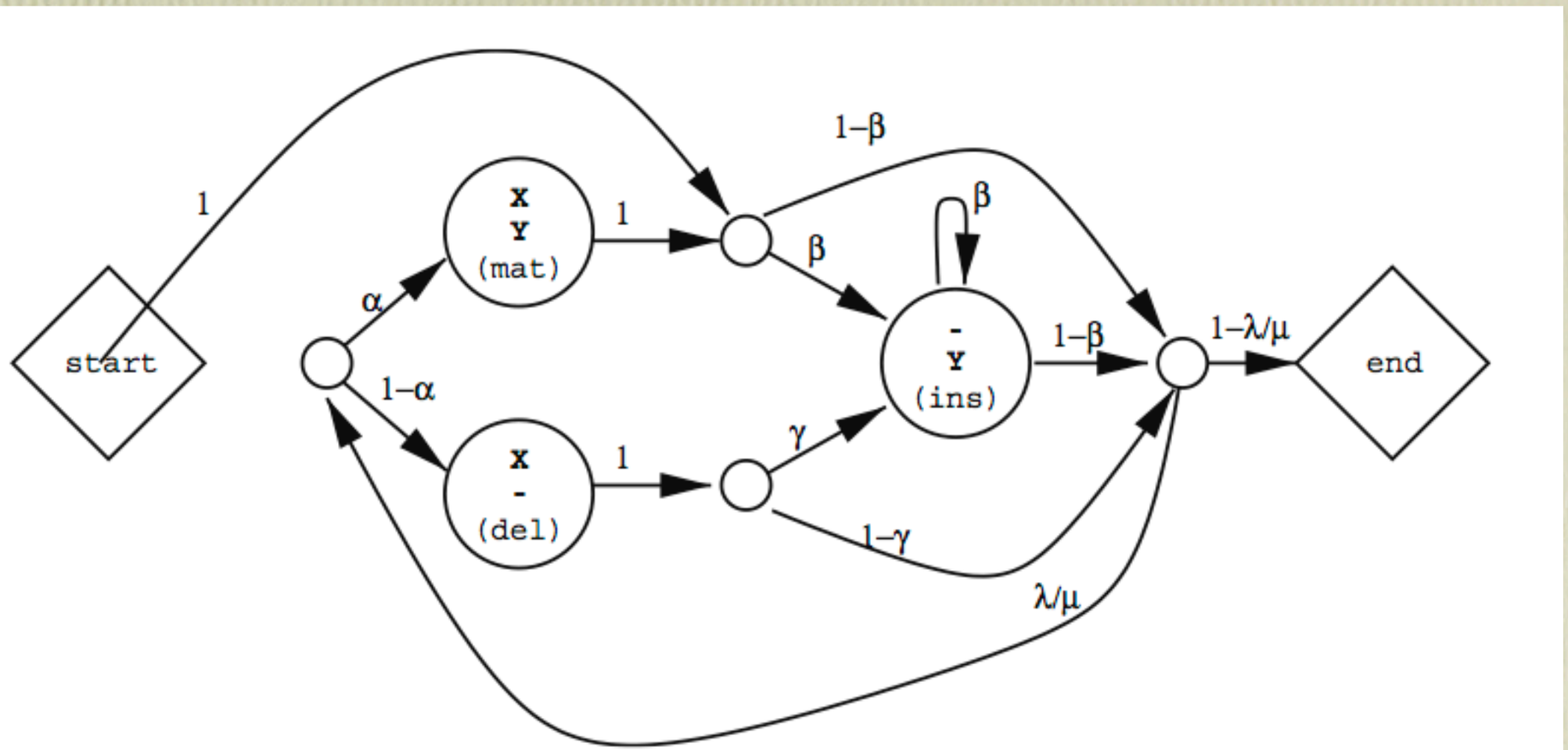
The observer can see only the emitted sequences and cannot see the path, even the co-emission pattern.

Pair-HMM dynamic pogramming

Viterbi: the most likely emitting path, namely, the most likely alignment.

Forward: the probability of emission, namely, the probability of observation.

The TKF91 model as HMM



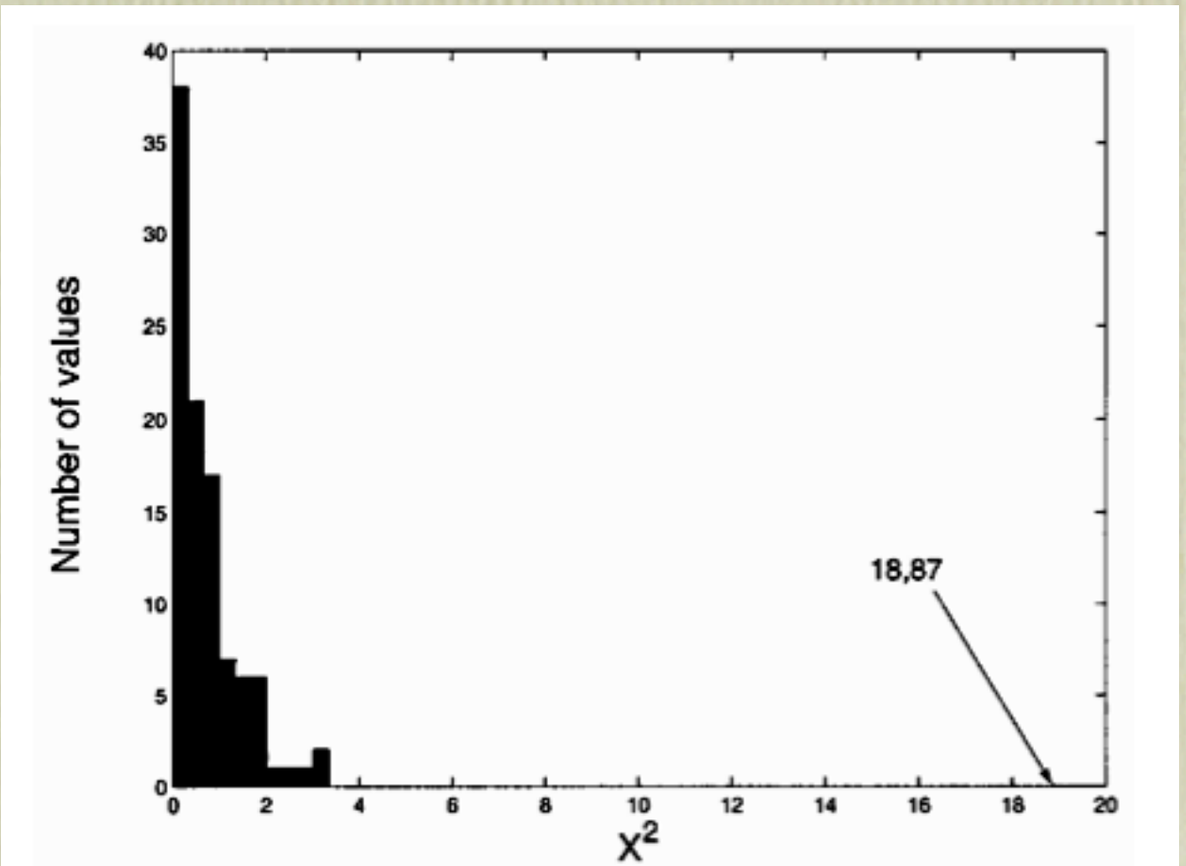
$$\alpha(t) = e^{-\mu t}$$

$$\beta(t) = \frac{\lambda(1 - e^{(\lambda-\mu)t})}{\mu - \lambda e^{(\lambda-\mu)t}}$$

$$\gamma(t) = 1 - \frac{\mu(1 - e^{(\lambda-\mu)t})}{(1 - e^{-\mu t})(\mu - \lambda e^{(\lambda-\mu)t})}$$

The TKF91 model is biologically unrealistic

Hein et al. (2000): goodness-of-fit test

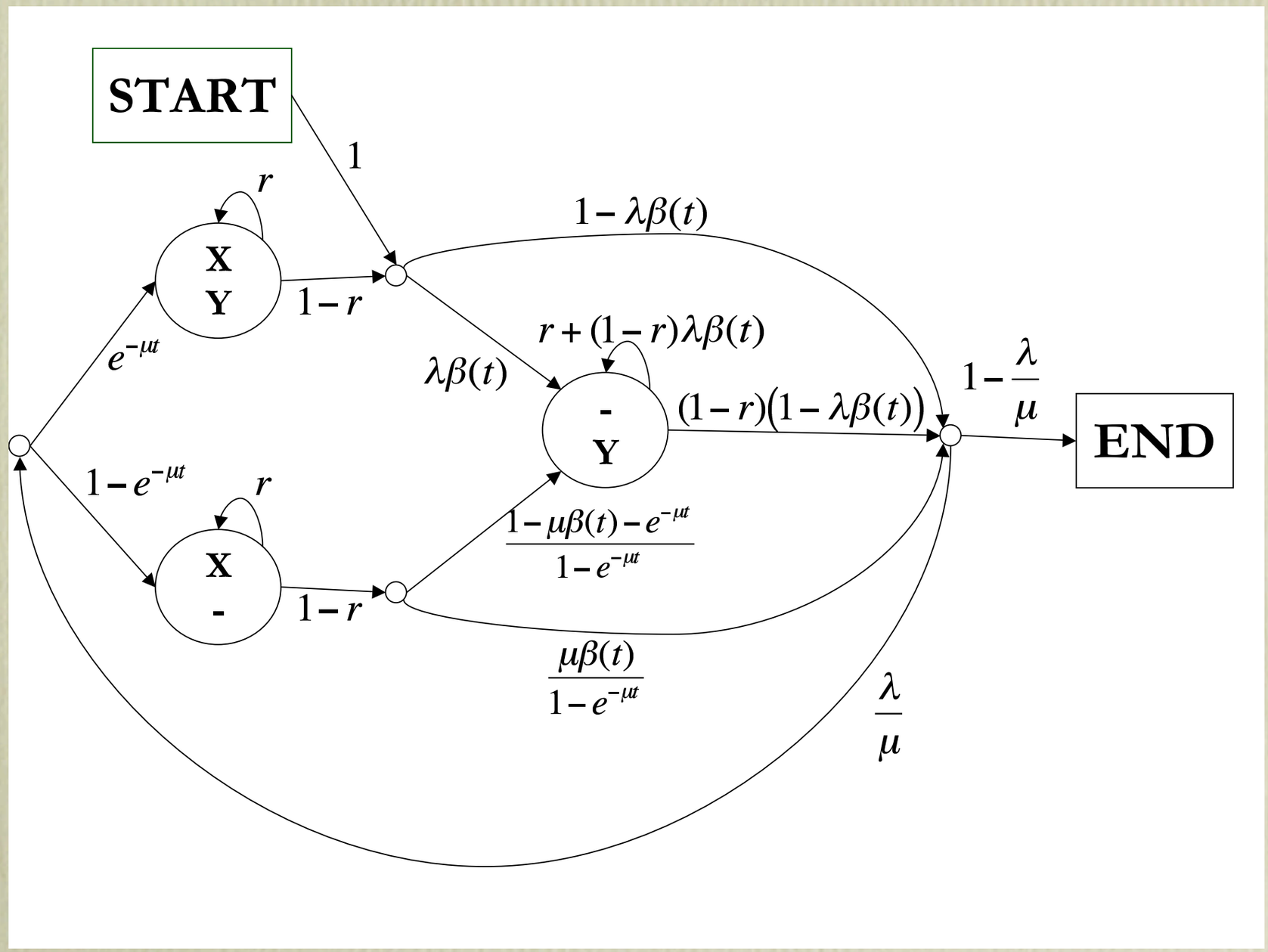


V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALT
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
DGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

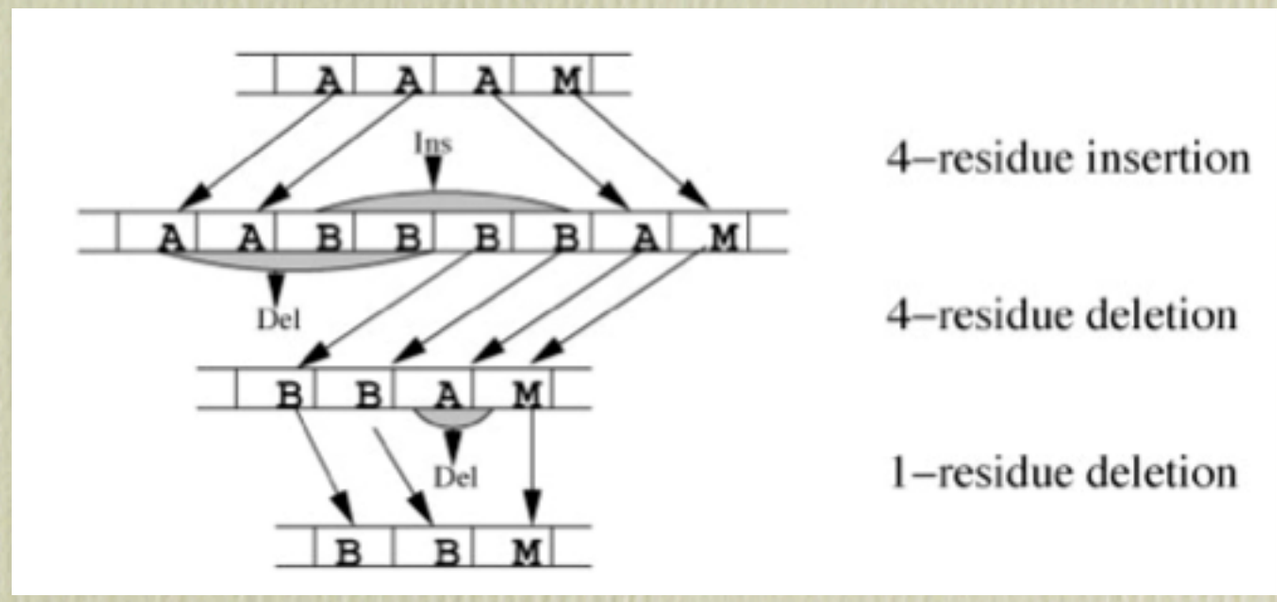
The TKF92 model

Unbreakable fragments



The ‘long indel’ model

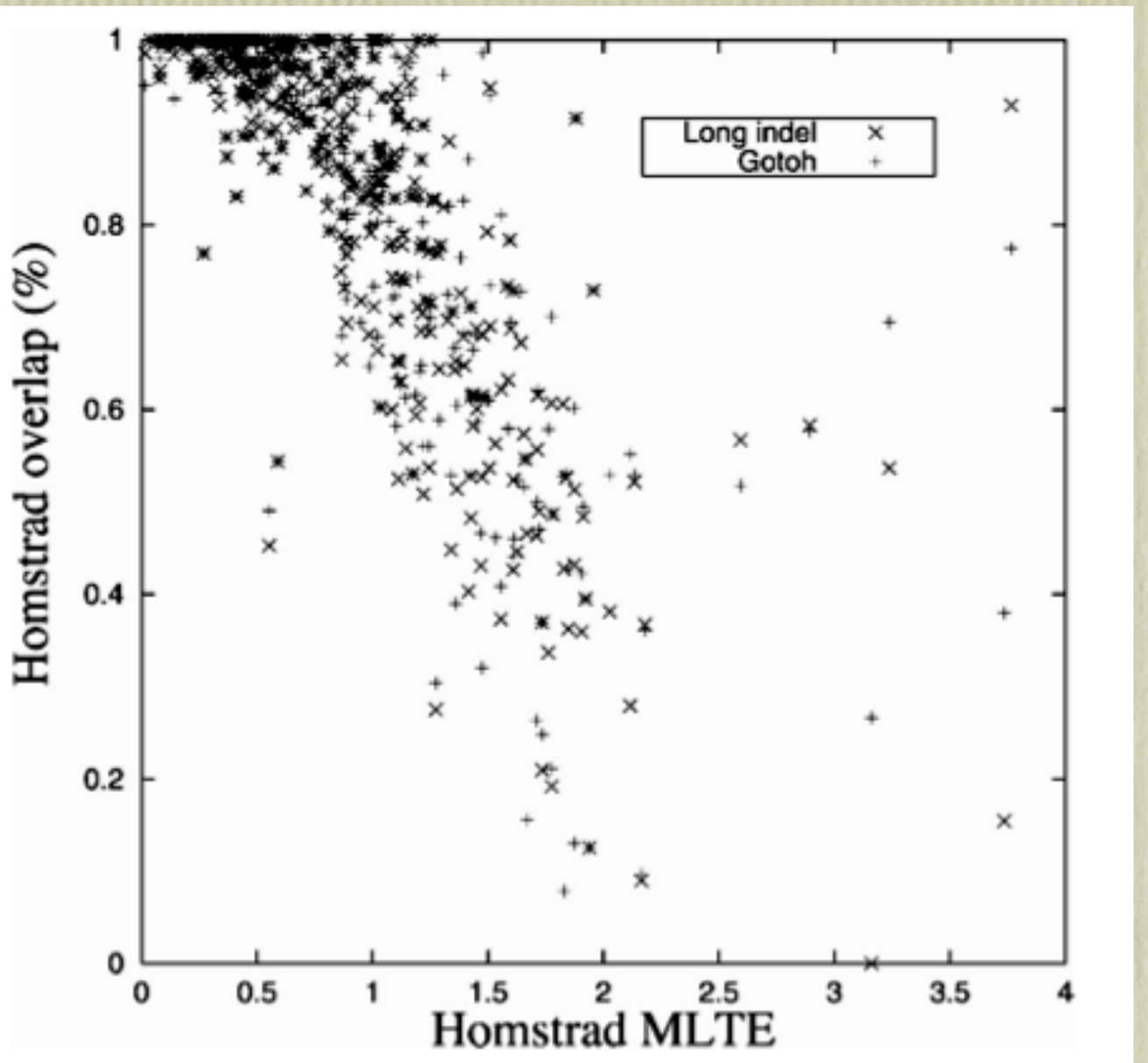
It is capable to model overlapping insertions and deletions



More involved dynamic programming in O(n^4) time

$$P_j^i = L_{i-1,j-1} p_t(A_i \to B_j) \prod_{k=1}^{j-1} q(B_k) + \sum_{n=0}^{i-2} \sum_{m=0}^{j-2} P_{j-m-1}^{i-n-1} N_{nm} p_t(A_i \to B_j) \prod_{k=j-m}^{j-1} q(B_k),$$

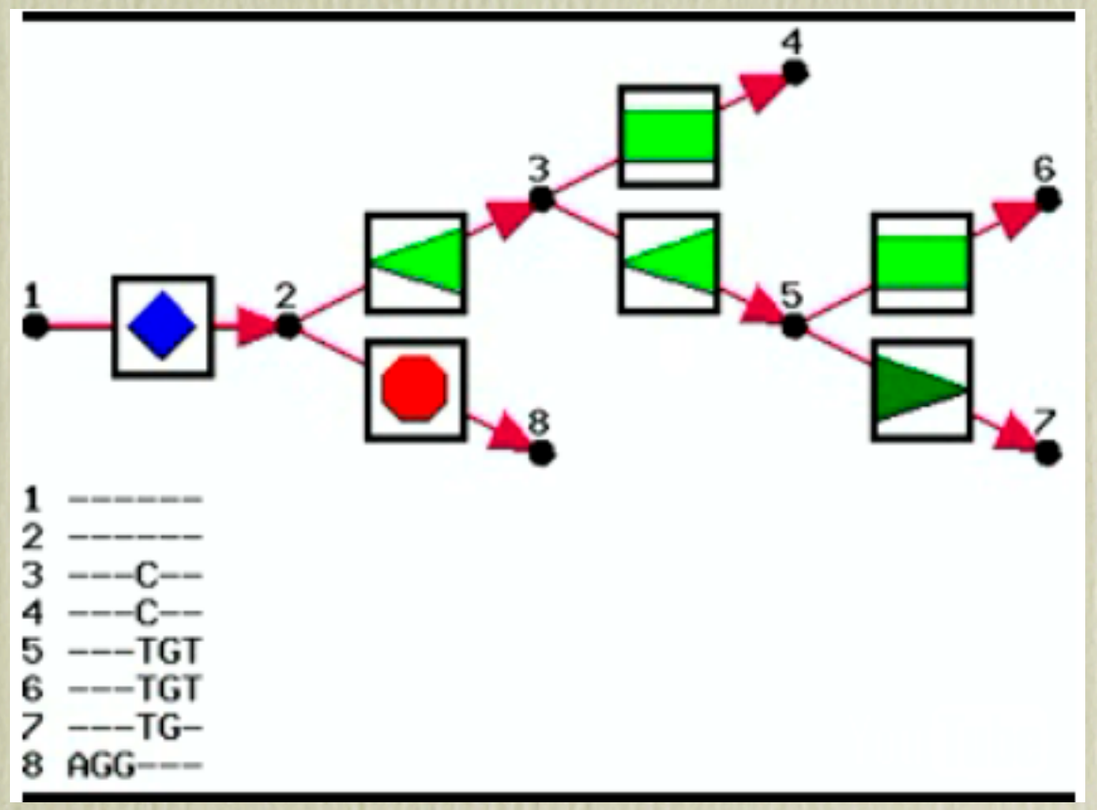
Alignment accuracy



Alignment Method	Training Set Optimization[a]	Test Set Overlap (%)
TKF91	ML	73.8
TKF92	ML	75.9
Gotoh (BLOSUM62)	NCBI defaults	80.9
Long indel	ML	81.1
Long indel, mixed geometric	Accuracy	82.1
Gotoh (BLOSUM62)	Accuracy	82.2

Multiple statistical alignment

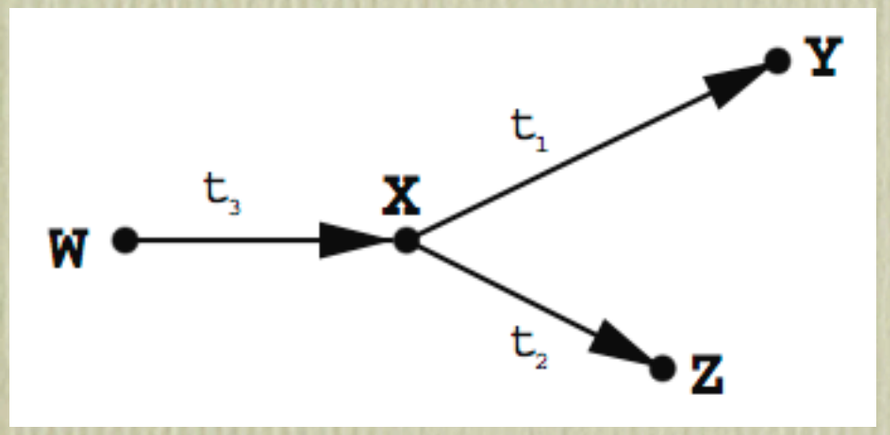
Transducer theory, Holmes (2003)

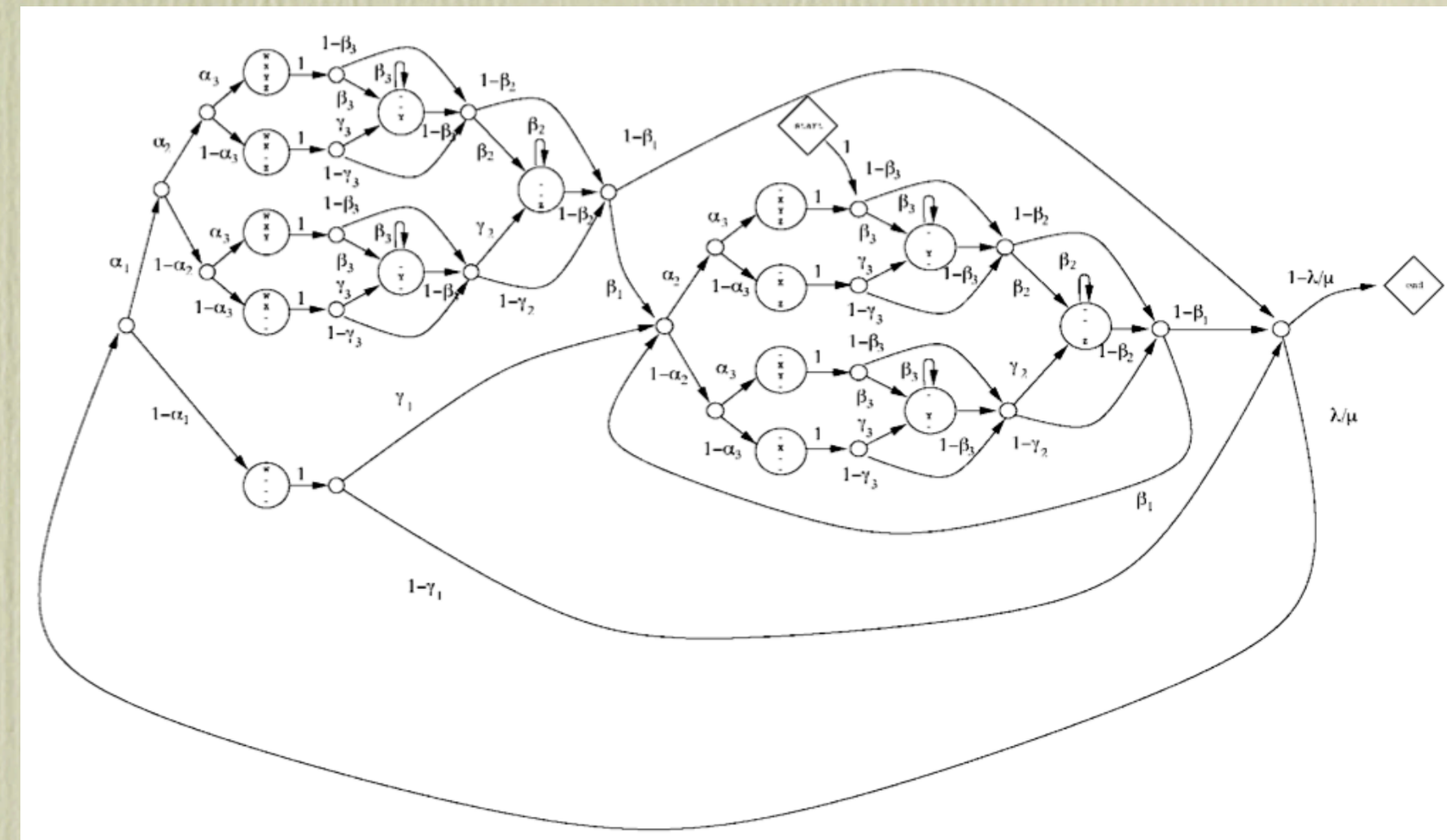


http://biowiki.org/PhyloFilm

http://www.youtube.com/watch?v=EcLj5MSDPyM

Multiple statistical alignment with multiple-HMMs





The number of states grows exponentially with the number of sequences

Algorithmic properties

- The running time of dynamic programming $O(5^n L^n)$ in case of Forward Multiple-HMMs
- In case of TKF91 model, it can be reduced to $O(2^n L^n)$-re
- Probably NP-hard

Markov chain Monte Carlo

T(Y|X) aperiodic, irreducible Markov chain,
$T(Y|X) \neq 0 \rightarrow T(X|Y) \neq 0$
$\forall X\ \pi(X)>0$ distribution

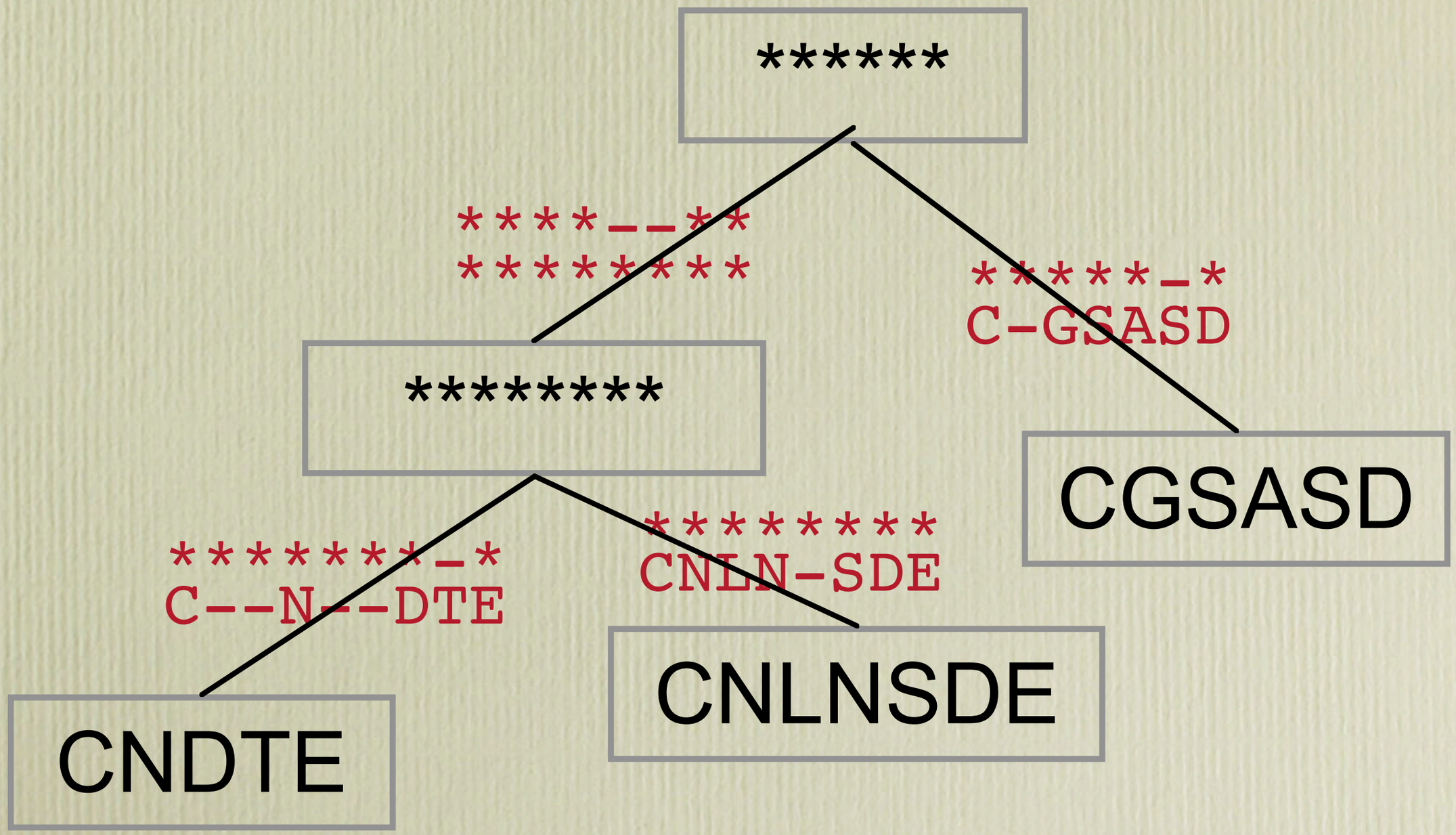
The following Markov chain converges to distribution π():

1. (proposal) Choose a random Y from T(·|X).
2. (acceptance) The next state in the Markov chain is Y with probability

$$\min\left\{1, \frac{T(X \mid Y)\pi(Y)}{T(Y \mid X)\pi(X)}\right\}$$

and remains X with complementer probability.

The state space of the Markov chain



Partial importance sampling of sequence alignments

```
ALITL---GGQST--QCC-STNQHVSCTGN
ALLTLTTLGGS-------CCSGN-HVSCTGK
---TLTSLGA---QST--QCTNQH-SCTLN
ALLGLTSLGA---QST--QCTNQHVSCTLN
```

- Stochastic iterative alignment
- Forward-backward sampling from a pair-HMM
- Iteratively on a sub-tree

Partial importance sampling of sequence alignments

```
ALITL---GGQST--QCC-STNQHVSCTGN
ALLTLTTLGGS-------CCSGN-HVSCTGK
---TLTSLGA---QST--QCTNQH-SCTLN
ALLGLTSLGA---QST--QCTNQHVSCTLN
```

- Stochastic iterative alignment
- Forward-backward sampling from a pair-HMM
- Iteratively on a sub-tree

Partial importance sampling of sequence alignments

```
ALITL---GG      QSTQCCS      TNQHVSCTGN
ALLTLTTLGG      SCCS         GN-HVSCTGK
---TLTSLGA      QSTQC        TNQH-SCTLN
ALLGLTSLGA      QSTQC        TNQHVSCTLN
```

- Stochastic iterative alignment
- Forward-backward sampling from a pair-HMM
- Iteratively on a sub-tree

Partial importance sampling of sequence alignments

```
ALITL---GG    QSTQCCS    TNQHVSCTGN
ALLTLTTLGG    -S--CCS    GN-HVSCTGK
---TLTSLGA    QSTQC--    TNQH-SCTLN
ALLGLTSLGA    QSTQC--    TNQHVSCTLN
```

- Stochastic iterative alignment
- Forward-backward sampling from a pair-HMM
- Iteratively on a sub-tree

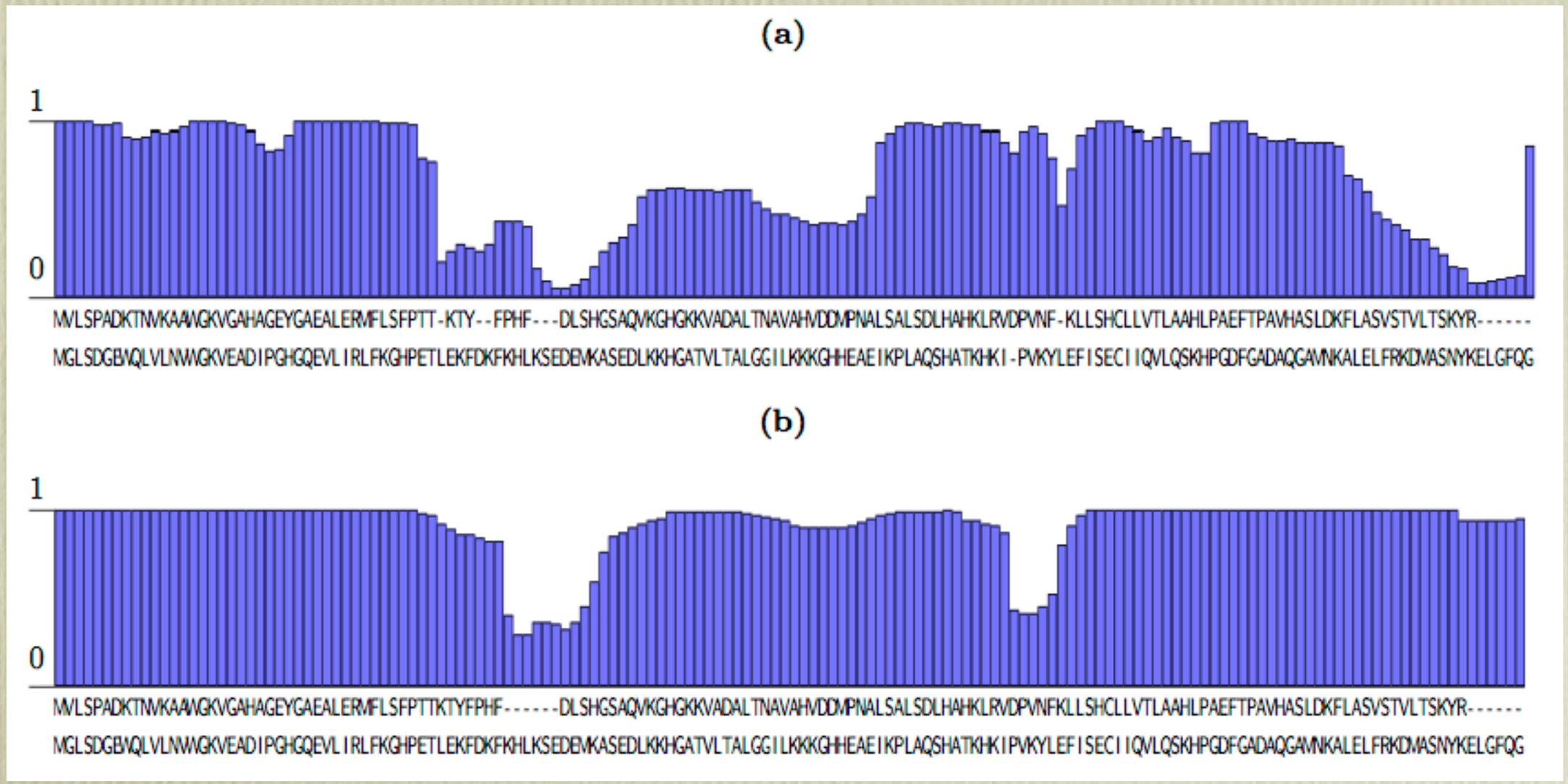
Partial importance sampling of sequence alignments

```
ALITL---GGQSTQCCSTNQHVSCTGN
ALLTLTTLGG-S--CCSGN-HVSCTGK
---TLTSLGAQSTQC--TNQH-SCTLN
ALLGLTSLGAQSTQC--TNQHVSCTLN
```

- Stochastic iterative alignment
- Forward-backward sampling from a pair-HMM
- Iteratively on a sub-tree

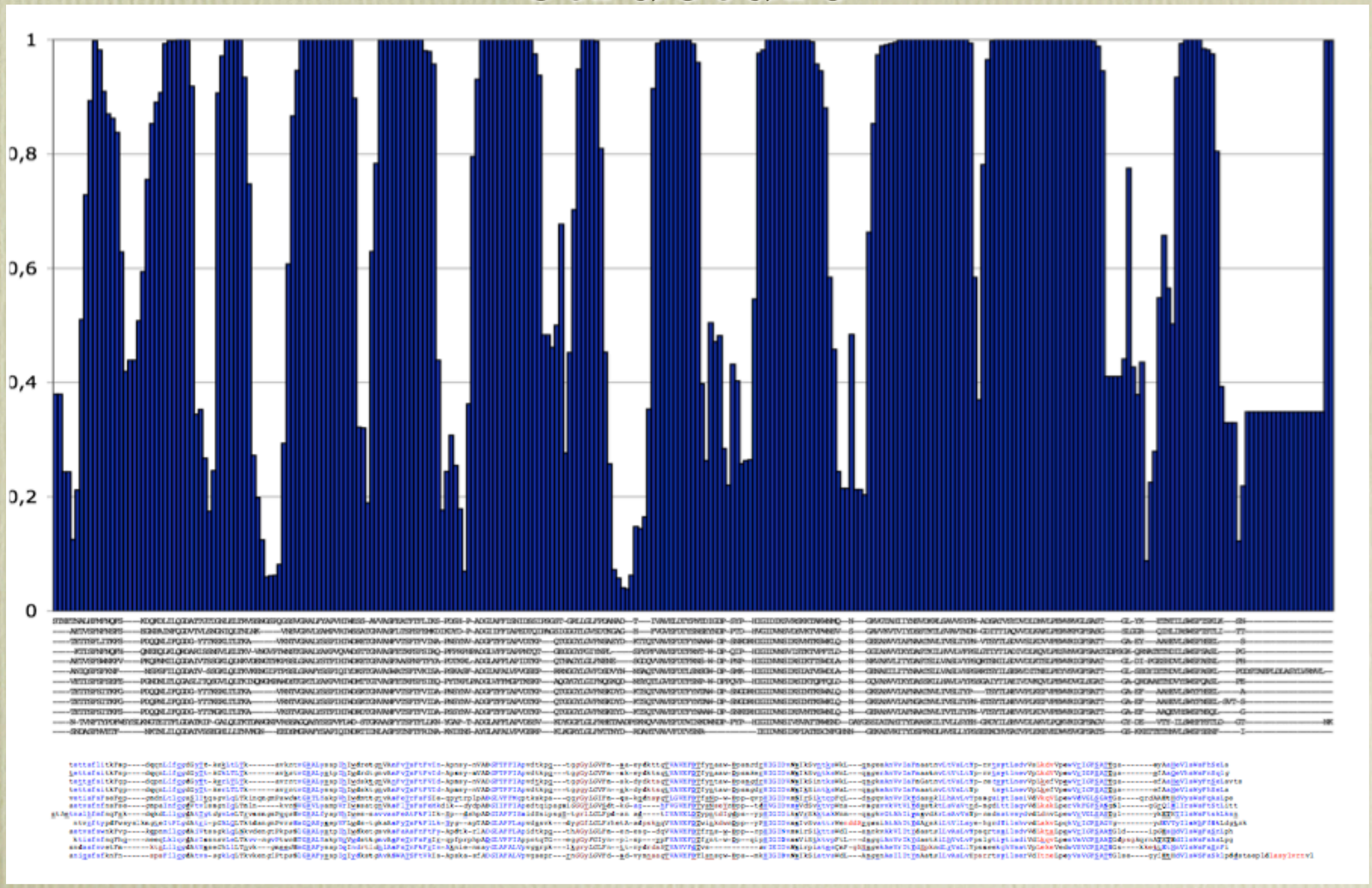
Posterior decoding

What is the probability that two characters are homologous given that the sequences are homologous



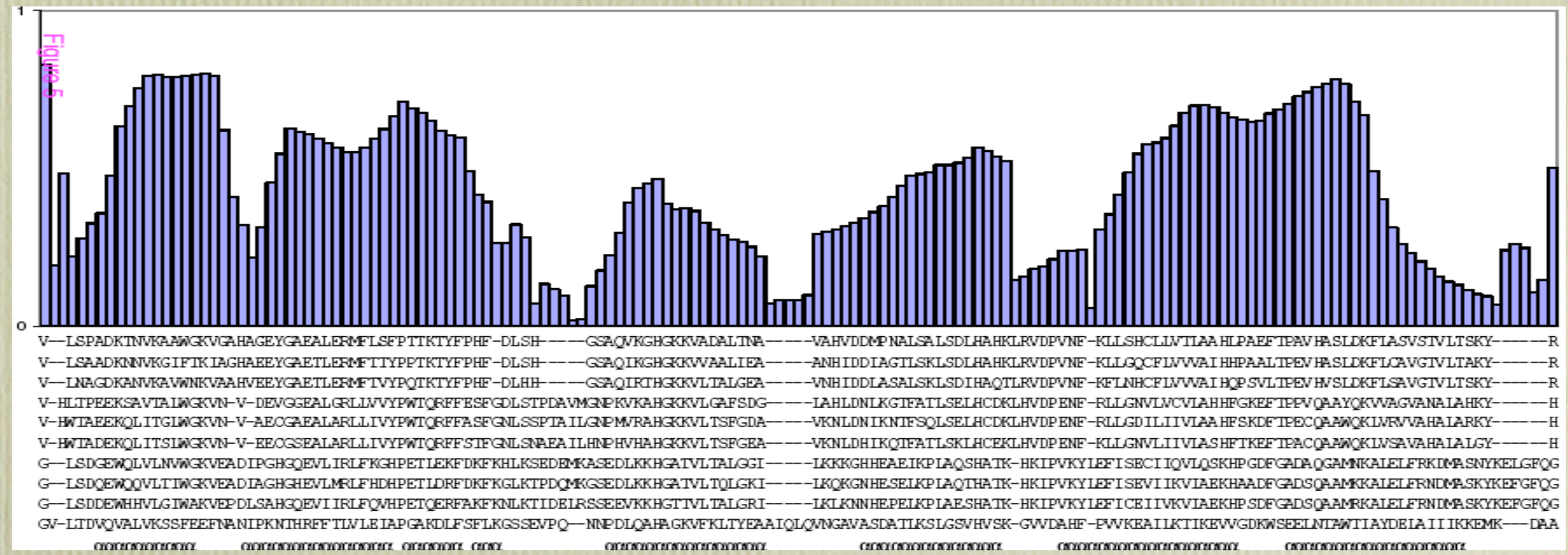
Viterbi alignment of Human alfa and beta globins in TKF91 and TKF92 models, together with posterior decoding.

Posterior decoding values correlate with structure



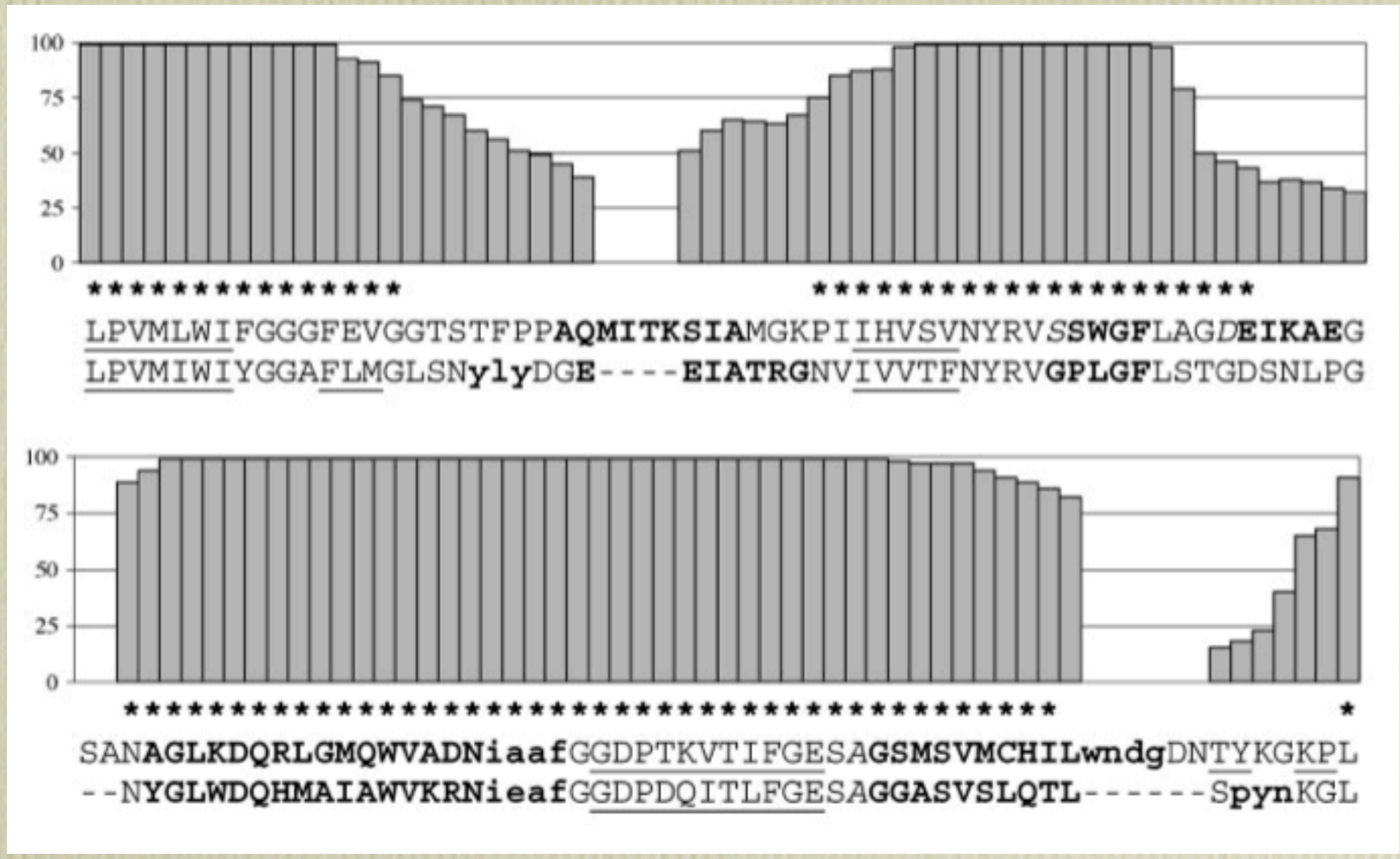
Legume lectin MPD alignment in the TKF92 model

Posterior decoding values correlate with structure



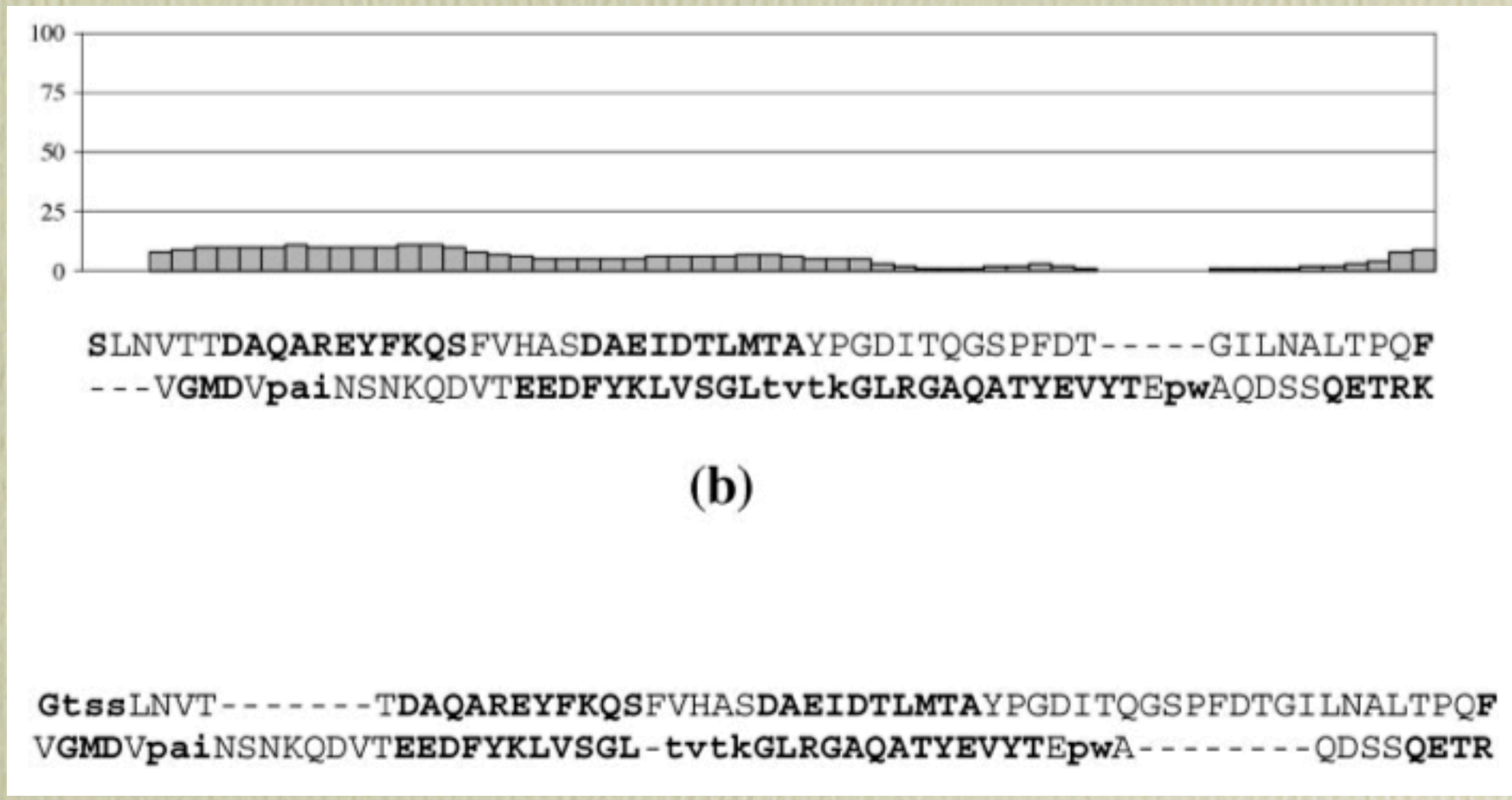
Multiple alignment of globins under the TKF91 model

Posterior probabilities correlate with the probability of correct prediction



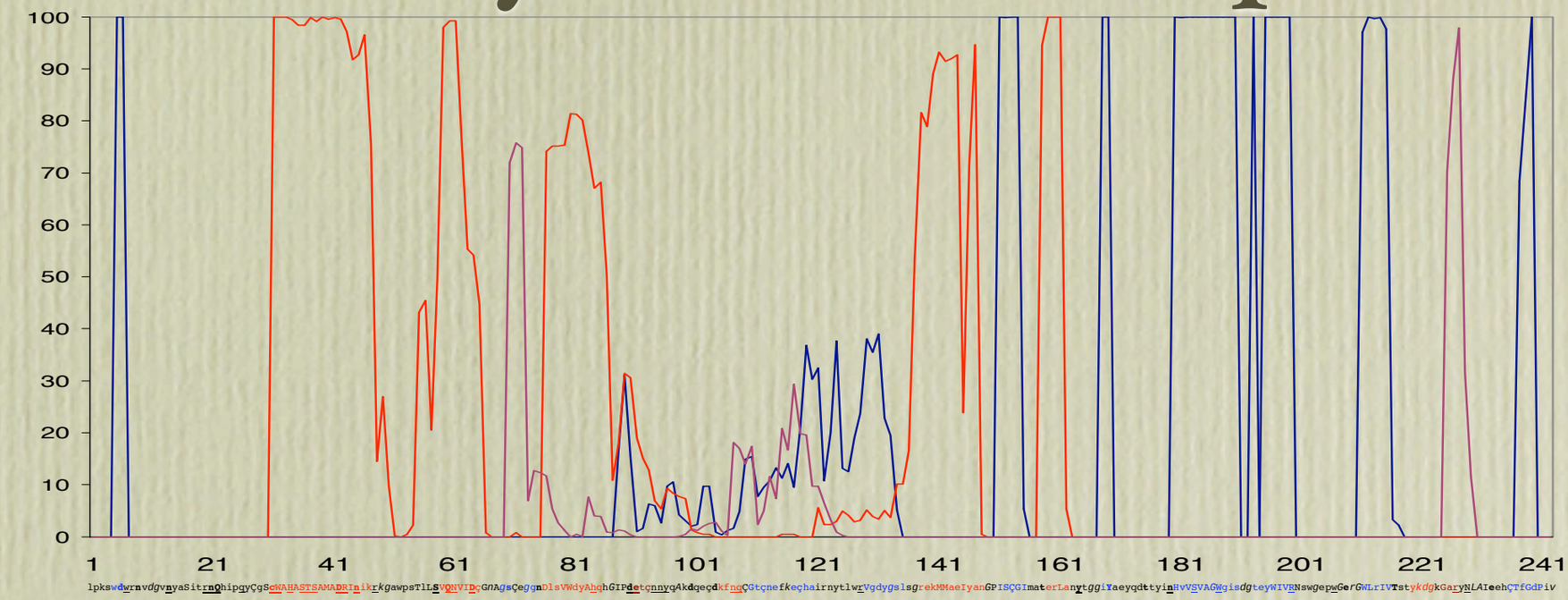
Part of the Candia rugosa and Bos taurus alfa-beta hydrolase Viterbi alignment under the 'long indel' model.

Posterior probabilities correlate with the probability of correct prediction



Part of the same alignment than on the previous slide.
Below: the HOMSTRAD structural alignment.

Posterior probabilities correlate with the probability of correct prediction



Papain family cystein proteinase, predicted structure

Papain family cystein proteinase, HOMSTRAD structural alignment + 3D structure

HOMSTRAD database

- Structural multiple sequence alignments
- 1031 family, 2-41 sequences/family, 8-92% PID
- Superimposed 3D structures are also downloadable in PBD format
- Alignments in JOY format

```
1gt91   ( 252 )    DVAGNADpa-TGYeVVIdgettviGgTsAVAPLFAALVArINqkLgkpVG
1ga6a   ( 261 )    DISFDAAqgT-GAlIynygqlqqiGGTsLASPIFVGLWARLQSanснsLG
1dbia   ( 202 )    dVVAPGv---dIvSTitgnryaymsGTsMASPHVAGLAALLASq--grnn
1thm    ( 201 )    dVAAPGs---siySTyptstyaslsGTsMATPHVAGVAGLLAsq--grsa
1bh6a   ( 197 )    eVMAPGv---sVySTypsntytslnGTSMASPHVAGAAALIlskyptlsA
1csee   ( 197 )    eVMAPGa---gVySTyptntyatlngTsMASPHVAGAAALIlskhpnlsA
1scja   ( 197 )    dVMAPGv---sIqSTlpggtygayngTcMATPHVAGAAALIlskhptwtn
1lw6e   ( 197 )    dVMAPGv---sIqSTlpgnkygaynGTsMASPHVAGAAALIlskhpnwtn
1gci    ( 197 )    DIVAPGv---nVqSTypgstyaslnGTsMATPHVAGAAALVkqknpswsn
1ea7a   ( 226 )    EISAPGs---sVySTwynggYntisgTSMATPHVSGLAAkIWaenpslsn
1ic6a   ( 200 )    dIFGPGt---dIlSTwiggstrsisGTsMATPHVAGLAAYLMTlg-ktta
```

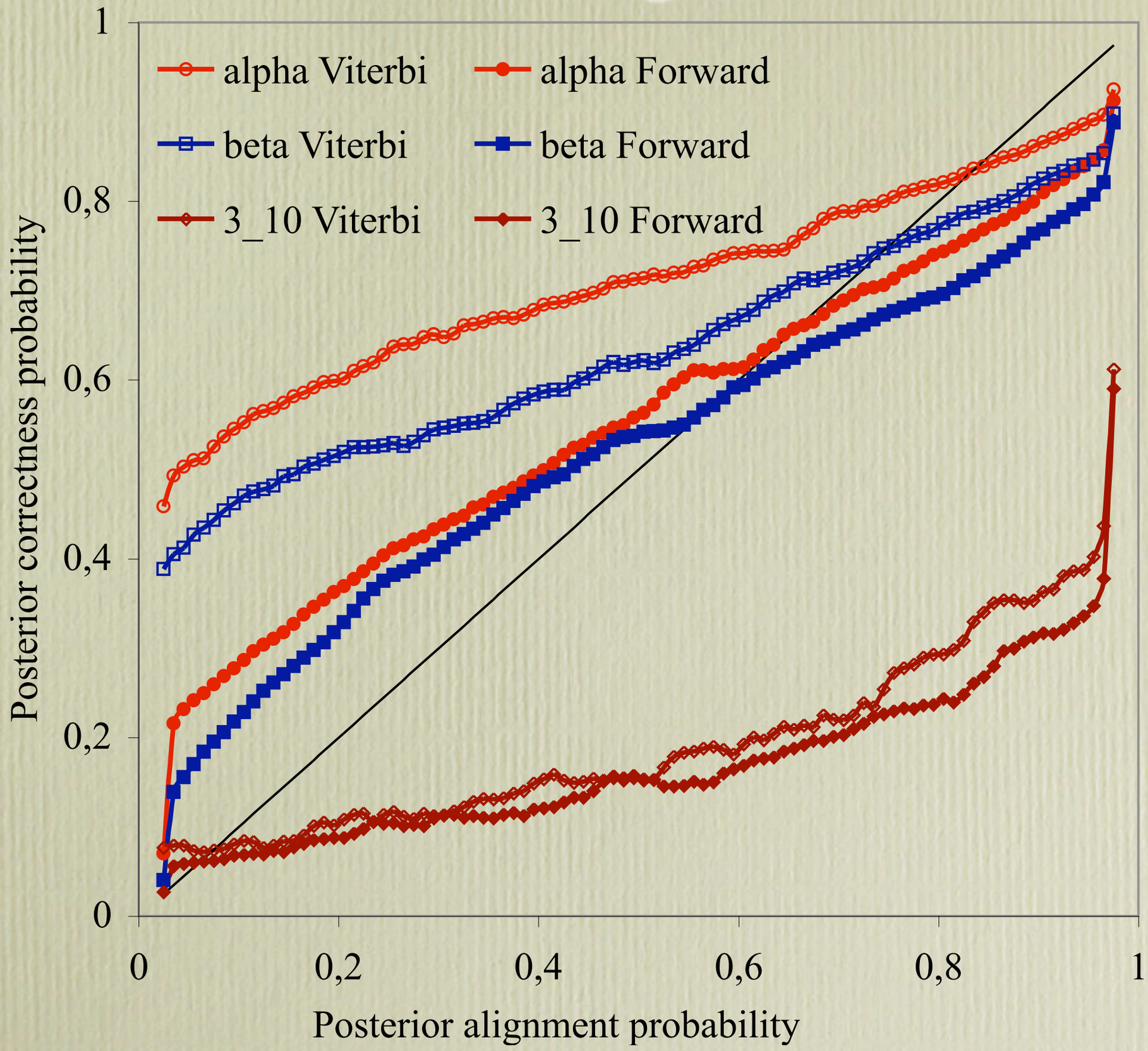
Predictions

- Both pairwise and multiple alignments
- Secondary structure predictions from Viterbi/MPD alignments, and also from the posterior distribution of alignments
- Multiple alignments only on 12 families
- 3D structure predictions for Viterbi/MPD alignments

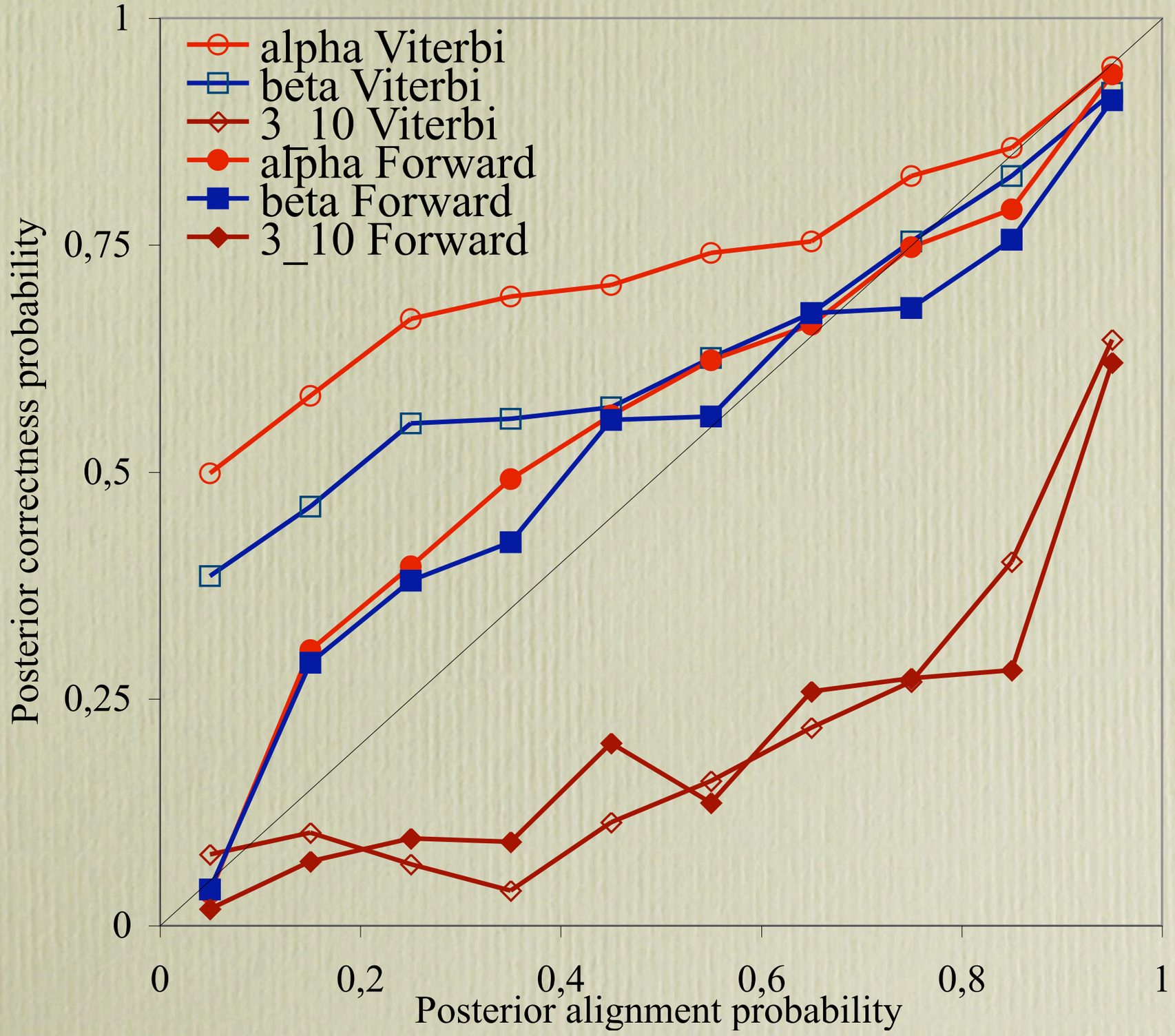
The 12 families

Family name	class	number of sequences	average length	average sequence id
Xylose isomerase	alpha beta barrel	6	388	69%
Annexin	all alpha	6	317	57%
Calcium-binding protein – parvalbumin-like	all alpha	7	107	56%
Starch binding domain	all beta	8	105	52%
Glycosyl hydrolase family 22 (lysozyme)	alpha+beta	12	126	51%
Legume lectin	all beta	12	234	50%
Papain family cysteine proteinase	alpha+beta	13	223	40%
Subtilase	alpha/beta	11	294	40%
Src homology 2 domains	alpha+beta	11	105	35%
C-type lectin	alpha+beta	8	126	27%
Halo-peroxidase	alpha/beta	9	286	25%
Response regulator receiver domain	alpha/beta	13	122	25%

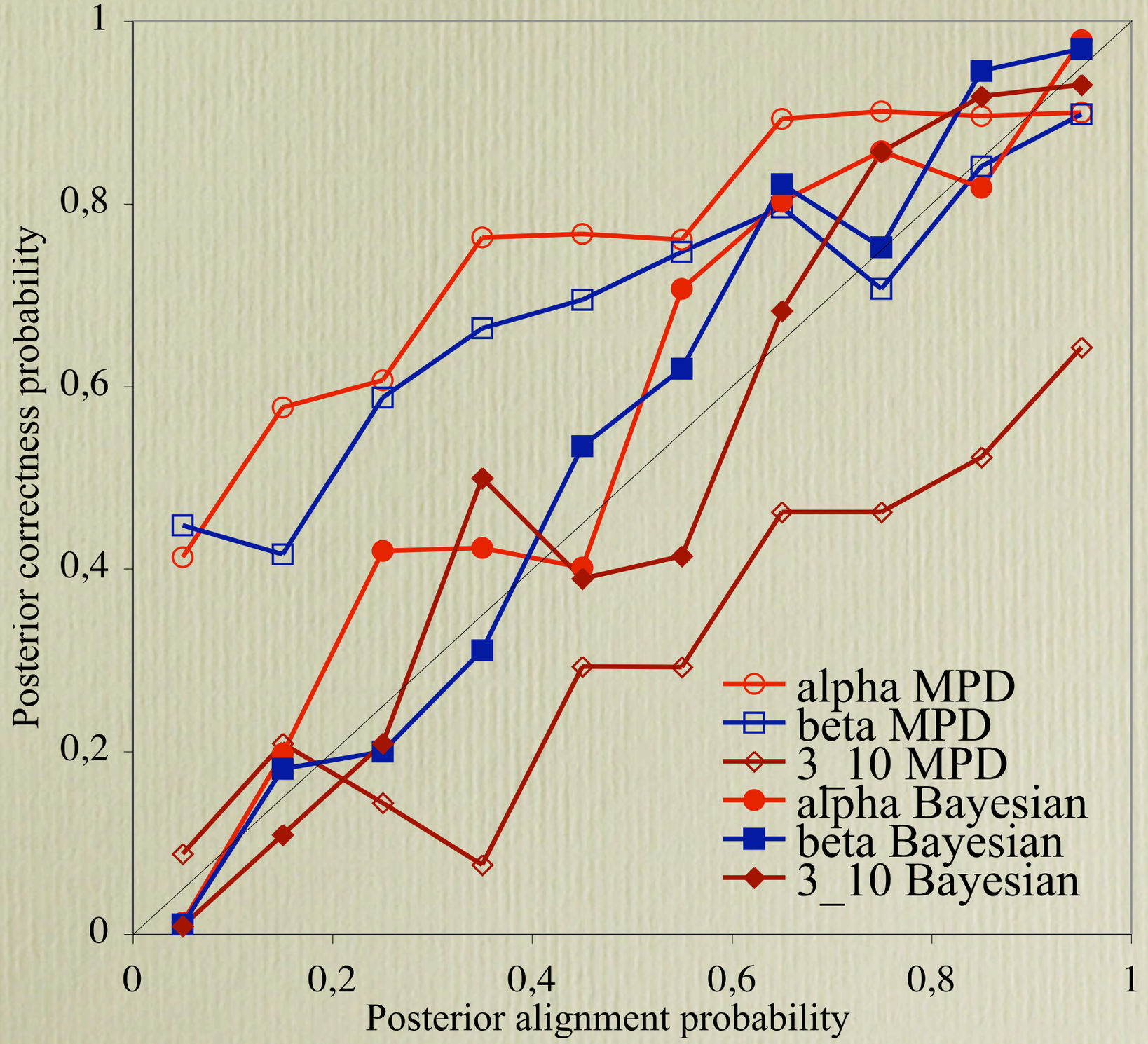
Pairwise alignments



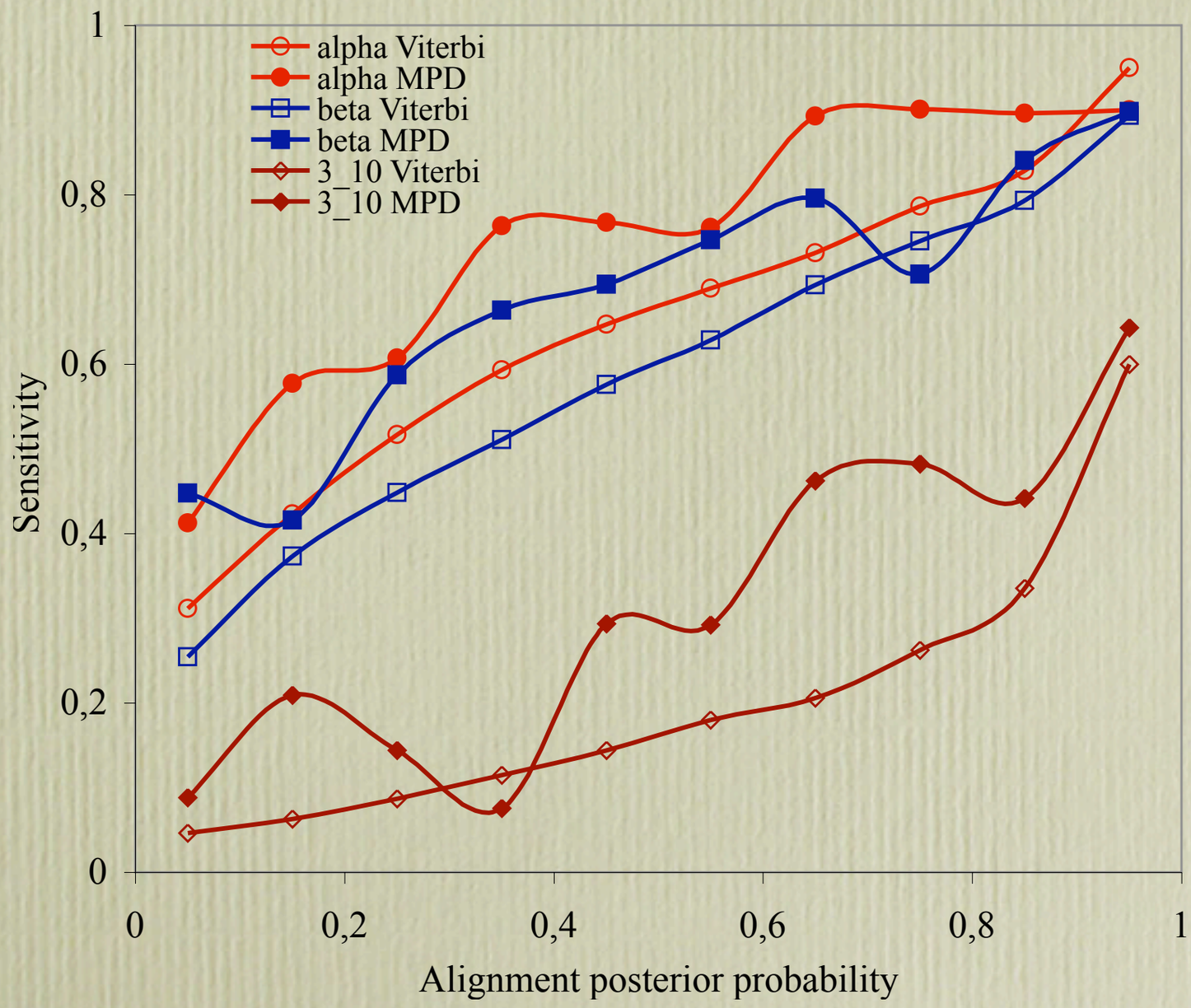
Pairwise alignments on the selected 12 families



Multiple alignments on the 12 families

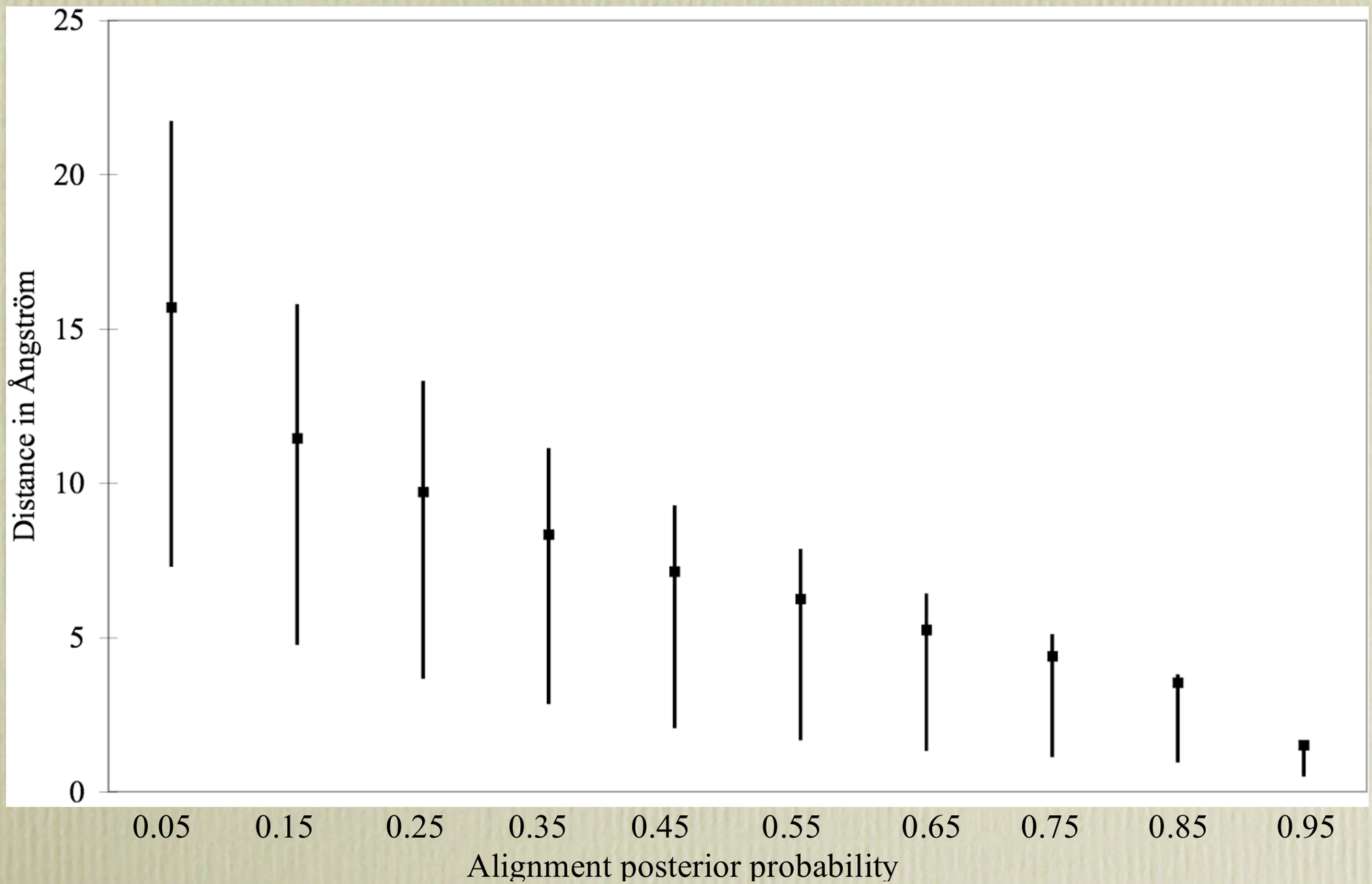


Sensitivities for the 12 selected alignments

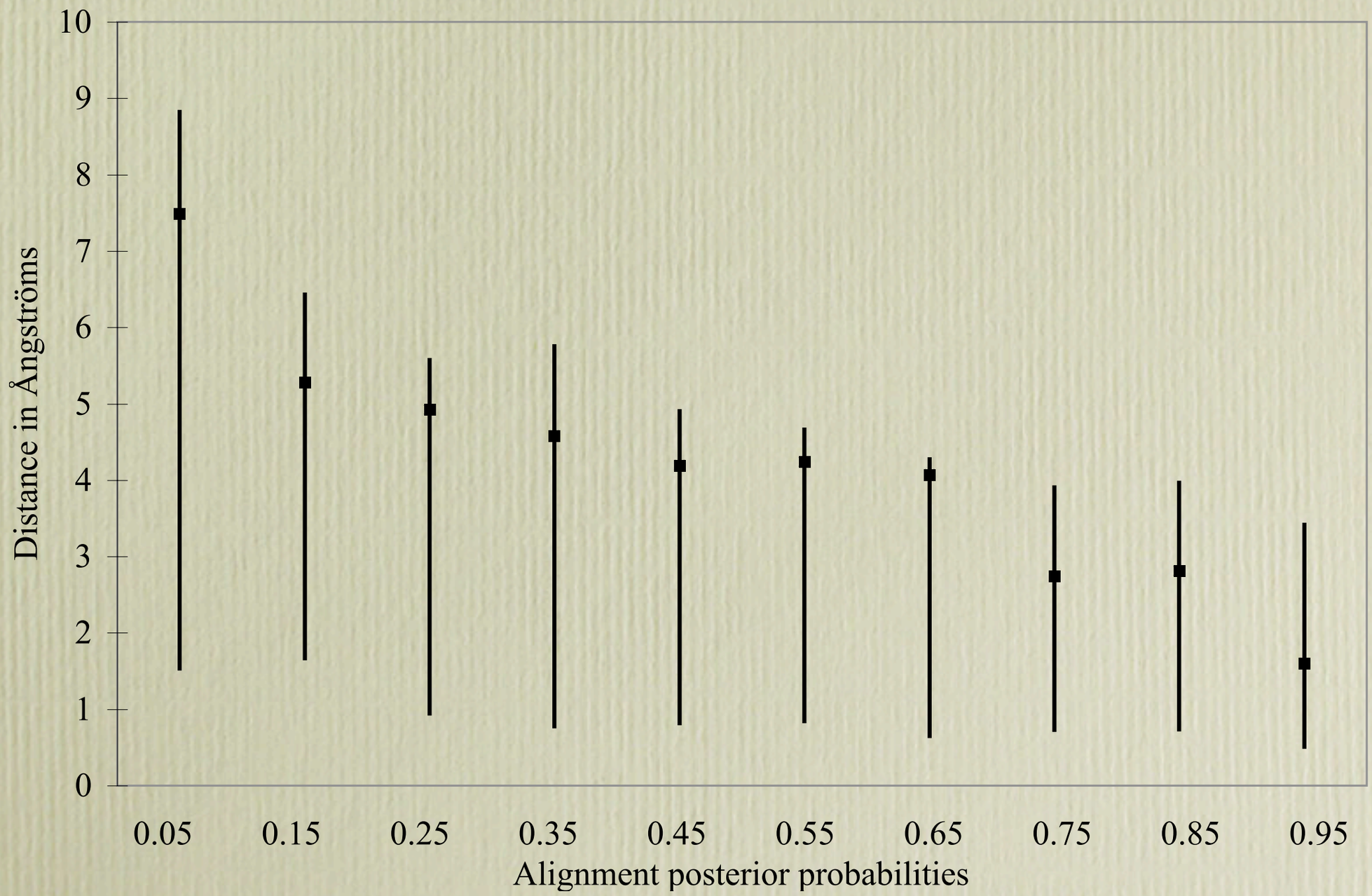


$$\frac{TP}{TP + FP}$$

Correlation between posterior probabilities and 3D structure distances



Correlation between posterior probabilities and 3D structure distances



Conclusions

- Posterior probabilities correlate with the actual probability that the predictions are correct.
- The best correlation achieved when predictions are based on the Bayesian distribution of multiple alignments.
- Potential tool for 3D predictions.