

# Hidden Markov Models for Sequence Alignment

Laboratory of Bioinformatics I  
Module 2

3 April, 2020

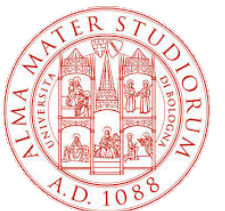
**Emidio Capriotti**

<http://biofold.org/>



**Biomolecules  
Folding and  
Disease**

Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna



# Alignment of Globins

Different positions are not equivalent

10 20 30 40 50 60 70 80

lqb1 pea/1-471 -GFTDKQ<sup>Q</sup>EALVNSSSE-FKQNLPG<sup>Y</sup>SILFY<sup>T</sup>IVLEKAPAAKGLFSFLKD---TAGVEDSPKLQAHAEQVFG<sup>L</sup>LV<sup>R</sup>DSAAQL

lqb1 vicfa/1-471 -GFTEKQ<sup>Q</sup>EALVNSSSQLFKQNP<sup>S</sup>SVLFY<sup>T</sup>II<sup>L</sup>QKAPTAKAMFSFLKD---SAGVVDS<sup>P</sup>KLGAHA<sup>E</sup>KEVFG<sup>M</sup>MV<sup>R</sup>DSAVQL

hbb speci/1-471 VHLSDGEKNAISTAWGKV--HAAEVGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGDLSSASAVMGNAK<sup>V</sup>KAHGKK<sup>V</sup>IDSFSNGLKHL

hbb speto/1-471 VHLTDGEKNAISTAWGKV--NAAEIGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGDLSSASAVMGNAK<sup>V</sup>KAHGKK<sup>V</sup>IDSFSNGLKHL

hbb equhe/1-471 VQLSGEEKAAVLALWDKV--NEEEVGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGDLSPAAVMGNPK<sup>V</sup>KAHGKK<sup>V</sup>LHSGEGVHHL

hbb sunmu/1-471 VHLSGEEKACVTGLWGKV--NEDEVGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGDLSSASAVMGNP<sup>K</sup>VKAHGKK<sup>V</sup>LHSLGEGVANL

hbb tupql/1-471 VHLSGEEKAAVTGLWGKV--DLEKVGGS<sup>L</sup>LSLLIVYPWTQ<sup>R</sup>FFDSFGDLSSPSAVMSNP<sup>K</sup>VKAHGKK<sup>V</sup>LTSFSDGLNHL

hbb calar/1-471 VHLTGEKKSAVTALWGKV--NVDEVGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFESFGDLSTPDAVMNNPK<sup>V</sup>KAHGKK<sup>V</sup>LGA<sup>F</sup>SDGLTHL

hbb mansp/1-471 VHLTPEEKTA<sup>V</sup>TTLWGKV--NVDEVGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGDLSSPDAVMGNPK<sup>V</sup>KAHGKK<sup>V</sup>LGA<sup>F</sup>SDGLNHL

hbb rabbit/1-471 VHLSSEKKS<sup>A</sup>VTALWGKV--NVEEVGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFESFGDLSSANAVMNNPK<sup>V</sup>KAHGKK<sup>V</sup>LAA<sup>F</sup>SEGLSHL

hbb ursma/1-471 VHLTGEKKS<sup>L</sup>VTLWGKV--NVDEVGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGDLSSADAIMNNPK<sup>V</sup>KAHGKK<sup>V</sup>LNS<sup>F</sup>SDGLKNL

hbb triin/1-471 VHLTPEEKAL<sup>V</sup>IGLWAKV--NVKEYGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFE<sup>H</sup>FGDLSSASAIMNNPK<sup>V</sup>KAHGK<sup>E</sup>VFTSFGDGLKHL

hbb ornan/1-471 VHLSGGEKS<sup>A</sup>VTNLWGKV--NINELGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFE<sup>A</sup>FGDLSSAGAVMGNP<sup>K</sup>VKAHGAK<sup>V</sup>LTSFGDALKNL

hbb tacac/1-471 VHLSGSEKT<sup>A</sup>VTNLWGHV--NVNELGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFESFGDLSSADAVMGNAK<sup>V</sup>KAHGAK<sup>V</sup>LTSFGDALKNL

hbe ponpy/1-471 VHFTAEEKAAV<sup>T</sup>SLWSKM--NVEEAGGA<sup>E</sup>ALGRLLV<sup>V</sup>YPWTQ<sup>R</sup>FFDSFGNLS<sup>S</sup>PSAILGNPK<sup>V</sup>KAHGKK<sup>V</sup>LTSFGDAIKNM

hbb colli/1-471 VHWSAEEKQ<sup>L</sup>ITSIWGKV--NVADCGAEAL<sup>A</sup>RLLIVYPWTQ<sup>R</sup>FFSSFGNLS<sup>S</sup>ATAISGNPNVKAHGKK<sup>V</sup>LTSFGDAVKNL

hbb larri/1-471 VHWSAEEKQ<sup>L</sup>ITGLWGKV--NVADCGAEAL<sup>A</sup>RLLIVYPWTQ<sup>R</sup>FFASFGNLS<sup>S</sup>PTAINGNPMVRAHGKK<sup>V</sup>LTSFGDAVKNL

hbb1 varex/1-471 VHWTAEEKQ<sup>L</sup>ICSLWGKI--DVGLIGGET<sup>L</sup>AGLLVIYPWTQ<sup>R</sup>QFS<sup>H</sup>FGNLS<sup>S</sup>PTAIAGNPRVKAHGKK<sup>V</sup>LTSFGDAIKNL

hbb2 xentr/1-471 VHWTAEEKAT<sup>I</sup>ASVWGKV--DIEQDGH<sup>D</sup>ALSRLV<sup>V</sup>YPWTQ<sup>R</sup>YFSSFGNLS<sup>N</sup>VSAVSGNVK<sup>V</sup>KAHGK<sup>N</sup>VLSAVGSAIQHL

hbb1 ranca/1-471 VHWTAEEKAV<sup>I</sup>NSVWQKV--DVEQDGHEA<sup>T</sup>RLFIYPWTQ<sup>R</sup>YFSTFGDLSSPA<sup>A</sup>IAGNPK<sup>V</sup>HAHGKK<sup>I</sup>LGAIDNAIHNL

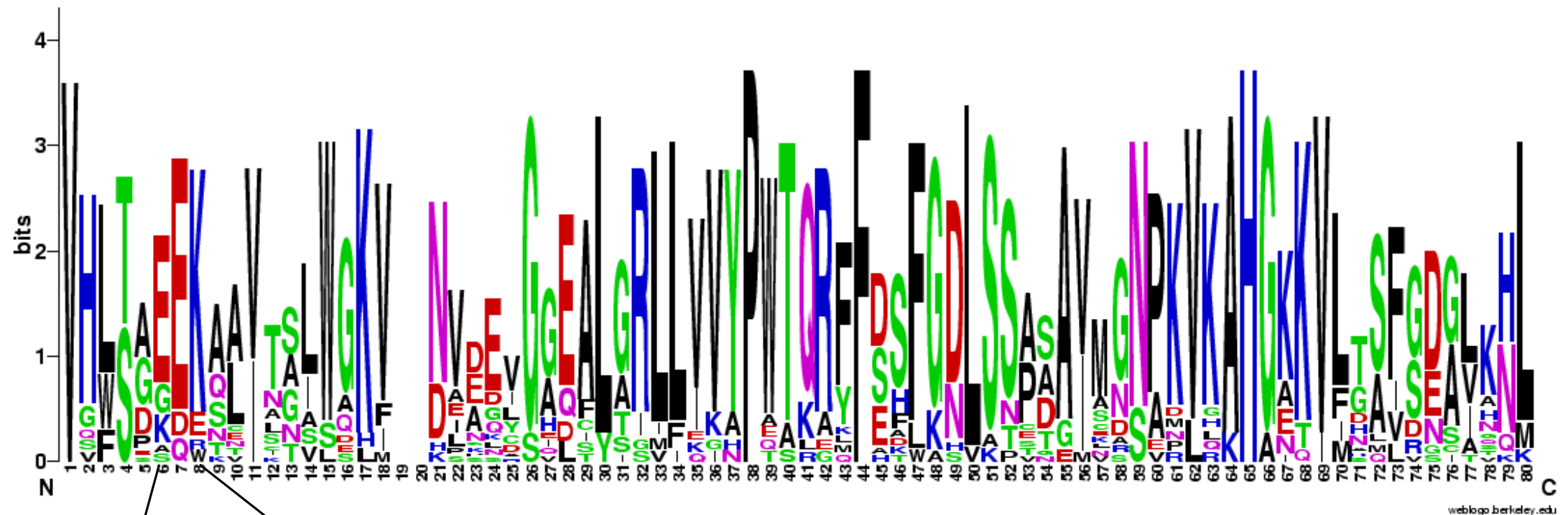
hbb2 tricr/1-471 VHLTAEDRKEIA<sup>A</sup>ILGKV--NVDSLGGQCL<sup>A</sup>RLIVNPNWSRRYF<sup>H</sup>DFGDLSSCDAICRNPK<sup>V</sup>LAHGAK<sup>V</sup>MRSIVEATKHL

hba4 salir/1-471 -SLSAKD<sup>K</sup>ANVKAIWGKILPK<sup>S</sup>DEIGEQA<sup>L</sup>SRMLV<sup>V</sup>YPQTKAYFSHWASVAP-----GSAPVKKHG<sup>I</sup>ITIMNQIDDCVGHM

myg\_escgi/1-471 -VLSDAEWQ<sup>L</sup>VLNIWAKVEADVAGHGQ<sup>D</sup>ILIRLFKGHPETLEKFDK<sup>E</sup>KHLKTEAEMKAS<sup>E</sup>DLKKHGNTVLTALGGILKKK

# Sequence Logo

## A more flexible alignment score is needed to align protein families

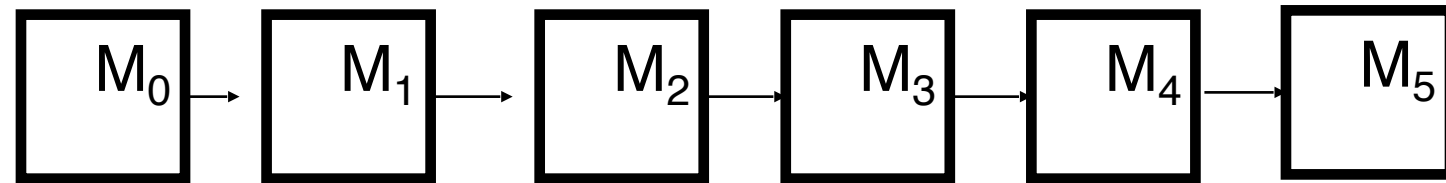


The substitution score may depend on the position.



# How to Align?

Each state represent a position in the alignment.



A	C	G	G	T	A
$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$

A	C	G	A	T	C
$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$

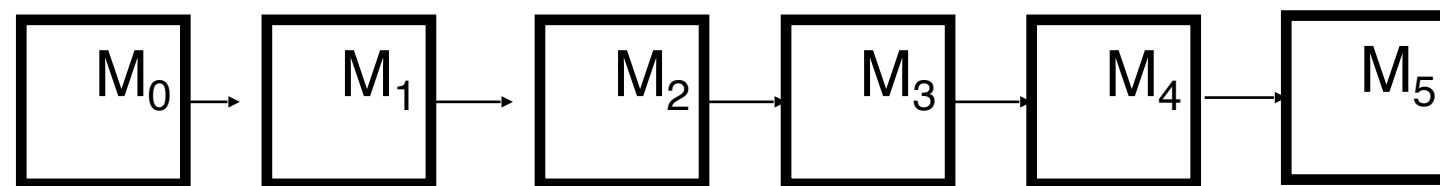
A	T	G	T	T	C
$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$

Each position has a peculiar composition

# From Sequences to Model

Given a set of sequences we can train a model by estimating the emission probability

A C G G T A  
A C G A T C  
A T G T T C



A	1	0	0	0.33	0	0.33
C	0	0.66	0	0	0	0.66
G	0	0	1	0.33	0	0
T	0	0.33	0	0.33	1	0

# Scoring a Sequence

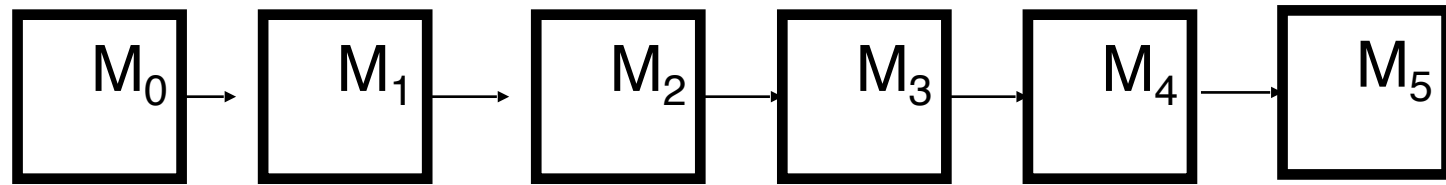
Given the model we can calculate the probability of the a new aligned sequence

	$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$
A	1	0	0	0.33	0	0.33
C	0	0.66	0	0	0	0.66
G	0	0	1	0.33	0	0
T	0	0.33	0	0.33	1	0
	A	C	G	A	T	C

$$P(s | M) = 1 \times 0.66 \times 1 \times 0.33 \times 1 \times 0.66$$

# Alignments with Gaps

A strategy to introduce gaps is needed



A	1	0	0	0 . 3 3	0	0 . 3 3
C	0	0 . 6 6	0	0	0	0 . 6 6
G	0	0	1	0 . 3 3	0	0
T	0	0 . 3 3	0	0 . 3 3	1	0

A  
M<sub>0</sub>

M<sub>2</sub>

G  
M<sub>3</sub>

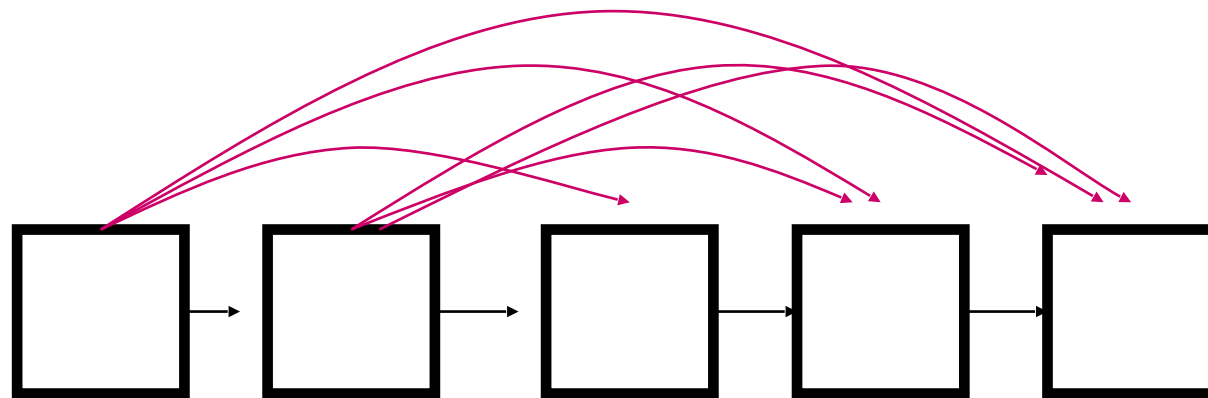
A  
M<sub>4</sub>

T  
M<sub>5</sub>

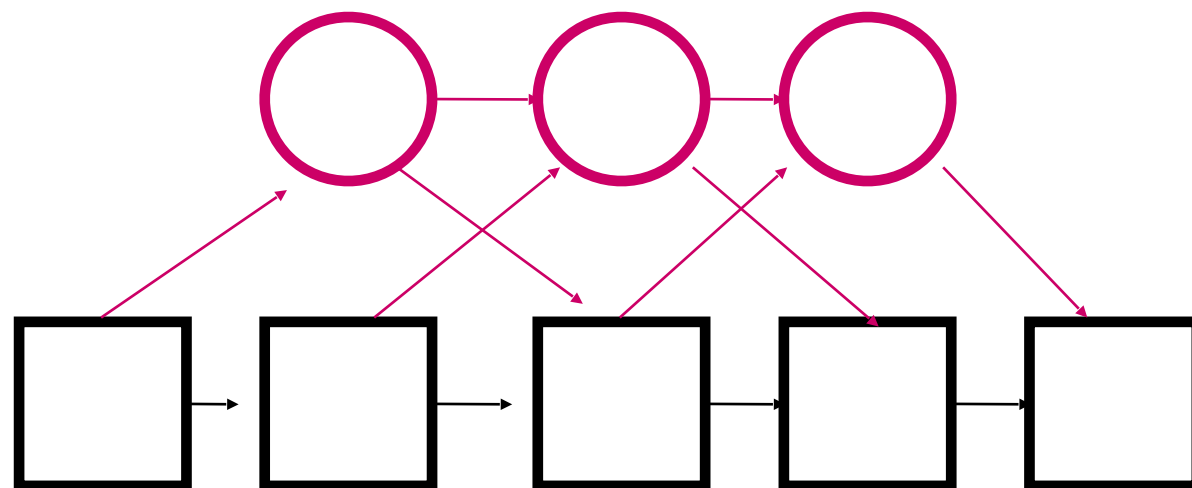
C  
M<sub>5</sub>

# Silent States

Different topology to model gaps



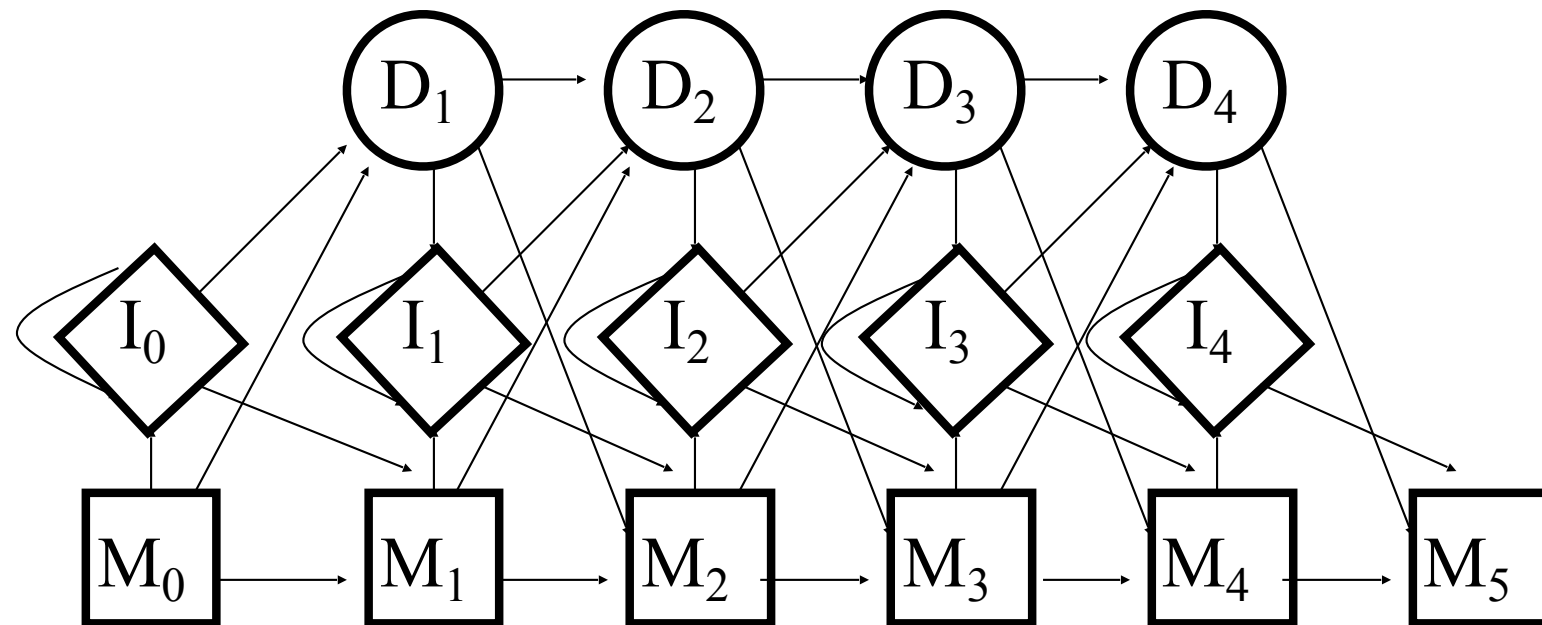
$N(N-1)/2$  transitions



To reduce the number of parameters we can use states that doesn't emit any character  
 $4N-8$  transitions



# Profile HMM



Delete states

Insert states

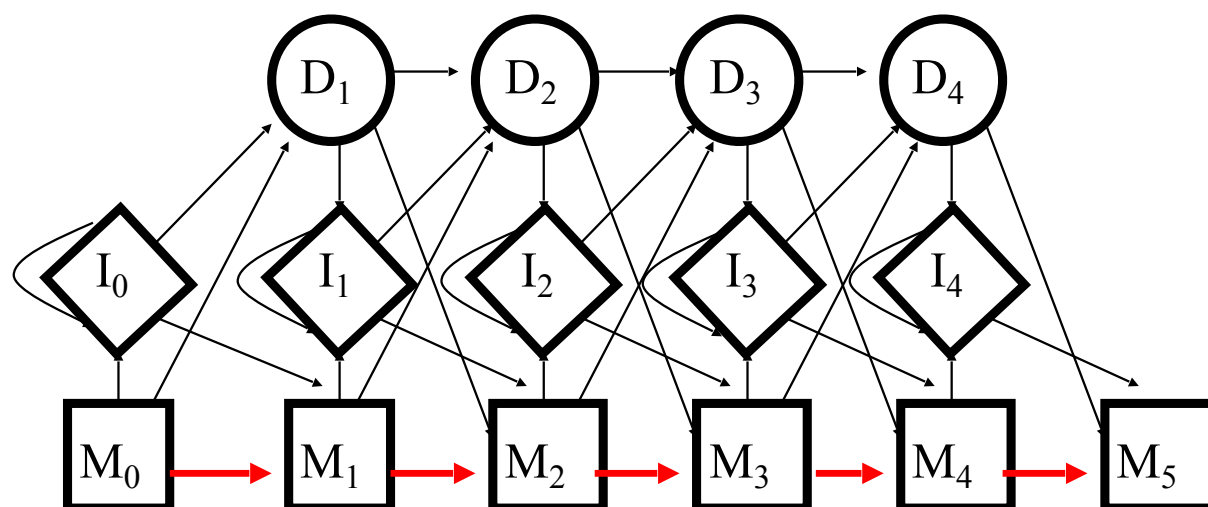
Match states

<b>A</b>	<b>C</b>	<b>G</b>	<b>G</b>	<b>T</b>	<b>A</b>
M <sub>0</sub>	M <sub>1</sub>	M <sub>2</sub>	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>

<b>A</b>	<b>C</b>	<b>G</b>	<b>C</b>	<b>A</b>	<b>G</b>	<b>T</b>	<b>C</b>
M <sub>0</sub>	I <sub>0</sub>	I <sub>0</sub>	M <sub>1</sub>	M <sub>2</sub>	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>

<b>A</b>		<b>G</b>	<b>A</b>	<b>T</b>	<b>C</b>
M <sub>0</sub>	D <sub>1</sub>	M <sub>2</sub>	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>

# Example of Alignment



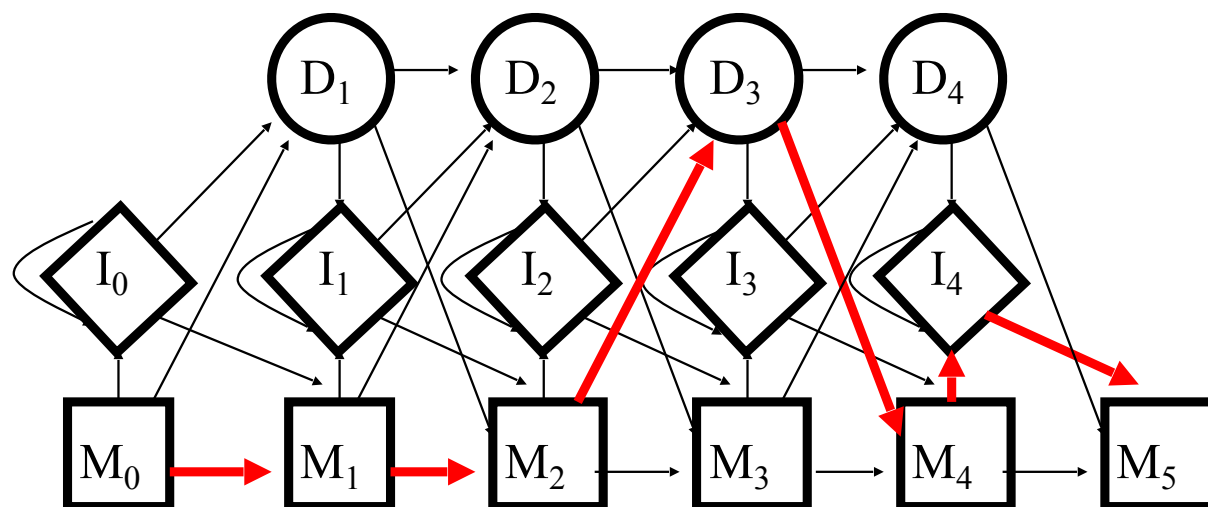
*Sequence 1*

**A S T R A L**

*Viterbi path*

**M<sub>0</sub> M<sub>1</sub> M<sub>2</sub> M<sub>3</sub> M<sub>4</sub> M<sub>5</sub>**

**A S T R A L**



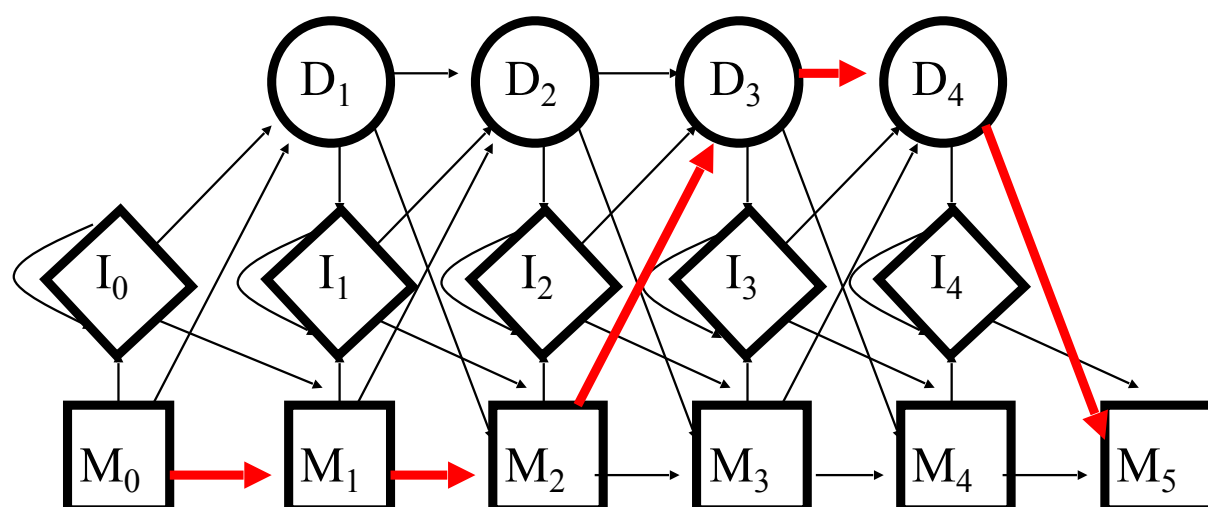
*Sequence 2*

**A S T A I L**

*Viterbi path*

**M<sub>0</sub> M<sub>1</sub> M<sub>2</sub> D<sub>3</sub> M<sub>4</sub> I<sub>4</sub> M<sub>5</sub>**

**A S T A I L**



*Sequence 3*

**A R T I**

*Viterbi path*

**M<sub>0</sub> M<sub>1</sub> M<sub>2</sub> D<sub>3</sub> D<sub>4</sub> M<sub>5</sub>**

**A R T I**

# Alignment Calculation

$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$		<i>Sequence 1</i>
A	S	T	R	A	L		
$M_0$	$M_1$	$M_2$	$D_3$	$M_4$	$I_4$	$M_5$	<i>Sequence 2</i>
A	S	T		A	I	L	
$M_0$	$M_1$	$M_2$	$D_3$	$D_4$	$M_5$		<i>Sequence 3</i>
A	R	T			I		

Grouping by vertical layers

	0	1	2	3	4	5
$s_1$	A	S	T	R	A	L
$s_2$	A	S	T		AI	L
$s_3$	A	R	T			I

Alignment

ASTRA-L  
AST-AIL  
ART---I

$-\log P(s | M)$  Is an alignment score

# Alignment of Globins

```

AAAAAAAAAAAAAAAAAAAA    BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
                                DDDD
-----VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF--DL
-----VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESFGDL
-----VLSEGEWQLVLHVWAKVEA--DIAGHGQDILIRLFKHHPETLEKFDREFKHL
-----LSADQISTVQASFDKVKG-----DPVGILYAVFKADPSIMAKFTQFAG-
PIVDTGSVAPLSAAEKTAKRSAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGL
-----GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-FLK-
-----GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-FSG-

```

```

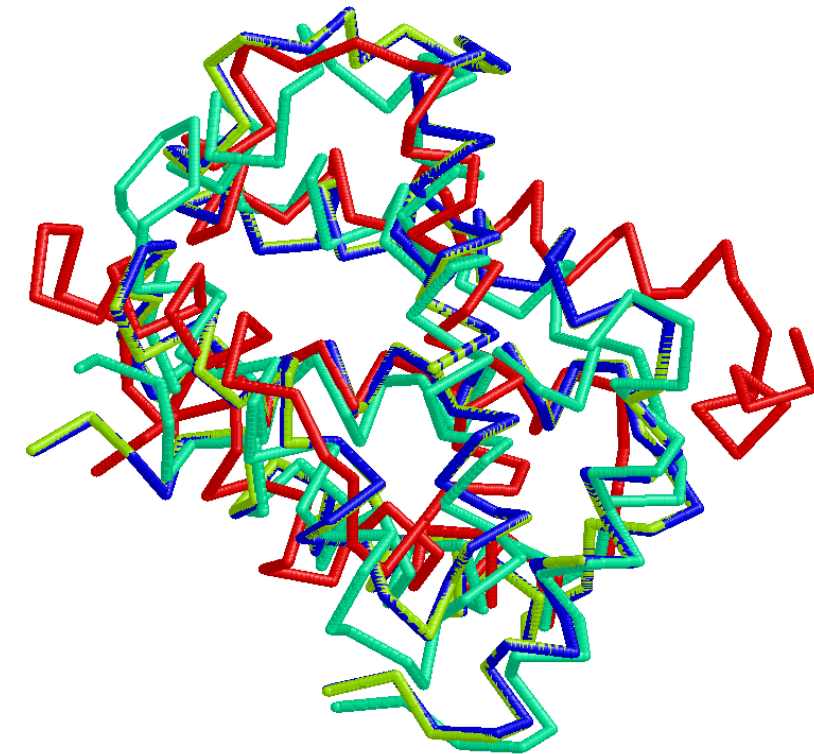
DDDDDDDDDEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
                                F          GG  GG
S-----HGSAQVKGHGKKVADALTNAVAHV--D--DMPNALSALSDLHAHKL--RVDPV
STPDVAVMGNPKVKAHGKKVLGAFSDGLAHL--D--NLKGTATLSELHCDKL--HVDPE
KSEAEMKASEDLKKHGVTVLTAIGAILKK---K-GHHEAELKPLAQSHATKH--KIPIK
KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG--VTHD
TTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF--QVDPQ
GTSEVPQNNPELQAHAGKVFCLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG---VADA
---AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKG YGNKHIKAQ

```

```

GGGGGGGGGGGGGGGGGGGG    HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR-----
NFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
YLEFISEAIIHVLHSRHPADFGADAQGAMSKALELFRKDIAAKYKELGYQG
QLNNFRAGFVSVMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-----
YFKVLA AVIADTVAAG-----DAGFEKLMSMICILLRSAY-----
HFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
YFEPLGASLLSAMEHRIGGKMNAAKDAWAAAYADISGALISGLQS-----

```



# Globins HMM

HMM are calculate from a training set of 400 unaligned sequences. After the HMM is built, it is used to obtain a multiple alignment of all the training sequences. This is the alignment of the 7 globins as aligned with the trained model.

```

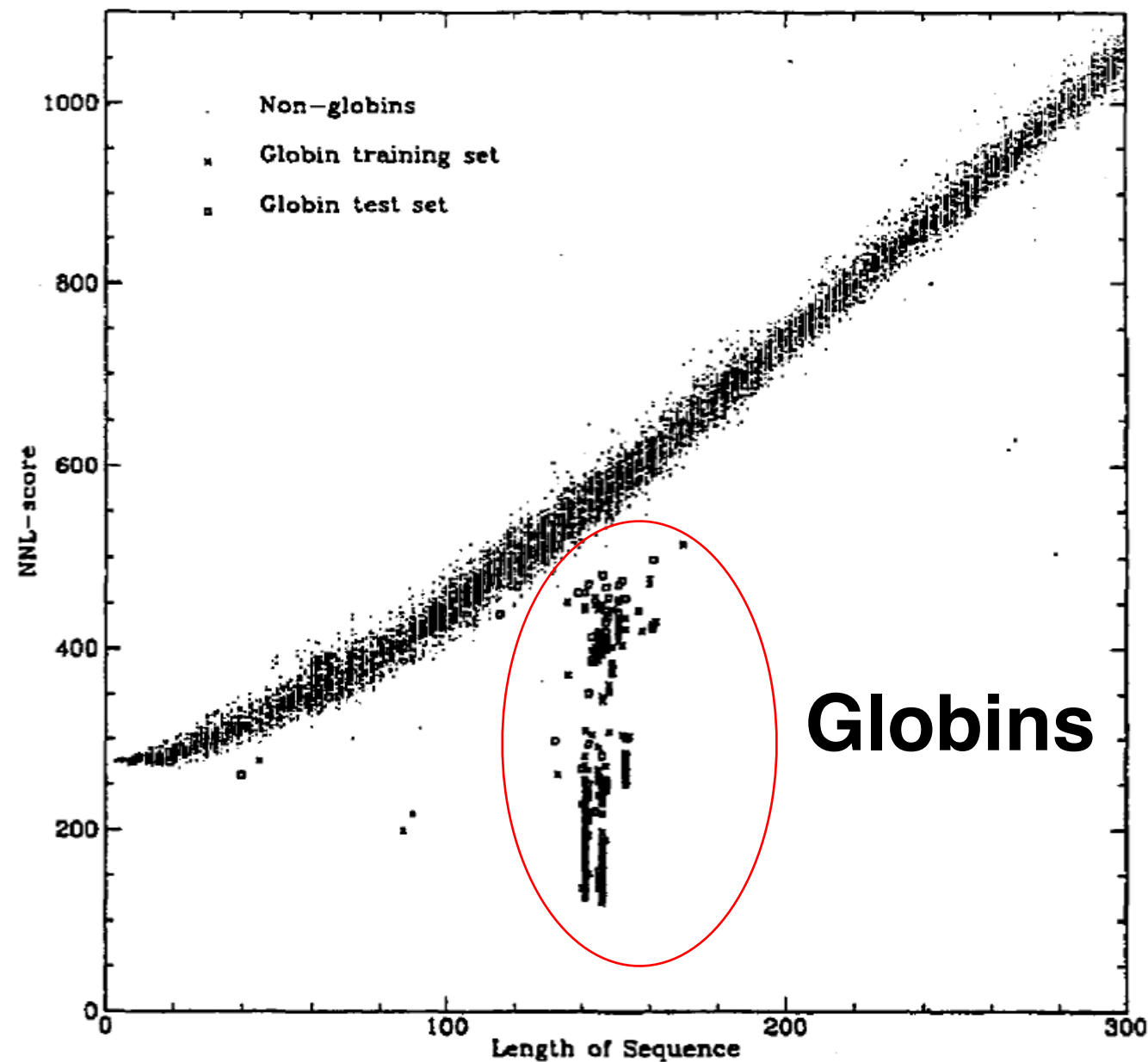
                AAAAAAAAAAAAAAAAAA   BBBB BBBB BBBB BBBB BBBBBB CCCCCCCCCCCC
                                   DDDD
                *****               *****
V.....LSPADKTNVKA AWGKVGA..HAGEYGAEALERMFLSFPTTKTYFPHFD-L
Vh.....LTPEEKSAVTALWGKV--..NVDEVGGEALGRLLVVYPWTQRFFESFGDL
V.....LSEGEWQLVHLVWAKVEA..DVAGHGQDILIRL FKSHPETLEKFD RFKHL
-.....LSADQISTVQASFDKV--..KGDPVGI--LYAVFKADPSIMAKFTQFAGK
PivdtgsvapLSAAEKT KIRSAWAPVYS..TYETSGVDILVKFFTSTPAAQEFPKFKGL
Ga.....LTESQAALVKSSWEEFN A..NIPKHTHRFFILVLEIAPAAKDLFSFLK-G
G.....LSAAQRQVIAATWKDIAGadNGAGVGKDCLIKFLSAHPQMA---AVFG-F

DDDDDDDEE  EEEEEEEEEEEEEEEEEEEEE
                                   F
                *****               *****
SHGSAQVKGH-GKK.----VADALTNAVAHVDD.....MPNALSALSDLHA...HKLRVD
STPDVVMGNPKVKA.HGKKVLGAFSDGLAHLDN.....LKGTFATLSELHC...DKLHVD
KTEA-EMKASEDLKkHGVTVL TALGAILKKKGH.....HEAELKPLAQSHA...TKHKIP
DLES-IKGTAPFET.HANRIVGFFSKIIGELPN.....IEADVNTFVASHK...PR-GVT
TTADQLKKSADVRW.HAERIINAVNDAVASMDDtek..MSMKLRDL SGKHA...KSFQVD
TSEVPQ-NNPELQA.HAGKVFKLVYEAAIQ LQVtgvvvT DATLKNLGSVHV...SK-GVA
SGAS----DPGVAA.LGAKVLAQIGVAVSHLGDegk..MVAQMKAVGV RHKgygNK-HIK

GGGGGGGGGGGGGGGGGGGGGGG   HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
*****
PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLT SKY.....R
PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKV VAGVANALAHKY.....H
IKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYkelgyqG
HDQLNNFRAGFVS YMKAH--TDF-AGAEAAWGATLD TFFGMIFSKM.....-
PQYFKVLAAVIADTVAA---GD-----AGFEKLMSMICILLRSAY.....-
DAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMnda...A
AQYFEPLGASLLSAMEHRIGGKMNA AAKDAWAAAYADISGALISGLq.....S
```

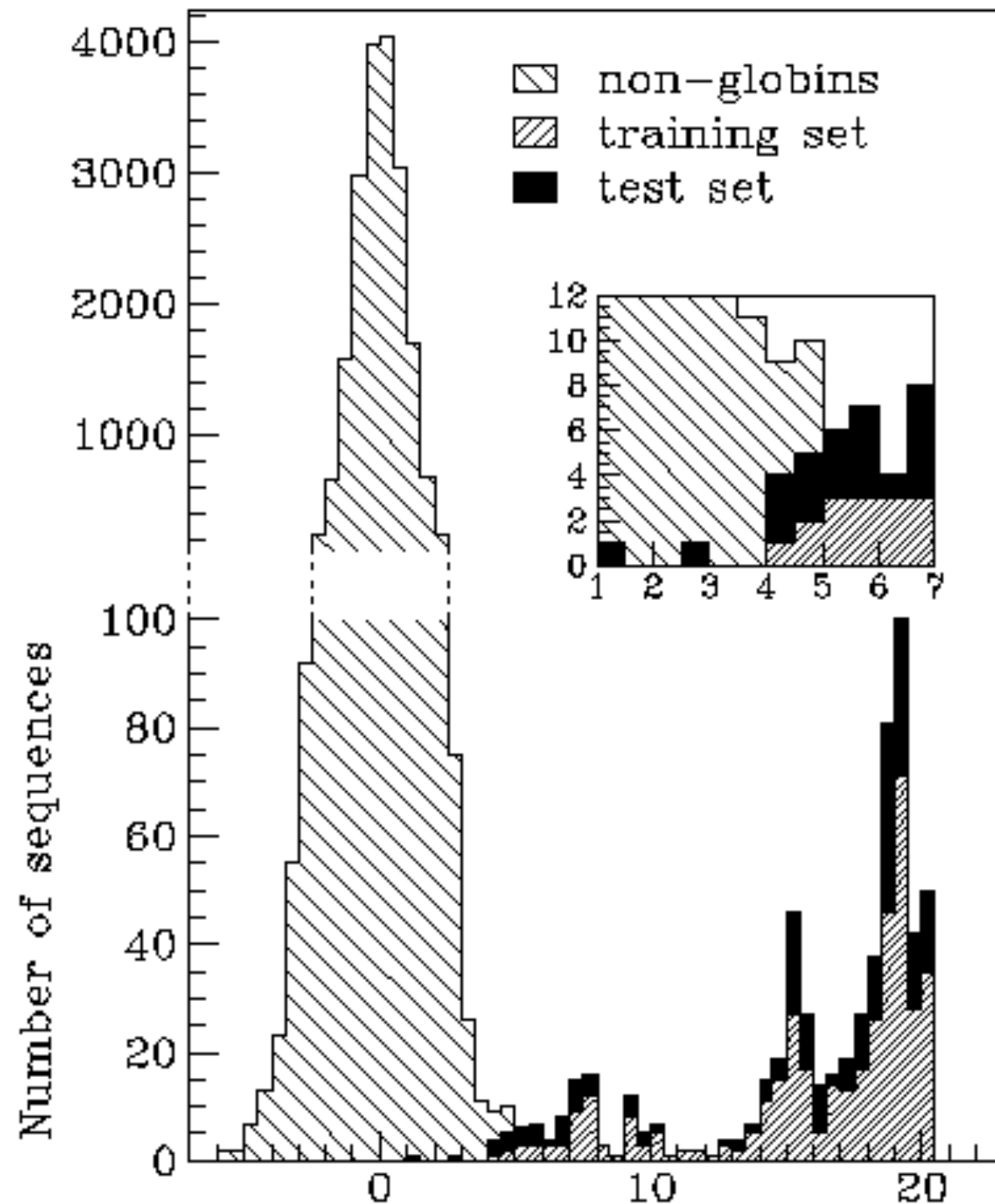
# Globin Classification

The NLL-score is calculated to discriminate between Globin and non-Globin protein sequences



$$\text{NLLscore} = -\log P(\text{sIM})$$

# Score distribution



$$\text{Z-score} = \frac{\text{NLL}(s) - \langle \text{NLL} \rangle}{\sigma(\text{NLL})}$$

With mean and standard deviation  
computed on sets of sequences with  
similar length



# Confusion Matrix

A 2x2 matrix for calculating the performance of prediction methods

		Condition (as determined by "Gold standard")	
		Condition positive	Condition negative
Test outcome	Test outcome positive	True positive	False positive (Type I error)
	Test outcome negative	False negative (Type II error)	True negative



# Overall Accuracy

How many predictions are correct on the overall?

*Accuracy (ACC):*

$$ACC = \frac{(TP + TN)}{(TP + FN + TN + FP)}$$

Is it an informative enough score?

# Dataset Unbalance

Accuracy can be strongly biased because of class unbalance. It is not very informative

	Class 1	Class -1
Prediction 1	90	10
Prediction -1	0	0

Acc = 0.9

ALL the examples are predicted in the class 1:

Very bad predictions

	Class 1	Class -1
Prediction 1	81	1
Prediction -1	9	9

Acc = 0.9

It seems a much more reasonable prediction

# Class Specific Measures

*Sensitivity (Sn) or  
True Positive Rate  
(TPR):*

$$Sn = \frac{TP}{TP+FN}$$

It answer to the question:

How many of the real positive examples  
are correctly predicted?

*Precision or Positive  
Predictive Value (PPV):*

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

It answer to the question:

How many of the positive predictions are correct?

It is sometimes referred as Specificity

# Matthews Correlation

*Matthews Correlation Coefficient (MCC):*

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

It answer to the question:

Is the prediction really correlated with the real classes?

It is 0 in case of random prediction

It is 1 only in case of perfect prediction

It is -1 only in case of completely wrong prediction

It is the Pearson's correlation coefficient for categorical classes

# MCC and Unbalance

MCC is not affected by dataset unbalance

	Class 1	Class -1
Prediction 1	90	10
Prediction -1	0	0

Acc = 0.9

All the examples are predicted in the class 1:

MCC = 0.0

Very bad predictions

	Class 1	Class -1
Prediction 1	81	1
Prediction -1	9	9

Acc = 0.9

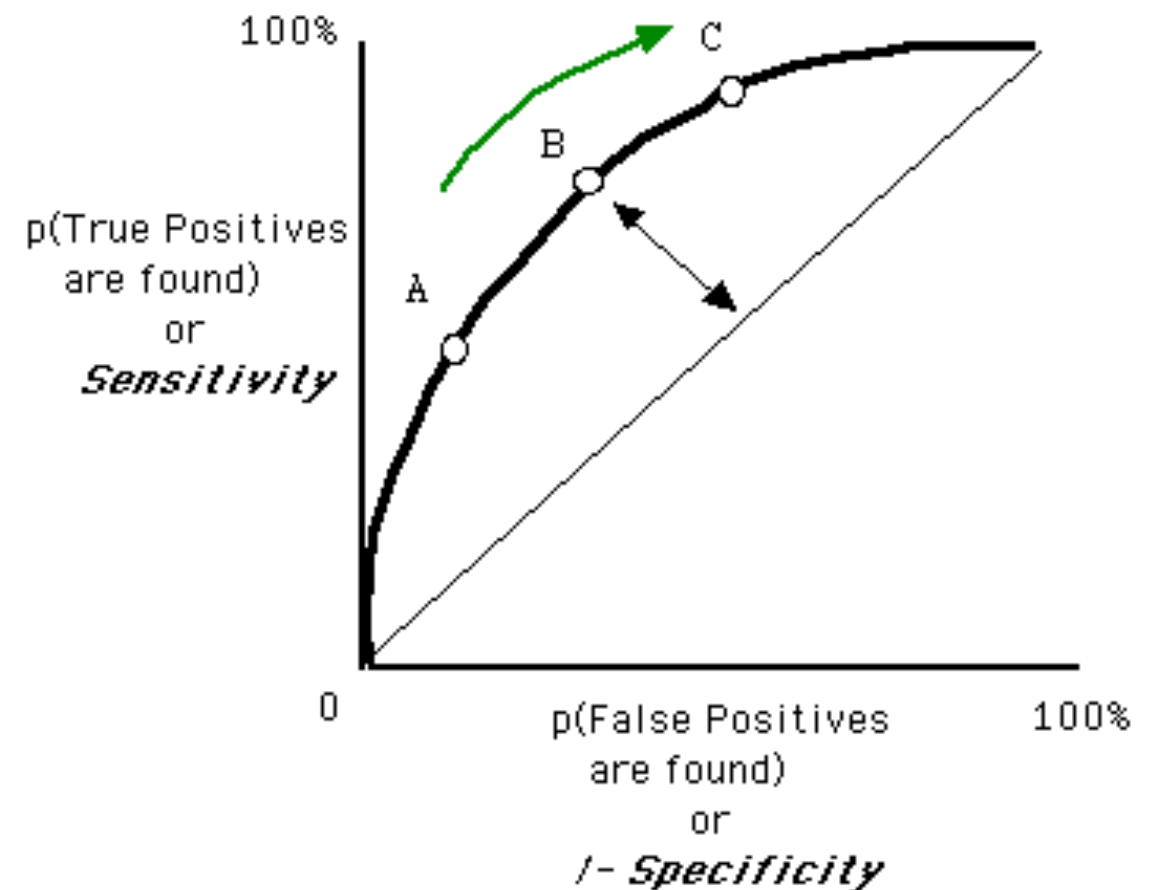
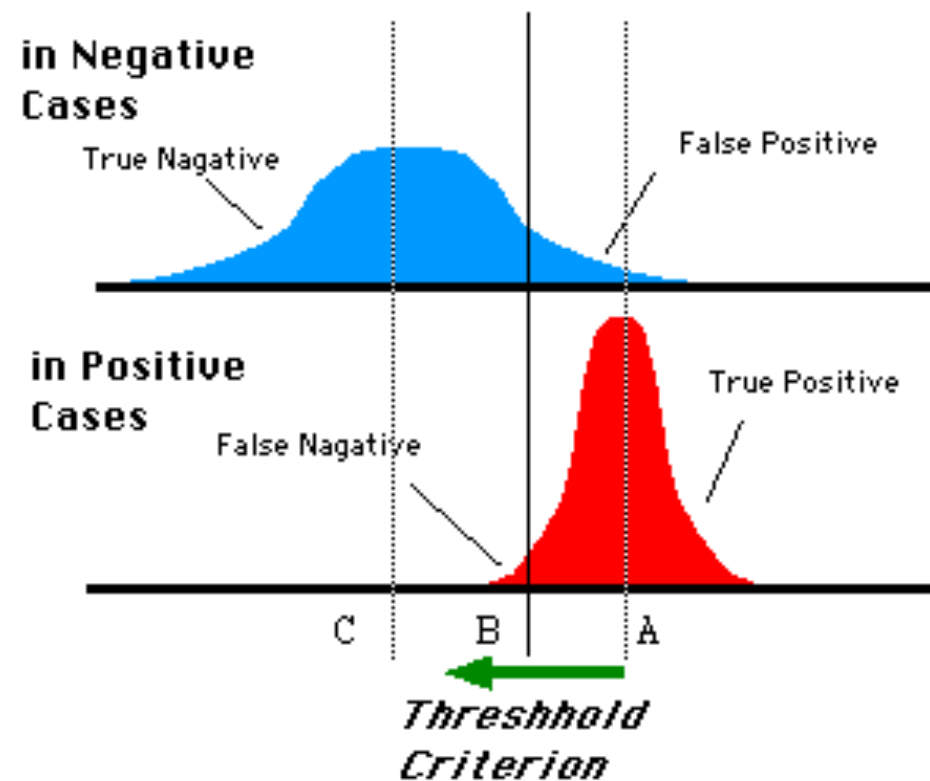
MCC = 0.62

Predictions are good

# ROC Curve

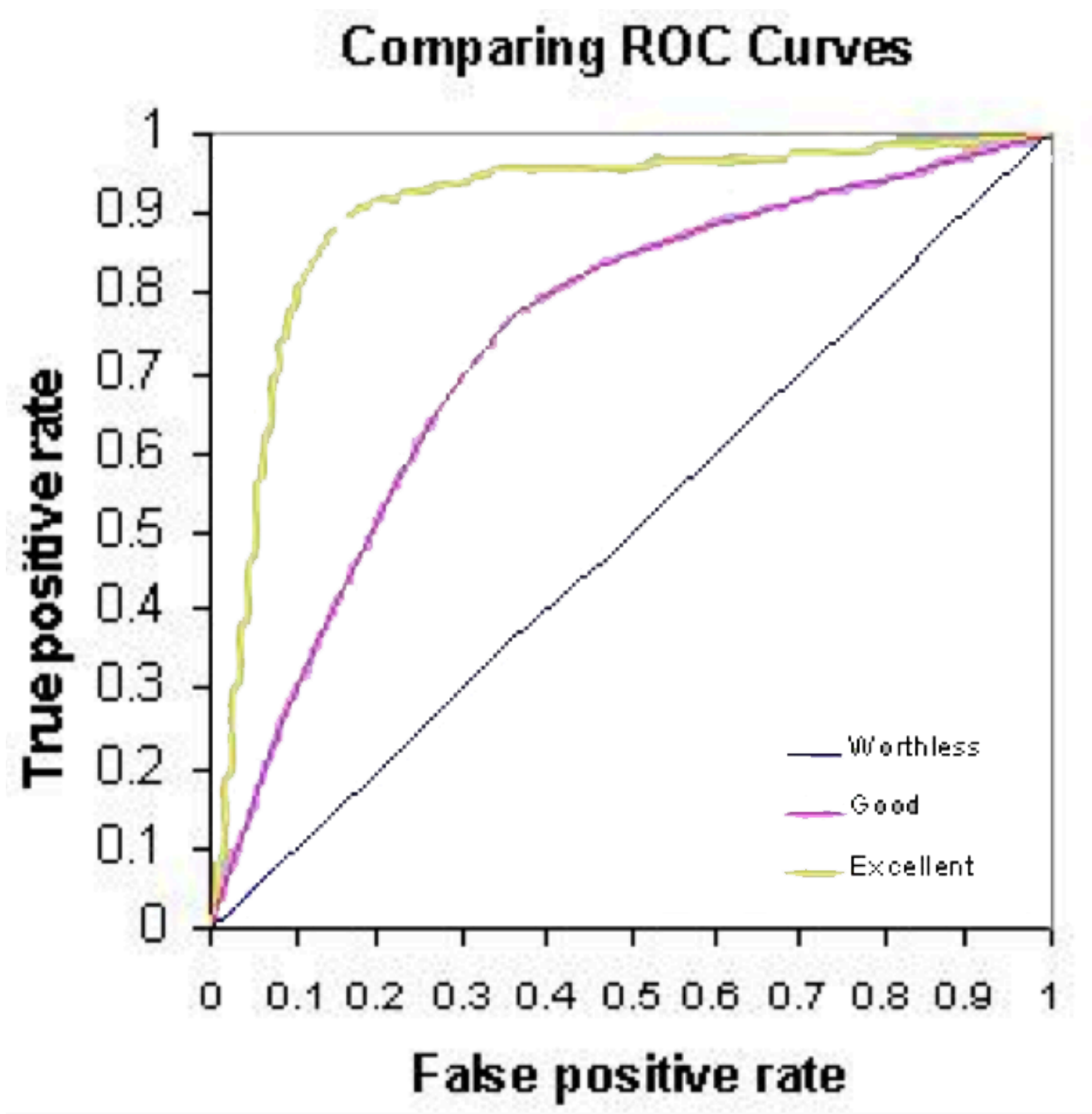
The Receiver Operating Characteristics depends on a parameter, TPR and FPR can be plotted at varying values of the parameter

Distributions of the Observed signal strength



# Area Under Curve

The Area Under the ROC Curve (AUC) is used to measure the performance of a predictor



AUC=0.5 → Random prediction

AUC=1 → Perfect prediction