

# Hidden Markov Models for Sequence Alignment

Laboratory of Bioinformatics I  
Module 2

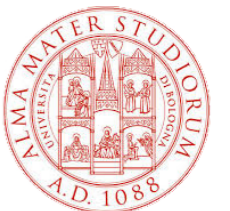
**Emidio Capriotti**

<http://biofold.org/>



**Biomolecules  
Folding and  
Disease**

Department of Pharmacy and  
Biotechnology (FaBiT)  
University of Bologna



# Alignment of Globins

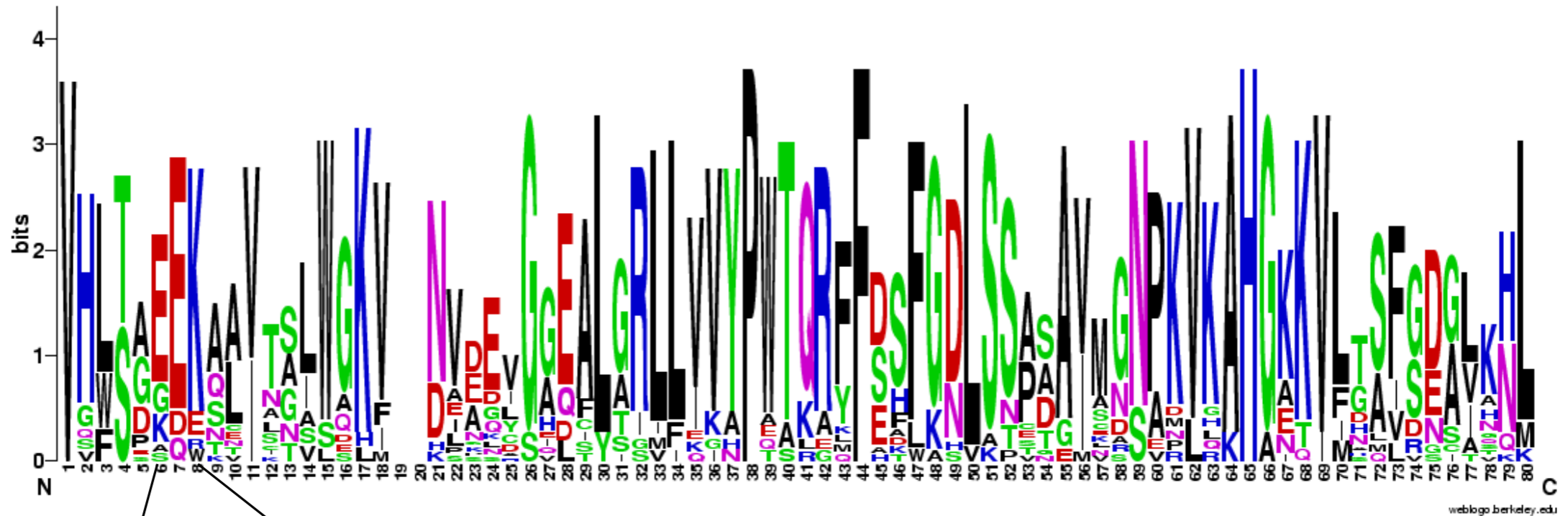
Different positions are not equivalent

```

      10      20      30      40      50      60      70      80
lqb1 pea/1-471 -GFTDKQQEALVMSSSE-FKQNLPGYSILFYTIVLEKAPAAKGLFSFLKD---TAGVEDSPKLOAHAEQVFGLVRDSAAQL
lqb1 vicfa/1-471-GFTEKQQEALVMSSSSQLFKQNPSNYSVLFYTIILQKAPTAKAMFSFLKD---SAGVVDSPKLGAHAEKVFGMVRDSAVQL
hbb speci/1-471 VHLSDGEKNAISTAWGKV--HAAEVGAEALGRLLVVYPWTQRFFDSFGDLSSASAVMGNAKVKAHGKKVIDSFSNGLKHL
hbb speto/1-471 VHLTDGEKNAISTAWGKV--NAAEIGAEALGRLLVVYPWTQRFFDSFGDLSSASAVMGNAKVKAHGKKVIDSFSNGLKHL
hbb equhe/1-471 VQLSGEEKAAVLALWDKV--NEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPAAVMGNPKVKAHGKKVLHSFGEGVHHL
hbb sunmu/1-471 VHLSGEEKACVTGLWGKV--NEDEVGAEALGRLLVVYPWTQRFFDSFGDLSSASAVMGNPKVKAHGKKVLHSLGEGVANL
hbb tupql/1-471 VHLSGEEKAAVTGLWGKV--DLEKVGGQSLGSLLIVVYPWTQRFFDSFGDLSSPSAVMSNPKVKAHGKKVLTSFSDGLNHL
hbb calar/1-471 VHLTGEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHL
hbb mansp/1-471 VHLTPEEKTAVTTLWGKV--NVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHL
hbb rabbit/1-471 VHLSSEEKSAVTALWGKV--NVEEVGGEALGRLLVVYPWTQRFFESFGDLSSANAVMNNPKVKAHGKKVLAAFSEGLSHL
hbb ursma/1-471 VHLTGEEKSLVTGLWGKV--NVDEVGGEALGRLLVVYPWTQRFFDSFGDLSSADAIMNNPKVKAHGKKVLNSFSDGLKNL
hbb triin/1-471 VHLTPEEKALVIGLWAKV--NVKEYGGEALGRLLVVYPWTQRFFEHFGDLSSASAIMNNPKVKAHGEKVFTSFGDGLKHL
hbb ornan/1-471 VHLSGGEKSAVTNLWGKV--NINELGGEALGRLLVVYPWTQRFFEAFGDLSSAGAVMGNPKVKAHGAKVLTSFGDALKNL
hbb tacac/1-471 VHLSGSEKTAVTNLWGHV--NVNELGGEALGRLLVVYPWTQRFFESFGDLSSADAVMGNAKVKAHGAKVLTSFGDALKNL
hbe ponpy/1-471 VHFTAEEKAAVTSLWSKM--NVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKVKAHGKKVLTSFGDAIKNM
hbb colli/1-471 VHWSAEEKQLITSIWGKV--NVADCGAEALARLLIVVYPWTQRFFSSFGNLSSATAISGNPNVKAHGKKVLTSFGDAVKNL
hbb larri/1-471 VHWSAEEKQLITGLWGKV--NVADCGAEALARLLIVVYPWTQRFFASFGNLSSPTAINGNPMVRAHGKKVLTSFGEAVKNL
hbb1 varex/1-471VHWTAEEKQLICSLWGKI--DVGLIGGETLAGLLVIVYPWTQRQFSHFGNLSSPTAIAGNPRVKAHGKKVLTSFGDAIKNL
hbb2 xentr/1-471VHWTAEEKATIASVWGKV--DIEQDGHDALSRLLVVYPWTQRYFSSFGNLSNVSAVSGNVKVKAHGNKVLSAVGSAIQHL
hbb1 ranca/1-471VHWTAEEKAVINSVWQKV--DVEQDGHEALTRLFIVVYPWTQRYFSTFGDLSSPAAIAGNPKVHAHGKKILGAIDNAIHNL
hbb2 tricr/1-471VHLTAEDRKEIAAILGKV--NVDSLGGQCLARLIVVNPWSRRYFHDFGDLSSCDAICRNPKVLAHGAKVMRSIVEATKHL
hba4 salir/1-471-SLSAKDKANVKAIWGKILPKSDEIGEQALSRMLVVYPQTKAYFSHWASVAP-----GSAPVKKHGITIMNQIDDCVGHM
myg_escgi/1-471 -VLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASEDLKKHGNTVLTALGGILKKK
```

# Sequence Logo

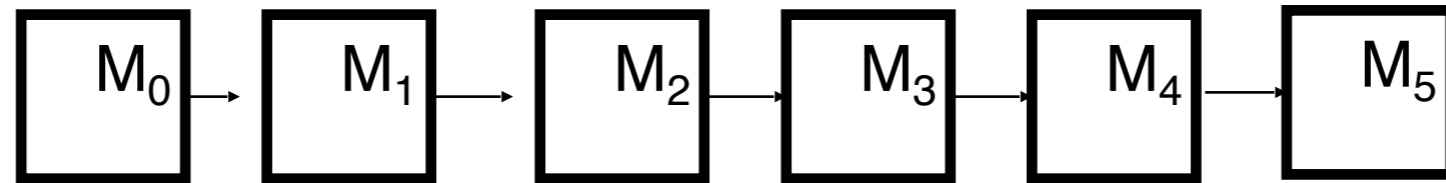
A more flexible alignment score is needed to align protein families



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
$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$

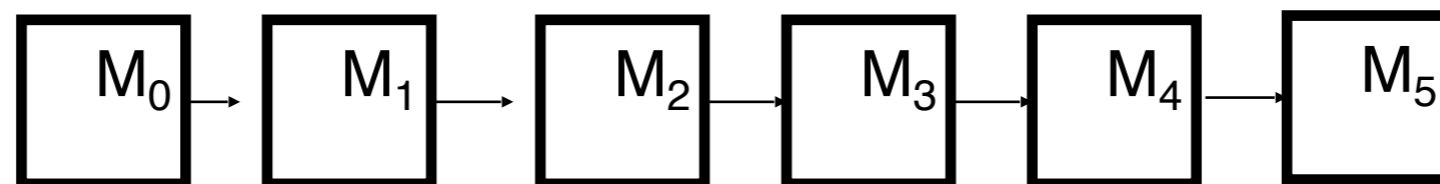
A	T	G	T	T	C
$M_0$	$M_1$	$M_2$	$M_3$	$M_4$	$M_5$

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

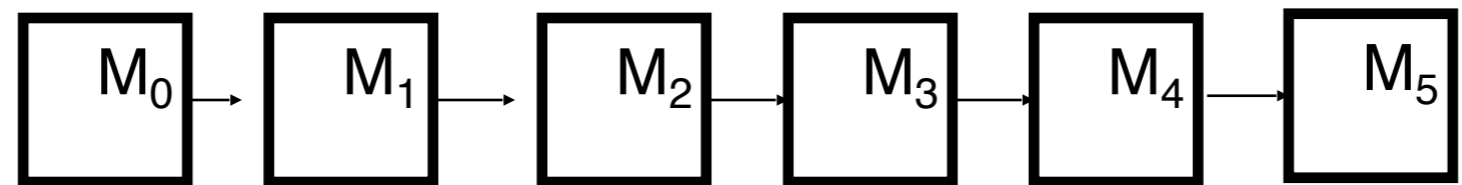
A C G G T A  
A C G A T C  
A T G T T C



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



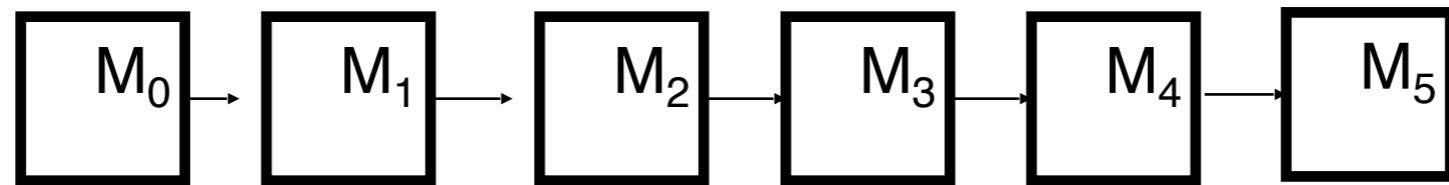
A	1	0	0	0.33	0	0.33
C	0	0.66	0	0	0	0.66
G	0	0	1	0.33	0	0
T	0	0.33	0	0.33	1	0

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

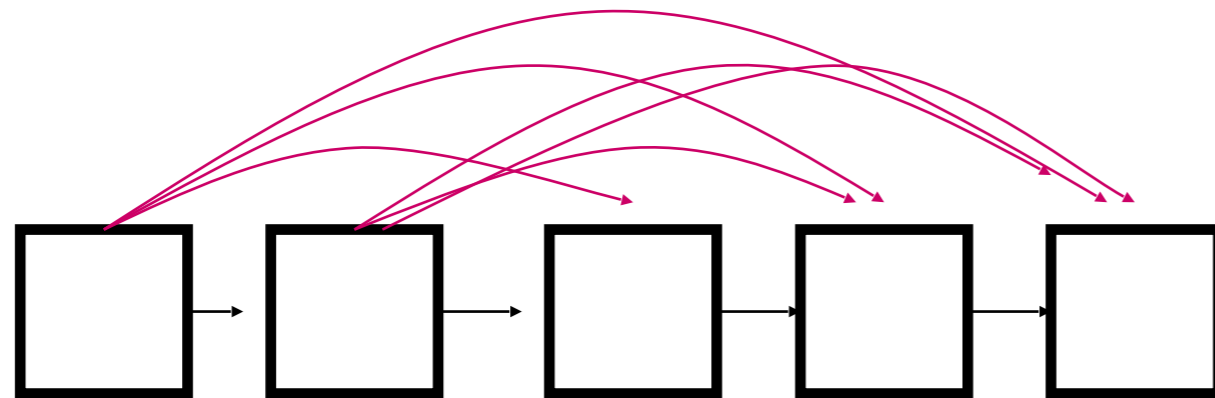


A	1	0	0	0.33	0	0.33
C	0	0.66	0	0	0	0.66
G	0	0	1	0.33	0	0
T	0	0.33	0	0.33	1	0

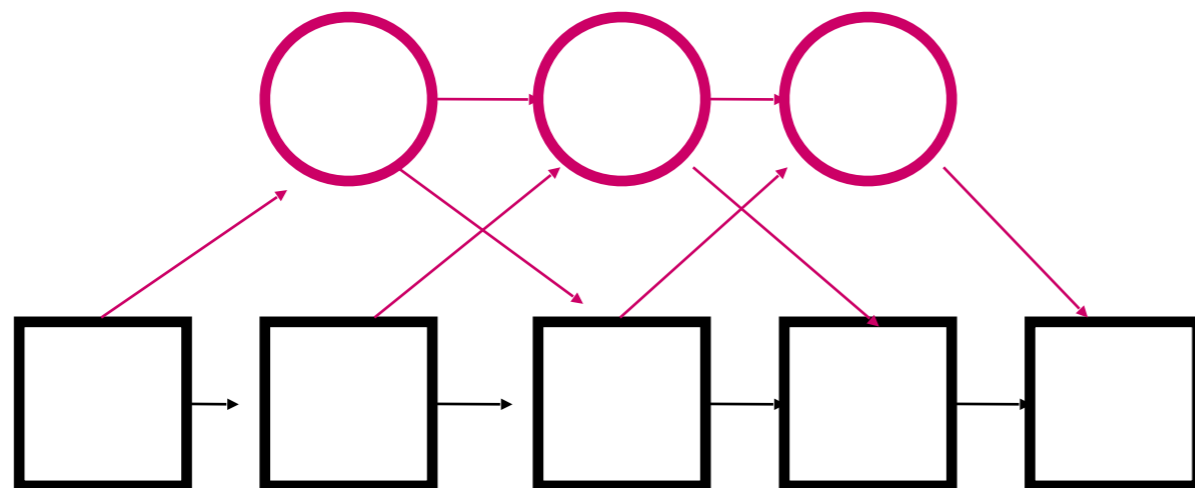
A		G	A	T	C
M <sub>0</sub>	M <sub>2</sub>	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>	M <sub>5</sub>

# Silent States

Different topology to model gaps



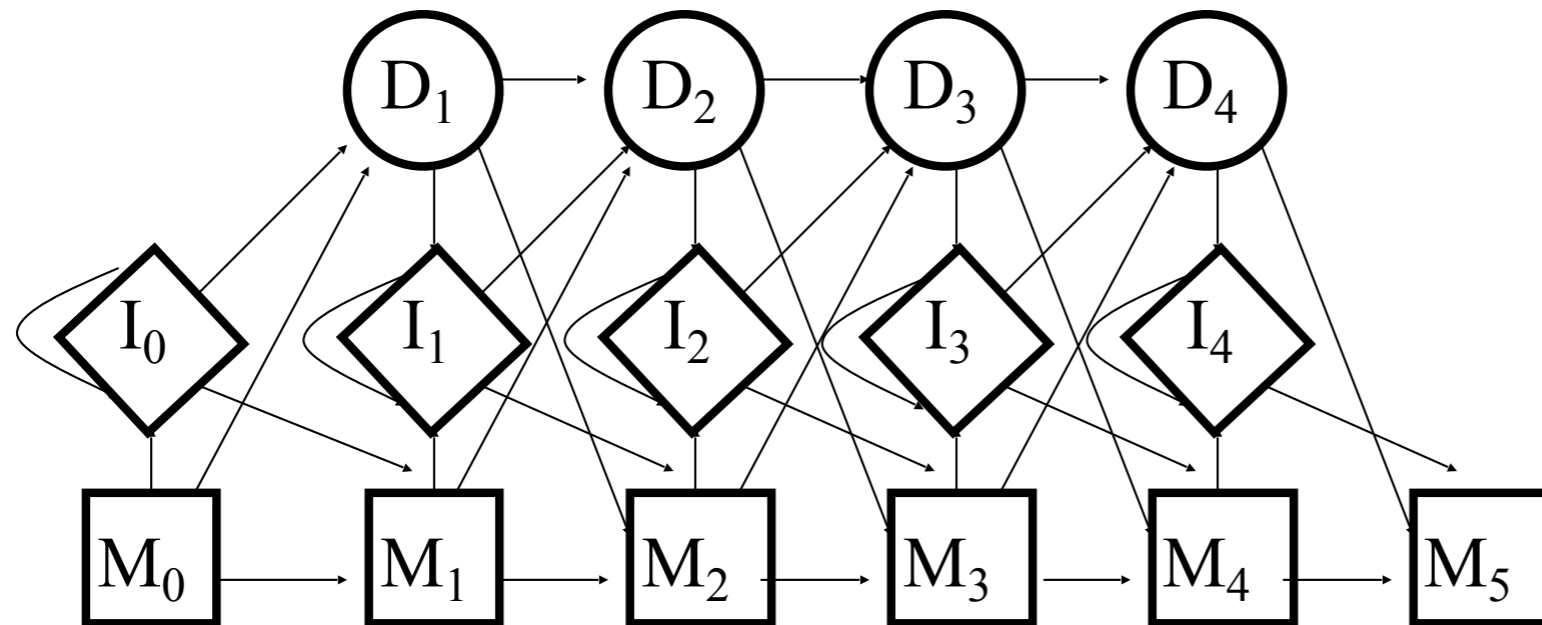
$N(N-1)/2$  transitions



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

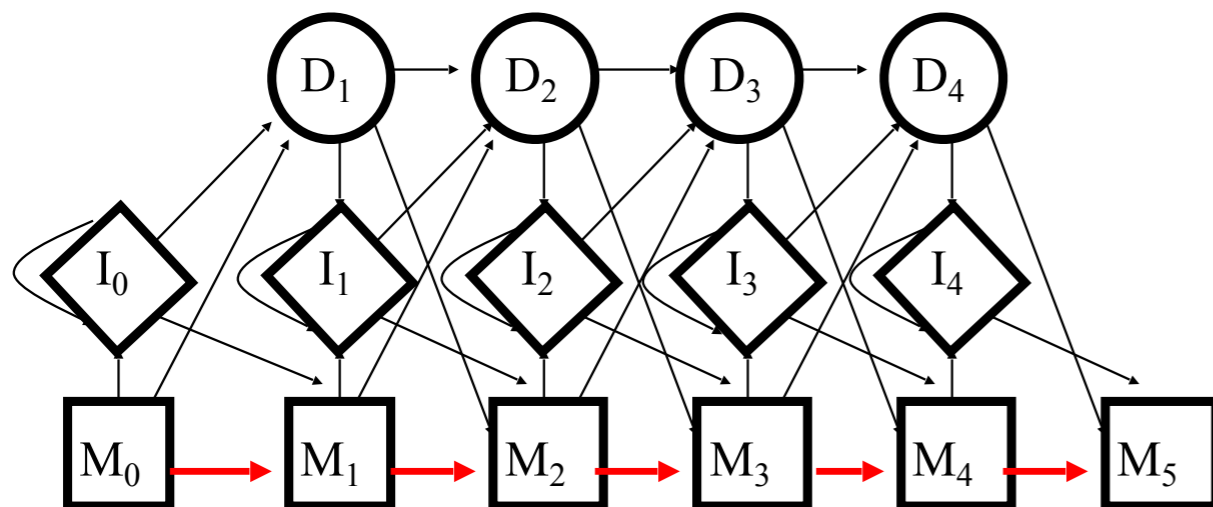
Match states

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

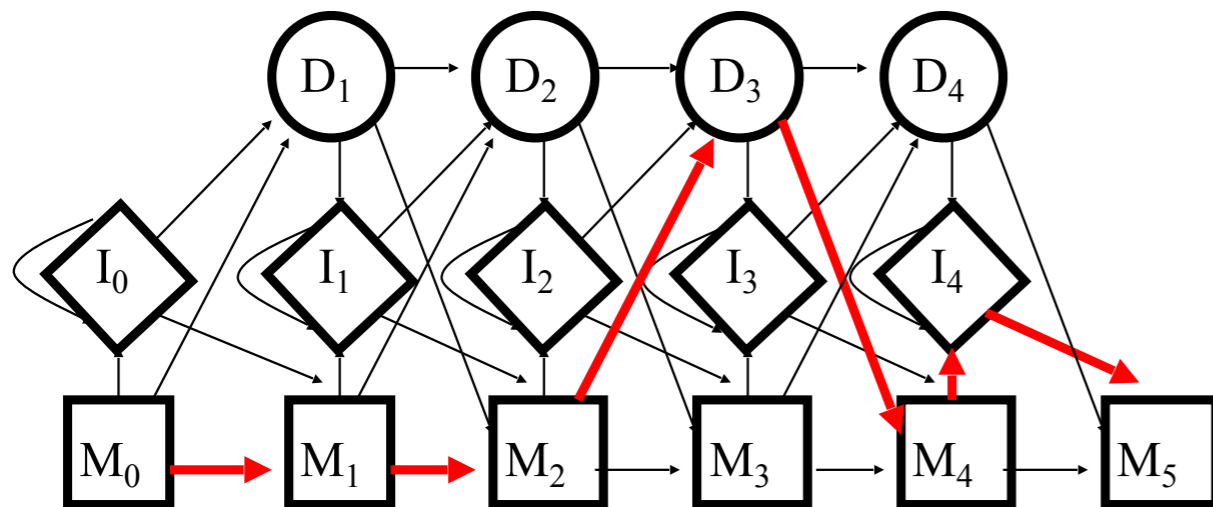
**A C G**    **C**            **A**            **G**            **T**            **C**  
 M<sub>0</sub> I<sub>0</sub> I<sub>0</sub>    M<sub>1</sub>          M<sub>2</sub>          M<sub>3</sub>          M<sub>4</sub>          M<sub>5</sub>

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

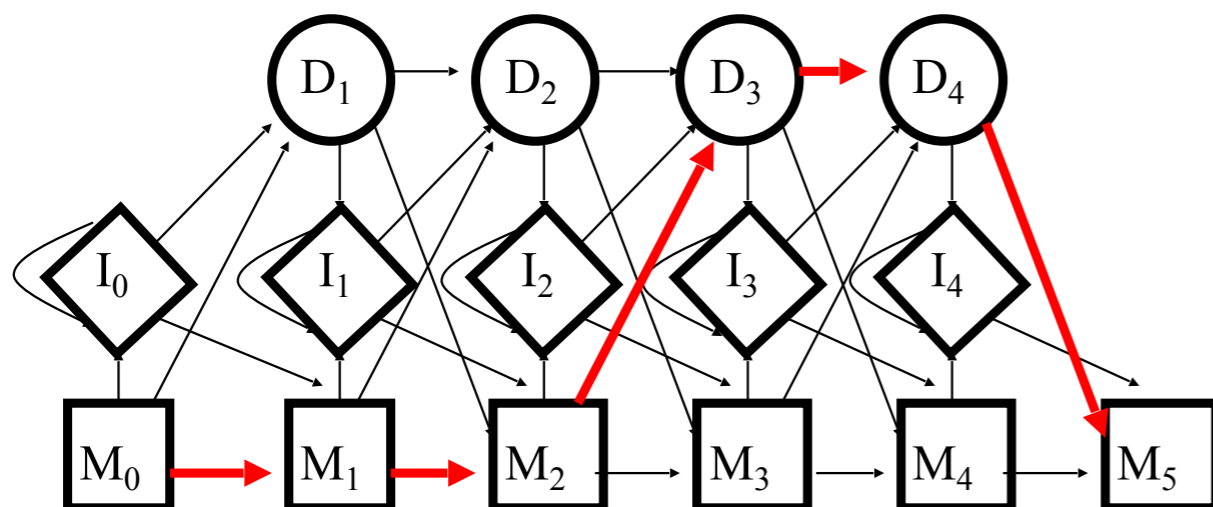
# Example of Alignment



**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

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



$-\text{Log } P(s | M)$  Is an alignment score

# Alignment of Globins

```

AAAAAAAAAAAAAAAAAAAA    BBBBCCCCCCCCCCCCCCCC
                                DDDD
-----VLSPADKTNVKAANGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF-DL
-----VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESFGDL
-----VLSEGEWQLVLHVWAKVEA--DIAGHGQDILIRLFKHHPETLEKFDREFKHL
-----LSADQISTVQASFDKVKG-----DPVGILYAVFKADPSIMAKFTQFAG-
PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGL
-----GALTESQAALVKSSWEEFNA--NIPKHTRFFILVLEIAPAAKDLFS-FLK-
-----GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-FSG-

```

```

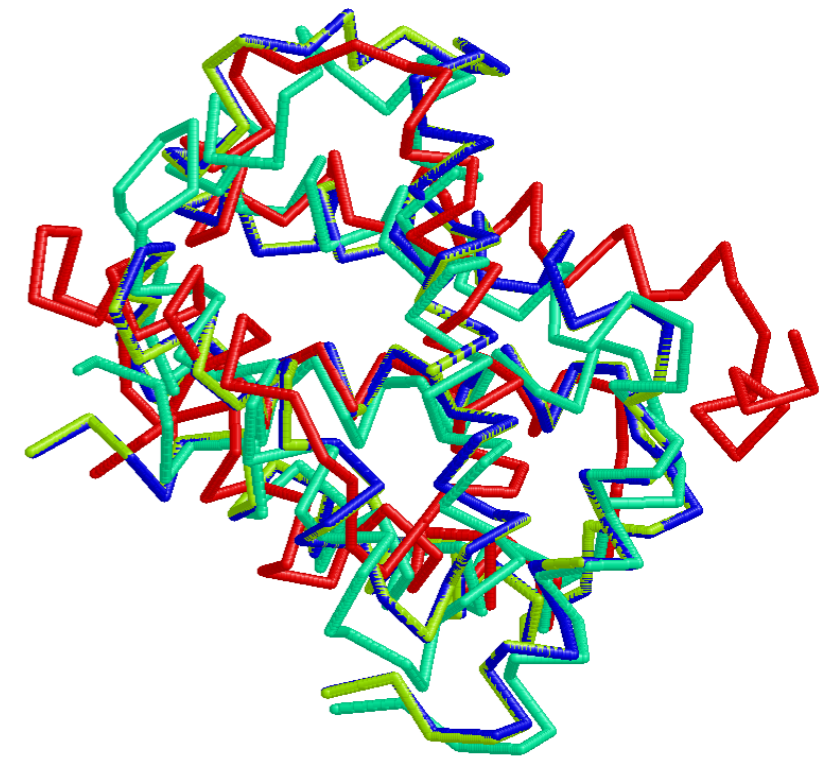
DDDDDDDEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
                                F          GG  GG
S-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL--RVDPV
STPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGT FATLSELHCDKL--HVDPE
KSEAEMKASEDLKKGHTVLTALGAILKK---K-GHHEAELKPLAQSHATKH--KIPIK
KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG---VTHD
TTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF--QVDPQ
GTSEVPQNNPELQAHAGKVFKLVEYAAIQVLTG VVVTDATLKNLGSVHVSKG---VADA
---AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKG YGNKHIKAQ

```

```

GGGGGGGGGGGGGGGGGGGG    HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
NFKLLSHCLLVTLAAHLPAEFTP AVHASLDKFLASVSTVLT SKYR-----
NRLLGNVLVCVLAHHFGKEFTP PVQAAYQKV VAGVANALAHKYH-----
YLEFISEAIIHVLHSRHPADFGADAQGAMSKALELFRKDIAAKYKELGYQG
QLNNFRAGFVS YMKAHT--DFA-GAEAANGATLDTFFGMIFSKM-----
YFKVLA AVIADTVAAG-----DAGFEKLMSMICILLRSAY-----
HFPVVK EAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
YFEPLGASLLSAMEHRIGGKMNA AAKDAWAAAYADISGALISGLQS-----

```



# 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.

```

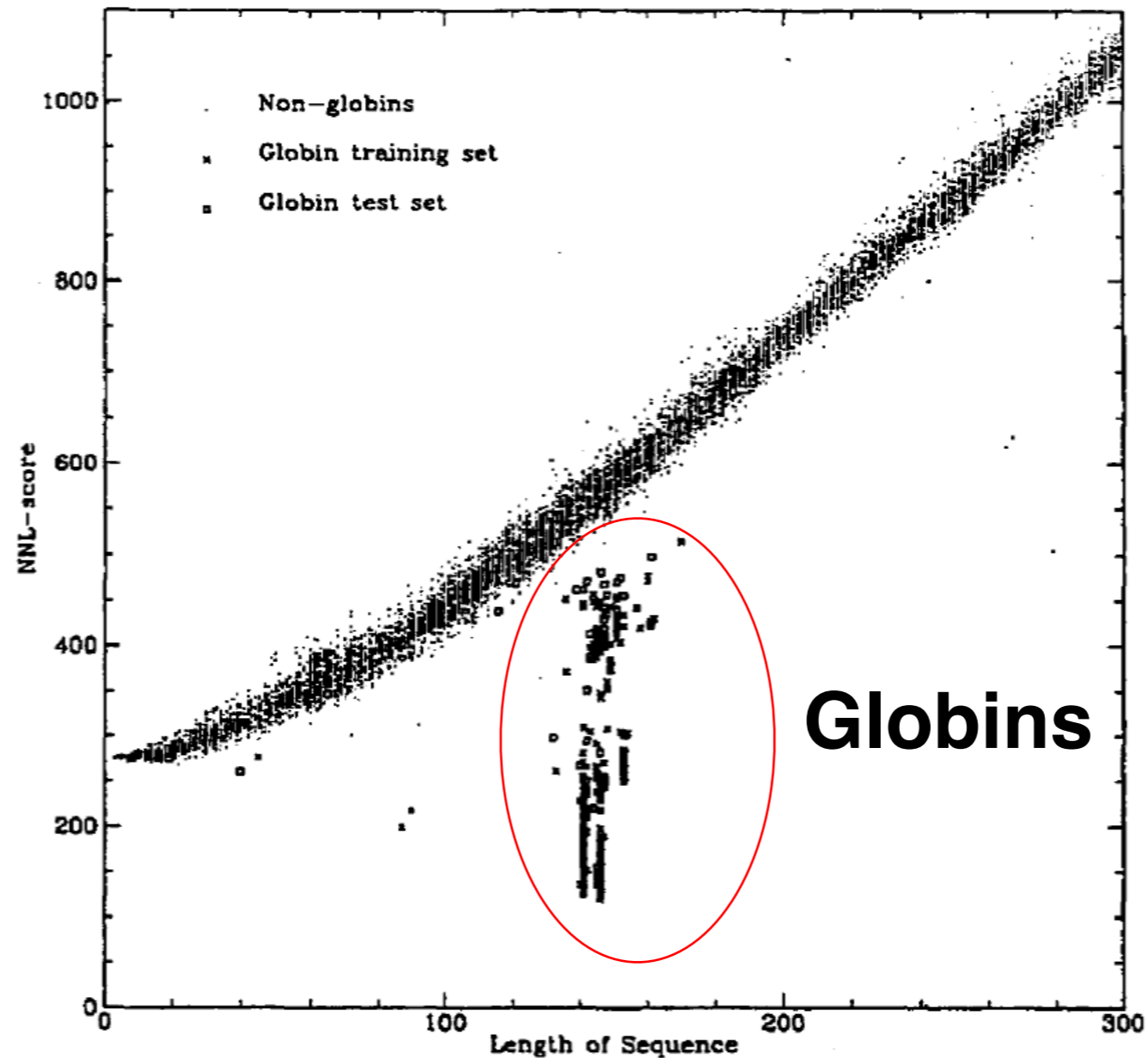
                AAAAAAAAAAAAAAAAAA   BBBBCCCCCCCCCCCCCCCC
                                DDDD
                *****
V.....LSPADKTNVKA AWGKVGA..HAGEYGAEALERMFLSFPTTKTYFPHFD-L
Vh.....LTPEEKSAVTALWGKV--..NVDEVGGEALGRLLVVYPWTQRRFFESFGDL
V.....LSEGEWQLVHLVWAKVEA..DVAGHGQDILIRLFKSHPETLEKFDKFKHL
-.....LSADQISTVQASFDKV--..KGDPVGI--LYAVFKADPSIMAKFTQFAGK
PivdtgsvapLSAAEKTIRSAWAPVYS..TYETSGVDILVKFFTSTPAAQEFFPKFKGL
Ga.....LTESQAALVKSSWEEFNA..NIPKHTHRFFILVLEIAPAAKDLFSFLK-G
G.....LSAAQRQVIAATWKDIAGadNGAGVVKDCLIKFLSAHPQMA---AVFG-F

DDDDDDDEE EEEEEEEEEEEEEEEEEEE
                                F
                *****
SHGSAQVKGH-GKK.----VADALTNAVAHVDD....MPNALSALSDLHA...HKLRVD
STPDAVMGNPKVKA.HGKKVLGAFSDGLAHLDN....LKGTFATLSELHC...DKLHVD
KTEA-EMKASEDLKkHGVTVLTA LGAILKKKGH....HEAELKPLAQSHA...TKHKIP
DLES-IKGTAPFET.HANRIVGFFSKIIGELPN....IEADVNTFVASHK...PR-GVT
TTADQLKKSADVRW.HAERIINAVNDAVASMDDtek..MSMKLRLSLGKHA...KSFQVD
TSEVPQ-NNPELQA.HAGKVFKLVYEA AIQLQVtgvvvTDATLKNLGSVHV...SK-GVA
SGAS----DPGVAA.LGAKVLAQIGVAVSHLGDegk..MVAQMKAVGVRHKgygNK-HIK

GGGGGGGGGGGGGGGGGGGGG   HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
*****
PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLT SKY.....R
PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY.....H
IKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYkelgyqG
HDQLNFRAGFVSYMKAH--TDF-AGAEAAWGATLDTFFGMIFSKM.....-
PQYFKVLA AVIADTVAA---GD-----AGFEKLMSMICILLRSAY.....-
DAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMnda...A
AQYFEPLGASLLSAMEHRIGGKMNA AAKDAWAAAYADISGALISGLq.....S
```

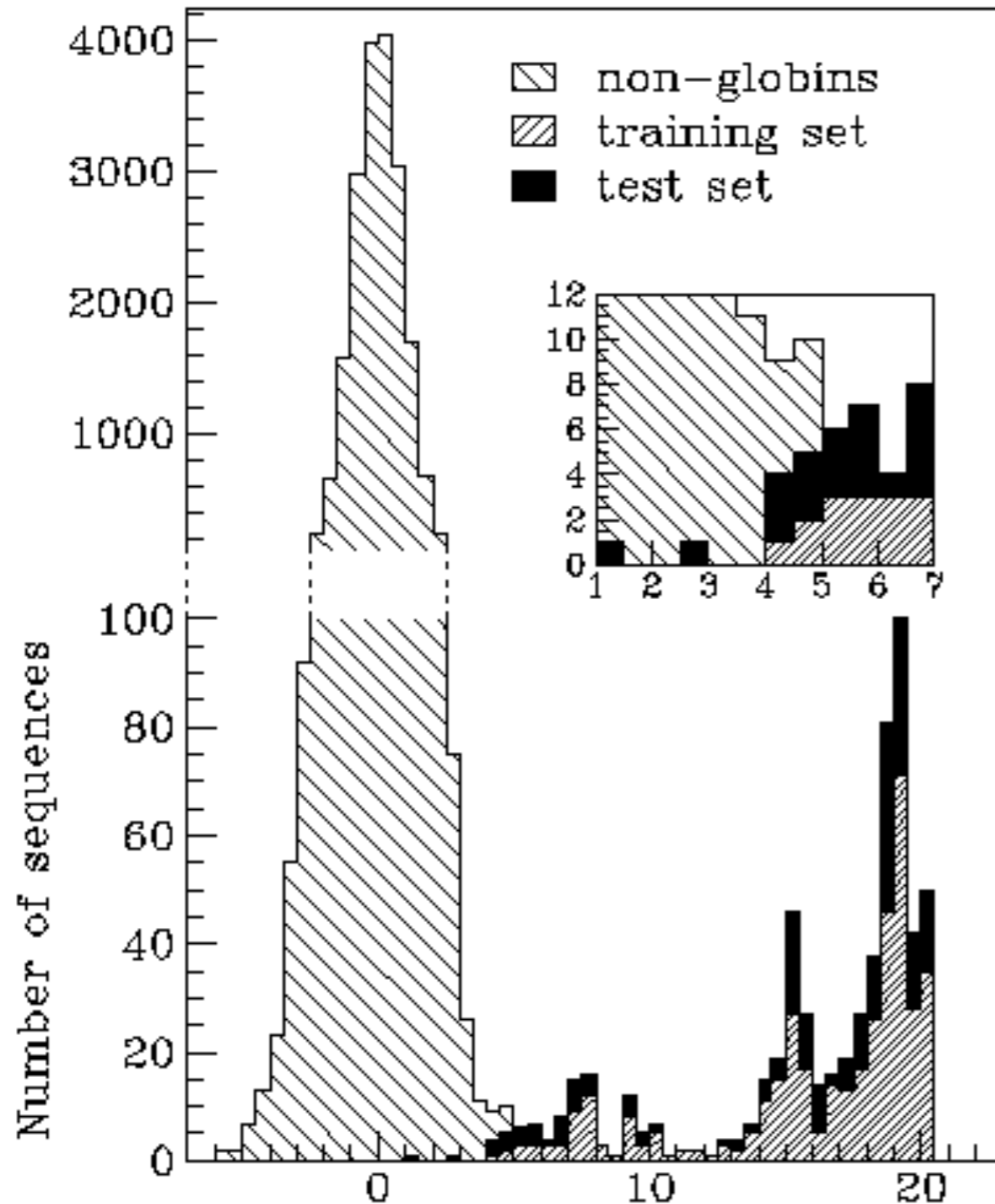
# Globin Classification

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	Total population		
	Test outcome positive	<b>True positive</b>	<b>False positive</b> (Type I error)
	Test outcome negative	<b>False negative</b> (Type II error)	<b>True negative</b>



# 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

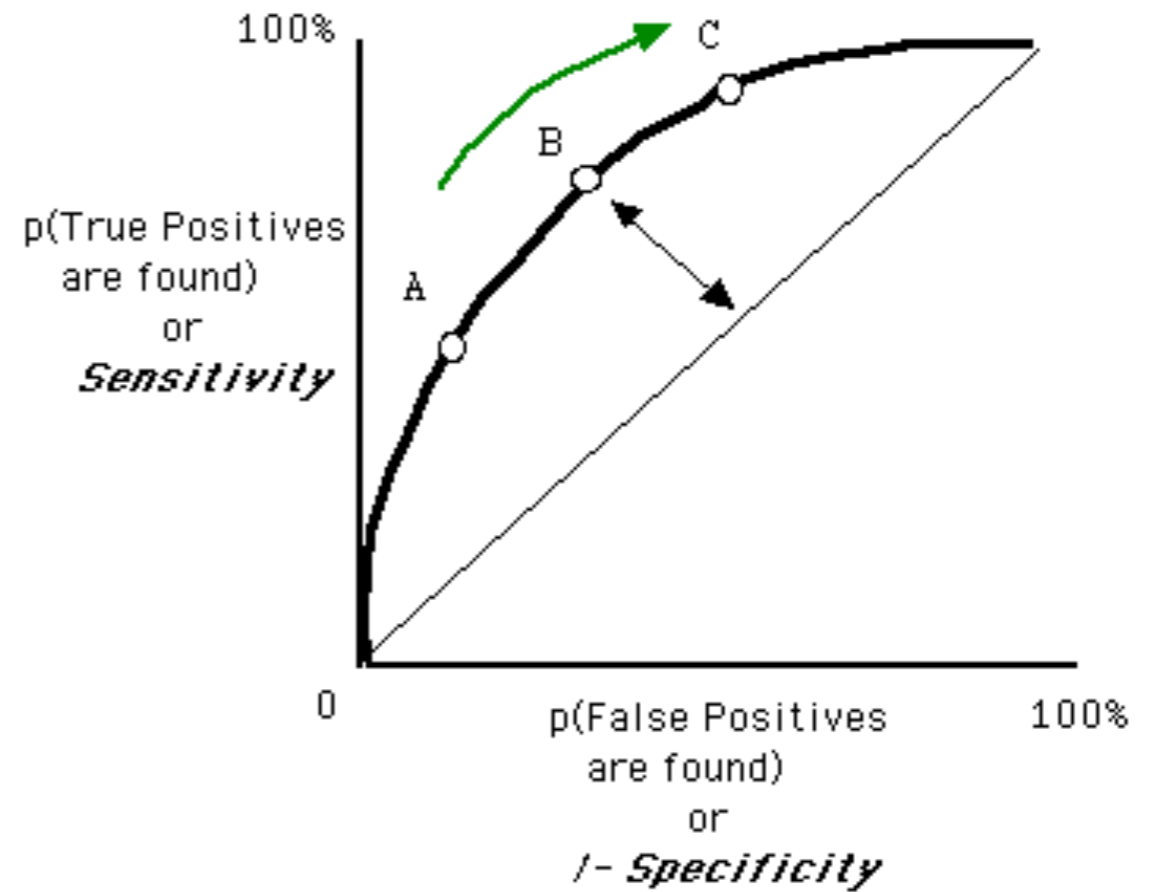
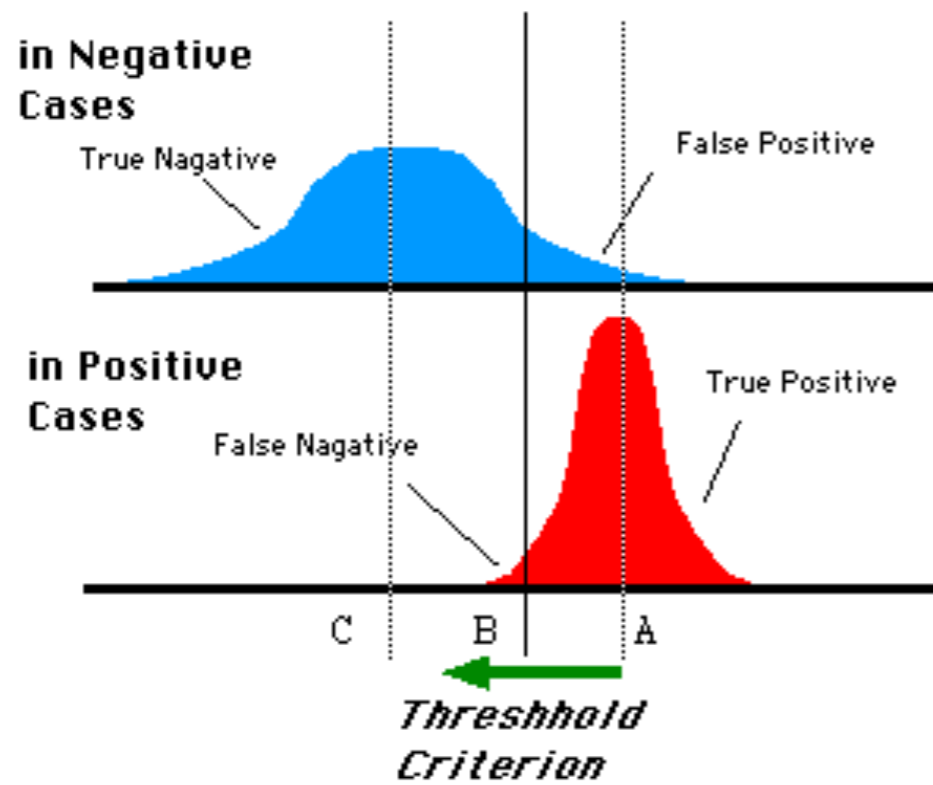
MCC = 0.62

Predictions are good

# ROC Curve

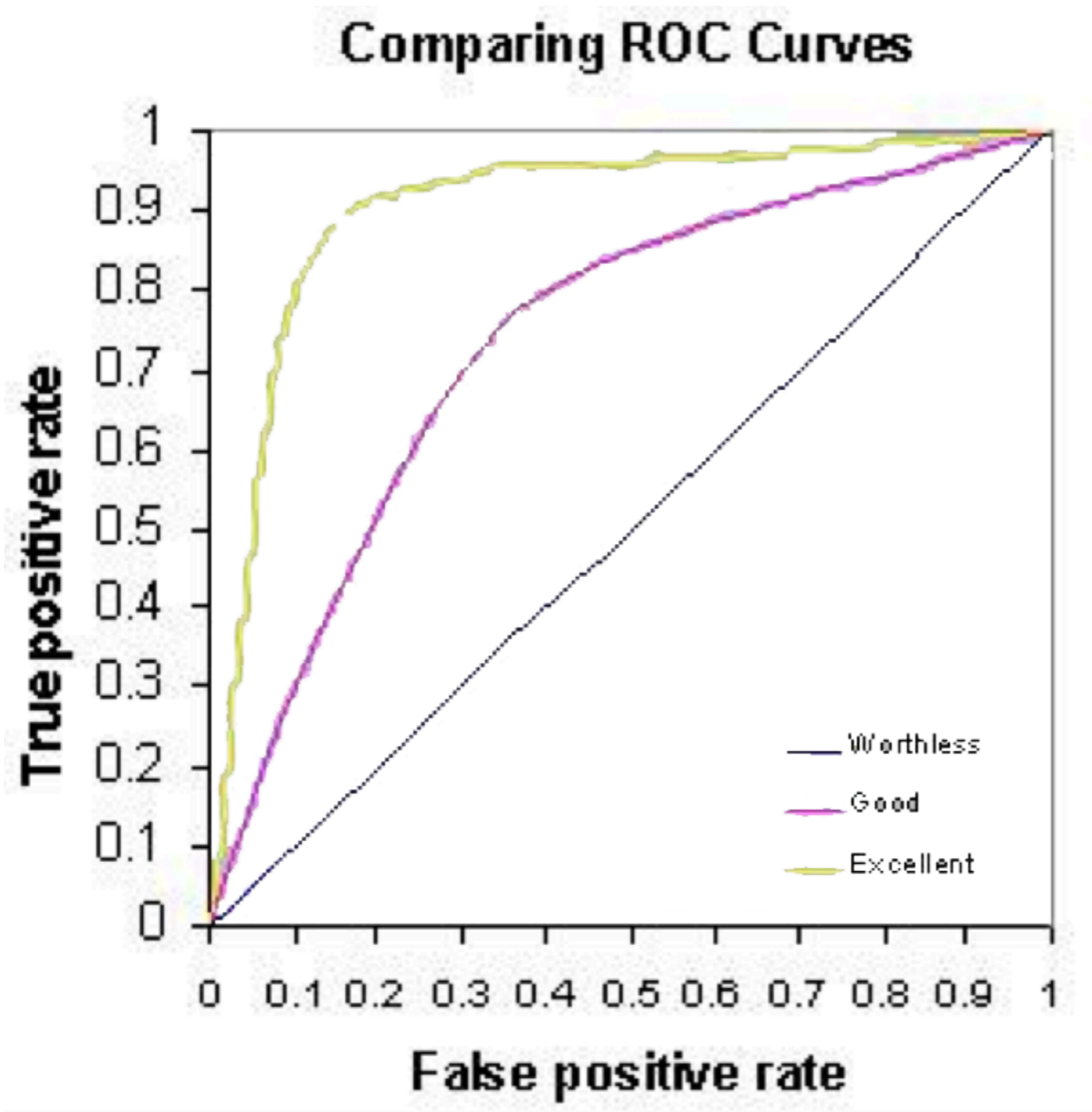
The Receiver Operating Characteristics depends on a parameter, TPR and FPR can be plotted at varying values of the parameter

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