Hidden Markov Models for Sequence Alignment

Laboratory of Bioinformatics I
Module 2

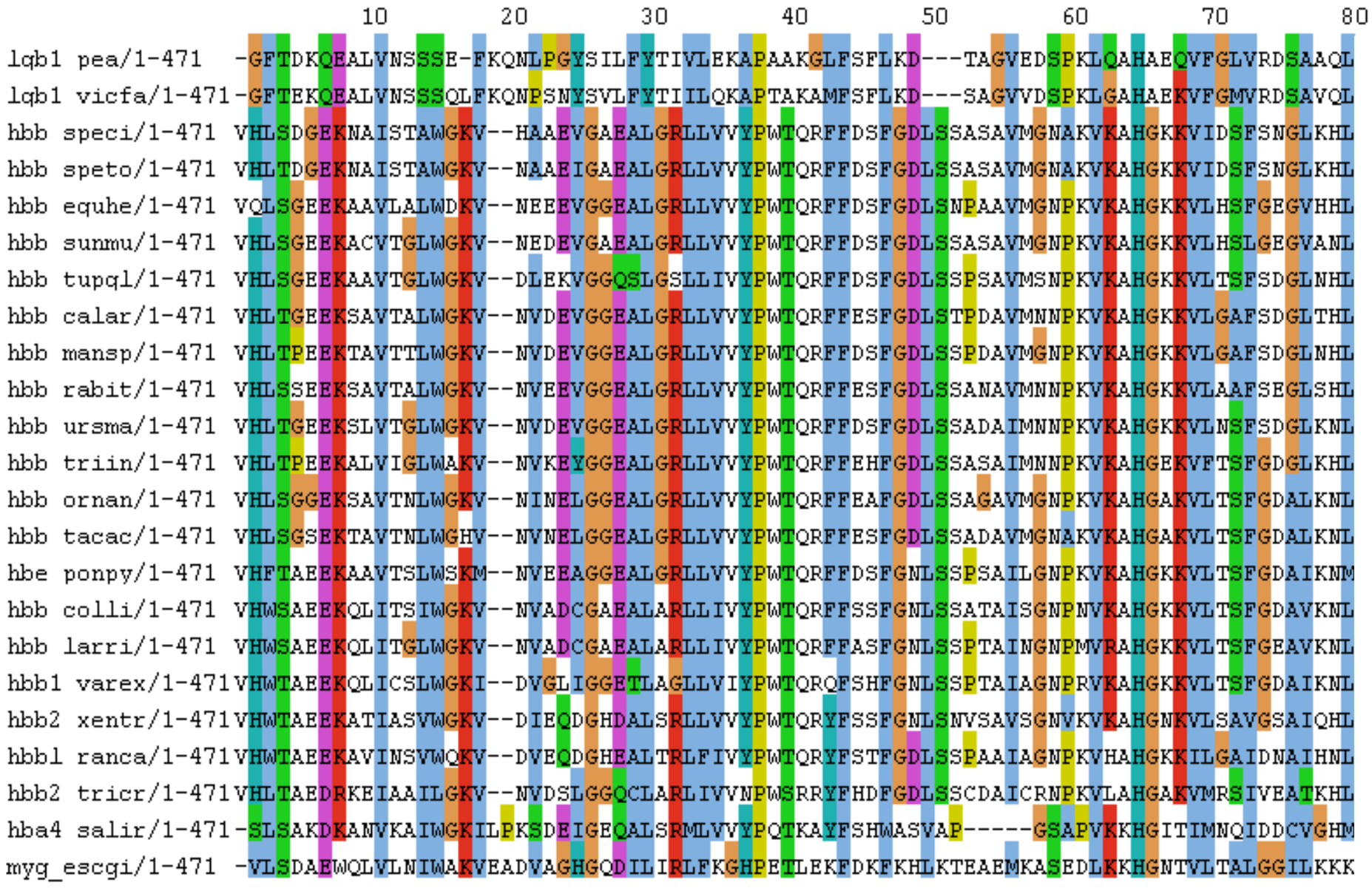
Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

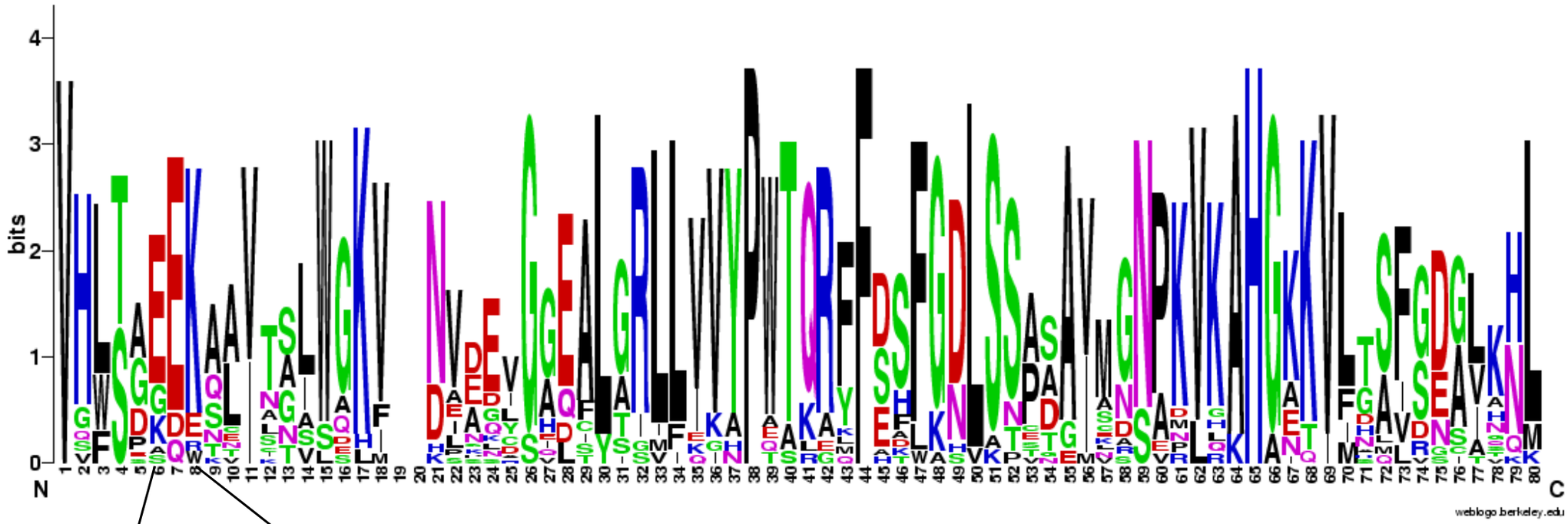
Alignment of Globins

Different positions are not equivalent



Sequence Logo

A more flexible alignment score is needed to align protein families

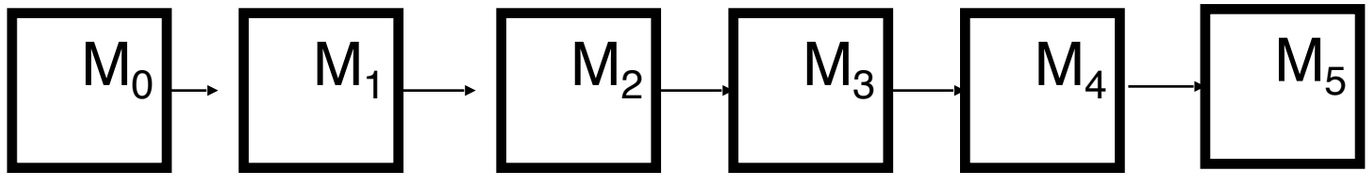


The substitution score may depend on the position.

http://weblogo.berkeley.edu/

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

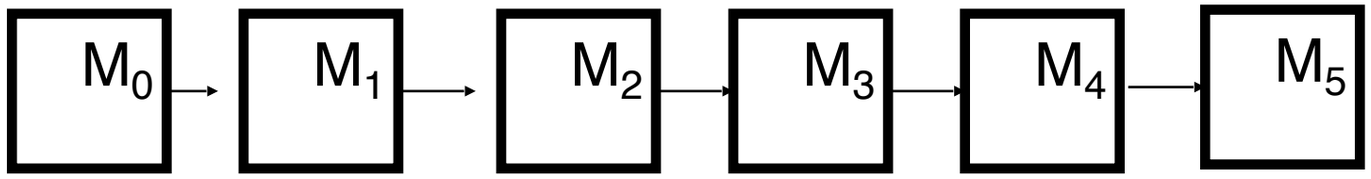
A	T	G	T	T	C
M_0	M_1	M_2	M_3	M_4	M_5

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

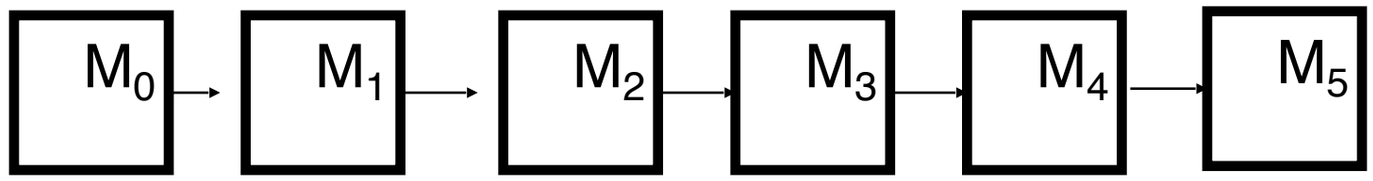
A	C	G	G	T	A
A	C	G	A	T	C
A	T	G	T	T	C



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

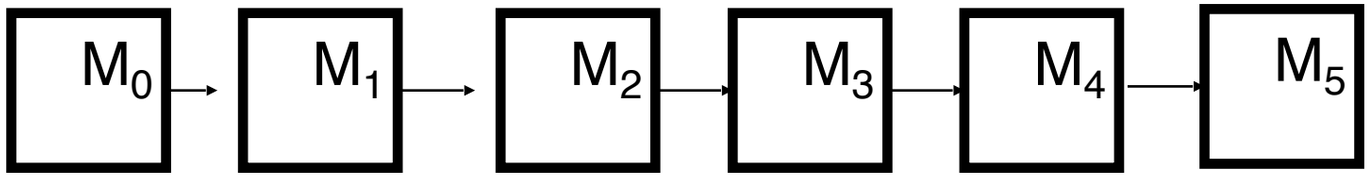


A	1	0	0	0.33	0	0.33
C	0	0.66	0	0	0	0.66
G	0	0	1	0.33	0	0
T	0	0.33	0	0.33	1	0
	A	C	G	A	T	C

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

Alignments with Gaps

A strategy to introduce gaps is needed

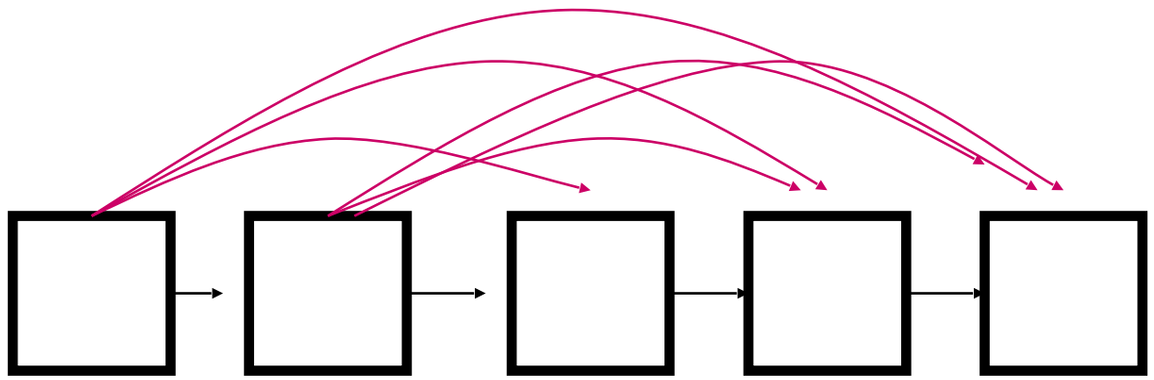


A	1	0	0	0.33	0	0.33
C	0	0.66	0	0	0	0.66
G	0	0	1	0.33	0	0
T	0	0.33	0	0.33	1	0

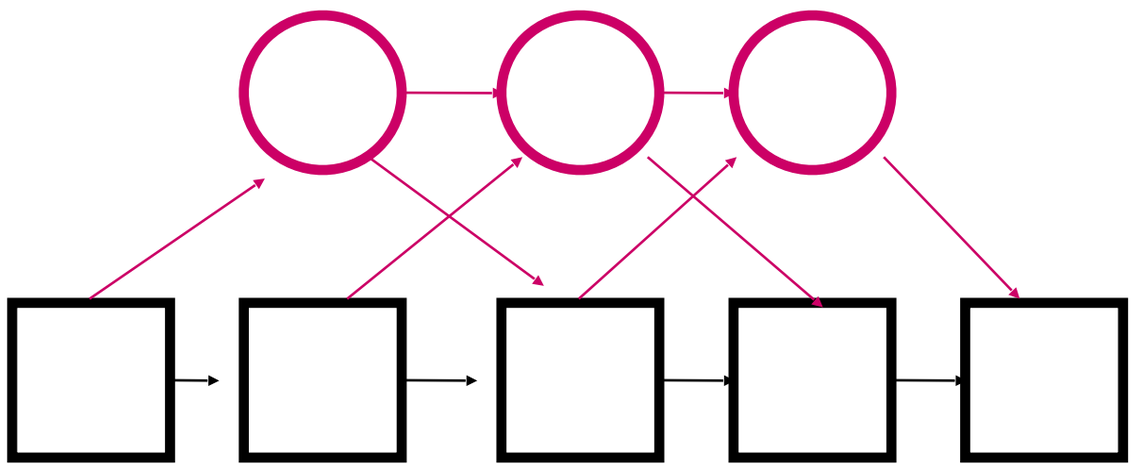
A		G	A	T	C
M_0	M_2	M_3	M_4	M_5	M_5

Silent States

Different topology to model gaps

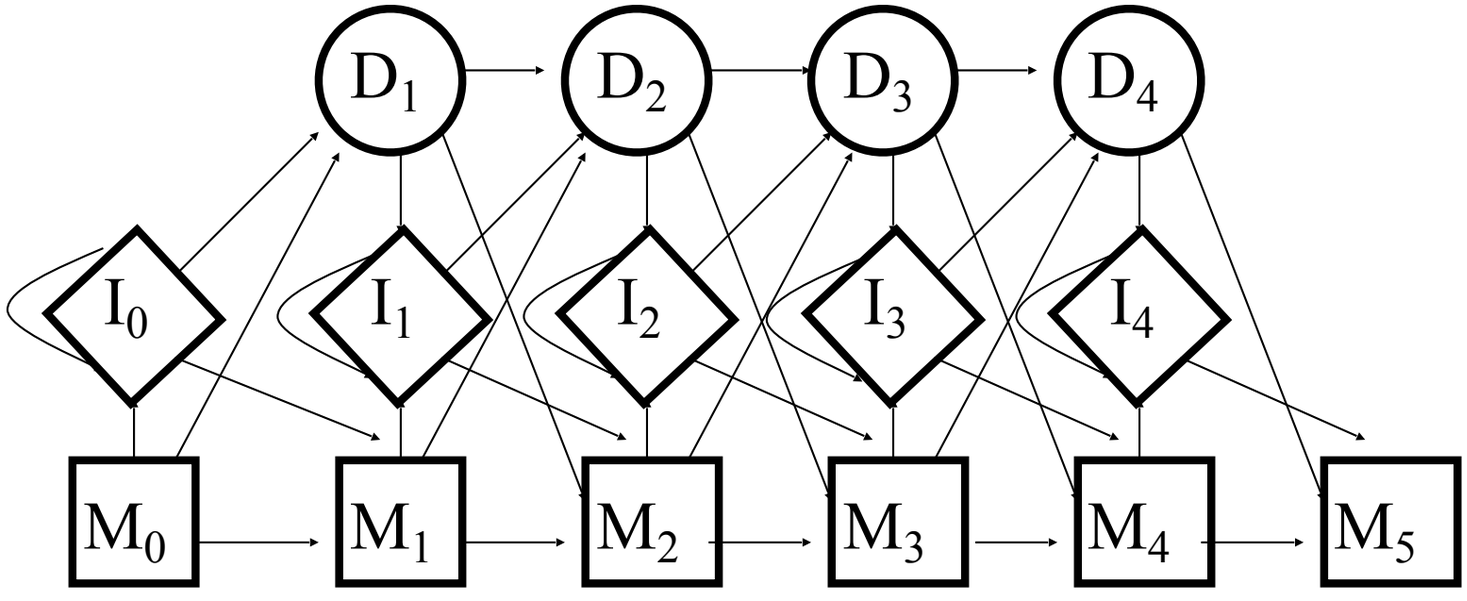


N(N-1)/2 transitions



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_0 M_1 M_2 M_3 M_4 M_5

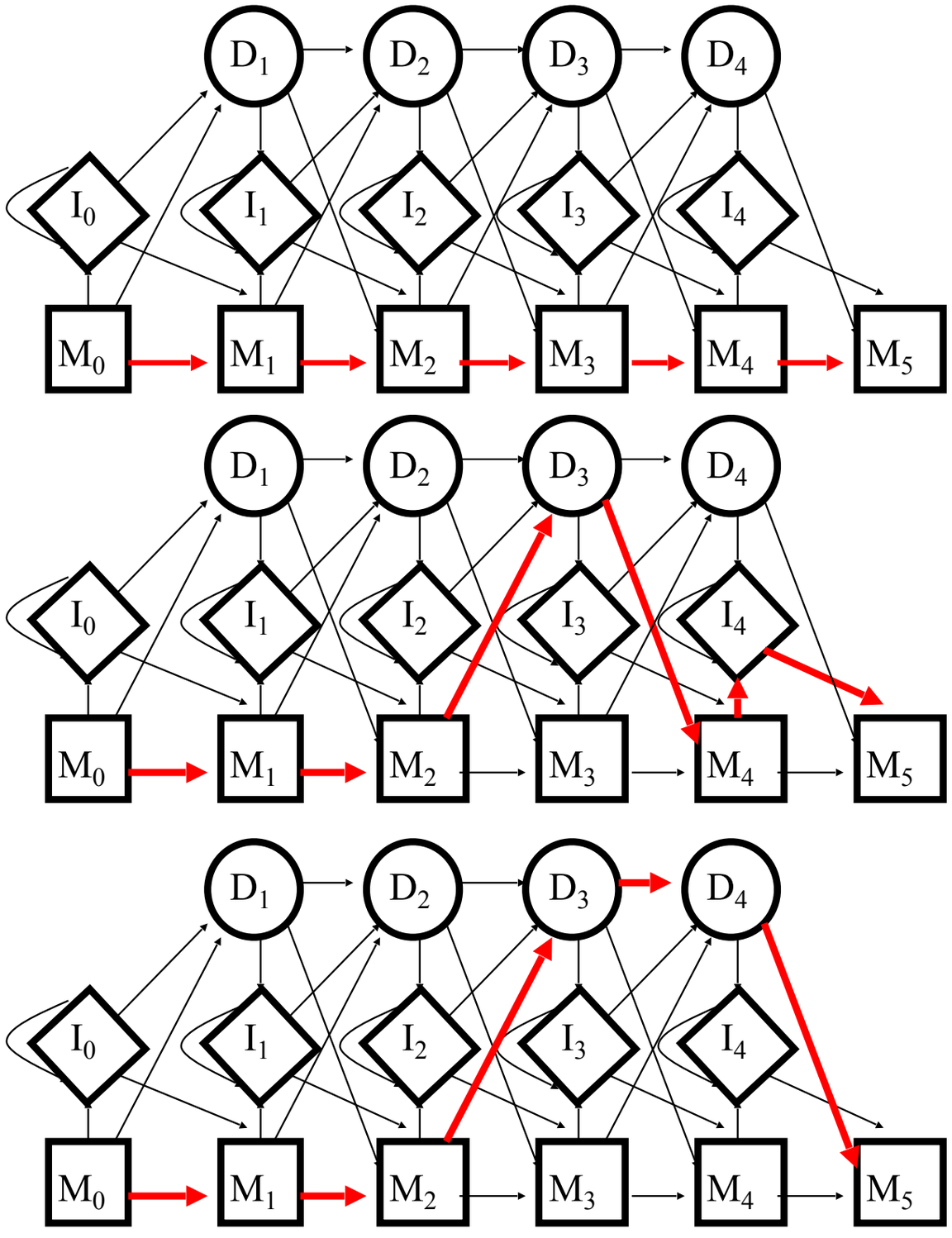
A C G C A G T C

M_0 I_0 I_0 M_1 M_2 M_3 M_4 M_5

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

M_0 D_1 M_2 M_3 M_4 M_5

Example of Alignment



Sequence 1

A S T R A L

Viterbi path

M_0 M_1 M_2 M_3 M_4 M_5

A S T R A L

Sequence 2

A S T A I L

Viterbi path

M_0 M_1 M_2 D_3 M_4 I_4 M_5

A S T A I L

Sequence 3

A R T I

Viterbi path

M_0 M_1 M_2 D_3 D_4 M_5

A R T I

Alignment Calculation

M_0	M_1	M_2	M_3	M_4		M_5	
A	S	T	R	A		L	*Sequence 1*
M_0	M_1	M_2	D_3	M_4	I_4	M_5	
A	S	T		A	I	L	*Sequence 2*
M_0	M_1	M_2	D_3	D_4		M_5	
A	R	T				I	*Sequence 3*

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

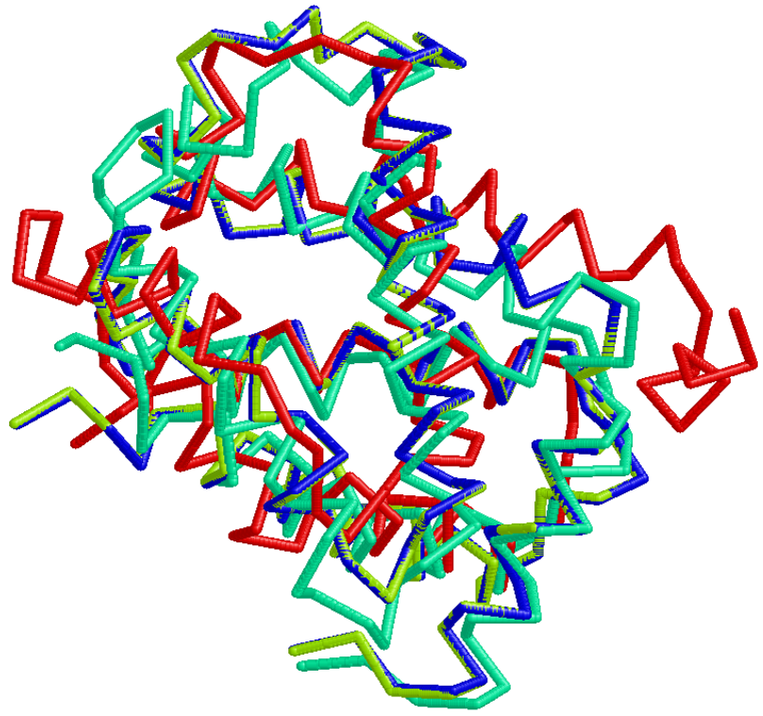
Alignment

```
ASTRA-L
AST-AIL
ART---I
```

-Log P(s I M) Is an alignment score

Alignment of Globins

```
            AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCC
                                                    DDDD
----------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF-DL
--------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESFGDL
----------VLSEGEWQLVLHVWAKVEA--DIAGHGQDILIRLFKHHPETLEKFDRFKHL
-----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQFAG-
PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGL
--------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-FLK-
----------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-FSG-

 DDDDDDDEEEEEEEEEEEEEEEEEEEE                FFFFFFFFFFF  FFGGG
                             F                          GG  GG
S-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL--RVDPV
STPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL--HVDPE
KSEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH--KIPIK
KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG---VTHD
TTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF--QVDPQ
GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG---VADA
---AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGNKHIKAQ

GGGGGGGGGGGGGGGG      HHHHHHHHHHHHHHHHHHHHHHHHHH

NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
NFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
YLEFISEAIIHVLHSRHPADFGADAQGAMSKALELFRKDIAAKYKELGYQG
QLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
YFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
HFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
YFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
```



Bashford et al. J Mol Biol 1987

Globins HMM

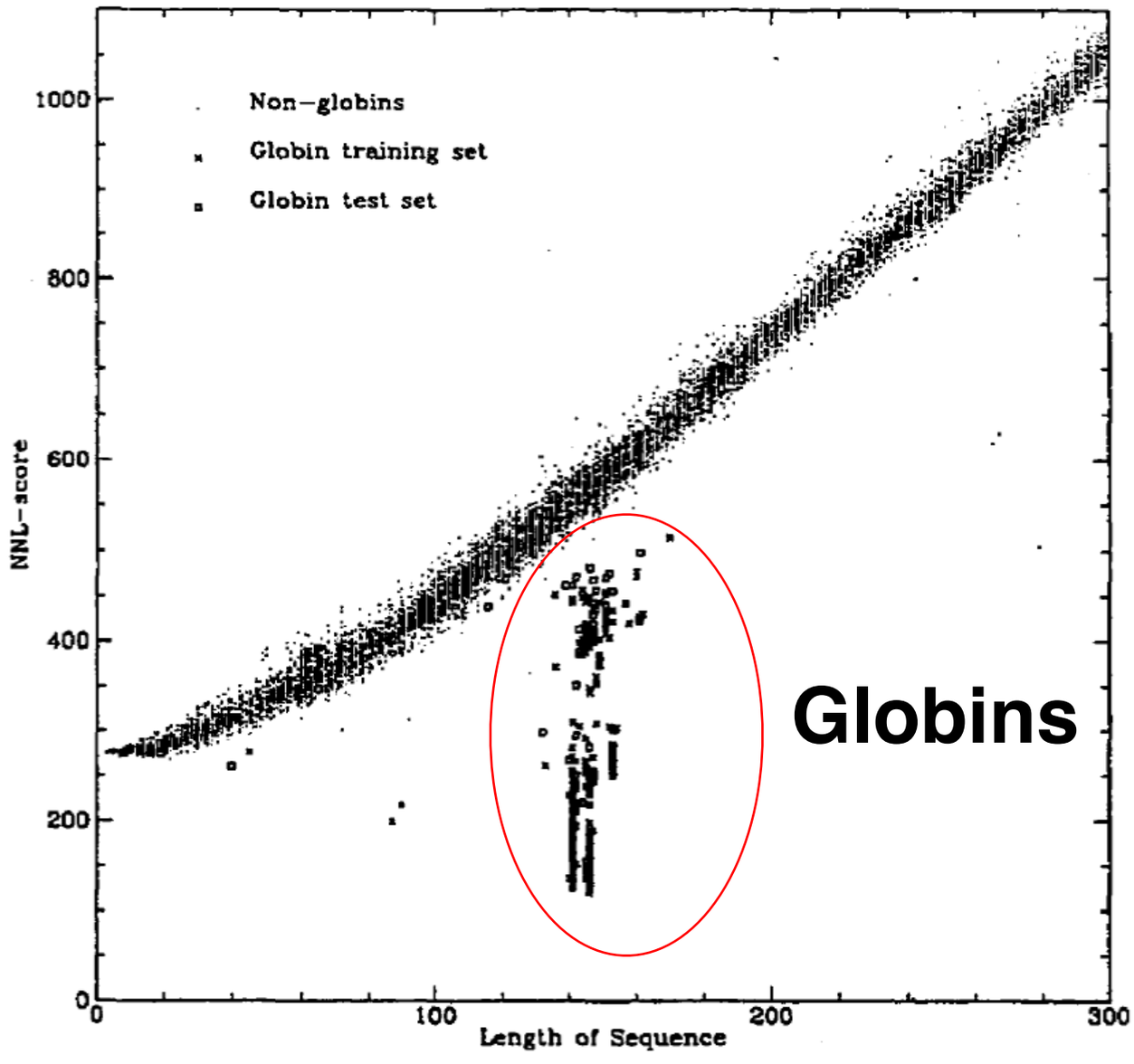
HMM are calculate from a training set of 400 unaligned sequences. After the HMM is built, it is used to obtain a multiple alignment of all the training sequences. This is the alignment of the 7 globins as aligned with the trained model.

```
                AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCC
                                                        DDDD
                ***************    ************************
V.........LSPADKTNVKAAWGKVGA..HAGEYGAEALERMFLSFPTTKTYFPHFD-L
Vh........LTPEEKSAVTALWGKV--..NVDEVGGEALGRLLVVYPWTQRFFESFGDL
V.........LSEGEWQLVLHVWAKVEA..DVAGHGQDILIRLFKSHPETLEKFDRFKHL
-.........LSADQISTVQASFDKV--..KGDPVGI--LYAVFKADPSIMAKFTQFAGK
PivdtgsvapLSAAEKTKIRSAWAPVYS..TYETSGVDILVKFFTSTPAAQEFFPKFKGL
Ga........LTESQAALVKSSWEEFNA..NIPKHTHRFFILVLEIAPAAKDLFSFLK-G
G.........LSAAQRQVIAATWKDIAGadNGAGVGKDCLIKFLSAHPQMA---AVFG-F

 DDDDDDDEE EEEEEEEEEEEEEEEEE            FFFFFFFFF   FFFFG
                          F                          GGGG
                 ***********            *********       **
SHGSAQVKGH-GKK.----VADALTNAVAHVDD.....MPNALSALSDLHA...HKLRVD
STPDAVMGNPKVKA.HGKKVLGAFSDGLAHLDN.....LKGTFATLSELHC...DKLHVD
KTEA-EMKASEDLKkHGVTVLTALGAILKKKGH.....HEAELKPLAQSHA...TKHKIP
DLES-IKGTAPFET.HANRIVGFFSKIIGELPN.....IEADVNTFVASHK...PR-GVT
TTADQLKKSADVRW.HAERIINAVNDAVASMDDtek..MSMKLRDLSGKHA...KSFQVD
TSEVPQ-NNPELQA.HAGKVFKLVYEAAIQLQVtgvvvTDATLKNLGSVHV...SK-GVA
SGAS----DPGVAA.LGAKVLAQIGVAVSHLGDegk..MVAQMKAVGVRHKgygNK-HIK

GGGGGGGGGGGGGGGGG     HHHHHHHHHHHHHHHHHHHHHHHHH
******************    ************************
PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY......R
PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY......H
IKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYkelgyqG
HDQLNNFRAGFVSYMKAH--TDF-AGAEAAWGATLDTFFGMIFSKM......-
PQYFKVLAAVIADTVAA---GD------AGFEKLMSMICILLRSAY......-
DAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMnda...A
AQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLq.....S
```

Krogh et al. J Mol Biol. 1994

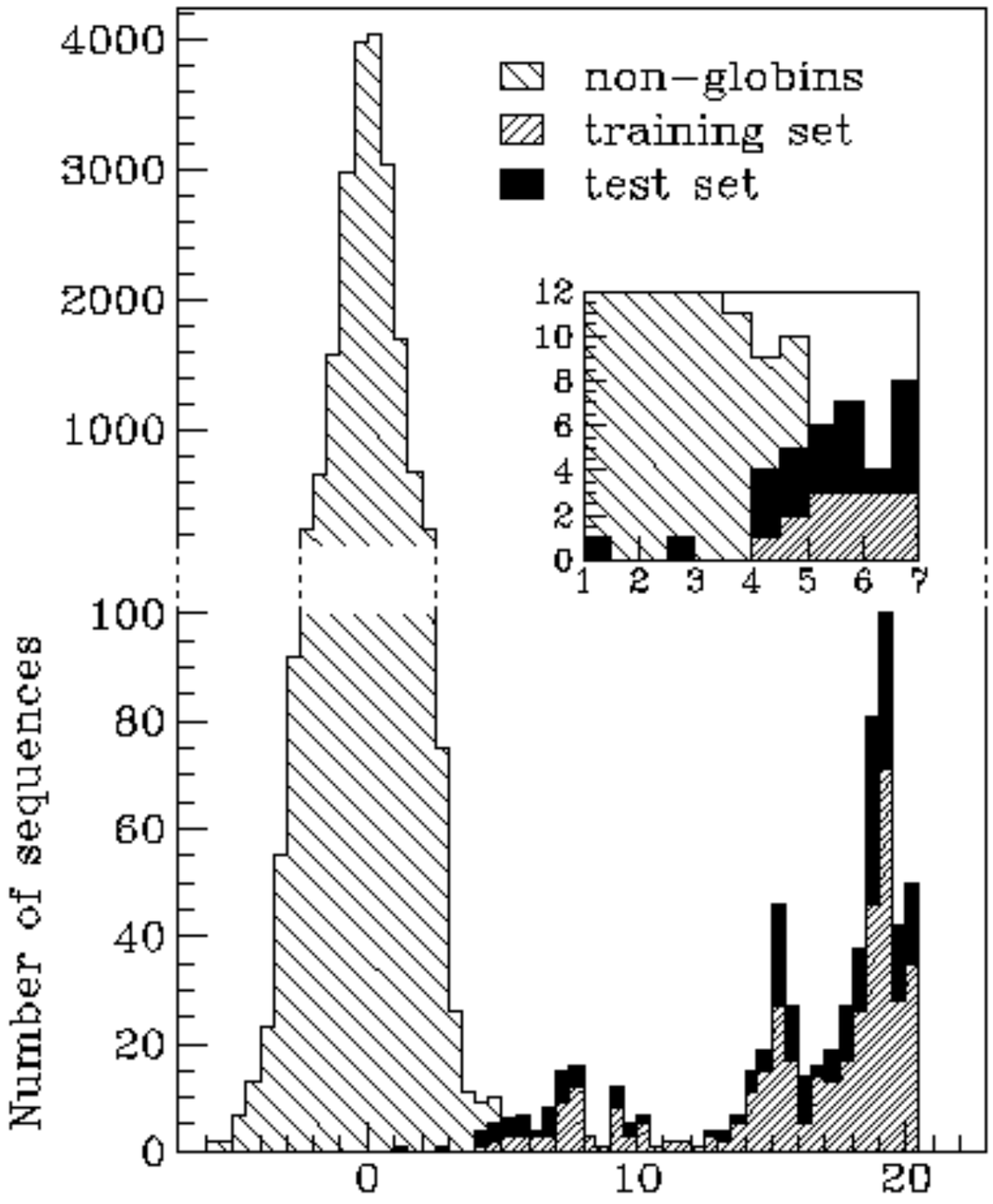
Globin Classification

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



NLLscore = —log P(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")	
	Total population	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

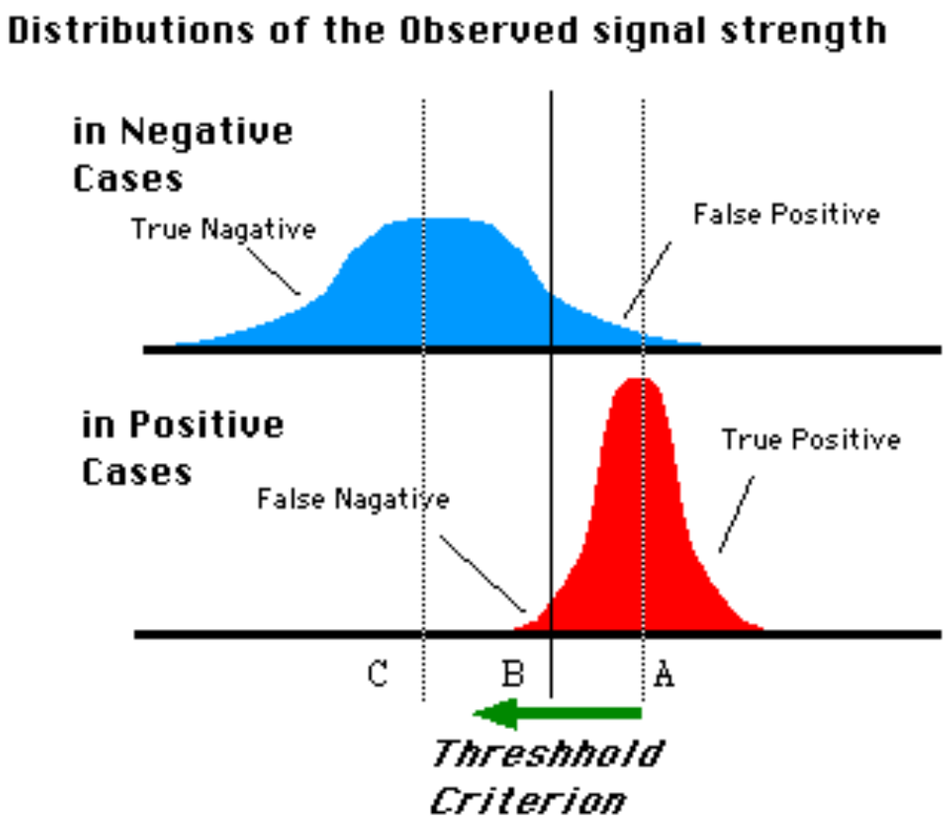
Acc = 0.9
All the examples are predicted in the class 1:
MCC = 0.0
Very bad predictions

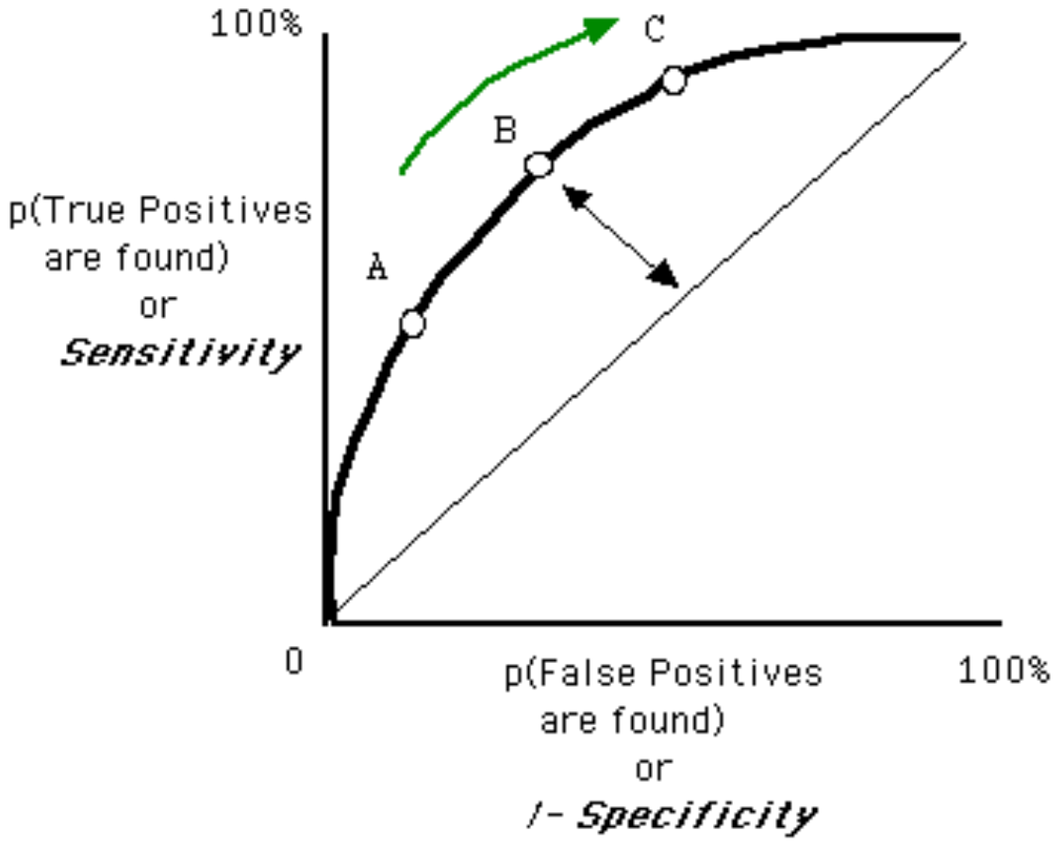
	Class 1	Class -1
Prediction 1	81	1
Prediction -1	9	9

Acc = 0.9
MCC = 0.62
Predictions are good

ROC Curve

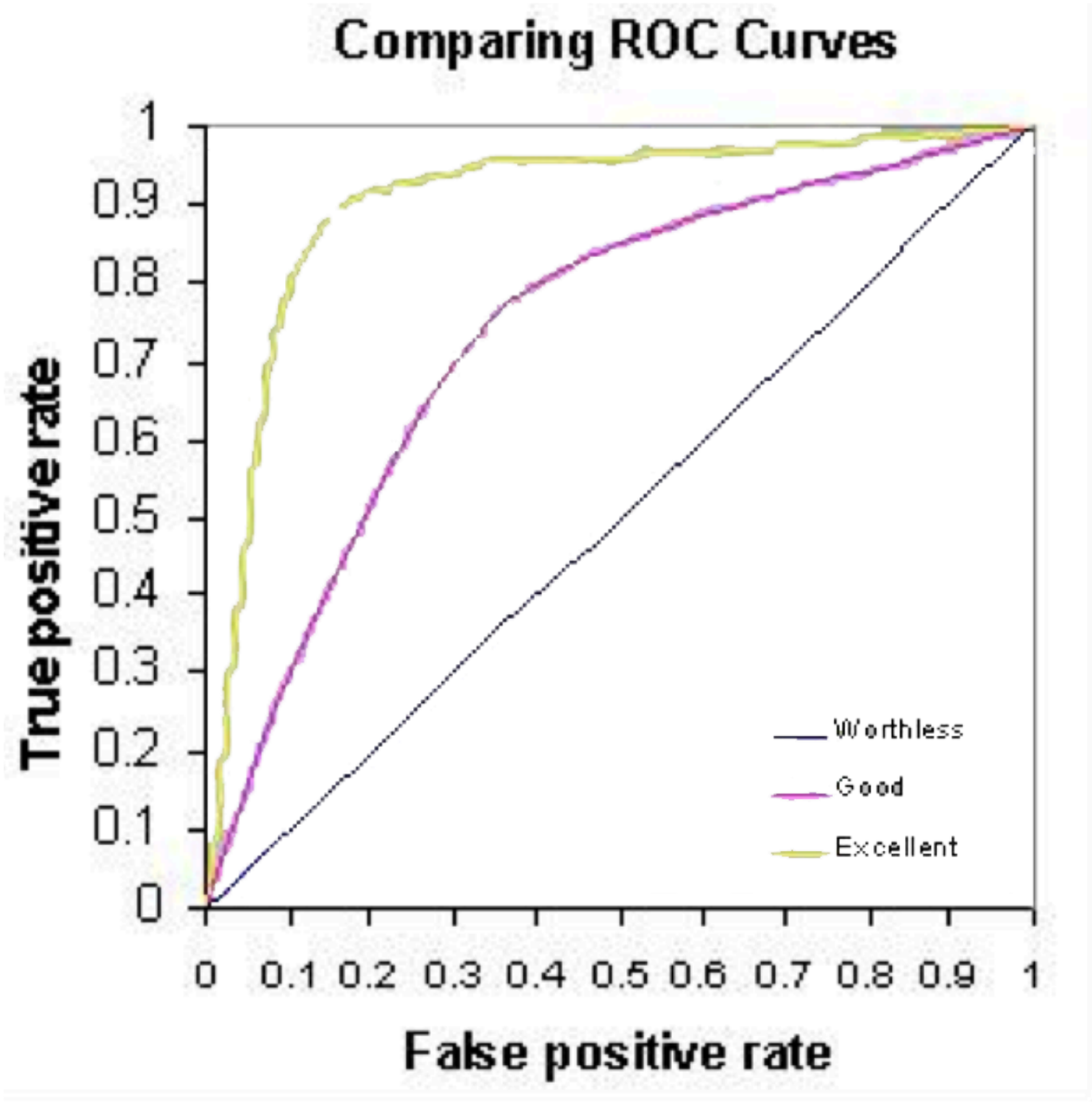
The Receiver Operating Characteristics depends on a parameter, TPR and FPR can be plotted at varying values of the parameter





Area Under Curve

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



AUC=0.5 → Random prediction

AUC=1 → Perfect prediction