

# BIOINFORMATICS 1

or why biologists need computers

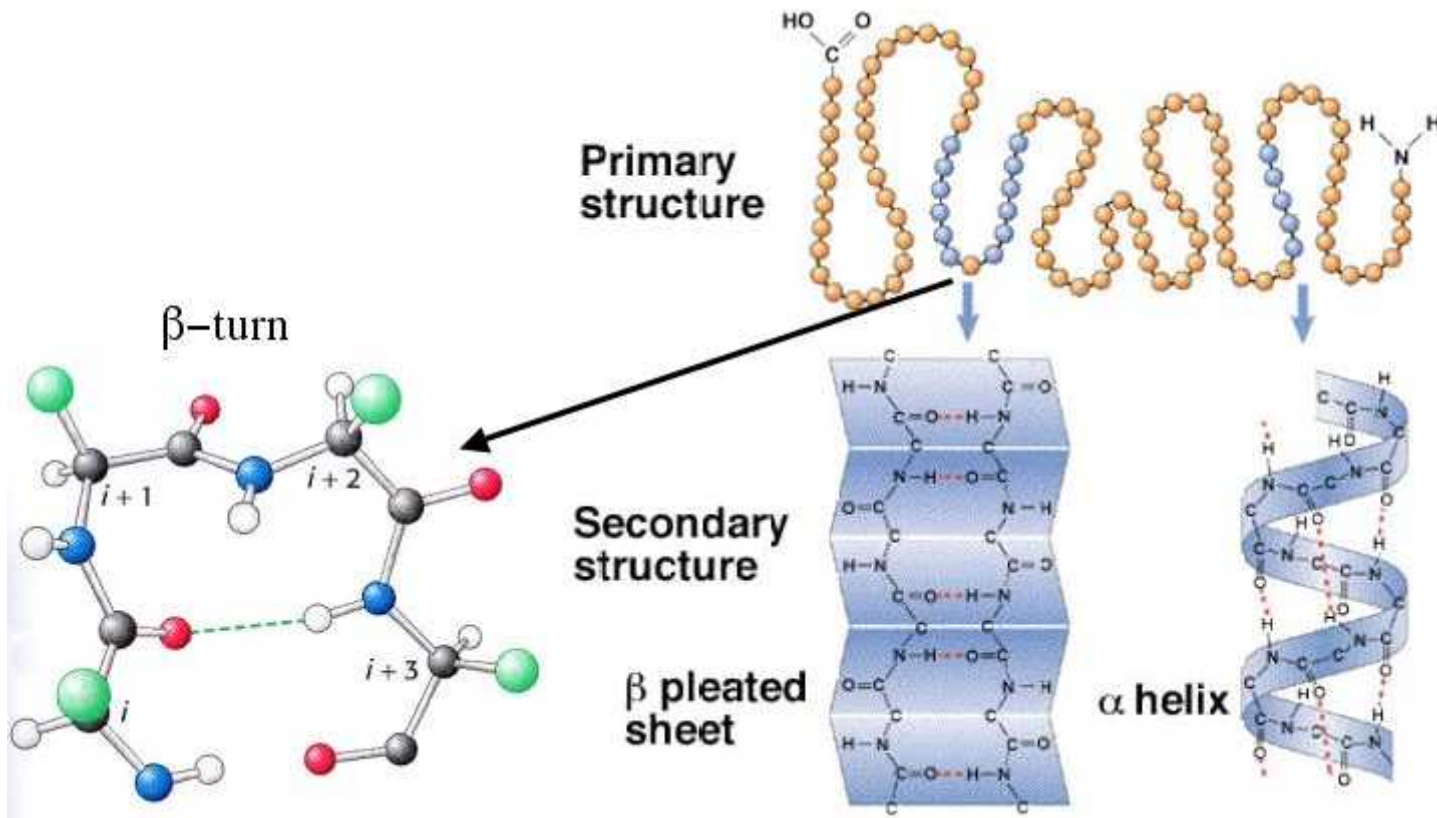
<http://www.bioinformatics.uni-muenster.de/teaching/courses-2011/bioinf1/index.hbi>



# (SOME) PROTEIN SEQUENCE ANALYSIS



...just a reminder



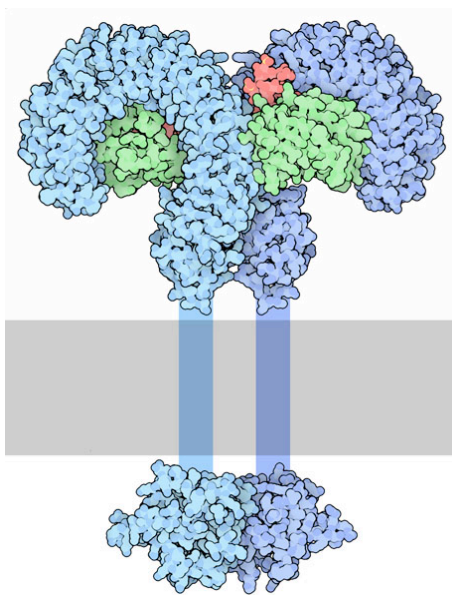


## PDB: Protein Data Bank:



[www.rcsb.org/pdb](http://www.rcsb.org/pdb)

- “primary database” – contains structures determined by experiments (Xray, NMR)
- Not only proteins – also complexes, peptides, nucleic acids, cofactors, ...
- Roughly 77,000 structures corresponding to ~ 44,000 sequences
- Many sequences in different variants, e.g. hemoglobin
- “Molecule of the month” – enjoy!



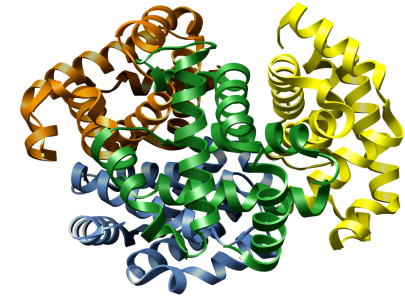
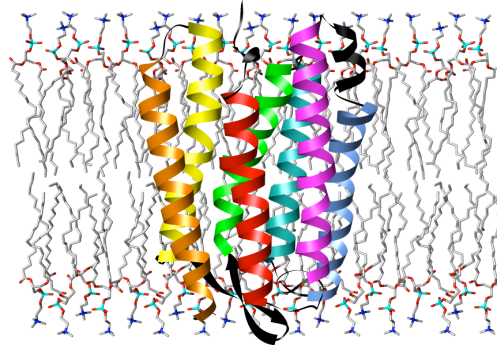
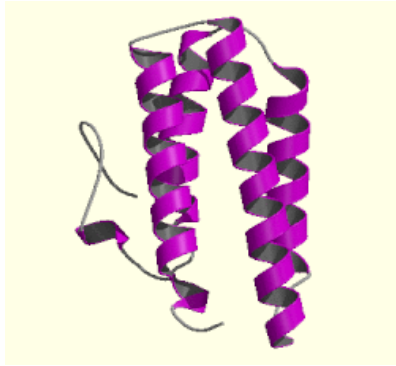
*Molecule of the Month – November 2011*  
*Toll-like receptor*


### Note:

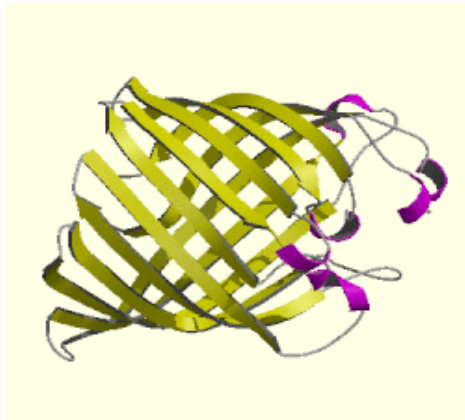
- one structure can contain many different sequences
- not all structures contain *full* protein sequences (usually just a fragment)
- many structures contain several peptides of identical sequence (homomeres)



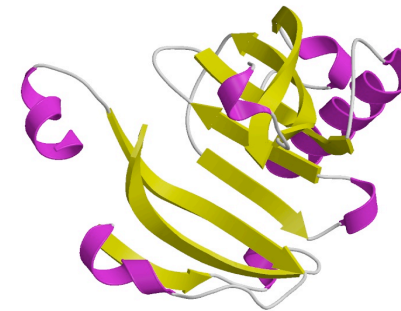
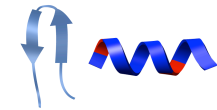
 All-alpha



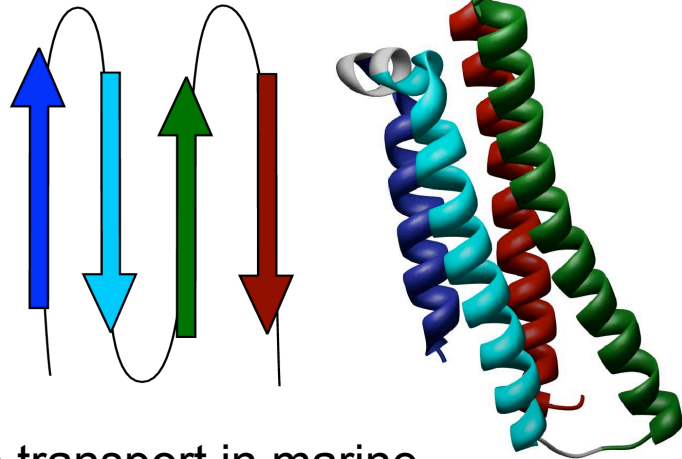
 All-beta



Mixed Alpha/Beta

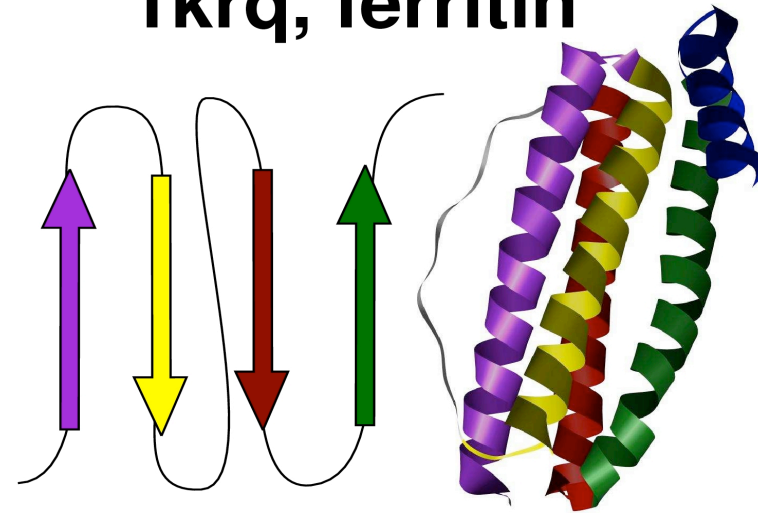


## 1lpe, hemerythrin

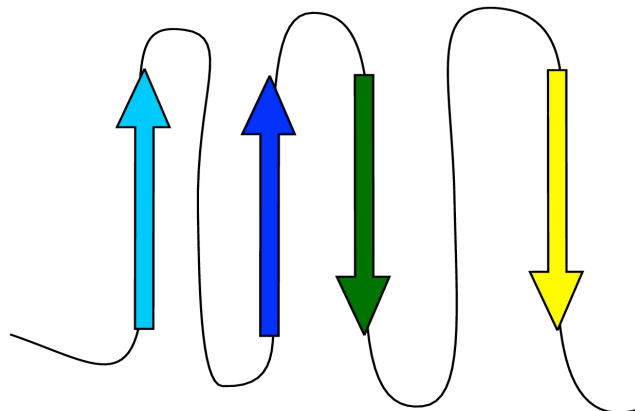


Oxygen transport in marine invertebrates; U-D-U-D

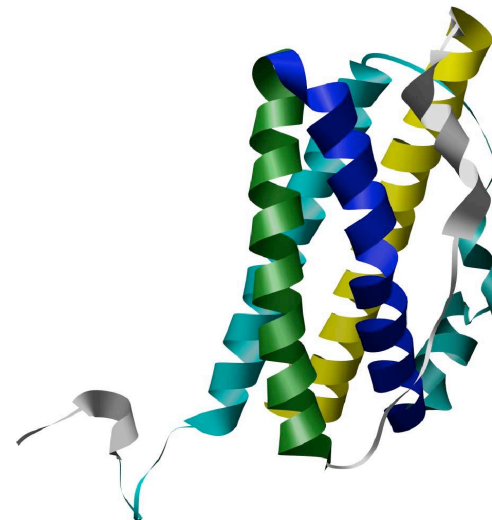
## 1krq, ferritin



Iron transport; multimeric; vertebrate liver; U-D-D-U



## 1f6f, placental lactogen



Hormone, e.g. the growth hormone  
U-U-D-D



- Protein folding problem
  - It is not completely understood how protein folds (why do they fold so quickly?)
- “Sequence – structure gap”
  - Making out a complete 3D structure based only on protein sequence is not possible
  - *Ab initio* algorithms: prediction of secondary structure and specific structural features (motifs, coiled coils etc.) – this already can give insights into protein function
  - Evolutionary comparisons – which regions are conserved?
  - Predictions based on comparison with homologous sequences (e.g. threading)
- Predictions of RNA structures



General principle in analyzing protein sequences to predict structure:

- Find properties of known structures (PDB)
  - the frequency of a given amino acid in an HLH motif
  - the tendency of an amino acid to form a  $\beta$ -sheet
- Come up with heuristic rules (e.g. ChouFasman)



# Chou-Fasman Parameters

helix potential ( $P_{\alpha}$ )			$\beta$ -sheet potential ( $P_{\beta}$ )		
Glu	1.51		Val	1.70	
Met	1.45		Ile	1.60	
Ala	1.42		Tyr	1.47	
Leu	1.21		Phe	1.38	
Lys	1.16		Trp	1.37	
Phe	1.13		Leu	1.30	
Gln	1.11		Cys	1.19	
Trp	1.08		Thr	1.19	
Ile	1.08		Gln	1.10	
Val	1.06		Met	1.05	
Asp	1.01		Arg	0.93	⋮
His	1.00		Asn	0.89	⋮
Arg	0.98	⋮	His	0.87	⋮
Thr	0.83	⋮	Ala	0.83	⋮
Ser	0.77	⋮	Ser	0.75	⋮
Cys	0.70	⋮	Gly	0.75	⋮
Tyr	0.69	⋮	Lys	0.74	⋮
Asn	0.67	⋮	Pro	0.55	⋮
Pro	0.57	⋮	Asp	0.54	⋮
Gly	0.57	⋮	Glu	0.37	⋮

||| strong former

|| former

| weak former

⋮ indifferent

⋮⋮ breaker

⋮⋮⋮ strong breaker

# Chou-Fasman Parameters

helix potential ( $P_{\alpha}$ )		$\beta$ -sheet potential ( $P_{\beta}$ )	
Glu	1.51	Val	1.70
Met	1.45	Ile	1.60
Ala	1.42	Tyr	1.47
Leu	1.21	Phe	1.38
Lys	1.16	Trp	1.37
Phe	1.13	Leu	1.30
Gln	1.11	Cys	1.19
Trp	1.08	Thr	1.19
Ile	1.08	Gln	1.10
Val	1.06	Met	1.05
Asp	1.01	Arg	0.93
His	1.00	Asn	0.89
Arg	0.98	His	0.87
Thr	0.83	Ala	0.83
Ser	0.77	Ser	0.75
Cys	0.70	Gly	0.75
Tyr	0.69	Lys	0.74
Asn	0.67	Pro	0.55
Pro	0.57	Asp	0.54
Gly	0.57	Glu	0.37

## Alpha-helices:

	strong former	+A
	former	+a
	weak former	
:	indifferent	
:::	breaker	-a
:::	strong breaker	-A

*BP*

# Chou-Fasman Parameters

helix potential ( $P_{\alpha}$ )		$\beta$ -sheet potential ( $P_{\beta}$ )	
Glu	1.51	Val	1.70
Met	1.45	Ile	1.60
Ala	1.42	Tyr	1.47
Leu	1.21	Phe	1.38
Lys	1.16	Trp	1.37
Phe	1.13	Leu	1.30
Gln	1.11	Cys	1.19
Trp	1.08	Thr	1.19
Ile	1.08	Gln	1.10
Val	1.06	Met	1.05
Asp	1.01	Arg	0.93
His	1.00	Asn	0.89
Arg	0.98	His	0.87
Thr	0.83	Ala	0.83
Ser	0.77	Ser	0.75
Cys	0.70	Gly	0.75
Tyr	0.69	Lys	0.74
Asn	0.67	Pro	0.55
Pro	0.57	Asp	0.54
Gly	0.57	Glu	0.37

Alpha-helices:

strong former	+A
former	+a
weak former	
: indifferent	
::: breaker	-a
::: strong breaker	-A

*BP*

# Chou-Fasman Parameters

helix potential ( $P_{\alpha}$ )		$\beta$ -sheet potential ( $P_{\beta}$ )	
Glu	1.51	Val	1.70
Met	1.45	Ile	1.60
Ala	1.42	Tyr	1.47
Leu	1.21	Phe	1.38
Lys	1.16	Trp	1.37
Phe	1.13	Leu	1.30
Gln	1.11	Cys	1.19
Trp	1.08	Thr	1.19
Ile	1.08	Gln	1.10
Val	1.06	Met	1.05
Asp	1.01	Arg	0.93
His	1.00	Asn	0.89
Arg	0.98	His	0.87
Thr	0.83	Ala	0.83
Ser	0.77	Ser	0.75
Cys	0.70	Gly	0.75
Tyr	0.69	Lys	0.74
Asn	0.67	Pro	0.55
Pro	0.57	Asp	0.54
Gly	0.57	Glu	0.37

Beta-sheets:

	strong former
	former
	weak former
:	indifferent
:::	breaker
::::	strong breaker

+B

+b

-b

-B

$\alpha$   
 $\beta$



# The Chou-Fasman algorithm

- Use a *window* of a certain size that slides along the sequences
- Window length differs for  $\alpha$ -helices and  $\beta$ -sheets
- First, let's consider alpha helices:
  - Find a region that has high  $\alpha$ -helix forming potential (*nucleation* region); that is, in the sliding window at least 4 out of 6 residues that are “+a” or “+A”
  - Move the sliding window. Keep prolonging the  $\alpha$ -helix as long as four consecutive amino acids in a window have average score  $>$  threshold
- Do the same for  $\beta$ -sheets (slightly different nucleation condition: 3 out of 5)
- In ambiguous cases (both  $\alpha$ -helix and  $\beta$ -sheet), take the value that is higher



## Chou-Fasman *ab initio* structure prediction

Note: simplified rules

Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

### 1. Sequence



## Chou-Fasman *ab initio* structure prediction

Note: simplified rules

Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

0.57 1.00 1.51 1.06 1.51 1.42 1.51 0.57 1.06 0.69 1.06 0.69 0.57

1. Sequence

2. Helix forming property of amino acids



## Chou-Fasman *ab initio* structure prediction

Note: simplified rules

Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A

1. Sequence

2. Helix forming property of amino acids



## Chou-Fasman *ab initio* structure prediction

Note: simplified rules

Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A

0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b

1. Sequence
2. Helix forming property of amino acids
3. Beta sheet forming property of amino acids



## Chou-Fasman *ab initio* structure prediction

Note: simplified rules

Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

0.57 1.00 1.51 1.06 1.51 1.42 1.51 0.57 1.06 0.69 1.06 0.69 0.57  
-A 0 +A +a +A +A +A -A +a -a +a -a -A

0.75 0.87 0.37 1.70 0.37 0.83 0.37 0.75 1.70 1.47 1.70 1.47 0.75  
-b 0 -B +B -B 0 -B -b +B +B +B +B -b



Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H							
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b

(i) Find the nucleation site: 5 out of 6 +a/+A



Gly	His	Glu	Val	Glu	Ala	Glu	Gly	Val	Tyr	Val	Tyr	Gly
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H						
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b

$$(1.06 + 1.51 + 1.42 + 1.51) / 4 = 1.38 > 1.03$$

(ii) Extend the nucleation site using a sliding window  
 if the average score in this region falls below 1.03, stop



Gly	His	Glu	Val	Glu	Ala	Glu	Gly	Val	Tyr	Val	Tyr	Gly
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H					
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b

$$(1.51 + 1.42 + 1.51 + 0.57) / 4 = 1.25 > 1.03$$

- (ii) Extend the nucleation site using a sliding window  
 if the average score in this region falls below 1.03, stop



Gly	His	Glu	Val	Glu	Ala	Glu	Gly	Val	Tyr	Val	Tyr	Gly
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H	H				
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b

$$(1.42 + 1.51 + 0.57 + 1.06) / 4 = 1.14 > 1.03$$

- (ii) Extend the nucleation site using a sliding window  
if the average score in this region falls below 1.03, stop



Gly-	His	-	Glu	-	Val	-	Glu	-	Ala	-	Glu	-	Gly	-	Val	-	Tyr	-	Gly
0.57	1.00		1.51		1.06		1.51		1.42		1.51		0.57		1.06		0.69		0.57
-A	0		+A		+a		+A		+A		+A		-A		+a		-a		-A
H	H		H		H		H		H		H		H		H		H		H
0.75	0.87		0.37		1.70		0.37		0.83		0.37		0.75		1.70		1.47		0.75
-b	0		-B		+B		-B		0		-B		-b		+B		+B		+B

STOP!

$$(1.51 + 0.57 + 1.06 + 0.69) / 4 = 0.96 < 1.03$$

- (ii) Extend the nucleation site using a sliding window  
if the average score in this region falls below 1.03, stop



Gly	His	Glu	Val	Glu	Ala	Glu	Gly	Val	Tyr	Val	Tyr	Gly
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H	H				
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b
								E	E	E	E	E

(i) Find the nucleation site for the beta sheets:  
3 out of 5 residues +b or +B

*BP*





Gly	His	Glu	Val	Glu	Ala	Glu	Gly	Val	Tyr	Val	Tyr	Gly
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H	H				
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b
							E	E	E	E	E	E



$$(0.75 + 1.70 + 1.47 + 1.70) / 4 = 1.41 > 1.0$$

(ii) Extend the nucleation site as long as four consecutive residues have average score greater than 1



Gly-	His	-	Glu	-	Val	-	Glu	-	Ala	-	Glu	-	Gly	-	Val	-	Tyr	-	Val	-	Tyr	-	Gly
0.57	1.00		1.51		1.06		1.51		1.42		1.51		0.57		1.06		0.69		1.06		0.69		0.57
-A	0		+A		+a		+A		+A		+A		-A		+a		-a		+a		-a		-A
H	H		H		H		H		H		H		H		H		H		H		H		H
0.75	0.87		0.37		1.70		0.37		0.83		0.37		0.75		1.70		1.47		1.70		1.47		0.75
-b	0		-B		+B		-B		0		-B		-b		+B		+B		+B		+B		-b
											E		E		E		E		E		E		E

$$(0.37 + 0.75 + 1.70 + 1.47) / 4 = 1.07 > 1.0$$

(ii) Extend the nucleation site as long as four consecutive residues have average score greater than 1



Gly	His	Glu	Val	Glu	Ala	Glu	Gly	Val	Tyr	Val	Tyr	Gly
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H	H				
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b
						E	E	E	E	E	E	E

STOP!

$$(0.83 + 0.37 + 0.75 + 1.70) / 4 = 0.91 < 1.0$$

- (ii) Extend the nucleation site as long as four consecutive residues have average score greater than 1



Gly- His - Glu - Val - Glu - Ala -						Glu - Gly - Val -			Tyr - Val - Tyr -Gly			
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H	H				
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b
						E	E	E	E	E	E	E

Resolve ambiguities for any overlapping region: calculate the average score for helices in the overlap and average score for b sheets in overlap. Whichever is greater, wins.



Gly- His - Glu - Val - Glu - Ala -						Glu - Gly - Val -			Tyr - Val - Tyr -Gly			
0.57	1.00	1.51	1.06	1.51	1.42	1.51	0.57	1.06	0.69	1.06	0.69	0.57
-A	0	+A	+a	+A	+A	+A	-A	+a	-a	+a	-a	-A
H	H	H	H	H	H	H	H	H				
0.75	0.87	0.37	1.70	0.37	0.83	0.37	0.75	1.70	1.47	1.70	1.47	0.75
-b	0	-B	+B	-B	0	-B	-b	+B	+B	+B	+B	-b
						E	E	E	E	E	E	E

Resolve ambiguities for any overlapping region: calculate the average score for helices in the overlap and average score for b sheets in overlap. Whichever is greater, wins.

Average for alpha-helices:  $(1.51 + 0.57 + 1.06) / 3 = 1.05$

Average for betha-sheets:  $(0.37 + 0.75 + 1.70) / 3 = 0.94$



Gly- His - Glu - Val - Glu - Ala - Glu - Gly - Val - Tyr - Val - Tyr -Gly

H H H H H H H H H

E E E E

Resolve ambiguities for any overlapping region: calculate the average score for helices in the overlap and average score for b sheets in overlap. Whichever is greater, wins.

Average for alpha-helices:  $(1.51 + 0.57 + 1.06) / 3 = 1.05$

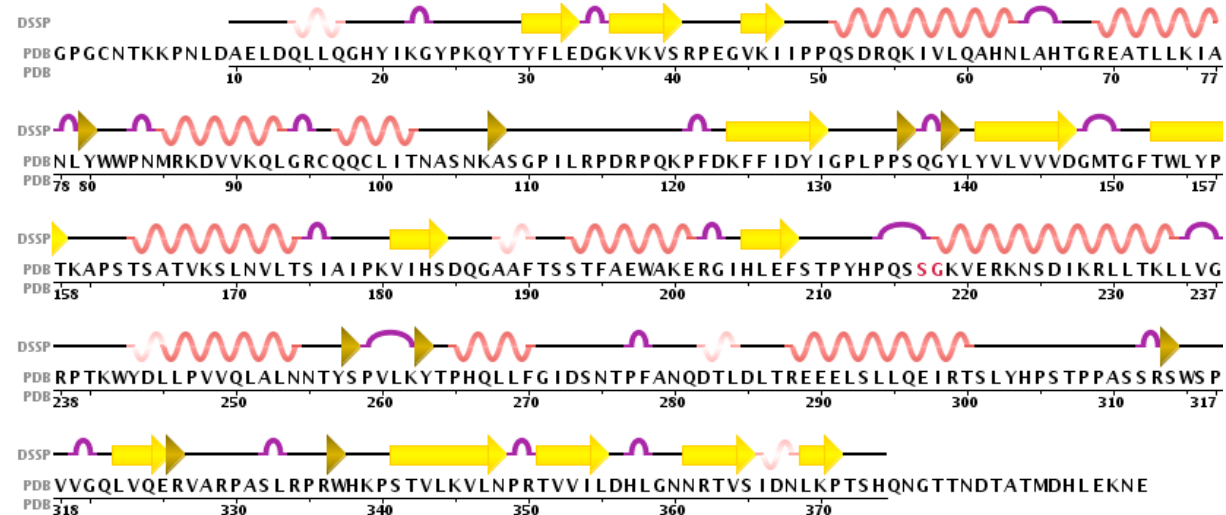
Average for betha-sheets:  $(0.37 + 0.75 + 1.70) / 3 = 0.94$

# The Chou-Fasman algorithm

Precision: ~ 55%

The actual algorithm is slightly more complex, e.g. it also predicts  $\beta$ -turns

Note also that this algorithm is just of historical and educational interest





## Other algorithms

Single sequences:

GOR: Similar to Chou-Fasman, precision  $\sim < 65\%$

NNPRED, BTPRED:  $< 70\%$

Using Multiple Sequence Alignments:

search for homologs, align them, match boundaries of secondary structures better

PHD, PREDATOR, JPRED:  $< 80\%$

SOPMA etc:  $< 80\%$

Consensus structure prediction

Run several algorithms and take the average





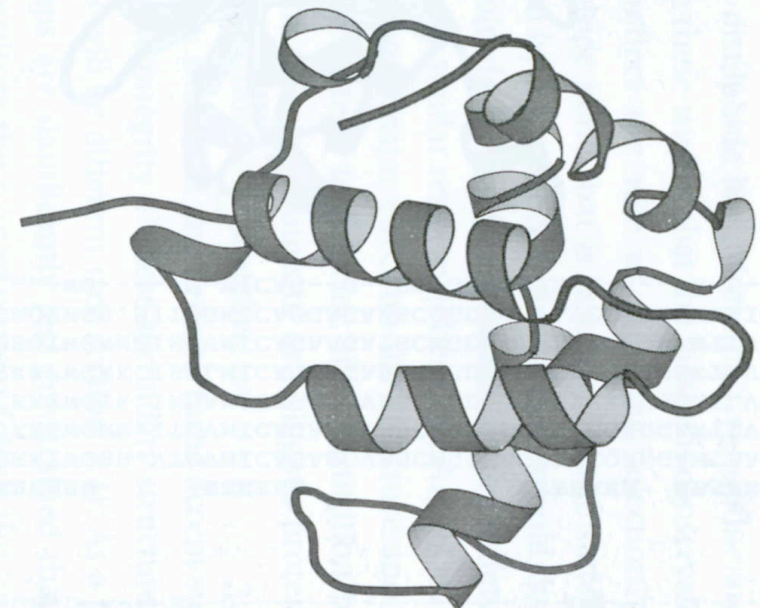
Prediction based on MSA:  
 Note that pronounced secondary structures are better conserved

```

HHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHHH  HHHHHH
hbb_human  VHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKV L GAFSDGLAHL DNLKGTFA TLSELHCDKLHV
hbb_pig     VHLSAEEKEAVLGLWGKV..NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMGNPKVKAHGKKV L QSFSDGLKHL DNLKGTFA KLSELHCDQLHV
hbb_horse   VQLSGEEKA AVLALWDKV..NEEEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMGNPKVKAHGKKV L HSFGE G VHHLDNLKGTFA ALSELHCDKLHV
hbb_bovin   ~MLTAE EKA AVTAFWGKV..KVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKVKAHGKKV L SFSNGMKHL DDLKGTFA ALSELHCDKLHV
hbb_chick   VHWTAE EKQLITGLWGKV..NVAECGAEALARLLIVYPWTQRFFASFGNLS SPTAILGNPMVRAHGKKV L TSFGDAVK NLDNIKNTFSQLSELHCDKLHV
hba_horse   ~VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF.DLSH....GSAQVKAHGKKVGDAL TLAVGH L DDLPGALS NLS DLHAHKLRV
hba_human   ~VLSPADKTNVKAAWGKVG AHAGEYGAEALERMFLSFPTTKTYFPHF.DLSH....GSAQVKGHGKKVADAL TNAVAHVDDMPNALS ALS DLHAHKLRV
hba_bovin   ~VLSAADKGNVKAAWGKVG GHAAEYGAEALERMFLSFPTTKTYFPHF.DLSH....GSAQVKGHGAKVAAAL TKAVEHL DDLPGALS ELS DLHAHKLRV
hba_pig     ~VLSAADKANVKAAWGKVG GQAGAHGAEALERMFLGFPTTKTYFPHF.NLSH....GSDQVKAHGQKVADAL TKAVGH L DDLPGALS ALS DLHAHKLRV
hba_chick   ~VLSAADKNNVKGIFTKIAGHAE EYGAE TLERMFTTYPPTTKTYFPHF.DLSH....GSAQIKGHGKKVVAAL IEAANHIDDIAGT LSKLS DLHAHKLRV
Consensus  -----K-----K-----G-E-L-R-----P-T---F--F--LS-----HG-KV-----D-----LS-LH---L-V
                                                    *
  
```

```

HHHHHHHHHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
hbb_human  DPENFRL L GNVLVCVLAH HFGKEFTPPVQAA YQKV VAGVANALAHKYH
hbb_pig     DPENFRL L GNVI VVVLARRLGHDFNP DVQAAFQKV VAGVANALAHKYH
hbb_horse   DPENFRL L GNVLVVVLARHFGKDFTPELQAS YQKV VAGVANALAHKYH
hbb_bovin   DPENFKLLGNVLVVVLARNFGKEFTPV LQADFQKV VAGVANALAHRYH
hbb_chick   DPENFRL LGDIL IIVLAAHFSKDFTPECQA AWQKLV RVVAHALARKYH
hba_horse   DPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLT SKYR
hba_human   DPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLT SKYR
hba_bovin   DPVNFKLLSHSLLVTLASHLP SDFTPAVHASLDKFLANVSTVLT SKYR
hba_pig     DPVNFKLLSHCLLVTLAAHHPDDFNPSVHASLDKFLANVSTVLT SKYR
hba_chick   DPVNFKLLGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR
Consensus  DP-NF-LL-----A-----P---A---K---V---L---Y-
  
```



(Patthy) *OP*



h = "alpha helix"  
 c = "random coil"  
 e = "extended strand" (probably part of a "beta sheet";  
 2 "strands" = 1 "sheet")  
 t = "turn" (hydrogen bond present)

	10	20	30	40	50	60	70
UNK_3360	MVLSPADKTNVKA	AWGKVG	AHAGEYGA	EALERMFLS	FPTTKTY	FPHFDL	SHGSAQVK
DPM	cchcchc	chhhhhh	chhhhh	thhhhhhh	hhhhhhhh	cccccccc	chcccc
DSC	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
GOR4	ccccccc	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	chhhhhh
HNNC	ccccccc	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
PHD	ccccccc	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	eeeecccc	hhhhhhh
Predator	ccccccc	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	heeecc
SIMPA96	cecccc	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	chhhhhh
SOPM	eecccc	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	ccccccc	tecccc
Sec. Cons.	cccccc	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	chhhhhh
PDB		HHHHHHHHHHHH	HHHHHHHHHHHH			HHHHHHHHHHHH	

	80	90	100	110	120	130	140
UNK_3360	VAHVDDMPNALS	SDLHAHKLR	VPVNFKLL	SHCLLVTL	AAHLPAEFT	PAVHASL	DKFLASV
DPM	hhhhhhh	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	eeeeech
DSC	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
GOR4	hhhhc	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	eeeece
HNNC	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
PHD	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
Predator	hhcccc	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	cccc
SIMPA96	hhcccc	chhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	cccc
SOPM	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
Sec. Cons.	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh	hhhhhhh
PDB	HH	HHHHTHHHHHH	HHHHHHHHHHHH	HHHHHHHHHHHH	HHHHHHHHHHHH	HHHHHH	HHHHHH



# Consensus prediction

- h = "alpha helix"
- c = "random coil"
- e = "extended strand" (probably part of a "beta sheet";  
2 "strands" = 1 "sheet")
- t = "turn" (hydrogen bond present)

	10	20	30	40	50	60	70
UNK_3360	MVLS	SPADKTNVKA	AWGKVG	AHAGEYGA	EALERMFLS	FPTTKTY	FPHFDLSHG
DPM	cc	ch	ch	hh	hh	hh	hh
DSC	ch	hh	hh	hh	hh	hh	hh
GOR4	cc	cc	cc	cc	cc	cc	cc
HNNC	cc	cc	cc	cc	cc	cc	cc
PHD	cc	cc	cc	cc	cc	cc	cc
Predator	cc	cc	cc	cc	cc	cc	cc
SIMPA96	cc	cc	cc	cc	cc	cc	cc
SOPM	ee	ee	ee	ee	ee	ee	ee
Sec. Cons.	cc	cc	cc	cc	cc	cc	cc

PDB      HHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHH    HHHHHHHHHHHHHHHHH

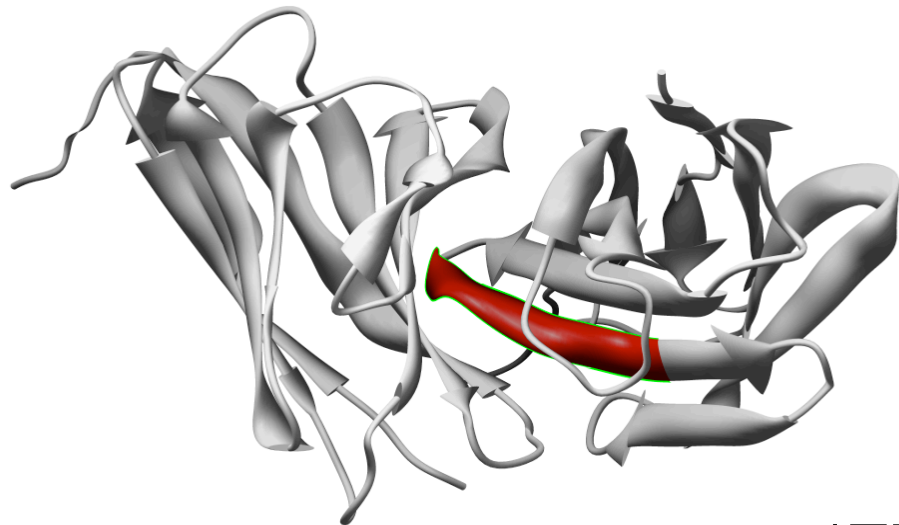
	80	90	100	110	120	130	140
UNK_3360	VAHVDD	MPNALSAL	SDLHAH	KLRVDP	VNFKLL	SHCLLV	TAAHLPA
DPM	hh	hh	hh	hh	hh	hh	hh
DSC	hh	hh	hh	hh	hh	hh	hh
GOR4	hh	hh	hh	hh	hh	hh	hh
HNNC	hh	hh	hh	hh	hh	hh	hh
PHD	hh	hh	hh	hh	hh	hh	hh
Predator	hh	hh	hh	hh	hh	hh	hh
SIMPA96	hh	hh	hh	hh	hh	hh	hh
SOPM	hh	hh	hh	hh	hh	hh	hh
Sec. Cons.	hh	hh	hh	hh	hh	hh	hh

PDB      HH    HHHHTHHHHHHHHHH    HHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHHH    HHHHHH

## Limits of *ab initio* prediction

- Folding depends on context
- Interactions with distant residues or other proteins can be essential
- The same fragment can have different structure in different proteins

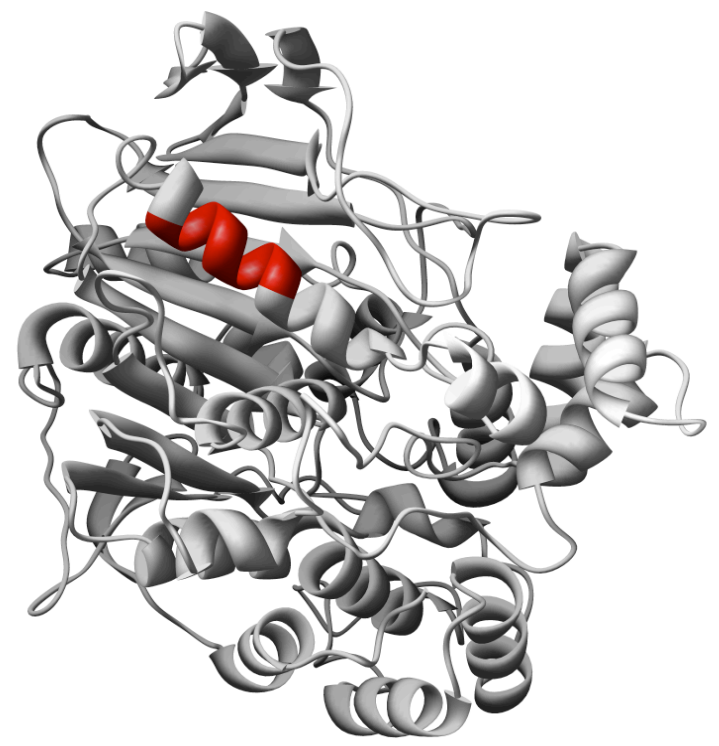




**1IGM**  
*Immunoglobulin heavy chain*

EVHLLLESGGNLVQPGGSLRLSCAA  
SGFTFNI FVMSWVRQAPG**KGLEWV**  
**S**GVFGSGGNTDYADAVKGRFTITR  
DNSKNTLYLQMNSLRAEDTAIYYC  
AKHRVSYVLTGFDSWGQGLVTVS  
SGSASAPTL

**1THG**  
*Lipase*



EAPTAVLNGNEVISGVLEGKVDTFKGI PFADPPLNDLRFKHPQPFTGSYQG  
LKANDFSPACMQLDPGNSLTLDDKALGLAKVIPEEFRGPLYDMAKGTVSMN  
EDCLYLNVFRPAGTKPDAKLPVMVWIYGGAFVYGSSAAYPGNSYVKESINM  
GQPVVVFSIN YRTGPF GFLGGDAITAEGNTNAGLHDQR**KGLEWVS**DNIANF  
GGDPDKVMIFGESAGAMSVAHQLIAYGGDNTYNGKKLFHSAILQSGGGLPY  
HDSSVGPDISYNRFAQYAGCDTSASANDTLECLRSKSSSVLHDAQNSYDL  
KDLFGLLPQFLGFGPRPDGNIIPDAAYELFRSGRYAKVPYISGNQEDEGTA  
FAPVALNATTTPHVKKWLQYIFYDASEASIDRVLSLYPQTL SVGSPFRTGI  
LNALTPQFKRVAAILSDMLFQSPRRVMLSATKDVNRWTYLSTHLHNLVPFL  
GTFHGNELIFQFNVNIGPANSYLRFYISFANHDPNVGTNLLQWDQYTDEG  
KEMLEIHMTDNVMRTDDYRIEGISNFETDVNLYG

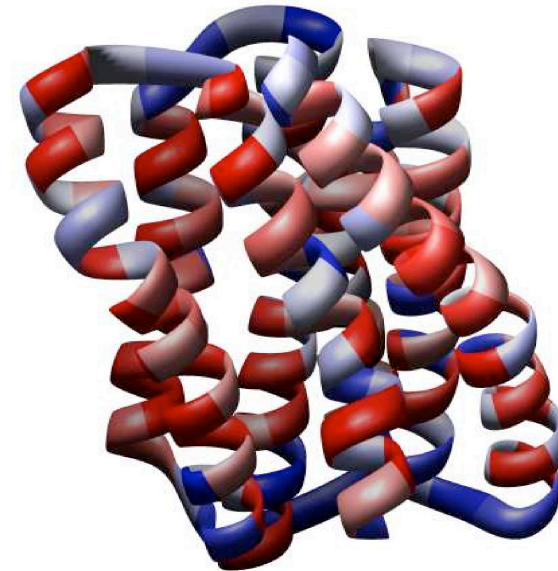
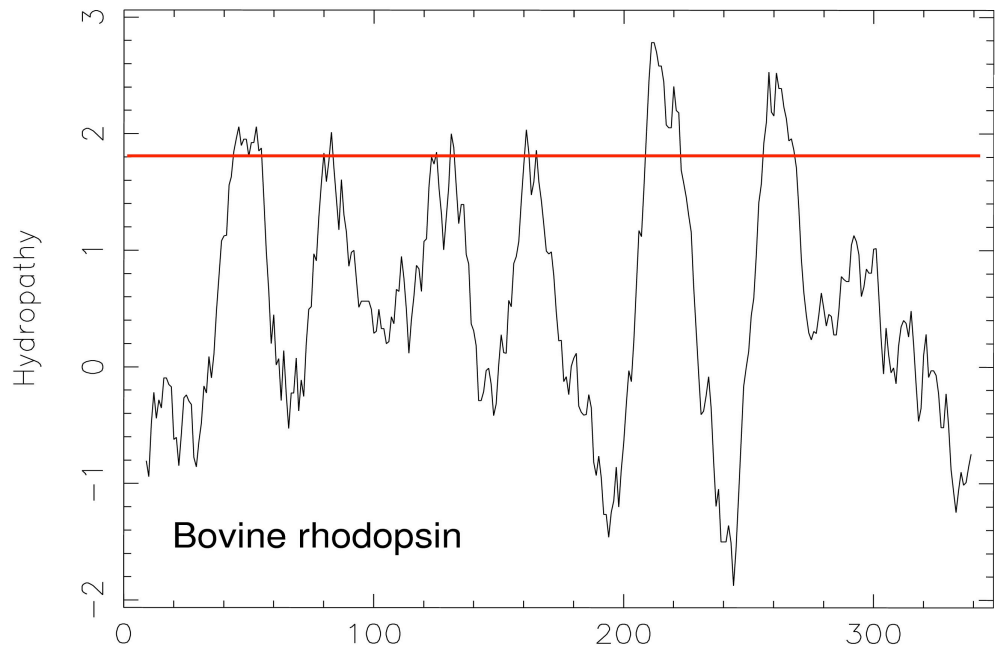
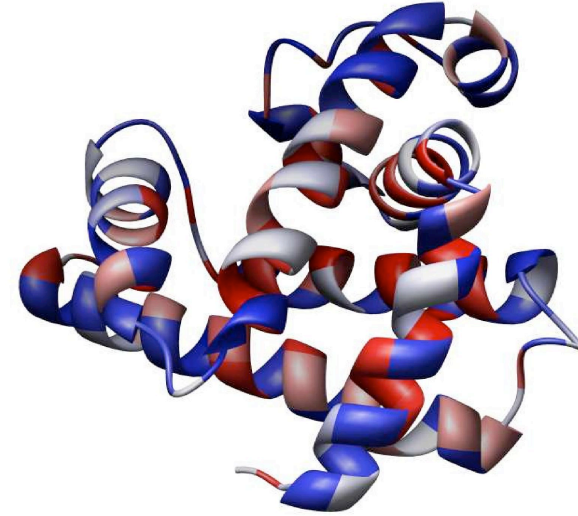
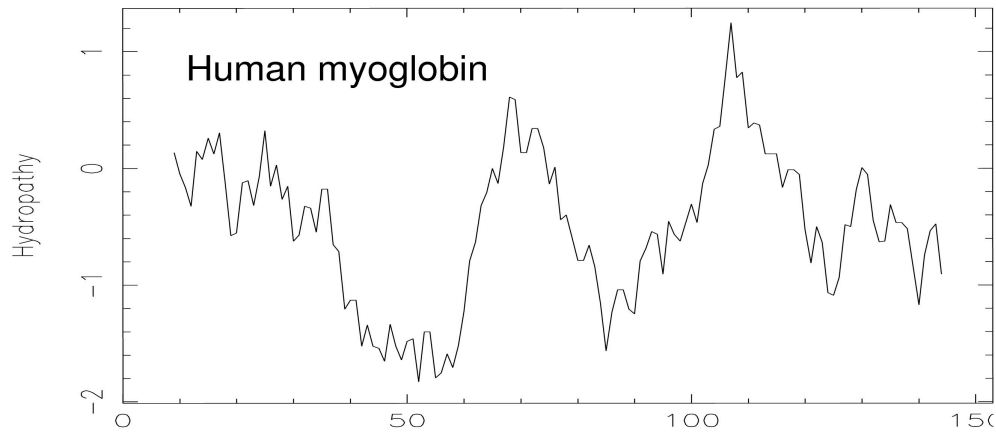


## Hydropathy profiles / plots (“Kyte-Doolittle”)

Measure the average hydrophobicity in a sliding window

## Amino acid scale values:

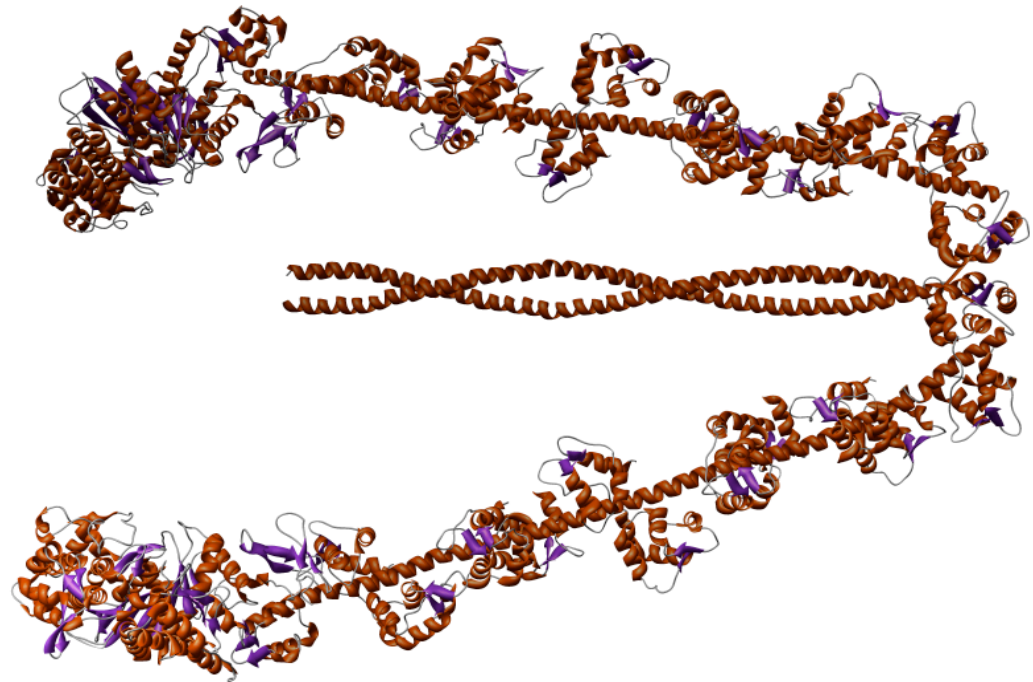
Ala:	1.800
Arg:	-4.500
Asn:	-3.500
Asp:	-3.500
Cys:	2.500
Gln:	-3.500
Glu:	-3.500
Gly:	-0.400
His:	-3.200
Ile:	4.500
Leu:	3.800
Lys:	-3.900
Met:	1.900
Phe:	2.800
Pro:	-1.600
Ser:	-0.800
Thr:	-0.700
Trp:	-0.900
Tyr:	-1.300
Val:	4.200



Scale: Kyte–Doolittle (1982)  
Window length: 17

# Coiled-coil Structure

Myosine – motor protein in eukaryotic cells; converts chemical energy into kinetic energy / movement

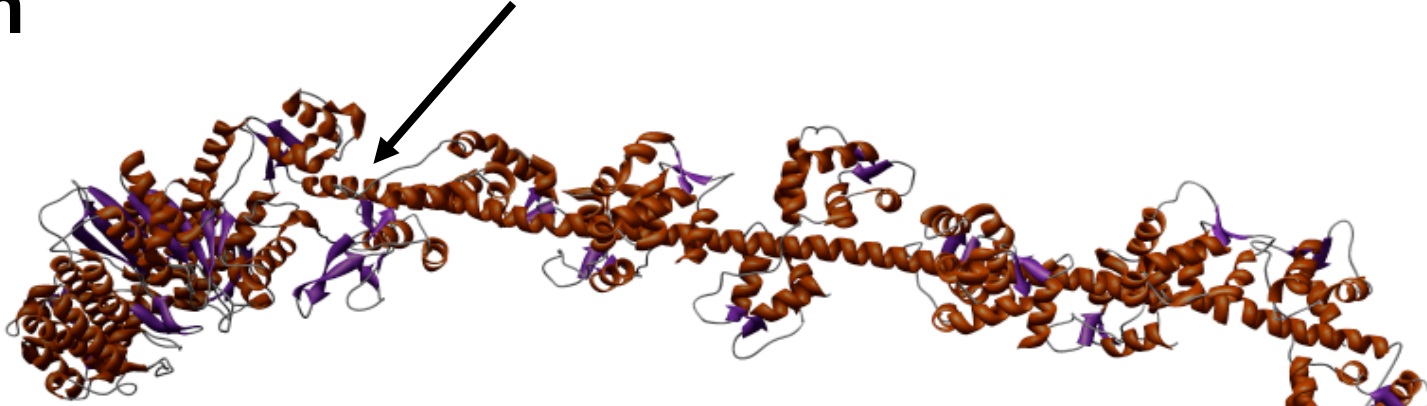




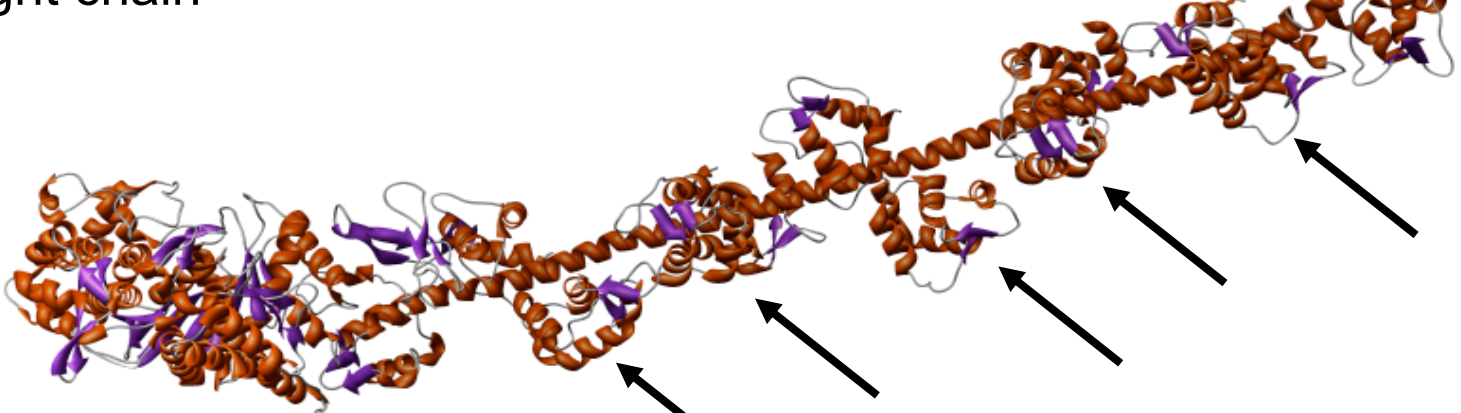


# Myosin

Myosin – heavy chain



Myosin – light chain

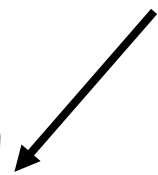


Calmodulin binding myosin

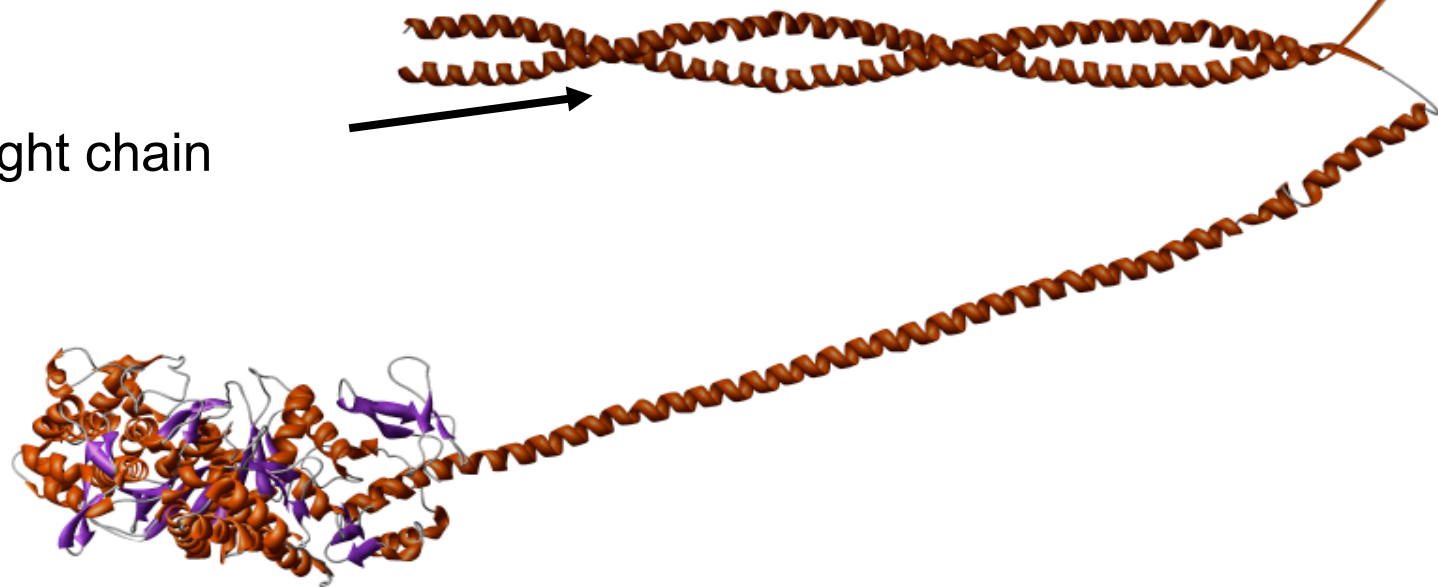


# Myosin

Myosin – heavy chain



Myosin – light chain  
*Coiled coil*



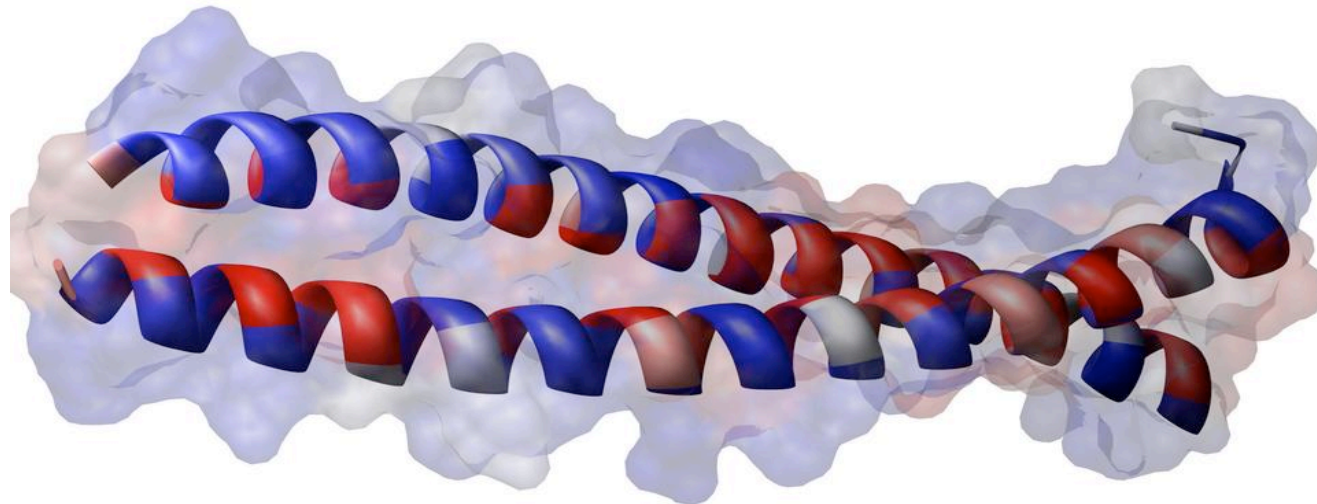
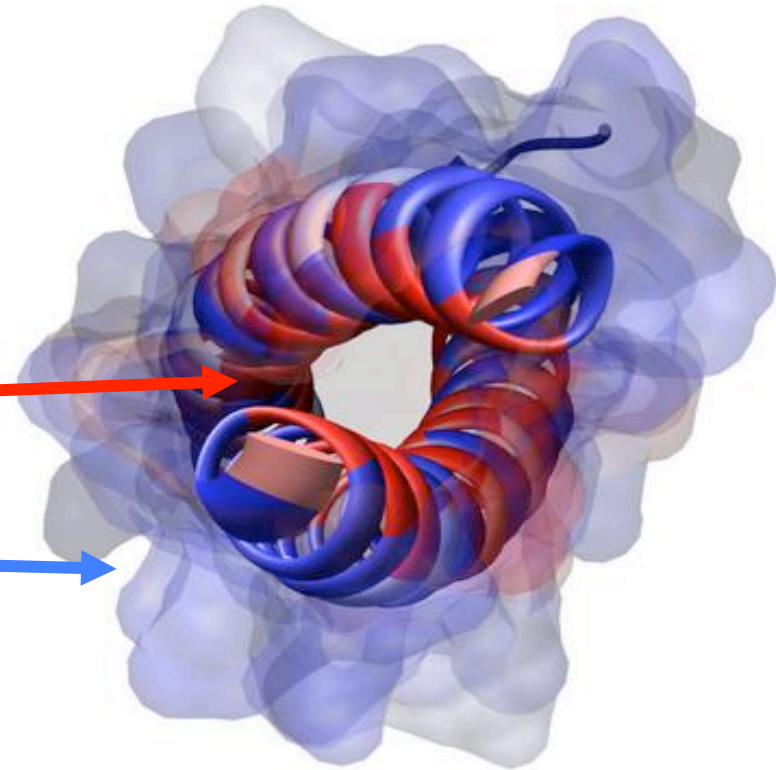
# Coiled-coil structure

2 alpha-Helices

**Hydrophobic** amino acid on the inner, contact surface

**Polar** amino acids on the outer surface

Holds the molecules together like a velcro tape (*Klettverschluss*)



PDB Id: 1gk6

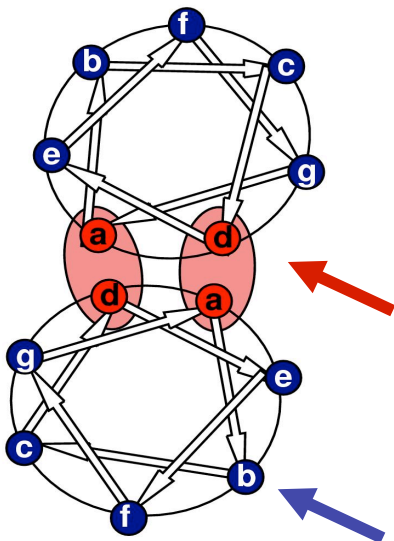
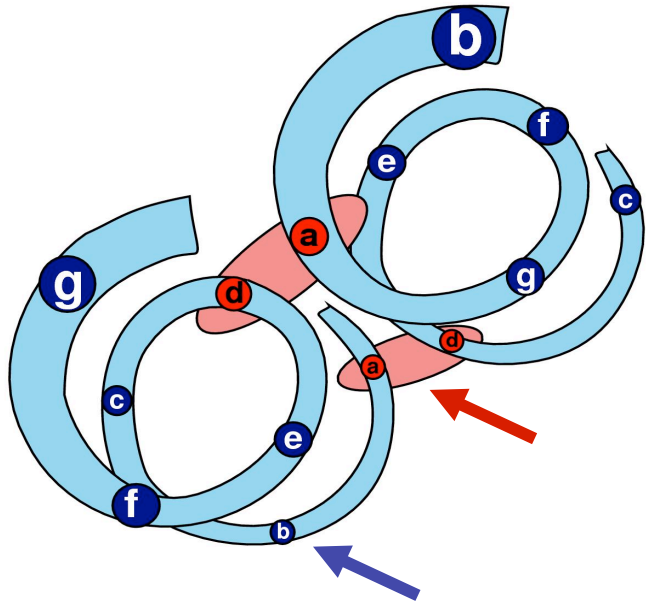


# Prediction of coiled-coils

Same principle – window sliding

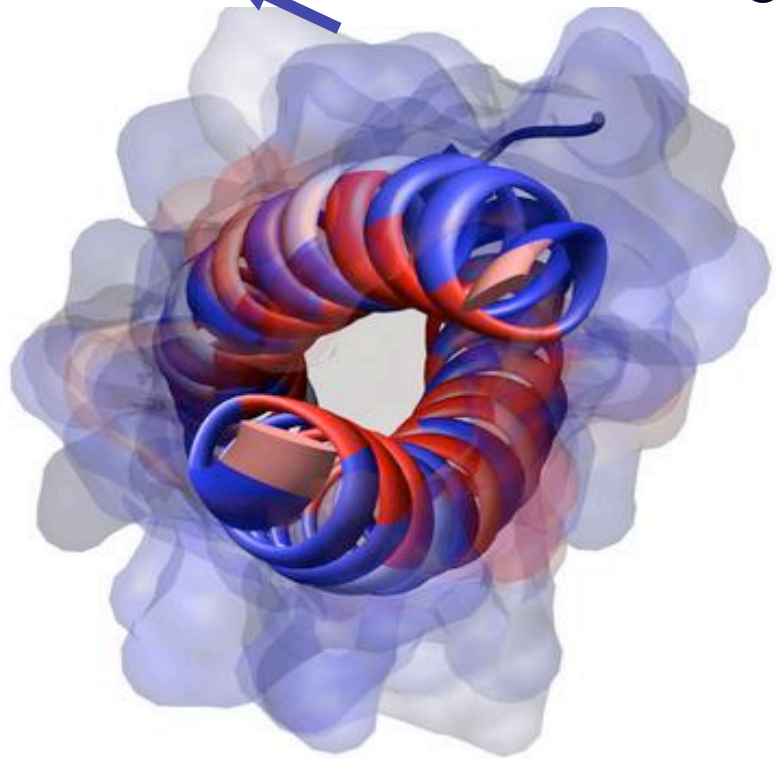
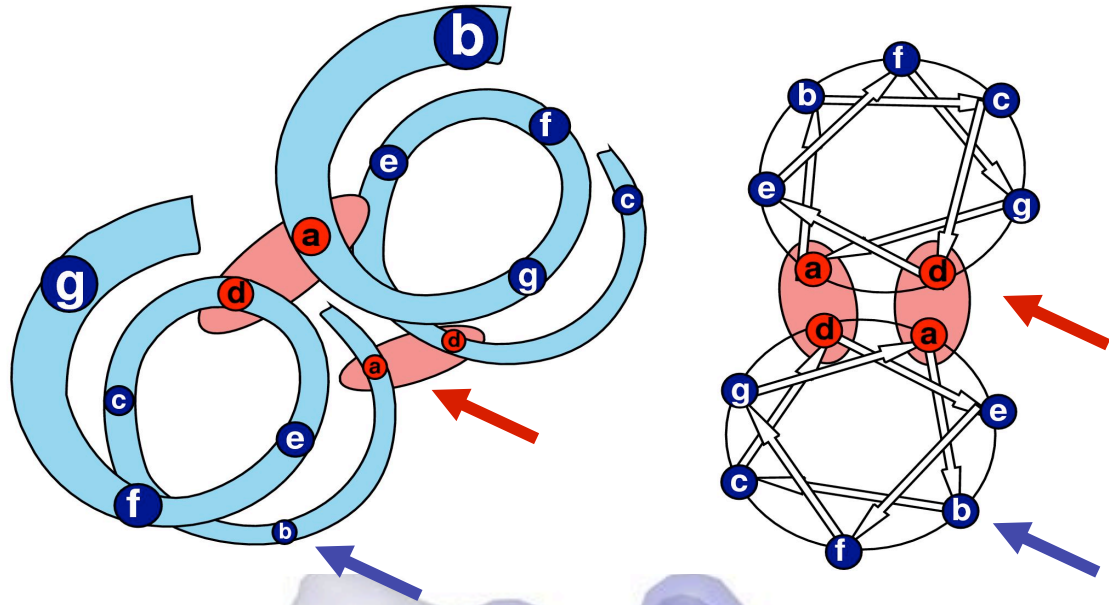
Difference: scores depend on the position in the heptade

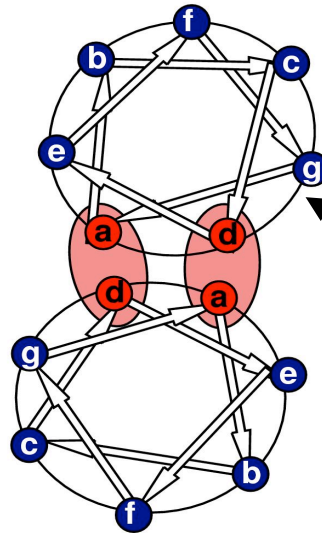
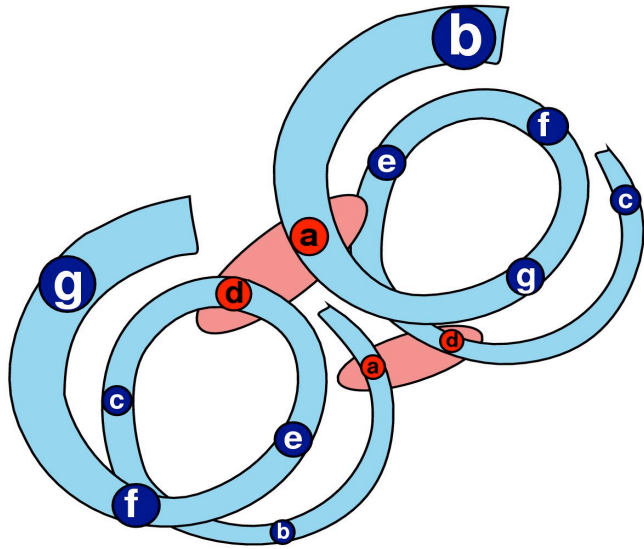




Hydrophobic, inner contact surface

Polar outer surface



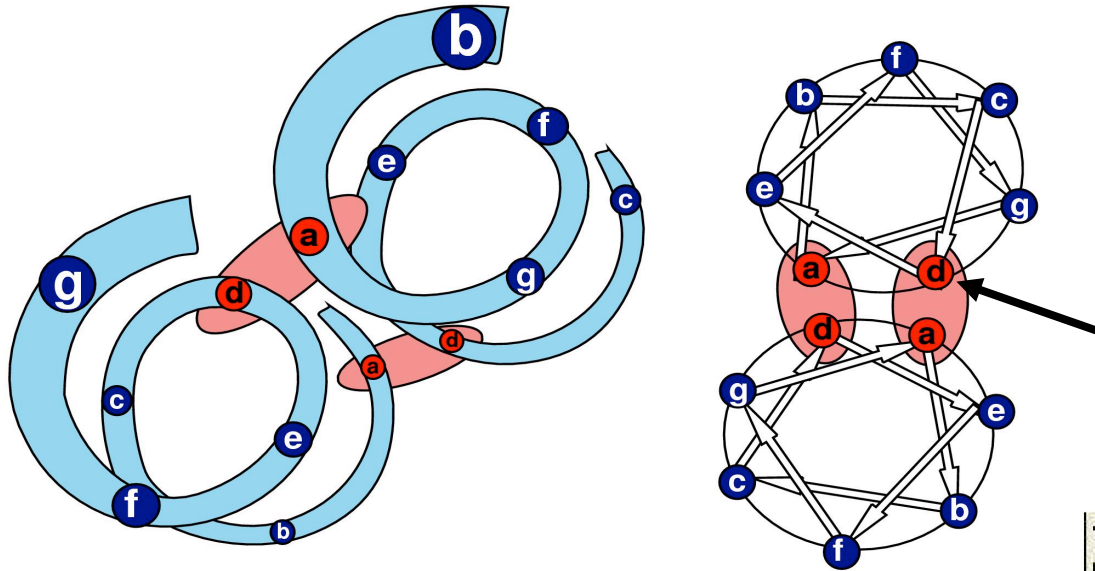


Position in the heptade

Amino acid sequence

Experimental matrix derived from alignments of coiled-coil proteins (log-odd scores)

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



Position in the Heptade



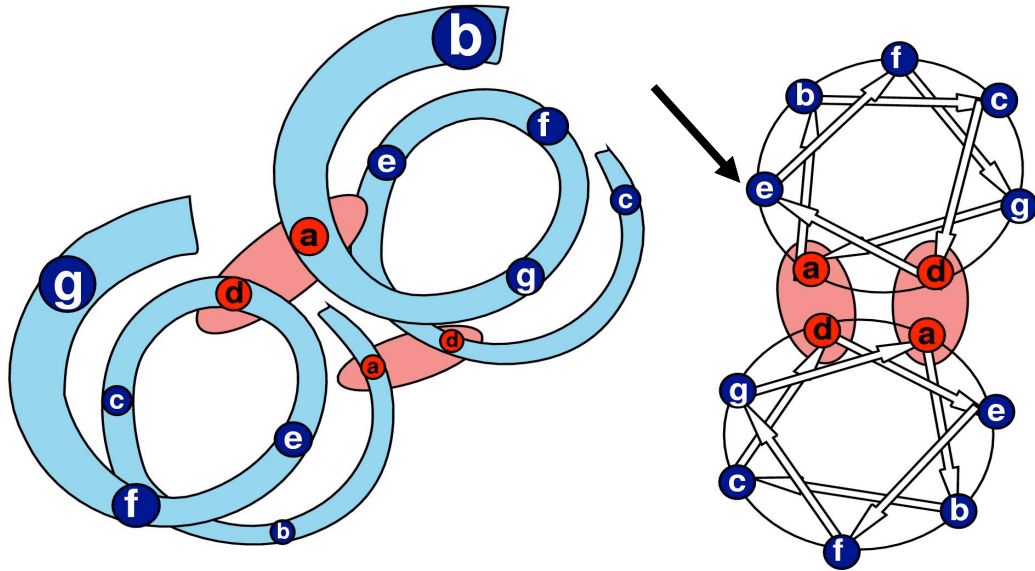
Amino acid



%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
<b>A</b>	1.283	1.364	1.071	<b>2.219</b>	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007

“A” (Alanin, **hydrophobic**) has a score of **2.219** if at position **d**





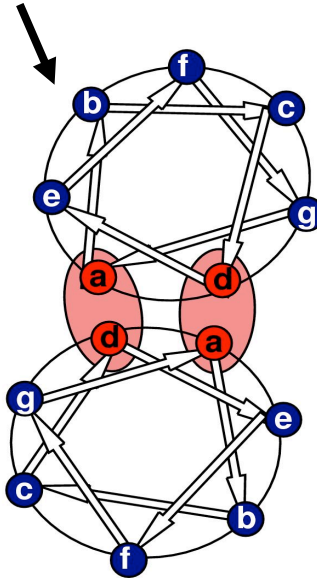
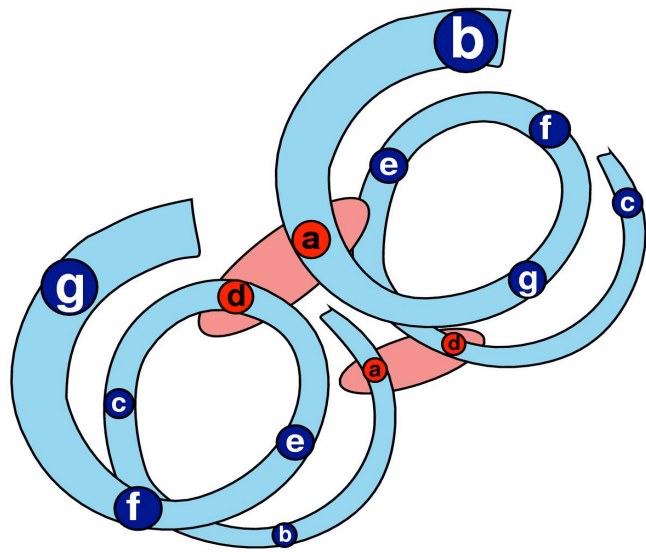
Position in the Heptade

Amino acid



%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.211	<b>0.490</b>	0.265	0.903
K	1.233	2.194	1.817	0.611	2.033	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007

“A” (Alanin, **hydrophobic**) has a score of **0.49** at position **e** (if it is a coiled coil, then we should not find hydrophobic residues on the outside of the alpha helix)

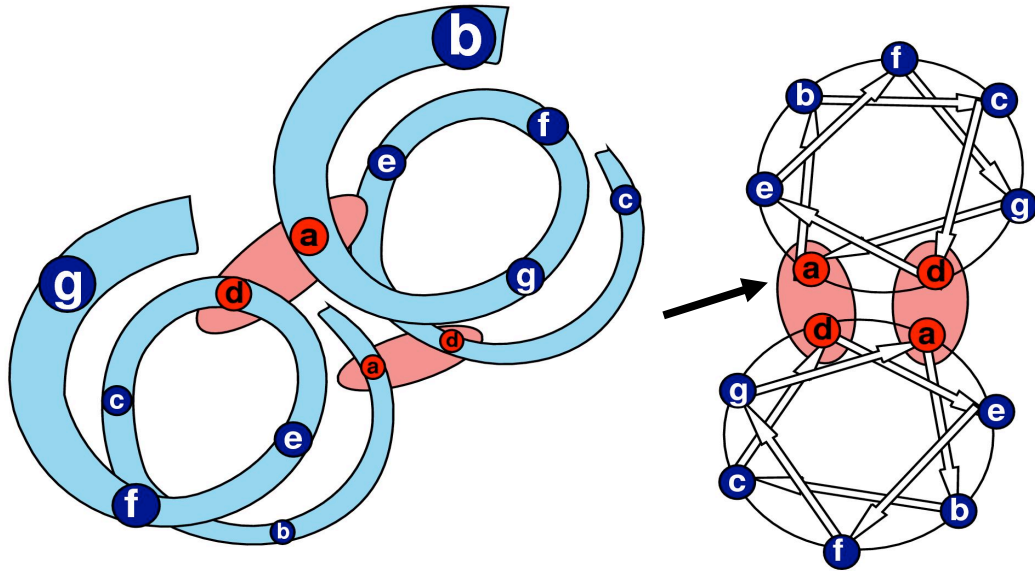


Amino acid

“D” (Aspartate, polar) scores **2.103** at position “b”

Position in the Heptade

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	2.851	2.998	0.789	4.868	2.735	3.812
D	0.061	<b>2.103</b>	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



Amino acid

“D” (Aspartate, polar) scores **0.068** at position “a”

Position in the Heptade

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	<b>0.068</b>	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



**a** **b** **c** **d** **e** **f** **g**

**V A L D L E A**

Calculate the score of a single heptade

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



**b** **c** **d** **e** **f** **g**

**a**

**V A L D L E A**

1.525

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



**c** **d** **e** **f** **g**

**a** **b**

**V A L D L E A**

1.525

1.364

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	2.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



**d** **e** **f** **g**

**a** **b** **c**

**V A L D L E A**

1.525

1.364

0.367

%	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	<b>e</b>	<b>f</b>	<b>g</b>
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.245	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.789	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



e f g

a b c d

V A L D L E A

1.525

1.364

0.367

0.182

%	a	b	c	d	e	f	g
L	2.998	0.269	0.367	3.852	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.265	0.903
K	1.233	2.194	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.783	4.868	2.735	3.812
D	0.068	2.103	1.646	0.182	0.664	1.581	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007





a b c d e f g  
**V A L D L E A**  
 1.525  
 1.364  
 0.367  
 0.182  
 0.510  
 2.735  
 0.903

%	<span style="color: red;">a</span>	<span style="color: blue;">b</span>	<span style="color: blue;">c</span>	<span style="color: red;">d</span>	<span style="color: blue;">e</span>	<span style="color: blue;">f</span>	<span style="color: blue;">g</span>
L	2.998	0.269	0.367	3.851	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.205	0.903
K	1.233	2.134	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.783	4.868	2.735	0.812
D	0.068	2.103	1.646	0.182	0.664	1.501	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



Calculate the geometric mean

a b c d e f g  
**V A L D L E A**

$$(1.525 * 1.364 * 0.367 * 0.182 * 0.510 * 2.735 * 0.903) = 0.172$$

$$0.172^{1/7} = \mathbf{0.780}$$

%	<span style="color: red;">a</span>	<span style="color: blue;">b</span>	<span style="color: blue;">c</span>	<span style="color: red;">d</span>	<span style="color: blue;">e</span>	<span style="color: blue;">f</span>	<span style="color: blue;">g</span>
L	2.998	0.269	0.367	3.85	0.510	0.514	0.562
I	2.408	0.261	0.345	0.931	0.402	0.440	0.289
V	1.525	0.479	0.350	0.887	0.286	0.350	0.362
M	2.161	0.605	0.442	1.441	0.607	0.457	0.570
F	0.490	0.075	0.391	0.639	0.125	0.081	0.038
Y	1.319	0.064	0.081	1.526	0.204	0.118	0.096
G	0.084	0.215	0.432	0.111	0.153	0.367	0.125
A	1.283	1.364	1.077	2.219	0.490	1.205	0.903
K	1.233	2.134	1.817	0.611	2.095	1.686	2.027
R	1.014	1.476	1.771	0.114	1.667	2.006	1.844
H	0.590	0.646	0.584	0.842	0.307	0.611	0.396
E	0.281	3.351	2.998	0.783	4.868	2.735	0.812
D	0.068	2.103	1.646	0.182	0.664	1.501	1.401
Q	0.311	2.290	2.330	0.811	2.596	2.155	2.585
N	1.231	1.683	2.157	0.197	1.653	2.430	2.065
S	0.332	0.753	0.930	0.424	0.734	0.801	0.518
T	0.197	0.543	0.647	0.680	0.905	0.643	0.808
C	0.918	0.002	0.385	0.440	0.138	0.432	0.079
W	0.066	0.064	0.065	0.747	0.006	0.115	0.014
P	0.004	0.108	0.018	0.006	0.010	0.004	0.007



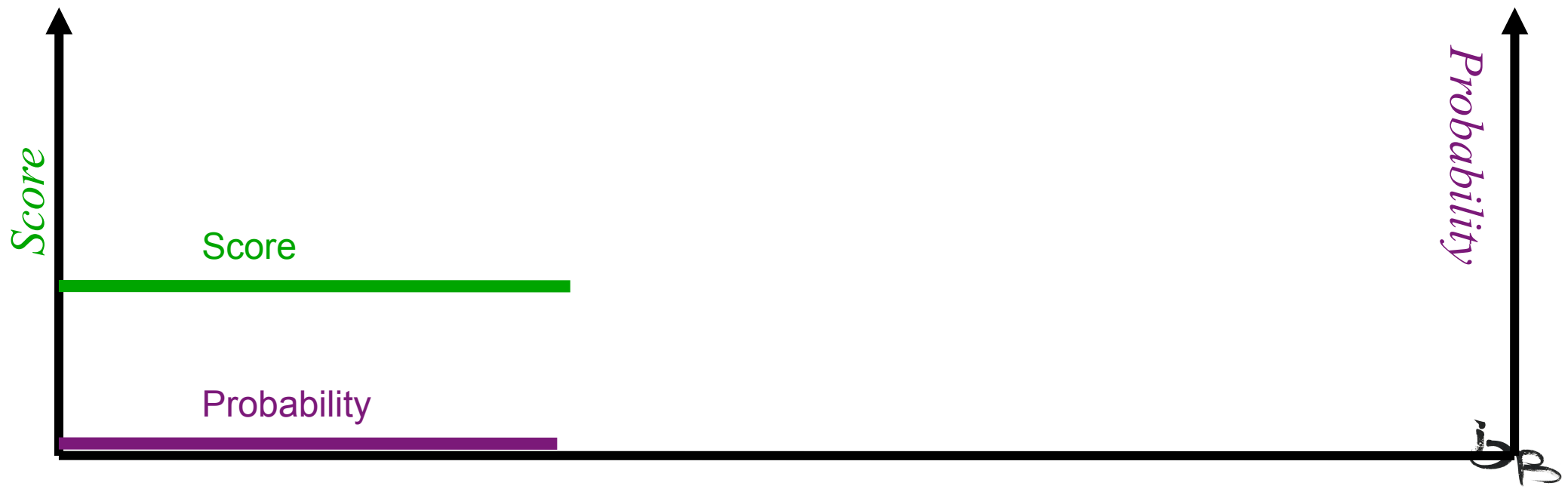
# Coils Algorithm: sliding window

0.780

First position

**a** **b** **c** **d** **e** **f** **g**

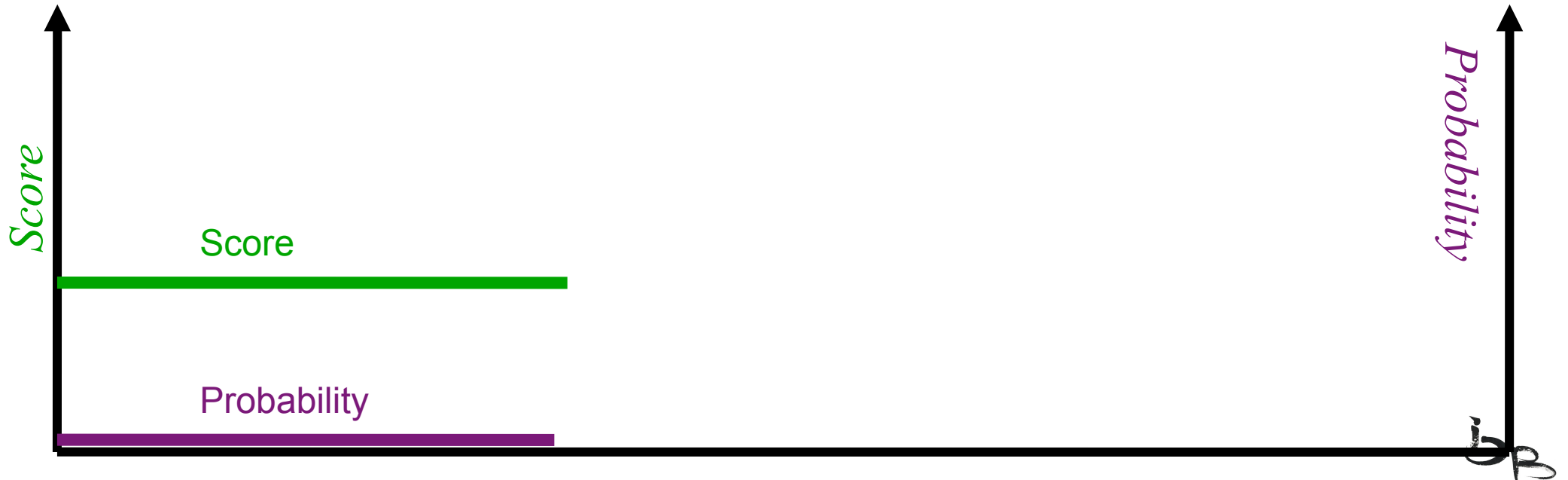
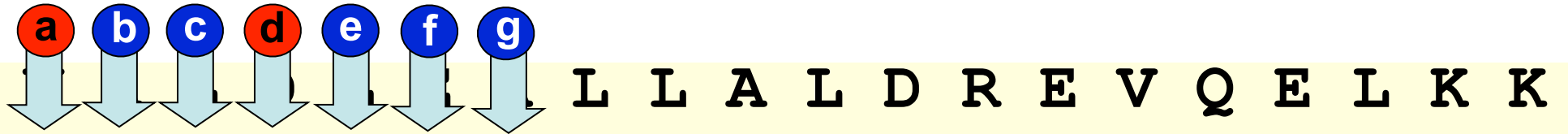
V A L D L E A L L A L D R E V Q E L K K



# Coils Algorithm: sliding window

0.780

Apply the values

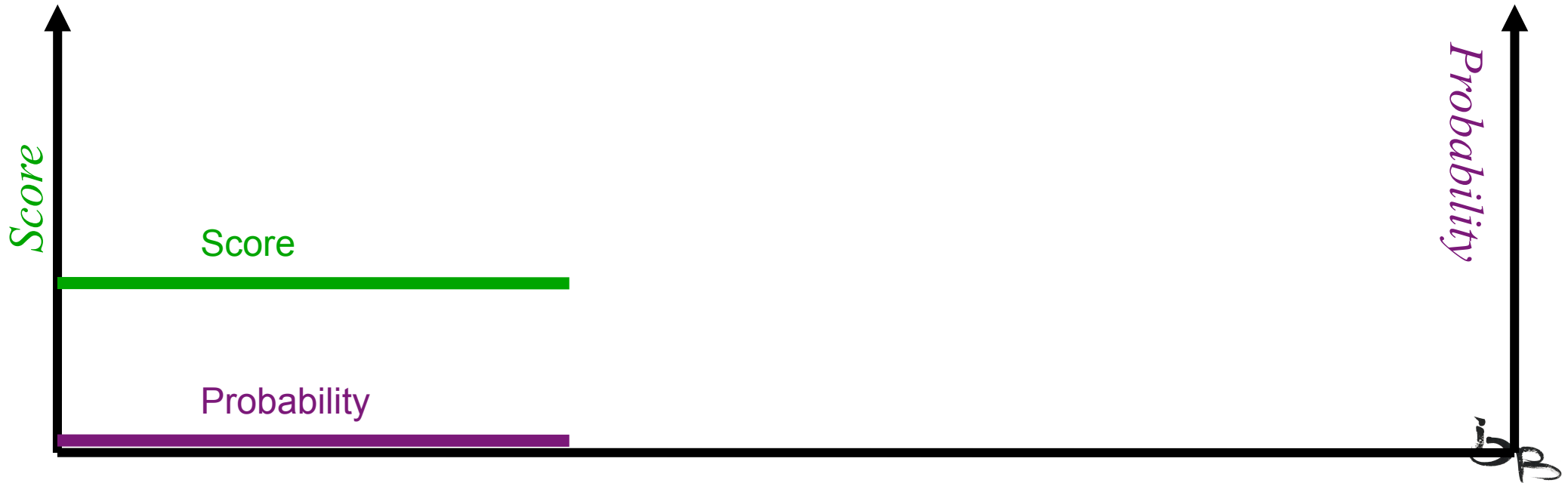




Apply the values

V A L D L E A L L A L D R E V Q E L K K  
a b c d e f g  
.....

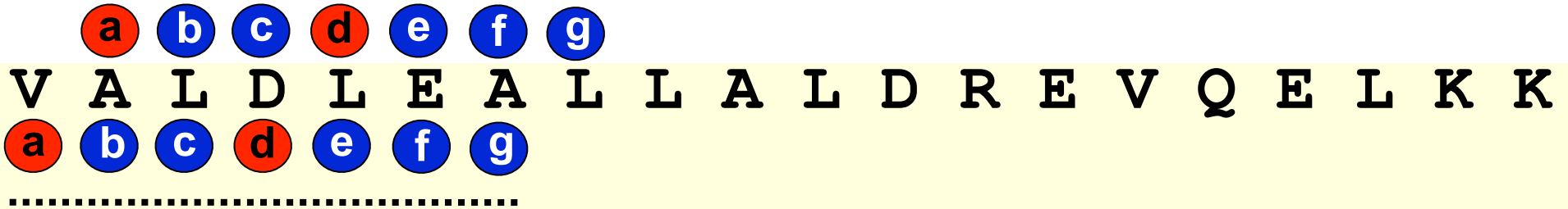
0.780



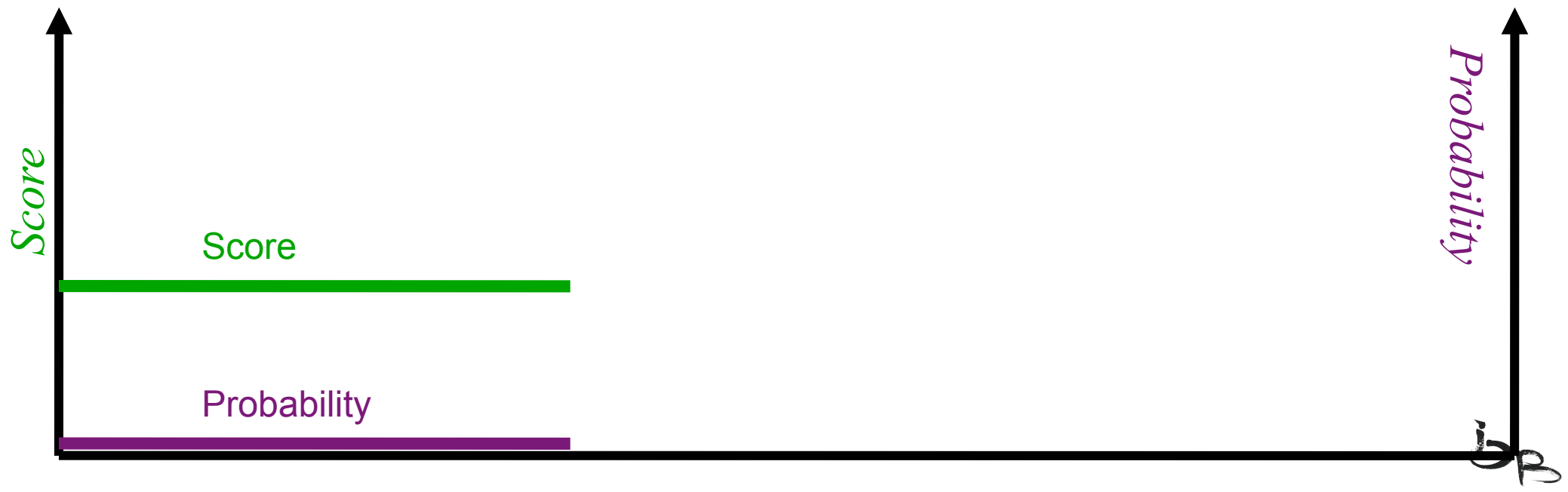


1.335

Second position



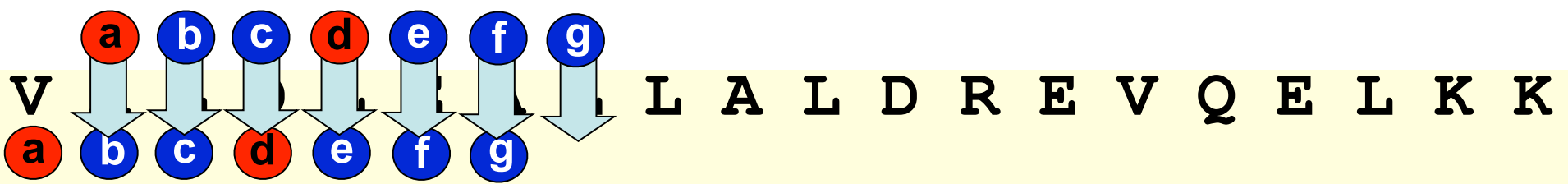
0.780





1.335

Apply the values



0.780

“L”: no previous value, 1.335 used

“ALDLEA”:  $1.335 > 0.780$ , therefore 1.335 replaces 0.780

“V”: only previous value, 0.780 kept

Score

Probability

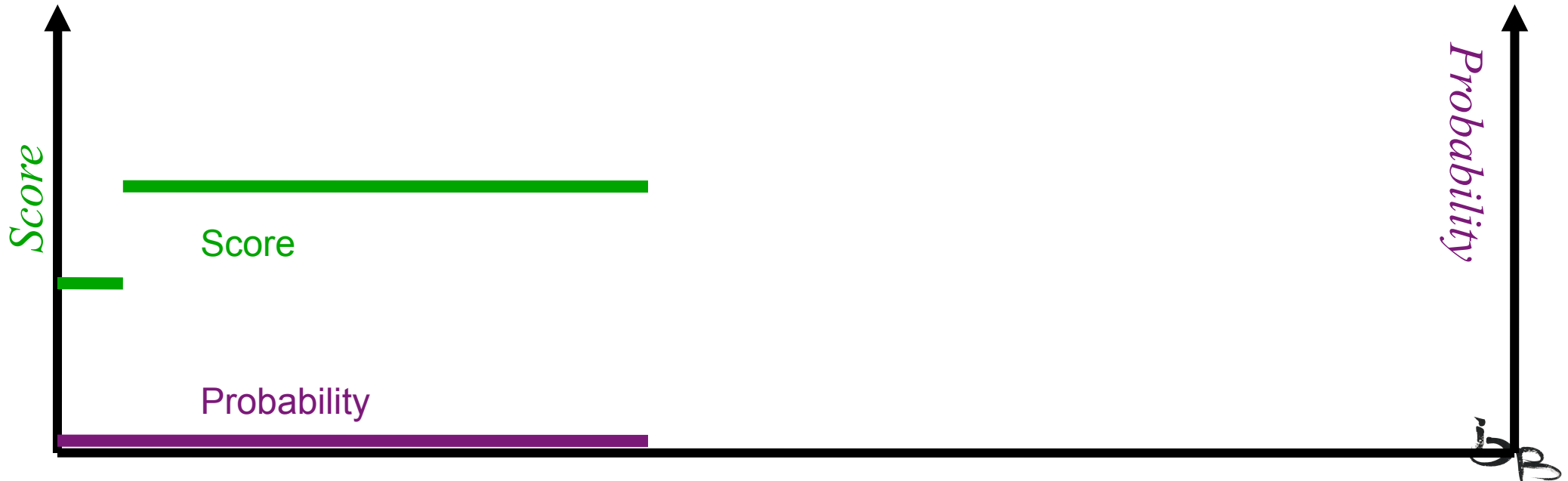


Apply the values

V A L D L E A L L A L D R E V Q E L K K  
a a b c d e f g  
.....

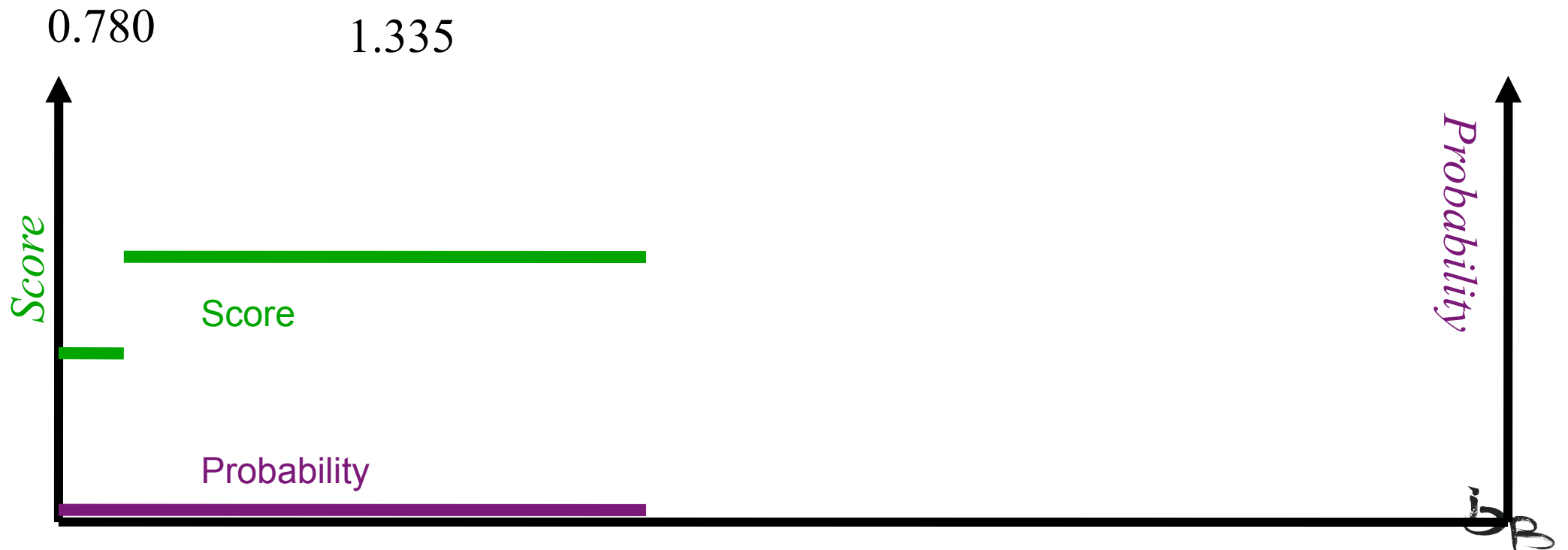
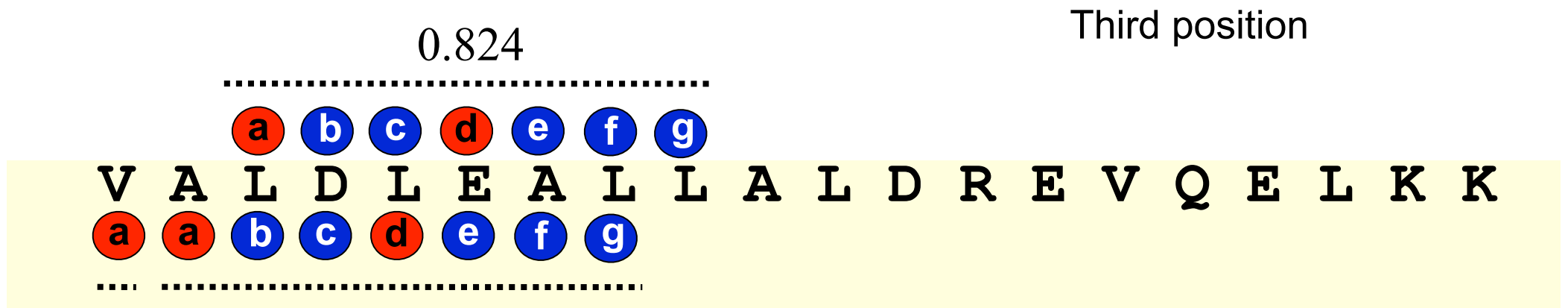
0.780

1.335



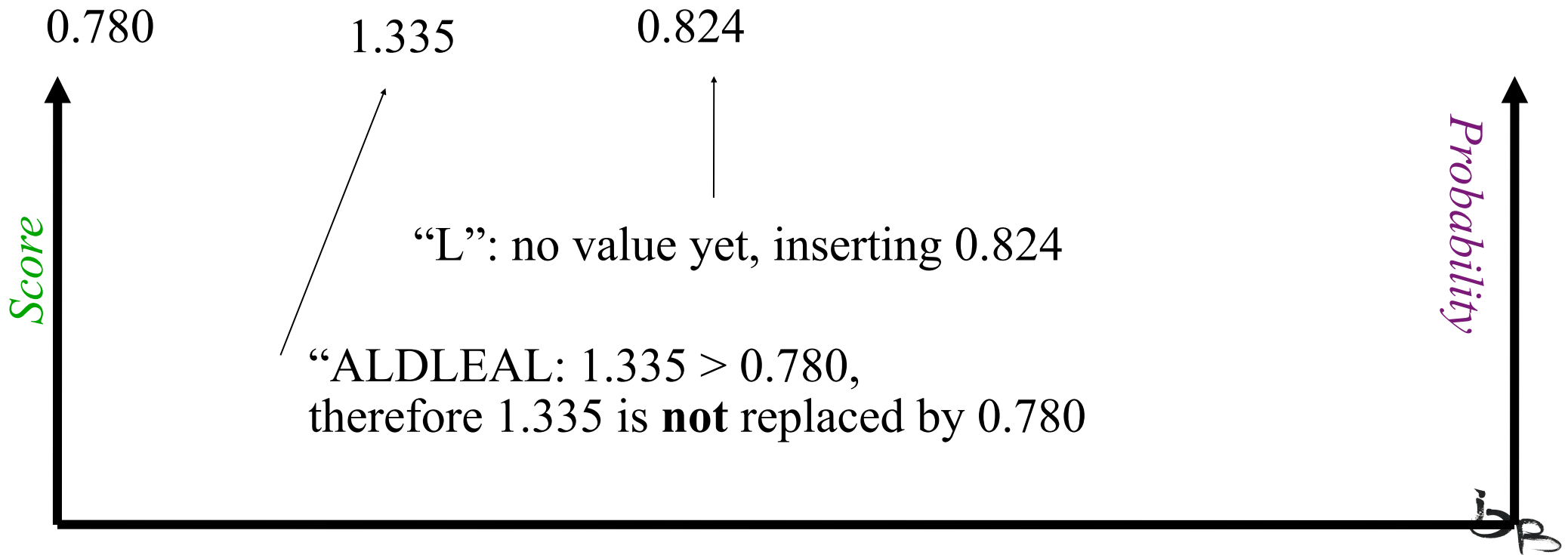
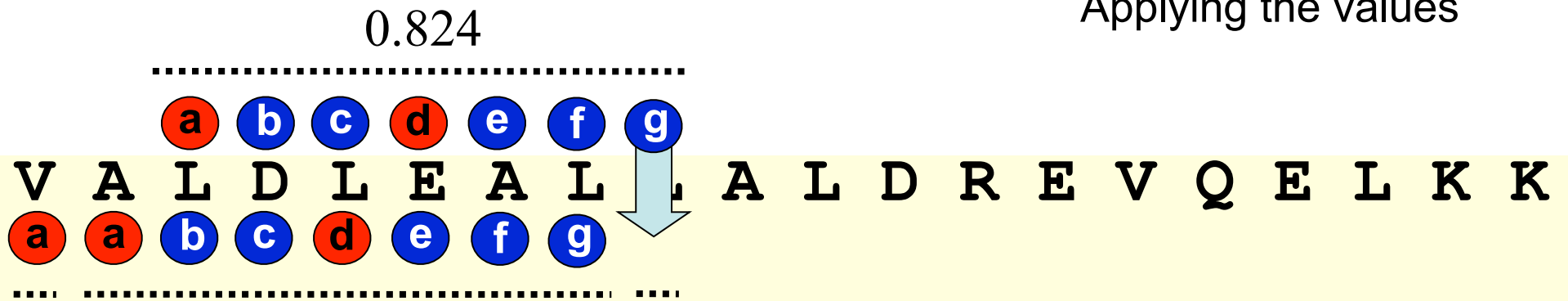


# Coils Algorithm: gesamte Prozedur





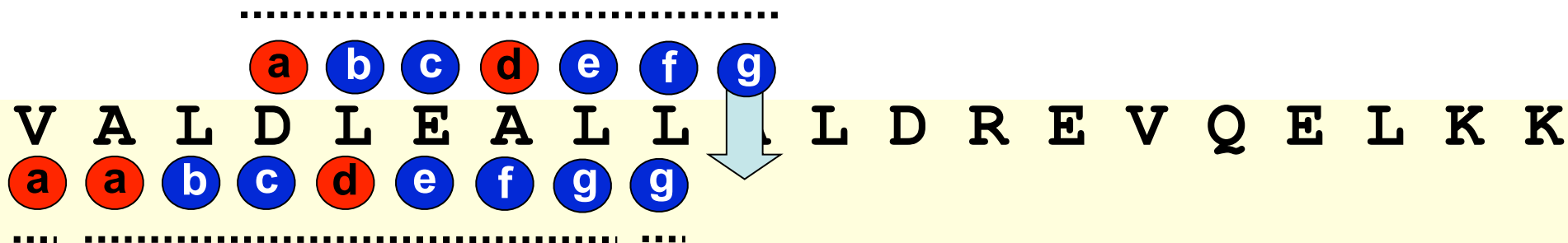
Applying the values





0.602

Fourth Position

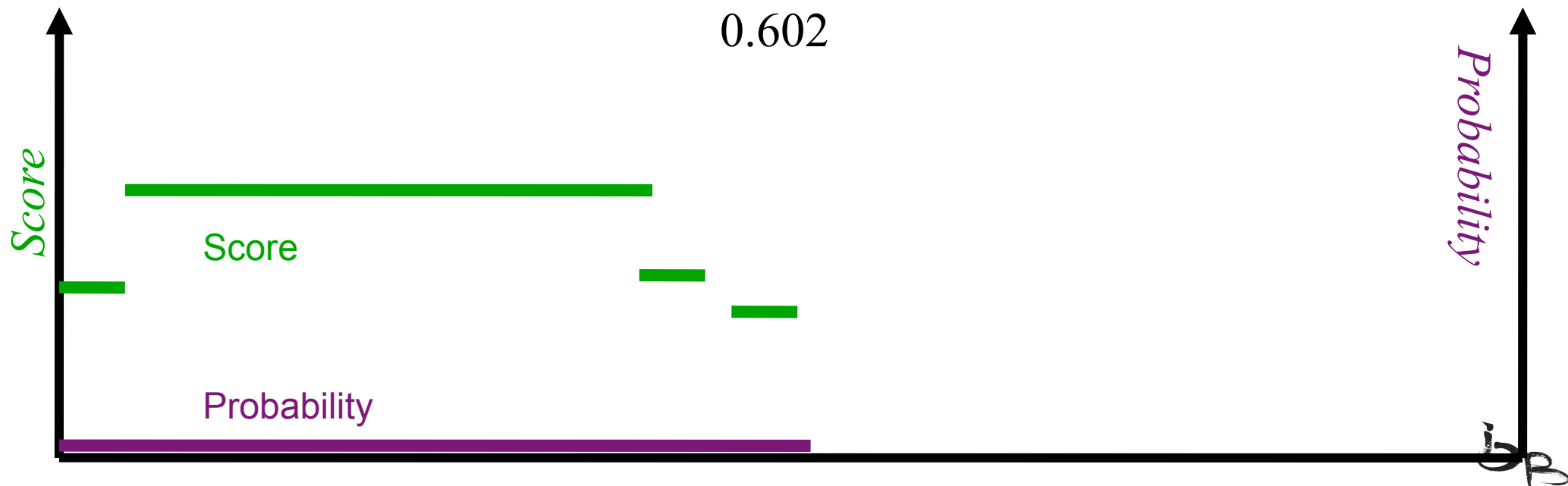


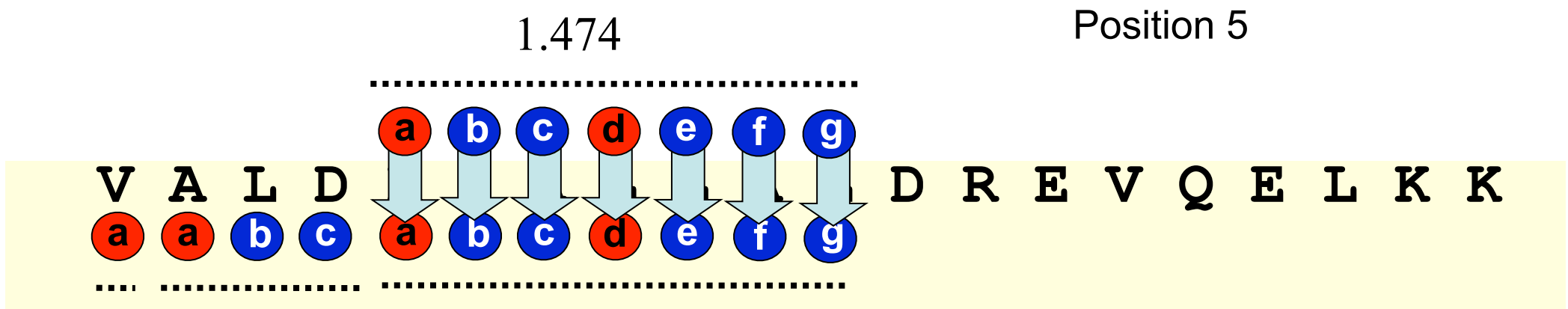
0.780

1.335

0.824

0.602

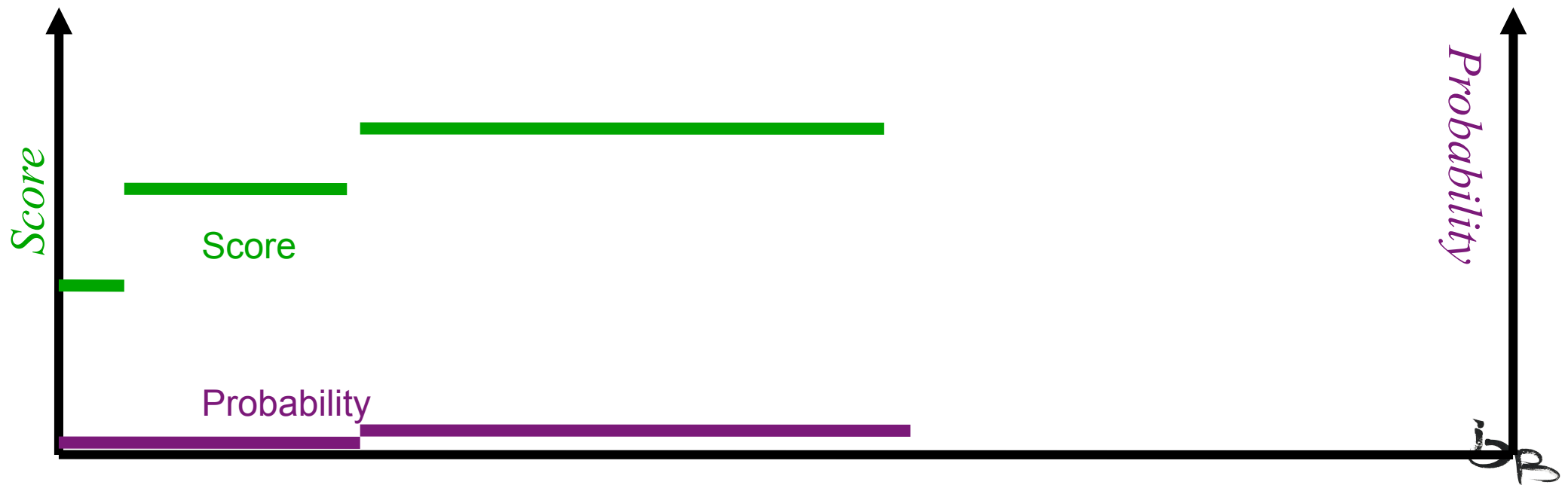




0.780

1.335

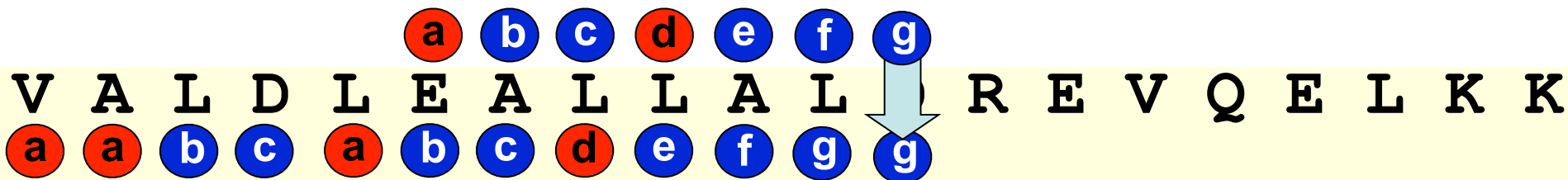
1.474





0.790

Position 6

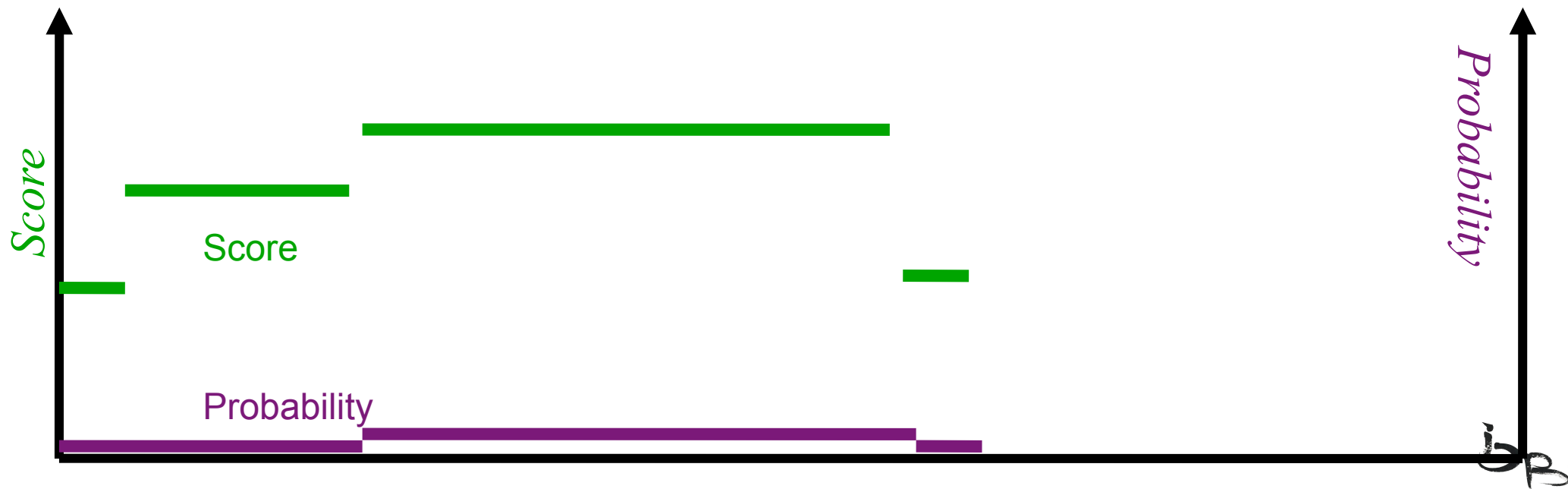


0.780

1.335

1.474

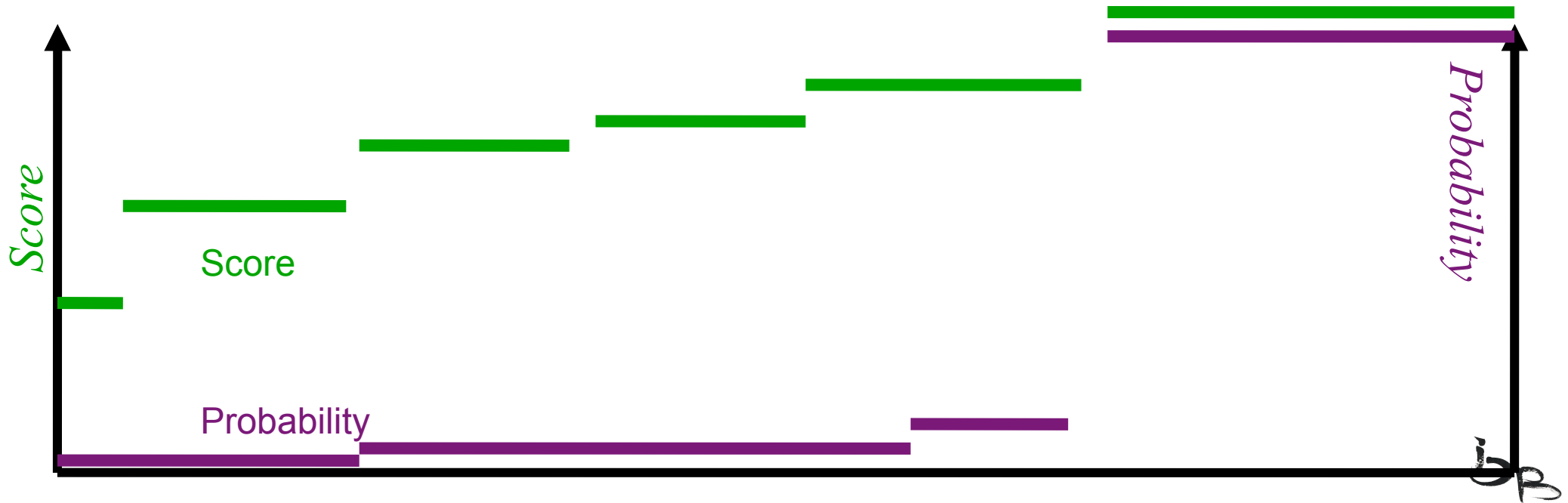
0.790





V A L D L E A L L A L D R E V Q E L K K  
a a b c a b c a b c a b c d a b c d e f

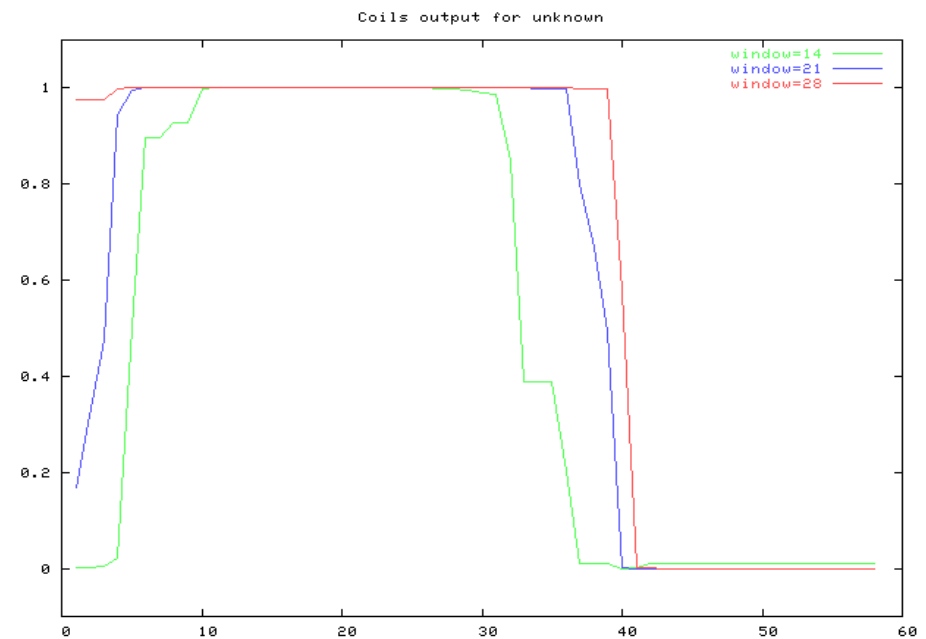
0.780 1.335 1.474 1.499 1.542 2.216





# Coils Algorithm: real results for this sequence

1	V	a	0.780	0.000	(	0.004	2.326)
2	A	a	1.335	0.015	(	0.604	2.036)
3	L	b	1.335	0.015	(	0.604	2.036)
4	D	c	1.335	0.015	(	0.604	2.036)
5	L	a	1.474	0.059	(	1.274	1.018)
6	E	b	1.474	0.059	(	1.274	1.018)
7	A	c	1.474	0.059	(	1.274	1.018)
8	L	a	1.499	0.075	(	1.424	0.874)
9	L	b	1.499	0.075	(	1.424	0.874)
10	A	c	1.499	0.075	(	1.424	0.874)
11	L	a	1.542	0.114	(	1.699	0.659)
12	D	b	1.542	0.114	(	1.699	0.659)
13	R	c	1.542	0.114	(	1.699	0.659)
14	E	d	1.542	0.114	(	1.699	0.659)
15	V	a	2.216	0.997	(	1.845	0.000)
16	Q	b	2.216	0.997	(	1.845	0.000)
17	E	c	2.216	0.997	(	1.845	0.000)
18	L	d	2.216	0.997	(	1.845	0.000)
19	K	e	2.216	0.997	(	1.845	0.000)
20	K	f	2.216	0.997	(	1.845	0.000)
21	R	g	2.216	0.997	(	1.845	0.000)



Coils Web Server  
(window: 14,21,28)

[http://www.ch.embnet.org/software/COILS\\_form.html](http://www.ch.embnet.org/software/COILS_form.html)

Ncoils Program (window = 7)

`ncoils -win 7 < test2.fasta`

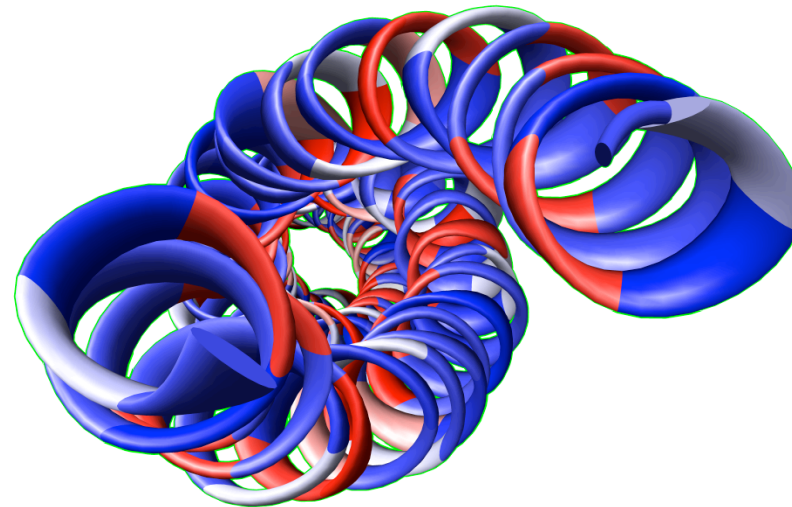
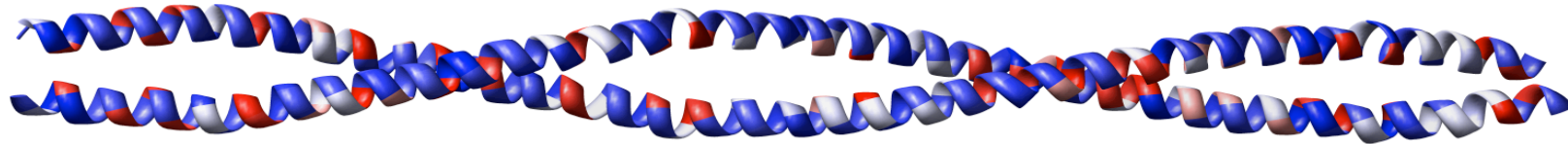
Window = 7 only for example purposes!





# Myosin – light chain

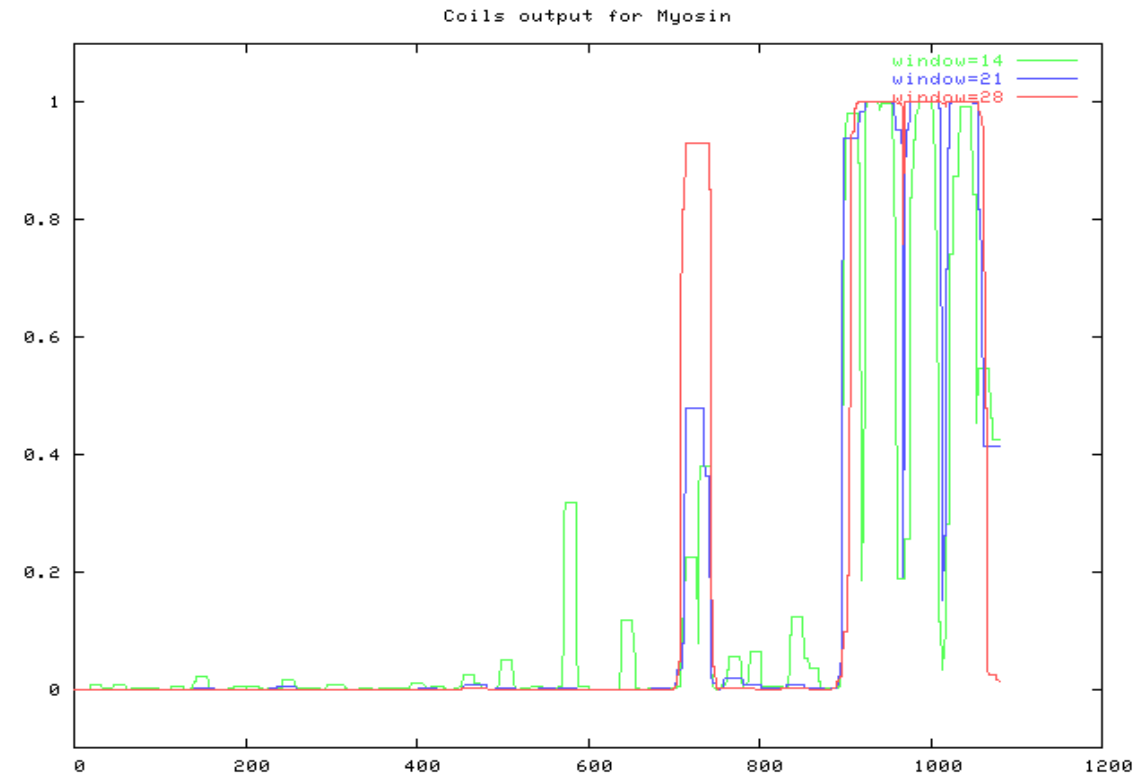
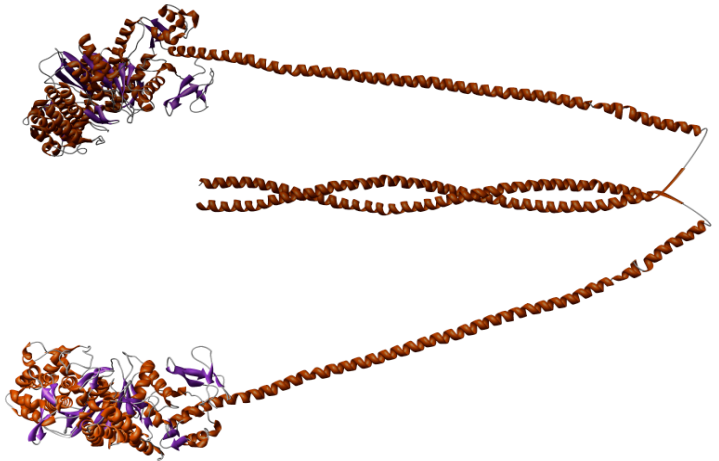
Position in the sequence: residues 953 - 1080







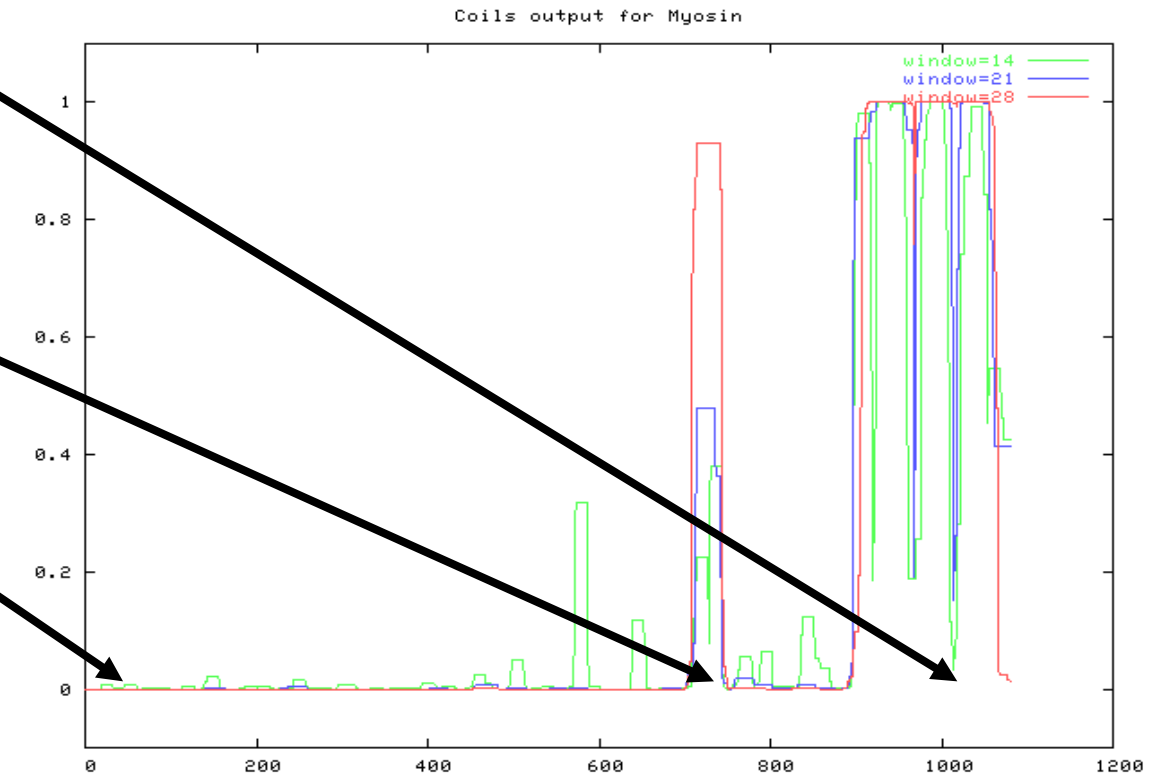
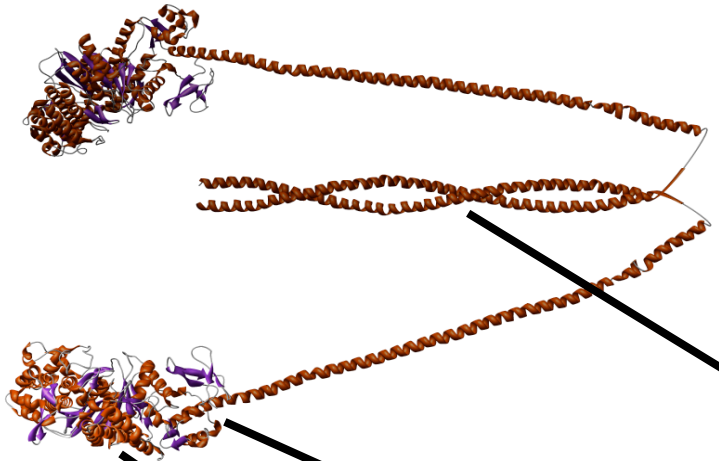
# Coils Algorithm: Myosin



UP

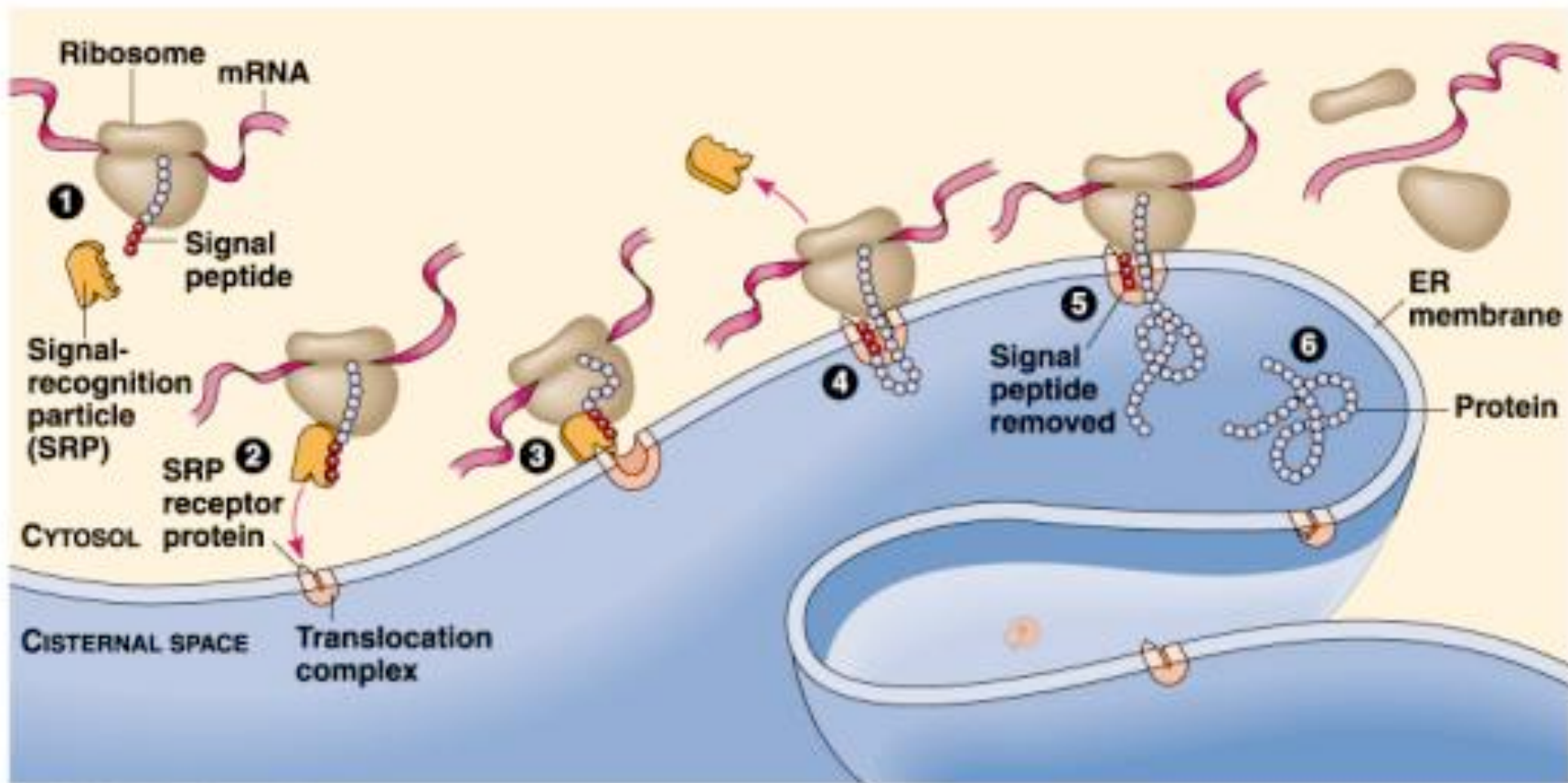


# Coils Algorithm: Myosin



UP

# Signal peptides

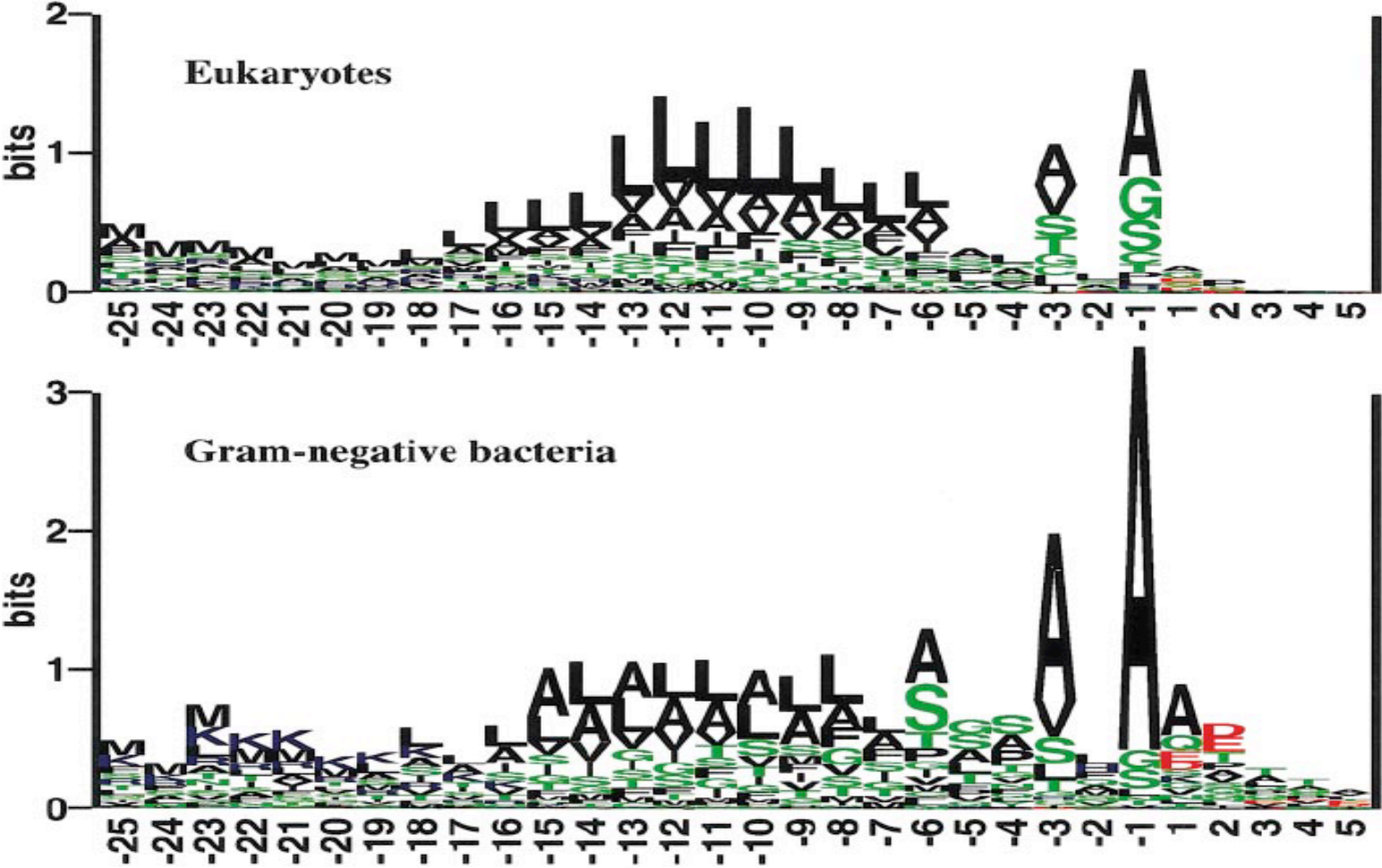




# Signal peptides - biological backgrounds

Signal peptides control the entry of virtually all proteins to the secretory pathway, both in eukaryotes and prokaryotes. They comprise the N-terminal part of the amino acid chain and are cleaved off while the protein is translocated through the membrane. The common structure of signal peptides from various proteins is commonly described as a positively charged n-region, followed by a hydrophobic h-region and a neutral but polar c-region. The  $(-3,-1)$  rule states that the residues at positions -3 and -1 (relative to the cleavage site) must be small and neutral for cleavage to occur correctly.

# Signal peptides - sequence logos

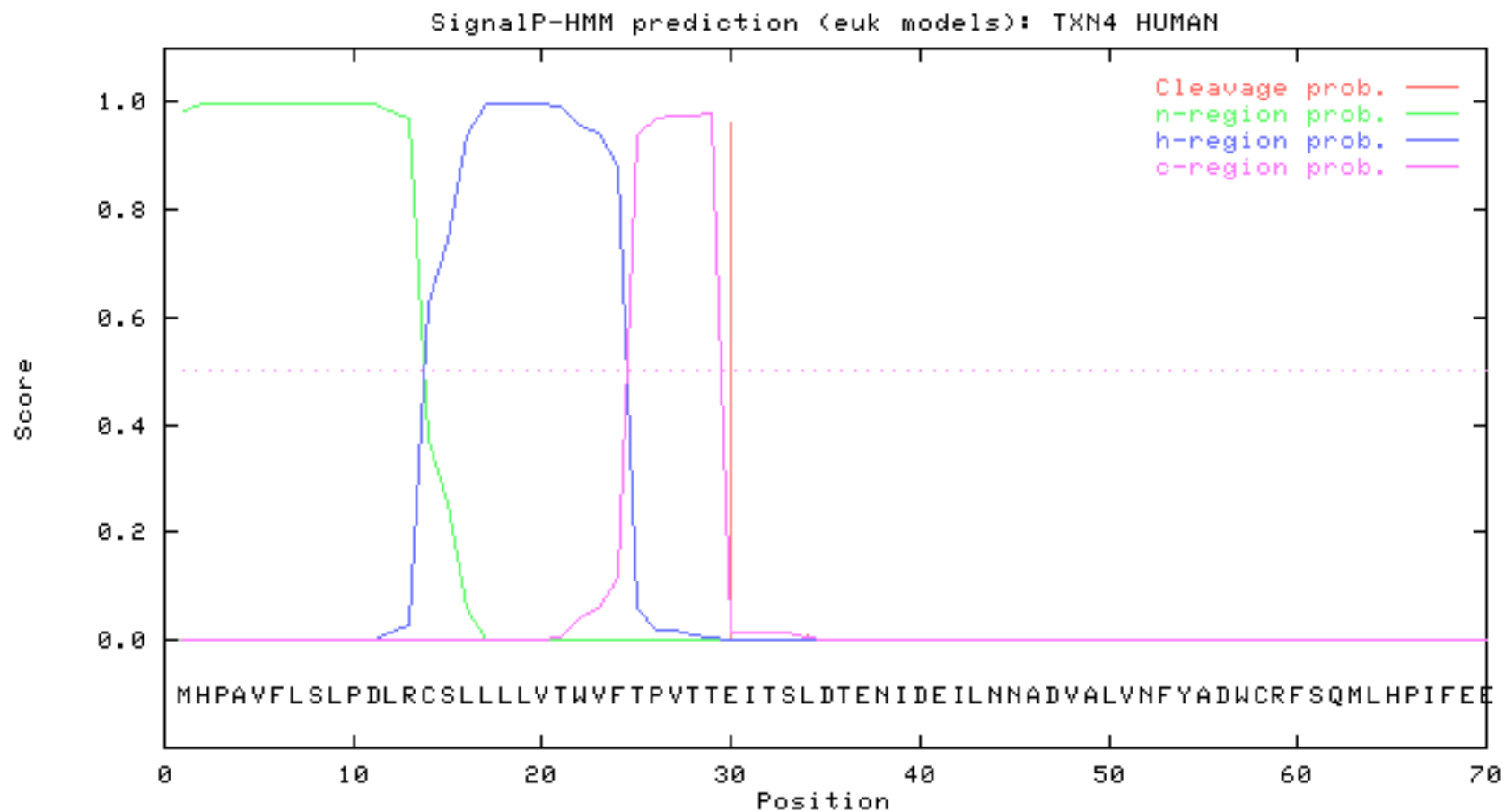




# SignalP

- predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms:
  - Gram-positive prokaryotes
  - Gram-negative prokaryotes
  - eukaryotes
- World Wide Web Prediction Server at Center for Biological Sequence Analysis:
  - <http://www.cbs.dtu.dk/services/SignalP-2.0/>
- prediction is based on a combination of several artificial neural networks and hidden Markov models.

# SignalP



>TXN4\_HUMAN

Prediction: Signal peptide

Signal peptide probability: 0.984

Signal anchor probability: 0.015

Max cleavage site probability: 0.962 between pos. 29 and 30

# BIOINFORMATICS CREDO

- Remember about biology
- Do not trust the data
- Use comparative approach
- Use statistics
- Know the limits
- Remember about biology!!!

