# Blast heuristics, Psi-Blast, and Sequence profiles

Morten Nielsen
Department of Health Technology,
DTU

# Outline



- Basic Local Alignment Search Tool
  - What are the Blast heuristics?
  - How does Blast calculate E-values?
  - What are the limits of Blast?
- Understand why BLAST often fails for low sequence similarity
- Psi-Blast
  - Why does it work so much better
  - See the beauty of sequence profiles
    - Position specific scoring matrices (PSSMs)
  - Use BLAST to generate sequence profiles
  - Use profiles to identify amino acids essential for protein function and structure

# Why alignment is slow

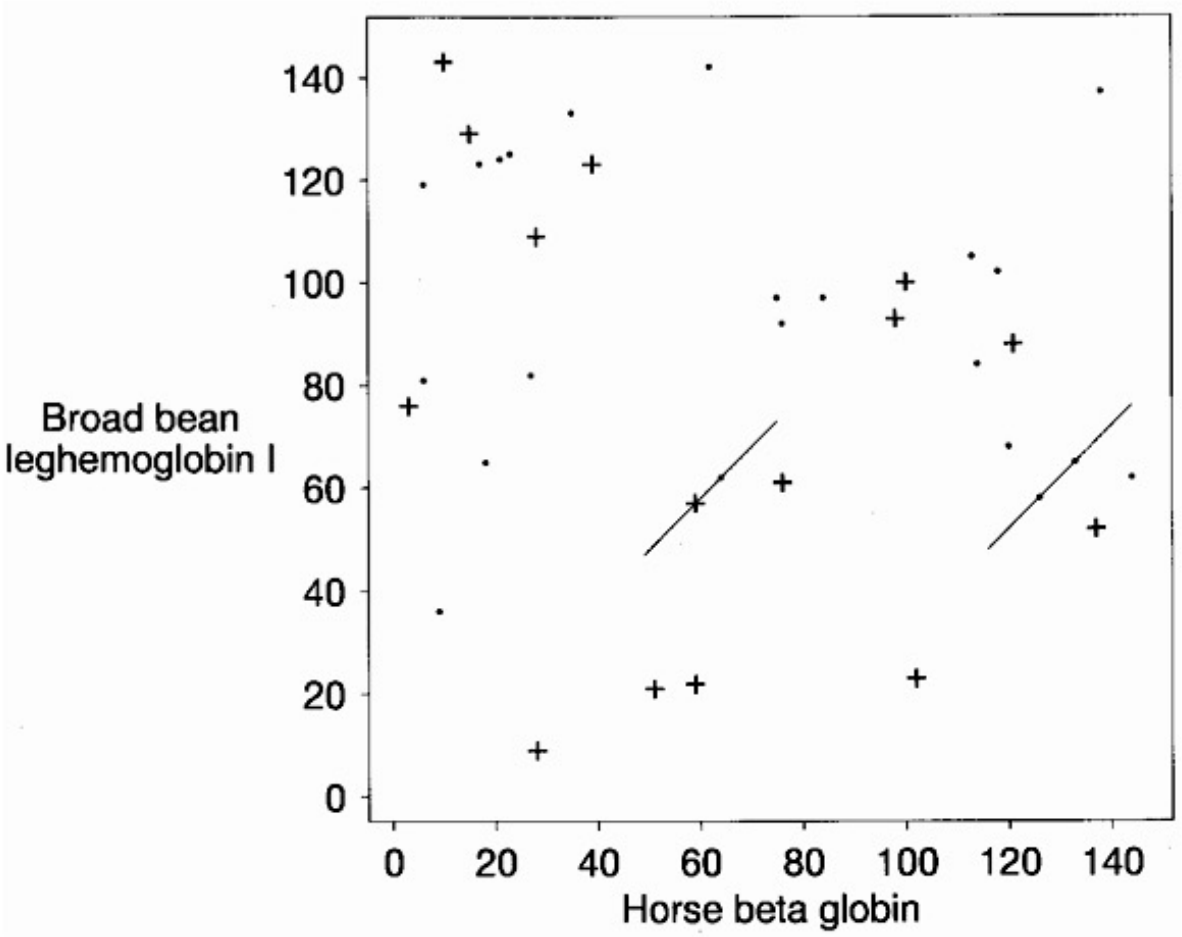


- 99% of the cpu time is spend aligning no-similar sequences
- The execution time for gapped alignments is 500 times that for un-gapped

# Blast heuristics

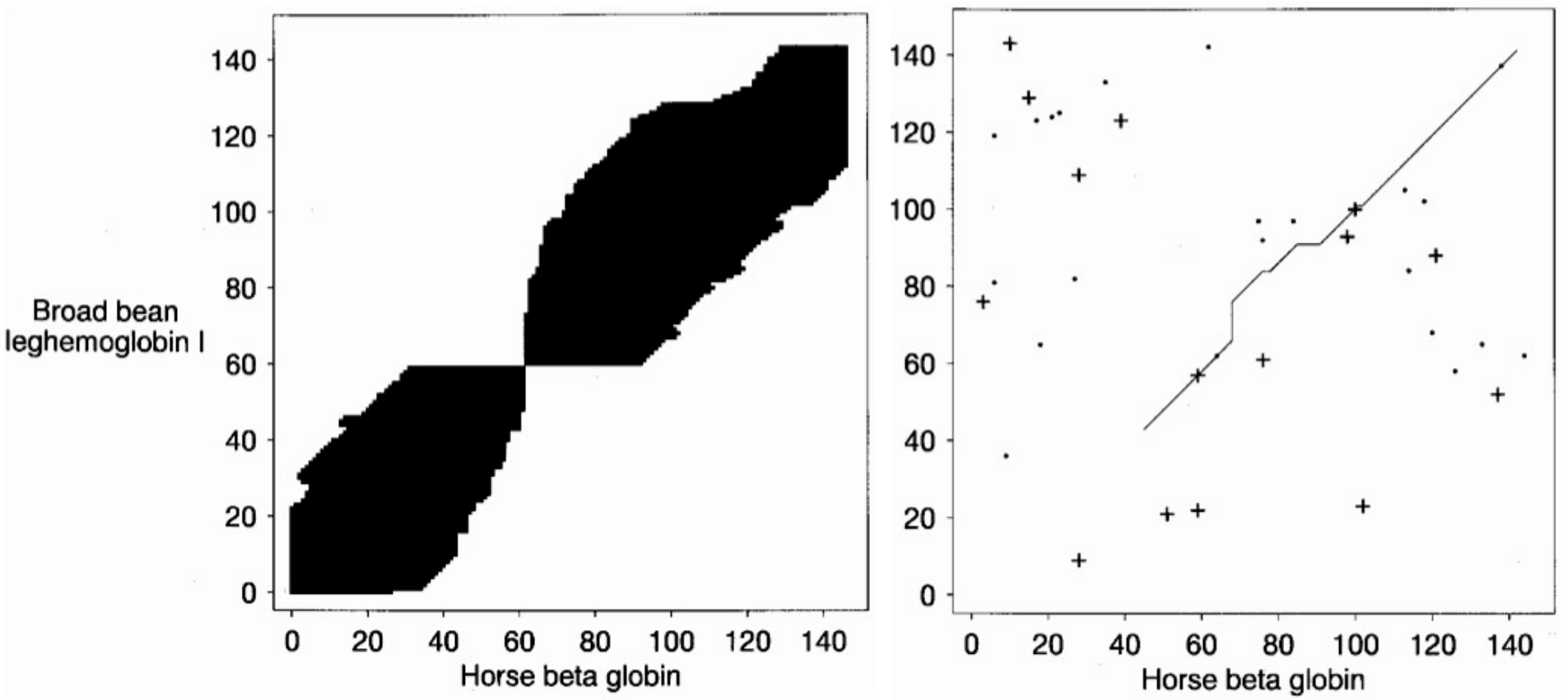


- Hits (High scoring segment pairs, HSP)
  - Triplets of amino acids that scores at least T



# Hit extension





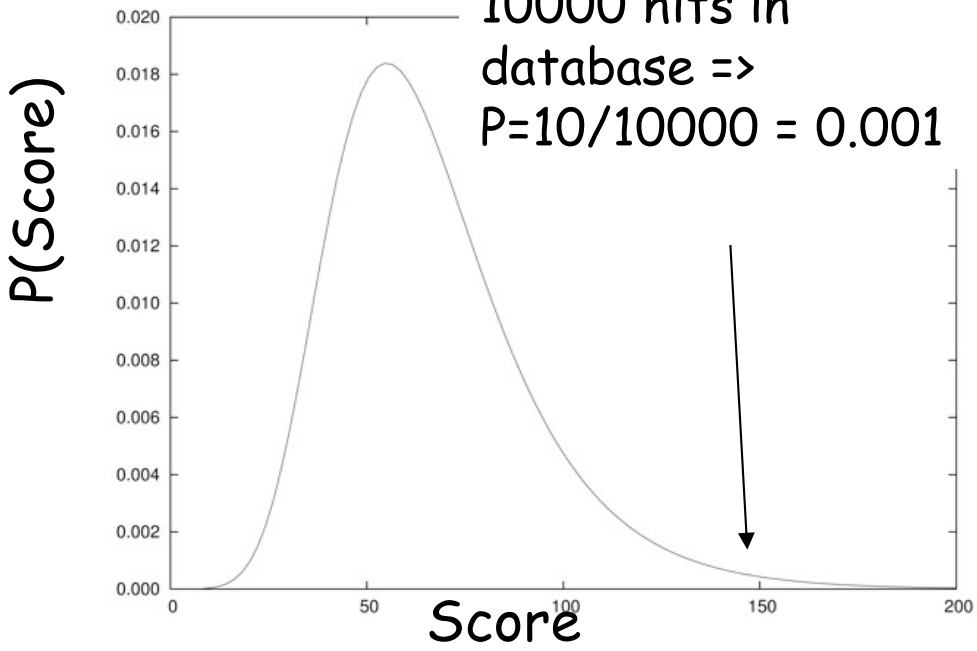

```
Leghemoglobin   43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------
                   F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin     45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE


Leghemoglobin   91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                   +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin     91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

# What are P and E values?



- E-value
  - Number of expected hits in database with score higher than match
  - Depends on database size
- P-value
  - Probability that a random hit will have score higher than match
  - Database size independent




Score 150
10 hits with higher score (E=10)
10000 hits in database =>
P=10/10000 = 0.001

# Blast



- Only align subset of sequences
- Only do gap extension at few seed sites
- Only extend gaps close to diagonal
- Approximate (and conservative) E-value estimates
- Details on the Blast algorithm
  - www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

# What goes wrong when Blast fails?



- Conventional sequence alignment uses a (Blosum) scoring matrix to identify amino acid matches in the two protein sequences

# Blosum scoring matrix

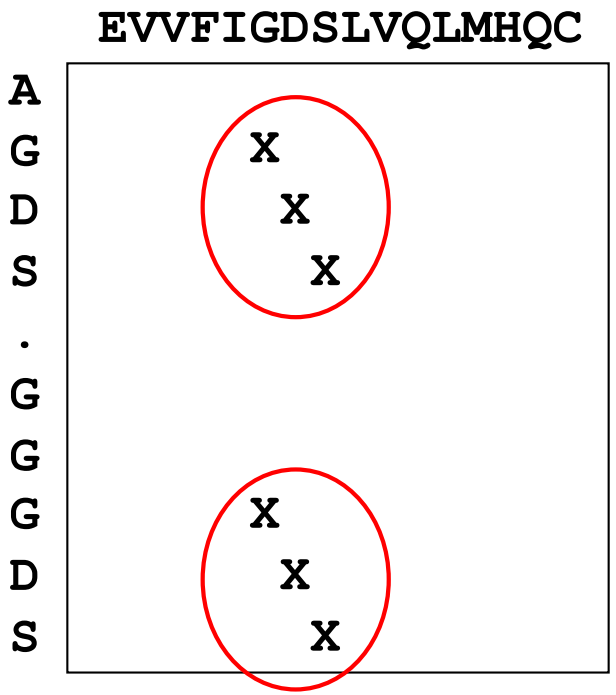
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

# What goes wrong when Blast fails?



- Conventional sequence alignment uses a (Blosum) scoring matrix to identify amino acids matches in the two protein sequences
- This scoring matrix is identical at all positions in the protein sequence!



# Alignment



- Blosum62 score matrix. Fg=1. Ng=0?

| | L | A | G | D | S | D |
|---|---|---|---|---|---|---|
| F | 0 | -2 | -3 | -3 | -2 | -3 |
| I | 2 | 1 | -4 | -3 | -2 | -3 |
| G | -4 | 0 | 6 | -1 | 0 | -1 |
| D | -4 | -2 | -1 | 6 | 0 | 6 |
| S | -2 | 1 | 0 | 0 | 4 | 0 |
| L | 4 | -1 | -4 | -4 | -2 | -4 |

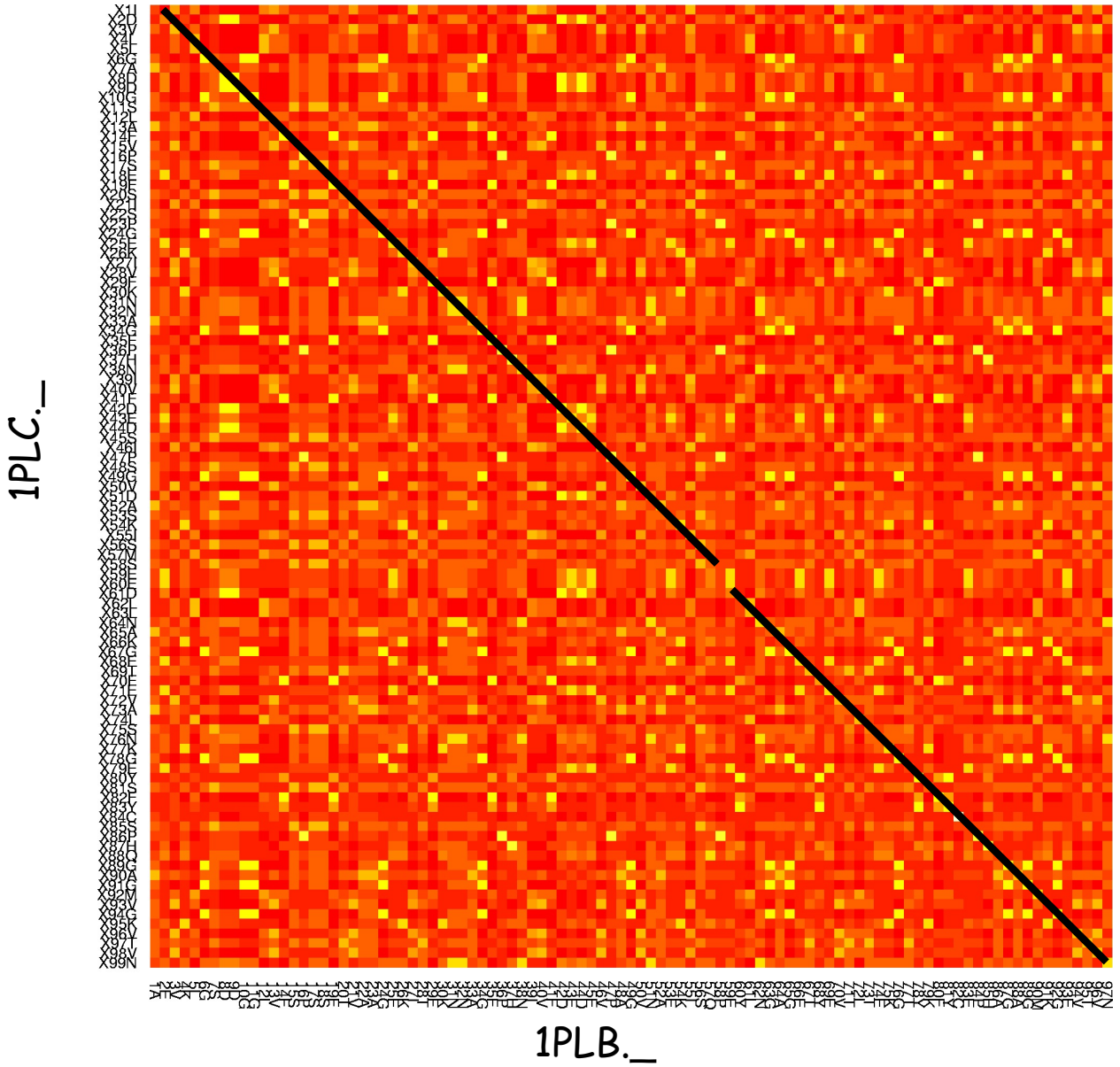
- Score =2+6+6+4-1=17
- Alignment

**LAGDS**

**I-GDS**

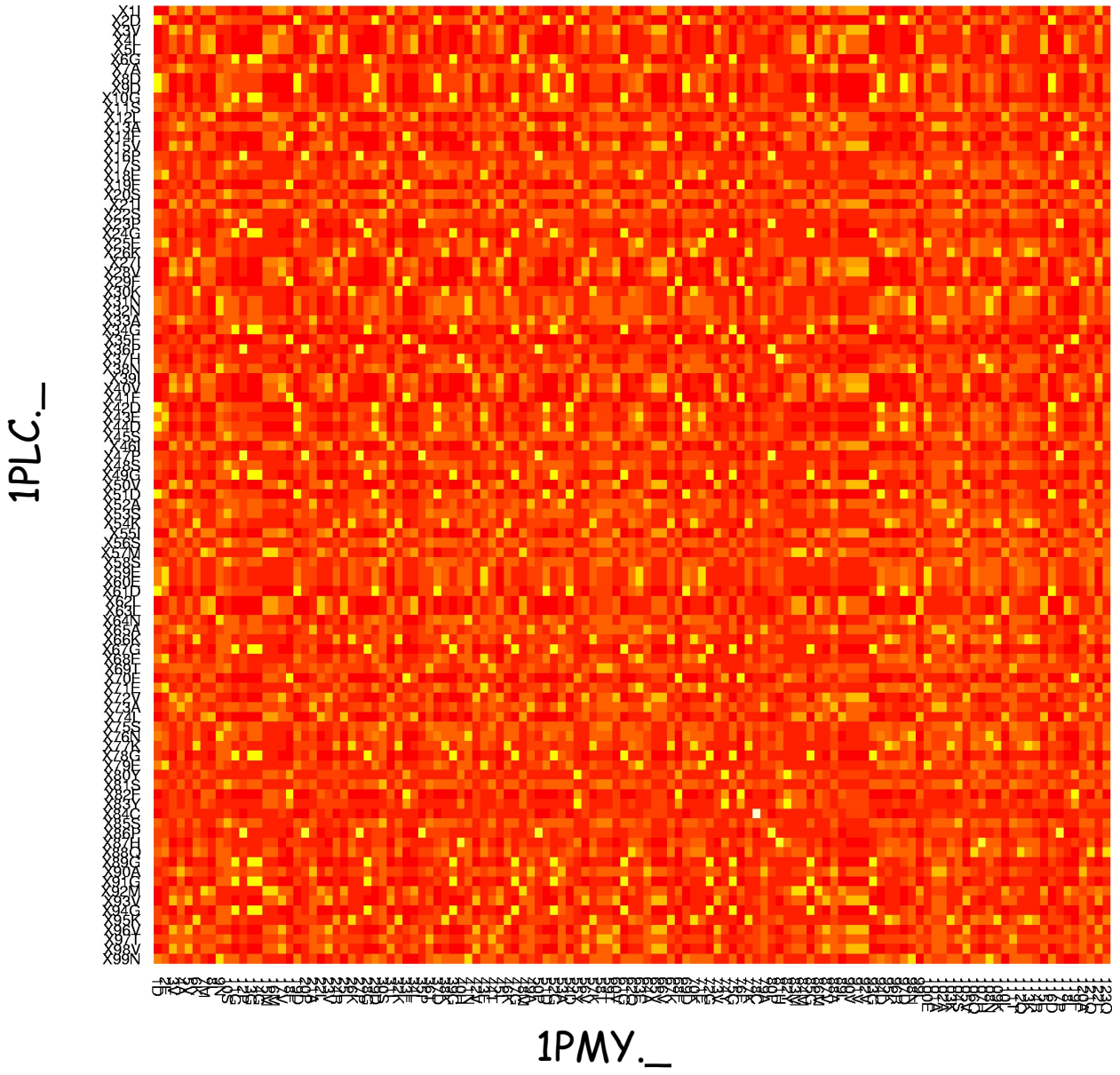
# When Blast works!





# When Blast fails!

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



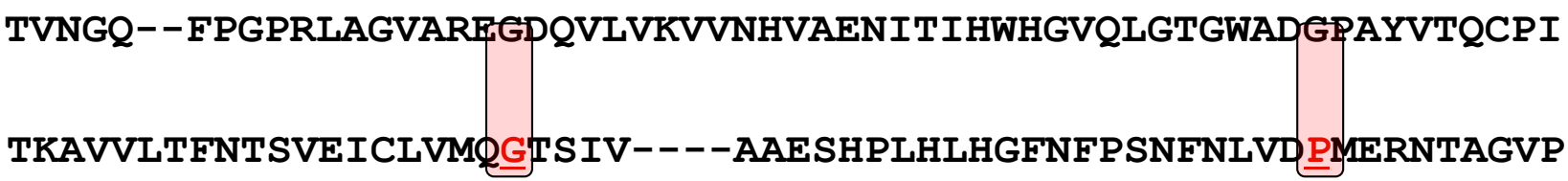
# Sequence profiles



- In reality not all positions in a protein are equally likely to mutate
  - Some amino acids (active cites) are highly conserved, and the score for mismatch must be very high
  - Other amino acids can mutate almost for free, and the score for mismatch should be lower than the BLOSUM score
- Sequence profiles can capture these differences

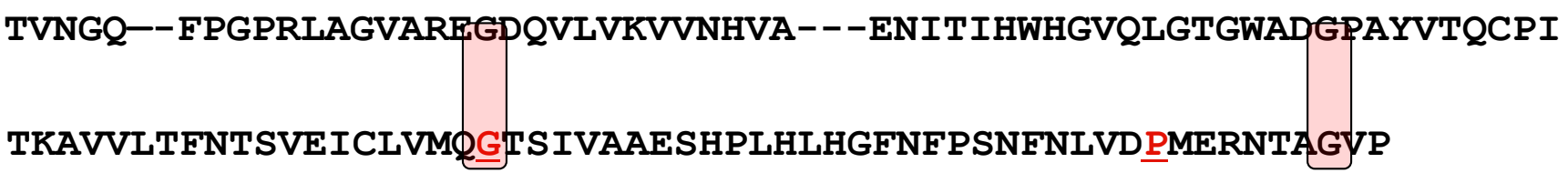
# Sequence profiles





# Sequence profiles - OR





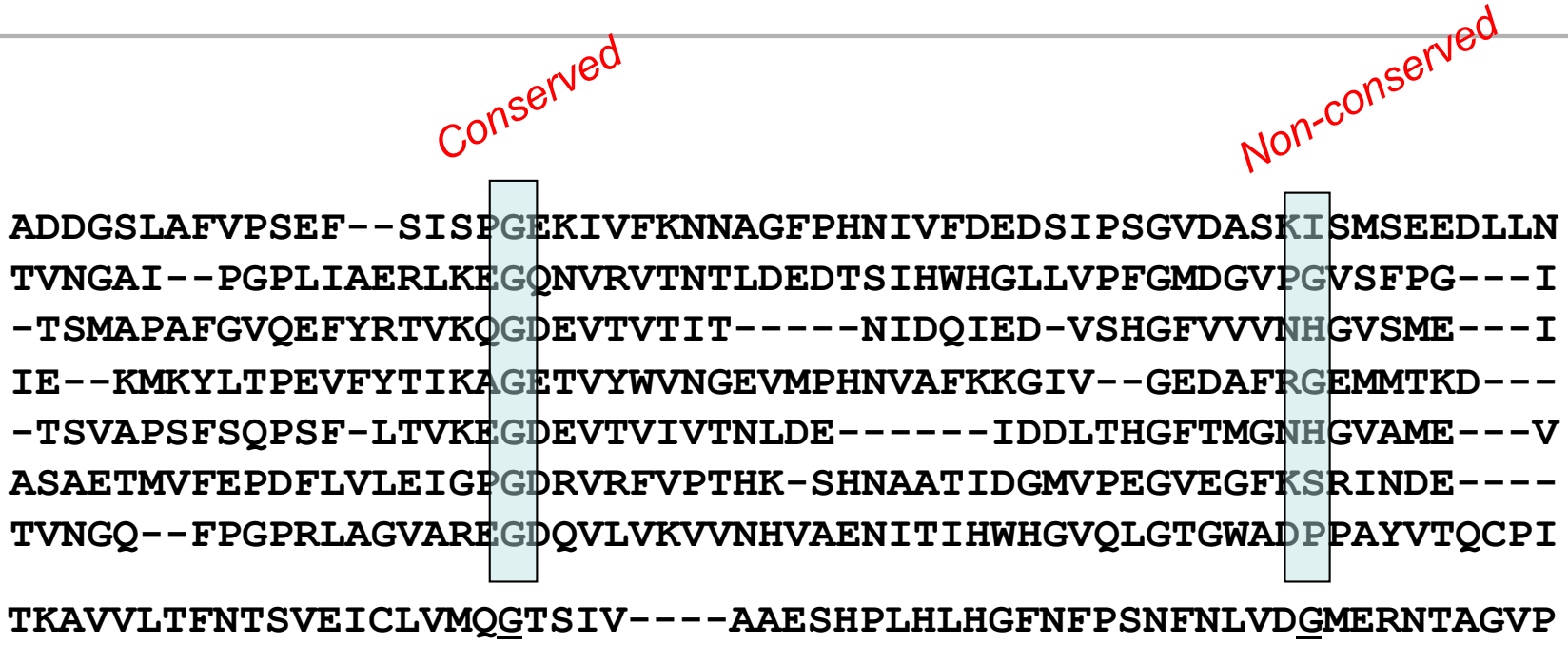
# Sequence profiles



**TVNGQ--FPGPRLAGVAREGDQVLVKVVNHVAENITIHWHGVQLGTGWADPPAYVTQCPI**

# Sequence profiles






Matching any thing but G => large negative score

Any thing can match

$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta}$$

$$g_b = \sum_a f_a \cdot q_{b|a}$$

# Visualization of Sequence logos



$$I = \sum_a p_a \log(\frac{p_a}{q_a})$$

$P_A = 6/10 = 0.6$
$P_G = 2/10 = 0.2$
$P_T = P_K = 1/10 = 0.1$
$P_C = P_D = \ldots P_V = 0.0$

$q_A = 0.07$
$q_G = 0.07$
$q_T = 0.05$
$q_K = 0.06$

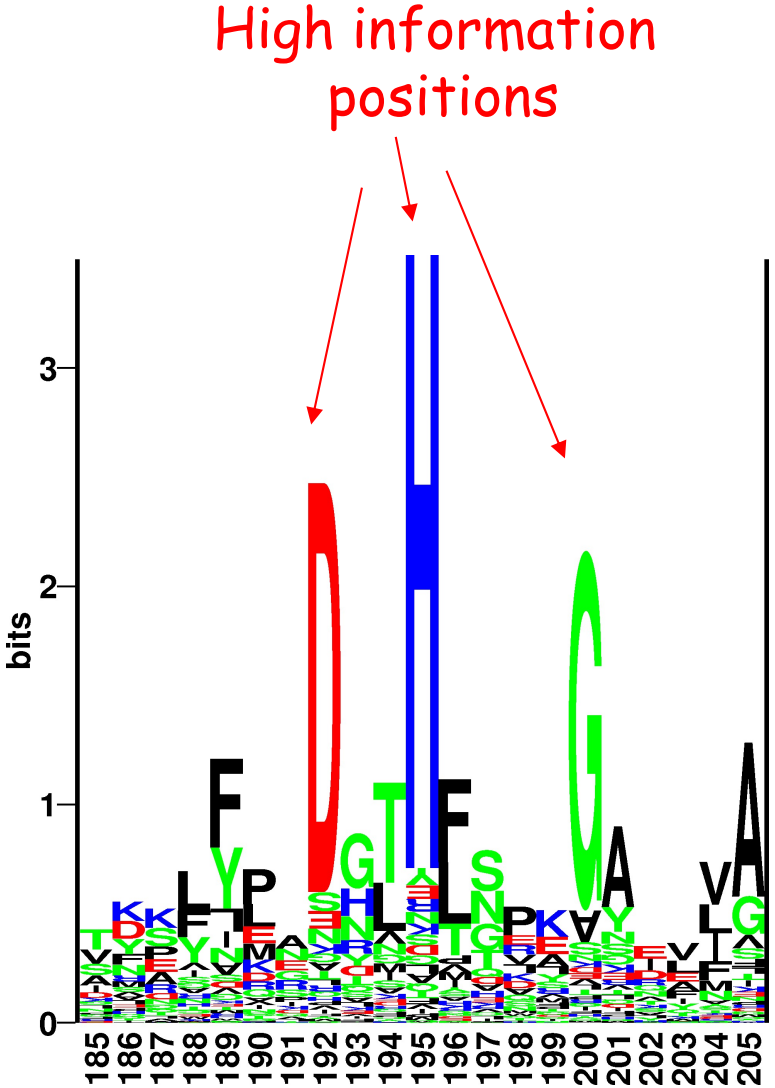
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

# Sequence logos (Kullback-Leibler)



$$I = \sum_a p_a \log(\frac{p_a}{q_a})$$

- Height of a column equal to I
- Relative height of a letter is p
- (Letters upside-down if $p_a < q_a$)



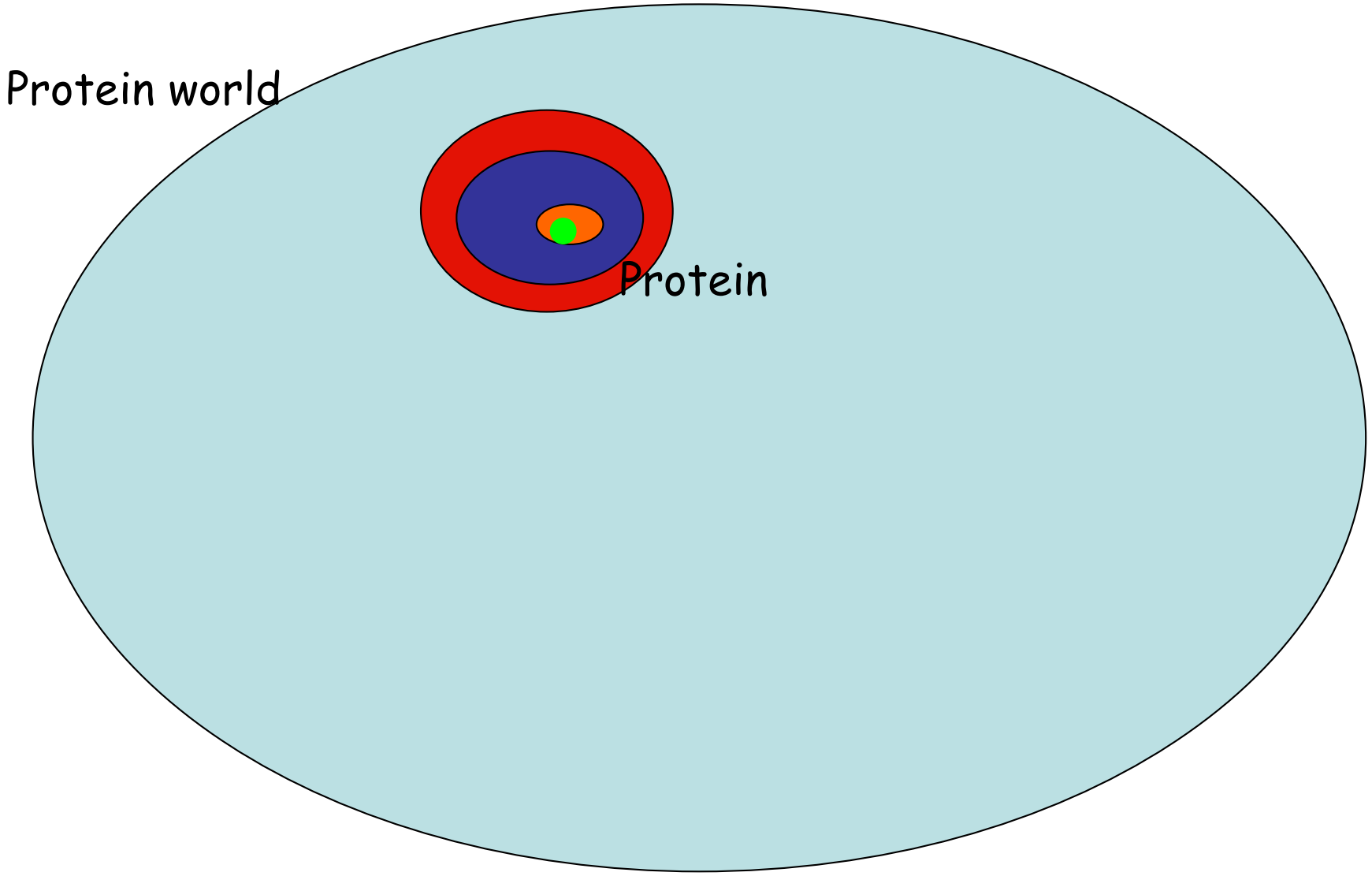
# How to make sequence profiles



1. Align (BLAST) sequence against large sequence database (Swiss-Prot)
2. Select significant alignments and make sequence profile
3. Use profile to align against sequence database to find new significant hits
4. Repeat 2 and 3 (normally 3 times!)

# Blast iterations





# How to make sequence profiles

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

## The blast command

## blastpgp -d db -e 0.00001 -j 4 -Q blastprofile -i fastafile -o out

Last position-specific scoring matrix computed

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | T | -1 | -3 | -2 | -3 | -3 | -2 | -2 | -3 | -3 | -3 | -4 | -2 | -3 | -4 | -3 | 5 | 6 | -5 | -4 | -3 |
| 2 | T | -3 | 1 | -1 | -4 | -4 | -2 | -3 | -4 | -4 | -4 | -4 | 5 | -4 | -5 | -4 | -2 | 6 | -5 | -4 | -3 |
| 3 | V | -3 | -4 | -6 | -6 | -4 | -5 | -6 | -5 | -6 | 6 | 2 | -5 | -2 | -2 | -6 | -5 | -3 | 1 | -4 | 4 |
| 4 | Y | 0 | -5 | -6 | -6 | 1 | -3 | -5 | -2 | -1 | 0 | 2 | -5 | -2 | 4 | -6 | -4 | -4 | 0 | 4 | 4 |
| 5 | L | -2 | -5 | -6 | -6 | 3 | -5 | -6 | -6 | -5 | 3 | 4 | -5 | 1 | 5 | -6 | -5 | -1 | -1 | -1 | 1 |
| 6 | A | 3 | -5 | -5 | -6 | 0 | -2 | -5 | -3 | -5 | 3 | 1 | -5 | 1 | 4 | -5 | -4 | -1 | -1 | -2 | 2 |
| 7 | G | -3 | -5 | -3 | -4 | -6 | -5 | -5 | 7 | -5 | -7 | -7 | -5 | -6 | -6 | -5 | -2 | -4 | -6 | -6 | -6 |
| 8 | D | -3 | -4 | 1 | 8 | -6 | -3 | -1 | -4 | -4 | -6 | -6 | -4 | -6 | -6 | -4 | -3 | -3 | -7 | -6 | -6 |
| 9 | S | -2 | -4 | -1 | -3 | -4 | -3 | -3 | -3 | -4 | -5 | -6 | -3 | -5 | -6 | -4 | 7 | -2 | -6 | -5 | -5 |
| 10 | T | -2 | -3 | -4 | -5 | -4 | -3 | -5 | -5 | -5 | 6 | 1 | -4 | -2 | -4 | -5 | -3 | 5 | -5 | -4 | 0 |
| 11 | M | 0 | -4 | -3 | -4 | -4 | -4 | -3 | -4 | -5 | -1 | -2 | -4 | 3 | -2 | -2 | 1 | 6 | -5 | -4 | 2 |
| 12 | A | 4 | -1 | 0 | 0 | 1 | 2 | 0 | -2 | 2 | -4 | -4 | -1 | -3 | -1 | -3 | 0 | -1 | 1 | -1 | -2 |

# Sequence profiles for a single sequence

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

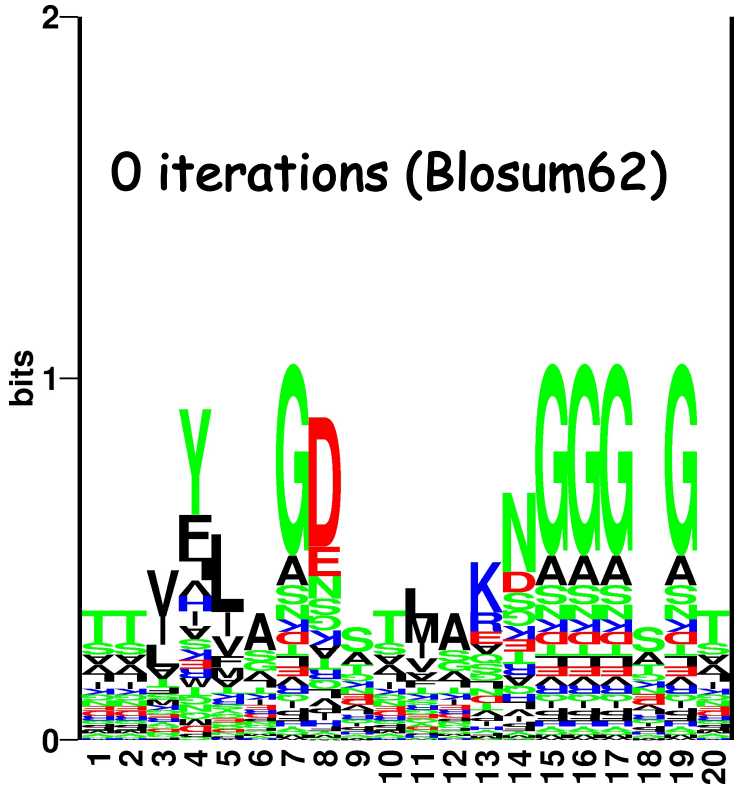
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.29 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.04 | 0.08 | 0.01 | 0.04 | 0.06 | 0.04 | 0.02 | 0.02 | 0.03 | 0.09 | 0.05 | 0.01 | 0.02 | 0.07 |
| R | 0.04 | 0.34 | 0.04 | 0.03 | 0.01 | 0.05 | 0.05 | 0.03 | 0.02 | 0.02 | 0.05 | 0.12 | 0.02 | 0.02 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 |
| N | 0.04 | 0.04 | 0.32 | 0.08 | 0.01 | 0.03 | 0.05 | 0.07 | 0.03 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.02 | 0.07 | 0.05 | 0.00 | 0.02 | 0.03 |
| D | 0.04 | 0.03 | 0.07 | 0.40 | 0.01 | 0.03 | 0.09 | 0.05 | 0.02 | 0.02 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.05 | 0.04 | 0.00 | 0.01 | 0.02 |
| C | 0.07 | 0.02 | 0.02 | 0.02 | 0.48 | 0.01 | 0.02 | 0.03 | 0.01 | 0.04 | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.04 | 0.00 | 0.01 | 0.06 |
| Q | 0.06 | 0.07 | 0.04 | 0.05 | 0.01 | 0.21 | 0.10 | 0.04 | 0.03 | 0.03 | 0.05 | 0.09 | 0.02 | 0.01 | 0.02 | 0.06 | 0.04 | 0.01 | 0.02 | 0.04 |
| E | 0.06 | 0.05 | 0.04 | 0.09 | 0.01 | 0.06 | 0.30 | 0.04 | 0.03 | 0.02 | 0.04 | 0.08 | 0.01 | 0.02 | 0.03 | 0.06 | 0.04 | 0.01 | 0.02 | 0.03 |
| G | 0.08 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 | 0.51 | 0.01 | 0.02 | 0.03 | 0.03 | 0.01 | 0.02 | 0.02 | 0.05 | 0.03 | 0.01 | 0.01 | 0.02 |
| H | 0.04 | 0.05 | 0.05 | 0.04 | 0.01 | 0.04 | 0.05 | 0.04 | 0.35 | 0.02 | 0.04 | 0.05 | 0.02 | 0.03 | 0.02 | 0.04 | 0.03 | 0.01 | 0.06 | 0.02 |
| I | 0.05 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.27 | 0.17 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.01 | 0.02 | 0.18 |
| L | 0.04 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.12 | 0.38 | 0.03 | 0.05 | 0.05 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.10 |
| K | 0.06 | 0.11 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.04 | 0.02 | 0.03 | 0.04 | 0.28 | 0.02 | 0.02 | 0.03 | 0.05 | 0.04 | 0.01 | 0.02 | 0.03 |
| M | 0.05 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 | 0.10 | 0.20 | 0.04 | 0.16 | 0.05 | 0.02 | 0.04 | 0.04 | 0.01 | 0.02 | 0.09 |
| F | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.03 | 0.02 | 0.06 | 0.11 | 0.02 | 0.03 | 0.39 | 0.01 | 0.03 | 0.03 | 0.02 | 0.0 | 0.06 |
| P | 0.06 | 0.03 | 0.02 | 0.03 | 0.01 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.04 | 0.01 | 0.01 | 0.49 | 0.04 | 0.04 | 0.00 | 0.01 | 0.03 |
| S | 0.11 | 0.04 | 0.05 | 0.05 | 0.02 | 0.03 | 0.05 | 0.07 | 0.02 | 0.03 | 0.04 | 0.05 | 0.02 | 0.02 | 0.03 | 0.22 | 0.08 | 0.0 | 0.02 | 0.04 |
| T | 0.07 | 0.04 | 0.04 | 0.04 | 0.02 | 0.03 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.05 | 0.02 | 0.02 | 0.03 | 0.09 | 0.25 | 0.01 | 0.02 | 0.07 |
| W | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.03 | 0.02 | 0.03 | 0.05 | 0.02 | 0.02 | 0.06 | 0.01 | 0.02 | 0.02 | 0.49 | 0.07 | 0.03 |
| Y | 0.04 | 0.03 | 0.02 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.05 | 0.04 | 0.07 | 0.03 | 0.02 | 0.13 | 0.02 | 0.03 | 0.03 | 0.03 | 0.32 | 0.05 |
| V | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.16 | 0.13 | 0.03 | 0.03 | 0.04 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.27 |

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

**TKAVVLTFNTSVEICLVMQGTSIV----AAESHPLHLHGFNFPSNFNLVDPMERNTAGVP**

# Sequence profiles (1K7C.A)

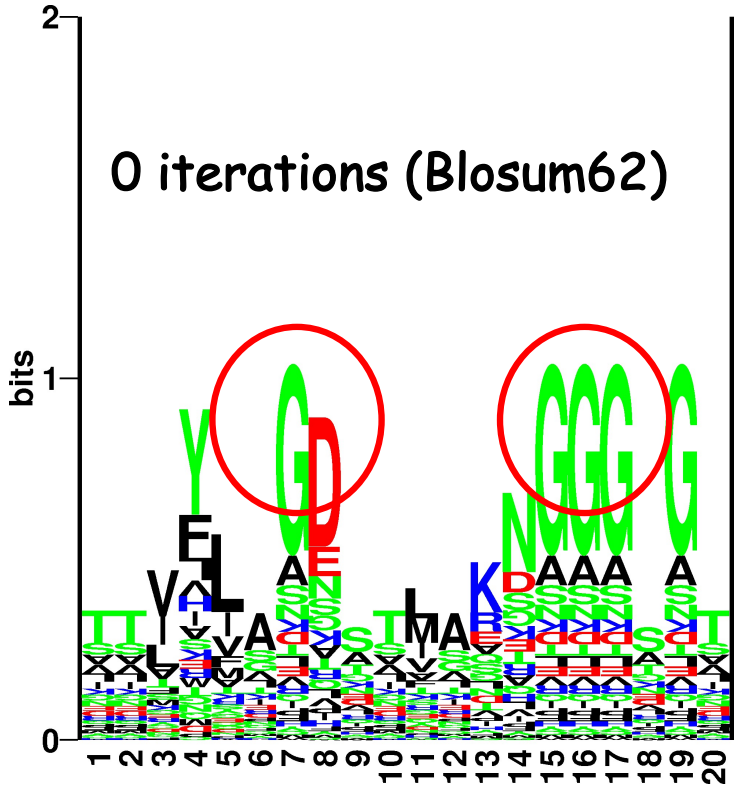





| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

# Sequence profiles (1K7C.A)

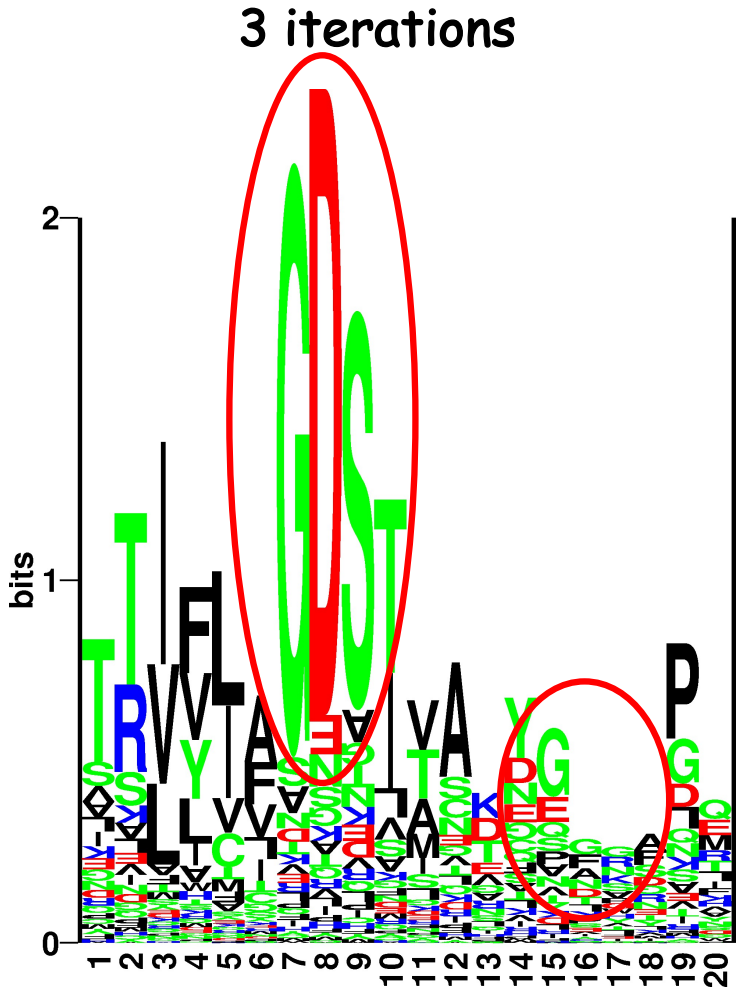





| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

# Sequence profiles (1K7C.A)






## Sequence profile

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | -1 | -2 | -1 | -2 | -2 | -2 | -2 | -3 | -3 | -2 | -3 | -2 | -2 | -4 | -2 | 2 | 6 | -4 | -3 | -1 |
| T | -2 | 3 | -2 | -3 | -3 | -2 | -3 | -4 | -3 | -3 | -4 | 2 | -3 | -4 | -3 | -1 | 6 | -5 | -4 | -3 |
| V | -3 | -5 | -6 | -6 | -3 | -5 | -5 | -6 | -5 | 6 | 2 | -5 | -1 | 0 | -5 | -4 | -3 | -5 | -3 | 4 |
| Y | -3 | -5 | -5 | -6 | -4 | -4 | -5 | -3 | 1 | -1 | 0 | -5 | -2 | 6 | -5 | -4 | -4 | 2 | 6 | 2 |
| L | -4 | -5 | -6 | -6 | 1 | -5 | -5 | -6 | -5 | 4 | 5 | -5 | 1 | 1 | -5 | -5 | -2 | 0 | -3 | 1 |
| A | 4 | -4 | -4 | -5 | 2 | -4 | -4 | -2 | -5 | 3 | -1 | -4 | -2 | 1 | -4 | -2 | -1 | -5 | -3 | 2 |
| G | -2 | -5 | -3 | -4 | -5 | -4 | -4 | 7 | -4 | -6 | -6 | -4 | -5 | -5 | -4 | -1 | -4 | -5 | -5 | -5 |
| D | -4 | -4 | 0 | 8 | -6 | -3 | -1 | -4 | -3 | -6 | -6 | -3 | -5 | -6 | -4 | -3 | -3 | -7 | -5 | -6 |
| S | -1 | -3 | -2 | -3 | -3 | -2 | -2 | -3 | -3 | -5 | -5 | -3 | -4 | -5 | -3 | 7 | -1 | -5 | -4 | -4 |
| T | -3 | -4 | -3 | -4 | -3 | -3 | -3 | -4 | -4 | 2 | -1 | -3 | -3 | -4 | -4 | -1 | 7 | -5 | -4 | -2 |
| M | 2 | -4 | -4 | -4 | -3 | -3 | -4 | -4 | -4 | -1 | -2 | -4 | 5 | -3 | -4 | -1 | 3 | -5 | -4 | 4 |
| A | 5 | -2 | -1 | -1 | 3 | 0 | -2 | -1 | -3 | -4 | -4 | -1 | -3 | -4 | -3 | 1 | -2 | -4 | 0 | -3 |
| K | -1 | 2 | 1 | 3 | -4 | 0 | 0 | 0 | -3 | -2 | -4 | 2 | -3 | -5 | -2 | -1 | 2 | -5 | -4 | 1 |
| N | -2 | -3 | 3 | 2 | 2 | 1 | 0 | -1 | -2 | -4 | -3 | -1 | -3 | -1 | -4 | -1 | -1 | -2 | 6 | -4 |
| G | -2 | -3 | -1 | -1 | -4 | 1 | 0 | 4 | -3 | -4 | -3 | -1 | -4 | -4 | 3 | 1 | -3 | -4 | 1 | -3 |
| G | 0 | -2 | 2 | 3 | -3 | 0 | 0 | 3 | 1 | -3 | -2 | 0 | -3 | -2 | -1 | 0 | 0 | -3 | 0 | -3 |
| G | 1 | 1 | 3 | -1 | -3 | -1 | 0 | 1 | 2 | -3 | -2 | 0 | -3 | -3 | -1 | 1 | 0 | 0 | 1 | -2 |
| S | 2 | 2 | 3 | 1 | -3 | -2 | 0 | 1 | 0 | -3 | -3 | -1 | -3 | -3 | -3 | 0 | -1 | -3 | 2 | -3 |
| G | 0 | -3 | -1 | -3 | -3 | -3 | -3 | 2 | -3 | -3 | -1 | -1 | -3 | -4 | 6 | -1 | -2 | -4 | -3 | -1 |
| T | -1 | 1 | -2 | -2 | -3 | 3 | 2 | -2 | -2 | -1 | -1 | 1 | 2 | -1 | 0 | -1 | 1 | 2 | 2 | -2 |

# Example. Are these two sequences alike?

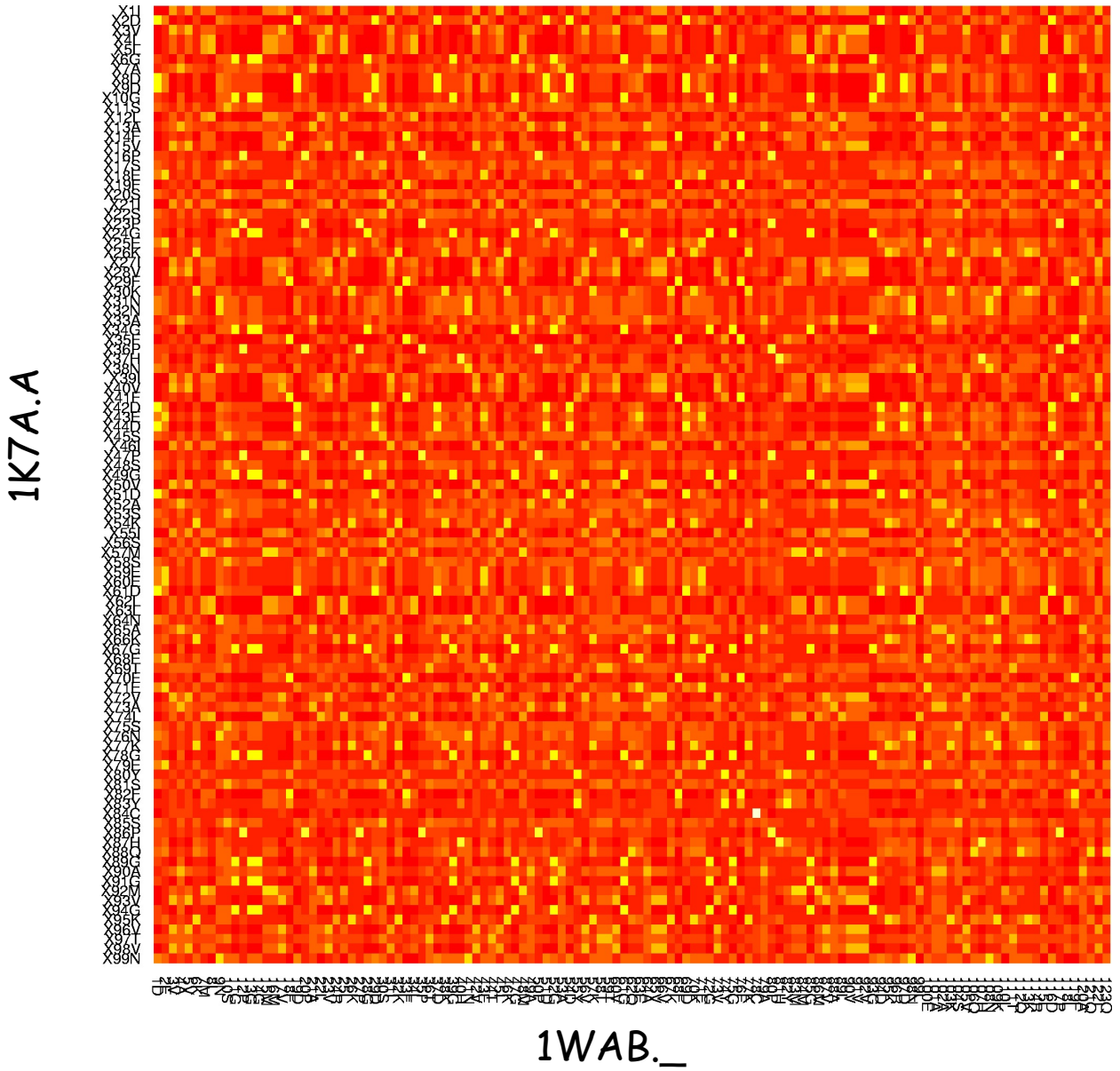
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

>1K7C.A
TTVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADV
VTAGDYVIVEFGHNDGGSLSTDNGRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKL
FTAKGAKVILSSQTPNNPWETGTFVNSPTRFVEYAELAAEVAGVEYVDHWSYVDSIYETL
GNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSLKSVLTTTSFEGTCL

>1WAB.A
ENPASKPTPVQDVQGDGRWMSLHHRFVADSKDKEPEVVFIGDSLVQLMHQCEIWRELFSP
LHALNFGIGGDSTQHVLWRLENGELEHIRPKIVVVWVGTNNHGHTAEQVTGGIKAIVQLV
NERQPQARVVVLGLLPRGQHPNPLREKNRRVNELVRAALAGHPRAHFLDADPGFVHSDGT
ISHHDMYDYLHLSRLGYTPVCRALHSLLLRLL

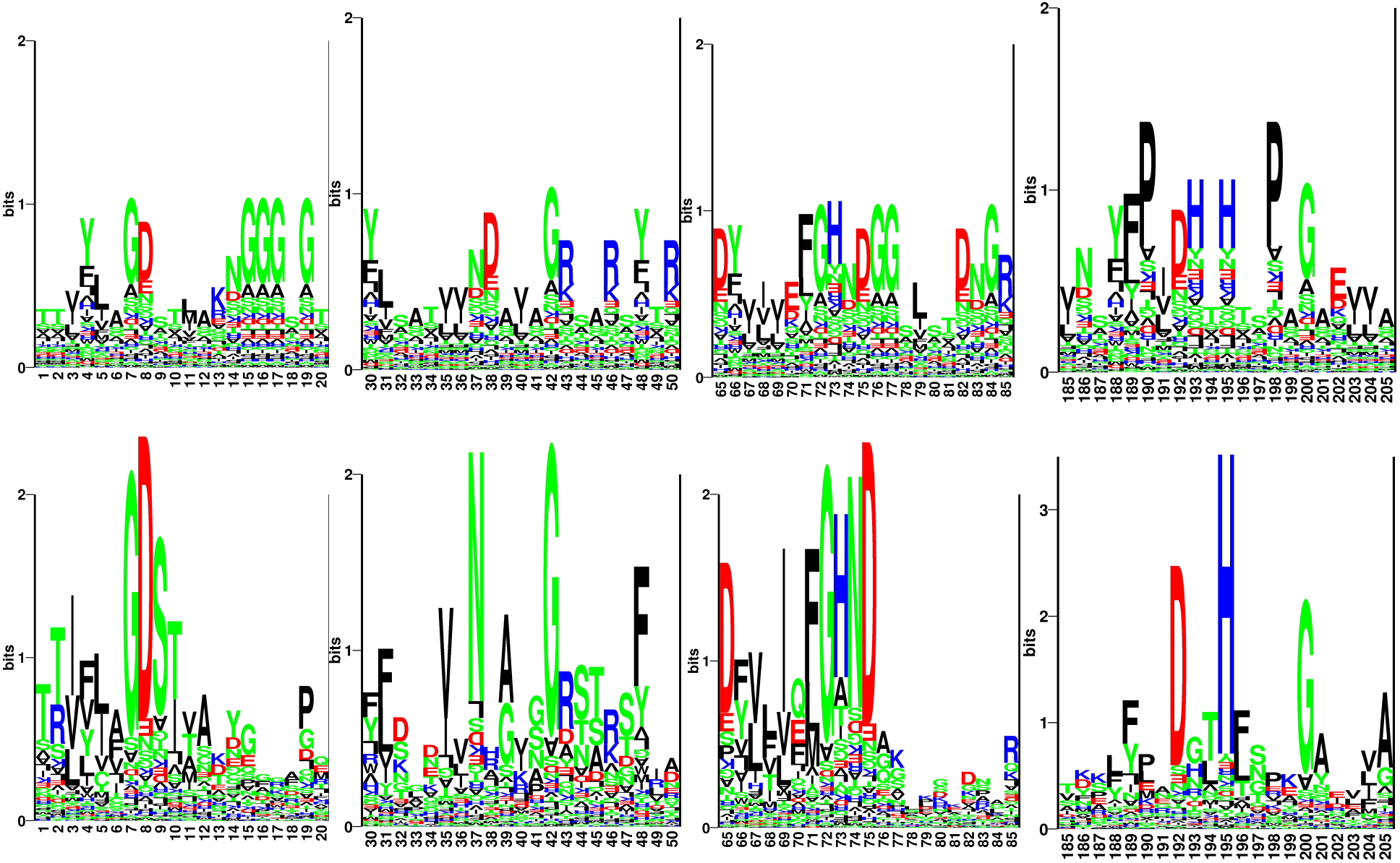
# When Blast fails!





# Example. (SGNH active site)

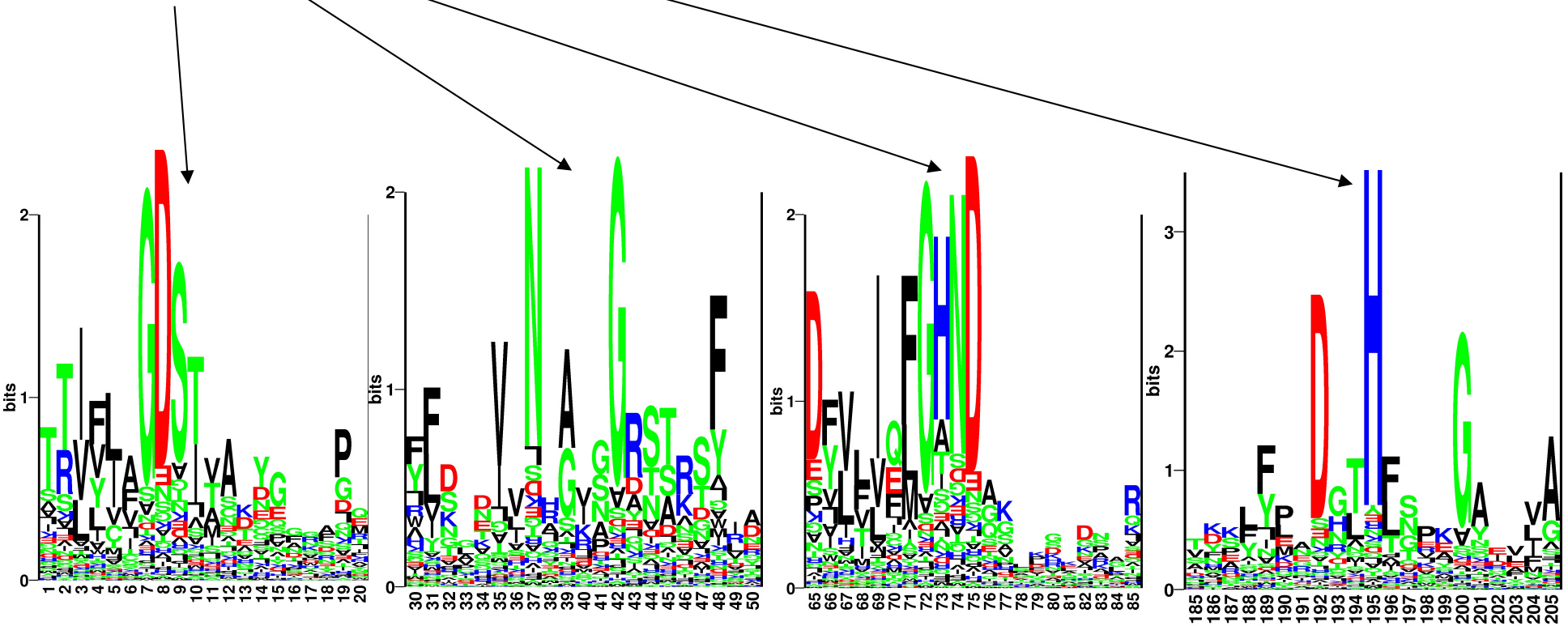




# Example. Where is the active site?

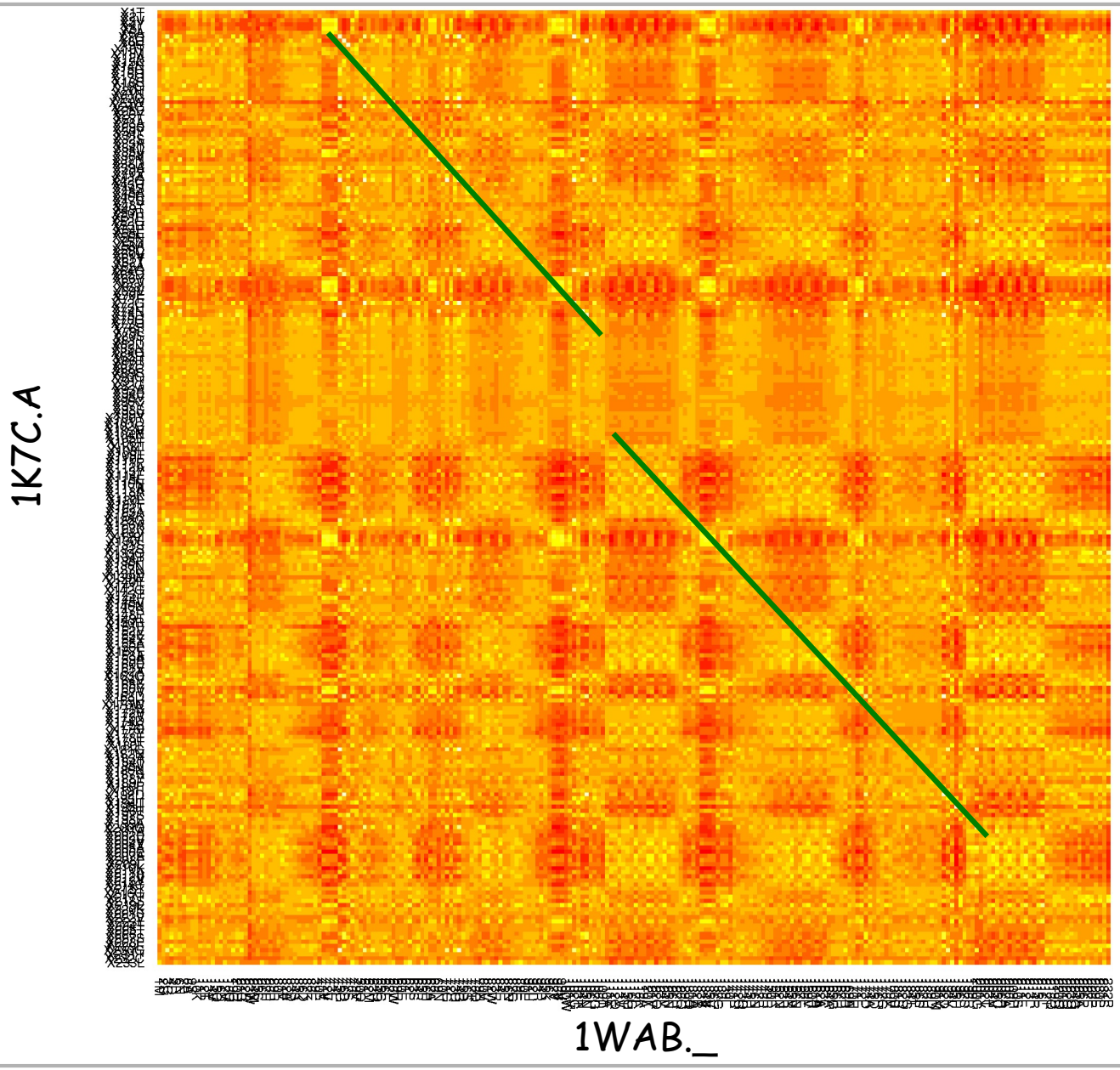


- Sequence profiles might show you where to look!
- The active site could be around
  - S9, G42, N74, and H195



# Profile-profile scoring matrix





# Profile-profile scoring matrix



Sequence profile

ARNDCQEGHILKMFPSTWYV

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1K7C.A | F | 71 | -4 | -5 | -5 | -6 | -5 | -5 | -5 | -3 | -4 | -1 | 0 | -5 | 4 | 8 | -3 | -4 | -2 | 0 | -3 |
| 1K7C.A | G | 72 | 0 | -5 | -3 | -4 | -5 | -4 | -4 | 7 | -4 | -6 | -6 | -4 | -5 | -6 | -3 | -4 | -5 | -5 | -5 |
| 1K7C.A | H | 73 | -2 | -3 | -1 | -4 | -5 | -2 | -3 | -2 | 10 | -1 | -4 | -3 | -1 | -4 | -3 | -2 | -5 | 0 | -4 |
| 1K7C.A | N | 74 | -4 | -3 | 8 | -1 | -5 | -2 | -3 | -3 | -2 | -6 | -6 | -3 | -5 | -5 | -2 | -2 | -6 | -5 | -5 |
| 1K7C.A | D | 75 | -4 | -4 | -1 | 8 | -6 | -3 | 0 | -4 | -3 | -6 | -6 | -3 | -5 | -6 | -3 | -3 | -7 | -5 | -6 |
| 1K7C.A | G | 76 | 2 | -2 | -1 | -3 | -4 | 5 | 2 | 3 | -3 | -4 | -4 | -1 | -3 | -5 | -1 | -3 | -4 | -4 | -2 |
| 1K7C.A | G | 77 | -1 | -1 | 1 | -2 | -3 | -1 | -2 | 3 | -2 | -3 | -2 | 5 | -3 | -3 | 1 | -1 | -3 | -1 | -3 |
| 1K7C.A | S | 78 | 1 | 2 | -1 | -1 | 3 | -2 | 1 | -2 | -2 | 0 | -1 | -1 | -2 | -3 | 1 | 0 | -3 | -1 | -1 |
| 1K7C.A | L | 79 | 0 | -2 | 1 | 2 | -3 | -1 | 1 | -1 | -2 | -1 | 0 | 0 | -1 | -2 | -1 | -1 | -3 | -1 | 0 |
| 1K7C.A | S | 80 | 1 | 0 | 1 | 3 | -2 | -1 | 1 | 1 | -2 | -3 | -3 | 1 | -2 | -3 | 2 | -1 | -3 | -2 | -2 |

# Example. Where is the active site?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

## Align using sequence profiles

ALN 1K7C.A 1WAB._ RMSD = 5.29522. 14% ID

```
1K7C.A TVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADVVTAGDYVIVEFGHNDGGSLSTDN
              S                                 G                                 N
1WAB._ EVVFIGDSLVQLMHQCE---IWRELFS---PLHALNFGIGGDSTQHVLW--RLENGELEHIRPKIVVVWVGTNNHG------

1K7C.A GRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKLFTAK--GAKVILSSQTPNNPWETGTFVNSPTRFVEYAEL-AAEVA
1WAB._ ----------------------HTAEQVTGGIKAIVQLVNERQPQARVVVLGLLPRGQ-HPNPLREKNRRVNELVRAALAGHP

1K7C.A GVEYVDHWSYVDSIYETLGNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSL
                                        H
1WAB._ RAHFLDADPG---FVHSDG--TISHHDMYDYLHLSRLGYTPVCRALHSLLLRL---L
```

# Profile-profile scoring matrix



Blosum profile

```
                                  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
1K7C.A G  84   0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3
1K7C.A R  85  -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3
1K7C.A T  86   0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0
1K7C.A D  87  -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3
1K7C.A C  88   0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1
....
1K7C.A C  96   0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1
1K7C.A Y  97  -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1
1K7C.A S  98   1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2
1K7C.A V  99   0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
1K7C.A Y 100  -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1
```

Sequence profile

```
1K7C.A G  84   1 -1  0 -1 -2 -1 -1  2 -1 -1  0 -1  0  1 -1 -1 -1  1  1  0
1K7C.A R  85  -1  6 -1 -1 -2 -1 -1 -2 -1 -2 -1  0 -1 -2 -1 -1 -1 -2 -1 -2
1K7C.A T  86   0 -1 -1 -1 -2  0 -1  3  3 -2 -2 -1 -1 -2  0  1  1 -2 -1 -2
1K7C.A D  87   0 -1 -1  2  1 -1  0 -1 -1  0  0 -1 -1 -1 -1  1  1 -1 -1  1
....
1K7C.A C  96   0 -1 -1 -1  6  1 -1 -1 -1 -1 -1  0 -1  0 -1 -1  1 -1  1 -1
1K7C.A Y  97   0  0 -1  0 -1  1  1 -1 -1  0 -1  0 -1  1  1  0  1 -1  2 -1
1K7C.A S  98  -1 -1  2 -1 -1  0 -1  0 -1 -1 -1 -1  0 -1 -1  2  2 -1  1  0
1K7C.A V  99   0  0 -1 -1 -1  0  1 -1 -1  0 -1  0 -1 -1  2  0  1  3 -1  1
1K7C.A Y 100   0  1  2 -1 -1 -1  0 -1 -1  0 -1  0 -1  0  3 -1  0 -1  2 -1
```

# And now you

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

# And now you



1) Calculate the alignment score of the two amino acids segments using the BLOSUM50 scoring matrix

**HTHT**
**YLHL**

How many of the alignment scores are positive (>=0)?

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

# And now you



**HTHT**
**YLHL**

2) Next, use the sequence profile calculated for 1K7C shown below to calculate the alignment score

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | T | -1 | -2 | -1 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | 3 | 6 | -4 | -3 | -2 |
| 2 | T | -2 | 0 | -2 | -3 | -3 | -2 | -3 | -4 | -4 | -3 | -3 | 1 | -3 | -4 | -3 | -1 | 7 | -5 | -4 | -2 |
| ... | | | | | | | | | | | | | | | | | | | | | |
| 193 | H | -3 | 3 | 4 | 0 | -5 | -2 | -3 | 1 | 6 | -4 | -4 | -2 | -4 | -2 | -2 | -2 | 0 | -5 | 2 | -5 |
| 194 | T | -3 | -4 | -3 | -4 | -3 | -3 | -4 | -4 | -4 | -2 | 0 | -3 | 1 | -1 | -4 | -1 | 7 | -5 | -4 | 1 |
| 195 | H | -4 | -3 | -2 | -4 | -6 | -2 | -2 | -5 | 10 | -6 | -5 | -3 | -4 | -4 | -5 | -3 | -4 | -5 | -1 | -6 |
| 196 | T | -3 | -4 | -4 | -5 | -4 | -4 | -4 | -5 | -1 | -2 | 2 | -4 | -1 | 6 | 2 | -2 | 4 | -3 | -1 | -1 |
| ... | | | | | | | | | | | | | | | | | | | | | |
| 232 | C | -1 | -3 | -2 | -3 | 9 | -3 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -3 | 2 | -1 | -3 | -3 | -2 |
| 233 | L | -2 | -3 | -4 | -4 | -2 | -3 | -4 | -4 | -4 | 3 | 5 | -3 | 1 | 0 | -4 | -3 | -2 | -2 | -2 | 1 |

Note, that the PSSM is calculated for 1K7C.

How many of the alignment scores are now positive (>=0)?

Can you understand why Psi-Blast is able to make a correct alignment of the two proteins?

# And now you



1) Calculate the alignment score of the two amino acids segments using the BLOSUM50 scoring matrix

```
HTHT
YLHL
```

How many of the alignment scores are positive (>=0)?

Scores: 2, -1, 8, -1.
Two are >= 0

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

# And now you



**HTHT**
**YLHL**

2) Next, use the sequence profile calculated for 1K7C shown below to calculate the alignment score

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | T | -1 | -2 | -1 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | 3 | 6 | -4 | -3 | -2 |
| 2 | T | -2 | 0 | -2 | -3 | -3 | -2 | -3 | -4 | -4 | -3 | -3 | 1 | -3 | -4 | -3 | -1 | 7 | -5 | -4 | -2 |
| ... | | | | | | | | | | | | | | | | | | | | | |
| 193 | H | -3 | 3 | 4 | 0 | -5 | -2 | -3 | 1 | 6 | -4 | -4 | -2 | -4 | -2 | -2 | -2 | 0 | -5 | 2 | -5 |
| 194 | T | -3 | -4 | -3 | -4 | -3 | -3 | -4 | -4 | -4 | -2 | 0 | -3 | 1 | -1 | -4 | -1 | 7 | -5 | -4 | 1 |
| 195 | H | -4 | -3 | -2 | -4 | -6 | -2 | -2 | -5 | 10 | -6 | -5 | -3 | -4 | -4 | -5 | -3 | -4 | -5 | -1 | -6 |
| 196 | T | -3 | -4 | -4 | -5 | -4 | -4 | -4 | -5 | -1 | -2 | 2 | -4 | -1 | 6 | 2 | -2 | 4 | -3 | -1 | -1 |
| ... | | | | | | | | | | | | | | | | | | | | | |
| 232 | C | -1 | -3 | -2 | -3 | 9 | -3 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -3 | 2 | -1 | -3 | -3 | -2 |
| 233 | L | -2 | -3 | -4 | -4 | -2 | -3 | -4 | -4 | -4 | 3 | 5 | -3 | 1 | 0 | -4 | -3 | -2 | -2 | -2 | 1 |

Note, that the PSSM is calculated for 1K7C.

How many of the alignment scores are now positive (>=0)?

Scores: 2, 0, 10, 2.
All 4 are >= 0

Can you understand why Psi-Blast is able to make a correct alignment of the two proteins?

# Summary



- Blast allows for extremely fast protein sequence alignment
  - HSP
  - E-value heuristics
- Psi-Blast allows for position specific scoring in alignment
  - Higher sensitivity maintaining high specificity