PSI-BLAST

Fishing in the (sequence) twilight zone

Introduction to Bioinformatics,
Faroe Islands 2024
Bent Petersen

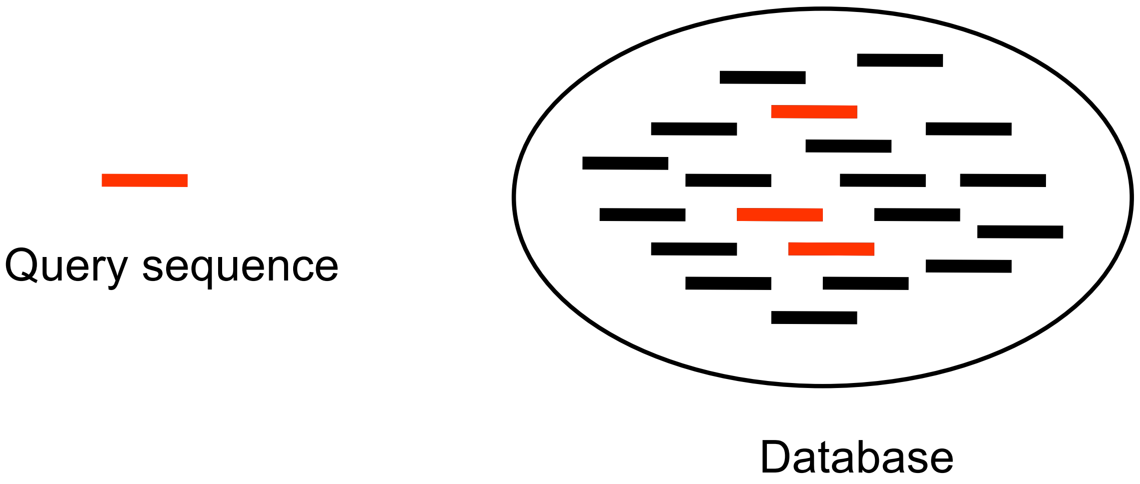
(With some borrowed concepts / slides from
Morten Nielsen, Rasmus Wernersson / Henrik Nielsen and Anders Gorm Pedersen)

Part 1

THE PROBLEM WITH PAIRWISE ALIGNMENTS

Reminder: how BLAST works

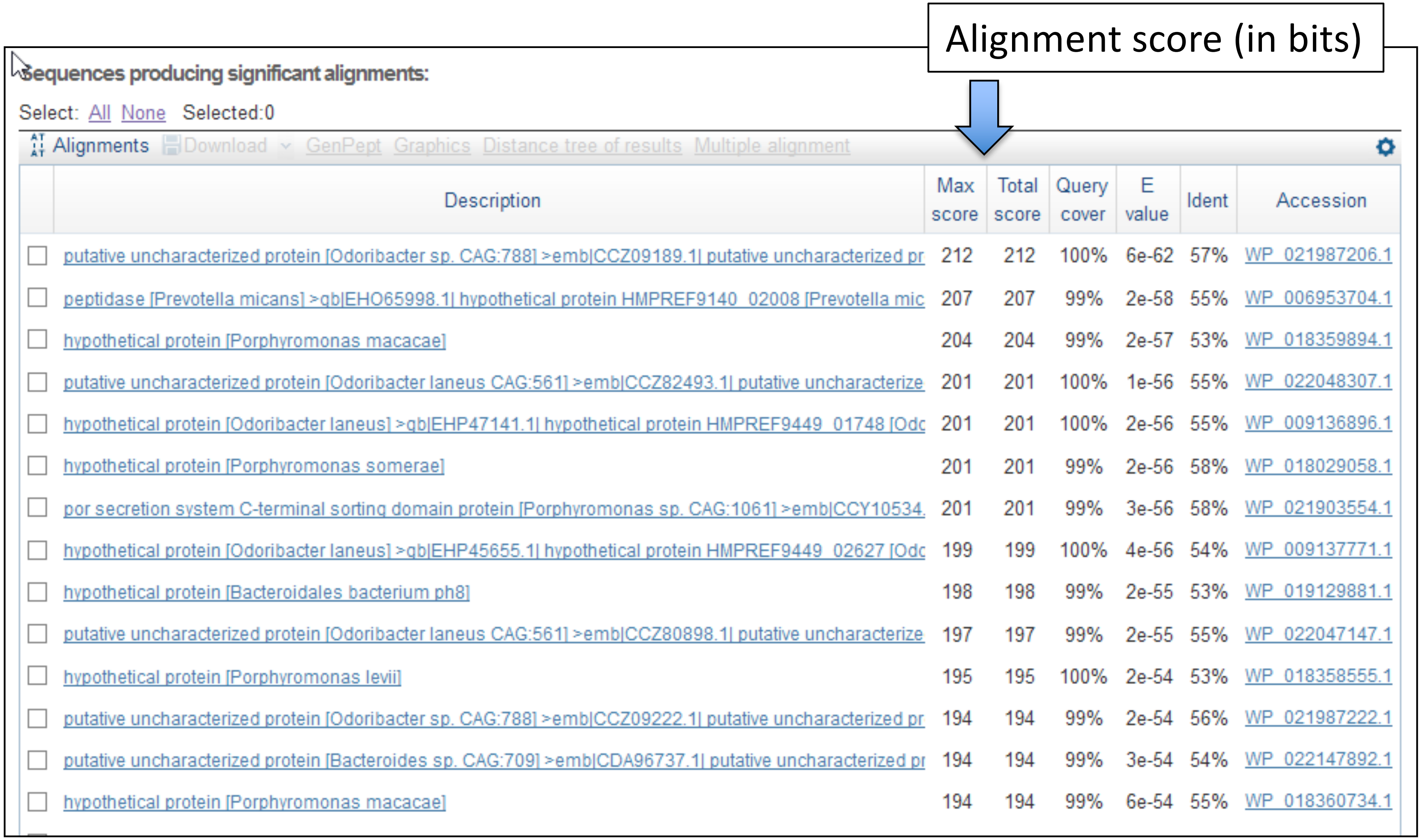
Use pairwise alignments to search databases for similar sequences



Query sequence

Database

BLASTP output



Alignment score (in bits)

Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	putative uncharacterized protein [Odoribacter sp. CAG:788] >emb|CCZ09189.1| putative uncharacterized pr	212	212	100%	6e-62	57%	WP_021987206.1
□	peptidase [Prevotella micans] >gb|EHO65998.1| hypothetical protein HMPREF9140_02008 [Prevotella mic	207	207	99%	2e-58	55%	WP_006953704.1
□	hypothetical protein [Porphyromonas macacae]	204	204	99%	2e-57	53%	WP_018359894.1
□	putative uncharacterized protein [Odoribacter laneus CAG:561] >emb|CCZ82493.1| putative uncharacterize	201	201	100%	1e-56	55%	WP_022048307.1
□	hypothetical protein [Odoribacter laneus] >gb|EHP47141.1| hypothetical protein HMPREF9449_01748 [Odc	201	201	100%	2e-56	55%	WP_009136896.1
□	hypothetical protein [Porphyromonas somerae]	201	201	99%	2e-56	58%	WP_018029058.1
□	por secretion system C-terminal sorting domain protein [Porphyromonas sp. CAG:1061] >emb|CCY10534.	201	201	99%	3e-56	58%	WP_021903554.1
□	hypothetical protein [Odoribacter laneus] >gb|EHP45655.1| hypothetical protein HMPREF9449_02627 [Odc	199	199	100%	4e-56	54%	WP_009137771.1
□	hypothetical protein [Bacteroidales bacterium ph8]	198	198	99%	2e-55	53%	WP_019129881.1
□	putative uncharacterized protein [Odoribacter laneus CAG:561] >emb|CCZ80898.1| putative uncharacterize	197	197	99%	2e-55	55%	WP_022047147.1
□	hypothetical protein [Porphyromonas levii]	195	195	100%	2e-54	53%	WP_018358555.1
□	putative uncharacterized protein [Odoribacter sp. CAG:788] >emb|CCZ09222.1| putative uncharacterized pr	194	194	99%	2e-54	56%	WP_021987222.1
□	putative uncharacterized protein [Bacteroides sp. CAG:709] >emb|CDA96737.1| putative uncharacterized pr	194	194	99%	3e-54	54%	WP_022147892.1
□	hypothetical protein [Porphyromonas macacae]	194	194	99%	6e-54	55%	WP_018360734.1

(Example from the BLAST exercise: At the protein level it was quite evident, that the unknown sequence was a serine peptidase)

BLASTP alignment

Alignment score

```
>ref|WP_006953704.1| peptidase [Prevotella micans]
Length=922

 Score =   207 bits (526),  Expect = 2e-58, Method: Compositional matrix adjust.
 Identities = 117/211 (55%), Positives = 145/211 (69%), Gaps = 14/211 (7%)

Query  2    GHGTHVAGTVAAVNNNGIGVAGVAGGNGSTNSGARLMSTQIFNSDGDYTNSETLVYRAIV  61
            GHGTHVAGTVAA NNNG+GVAG+AGG+GSTNSG RL+S QIF    +  ++E     AI
Sbjct  279  GHGTHVAGTVAARNNNGLGVAGIAGGDGSTNSGVRLLSCQIFRKSKEEGSAEA----AIK  334

Query  62   YGADNGAVISQNSWGSQSL-TIKELQKA---AIDYFIDYAGMDETGEIQT-GPMRGGIFI  116
            Y ADNGAVI+Q SWG  S   +KEL K+   AIDYFI +AG D  G  ++  PM+GG+ I
Sbjct  335  YAADNGAVIAQCSWGYASKENVKELPKSLKEAIDYFITFAGCDAHGAQRSDSPMKGGVMI  394

Query  117  AAAGNDNVSTPNMPSAYERVLAVASMGPDFTKASYSTFGTWTDITAPGGDIDKFDLSEYG  176
             AAGN+N++    P+AYE+V++VAS   +F KASYS +  W  I+APGGD D F L + G
Sbjct  395  FAAGNENMNFKEFPAAYEKVISVASTAWNFQKASYSNYADWVSISAPGGDQDAFGL-KAG  453

Query  177  VLSTYADNY----YAYGEGTSMACPHVAGAA  203
            VLST         Y Y +GTSMACPHV+G A
Sbjct  454  VLSTMPKKIASSGYGYMQGTSMACPHVSGIA  484
```

(Example from the BLAST exercise: At the protein level it was quite evident, that the unknown sequence was a serine peptidase)

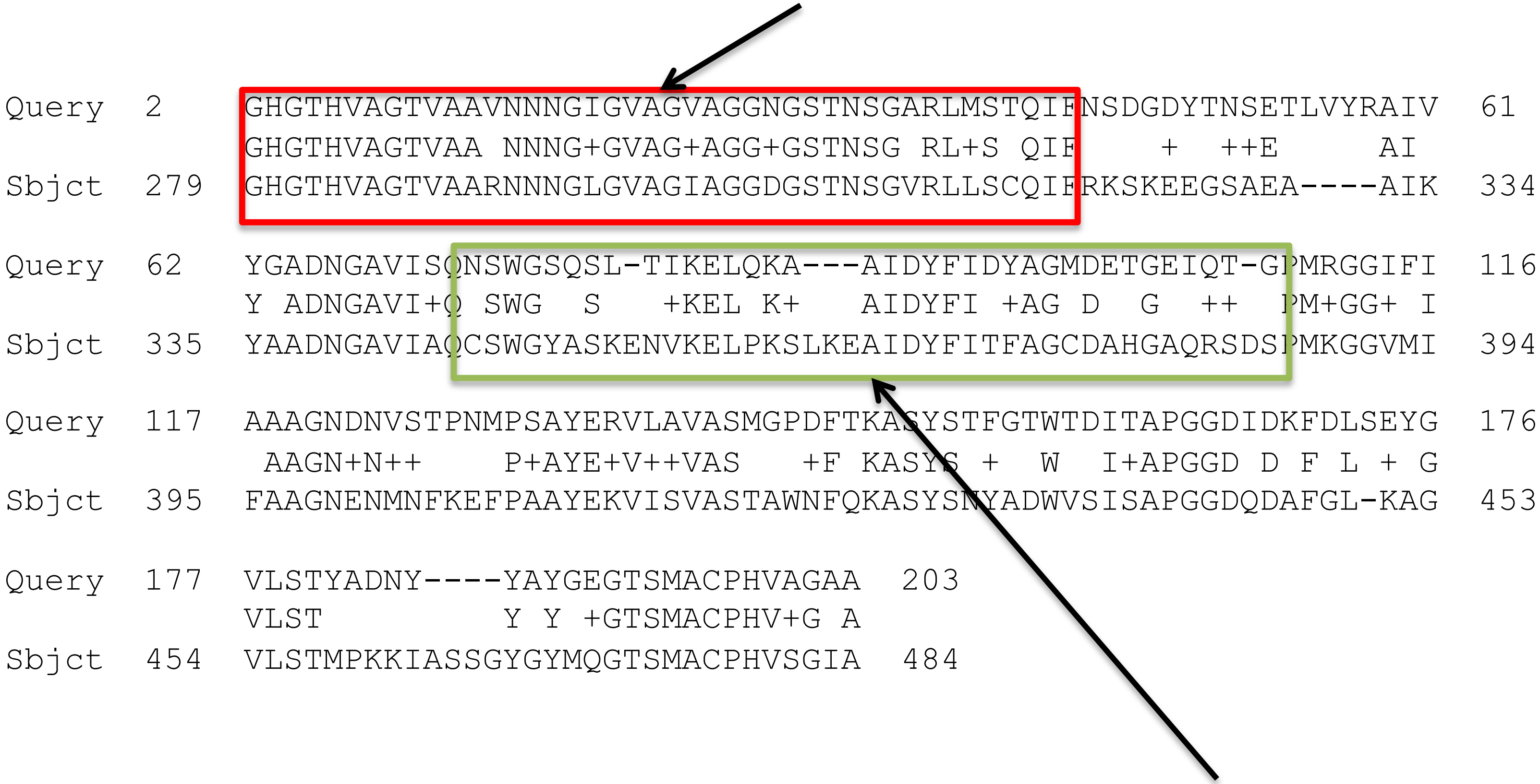
Alignment matrix: BLOSUM62

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

Not all positions are biological equal

Conserved region:

Is likely important for the function of the enzyme



```
Query  2     GHGTHVAGTVAAVNNNGIGVAGVAGGNGSTNSGARLMSTQIFNSDGDYTNSETLVYRAIV  61
             GHGTHVAGTVAA NNNG+GVAG+AGG+GSTNSG RL+S QIF    +  ++E     AI
Sbjct  279   GHGTHVAGTVAARNNNGLGVAGIAGGDGSTNSGVRLLSCQIFRKSKEEGSAEA----AIK  334

Query  62    YGADNGAVISQNSWGSQSL-TIKELQKA---AIDYFIDYAGMDETGEIQT-GPMRGGIFI  116
             Y ADNGAVI+Q SWG  S   +KEL K+   AIDYFI +AG D  G  ++  PM+GG+ I
Sbjct  335   YAADNGAVIAQCSWGYASKENVKELPKSLKEAIDYFITFAGCDAHGAQRSDSPMKGGVMI  394

Query  117   AAAGNDNVSTPNMPSAYERVLAVASMGPDFTKASYSTFGTWTDITAPGGDIDKFDLSEYG  176
              AAGN+N++    P+AYE+V++VAS   +F KASYS +  W  I+APGGD D F L + G
Sbjct  395   FAAGNENMNFKEFPAAYEKVISVASTAWNFQKASYSNYADWVSISAPGGDQDAFGL-KAG  453

Query  177   VLSTYADNY----YAYGEGTSMACPHVAGAA  203
             VLST         Y Y +GTSMACPHV+G A
Sbjct  454   VLSTMPKKIASSGYGYMQGTSMACPHVSGIA  484
```

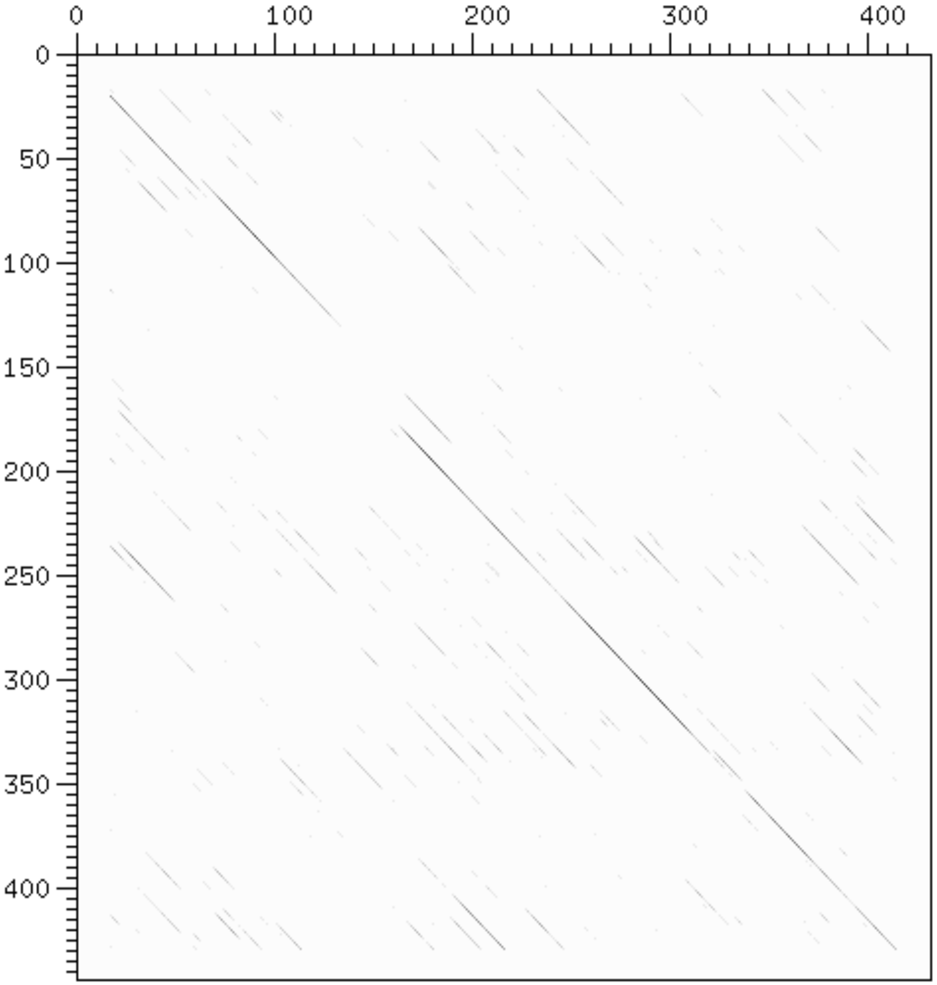
Variable region:

Is likely not that important for the function of the enzyme

Scoring of pairwise alignments

- In a normal pairwise alignment the same scores (the same matrix) is used for all positions
- As we saw before the selection pressure on the different parts of the sequence is not equal, and ideally we should take this into account
- IMPORTANT: if the sequences is of high enough similarity, this is usually not a big issue

Reminder: Dot-plot

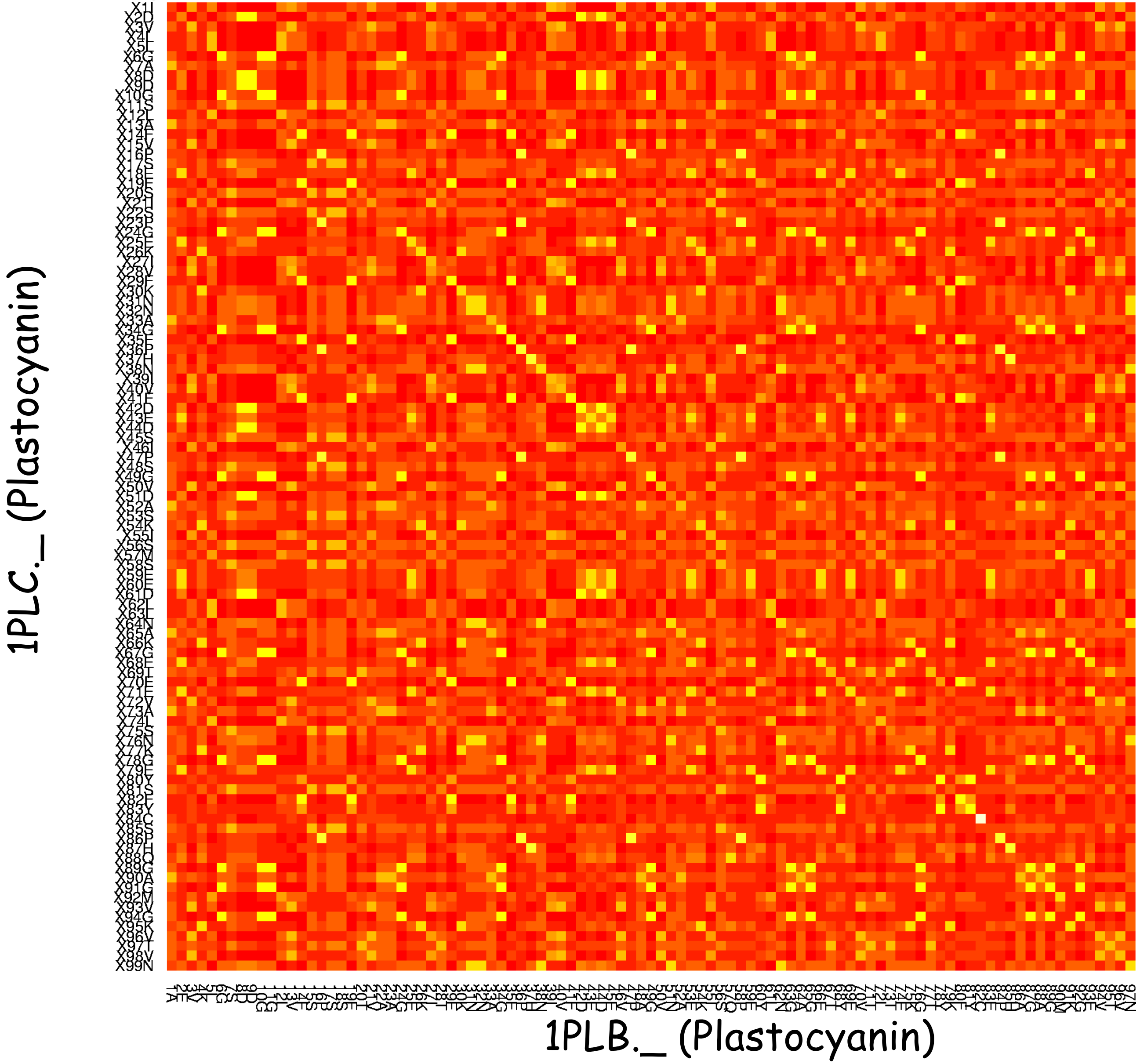


1. Place two sequences along axes of plot

2. Place dot at grid points where two sequences have identical residues

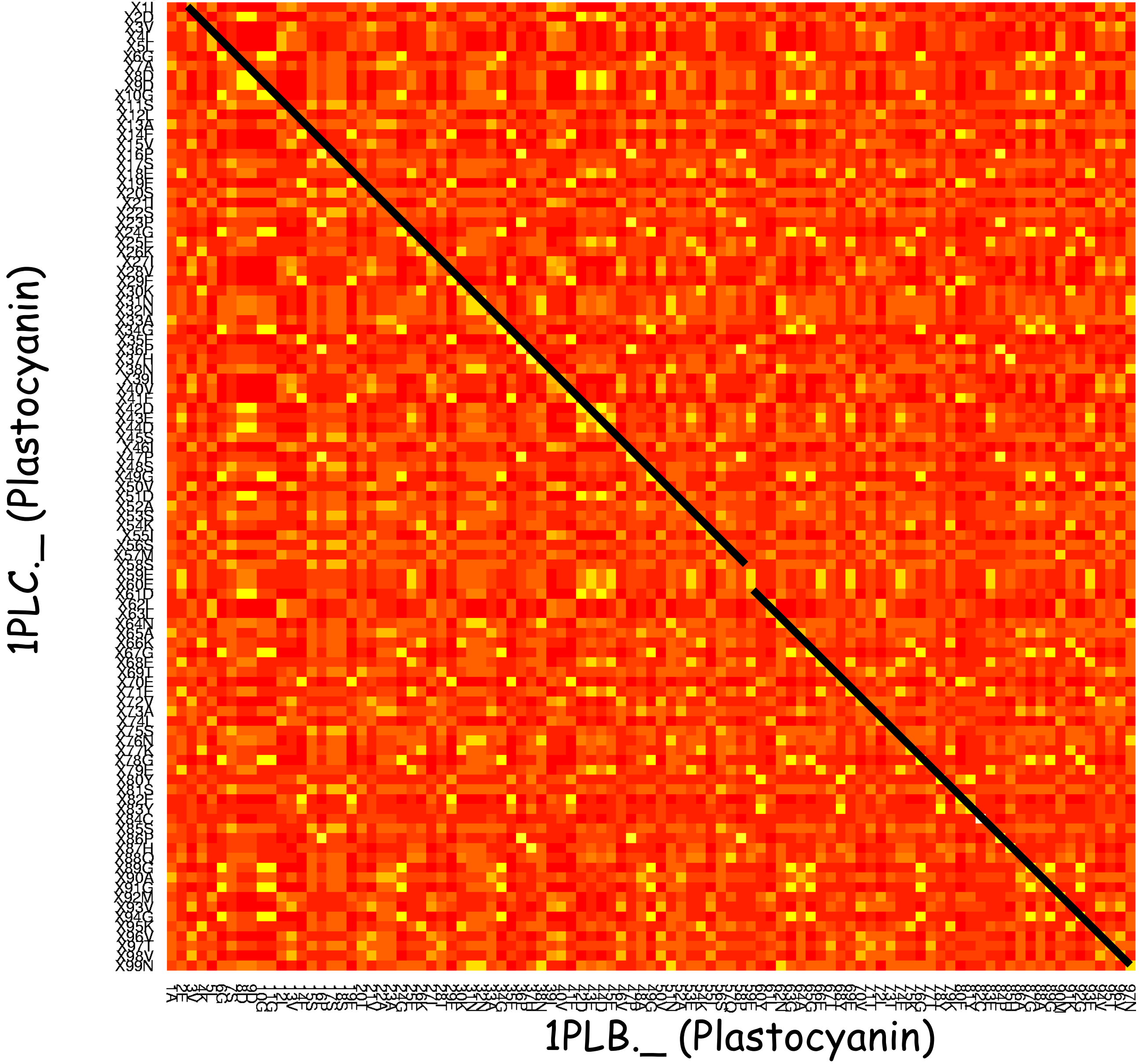
3. Diagonals correspond to conserved regions

Dot-plot with BLOSUM colors



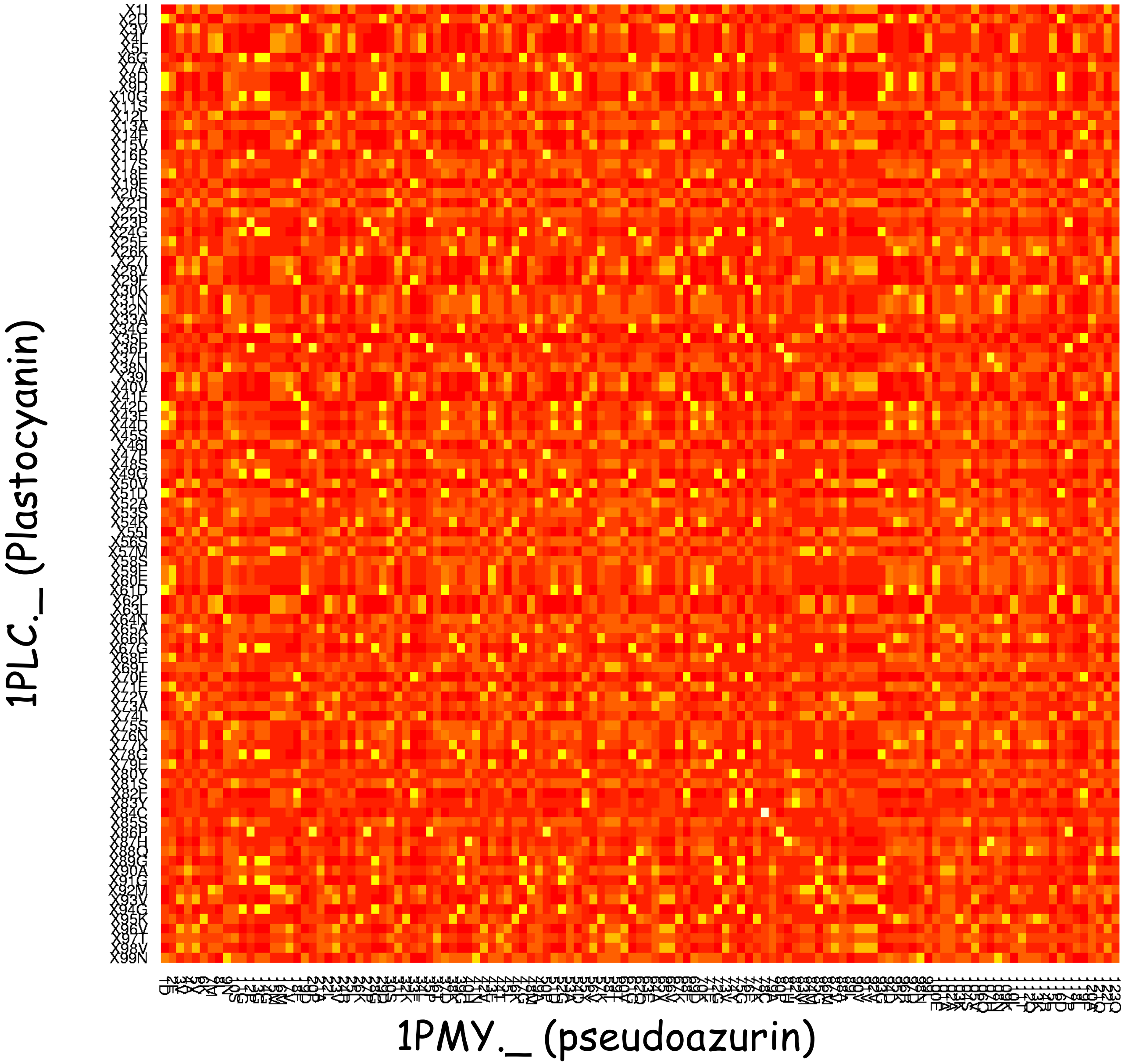
Relationship can be detected using BLASTP

Dot-plot with BLOSUM colors



Relationship can be detected using BLASTP

Color dot-plot of low-similarity sequences



Relationship CANNOT be detected using BLASTP

Part 2

THOUGHTS ABOUT HOW TO SOLVE THE PROBLEM

Idea catalog

- We would like to build a **scoring model** for pairwise alignments that more closely resembles what happens in **real sequence evolution**
  - Highly conserved sites/regions should have a high weight
  - Non-conserved regions should have a low weight (be allowed to vary without counting too much against the alignment score)

- IMPORTANT: Different protein families are under different selection pressure, so our model needs to account for this

Protein families

- Tools we can use, to identify the selective pressure on protein families:
  - Data sets of truly related proteins
  - Multiple alignment
  - Logo plots
  - Weight matrices

Protein family data sets

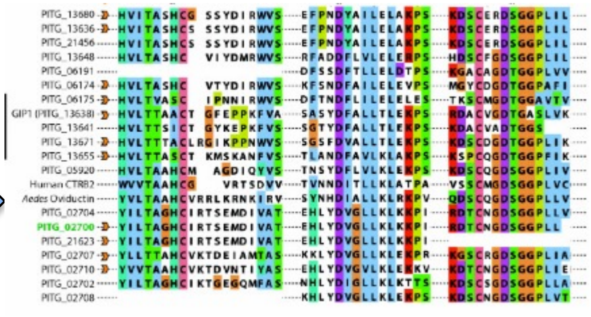
- How we can build such data sets:
  - – Already known collections (literature, curated data sets)
    - Limitation: What have other people looked at before
  - – “Text based” search in protein data bases (e.g. UniProt)
    - Limitation: Coverage, how well are the sequences described
  - – BlastP (!)
    - Limitation: We only expect to find sequences of moderate to high similarity

Signal across multiple sequences

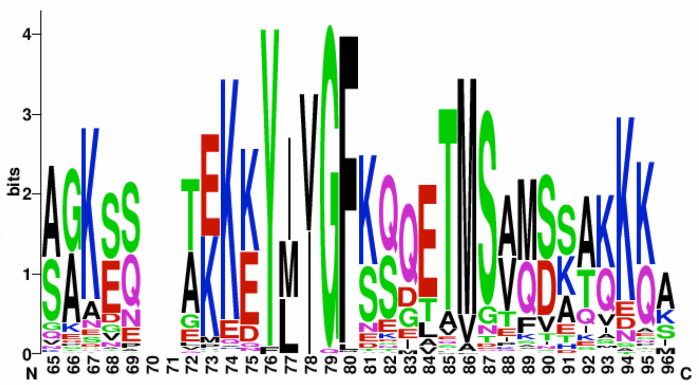
Multiple seq alignment (“MSA”)
(e.g. using MAFFT)

LOGO plot
(e.g using WebLogo)

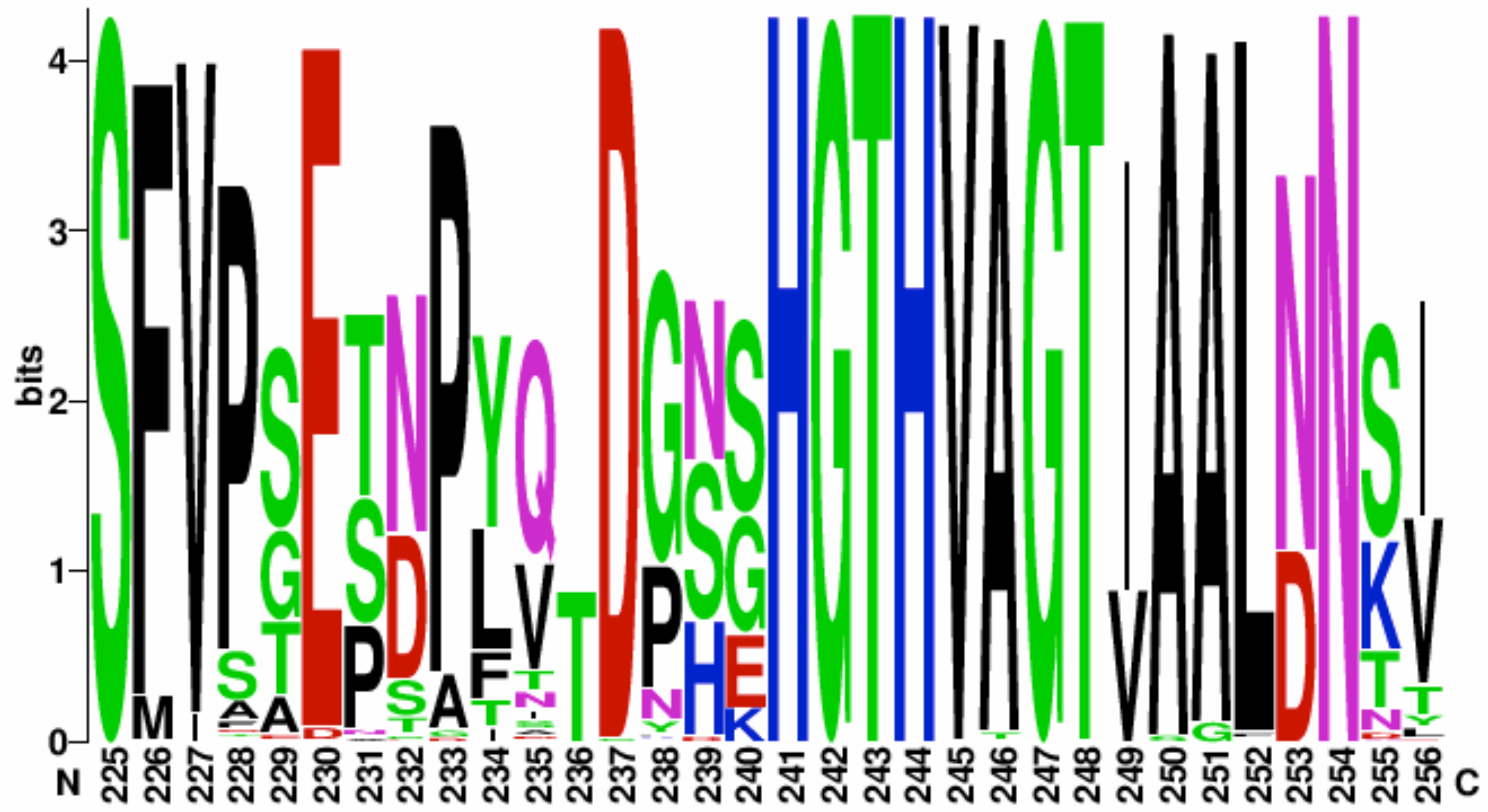
Seq 1
Seq 2
Seq 3
..
Seq n



```
PITG_13680 --HVITASHCG SSYDIRWVS---EFPNDYAILELAKPS---KDSCERDSGGPLIL---
PITG_13636 --HVITASHCS SSYDIRWVS---EFPNDYAILELAKPS---KDSCERDSGGPLIL---
PITG_21456 ---HVITASHCS SSYDIRWVS---EFPNDYAILELAKPS---KDSCERDSGGPLIL---
PITG_13648 ---HVLTASHC  VIYDMRWVS---RFADDFLVLELERPS---HDSCFGDSGGPLIL---
PITG_06191 ---                   ---DFSSDFTLLELDTPS---KGACAGDTGGPLVV---
PITG_06174 --HVLTASHC  VTYDIRWVS---KFSNDFAILELEVPS---MGYCDGDTGGPAFI---
PITG_06175 --HVLTVASC  IPNNIRWVS---DFTNDFLILELELES---TKSCMGDTGGAVTV---
GIP1 (PITG_13638) HVLTTAACT GFEPPKFVA---SASYDFALLTLEKPS---RDACVGDTGASLVK---
PITG_13641 ---HVLTTSICT GYKEPKFVS---SGTYDFALLTLEKPS---KDACVADTGGS   ---
PITG_13671 --HVLTTTACLRGIKPPNWVS---SGSFDVALLTLEKPS---KDSCDGDTGGPLIK---
PITG_13655 --HVLTTASCT KMSKANFVS---TLANDFAVLKLAKPS---KSPCQGDTGGPFIK---
PITG_05920 ---HVLTAAHCM  AGDIQYVS---TNSYDFLVLKLEKPS---KDSCQGDSGGPLIV---
Human CTRB2 ---WVVTAAHCG   VRTSDVV---TVNNDITLLKLATPA---VSSCMGDSGGPLVC---
Aedes Oviductin ---YVLTAAHCVRRLKRNKIRV---SYNHDIALLKLRKPV---QDSCQGDSGGPLLV---
PITG_02704 ---YILTAGHCIRTSEMDIVAT---EHLYDVGLLKLKKPI---RDTCNGDSGGPLLV---
PITG_02700 --YILTAGHCIRTSEMDIVAT---EHLYDVGLLKLKKPI---RDTCNGDSGGPLL ---
PITG_21623 --YILTAGHCIRTSEMDIVAT---EHLYDVGLLKLKKPI---
PITG_02707 --YLLTTAHCVKTDEIAMTAS---KKLYDVGLLKLEKPR---KGSCRGDSGGPLIA---
PITG_02710 --YVVTAAHCVKTDVNTIYAS---EHLYDVGVLKLEKKV---KDTCNGDSGGPLIE---
PITG_02702 ---YILTAGHCIKTGEGQMFAS---NHLYDIGLLKLKTTS---KDSCSGDSGGPLLA---
PITG_02708 ---                   ---KHLYDVGLLKLEKPS---KDSCNGDSGGPLVT---
```

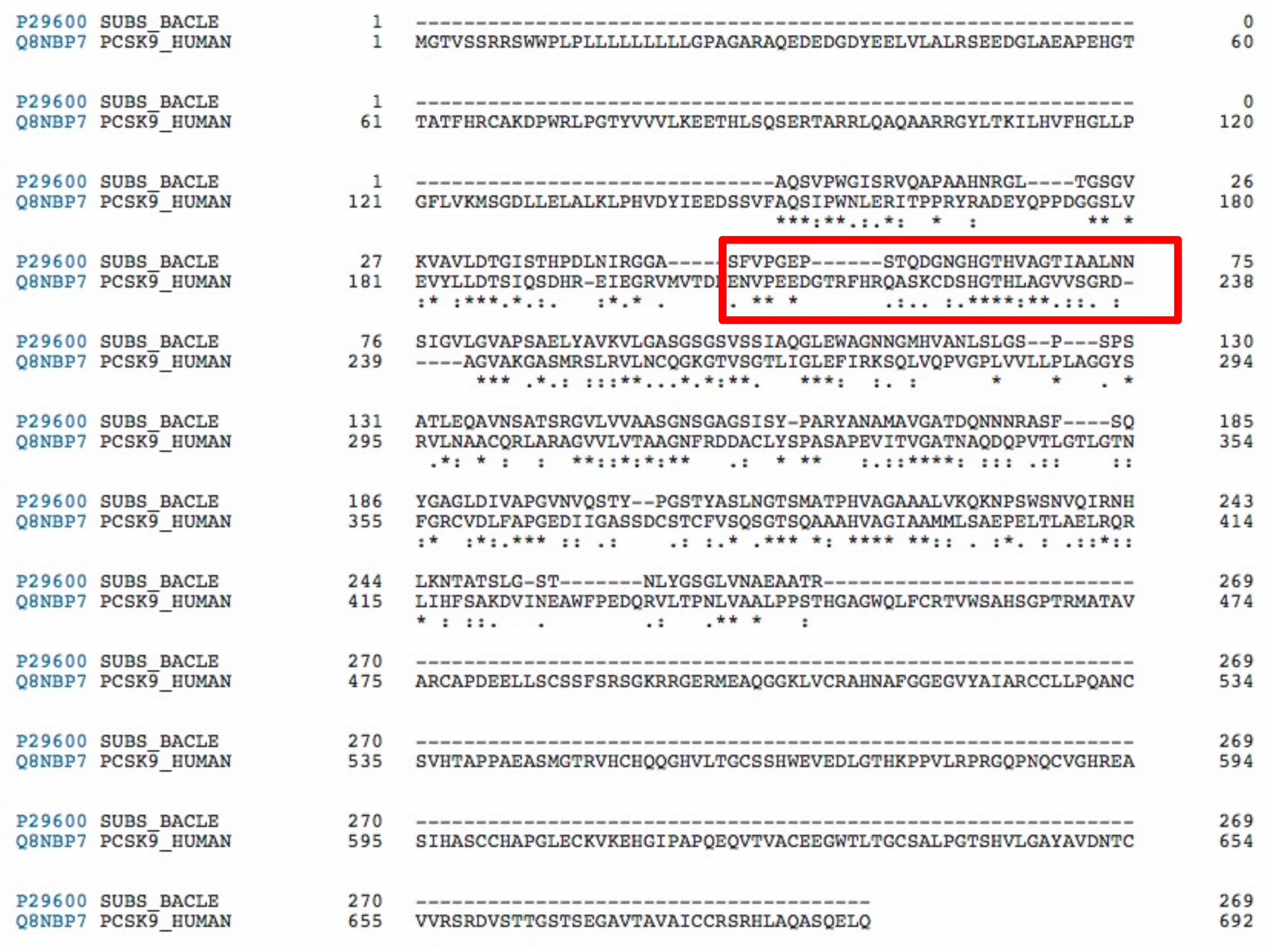


LOGO example



Small section of a LOGO from 1500 aligned bacterial serine proteases

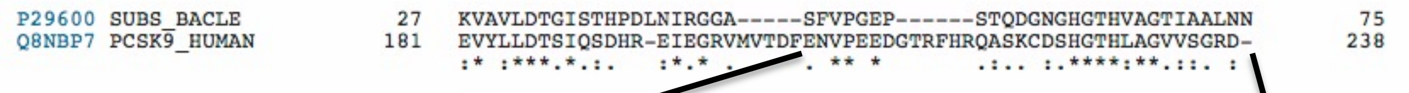
Going back to pairwise alignments



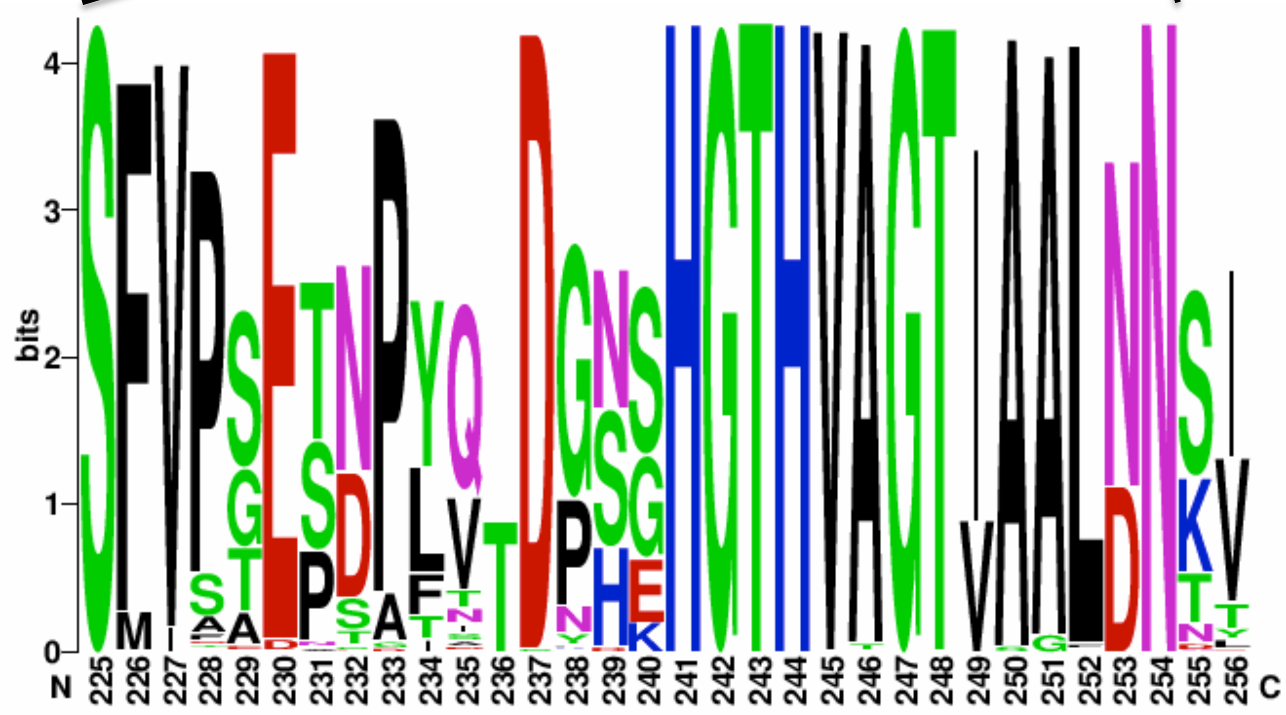
```
P29600 SUBS_BACLE     1   ------------------------------------------------------------      0
Q8NBP7 PCSK9_HUMAN    1   MGTVSSRRSWWPLPLLLLLLLLLGPAGARAQEDEDGDYEELVLALRSEEDGLAEAPEHGT     60

P29600 SUBS_BACLE     1   ------------------------------------------------------------      0
Q8NBP7 PCSK9_HUMAN   61   TATFHRCAKDPWRLPGTYVVVLKEETHLSQSERTARRLQAQAARRGYLTKILHVFHGLLP    120

P29600 SUBS_BACLE     1   ------------------------------AQSVPWGISRVQAPAAHNRGL----TGSGV     26
Q8NBP7 PCSK9_HUMAN  121   GFLVKMSGDLLELALKLPHVDYIEEDSSVFAQSIPWNLERITPPRYRADEYQPPDGGSLV    180
                                                        ***:**.:.*:  *  :         ** *

P29600 SUBS_BACLE    27   KVAVLDTGISTHPDLNIRGGA-----SFVPGEP------STQDGNGHGTHVAGTIAALNN     75
Q8NBP7 PCSK9_HUMAN  181   EVYLLDTSIQSDHR-EIEGRVMVTDFENVPEEDGTRFHRQASKCDSHGTHLAGVVSGRD-    238
                          :* :***.*.:.   :*.* .     . ** *       .:.. :.****:**.::. :

P29600 SUBS_BACLE    76   SIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMHVANLSLGS--P---SPS    130
Q8NBP7 PCSK9_HUMAN  239   ----AGVAKGASMRSLRVLNCQGKGTVSGTLIGLEFIRKSQLVQPVGPLVVLLPLAGGYS    294
                               *** .*.: :::**...*.*:**.   ***:  :. :      *    *   . *

P29600 SUBS_BACLE   131   ATLEQAVNSATSRGVLVVAASGNSGAGSISY-PARYANAMAVGATDQNNNRASF----SQ    185
Q8NBP7 PCSK9_HUMAN  295   RVLNAACQRLARAGVVLVTAAGNFRDDACLYSPASAPEVITVGATNAQDQPVTLGTLGTN    354
                           .*: * :  :  **::*:*:**   .:  * **   :.::****: ::: .::    ::

P29600 SUBS_BACLE   186   YGAGLDIVAPGVNVQSTY--PGSTYASLNGTSMATPHVAGAAALVKQKNPSWSNVQIRNH    243
Q8NBP7 PCSK9_HUMAN  355   FGRCVDLFAPGEDIIGASSDCSTCFVSQSGTSQAAAHVAGIAAMMLSAEPELTLAELRQR    414
                          :*  :*:.*** :: .:    .: :.* .*** *: **** **:: . :*. : .::*::

P29600 SUBS_BACLE   244   LKNTATSLG-ST-------NLYGSGLVNAEAATR--------------------------    269
Q8NBP7 PCSK9_HUMAN  415   LIHFSAKDVINEAWFPEDQRVLTPNLVAALPPSTHGAGWQLFCRTVWSAHSGPTRMATAV    474
                          * : ::.   .        .:   .** *   :

P29600 SUBS_BACLE   270   ------------------------------------------------------------    269
Q8NBP7 PCSK9_HUMAN  475   ARCAPDEELLSCSSFSRSGKRRGERMEAQGGKLVCRAHNAFGGEGVYAIARCCLLPQANC    534

P29600 SUBS_BACLE   270   ------------------------------------------------------------    269
Q8NBP7 PCSK9_HUMAN  535   SVHTAPPAEASMGTRVHCHQQGHVLTGCSSHWEVEDLGTHKPPVLRPRGQPNQCVGHREA    594

P29600 SUBS_BACLE   270   ------------------------------------------------------------    269
Q8NBP7 PCSK9_HUMAN  595   SIHASCCHAPGLECKVKEHGIPAPQEQVTVACEEGWTLTGCSALPGTSHVLGAYAVDNTC    654

P29600 SUBS_BACLE   270   --------------------------------------                          269
Q8NBP7 PCSK9_HUMAN  655   VVRSRDVSTTGSTSEGAVTAVAICCRSRHLAQASQELQ                          692
```

Alignment: Bacterial serine peptidase (“Savinase”) vs. human PCSK9

Going back to pairwise alignments



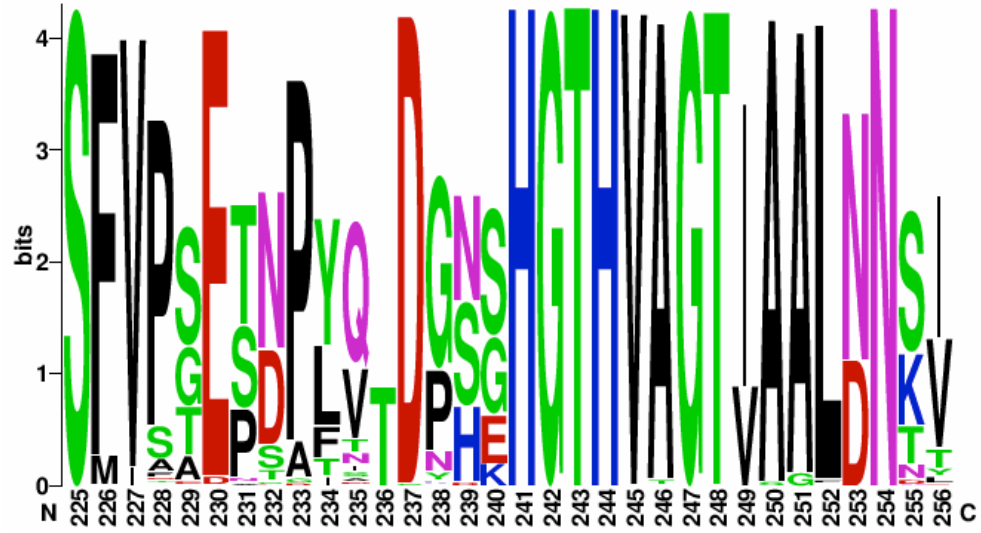
```
P29600 SUBS_BACLE          27   KVAVLDTGISTHPDLNIRGGA-----SFVPGEP------STQDGNGHGTHVAGTIAALNN      75
Q8NBP7 PCSK9_HUMAN        181   EVYLLDTSIQSDHR-EIEGRVMVTDFENVPEEDGTRFHRQASKCDSHGTHLAGVVSGRD-     238
                                :* :***.*.:.   :*.* .     . ** *       .:.. :.****:**.::. :
```



Goal: combine observations from large data set (1500 sequences) into the scoring scheme for the pairwise alignment

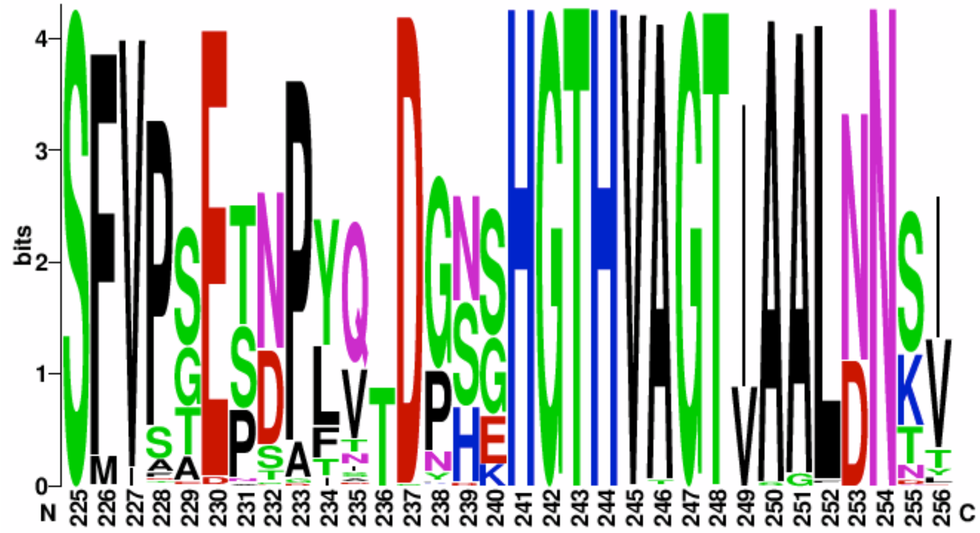
Naïve approach

- A naïve approach that would actually work:
  - When calculating the alignment score, look at how much information is in the LOGO plot (from the large data set) at the corresponding position.
  - Then scale the score from the BLOSUM62 matrix according to this.
  - That would mean that highly conserved regions would count more and variable regions would count less in the alignment score.



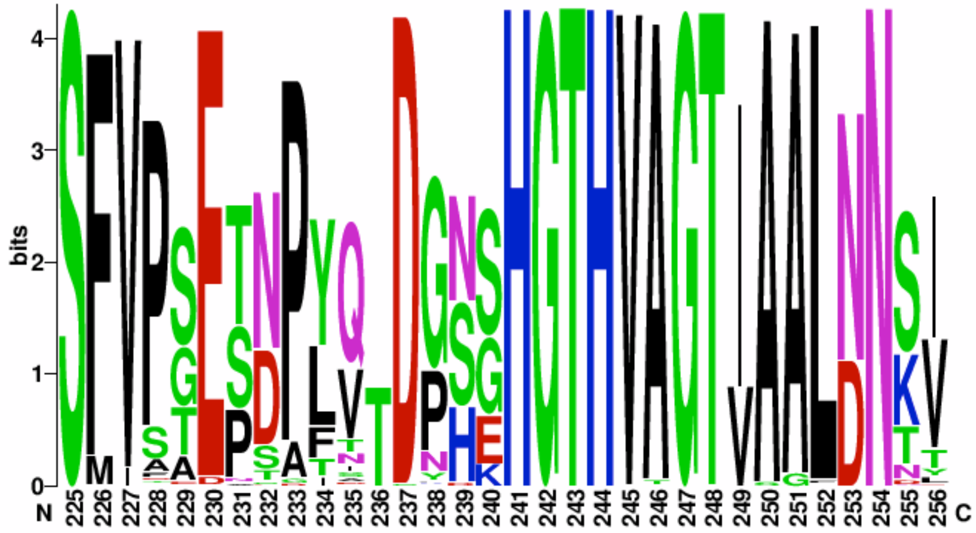
But we can actually do better

- Some things the naïve approach do not cover:
  - From the LOGO plot, a clear preference for certain amino acids at certain positions is seen.
  - We would like to build this into the model.



Weight matrices to the rescue

- Weight matrices:
  – Built from large data sets of aligned sequences.
  – Is essentially log2(observed/expected) AA frequencies (the pseudo-frequencies is a trick to cope with small data sets).
  – A score for how well new sequences match the pattern in the matrix can easily be calculated.



How to construct a WM

- A weight matrix is given as

  $W_{ij} = \log_2(p_{ij}/q_j)$

  – where i is a position in the motif, and j an amino acid. q_j is the background frequency for amino acid j.

  – if p_{ij} = 0, we cannot apply the logarithm, so we have to add pseudocounts.

Notice the LOG transform

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

- W is a L x 20 matrix, L is motif length
- Wij > 0, Amino acid is seen **more** often than expected from random
- Wij < 0, Amino acid is seen **less** often than expected from random

Scoring a sequence

- Score sequences to weight matrix by looking up and adding L values from the matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

RLLDDTPEV	**11.9**
GLLGNVSTV	
ALAKAAAAL	

Which peptide is most likely to bind?
Which peptide second?

Scoring a sequence

- Score sequences to weight matrix by looking up and adding L values from the matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0.6	0.4	-3.5	-2.4	-0.4	-1.9	-2.7	0.3	-1.1	1.0	0.3	0.0	1.4	1.2	-2.7	1.4	-1.2	-2.0	1.1	0.7
2	-1.6	-6.6	-6.5	-5.4	-2.5	-4.0	-4.7	-3.7	-6.3	1.0	5.1	-3.7	3.1	-4.2	-4.3	-4.2	-0.2	-5.9	-3.8	0.4
3	0.2	-1.3	0.1	1.5	0.0	-1.8	-3.3	0.4	0.5	-1.0	0.3	-2.5	1.2	1.0	-0.1	-0.3	-0.5	3.4	1.6	0.0
4	-0.1	-0.1	-2.0	2.0	-1.6	0.5	0.8	2.0	-3.3	0.1	-1.7	-1.0	-2.2	-1.6	1.7	-0.6	-0.2	1.3	-6.8	-0.7
5	-1.6	-0.1	0.1	-2.2	-1.2	0.4	-0.5	1.9	1.2	-2.2	-0.5	-1.3	-2.2	1.7	1.2	-2.5	-0.1	1.7	1.5	1.0
6	-0.7	-1.4	-1.0	-2.3	1.1	-1.3	-1.4	-0.2	-1.0	1.8	0.8	-1.9	0.2	1.0	-0.4	-0.6	0.4	-0.5	-0.0	2.1
7	1.1	-3.8	-0.2	-1.3	1.3	-0.3	-1.3	-1.4	2.1	0.6	0.7	-5.0	1.1	0.9	1.3	-0.5	-0.9	2.9	-0.4	0.5
8	-2.2	1.0	-0.8	-2.9	-1.4	0.4	0.1	-0.4	0.2	-0.0	1.1	-0.5	-0.5	0.7	-0.3	0.8	0.8	-0.7	1.3	-1.1
9	-0.2	-3.5	-6.1	-4.5	0.7	-0.8	-2.5	-4.0	-2.6	0.9	2.8	-3.0	-1.8	-1.4	-6.2	-1.9	-1.6	-4.9	-1.6	4.5

RLLDDTPEV	11.9	84nM
GLLGNVSTV	14.7	23nM
ALAKAAAAL	4.3	309nM

Which peptide is most likely to bind?

Which peptide second?

Where have we seen this before?

Estimation of the BLOSUM 62 matrix

- Use the BLOCKS database (ungapped alignments of especially conserved regions of multiple alignments)
- For each alignment in the BLOCKS database the sequences are grouped into clusters with at least 62% identical residues (for BLOSUM 62)
- All pairs of sequences are compared *between* clusters, and the **observed pair frequencies** are noted

```
ID    FIBRONECTIN_2; BLOCK
COG9_CANFA  GNSAGEPCVFPFIFLGKQYSTCTREGRGDGHLWCATT
COG9_RABIT  GNADGAPCHFPFTFEGRSYTACTTDGRSDGMAWCSTT
FA12_HUMAN  LTVTGEPCHFPFQYHRQLYHKCTHKGRPGPQPWCATT
HGFA_HUMAN  LTEDGRPCRFPFRYGGRMLHACTSEGSAHRKWCATTH
MANR_HUMAN  GNANGATCAFPFKFENKWYADCTSAGRSDGWLWCGTT
MPRI_MOUSE  ETDDGEPCVFPFIYKGKSYDECVLEGRAKLWCSKTAN
PB1_PIG     AITSDDKCVFPFIYKGNLYFDCTLHDSTYYWCSVTTY
SFP1_BOVIN  ELPEDEECVFPFVYRNRKHFDCTVHGSLFPWCSLDAD
SFP3_BOVIN  AETKDNKCVFPFIYGNKKYFDCTLHGSLFLWCSLDAD
SFP4_BOVIN  AVFEGPACAFPFTYKGKKYYMCTRKNSVLLWCSLDTE
SP1_HORSE   AATDYAKCAFPFVYRGQTYDRCTTDGSLFRISWCSVT
COG2_CHICK  GNSEGAPCVFPFIFLGNKYDSCTSAGRNDGKLWCAST
COG2_HUMAN  GNSEGAPCVFPFTFLGNKYESCTSAGRSDGKMWCATT
COG2_MOUSE  GNSEGAPCVFPFTFLGNKYESCTSAGRNDGKVWCATT
COG2_RABIT  GNSEGAPCVFPFTFLGNKYESCTSAGRSDGKMWCATS
COG2_RAT    GNSEGAPCVFPFTFLGNKYESCTSAGRNDGKVWCATT
COG9_BOVIN  GNADGKPCVFPFTFQGRTYSACTSDGRSDGYRWCATT
COG9_HUMAN  GNADGKPCQFPFIFQGQSYSACTTDGRSDGYRWCATT
COG9_MOUSE  GNGEGKPCVFPFIFEGRSYSACTTKGRSDGYRWCATT
COG9_RAT    GNGDGKPCVFPFIFEGHSYSACTTKGRSDGYRWCATT
FINC_BOVIN  GNSNGALCHFPFLYNNHNYTDCTSEGRRDNMKWCGTT
FINC_HUMAN  GNSNGALCHFPFLYNNHNYTDCTSEGRRDNMKWCGTT
FINC_RAT    GNSNGALCHFPFLYSNRNYSDCTSEGRRDNMKWCGTT
MPRI_BOVIN  ETEDGEPCVFPFVFNGKSYEECVVESRARLWCATTAN
MPRI_HUMAN  ETDDGVPCVFPFIFNGKSYEECIIESRAKLWCSTTAD
PA2R_BOVIN  GNAHGTPCMFPFQYNQQWHHECTREGREDNLLWCATT
PA2R_RABIT  GNAHGTPCMFPFQYNHQWHHECTREGRQDDSLWCATT
```

BLOSUM score = log2(observed pair freq/expected pair freq)

IMPORTANT: This means that BLOSUM is **not** position specific – it is a kind of an averaged across all alignment positions.

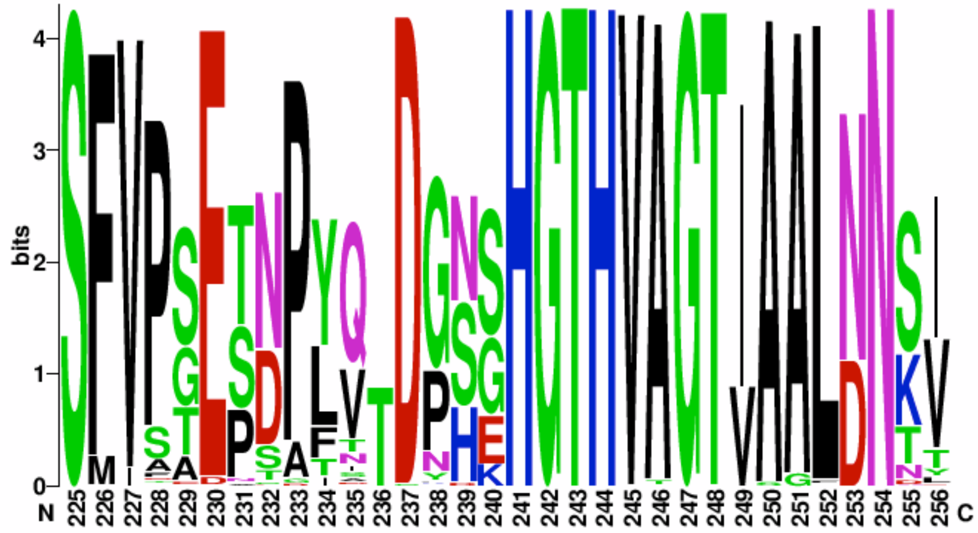
Idea: merge BLOSUM and WMs

- Pairwise alignment:
  - Alignment score = sum(BLOSUM(for each AA pair))
  - + penalty for gaps
  - IMPORTANT: 2 sequences

- Weight matrix:
  - WM score = sum(WM_score(for each AA, for each position))
  - IMPORTANT: single sequence

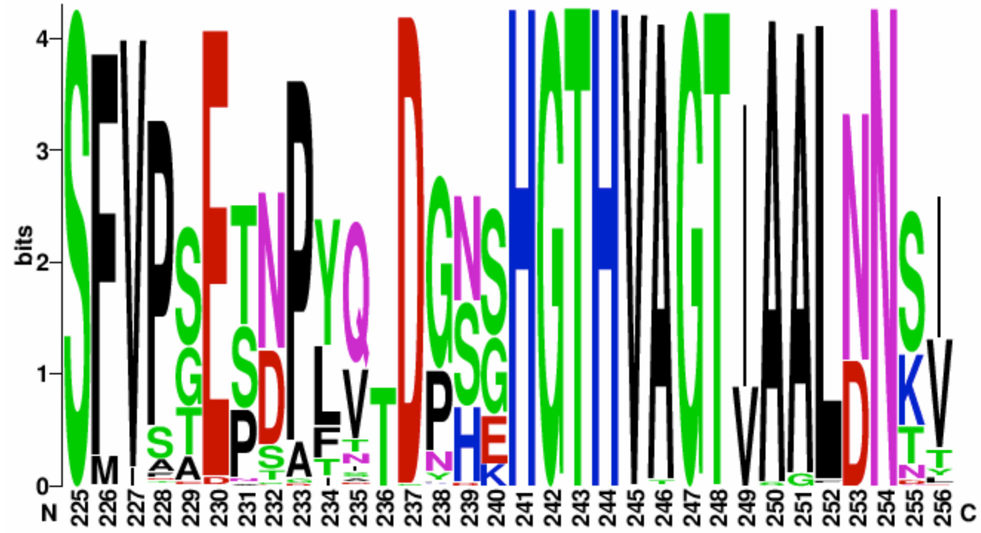
Idea: merge BLOSUM and WMs

- “New BLOSUM”:
  - Use protein family data set to estimate AA pair frequencies **per position**.
  - We need to apply the **pseudo-count** approach to account for AAs we do not observe.



Idea: merge BLOSUM and WMs

- "New alignment":
  - Look up alignment score per position
  - Sum up score + penalize for gaps the usual way

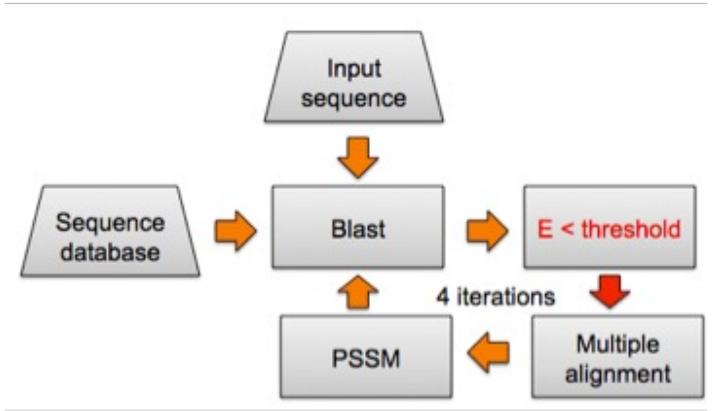


Part 3

HOW PSI-BLAST ACTUALLY WORKS

PSI-BLAST

- **P**osition-**S**pecific **I**terative BLAST
- Start with one sequence (as with BLASTP)
- Build protein family model on the fly:
  - **Step 0:** Start with an alignment model build purely on BLOSUM 62*
  - **Step 1:** Find set of related sequences
  - **Step 2:** Build refined **position specific** alignment model based on the identified related sequences
  - **Step 3:** Re-**iterate** step 1-2 until model does not improve anymore (in practice 3-4 iterations)

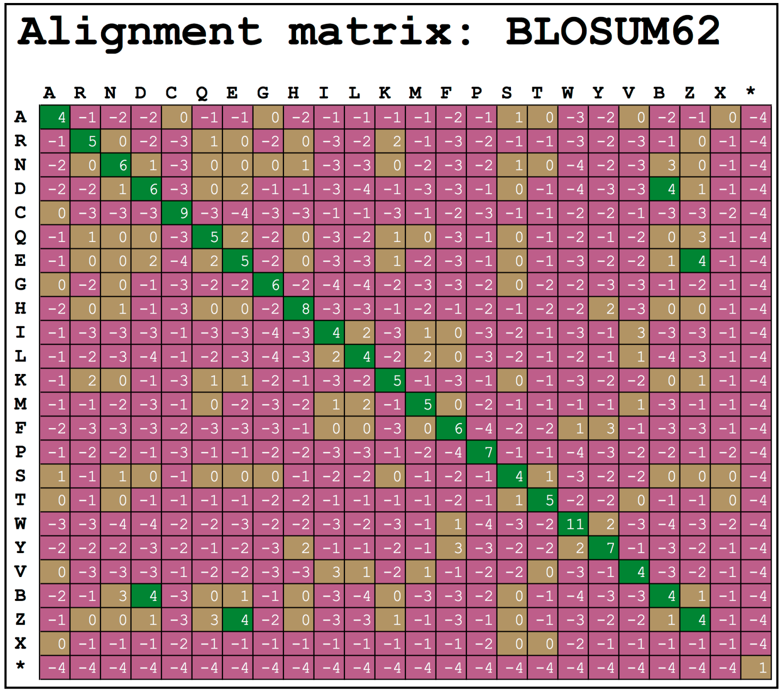


*The NCBI server actually "cheats" a bit here and just run BLASTP in step 0 for speed reasons

PSSM

- PSSM (pronounced “**P**o**SS**o**M**”):
  - **P**osition-**S**pecific **S**coring **M**atrix

- Start by creating a n*20 matrix
  - n = length of input sequence
- For each AA in the input sequence look up the corresponding row in the BLOSUM62 matrix and copy in the values

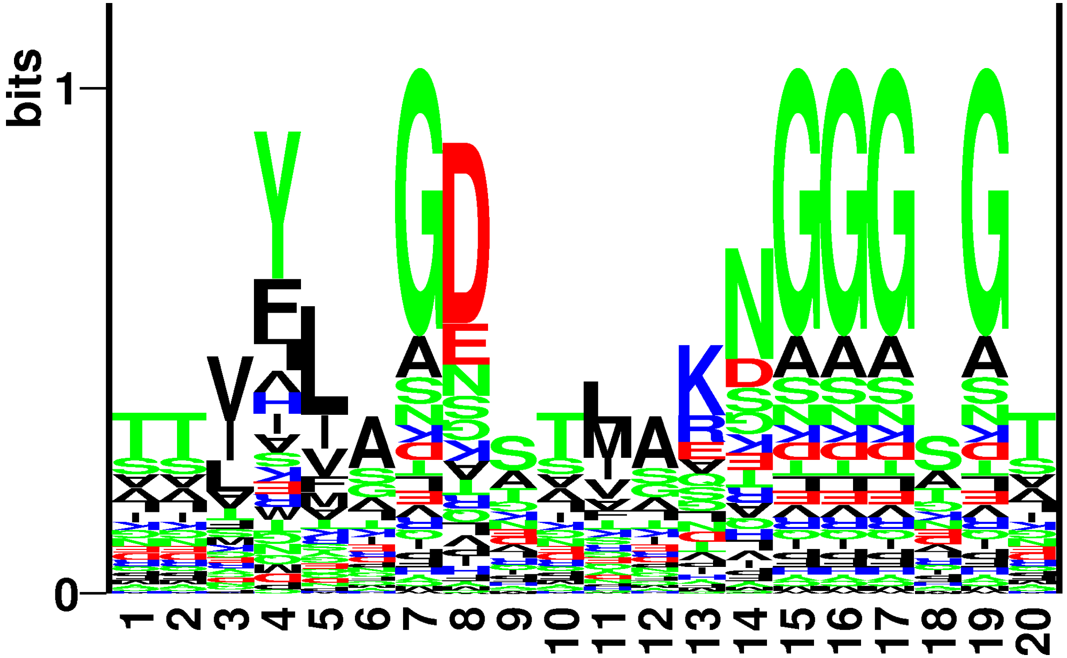


Alignment matrix: BLOSUM62

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

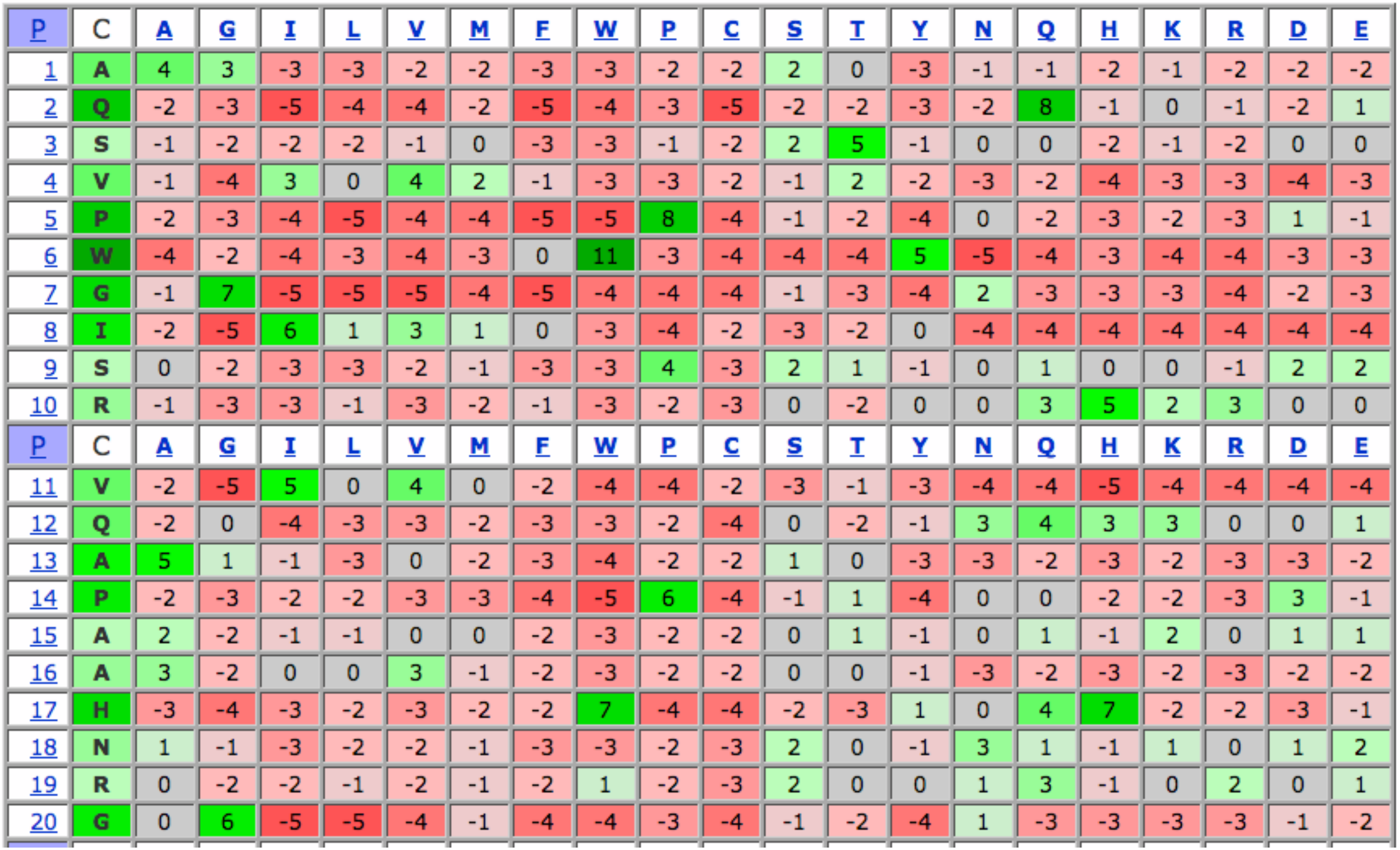
PSSM visualization

- Trick:
  - The PSSM can be visualized as a LOGO plot
  - Here’s what it can look like initially (after the trivial seeding with BLOSUM62):



PSSM adjusted after each iteration

Seed: Savinase (p29600) – database: NR

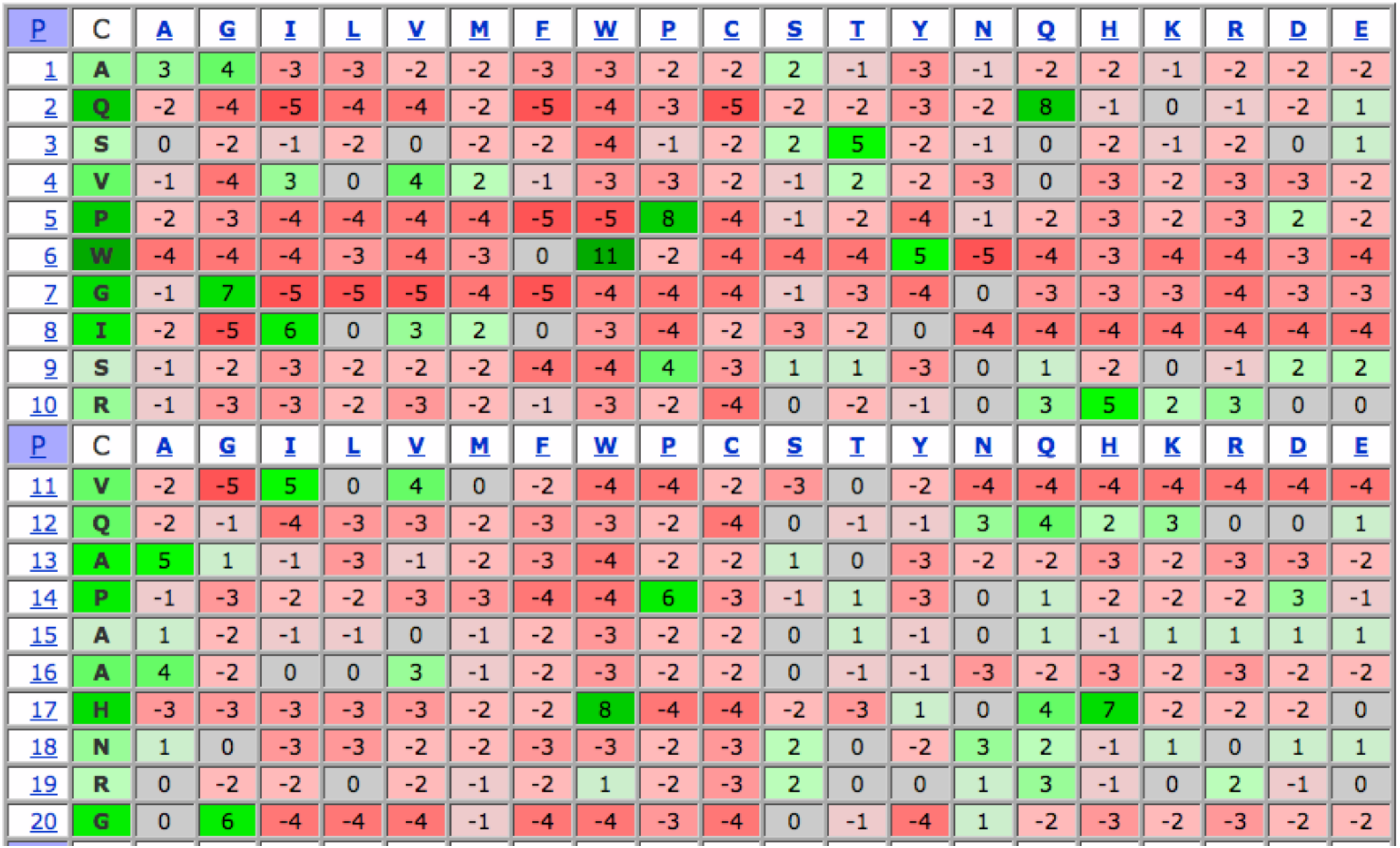


P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	A	4	3	-3	-3	-2	-2	-3	-3	-2	-2	2	0	-3	-1	-1	-2	-1	-2	-2	-2
2	Q	-2	-3	-5	-4	-4	-2	-5	-4	-3	-5	-2	-2	-3	-2	8	-1	0	-1	-2	1
3	S	-1	-2	-2	-2	-1	0	-3	-3	-1	-2	2	5	-1	0	0	-2	-1	-2	0	0
4	V	-1	-4	3	0	4	2	-1	-3	-3	-2	-1	2	-2	-3	-2	-4	-3	-3	-4	-3
5	P	-2	-3	-4	-5	-4	-4	-5	-5	8	-4	-1	-2	-4	0	-2	-3	-2	-3	1	-1
6	W	-4	-2	-4	-3	-4	-3	0	11	-3	-4	-4	-4	5	-5	-4	-3	-4	-4	-3	-3
7	G	-1	7	-5	-5	-5	-4	-5	-4	-4	-4	-1	-3	-4	2	-3	-3	-3	-4	-2	-3
8	I	-2	-5	6	1	3	1	0	-3	-4	-2	-3	-2	0	-4	-4	-4	-4	-4	-4	-4
9	S	0	-2	-3	-3	-2	-1	-3	-3	4	-3	2	1	-1	0	1	0	0	-1	2	2
10	R	-1	-3	-3	-1	-3	-2	-1	-3	-2	-3	0	-2	0	0	3	5	2	3	0	0
P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	V	-2	-5	5	0	4	0	-2	-4	-4	-2	-3	-1	-3	-4	-4	-5	-4	-4	-4	-4
12	Q	-2	0	-4	-3	-3	-2	-3	-3	-2	-4	0	-2	-1	3	4	3	3	0	0	1
13	A	5	1	-1	-3	0	-2	-3	-4	-2	-2	1	0	-3	-3	-2	-3	-2	-3	-3	-2
14	P	-2	-3	-2	-2	-3	-3	-4	-5	6	-4	-1	1	-4	0	0	-2	-2	-3	3	-1
15	A	2	-2	-1	-1	0	0	-2	-3	-2	-2	0	1	-1	0	1	-1	2	0	1	1
16	A	3	-2	0	0	3	-1	-2	-3	-2	-2	0	0	-1	-3	-2	-3	-2	-3	-2	-2
17	H	-3	-4	-3	-2	-3	-2	-2	7	-4	-4	-2	-3	1	0	4	7	-2	-2	-3	-1
18	N	1	-1	-3	-2	-2	-1	-3	-3	-2	-3	2	0	-1	3	1	-1	1	0	1	2
19	R	0	-2	-2	-1	-2	-1	-2	1	-2	-3	2	0	0	1	3	-1	0	2	0	1
20	G	0	6	-5	-5	-4	-1	-4	-4	-3	-4	-1	-2	-4	1	-3	-3	-3	-3	-1	-2

After iteration 2

PSSM adjusted after each iteration

Seed: Savinase (p29600) – database: NR

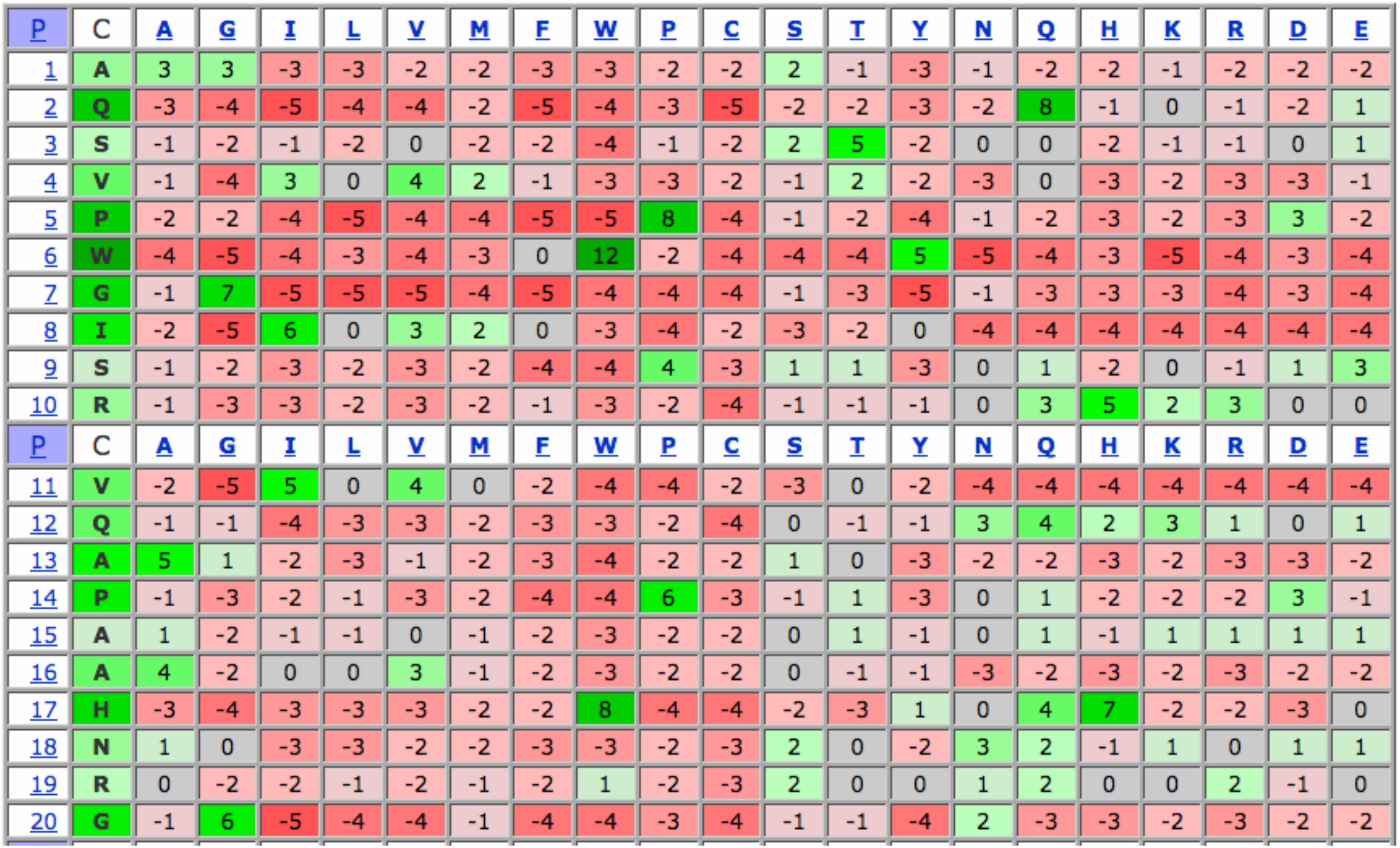


P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	A	3	4	-3	-3	-2	-2	-3	-3	-2	-2	2	-1	-3	-1	-2	-2	-1	-2	-2	-2
2	Q	-2	-4	-5	-4	-4	-2	-5	-4	-3	-5	-2	-2	-3	-2	8	-1	0	-1	-2	1
3	S	0	-2	-1	-2	0	-2	-2	-4	-1	-2	2	5	-2	-1	0	-2	-1	-2	0	1
4	V	-1	-4	3	0	4	2	-1	-3	-3	-2	-1	2	-2	-3	0	-3	-2	-3	-3	-2
5	P	-2	-3	-4	-4	-4	-4	-5	-5	8	-4	-1	-2	-4	-1	-2	-3	-2	-3	2	-2
6	W	-4	-4	-4	-3	-4	-3	0	11	-2	-4	-4	-4	5	-5	-4	-3	-4	-4	-3	-4
7	G	-1	7	-5	-5	-5	-4	-5	-4	-4	-4	-1	-3	-4	0	-3	-3	-3	-4	-3	-3
8	I	-2	-5	6	0	3	2	0	-3	-4	-2	-3	-2	0	-4	-4	-4	-4	-4	-4	-4
9	S	-1	-2	-3	-2	-2	-2	-4	-4	4	-3	1	1	-3	0	1	-2	0	-1	2	2
10	R	-1	-3	-3	-2	-3	-2	-1	-3	-2	-4	0	-2	-1	0	3	5	2	3	0	0
11	V	-2	-5	5	0	4	0	-2	-4	-4	-2	-3	0	-2	-4	-4	-4	-4	-4	-4	-4
12	Q	-2	-1	-4	-3	-3	-2	-3	-3	-2	-4	0	-1	-1	3	4	2	3	0	0	1
13	A	5	1	-1	-3	-1	-2	-3	-4	-2	-2	1	0	-3	-2	-2	-3	-2	-3	-3	-2
14	P	-1	-3	-2	-2	-3	-3	-4	-4	6	-3	-1	1	-3	0	1	-2	-2	-2	3	-1
15	A	1	-2	-1	-1	0	-1	-2	-3	-2	-2	0	1	-1	0	1	-1	1	1	1	1
16	A	4	-2	0	0	3	-1	-2	-3	-2	-2	0	-1	-1	-3	-2	-3	-2	-3	-2	-2
17	H	-3	-3	-3	-3	-3	-2	-2	8	-4	-4	-2	-3	1	0	4	7	-2	-2	-2	0
18	N	1	0	-3	-3	-2	-2	-3	-3	-2	-3	2	0	-2	3	2	-1	1	0	1	1
19	R	0	-2	-2	0	-2	-1	-2	1	-2	-3	2	0	0	1	3	-1	0	2	-1	0
20	G	0	6	-4	-4	-4	-1	-4	-4	-3	-4	0	-1	-4	1	-2	-3	-2	-3	-2	-2

After iteration 3

PSSM adjusted after each iteration

Seed: Savinase (p29600) – database: NR

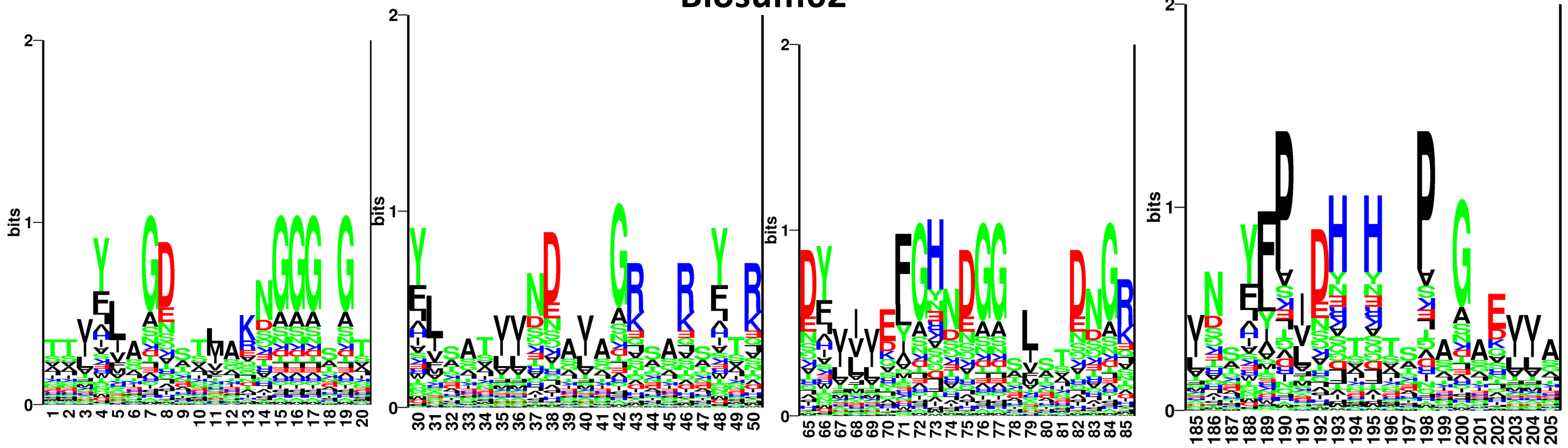


P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	A	3	3	-3	-3	-2	-2	-3	-3	-2	-2	2	-1	-3	-1	-2	-2	-1	-2	-2	-2
2	Q	-3	-4	-5	-4	-4	-2	-5	-4	-3	-5	-2	-2	-3	-2	8	-1	0	-1	-2	1
3	S	-1	-2	-1	-2	0	-2	-2	-4	-1	-2	2	5	-2	0	0	-2	-1	-1	0	1
4	V	-1	-4	3	0	4	2	-1	-3	-3	-2	-1	2	-2	-3	0	-3	-2	-3	-3	-1
5	P	-2	-2	-4	-5	-4	-4	-5	-5	8	-4	-1	-2	-4	-1	-2	-3	-2	-3	3	-2
6	W	-4	-5	-4	-3	-4	-3	0	12	-2	-4	-4	-4	5	-5	-4	-3	-5	-4	-3	-4
7	G	-1	7	-5	-5	-5	-4	-5	-4	-4	-4	-1	-3	-5	-1	-3	-3	-3	-4	-3	-4
8	I	-2	-5	6	0	3	2	0	-3	-4	-2	-3	-2	0	-4	-4	-4	-4	-4	-4	-4
9	S	-1	-2	-3	-2	-3	-2	-4	-4	4	-3	1	1	-3	0	1	-2	0	-1	1	3
10	R	-1	-3	-3	-2	-3	-2	-1	-3	-2	-4	-1	-1	-1	0	3	5	2	3	0	0
P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	V	-2	-5	5	0	4	0	-2	-4	-4	-2	-3	0	-2	-4	-4	-4	-4	-4	-4	-4
12	Q	-1	-1	-4	-3	-3	-2	-3	-3	-2	-4	0	-1	-1	3	4	2	3	1	0	1
13	A	5	1	-2	-3	-1	-2	-3	-4	-2	-2	1	0	-3	-2	-2	-3	-2	-3	-3	-2
14	P	-1	-3	-2	-1	-3	-2	-4	-4	6	-3	-1	1	-3	0	1	-2	-2	-2	3	-1
15	A	1	-2	-1	-1	0	-1	-2	-3	-2	-2	0	1	-1	0	1	-1	1	1	1	1
16	A	4	-2	0	0	3	-1	-2	-3	-2	-2	0	-1	-1	-3	-2	-3	-2	-3	-2	-2
17	H	-3	-4	-3	-3	-3	-2	-2	8	-4	-4	-2	-3	1	0	4	7	-2	-2	-3	0
18	N	1	0	-3	-3	-2	-2	-3	-3	-2	-3	2	0	-2	3	2	-1	1	0	1	1
19	R	0	-2	-2	-1	-2	-1	-2	1	-2	-3	2	0	0	1	2	0	0	2	-1	0
20	G	-1	6	-5	-4	-4	-1	-4	-4	-3	-4	-1	-1	-4	2	-3	-3	-2	-3	-2	-2

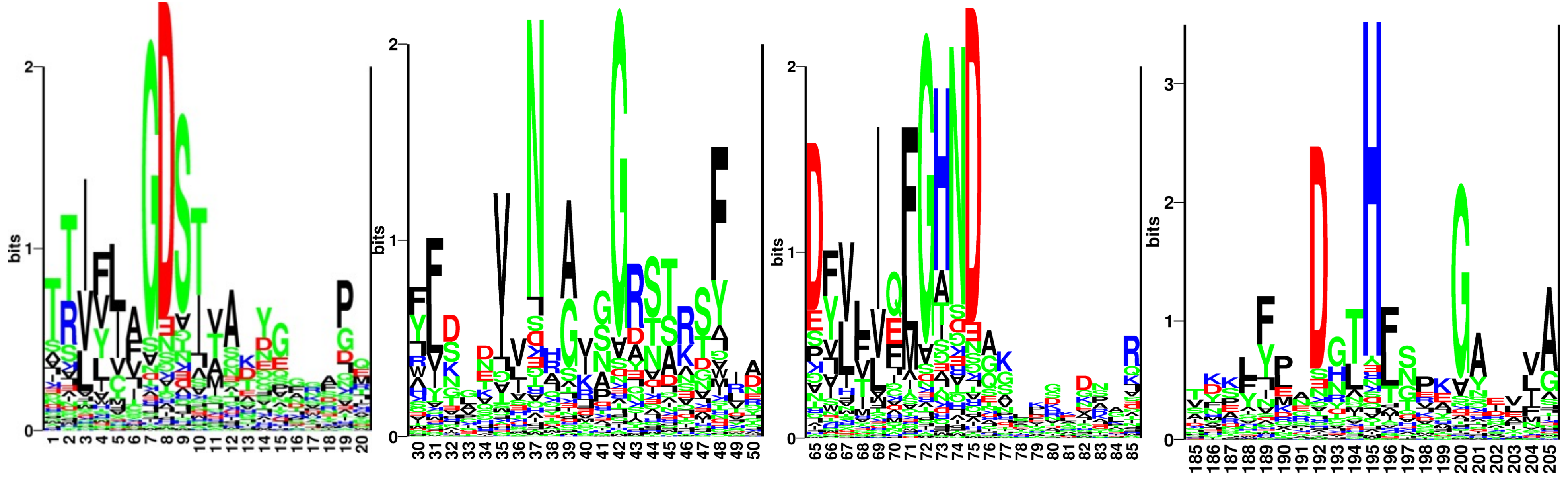
After iteration 4

Example (SGNH active site)

Blosum62



PSSM



Saving a PSSM for later use

- Very important:
  - The PSSM you have arrived at after all your iterations can be saved for later use
- Uses:
  - **Scenario 1:** Visualize your PSSM to assess the patterns picked up.
  - **Scenario 2:** Run your search again (perhaps ½ year later) without having to go through all the iterations.
  - **Scenario 3:** Search **a different database** using your PSSM
    - For example: train a rock solid PSSM for detecting prokaryotic serine peptidases on the big “NR” database, then save it and use it to hunt for human/mouse remote homologs.
    - You’ll HAVE TO do it this way, as it’s highly unlikely to find sufficiently good homologs to build the model in the restricted data set.

PSI-BLAST summary

- Is much better at finding remote homologs compared to BLASTP
  - If used correctly!
  - Remember to build your PSSM on the best possible data set, and potentially re-apply it in the actual data set you want to search
- Great for building data sets of related sequences
  - In the NCBI interface you can save all found sequences as a single pre-aligned multi FASTA file