Multiple Sequence Alignment

2020/10/27

1

What is multiple sequence alignment (MSA)?

2020/10/27

To align 3 or more related sequences to achieve optimal matching of the sequences.

Example:

```
seq1 CASVC
seq2 CTSIC
seq3 CAGVC
seq4 CTGIC
     *  :*
```

2

How do we find related sequences that are appropriate to align in a MSA?

2020/10/27

By database similarity searching.

Example:

```
query VKQDFDRMRYQGTWYAV      query VKQDFDRMRYQGTWYAV
seq_1 VQQDFNRTRYQGTWYAV      seq_2 VMQNFDRSRYTGRWYAV
```

MSA:

```
query VKQDFDRMRYQGTWYAV
seq_1 VQQDFNRTRYQGTWYAV
seq_2 VMQNFDRSRYTGRWYAV
      * *.*.* **:*.****
```

3

Why is multiple sequence alignment important?

2020/10/27

- Reveals more biological information than many pairwise alignments:
  - Allows identification of sequence motifs in the whole protein family.
  - Conserved and functionally critical amino acid residues can be identified.

```
seq1 CASVC
seq2 CTSIC
seq3 CAGVC
seq4 CTGIC
```

- Provides the basis for the most sensitive sequence database searching algorithm– PSI-BLAST– for detecting distantly related homologs.
- It is an essential prerequisite for phylogenetic analysis (which is the inference of evolutionary relationships).
- It is an essential prerequisite for prediction of protein structure.

4

Practical uses of multiple sequence alignment:

Predict the function of a newly sequenced protein based on similarity to other members of a family. (Most proteins have been identified by sequencing of genomic DNA or cDNA; functions are assigned to these proteins based on homology, rather than results of biochemical or cell biological assays.)

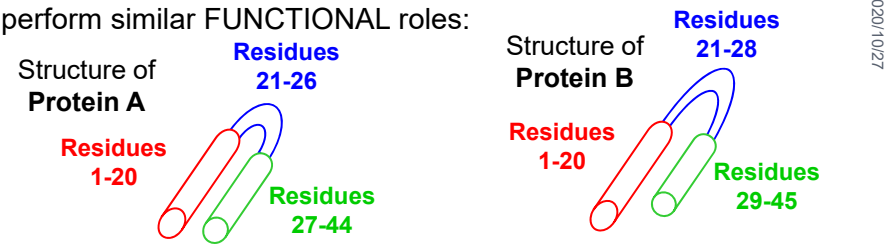
- Predict the structure of a protein based on similarities to other members of a family.
- Find discrepancies in sequences of cDNA clones.
- Design degenerate PCR primers.
- Identify the consensus sequence for gene regulatory regions.

2020/10/27

5

The Best Way to Compare Two Proteins

Comparing protein STRUCTURES allows us to determine which residues are in similar positions in 3D space and, hence, are likely to perform similar FUNCTIONAL roles:



2020/10/27

We can then generate a SEQUENCE alignment by comparing STRUCTURES. The extra information from 3D structures allows for a more reliable alignment of residues that are functionally equivalent:

```
Sequence alignment based on structural comparison:
ProtA: xxxxxxxxxxxxxxxxxxxxxxx--xxxxxxxxxxxxxxxxxxxx
ProtB: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx-
```

6

BUT– only a small fraction of proteins have known structures, so a sequence alignment based on structures usually isn't possible. Alignment methods must then rely solely on comparison of SEQUENCES.

2020/10/27

(A sequence alignment generated by comparing STRUCTURES is the "gold standard" against which sequence alignment algorithms are judged.)

A sequence alignment based on STRUCTURES is the most important type of alignment for predicting FUNCTION.

HOWEVER, this type of alignment does not necessarily correspond to the evolutionary alignment implied by divergence from a common ancestor protein. (More info when we cover phylogenetic analysis.)

7

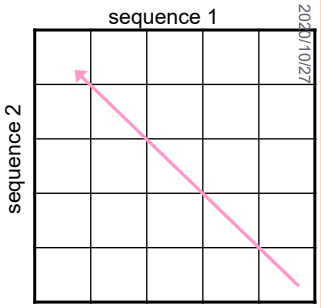
A quick review of dynamic programming for pairwise sequence comparison.

Aligning two sequences of 5 residues each using dynamic programming:

Each cell in the matrix contains a score for two aligned residues, one from each sequence.

5^2 comparisons are made (excluding gaps), so there are 5^2 cells (25) in the matrix.

The pink arrow shows how one obtains the best alignment by tracing back from the final cell in the matrix.



2020/10/27

8

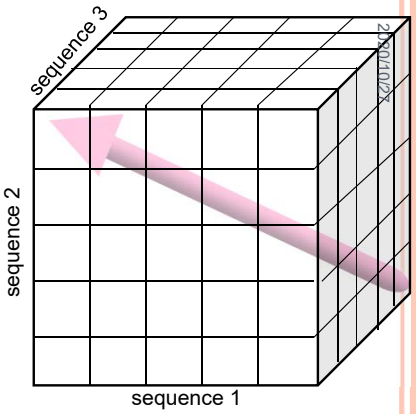
The dynamic programming algorithm can be used for multiple sequence alignment.

Aligning three sequences of 5 residues each using dynamic programming:

Each cell/cube in the matrix contains a score for three aligned residues, one from each sequence.

5^3 comparisons are made (excluding gaps), so there are 5^3 cells (125) in the matrix.

The pink arrow shows how one obtains the best alignment by tracing back from the final cell in the matrix.



Aligning N sequences would require an N-dimensional matrix and 5^N comparisons.

9

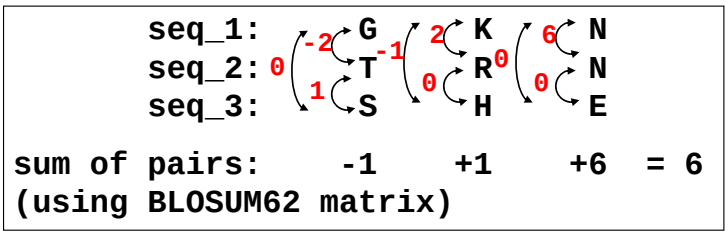
2020/10/27

How are scores calculated for positions in the matrix?

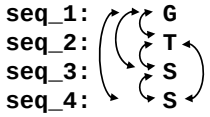
Sum of pairs method– the value in a particular matrix cell is the sum of the scores of every possible pair of residues at that position. (This is the simplest method of scoring.)

Example:

```
seq_1:  GKN
seq_2:  TRN
seq_3:  SHE
```



Sum of pairs for 4 sequences:



10

2020/10/27

Full dynamic programming is not practical for multiple sequence alignment.

Suppose we want to align 10 sequences, each of length 300 residues:

need a 10-dimensional matrix with 300 cells in each dimension; 300^{10} comparisons will be made (which is 5.9 X 10^{24}).

2020/10/27

Computation time and memory space required increase exponentially with the number of sequences.

MSA by Carrillo & Lipman (1988) is a multiple sequence alignment program that uses the dynamic programming algorithm. However, it does not calculate scores for all cells in the matrix– uses a shortcut to determine which cells are worth considering.

It can optimally align 5-7 protein sequences of reasonable length (200-300 residues).

11

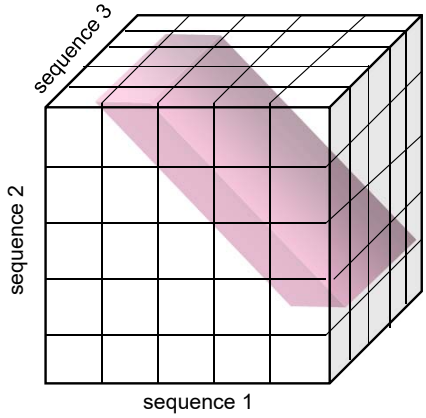
Divide-and-Conquer Multiple Sequence Alignment (DCA) is a program for producing fast, high-quality multiple sequence alignments; it uses a semi-exhaustive algorithm.

2020/10/27

Available at: http://bibiserv.techfak.uni-bielefeld.de/dca/

12

Heuristic algorithms speed up the alignment process by taking shortcuts to limit the space within the matrix within which optimal alignments are likely to be found:



2020/10/27

Example: A heuristic algorithm may determine that optimal alignments are likely to be found only in the pink region. Therefore, only these cells in the matrix are filled in. This saves time.

Heuristic algorithms are not guaranteed to find the mathematically optimal alignment for a set of sequences.

HOWEVER, the mathematically optimal alignment rarely makes more biological sense than alignments produced by good heuristic methods!

13

Categories of Heuristic Algorithms for Multiple Sequence Alignment:

2020/10/27

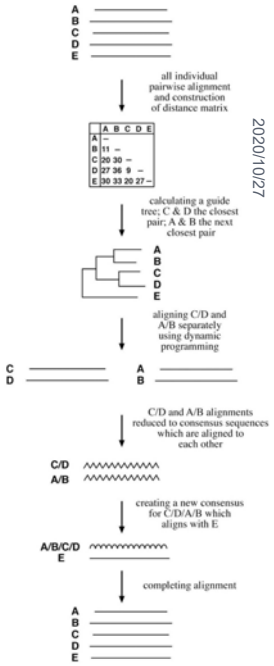
1. **Progressive** alignment methods
2. **Iterative** alignment methods

14

Progressive Alignment Methods

- Most are based on the progressive alignment algorithm published by Da-Fei Feng & Russell Doolittle in 1987.
- How they work: first calculate a pairwise alignment for the two most closely related sequences and <u>progressively</u> add more sequences to the alignment.

(Left figure shows the idea)



2020/10/27

Steps in the Feng-Doolittle Progressive Alignment Method

(Other progressive alignment algorithms differ slightly in the details of these steps.)

1. Perform pairwise alignments between ALL possible pairs of sequences. Use Needleman-Wunsch global dynamic programming algorithm. Calculate a similarity score for each pairwise alignment.

2020/10/27

Sequences To Be Aligned:
seq_1
seq_2
seq_3
seq_4
seq_5

→

Pairwise Alignment	Similarity Score
seq_1 & seq_2	84
seq_1 & seq_3	14
seq_1 & seq_4	8
seq_1 & seq_5	12
seq_2 & seq_3	17
seq_2 & seq_4	9
seq_2 & seq_5	19
seq_3 & seq_4	9
seq_3 & seq_5	27
seq_4 & seq_5	7

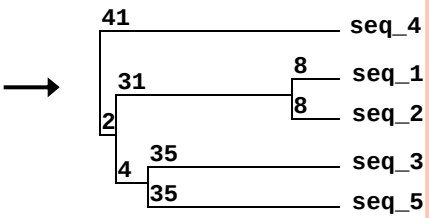
16

2. Use the similarity scores to calculate evolutionary distances; carry out a cluster analysis using these distances to construct a "guide tree." Some methods use neighbor-joining method to construct the tree; others use unweighted pair group method of arithmetic averages (UPGMA). (More on construction of trees later.) The guide tree will be used to determine the order in which the sequences will be progressively aligned.

2020/10/27

Pairwise Alignment	Similarity Score
seq_1 & seq_2	84
seq_1 & seq_3	14
seq_1 & seq_4	8
seq_1 & seq_5	12
seq_2 & seq_3	17
seq_2 & seq_4	9
seq_2 & seq_5	19
seq_3 & seq_4	9
seq_3 & seq_5	27
seq_4 & seq_5	7

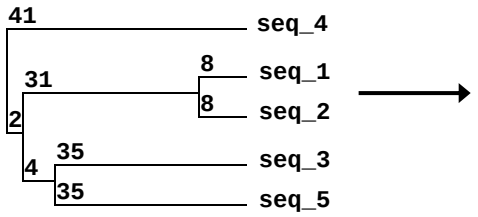
Guide Tree



The branching order of the tree reflects the relationships between sequences in terms of similarity; the branch lengths are proportional to evolutionary distance.

17

3. Select the two most closely related sequences from the guide tree and create a pairwise alignment.



Pairwise Alignment:

```
seq_1:  XXXXX
        |||||
seq_2:  XXXXX
```

2020/10/27

4. Align the next most similar sequence to the existing pairwise alignment. The new sequence is added by aligning it pairwise to each sequence in the group in turn. The highest scoring pairwise alignment determines how the new sequence will be aligned to the group.

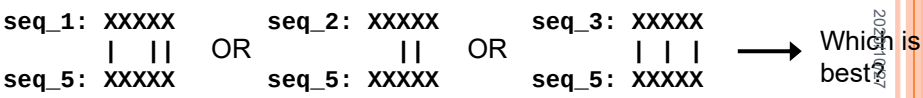
```
seq_1:  XXXXX          seq_2:  XXXXX
        || ||    OR            | ||
seq_3:  XXXXX          seq_3:  XXXXX
```

→ Which is better?

(A Note on Scoring: Each score from pairwise comparison of two residues may be multiplied by a factor based on the evolutionary relationships represented in the guide tree. Thus, the final sum-of-pairs score will be "weighted" so that more distantly related sequences have a greater contribution to the score.)

18

5. Align the next most similar sequence to the existing multiple sequence alignment. This is done in the same manner described in step 4.



6. Repeat step 5 until all sequences have been added to the alignment and a full multiple sequence alignment is obtained.

Multiple Sequence Alignment:

```
seq_1:  XXXXX
seq_2:  XXXXX
seq_3:  XXXXX
seq_5:  XXXXX
seq_4:  XXXXX
```

19

2020/10/27

CLUSTALW

- Probably the most widely used multiple sequence alignment program.
- Performs global multiple sequence alignment using the progressive method (neighbor-joining method for construction of guide tree).
- CLUSTALW is a newer and better version (1994) of the original CLUSTAL (1988). (Presumably, the name comes from the fact that building a guide tree is a type of "cluster" analysis.)
- Available online at EMBL-EBI: http://www.ebi.ac.uk/Tools/clustalw2/index.html
- CLUSTALX provides a window-based user interface for CLUSTALW. Available for download from: ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/

20

2020/10/27

Important features of CLUSTALW:

- Does not rely on a single scoring matrix:
  - Applies different matrices depending on degree of similarity between sequences.
  - For example, BLOSUM80 for initially aligned closely related sequences; BLOSUM50 later when more divergent sequences are added to the alignment.
- Weighting scheme used in scoring alignments:
  - Indicated by "W" in the name.
  - More distantly related sequences have a greater contribution; highly similar sequences don't dominate the alignment.
- Does not rely on a single set of gap penalties:
  - When sequences of structurally related proteins are aligned, gaps occur preferentially in loop regions between elements of secondary structure. Gaps are also more likely to occur next to some amino acids than others.
  - CLUSTALW uses this information to try to place gaps between conserved domains, rather than within these domains (conserved domains are likely to represent elements of secondary structure). The program does this by changing the initial gap penalties as more sequences are added to the alignment– e.g., penalties are decreased within stretches of certain amino acids that are more likely to be part of a loop region.

2020/10/27

21

Problems with progressive alignment methods:

- Errors made early on in the initial alignments will be propagated as new sequences are added to the alignment:
  - Gaps introduced in the initial alignments remain– "once a gap, always a gap." (Gaps remain because it is assumed that the two most closely related sequences, that are initially aligned, should be weighted most heavily in introducing gaps.)
- Difficult to choose a scoring matrix and gap penalties that are appropriate for the entire set of sequences:
  - Some sequences in the set may be very closely related while others are distantly related.
  - CLUSTALW addresses this problem because it does use multiple scoring matrices and changes the gap penalties.
- Only suitable for sequences of similar lengths since it is a global alignment method. (Truncating the sequences will sometimes solve this problem– you will see an example of this on the next homework assignment.)
- Slow– it's time-consuming to generate the first all-against-all pairwise alignments that are necessary for generating the guide tree:
  - $N(N-1)/2$ alignments must be generated.
  - Example: for 100 sequences, 4950 pairwise alignments must be made.

2020/10/27

22

T-Coffee:

Tree-based **C**onsistency **O**bjective **F**unction **F**or alignm**E**nt **E**valuation

T-Coffee addresses some of the shortcomings of CLUSTALW and other progressive alignment algorithms:

2020/10/27

1. Performs global and local pairwise alignments for all possible pairs of sequences.
2. The consistency of the global and local alignments is evaluated, and the pairwise alignments are refined.
3. A guide tree is derived from the pairwise alignments.
4. A full multiple sequence alignment is generated progressively based on the guide tree.

- Since T-Coffee optimizes the initial alignments, it minimizes the initial alignment errors which will be propagated to the final alignment.
- T-Coffee shown to outperform CLUSTALW for moderately divergent sequences, but it is much slower.

http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html

http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi

23

Categories of Heuristic Algorithms for Multiple Sequence Alignment:

2020/10/27

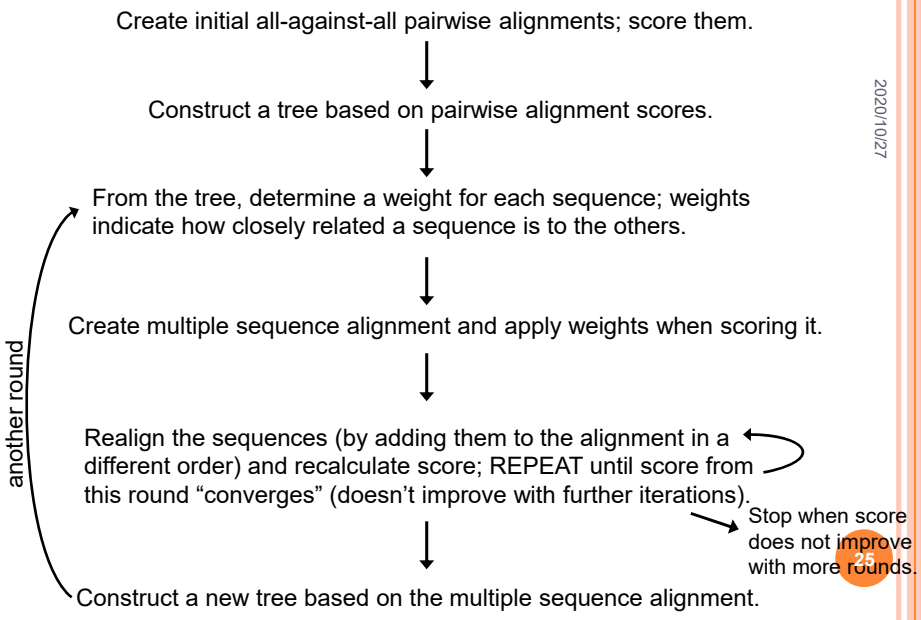
1. **Progressive** alignment methods
2. **Iterative** alignment methods

Iterative alignment methods attempt to correct a problem of progressive methods:

Errors in the initial alignments of the most closely related sequences are propagated to the final multiple sequence alignment.

24

Iterative Alignment Methods



2020/10/27

25

There are many programs available for multiple sequence alignment.

- All have their strengths and weaknesses; choice of a program will depend on various factors that are important to the user:
  - Speed
  - Ease of use
  - Characteristics of sequences being aligned (similar vs. different lengths; closely related vs. divergent)
- Best to try several methods on the same set of sequences.
- CLUSTALW is popular because it is easy to use and quickly provides an adequate alignment of a large number of closely related sequences.
- Another progressive method: PRALINE
- Two iterative methods: DIALIGN2, PRRN

2020/10/27

26

One application of multiple sequence alignment (MSA):
Using a multiple sequence alignment to identify related sequences / family members.

2020/10/27

- **Why can a MSA be used to find related sequences?**
  Embedded within a multiple sequence alignment is intrinsic sequence information that represents the common characteristics of that particular collection of sequences.

- **One very simple way to use a MSA to find related sequences:**
  Convert the multiple sequence alignment to a single "consensus" sequence, and use the consensus sequence to search databases.

```
seq_1:  AADK
seq_2:  AVDR
seq_3:  AAER
seq_4:  LADR
seq_5:  VADK
consensus:  AADR
```

27

Note that the consensus sequence isn't identical to any of the sequences from which it was derived, but it contains information derived from all of them.

BUT: a consensus sequence doesn't capture very much information about the sequences from which it was derived, so it wouldn't work well to identify related sequences in a database.

2020/10/27

```
seq_1:  AADK
seq_2:  AVDR
seq_3:  AAER
seq_4:  LADR
seq_5:  VADK
consensus:  AADR
```

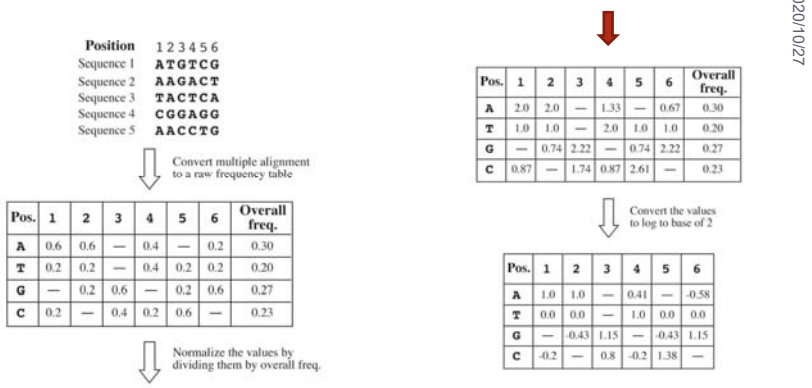
This position is assigned 'A', but how OFTEN does 'A' occur at this location, and which other amino acids can occur at this location???

THE SOLUTION: a position-specific scoring matrix (PSSM) contains much more information than a consensus sequence.

- A PSSM is a table/matrix that contains probability information for the residues at each position in an ungapped multiple sequence alignment.
  (Tells us the probability of finding 'A' at the first position, AND the probability of finding 'L' at first position, etc.)
- A PSSM can be used to search sequence databases.

28

An Example of Construction of PSSM



Position	1	2	3	4	5	6
Sequence 1	A	T	G	T	C	G
Sequence 2	A	A	G	A	C	T
Sequence 3	T	A	C	T	C	A
Sequence 4	C	G	G	A	G	G
Sequence 5	A	A	C	C	T	G

Convert multiple alignment to a raw frequency table

Pos.	1	2	3	4	5	6	Overall freq.
A	0.6	0.6	—	0.4	—	0.2	0.30
T	0.2	0.2	—	0.4	0.2	0.2	0.20
G	—	0.2	0.6	—	0.2	0.6	0.27
C	0.2	—	0.4	0.2	0.6	—	0.23

Normalize the values by dividing them by overall freq.

Pos.	1	2	3	4	5	6	Overall freq.
A	2.0	2.0	—	1.33	—	0.67	0.30
T	1.0	1.0	—	2.0	1.0	1.0	0.20
G	—	0.74	2.22	—	0.74	2.22	0.27
C	0.87	—	1.74	0.87	2.61	—	0.23

Convert the values to log to base of 2

Pos.	1	2	3	4	5	6
A	1.0	1.0	—	0.41	—	-0.58
T	0.0	0.0	—	1.0	0.0	0.0
G	—	-0.43	1.15	—	-0.43	1.15
C	-0.2	—	0.8	-0.2	1.38	—

2020/10/27

29

Interpreting a PSSM

Each value in the PSSM is the ratio of the observed counts of that amino acid at a given position in the multiple sequence alignment, divided by the expected count of that amino acid at that position in the alignment. This value is then converted to a log scale (usually base 2).

The values in the PSSM reflect the probability of any given amino acid occurring at each position in the alignment, and the effect of a conservative or nonconservative substitution at each position in the alignment.

Multiple sequence alignment:

```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
consensus: GPHIVVATPG
```

2020/10/27

Corresponding PSSM (explained on next slide):

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	17	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22
P	18	13	0	0	-22	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8
H	5	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7
I	-1	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8
V	3	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2
V	5	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0
A	54	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20
T	40	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30
P	31	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48
G	70	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70

30

Interpreting a PSSM

The residue 'G' occurs most frequently at position 1 in the alignment, so a 'G' is shown in position 1 of the consensus sequence.

The first row of the PSSM corresponds to position 1 of the alignment. Notice that the amino acid 'G' is given the highest number in this row: 31. This reflects the fact that G is the residue most likely to occur in position 1.

The 2nd row of the PSSM corresponds to position 2 in the alignment. The amino acid 'P' is given the highest number in this row, reflecting the fact that P is the residue most likely to occur in position 2. Etc.

Multiple sequence alignment:

```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
consensus: GPHIVVATPG
```

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	17	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22
P	18	13	0	0	-22	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8
H	5	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7
I	-1	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8
V	3	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2
V	5	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0
A	54	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20
T	40	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30
P	31	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48
G	70	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70

Interpreting a PSSM

Position 8 in the alignment is always 'T'.
Position 10 in the alignment is always 'G'.
The high values in the PSSM (150) reflect this.

Position 9 is almost always 'P'. The high value in the PSSM (89) reflects this, but the score is not as high as it would have been if 'P' always occurred at this position.

Amino acids with negative values are not likely to be found at that position in the multiple sequence alignment.

Multiple sequence alignment:

```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
consensus: GPHIVVATPG
```

Corresponding PSSM:

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	17	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22
P	18	13	0	0	-22	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8
H	5	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7
I	-1	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8
V	3	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2
V	5	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0
A	54	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20
T	40	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30
P	31	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48
G	70	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70

The information content in a PSSM:

A formal method called information theory can be used to describe the amount of information in each row in a PSSM.

2020/10/27

Consider the multiple sequence alignment here:

Columns 8 and 10 contain more information than any others because the amino acid at that position is conserved in each case. Column 9 also contains a lot of information, but not quite as much as 8 and 10.

Column 1 is highly variable, so it contains less information than the others.

```
12345678910
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

33

Suppose we want to know which amino acid belongs at a particular position in a multiple sequence alignment of a family of related proteins.

2020/10/27

The amount of "uncertainty" in the identity of the amino acid at that position is given by:

$-\log_2$ (probability of a certain amino acid)

$-\log_2 (1/20) = 4.32$ bits

An uncertainty of 4.32 bits means we have no information at all regarding the likelihood of one amino acid versus another at a given location. Suppose we aligned 20 different sequences and found that every amino acid occurred once at a given position in the alignment. The uncertainty at that position would be 4.32 bits (the maximum value).

The data in a PSSM provide information that reduces this uncertainty to values less than 4.32. If only 'T' occurs at a given position when we align different sequences, the uncertainty at that position is reduced to zero bits.

34

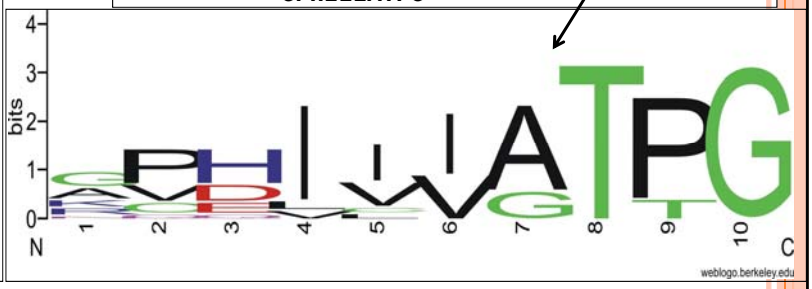
The higher the information content of a PSSM (the lower the uncertainty), the more useful the PSSM.

The information content (amount of uncertainty) in a PSSM can be represented by a graph called a **sequence logo** (shown below).

2020/10/27

```
12345678910
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

Get consensus sequence by reading letters across top of logo:
GPHIIIATPG



Each column in the logo represents a column in the multiple sequence alignment and in the PSSM. The amino acids shown in a given column of the logo are those that appear in that column in the alignment.

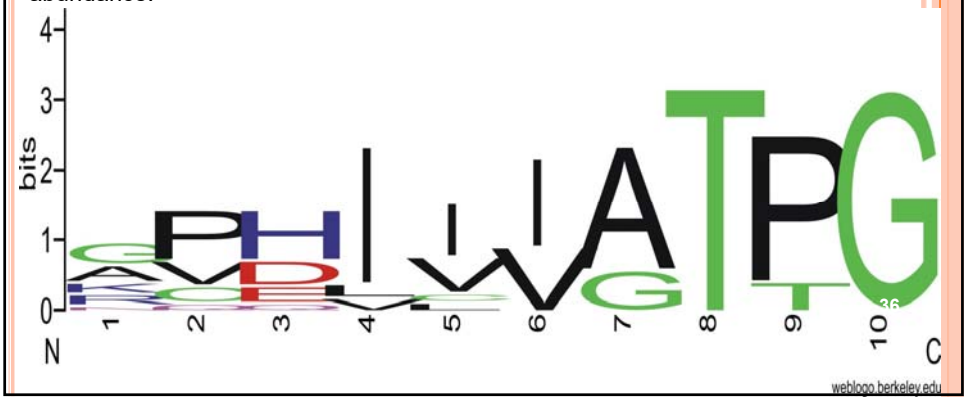
35

Interpreting a Sequence Logo

The **total height** of each column gives the decrease in uncertainty provided by the information content at that position in the PSSM. Taller columns contain the most information because the uncertainty has been reduced the most. The height of a column also reflects the diversity of amino acids that can occur at that position– taller columns for conserved positions.

2020/10/27

The **height of each amino acid** within a given column is proportional to its frequency of occurrence at that position in the PSSM. They are stacked in increasing order of abundance.



36

A "**profile**" is another type of matrix derived from a gapped multiple sequence alignment (PSSMs are derived from ungapped alignments). Therefore, a profile is simply a scoring matrix like a PSSM, but it also contains information regarding insertions and deletions in a sequence family.

Books and literature often use the terms 'profile' and 'PSSM' interchangeably, but technically a PSSM doesn't include information about gaps, while a profile does.

2020/10/27

37

How can profiles and PSSMs be used?

- To determine if a sequence of interest contains the sequence motif represented by the PSSM.

```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
consensus: GPHIVVATPG
```

```
New protein sequence:
XXXXXXXXXXXXXXXXXXXXXXXXXXX...
```

Does the new protein contain the sequence pattern (motif) that all the proteins in this multiple sequence alignment contain? The method for determining this is described on the next two slides.

- To search a sequence database to find new members of the family represented by the PSSM. (Described on subsequent slides)

2020/10/27

38

Determining if a sequence of interest contains the sequence motif represented by the PSSM: we can determine how well any sequence of the same length as the PSSM "fits" the PSSM.

How is this done?

Example: Is the sequence **'GGDIVVGTPG'** another example of the sequence motif represented by this PSSM?

2020/10/27

- Score this new sequence using the PSSM (follow red #s in PSSM):
  31 + 13 + 29 + 63 + 50 + 58 + 44 + 150 + 89 + 150 = 677
- Calculate the probability of the motif occurring in the sequence:
  $2^{677} = 6.27 \times 10^{203}$
  The sequence is 6.27×10^{203} times more likely to fit the motif because it is related to the other sequences, than to fit by chance.

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	17	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22
P	18	13	0	0	-22	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8
H	5	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7
I	-1	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8
V	3	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2
V	5	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0
A	54	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20
T	40	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30
P	31	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48
G	70	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70

39

Determining if a sequence of interest contains the sequence motif represented by the PSSM:

Suppose we want to determine if a new protein contains the sequence motif, but we don't know the location of the motif in the new protein sequence.

2020/10/27

Use the PSSM to "scan" the new protein's sequence:

X X X X X X X X X X...→score1	Calculate score for occurrence of motif beginning at residue 1
X **X X X X X** X X X X...→score2	Calculate score for occurrence of motif beginning at residue 2
⋮ Continue scanning until end of sequence is reached.	
...X X X X X **X X X X X** →scoreN	Calculate score for occurrence of motif at last possible position

The highest scoring location is the most likely position of the motif in the sequence. The significance of the score can be evaluated by calculating the value of 2^{score} (see previous slide).

40

To search a sequence database to find new members of the family represented by the PSSM:

- The PSSM is used to scan each sequence in the database using the method just described:

2020/10/27

X X X X X X X X X X...→score1	Calculate score for occurrence of motif beginning at residue 1
X X X X X X X X X X X...→score2	Calculate score for occurrence of motif beginning at residue 2
⋮ Continue scanning until end of sequence is reached.	
...X X X X X X X X X X →scoreN	Calculate score for occurrence of motif at last possible position

- Advantage of searching a database with a PSSM rather than a single member of the family: **more sensitive**– will be able to identify additional family members that might be missed if using a single sequence to search the database.

41

- PSI-BLAST is one program that does this.

PSI-BLAST: Position-Specific-Iterated BLAST (ψ-BLAST)

Steps of the algorithm:

2020/10/27

1. Query a protein database using a single protein sequence (same as a BLASTP search).

2. Any related sequences found in the search (those with E-values below a cutoff / threshold) are aligned with the initial query sequence to create a multiple sequence alignment, and a profile is created from this alignment.

3. The profile is used to again search the database– so the search has been expanded to include sequences that match the variations found in the multiple sequence alignment at each sequence position.

4. Any newly discovered sequences (hits from the search) that are similar to those used to create the profile are then added to the multiple sequence alignment and an updated/refined profile is created.

5. The refined profile is used to search the database again to find more distantly related sequences. Steps 4 and 5 are repeated until no new sequences are found, or until the user chooses to stop the iterations (usually it is best to stop after a few iterations).

42

PSI-BLAST: Position-Specific-Iterated BLAST

PROS:

- Easy to use.
- Fast.

CONS:

- Determining the significance of the alignments found–
  The sequences found may be distantly related family members, or matches may just be due to chance.
- Newly found sequences that match the query influence the finding of more sequences like themselves–
  There is no guarantee that subsequently discovered sequences are in the same family.

2020/10/27

43

The usefulness of a PSSM or profile depends on the sequences from which it was derived:

If the set of sequences is small, the values in the PSSM / profile may not adequately represent all sequences in the family.

2020/10/27

44