Protein Motifs and Domains

2021/11/16

1

Protein motifs and domains are consensus sequence patterns.

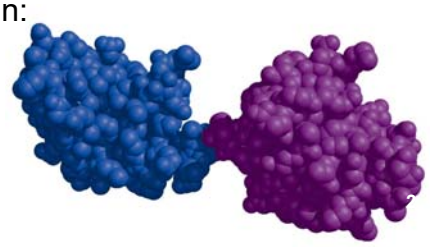
2021/11/16

Motif– a short conserved sequence pattern; can be just a few amino acid residues, up to ~20.

Examples: **Y-X-Y** and **C-X_4-C-X_{12}-H-X_3-H**

Domain– a longer conserved sequence pattern which adopts a particular three-dimensional structure and is an independent functional and structural unit; typically 40-700 residues.

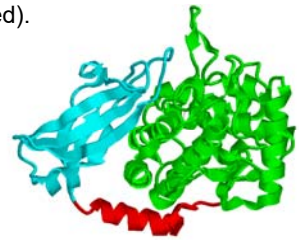
Example of a two-domain protein:



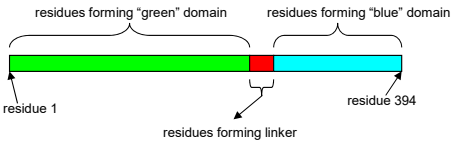
This protein (troponin C) is composed of a single amino acid chain, but each half of the chain forms an independent structural and functional unit– a **domain**.

More examples of protein domains

This protein (hemocyanin) has two distinct domains (blue and green) which are connected by a short linker (red).

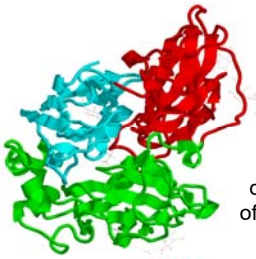


The amino acid chain of hemocyanin can be represented like this:

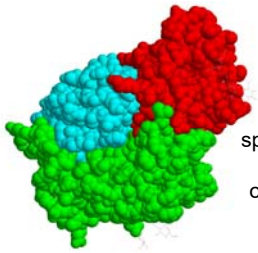


This enzyme (laccase) has three distinct domains (each colored differently).

2021/11/16



ribbon diagram of laccase



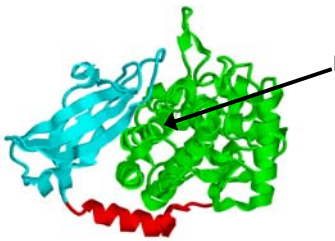
space-filling diagram of laccase

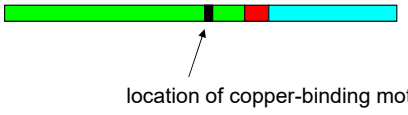
Domains may contain motifs. Hemocyanin as an example:

The "green" domain of hemocyanin contains this copper-binding motif:

$$H\text{-}X_3\text{-}H\text{-}X_{22\text{-}37}\text{-}H$$

2021/11/16



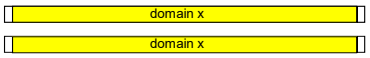


4

A group of proteins that share a domain in common constitute a FAMILY. Family members are evolutionarily related (homologous) and their domains have sequence similarity.

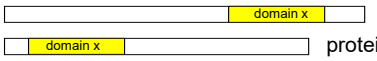
2021/11/16

Family members can share a domain in common in a number of ways:



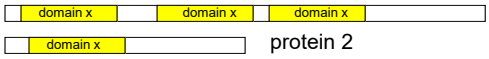
protein 1
protein 2

A domain may extend essentially across the length of a protein.



protein 1
protein 2

Domains may contain highly related stretches of amino acids that form only a subset of each protein's sequence.



protein 1
protein 2

A domain may be repeated within a single protein.

5

(Figure 8.2 from *Bioinformatics and Functional Genomics* by J. Pevsner)

NOTE: Many short motifs are NOT specific to a particular protein family. Thus, their occurrence does not indicate homology.

2021/11/16

Example:
protein kinase C phosphorylation site has this 3-residue motif:

S/T – X - R/K

(S or T, followed by any residue, followed by R or K)

This is a common motif that occurs in many unrelated proteins.

6

Motifs and domains are FUNCTIONAL elements of proteins.

Some examples of biochemical functions of domains:

- An enzyme's catalytic domain has the function of catalyzing the conversion of a reactant into a product.
- A structural protein domain has the function of influencing the shape of a cell.
- The binding domain of a transport protein has the function of carrying a ligand from one location to another.

Some examples of the functions of motifs:

- The Y's of this tyrosine motif have the function of interacting with specific residues of a protein to stabilize its structure.

  Y-X-Y

- The H's and C's of this zinc finger motif have the function of binding zinc ions. **$C\text{-}X_4\text{-}C\text{-}X_{12}\text{-}H\text{-}X_3\text{-}H$**

2021/11/10

7

How are motifs and domains identified in protein families?

By aligning family members in a global multiple sequence alignment. Motifs and/or domains can then be identified as conserved regions of the alignment. Sometimes it is easy to align the sequences, and these conserved regions are obvious and can be identified "by eye."

At the right is an example of a multiple sequence alignment of a family of proteins. A conserved copper-binding motif is known to exist in these proteins. Examine the alignment carefully– can you identify the region containing the motif?

(See next slide for answer.)

```
Seq1 APIPPPDLKSCGVAHIDDKGTEVSY--SCCPPVPDDIDSVPYYKFPPMTKLR-IRPPAHA 57
Seq2 APAPPPDLSSCSIARINEN-QVVPY--SCCAPKPDDMEKVPYYKFPSMTKLR-VRQPAHE 56
Seq3 APVPIPDLTKCVI-P---PSGAPVP-INCCPPFSK--DIIDFKYP-SFEKLR-VRPAAQL 51
Seq4 SPISPPDLSKCVP-PSDLPSGTTPPNINCCPPYST--KITDFKFP-SNQPLR-VRQAAHL 55
Seq5 APIQAPDLGDCHQ-PVDVPATAPAI--NCCPTYSAGTVAVDFAPPPASSPLR-VRPAAHL 56
Seq6 APILAPDLSTCGP-PADLPASARPT--VCCPPYQS--TIIDFKLPPRSAPLR-VRPAAHL 54
Seq7 APIQAPDISKCG--TATVPDGVTPT--NCCPPVTT--KIIDFQLPSSGSPMR-TRPAAHL 53
Seq8 APIQAPEISKCVVPPADLPPGAVVD--NCCPPVAS--NIVDYKIP-VVTTMK-VRPAAHT 54
Seq9 APIL-PDVEKCTLSDALWDGSVGDH---CCPPPFDLNITKDFEFKNYHNHVKKVRRPAHK 56
     :*   *::  *                **..        :        ::  * .*  *

Seq1 A--DEEYVAKYQLATSRMRELDK-DPFDPLGFKQQANIHCAYCNGAYKIGGK---ELQVH 111
Seq2 A--NEEYIAKYNLAISRMKDLDKTQPLNPIGFKQQANIHCAYCNGAYRIGGK---ELQVH 111
Seq3 V--DDDYFAKYNKALELMRALPDDDPRS---FSQQAKIHCAYCVGGYKQLGYPEIELSVH 106
Seq4 V--DNEFLEKYKKATELMKALPSNDPRN---FTQQANIHCAYCDGAYSQIGFPDLKLQVH 110
Seq5 A--DRAYLAKYERAVSLMKKLPADDPRS---FEQQWRVHCAYCDGAYDQVGFPGLEIQIH 111
Seq6 V--DADYLAKYKKAVELMRALPADDPRN---FVQQAKVHCAYCDGAYDQIGFPDLEIQIH 109
Seq7 V--SKEYLAKYKKAIELQKALPDDDPRS---FKQQANVHCTYCQGAYDQVGYTDLELQVH 108
Seq8 M--DKDAIAKFARAVDLMRALPGDDPRN---FYQQALVHCAYCNGGYDQVNFPDQEIQVH 109
Seq9 AYEDQEWLNDYKRAIAIMKSLPMSDPRS---HMQQARVHCAYCDGSYPVLGHNDTRLEVH 113
          .   . .:  *    : *   :* .   . **  :**:** *.*   .    ::.:*

Seq1 FSWLFFPFHRWYLYFYERILGSLINDPTFALPYWNWDHPKGMRIPPMFDREGSSLYDEKR 171
Seq2 NSWLFFPFHRWYLYFHERIVGKFIDDPTFALPYWNWDHPKGMRFPAMYDREGTSLFDVTR 171
Seq3 NSWLFLAFHRWYIYFYERILGSLINDPTFAIPFWNFDAPDGMQIPSIFTNPNSSLYDLKR 166
Seq4 GSWLFFPFHRWYLYFYERILGSLINDPTFALPFWNYDAPDGMQLPTIYADKASPLYDELR 170
Seq5 SCWLFFPWHRMYLYFHERILGKLIGDETFALPFWNWDAPDGMSFPAMYANRWSPLYDPRR 171
Seq6 NSWLFFPWHRFYLYSNERILGKLIGDDTFALPFWNWDAPGGMQFPSIYTDPSSSLYDKLR 169
Seq7 ASWLFLPFHRYYLYFNERILAKLIDDPTFALPYWAWDNPDGMYMPTIYASSPSSLYDEKR 168
Seq8 NSWLFFPFHRWYLYFYERILGKLIGDPSFGLPFWNWDNPGGMVLPDFLNDSTSSLYDSNR 169
Seq9 ASWLFPSFHRWYLYFYERILGKLINKPDFALPYWNWDHRDGMRIPEIFKEMDSPLFDPNR 173
      .*** .:** *:*  ***:..:*..  *.:*:* :*   ** :* :     :.*:*  *

Seq1 NQNHRNGTIIDLGHFGKDVRTPQL------
Seq2 DQSHRNGAVIDLGFFGNEVETTQL------
Seq3 DSRHQPPRIIDLNYNKDTEDPGPNYPPSAE
Seq4 NASHQPPTLIDLNFCDIGSDIDRN------
Seq5 NQAHLPPFPLDLDYSGTDTNIPKD------
Seq6 DAKHQPPTLIDLDYNGTDPTFSPE------
Seq7 NAKHLPPTVIDLDYDGTEPTIPDD------
Seq8 NQSHLPPVVVDLGYNGADTDVTDQ------
Seq9 NTNHLD-KMMNLSFVSDEEGSDVN----ED
     :  *     ::*..
```

8

2021/11/10

The copper-binding motif is within the red box. It is located within a conserved section of sequence which is marked with a yellow box (note the " * : . " symbols below the alignment which indicate conserved residues).

Sometimes (as in this example) it is easy to align family members and identify conserved regions that are likely to be important to the function of the protein.

However, for distantly related sequences, it may be very difficult to even align the sequences properly, let alone detect conserved sequence patterns. These situations require the use of sensitive statistical methods.

Why is it useful to identify motifs and domains in families of proteins?

- To identify the functionally important residues and patterns in a given domain.
- To predict the function of a new protein by comparing its sequence to the sequences of domains with known functions.

How are motifs and domains in protein families represented?

1. Regular expressions/patterns

A multiple sequence alignment is converted to a consensus sequence called a regular expression or pattern. Example:

```
Multiple sequence alignment:
       seq1 GEW
       seq2 GTW
       seq3 GTY
       seq4 GRW
       seq5 GKW
       seq6 GAW
----------------------------

Regular expression: G-X-[WY]
(G, followed by any residue,
followed by W or Y)
```

2021/11/16

11

Interpreting regular expressions:

Example: `E-X(2)-[FHM]-X(4)-{P}-L`

2021/11/16

Interpretation:

First residue of the pattern is E;
followed by any 2 residues;
followed by F, or H, or M;
followed by any 4 residues;
followed by any residue except P;
followed by L.

Limitations of regular expressions:

They do not take into account sequence probability information about the multiple sequence alignment. For instance, in the above example, we don't know how often F, H, and M each occur at the 4th position in this motif. H may be much more common than F or M, but we have no way of knowing this from the regular expression.

12

How are motifs and domains in protein families represented?

1. **Regular expressions/patterns**
2. **Statistical models:**

   PSSMs, profiles, and profile hidden Markov models.

2021/11/16

Recall that position-specific scoring matrices (PSSMs) and profiles are numerical representations of a multiple sequence alignment that contain information about the probability of observing a specific residue at a given location in the alignment.

See Powerpoint notes from "Multiple Sequence Alignment" to review PSSMs and profiles.

Profile hidden Markov models are explained next…

13

Profile hidden Markov models (HMMs) are similar to PSSMs and profiles because they describe the likelihood that a specific amino acid residue occurs at a given position in an alignment.

2021/11/16

Consider this multiple sequence alignment of a short motif:

```
seq1  GTWYA
seq2  GLWYA
seq3  GRWYE
seq4  GTWYE
seq5  GEWFS
```

If we were to construct a PSSM for this sequence alignment, we would set up a 20X5 matrix– 5 matrix rows (one for each column in the alignment), and 20 matrix columns for the 20 possible amino acids. Each position in the matrix would have a number indicating the probability of finding a particular amino acid at that column in the alignment. (See Multiple Sequence Alignment Powerpoint notes for examples of PSSMs.)

A profile HMM contains the same type of probability information for various amino acids at all positions in the alignment, but this information is presented in a specific type of **diagram**, rather than in a **matrix**. The next slide shows one of these diagrams.

14

To keep things simple, assume that gaps are not allowed in this alignment. Below is a diagram representing a profile HMM for this sequence alignment.

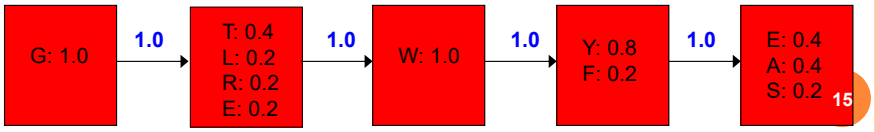
```
seq1 GTWYA
seq2 GLWYA
seq3 GRWYE
seq4 GTWYE
seq5 GEWFS
```

2021/11/16

Each box represents a position in the alignment and is called a "**match state**."

Each match state contains the probabilities of observing various residues at that position in the alignment– for example, the probability of a T in the second position is 2/5, or 0.4. These probabilities are called "**emission probabilities**."

Each arrow represents a transition from one match state to the next. Each transition has an associated probability called a "**transition probability**."



15

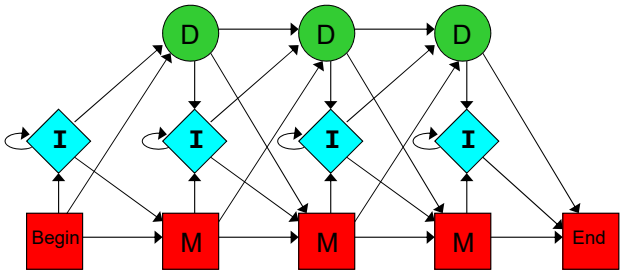
Profile HMMs account for gaps in an alignment:

Below is a generic structure of a profile HMM that includes information about insertions and deletions (gaps) in the multiple sequence alignment from which it was generated.

2021/11/16

The squares represent **match states**, the diamonds represent **insert states**, and the circles represent **delete states**.

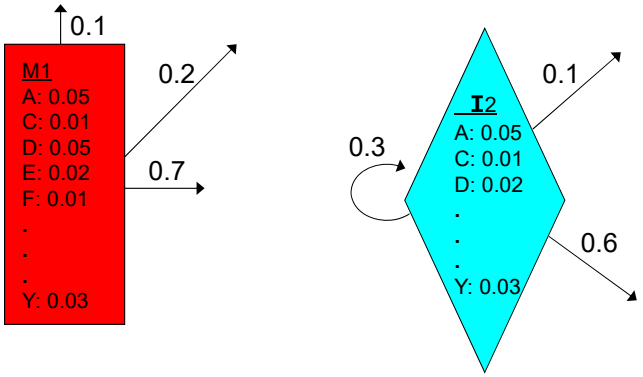
The **arrows** represent transitions from one state to the next. (Note the circular arrow on each insert state. This allows for insertion of more than one residue between match state positions.)



16

Each match state, insert state, and delete state has an associated set of **emission probabilities** for the 20 amino acids that are based on the observed frequencies of the amino acids at that position in the alignment. For example:

2021/11/16



Each arrow/transition has an associated **transition probability** indicating the probability of transitioning to another state. The sum of the probabilities of transitions leaving each state is one.

17

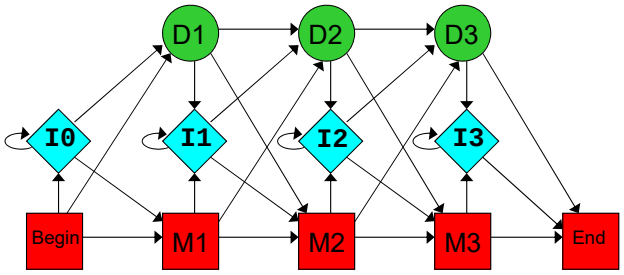
Here is a profile HMM derived from a sequence alignment of a short motif:

```
seq1  G-TW
seq2  G-TW
seq3  G-T-
seq4  G-TW
seq5  GATW
      *  **
```

The alignment is color coded to correspond to its profile HMM. There are three match states, corresponding to the three conserved amino acids in the motif (G,T,W). Seq5 has an insert (A) between the first two match states. Seq3 has a delete in place of the third match state.

2021/11/16

The path of **seq1** through the HMM would be:
Begin→M1→M2→M3→End
The path of **seq3** through the HMM would be:
Begin→M1→M2→D3→End
The path of **seq5** through the HMM would be:
Begin→M1→**I1**→M2→M3→End



NOTE:
Profile HMMs are different from regular profiles because they distinguish between inserts and deletes when accounting for gaps in an alignment.

18

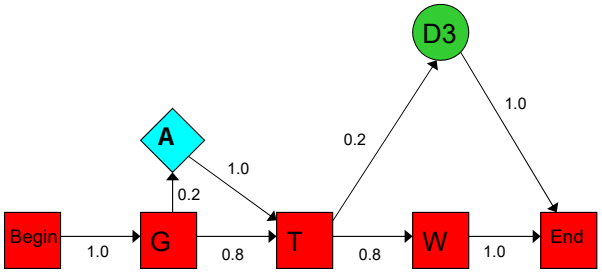
Here is a profile HMM derived from a sequence alignment of a short motif:

```
seq1  G-TW
seq2  G-TW
seq3  G-T-
seq4  G-TW
seq5  GATW
      *  **
```

The alignment is color coded to correspond to its profile HMM. There are three match states, corresponding to the three conserved amino acids in the motif (G,T,W).
Seq5 has an insert (A) between the first two match states.
Seq3 has a delete in place of the third match state.

2021/11/16

The path of **seq1** through the HMM would be:
Begin→M1→M2→M3→End
The path of **seq3** through the HMM would be:
Begin→M1→M2→D3→End
The path of **seq5** through the HMM would be:
Begin→M1→**I1**→M2→M3→End



NOTE:
Profile HMMs are different from regular profiles because they distinguish between inserts and deletes when accounting for gaps in an alignment.

19

Construction of a profile HMM:

Constructing a profile HMM from a set of related sequences is called "**training**" the model.

2021/11/16

The sequences used are the "**training set**." (They do not necessarily have to be aligned prior to constructing the HMM.)

Usually 50 or more related sequences are needed to train the model, but sometimes as few as 20 sequences will work.

20

To construct a profile HMM from an alignment, we must assign:

(1) the length of the model (how many match states)

(2) the probability parameters (emission and transition probabilities)

An example:

```
Seq1 VGA--HAGEY
Seq2 V----NVDEV
Seq3 VEA--DVAGH
Seq4 VKG------D
Seq5 VYS--TYETS
Seq6 FNA--NIPKH
Seq7 IAGADNGAGV
     ***  *****
```

2021/11/16

(1) Length of the model:

A length of **8 match states** is appropriate for the HMM that will represent this alignment. The 8 match states correspond to the columns indicated by the asterisks.

A simple rule of thumb is that columns with more than 50% gap characters should be modeled as insert states rather than match states. Therefore, columns 4 and 5 of the alignment are not included as match states because all sequences except Seq7 have gap characters in these columns. Instead, Seq7 will have two insert states in a row between match states 3 and 4.

21

Same example:

```
Seq1 VGA--HAGEY
Seq2 V----NVDEV
Seq3 VEA--DVAGH
Seq4 VKG------D
Seq5 VYS--TYETS
Seq6 FNA--NIPKH
Seq7 IAGADNGAGV
     ***  *****
```

2021/11/16

(2) Probability parameters:

The values of the emission and transition probabilities are based on the number of times a particular amino acid or gap appears in a given column. In the initial phase of training the model, estimates are made for these probabilities. The model is trained by iteratively refining it and updating the probabilities. This may take up to 10 rounds.

22

Same example:

```
Seq1 VGA--HAGEY
Seq2 V----NVDEV
Seq3 VEA--DVAGH
Seq4 VKG------D
Seq5 VYS--TYETS
Seq6 FNA--NIPKH
Seq7 IAGADNGAGV
     ***  *****
```

2021/11/16

A major problem is that there may not be very many sequences (family members) in the training set, so some legitimate transitions or emissions may not be represented in the alignment and would receive a zero probability. Then these transitions and emissions would not be allowed when the HMM was used in the future. To avoid zero probabilities, **pseudocounts** are added to the observed frequencies of each amino acid. The simplest pseudocount method is Laplace's rule in which one is added to each observed frequency.

For example, V appears 5 times in column 1. According to Laplace's rule:
the count for V in column 1 would become 6 (5 real counts + 1 pseudocount);
the count for F in column 1 would become 2 (1 real count + 1 pseudocount);
the count for **I** in column 1 would become 2 (1 real count + 1 pseudocount) ;
the count for any other residue in column 1 would be 1 (0 counts + 1 pseudocount).

```
Seq1 VGA--HAGEY
Seq2 V----NVDEV
Seq3 VEA--DVAGH
Seq4 VKG------D
Seq5 VYS--TYETS
Seq6 FNA--NIPKH
Seq7 IAGADNGAGV
```

The final profile HMM for the sequence alignment:

2021/11/16

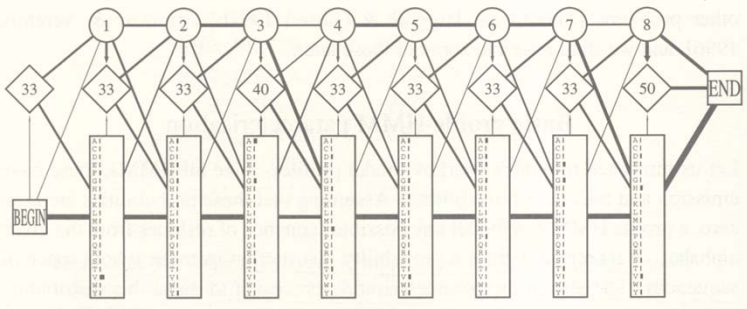


Figure 5.4 *A hidden Markov model derived from the small alignment shown in Figure 5.3 using Laplace's rule. Emission probabilities are shown as bars opposite the different amino acids for each match state, and transition probabilities are indicated by the thickness of the lines. The* I → I *transition probabilities times 100 are shown in the insert states. (Figure generated automatically using the* SAM *package.)*

(From *Biological Sequence Analysis*, by R. Durbin et al., 1998.)

How do we find motifs and domains that may be present in the sequence of a new protein?

Regular expressions, PSSMs, profiles, and profile HMMs represent motifs and domains found in protein families. Therefore:

2021/11/16

- An individual sequence can be compared to a regular expression, PSSM, profile, or profile HMM to see if the new sequence "fits" the previously characterized domain or motif that is represented. This process was explained for PSSMs in the "Multiple Sequence Alignment" PPT notes.

- Better yet, an individual sequence can be compared to an entire database of regular expressions, PSSMs, profiles, or profile HMMs to see whether it belongs to any of the previously characterized families.

NOTE: The statistical models have much more predictive power than regular expressions. Profile HMMs are especially powerful. In fact, one of the main purposes of developing profile HMMs is to use them to detect potential membership in a family by obtaining a match of a sequence to the profile HMM.

25

Scoring a match of a new sequence to a profile HMM:

new_seq: GTVW

```
seq1  G-TW
seq2  G-TW
seq3  G-T-
seq4  G-TW
seq5  GATW
      *  **
```

The best path through the model for the new sequence '**GTVW**' appears to be:

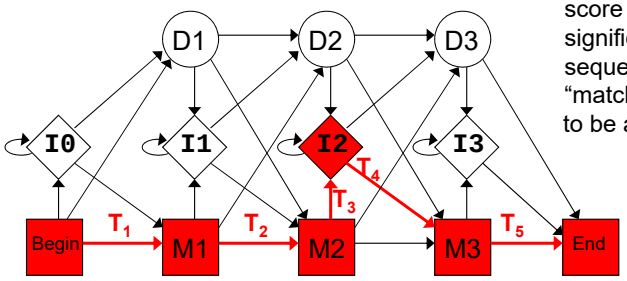
Begin→M1→M2→**I**2→M3→End

2021/11/16

The path is highlighted in red below. To score this path, we add the log odds scores for the relevant emission (E) and transition (T) probabilities:

Score = $T_1 + E_{M1:G} + T_2 + E_{M2:T} + T_3 + E_{I2:V} + T_4 + E_{M3:W} + T_5$

The score is then compared to the score for a random sequence scored against the same profile HMM. If the score of the new sequence is significantly better, the new sequence can be considered a "match" to the HMM and is likely to be another family member.



26

Determining if a sequence of interest contains a motif or domain represented by a profile HMM:

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

2021/11/16

We would use the profile HMM 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/domain in the sequence.

Databases of motifs and domains

The following are databases of regular expressions, PSSMs, profiles, and/or profile HMMs derived from alignments of motifs and domains found in protein families. You can submit a protein sequence to any of these databases in order to determine if the sequence contains one of the motifs or domains represented in the database.

2021/11/16

PRINTS (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/PRINTS.html)
Uses PSSMs.
Breaks a sequence down into small, nonoverlapping motifs; each protein family is said to have a characteristic "fingerprint," or set of these motifs.
Beware: database is small.

BLOCKS (http://blocks.fhcrc.org/blocks/)
Uses PSSMs derived from the most conserved, ungapped regions of alignments. These ungapped aligned regions are called "blocks."
Note: BLOCKS database is no longer being updated.

ProDom (http://prodom.prabi.fr/prodom/current/html/form.php)
Domain alignments were built using PSI-BLAST.

Databases of motifs and domains

Pfam (http://pfam.sanger.ac.uk/search)
Uses profile HMMs.
Two-part database: Pfam-A (curated) and Pfam-B (automatically generated).

SMART (http://smart.embl-heidelberg.de/)
Uses profile HMMs.
Alignments of domains checked manually by curators.

InterPro (http://www.ebi.ac.uk/interpro/)
An integrated database designed to unify multiple databases, including PROSITE, Pfam, PRINTS, ProDom, SMART, and others.
Note: searching InterPro may produce different results than searching the individual databases that are part of InterPro.

CDART (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps)
Uses profiles.
Includes the SMART and Pfam databases.
Note: searching CDART may produce different results than searching SMART or Pfam individually.

2021/11/16

29

Databases of motifs and domains

PROSITE (http://www.expasy.org/prosite/)
Mainly uses regular expressions, some profiles.
Beware: some sequence patterns in the database are too short to be specific, and the database is small; results should be treated with caution.

Emotif (http://motif.stanford.edu/distributions/emotif/)
Uses regular expressions.

Each database has its strengths and weaknesses. HMM-based methods for finding motifs/domains are the most sensitive, and methods using regular expressions are the least sensitive. It is always best to search multiple databases when looking for motifs/domains in a new sequence.

If you don't find any motifs in a sequence when searching a particular database, it could be due to limited coverage of the database. Try other databases before concluding that the sequence has no known motifs.

Keep in mind that there are many sequences with misannotated motifs/domains in databases.

2021/11/16

30

Using a profile HMM to find new members of a protein family:

In the “Multiple Sequence Alignment” module you learned that a PSSM or profile can be used to scan a database of individual sequences in order to find other sequences that are members of the family represented by the PSSM/profile. This is done by scoring every sequence in the database against the PSSM/profile to find any that produce high scores.

A profile HMM can also be used to scan sequence databases in the same manner to find other family members.

2021/11/16

31

Identifying motifs and domains using statistical methods:

As mentioned earlier, for distantly related sequences it may be very difficult to align the sequences and detect conserved regions. Three statistical methods can be used in these cases:

Expectation maximization algorithm
Gibbs sampler algorithm
Profile HMMs

2021/11/16

These methods do not rely on a previously produced multiple sequence alignment. They identify patterns (conserved motifs/domains) in a set of sequences by producing trial alignments and then improving the alignments using statistical methods.

From the final alignment, each of these methods produces a scoring matrix that may be used to search other sequences for the same pattern.

32

Expectation Maximization (EM) Algorithm

EM is a two-stage iterative process. An initial guess is made as to the location and size of a sequence pattern (a motif or domain) in each sequence in a set of related sequences. These regions are aligned to create a "trial" alignment for the set of sequences. Using the trial alignment, the residue composition of each column in the alignment is first calculated and used to create a PSSM.

2021/11/16

Step 1. Expectation
Using the values in the PSSM, the probability of finding the pattern at every possible position in each sequence is calculated.

Step 2. Maximization
The probabilities from step 1 are used to weight the values in the PSSM, essentially providing new information about the likely location of the pattern in each sequence. The values in the PSSM are updated using these weights.

Steps 1 and 2 are repeated until the values in the PSSM don't change with continued iterations.

33

The Gibbs Sampler Algorithm

Gibbs sampling is an iterative process similar in principle to EM, but the algorithm is different. At each iteration, one sequence is removed and a trial alignment is built from the remaining sequences. Like EM, Gibbs sampling searches a set of sequences for the statistically most probable motifs, and can find the optimal width and number of these motifs in each sequence.

2021/11/16

Profile HMMs

A profile HMM can be built from a set of unaligned sequences. An initial guess is made as to the length of the model (number of match states) and the probability parameters. Then the model is trained by iteratively updating the probability parameters (a variety of algorithms can be used for this process).

34

An Example of EM-algorithm

Initialization: begin with a set of aligned sequences (randomly).

1:	C G G G T A A G T
2:	A A G G T A T G C
3:	C A G G T G A G G

2021/11/16

35

Step 1. Maximization

Calculating weight matrix of

1:	C G G G T A A G T
2:	A A G G T A T G C
3:	C A G G T G A G G

Now one counts the number of nucleotides contained in all sequences:

A:	1 2 0 0 0 2 2 0 0	7
C:	2 0 0 0 0 0 0 0 1	3
G:	0 1 3 3 0 1 0 3 1	12
T:	0 0 0 0 3 0 1 0 1	5

2021/11/16

36

Now one needs to sum up the total: 7+3+12+5 = 27; this gives us a "dividing factor" for each base, or the equivalent probability of each nucleotides.

A:	7/27 ≈ 0.26
C:	3/27 ≈ 0.11
G:	12/27 ≈ 0.44
T:	5/27 ≈ 0.19

Now one can "redo" the weight matrix (WM) by dividing it by the total number of sequences (in our case 3):

A: 0.33 0.66 0.00 0.00 0.00 0.66 0.66 0.00 0.00
C: 0.66 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.33
G: 0.00 0.33 1.00 1.00 0.00 0.33 0.00 1.00 0.33
T: 0.00 0.00 0.00 0.00 1.00 0.00 0.33 0.00 0.33

2021/11/16

37

Next, one divides the entries of the WM at position with the probability of the base .

A: 1.29 2.57 0.00 0.00 0.00 2.57 2.57 0.00 0.00
C: 6.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 3.00
G: 0.00 0.75 2.25 2.25 0.00 0.75 0.00 2.25 0.75
T: 0.00 0.00 0.00 0.00 5.40 0.00 1.80 0.00 1.80

In general one would now multiply the probabilities. In our case one would have zero for every one. Due to this we define $log_{10}0 := -10$ and take the (base 10) logarithm:

A :	0.11	0.41	−10	−10	−10	0.41	0.41	−10	−10
C :	0.78	−10	−10	−10	−10	−10	−10	−10	0.48
G :	−10	−0.12	0.35	0.35	−10	−0.12	−10	0.35	−0.12
T :	−10	−10	−10	−10	0.73	−10	0.26	−10	0.26

2021/11/16

38

Step 2. Expectation

Identify a new set of aligned sequences.

2021/11/16

Using weight matrix (WM), one can use an example of a promoter sequence to determine its score. To do this, one has to add the numbers found at the position of the logarithmic WM. For instance, if one takes the AGGCTGATC promoter:
0.11 - 0.12 + 0.35 - 10 + 0.73 - 0.12 + 0.41 - 10 + 0.48 = -18.17
This is then divided by the number of entries (in our case 9) yielding a score of -2.02.

Step 3. Repeat Step 1 and 2.

39