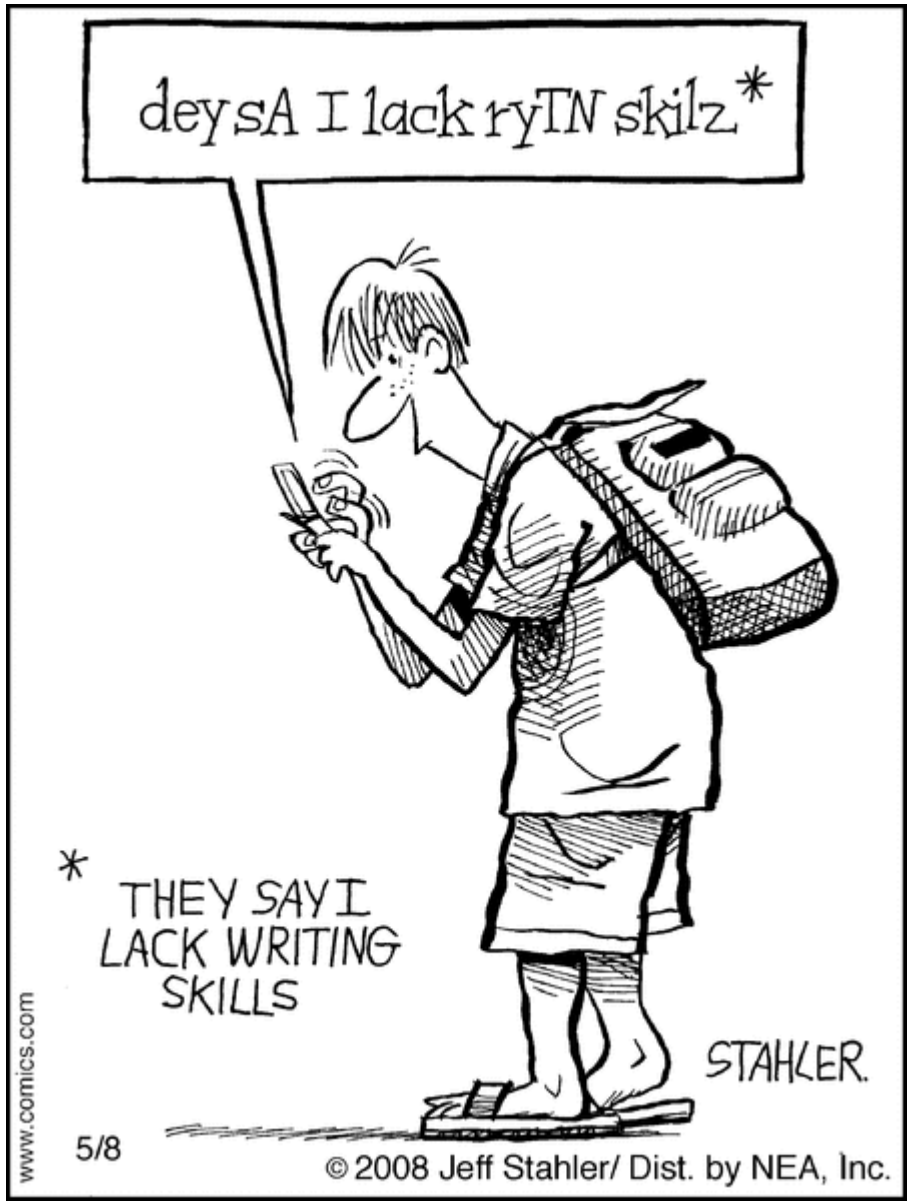
EE550
Computational Biology

Week 10 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Pattern searching in functional protein groups
  - The protein recognition problem
  - Regular expressions
  - Profiles and PSSMs
  - Fingerprints
  - Blocks



Source: http://www.gocomics.com/moderately-confused/2008/05/08

Protein Recognition Problem

- Sequence alignment algorithms can be used to determine the functional properties of a newly sequenced protein
  - The sequence is aligned to all protein sequences in a database
  - The database proteins with the most similar sequences are determined
  - As sequence is predictive of function, the newly sequenced protein can be hypothesized to carry out similar functions in the cell
- The results, however, are subject to noise and errors
  - Chance alignments against sequence databases containing millions of sequences
  - Homologue problem in protein-protein interactions
    - Homologues of interacting proteins do not necessarily interact
- An alternative approach focuses on the presence of sequence fragments that are indicative of a functional protein group
  - If sequence fragments commonly found in a certain functional protein group are identified on the newly sequenced protein, the odds are it performs functions similar to those related to that protein group
  - Still, everything is stochastic!!

Sequence Motifs and Domains

- Sequence fragments that are observed much more frequently among the sequences of a protein group are termed **sequence motifs**
  - These are conserved regions that presumably signify structural and functional properties common to all members of the group
  - Presence of sequence motifs in the sequence of a novel protein provides a strong indication that the protein may also be a member of the corresponding groups
- Sequence fragments that characterize members of a functional or structural protein group are termed **protein domains**
  - Domains preserve their three-dimensional conformation in different proteins
- Identification of sequence motifs in a given sequence is a **pattern searching/matching problem**
  - Patterns encode the sequence motifs

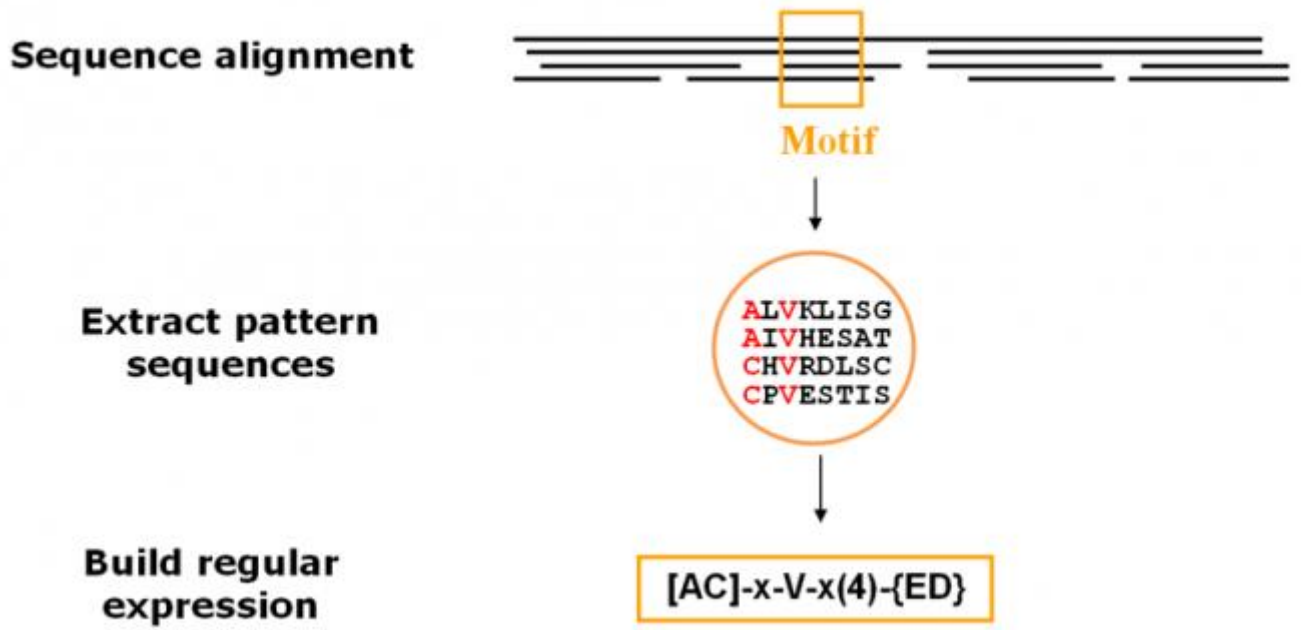
Patterns – Regular Expressions

- Regular expressions (regexs) encode the most conserved regions of motifs allowing variation over the less conserved ones
- In a regular expression encoding of a sequence motif, the required, permitted, and prohibited amino acids are indicated
  - Absolutely conserved sites (represented by the amino acid letter)
  - Sites with a few amino acids with similar properties (represented by the possible amino acid letters in square brackets; [ ])
  - Sites that are not conserved (represented by an x, and followed by a numeric range if the lengths of such regions vary as well)
  - Sites that are prohibited to certain amino acids (represented by the amino acid letters in curly brackets; { })
  - Repeat sites with similar regex encoding (represented by a number in parentheses, indicating the number of times the previous encoding pattern is repeated)
- Regexs are obtained from results of multiple sequence alignment of proteins of a select functional group
  - using local alignments instead of global alignments

Regular Expressions

- Note that the list of allowed and disallowed amino acids at a site has to do with how similar amino acids are
  - Amino acids can be clustered into the following groups based on their physico-chemical properties
    - Basic : K, R, H
    - Acid and amides : E, D, Q, N
    - Small : P, T, S, G, A
    - Cysteine : C
    - Hydrophobic : V, L, I, M, F
    - Large and aromatic : W, Y
  - Alternatively, a scoring matrix can be used to determine whether a site is conserved within an interchangeable set or simply not conserved

Regular Expressions



Source: https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-ar-2

Regular Expressions

- Example
  - Consider the following multiple sequence alignment result

```
…Q E R V E E L S L V R V D D T I S Q…
…Q E R V E E L S L V R V D D A I S Q…
…Q E K I E E L S L V R V D D T V S Q…
…Q E R I E E L S L V R V D D T I S Q…
…Q E K I E E L S L V R V D D T V S Q…
…Q E R V E Q L S L V R V D D T I S Q…
…Q E R I E E L S L V R V D D T I S Q…
…Q E R I E E L S L V R V D D T I S Q…
…Q E R V E E L S L V R V D D T I S Q…
…Q E R I E E L S L V R V D D T I S Q…
```

  - The corresponding regex is given by

    Q-E-[RK]-[VI]-E-[EQ]-L-S-L-V-R-V-D-D-[AT]-[VI]-S-Q

Regular Expressions

- Example (continued):

  Q-E-[RK]-[VI]-E-[EQ]-L-S-L-V-R-V-D-D-[AT]-[VI]-S-Q

  – Note that
    - R and K are both basic residues,
    - V and I are both hydrophobic,
    - E and Q are acidic, and
    - A and T are small
  – Note:
    - Basing regex construction upon the multiple sequence alignment results creates a circularity problem
      – Regexes represent patterns over sites that have aligned well
        » According to a specified substitution structure
        » The regexes inevitably reflect this structure
      – Regions that have not aligned well do not produce any regexes, or motifs in a general sense

Regular Expressions

- Given a set of regular expressions, the remaining task is to locate them on a given amino acid sequence
  - if they exist
- This entails searching the sequence for potential matches
- A match is identified when a sequence fragment fits the pattern encoding provided by the regex
  - Given the regex Q-E-[RK]-[VI]-E-[EQ]-L :
    - The fragment QEKVEEL is a match
    - The fragment QEHVEEL is a mismatch

Regular Expressions

- Remarks:
  - Regular expressions for conserved amino acid patterns are derived from multiple sequence alignment of sequences belonging to a given functional protein group
  - While the regex may characterize the conserved region information well on the current group definition, it may change in the future when the group definition is revised
    - Relocation of existing members
    - Addition of new members
  - A very similar fragment may temporarily be deemed a mismatch simply because the sequence slightly deviates from the regex definition
    - The deviation may be as simple as carrying an alternative amino acid at a site
    - If such a possibility has not been encountered before across the existing sequence data, it will not be recognized as a match

Regular Expressions

- Remarks:
  - In addition, some sequence fragments may be conserved across many protein groups
    - This is especially true for short patterns
      - The number of specific amino acid patterns of length L is 20^L
      - In a sequence of length N, there are a total of $N - L + 1$ sequence segments of length L
      - In a database of M sequences, each of length N, the average number of times a given sequence fragment will be observed is thus $M(N - L + 1)/20^L$
  - Therefore, matches of short regexes must be evaluated carefully
  - All these considerations impose a trade-off for regex generation:
    - The more specific (long and well-determined) the less likely to find a match, let alone an unrelated match
    - The more relaxed (short and with many alternatives) the more numerous the unrelated matches

Regular Expressions

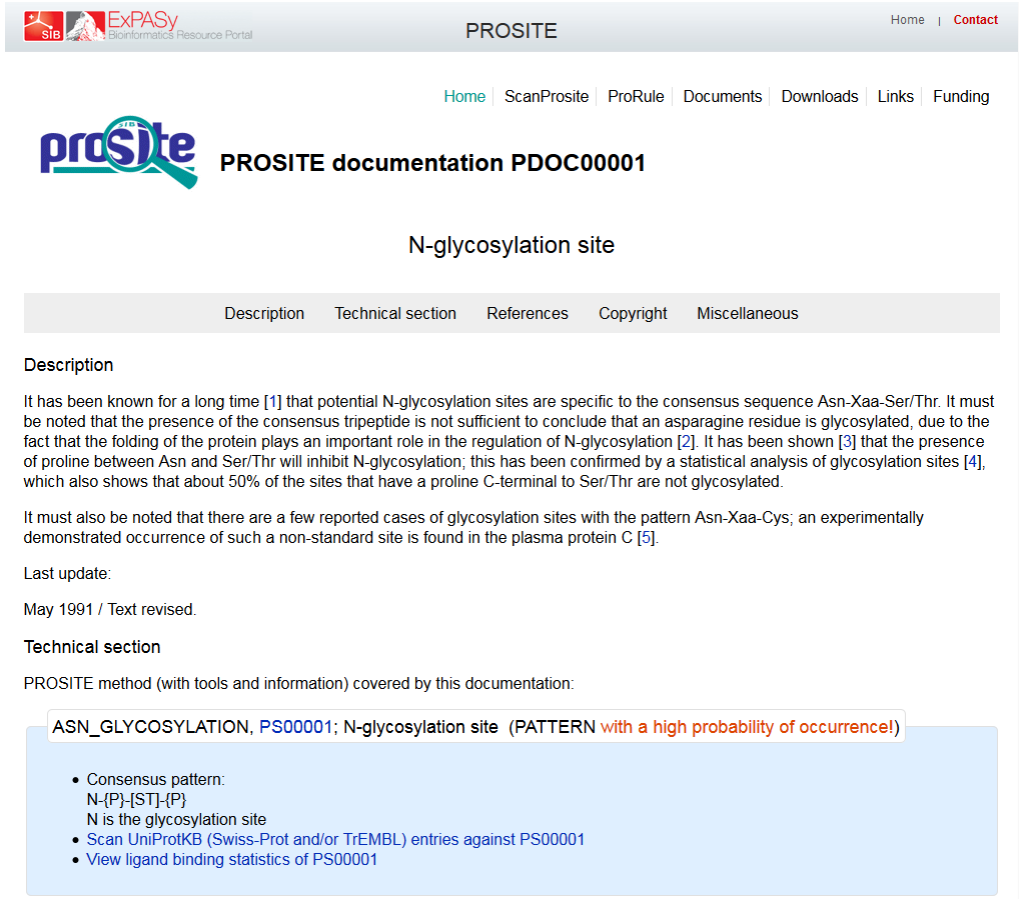
- Example: N-glycosylation sites
  - The regular expression for N-glycosylation is

N-{P}-[ST]-{P}

≡

asp-not pro-ser or thr-not pro

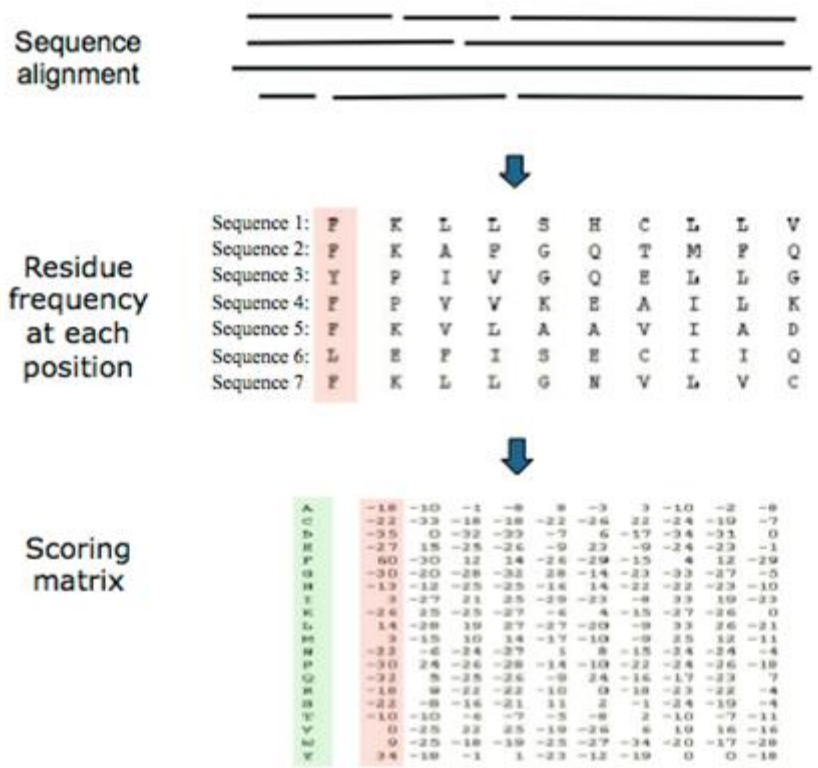
  - asparagine is the N-glycosylation site
  - either a serine or a threonine residue is required for glycosylation
  - a proline residue between the asparagine and serine/threonine prohibits glycosylation



Profiles – Position-Specific Scoring Matrices

- Regexs produce the lists of sites that are conserved, partially conserved, or not conserved
- A straightforward strategy to generalize such a list of conservation patterns is to note the frequencies of amino acids observed at each site in a $L \times 20$ matrix
  - L denoting the length of the motif
- Such a frequency matrix can then be used to search for the same motif in other sequences
  - Frequency matrices can be slid over a given amino acid sequence
  - The window positions that produce the highest similarity indicate the matches to the pattern in consideration

Profiles – Position-Specific Scoring Matrices



Source: https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-are-

Profiles – Position-Specific Scoring Matrices

- Example
  - Consider the multiple sequence alignment in the previous example
  - The corresponding frequency matrix is given by

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

Profiles – Position-Specific Scoring Matrices

- Remarks:
  - Frequency matrices computed from a handful of sequences are inevitably very sparse
  - This implies a severe limitation in characterizing possible amino acid alternatives at different sites for motifs of small protein groups
  - Carrying out multiple iterations of profile construction alleviates this situation
    - Each time, additional instances of the motif are identified via a relaxed inclusion criterion and incorporated into the profile
    - As the number of motifs increases, the identification criterion becomes more strict
  - An alternative is to incorporate the amino acid conservation information from scoring matrices
    - even though the scoring matrices are obtained from much larger and much more general (albeit homologue) sequence datasets

Profiles – Position-Specific Scoring Matrices

- Searching sequences for instances of profiles entails contrasting the observed amino acids within a sliding window along the sequences to frequency matrices or PSSMs
  - This requires computing a likelihood for the window in question subject to the profile in consideration
  - The likelihood can then be converted into a matching score
- Once all sequences are searched, the matches are presented as a list of decreasing matching score

Profiles – Position-Specific Scoring Matrices

- PSSMs:
  - Consider a frequency matrix $F(i,j)$ for $i = 1,2,\dots,20$, and $j = 1,2,\dots,L$ characterizing the amino acid occurrences in a given motif
  - A matching score for a window W of length L on a sequence S with

$$W_{\Delta i}(i) = S(i + \Delta i)$$

can then be computed by

$$\sum_j F\left(I\left(W_{\Delta i}(j)\right), j\right)$$

where $I\left(W_{\Delta i}(j)\right)$ produces the row index of the amino acid observed at the j'th site on the window W

Profiles – Position-Specific Scoring Matrices

- PSSMs (continued):
  - Now, suppose also that the matrix F has been observed from a total of M instances
    - So that

$$\sum_{i=1}^{20} F(i,1) = \sum_{i=1}^{20} F(i,2) = \cdots = \sum_{i=1}^{20} F(i,L) = M$$

  - This allows converting frequencies into probabilities with

$$\frac{F(i,j)}{M}$$

calculating the probability of observing the character indexed by i at the j'th position on the sliding window at a matching position

  - These probabilities can then be used to calculate a likelihood of observing a match on the current window using

$$L(W) = \Pr\{\text{"profile match over } W\text{"}\} = \prod_{j=1}^{L} \frac{F\big(I\big(W(i)\big), j\big)}{M}$$

or the log-likelihood

$$\log\big(L(W)\big) = \sum_{j=1}^{L} \log\left(\frac{F\big(I\big(W(i)\big), j\big)}{M}\right)$$

  - The logarithm can be expressed as scores which allows calculating the log likelihood as summations over scores

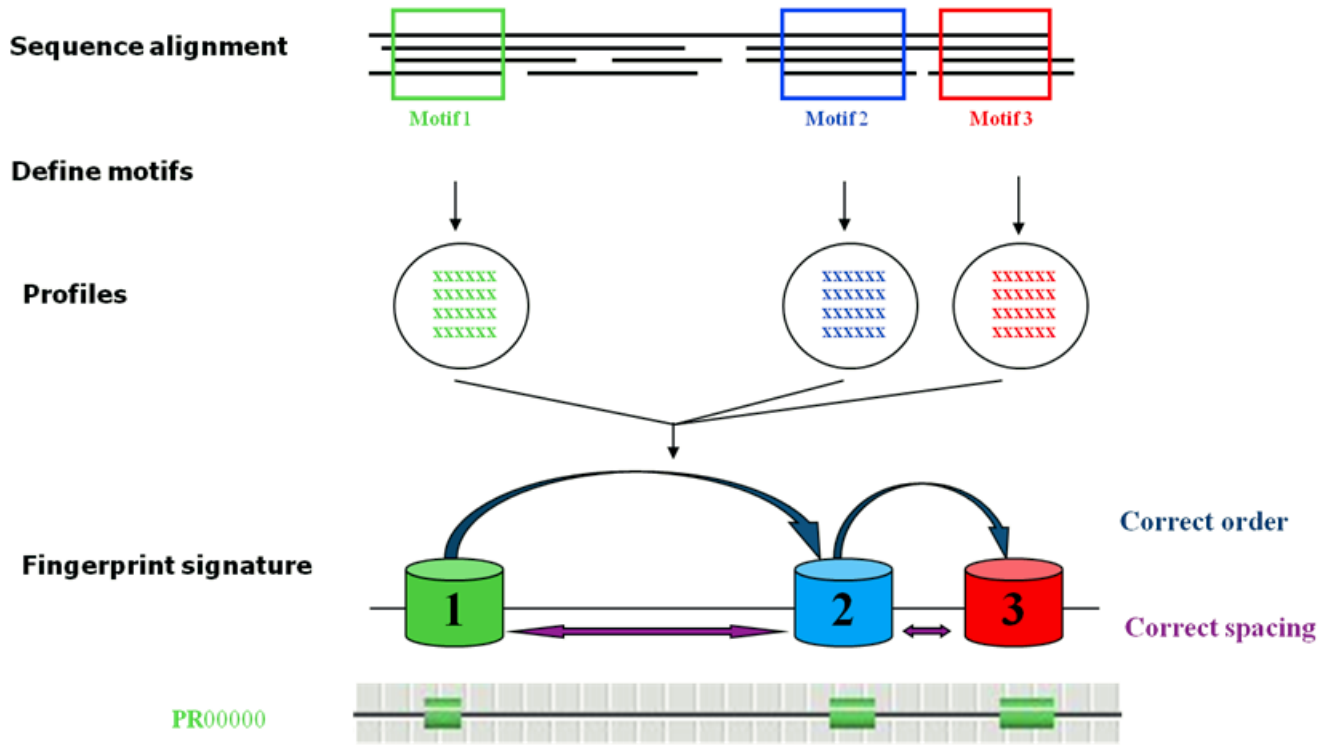
Profiles – Position-Specific Scoring Matrices

- Remarks:
  - The searches performed while generating the profiles and while locating them on a given sequence or list if sequences are different
    - While generating the profiles, a very high-fidelity match to the motif is required, and only the sequence fragments with substantial agreement to the growing profile are included in the hit lists
    - Conversely, when searching for motif occurrences, the hit lists are generated in a more permissive manner, so as not to miss potential matches
  - In addition, when generating a hit list, certain measures on the reliability of the hits are also produced indicating the statistical significance of the hits
    - Probability P
    - Expected value E

Fingerprints

- Regular expressions and profiles are fine for capturing the composition of short sequence fragments in a group of sequences
- However, members of a typical protein group tend to possess a **multitude of motifs** in a specific **order** and a specific **spacing** ➔ fingerprints

Fingerprints



Source: https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-ar-0

Fingerprints

- Generating a fingerprint signature for a given group of proteins involves the following steps:
  - multiple sequence alignment
  - motif generation over highly conserved regions
    - typically using profiles
  - identification of the correct order and spacing of the established motifs
- Evaluation of the prospective members to the group then requires observing the same motifs in the same order and spacing

Blocks

- Regexes encode the pattern information via semantic rules
- Profiles store the site-specific amino acid frequencies
- Another alternative is to encode the relative positions of amino acid **triplets** along amino acid sequences
  - Amino acid pairs occur plentifully without much regard to specificity
  - The complexity of tracking four amino acid line-ups is prohibitive
  - Amino acid triplets provide the right combination of specificity and analytical complexity
- The conserved motifs are thus represented by spaced out amino acid triplets called **blocks**
  - The blocks that are conserved among the sequences of a protein group are to be determined by searching and counting the occurrences of different blocks
  - Similar triplets are determined by PSSMs and grouped under the same block

Blocks

- Example:
  - Consider the amino acid sequence

Q-E-R-V-E-E-L-S-L-V-R-V-D-D-T-I-S-Q-P-P

  - From this sequence, the following instances of blocks can be identified:
    - AA_1-AA_2-AA_3: Q-E-R, E-R-V, R-V-E, V-E-E, E-E-L, …
    - AA_1-AA_2-x-AA_3: Q-E-x-V, E-R-x-E, R-V-x-E, V-E-x-L, …
    - AA_1-AA_2-x-x-AA_3: Q-E-x-x-E, E-R-x-x-E, R-V-x-x-L, …
    - …
    - AA_1-x-AA_2-AA_3: Q-x-R-V, E-x-V-E, R-x-E-E, V-x-E-L, …
    - AA_1-x-AA_2-x-AA_3: Q-x-R-x-E, E-x-V-x-E, R-x-E-x-L, …
    - AA_1-x-AA_2-x-x-AA_3: Q-x-R-x-x-E, E-x-V-x-x-L, R-x-E-x-x-S, …
    - …
  - This process is to be repeated on all sequences in the protein group
  - The blocks that are most commonly observed across the group can then be identified
    - Using an amino acid substitution matrix to calculate the similarities between different blocks of the same structure
  - A second pass over the sequence data looking for non-overlapping blocks eliminates the spurious blocks

Blocks

- Whether a novel protein is a member of a functional group can be determined by locating the blocks associated with the functional group along the novel protein's sequence
  - More than one block may be associated with the functional group
  - Recognition of several blocks associated with the same group provides a strong indication of the novel protein's membership
- In cases where there are multiple blocks characterizing one functional group, their locations with respect to each other on the sequences may also be of significance
  - introducing additional sophistication to the recognition process, and
  - approaching a fingerprint-like representation for the common sequence characteristics for the group

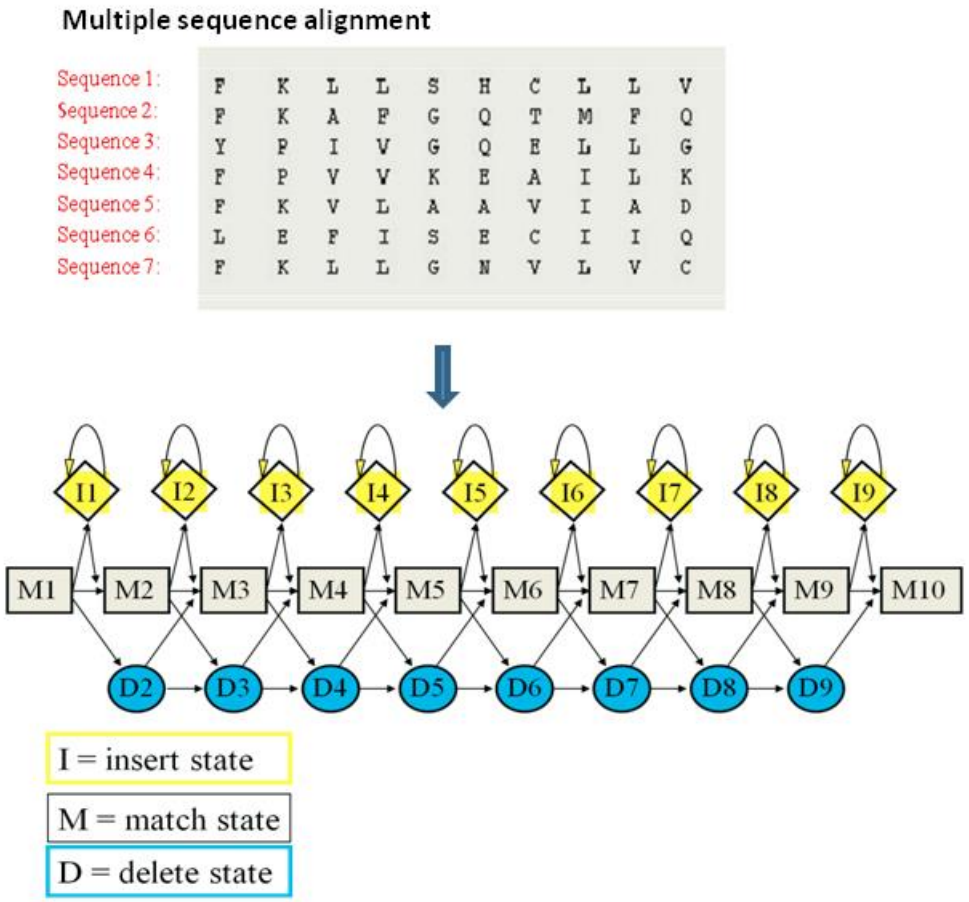
Blocks

- Further remarks:
  - During the search, blocks can be evaluated using position-specific scoring matrices
    - given the position of the first amino acid of a block on the sequence, the rates at which different amino acids are to be observed at the second and third sites
  - A profile, in contrast to blocks, seeks to identify a position-specific substitution rate for all matched regions in a protein group
    - The matched regions are determined using multiple sequence alignment
  - Note also that profiles require aligned sequence fragments; whereas blocks do not

Hidden Markov Models

- A probabilistic characterization of common sequence features among the members of a protein group is obtained using hidden Markov models – HMMs
- This allows representing the motifs using a combination of substitution, insertion and deletion events with respective probabilities

Hidden Markov Models



Source: https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types/what-ar-1

Protein Recognition Example

- Task:
  - Given the amino acid sequence of the CAG pathogenicity island protein 23 of *Helicobacter pylori*
    - http://www.uniprot.org/uniprot/Q48252
  - Use web resources to determine the functional properties of the corresponding protein
    - Alignment search (using protein BLAST at the NCBI protein database)
      - https://www.ncbi.nlm.nih.gov/protein/
    - Regexes and profiles (using the PROSITE database)
      - https://prosite.expasy.org/
    - Fingerprints (using the PRINTS database)
      - http://130.88.97.239/PRINTS/index.php
    - Validation using protein family search in the InterPro database
      - https://www.ebi.ac.uk/interpro/

Summary

- Functional protein groups tend to be characterized by segments of amino acid sequences that are conserved among the group members
- These sequence segments can be encoded using different strategies
- The protein recognition problem entails identifying such conserved sequence fragments on a novel protein sequence
- The more conserved regions pertaining to a specific functional group located on the sequence the stronger the evidence for its membership in the corresponding group