CSI5126. Algorithms in bioinformatics

Deterministic Sequence Motifs

Marcel Turcotte

School of Electrical Engineering and Computer Science (EECS)
University of Ottawa

Version 22 novembre 2017
Summary

This module focuses on **sequence motif discover**. A general framework to classify approaches is presented. This lecture focuses on **deterministic** motifs, whereas the next one focuses on **probabilistic** motifs. Several representations are examined.

**Reading**

Deterministic sequence motifs, manual and automated methods*

- Median String Problem, PRINTS, FINGERPRINTS;
- Regular motifs, PROSITE.

* Hidden agenda: talking about the use of information content to compare motifs.
Motifs

Merriam-Webster Online:

1: a usually recurring salient thematic element (as in the arts); especially: a dominant idea or central theme

2: a single or repeated design or color

Let’s define a **pattern** simply as a set of properties (such as amino acids, secondary or tertiary structure elements) that are common to **some** members of a **family**.

A **motif** is pattern that is common to **most** members of a **family** (input set).
“What is a motif in a biological sequence? One possible meaningful definition is to look for structural or functional implications of a segment and if this can be (unambiguously) associated with the segment, then the segment qualifies to be a motif.” [3]
How to define a **family**?

Here a **family** will be an ensemble of macromolecules, aligned or not, which are thought to be **related**.

**Homologous sequences** are an example of related sequences.

But, **sequences don’t need to be similar**. Experimental evidences may suggest that an ensemble of protein coding genes are always translated (expressed) simultaneously, the gene sequences and their surrounding region can be quite different from one another, yet we might be interested in finding if the occurrence of a common motif could explain this experimental fact; the presence of a protein binding site, for example.
Regulatory Motifs

Find the **common substring** amongst the following 10 strings.

```
ATTGC GGGCACGCGGCGCATCCCGAAACGGAAGCCGATGAT
AGCTCTCCGGGACTCGTAGCCAACGCATCCC AATCTAGA TAATAGTGGCAATCA
ATGTCGACTACGCA GGTTTCG CATCCC AACACAGCCC GGGA
TTACGAGTAGCCTCTG AAAACTCCGC ATCCC TAAAGGGTGCCAAGA ATTAAGT
GACATCACACTACGC GCATCCC ACGTGTA TTTTCTT
ATGGGACG GC GTACGGGC GCATCCC TCTTT TGGCAGGGC
CATTTGTA A T GTGGA ACCACC GCATCCC TTAAGAC ACCAGATACG CGG
AGGGTCGC GTAC T GTGAAGCGC ATCCC GAGTGCA AAGATGA AA
GTGTTTTAAAACAGCGC ATCCC AACC GCAGC GCAGT
TG TACC GACCCCCC GCATCCCGTGAGTG TAA TTAAT TTA
```

**Regulatory regions or sequences**: A DNA base sequence that controls gene expression.
Regulatory Motifs

Find the **common substring** amongst the following 10 strings.

```
attgcgggacgcggCGCATCCCGaaacggaagccgatgat
agctctccgggactcgtagccaaCGCATCCCaatctagataatagtggcaatca
atgtcgactacgcaggttCGCATCCCaaacagccccggga
ttacgagtagcctctgaaactcCGCATCCCtaaggggtgccagaattaagt
gacatcacactacgCGCATCCCAcgtgtatatttctt
atgggacggcgtacggCGCATCCCtttttgcgagggcg
catttgtaattgtggaccacCGCATCCCcttagacaccagatacgcgg
agggtcgcgtactgtaagCGCATCCCGagtgcgaagagatgaaa
gtcatcaccagCGCATCCCaaccgcagccgtag
tggtaccggccccccCGCATCCCGtggtgtatatcaatatta
```

**Regulatory regions or sequences**: A DNA base sequence that controls gene expression.
Regulatory Motifs

Find the **common substring** amongst the following 10 strings.

```
attgcgggacgcggCGCATCCCGaaacggaagccgatgat
agctctccgggactcgtagccaaCGCATCCCaatctagataatagtggaatca
atgtcgactacgcaggttCGCATCCCaaacagccgggga
attacgagtagcctctgaaactcCGCATCCCaagggtgccaaagaattaagt
gacatcacactacgCGCATCCCaacgtgtatattttctt
atgggacggcgtacggCGCATCCCarttttgccgaggcg
catttgtaatttgaccaccacCGCATCCCtagacaccagataacgcgg
agggtcgcgtactgtaagCGCATCCCgagtgcaagatgaaa
gtctgttaaacagCGCATCCCaaccgcagccgtag
tgtaccgaccccccCGCATCCCgtgagtgtaattcaattta
```

**Regulatory regions or sequences**: A DNA base sequence that controls gene expression.
Gene expression: regulating the transcription

Simple sequence elements serve as binding sites for regulatory proteins (factors).

For example, in *Saccharomyces cerevisiae* (yeast), the protein **GAL4** is a transcriptional activator, it binds the following wild card containing sequence, `G.CAAAA.CCGC.GGCGG.A.T`, and activates transcription from a nearby promoter.
Structural Motifs

The **WW domain**: a protein module that binds proline-rich or proline-containing ligands.

The WW domain is a protein-protein interaction module composed of 35-40 amino acids. It is the smallest, monomeric, triple-stranded, anti-parallel beta-sheet protein domain that is stable in the absence of disulfide bonds, cofactors or ligands.

- **Two** conserved *tryptophans* (*W*) spaced 20-22 amino acids apart;
- A block of **two or three aromatic amino acids** located centrally between the two signature tryptophans, and
- A **conserved proline** located **three amino acids** carboxyterminal **to the second conserved tryptophan**.

⇒ Bork and Sudol (1994), TIBS 19 (94), 531-533)
Structural Motifs (cont.)

Representative structures of WW domains based on their sequences

**Class 1:** YAP65

**Class 2:** FBPWW28

**Class 3:** YQJ8WW

1. `-VPLPAGWEMAKT-SSGQRYFLNHIDQTWQDPRKAMLS`
2. `GATAVSEWTEYKTA-DGKTYYYNRTLESTWWEKPQELK--`
3. `--VRLPPGWEIIHE--NGRPLYYNAEQ-KTKLHYPPSGSS--`

Structural Motifs (cont.)
Motivation

➢ To help **understanding protein families**. How?
Features that are conserved are good indicators of important structural or functional positions;

➢ Computational models can allow to **find new members**;

➢ Can serve as the basis for **classification schemes**;

➢ Sometimes allow to detect **sequencing errors**;

➢ As an alternative method to **detect remote homologues/analogue**;

➢ Sometimes it is **difficult** or **not realistic** to compute a **multiple sequence alignment**, pattern discovery can help identify common patterns.

- How to represent patterns?
- How to search for a pattern?
- How to discover patterns automatically?

Let’s distinguish between two kinds of motifs/patterns: deterministic and probabilistic.
How to define a motif?

The most basic pattern is a **substring** (aka rigid pattern).

We have seen algorithms to process strings: exact and approximate string matching.

- Search algorithm. Fast algorithms exist to check for the presence of a motif, Boyer & Moore for example;
- Motif discovery. The longest common substring of $K$ strings can be found with help of generalized suffix trees;
- Mismatches can be allowed, mismatch check algorithm;
- Insertions/deletions and weighted alphabet scoring scheme (string edit distance) are also possible.

⇒ BLOCKS and PRINTS are examples of databases that contain substrings.
Automated approaches to detect conserved substrings

Overrepresented $l$-mers.

Find an efficient algorithm to enumerate conserved or overrepresented $l$-mers ($l$-words appearing $k$ times in the input string (genome), or $l$-words appearing in at least $k$ input strings (genes)).

What are the pros/cons of these approaches (or representation)?
Motifs are **not 100 % conserved**

Here, the **consensus sequence** is not found in any of the input sequences.
Inferring motifs automatically: Median string problem

**Input**: $K$ input sequences and the length $l$ of the motif to be found.

**Problem**: Find a string $v$ (of length $l$) minimizing

$$
\sum_{k=1}^{K} \min_{i_k \in [1..|S_k|]} d_{\text{Hamming}}(v, S_k[i_k, i_k + l - 1])
$$
Inferring motifs automatically: Median string problem (cont.)

1. attgcgggacgcggGCGATTCCgaaacggaagccgatgat
   GCGCATCCC 8
   GCGCATCCC 8
   GCGCATCCC 5
   GCGCATCCC 7
   ...
   GCGCATCCC 1
   ...
   ...

Marcel Turcotte

CSI5126. Algorithms in bioinformatics
Inferring motifs automatically: **Median string problem** (cont.)

\[
\begin{align*}
\text{K. agctctccggyactcgtagccaaCGGATCCGaatctagataatagtggaatca} \\
\text{CGCATCCC 4} \\
\text{CGCATCCC 6} \\
\text{CGCATCCC 5} \\
\text{CGCATCCC 8} \\
\ldots \\
\text{CGCATCCC 2} \\
\ldots
\end{align*}
\]
Inferring motifs automatically : Median string problem (cont.)

Given \( \nu \), calculating

\[
\sum_{k=1}^{K} \min_{i_k \in [1..|S_k|]} d_{\text{Hamming}}(\nu, S_k[i_k, i_k + l - 1]).
\]
Inferring motifs automatically: **Median string problem** (cont.)

However, there are $4^l$ choices of $v$. For small values of $l$, an exhaustive search can be considered, for instance there are 65,536 8-mers.
Exhaustive search
Branch-and-bound

Set $best$ to $\infty$.

Traverse the search tree (depth-first using a stack or best-first using a priority queue).

If current node is a leaf and the total distance of the motif represented by the leaf and the $K$ input sequences is less than $best$ then set $best$ to the score of this motif and memorize the current motif.

If the current node is an internal node and its total distance is larger than $best$ than prune this sub-tree.
How to improve this approach? Finding more aggressive bounds.


Branch-and-bound
Practical application: PRINTS

- “PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family”;
- Release 39.0 of PRINTS (02.02.2009) contains 1950 entries;
- bioinf.man.ac.uk/dbrowser/PRINTS/
Practical application : PRINTS (cont.)

```
261 271 281 291
--- --- --- ---
SHEKEEMAALKNLKHMRAQAAAGSAMX-L
SHEKEEMAALKNLKHMRAQAAAGSAMX-L
RAVAAQ--QQSKSTQ--KAE-REV
KBAAR--QQQSAATTQ--KAE-REV
AHEKQLQRAKMMVNASLRAAMQQQSAACR-L
RAVAK--QQQSAATTQ--KAE-KEV
KBAAA--QQQSAATTQ--KAE-KEV
KAAA--QQQSAATTQ--KAE-KEV
RAVAK--QQQSAATTQ--KAE-KEV
AHEKNNREQAAMNVALGSAANQGTSABCK-L
KLAAK--AQADSASTQ--KAE-KEV
AHEKQLQRAKMMVNASLRAAMQQQSAACR-L
AHEKGMRDQAKKMGLSK-RNEAQKTSACR-L
AHEKNMREQAAMNVALGSSRN-QNTSACR-L
FHEKALFRAKMMVBSLRNVDKSTAIIR-I
RAVAA--QQQSKSTQ--KAE-KEV
RAVAA--QQQSAATTQ--KAE-KEV
HQVAQ--QQQSAATTQ--KAE-KEV
KAAA--QQQSAATTQ--KAE-KEV
AHEKNNREQAAMNVALGSSARN-QNTSACR-L
SAHEKAMREQAAMNVLKSSRSE-AKSAEGR-K-L
KAAA--QQQSAATTQ--KAE-KEV
FSHEKALFRAKMMVBSLRNVDKSTAIIR-I
RTVAQ--QQQSKSTQ--KAE-KEV
KAAA--QQQSAATTQ--KAE-KEV
RAVAA--QQQSAATTQ--KAE-KEV
HQVAQ--QQQSKSTQ--KAE-KEV
SAHEKAMREQAAMNVLKSSRSE-ANTSAACR-L
KAAA--QQQSAATTQ--KAE-KEV
FSHEKALFRAKMMVBSLRNVDKSTAIIR-I
SAHEKAMREQAAMNVLKSSRSE-AKSAEGR-K-L
KAAA--QQQSAATTQ--KAE-KEV
RAAA--QQQSAATTQ--KAE-KEV
FSHEKALFRAKMMVBSLRNVDKSTAIIR-I
RAVAA--QQQSAATTQ--KAE-KEV
RAVAR--QQQSAATTQ--KAE-KEV
AHEKAMREQAAMNVLKSSRSE-CDKSAEGR-L
RAAA--QQQSAATTQ--KAE-KEV
KAAA--QQQSAATTQ--KAE-KEV
SAHEKAMREQAAMNVLKSSRSE-AKSAEGR-K-L
AHEKAMREQAAMNVLKSSRSE-CDKSAEGR-L
RAVAA--QQQSKSTQ--KAE-KEV
KBAAR--QQQSAATTQ--KAE-KEV
```
Practical application: **PRINTS** (cont.)

```
321 331 341 351
- --------------- --------------- --------------- ----------------
QPGPEWTVYAPAASLAPLFAKASAIHNPIYGVSHFQPI
QPGPEWTVYAPAASLAPLFAKASAIHNPIYGVSHFQPI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
FTHQGSEFGDVFMTLPFAFACKTSAYNIDYIDMNQPI
VPSGSLRTLPATWSVFAKANCRPNSIPYIVGISHPYI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
FTHQGSGFNPFGTFMTLPFAFACKTSAYNIDYIDMNQPI
FTHQGSGFNPFGTFMTLPFAFACKTSAYNIDYIDMNQPI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
IFPT-VKINPLPTILWSLFAKANVFYNPIYTVMNPQPI
VSHPGHEFDLRAIMPSLGSKASTYVNPYPYLMNQPI
VPSGSLRTLPATWSVFAKANCRPNSIPYIVGISHPYI
MFAR-YSLSPTYTYWGYVFKANCRPNSIPYIVGISHPYI
IFNL-VKISPLETTSWSLFAKANCRPNSIPYIVGISHPYI
AFGDKSLLTPGATMIPACTYKLVAIDDFVPYIAHSPYI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
VNRNHRGVDLRVLVTPFAFCKSAIYNPIYIMNQPI
AVNPYGAVHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
FFNKGADSFSAKFRMAITAFAFCSKSAIYNPIYTVYLMNQPI
IFPT-TKISPLETTWSLFAKANCRPNSIPYIVGISHPYI
LQFK-GLTLRPLTIFGWYVFKANCRPNSIPYIVGISHPYI
FFNKGADSFSAKFRMAITAFAFCSKSAIYNPIYTVYLMNQPI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
VNRNHRGVDLRVLVTPFAFCKSAIYNPIYIMNQPI
AVNPYGAVHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
IFPT-WKISPLETTWSLFAKANCRPNSIPYIVGISHPYI
LQFK-GLTLRPLTIFGWYVFKANCRPNSIPYIVGISHPYI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
VNRNHRGVDLRVLVTPFAFCKSAIYNPIYIMNQPI
AVNPYGAVHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
IFPT-WKISPLETTWSLFAKANCRPNSIPYIVGISHPYI
LQFK-GLTLRPLTIFGWYVFKANCRPNSIPYIVGISHPYI
AANPGYAFHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
VNRNHRGVDLRVLVTPFAFCKSAIYNPIYIMNQPI
AVNPYGAVHPPLAAALFYAFYAKSATIYNPIYVPNRMQPI
```

Marcel Turcotte  
CS15126. Algorithms in bioinformatics
Practical application: PRINTS (cont.)

<table>
<thead>
<tr>
<th>361</th>
<th>371</th>
<th>381</th>
<th>391</th>
<th>401</th>
<th>411</th>
</tr>
</thead>
<tbody>
<tr>
<td>HENLVAVVAVVER-QEGRERDAAMMKPAHNAQVREK</td>
<td>IASNFPLTQYKDEKTDKFAARITPAGVEQGGKTDAAMMKPAHNAQVREK</td>
<td>MOLPGKVKVEDGE----ASTTTSEVSSVSSVSSVAPA</td>
<td>MITTCCGKNFEEGEE-GASTTASSVSSVSSVSSVAPA</td>
<td>LYQRFPSLACGSGESDVKEASATTMREEKFPEA</td>
<td>ILQKPGKVKVEDGE----LSSAKTEVSVSSVSSVAPA</td>
</tr>
<tr>
<td>MLTTICGKNFGEED---ASAMKRTTSTQ------VAPA</td>
<td>MVTTICGKNFGEED---ASTTVSTTSTQ------VAPA</td>
<td>ILQLPGKVKVEDGE----LSSATKTEVSVSSVSSVAPA</td>
<td>VLPQRFSPALASSPGADTTSTTTVCTGETKEFAA</td>
<td>MMKV---CGKNOEDEA----STSSQTVQSVSSVAPA</td>
<td>MMKV---CGKNIEDEA----STSSQTVQSVSSVAPA</td>
</tr>
<tr>
<td>MEKKLPCSLKKT-ESDDVSEASASTTSAAEKAESA</td>
<td>LFQKPSLACGGSDESVKEASATTMREEKFPEA</td>
<td>LQKRPWVLGVEKKSAS---AQSTTTQRTQQTGEFAA</td>
<td>LFQKPSLACTTTEPTGATMSTTTTHTVTGEKFAA</td>
<td>AKKKCQCVGFVKGVDGKESE-AQSQATSSAEASKA</td>
<td>MLTITCGKNFGEDE--SSTVSTKTEVSVSSVAPA</td>
</tr>
<tr>
<td>FAKPSLACGGSDESVKEASATTMREEKFPEA</td>
<td>LQKRPWVLGVEKKSAS---AQSTTTQRTQQTGEFAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
<tr>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
<td>LQKRPWVLAMEKAEPE-SSAVASTSTTTQPQQTAA</td>
</tr>
</tbody>
</table>

Marcel Turcotte  
CSI5126. Algorithms in bioinformatics
The degree of conservation along a multiple sequence alignment (MSA) varies;

An MSA often consists of a number of blocks with a high degree of conservation, interspersed by more variable regions;

Each entry in PRINTS consists of a collection of ungapped, unweighted local alignments;

In PRINTS, 3 conserved segments of the OPSIN alignment serve to represent the OPSIN motif.
Visual pigments are the light-absorbing molecules that mediate vision [1,2]. They comprise an apoprotein (opsin), covalently linked to the chromophore cis-retinal. Vision is effected through the absorption of a photon by the chromophore, which is isomerised to the all-trans form, promoting a conformational change in the protein.
PRINTS Entry/Diagnostic

SUMMARY INFORMATION

123 codes involving 3 elements
7 codes involving 2 elements

COMPOSITE FINGERPRINT INDEX

3| 123 123 123
2| 5 3 6

True positives:
OPSD_CHICK  OPSD_CANFA  OPSD_TRIMA  OPSD_RABIT
OPSD_MOUSE  OPSD_CRIGR  OPSD_PIG   OPSD_MACFA

....
## PRINTS Entry/Motifs

**OPSIN1**  
Length of motif = 13  
Motif number = 1

<table>
<thead>
<tr>
<th>Opsin motif I - 1</th>
<th>PCODE</th>
<th>ST</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>YVTQHKKLRTPL</td>
<td>OPSD_BOVIN</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>YVTQHKKLRTPL</td>
<td>OPSD_HUMAN</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>YVTQHKKLRTPL</td>
<td>OPSD_SHEEP</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>AATMFKKLRHPL</td>
<td>OPSG_HUMAN</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>AATMFKKLRHPL</td>
<td>OPSR_HUMAN</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>YIFATKSLRTPA</td>
<td>OPS1_DROME</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>VATFRKLRQPL</td>
<td>OPSB_HUMAN</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>YIFGKTSLRTPA</td>
<td>OPS2_DROME</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>WVFSAAKLRTPS</td>
<td>OPS3_DROME</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>WIFSTSKLRTPS</td>
<td>OPS4_DROME</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>YLFSTKSLQTPA</td>
<td>OPSD_OCTDO</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>YLFSTKSLQTPA</td>
<td>OPSD_LOLFO</td>
<td>57</td>
<td>57</td>
</tr>
</tbody>
</table>

...
### PRINTS Entry/Motifs

**OPSIN2**  
Length of motif = 13  
Motif number = 2

Opsin motif II - 1

<table>
<thead>
<tr>
<th>Motif</th>
<th>PCODE</th>
<th>ST</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWSRYIPEGMQCS</td>
<td>OPSD_BOVIN</td>
<td>174</td>
<td>101</td>
</tr>
<tr>
<td>GWSRYIPEGLQCS</td>
<td>OPSD_HUMAN</td>
<td>174</td>
<td>101</td>
</tr>
<tr>
<td>GWSRYIPQGMCQS</td>
<td>OPSD_SHEEP</td>
<td>174</td>
<td>101</td>
</tr>
<tr>
<td>GWSRYWPHGLKTS</td>
<td>OPSG_HUMAN</td>
<td>190</td>
<td>101</td>
</tr>
<tr>
<td>GWSRYWPHGLKTS</td>
<td>OPSR_HUMAN</td>
<td>190</td>
<td>101</td>
</tr>
<tr>
<td>GWSRYVPEGNLTS</td>
<td>OPS1_DROME</td>
<td>187</td>
<td>101</td>
</tr>
<tr>
<td>GWSRFIPEGLQCS</td>
<td>OPSB_HUMAN</td>
<td>171</td>
<td>101</td>
</tr>
<tr>
<td>GWSAYVPEGNLTA</td>
<td>OPS2_DROME</td>
<td>194</td>
<td>101</td>
</tr>
<tr>
<td>TWGRFVPEGYLTs</td>
<td>OPS3_DROME</td>
<td>194</td>
<td>100</td>
</tr>
<tr>
<td>FWDRFVPEGYLTs</td>
<td>OPS4_DROME</td>
<td>190</td>
<td>100</td>
</tr>
<tr>
<td>NWGAYVPEGILTS</td>
<td>OPSD_OCTDO</td>
<td>174</td>
<td>103</td>
</tr>
<tr>
<td>GWGAYTLEGVLNC</td>
<td>OPSD_LOLFO</td>
<td>173</td>
<td>103</td>
</tr>
</tbody>
</table>

...
### PRINTS Entry/Motifs

**OPSIN3**  
Length of motif = 13  
Motif number = 3

Opsin motif III - 1

<table>
<thead>
<tr>
<th>Motif</th>
<th>PCODE</th>
<th>ST</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIFMTIAPAFFAKT</td>
<td>OPSD_BOVIN</td>
<td>285</td>
<td>98</td>
</tr>
<tr>
<td>PIFMTIAPAFFAKS</td>
<td>OPSD_HUMAN</td>
<td>285</td>
<td>98</td>
</tr>
<tr>
<td>PIFMTIAPAFFAKS</td>
<td>OPSD_SHEEP</td>
<td>285</td>
<td>98</td>
</tr>
<tr>
<td>PLMAALPAFFAKS</td>
<td>OPSG_HUMAN</td>
<td>301</td>
<td>98</td>
</tr>
<tr>
<td>PLMAALPAYFAKS</td>
<td>OPSR_HUMAN</td>
<td>301</td>
<td>98</td>
</tr>
<tr>
<td>PLNTIWGACFAKS</td>
<td>OPS1_DROME</td>
<td>308</td>
<td>108</td>
</tr>
<tr>
<td>LRLVTIPSFFSKS</td>
<td>OPSB_HUMAN</td>
<td>282</td>
<td>98</td>
</tr>
<tr>
<td>PLTTIWGATFAKT</td>
<td>OPS2_DROME</td>
<td>315</td>
<td>108</td>
</tr>
<tr>
<td>PGATMIPACACKM</td>
<td>OPS3_DROME</td>
<td>317</td>
<td>110</td>
</tr>
<tr>
<td>QGATMIPACTCKL</td>
<td>OPS4_DROME</td>
<td>313</td>
<td>110</td>
</tr>
<tr>
<td>PYAAELPVLFAKA</td>
<td>OPSD_OCTDO</td>
<td>295</td>
<td>108</td>
</tr>
<tr>
<td>PYAAQLPVMFAKA</td>
<td>OPSD_LOLFO</td>
<td>294</td>
<td>108</td>
</tr>
</tbody>
</table>

...
Deriving a motif

- “(...) from a small multiple sequence alignment, conserved motifs are identified and excised **manually** for database searching (...)”;
- “Results are examined **manually** (...)”;
- “(...) if there are more matches than were in the initial alignment, the additional information from these new sequences is added to the motifs.”;
- “(...) the database is searched again.”;
- “This iterative process is repeated until no further complete fingerprint matches can be identified.”
PRINTS : Summary

Pros :

- Since raw alignments are stored, they can be used to derive regular expressions, profiles, etc.;
- **High signal-to-noise ratio** (curated database);
- Combination of local motifs together with the iterative process helps detecting more remote homologues.

Cons :

- **Human intervention (construction/interpretation)** high;
- Lack of a theory for composite motifs.
Substring Motifs: Cons

- Selecting **appropriate parameters**: number of mismatches, gap penalty, etc.;
- **Pairwise sequence comparison might not be applicable**: sequences do not align on their entire length or are too divergent;
- Sometimes we would like to emphasize that certain **identities are mandatory**. (See WW domain for instance)
Motifs: **Regular Expressions**

**Regular expressions** are often used to represent key residues composing a motif.

A large database of regular expressions exists: **PROSITE**.

Methods have been developed to derive automatically PROSITE signatures: see **PRATT** (Pattern driven) and **eMOTIF** (data driven).

⇒ Consult the appendix for a brief summary of regular expressions and finite state automaton.
How to?

Most of the regular expressions found in PROSITE have been created by hand.

- Build a multiple alignment;
- Reduce the alignment to a consensus regular expression;
- Refine the expression base database search results.

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Regular expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADLGAVFALCDRYFQ</td>
<td>[AS]-D-[IVL]-G-x4-{PG}-C-[DE]-R-[FY]2-Q</td>
</tr>
<tr>
<td>SDVGPRSCFCERFYQ</td>
<td></td>
</tr>
<tr>
<td>ADLGRTQNRCDRYYQ</td>
<td></td>
</tr>
<tr>
<td>ADIGQPHSLCERYFQ</td>
<td></td>
</tr>
<tr>
<td>* * * * *</td>
<td></td>
</tr>
</tbody>
</table>
How to? (cont.)

Prosite     Perl

\{PG\}     \[^PG]\]
x4         \{4\}

"-" are simply spacers.

⇒ http://www.expasy.ch/prosite/
Sometimes, such patterns are published (might not be in the form of a regular expression, but as a list of functionally important residues and their spacing);

- Starts with a group or family of sequences;
- Identify regions of the alignment that are important for function, ideally these are supported by experimental evidences, such as: enzyme catalytic site, prosthetic group (heme, etc.) attachment sites, metal binding sites, disulfide bonds, binding a molecule (ATP, Calcium, DNA, etc.);
- Identify core residues in the region, < 4 or 5 conserved residues, scan a sequence database with the core pattern, normally this would also match non-members, then the pattern is further extended.
Experimental data might suggest that the histidine participates to the active site, a first pattern is constructed $\text{ATH[DE]}$, which is used to scan a sequence database, if no false positive, then fine, otherwise extend pattern, may involve starting from a new core pattern.
URL = www.expasy.ch/prosite

Approximately 146Mb, updated twice per year.

Typically, a rule involves 10-20 conserved residues.

Pros/Cons:
- Biased towards sensitivity at the expense of specificity *(many false positives)*;
- Documented (biological properties of the family/domain);
- Maintained;
- Tightly linked to the development of SwissProt.

⇒ Now also part of InterPro: www.ebi.ac.uk/interpro.
Prosite Motifs

What are they?

- Short universal motifs:
  - N-glycosylation site $N-\{P\}-[ST]-\{P\}$
  - Phosphorylation site $[ST]-x-[RK]$  
  - Another phosphorylation site $[ST]-x(2)-[DE]$ 
  - Asp or Asn hydroxylation site $C-x-[DN]-x(4)-[FY]-x-C-x-C$

- Some have a structural basis, WW, helix-turn-helix;
- Families.

⇒ How many hits for the pattern $N-\{P\}-[ST]-\{P\}$ would occur by chance when matched against SwissProt? SwissProt is a popular sequence database.
SCOP: Protein Structure Classification

⇒ Class
  ⇒ Fold
    ⇒ Superfamily
      ⇒ Family
        ⇒ Domain

PROSITE Matches/SCOP

6 % Universal, phosphorylation, amidation, etc.
17 % Specific to a class.
8 % Specific to a fold.
17 % Specific to a superfamily.
12 % Specific to family.
40 % Specific to a sub-set of a family.
Automated approaches

Issues related to automated pattern discovery:

- Search space
  - Valid regular expressions
- Algorithm
  - Pattern driven (PRATT)
  - Data driven (eMOTIF)
- Evaluation function (a measure of surprise)
Preliminaries: **information theory**

The **information content** measures the reduction of the **uncertainty** (also called **entropy**) after some message has been received. In the case of regular expression motifs, the interpretation is “how much information is gained by knowing that a sequence segment matches a given regular expression”.

Merriam-Webster Online about “entropy”:

1: ... usually considered to be a measure of the system's disorder ...

3: Chaos, disorganization, randomness.
Uncertainty

Information is based on the notion uncertainty about an event — what symbol do you expect to find at a given position of the sequence?

Uncertainty is defined as follows,

\[ H = - \sum_{i=1}^{M} P_i \log_2 P_i \]
Consider a sample space that has **two outcomes**, one occurring with probability $p$, and the other outcome occurring with probability $1 - p$. 
Uncertainty (cont.)

The above picture shows how the entropy varies as a function of $p$. 
Uncertainty (cont.)

In particular, you can clearly see that the entropy is maximum when the events are all equiprobable, its value is then \( \log_2 M \) bits, where \( M \) is the number of outcomes (the cardinality of the sample space \( M = |S| \)). Here, the entropy maximum is \( \log_2 2 = 1 \) bit.

Notice also that the entropy approaches zero, whenever the probability of one of the events approaches 1 (and hence, the probabilities of the other events approach 0).

This models quite well the concept of uncertainty (entropy). When all the outcomes are equiprobable you can’t predict the result of an experiment, but any bias towards one of the outcomes reduces the uncertainty.
Uncertainty (cont.)

Then entropy is maximal when the $M$ outcomes are equally likely, and zero when only one outcome out of $M$ occurs.

Consider the case where all the outcomes are equiprobable, $P_i = \frac{1}{M}$ for all $i \in 1 \ldots M$.

$$- \sum_{i=1}^{M} P_i \log_2 P_i =$$

$$- \sum_{i=1}^{M} \frac{1}{M} \log_2 \frac{1}{M} =$$

$$- M \times \frac{1}{M} \log_2 \frac{1}{M} =$$

$$- \log_2 \frac{1}{M} = \log_2 M$$
Finally, consider the case where one outcome occurs with probability 1, and the other \( M - 1 \) outcomes occur with probability 0.

\[
-(1 \times \log_2 1 + \sum_{i=1..M, P_i \neq 1} 0 \times \log_2 0) \\
-(1 \times 0 + \sum_{i=1..M, P_i \neq 1} 0)
\]

the uncertainty is zero as expected.

\( \Rightarrow \) It is customary to let \( 0 \log 0 = 0 \).
Information content

The information content is defined as,

\[ I = H_{\text{before}} - H_{\text{after}} \]

i.e. the difference of entropy between two probability distributions.
Information content

Consider the case of DNA strings, \( \Sigma = \{A, C, G, T\} \), where all four bases are equiprobable, i.e. \( P_i = 0.25 \).

Considering a wild card, [ACGT], no information is gained.

\[
I = H_{\text{before}} - H_{\text{after}} = \left( - \sum_{1}^{4} \frac{1}{4} \times \log_2 \frac{1}{4} \right) - \left( - \sum_{1}^{4} \frac{1}{4} \times \log_2 \frac{1}{4} \right) = 2 - 2 = 0
\]
When a regular expression contains a single character, say C, then the amount of information gained is maximal $\log_2 4 = 2$ bits.

$$I = H_{\text{before}} - H_{\text{after}} = 2 - 0 = 2$$
Information content

In the case of a **character class** containing two elements, \([AG]\),

\[
l = \left( - \sum_{1}^{4} \frac{1}{4} \times \log_{2} \frac{1}{4} \right) - \left( - [2 \left(\frac{1}{2} \log_{2} \frac{1}{2}\right) + 2(0 \log_{2} 0)] \right)
\]

1 bit of information is gained.
Information content

The information content for a regular expression will be the sum of the information content at each position,

\[ I_{G[GA]C[ACGT]} = 2 + 1 + 2 + 0 = 5. \]
Exercise

Consider an organism whose genome has the following nucleotide frequencies: \( P_A = \frac{1}{6}, P_C = \frac{1}{3}, P_G = \frac{1}{3}, P_T = \frac{1}{6} \). Calculate the information content of the following expression \( G[GA]C[ACGT] \).

\[
I_{G[GA]C[ACGT]} = I_G + I_{[GA]} + I_C + I_{[AGT]}
\]
Comparing motifs, signals, active sites, etc.

40 yeast TATA sites

The -10 region of 350 E. coli promoters

- www.lecb.ncifcrf.gov/˜toms/sequencelogo.html
- weblogo.berkeley.edu
FYI: Claude Shannon – Father of the Information Age

Half hour video presenting Claude Shannon’s work. “This fascinating program explores his life and the major influence his work had on today’s digital world through interviews with his friends and colleagues.” (includes comments from Andrew Viterbi, Ian Blake, and others)

- www.ucsd.tv/search-details.asp?showID=6090
- cm.bell-labs.com/cm/ms/what/shannonday/paper.html
Motifs: Regular Expressions

Regular expressions are often used to represent key residues forming motifs.

A large database of regular expressions exists: PROSITE.

Methods have been developed to derive automatically PROSITE signatures: see PRATT and eMOTIF.
Things we like about REs!

Allow to model mandatory amino acids.

Easy to interpret in terms of biological concepts, such as binding sites, etc.
Issues

- **Human intervention high**, often derived from literature;
- **Subjective choice** of the region in some cases;
- Entries must be revised as **new sequences become available**;
- Too rigid! **Does not allow for mismatches**;
- Compromise between sensitivity/sensibility, flexibility/noise;
- Will not perform well on new entries (**overfitting**);
- Short motifs can **occur by chance**.
Pattern Discovery

Approaches to derive patterns automatically can be classified as “pattern driven” (PRATT) or “data (sequence) driven” (eMOTIF).

Issues:

- Search algorithm;
- Performance measure or fitness function.
Pattern driven approaches (PRATT [1])

**Input**: A set of related but unaligned sequences.

**Problem**: Constructs automatically regular expressions (patterns) consisting of single letter (A), character classes ([KER]) and range patterns (x-(i,j)).

For example,

A-x-[KER]-x(2)-D-[ILV]-E-x(4)-[KR]

Based on an exhaustive search from the most general motifs to the most specific ones.

This is done in **two steps**:

- Single letter patterns search, A-x(4)-D-x-E;
- Pattern refinement,
  A-x-[KER]-x(2)-D-[ILV]-E-x(4)-[KR].

www.ii.uib.no/~inge/Pratt.html
Step 1: Single Letter Pattern Search

Starting with the empty pattern (most general motif) all possible extensions of a motif are considered.

The process is repeated recursively unless a pattern does not match the required minimum number matches $c$ (coverage, support).

This is a tree-based search with pruning based on coverage.

Specifically, a regular expression $\alpha$ (corresponding to a node of the search tree) is extended with all the possible suffixes of the form $-x(i,j) - \beta$ for $0 \leq i \leq j \leq t$ and $\beta \in \Sigma$.

Notice that $i$ and $j$ can both be of length zero, which corresponds to an extension of a single letter.

For some small $t$ and large $c$ it’s possible to exhaustively search the space of all possible motifs.
Step 1: Single Letter Pattern Search (cont.)

\[ \alpha \]

\[ \alpha-x(i,j)-\beta \]

\[ \alpha-x(0,0)-A \]
\[ \alpha-x(0,0)-C \]
\[ \alpha-x(0,0)-G \]
\[ \alpha-x(0,0)-T \]
\[ \alpha-x(0,1)-A \]
\[ \alpha-x(0,1)-C \]
\[ \alpha-x(0,1)-G \]
\[ \alpha-x(0,1)-T \]
\[ \alpha-x(1,1)-A \]
\[ \alpha-x(1,1)-C \]
\[ \alpha-x(1,1)-G \]
\[ \alpha-x(1,1)-T \]

...
**Step 1 : Single Letter Pattern Search (cont.)**

Why is the introduction of the character classes delayed until the refinement step?

Character classes are not introduced earlier because there are too many of them!

How many?

\[ 2^{|\Sigma|}. \]

Consider the case where \( \Sigma \) represents all 20 amino acids

\[ 2^{20} = 1,048,576. \]

Since the extensions represent the branching factor of the tree-based search, it cannot be afforded.
Step 1 (contd)

Given the pattern $P$, the children of $P$ are as follows.

$$P-x(0,0)-A \ldots P-x(0,0)-Y$$
$$P-x(0,1)-A \ldots P-x(0,1)-Y \ldots P-x(0,5)-A \ldots P-x(0,5)-Y$$
$$\ldots$$
$$P-x(5,5)-A \ldots P-x(5,5)-Y$$

The resulting patterns are checked against the set of sequences and retained if they match enough sequences.
Wildcard positions, $x(i, j)$, such that $i = j$ are considered for refinement.

Nota: The information that I could obtained is vague.

As far as I understand, the list of groups is supplied by the user (a nice way to derive groups will be presented along with the presentation on eMOTIF [2]).

In the meantime, you can image that the groups are obtained from the Venn diagrams based on the properties of amino acids, tiny$=\{SGA\}$, small$=\{SGAPTNCV\}$, ...

For a given pattern, all the sequences that it matches are retrieved.

For all the positions $k$ of all the range patterns $-x(i, i)$ find a group, if it exists, that is a superset of the amino acids found this position.
Step 2: pattern refinement (cont.)

- hydrophobic
- aromatic
- aliphatic
- small
- tiny
- positive
- polar
- charged
- negative
Example. Consider the following user defined groups.

[FAMILYVW] [KREND] [PGSTQ] [HC]

The expression $C-x(3,3)-C$ matches the following three sequences.

CDFGC
CEIMC
CRIMC

The amino acids at the second position are a subset of the group [KREND], those at the third position are a subset of the group [FAMILYVW], but there are no groups containing M and G. The following expressions can be derived $C-x(3,3)-C$, $C-[KREND]-x(2,2)-C$, $C-x(1,1)-[FAMILYVW]-x(1,1)-C$ and $C-[KREND]-[FAMILYVW]-x(1,1)-C$. 
Step 2: pattern refinement (cont.)

Given \( k \) wild cards, \( 2^k \) expressions can derived (not all of them will have the minimum coverage), “a heuristic refinement algorithm” is used.
Scoring Pattern

PRATT has three scoring schemes:

- Positive Predictive Value (PPV) (requires a set of negative examples)
- Information Content (default)
- Minimum Description Length (MDL) (takes into account the number of matches and the complexity of the motif)

Alternatively, a Z-score (aka standard score, normal score) could be used as a measure of surprise,

\[
z(w) = \frac{f(w) - E(w)}{N(w)}
\]

where \(f(w)\) is the number of observed occurrences, \(E(w)\) is the expected number of occurrences, and \(N(w)\) is a normalization factor.
Pros:

- Automated approach;
- Uses unaligned sequences.

Cons:

- Unsatisfactory solution to the over-fitting problem.
Data driven approaches (eMOTIF)

- Automatically defined motifs;
- Strategies to overcome the rigidity of REs:
  - Classes of amino acids;
  - Regular expressions with approximate matching; agrep (allow 0, 1, 2, 2 or 4 mismatch(es));
  - Variable specificity.

The eMOTIFS are derived from the multiple sequence alignments in the BLOCKS+ database, the PRINTS database, and the eBLOCKS database.

Originally constituted of 50,000 motifs from 7,000 alignments.

⇒ motif.stanford.edu
**Input data**

```
MFRRKAFHWYTGEMEGMEFTEAESNMNDPVAEYQQY
MFRRKAFHWYTGEMEMEFTAEASENMNDPVAEYQQY
MFGKRAFVHYVPEGMEENEFTDARQDLYELEVDAQNL
MFKKRAFVHWYVPEGMEFACEARENIAVLERDFEEV
MFRRKAFHWYTGEMEFTAEASENMNDLAVSEYQQY
MFRRKAFHWYTGEMEFTEVRAVMNDLVAEYQQY
MFRRKAFHWYTGEMEFTAEASENMNDLAVSEYQQY
MFRRKAFHWYTGEMEFTAEASENMNDLAVSEYQQY
MFRRKAFHWYTGEMEFTAEASENMNDLAVSEYQQY
MFRRKAFHWYTGEMEFTAEASENMNDLAVSEYQQY
MFRRKAFHWYTGEMEFTAEASENMNDLAVSEYQQY
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
MYAKRAFVHYVPEGMEGEFSEAEDLALEEKDYYEEV
```
Creating a Motif: *ad hoc*

Each position consists of a character class that contain all the observed amino acids at that position. The motif for that block would start with M[FY][AGKR].
Creating a Motif: *ad hoc* (cont.)

```
MFRRKAFHWYTGEFMDEMENFTEAESNMNDPVAEYQQY
MFRRKAFHWYTGEFMDEMENFTEAESNMNDPVAEYQQY
MFKKAFWYVYVGYGEMEAFEYRANvlDVRFFE EV
MFKRKAFHWYTGEFMDEMENFTEAENMDLVSEYQQY
MFKRKAFHWYTGEFMDEMENFTEVRELMDALYQQY
MFKRKAFHWYTSEGINDELFSEAEENMSNDLVSEYQQY
MFKRKGFHWYTGEFMENPVEFSEASEQSLDILLEYQQY
MFRRKAFHWFTGEFMDEMENFTEAENMSNDLVSEYQQY
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MYAKRAVHYVYVGYGEMEFEADRLAEKDYEEV
MFRRKAFHWFTGEFMDEGFSEAEADIALAEKDFEEY
YGRG V YVS E L T VKENLDPISYEYQQQL
K MQ MNE VL V
R N RS VY A
V R
V
```
What do you think?
Remarks

The *ad hoc* motif is too specific. For example, position 3 contains amino acids that have nothing in common.

Evolution does not constrain this position.

It can be expected that most mutations at that position would be tolerated; including mutations to an amino acid type other than \[\text{[AGKR]}\].

Because RE are deterministic (match/not match), several true positive will be missed.

Over-fitting problem!
eMOTIF : Substitution groups

**Input** : Columns of multiple sequence alignments from BLOCKS and HSSP.

Of all $2^k$ subsets, select all the groups with the following properties:

1. All the amino acids from this group substitute frequently with other amino acids from the same group (**compactness**);

2. All the amino acids that are not part of the group substitute with members of the group with low frequencies (**isolation**);

20 groups were found.
$$eMOTIF : \textbf{Substitution groups}$$

⇒ Amino acids of the same group are more likely to substitute for one another.
Creating a Motif: most specific motif

Each position consists of the most specific substitution group that contains all the amino acid types observed at that position.

Observation. For a given set of input sequences the most specific motif is unique.
Creating a Motif: most specific motif (cont.)

MFRRKAFHLHWYTGEVMDEMEEFTEAESNMNDPVAYQY
MFGKAFVHHYVGEFMEEFTEASNMNDLYELEVDYANL
MFKKKAFVHWGEMEFEFTEARENNVLERDFEEV
MFRRKAFHLYWYTGEVMDEMEEFTEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL
MFRRKAFHLYWYTGEVMDEMEEFTEARENNVLERDFEEV
MFRRKAFHLYWYTSEMEDELFSEAESNMNDLYELEVDYANL

MF.KKAFIFHWF..EGMDE.EFSE.E.DI......DFEEF  
Y RR L Y E T K N L EYQQI  
V Q M L  
R V V  
Y

Marcel Turcotte
CSI5126. Algorithms in bioinformatics
Remarks

Consider position 3, there is no group that contains “G”, “R”, “S” and “A”, therefore a wild-card is inserted at that position.

Consider position 8, although “I” is not observed at that position, we can expect that other members of this family would have an “I” at that position since “L” and “V” often substituted by “I”.

The most specific motif is more general than the ad hoc motif.

The most specific motif is sensitive to noise, consider the 8th position from the right, all the sequences have an “L” at that position but the first one has a “P”. This could be the result of an experimental error. However the most specific motif will have a wild-card because of that.

Consequently, the RE may be too general and will produce many false positive results!
Exploring the space of RE motifs

Because some RE may be too general and will produce many false positive results, we would like to explore the space of possible REs for finding new ones that are more specific (but also cover fewer sequences).
Coverage/Sensitivity

eMOTIF proposes an ensemble of motifs with different coverage and sensitivity.

⇒ Ideal motif would be found in the bottom-right corner.
eMOTIF exhaustively generates all possible motifs using the allowable substitution groups.
Probability that a motif matches a random sequence

Assumptions: AA are independent and identically distributed.
AA distribution estimated from the observed frequencies from SWISSPROT.

\[ P(M[FWY].[KR] \ldots [FYW]) \]

\[ p(M) \times [p(F) + p(W) + p(Y)] \times 1 \times [p(K) + p(R)] \ldots [p(F) + p(W) + p(Y)] \]

Wild card characters (.) matches with probability 1.
The amino acids probabilities are estimated from a large database.
Choosing the right RE

When using an RE for detecting new members of a sequence family, the expected number of random sequences matching the RE should be less than 1.

The expected number of matches depends on the size of the database!

\[ P_{RE} \times N \]

where \( N \) is the size of the database.

You should select an RE with probability less \( \frac{1}{N} \).

Obviously, such RE will match fewer sequences!
Disjunction of REs can be used to represent a family

Find an RE with probability $\frac{1}{N}$ of matching a random sequence for a database of size $N$.

Remove all the sequences that it matches and apply the algorithm to the remaining sequences.

A family is therefore represented by a disjunction of REs (high specificity and coverage).

**AKA sequential covering in machine learning.**
Size of the space of motifs

The space of all the possible motifs is huge: \((m + 20)^n\), where \(m\) is the number of character classes that are used to construct the motifs and \(n\) is the number of columns, e.g. \((20 + 20)^{38} \approx 10^{60}\).
Exploring the space of all possible motifs: Solution 1

Each subset of sequences induces a most specific motif.

Let’s generate the most specific motif for all the subsets of the input sequences.

For 10 sequences there are $1,024 (= 2^{10})$ most specific motifs, which is much less than $((20 + 20)^{10} \approx 10^{16}$.

The number of motifs is independent of the number of columns and the number of groups!

However, for 158 sequences, there are $10^{48}$ subsets ...
Observation 1

Several REs select the same subset of sequences.

For example, at position 1, M, [MIVL], [MIVKF], [MIVKRY] and the wild-card all select the same subset (i.e. all the sequences).

Out the several motifs that select the same subset of sequences, eMOTIF records only the most specific one.
Observation 2

Many subsets induce the same RE.

1. AADACAAAA
2. AAAABAAAA
3. AAAACAAAA

AAAABAAAA
   D C

1. AADACAAAA
2. AAAABAAAA

AAAABAAAA
   D C

1. AADACAAAA
3. AAAACAAAA

AAAACAAAA
   D

2. AAAABAAAA
3. AAAACAAAA

AAAABAAAA
   C
Observation 3

An arbitrary motif matches a subset of the input sequence, and the subset of sequences induces a most specific (canonical) motif.
Observation 4

Not all subsets are interesting, typically only those that contain at least 30% of the input sequences.
Summary

- A given motif (RE) selects a subset of sequences.
- A given subset of sequences specifies a **most specific motif**.
- The most specific motif of a subset is called a **canonical motifs**.
- eMOTIFS explores the space of canonical motifs.
- In the Tubulin subuni example, $10^{56}$ possible motifs, $10^{48}$ subsets, yet only 39000 motifs that select different subsets!
eMOTIF : Algorithm

1. Starting at position $p = 1$ and $\text{set} = \text{all}$;
2. Record the most specific RE for the set;
3. For position $p$, find all the groups that match a minimum number of sequences (typically 30 %);
4. For all the of groups that match the same subset of sequences only keep the most specific one;
5. $p = p + 1$;
6. For each remaining group, set $\text{set}$ to this subset and goto 2;
eMOTIF : Algorithm (cont.)

```
MFRRKAF LHWYT GE GME EFT A E S N M D F A E Y Q Q Y
MFGRKAF VLHV YGE M E E NT D A R Q D L Y E L E V D Y A N L
MFKAF V H YV G E M E E GE F T A E R N I A V L E R D F E E V
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT E V R A N M D L V A E Y Q Q Y
MFRRKAF LHWYT SE G M E L F S A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE G M E P V F S A Q S D L E D L I E Y Q Q Y
MFRRKAF LHWFT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D E M T Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L M S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V A E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V H E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V Y E Y Q Q Y
MFRRKAF LHWYT S E G M E T E F A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRKAF LHWYT GE GME EFT A E S N M D L V S E Y Q Q Y
MFRRQAF LHWYT SE G M E T E F A E S N M D L V S E Y Q Q Y
MFRRQAF LHWYT GE G M E E G D F A E A D N V S D L L S E Y Q Q Y
MFVKRAVF VH VGE M E E G F A E A R D L L A L E K D Y E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E A D N V S D L L S E Y Q Q Y
MFVKRAVF VH VGE M E E G F A E A R D L L A L E K D Y E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
MYAKRAVF VH VGE M E E G F A E R D L A L E K D Y E E V
```

Marcel Turcotte

CSI5126. Algorithms in bioinformatics
eMOTIF: Algorithm (cont.)

position

1

2

3

4

vlim

\( m^{34} \)

\( f^{22} \)

\( y^{12} \)

\( fyw^{34} \)

ag

st

\( a^{10} \)

\( s^{8} \)

\( pagst^{12} \)

\( \ldots \)
eMOTIF : Algorithm (variant)

1. Starting at position $p = 1$ and $set = \text{all}$;
2. Record the most specific RE for the set;
3. For position $p$, find all the groups that match a minimum number of sequences (typically 30 %);
4. For all the of groups that match the same subset of sequences only keep the most specific one;
5. $p = p + 1$;
6. For each remaining group, if subset not visited, set subset to visited, set set to this subset and goto 2;
Selecting a cutoff

Use the pattern that has a maximum coverage for a specificity $(10^{-10}, 10^{-9}, 10^{-8} \ldots)$. 
Predictive accuracy : eMOTIF vs PROSITE

Experiment : 410 PROSITE motifs from 1991 were selected. 410 motifs are used to retrieve sequences, which are aligned, and the used to derived eMOTIFs. Collected sequences that have been determined after 1991 and used PROSITE and eMOTIFs to classify them. Those sequences were not used to derive PROSITE motifs or eMOTIFs.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROSITE</td>
<td>6,598</td>
<td>880</td>
<td>9,068</td>
</tr>
<tr>
<td>eMOTIFs</td>
<td>4,619</td>
<td>12</td>
<td>11,047</td>
</tr>
</tbody>
</table>

eMOTIF has a higher precision (few errors) but it is less sensitive (many true positive are missed). 70 times less false positive and 1.4 less false negative. Only 4 motifs were less precise. There were 100 RE that had same or better coverage than PROSITE.
Issues

Nice definition of substitution groups.
Good tradeoff specificity vs coverage.
Requires a multiple sequence alignment as input.
Still based on regular expressions.
Still yes or no answer!
sub formMotifs {
    my $column = shift;
    my @set = @_;

    if ($column >= $col) { return; }
    $re = findMostSpecificMotif(@set);
    my %subsets = ();
    for $group (@groups, '.') {
        @subset = findSubset($column, $group, @set);
        next unless @subset > ($match * $row);
        $subset = makeKey(@subset);
        next if defined $subsets{$subset};
        $subsets{$subset} = 1;
        formMotifs($column + 1, @subset);
    }
}
formMotifs(0, 0 .. ($row-1));
Issues (cont.)

sub findMostSpecificGroup {

    # naive implementation, will do for now

    my @as = @_;  
    my @xs = ();

    GROUP:
    foreach $group (@groups) {
        foreach $aa (@as) {
            next GROUP if index($group, $aa) == -1;
        }
        push @xs, $group;
    }
}
return "." unless @xs;
return "[".(sort { length($a) < length($b) } @xs)[0]."]";
}
sub findSubset {
    my $column = shift;
    my $group = shift;
    my @set = @_;  

    @subset = ();
    for $i (@set) {
        $aa = substr $seqs[$i], $column, 1;
        push @subset, $i if (index($group, $aa) > -1) || ($group eq '.');
    }
    return @subset;
}
sub makeKey {
    my @set = @_; 

    @key = ("0") x 256; 

    for $i (@set) {
        $key[$i] = "1";
    }

    $key = join "", @key;
    $key = pack "b256", $key;

    return $key;
}
Issues (cont.)
if ($groups eq "small") {
    @groups = qw(AG ST KR FWY HKR ILV ILMV EDNQ AGPST);
} else {
    @groups = qw(IV IVL MIVL MIVLF FY LFY IVLFY MIVLFY FWY YF RK);
}

push @groups, 'A';
push @groups, 'R';
push @groups, 'N';
push @groups, 'D';
push @groups, 'C';
push @groups, 'Q';
push @groups, 'E';
push @groups, 'G';
push @groups, 'H';
push @groups, 'I';
push @groups, 'L';
push @groups, 'K';
push @groups, 'M';
push @groups, 'F';
push @groups, 'P';
push @groups, 'S';
push @groups, 'T';
push @groups, 'W';
push @groups, 'Y';
push @groups, 'V';

# Sorted by length so that the most specific group is
# considered first
Issues (cont.)

@groups = sort { length($a) <=> length($b) } @groups;
Example

The WW domain is a protein module that binds proline-rich or proline-containing ligands. The WW domain is a protein-protein interaction module composed of 35-40 amino acids. It is the smallest, monomeric, triple-stranded, anti-parallel beta-sheet protein domain that is stable in the absence of disulfide bonds, cofactors or ligands.

- Two conserved tryptophans (W) spaced 20-22 amino acids apart;
- A block of two or three aromatic amino acids located centrally between the two signature tryptophans, and
- A conserved proline located three amino acids carboxyterminal to the second conserved tryptophan.

⇒ Bork and Sudol (1994), TIBS 19 (94), 531-533

 Marcel Turcotte

 CSI5126. Algorithms in bioinformatics
Representative structures of WW domains based on their sequences

Class 1: YAP65

Class 2: FBPWW28

Class 3: YQJ8WW

1. \(-\text{VLPAGWEMAKT-SSQRYFLNHIDQTFTWQDPKRAML}\)
2. \(-\text{GATAVSEWTEYKTA-DGKYYNNRTESTWEKPOELK-}\)
3. \(-\text{VRLPPGWEIIHE--NGRPLYYNAEQ-KTKLHYPPS}\)

Example (cont.)
## Representative sequences of the WW domain

<table>
<thead>
<tr>
<th>name/species</th>
<th>pos</th>
<th>sequence</th>
<th>acc.no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yap_Human</td>
<td>171</td>
<td>VPLPAGWEMAKTSS.GQRYFLNHIDQTTTWQDPRKAMLS</td>
<td>P46937</td>
</tr>
<tr>
<td>Yap_Chick-1</td>
<td>169</td>
<td>VPLPPGWEMAKTPS.GQRYFLNHIDQTTTWQDPRKAMLS</td>
<td>P46936</td>
</tr>
<tr>
<td>Yap_Mouse-1</td>
<td>156</td>
<td>VPLPAGWEMAKTSS.GQRYFLNHIDQTTTWQDPRKAMLS</td>
<td>P46938</td>
</tr>
<tr>
<td>Yap_Mouse-2</td>
<td>215</td>
<td>GPLPDGWEQAMTQD.GEVYYINHKNKTTSWLDPRLDPDF</td>
<td>P46938</td>
</tr>
<tr>
<td>Ned4_Mouse-1</td>
<td>40</td>
<td>SPLPPGWEEQVDIL.GRTYYVNHERSRTQWKRSPDQD</td>
<td>P46935</td>
</tr>
<tr>
<td>Ned4_Mouse-1</td>
<td>218</td>
<td>SPLPPGWEEQVDIL.GRTYYVNHERSRTQWKRSPDQD</td>
<td>P46934</td>
</tr>
<tr>
<td>Rsp5_Yeast-1</td>
<td>229</td>
<td>GRLPPGWERRTDFN.GRTYYVDHNRTRTVWKRPTLDQTE</td>
<td>P39940</td>
</tr>
<tr>
<td>Ned4_Mouse-2</td>
<td>196</td>
<td>SGLPPGWEKQDDR.GRSYYVHDHNSKTTSWKTQMDDP</td>
<td>P46935</td>
</tr>
<tr>
<td>Rsp5_Yeast-2</td>
<td>331</td>
<td>GELPSGWEQRFTP.EGRAYVDFHNRTRTVWDPFRQY</td>
<td>P39940</td>
</tr>
<tr>
<td>Ned4_Mouse-3</td>
<td>251</td>
<td>GPLPPGWEEHTHD.GRVFFINHNIKTQWEDPRLQNA</td>
<td>P46935</td>
</tr>
<tr>
<td>Rsp5_Yeast-3</td>
<td>387</td>
<td>GPLPPGWEMRILNT.ARVFYVHDHRTTVWDPRLSSL</td>
<td>P39940</td>
</tr>
<tr>
<td>Dmd_Human</td>
<td>3055</td>
<td>TSVQGPWERAISPKNVPYYINHETQTTCDHPKMTELY</td>
<td>P11532</td>
</tr>
<tr>
<td>Dmd/Torca</td>
<td>253</td>
<td>TSVQGPWERAISPKNVPYYINHETQTTCDHPKMTELY</td>
<td>M37645</td>
</tr>
<tr>
<td>Utro_Human</td>
<td>2812</td>
<td>TSVQLPQWQRSISHN.KVPYYINHIQTQTTCDHPKMTELY</td>
<td>P46939</td>
</tr>
<tr>
<td>Ykb2_Yeast-1</td>
<td>1</td>
<td>...MSIWKEAKDAS.GRIYYNTLKSTWEKPKESIQ</td>
<td>P33203</td>
</tr>
<tr>
<td>Ykb2_Yeast-2</td>
<td>39</td>
<td>LLLRENGKAAKTAD.GKVYYNPTRETSTWIPAFKKV</td>
<td>P33203</td>
</tr>
<tr>
<td>Yo61_Caeel-1</td>
<td>78</td>
<td>PSVESDWSVHNEK.GTPYHNRVTQTSWIKPDLKTP</td>
<td>P34600</td>
</tr>
<tr>
<td>Yo61_Caeel-2</td>
<td>123</td>
<td>QPQQQWKEFMSD.GKPYYNTLKTQWVKPDGEIT</td>
<td>P34600</td>
</tr>
<tr>
<td>Amoe/Acaca</td>
<td>?</td>
<td>MASHQGWQYPFTEAGNAYYNEVSGETSDPSSLQSH</td>
<td>M60954</td>
</tr>
<tr>
<td>FE65_Rat</td>
<td>42</td>
<td>SDLPGAMRQVQTS.GTYWYHI.PTGGTQWEEPGRASPS</td>
<td>P46933</td>
</tr>
<tr>
<td>Essl_Yeast</td>
<td>29</td>
<td>TGLPTPWTVRSKSKKREYFFNPETKHSQWEEPEGNTKD</td>
<td>P22696</td>
</tr>
</tbody>
</table>

⇒ Bork and Sudol (1994), TIBS 19 (94), 531-533
PROSITE : PS01159

ID WW_DOMAIN_1; PATTERN.
AC PS01159;
DT NOV-1995 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE WW/rsp5/WWP domain signature.
PA W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P.
NR /RELEASE=38,80000;
NR /TOTAL=46(33); /POSITIVE=37(24); /UNKNOWN=1(1); /FALSE_POS=8(8);
NR /FALSE_NEG=0; /PARTIAL=1;
CC /TAXO-RANGE=??E??; /MAX-REPEAT=4;
DR P46942, DB10_NICSY, T; P11533, DMD_CHICK , T; P11532, DMD_HUMAN , T;
DR P11531, DMD_MOUSE , T; P54353, DOD_DROME , T; Q13474, DRP2_HUMAN, T;
DR P22696, ESS1 YEAST, T; P46933, FE65_RAT , T; Q12647, GUNB_NEOPA, T;
DR P46940, IQGA_HUMAN, T; Q13526, PIN1_HUMAN, T; P46939, UTRQ_HUMAN, T;
DR P46936, YA65 CHICK, T; P46937, YA65_HUMAN, T; P43582, YFBO_YEAST, T;
DR P46941, YLE5_CAEEL, T; P33203, PR40 YEAST, T; Q09685, YA12_SCHPO, T;
DR P46938, YA65_MOUSE, T; P34600, YO61_CAEEL, T; P46935, NED4_MOUSE, T;
DR Q92462, PUB1_SCHPO, T; P39940, RSP5 YEAST, T; P46934, NED4_HUMAN, T;
DR P11530, DMD_RAT , P;
DR P40318, SSM4 YEAST, ?;
DR P53868, ALG9 YEAST, F; P12807, AMO_PICAN , F; P47332, LGT_MYCGE , F;
DR P75547, LGT_MYCPN , F; Q00019, RHGB_ASPAC, F; Q07307, UAPA_EMENI, F;
DR P48777, UAPC_EMENI, F; P53076, YGX7 YEAST, F;
DO PDOC50020;
//

Marcel Turcotte
CSI5126. Algorithms in bioinformatics
The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline- motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

- Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin form tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and...
organization of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

- Utrophin, a dystrophin-like protein of unknown function.
- Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].
- Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].
- Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>), followed by a histidine-rich region, 3 WW domains and a HECT domain.
- Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.
- Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).
- Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.
- IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.
- Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with
similarity to MYO2- type myosin, each containing two WW-domains at the N-terminus.
- Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.
- Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, we have developed a profile which spans the whole homology region as well as a pattern.

- Consensus pattern:
  \[ W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P \]
- Sequences known to belong to this class detected by the pattern: ALL.
- Other sequence(s) detected in SWISS-PROT: 8.
- Sequences known to belong to this class detected by the profile: ALL.
- Other sequence(s) detected in SWISS-PROT: NONE.
- Note: this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.
- Expert(s) to contact by email:
  Peer Bork; bork@embl-heidelberg.de
  Sudol M.; m_sudol@smtplink.mssm.edu

- Last update: July 1999 / Text revised.

[ 1] Bork P., Sudol M.
[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P.
[5] Chen H.I., Sudol M.
    Huebner K., Lehman D.
Things we like about REs!

Allow to model mandatory amino acids.
Easy to interpret in terms of biological concepts, such as binding sites, etc.
Issues

➤ Too rigid! Does not allow for mismatches;
➤ Will not perform well on new entries (overfitting);
➤ Compromise between sensitivity/sensibility, flexibility/noise;
➤ Entries must be revised as new sequences become available;
➤ Human intervention high, often derived from literature;
➤ Subjective choice of the region in some cases;
➤ Short motifs can occur by chance.
Issues (cont.)


Given $r$ and $s$ two regular expressions.

Perl representation

- $a \rightarrow a$
- $rs \rightarrow rs$
- $r + s \rightarrow r|s$
- $r^i \rightarrow r\{i\}$
  - $r\{i,j\}$, range quantifier
  - $r\{i,\}$, at least $i$ times
- $r^* \rightarrow r^*$
- $r^+ \rightarrow r^+$
- $[a-z]$
- $[^a-z]$
The language accepted by $M$, designated by $L(M)$, is
$\{x | \delta(q_0, x) \in F\}$.

A language is regular if accepted by an FSA.

The languages accepted by FSA can be described by simple expressions called regular expressions.
Regular Expressions

\[ \epsilon \]
the empty string

\[ a, \ \forall \ a \in \Sigma \]
a single character is a regular expression, eg a

\[ L_1L_2 \]
concatenation, eg ab

\[ L_1 + L_2 \]
union, eg : a+b matches a or b

\[ L^i = LL^{i-1} \]
fixed number of repeats

\[ L^0 = \{ \epsilon \} \]
Base case of the recursion

\[ L^* = \bigcup L^i, \ 0 \leq i \leq \infty \]
(Kleene) closure

\[ L^+ = \bigcup L^i, \ 1 \leq i \leq \infty \]
positive closure

⇒ The languages accepted by finite state automata are precisely the languages denoted by regular expressions.
Regular Expressions (cont.)

\[(P^+ G) + (G^+ P)\]
Finite State Automaton

Following Hopcroft & Ullman 1979, p. 16†
“A finite (state) automaton consists of finite set of states and a set of transitions from state to state that occur on input symbols chosen from and alphabet \( \Sigma \). For each input symbol there is exactly one transition out of each state (possibly back to the state itself).”

\[(Q, \Sigma, \delta, q_0, F)\]

where,
- \( Q \) is a set of states
- \( \Sigma \) is a finite input alphabet
- \( \delta \) is a set of transitions
- \( q_0 \) is the initial state, \( q_0 \in Q \)
- \( F \) is the set of final states such that \( F \subseteq Q \).

Finite Automaton

N-glycosylation

Start

q0 \rightarrow q1 \rightarrow q2 \rightarrow q3 \rightarrow q4

N, \leq P, S,T, \leq P
“It has been known for a long time [1] that potential N-glycosylation sites are specific to the consensus sequence Asn-Xaa-Ser/Thr. It must be noted that the presence of the consensus tripeptide is not sufficient to conclude that an asparagine residue is glycosylated, due to the fact that the folding of the protein plays an important role in the regulation of N-glycosylation [2]. It has been shown [3] that the presence of proline between Asn and Ser/Thr will inhibit N-glycosylation; this has been confirmed by a recent [4] statistical analysis of glycosylation sites, which also shows that about 50% of the sites that have a proline C-terminal to Ser/Thr are not glycosylated.”

⇒ www.expasy.ch/prosite
Finite State Automaton

(\( Q = \{ q_0, q_1, q_2, q_3, q_4 \}, \Sigma, \delta, q_0, F = \{ q_4 \} \))

\[
\delta = \begin{cases} 
\delta(q_0, N) = q_1, \\
\delta(q_1, \neq P) = q_2, \\
\delta(q_2, S) = q_3, \\
\delta(q_2, T) = q_3, \\
\delta(q_3, \neq P) = q_4.
\end{cases}
\]

where \( \Sigma \) is all 20 amino acids.

\[ \Rightarrow \delta(q_1, \neq P) = q_2, \text{ is a short hand notation for the 19 transitions : } \delta(q_1, A) = q_2, \ldots \delta(q_1, W) = q_2; \text{ all except P.} \]
Finite State Automaton

We can extend the definition of $\delta$ to strings,

$$\delta(q, wa) = \delta(\delta(q, w), a)$$

A string $x$ is accepted by an automaton $M = (Q, \Sigma, \delta, q_0, F)$ if $\delta(q_0, x) = p$ for some $p \in F$.

The language accepted by $M$, designated by $L(M)$, is

$$\{x|\delta(q_0, x) \in F\}.$$  

A language is regular if accepted by an FSA.
Finite State Automaton

N-glycosylation

Accepted: NASA, NKTE, NCST, ...
Not accepted: GASA, NPSA, NKTP, ...

⇒ How many distinct peptides (short protein sequences) are accepted by the above FSA?
References


Pensez-y !

L’impression de ces notes n’est probablement pas nécessaire!