Sequence Entropy

Algorithms in Sequence Analysis

ASA - Sequence Entropy
22 Nov 2012
Significance of Alignment Positions

- Observed occurrence of amino acids at some position in an alignment that deviates from expected may indicate some (functional) significance.

- What ‘deviates from expected’?
  - unlikely occurrences

- What is unlikely?
  - only (relatively) few possibilities to obtain observed result.
Aquaporin: Motifs

- **NPA**: stabilizes loops B and E
- **G(a)xxxG(a)xxG(a)**:
  - Crossing of right-hand helical bundles

Andreas Engel and Henning Stahlberg, in: *Current Topics in Membranes* (2001), Hohmann, Agre & Nielsen (Eds.) Academic Press
Leucine Zipper

- **Dimer**
  - Leu interactions
  - binds DNA by a fork-shaped structure
- **‘coiled-coil’ structure:**
  - leucines on one side of helix
  - 7-residue repeat; one helix turn is 3.6 residues

```
256  
K
V E E L L S K
N Y H L E N E
V A R L K K L
V G
```

Images: [www.wikipedia.org/wiki/Leucine_zipper](http://www.wikipedia.org/wiki/Leucine_zipper)
Counting...

- Number of possibilities for finding some combination of amino acids:
  - which types?
  - how much of each?

- Examples:
  - WWW 3 W \(\Rightarrow\) only 1 way
  - RHH 1 R, 2 H \(\Rightarrow\) three ways
  - SHQ 1 S, 1 H, 1 Q \(\Rightarrow\) six ways
Counting... (2)

- ‘Real’ examples:
  - WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
    - 33 W ⇒ only 1 way
  - RRRRRRRRRRRRRRRRRRHHHHHHHHHHHHHHHHHH
    - 16 R, 17 H ⇒ ? ways (~ $2^{33} \approx 10^9$)
  - SSSSSHSSCCCCCCCCCEEQQEEEEEEEEQEEEE
    - 7 S, 1 H, 8 C, 14 E, 3 Q ⇒ ??? ways (~ $5^{32} \approx 10^{23}$)

- ‘many’ ways

⇒ but, we can calculate that!
Counting… (3)

- Number of possibilities for finding given numbers $N_x$ of aminoacids types $x$:

$$
\Omega = \frac{\left( \sum_x N_x \right)!}{\prod_x \left( N_x! \right)}
$$

- Example:

  $N_1 = 3, \quad N_2 = 4$

  $\frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7}{(1 \times 2 \times 3) \times (1 \times 2 \times 3 \times 4)} = \frac{5040}{144} = 35$

- Another example:

  $N_1 = 3, \quad N_2 = 2, \quad N_3 = 4$

  $\frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8 \times 9}{(1 \times 2 \times 3) \times (1 \times 2) \times (1 \times 2 \times 3 \times 4)} = \frac{362880}{288} = 1260$

⇒ Problem: evaluates to huge numbers even for modest numbers of sequences…
### Counting... (4)

<table>
<thead>
<tr>
<th>Aminoacids</th>
<th>Counts</th>
<th>Ω</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW</td>
<td>3 W</td>
<td>3!/3!</td>
</tr>
<tr>
<td>RHH</td>
<td>1 R, 2 H</td>
<td>(1+2)! / (2!*1!)</td>
</tr>
<tr>
<td>SHQ</td>
<td>1 S, 1 H, 1 Q</td>
<td>(1+1+1)! / (1!*1! *1!)</td>
</tr>
<tr>
<td>WWWWWW</td>
<td>33 W</td>
<td>33!/33!</td>
</tr>
<tr>
<td>WWWWWW</td>
<td>16 R, 17 H</td>
<td>(16+17)! / (16!*17!)</td>
</tr>
<tr>
<td>HHHHHHHHHHHHH</td>
<td>7 S, 1 H, 8 C, 14 E, 3 Q</td>
<td>(7+1+8+14+3)!/ (7!*1!*8!*14!*3!)</td>
</tr>
</tbody>
</table>
Sequence Entropy

- Entropy: \( S = \ln \Omega = \ln \left( \sum_{x} \frac{N_x}{\prod_{x} (N_x !)} \right) \)
- Using Stirlings approximation:
  \( \ln x! \approx x \ln x - x \quad \text{for } x \gg 1 \)
- and using probabilities \( p \) in stead of counts \( N \)

- we get:
  \[ \frac{S}{\sum_{x} N_x} = - \sum_{x} p_x \log p_x \]
Shannon’s ‘Information Entropy’:


“Can we define a quantity which will measure, in some sense, how much information is ‘produced’ by such a process, or better, at what rate information is produced?”

• He was thinking about the Transmission of Information, i.e., from a Source through some Channel to a Destination.
Choice, Uncertainty and Entropy

- A set of ‘events’ with probabilities $p_1, p_2, \ldots, p_n$

- Is there a measure that indicates how much ‘choice’ is possible, given those probabilities?

- If there is, it should be:
  - continuous for all $p_i$
  - monotonic in $n$ if all probabilities are equal
  - additive for ‘sub-events’
Additivity:

\[ H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right) \]
Solution: Entropy

\[ H = - \sum_{i=1}^{n} p_i \log p_i \]

• the entropy of a set of probabilities \( p_i \)
• measures information, choice and uncertainty
• zero only if only one \( p_i \) is not zero
  • there is only one choice
• maximal if all \( p_i \) are equal
  • most ‘uncertain’ situation: all options are possible
Information Content

• Shannon was thinking about the *Transmission of Information*, i.e., from a Source through some Channel to a Destination.

• …but it applies equally well to any type of ‘message’

• We can use it to measure the level of conservation in columns in an alignment
Simple Example: Sequence Entropy

$p_1 = p_2 = \frac{1}{2}$

$H = -\sum_{i=1}^{n} p_i \log p_i$

$p_1 = 0$

$p_2 = 0$

$p_2 = f(‘A’)$

$p_1 = f(‘L’)
## Counting... (4)

<table>
<thead>
<tr>
<th>Aminoacids</th>
<th>Counts</th>
<th># states $\Omega$</th>
<th>$\ln(\Omega)/M$</th>
<th>$-\sum p \log p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW</td>
<td>3 W</td>
<td>1</td>
<td>0.37</td>
<td>0.64</td>
</tr>
<tr>
<td>RHH</td>
<td>1 R, 2 H</td>
<td>3</td>
<td>0.60</td>
<td>1.10</td>
</tr>
<tr>
<td>SHQ</td>
<td>1 S, 1 H, 1 Q</td>
<td>6</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>WWWW</td>
<td>33 W</td>
<td>1</td>
<td>1.2*10^9</td>
<td>0.63</td>
</tr>
<tr>
<td>RRRR</td>
<td>16 R, 17 H</td>
<td>2.1*10^13</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>SSSS</td>
<td>7 S, 1 H, 8 C, 14 E, 3 Q</td>
<td>8.2*10^16</td>
<td>1.18</td>
<td>1.36</td>
</tr>
</tbody>
</table>
Identification of Functional Sites

- Functional differences between Protein (sub-)families
- Current practice:
  - use Multiple Sequence Alignment
  - look for Conserved Sites within (sub-)families
    - (ignore sites that are overall conserved)
- Example Binders vs. Non-Binders:
  - sites crucial for binding: conserved (?)
  - sites determining ‘non-binding’: not conserved (!)

⇒ Take into account Non-Conserved Sites as well!
  - comparing Amino-acid Compositions
Identification of Functional Sites

• Sequence Harmony:
  • Conservation *versus* Differences
  • test-cases:
    • TGF-β signalling pathway (and others)
    • HIV Differential Progression/Replication

• Multi-RELIEF:
  • Feature-selection for Specificity
  • test-cases:
    • GPCRs (and others)

• Comparison of methods:
  • Detection of different specificity types
Conservation and (functional) Differences:

- Most methods rely on conservation
- But functionally specific sites are not always conserved:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Known</th>
<th>Conserved in Groups</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Both</td>
<td>One</td>
<td>Not</td>
<td></td>
</tr>
<tr>
<td>Ras/Ral</td>
<td>12</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Rab5/6</td>
<td>28</td>
<td>10</td>
<td>14</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>MIP</td>
<td>23</td>
<td>0</td>
<td>7</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>SMAD</td>
<td>29</td>
<td>10</td>
<td>14</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>TOTAL:</td>
<td>92</td>
<td>21</td>
<td>46</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
Conditional Probability / Prior Knowledge

• Suppose we observe two things (A and B), and we suspect a relation (A causes B)

• The fundamental question is then: How likely is A when we know B?

(N.B., this is Bayes statistics…)

• Or, what is the uncertainty (entropy) of A knowing B?
Conditional Entropy

- Entropy of joint occurrence of value $i$ for event $x$ and value $j$ for event $y$:

\[
H(x,y) = - \sum_{i,j} p(i,j) \log p(i,j)
\]

- and

\[
H(x) = - \sum_{i} p(i,j) \log \sum_{j} p(i,j) \quad \text{and} \quad H(y) = - \sum_{j} p(i,j) \log \sum_{i} p(i,j)
\]

- so that

\[
H(x,y) \leq H(x) + H(y)
\]

- i.e., the entropy of a joint event is less than or equal to the sum of the individual entropies

- it is equal *only* if the events are independent
Co-occurrence in practice

- Measure of ‘mutual information’, by relative entropy:

\[ H_y(x) = \sum_{i,j} p(i,j) \log \frac{p(i,j)}{\sum_i p(i,j)} \]

- more often written like:

\[ H(x \mid y) = \sum_x p(x) \log \frac{p(x)}{p(y)} \]

- i.e., what is the entropy in \( x \), given \( y \)
Relative Entropy in Sequence Analysis

- Many biological problems relate to questions like:
  "Why do these proteins do this, and those proteins not?"
- or
  "Why do these patients get sick, and those not?"

- The answer can be related to similarities and differences between sequences
  - Similarities (conservation) relate to functionally critical positions
  - Differences can explain functional differences
Comparing groups of Sequences

- For each position \( i \) in an alignment, we calculate the relative entropy for group A vs. B from the frequencies \( p \) (observed probabilities) of all aminoacid types \( x \), as follows:

\[
H_{i}^{A/B} = \sum_{x=1}^{n} p_{i,x}^{A} \log \frac{p_{i,x}^{A}}{p_{i,x}^{B}}
\]

- how likely are the frequencies in B \( (p^{B}) \), given those in A \( (p^{A}) \)
Simple Example: Relative Entropy ‘A/B’

\[ \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^B} \]

\[ \sum_x p_{i,x}^B \log \frac{p_{i,x}^B}{p_{i,x}^A} \]
Relative Entropy

- Captures similarities and differences
- Is infinite for completely dissimilar positions
  - e.g., A vs. L
- Not symmetrical:
  \[
  H_y(x) \neq H_x(y)
  \]
- Maybe not easy for selecting dissimilar positions
Simple Example: Relative Entropy ‘A/AB’

\[ \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^{AB}} \]

\[ \sum_x p_{i,x}^B \log \frac{p_{i,x}^B}{p_{i,x}^{AB}} \]
Relative Entropy (A/AB)

- Also captures similarities and differences
- Is no longer infinite for completely dissimilar positions
- Only symmetrical for equal size groups
  - in practice, not symmetrical
- Maybe still not too easy for selecting dissimilar positions
Entropy vs. Sequence Harmony: Example

\[ \sum_x p_{i,x} \log_2 \frac{p^A_{i,x}}{p^A_{i,x} + p^B_{i,x}} \]

\[ \sum_x p_{i,x} \log_2 \frac{p^B_{i,x}}{p^A_{i,x} + p^B_{i,x}} \]
Measuring Overlapping Distributions

- Weigh both groups equally; take $p^A + p^B$ instead of $p^{AB}$:

$$SH_{i}^{A/B} = \sum_{x} p_{i,x}^{A} \log \frac{p_{i,x}^{A}}{p_{i,x}^{A} + p_{i,x}^{B}}$$

- Fixed interval $[0,1]$, but not completely symmetrical

- We call it 'Sequence Harmony' (SH)
Sequence Harmony

- Introduce symmetry by averaging:
  \[
  \text{SH}_{i}^{A/B} = \frac{1}{2} \left( \text{SH}_{i}^{A/AB} + \text{SH}_{i}^{B/AB} \right)
  \]

- May seem a trivial choice, but:
  \[
  \text{SH}_{i}^{A/B} = \frac{1}{2} \left( \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B} + \sum_x p_{i,x}^B \log \frac{p_{i,x}^B}{p_{i,x}^A + p_{i,x}^B} \right)
  \]
  \[
  = \frac{1}{2} \left( \sum_x p_{i,x}^A \log p_{i,x}^A + \sum_x p_{i,x}^B \log p_{i,x}^B - \sum_x \left( p_{i,x}^A + p_{i,x}^B \right) \log \left( p_{i,x}^A + p_{i,x}^B \right) \right)
  \]
  \[
  = \frac{1}{2} \left( -H(A) - H(B) - \sum_x \left( p_{i,x}^A + p_{i,x}^B \right) \log \left( p_{i,x}^A + p_{i,x}^B \right) \right)
  \]
  \[
  = -\frac{1}{2} \left( H(A) + H(B) \right) + H(A+B) - \log 2
  \]

Pirovano, Feenstra & Heringa. 2006 “Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites”, Nucleic Acids Res. 24 6540-6548
Entropy vs. Sequence Harmony: Example

\[
\sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B} + H(A) + H(B) + H(A+B) - \log 2
\]
Smad-MH2 Alignment & Sequence Harmony

## Finding Low-harmony sites in Smad-MH2

### Table: Low-harmony sites

<table>
<thead>
<tr>
<th>Pos.</th>
<th>Sec.str.</th>
<th>SH</th>
<th>AR</th>
<th>BR</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>L263</td>
<td>B1'</td>
<td>0</td>
<td>La</td>
<td>Vfm</td>
<td>SARA</td>
</tr>
<tr>
<td>T267</td>
<td>B1'</td>
<td>0</td>
<td>Tm</td>
<td>Acen</td>
<td>SARA</td>
</tr>
<tr>
<td>S269</td>
<td>loop</td>
<td>0</td>
<td>CSh</td>
<td>Eq</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>A272</td>
<td>loop</td>
<td>0</td>
<td>A</td>
<td>Kqls</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>F273</td>
<td>loop</td>
<td>0</td>
<td>F</td>
<td>Hy</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>Q284</td>
<td>B2</td>
<td>0</td>
<td>Qt</td>
<td>N</td>
<td>TβR-I</td>
</tr>
<tr>
<td>Q294</td>
<td>loop</td>
<td>0.16</td>
<td>Q</td>
<td>Sq</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>P295</td>
<td>B3</td>
<td>0</td>
<td>P</td>
<td>Trl</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>L297</td>
<td>B3</td>
<td>0.11</td>
<td>LMi</td>
<td>Vi</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>T298</td>
<td>B3</td>
<td>0</td>
<td>T</td>
<td>Li</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>S308</td>
<td>L1</td>
<td>0</td>
<td>Sa</td>
<td>N</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>-</td>
<td>L1</td>
<td>0</td>
<td>-</td>
<td>Nsd</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>E309</td>
<td>L1</td>
<td>0</td>
<td>E</td>
<td>Krs</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>A323</td>
<td>H1</td>
<td>0</td>
<td>Ae</td>
<td>S</td>
<td>ALK1/2</td>
</tr>
<tr>
<td>V325</td>
<td>H1</td>
<td>0</td>
<td>V</td>
<td>I</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>M327</td>
<td>H1</td>
<td>0</td>
<td>LMq</td>
<td>N</td>
<td>ALK1/2</td>
</tr>
<tr>
<td>R334</td>
<td>Loop</td>
<td>0.18</td>
<td>Rk</td>
<td>K</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>R337</td>
<td>B5</td>
<td>0</td>
<td>R</td>
<td>H</td>
<td>?? (putative)</td>
</tr>
</tbody>
</table>
Finding Low-harmony sites in Smad-MH2

<table>
<thead>
<tr>
<th>Pos.</th>
<th>Sec.str.</th>
<th>SH</th>
<th>AR</th>
<th>BR</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>L263</td>
<td>B1’</td>
<td>0</td>
<td>La</td>
<td>Vfm</td>
<td>SARA</td>
</tr>
<tr>
<td>T267</td>
<td>B1’</td>
<td>0</td>
<td>Tm</td>
<td>Acen</td>
<td>SARA</td>
</tr>
<tr>
<td>S269</td>
<td>loop</td>
<td>0</td>
<td>CSh</td>
<td>Eq</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>A272</td>
<td>loop</td>
<td>0</td>
<td>A</td>
<td>Kqls</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>F273</td>
<td>loop</td>
<td>0</td>
<td>F</td>
<td>Hy</td>
<td>?? (putative)</td>
</tr>
<tr>
<td>Q284</td>
<td>B2</td>
<td>0</td>
<td>Qt</td>
<td>N</td>
<td>TβR-I</td>
</tr>
<tr>
<td>Q294</td>
<td>loop</td>
<td>0.16</td>
<td>Q</td>
<td>Sq</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>P295</td>
<td>B3</td>
<td>0</td>
<td>P</td>
<td>Trl</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>L297</td>
<td>B3</td>
<td>0.11</td>
<td>LMi</td>
<td>Vi</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>T298</td>
<td>B3</td>
<td>0</td>
<td>T</td>
<td>Li</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>S308</td>
<td>L1</td>
<td>0</td>
<td>Sa</td>
<td>N</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>-</td>
<td>L1</td>
<td>0</td>
<td>–</td>
<td>Nsd</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>E309</td>
<td>L1</td>
<td>0</td>
<td>E</td>
<td>Krs</td>
<td>c-Ski/SnoN</td>
</tr>
<tr>
<td>A323</td>
<td>H1</td>
<td>0</td>
<td>Ae</td>
<td>S</td>
<td>ALK1/2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Predict</th>
<th>Specificity</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAS</td>
<td>6</td>
<td>21%</td>
<td>3%</td>
</tr>
<tr>
<td>TreeDet</td>
<td>21</td>
<td>52%</td>
<td>21%</td>
</tr>
<tr>
<td>SDPpred</td>
<td>12</td>
<td>31%</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Sequence Harmony</strong></td>
<td><strong>(SH=0)</strong></td>
<td><strong>79%</strong></td>
<td><strong>28%</strong></td>
</tr>
<tr>
<td></td>
<td><strong>(SH&lt;0.2)</strong></td>
<td><strong>93%</strong></td>
<td><strong>33%</strong></td>
</tr>
</tbody>
</table>

www.few.vu.nl/~feenstra/articles/NAR 2006 Sequence Harmony.pdf
Smad-MH2: Low Harmony Patches
Smad-MH2: Functional Clusters

- **TβR-I/BMPR-I**
- **TβR-I/ALK1/2**
- **FAST1, Mixer, SARA**
- **SARA/Mixer**
- **co-repressors**
- **c-Ski/SnoN**

**receptor-binding**

**retention & transcription factors**
Example Smad dataset - Different Methods Select Different Sites:

<table>
<thead>
<tr>
<th>Method</th>
<th>%TP</th>
<th>Conserved in Groups (known function)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Both</td>
</tr>
<tr>
<td>multi-R</td>
<td>97%</td>
<td>13(10)</td>
</tr>
<tr>
<td>SH</td>
<td>93%</td>
<td>13(10)</td>
</tr>
<tr>
<td>TreeDet</td>
<td>52%</td>
<td>13(10)</td>
</tr>
<tr>
<td>SDPpred</td>
<td>31%</td>
<td>12(9)</td>
</tr>
<tr>
<td>All sites</td>
<td>60%</td>
<td>13(10)</td>
</tr>
</tbody>
</table>
Analyzing multiple groups

- Relative Entropy and Sequence Harmony are defined to compare a set of groups (N=2)
- Relief is also defined for two groups
  - problem for multiple groups (N>2)
- Another solution: Two-entropy plots
  - Total family entropy vs. sum of sub-family entropy
  - similar to Sequence Harmony
- Several other methods (mentioned later)
Usage

relief(data, nosample, threshold, vnom)

Arguments

type{data, nosample, threshold, vnom}

data

the dataset for which feature selection will be carried out

nosample

number of instances drawn from the original dataset

threshold

the cutoff point to select the features

vnom

a vector containing the indexes of the nominal features

Details

The general idea of this method is to choose the features that can be most distinguished between classes. These are known as the relevant features. At each step of an iterative process, an instance \( x \) is chosen at random from the dataset and the weight for each feature is updated according to the distance of \( x \) to its Nearmiss and NearHit.

Value

relevant

a table that gives the frequency with which the feature was selected as relevant over the trials performed, and the average weight of the feature.

a plot

a plot of the weights of the features

Guyon and Elisseeff, 2003
RELIEF for Sequence Analysis

• residue position conserved within each class but divergent between classes:
  • RELIEF weight will be high

• residue position divergent within subfamilies but conserved between subfamilies:
  • negative weight

• position completely conserved or non-overlapping
  • zero weight
RELIEF for Sequence Analysis

- Input: multiple sequence alignment containing two subgroups of sequences
- Iteratively randomly select a sequence:
  - find nearest neighbour in the same group (near hit ) and nearest neighbour in the other group (near miss )
    - selected using Hamming distance (i.e. number of unequal positions) between (complete) sequences as aligned in MSA
  - Calculate difference vector between near hit and near miss:
    - e.g. AA|V gives AV-AA = 1 (well placed); AV|A gives AA-AV=-1 (misplaced); AA|A gives 0 (indifferent)
  - Update feature weight vector with difference vector.
- Output feature weight vector with frequencies of relevant features (i.e. how often residues in sampled sequences were “well placed”)

[44] 22 Nov 2012 ASA - Sequence Entropy
RELIEF Pseudocode

% input: X (two classes of aligned proteins)
% output: weights assigned to each site
% features are sites
% examples are sequences

nr_{\text{feat}} = \text{total number of features};
weights = \text{zero vector of size } nr_{\text{feat}};

for all exa in X do
    hit(exa) = \text{NNB of exa from same class};
    miss(exa) = \text{NNB of exa from opposite class};
    weights += (exa-miss(exa)) - (exa-hit(exa));
end;
return weights;
### Example of how Relief works:

<table>
<thead>
<tr>
<th>Difference vector</th>
<th>Exemplar</th>
<th>Hit</th>
<th>Distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>**C1</td>
<td>S1**</td>
<td>D V Q A V A Y E E P Y H W C S</td>
<td>D A Q A V G T E E P K C W C S</td>
</tr>
<tr>
<td>exemplar →</td>
<td>0 1 0 0 0 1 1 0 0 0 1 1 0 0 0</td>
<td>0 0 0 1 0 1 0 0 0 0 0 1 1 0 0 0</td>
<td>4</td>
</tr>
<tr>
<td>hit →</td>
<td>D A Q P V A T E E P Y H W C S</td>
<td>D A Q P V A T E E P Y H W C S</td>
<td>7</td>
</tr>
<tr>
<td>**C2</td>
<td>S4**</td>
<td>D L Q P V T Y C E P A F W C S</td>
<td>D L Q P V A T Y C E P A F W C S</td>
</tr>
<tr>
<td>miss →</td>
<td>0 1 0 0 0 1 1 0 0 0 1 1 0 0 0</td>
<td>0 1 0 0 0 1 1 0 0 0 1 1 0 0 0</td>
<td>6</td>
</tr>
</tbody>
</table>
Example of how Relief works:

<table>
<thead>
<tr>
<th>Difference vector</th>
<th>ex – hit</th>
<th>hit</th>
<th>(ex–miss) – (ex–hit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>S1</td>
<td>0 1 0 0 0 1 1 0 0 0 1 1 0 0 0</td>
<td></td>
</tr>
<tr>
<td>exemplar →</td>
<td>D V Q A V A Y E E E P Y H W C S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ex – hit →</td>
<td>ex – hit</td>
<td>D A Q A V G T E E P K C W S</td>
<td></td>
</tr>
<tr>
<td>hit →</td>
<td>ex – miss</td>
<td>D A Q P V A T E E P Y H W C S</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>S4</td>
<td>0 1 0 1 0 1 1 1 0 0 1 1 0 0 0</td>
<td>= 7</td>
</tr>
<tr>
<td>C2</td>
<td>S5</td>
<td>0 1 0 1 0 1 1 1 0 0 1 1 0 0 0</td>
<td>= 7</td>
</tr>
<tr>
<td>ex – miss →</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miss →</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Distances:
- (ex–miss) – (ex–hit) = 0 1 0 0 0 0 0 1 1 0 0 0 -1 0 0 0 0

- 0 1 0 0 0 1 1 0 0 0 1 1 0 0 0
- 0 0 0 1 0 1 0 0 0 0 1 1 0 0 0
- 0 1 0 1 0 1 1 1 0 0 1 1 0 0 0
- 0 1 0 1 0 1 1 1 0 0 1 1 0 0 0
- 0 1 0 1 0 1 1 1 0 0 0 1 0 0 0
- 0 1 0 1 0 1 1 1 0 0 0 1 0 0 0

- D V Q A V A Y E E E P Y H W C S
- D A Q A V G T E E P K C W S
- D A Q P V A T E E P Y H W C S
- D A Q P V A T E E P Y H W C S
- D A Q P V A T E E P K C W S
- D A Q P V A T E E P K C W S

- 5
- 4
- 7
- 7
- 6
Multi-RELIEF for MSAs with more than two subgroups

- Extensions of RELIEF to handle multiple classes have been proposed (Kononenko, 1994; Robnik-Sikonja and Kononenko, 2003; Sun and Li, 2006).

- Kononenko (1994) introduced RELIEF-F where the weight vector is updated by the sum of miss(seq) weighted by the estimated a priori probabilities of the classes.

- We present a new ensemble-based approach using random sub-sampling of pairs of classes.
Multi-RELIEF

- Iteratively apply RELIEF to random pair of classes:
  - take average over positive $W_i(s)$’s
    - high score if position $s$ discriminates at least one pair
  - take average over all $W_i(s)$’s if none is positive
    - score can become negative (see next slide)
- Maximum weight assigned whenever:
  - $s$ fully discriminates (at least) two specific classes
  - does not differentiate (i.e. $W_i(s) \leq 0$) any other pair of classes.
    $\rightarrow$ robust against some misclassified sequences.
- Random sampling of pairs of classes for efficiency reasons
- Random sub-sampling of sequences for efficiency and for handling unbalanced classes (with highly different numbers of sequences)
Multi-RELIEF weights

\[\text{multi}_W(s) = \begin{cases} 
\frac{1}{N^+} \sum_i \{W_i(s) > 0 \ \forall \ i\} & \text{for } N^+ > 0 \\
\frac{1}{N^-} \sum_i \{W_i(s) < 0 \ \forall \ i\} & \text{for } N^+ = 0 \land N^- > 0 \\
0 & \text{for } N^+ = 0 \land N^- = 0 
\end{cases}\]

using \(N^+ = |\{W_i(s) > 0 \ \forall \ i\}|\) and \(N^- = |\{W_i(s) < 0 \ \forall \ i\}|\)
Multi-RELIEF Pseudocode

%input: X_1,...,X_m (m classes of aligned proteins)
%parameters: nr_{iter}, nr_{sample}
%output: multi_W (weights assigned to positions)

nr_{positions} = total number of positions;
weights = zero vector of size nr_{positions};
for i=1: nr_{iter}
    select randomly two classes
    X = select randomly nr_{sample} sequences
        from each selected class
    W_i = apply RELIEF to X
end;
for s=1: nr_{positions}
    multi_W(s) = (average across positive W_i(s)'s);
end;
return multi_W
### Weights computed by SH and mR

<table>
<thead>
<tr>
<th>Alignment position</th>
<th>Distance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8</td>
<td>1 2 3 4 1 2 3 1</td>
</tr>
</tbody>
</table>

#### Group 1

| seq1  | R   | E   | L   | A   | A   | K   | K   | A   |
| seq2  | R   | E   | L   | A   | F   | K   | K   | I   |
| seq3  | R   | E   | A   | A   | Y   | R   | K   | L   |
| seq4  | R   | E   | A   | A   | Y   | R   | K   | L   |

| SH     | 0.42 | 0    | 0    | 0.57 | 0.87 | 0.99 | 1    | 0    |
| mR     | 1    | 1    | 0.67 | 1    | -0.42 | -0.19 | 0    | 0.5  |

- **Simple example with three groups:**
  - 1, 2: conserved within groups
  - 3: completely different between groups (not conserved)
  - 4: some overlap
  - 5, 6: different residues within a sub-family, but conservation between sub-families
  - 7: completely conserved
  - 8: partly conserved, but completely different between groups
## Properties of the Datasets

<table>
<thead>
<tr>
<th>dataset</th>
<th>nr of classes</th>
<th>avg (std) class size</th>
<th>max, min class size</th>
<th>nr of sites</th>
<th>site information</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCR</td>
<td>77</td>
<td>26.8 (34)</td>
<td>189, 3</td>
<td>214</td>
<td>ligand</td>
</tr>
<tr>
<td>GPCR190</td>
<td>39</td>
<td>4.9 (3.8)</td>
<td>21, 2</td>
<td>214</td>
<td>ligand &amp; DNA</td>
</tr>
<tr>
<td>LacI</td>
<td>15</td>
<td>3.6 (2.5)</td>
<td>12, 2</td>
<td>339</td>
<td>ligand &amp; DNA</td>
</tr>
<tr>
<td>Ras/Ral</td>
<td>2</td>
<td>44.5 (24.5)</td>
<td>69, 20</td>
<td>218</td>
<td>protein</td>
</tr>
<tr>
<td>Rab5/6</td>
<td>2</td>
<td>5.0 (1)</td>
<td>4, 6</td>
<td>163</td>
<td>protein</td>
</tr>
<tr>
<td>AQP/GLP</td>
<td>2</td>
<td>30.0 (18)</td>
<td>48, 12</td>
<td>430</td>
<td>protein</td>
</tr>
<tr>
<td>Smad</td>
<td>2</td>
<td>10.0 (2)</td>
<td>12, 8</td>
<td>211</td>
<td>protein</td>
</tr>
</tbody>
</table>

- 7 datasets as above
- 15 other datasets from:
GPCR
LacI
MIP
(AQP/GLP)
## Related methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Website</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROUST2</td>
<td><a href="http://www.russell.embl-heidelberg.de/proust2">www.russell.embl-heidelberg.de/proust2</a></td>
<td>Hannenhalli &amp; Russell, 2000</td>
</tr>
<tr>
<td>ProteinKeys</td>
<td><a href="http://www.proteinkeys.org/proteinkeys">www.proteinkeys.org/proteinkeys</a></td>
<td>Reva et al., 2007</td>
</tr>
<tr>
<td>SDPpred</td>
<td>bioinf.fbb.msu.ru/SDPpred</td>
<td>Kalinina et al., 2004b</td>
</tr>
<tr>
<td>SDPsite</td>
<td>bioinf.fbb.msu.ru/SDPsite</td>
<td>Kalinina et al., 2009</td>
</tr>
<tr>
<td>TreeDet</td>
<td>treedetv2.bioinfo.cnio.es/treedet</td>
<td>Carro et al., 2006</td>
</tr>
<tr>
<td>Xdet</td>
<td>pdg.cnb.uam.es/pazos/mtreedet</td>
<td>Pazos et al., 2006</td>
</tr>
</tbody>
</table>
Output: example
Average Coverage/Precision for all methods

- Xdet
- Xdet supervised
- multi-Relief
- PROUST-II
- Sequence Harmony
- SDPpred
- ProteinKeys

Coverage TP/(TP+FN)

Precision (1-Error) FP/(FP+TN)
Average Precision-Recall for all methods

Precision TP/(TP+FP)
Recall TP/(TP+FN)

Seq. Harm. Z≤-9
Sequence Harmony
Xdet supervised
SDPpred
multi-Relief Z≥6
multi-Relief
PROUST-II
ProteinKeys
Xdet
Average Ranks for all methods (low=good!)

![Box plot showing average ranks for different methods](image)
### Areas Under Curve in the Precision-Recall-plots

<table>
<thead>
<tr>
<th>Dataset</th>
<th># positives</th>
<th>mR Z ≥ 6</th>
<th>mR Z ≤ −9</th>
<th>SH. Z ≤ −9</th>
<th>Protein Keys</th>
<th>PROUS T-II</th>
<th>SDP pred v.2</th>
<th>Xdet</th>
<th>Xdet sup</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>cbm9</td>
<td>7</td>
<td>0.161</td>
<td>0.161</td>
<td>0.074</td>
<td>0.074</td>
<td>0.049</td>
<td>0.349</td>
<td>0.122</td>
<td>0.352</td>
<td>0.209</td>
</tr>
<tr>
<td>cd00120</td>
<td>3</td>
<td>0.058</td>
<td>0.058</td>
<td>0.054</td>
<td>0.054</td>
<td>0.008</td>
<td>0.079</td>
<td>0.126</td>
<td>0.106</td>
<td>0.106</td>
</tr>
<tr>
<td>cd00264</td>
<td>3</td>
<td>0.006</td>
<td>0.006</td>
<td>0.003</td>
<td>0.003</td>
<td>0.087</td>
<td>0.012</td>
<td>0.017</td>
<td>0.080</td>
<td>0.019</td>
</tr>
<tr>
<td>cd00333</td>
<td>12</td>
<td>0.301</td>
<td>0.301</td>
<td>0.287</td>
<td>0.287</td>
<td>0.203</td>
<td>0.055</td>
<td>0.376</td>
<td>0.366</td>
<td>0.346</td>
</tr>
<tr>
<td>cd00363</td>
<td>6</td>
<td>0.010</td>
<td>0.010</td>
<td>0.008</td>
<td>0.008</td>
<td>0.010</td>
<td>0.011</td>
<td>0.012</td>
<td>0.011</td>
<td>0.012</td>
</tr>
<tr>
<td>cd00365</td>
<td>10</td>
<td>0.055</td>
<td>0.055</td>
<td>0.119</td>
<td>0.119</td>
<td>0.010</td>
<td>0.016</td>
<td>0.126</td>
<td>0.103</td>
<td>0.189</td>
</tr>
<tr>
<td>cd00423</td>
<td>4</td>
<td>0.204</td>
<td>0.204</td>
<td>0.080</td>
<td>0.080</td>
<td>0.002</td>
<td>0.049</td>
<td>0.234</td>
<td>0.196</td>
<td>0.171</td>
</tr>
<tr>
<td>cd00985</td>
<td>3</td>
<td>0.329</td>
<td>0.329</td>
<td>0.198</td>
<td>0.198</td>
<td>0.034</td>
<td>0.058</td>
<td>0.509</td>
<td>0.387</td>
<td>0.534</td>
</tr>
<tr>
<td>CN-myc</td>
<td>11</td>
<td>0.037</td>
<td>0.037</td>
<td>0.067</td>
<td>0.067</td>
<td>0.027</td>
<td>0.122</td>
<td>0.162</td>
<td>0.086</td>
<td>0.010</td>
</tr>
<tr>
<td>GPCR190</td>
<td>21</td>
<td>0.246</td>
<td>0.252</td>
<td>0.486</td>
<td>0.517</td>
<td>0.377</td>
<td>0.308</td>
<td>0.508</td>
<td>0.125</td>
<td>0.275</td>
</tr>
<tr>
<td>GPCR</td>
<td>21</td>
<td>0.347</td>
<td>0.347</td>
<td>0.489</td>
<td>0.489</td>
<td>0.505</td>
<td>–</td>
<td>0.508</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GST</td>
<td>9</td>
<td>0.156</td>
<td>0.156</td>
<td>0.242</td>
<td>0.242</td>
<td>0.483</td>
<td>0.446</td>
<td>0.615</td>
<td>0.117</td>
<td>0.402</td>
</tr>
<tr>
<td>IDH/IMDH</td>
<td>14</td>
<td>0.050</td>
<td>0.050</td>
<td>0.048</td>
<td>0.048</td>
<td>0.065</td>
<td>0.089</td>
<td>0.196</td>
<td>0.100</td>
<td>0.129</td>
</tr>
<tr>
<td>LacI</td>
<td>28</td>
<td>0.266</td>
<td>0.282</td>
<td>0.124</td>
<td>0.207</td>
<td>0.301</td>
<td>0.111</td>
<td>0.146</td>
<td>0.190</td>
<td>0.207</td>
</tr>
<tr>
<td>MDH/LDH</td>
<td>1</td>
<td>0.063</td>
<td>0.063</td>
<td>0.125</td>
<td>0.125</td>
<td>0.005</td>
<td>0.015</td>
<td>0.250</td>
<td>0.033</td>
<td>0.250</td>
</tr>
<tr>
<td>AQP/GLP</td>
<td>23</td>
<td>0.213</td>
<td>0.216</td>
<td>0.249</td>
<td>0.268</td>
<td>0.119</td>
<td>0.187</td>
<td>0.242</td>
<td>0.169</td>
<td>0.208</td>
</tr>
<tr>
<td>nucl cycl.a</td>
<td>2</td>
<td>0.417</td>
<td>0.417</td>
<td>0.413</td>
<td>0.413</td>
<td>0.011</td>
<td>0.305</td>
<td>0.413</td>
<td>0.054</td>
<td>0.092</td>
</tr>
<tr>
<td>rab5/6</td>
<td>28</td>
<td>0.540</td>
<td>0.539</td>
<td>0.602</td>
<td>0.602</td>
<td>0.364</td>
<td>0.455</td>
<td>0.416</td>
<td>0.350</td>
<td>0.346</td>
</tr>
<tr>
<td>ras/ral</td>
<td>12</td>
<td>0.666</td>
<td>0.666</td>
<td>0.540</td>
<td>0.540</td>
<td>0.092</td>
<td>0.378</td>
<td>0.357</td>
<td>0.398</td>
<td>0.545</td>
</tr>
<tr>
<td>ricin</td>
<td>21</td>
<td>0.186</td>
<td>0.186</td>
<td>0.194</td>
<td>0.194</td>
<td>0.276</td>
<td>0.256</td>
<td>0.201</td>
<td>0.173</td>
<td>0.193</td>
</tr>
<tr>
<td>serine</td>
<td>2</td>
<td>0.078</td>
<td>0.078</td>
<td>0.261</td>
<td>0.261</td>
<td>0.006</td>
<td>0.750</td>
<td>0.542</td>
<td>0.105</td>
<td>0.750</td>
</tr>
<tr>
<td>Smad</td>
<td>29</td>
<td>0.719</td>
<td>0.721</td>
<td>0.713</td>
<td>0.703</td>
<td>0.748</td>
<td>0.723</td>
<td>0.522</td>
<td>0.688</td>
<td>0.677</td>
</tr>
<tr>
<td>Aver Wt’d</td>
<td></td>
<td>0.310</td>
<td>0.312</td>
<td>0.330</td>
<td>0.342</td>
<td>0.287</td>
<td>0.258</td>
<td>0.333</td>
<td>0.234</td>
<td>0.279</td>
</tr>
</tbody>
</table>
Summary

• Sequence Entropy ↔ Information Content
• Relative Entropy ↔ Mutual Information
• Sequence Harmony
• multi-Relief
• discrimination of methods' performance