Sequence Entropy

Sequence Analysis

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

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>Both</th>
<th>One</th>
<th>Not</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ras/Ral</td>
<td>12</td>
<td></td>
<td>1</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Rab5/6</td>
<td>28</td>
<td></td>
<td>10</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>MIP</td>
<td>23</td>
<td></td>
<td>0</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>SMAD</td>
<td>29</td>
<td></td>
<td>10</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>92</td>
<td></td>
<td>11</td>
<td>46</td>
<td>25</td>
</tr>
</tbody>
</table>

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

Identification of Functional Sites
- Sequence Harmony:
  - Conservation versus Differences
    - test-cases:
      - TGF-beta signaling pathway (and others)
      - HIV Differential Progression/Replication
    - Multi-RELIEF:
      - Feature selection for Specificity
    - test-cases:
      - GPCR’s (and others)
  - Comparison of methods:
    - Detection of different specificity types

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

Examples:
- RAMS 3 W ⇒ only 1 way
- RHE 1 R, 2 H ⇒ three ways
- RJQ 1 S, 1 H, 1 Q ⇒ six ways
Counting... (2)

- 'Real' examples:
  - 32 W  ⇒ only 1 way
  - \( \text{HH} \)  ⇒ \( 7 \) ways (\( 2^2 = 10 \))
  - \( \text{HS} \)  ⇒ \( 7 \) ways (\( 2^2 = 10 \))
  - \( \text{SS} \)  ⇒ \( 7 \) ways (\( 2^2 = 10 \))
  - \( \text{SS} \)  ⇒ \( 7 \) ways (\( 2^2 = 10 \))
  - \( 7, 1, 4, 3, 2, 0 \)  ⇒ \( 7 \) ways (\( 2^2 = 10 \))

- 'many' ways
  ⇒ but, we can calculate that!

Counting... (3)

- Number of possibilities \( \Omega \) for finding given numbers \( N_x \) of amino acids types \( x \):

\[
\Omega = \frac{\prod N_x!}{\prod N_x!}
\]

⇒ New problem: now we calculate really huge numbers even for modest numbers of sequences.

Counting... (4)

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Counts</th>
<th>( \ln(\Omega) )</th>
<th>( \text{size } \Omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>3 W</td>
<td>( \frac{\ln(3)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
<tr>
<td>HHH</td>
<td>1 R, 2 H</td>
<td>( \frac{\ln(2)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
<tr>
<td>HHH</td>
<td>1 S, 1 H, 1 G</td>
<td>( \frac{\ln(3)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
<tr>
<td>HHH</td>
<td>33 W</td>
<td>( \frac{\ln(3)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
<tr>
<td>HHH</td>
<td>16 R, 17 H</td>
<td>( \frac{\ln(3)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
<tr>
<td>HHH</td>
<td>7 S, 1 H, 1 G</td>
<td>( \frac{\ln(3)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
<tr>
<td>HHH</td>
<td>5 S, 1 H, 1 G</td>
<td>( \frac{\ln(3)}{\ln(2)} )</td>
<td>( 2^3 = 8 )</td>
</tr>
</tbody>
</table>

Sequence Entropy

- Entropy:

\[
S = \ln(\Omega) = \ln \left( \frac{\prod N_x!}{\prod N_x!} \right)
\]

- We can use Stirling's approximation:

\[
\ln n! \approx n \ln n - n
\]

⇒ so we get:

\[
S = \sum p_x \log p_x
\]

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:
\[ H(Y_2, Y_1, Y_3) = H(Y_2 | Y_3) + H(Y_1 | Y_3) \]

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
\[ p_1 = f(A) \quad p_1 = f(L) \]

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
  \[ H(x) = \sum_{i} p(i) \log p(i) \]
- so that
  \[ H(x,y) = H(x) + H(y) \]
- 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:
  \[ H_{xy}(x) = \sum_{ij} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]
- More often written like:
  \[ H_{x|y}(x) = \sum_x p(x|y) \log \frac{p(x|y)}{p(y)} \]
- 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 these 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 amino acid types x, as follows:
  \[ H_{i}^{AB} = \sum_{x=1}^{n} p_{i,x}^{A} \log \frac{p_{i,x}^{A}}{p_{i,x}^{B}} \]

Simple Example: Relative Entropy ‘A/B’

Relative Entropy
- Captures similarities and differences
- Is infinite for completely dissimilar positions
  - e.g., A vs. I
- Not symmetrical:
  \[ H_{x}(y) \neq H_{y}(x) \]
- Maybe not easy for selecting dissimilar positions

Simple Example: Relative Entropy ‘A/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

**Measuring Overlapping Distributions**
- Weigh both groups equally:
  - take $p^* = p^3$ instead of $p^{1/3}$:
  \[
  SH_t = \sum p_i \log \frac{p_i^*}{p_i^{1/3}}
  \]
- Fixed interval $[0,1]$, but not completely symmetrical

**Entropy vs. Sequence Harmony: Example**
\[
\sum p_i^a \log \frac{p_i^a}{p_i^{1/3}} + \sum p_i^b \log \frac{p_i^b}{p_i^{1/3}}
\]

**Sequence Harmony**
- Introduce symmetry by averaging:
  \[
  SH^{Aa} = \frac{1}{2} SH(a) + \frac{1}{2} SH(b)
  \]
- May seem a trivial choice, but:
  \[
  SH^{Aa} = \frac{1}{2} \sum p_i^a \log p_i^a - \sum p_i^b \log p_i^b - \frac{1}{2} \sum p_i^a \log p_i^a - \frac{1}{2} \sum p_i^b \log p_i^b
  \]
- For N=2:
  \[
  SH^{Aa} = \frac{1}{2} \left( H(A) + H(A) + H(B) + H(B) - \log 2 \right)
  \]

**Analyzing multiple groups**
- Relative Entropy and Sequence Harmony are defined to compare a set of groups (N=2)
  - problem for multiple groups (N>2)
- One solution: Entropy-variability plots
  - Variability is number of different aminoacid types in a certain alignment position
  - Problem: Variability always tends towards maximum (20) for larger number of sequences
- Another solution: Two-entropy plots
  - Total family entropy vs. sum of sub-family entropy
  - Other methods...
Two-Entropies analysis of GPCRs
- Goal: to identify the function of individual positions

G-protein Coupled Receptors (GPCRs)
- Huge family of integral cell-membrane proteins
  - crucial in signal transduction
  - 70 subfamilies
  - 1995 Class A GPCRs
  - Three main regions:
    - extracellular side
    - transmembrane (TM)
    - cytoplasmic side

Two-entropies: upper vs. lower domain

Two-entropies: solvent accessibility

Two-entropies: ligand binding
- Subfamily specific binding
- Common activation mechanism

Red: upper domain
Blue: lower domain

Red: RSA < 15%
Blue: RSA > 15%

Red: <4 Å from retinal
Blue: >4 Å from retinal

GPCR Two-entropies analysis
Red: ligand binding
Green: coupling & activation
Blue: others
**Smad-MH2 Alignment & Functionally Specific Sites**

- 29 known sites of functional specificity
- based mostly on site-specific mutants and characterized by affinity for binding to BMPR-I vs. TBR-1 receptor types

<table>
<thead>
<tr>
<th>Method</th>
<th>Predict</th>
<th>Specificity</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAS</td>
<td>5</td>
<td>21%</td>
<td>8%</td>
</tr>
<tr>
<td>TreeDet</td>
<td>27</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>SDPpred</td>
<td>12</td>
<td>21%</td>
<td>17%</td>
</tr>
</tbody>
</table>

**Finding Low-harmony sites in Smad-MH2**

<table>
<thead>
<tr>
<th>Method</th>
<th>Predict</th>
<th>Specificity</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAS</td>
<td>5</td>
<td>21%</td>
<td>8%</td>
</tr>
<tr>
<td>TreeDet</td>
<td>27</td>
<td>99%</td>
<td>1%</td>
</tr>
<tr>
<td>SDPpred</td>
<td>12</td>
<td>21%</td>
<td>17%</td>
</tr>
</tbody>
</table>

**Conclusions Smad-MH2**

- 45 Sites of Low Sequence Harmony in Smad-MH2
  - different between the A1 (TGF-β) and B1 (E3) sub-type Smads
  - Low Harmony sites in Smad-MH2 are functionally relevant
  - Other methods cannot select all known sites!

> Functional Sites are Interaction Surfaces on Protein Surface:  
  - Note: Analyze Interaction Partners in the Pathway

- 14 Low Harmony Sites in Smad-MH2 of unknown function
  - 11 putative functions from structural considerations
  - promising candidates that determine TGF-β/BMP specificity
  - confirm (or refute) putative functions?

**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>
**RELIEF**

- Input: multiple sequence alignment containing two subgroups of sequences.
- Iteratively randomly select a sequence.
- Randomly select another sequence.
- 

**RELIEF for Sequence Analysis**

- Select closest neighbor in the same group and second closest neighbor in the other group.
- Calculate a weighted ranking distance for each feature.
- Update feature weight vector.
- Output feature weight vector with frequencies.

**RELIEF Pseudocode**

```plaintext
% input: X [two classes of aligned proteins]
% output: weights assigned to each site
% features are sites
% examples are sequences
nr_u = total number of features
weights = zero vector of size nr_u;
for all exa in X do
    hit(exa) = NMR of exa from same class;
    exit(hit(exa));
    weights += (exa-hit(exa)) - (hit(exa)-exa);
end;
return weights;
```

**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 weights

$$\text{weights}(s) = \begin{cases} \frac{1}{N} \sum_{i=1}^{N} I(W_i(s) > 0 \lor 0) & \text{for } S^+ > 0 \\ \frac{1}{N} \sum_{i=1}^{N} I(W_i(s) < 0 \lor 0) & \text{for } S^- > 0 \\ 0 & \text{for } S^- = 0 = S^+ = 0 \end{cases}$$

using $N^+ = |\{W_i(s) > 0 \lor 0\}|$ and $N^- = |\{W_i(s) < 0 \lor 0\}|$

Multi-RELIEF Pseudocode

Input: $s_1, s_2, \ldots, s_m$ (m classes of aligned proteins)
Parameters: $m_{\text{rel}}, m_{\text{max}}$
Output: $\text{multi}(s)$ (weights assigned to positions)

1. Initialize $\text{weights} = \text{zero vector of size } m_{\text{rel}}$
2. for all $s_i$
   1. $m_{\text{rel}}$ randomly two classes
   2. $s_1$ select randomly $m_{\text{max}}$ sequences
      from each selected class
   3. $s_2$ $W_i$ = apply RELIEF to $I$
   4. for all $s_i$
      1. $m_{\text{rel}}$ $W_i$ = average across positive $W_{i}(s_i)$'s
   5. return $\text{multi}(s)$

Weights computed by multi-RELIEF

- a toy example
- with four classes
- five types of sites:
  - a = conserved
  - b = partially specific
  - c = fully specific
  - d = non-conserved
  - e = conserved

Multi-RELIEF + 3D

- Functional specificity does not evolve for single residues, but involves a cluster of residues in the protein structure:
- multi-RELIEF can exploit 3D-structural information
- residue neighbour list
- A simple heuristic:
  - multi-RELIEF weights for each position $s$ are incremented by average weight of 3D neighbours of $s$
  - score will be boosted if neighbours have high scores
  - 3D neighbours share surface area:
    - calculated by the web server at http://highwear.imm.dtu.dk/3D
      (Sekowsky et al., 1999), with sequential neighbours removed

Properties of the Datasets

<table>
<thead>
<tr>
<th>dataset</th>
<th>m of classes</th>
<th>n of sites</th>
<th>max, min</th>
<th>n of classes</th>
<th>n of sites</th>
<th>info</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCR</td>
<td>77</td>
<td>3 (6-9)</td>
<td>100, 3</td>
<td>214</td>
<td>18</td>
<td>protein, ligand</td>
</tr>
<tr>
<td>GPCR100</td>
<td>39</td>
<td>4 (3-6)</td>
<td>21, 2</td>
<td>214</td>
<td>18</td>
<td>protein, ligand</td>
</tr>
<tr>
<td>Lali</td>
<td>15</td>
<td>3 (6-2)</td>
<td>12, 2</td>
<td>399</td>
<td>18</td>
<td>ligand, DNA</td>
</tr>
<tr>
<td>Ras/Pal</td>
<td>2</td>
<td>4 (5-24.5)</td>
<td>20, 20</td>
<td>218</td>
<td>18</td>
<td>protein</td>
</tr>
<tr>
<td>Rab/S</td>
<td>2</td>
<td>5 (0)</td>
<td>4, 6</td>
<td>162</td>
<td>18</td>
<td>protein</td>
</tr>
<tr>
<td>AGP/GLP</td>
<td>2</td>
<td>30 (0-10)</td>
<td>40, 12</td>
<td>450</td>
<td>18</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>18</td>
<td>protein</td>
</tr>
</tbody>
</table>

Related methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Authors</th>
<th>Test set compiled used</th>
</tr>
</thead>
<tbody>
<tr>
<td>TreeDes</td>
<td>Dat et al. 2003</td>
<td>protein, distance to ligand</td>
</tr>
<tr>
<td>SDHybrid</td>
<td>Kalinina 2004</td>
<td>Lali, ligand distance</td>
</tr>
<tr>
<td>Mimir &amp; O'Gall 2002</td>
<td>Ligand, ligand distance</td>
<td></td>
</tr>
<tr>
<td>TEA</td>
<td>Ye et al. 2005</td>
<td>GPCPs, ligand distance</td>
</tr>
<tr>
<td>SH</td>
<td>Pirrino et al. 2006, Fumagalli et al. 2007</td>
<td>Smad, Rab1/Rab2, Rab11, MIP, mutation data, ligand distance</td>
</tr>
</tbody>
</table>