Proteins in a subfamily usually share a specific function that is not common to the entire family. We investigate the use of clustering trees to identify such subfamilies.

### Protein function prediction

Several computational methods have been designed to assist scientists in the context of protein function prediction:

- **Homology-based methods**
  - Error prone: error propagation; proteins can change functions

- **Supervised learning approach**
  - Large amount of training data needed

- **Phylogenomic methods**
  - Use phylogenetic information
  - Example: SCI-PHY (Brown et al. 2007)

### Top-down hierarchical clustering: Clus-φ

**Divisive clustering algorithm Clus-φ**

- **Start with one cluster containing all sequences**
- **Repeat**
  - **Split cluster**
  - Clus-φ uses the minimum evolution hypothesis to choose the best split: it constructs a tree with minimal total branch length (cfr. Neighbor Joining).
- **Until there is only one sequence per cluster**

### Applying Clus-φ to identify subfamilies

**Problem: how to extract clusters from the tree?**

- Post-pruning based on minimum description length (minimizing encoding cost) – (cfr. SCI-PHY)

**Procedure:**

1. Start with complete tree
2. Merge 2 leaves that most reduces the encoding cost
3. Repeat until one node left
4. Choose clustering with minimum encoding cost

**Advantages over existing phylogenomic methods:**

- Allows to identify functional sites
- Allows to directly classify new sequences
- No need to build the complete tree if only high level grouping of subfamilies is wanted

### Experiments

#### 1 - Feasibility

**Question:** Are polymorphic positions useful to determine protein subfamily identification?

**Setting:** We added the subfamily information to the data and induced a classification tree using Clus (i.e., supervised classification task)

<table>
<thead>
<tr>
<th>Subfamilies</th>
<th>Classification tree # leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3A</td>
<td>16</td>
</tr>
<tr>
<td>3A'</td>
<td>15</td>
</tr>
<tr>
<td>3A''</td>
<td>14</td>
</tr>
<tr>
<td>3B</td>
<td>12</td>
</tr>
<tr>
<td>3B'</td>
<td>11</td>
</tr>
</tbody>
</table>

**Results:**

- Subfamilies can be perfectly separated from one another using compact trees.
- On average, the classification tree is 1.2 times larger than the ideal case.

#### 2 - Evaluating the predicted protein subfamilies

**Question:** Do the predicted clusters correspond to the protein subfamilies defined by the biologist?  

**Setting:** We compared our results with the ones given by SCI-PHY and Neighbor-Joining – NJ (with post-pruning based on encoding cost)

**Evaluation measures:** We used several clustering evaluation measures proposed in the literature (purity, edit distance, VI distance, category utility)

**Results:**

- Clus-φ vs NJ: Clus-φ has in general better results than NJ
- Clus-φ vs SCI-PHY: no clear winner (each method has better results for about half of the comparisons)

#### 3 - Evaluating the underlying tree

**Question:** How meaningful is the tree structure?

**Setting:** We compared the Clus-φ trees with the SCI-PHY and NJ trees

**Evaluation measures:** We propose 2 measures to evaluate the trees:

- Number of protein subfamily switches (cfr. Fitch score)
- Edited tree size (most compact tree with pure leaves)

**Results:**

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of protein subfamily switches</th>
<th>Edited tree size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clus-φ</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>SCI-PHY</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>NJ</td>
<td>14</td>
<td>20</td>
</tr>
</tbody>
</table>

Clus-φ vs SCI-PHY: Clus-φ has mostly a smaller edited tree but a higher number of protein subfamily switches

Clus-φ vs NJ: Clus-φ trees are better evaluated according to the 2 measures

Visual inspection of a biologist of 5 datasets - 4 wins

For Clus-φ tree / 1 undecided case