AlignMask

- Improving masking with multiple sequence alignment?

Wei Qian & Jinfeng Xiao

05/01/2018
CS 581 Project Presentation
Motivation - Masking is Proposed

[Image of a phylogenetic tree with various plant species]

- Erigeron_tenuis
- Kippista_ssaedifolia
- Ursina_speciosa
- Matricaria_chamomila
- Senecio_vulgaris
- Synedrella_nodiflora
- Xerochrysum_bracteatum
- Lactuca_serriola
- Nassauvia_revoluta
- Calycera_crasifolia

[Additional images and links for further details]
Motivation - Current Masking Methods

[Graph showing amino acid alignments with different methods for Fungi, Eukaryotes, and Bacteria.]
Motivation - Heuristic Based on Column Annotated Score (CAS)
Motivation - Heuristic Based on Column Annotated Score (CAS)
Motivation - Heuristic Based on Column Annotated Score (CAS)

... .92 .01 .34 .58 ...
... A C - D ...
... A - F B ...
... A - E D ...
... ... ... ...
...
Motivation - Heuristic Based on Column Annotated Score (CAS)
Approach - PASTA: Multiple Sequence Generations

FIG. 1. Algorithmic design of PASTA. The first six boxes show the steps involved in one iteration of PASTA. The last two boxes show the meaning of transitivity for homologies defined by a column of an MSA, and how the concept of transitivity can be used to merge two compatible and overlapping alignments. MSA, multiple sequence alignment.
Approach - WeaveAlign: Consensus Generation

a)

\[
\begin{array}{cccc}
- & A & B & - & C \\
A & - & B & C & - \\
- & A & B & - & C
\end{array}
\rightarrow \begin{array}{cccc}
0 & 1 & 3 & 4 & 5 \\
1 & 2 & 3 & 5 & 6 \\
1 & 2 & 3 & 5 & 6 \\
0 & 1 & 3 & 4 & 5
\end{array}
\]

b)

\[
\begin{array}{cccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 \\
1 & 2 & 3 & 4 & 5 & 6 & 6 \\
1 & 2 & 3 & 4 & 5 & 6 & 6
\end{array}
\]
Pairwise Column Annotation Score (pCAS)

Given a library of $n$ input MSAs $\{A_1, \ldots, A_n\}$ and a consensus $A_c$. The pairwise column annotation score $pCAS$ for the $j^{th}$ column of $A_c$ and the $i^{th}$ input MSA $A_i$ is defined as

$$pCAS_j(A_c, A_i) = \frac{|H_{cj} \cap H_i|}{|H_c|}$$

where $H_i$ is the set of homology pairs in $A_i$, $H_{cj}$ is the set of homology pairs in the $j^{th}$ column of $A_c$, and $|H|$ denotes the size of a set of homology pairs.
\begin{align*}
s_1 &= AC \\
s_2 &= -C \\
s_3 &= A- \\
s_4 &= T- \\
s_5 &= G- \\
s_1 &= N_{1,1}N_{1,2} \\
s_2 &= -N_{2,1} \\
s_3 &= N_{3,1} - \\
s_4 &= N_{4,1} - \\
s_5 &= N_{5,1} - \\
s_1 &= N_{1,1}N_{1,2} \\
s_2 &= -N_{2,1} \\
s_3 &= -N_{3,1} \\
s_4 &= -N_{4,1} \\
s_5 &= N_{5,1} -
\end{align*}
Pipeline

**Approach**

- **PASTA**: MSAs Generation
- **WeaveAlign**: Consensus Generation
- **FastSP**: Column Annotation
- **SP Score Mask**
- **Random Mask**
- **RAxML**: Tree Estimation
- **RF Score Evaluation**
Table 1. Statistic for the Experiment Dataset

<table>
<thead>
<tr>
<th></th>
<th>1GADBL_100</th>
<th>coli_epi_100</th>
<th>RV100_BBA0117</th>
<th>RV100_BBA0154</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sequences</td>
<td>561</td>
<td>320</td>
<td>460</td>
<td>303</td>
</tr>
<tr>
<td># of sites</td>
<td>490</td>
<td>150</td>
<td>110</td>
<td>1275</td>
</tr>
</tbody>
</table>
Result - Does it work?

Not really...
Result - Does it work?

Score seems to be meaningful...

But really?
Result - CAS and Non-Gap Frequency
Alternative CAS

What we have seen:

\[ pCAS_j(A_c, A_i) = \frac{|H_{cj} \cap H_i|}{|H_c|} \]

Alternative:

\[ apCAS_j(A_c, A_i) = \frac{|H_{cj} \cap H_i|}{|H_{cj}|} \]
\[ s_1 = AC \quad s_1 = N_{1,1}N_{1,2} \]
\[ s_2 = -C \quad s_2 = -N_{2,1} \]
\[ s_3 = A- \quad s_3 = N_{3,1}- \]
\[ s_4 = T- \quad s_4 = N_{4,1}- \]
\[ s_5 = G- \quad s_5 = N_{5,1}- \]
Soft Filtering

- NOW: filtering and tree estimation is two separate processes

- CAS can be used as column weight in the likelihood calculation or edit distance calculation
Impact of Alignment Method

- Maybe PASTA / WeaveAlign already provide a good alignment

- Test with another alignment methods?
Discussion

RAxML Hyper-Parameter Tuning

- Simplest substitution model for amino acid

- Fair comparison if result if not the optimal?
Larger Dataset

- Might be in a different problem domain