GYM Training Set Selection Part-2

Jerry Deng

Yanli Sun

3. Algorithm

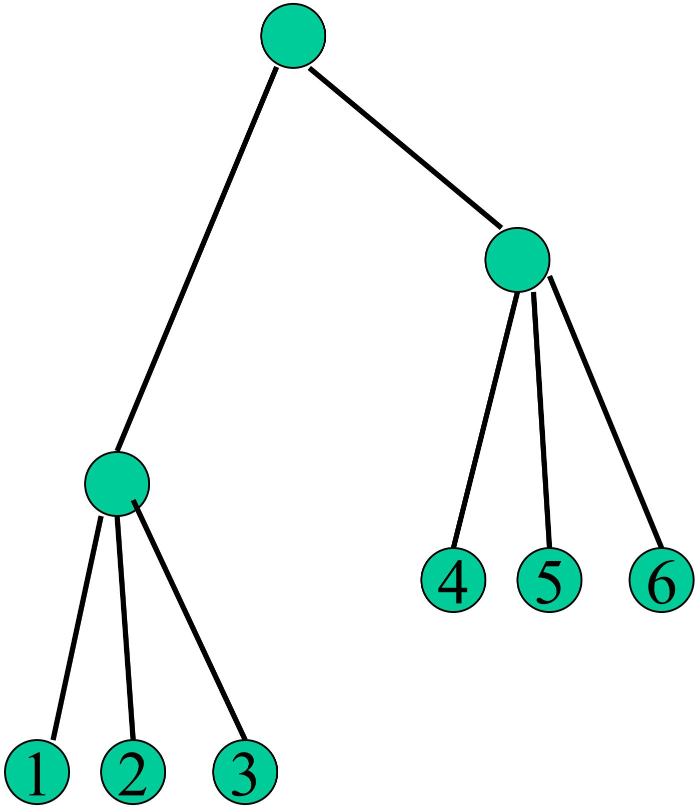
- To remove some very similar (or duplicated) sequences from the training set in order to reduce both pure and partial spurious patterns in pattern dictionary.
- Use Phylogenetic Tree to figure out similarity among sequences.

3.1. Similarity Measurement

- Alignment Score
  - either pairwise or multiple
  - unable to give further information such as evolutionary and classification information.
- Phylogenetic Tree
  - evolution family classification
  - evolution distance.
  - Black box

3.2. Phylogenetic Tree

- Sequences in one family are closer than those from different families.
- For an internal node, the farther it is away from the root, the more likely that all its descents are closer.

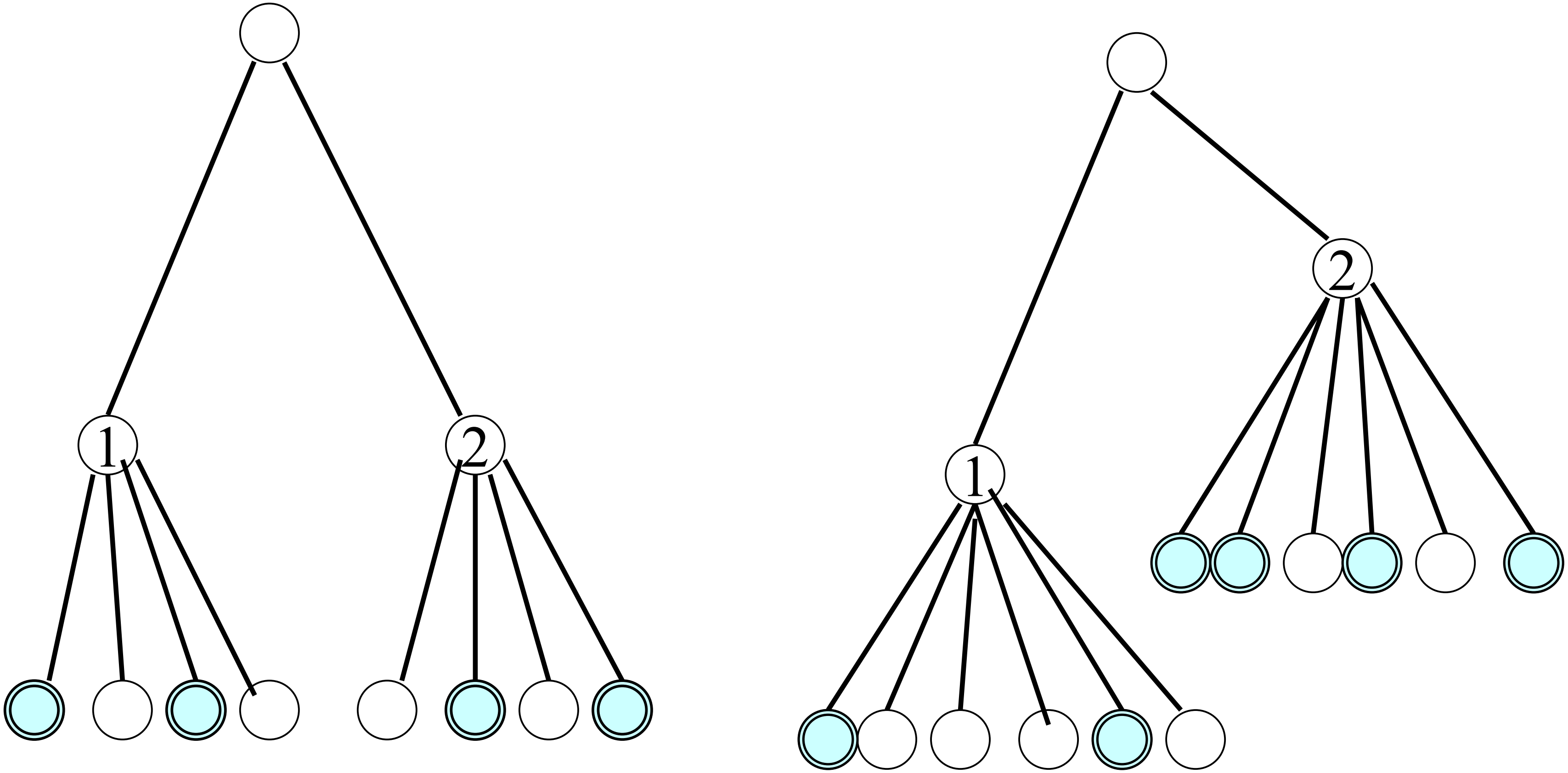


3.3. Node Score

- To precisely measure the evolutionary similarity, A score was recursively calculated for each node.
  - The root's value is 0
  - For a non-root node:

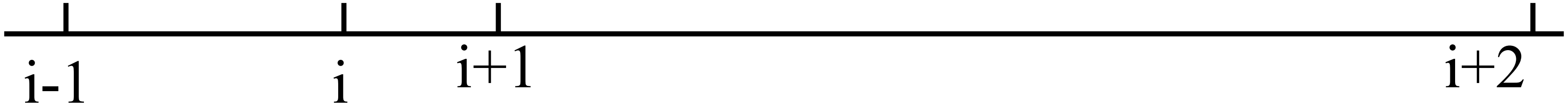
$$(Index+B*(1- DistanceP)* (A*Degree)^{(TD\text{-}depth)}$$
$$+G*(1-Distance)$$

  - Degree: The maximum branching degree present in the tree.
  - TD: Total depth of the tree
  - Depth: The depth of current node
  - Index: The sequential index among siblings
  - DistanceP: The total distance between root and its parent node.
  - Distance: The distance from the underlying node to its parent
  - A: Constant factor, default is 1.5
  - B: Constant factor, default is 1
  - G: Constant factor, default is 1


1
2
1
2

3.4. Selection

- The difference between the scores of two nodes reflects the evolutionary similarity between them.
- Once the scores have been obtained for all sequences (nodes), they are put into a list in order by traversing the tree in a depth-first manner.
- Repeatedly find out the pair of nodes of the smallest difference, remove one of them from the list, until the number of sequences contained reaches the desired number.



4. Implementation

- Three steps:
  - tree parsing
  - calculating scores
  - selecting training set
- Application
  - Input
    - tree script
    - desired size of training set
  - Output
    - training set

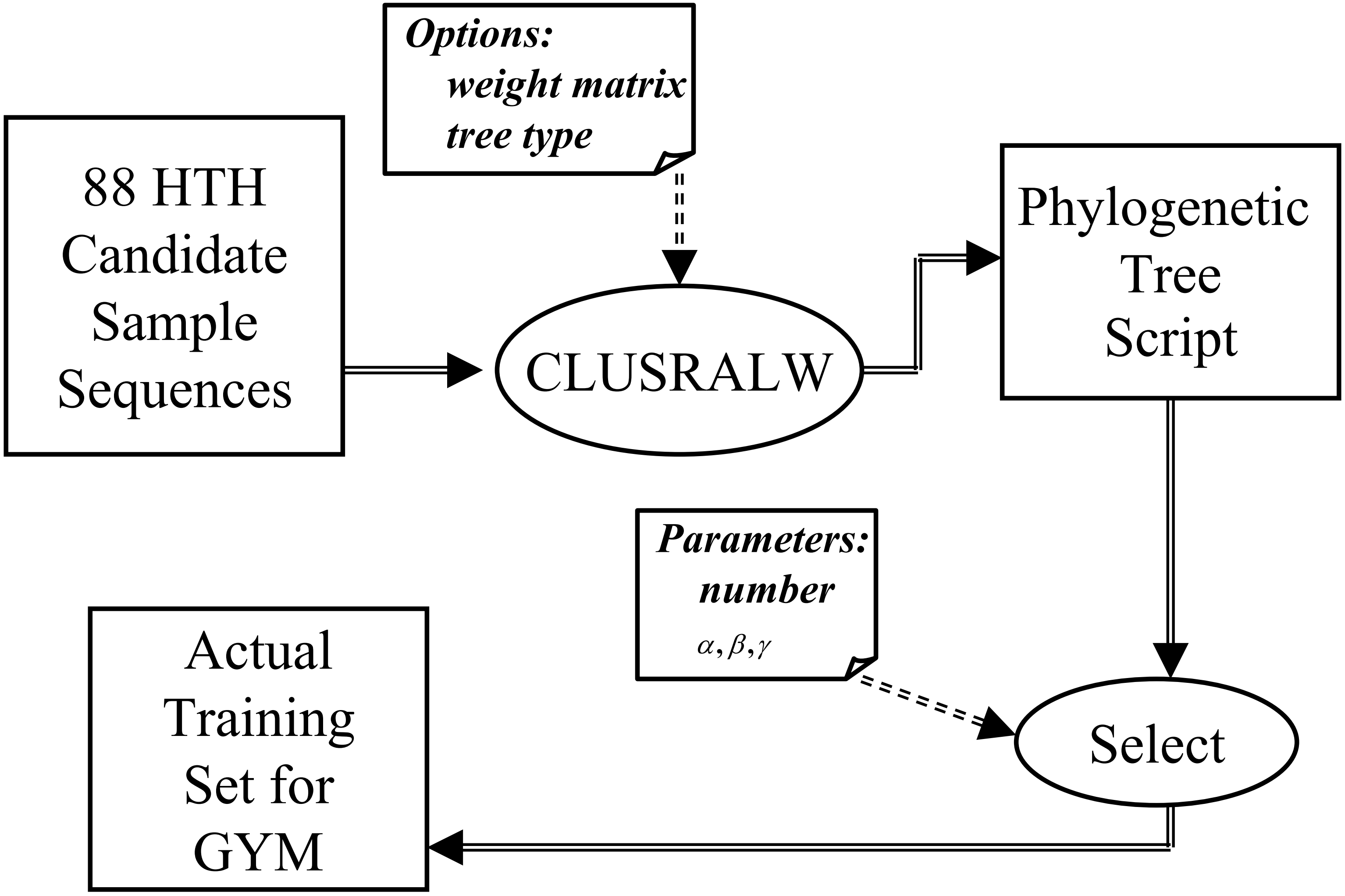
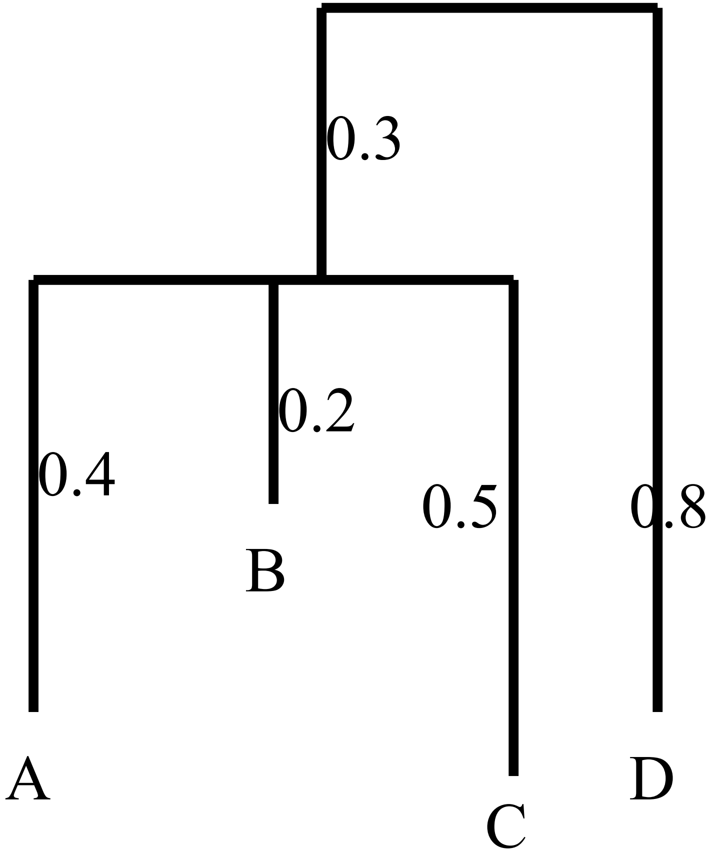


Illustration of Data Flow

```
(
   (   A:0.4,
       B:0.2,
       C:0.5
   ):0.3,
   D:0.8
);
```



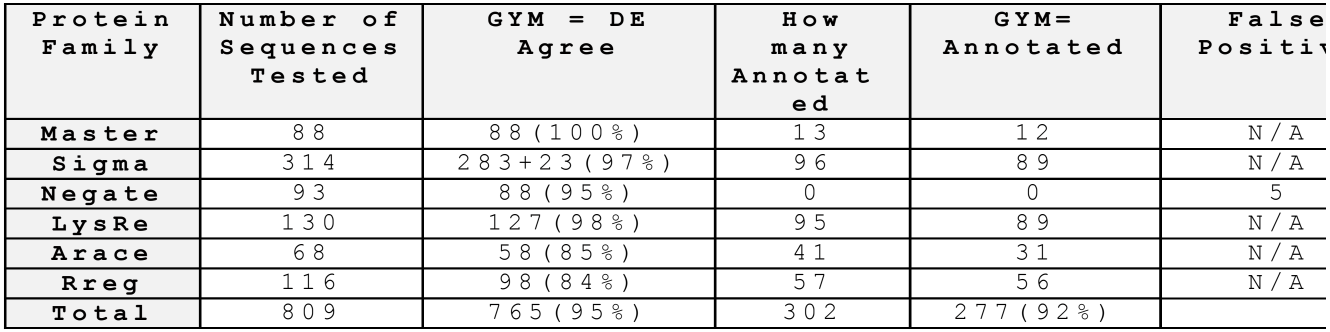
```
Tree:= (NodeList);

NodeList:= Node,NodeList  |
           Node
Node:= (NodeList):distance |
        seq_num:distance
```

Script Passing

5. Test Result

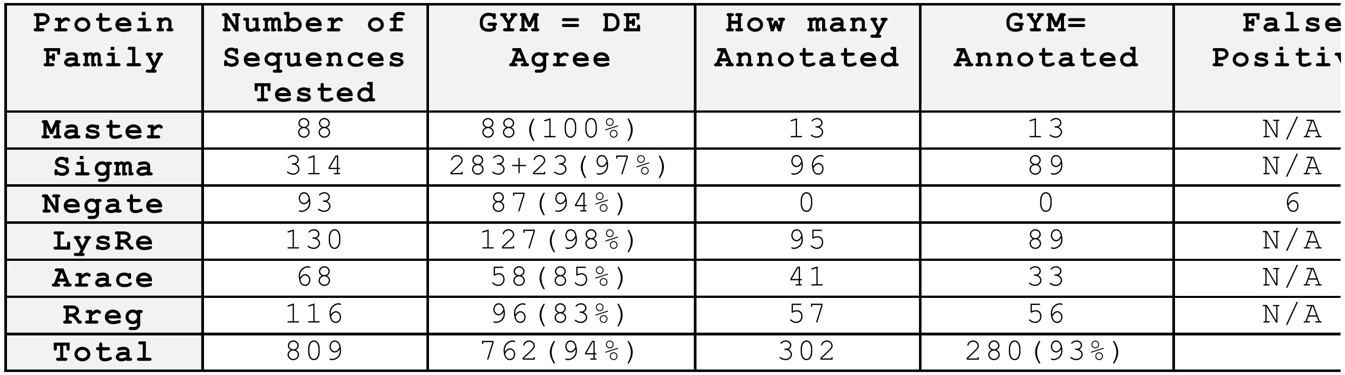
We testing was performed on

- two different phylogenetic trees from CLUSTAL based on GYM's Master Set in our testing.
  - One was generated with the default setup
  - other one was generated wit BLOSUM matrix and PHILIP tree type
- Various number of sequences in the training set out of 88.



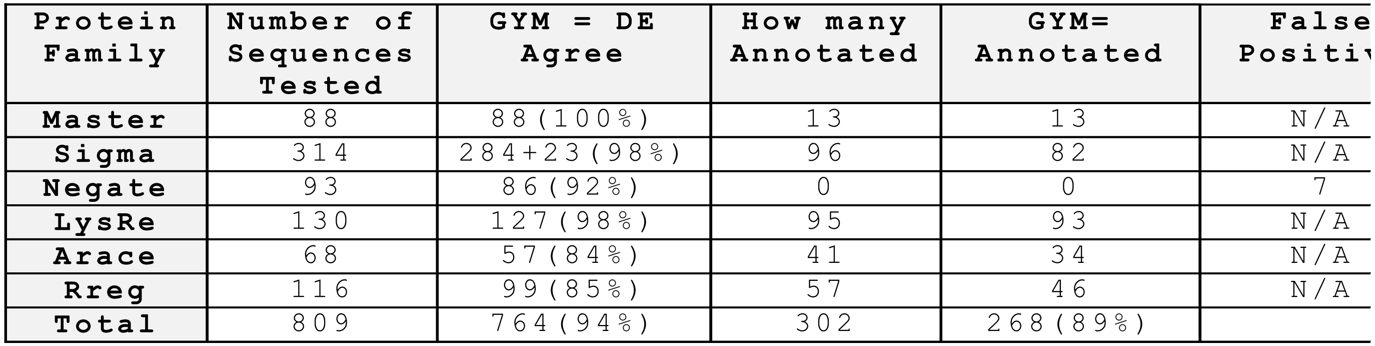
Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positi
Master	88	88(100%)	13	12	N/A
Sigma	314	283+23(97%)	96	89	N/A
Negate	93	88(95%)	0	0	5
LysRe	130	127(98%)	95	89	N/A
Arace	68	58(85%)	41	31	N/A
Rreg	116	98(84%)	57	56	N/A
Total	809	765(95%)	302	277(92%)	

BLOSUM,
PHILIP,
82



Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positi
Master	88	88(100%)	13	13	N/A
Sigma	314	283+23(97%)	96	89	N/A
Negate	93	87(94%)	0	0	6
LysRe	130	127(98%)	95	89	N/A
Arace	68	58(85%)	41	33	N/A
Rreg	116	96(83%)	57	56	N/A
Total	809	762(94%)	302	280(93%)	

BLOSUM,
PHILIP,
84



Protein Family	Number of Sequences Tested	GYM = DE Agree	How many Annotated	GYM= Annotated	False Positi
Master	88	88(100%)	13	13	N/A
Sigma	314	284+23(98%)	96	82	N/A
Negate	93	86(92%)	0	0	7
LysRe	130	127(98%)	95	93	N/A
Arace	68	57(84%)	41	34	N/A
Rreg	116	99(85%)	57	46	N/A
Total	809	764(94%)	302	268(89%)	

GYM
Original
Result

From above testing results, we can see:

- slightly increased detection rate on some protein families (e.g. Sigma, Rege and Lysr) where HTH motif existences are verified
- decreased false positive rate on Negates family where HTH motif is unlikely to exist.

With just a few (4~6) very similar HTM motif sequences removed

6. Conclusion

- This project presents an effective approach for training set refinement in pattern mining by means of similarity control among sequences.
- For pattern mining, like the two sides of a coin, similarity represents the trade-off between the sensitivity of both true positives and false positives.
- In practice, the optimal similarity control can only be achieved by experiments. Theoretically, there is no algorithm that can automatically figure out the optimal similarity threshold without further biological knowledge.