A New Approach for Clustering of Gene Sequences

Tanjil Ahmed[1], Rachel Scwartz[2]

[1] Department of Geosciences, University of Rhode Island
[2] Department of Biological Sciences, University of Rhode Island

ABSTRACT

In modern days, with the development of biotechnology results in the vast amount of raw biological data. Thousands of sequence data appear every day. So to identify the similar sequence into a group is necessary for different bioinformatics problems like phylogeny construction, genomic sequence analysis, gene finding, gene mapping. Due to the massive amount of sequences, computational complexity for clustering sequences increases day by day. Here a new approach is proposed for clustering gene sequences.

CLUSTERING

Clustering is the process of identifying similar characteristics and grouping data with similar characteristics together. These groups are called clusters. Clustering approach can be viewed as extrinsic or intrinsic. Extrinsic technique implies traditional classification supervised learning algorithm in which a particular input training is used. Intrinsic algorithms do not use any a priori category labels but depend only on a similarity measure. The proposed algorithm falls into the intrinsic class. For each cluster, representative genes are identified by approximated weighted length technique. It is called CURE (Clustering Using REpresentatives) approach

PROPOSED ALGORITHM: RaTa (Definition)

n_i = frequency of i type element in input database, $i \{A,C,T,G\}$
N = total occurrence of each type of element = $\sum n_i$
w_i = weight of i type element of gene sequence = n_i/N
W_g = weight of the gene sequence g = $\sum w_i$
l_g = lenght of the gene sequence g
W_{cmax} = weight of the highest length gene of cluster c
L_{cmax} = max(l_g) for each gene g in cluster c
ref_c = reference gene of cluster c
$S(g, ref_c)$ = function denoting dissimilarity measure between g and $ref_c = | \times l_{ref} - W_g \times l_g |$
A_T = currently assigned cluster number
$MinDism$ = dissimilarity measure between input sequence g and cluster $A_T = S(g, ref_{A_c})$
H_c = threshold value of cluster c = $W_{cmax} \times L_{cmax} \times 55\%$
s_c = current size (number of sequences) of cluster c
$avgval_x$ = average length per sequence of cluster x
$= \frac{\text{total length of sequences}}{\text{total number of sequences}}$
V_c = weighted-length of the cluster c
$= \sum W_g \times l_g$, gene g in cluster c

PROPOSED ALGORITHM: RaTa (Method)

Input: String of sequences (FASTA format).
Output: Clusters of sequences.

Begin
1. Initialize *MinDism* to INFINITY and A_T to 0
2. Assign weight w_i to each different element i of gene sequence in input database.
3. Make the first sequence as first cluster and mark it as reference gene of that cluster.
4. Calculate V_1 and H_1 according to definition.
5. When a new sequence g arrives
   5.1. For each cluster i
      5.1.1. Find $S(g, ref_i)$
      5.1.2. if $S(g, ref_i) < H_i$ and $S(g, ref_i) < Min\_Dis$
         5.1.2.1. $MinDism = S(g, ref_i)$
         5.1.2.2. $A_T = i$.
   5.2. If A_T not equal to zero
      5.2.1. assign the sequence to cluster A_T
      5.2.2. For cluster A_T
         5.2.2.1. For new sequence g
            5.2.2.1.1. Calculate W_g
            5.2.2.1.2. $temp1 = W_g \times l_g$
            5.2.2.1.3. $temp2 = V_{A_T} + temp1$
            5.2.2.1.4. $S_{A_T} = S_{A_T} + 1$
            5.2.2.1.5. $avgval_{A_T} = \frac{temp2}{S_{A_T}}$
            5.2.2.1.6. $V_{A_T} = temp2$
            5.2.2.1.7. find the sequence in the cluster whose gene value is closest to and make it reference gene sequence.
   5.3. If A_T equal to zero
      5.3.1. create a new cluster d and assign the sequence to it
      5.3.2. make the sequence as reference gene
      5.3.3. calculate V_d
      5.3.4. $A_T = d$
6. Calculate
7. $MinDism = INFINITY$, $A_T = 0$
8. Repeat steps 5 - 7 until all the available sequences are clustered.

END

REFERENCES

1. **"Interactive clustering model for explanation of Genomic Data", www.cse.msstate.edu**
2. **D. Jiang, C. Tang, A. Zhang, "Cluster analysis for gene expression data: A survey", *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370-1386, Nov. 2004.**

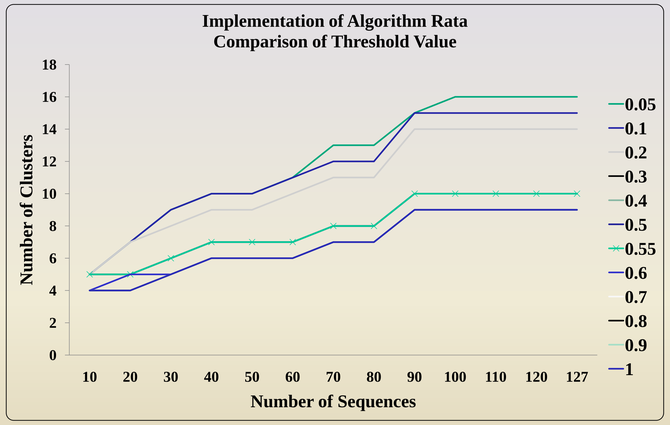
RESULTS & OBSERVATION

Space Complexity:

Here maximum possible number of clusters is n. Again the maximum possible gene sequences in a cluster are also n and the maximum length of each sequence is L. So, the space complexity is $O(n^2L)$.

Time Complexity:

Line (1-4) need O(1) time. Step 5 needs O(n) time. All other lines require constant time operation. So for n sequence, the total running time of the algorithm is $O(n^2)$.



Implementation of algorithm shows (see graph) number of cluster increases with number of sequences; But remains almost static at some stages. Curves are obtained by varying the values of the threshold value H from 0.05 to 1. Curve will be smooth based on number of input sequences and intervals. Marked line refer average threshold value 0.55, which mentioned in algorithm

Conclusion and Future Works

Algorithm output can be used as input for phylogenetic tree constructions

Future scope includes:

1. **Inclusion of dynamic clustering**
2. **Proof of optimality of the algorithm in parallel computation**

Contact

Tanjil Ahmed
Graduate Research Assistant
University of Rhode Island
E-mail: tanjil@uri.edu