Clustering trees for protein subfamily identification, classification and characterization

Eduardo De Paula Costa[1], Celine Vens[1], Hendrik Blockeel[1,2]

[1]Department of Computer Science, KU Leuven, Belgium
[2]Leiden Institute of Advanced Computer Science, Universiteit Leiden, The Netherlands

We consider the task of protein subfamily identification: given a set of sequences that belong to one protein family, identify subfamilies of functionally closely related sequences. This is in essence a clustering task. Most current methods for subfamily identification use a bottom-up clustering method to construct a cluster hierarchy, and then cut the hierarchy at the most appropriate locations to obtain a single partitioning. Such approaches rely on the assumption that functionally similar proteins have sequences with a high overall similarity, but do not exploit the fact that these sequences are likely to be highly conserved at particular positions. This raises the question to what extent clustering procedures can be improved by making them exploit this property.

We have proposed an alternative clustering procedure [1] that uses the “top-down induction of clustering trees” approach [2]. This approach differs from bottom-up clustering methods in that it forms clusters whose elements do not only have high overall similarity, but also have particular properties in common. In the case of subfamily identification, these properties are the amino acids found at particular positions. Apart from possibly yielding higher quality clusterings, this approach has two important additional advantages. First, it allows for easy classification of new sequences into a subfamily. Starting at the root node, a new sequence is moved down the tree by checking its residues at the identified positions, until it is classified into one of the predicted subfamilies. Second, the identified tests result in a candidate list of functionally important sites, i.e., positions that are likely to play a role in the subfamily-specific functions.

The method [1] starts from a multiple sequence alignment, and tries to split the sequences into subsets such that (1) sequences within a subset are similar, and (2) the split is defined by a test "p∈S", with p a position and S a set of amino acids. The splitting criterion was designed specifically for the phylogenetic context [3], and can be seen as a top-down counterpart of the joining criterion used by Neighbor Joining. After dividing a set into two subsets, the same principle can be used to further subdivide the subsets, up to the level of singletons. This yields a hierarchical tree, which is then cut at particular locations. The criterion used to extract clusters is based on encoding cost [4], i.e., the cost to encode a clustering given the homogeneity of the clusters and the number of clusters. The resulting clusters, which correspond to the predicted protein subfamilies, are then output along with the underlying tree, which explicates how the clusters were split and which tests were used.

We have evaluated the proposed approach on 11 publicly available datasets, using a wide range of evaluation measures. Our results can be summarized as follows. First, splits based on polymorphic positions (i.e., positions that have more than one amino acid residue) are highly discriminative between protein subfamilies. Second, using such splits to guide a clustering procedure improves protein subfamily identification over the state-of-the-art method SCI-PHY [4]. Third, the identified positions yield accurate classification of new sequences, with accuracy close to that of constructing a profile HMM for each predicted subfamily. Finally, the resulting clustering tree identifies functionally important sites at prominent nodes.

References
[1] E. De Paula Costa, C. Vens, H. Blockeel “Top-down clustering for protein subfamily identification”, Evolutionary Bioinformatics , vol. 9, pp. 185-202, 2013.
[2] H. Blockeel, L. De Raedt, and J. Ramon, “Top-down induction of clustering trees,” in Proc. of the 15th International Conference on Machine Learning, 1998, pp. 55–63.
[3] C. Vens, E. P. Costa, and H. Blockeel, “Top-down induction of phylogenetic trees,” in Proc. of the 7th Eur. Conf. on Evol. Computation, Machine Learning and Data Mining in Bioinformatics. LNCS, vol. 6023. Springer, 2010, pp. 62–73.
[4] D. Brown, N. Krishnamurthy, and K. Sjolander, “Automated protein subfamily identification and classification,” PLoS computational biology, vol. 3, no. 8, p. e160, 2007.