Synthetic Test Data Generation for Hierarchical Graph Clustering Methods *

László Szilágyi[1,2], Levente Kovács[3], and Sándor Miklós Szilágyi[4]

[1] Dept. of Control Engineering and Information Technology,
Budapest University of Technology and Economics, Hungary
[2] Sapientia - Hungarian Science University of Transylvania, Romania
lazacika@yahoo.com
[3] Óbuda University of Budapest, Hungary
[4] Dept. of Informatics, Petru Maior University of Tîrgu Mureş, Romania

Abstract. Recent achievements in graph-based clustering algorithms revealed the need for large-scale test data sets. This paper introduces a procedure that can provide synthetic but realistic test data to the hierarchical Markov clustering algorithm. Being created according to the structure and properties of the SCOP95 protein sequence data set, the synthetic data act as a collection of proteins organized in a four-level hierarchy and a similarity matrix containing pairwise similarity values of the proteins. An ultimate high-speed TRIBE-MCL algorithm was employed to validate the synthetic data. Generated data sets have a healthy amount of variability due to the randomness in the processing, and are suitable for testing graph-based clustering algorithms on large-scale data.

Keywords: bioinformatics, fast Markov clustering, synthetic test data.

1 Introduction

Bioinformatics is one of the fields where there is an excessive need for clustering algorithms that are capable to handle large-scale protein sequence or interaction networks [2]. Graph-based algorithms usually need to store the matrix representation of the graph [5], which becomes prohibitively costly in memory storage above 10^4 nodes. Sparse matrix models made it possible to extend this limit towards one million nodes [12]. However, there are few publicly available large data sets, and even those existing ones are not suitable for a wide variety of algorithms.

Using synthetic test data is a frequently employed method, even if real data is also available (e.g. [4,13]). Our main goal is to provide synthetic but realistic test data for clustering algorithm designed to process large-scale protein sequence data sets. Our principal target is the Markov clustering algorithm, and

* Research supported by the Hungarian National Research Funds (OTKA), Project no. PD103921, the János Bolyai Fellowship Program of the Hungarian Academy of Sciences. Decision on support by Sapientia Institute for Research Programs is pending.

L.C. Kiong et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 303–310, 2014.

more exactly the TRIBE-MCL [5], which groups protein sequence data based on pairwise similarity measures stored in a similarity matrix. Synthetic data is created in two steps: first the main properties and attributes of the 11944-node similarity graph and corresponding similarity matrix of the SCOP95 data set [8] are identified, and then a random data set is generated, which has similar properties and the desired size. Properties considered by the proposed method include: four-level hierarchy of SCOP95, distribution of protein family sizes, density and distribution of nonzero values in various locations of the similarity matrix.

The rest of this paper is structured as follows. Section 2 presents background information on the structure and properties of the SCOP95 data set, and the TRIBE-MCL algorithm that will be used for test purposes. Section 3 presents the details of the proposed property identification and synthetic test data generation process. Section 4 produces a numerical analysis to support the validity of the produced synthetic data. Conclusions are given in the last section.

2 Background

2.1 The SCOP95 Database

The Structural Classification of Proteins (SCOP) database [10] contains protein sequences in order of tens of thousands, which are organized in a four-level hierarchy composed by classes, folds, superfamilies and families [2]. This hierarchy can be employed as ground truth for protein sequence clustering algorithms. The SCOP95 database that we use as input is a subset of SCOP (version 1.69), which contains 11944 proteins, exhibiting a maximum similarity of 95% among each other. Pairwise similarity matrices (e.g. BLAST [1], Smith-Waterman [9], Needleman-Wunsch [7]) are also available at the Protein Classification Benchmark Collection [8]. Our purpose is best served by the BLAST matrix due to its sparse nature: in its symmetrized version it has a density of 0.00387 indicating that an average node in the similarity graph is connected to 45 other nodes.

2.2 TRIBE-MCL Markov Clustering

TRIBE-MCL is an efficient clustering method based on Markov chain theory introduced by Enright et al [5]. TRIBE-MCL assigns a graph structure to the protein set such a way that each protein has a corresponding node. Edge weights are stored in the so-called similarity matrix S, which acts as a stochastic matrix. At any moment, edge weight s_{ij} reflects the posterior probability that protein i and protein j have a common evolutionary ancestor. TRIBE-MCL is an iterative algorithm, performing in each loop two main operations on the similarity matrix: inflation and expansion. Inflation raises each element of the similarity matrix to power r, which is a previously established fixed inflation rate. Due to the constraint $r > 1$, inflation favors higher similarity values in the detriment of lower ones. Expansion, performed by raising matrix S to the second power, is aimed to favor longer walks along the graph. Further operations like column or row normalization, and matrix symmetrization are included to serve the stability and

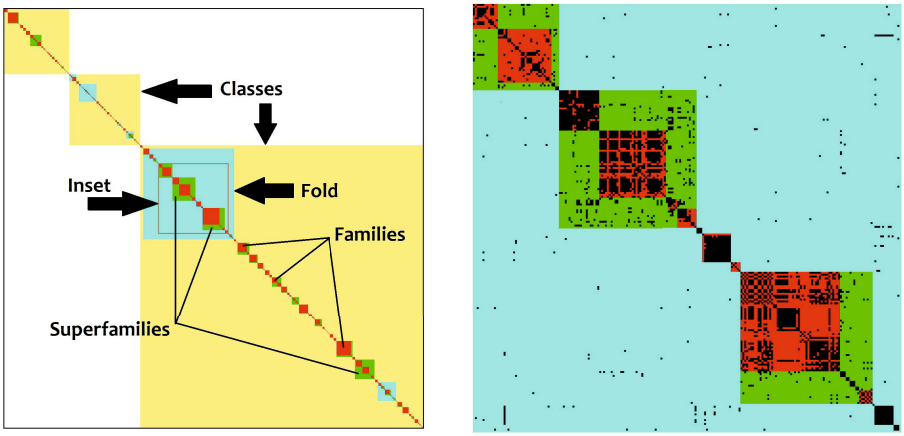

Fig. 1. A selected part (classes e-g) of the BLAST similarity matrix of the SCOP95 data set indicating the hierarchy of classes, folds, superfamilies and families (left), and the magnified view of superfamilies g.3.6-g.3.14 within the inset (right). Black pixels on the right image represent the nonzero values in the matrix.

robustness of the algorithm, and to enforce the probabilistic constraint. Similarity values that fall below a previously defined threshold value ε are rounded to zero. Clusters are obtained as connected subgraphs in the graph. Further details on TRIBE-MCL operations are available in [5,11].

3 Methods

3.1 Identification of SCOP95's Properties

The identification of the SCOP95 matrix properties is performed in several steps. Proteins are grouped into families of various sizes (1-557 members), each represented by a square block situated on the matrix diagonal. These diagonal blocks are not very sparse as they contain over two thirds of all nonzero values in the matrix. Superfamilies are small groups of families represented by larger diagonal blocks that include the blocks of contained families. Similarity values within the superfamily blocks but outside the family blocks are significantly less dense, and they become sparser as the distance from the diagonal grows. The structure of the similarity matrix is depicted in Fig. 1.

According to the recently developed theory of natural networks [3], the number of connections the graph nodes have follows a negative power distribution. This is also valid in case of the SCOP95 graph: both the distribution of connections and the distribution of family sizes share this attribute. Figure 2 exhibits the approximation of this distribution in the range of families with up to 80 proteins. A few larger families are also present in SCOP95, their distribution also follows the rule of the power distribution.

Table 1. Identified parameter values for families of various sizes

Proteins in family	Average density	Families of full density	Average density in not fully dense families	Average similarity value
2	0.827	82.7 %	0.000	0.440
3	0.809	70.4 %	0.353	0.421
4-5	0.733	51.5 %	0.452	0.382
6-10	0.659	31.3 %	0.507	0.358
11-19	0.617	14.9 %	0.547	0.318
20-99	0.485	6.67 %	0.470	0.251
100+	0.807	0.00 %	0.807	0.315

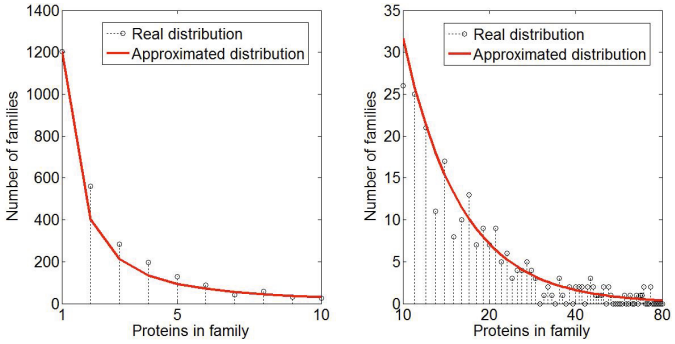

Fig. 2. Family sizes follows a negative power distribution: the amount of families of size n is proportional with n^{-k}

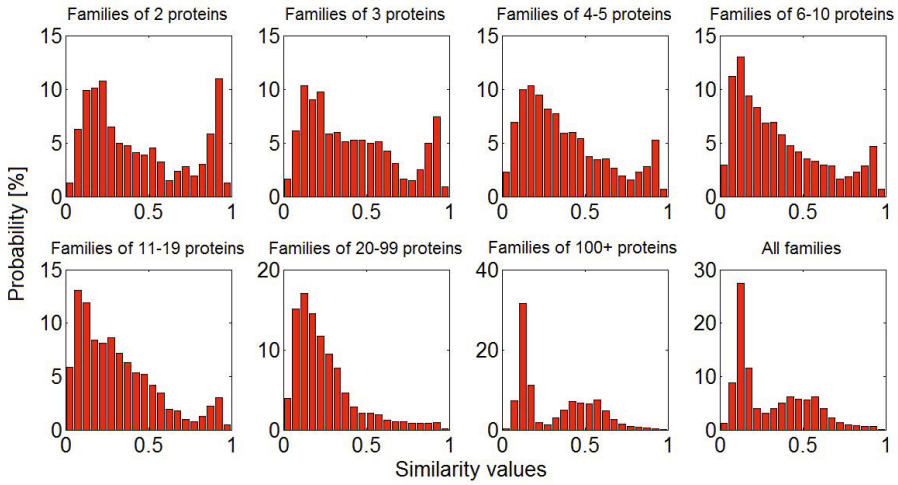

Fig. 3. Distribution of nonzero similarity values in the SCOP95 matrix, in case of protein couples situated in the same family: various distributions for all family sizes

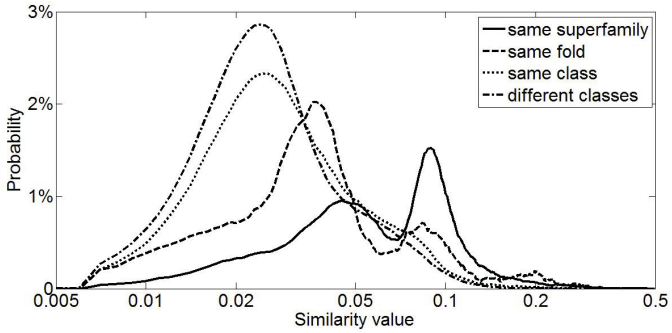

Fig. 4. Distribution of nonzero similarity values in the SCOP95 matrix, in case of protein couples situated in different families, but in the same superfamily, fold, or class, and also for proteins from different classes

Table 2. Identified densities in various parts of the SCOP95 BLAST similarity matrix

Same family	Different families same superfamily	Different superfamilies same fold	Different folds same class	Different classes
0.7211	0.0376	0.00274	0.00161	0.00103

Table 1 exhibits some identified parameters concerning matrix density and protein families. Average density gives us the probability that the similarity value s_{ij} with $i \neq j$ but proteins i and j chosen from the same family is a nonzero. Some part of the families are represented by fully dense blocks in the similarity matrix. Larger families of such property are usually rare. The average nonzero s_{ij} value ($i \neq j$) present in the families are also indicated in Table 1. The unit values situated on the diagonal are not counted into these averages. The distribution of the nonzero similarities within families of various sizes is exhibited in Fig. 3. These similarity values cover the whole range between 0 and 1. On the other hand, Fig. 4 shows the distribution of nonzero similarities between proteins of different families of the same superfamily, proteins of different superfamilies situated in the same fold, proteins of different folds situated in the same class, and proteins of different classes, respectively. These parts of the similarity matrix have a decreasing density in the enumerated order. Such similarity values rarely exceed 1/3.

The properties enumerated above are all taken in consideration when new matrices are generated.

3.2 Generating New Large Matrices

When a new matrix is generated, there is a single input parameter to set, namely the number of proteins (N) in the synthetic data set. This number N is supposed

to be greater than 1000. There is no sense to define an upper limit, it will be forced by technical constraints. The main goal is to be able to create matrices describing pairwise similarities of 10^6 proteins using an ordinary PC. One important tool in matrix generation is a good-quality random number generator [6]. The main steps of matrix generation are enumerated below:

1. First thing to create is a series of random numbers n_1, n_2, ... n_F such a way that they follow the identified power function distribution of family sizes, and $\sum_{i=1}^{F} n_i = N$. The new synthetic protein data set will consist of F families, and family with index i will contain exactly n_i proteins.
2. The second thing is to decide the hierarchy of families, which is performed sequentially. Initially we need to create a class, a fold in the class, a superfamily in the fold, and assign the first family to the newly created superfamily. For each further family there is a p_c probability to add a new class; if no new class is created then there is a p_f probability to add a new fold in an existing class; if no new fold is created then there is a p_s probability of add a new superfamily into an existing fold and place the new family there. Otherwise the new family is included into the existing superfamilies, following the popularity rule introduced in network theory [3]. The new family will be assigned to popular classes, folds and superfamilies with higher probability. The current version uses $p_c = 0.99$, $p_f = 0.8$, $p_s = 0.75$, but there is also an upper limit for the number of classes, which logarithmically grows with N.
3. Nonzero similarity values are generated as random numbers in the $(0, 1]$ interval. Diagonal values are all 1 by default. Further nonzeros within the blocks representing families follow the identified distribution functions shown in Fig. 3. Nonzero values in other regions of the matrix follow the corresponding distribution function indicated in Fig. 4. The density of nonzeros in various regions of the matrix corresponds to the values given in Table 2. The generated matrix is perfectly symmetrical.
4. Finally the randomly generated nonzeros are sorted according to row and column and are transferred into the output file. The header of the output file contains information on the number of proteins and the hierarchical structure of the generated data set so that it can serve as ground truth at testing.

4 Results and Discussion

The first-order validation of the proposed method consisted of inspection of properties and attributes of output matrices, and functional testing using the TRIBE-MCL algorithm.

For the sake of property inspection, we have created synthetic test data of sizes varying from 10,000 to 250,000 proteins, 25 instances of each. Table 3 indicates the average and standard deviation of the hierarchy attributes, namely the number of classes, folds, superfamilies, and families, and finally the density if the matrix as well. The table reflects that the randomness within the generation process produces a considerable amount of variance among matrices of the same

Table 3. Properties of the generated synthetic protein graphs, for various sizes of the data set

Proteins	Classes	Folds	Superfam.	Families	Density ($\times 10^{-3}$)
10k	7	280 ± 25	557 ± 45	1389 ± 129	5.06 ± 1.80
15k	7	392 ± 23	770 ± 42	1926 ± 105	4.16 ± 1.17
20k	8	515 ± 45	1024 ± 58	2526 ± 137	3.56 ± 0.59
30k	9	726 ± 28	1441 ± 57	3593 ± 103	3.03 ± 0.37
50k	11	1136 ± 34	2254 ± 68	5677 ± 184	2.44 ± 0.36
70k	12	1550 ± 65	3087 ± 124	7657 ± 295	2.15 ± 0.16
100k	13	2138 ± 53	4283 ± 91	10671 ± 227	1.72 ± 0.13
150k	14	3066 ± 95	6111 ± 195	15270 ± 404	1.63 ± 0.15
200k	15	3992 ± 122	7964 ± 233	19915 ± 624	1.54 ± 0.15
250k	16	4900 ± 88	9762 ± 154	24360 ± 420	1.47 ± 0.17

size. Each generated matrix is different of all others and can be used to test the TRIBE-MCL algorithm.

In order to run a set of simple functional tests, we have created test matrices of sizes between 10,000 and 50,000 varying in small steps, 15 instances of each size. All these matrices were fed to our ultimate high-speed and memory saving version [11] of the TRIBE-MCL algorithm, using inflation rate $r = 2.0$ and similarity threshold $\varepsilon = 10^{-3}$. Simple statistical parameters were extracted from the total runtime values, and are exhibited in Fig. 5. Generated matrices contain a considerable amount of variability, which seemingly reduces as the size of the matrix grows. Considerably wider test suites and detailed results using data generated by the proposed method are exhibited in [11].

The main limitation of the proposed method is the fact that it builds on information extracted from a single protein data set designed to test clustering

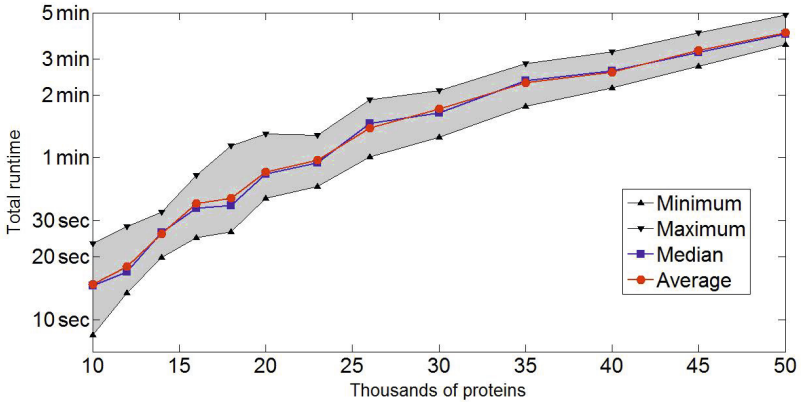

Fig. 5. Total runtime of TRIBE-MCL clustering process plotted against the protein count in the synthetic protein graph: minimum, maximum, median and average values

algorithms. Further efforts will be made to provide the user the opportunity to manually tune the attributes of the output data set and similarity matrix.

Creating synthetic protein data sets and corresponding pairwise similarity matrices of up to 250 thousand items can be performed on an ordinary Pentium4 PC with 2GB RAM in less than one minute. Creating larger data sets of up to 10^6 items is also possible, but it requires more memory and time.

5 Conclusions

In this paper we proposed a novel method to create test data to hierarchical clustering methods based on pairwise similarity measures. The proposed method was applied to generate synthetical protein data sets, their four-level hierarchical structure and sparse similarity matrix that contains BLAST-like pairwise alignment scores. Test matrices were fed to the TRIBE-MCL algorithm, which proved the validity of the synthetical data. The proposed method can efficiently support the validation process of hierarchical clustering algorithms on large-scale data.

References

1. Altschul, S.F., Madden, T.L., Schaffen, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search program. Nucl. Acids Res. 25, 3389–3402 (1997)
2. Andreeva, A., Howorth, D., Chadonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., Murzin, A. G.: Data growth and its impact on the SCOP database: new developments. Nucl. Acids Res. 36, D419–D425 (2008)
3. Barabási, A.L.: Linked: The New Science of Networks. Perseus Book Group, New York (2002)
4. BrainWeb: Simulated Brain Database, `http://brainweb.bic.mni.mcgill.ca/brainweb/`
5. Enright, A.J., van Dongen, S., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res. 30, 1575–1584 (2002)
6. Knuth, D.A.: Random number generator. US Patent 3548174A (1970)
7. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453 (1970)
8. Protein Classification Benchmark Collection, `http://net.icgeb.org/benchmark`
9. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197 (1981)
10. Structural Classification of Proteins database, `http://scop.mrc-lmb.cam.ac.uk/scop`
11. Szilágyi, L., Szilágyi, S.M., Hirsbrunner, B.: A fast and memory-efficient hierarchical graph clustering algorithm. In: Loo, C.K., Yap, K.S., Wong, K.W., Teoh, A., Huang, K. (eds.) ICONIP 2014, Part II. LNCS, vol. 8835, Springer, Heidelberg (2014)
12. Szilágyi, S.M., Szilágyi, L.: A fast hierarchical clustering algorithm for large-scale protein sequence data sets. Comput. Biol. Med. 48, 94–101 (2014)
13. Várady, P., Benyó, Z., Benyó, B.: An open architecture patient monitoring system using standard technologies. IEEE Trans. Inform. Technol. Biomed. 6, 95–98 (2002)