Engineering Innovations
ISSN: 2813-1002, Vol. 4, pp 37-46
doi:10.4028/p-8a27xd
© 2023 The Author(s). Published by Trans Tech Publications Ltd, Switzerland.

Submitted: 2022-09-30
Revised: 2022-10-11
Accepted: 2022-10-13
Online: 2023-02-15

A Novel Epitope Dataset: Performance of the MCL-Based Algorithms to Generate Dataset for Graph Learning Model

Binti Solihah[1,a], Aina Musdholifah[2,b,] Azhari Azhari [2, c]

[1]Department of Informatic, FTI, Universitas Trisakti, Indonesia

[2]Department of Computer Science and Electronics, FMIPA, Universitas Gadjah Mada, Indonesia

[A]binti@trisakti.ac.id, [b]aina_m@ugm.ac.id, [c]arisn@ugm.ac.id

Keywords: Epitope Dataset, PS-MCL, graph clustering, graph learning model.

Abstract. Naturally, the epitope dataset can be presented as a graph. Dataset preparation in the previous methods is part of model development. There are many graph-based classification and regression methods. Still, it is difficult to identify their performance on the conformational epitope prediction model because datasets in a suitable format are unavailable. This research aims to build a dataset in a suitable format to evaluate kernel graph and graph convolution network. This dataset, which results from graph clustering on graph antigens, can be used to identify the performance of many graph neural network-based algorithms for conformational epitope prediction. The Ag-Ab complexes that meet the criteria for forming a conformational epitope prediction dataset from previous studies were downloaded from the Protein Data Bank. Raw datasets in the form of specific exposed antigen chain residues are labeled as epitope or non-epitope based on their proximity to the paratope. The engineering features in the raw dataset are derived from the structure of the antigen-antibody complex and the propensity score. Aggregating atomic-level interactions into residual levels create an initial graph of the antigen chain. The MCL, MLR-MCL, and PS-MCL are graph clustering algorithms to obtain labeled sub-clusters from the initial graph. A balance factor parameter is set to several values to identify the optimal dataset formation based on minimal fragmentation. The output of the MCL algorithm is used as a baseline. As a result of the fragmentation analysis that occurs, the MLR-MCL algorithm gives the best model performance at a balance factor equal to 2. PS-MCL gives the best performance at a value of 0.9. Based on the minimum fragmentation, the MLR-MCL algorithm provides the best model performance compared to MCL and PS-MCL. The dataset in a format according to benchmarking dataset can be used to identify the characteristics of antigen subgraphs formed from the graph clustering process and to explore the performance of graph-based learning conformational epitope prediction models such as graph convolution networks.

Introduction

The conformational epitope is the antigen part bound by the antibody in the recognition phase of the immune system [1] Identification of this part of the antigen required in developing drugs, vaccines, and therapeutic vaccines [2, 3] using a protein engineering approach [4]. The conformational epitope consists of amino acid residues that have spatial proximity in the binding state and consist of segments in the primary sequence. Epitopes in an antigen chain can be single, multiple, or overlapping. The residues which conform epitope may form a network that has not identified what is behind the conformation. However, the statistical analysis showed that specific pairs of amino acids differed significantly from other pairs on the epitope [5, 6].

Graphs can abstract complex systems with interactions between objects, such as molecular graphs, social networks, and biological networks. Graph generation and transformation using machine learning can potentially solve complex problems such as protein folding, statistical modeling, and molecule design [11]. The use of the graph approach in the prediction of conformational epitopes is still minimal. Zhao et al. initiated this approach by implementing MCL to obtain subgraphs of antigens based on edge weights and topology graphs [12]. Another study by the same researcher used the coupling graph to improve the model's performance by utilizing the frequent coupling subgraph pattern [13]. The other model applies graph partitioning to obtain a subgraph of the residual graph surface and continues with seed expansion to obtain a candidate epitope [14]. Wang et al. introduced conformations in graph formation [15]. Zhao applied a graph kernel for classification, but Wang used the Graph Convolution network.

The availability of epitope prediction datasets in graph representation is needed to explore many graphs computational approaches to build epitope prediction models. Graph dataset has elements in the form of nodes, edges, node attributes, and edge attributes. SNAP dataset [16], MoleculNet [17], TUDataset [18], GraphGT [19], Open Graph Benchmark (OGB) [20] is a graph dataset originating from many different domains. SNAP contains 80 different datasets and is a system for generating, manipulating, and analyzing graphs and networks. MoleculeNet is the molecular dataset built on DeepChem for classification and regression. There are 17 datasets from the domains of quantum mechanics, physics chemistry, biophysics, and physiology on MoleculeNet. OGB's molecular graph is an improvement from the MoleculNet version. It has a data loader feature and additional features on Molecular Graph. The addition of features to the nodes and edges of the Molecular Graph in OGB can improve the performance of model [20]. The TUDataset contains 120 datasets and baseline methods with the Python interface. GraphGT contains 36 datasets from 9 domains, including protein, brain network, vision, transportation science, social network, and molecule. GraphGT provides a Python API for accessing data and is available for download via https://graphgt.github.io/.

Graph learning research such as Deep Walk [23] and Graph Convolution Network [24] has received attention from many researchers and is growing fast. Many graph learning approaches are applied in node level classification, link edge, and graph classification. This approach is naturally appropriate for many real-world problems. The graph-based computational approach opens the opportunity to produce a better conformational epitope prediction model than the previous methods. The application of the graph method for the prediction of the conformational epitope is still minimal compared to the methods that have been developed. One of the causes is the availability of datasets in graph representations that follow benchmarking standards in graph learning. This study aims to create a dataset in a graph representation whose nodes and edges are enriched with features to support the prediction of conformational epitopes. Residues with proximity at the atomic level are aggregated to form a graph with nodes enriched with features calculated at the residual level. Statistically, the features at the epitope residues are significantly different from the non-epitope residues. The antigen graph is then clustered using the MCL graph clustering algorithm and its derivatives to obtain the non-epitope and epitope subclusters. Very few sub-clusters were detected containing epitope residues, so for optimization of sub-cluster formation, a threshold value was used in the form of the percentage of epitope residues in each category. The resulting dataset format follows the TUDataset format, whose explanation is in the URL File Format | TUDataset (chrsmrrs.github.io)

The following section will discuss the data preparation and methods for generating epitope predictive datasets in graph representation. Next will be presented the results of the experiment, an analysis of the results, and conclusions.

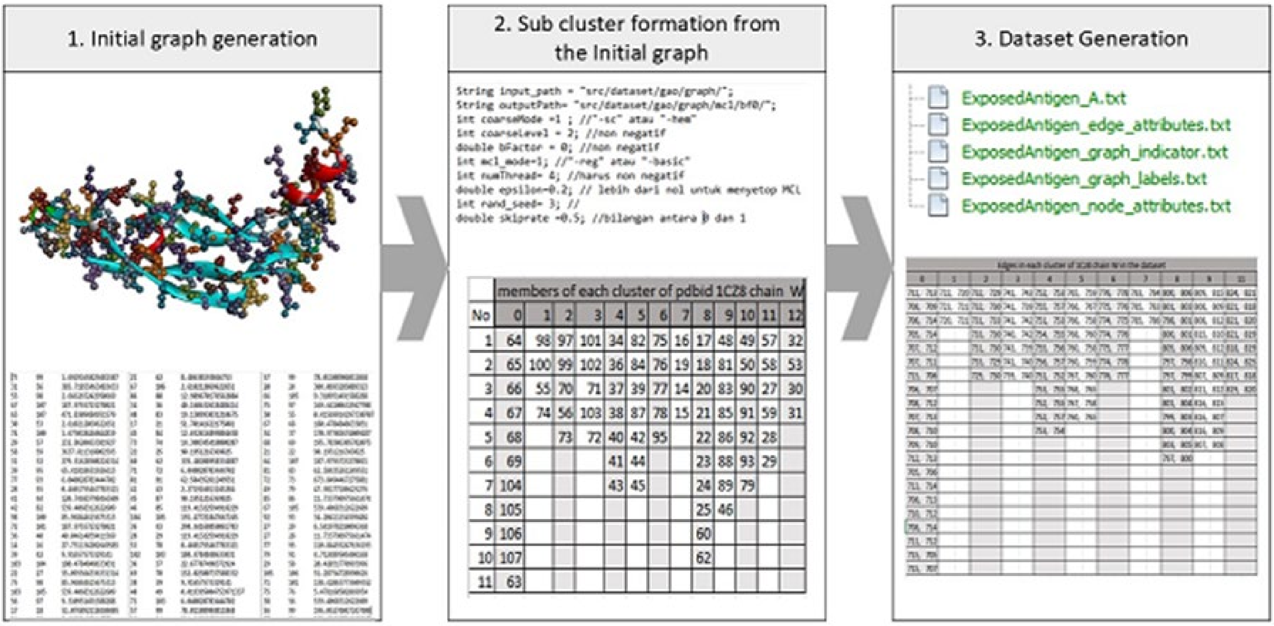
Dataset Collection

The dataset was collected and preprocessed using the procedure described in Solihah et al. [25]. Raw data in the form of 3d chain antigen structures containing information on residues exposed at the atomic level and labeled epitope and non-epitope was taken from Gao et al. [7]. These residues are enriched with features of structural analysis with PSAIA, which include ASA, RSA, and Protrusion Index [26], Contact Number [27], QSE [28], HSE [29], AAIndex [30], B Factor [31] and Log Odd Ratio [32]. The dimension of the vector properties is 601. PCA is used to reduce its dimensions to a vector of length 25.

Experimental Procedure

The experimental procedure is described in the block diagram in Fig 1. There are three steps to construct a dataset: (1) Initial graph generation; (2) Sub-cluster formation from the initial graph; (3) Dataset generation.

Initial graph generation. This stage aims to obtain an initial undirected graph representation of residues exposed to an antigen chain. An edge is formed if the distance between the residue and other residues is within the allowable distance. Edge weights were calculated based on the log odds ratio of the frequency pairs on the epitope compared to the frequency of the pairs on the non-epitope.
Sub-cluster formation. The formation of the dataset begins with applying MCL-based clustering to each input in the form of a graph from the data preparation stage, as describe in Fig 2 . The MCL algorithm used is MCL (21), MLR-MCL [22], and PS-MCL [11]. At this stage, an MCL-based algorithm is identified that produces the best cluster. Analysis of the number of clusters formed and the purity of the clusters became the reference parameters in determining the quality of the clusters. The balancing factor parameter (r) and coarse mode (shotgun coarsening and heavy edge matching) are the parameters that are optimized to get the best cluster.
Dataset generation. Save the labeled cluster data into a file format following the file format on the TUDataset. Storage procedure is defined to create the graph dataset with a file format such as TUDataset.


1. Initial graph generation
2. Sub cluster formation from the Initial graph
3. Dataset Generation

```
String input_path = "src/dataset/gao/graph/";
String outputPath= "src/dataset/gao/graph/mcl/bf0/";
int coarseMode =1 ; //"-sc" atau "-hem"
int coarseLevel = 2; //non negatif
double bFactor = 0; //non negatif
int mcl_mode=1; //"-reg" atau "-basic"
int numThread= 4; //harus non negatif
double epsilon=0.2; // lebih dari nol untuk menyetop MCL
int rand_seed= 3; //
double skiprate =0.5; //bilangan antara 0 dan 1
```

	members of each cluster of pdbid 1C28 chain W												
No	0	1	2	3	4	5	6	7	8	9	10	11	12
1	64	98	97	101	34	82	75	16	17	48	49	57	32
2	65	100	99	102	36	84	76	19	18	81	50	58	53
3	66	55	70	71	37	39	77	14	20	83	90	27	30
4	67	74	56	103	38	87	78	15	21	85	91	59	31
5	68		73	72	40	42	95		22	86	92	28	
6	69				41	44			23	88	93	29	
7	104				43	45			24	89	79		
8	105								25	46			
9	106								60				
10	107								62				
11	63												

ExposedAntigen_A.txt
ExposedAntigen_edge_attributes.txt
ExposedAntigen_graph_indicator.txt
ExposedAntigen_graph_labels.txt
ExposedAntigen_node_attributes.txt

Fig. 1. The step by step of experimental procedures

MCL-based Sub Graph Clustering

The MCL algorithm works on data in graph representation. A graph is expressed as a set of nodes and edges, where edges connect the nodes in a graph. MCL is a flow-based graph clustering method that applies the random walk concept. Visits within the caster have a higher probability than visits to other clusters. The existence of edges in flow-based graph clustering states the probability of a flow occurring from node to node. If a node is connected to 3 other nodes, then the transition probability of that node to another node is 1/3 (0.33%). In a weighted graph, edges have weights so that the transition probability matrix Mij is a column-normalized matrix of the adjacency matrix plus self-loop (Eq. 1).

$$M_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{kj}} \tag{1}$$

where n is the number of nodes in the graph. The MCL algorithm is built by three basic operations: expand, inflate, and prune. Expansion is done by raising the transition probability matrix. The expansion operation allows flows to be connected to different clusters in the graph. The inflate operation powers the transition probability matrix M by a real number r followed by column normalization. Inflation is used to strengthen and weakening flow (Eq. 2).

$$(\Gamma_r M)_{pq} = \left(M_{pq}\right)^r / \sum_{t=1}^{k} (M_{iq})^r \tag{2}$$

MCL is one of the superior graph clustering algorithms in the bio-networking field because the clustering results are in the form of small sub-graphs (1 to 3 nodes). The MCL algorithm has attracted researchers' interest in developing its derivatives, such as Regularized MCL, MLR-MCL, and PS-MCL. If viewed from the number of constituent residues, the conformational epitopes of B cells are categorized as clusters with sizes between 5 and 30 constituent residues. In terms of the number of constituent residues, the epitope is suitable if identified by utilizing the derivative of the MCL algorithm compared to the original MCL, which was identified as causing fragmentation.

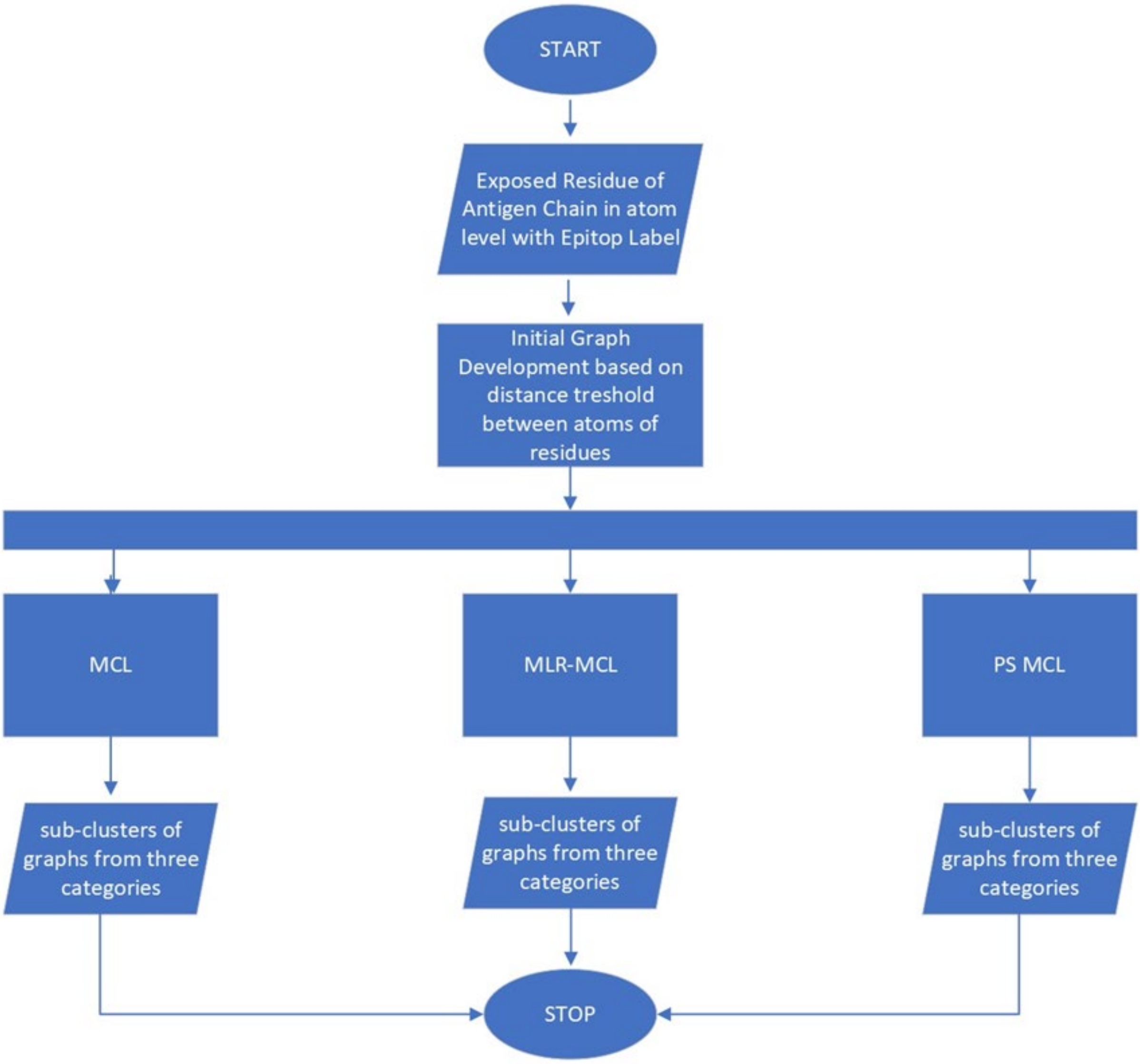

START
Exposed Residue of Antigen Chain in atom level with Epitop Label
Initial Graph Development based on distance treshold between atoms of residues
MCL
MLR-MCL
PS MCL
sub-clusters of graphs from three categories
sub-clusters of graphs from three categories
sub-clusters of graphs from three categories
STOP

Fig. 2. MCL-based sub graph clustering

The dataset formation stage is presented in Fig. 3. Input is in the form of 3d structure data from the exposed residues of the antigen chain, each with the "epitope" or "non-epitope" label. The output is a collection of sub-graphs with the "epitope," "mixed," or "non-epitope" label, which is stored in separate files. Details of the dataset formation process are as follows:

1. Build a graph from the input data following the graph formation procedure in Gao et al.[7]
2. Calculate the edge weights following the procedure in Gao [7]
3. Calculate the ASA, RSA, PSAIA, and AAIndex features of the residues in the graph and store them as attributes of each node of the graph
4. Perform graph clustering using PS-MCL algorithm with parameters that optimize cluster results to obtain three subgraph categories. The first category is non-epitope clusters

containing < 30% epitope residues. The second categories consist of mixed clusters containing epitope residues >= 30% and < 70%. The third category is epitope clusters containing >= 70% epitope.

5. Write down all the subgraphs resulting from the process of step No. 4 into the four files.

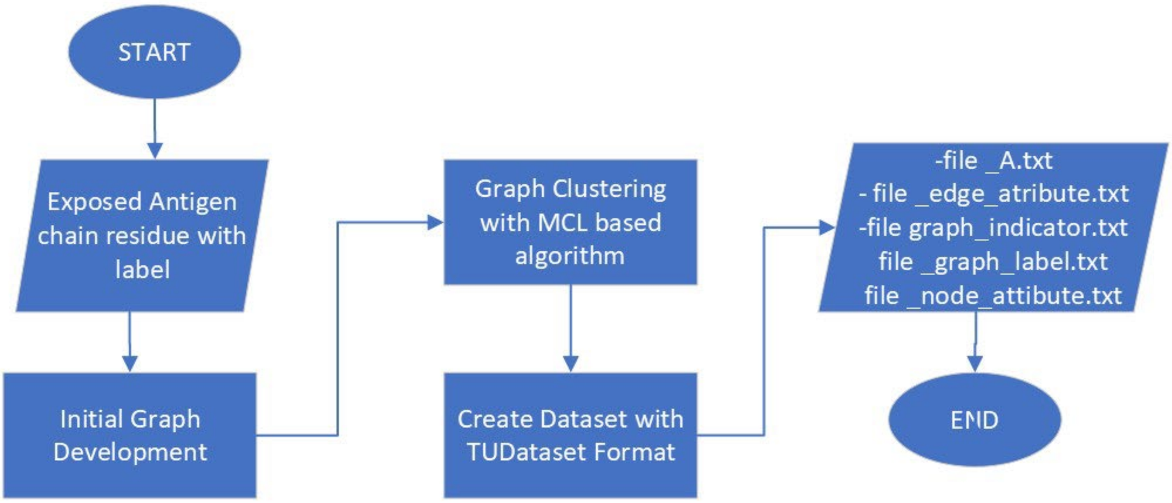

START
Exposed Antigen chain residue with label
Initial Graph Development
Graph Clustering with MCL based algorithm
Create Dataset with TUDataset Format
-file _A.txt
- file _edge_atribute.txt
-file graph_indicator.txt
file _graph_label.txt
file _node_attibute.txt
END

Fig. 3. Epitope dataset construction

Statistical Analytic Procedures

Among the critical issues in the application of the MCL algorithm is the formation of small clusters or fragmentation [11]. At the initial stage, min and max statistical analyses are conducted to answer the following research questions related to the problem of fragmentation:

Q1. How does the selection of the balance factor affect the distribution of clusters and nodes of the formed cluster?

Q2. does the balance factor on the MLR-MCL and PS-MCL algorithms have the same effect on the distribution of clusters and nodes in the cluster?

Q3. Which factor balance gives clustering results with the minimum number of small clusters and the minimum number of nodes?

Experiment Result

We applied the MCL, MLR-MCL, and PS-MCL algorithms to the initial graph antigen data at the residual level and identified the distribution of the number of clusters and the number of nodes formed in each cluster. The node distribution in each cluster generated by the MCL algorithm is used as a baseline to measure the fragmentation in the cluster results of the MLR-MCL and PS-MCL algorithms. Table 1 presents the results of graph clustering with the MCL algorithm. The number of small-sized clusters formed is 120 clusters composed of 325 nodes. The number of clusters with more than three nodes is 3641 clusters and is composed of 32667 nodes. The percentage of the small clusters, as much as 3.19 percent, is express fragmentation. The percentage of nodes that form fragmentation compared to total nodes is 0.99 percent.

Table 1. The distribution clusters and nodes by MCL Algorithm

number of clusters <= 3	number of clusters > 3	number of nodes in the cluster <=3	number of nodes in the cluster > 3
120	3641	325	32667

In the next experiment, the effect of the balance factor on fragmentation was identified. As presented in Table 2 and Table 3, several balance factor values were tested for the MLR-MCL and PS-MCL algorithms to determine the best balance factor that produces minimal fragmentation. The balance factor values of 1, 1.5, and 2 were selected following the experiment of Lim et al.[]. Another balance factor value was tried to identify the optimal fragmentation probability at a balance factor of less than 1.

Based on the experimental results presented in Table 2, in the PS-MCL algorithm, the number of small clusters with a balance factor of less than one is less than the number of small clusters with a balance factor of more than 1. From all experiments conducted, the minimum number of clusters is Small size is obtained at a balance factor of 0.9, where the number of small clusters formed is 192 clusters with the number of nodes that compose as many as 521. The percentage of small casters compared to all clusters formed is 4.13 percent, while the percentage of constituent nodes compared to the total number of nodes is 1.58 percent. Compared with the percentage of fragmentation in the MCL algorithm, the percentage of fragmentation in PS-MCL is greater than that in MCL. At balance factor = 1, fragmentation occurs with a percentage of 16.5 percent and is the largest percentage of fragmentation in PS-MCL.

Table 2. The distribution of number of clusters and nodes in various balance factors of PS_MCL

balance factor	**number of clusters <= 3**	**number of clusters > 3**	**number of nodes in the cluster <=3**	**number of nodes in the cluster > 3**
0	246	**3076***	666	32326
0.1	296	3392	781	32211
0.2	590	3598	844	32148
0.3	392	3783	912	32080
0.4	405	3945	944	32048
0.5	381	4063	952	32040
0.6	331	4191	840	32152
0.7	271	4305	716	32276
0.8	238	4389	644	32348
0.9	**192***	**4456****	**521***	**32471****
1	**730****	3682	**5344****	**27648***
1.5	690	3722	1596	31396
2	690	3722	1596	31396

note: *: minimum; **: maximum

The experimental results of several balance factor values in the MR-MCL are presented in Table 3. The number of small clusters produced in various balance factors of the MLR-MCL algorithm is generally smaller than in the PS-MCL algorithm. If in the PS-MCL, the optimal fragmentation is at the balance factor = 0.9, then in the MLR-MCL, the optimal fragmentation is obtained at the balance factor = 2. The fragmentation that occurs in the balance factor between 0 to 1 is negatively correlated with the balance factor. In MLR-MCL, the distribution of clusters and their constituent nodes at balance factor = 0 is the same as at balance factor = 1. The minimum number of small-sized clusters formed occurs at balance factor = 2, 16 clusters (0.38 percent) with the number of nodes that make up 39 (0.11 percent).

Table 3. The distribution of number of clusters and nodes in various balance factors of MLR-MCL

balance factor	**number of clusters <= 3**	**number of clusters > 3**	**number of nodes in the cluster <=3**	**number of nodes in the cluster > 3**
0	49	3982	136	32856
0.1	173	**3301***	472*	**32520***
0.2	175	3469	457	32535
0.3	**176***	3594	429	32563
0.4	147	3702	361	32631
0.5	130	3765	333	32659
0.6	109	3821	288	32704
0.7	88	3878	239	32756
0.8	74	3916	203	32789
0.9	60	3950	163	32829
1	49	3982	136	32856
1.5	26	4124	69	32923
2	**16***	**4177****	**39***	**32956****

*:min , ** : max

Discussion

The model's best performance in forming graph datasets for predicting the conformational epitope is inversely proportional to the fragmentation that occurs. Fragmentation is represented by the number of small clusters (1 to 3 nodes) and the nodes involved in their formation. The main factor that influences the occurrence of fragmentation is the balance factor. The best model performance from the MLR-MCL and PS-MCL models is presented in Table 4. The best model performance is obtained from the MLR-MCL model with a balance factor of 2, while the PS-MCL model occurs at a balance factor of 0.9. In this case, MLR-MCL is better than PS-MCL. The number of small clusters in PS-MCL is much more than in MLR-MCL by 12 times, with the number of nodes contributing more than 13 times. This condition can be observed on the gray and blue bars in Fig 4.

Table 4. The best performance of PS-MCL and MLR-MCL

Algorithm	**balance factor**	**number of clusters <= 3**	**number of clusters > 3**	**number of nodes in the cluster <=3**	**number of nodes in the cluster > 3**
PS-MCL	0.9	**192**	**4456**	**521**	**32471**
MLR-MCL	2	**16**	**4177**	**39**	**32956**
MCL		**120**	3641	325	32667

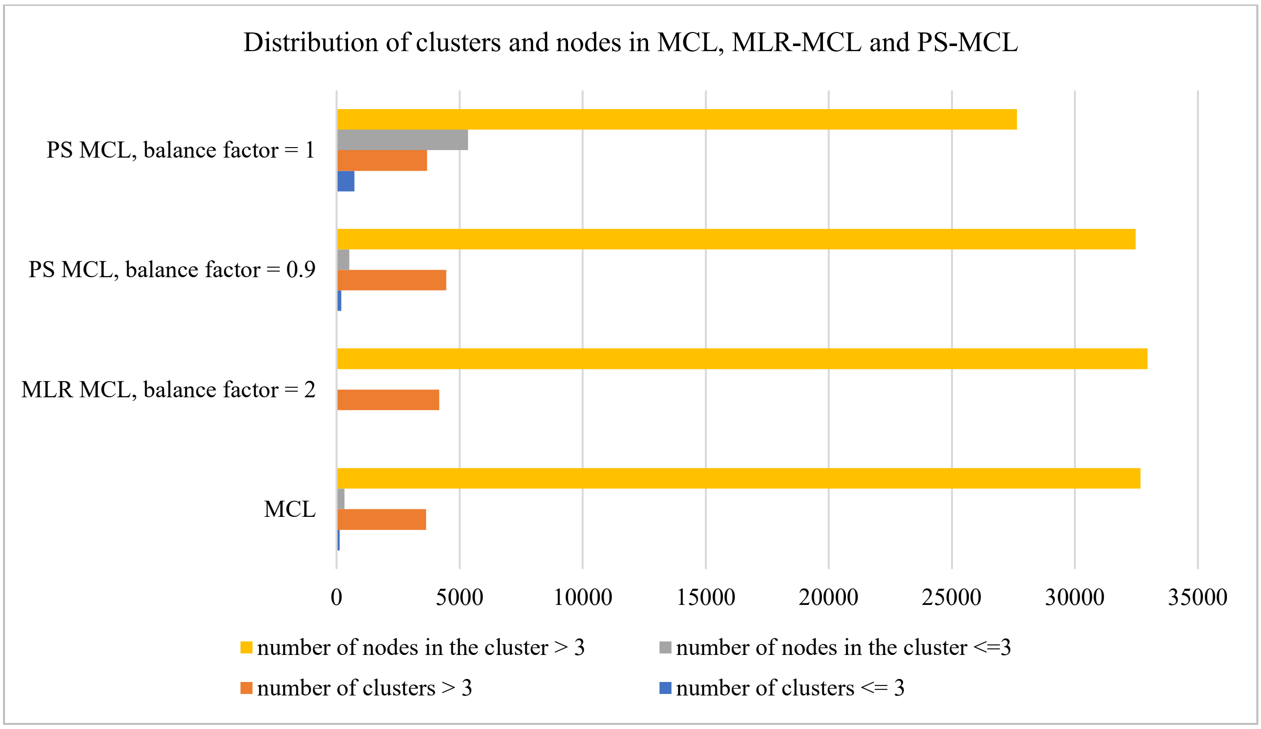


Fig. 4. Performance comparison on fragmentation

MCL implementation. PS-MCL and MLR-MCL in other datasets were performed by Lim et al.[11]. According to Lim et al., the number of nodes that build small clusters in PS-MCL is less than in MLR-MCL. The experimental results on the antigen dataset showed different results, where the number of nodes that formed fragmentation in MLR-MCL was less than in PS-MCL.

This study still does not yet identify the distribution of epitope and non-epitope residues in the formed sub-clusters. Previous research shows that if the sub-cluster contains epitope nodes greater than non-epitope, then it is categorized as an epitope sub-cluster. This proportion has not been proven to capture critical features of an epitope or non-epitope cluster. Experiments are needed to show the classification model's performance from the dataset formed based on various percentages of the epitope in the sub-cluster labeled with the epitope. In addition, the distribution of epitope residues in small clusters is not yet known whether they dominate or not.

The experimental results of the fragmentation measurement of the MCL, PS-MCL, and MLR-MCL algorithms on the antigen dataset show that the MLR-MCL approach is better than the MCL and PS-MCL approach. Experimental results with PS-MCL showed greater fragmentation than MCL. When viewed from the number of clusters measuring more than three nodes, both PS-MCL and MLR-MCL resulted in a higher number of clusters. It can be roughly stated that the cluster size in MCL is larger than in PS-MCL and MLR-MCL.

Conclusion

There are many graph-based classification and regression methods. Still, it is difficult to identify their performance on the conformational epitope prediction model because datasets in a suitable format are unavailable. This research aims to build a dataset in a suitable format to evaluate kernel graph and graph convolution network. The MCL, MLR-MCL, and PS-MCL are graph clustering algorithms to obtain labeled sub-clusters from the initial graph. A balance factor parameter is set to several values to identify the optimal dataset formation based on minimal fragmentation. The output of the MCL algorithm is used as a baseline. As a result of the fragmentation analysis that occurs, the MLR-MCL algorithm gives the best model performance at a balance factor equal to 2. At the balance factor 0.1, MLR-MCL generates 176 small clusters which use 472 nodes to form a small cluster. PS-MCL gives the best performance at a value of 0.9. At balance factor 1, PS-MCL produces 730 small clusters with5344 nodes. Based on the minimum fragmentation, the MLR-MCL algorithm provides the best model performance compared to MCL and PS-MCL. Each subgraph must be labeled epitope and non-epitope to be used for classification. It is necessary to identify a threshold of the percentage of residue epitope in the subgraph to label the subgraph as epitope or non-epitope. Previous researchers used a limit on the number of epitopes more or equal to 3%.

References

[1] A. K. Abbas, A. H. Lichtman, S. Pillai, and D. L. Baker, “Antibodies and Antigens,” Cell. Mol. Immunol., vol. 1, pp. 75–96, 2010.

[2] A. Wadood et al., “Epitopes based drug design for dengue virus envelope protein: A computational approach,” Comput. Biol. Chem., vol. 71, pp. 152–160, 2017.

[3] S.H. Guo, C. Wang, H.Y. Yang, N.N. Zhang, X. Zhuang, D.Z. Cui,. “Prediction of Antigenic Epitopes for Coat Protein of Potato virus; AMR p.183–185:1204–8, 2011 https://doi.org/10.4028/www.scientific.net/amr.183-185.1204.

[4] S. Shalkharov, Z. Shalkharova, K. Rysbekov, Shalkharova, Y. Paromova, and Y. Petrova, “Biomedical Engineering as a Modern Component of Science in Biology and Medicine”, Journal of Biomimetics, Biomaterials and Biomedical Engineering, Vol. 53, pp. 67–75, 2021 Trans Tech Publications, Ltd. https://doi.org/10.4028/www.scientific.net/jbbbe.53.67

[5] N. D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni, and T. Pupko, “Computational characterization of B-cell epitopes,” Mol. Immunol., vol. 45, pp. 3477–3489, 2008.

[6] J. V. Kringelum, M. Nielsen, S. Padkjaer, and O. Lund, “Structural analysis of B-cell epitopes in antibody: protein complexes,” Mol. Immunol., vol. 53, no. 1–2, pp. 24–34, 2013.

[7] C. Gao, Y. Wang, J. Luo, Z. Zhou, Z. Dong, and L. Zhao, “Flexibility-aware graph-based algorithm improves antigen epitopes identification,” bioRxiv, p. 2021.05.17.444445, 2021.

[8] P. Haste Andersen, M. Nielsen, and O. Lund, “Prediction of residues in discontinuous B-cell epitopes using protein 3D structures,” Protein Sci., vol. 15, no. 11, pp. 2558–2567, 2006.

[9] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, “epitope prediction using conformational epitopes,” Nucleic Acids Res., vol. 45, no. May, pp. 24–29, 2017.

[10] G. A. Dalkas and M. Rooman, “SEPIa , a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence,” BMC Bioinformatics, vol. 18, no. 95, pp. 1–12, 2017.

[11] Y. Lim, I. Yu, D. Seo, U. Kang, and L. Sael, “PS-MCL: Parallel shotgun coarsened Markov clustering of protein interaction networks,” BMC Bioinformatics, vol. 20, no. Suppl 13, pp. 1–12, 2019.

[12] L. Zhao, L. Wong, L. Lu, S. C. H. Hoi, and J. Li, “B-cell epitope prediction through a graph model,” BMC Bioinformatics, vol. 13, no. Suppl 17, pp. 1–12, 2012.

[13] L. Zhao, S. C. H. Hoi, Z. Li, L. Wong, H. Nguyen, and J. Li, “Coupling graphs, efficient algorithmsand B-cell epitope prediction,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 11, no. 1, pp. 7–16, 2014.

[14] L. Zhao, S. Wu, J. Jiang, W. Li, J. Luo, and J. Li, “Novel overlapping subgraph clustering for the detection of antigen epitopes,” Bioinformatics, vol. 34, no. 12, pp. 2061–2068, 2018.

[15] Y. Wang et al., “Flexibility-aware graph model for accurate epitope identification,” Comput. Biol. Med., vol. 149, no. August, p. 106064, 2022.

[16] J. Leskovec and R. Sosič, “SNAP: A general-purpose network analysis and graph-mining library,” ACM Trans. Intell. Syst. Technol., vol. 8, no. 1, 2016.

[17] Z. Wu et al., “MoleculeNet: A benchmark for molecular machine learning,” Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018.

[18] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann, “TUDataset: A collection of benchmark datasets for learning with graphs,” 2020.

[19] Y. Du, X. Guo, H. Cao, S. Hu, and J. Jiang, “GraphGT : Machine Learning Datasets for Graph Generation and Transformation,” no. NeurIPS, pp. 1–29, 2021.

[20] W. Hu et al., “Open graph benchmark: Datasets for machine learning on graphs,” Adv. Neural Inf. Process. Syst., vol. 2020-December, no. NeurIPS, pp. 1–34, 2020.

[21] S. van Dongen, “Graph stimulation by flow clustering,” Graph Stimul. by flow Clust., vol. PhD thesis, p. University of Utrecht, 2000.

[22] V. Satuluri and S. Parthasarathy, “Scalable graph clustering using stochastic flows: Applications to community discovery,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 737–745, 2009.

[23] B. Perozzi, R. Al-Rfou, and S. Skiena, “DeepWalk: Online learning of social representations,” Proc. ACM SIGKD

[24] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc., pp. 1–14, 2017. D Int. Conf. Knowl. Discov. Data Min., pp. 701–710, 2014.

[25] B. Solihah, A. Azhari, and A. Musdholifah, “The Empirical Comparison of Machine Learning Algorithm for the Class Imbalanced Problem in Conformational Epitope Prediction,” JUITA J. Inform., vol. 9, no. 1, p. 131, 2021.

[26] J. Mihel, M. Šiki, S. Tomi, B. Jeren, and K. Vlahovi, “PSAIA – Protein Structure and Interaction Analyzer,” BMC Struct. Biol., vol. 11, pp. 1–11, 2008.

[27] K. Nishikawa and T. Ooi, “PREDICTION OF THE SURFACE-INTERIOR DIAGRAM OF GLOBULAR PROTEINS BY AN EMPIRICAL METHOD.pdf,” Int J Pept. Protein Res, vol. 16, pp. 19–32, 1980.

[28] P. Li, G. Pok, K. S. J. Ã, H. S. Shon, and K. H. Ryu, “R ESEARCH A RTICLE QSE: A new 3-D solvent exposure measure for the analysis of protein structure,” Proteomics, vol. 11, pp. 3793–3801, 2011.

[29] T. Hamelryck, “An Amino Acid Has Two Sides: A New 2D Measure Provides a Different View of Solvent Exposure,” ProteinsStructure, Funct. Bioinforma., vol. 59, no. September 2004, pp. 38–48, 2005.

[30] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, “AAindex : amino acid index database, progress report 2008,” Nucleid Acids Res., vol. 36, no. November 2007, pp. 202–205, 2008.

[31] J. Ren, Q. Liu, J. Ellis, and J. Li, “Tertiary structure-based prediction of conformational B-cell epitopes through B factors,” Bioinformatics, vol. 30, pp. 264–273, 2014.

[32] H. R. Ansari and G. P. S. Raghava, “Identification of conformational B-cell Epitopes in an antigen from its primary sequence,” Immunome Res., vol. 6, no. 1, p. 6, 2010.