International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 15 (2023) pp. 460-468

Received: 14 March 2023; Accepted: 5 June, 2023; Published: 23 June, 2023

Covariance Search Model for Identifying ncRNA using Particle Swarm Optimized Agglomerative Clustering

Lustiana Pratiwi[1], Yun-Huoy Choo[1]✉, Azah Kamilah Muda[1], and Satrya Fajri Pratama[2]

[1]Fakulti Teknologi dan Maklumat Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
lustiana@gmail.com, {huoy, azah}@utem.edu.my

[2]Department of Computing, College of Business, Technology and Engineering, Sheffield Hallam University, United Kingdom
s.pratama@shu.ac.uk

***Abstract*: Recent studies have shown that the functional discovery of noncoding ribonucleic acids (ncRNA) is gradually gaining interest among bioinformatics experts. Families of ncRNAs are responsible for various biological functions, including gene expression regulation and catalytic activities, which have yet to be discovered. These discoveries have expanded the scope of the ncRNA study, including finding functional subgroups. Hence, cross-fertilisation solutions derived from computational intelligence principles and algorithms have begun to produce promising outcomes. For the Covariance Model (CM) in ncRNA identification, data clustering is one of the most common strategies in various fields. Based on sequence similarity, hierarchical clustering is the most common method for classifying a set of human ncRNAs into distinct families. However, standard techniques have several drawbacks, such as the sequence structures of each family getting considerably diluted as the number of sequence characteristics in the known family dataset grows. This study optimises the hierarchical clustering approach for identifying ncRNA families using Particle Swarm Optimization (PSO).**

***Keywords*:** Covariance Model, ncRNA Identification Agglomerative Clustering, PSO.

I. Introduction

Several areas of bio-computational technology [1]-[5] have focused on differentiating numerous types of noncoding ribonucleic acids (ncRNA) based on the execution of their varied functions. For instance, when the relationship between RNA structure and function is vital, knowing the usual structure of homologous RNAs is advantageous to identify functional signatures. Moreover, it is desirable to search a genome for ncRNAs. Identification strategies for protein-coding genes frequently fail when applied to ncRNAs. Hence, identifying ncRNA remains an open topic in bioinformatics. However, note that the two bases do not need to covary, as point mutations such as G–C to G–U are evidence of base pairing [6].

Hence, approaches that only look for covariation need to include important information. The covariance model (CM) 's primary advantage is incorporating gene family-specific information to improve accuracy [7]. In ncRNA identification, CM has proven to be highly effective at locating possible members of existing families and has provided excellent precision in the database of genome sequences [8]-[10]. Existing sequence alignments or unaligned example sequences are utilised to automatically generate the annotation of various secondary structure alignments within a hairpin loop based on an ordered tree [11]-[14].

However, it has a substantial disadvantage, namely high computational complexity, which limits its practical application [6]-[12]. Identifying ncRNA has been hampered by uncertainty in determining which sequences comprise a family and a need for adequate numbers of known sequences to estimate model parameters accurately, in addition to challenges associated with a family-specific search that necessitates extensive processing. Hierarchical clustering is the most common utilised mathematical approach which arranges genes into tiny clusters and clusters into higher-level systems to increase CM performance [13]. Using hierarchical clustering, the dataset is partitioned into a series of subsets. It divides the data into a nested tree structure, where the levels of the tree indicate similarity or dissimilarity among the clusters at different levels and where it has been demonstrated to reduce the search time required to identify members of all original ncRNA families using Dot Bracket Notations [5], [7].

Hierarchical clustering is an effective and valuable technique for analysing genomic data and may cluster known ncRNA gene families [13]-[17]. Numerous prior research [15] have utilised hierarchical clustering to aid in the identification process during the merging and clustering of family units. It is desirable to reduce the high computational burden imposed by covariance model (CM)-based non-coding RNA (ncRNA) gene discovery when searching sequencing data using a large number of ncRNA families [18], [20]-[23]. The search for a gigabyte database of sequences for all known ncRNA gene families could take weeks or even years, which is impractical for CM.

Hierarchical clustering has effectively decreased the time required to locate members of all original ncRNA families via

Dot Bracket Notations [5], [19]. A tree-based approach based on the Base Pair Conflicts algorithm is used to pick the original combined CM from stem-loop structural elements (the existence of a base pair in the other secondary structure). On the other hand, the determination process of combined CMs significantly depends on the quantity of original CM since more structural information leads to adding more original CMs to the combined CM. When the number of original CMs increases, the sequence characteristics of each family will be significantly diluted, resulting in premature clustering [5], [24].

The paper is structured as follows: the following section outlines the proposed methodology, Section III gives the results and their discussion, and the final section concludes the study.

II. Research Methodology

In this part, the two primary steps involved in this investigation are described in detail. The mechanism for constructing the hierarchical clustering algorithm for the CM will be highlighted in Section A. In contrast, Section B will detail the activities to hybridise the proposed approaches with Particle Swarm Optimization to develop them (PSO). These sections aim to identify ncRNA using the Covariance Search Model.

A. Hierarchical Agglomerative Clustering for Covariance Model (HACCM)

The first challenge is addressed by applying a clustering technique to known ncRNA gene family clusters. Clustering is a commonly used data analysis approach in numerous disciplines, such as data mining, machine learning, image analysis, and bioinformatics. It is a practical approach to the initial challenge [15], [19].

This is because hierarchical clustering can yield a dendrogram that aids in the organisation of the combined CMs [5]. Each non-leaf node in the dendrogram represents the CM of its child nodes, while each leaf node represents the CM of an ncRNA gene family. Figure 1 displays a hypothetical dendrogram structure example. Most hierarchical clustering algorithms require a measure of dissimilarity between clusters based on tree structure to determine whether clusters should be combined (for agglomerative) or where a cluster should be divided (for divisive) [16].

The distance function often determines this. The most prevalent distance functions include the Euclidean distance, the Manhattan distance, and the Hamming distance, among others. Nevertheless, only a few of these strategies apply to the scenario. This cluster examines ncRNA gene families, each represented by its secondary structure in Dot-Bracket Notation. Dot-Bracket Notation is commonly employed to describe the secondary structure of RNA [12].

It indicates paired bases with matching brackets and unpaired bases with dots. This study defines a unique distance function for dealing with Dot-Bracket Notation secondary structure data. Before describing the hierarchical clustering algorithm for ncRNA genes, this study must clarify two definitions [5].

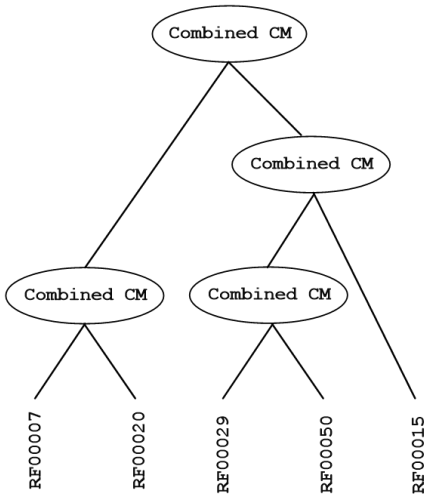

Figure 1. The identification of five ncRNA families from the *Rfam* database using agglomerative clustering [5]

- Definition 1 (Base Pair Conflict): Given two RNA secondary structures, a base pair (m, n) from one secondary structure is called Base Pair Conflict if there exists a base pair (i, j) in the other secondary structure such that $m < i < n < j$, or $i < m < j < n$.
- Definition 2 (Structure Distance): Given two RNA secondary structures, the Structure Distance between them is the average number of Base Pair Conflicts in each secondary structure.

The definitions of Base Pair Conflict and pseudoknots are comparable. Nonetheless, Base Pair Conflict refers to base pairs in two distinct structures, whereas pseudoknots refer to those in a single structure [5], [15]. Due to the inability of CM to consider pseudoknots, which should be avoided in the combined structure, only one of the two conflicting base pairs can be retained in the combined structure.

In contrast to most distance functions used in clustering algorithms, which measure dissimilarity between observations, the distance function utilised in this study, Structure Distance, calculates the compatibility of two RNA secondary structures. The lower this value, the greater the compatibility between the two secondary structures. Compatibility between two secondary structures indicates how much structural information can be retained when combined.

Since the goal is to create a combined CM capable of capturing as much information as possible from both original CMs, two CMs with more compatible secondary structure components would be ideal for combining. There are generally two types of hierarchical clustering strategies: agglomerative, also known as a bottom-up approach, and divisive, also known as a top-down approach.

This study clusters ncRNA gene families using an agglomerative approach. The fundamental hierarchical clustering procedure is as follows [25], [26]:

1. Assign each ncRNA gene family to its cluster, then construct the distance matrix by calculating the structural distance between each family.

2. Identify the closest pair of clusters (the smallest element in the distance matrix) and combine their secondary structures to create a new family corresponding to a non-leaf node in the dendrogram. The cluster pair should then be removed from the dendrogram.
3. Determine the structure distance between the new cluster and each of the previous clusters.
4. Repeat steps 2 and 3 until only a single cluster corresponds to the root of the dendrogram.

Determining how to select base pairs from the two original secondary structures and insert them into the new structure is the most important component of combining secondary structures. This study cannot simply choose and link all base pairs between the two original structures. This may result in the combined structure capturing all structural properties of both original gene families without information loss, but it will significantly increase the complexity of the CM and make little difference when compared to searching with the two original CMs separately [15]-[19].

This study suggests three selection criteria for base pairs from two secondary structures. This study proposes three criteria for selecting base pairs from two secondary structures. First, the selection of base pairs should be as comprehensive as feasible. Since more base pairs are selected, more secondary structure components are kept, and it is more likely that a target sequence will be located while scanning the genome database. Second, the base pairs selected from one CM must be compatible with those selected from the other CM.

This implies no pseudoknots in the combined secondary structure, as CM cannot handle pseudoknots. Thirdly, roughly the same number of base pairs should be chosen for each CM, which means this study seeks to achieve a balance between the two original secondary structures. A greedy algorithm selects base pairs from two secondary structures to form a new secondary structure that satisfies the abovementioned criteria: Hierarchical Agglomerative Clustered Covariance Model (HACCM).

Therefore, the fundamental concept of this study is to choose a base pair from one structure that has the fewest conflicts (pseudoknots) with base pairs from the other structure. This base pair selection will result in the fewest base pair deletions in the other structure.

B. PSO and HACCM Hybridization

Particle Swarm Optimization (PSO) is chosen as one of the warm intelligence techniques to optimise HACCM in this study. One of the most important aspects to consider when selecting PSO is its straightforward yet effective implementation in biological modelling and other related problem [16], [27], [28]. The central concept of PSO-HACCM is that the fitness function of PSO is modified by implementing confusion matrix-based performance measurement techniques that are calculated using bit-scores.

This will enable the optimal interaction between PSO and HACCM, expanding the search space [25], [26], [29]. Additionally, multiple instances of HACCM are executed concurrently, each in a PSO particle. This study assigns particle position based on the sum of bit-score values obtained from the *cmsearch* program in the Infernal package. In the meantime, the fitness value is determined in advance, as there are multiple possible confusion matrix-based performance measurement techniques.

Using the confusion matrix performance measurement of bit-scores, a comparative study will be conducted to determine the most appropriate method for calculating the fitness value for the proposed method. Each particle will examine a distinct set of CM family clusters and produce unique results as the examined CM family cluster set. Its results are recorded to prevent multiple examinations of the same set by various particles.

In addition to modifying the fitness function, the strategy for updating particle velocity is modified, as shown in Figure 2. Similar to the original PSO [25], the velocity of the PSO v_i in the $(t+1)$th iteration is affected by the inertia ratio I_i, cognitive acceleration ratio C_i, and social acceleration ratio S_i, with a slight modification such that

$$v_i(t+1) = I_r + C_r + S_r, \quad (1)$$

$$I_i = I \times v_i(t), \quad (2)$$

$$C_i = C \times rand() \times (p_i - x_i(k)), \quad (3)$$

$$S_i = S \times rand() \times (p_{best} - x_i(k)), \quad (4)$$

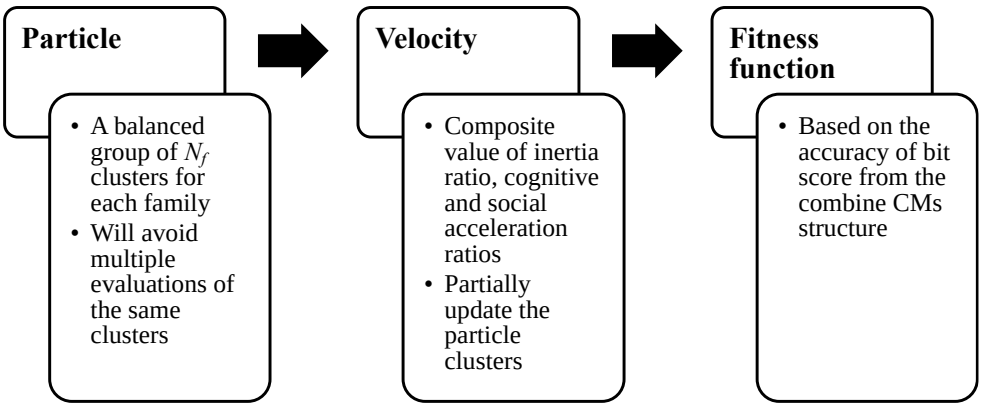

Figure 2. Optimisation process using PSO

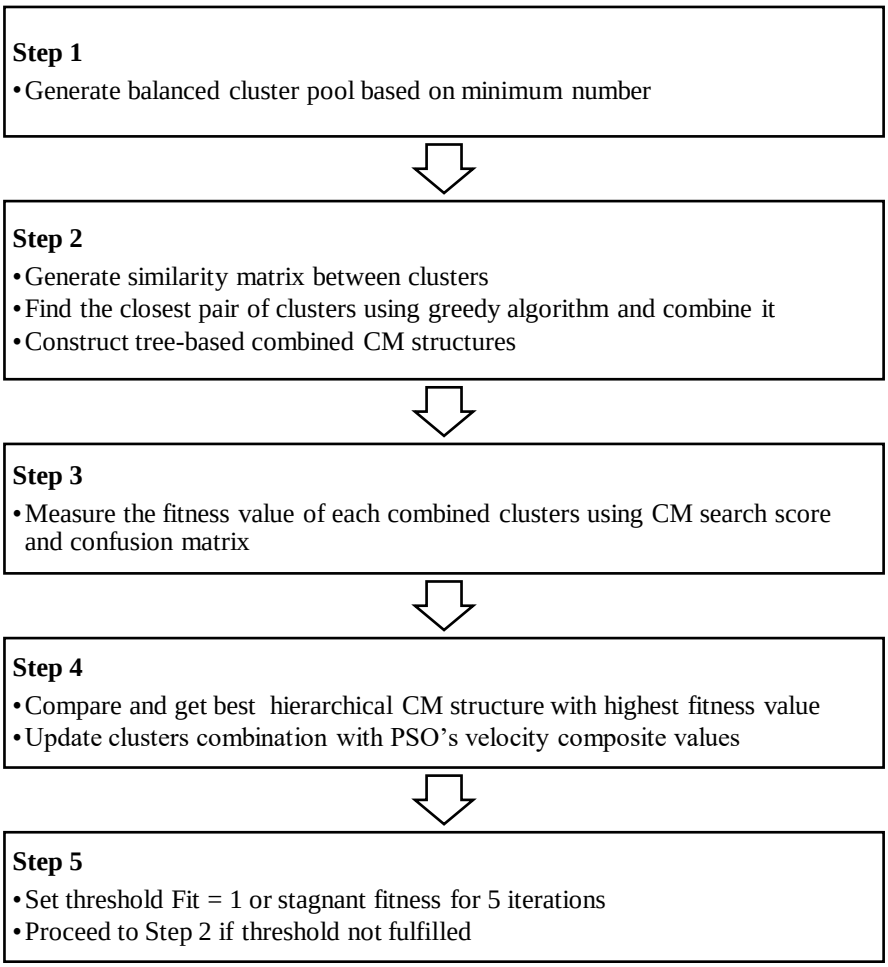

Figure 3. The flowchart to construct hybrid PSO and HACCM

where I , C , and S represent, respectively, the inertia weighting, cognitive acceleration, and social acceleration coefficients. I is set to 0.729844 in this investigation, while C and S are both set to 1.49618 [24]. As the particle is composed of N_f clusters, the implementation of these ratios is that maximum $\left\lfloor \frac{I_r \times N_f}{v_i(t)} \right\rfloor$ clusters are reselected from the clusters pool, maximum $\left\lfloor \frac{C_r \times N_f}{v_i(t)} \right\rfloor$ clusters are reselected from the particle's personal best, and maximum $\left\lfloor \frac{S_r \times N_f}{v_i(t)} \right\rfloor$ clusters are reselected from the global best particle. The PSO-HACCM algorithm is depicted in Figure 3.

This study employs accuracy as the fitness value candidate because it does not rely exclusively on true and false positives, unlike the *F-score*, where the sum of these values can be zero. However, other criteria should be considered if the fitness values of two or more PSO particles are identical. In this investigation, the tiebreaker will be the total of the literal values of each family's bit-scores or similarity scores. Consequently, the recommended fitness function for this investigation is as follows:

$$F_i = \alpha \times \frac{Acc_i(t) - Acc_i(0)}{Acc_i(0)} + \beta \times \frac{SS_i(t) - SS_i(0)}{SS_i(0)} \quad t = 1,2 \ldots. \qquad (5)$$

where the Acc_i is the accuracy of ith member (particle) in the tth iteration, and SS_i is the sum of bit-scores from the families of ith member, with $Acc_i(0)$ and $SS_i(0)$ are the accuracy and sum of bit-scores of the original HACCM. α and β are the parameters used to determine the importance of classification accuracy and the subset size, where the $\alpha \in [0,1]$ and $\beta = 1 - \alpha$. In this study, the α is set to 0.9, while the β is set to 0.1.

III. Result and Discussion

This section describes the simulation results of a comparative study between the Hierarchical Agglomerative Clustering for Covariance Model (HACCM) and Particle Swarm Optimization HACCM (PSO-HACCM) by following the data preparation in Figures 4 and 5. Five sets of ncRNA gene families were selected from the *Rfam* database to test the CMs combination method. The selected gene families have a roughly similar average length, so their CM combination would not be biased towards either. Therefore, the number of selected sequences will be set to the lowest number of sequences accessible, 51.

After successfully obtaining the ncRNA family dataset from the *Rfam* database, the training and testing datasets must be prepared and processed for use by the Infernal package's utilities. The suggested technique randomly divides each gene family's selected sequences into three groups of three sequences. In contrast, the remaining unselected sequences will be used to validate the generic combined CM produced by existing and suggested methodologies using the testing dataset.

The simulation results present the accuracies and the sum of bit-scores of the best member from 50 executions using three members for each technique and the total processing time to complete all. Based on the results shown in Tables 1 and 2 with the percentage of the differences between the two techniques, shown in Tables 1 and 2, PSO-HACCM produces the best accuracy, the sum of bit-scores, and processing time compared to the original HACCM.

This study prefers to use the bit-score instead of the E-value, which is based on the bit-score, because the bit-score doesn't depend on the sequence database size, only on the covariance model and the target sequence. The bit-score is the log-odds score for the hit. Each technique is executed for 50 iterations to ensure the performance stability of the existing and proposed techniques.

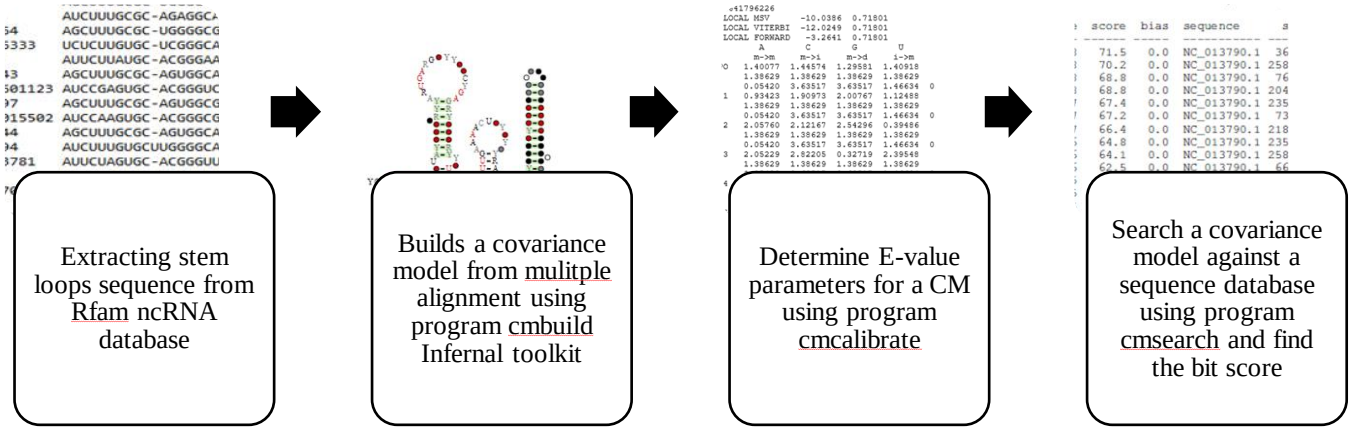

Figure 4. Data collection and preparation from *Rfam* using Infernal

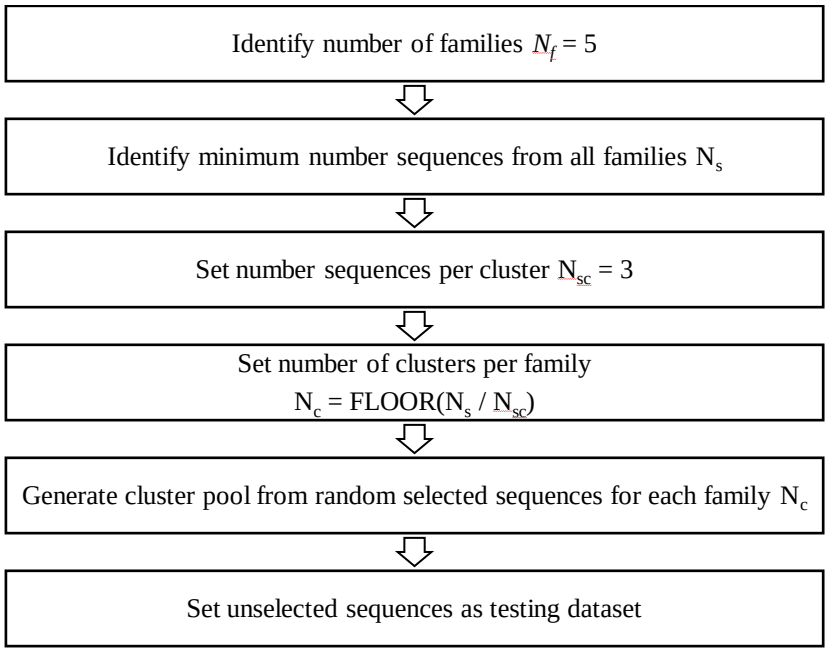

Figure 5. The flowchart to construct hybrid PSO and HACCM

Table 1. The accuracy, sum of bit-scores, and processing time results of HACCM and PSO-HACCM techniques from fifty executions

Accuracy		**Sum of Bit-Scores**		**Processing Time (s)**	
HACCM	PSO-HACCM	HACCM	PSO-HACCM	HACCM	PSO-HACCM
80%	85%	302.1	329.7	392.5	358.8
85%	85%	277.8	347.3	412.7	404.0
85%	95%	290.9	378.5	377.1	322.9
80%	90%	310.6	294.3	379.3	353.5
85%	85%	297.5	337.9	338.7	328.3
80%	100%	324.6	434.8	415.3	325.0
80%	85%	268	279.2	422.1	363.7
80%	85%	285	334.1	476.1	354.1
80%	85%	330.8	306.4	432.7	357.1
85%	90%	366.6	347.3	398.4	377.5
100%	100%	369.8	384.6	353.3	346.1

Table 2. Descriptive test results of HACCM and PSO-HACCM

Descriptive	**Technique**	**Accuracy**	**Sum of Bit-Score**	**Processing Time**
Mean	HACCM	82.8%	316	391
	PSO-HACCM	91.0%	359	357
Median	HACCM	80.0%	313	391
	PSO-HACCM	90.0%	362	355
Mode	HACCM	80.0%	302	328
	PSO-HACCM	95.0%	306	314

Table 3. Differences between PSO-HACCM and HACCM techniques

Compared Technique	**Accuracy**		**Sum of Bit-Score**		**Processing Time**	
	Value	**%**	**Value**	**%**	**Value**	**%**
PSO-HACCM vs. HACCM	4.8%	+10.2%	43	+13.6%	34	+8.7%

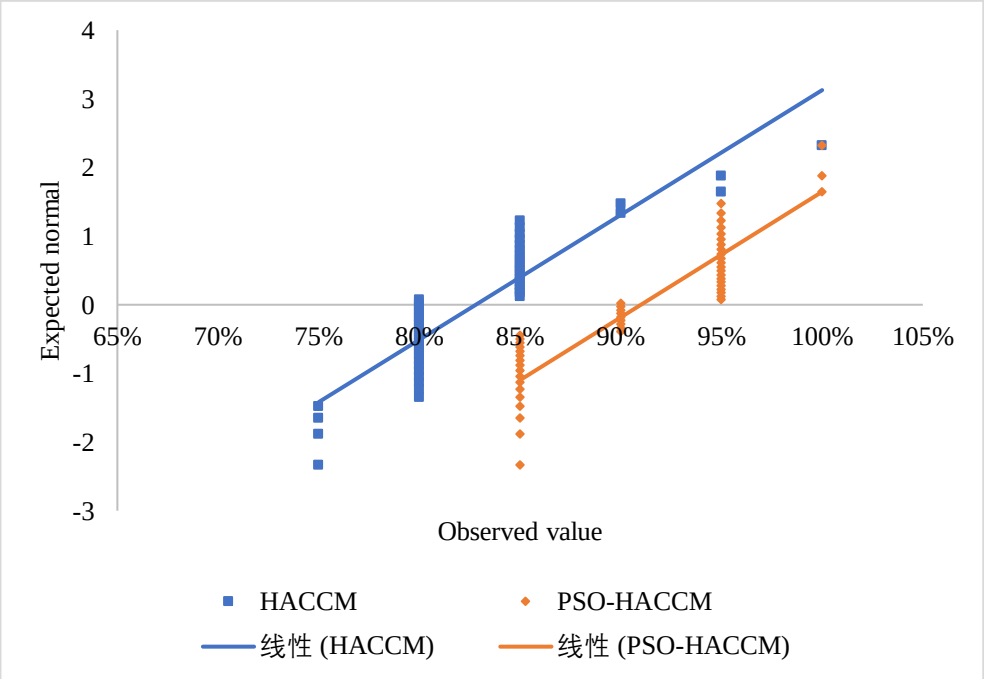

Figure 4. Normal Q-Q plot of the accuracy of HACCM and PSO-HACCM

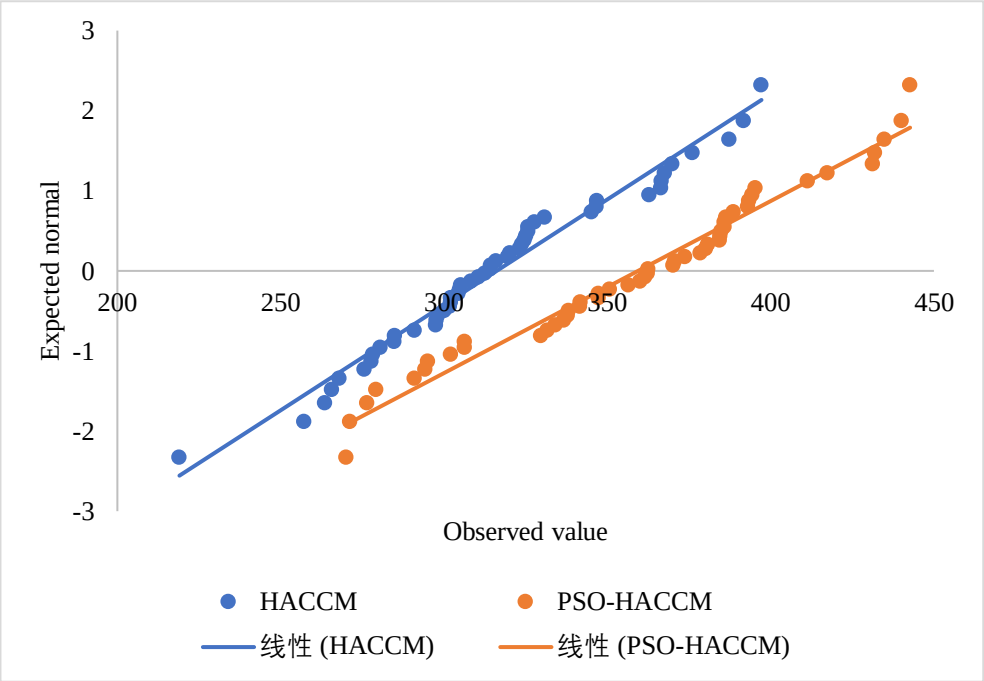

Figure 5. Normal Q-Q plot of the bit-scores of HACCM and PSO-HACCM

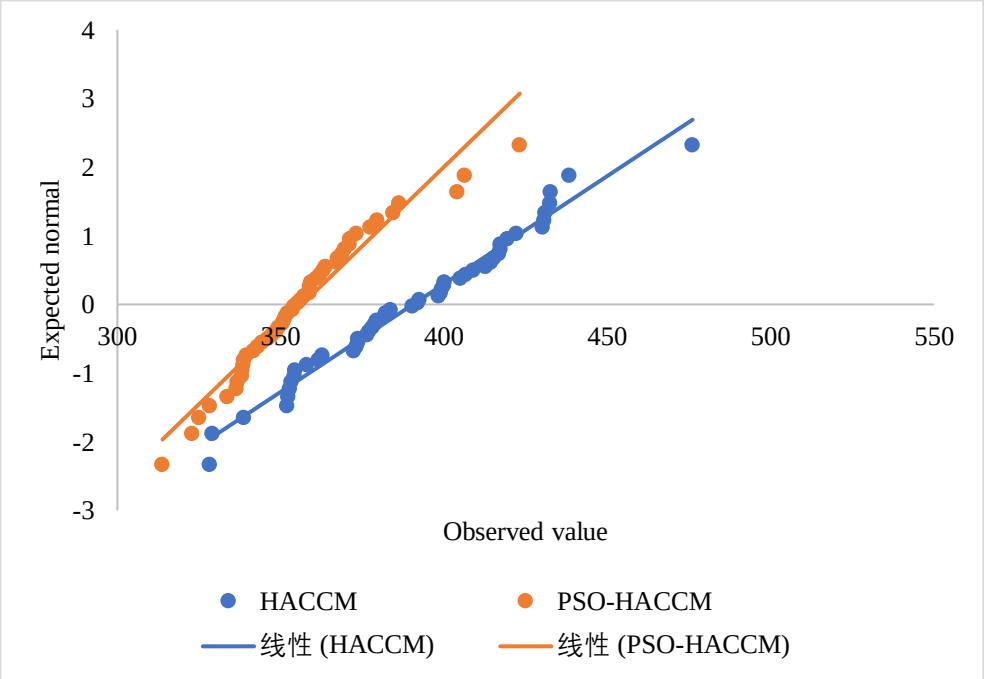

Figure 6. Normal Q-Q plot of the processing time of HACCM and PSO-HACCM

Table 3 shows that PSO-HACCM has 10.2% higher accuracy, 13.6% higher sum of bit score, and 8.7% faster than HACCM due to its cognitive and social accelerations capabilities. The results are also presented in graphical format using normal Q-Q plots in Figures 4, 5, and 6. However, an in-depth comparison of these techniques must be conducted to validate whether the difference between PSO-HACCM and original HACCM is statistically significant.

Both techniques have been tested for normality to determine whether their data is normally distributed using Shapiro–Wilk *W* tests of normality. Then it is concluded that the accuracies of PSO-HACCM and original HACCM techniques are normally distributed in Table 4. Therefore, the *t-test* can be conducted to validate the identification accuracy of these techniques, which is based on the post hoc comparisons using the Tukey HSD test shown in Table 5, indicating that the mean score for the identification accuracy of PSO-HACCM is statistically significantly better than HACCM $[t(147) = -4.31, p < 0.001]$.

Table 4. Shapiro–Wilk W tests of normality for assumption checks for HACCM and PSO-HACCM

W	**Sig.**
0.988	0.222

Table 5. Post hoc test results of identification accuracy using Tukey HSD for HACCM and PSO-HACCM

Post-hoc Test	Mean Difference	*t*-value	*df*	Sig.
Tukey HSD	-9.07	-4.31	147	< 0.001

IV. Conclusion

A hierarchical agglomerative clustering (HAC) method heuristically merges and clusters sequence features of each family, which are selected randomly to form the combined covariance model (CCM). Grouping several sub-families permits the modelling of the observed variation between the sub-families and may enable the discovery of new family members with different combinations of the various features than those observed in the initial training set. However, essential characteristics and features of a given subfamily may be diluted by mixing with other sub-families that lack those characteristics.

Therefore, the objective of this study is to propose this technique as the improvement in constructing the best candidate of combined CMs and to measure the performance of swarm intelligence-based hierarchical clustering for ncRNA identification. Hence, the proposed technique acts as an enhanced mechanism to ensure a balanced dataset is used to generate HACCM.

Comprehensive performance measurements, such as the comparison of accuracy, the sum of bit-scores, and processing time, have also been conducted, which corresponds to the purpose of this study, which is to formulate a balanced and optimised combined CMs hierarchical model structure in identifying ncRNA families. Thus, based on the performance measurement and statistical validations, the proposed PSO-HACCM performs better than the HACCM technique in terms of optimal performance, identification accuracy, the sum of bit-scores, and processing time.

Acknowledgement

Authors would also wish to express their gratitude towards Skim Zamalah UTeM provided by Universiti Teknikal Malaysia Melaka (UTeM). The authors would also like to thank UTeM for the research facilities provided.

References

[1] C. Biology and T. B. Laboratories, "Introns and the RNA World," *RNA World*, pp. 221–232, 1999.

[2] H.-H. Tseng, Z. Weinberg, J. Gore, R. R. Breaker, and W. L. Ruzzo, "Finding non-coding RNAs through genome-scale clustering.," *J. Bioinform. Comput. Biol.*, vol. 7, no. 2, pp. 373–88, 2009.

[3] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models," *Nucleic Acids Res.*, vol. 22, no. 11, pp. 2079–2088, 1994.

[4] S. F. Smith, "Covariance searches for ncRNA gene finding," *Proc. 2006 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol. CIBCB'06*, pp. 320–326, 2006.

[5] W. Jiang and K. C. Wiese, "Combined covariance model for non-coding RNA gene finding," *IEEE SSCI 2011 - Symp. Ser. Comput. Intell. - CIBCB 2011 2011 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol.*, pp. 22–26, 2011.

[6] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, "Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering," *PLoS Comput. Biol.*, vol. 3, no. 4, pp. 680–691, 2007.

[7] Y. Saito, K. Sato, and Y. Sakakibara, "Fast and accurate clustering of noncoding RNAs using ensembles of sequence alignments and secondary structures.," *BMC Bioinformatics*, vol. 12 Suppl 1, p. S48, 2011.

[8] T. Hermann and E. Westhof, "Non-Watson-Crick base pairs in RNA-protein recognition," *Chemistry and Biology*, vol. 6, no. 12. 1999.

[9] A. MacHado-Lima, H. A. Del Portillo, and A. M. Durham, "Computational methods in noncoding RNA research," *J. Math. Biol.*, vol. 56, no. 1–2, pp. 15–49, 2008.

[10] S. Zhang, I. Borovok, Y. Aharonowitz, R. Sharan, and V. Bafna, "A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements," *Bioinformatics*, vol. 22, no. 14, pp. 1–11, 2006.

[11] S. E. Butcher and A. M. Pyle, "The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks," *Acc. Chem. Res.*, vol. 44, no. 12, pp. 1302–1311, 2011.

[12] S. Crowder, J. Holton, and T. Alber, "Covariance analysis of RNA recognition motifs identifies functionally linked amino acids.," *J. Mol. Biol.*, vol. 310, no. 4, pp. 793–800, 2001.

[13] Z. Yao, Z. Weinberg, and W. L. Ruzzo, "CMfinder - A covariance model based RNA motif finding algorithm," *Bioinformatics*, vol. 22, no. 4, pp. 445–452, 2006.

[14] S. R. Eddy, "A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.," *BMC Bioinformatics*, vol. 3, p. 18, 2002.

[15] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[16] S. Alam, G. Dobbie, P. Riddle, and M. a. Naeem, "Particle Swarm Optimization Based Hierarchical Agglomerative Clustering," *Web Intell. Intell. Agent Technol. (WI-IAT), 2010 IEEE/WIC/ACM Int. Conf.*, vol. 2, pp. 64–68, 2010.

[17] G. Nowak and R. Tibshirani, "Complementary hierarchical clustering," *Biostatistics*, vol. 9, no. 3, pp. 467–483, 2008.

[18] J. A. Smith, "RNA search with decision trees and partial covariance models," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 6, no. 3, pp. 517–527, 2009.

[19] F. Murtagh and P. Contreras, "Methods of Hierarchical Clustering," *Computer (Long. Beach. Calif).*, vol. 38, no. 2, pp. 1–21, 2011.

[20] J. Augen, "Bioinformatics and Transcription," in *Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine*, 2005, p. 408.

[21] S. Wang, S. Hou, J. Wu, and J. Wei, "Clustering of ncRNA based on structural and semantic similarity," *J. Bionanoscience*, vol. 7, no. 1, pp. 20–25, 2013.

[22] D. Li *et al.*, "Experimental RNomics and genomic comparative analysis reveal a large group of species-specific small non-message RNAs in the silkworm Bombyx mori," *Nucleic Acids Res.*, vol. 39, no. 9, pp. 3792–3805, 2011.

[23] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: An RNA family database," *Nucleic Acids Research*, vol. 31, no. 1. pp. 439–441, 2003.

[24] S. Das, A. Abraham, and A. Konar, "Swarm intelligence algorithms in bioinformatics," *Stud. Comput. Intell.*, vol. 94, no. 2008, pp. 113–147, 2008.

[25] S. Alam, G. Dobbie, Y. S. Koh, P. Riddle, and S. Ur Rehman, "Research on particle swarm optimization based clustering: A systematic review of literature and techniques," *Swarm Evol. Comput.*, vol. 17, pp. 1–13, 2014.

[26] C. Chen, "Hierarchical Particle Swarm Optimization for Optimization Problems," *Tamkang Journal of Science and Engineering.*, vol. 12, no. 3, pp. 289–298, 2009.

[27] A. Abdullah, S. Deris, M. S. Mohamad and S. Z. M. Hashim, "A New Particle Swarm Evolutionary Optimization for Parameter Estimation of Biological Models". *IJCISIM*, vol. 3, 2013.

[28] N. I. Anuar and M. H. F. M Fauadi, "A Study on Multi-

Objective Particle Swarm Optimization in Solving Job-Shop Scheduling Problems". *IJCISIM*, vol. 13, 2021.

[29] C.-Y. Chen and F. Ye, "Particle swarm optimization algorithm and its application to clustering analysis," *Networking, Sensing and Control, 2004 IEEE International Conference on*, vol. 2. pp. 789-794 Vol.2, 2004.

Author Biographies

Lustiana Pratiwi was born in Jakarta, Indonesia on 23 June 1988. She received her BSc and MSc from Universiti Teknikal Malaysia Melaka (UTeM) in 2010 and 2013, respectively. She is also pursuing her PhD in Information and Communication Technology at the same university. Her research interests are in programming techniques, including Particle Swarm Optimization (PSO), data mining, bioinformatics, and software engineering.

Yun-Huoy Choo obtained her Bachelor's degree of Science and Computer with Education majoring in Mathematics (2000) and her Master of Information Technology (2002) from the University of Technology Malaysia (UTM). She completed her PhD in 2008 from the National University of Malaysia (UKM), specialising in Data Mining. She is currently lecturing at Universiti Teknikal Malaysia Melaka (UTeM). Her research interest includes rough set theory, fuzzy sets theory, association rules mining, discretisation and feature selection, vagueness, and uncertainty in specific, besides the fundamentals and application of data mining in general.

Azah Kamilah Muda received the BSc and MSc from Universiti Teknologi Malaysia (UTM), Malaysia, in 1997 and 1999, respectively. From 1997 to 2002, she joined the Department of Software Engineering at the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia as a lecturer. She obtained her PhD in bio-inspired pattern recognition from Universiti Teknologi Malaysia in 2009. Her research interests are Pattern Recognition, Image Processing, Soft & Biologically Inspired Computing, System Identification, and Artificial Intelligence.

Satrya Fajri Pratama is a lecturer at the Department of Computing, College of Business, Technology and Engineering, Sheffield Hallam University, United Kingdom, having previously served as a senior lecturer at Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka from 2019 to 2022. He was born in Bandung on 11 October 1988. He completed his BCS in 2010, MSc in 2013, and PhD in 2017, all from Universiti Teknikal Malaysia Melaka. His research interests are software engineering, computational intelligence, cheminformatics, enterprise and mobile application, cloud computing, and the Internet of Things.