RESEARCH ARTICLE

Editorial Process: Submission:12/21/2017 Acceptance:10/22/2018

Distributed ICSA Clustering Approach for Large Scale Protein Sequences and Cancer Diagnosis

Thenmozhi K[1]*, Karthikeyani Visalakshi N[2], Shanthi S[3], Pyingkodi M[3]

Abstract

Objective: With the over saturating growth of biological sequence databases, handling of these amounts of data has increasingly become a problem. Clustering has become one of the principal research objectives in structural and functional genomics. However, exact clustering algorithms, such as partitioned and hierarchical clustering, scale relatively poorly in terms of run time and memory usage with large sets of sequences. **Methods:** From these performance limits, heuristic optimizations such as Cuckoo Search Algorithm with genetic operators (ICSA) algorithm have been implemented in distributed computing environment. The proposed ICSA, a global optimized algorithm that can cluster large numbers of protein sequences by running on distributed computing hardware. **Results:** It allocates both memory and computing resources efficiently. Compare with the latest research results, our method requires only 15% of the execution time and obtains even higher quality information of protein sequence. **Conclusion:** From the experimental analysis, We noticed that the cluster of large protein sequence data sets using ICSA technique instead of only alignment methods reduce extremely the execution time and improve the efficiency of this important task in molecular biology. Moreover, the new era of proteomics is providing us with extensive knowledge of mutations and other alterations in cancer study.

Keywords: Proteomics- cancer diagnosis- molecular biology- distributed clustering- mutation- genetic algorithm

Asian Pac J Cancer Prev, **19 (11)**, 3105-3109

Introduction

In general, widely adopted paradigm in cancer diagnosis and treatment is that early cancer detection which increases likelihood of survival (Zhu et al., 2011). Nevertheless, insufficient treatment prohibits detection of certain types of cancer until last stages. At the molecular level, the main targets for drugs are proteins and nucleic acids (Singh et al., 2006). Most of the cancer diagnosis utilizes DNA analysis to characterize and detect disease. But, a few applications need protein analysis for greater accuracy. Due to that transcription at the gene level does not required with respect to expression at the protein level. Protein sequence analysis gives big chance for diagnosing, stratifying, and monitoring disease. This analysis must meet certain needs, in order to be clinically useful. However, numerous genome-sequencing projects have caused a rapid growth of the protein sequence databases (Wolf et al., 2001). The cuckoo search and support vector machine introduced to early detection of breast cancer tissues in mammographic breast images. The early detection of accuracy rate is better but it is increases the number of clusters as well as specificity was discussed in Deepika and Rajurkar, (2017).

With the over saturating growth of biological sequence databases, handling of these much amounts of data has increasingly become a problem. However, manual annotation of sequences is difficult and expensive (Day and Edelsbrunner, 1984). Therefore, now systematic approach allows us for the automatic functional classification of genome. One of the supervised machines learning method, classifiers performs fairly well at prediction if the training dataset is well prepared. The unsupervised learning approach of these data into functional groups or families, clustering, has become one of the principal research objectives in structural and functional genomics. However, most of the familiar clustering algorithms, such as partition and hierarchical clustering, scale relatively poorly in terms of run time and memory usage with large sets of sequences.

In order to overcome the problem associated with centralized (standalone) clustering algorithm, in literature authors has been introduced various distributed clustering approaches to handle the protein sequences in recent years. Distributed clustering algorithm is fully decentralized and makes all the different clusters grow concurrently. One of these, distributed hierarchical clustering (HPC-CLUST) has received considerable attention for large sets of protein

*[1]Department of Computer Applications, Selvam College of Technology, Namakkal, [2]Department of Computer Science, Government Arts and Science College, Kangayam, [3]Department of Computer Applications, Kongu Engineering College, Erode, Tamilnadu, India. *For Correspondence: thenmegu@gmail.com*

sequence analysis. It begins by taking every sequence separately and merging the two closest ones into a cluster (Matias and Mering, 2014).

Li and Godzik (2006) proposed a fast software for clustering and comparing large sets of protein or nucleotide sequences (CD-HIT). Schloss et al., (2009) developed software package that allows users to use a single piece of software to analyze community sequence data. This method provides to user to screen, trim, assign sequences; operational taxonomic units. Sun et al., (2009) proposed a method called estimating species richness using large collections of 16S rRNA pyrosequences. Massively it parallel pyrosequencing technology enables ultra-deep sequencing of complex microbial populations rapidly and inexpensively. The Ribosomal Database Project (RDP) gives author with quality-controlled bacterial and archaeal small subunit rRNA alignments and analysis platform (Cole et al., 2009). However, the existing implementations such as HPC-CLUST (Matias and Mering, 2014), CD-HIT (Li and Godzik, 2006), MOTHUR (Schloss et al., 2009), ESPRIT (Sun et al., 2009) or RDP online clustering (Cole et al., 2009), all struggle with large sets of sequences.

Materials and Methods

In the context of protein sequence data, development of algorithms can contribute towards the identification of cluster with coherent values. Finding the distributed clusters in a large protein data is a much more complex problem than clusters (Enright, 2002). The search space for the distributed clustering problem is 2m+n where m and n are the number of local and global level respectively. Usually m+n are more than 2000. In fact, it is proven to be a NP-hard problem. For that reason, meta-heuristic search algorithms are used to approximate the problem by finding sub optimal solutions. The primary objective of this work is to derive the meta-heuristic approaches for the identification of coherent clusters from protein sequence data with minimum computation time and memory usage.

Since, various approaches have been inspired from biological system, nature behaviour and physical processing (Jong, 2005). These methodologies have been successfully implemented on plenty of optimization problems, especially the protein sequence analysis. Among these metaheuristics are genetic algorithm, simulated annealing, cuckoo search, firefly algorithm, ant colony optimization, tabu search, particle swarm optimization (Yang, 2008). For solving a wide range of complex problems including combinatorial ones, such as the one studied here, within reasonable computing time, evolutionary algorithms have been proved to be successful. The results of such studies have demonstrated that evolutionary algorithms are capable of outperforming other principled approaches.

Improved Cuckoo Search Algorithm (ICSA)

Cuckoo search algorithm (CS) is a novel population based stochastic search meta-heuristic algorithm (Yang and Deb, 2009). It is mimics by natural mechanisms, the breeding behavior of some cuckoo species; lays their eggs in the nests of host birds. However, CS is a randomized method which is able to find the nearest local optimum. In benchmark CS, Maintenance of diversity within the population can be a problem, and some successful algorithms explicitly use mechanisms to preserve diversity. In order to accelerate the diversity process of potential local minima in our method, the cuckoo egg is modified by genetic operators.

Genetic Algorithm (GA) is a stochastic search technique (Goldberg, 2009). It is used to solve the engineering optimization problems and mathematical problems. It mimics Darwin's evolutionary theory; genetic inheritance and survival of the fittest. GA search efforts to detect the optimal solution to the problem by genetically breeding a population of individuals over a series of generations. GA process based on the sequential application of two operators: cross over and mutation.

There are two types of cuckoo bird: common and European cuckoo. In Europe, the cuckoo bird is associated with spring. Most Europe species of cuckoo are deskbound, whereas few take on regular seasonal migrations and others carry out biased migrations over their region. Based on this, if egg type is common (C_cuckoo) then make couple of eggs by using crossover operator with the two best eggs in the nest and choose the best one among them. Else (E_cuckoo) make couple of eggs using crossover with uniform mutation operator with any two eggs in the nest and choose the best one among them. Based on the evolutionary theory the genetic operator increases the chance of survival cuckoo's genes. It allows the cuckoo to spend more time for breeding and laying more eggs. Finally, it leads to the global optimum value.

Pseduocode of ICSA

```
Generate an initial population of n host nests
while (t<MaxGeneration)
    for each nest Get a cuckoo type randomly (say, i)
        Check the type of the cuckoo
        if type = C_cuckoo
            make two eggs with the help of crossover operator with the two best eggs in the nest and choose the best one among them
        else if type = E_cukcoo
            maketwo eggs using crossover with uniform mutation operator with any two eggs in the nest and choose the best one among them
        else
            Create an egg with random solution
        end if
        Evaluate its fitness Fi
        Choose an egg with the worst solution in the nest (say, j);
        if (Fi >Fj)
            Replace j by the new solution i;
        end if
        Rank the eggs based on the solution;
        Abandon a fraction (pa) of the eggs in the nest;
        Keep the best solutions;
    end for
    Rank the eggs in all nests using fitness value and find the current best;
end while
```

Results

An experiment conducted on large-scale protein data base to show the success of the novel proposed ICSA algorithm. The experiment was conducted on a Dell Blade M605 compute nodes with 2 quad-core Opteron 2.33 GHz processors and 24 GB of random access memory on Linux 5.2 system environment. In this work, the essential parameters of Cuckoo search and Genetic operators are adopted from benchmark algorithms (Yang and Deb, 2009; Goldberg, 2009). Table 1 shows the parameter setting of ICSA.

Table 1. Parameters for Proposed ICSA Algorithm

Parameters	Values
No. of nest	50
Crossover	0.9
Mutation	0.001
Maximum Iteration	100
Levy distributions	1.5
Probability of abortion nest	0.25

The ICSA is implemented for an online dataset of publicly available full-length 16S protein sequences from PDB sequence. It were aligned using INFERNAL v1.0.2 with a 16S model for bacteria from the ssu-align package (Nawrocki et al., 2009). Importantly, INFERNAL employs a profile alignment strategy that scales linearly O(N) with the number of sequences, and can be inconsequentially parallelized. The final dataset consisted of 1,105,195 sequences. Moreover, it has 833,013 unique sequences of 1,301 in aligned length.

In order to improve the memory and computational efficiency, ICSA compresses the sequences such that each protein employs only 4 bits rather than usual 8 bits. Other improvements have been done in the distance calculation by making use of 64 bit variables instead of 32 bit. Therefore, performing the comparison of 16 proteins at a time is possible. Figure 1 and 2 depicts the differences in runtime between clustering algorithms running on a single CPU and eight CPU respectively. Here, ICSA and CD-HIT were competent of nearly optimal parallelization; resulting very less wall clock runtime is required. However, CD-HIT performance is reduced when size of sequences are increased. In contrast, HPC-CLUST and MOTHUR tools are needed much wall clock runtime in both threads. ICSA is proficient of parallelizing the more costly steps of the clustering process, out coming in a huge reduce in wall clock runtime.

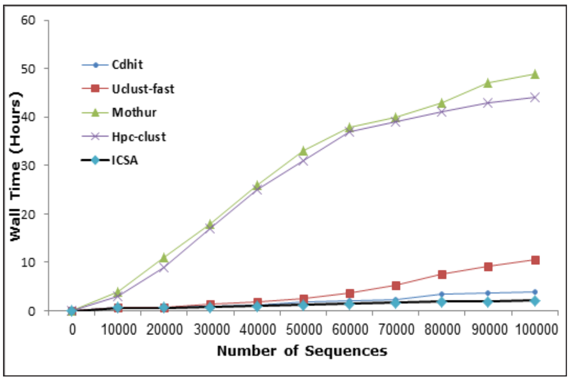

Figure 1. Performance Analysis for Number of Sequences Vs Wall Clock Runtime in 1 CPU

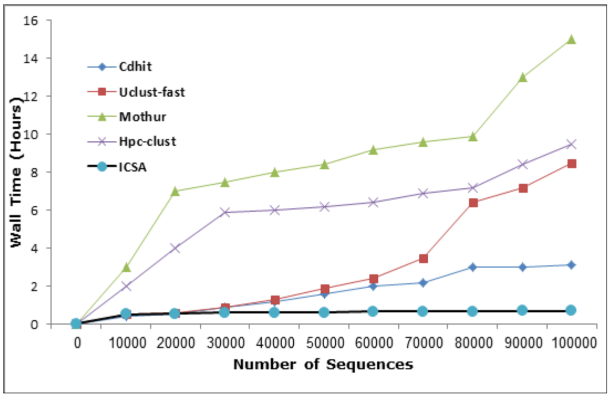

Figure 2. Performance Analysis for Number of Sequences Vs Wall Clock Runtime in 8 CPU

For example in Table 2, HPC-CLUST the number of true positive predictions increased by 493. However, this increases come with by an additional 174 false negative and 92 false positive predictions, thus decreasing the overall performance. Not all sequences were identical for both cases. Correspondingly, there were 36,000 sequences that were present in clusters but were not present in "40" clusters.

The performance of the algorithm for these 6,300 new sequences was 493 true positives, 174 false negatives and 92 false positives. This corresponds to 12% in specificity. Thus the performance of the ICSA was substantially better than the other methods both in terms of covered sequences and quality of annotation. This enhancement is of huge importance for an automatic annotation of protein sequences.

Table 2. Clustering of PDB Sequences Using ICSA, HPC-CLUST, MOTHUR, UCLUST-FAST, CD-HIT Algorithms

Methods	Clusters	True positive	False positive	False negative	Specificity
ICSA	700	8507	586	489	92.6
HPC-CLUST	740	9000	412	392	95.1
MOTHUR	720	9297	401	378	95.5
UCLUST-FAST	880	9508	390	361	97.8
CD-HIT	900	9603	315	322	96.7

Discussion

Data clustering used to detect the similar or dissimilar data from huge data. Fuzzy cuckoo optimization

Algorithm was introduced to get a better performance. Fuzzy used to get optimal solution and Cuckoo Optimization algorithm used to solves the problem of data clustering. This searching optimization gives the quality of solution and average number of function. But, the specificity was not improved in early discussed paper (Amiri and Mahmoudi, 2016). The earlier cancer detection of cuckoo search optimization and support vector machine in (Prabukumar, et al., 2017), the data segmented by Fuzzy C mean, then features selected by Cuckoo search. Not only used for earlier detection, mainly applied for the efficient accuracy of cancer detection. Elyasigomari et al., (2017), reported to avoid the dimensionality problem, it introduced the Minimum Redundancy and Maximum Relevance (MRMR) features selection method for four microarray datasets. Performance validated by cross validation measure. The appropriated gene data validate by cross-validation method and compared with evolutionary algorithm. The selected genes are relevant to cancer type at the same time it takes more time for cancerous gene detection.

The Cuckoo Optimization with Genetic Algorithm used for data clustering to detect the cancerous data (Elyasigomari, et al., 2015). The k means, hierarchical algorithms does not get the efficient accuracy than COA-GA. Particle Swarm optimization (PSO) performs the better classification and also it reaches a best solution in minimum iterations. But it not achieves the false positive rate. K nearest neighbors and support vector machines was used to detect the performance of cuckoo search with the rough set. Rough set is the new computational intelligent techniques, which is used to evaluate the uncertainty in terms of cancer detection. The rough set based fitness function distinct on two factors which is reducing the number of features and classification quality. In terms of removing irrelevant, redundant or noisy features while sustain the classification accuracy in (Abd El Aziz and Hassanien, 2018), however the false positive rate and specificity is not improved. The proposed ICSA techniques used to cluster the huge protein sequence data to detect the cancerous sequence in terms of efficient memory, minimum computing resource and reduce the time of execution.

The proposed ICSA has been highly improved for computation speed and memory efficiency. Even when running on a single machine, this is much faster than the exact clustering implementations tested here. Compared with HPC-CLUST, it returns better clustering results. Whereas compared with MOTHUR, HPC-CLUST yields identical or nearly identical results. UCLUST-FAST and CD-HIT use a different approach to clustering; they are not directly creating the impact in the results. Clustering the entire dataset to 98% identity threshold needed a total of 2 h and 30 min on a compute cluster of 16 nodes with 8 cores each (128 total cores). Due to parallelization, the solution initialization and fitness computation took only 42 min, with respect to >20,000 min CPU time. The remaining 1 h and 48 min was spent initializing and clustering the nodes. The entire memory used for the computation was 1.9 GB per node.

In conclusion, we have implemented that the newly proposed ICSA algorithm can process several hundreds of thousands of sequences within 42 min by using a relatively small computing environment. To our knowledge, no existing methods can efficiently handle such large data. ICSA, however, does provide a promising direction for analyzing large collection of protein sequence data. From the result, more than 650 genes have been found as strongly implicated in the process of transforming normal cells to cancer cells. Moreover, the new era of proteomics give us with wide knowledge of mutations and other alterations in cancer study. In future, this distributed concept is functional by soft computing methodologies combined with spectral clustering for huge data which improves the effect for the early detection of cancer using protein sequence and structural motifs for detect the active and inactive sequence of cancer.

Statement conflict of Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names:

1. K. Thenmozhi
2. N. Karthikeyani Visalakshi
3. S. Shanthi
4. M.Pyingkodi

Acknowledgments

The authors would like to thank the Institution for Selvam College of Technology, Namakkal, and Kongu Engineering College, Erode, Tamil Nadu, India for much assistance to the publication of this paper.

References

Abd El Aziz M, Hassanien AE (2018). Modified cuckoo search algorithm with rough sets for feature selection. *Neural Comp Appl*, **29**, 925–34.

Amiri E, Mahmoudi S (2016). Efficient protocol for data clustering by fuzzy Cuckoo Optimization Algorithm. *Appl Soft Comp*, **41**, 15-21.

Cole JR, Wang Q, Cardenas E, et al (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*, **37**, 141–5.

Day W, Edelsbrunner H (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *J Classif*, **1**, 7–24.

Deepika KN, Rajurkar AM (2017). Cuckoo search: An optimized way for mammogram feature selection. *Int J Eng Technol Sci Innov*, **20**, 81-6.

Elyasigomari V, Mirjafari MS, Screen HRC, et al (2015). Cancer classification using a novel gene selection approach by meansof shuffling based on data clustering with optimization. *Appl Soft Comp*, **35**, 43-51.

Elyasigomari V, Lee DA, Screen HR, et al (2017). Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. *J Biomed Inform*, **67**, 11–20.

Enright AJ, Van Dongen S, Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, **30**, 1575–84.

Goldberg DE (1989). Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, pp 1-432.

Jong KD (2005). Genetic algorithms: A 30 year perspective, in Perspectives on Adaptation in Natural and Artificial Systems, Oxford University Press, Oxford, UK, pp 1-18.

Li W, Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-9.

Matias Rodrigues JF, von Mering C (2014). HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics*, **30**, 287–8.

Nawrocki EP, Kolbe DL, Eddy SR (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–7.

Patrick DS, Sarah LW, Thomas R, et al (2009). Introducing MOTHUR: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*, **75**, 7537–41.

Prabukumar M, Agilandeeswari L, Ganesan K (2017). An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier. *J Ambient Intelligence Humanized Comp*, **1**, 1–27.

Singh S, Malik BK, Sharma DK (2006). Molecular drug targets and structure based drug design: A holistic approach. *Bio Information*, **1**, 314–20.

Sun Y, Cai Y, Liu L, et al (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res*, **37**, e76.

Wolf YI, Rogozin IB, Kondrashov AS, et al (2001). Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. *Genome Res*, **11**, 356–72.

Yang XS (2008). Nature-Inspired Metaheuristic Algorithms. Luniver Press, UK, pp 1-75.

Yang XS, Deb S (2010). Cuckoo search via lévy flights', Proceedings of World Congress on Nature and Biologically Inspired Computing. India, IEEE Publications, USA, pp 210-4.

Zhu A, Lee D, Shim H (2011). Metabolic PET imaging in cancer detection and therapy response. *Semin Oncol*, **38**, 55–69.