Conceptual Clustering Of RNA Sequences With Codon Usage Model

GJCST Classification
J.3.H.3.3

Barilee B. Baridam[1], Olumide Owolabi[2]

***Abstract*-This paper proposes a conceptual clustering approachfor RNA sequences using codons. It is shown thatemploying the codons (codon usage model) in the conceptualclustering of RNA sequences has high efficiency and robustnesscompared to conventional clustering methods. In cases wherethere are hidden structural patterns, homology search algorithmsare inefficient in locating similar sequences and as a result arenot reliable in the task of biological sequence clustering. As isshown by empirical results in this paper, conceptual clusteringusing the codons is able to discover similar sequences in adatabase of sequences with hidden structural homologues. Thecodon usage and cohesiveness model introduced in this papercan be efficiently employed in clustering biological sequence datawhere conventional homology search algorithms fail.**

Keywords- Conceptual clustering, cohesiveness, codon usage,formal concept analysis, RNA sequences.

I. Introduction

Conventional Data analysis employs context-free similarity measures, that is, similarity based on the properties of the objects without considering the environment where the objects are found. On the other hand, contextsensitive similarity measures are not only based on the properties of the objects but also the properties of the surrounding environment. All these similarity measures (context-free andcontext-sensitive) are concept-free. Similarity search based on a set of concepts describing objects, and not just on properties and environment, are what is employed in this paper.

Although biological data can be clustered using context-free similarity measures (Lee & Crawford 2005), the clustering of biological sequence data with context-sensitive similarity measures may not be appropriate. This is because the environment has little or no effect on already sequenced biological data. However, context-free homology searches can only yield less than 60% found genes and only a few of the searches can result in assigning the correct structure of the genes (Math´e, Peresetsky, D´ehais, Van Montagu & Rouz´e 1999). Therefore, biological clustering using conceptual clustering, clustering based on sets of concepts, by employing the codon usage (CU) model becomes appropriate to cluster sequences with hidden biological patterns.

Conceptual clustering is employed in this paper for the task of clustering RNA sequences. The goal is to employ codons,otherwise referred to as the CU model, and the cohesiveness model (the degree of codon cohesion) in clustering RNA sequences. Conceptual cohesiveness, from which codon cohesiveness is derived, is a measure of similarity between two points based on a set of concepts available for describing the two points (Michalski & Stepp 1986). The method has the ability to cluster sequences which would not ordinarily be clustered with conventional categorical clustering methods like CLUSEQ - CLUstering for SEQuences, ED - Edit Distance, and EDBO - Edit Distance with Block Operations (Yang & Wang 2003), (Levenshtein 1965), (Lopresti & Tomkins 1997).

The remainder of this paper is arranged as follows: A brief look at formal concept analysis followed by related work, themethods employed in this paper for the clustering of biologicalsequence data, followed by some experimental results, andlastly conclusions and future research.

II. Conceptual Clustering

Conceptual clustering is a machine-learning paradigm forunsupervised classification that aims at generating a conceptdescription for each generated class. This section considers

formal concept analysis (FCA) and the Galois or conceptlattice.

A. *Formal Concept Analysis*

FCA aims at the automatic derivation of ontology based on a collection of objects and their properties. FCA, introduced byRudolf Wille and his students in 1984, is a direct applicationof the applied lattice and order theory developed by Birkhoffand others in the 1930s (Birkhoff 1930). FCA attempts tofind all the natural clusters of properties and all the naturalclusters of objects in the input data. The set of all objects thatshare a common subset of properties or attributes is referredto as a natural object cluster, while the set of all propertiesor attributes shared by one of the natural object clusters isreferred to as a natural property cluster.

i. Concepts Definition

From the description of FCA, conceptanalysis employs a set of objects and a set of propertiesor attributes belonging to all or some of the objects. For everyset of objects O, set of properties P and an indication of whichobject has which attribute, a concept can be defined to be a pair(Oi; Pi) such that the following conditions hold(Vinner 1983):

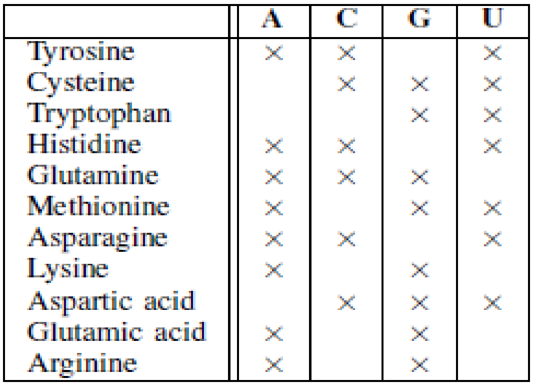
1) $O_i \subseteq O$
2) $P_i \subseteq P$

B. B. Baridam is with the Department of Computer Science, University of Pretoria, South Africa, 0083. E-mail: bbaridam@cs.up.ac.za
O. Owolabi is the Director of Computer Science Centre, University of Abuja, Nigeria. E-mail: olumideo@uniabuja.edu.ng

1) Every object in $O_i$has every attribute in Pi
2) For every object in O that is not in O_i, there is an attribute in Pi that the object does not have
3) For every object in P that is not in P_i, there is an attribute in O_i that does not have that attribute.

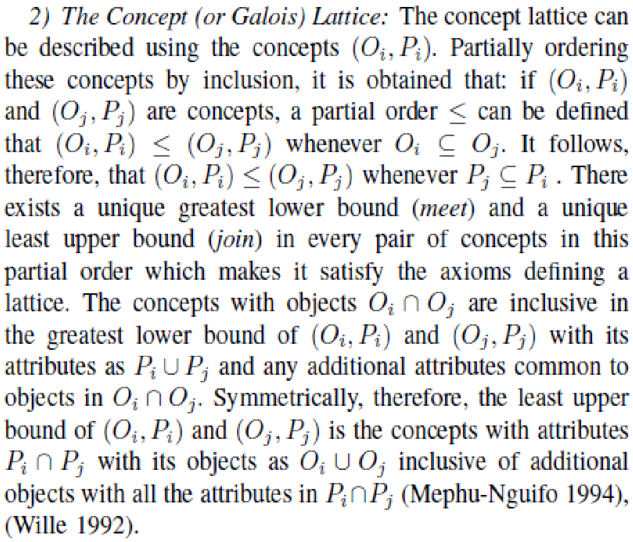
From the definition above, it can be said that a concept isa pair containing both a natural property cluster and its correspondingobject cluster. The mathematical axioms defining

TABLE I
CONCEPT REPRESENTATION WITH NUCLEOTIDES



	A	C	G	U
Tyrosine	×	×		×
Cysteine		×	×	×
Tryptophan			×	×
Histidine	×	×		×
Glutamine	×	×	×	
Methionine	×		×	×
Asparagine	×	×		×
Lysine	×		×	
Aspartic acid		×	×	×
Glutamic acid	×		×	
Arginine	×		×	

A lattice based on these concepts are referred to as concept lattice or as a general term, Galois lattice.



2) The Concept (or Galois) Lattice: The concept lattice can be described using the concepts (O_i, P_i). Partially ordering these concepts by inclusion, it is obtained that: if (O_i, P_i) and (O_j, P_j) are concepts, a partial order $\leq$ can be defined that $(O_i, P_i) \leq (O_j, P_j)$ whenever $O_i \subseteq O_j$. It follows, therefore, that $(O_i, P_i) \leq (O_j, P_j)$ whenever $P_j \subseteq P_i$. There exists a unique greatest lower bound (*meet*) and a unique least upper bound (*join*) in every pair of concepts in this partial order which makes it satisfy the axioms defining a lattice. The concepts with objects $O_i \cap O_j$ are inclusive in the greatest lower bound of (O_i, P_i) and (O_j, P_j) with its attributes as $P_i \cup P_j$ and any additional attributes common to objects in $O_i \cap O_j$. Symmetrically, therefore, the least upper bound of (O_i, P_i) and (O_j, P_j) is the concepts with attributes $P_i \cap P_j$ with its objects as $O_i \cup O_j$ inclusive of additional objects with all the attributes in $P_i \cap P_j$ (Mephu-Nguifo 1994), (Wille 1992).

Biological sequence clustering, using conceptual clustering based on the CC model, becomes appropriate, therefore, to capture hidden biological (structural) pattern in sequence data.Following the rule for conceptual clustering, the objects andtheir attributes (properties) are derived as explained below.The objects are derived from the nucleotides in peptideformation during RNA translation using the basic RNA nucleotides- A, C, G and U. The nucleotides are the attributes.These peptides are Tyrosine, Cysteine, Tryptophan, Histidine,Glutamine, Methionine, Asparagine, Lysine, Aspartic acid,Glutamic acid and Arginine.

A tabular representation of these peptides showing their properties (attributes) based on their nucleotide formation, is given in Table I. A cross (X) in the cells indicates the presence of an attribute, while a space indicates none. Note that the bases are in triplets, referred to as a codon, and that several contiguous bases (codons) may form a particular peptide and so a base can be repeated twice or three times, depending on the peptide involved, e.g. Lysine and Arginine with AAA, AAG and AGA, AGG, respectively.

Table I serves as a guide in the clustering of nucleic acidsequences. In the clustering task, sequences are represented asobjects while peptides are the attributes.

III. RELATED WORK

Several algorithms have been proposed for conceptual clusteringsince the idea was developed in the 1980s. Carpinetoand Romano (Carpineto & Romano 1993), introduced GALOISwhich is an order-theoretic approach to conceptualclustering. From experimental results presented, Carpineto andRomano argued that GALOIS performs better than other methods.Michalski and Stepp (Michalski & Stepp 1986) developedthe conjunctive conceptual clustering program CLUSTER/2in which the predefined concept class consists of conjunctivestatements involving relations on selected object attributes.The method was experimented on a large collection of Spanishfolk songs. The result proved the efficiency of CLUSTER/2in the clustering task. Kolodner (Kolodner 1983) proposedthe CYRUS algorithm, which was also an improvement onexisting methods. An earlier paper by Michalski (Michalski1980) introduced the idea of partitioning data into conjunctiveconcepts to handle knowledge acquisition through conceptualclustering. Furthermore, Lebowitz (Lebowitz 1987) proposedthe UNIMEM algorithm for incremental concept formation inconceptual clustering problems as a system that learns from observation by noticing regularities among examples and organizingthem into a generalization hierarchy. In the same year,Fisher (Fisher 1987) came up with the COBWEB algorithm forknowledge acquisition via incremental conceptual clustering.The most recent algorithms in this field were proposed byJonyer et al. (Jonyer, Cook & Holder 2001) and Talavera andB´ejar (Talavera & B´ejar 2001), namely SUBDUE and GCF,respectively. Talavera and Bjar employed probabilistic conceptsin performing a generality-based conceptual clustering.Despite the successful implementation of conceptual clusteringin data analysis (Kuminek & Kazman 1997),(Ketterlin,Ganc¸arski & Korczak 1995), it has not been employed as muchin the field of bioinformatics to date. The most recent workon the application of conceptual clustering in the clustering ofbiological data is the work done by McClean et al. (McClean,Scotney & Robinson 2001) on the conceptual clustering ofheterogeneous gene expression sequences. Other work thatmay look like conceptual clustering, though not explicitlystated, was done by Math´e et al. (Math´e et al. 1999). In theclassification of Arabidopsis thalianagene sequences, codonusage was employed by Math´e et al. in the classificationof coding sequences into two groups. The result was animprovement in the quality of gene prediction compared toexisting methods.It is important to note that other than the work presentedby Math´e et al. (Math´e et al. 1999) none of the methodsmentioned above considered the application of conceptualclustering in the clustering of biological sequences, althoughthe work presented by Math´e et al. is limited to a particularset of gene sequences.

IV. THE CODON COHESIVENESS MODEL

TABLE II
CLUSTERS GENERATED BY CLUSTAL

Cluster	Sequence
1	7,16,5,15,3,6,1
2	13,20,12,2,17,4
3	14,11,9,19,10,18,8

The codon cohesiveness model employs what is referredto here as codon usage in determining the frequency of each codon in a given sequence. The codon usage (CU) of a given sequence is defined as:

$$CU = \frac{f_c}{S_l}F_l \qquad (1)$$

where f_c= the relative codon frequencies, S_l = the sequence length and F_l = the feature (codon) length. The feature lengthis a constant and is equals 3, since there are just three basesthat form a codon.

The codon cohesiveness (CC) or the degree of cohesion is now defined based on the CU as follows:

$$CC = \sum_{i=0}^{N} \frac{f_{c_i}}{S_l}.F_l \;= \sum_{i=0}^{N} CU_i \qquad (2)$$

The values of CU and CC are between 0 and 1. CCdetermines to what extent the sequence to be clustered is closeto the peptide group - the attribute.Codon cohesiveness is used to group similar sequencesbased on the occurrence of codons. Sequences with higheroccurrence of a peptide group are grouped in the same cluster.

V. EXPERIMENTAL RESULTS

The method was tested on 20 Rickettsia typhi str. sequences from the Wilmington complete genome. Patternelement-wise search was used in detecting available codonsin the sequences. When the edit distance was employed inthe search, it was found that none of the sequences was atleast 60% similar, based on the homology principle (Claverie& Notredame 2007), and so the clustering result was notuseful. Also, clustering Rickettsia typhi str. sequences with edit distance violates the rule that nucleic acid sequences canonly be considered homologue if and only if they are or morethan 70% similar (Claverie & Notredame 2007).

Overlaps are encountered with this clustering technique. The solution used to overcome the problem of overlaps is the CC model. In the result obtained in Table IV, sequences with at least 30% amino acid occurrence are grouped based on their CC values. When this was done, 6 clusters were generated as indicated in Table II using the peptide formation grouping.Of all the sequences clustered, sequences 1, 2, 4, 6, 15 and 17 have some similarities. However, they could not be grouped based on the values of the CU model. The CU values and the resultant CC values for these sequences are less than 20%.However, they cannot be considered as outliers since they manifest some measure of similarity. Recall that the highest CU or CC values renders a sequence clusterable. However,sequence 3 could not be grouped although it has the highest CU and CC values. The method employed here reveals that sequence 3 has a STOP signal. This makes it different from the rest of the sequences tested. It will not be out of place to consider sequence 3 outlier.

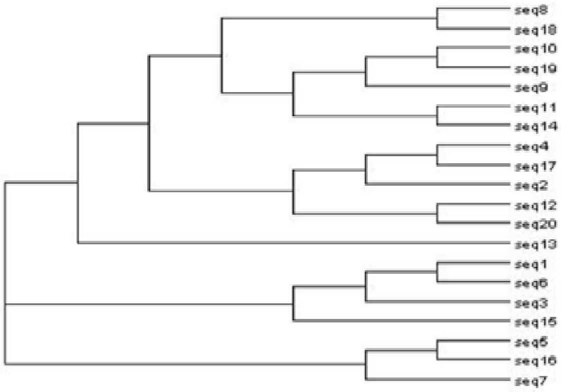


Fig. 1. Generated phylogenetic tree of the sequences

With crisp clustering (sequences belong to one and only one cluster), six clusters were generated as indicated in Table IV.From the results it is evident that the method employed in this paper produces clusters of even shape based on their codons.CLUSTAL produced three clusters with crisp clustering. The result of CLUSTAL clustering is indicated in Table II.Employing fuzzy clustering, Table III produces more clusters of sequences 11 and 14; 11, 13 and 18; 9, 10 and 19; 8,10 and 13; 5, 11 and 16; 12 and 20; 5, 7 and 16, forming separated clusters.The result was compared with a constructed phylogenetic tree of the sequences. A phylogenetic tree (Figure V) is used to show how related the sequences are based on their genetic composition, thus defining or at the very least, giving the idea of the composition of clusters that may be formed by any clustering or similarity search algorithm. Note that phylogenetic trees are constructed mostly using multiple-alignment algorithms. Note also that alignment algorithms introduce gaps to achieve sequence alignments (Corpet 1988), (Gondro & Kinghorn 2007), (Notredame & Higgins 1995). To prove the inefficiency of such methods, gaps are penalized. The clustering done in this paper does not consider the introductionof gaps, hence, the result is somewhat different and better than the one achieved with other methods that use aligned sequences.

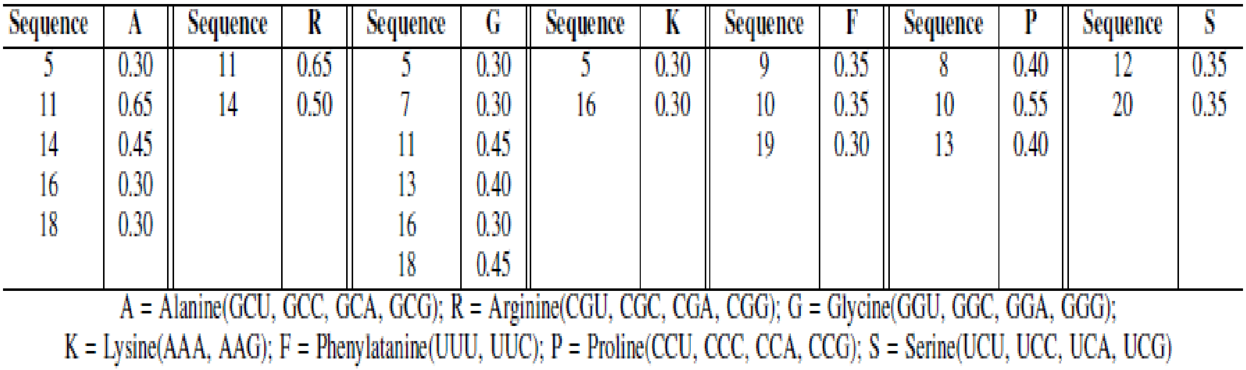
VI. CONCLUSION

Conceptual clustering is successfully employed in this paper to cluster RNA sequences through the application of the geneticcode triplet bases arrangement referred to as codon. Themethod is a strong deviation from popular clustering methods.The result obtained from the method is promising and couldbe extended to other areas of biological sequence clustering.Further research on this work could involve the clustering ofother biological sequences, for example amino acids.

VII. References

1) Birkhoff, G. D. (1930), _Fomal theory of irregular linear difference equations', Acta Mathematica 54(1), 205–246.
2) Carpineto, C. & Romano, G. (1993), Galois: An order-theoretic approach to conceptual clustering, in _Proceedings of10th International Conference on Machine Learning', Amherst, pp. 33–40.
3) Claverie, J. & Notredame, C. (2007), Bioinformatics for dummies, 2nd edn,Wiley, Indiana.
4) Corpet, F. (1988), _Multiple sequence alignment with hierarchical clustering',Nucleic Acids Research 16(22), 10881–10890.
5) Fisher, D. H. (1987), _Knowledge acquisition via incremental concept clustering',Machine Learning 2, 139–172.
6) Gondro, C. & Kinghorn, B. P. (2007), _A simple genetic algorithm for multiplesequence alignment', Genetics and Molecular Research 6, 964–982.
7) Jonyer, I., Cook, D. J. & Holder, L. B. (2001), _Graph-based hierarchicalconceptual clustering', Journal of Machine Learning Research 2, 19–43.
8) Ketterlin, A., Ganc¸arski, P. & Korczak, J. J. (1995), Conceptual clustering instructured databases: A practical approach, in _Proceedings of KDD',pp. 180–185.
9) Kolodner, J. L. (1983), _Reconstructive memory: A computer model', CognitiveScience 7, 281–328.
10) Kuminek, J. & Kazman, R. (1997), Accessing multimedia through conceptclustering, in _Proceedings of ACM CHI', pp. 19–26.
11) Lebowitz, M. (1987), _Experiments with incremental concept formation',Machine Learning 2, 103–138.
12) Lee, S. & Crawford, M. M. (2005), _Unsupervised multistage image classificationusing hierarchical clustering with a Bayesian similarity measure',IEEE Transactions on Image Processing 14(3), 312–320.
13) Levenshtein, V. I. (1965), _Binary codes capable of correcting deletions, insertions, and reversals', Doklady Akademii Nauk SSSR 163(4), 845–848.
14) Lopresti, D. & Tomkins, A. (1997), _Block edit models for approximate stringmatching', Theoretical Computer Science 181, 159–179.
15) Math´e, C., Peresetsky, A., D´ehais, P., Van Montagu, M. & Rouz´e, P. (1999),_Classification of Arabidopsis thaliana gene sequences: Clustering ofcoding sequences into two groups according to codon usage improvesgene prediction', Journal of Molecular Biology 285, 1977–1991.
16) McClean, S., Scotney, B. & Robinson, S. (2001), _Conceptual clustering ofheterogeneous gene expression sequences', Artificial Intelligence Review20, 53–73.
17) Mephu-Nguifo, E. (1994), Galois lattice: a framework for concept learning.design,evaluation and refinement, in _Proceedings of Sixth International Conference on Tools with Artificial Intelligence', New Orleans, LA,USA, pp. 461–467.
18) Michalski, R. S. (1980), _Knowledge acquisition through conceptual clustering:A theoretical framework and an algorithm for partitioning datainto conjunctive concepts', International Journal of Policy Analysis andInformation Systems 4, 219–244.
19) Michalski, R. S. & Stepp, R. E. (1986), Learning from observation: Conceptualclustering, in R. S. Michalski, J. G. Carbonell & T. M. Mitchell,eds, _Machine learning - An artificial intelligence approach', MorganKaufmann, Los Altos, CA, pp. 471–498.
20) Notredame, C. & Higgins, D. G. (1995), _SAGA a genetic algorithm for multiple sequence alignment', Nucleic Acids Research 174, 1515.
21) Talavera, L. & B´ejar, J. (2001), Generality-based conceptual clustering withprobabilistic concepts, in _IEEE Transactions on Pattern Analysis andMachine Intelligence', Vol. 23, Amherst, pp. 196–206.
22) Vinner, S. (1983), _Concept definition, concept image and the notion of function', International Journal of Mathematical Education in Science and Technology 14(3), 293–305.
23) Wille, R. (1992), _Concept lattices and conceptual knowledge systems',Computers and Mathematics with Applications 23(6–9), 493–515.
24) Yang, J. & Wang, W. (2003), CLUSEQ: efficient and effective sequenceclustering, in _Proceeding of 19th International Conference Data Engineering',pp. 101–1125.

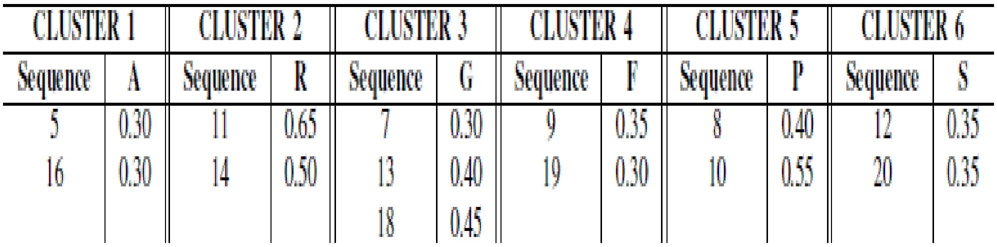
TABLE IIICALCULATED CC OF SEQUENCES



Sequence	A	Sequence	R	Sequence	G	Sequence	K	Sequence	F	Sequence	P	Sequence	S
5	0.30	11	0.65	5	0.30	5	0.30	9	0.35	8	0.40	12	0.35
11	0.65	14	0.50	7	0.30	16	0.30	10	0.35	10	0.55	20	0.35
14	0.45			11	0.45			19	0.30	13	0.40		
16	0.30			13	0.40								
18	0.30			16	0.30								
				18	0.45								

A = Alanine(GCU, GCC, GCA, GCG); R = Arginine(CGU, CGC, CGA, CGG); G = Glycine(GGU, GGC, GGA, GGG);
K = Lysine(AAA, AAG); F = Phenylatanine(UUU, UUC); P = Proline(CCU, CCC, CCA, CCG); S = Serine(UCU, UCC, UCA, UCG)

TABLE IV CLUSTERS GENERATED BASED ON CC VALUES



CLUSTER 1		CLUSTER 2		CLUSTER 3		CLUSTER 4		CLUSTER 5		CLUSTER 6	
Sequence	A	Sequence	R	Sequence	G	Sequence	F	Sequence	P	Sequence	S
5	0.30	11	0.65	7	0.30	9	0.35	8	0.40	12	0.35
16	0.30	14	0.50	13	0.40	19	0.30	10	0.55	20	0.35
				18	0.45						