Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Feature Selection using Relative Reduct hybridized with Improved Harmony Search for Protein Sequence Classification

M. Bagyamathi
Assistant Professor in Computer Science,
Gonzaga College of Arts and Science for Women,
Krishnagiri, Tamil Nadu, India.

Dr. H. Hannah Inbarani
Assistant Professor in Computer Science,
Periyar University,
Salem, Tamil Nadu, India.

***Abstract*—** Recent advances in future generation sequencing technologies have resulted in a tremendous raise in the rate at which protein sequence data are being obtained. Protein sequence analysis is a significant problem in functional genomics. Feature selection techniques are capable of dealing with this high dimensional space of features. In this paper, we propose a feature selection algorithm that combines the Improved Harmony Search algorithm with Rough Set Relative Reduct for Protein sequences for faster and better search capabilities. The feature vectors are extracted from protein sequence database, based on amino acid composition and K-mer patterns or K-tuples and then feature selection is carried out from the extracted feature vectors. The proposed algorithm is compared with Improved Harmony Search hybridized with Rough Set Quick Reduct approach. The experiments are carried out on protein primary single sequence data sets which are derived from PDB on SCOP classification, based on the structural class predictions such as all α, all β, all α+β and all α / β. The feature subset of protein sequences predicted by both existing and proposed algorithms are analyzed with the decision tree classification algorithms.

***Keywords*—** *Data Mining; Bioinformatics; Feature Selection; Protein Sequence; Rough Set; Relative Reduct; Harmony Search; Protein sequence classification.*

I. INTRODUCTION

Feature Selection (FS) is an important part of knowledge discovery. FS is used to improve the classification accuracy and reduce the computational time of classification algorithms [2]. FS is divided into the supervised and unsupervised categories. When class labels of the data are known, supervised feature selection can be applied, otherwise unsupervised feature selection is appropriate.Proteins play a fundamental role in all living organisms and are involved in a variety of molecular functions and biological processes [4]. Proteins are composed of one or more chains of amino acids and show several levels of structure. In fact, according to their chain folding pattern, proteins are usually folded into four structural classes such as all α, all β, all α + β and all α / β [3]. In this paper, the features are extracted from protein primary sequence, based on amino acid composition and K-mer patterns or K-tuples [1].

The rest of the paper is structured as sections II to VI. Section II describes the proposed framework, Section III specifies about the feature extraction method from protein sequences, Section IVdescribes the feature selection algorithms. The experimental analysis with the results and discussion were described in Section V and the paper concludes with future work in this area in Section VI.

II. THE PROPOSED FRAMEWORK

There are several strategies available for classifying the protein sequences. The proposed model predicts the optimal number of features that improves the classification performance. In this study, the protein primary sequences are collected from Protein Data Bank (PDB) in fasta format [3]. The fasta sequence file is used as input data to the PseAAC-builder, a Web server, that constructs the protein feature space using amino acid composition and amino acid K- tuples or K-mer patterns [5].The generated features are real valued, but the rough set theory best in dealing with discrete values. Hence the real valued data are to be discretized. The discretized values are the actual extracted feature set of this study[6]. In the last step, the feature subset predicted by the various feature selection algorithms are evaluated with classification techniques using the WEKA tool [7].

III. FEATURE EXTRACTION

Protein sequences are consecutive amino acid residues, and we regard them as text strings with an alphabet A of size |A| = 20. Many feature extraction methods have been developed in the past several years. Typically, these methods can be classified into two categories. One is based on amino acid composition [1]. The other one is an extension of the atomic length from only one amino acid to K amino acid tuple, where K is an integer and larger than one. We refer to it as 'K-tuple', such as 2-tuple in [10].

In this paper, the features are extracted from protein primary sequence, based on both amino acid composition and K-mer patterns or K-tuples [1]. In Rough set method, the decision table is constructed for dimensionality reduction, which consists of conditional attributes and decision attributes, A = (U, A U {d}) [9]. The features extracted from protein primary sequence are considered as conditional attributes. In this paper, conditional attributes set A consists of K-mer patterns or K-tuples of compositional values of the 20 amino acid in protein primary sequences. The four structural classes such as all α, all β, all α + β and all α / β are considered as decision attribute d as shown in Table 1.

The protein feature vector constructed using amino acids composition that represents a simple sequence that is widely used in prediction of various structural aspects. When K=1,the features are constructed from 20 amino acids A, C,D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y which are denoted as A1, A2,…, A19, and A20, and the number of occurrences of A_i in the sequence that is denoted as x_i, the composition vector is defined as (x_1/L, x_2/L,…, x_i/L), where L is the length of the sequence [13,15]. However, the composition vector is insufficient to represent a sequence, since it only counts the frequencies of individual amino acids. Therefore, along with the 1-tuple feature set, 2-tuple features (when K=2) are constructed to generate the frequencies of amino acid pairs (dipeptides) which provide more information since they reflect interaction between local amino acid pairs. Based on the frequency of collocation of amino acid pairs in the sequence, all dipeptides in the sequence can be counted.

Since there are 400 possible dipeptides (AA, AC, AD,…,YY), a feature vector of that size is used to represent occurrence of these pairs in the sequence[14]. As a result, we propose representation that includes total of 400 + 20 + 1= 421 features (420 conditional attributes and 1 decision attribute) that can be shown in Table 1.

IV. FEATURE SELECTION

A. *Basics of Rough Set Theory*

Rough Set Theory (RST) has been used as a tool to discover data dependencies and to diminish the number of attributes contained in a dataset using the data alone, requiring no additional information [18].

Let *I= (U, A U {d})* be an information system, where *U* is the universe with a non-empty set of finite objects, A is a non-empty finite set of conditional attributes, and d is the decision attribute (decision table),$\forall a \in A$, there is a corresponding function f_a: $U \rightarrow V_a$, where V_a is the set of values of *a* [20]. If$P \subseteq A$, there is an associated equivalence relation:

$$IND\ P = \{ x,y \in UXU \mid \forall a \in P,\ f_a(x) = f_a(y)\} \quad (1)$$

The partition of *U* generated by *IND(P)* is denoted *U/P*. If $x,y \in IND(P)$, then x and y are

Table 1: Decision Table (amino acid composition of 2-tuple feature vector

Object	A	C	.	Y	AA	AC	.	YY	Class
1	5.42	1.81	.	2.71	5.42	8.13	.	8.13	1
2	6.15	1.54	.	1.54	5.38	6.92	.	7.69	1
3	12.5	0	.	0	4.69	8.59	.	8.59	2
4	8.57	1.43	.	1.43	10	2.86	.	11.43	2
5	12.5	0	.	0	4.69	8.59	.	8.59	3
6	8.96	1.49	.	1.49	10.45	2.99	.	10.45	3
7	12.5	0	.	0	4.69	8.59	.	8.59	4
8	8.7	1.45	.	1.45	10.14	2.9	.	11.59	4

indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted as $[x]_P$. Let $X \subseteq U$, the P-lower approximation $\underline{P}X$ and P-upper approximation $\overline{P}X$ of set X can be defined as:

$$\underline{P}X = \{x \in U \mid x_P \subseteq X\} \quad (2)$$

$$\overline{P}X = \{x \in U \mid x_P \cap X \neq \varphi\} \quad (3)$$

Let $P, Q \subseteq A$be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_P(Q) = \bigcup_{x \in U|Q} \underline{P}X \quad (4)$$

$$NEG_P(Q) = U - \bigcup_{x \in U|Q} \overline{P}X \quad (5)$$

$$BND_P(Q) = \bigcup_{x \in U|Q} \overline{P}X - \bigcup_{x \in U|Q} \overline{P}X \quad (6)$$

The positive region of the partition *U/Q* with respect to *P*, $POS_P(Q)$, is the set of all objects of *U* that can be certainly classified to blocks of the partition *U/Q* by means of *P*. *Q* depends on *P* in a degree k $(0 \leq k \leq 1)$ denoted by $P \Rightarrow_k Q$

$$K = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \quad (7)$$

Where *P* is a set of all conditional attributes, *Q* is the decision attributes, and γ_P Q is the quality of classification. If *k*=1, *Q* depends totally on *P*; if $0 < k < 1$, *Q* depends partially on *P*; and if *k*=0 then *Q* does not depend on *P*. The goal of attribute reduction is to remove redundant attributes so that the reduced set provides the same quality of classification as the original [19]. The set of all reducts is defined as:

$$Red\ C = \{R \subseteq C \mid \gamma_R\ D = \gamma_C\ D\ , \forall B \subset R, \gamma_B\ D \neq \gamma_C\ D\} \quad (8)$$

B. *Rough Set Improved Harmony Search Quick Reduct (RS-IHS-QR) Algorithm*

Harmony search (HS) is a relatively new population-based meta heuristic optimization algorithm, that imitates the music improvisation process where the musicians improvise their instruments' pitch by searching for a perfect state of harmony. It was able to attract many researchers to develop HS based solutions for many optimization problems [16]. This method HSA is developed by Mahdavi et al. 2007 [17]. In HSA , HMCR, PAR, bw, but PAR and bw are very important parameters in fine-tuning of optimized solution vectors. The traditional HS algorithm uses fixed value for both PAR and bw. In the HS method, PAR and bw values are adjusted in Step 1 and cannot be changed during new generations [22]. The main drawback of this method is that the number of iterations increases to find an optimal solution. To improve the performance of the HS algorithm and to eliminate the drawbacks that lies with fixed values of PAR and bw, IHSA uses variables PAR and bw in improvisation step (Step 3) [21]. PAR and bw change dynamically with generation number and expressed as follows:

$$PAR(gn) = PAR_{min} + \frac{PAR_{max} - PAR_{min}}{NI} * gn \quad (9)$$

Where, *PAR(gn)* = Pitch Adjusting Rate for each generation PAR_{min} = Minimum Pitch Adjusting Rate, PAR_{max} = Maximum Pitch Adjusting Rate, *NI* = Number of Improvisations and *gn* = Generation Number

$$bw(gn) = bw_{max} * exp(c * gn);$$

$$c = ln\ [(bw_{min} / bw_{max})] / NI \quad (10)$$

Where, *bw(gn)* = Bandwidth for each generation bw_{min} = Minimum bandwidth bw_{max} = Maximum bandwidth.

The Pseudocode of the Improved HS using Rough Set based Quick Reduct is given in [6].

C. *Rough Set Improved Harmony Search Relative Reduct (RS-IHS-RR) Algorithm*

An Improved HS algorithm was developed by Mahdavi et al. 2007 [11], [17]. In the proposed algorithm, HMCR, PAR, bw parameters are considered, but PAR and bw are very important parameters in fine-tuning of optimized solution vectors. The traditional HS algorithm uses a fixed value for both PAR and bw. In the HS method, PAR and bw values are adjusted and cannot be changed during new generations [12]. The main drawback of this method is that the numbers of iterations increases to find an optimal solution. To improve the performance of the HS algorithm and to eliminate the drawbacks that lies with fixed values of PAR and bw, IHS algorithm uses variables PAR and bw in improvisation step (Step 3) [6].The pitch adjustment rate parameter causes a musician to selecta value neighboring to its current choice. To achieve better results, we define lower limit andupper limit of the PAR and bw. PAR and bw change dynamically with generation number and expressed as follows:

$$PAR(gn) = PAR_{min} + \frac{PAR_{max} - PAR_{min}}{NI} * gn \quad (11)$$

$$bw(gn) = bw_{max} * exp(c * gn);$$

$$c = ln\ [(bw_{min} / bw_{max})] / NI \quad (12)$$

The Pseudocode of the RS-IHS-RR algorithm is given in [23].

V. EXPERIMENTAL ANALYSIS

A. Data Source

In this paper, the protein primary sequence datasets are derived from PDB (http://www.rcsb.org/pdb) on SCOP classification. The Structural Classification of Proteins (SCOP) database is largely a manual classification of protein structural domains based on similarities of their structures and amino acidsequences[8]. The data set consists of sequences with 7623 of all α, 10672 of all β, 11048 of all α + β and 11961 of all α / β [3]. Among one thousand sequences with combinations of all α , all β, all α + β and all α / β, each 250 sequences are taken for this study.

B. Results

A sequence of experiments was conducted to show the efficacy of proposed feature selection algorithm. All experiments have been run on a machine with 3.0 GHz CPU and 2 GB of RAM. We implement proposed Improved Harmony Search hybridized with

Table 2: Number of features selected by feature selection algorithms

Protein Data Set	*Number of features extracted using K-tuple sequences*			*Number of features selected*	
	K	*Number of Conditional features*	*Number of decision features*	*Improved HSQR*	*Improved HSRR*
1000 objects	1	20^1	20	1	11	8
	2	20^2	420	1	32	26

Rough Set Relative Reduct feature selection algorithm in Matlab 2012a. The operating system is Windows Vista. For experimental studies, we have considered 1000 objects from SCOP classification of Protein Data Bank. The following section describes the implementation results of this study.

The feature subset length and the classification quality are the two criteria that are considered to assess the performance of algorithms. Comparing the first criterion, number of selected features shown in Table 2, the proposed algorithm Improved Harmony Search Relative Reduct outperforms the Improved Harmony Search Quick Reduct algorithm in selecting smaller subset of features in both the tuples which is compared in Fig. 1.

Second, we compare the other criterion, predictive accuracy. The proposed algorithm Improved HS Relative Reduct algorithms revealed best accuracy than

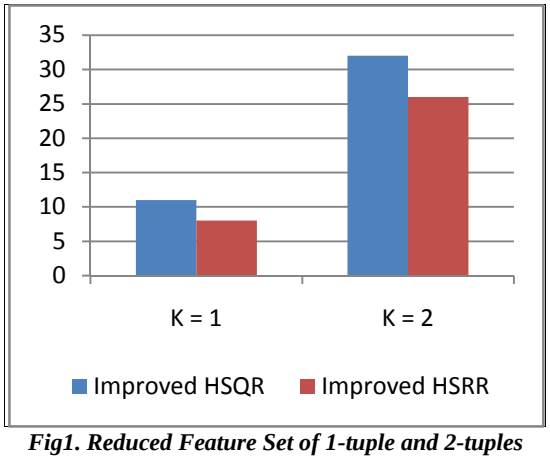

Fig1. Reduced Feature Set of 1-tuple and 2-tuples

Table 3: Classification accuracy of 1-tuple (K=1) protein sequence features

Classification Method	*Predictive Accuracy(%)*	
	Improved HSQR	*Improved HSRR*
IBK	91.1	91.5
Kstar	90.0	91.2
Randomforest	88.5	88.9
J48	81.3	82.5
JRip	78.3	77.5

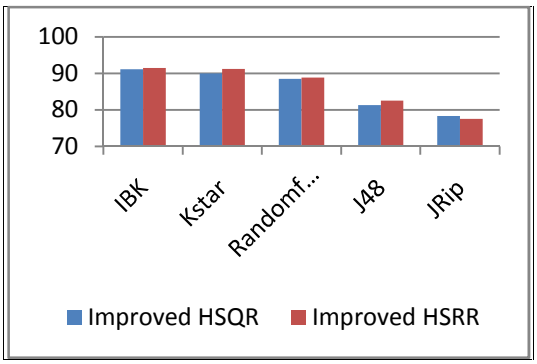

Fig 2. Classification accuracy of 1- tuple (k=1) protein sequence features

an existing algorithm Improved HS Quick Reduct algorithm. The predictive accuracy results of the existing and proposed algorithms of 1-tuple and 2-tuples are compared in Table 3 and 4. Figs. 4 – 8 show the predictive accuracy for each of the feature selection algorithms considered in this study.

Table 4: Classification accuracy of 2-tuple (K=2) protein sequence features

Classification Method	*Predictive Accuracy (%)*	
	Improved HSQR	*Improved HSRR*
IBK	91.3	92.0
Kstar	90.8	91.5
Randomforest	91.7	91.9
J48	90.5	92.2
JRip	91.8	92.4

C. Discussion

Experimental results show that the use of irrelevant features hurts classification accuracy and Feature Selection technique is used to reduce redundancy in the information provided by the selected features. Using only a small subset of selected features, the proposed Improved HS Relative Reduct algorithms

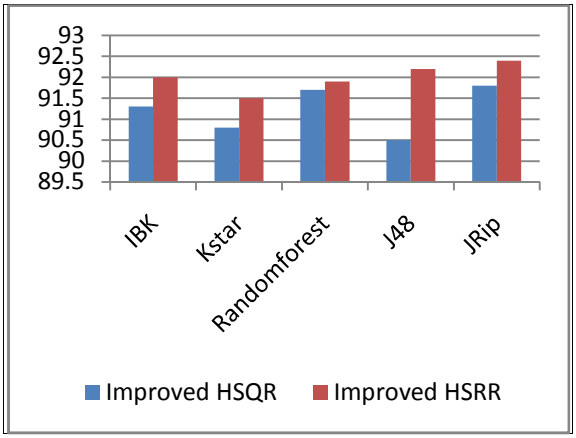

Fig 3. Classification accuracy of 2- tuple (k=2) protein sequence features

obtained better classification accuracy than the existing algorithm compared in this study.To compare the performance of the above feature selection algorithms, classification techniques such as IBK, Kstar, Randomforest, J48 and JRip are applied in this work.Theselected feature subset of protein sequences are used as the input of the classifiers. All experiments were carried out using a ten-fold cross validation approach.

The results strongly suggest that proposed method can assist in solving the high-dimensionality problem, and accurately classifies the protein sequences to its corresponding structures and can be very useful for predicting the function of the protein. The proposed algorithm outperforms an existing algorithm with all the classifiers considered in this study, which is shown in Fig. 2 and Fig. 3.

VI. CONCLUSION

The aim of this study is to reduce the dataset by eliminating the irrelevant features. In this work, the rough set theory is hybridized with the Improved Harmony Search Relative Reduct algorithm to classify the protein sequences. The RS-IHS-RR algorithm has a number of advantages over the existing algorithm. The reduced and relevant features help in classifying the protein sequences very efficiently. The experimental results show that how the meta-heuristics approaches increases the predictive accuracy for the given dataset. Hence the analysis section clearly proved the efficiency and effectiveness of Harmony Search and RST based approaches. As a future work, this model can also be extended to hybridize advanced swarm intelligence techniques such as bees colony optimization, fish swarm, cuckoo search optimization etc.

References

[1] Chandran C.P,"Feature Selection from Protein Primary Sequence Database using Enhanced Quick Reduct Fuzzy-Rough Set", In: International Conference on Granular Computing , GrC 2008, Hangzhou, China, pp. 111-114,doi: 10.1109/GRC.2008.4664758.

[2] Chandrasekhar T, Thangavel K, and Sathishkumar E.N. "Verdict Accuracy of Quick Reduct Algorithm using Clustering and Classification Techniques for Gene Expression Data", IJCSI International Journal of Computer Science Issues, 2012, Vol. 9, No. 1, pp. 357-363.

[3] Cao Y, Liu S, Zhang L, Qin J, Wang J, and Tang K,"Prediction of protein structural class with Rough Sets, BMC Bioinformatics, 2006, Vol. 7, No. 1, pp. 20. doi:10.1186/1471-2105-7-20.

[4] Nemati S, Basiri ME, Ghasem-Aghaee N, and Aghdam MH,"A novel ACO–GA hybrid algorithm for feature selection in protein function prediction", Expert Systems with Applications, 2009, Vol. 36, No. 10, pp. 12086–12094.

[5] Du P, Wang X, Xu C, and Gao Y, "PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions", Analytical Biochemistry, 2012, Vol. 425, No. 2, pp. 117-119.

[6] Bagyamathi M and Inbarani HH, "A Novel Hybridized Rough Set and Improved Harmony Search Based Feature Selection for Protein Sequence Classification", Big Data in Complex Systems: Challenges and Opportunities, Studies in Big Data, 2015, Vol. 9, pp. 173 - 204, Springer-Verlag.

[7] Hall M, Frank E, Holmes G, Pfahringer G, Reutemann P, and Witten I.H, "The WEKA data mining software: an update", ACM SIGKDD Explorations Newsletter, 2009, Vol. 11, No. 1, pp. 10-18, doi:10.1145/1656274.1656278

[8] Chinnasamy A, Sung W.K, and Mittal A, "Protein Structure and Fold Prediction Using Tree-Augmented Bayesian Classifier", Journal of Bioinformatics and Computational Biology, 2005, Vol. 3, No. 4, pp. 803. doi: 10.1142/S0219720005001302.

[9] Pawlak Z, "Rough Sets: Present State and The Future, Foundations of Computing and Decision Sciences", 1993, Vol. 18, No. 3-4, pp. 157–166.

[10] Park K.J, and Kanehisa M, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs", Bioinformatics, 2003, Vol. 19, No. 13, pp. 1656–1663.

[11] Geem ZW, "Improved harmony search from ensemble of music players", In proceedings of 10th International Conference on Knowledge-based intelligent information and engineering systems – KES 2006, LNCS 4251, pp. 86–93. Springer Heidelberg.

[12] Al-Betar M, Khader A and Liao I, "A harmony search with multi-pitch adjusting rate for the university course timetabling", In Geem Z (ed) Recent advances in Harmony search algorithm, 2010, Springer Berlin Heidelberg, Vol. 270, pp. 147–161.

[13] Chen C, Tian YX, Zou XY, Cai PX, and Mo JY, "Using pseudoamino acid composition and support vector machine to predict protein structural class",Journal of theoretical biology, Elsevier, 2006, Vol. 243, No. 3, pp. 444–448.

[14] Shi JY, Zhang SW, Pan Q, Cheng YM, and Xie J, "Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition", Amino Acids 2007, Vol. 33, No. 1, pp. 69–74.

[15] Gu Q, Ding Y, Jiang X, and Zhang T, "Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection", Amino Acids, 2010, Vol. 38, No. 4, pp. 975–983. Springer-Verlag.

[16] Geem ZW, "Improved harmony search from ensemble of music players", In proceedings of 10th International Conference on Knowledge-based intelligent information and engineering systems – KES 2006, LNCS 4251, pp. 86–93, Springer Heidelberg. doi:10.1007/11892960_11.

[17] Mahdavi M, Fesanghary M, and Damangir E, "An improved harmony search algorithm for solving optimization problems", Applied Mathematics and Computation, 2007, Vol. 188, No. 2, pp. 1567–1579.

[18]Anaraki JR and Eftekhari M, "Rough set based feature selection: A Review", Fifth Conference on Information and Knowledge Technology (IKT), 28-30 May 2013, pp. 301-306. IEEE.

[19] Pawlak Z, "Rough Sets and Intelligent Data Analysis", Information Sciences, 2002, Vol. 147, No. 1-4, pp. 1–12.

[20] Velayutham C and Thangavel K, " Unsupervised Quick Reduct Algorithm Using Rough Set Theory", Journal of Electronic Science and Technology, 2011, Vol. 9, No. 3, pp. 193 – 201.

[21] Chakraborty P, Roy GG, Das S, Jain D, and Abraham A, "An improved harmony search algorithm with differential mutation operator", FundamentaInformaticae, 2009, Vol. 95, No. 4, pp. 1–26. doi:10.3233/FI-2009-181.

[22] Al-Betar M, Khader A, and Liao I, "A harmony search with multi-pitch adjusting rate for the university course timetabling", In Geem Z (ed) Recent advances in Harmony search algorithm, 2010, Springer-Verlag, Berlin, Heidelberg, pp. 147–161.

[23] Bagyamathi M and Inbarani HH, "Feature Selection using Improved Harmony Search Hybridized with Relative Reduct for Medical Data Classification", International Journal of Applied Engineering Research (IJAER), 2015, Vol. 10, No. 20, pp. 19476-19480.

[24]Inbarani HH, Bagyamathi M, Azar AT, "A novel hybrid feature selection method based on rough set and improved harmony search", Neural Computing and Applications, 2015, pp.1-22. Springer London.