Journal of Theoretical and Applied Information Technology





www.jatit.org

# ROUGH SET PROTEIN CLASSIFIER

**RAMADEVI YELLASIRI[1], C.R.RAO[2]**

[1]Dept. of CSE, Chaitanya Bharathi Institute of Technology, Hyderabad, INDIA.

[2]DCIS, School of MCIS, University of Hyderabad, Hyderabad, INDIA.

Email: rdyellasiri@yahoo.co.in

## ABSTRACT

Classification of voluminous protein data based on the structural and functional properties is a challenging task for researchers in bioinformatics field. In this paper a faster, accurate and efficient classification tool Rough Set Protein Classifier has been developed which has a classification accuracy of 97.7%. It is a hybridized tool comprising Sequence Arithmetic, Rough Set Theory and Concept Lattice. It reduces the domain search space to 9% without losing the potentiality of classification of proteins.

**Keywords:** *Sequence Arithmetic, Rough Set Theory, Concept Lattice, Neighborhood Analysis*

## 1. INTRODUCTION

In the recent times, a large amount of new protein sequences are progressively being accumulated in various databases. The significant task for researchers in bioinformatics is to classify these proteins into families based on their structural and functional properties, thereby predicting the functions of these new protein sequences. Protein sequences are of varying length comprising of twenty Amino Acids (AA). Proteins are categorized into families, families into classes and classes into subclasses based on their functional and structural properties. The composition and frequency of occurrence of AA and Neighborhood Analysis of AA is not addressed in existing classification algorithms.

In the present paper, protein hierarchy (family, class, and subclass) is regarded and level-by-level reduction of domain search space is emphasized. An innovative technique viz., Sequence Arithmetic (SA) to identify family information and utilize it for reducing the domain search space is proposed. Rules are generated and stored in Sequence Arithmetic database. A new approach to compute predominant attributes (approximate reducts) and use them to construct decision tree called Reduct based Decision Tree (RDT) is proposed. Decision rules generated from the RDT are stored in RDT Rules Database (RDTRD). These rules are used to obtain class information. The infirmity of RDT is overcome by extracting spatial information by means of Neighborhood Analysis (NA). Spatial information is converted into binary information using threshold. It is utilized for the construction of Concept Lattice (CL). The Associated Rules from the CL are stored in Concept lattice Association Rule Database (CARD). Further, the domain search space is confined to a set of sequences within a class by using these Association Rules.

The Rough Set Protein Classifier (RSPC) is the integration of Sequence Arithmetic, Reduct based Decision Tree, Neighborhood Analysis and Concept Lattice modules. RSPC accelerates the process of classification/identification by reducing the domain search space stage by stage. Myoglobin and G-Protein Coupled Receptors (GPCR) protein sequences are considered for experimentation.

A brief introduction to proteins and classification algorithms is given in Section 2. RSPC is introduced, followed by the description of modules in Section 3. Section 4 gives the complexity of the modules. The performance evaluation of RSPC is given in Section 5. The paper is concluded in Section 6.

## 2. PROTEIN CLASSIFICATION ALGORITHMS

Proteins are building blocks of cells and tissues; they play a vital role in executing and regulating many biological processes. AA are the building blocks of all proteins made up of twenty (+2 for

derived AA) Amino Acids. Protein sequences are of varying length with an average of 350 AA.

PROSITE, PRINTS and PFAM are protein classification methods and they utilize multiple sequence alignments to build Hidden Markov Models. PROSITE is a database consisting of information of significant sites, patterns, and profiles that specify different protein families [1]. PRINTS is a database of protein fingerprints [2]. Fingerprints are sets of short sequence motifs (patterns) conserved among members of a protein family. PFAM is a database of alignments and profile-Hidden Markov Models of protein families [3]. They can predict protein functions only if sequences are sufficiently conserved. Although sequences share some structural similarities, these methods may fail, lest there is an inadequate sequence similarity. In general, they require large amount of time for building models as well as for predicting functions based on them. Precisely, there is an acute need for development of faster, more sensitive and accurate methods to meet the present challenges of classification. The following section discusses about the new hybridized technique RSPC and its different modules.

Protein classification algorithms use multiple sequence alignment, profiles, motifs, short sequences etc., [4]. Profiles and motifs need expert knowledge for attributing structural or functional aspects of a protein. It is possible to derive results through the existing methods based on profiles and motifs with the aid of a good expert system. A motif is a consecutive string of AA which frequently occur in a given protein sequence whose general character is repeated, or conserved, in all sequences in a multiple alignment at a particular position.

## 3. DESIGN OF ROUGH SET PROTEIN CLASSIFIER

Rough Set Protein Classifier consists of four modules: i) Sequence Arithmetic Module, ii) RDT Module iii) Neighborhood Association Module and iv) Concept Lattice Module. The input to the RSPC is the query sequence ‘y’ and output is the set of sequences to which ‘y’ is closely related as shown in Figure 1.

### 3.1 Sequence Arithmetic (SA) Module

It is based on set theory. The occurrence and frequencies of AA are considered and three characteristics sets such as A-Set, D-Set and P-Set are defined for protein sequences. They are also generated for classes of proteins and family of proteins. This information is stored in the Sequence Arithmetic Knowledge Database, family information is extracted by mapping these sets to the information present in the Sequence Arithmetic Knowledge Database as shown in Figure 2. The output of the Sequence Arithmetic Module is family information < F >. Therefore, the search space is confined to only < F >, thus reduction in the domain search space is inversely proportional to **‘f’** (number of families under study).

### 3.2 Reduct based Decision Tree (RDT) Module

A-Table, D-Table and P-Table are generated from the corresponding characteristics sets and they are used to obtain class information within a family. A new approach is proposed to obtain the Rank ordered Predominant Attributes from the decision table (P-Set or A-Set or D-Set with a ClassID (decision attribute concatenated as the first column) and a decision tree is created based on these Predominant Attributes [5]. These Predominant Attributes are known as Reduct. Rules are generated from the decision tree and stored in the Reduct based Decision Tree Rules Database. The decision tree for corresponding family will help in classifying the queried protein into a class or set of classes respectively. The infirmity of this module is that it cannot be applied if all the entries of the table are one, therefore neighborhood associations has to be studied.

### 3.3 Neighborhood Association Module

Localized information is extracted through neighborhood analysis in this module. For varying distance ‘d’, Neighborhood Association Matrix for each AA with least frequent occurring AA as centre is computed. It is used to construct a Binary Association Matrix by considering binarization threshold ‘T’. Experimental results recommended ‘T’ as 0.05 and neighborhood distance‘d’ as 5.

Predominant Attribute set is derived using Reduct based Decision Tree module by attributing class information to the Binary Association Matrix and treating it as a decision table. Predominant Attributes are used to construct decision tree for identifying subclass. The decision tree based on the Binary Association Matrix for a family or for a class will classify the questioned protein into a subclass. The size of the subclass may not be of manageable size. Thus, Binary Association Matrix is further used for constructing concept lattice and

hence, attaching the given unknown protein to a set of proteins.

### 3.4 Concept Lattice Module

The number of proteins to which the questioned protein needs to be compared will not be small by this stage. To further reduce the complexity of comparisons string by string, concept lattice module [6] is adapted.

The concept lattice is constructed based on the Binary Association Matrix. Concept lattice Association Rules are generated from the node information of the concept lattice. They are stored in the Concept lattice Association Rules Database along with familyID and ClassID. The discovered family / class / subclass information for the queried protein is compared with few similar proteins given by the set of proteins in the Concept Lattice Association Rules Database.

The derived information of the questioned protein is processed through the above modules to classify the protein, resulting in the reduction of search space. For classification of family / class / subclass, string comparisons are avoided.

## 4. COMPLEXITY OF DIFFERENT MODULES

Performance evaluation of developed modules is discussed in this section. Compositions and frequency occurrence of AA were studied while creating databases. Horizontal fragmentation of the domain search space (in levels) was achieved while traversing from one module to another following protein hierarchy. Spatial information was obtained by performing neighborhood analysis.

### 4.1 Creation of Different Databases

Sequence Arithmetic Knowledge Database was created based on the A-Set, D-Set and P-Set generated for the datasets. The Sequence Arithmetic rules were derived and stored in the Sequence Arithmetic Knowledge Database.

RDT Rules were generated for different classes of the datasets and stored in RDT Rules Database. RDT Rules were used to obtain the class information. Rules for Myoglobin were constructed based on the P-Table generated in Reduct based Decision Tree module, whereas for G-Protein Coupled Receptor, neighborhood associations was performed (as only one set is generated).

Concept lattice Association Rules were generated separately for each class with different Least Frequent occurring Character as centers. These rules are stored in Concept lattice Association Rules Database. The creation of the database was an offline process.

### 4.2 Time Complexity (Queried sequence)

For an unknown sequence ‘y’ of size ‘n’, a single scan is required for obtaining the A(y) and D(y). If the number of families present in the database is ‘f’ then the number of comparisons will be O(f). Therefore the worst case complexity was of the order of O(n) + O(f) .

Let ‘C’ be the number of classes in a given family then complexity of RDT will be O(log C). At the end of SA and RDT module the complexity is O(n) + O(f) + O(log C). The complexity of Neighborhood Analysis module followed by RDT module is O(22n) + O(log C).

If ‘r’ is the number of proteins in classes then the CL will have $2^r$ nodes. Therefore the number of comparisons required will be $O(2^r)$. Thus, the complexity is O(n) + O(f) + O(log C) + $O(2^r)$.

The empirical analysis concludes n ≈ 400, f ≈ 1.5, C ≈ 27 to 30 and r ≈ 8 (on an average).

## 5. PERFORMANCE EVALUATION

The performance of various modules and RSPC in terms of time and space is discussed in the subsections to follow.

### 5.1 Sequence Arithmetic Module

The complexity of the Sequence Arithmetic module depends on A-Set, D-Set and P-Set and the mode of storing them in Sequence Arithmetic Knowledge Database.

For a family $T_{SA}$ = O(nm) + O(3n*22) = O(nm).

Let ‘$\tau$’ be the complexity of searching for a string in a database with ‘N’ sequences using BLAST [7]. Family information is obtained by applying Sequence Arithmetic module; let the number of sequences be $N_1$ in the family. Therefore, $N_1 < N$.

With SA it will be $T_{SA}$ + O($N_1 \tau$) which will be ≤ O(N $\tau$). Therefore the complexity of Sequence Arithmetic is O($N_1 \tau$) + O(nm).

In the absence of Sequence Arithmetic module, a protein has to be with the complexity of $N^{\tau}$ (where $\tau$ is BLAST complexity). This complexity can be reduced by adopting multistage classifiers.

The application of the Sequence Arithmetic Rules to Myoglobin and GPCR families showed that there is 50% reduction in the search space.

### 5.2 Reduct Based Decision Tree Module

The performance evaluation of the Reduct based Decision Tree algorithm with the existing classification technique (WEKA tool [8] is used for experimentation of other methods and Reduct based Decision Tree was constructed according to the procedure) on the standard databases was performed and the results are shown in Table 1. It is observed that Reduct based Decision Tree performed better than other methods in terms of Kappa statistics [5]. In case of large dataset (GPCR), Reduct based Decision Tree (RDT) and RandomForest are equally efficient.

**Table 1: Comparison of RDT with other Classification Methods**

| | Kappa Statistics | | |
|---|---|---|---|
| **Classification Technique** | **Sunburn** | **Weather** | **GPCR** |
| Bayes Network | 0 | 0 | 0.306 |
| ComplementNaiveBayes | 0.142 | 0 | 0.29 |
| NaiveBayesMultinomial | 0 | 0 | 0.013 |
| Logistic | 0 | 0.588 | 0.362 |
| RBFNetwork | 0 | 0.256 | 0.416 |
| SimpleLogistic | 0 | 0 | 0.359 |
| SMO | 0 | 0 | 0.338 |
| BFTree | 0 | 0 | 0.574 |
| J48 | 0 | 0.143 | 0.582 |
| J48graft | 0 | 0.143 | 0.585 |
| LMT | 0 | 0 | 0.556 |
| NBTree | 0 | 0 | 0.445 |
| RandomForest | 0 | 0.429 | 0.652 |
| RandomTree | 0 | 0.378 | 0.595 |
| SimpleCart | 0 | 0 | 0.572 |
| RDT | 0.5 | 0.58 | 0.62 |

Number of rules generated with Reduct based Decision Tree (RDT) and Decision Tree (ID3) as a result of five-fold test is given in Table 2. It shows that Reduct based Decision Tree performed better than ID3.

**Table 2: Results of Five-fold test of GPCR dataset**

| | Classified | | Misclassified | |
|---|---|---|---|---|
| **Set** | **RDT** | **ID3** | **RDT** | **ID3** |
| Set 1 | 80% | 78% | 20% | 22% |
| Set 2 | 80% | 81% | 20% | 19% |
| Set 3 | 82% | 80% | 18% | 20% |
| Set 4 | 84% | 84% | 16% | 16% |
| Set 5 | 82% | 80% | 18% | 20% |

### 5.3 Cross-validation of RSPC

The dataset with 3896 GPCR sequences and available 25 non-GPCR sequences were tested. The results comprise of True Positives (TP)-3896, False Positives (FP)-6, True Negatives (TN)-19 and False Negatives (FN)-NIL. Therefore, the accuracy was 99.85%.

Ten different test datasets consisting of 200 sequences from GPCR (selected randomly) and 24 non-GPCR sequences were added to each set thus making the set size of 224 sequences for testing. The average results were TP = 200, FP = 5, TN = 19 and FN = 0. Thus the average accuracy rate was 97.7% [9]. The accuracy rate of Support Vector Machines (SVM) [10] was compared with RSPC as reported in Table 3.

**Table 3 Accuracy rate of the cross-validation for classifying the GPCR**

| Classification Methods | | | | | | |
|---|---|---|---|---|---|---|
| **SVM-lin** | **SVM-Poly** | **SVM-Sig** | **SVM-rbf** | **SVM-pw** | **SVM-Fish** | **RSPC** |
| 0.905 | 0.937 | 0.897 | 0.97 | 0.99 | 0.992 | 0.977 |

SVM-lin : SVM with linear function.
SVM-Poly : SVM with Polynomial function.
SVM-Sig : SVM with Sigmoid function.
SVM-rbf : SVM with radial basis function.
SVM-pw : SVM with pair wise alignment.
SVM-Fish : SVM with Fisher approach.
RSPC : Rough Set Protein classifier.

### 5.4 Analysis of Rough Set Protein Classifier

Samples of 60 sequences out of 1440 sequences were randomly selected and subjected to the RSPCC. In the Sequence Arithmetic, it is observed that a sequence has to be compared with 885 sequences on an average instead of 1440, thus a reduction of about 40% is observed. The domain search space is confined to about 60%.

In the Reduct based Decision Tree module and Neighborhood Analysis module, the search is confined to the families obtained in the sequence arithmetic module, resulting in reduction of the search proteins set to 530. The average numbers of classes were reduced from 36 to 25 classes. Observation shows that one has to search 530 sequences instead of 1440 sequences, thus the search space is further confined to 37%.

The domain search space was reduced by 63% at the end of RDT module, and the average number of concept lattices to be compared was 25 (one for each class). Exploiting the power of concept lattice further reduced the search space to 9%. All the search space reduction is expected to preserve the accuracy of identification intact. Hence, the RSPC will help in reducing the search space identification of a queried protein from 100% to 9% (without compromising the potentiality for the identification the protein).

### 5.5 Analysis of Search Space Reduction

The impact of RSPC modules in search space is discussed in the above sections. Presence of the queried protein in the reduced space was carried out. It is observed in all the stages the queried protein is one among the reduced set, which indicate the confidence of identifying the exact protein is 100% as in Figure 3. In Figure, 3 X-axis indicates the modules and Y-axis the percentage of the search space. The column with maroon colour depicts the confidence of belongingness of queried protein with respect to the respective module. The experimental evidence shows that RSPC reduces the search significantly without loss of identification.

## 6. CONCLUSIONS

The RSPC developed is a hybridized tool which consists of Sequence Arithmetic and Rough Set Theory. The frequencies of AA were considered for classification and RSPC performed better with 97.7% classification accuracy. It reduces the domain search space from 100% to 9% without losing the potentially of classification of proteins.

## REFERENCES


[1]. Hulo N., Sigrist C., Saux V. L., Langendijk-Genevaux P., Bordoli L., Gattiker A., Castro E. D., Bucher P., and Bairoch A.. : "Recent improvements to the PROSITE database". *Nucleic Acids Research*, 32, 2004. pp. 134–137.

[2]. Attwood T., Bradley P, Flower D, Gaulton A, Maudling N, Mitchell A., Moulton G., Nordle A., Paine K., Taylor P., Uddin A., and Zygouri C.: "PRINTS and its automatic supplement". *Nucleic Acids Research*, 31(1), 2003. pp. 400–402.

[3]. Bateman A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. Reddy : "The Pfam protein families database". *Nucleic Acids Research*, 32(1), 2003. pp. D138–D141.

[4]. David Mount: "Bioinformatics: Sequence and Genome Analysis", CSHL Press, 2001.

[5]. Ramadevi Y, C.R.Rao: "Reduct based Decision Tree (RDT)". *International Journal of Computer Science and Engineering Science 2008*.2(4).pp. 245-250.

[6]. Robert Godin, Rokia Missaoui, Hasan Alaoui : "Incremental Concept Formation Algorithm based on Galois Lattices". *Computational Intelligence*, 1995.11(2). pp. 246-267.

[7]. Altschul S. F., Madden T.L., Schäffer A. A.,Zhang J., Zhang Z., Miller W. and Lipman D. J.: "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs". *Nucleic Acids Research* 25, 1997. pp. 3389-3402.

[8]. David Scuse, Peter Reutemann: "WEKA Experimenter Tutorial", Version 3-4, January 26, 2007.

[9]. Ramadevi Y, C.R.Rao: "Decision Tree Induction Using Rough Set Theory-Comparative Study". *Journal of Theoretical and Applied Information Technology*- Vol 3, 2007.

[10]. Pooja Khati: "Comparative Analysis of Protein Classification Methods". *Master's Thesis*, University of Nebraska, Lincoln, December, 2004.

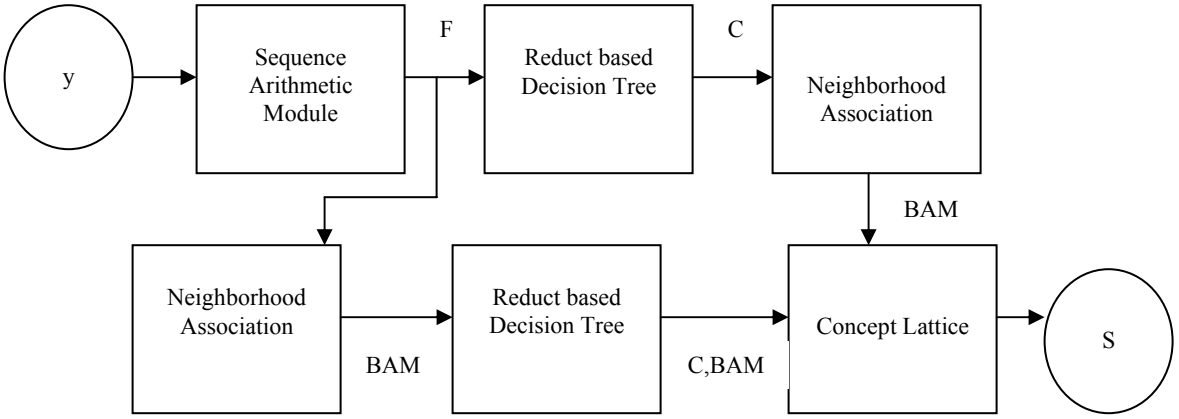


**Figure 1: Rough Set Protein Classifier**

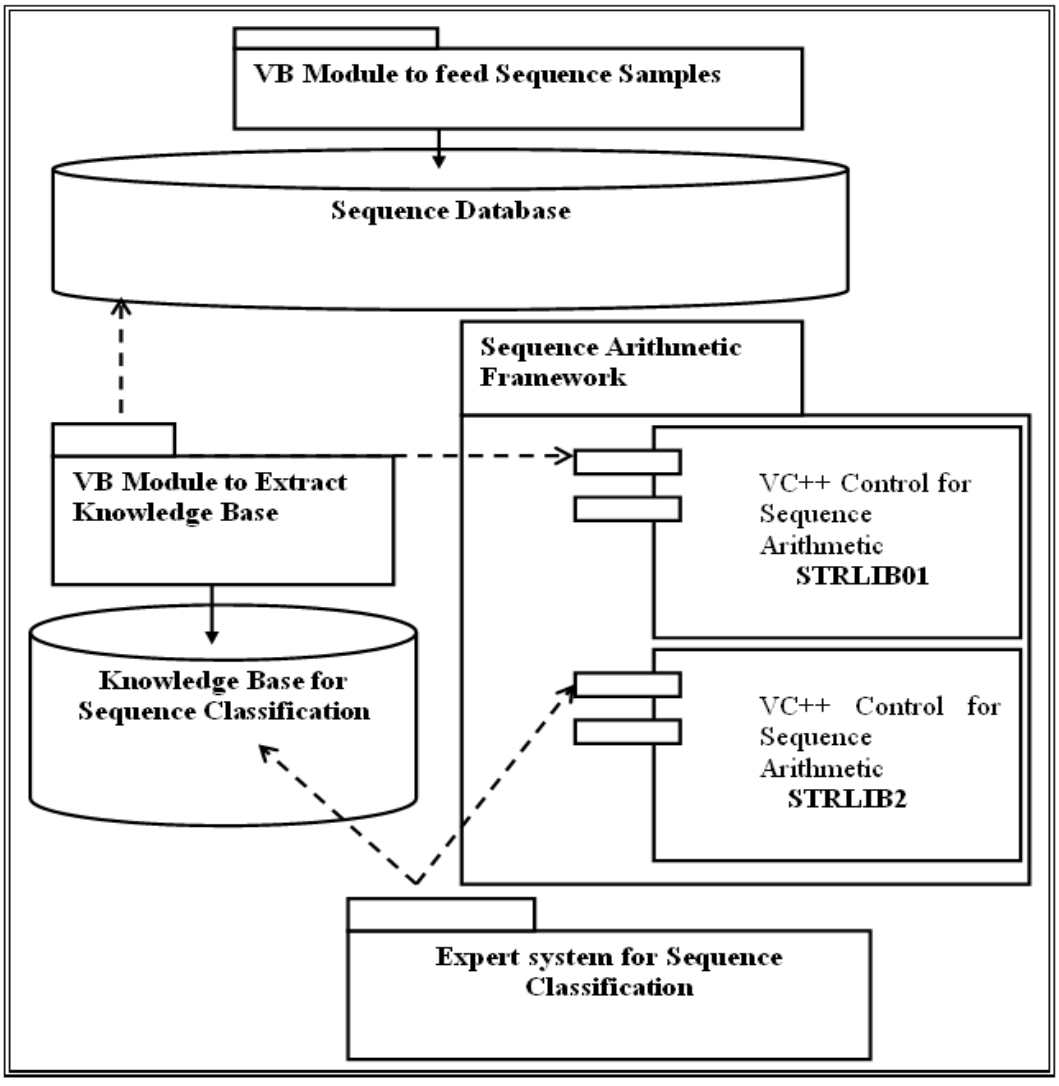


**Figure 2: Sequence Arithmetic Architecture**

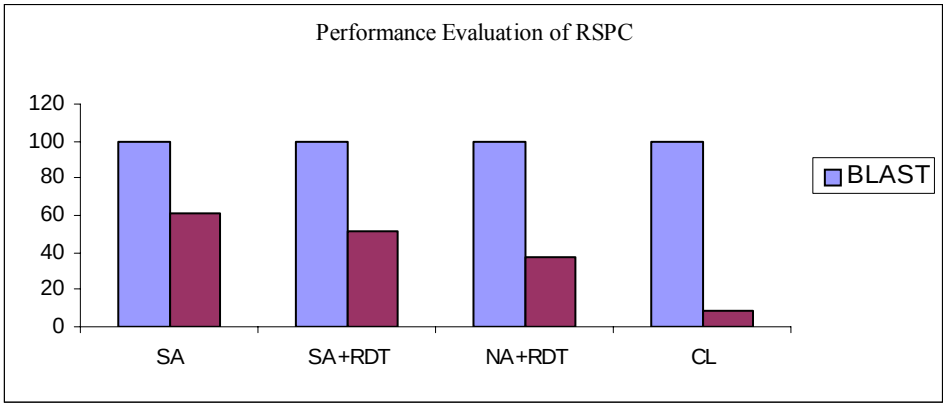


**Figure 3: Performance Evaluation of RSPC**