

# SRPVS: A New Motif Searching Algorithm for Protein Analysis

Xiaolu Huang and Hesham Ali  
Department of Computer Science  
University of Nebraska at Omaha  
Omaha, NE 68182

xhuang@unmc.edu, hesham@unomaha.edu

Anguraj Sadanandam and Rakesh Singh  
Department of Pathology and Microbiology  
University of Nebraska Medical Center  
Omaha, NE 68198  
{asadanandam, rsingh}@unmc.edu

## Abstract

*In some protein sequence regions, when two sequences share similar amino acid composition, they also share the same biological structure regardless of the sequence order. Traditional protein analysis tools, since they are sequence order dependent, cannot detect such a sequence order relaxing similarity. In this study, a more flexible protein comparison algorithm, the Similar enRiched Parikh Vector Searching (SRPVS) algorithm is designed to detect sequence similarity in a local-sequence-order-flexible manner. In SRPVS, a peptide sequence is broken into a group of Parikh vectors of predefined word sizes, and then Similar enRiched Parikh Vectors (SRPV) are searched between the two sequences and an Order Score is assigned to each pair of SRPV to reflect the order difference between the two sequences. A test has shown that SRPVS can detect shuffled protein sequence regions that share biological structure between two protein sequences.*

## 1. Introduction

It is well known that a protein of biological interest with unknown structure could very likely have sequence and structure similarity with its known structured biological homologues. In some protein sequence regions when two sequences share similar amino acid composition, they also share the same biological structure even when the order of the amino acids in the sequence is not preserved. This type of similarity has been mentioned in a recent study on tumor marker ligand study through *in vivo* phage display peptides [1]. Such an order flexible similarity will be referred to as composition similarity in the rest of this work. Almost all traditional protein sequence analysis tools, such as PSI-BLAST, are sequence order dependent. These tools cannot detect composition

similarities with shuffled orders between two protein sequences.

In this study, a more flexible protein sequence comparison algorithm, Similar enRiched Parikh Vector Searching (SRPVS) algorithm is designed to detect composition similarity among sequences while allowing flexibility in dealing with the local sequence orders. The Parikh vector is an ideal data structure to store information of a protein sequence within a certain word size regardless of the sequence order within that window size. Given an ordered alphabet  $A$  of finite  $k$  elements, with redundant elements permissible, a Parikh vector of a word  $w$  on the alphabet  $A$  is the integer vector  $v = (n_1, n_2, \dots, n_i, \dots, n_k)$  where  $n_i$  is the number of occurrences of the  $i^{\text{th}}$  letter of  $A$  in  $w$  [2]. An enriched Parikh vector is a data structure that contains an additional integer array storing order information of a PV to an original PV structure. In this way, for a word-sized subsequence, enRiched PV (RPV) is capable of storing the amino acid composition information and amino acid order information separately. The preliminary test has shown that SRPVS is capable of detecting shuffled protein sequence regions that share the same biological structure between two protein sequences or subsequences.

## 2. SRPVS Algorithm

In SRPVS, a peptide sequence is broken into groups containing all possible RPVs of different predefined word sizes. SRPVS searches for Similar enRiched Parikh Vectors (SRPV) between two sequences with predefined Sequence Composition Threshold (SCT) levels. Then for each pair of SRPVs, the algorithm assigns an Order Score (OS) that measures the difference in order between the two RPV word-sized sequences. SRPVS outputs the SRPV pairs whose composition similarity is greater than the SCT and whose OS meets the predefined condition of the

OS Threshold (OST) input value - either greater or less than the OST, depending on the research question.

#### Algorithm 1 (main SRPVS)

**Input:** polypeptide  $seq_1$  of size  $s_1$  and  $seq_2$  of size  $s_2$  and a known set  $A$  of size  $m$ .

**Output:** all similar Parikh vectors between the two sequences with locations and contents information.

```

1: initialize PV(seq_id, seq_size, word_size)
2: Get all PV for all subsequences in seq
3:  $s_1 \leftarrow$  size of  $seq_1$ 
4:  $s_2 \leftarrow$  size of  $seq_2$ 
5: for  $x_1 = 1, \dots, s_1 - I$  do //  $I$ : largest word size
6:   for  $y_1 = 2, \dots, I$  do
7:     for  $x_2 = 1, \dots, s_2 - I$ , do
8:       for  $y_2 = 2, \dots, I$ , do
// the detail of Sim function is in Algorithm 2
9:       if Sim (PV (1,  $x_1, y_1$ ), PV (2,  $x_2, y_2$ )) then
10:        add  $x_1, y_1$  and  $x_2, y_2$  to PVList
11:        increment PVList counter
12: print PVList

```

#### Algorithm 2 (Sim function)

**Input:** two enriched Parikh vectors, SCT.

**Output:** the OS if are SRPVS, -1 otherwise.

```

1: if sum < SCT, then //if the two are not SRPVS
2:   return -1;
3: else do //if the two are similar Parikh vectors
4:   for j from 0 to size1 in PV1, do
5:     saa  $\leftarrow$  amino acid at location j in PV1
6:     for k from j+1 to s1 in PV1, do
7:       faa  $\leftarrow$  residue followed saa in PV1
8:       if saa is before faa in PV2, faa, then do
9:         similar++;
10:        fscore  $\leftarrow$  sum (sum-1) / 2;
11:   OS  $\leftarrow$  similar/fscore
//OS=1 if the order is the same, OS = 0 if reversed
12: return OS;

```

### 3. Materials and Methods

The SRPVS algorithm was coded in java and is run under Linux with space complexity  $O(s_1 + s_2)I$  and time complexity  $O(s_1 s_2 I)$ . The SRPVS program is applied to blocks database version 5 [3]. The program compared the SRPV ratios from all columns (Acol) and the SRPV ratios from the same column (Scol) and the. Acol represents ratio of SRPV counts to all existing RPV pairs within a block and Scol is similar to Acol except that each SRPV pair represents two regions start at the same column and end at the same column. Both Scol and Acol similar RPV pairs conforms to the following conditions:

- Same word size;
- SCT = 1 (same composition in two RPVs);
- OST < 0.6 (exclude well-aligned Similar RPVs that can also be detected by alignment methods);
- alphabet of 20 amino acids;
- RPV word size range 5 to 10 (size 5 to 10 is a typical motif size range);
- SRPV ratios are calculated at RPV size level.

### 4. Results and Discussion

The ratio results of Scol and Acol in the block of glu\_carboxylation are in logarithmic scale (Figure 1).

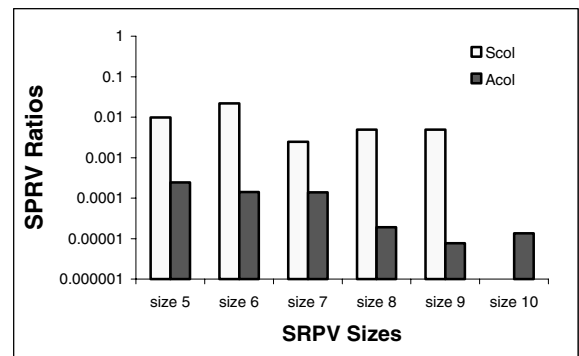


Figure 1. Scol vs. Acol within the block of glu\_carboxylation.

The ratios of SRPV in the same columns are significantly higher than the ratios of SRPV in all columns in the block of glu\_carboxylation except for that of the word size 10. Since amino acids in the same column within a block are considered to have same functionality for that block, so higher SRPV ratio in same columns than that in all columns suggests that two regions with same composition but shuffled order may share same biological function or structure, and such a similarity cannot be detected by traditional alignment tools. The inconsistency in word size 10 might be due to data insufficiency.

### 5. References

- [1] Arap W et al. "Steps toward mapping the human vasculature by phage display", *Nature Medicine*, vol. 8 (2), February 2002, pp. 121-127.
- [2] Parikh RJ. "On Context-Free Languages", *JACM*, vol. 13 (4), 1966, pp. 570 – 581.
- [3] Steven Henikoff and Jorja G. Henikoff, "Amino acid substitution matrices from protein blocks", *Biochemistry*, vol. 89, 1992, pp.10915-1091.