PROMETHEUS
Secondary Structural Elements' Profiles of Proteins

Mustafa TOPRAK[2], Canan HAS[1] and Jens ALLMER[1]

[1]*Molecular Biology and Genetics Department, Izmir Institute of Technology, Urla – Izmir, Turkey*

[2]*Computer Engineering Department, Izmir Institute of Technology, Urla – Izmir, Turkey*

INTRODUCTION

The tertiary structure of the native fold of a protein is essential for its function and although half a century has been spent on the prediction of tertiary structure from primary structure only little has been achieved for ab initio prediction. Predicting the secondary structure may be useful as a stepping stone to predict tertiary structure. Unfortunately, the accuracy of secondary structure prediction algorithms is approximately 80% according to the critical assessment of protein structure prediction (CASP) competition. We identified two interesting problems and set forth to tackle them. The first a comprehensive well designed benchmark dataset and the second a generalized prediction algorithm for secondary structure.

Here, we introduce a method for profiling and determining the secondary structure of proteins. As the benchmark dataset, structural sequences of Athena benchmark database [Master Thesis, Canan Has, İYTE-2011] which includes DSSP secondary structure assignments of proteins were used. Alpha helix motifs were extracted and their corresponding amino acid sequences with reduced alphabet representation were grouped into length categories. Each length group was clustered by K-means clustering. For the next step, motif profiles will be produced. These profiles later will be used to assign alpha helices to structurally unknown proteins.

MATERIALS AND METHODS

In spite of the other methods that are focusing on the atomic coordinates given by X-Ray crystallography, our algorithm tries to focus on the interpretation of the patterns extracted from structural motifs of the proteins. Four steps are applied: Conversion of amino acid sequences by three reduced alphabets, disposition of amino acids to occur in structural elements, biochemical properties, clustering of extracted motifs.

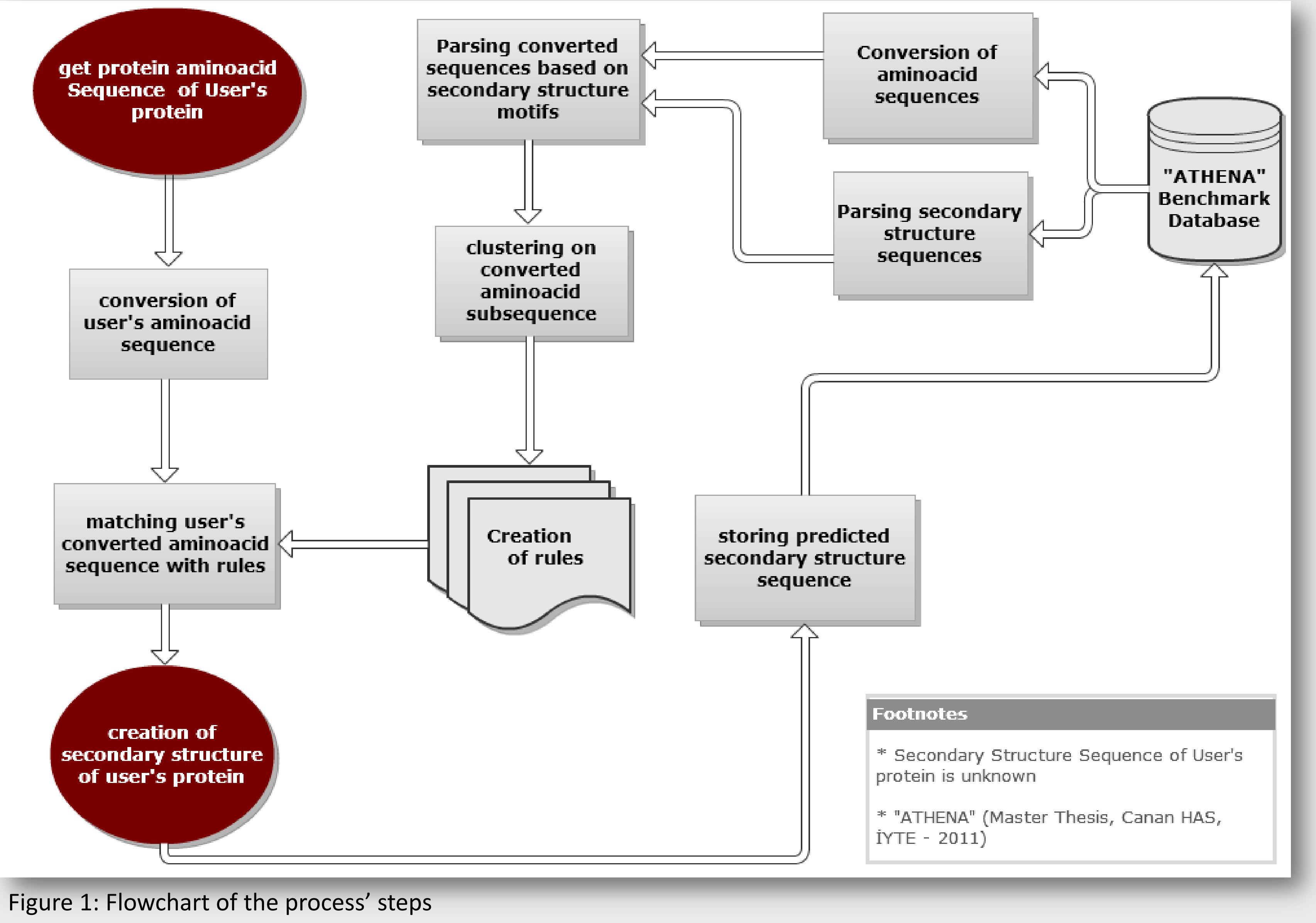
get protein aminoacid Sequence of User's protein
conversion of user's aminoacid sequence
matching user's converted aminoacid sequence with rules
creation of secondary structure of user's protein
Parsing converted sequences based on secondary structure motifs
clustering on converted aminoacid subsequence
Creation of rules
Conversion of aminoacid sequences
Parsing secondary structure sequences
"ATHENA" Benchmark Database
storing predicted secondary structure sequence
Footnotes
* Secondary Structure Sequence of User's protein is unknown
* "ATHENA" (Master Thesis, Canan HAS, İYTE - 2011)

Figure 1: Flowchart of the process' steps

Alphabet	a	b	c	d	e	f	g
Murphy99	LVIMC	AGSTP	FYW	EDNQKRH	-	-	-
Andersen04	PG	EKRQ	DSNTHC	IV	WFT	-	-
Barcadit09	ACILMV	FWY	DE	KHR	STNQ	G	P

Figure 2: List of reduced alphabets and their items. In Murphy99, amino acid groups are obtained according to BLOSUM 50 similarity matrix. Andersen40 shows flexibility in group a, large-polarity in group b, small-polarity in group c, aliphacity in group d, aromacity in group e, hydrophobicity in group e. Barcadit90 groups amino acids according to hydrophobicity (group a), aromacity-hydrophobicty (group b) , negative-charge (group c), positive charge (group d), polarity (group e), flexibility (group f), rigidity (group g).

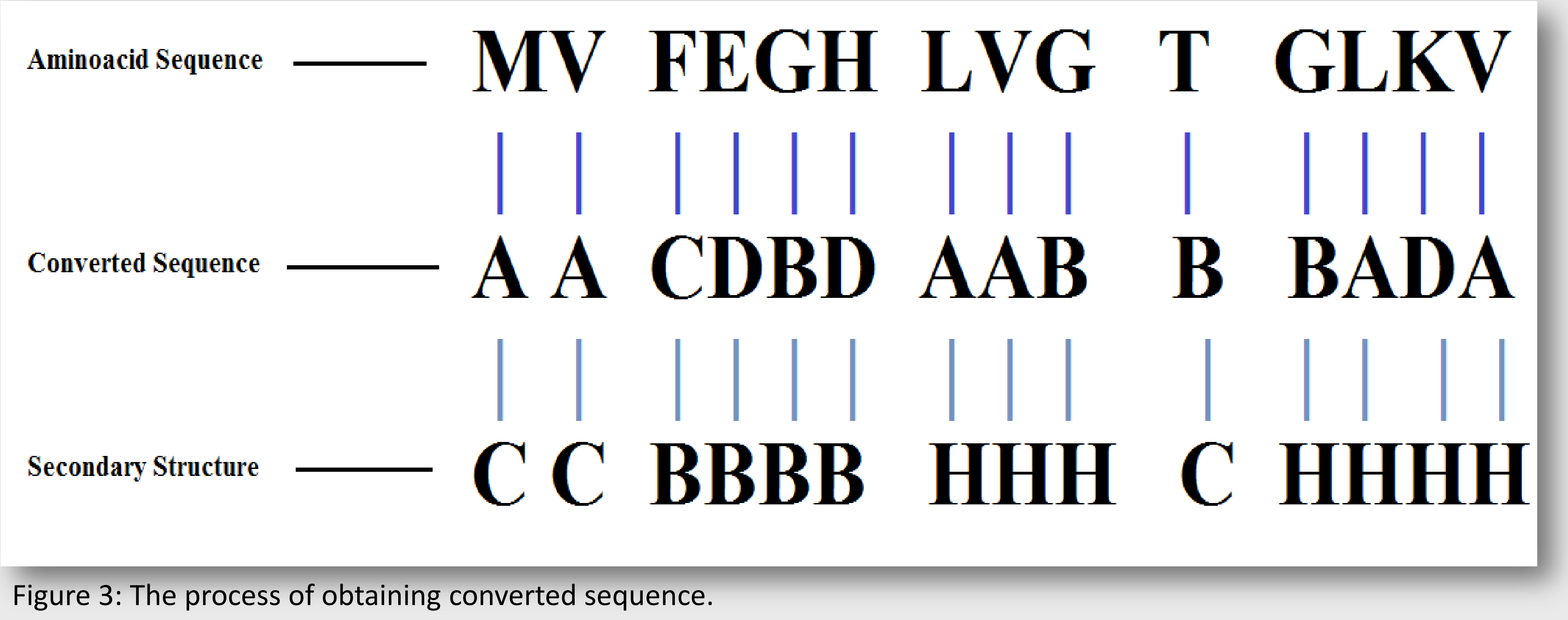

Figure 3: The process of obtaining converted sequence.

RESULTS

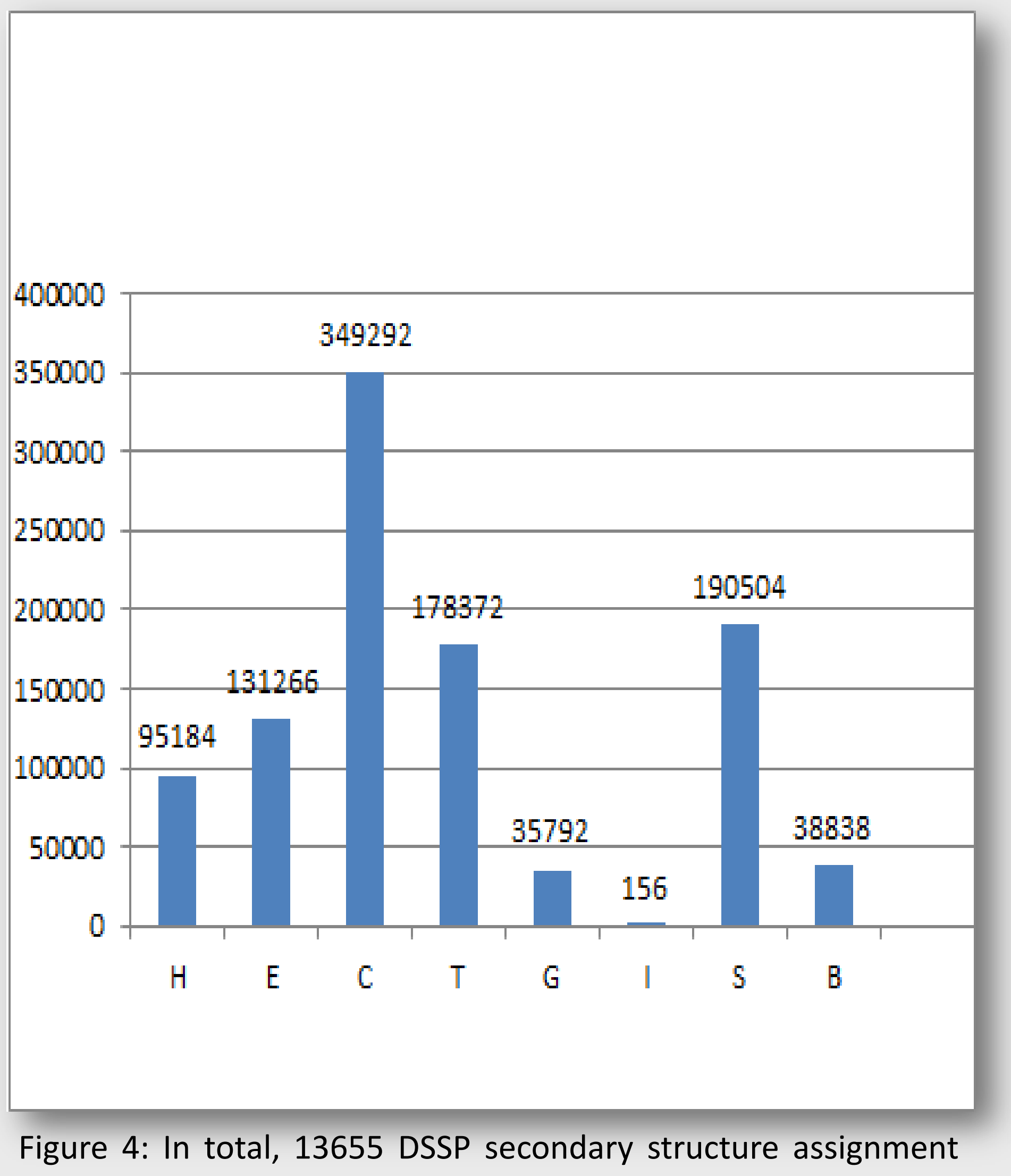

Figure 4: In total, 13655 DSSP secondary structure assignment sequences were used. According to DSSP structure classification; H=helix, E=sheet; C=coil/unassigned; T=turn; G= 3-10 helix; B=beta-bridge; I=pi-helix; S=bend. The distribution of these structural elements is shown.

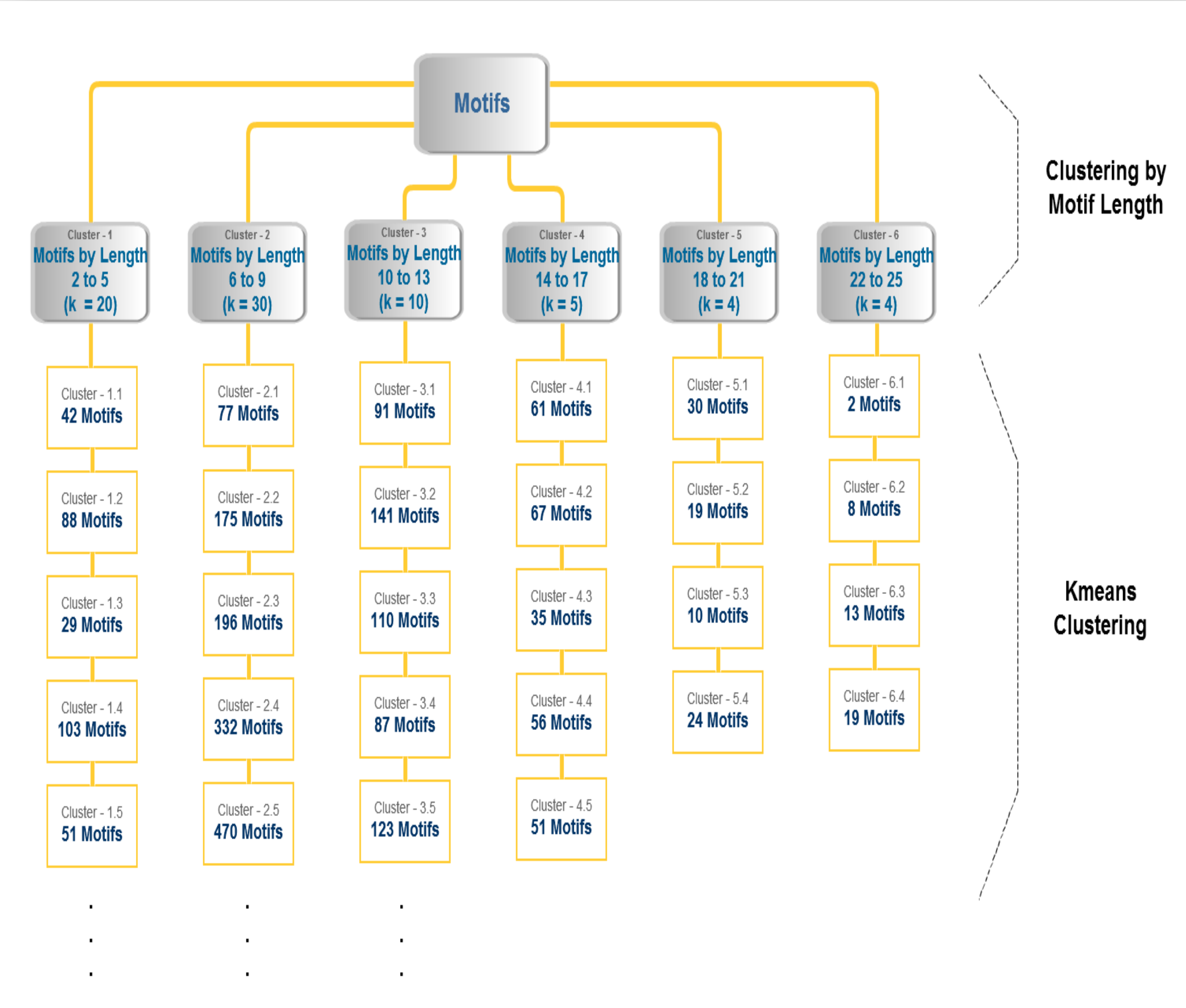

Figure 5: Two level clustering model is shown. At the first level, motifs were grouped into clusters according to their lengths. Six length-groups were obtained. For each group, k-means clustering were performed. For each length group, 4 to 5 cluster groups were observed. The number of motifs in each cluster is illustrated.

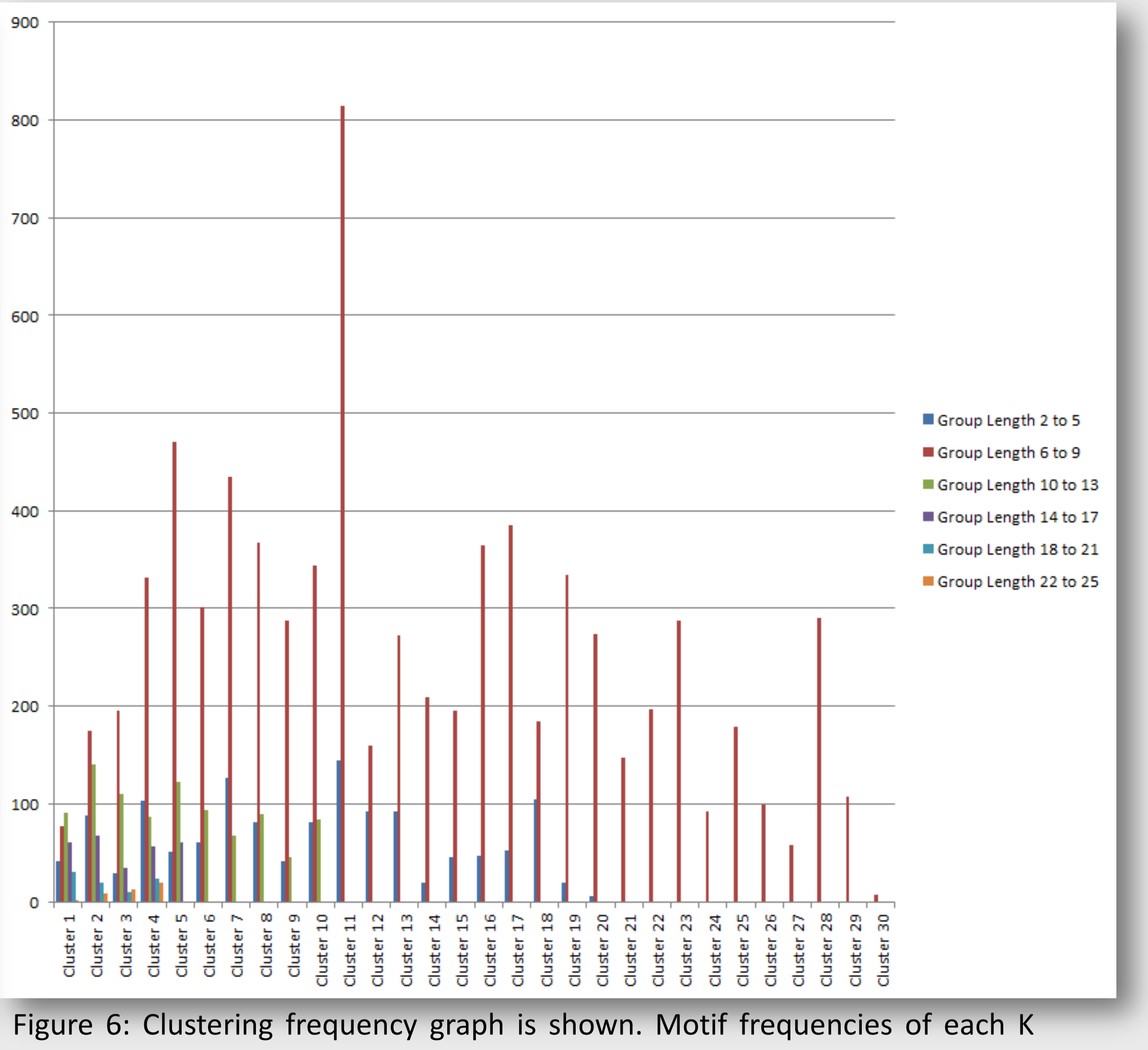

Figure 6: Clustering frequency graph is shown. Motif frequencies of each K means cluster of length groups.

FUTURE STUDIES

We aim to determine the optimum number of clusters for each motif length group. Afterward, motif profiles will be gathered by running multiple sequence alignment algorithm for the resulting clusters. The resulting alignments will be turned into profiles. Information in profiles then will be used to map motifs to sequences of known structure to determine our algorithms success rate. We expect that the knowledge acquired from profiles will convey more significant alpha helix assignment results than so far possible with other algorithms.