Protein Structure-Function Analysis with Self-Organizing Maps

Seonjoo Lim **Stephen Jaegle** **Lutz Hamel**
Department of Computer Science and Statistics
University of Rhode Island
Kingston, Rhode Island, USA
hamel@cs.uri.edu

Abstract

Here we describe an approach for protein structure-function analysis using self-organizing maps based on the structure of a protein's functional site. Our current approach differs from other approaches in that we directly unfold the 3D structure of the functional center of a protein into a suitable high-dimensional feature vector and then use self-organizing maps to discover similarities/dissimilarities between the corresponding feature vectors. We successfully applied our techniques to two large protein families: the protein kinases and the Ras superfamily. Even though a number of different approaches using self-organizing maps for the conformational analysis of molecules have been proposed, our approach is novel in that we apply it to protein families, use an efficient feature vector construction, and use a recently developed self-organizing map package that provides statistical support for evaluating the resulting map quality.

1 Introduction

The function of a protein is mainly determined by structural features, especially the functional site of the protein [1]. Therefore, common functionalities among proteins can be inferred from their structural similarities. Furthermore, in large protein families we can observe that small differences in the structure of the well preserved functional sites denote differences in functionality between the proteins. Here we describe an approach for protein structure-function analysis using self-organizing maps (SOMs) [2]. Our approach is based on the structure of a protein's functional site. The approach we employ here differs from other approaches, *e.g.* [3, 4, 5], in that we unfold the 3D structure of the functional center of a protein into a suitable high-dimensional feature vector and then use self-organizing maps to discover similarities/dissimilarities between the corresponding feature vectors. This approach to constructing feature vectors is substantially more efficient than the approach first outlined in [6]. Perhaps the work most closely related to ours is [7] where the authors classify protein motifs using SOMs. However, their work differs substantially from ours in that instead of directly encoding 3D spatial information of the motifs in question the authors compute a feature vector for a motif by looking at the angles at the α-carbon atoms along the backbone of a protein.

We successfully applied our technique to two large protein families: the protein kinases [8] and the Ras superfamily [9]. Our proposed approach seems to be novel in that we apply it to protein families, use an efficient feature vector construction, and use a recently developed self-organizing map package that provides statistical support for evaluating self-organizing map quality [10].

The remainder of this paper is structured as follows. In Section 2 we describe our methodology for aligning functional sites, extracting feature vectors, and computing SOM based models. We look at details of model building and cluster analysis in Section 3. In particular, we discuss the application of our technique to the two different protein families. Finally, we discuss our conclusions and further work in Section 4.

2 Preprocessing the Protein Structure Information

The major steps for preprocessing protein data are summarized in Figure 1. First, the protein structures for proteins under investigation are pulled from the Protein Data Bank (PDB) [11]. The proteins are then aligned using FATCAT [12]. From the aligned proteins we then extract only the functional site structures for our functional site based analysis. In order to filter out the functional sites, key structural information is used, like the consensus of a motif or the positional information (e.g., residue number) of a binding site for each protein. Next, the structures are simplified by extracting only the α-carbons from these functional sites. This provides information on the backbone structure of the functional sites by excluding the side chains. Finally, each functional site is represented by the 3D-coordinates of its α-carbons, and the coordinate data of all the α-carbons is unfolded into a linear vector – the feature vector of the functional site.

Figure 2 shows a set of feature vectors as rows. Each row

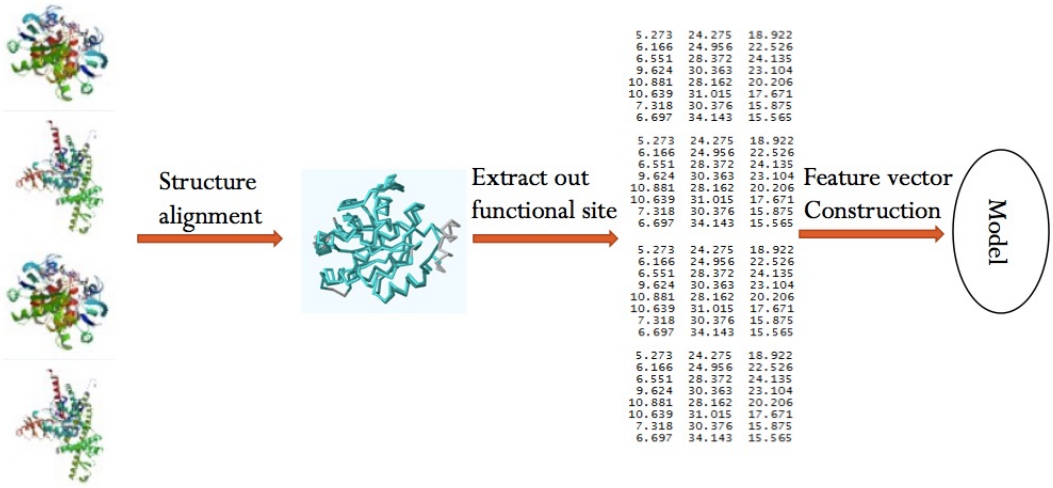

Figure 1: Steps in the protein structure-function analysis.

	family	ids	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	HRas	121p	5.273	24.275	18.922	6.166	24.956	22.526	6.551	28.372	24.135
2	HRas	1qra	5.252	24.249	18.962	6.297	24.972	22.523	6.522	28.442	23.963
3	HRas	1ctq	5.278	24.256	18.976	6.262	24.946	22.534	6.503	28.426	23.990
4	HRas	1p2s	5.170	24.126	19.013	6.269	24.884	22.596	6.456	28.492	23.750
5	HRas	1agp	5.330	24.484	18.940	6.208	24.978	22.556	6.508	28.340	24.146
6	KRas	4dsn	5.348	24.380	18.886	6.154	24.970	22.557	6.604	28.328	24.302
7	RhoA	1a2b	5.149	24.282	18.878	6.373	24.932	22.453	6.699	28.276	24.223
8	RhoA	1cc0	5.153	24.122	18.987	6.432	24.981	22.549	6.441	28.517	23.924
9	RhoA	1cxz	5.202	24.334	18.906	6.462	24.972	22.434	6.659	28.312	24.221
10	RhoA	1dpf	5.378	24.265	18.894	6.376	25.021	22.497	6.459	28.517	24.011
11	RhoA	1ftn	5.153	24.122	18.988	6.432	24.981	22.548	6.441	28.517	23.925
12	Rab1A	2fol	5.217	24.048	18.990	6.134	24.992	22.576	6.489	28.664	23.592
13	Rab1A	2wwx	5.564	24.206	18.775	6.150	25.179	22.363	6.372	28.705	23.726
14	Rab1A	3sfv	5.264	24.036	19.057	6.325	25.017	22.610	6.309	28.671	23.678
15	Rab1A	3tkl	5.165	24.113	19.010	6.364	25.004	22.511	6.405	28.514	23.895
16	Rab1B	3jza	5.434	24.280	18.825	6.028	25.179	22.421	6.506	28.721	23.680

G(x,y,z) X(x,y,z) X(x,y,z)

Figure 2: Feature vector construction, unfolded 3D-coordinates.

represents the residues of a functional site of a protein from the Ras family. For demonstration purposes the functional sites have been truncated to three amino acids. In our actual setting we consider eight residues. Here each feature vector is denoted by two labels (a family name and a PDB ID), and three sets of attributes representing the 3D coordinates of the α-carbon of the three residues (GXX). The three residues are the first three of the eight residues making up the phosphate binding loop (p-loop) motif GXXXXGK[S/T]. The p-loop is the active site in the Ras family. In other words, a feature vector for a protein is the sequential listing of the 3D coordinates of the α-carbons appearing in its functional site.

3 Model Building and Evaluation

For our experiments we used proteins from two large protein families: the Ras family and the protein kinase family. We preprocessed the proteins as described in Section 2 and then constructed self-organizing maps for each protein family. The maps shown are fully converged, i.e., the clusters, their size, and their relative position to each other are statistically meaningful [13]. We commence this section by briefly reviewing self-organizing maps.

3.1 Self-Organizing Maps

A self-organizing map [2] is a kind of artificial neural network that implements competitive learning, which can be considered a form of unsupervised learning. On the map itself, neurons are arranged along a rectangular grid with dimensions x_{dim} and y_{dim}. Learning proceeds in two steps for each training instance $\vec{x}_k$, $k = 1, 2, 3, \ldots, M$, with M the number of training instances:

1. The **competitive step** where the best matching neuron for a particular training instance is found on the rectangular grid,

$$c = \underset{i}{\operatorname{argmin}}(||\vec{m}_i - \vec{x}_k||)$$

where $i = 1, 2, \ldots, N$ is an index over the neurons of the map with $N = x_{dim} \times y_{dim}$ the number of neurons on the grid, and $\vec{m}_i$ is a neuron indexed by i. Finally, c is the index of the best matching neuron $\vec{m}_c$ on the map.

2. The **update step** where the training instance $\vec{x}_k$ influences the best matching neuron $\vec{m}_c$ and its neighborhood. The update step can be represented by the following update rule for the neurons on the map,

$$\vec{m}_i \leftarrow \vec{m}_i - \eta \vec{\delta}_i h(c, i)$$

for $i = 1, 2, \ldots, N$. Here $\vec{\delta}_i = \vec{m}_i - \vec{x}_k$, η is the learning rate, and $h(c, i)$ is a loss function with,

$$h(c, i) = \begin{cases} 1 & \text{if } i \in \Gamma(c), \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(c)$ is the neighborhood of the best matching neuron $\vec{m}_c$ with $c \in \Gamma(c)$. Typically the neighborhood is a function of time and its size decays during training. Initially the neighborhood for neuron $\vec{m}_c$ includes all other neurons on the map,

$$\Gamma(c)|_{t=0} = \{1, 2, \ldots, N\}.$$

As training proceeds the neighborhood for $\vec{m}_c$ shrinks down to just the neuron itself,

$$\Gamma(c)|_{t \gg 0} = \{c\}.$$

Here, as before, $N = x_{dim} \times y_{dim}$ is the number of neurons on the map. This means that initially the update rule for each best matching neuron has a very large field of influence which gradually shrinks to the point that the field of influence just includes the best matching neuron itself.

The two training steps above are repeated for each training instance until the given map converges.

Here we use our `popsom` package [10] which supports statistical convergence criteria [13] and detailed cluster visualizations in terms of our starburst plots [14]. Figure 3 shows a scatter plot of the hepta problem in Ultsch's fundamental clustering problem suite [15]. The data set consists of seven distinct clusters embedded in three dimensional space. Notice that there is a single, very tight cluster at (0,0) and then we have six clusters surrounding this center cluster. Figure 4 shows a SOM starburst plot of this data set. The seven clusters can easily be identified on the map by their starbursts. Also easily visible is the fact that clusters themselves are identified by their light color and cluster boundaries are identified by darker colors. The easily identified borders mean that the clusters are indeed distinct clusters. Their relative position is also meaningful to a point, given that this is a 2D rendering of a higher dimensional space. Here we see the cluster with

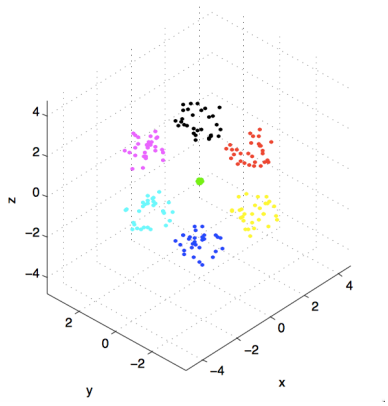

Figure 3: The FCPS Hepta data set.

Table 1: Hierarchy of the STE Kinase Family and corresponding Binding Sites.

Family	Subfamily	PDB ID	Binding Site
STE 7	MAP2K4	3ALO	108-116
STE 11	MAP3K5	4BF2 3VW6	686-694
STE 20	PAK6 PAK4	4KS7 2J0I, 4JDI	413-421

class label 1 towards the center of the map. This is a representation of the tight center cluster in the original plot. We can also see that it consumes somewhat less map real estate than the other clusters meaning that the cluster is very tight. All these observations are justified due to the fact that the map has converged and therefore positioning and distance amongst clusters is statistically meaningful.

3.2 The Protein Kinase Family

Protein kinases catalyze proteins by attaching phosphate groups to them. For example, protein kinase helps bind ATP to proteins so that they can be phosphorylated and produce ADP. Here we consider the sterile (STE) group which is one of ten human kinase families [8]. Three main families in the STE group operate on each other sequentially: STE 20 activates STE11 and STE11 activates STE 7. Table 1 shows the STE families with their subfamilies and their corresponding members. Also shown is the binding site for each member. Note that the length of the respective binding sites is eight residues. As mentioned above, that means the corresponding feature vectors have a length of twenty four.

Figure 5 shows that the SOM algorithm was able to recover the subfamilies given in Table 1 as separate clusters. Here we make use of the fact that cluster relative positioning and the coloring of the map is statistically meaningful because the map has converged. We can observe that the clusters for

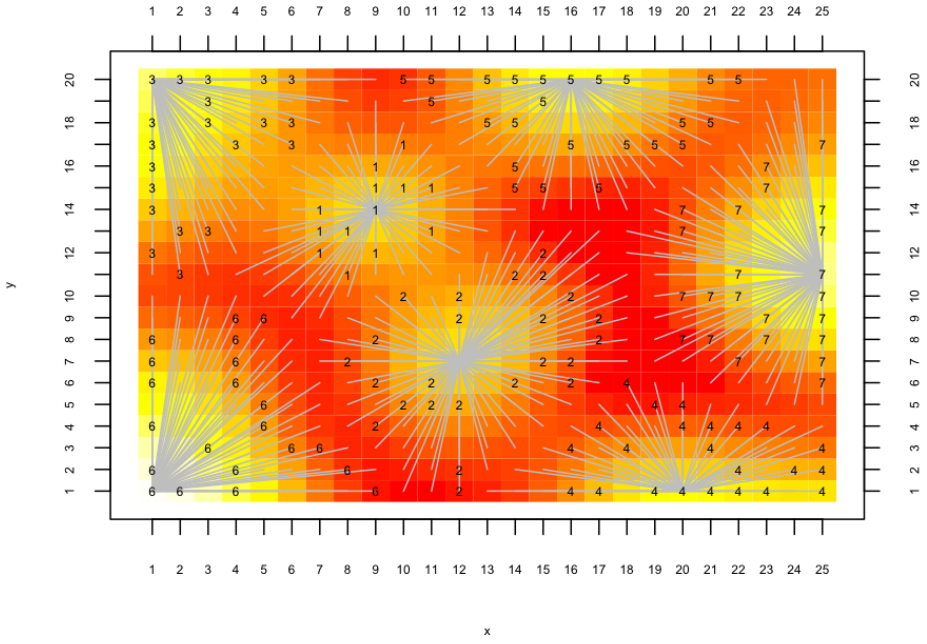

Figure 4: A SOM starburst plot of the Hepta data set.

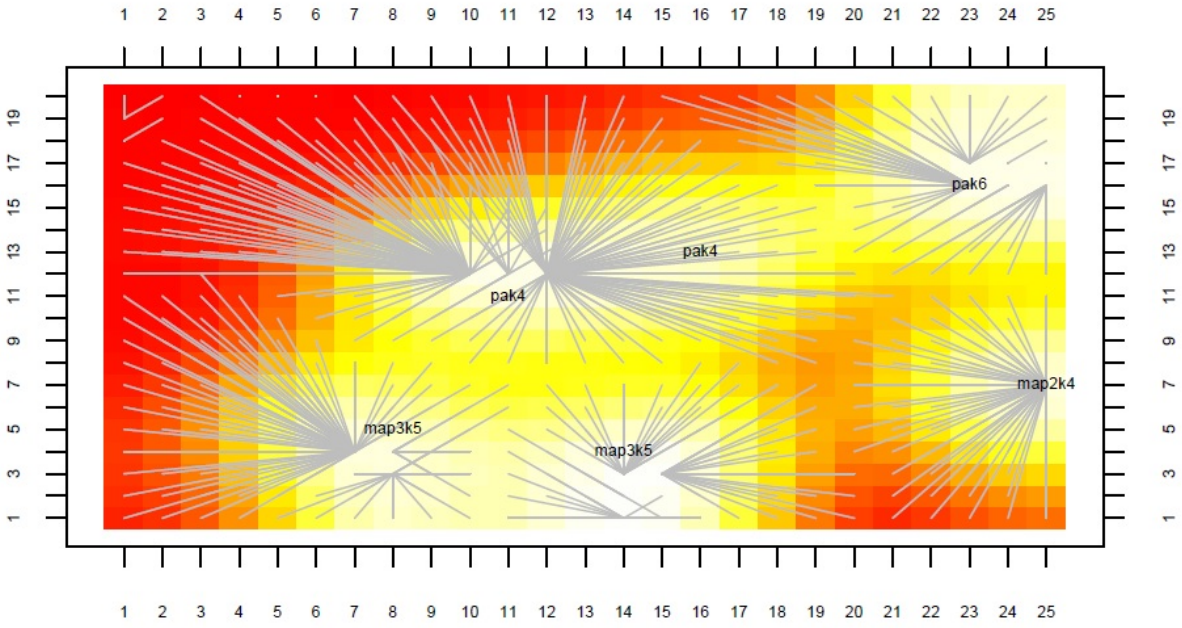

Figure 5: A SOM depicting the protein kinase STE Group.

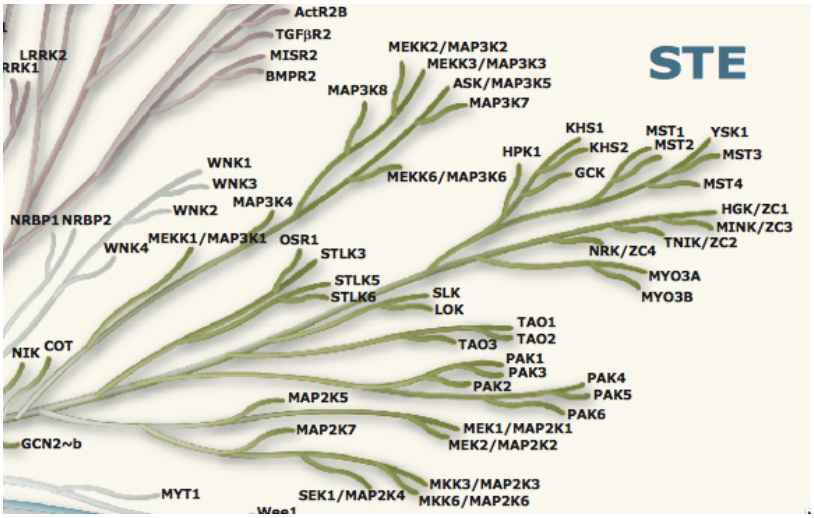

Figure 6: The STE branch of the evolutionary tree of the human kinase complement [8].

PAK4 and MAP3K5 are distinct but close together. We can therefore infer that the active site structures for those proteins are very similar. The same holds for the clusters PAK4 and PAK6. It is perhaps remarkable that the active site structure for MAP2K4 seems to share more similarity with PAK6 than with MAP3K5. It is tempting to see if an analysis solely based on the functional site of these proteins can recover the evolutionary relationships between the protein families shown in Figure 6 which is the STE branch of the evolutionary tree of the human kinase complement published as a supplement to [8]. Now, if we consider Figure 6 we can find MAP2K4, PAK4, and PAK6 on the lower branches of the tree. We can find MAP3K5 on an upper branch near the STE label. Tracing the evolutionary lines in Figure 6 it is clear that SOM's cluster structure captures the the evolutionary relationships with the exception of perhaps the MAP3K5/PAK4 relationship which on the SOM appears to be much closer than the tree indicates. One interpretation is that the structure of the active site of these two protein families has been highly preserved and that evolutionary differences manifest themselves in different protein domains.

3.3 The Ras Superfamily

The Ras superfamily of small GTPases is a large and diverse group of proteins that act as molecular switches for regulating cellular functions [9]. This superfamily is divided into five major families based on their structural and functional similarities: Rho, Ras, Rab, Ran, and Arf [16]. The protein members of the Ras superfamily have 40% - 85% of high primary sequence identity, while each subfamily has individual functions and different targets [17]. All members of the Ras superfamily have highly conserved common structural cores and function as GDP/GTP-regulated molecular switches. For example, a GTP-binding protein binds to either guanosine diphosphate (GDP) or guanosine triphosphate (GTP) so the protein becomes either inactive or active, respectively [18].

Table 2: Hierarchy of the Ras superfamily and the list of proteins used in the analysis.

Family	Subfamily	PDB ID
Ras	HRas KRas	121P, 1QRA, 1CTQ, 1P2S, 1AGP 4DSN
Rho	RhoA	1A2B, 1CC0, 1CXZ, 1DPF, 1FTN
Rab	Rab1A Rab1B	2FOL, 2WWX, 3SFV, 3TKL 3JZA
Arf	Arf1 Arf2 Arf3 Arf4	1HUR 1U81 1RE0 1Z6X
Ran		1I2M, 1IBR, 1RRP, 3CH5, 3EA5, 3GJ3

There is a particular motif in the proteins of the Ras superfamily that determines the features of each subfamily. Each subfamily acts as a molecular switch for a unique target or intervenes in a cell process, such as cell proliferation. Members of this superfamily conserve five G domains which are fundamental subunits: G1-G5 [9]. G domains are highly conserved regions related to nucleotide binding, a process that is involved with the GDP/GTP cycle. The G1 domain contains the phosphate binding loop (p-loop), which is a common motif in GTP binding proteins with a consensus of GXXXXGK[S/T], where X denotes any amino acid and S/T means S or T. Interestingly enough, the length of the active site of the protein family measure also eight residues. Table 2 shows the hierarchical relationship of the Ras superfamily and the list of PDB IDs chosen for analysis in this paper.

Figure 7 shows a SOM constructed for the Ras superfamily.

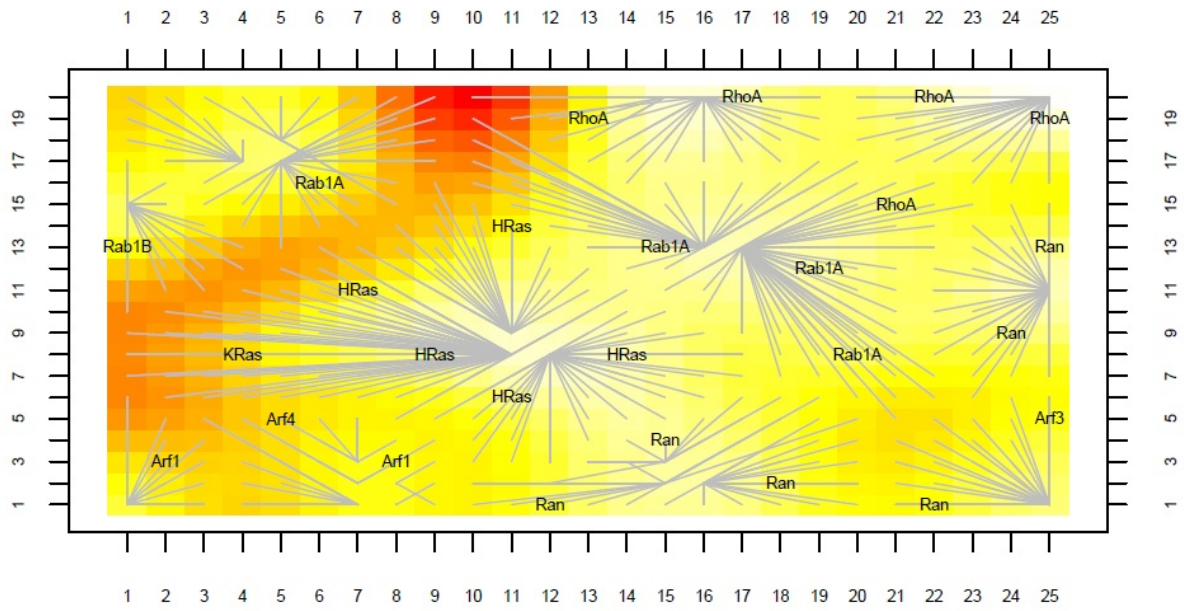

Figure 7: A SOM depicting the Ras Superfamily.

The major clusters for Rho (top-right), Rab and Ran (center-right), and Arf and Ras (bottom-center-left), are easily identified. What is curious is that there is a separate Rab cluster on the top-left of the map separated from the remaining clusters by a fairly dark border. This means that structurally these Rab proteins look substantially different from the remaining proteins. We intend to follow up and investigate.

We can now investigate whether an structure-function analysis solely based on the active site of the proteins preserves the evolutionary relationship of the proteins. Figure 8 shows a consensus tree of the Ras superfamily published in [19]. From the tree it is easily identified that the families Rab, Ran, and Rho are closely related to each other and that the families Arf and Ras form another cluster. Going back to our map in Figure 7 we can see that the relative positioning of the clusters on the map preserves these evolutionary relationships. We can observe one outlier - a sole Arf protein shows up in the Ran cluster at the bottom right of the corner.

4 Conclusions and Further Work

Here we described our approach for protein structure-function analysis using self-organizing maps based on the structure of a protein's functional site. Our current approach differs from other approaches in that we directly encode the 3D structure of the functional center of a protein into a suitable high-dimensional feature vector and then use self-organizing maps to discover similarities/dissimilarities between the corresponding feature vectors. We used our `popsom` package which supports statistical convergence and quality measures and advanced visualization techniques for self-organizing map construction and evaluation. We successfully applied our techniques to two large protein families: the protein kinases and the Ras superfamily. We have shown that SOM preserves protein intra-family relationships as clusters and inter-family relationships with the relative positioning of family cluster to each other on a map. We have shown that evolutionary relationships between protein families are to a large degree preserved within the active site of the proteins.

Future research will focus on applying this technique to other protein families and structures. We also intend to investigate the outliers we found on the map for the Ras super family.

References

[1] H. A. Maghawry, M. G. Mostafa, M. H. Abdul-Aziz, and T. F. Gharib, "Structural protein function prediction-a comprehensive review.," *International Journal of Modern Education & Computer Science*, vol. 7, no. 10, 2015.

[2] T. Kohonen, *Self-organizing maps*. Springer Berlin, 2001.

[3] M. T. Hyvönen, Y. Hiltunen, W. El-Deredy, T. Ojala, J. Vaara, P. T. Kovanen, and M. Ala-Korpela, "Application of self-organizing maps in conformational analysis of lipids," *Journal of the American Chemical Society*, vol. 123, no. 5, pp. 810–816, 2001.

[4] T. Murtola, M. Kupiainen, E. Falck, and I. Vattulainen, "Conformational analysis of lipid molecules by self-organizing

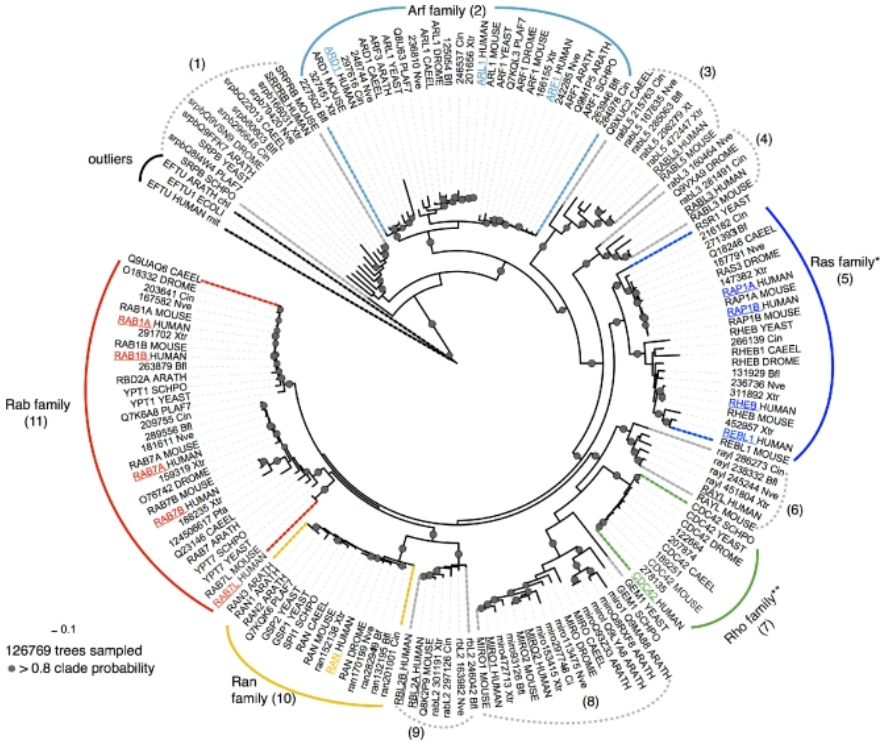

Figure 8: A consensus tree of the Ras Superfamily.

maps," *The Journal of chemical physics*, vol. 126, no. 5, p. 054707, 2007.

[5] D. Fraccalvieri, A. Pandini, F. Stella, and L. Bonati, "Conformational and functional analysis of molecular dynamics trajectories by self-organising maps," *BMC bioinformatics*, vol. 12, no. 1, p. 1, 2011.

[6] L. Hamel, G. Sun, and J. Zhang, "Toward protein structure analysis with self-organizing maps," in *Computational Intelligence in Bioinformatics and Computational Biology, 2005. CIBCB'05. Proceedings of the 2005 IEEE Symposium on*, pp. 1–8, IEEE, 2005.

[7] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg, and P. Wrede, "Local structural motifs of protein backbones are classified by self-organizing neural networks," *Protein engineering*, vol. 9, no. 10, pp. 833–842, 1996.

[8] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002.

[9] K. Wennerberg, K. L. Rossman, and C. J. Der, "The ras superfamily at a glance," *Journal of cell science*, vol. 118, no. 5, pp. 843–846, 2005.

[10] L. Hamel, B. Ott, and G. Breard, *popsom: Self-Organizing Maps With Population Based Convergence Criterion*, 2015. R package version 3.0.

[11] P. W. R. et al., "The RCSB Protein Data Bank: new resources for research and education," *Nucleic acids research*, vol. 41, no. D1, pp. D475–D482, 2013.

[12] Y. Ye and A. Godzik, "Flexible structure alignment by chaining aligned fragment pairs allowing twists," *Bioinformatics*, vol. 19, no. suppl 2, pp. ii246–ii255, 2003.

[13] L. Hamel, "Som quality measures: An efficient statistical approach," in *Advances in Self-Organizing Maps and Learning Vector Quantization*, pp. 49–59, Springer, 2016.

[14] L. Hamel and C. W. Brown, "Improved interpretability of the unified distance matrix with connected components," in *7th International Conference on Data Mining (DMIN'11)*, pp. 338–343, 2011.

[15] A. Ultsch, "Clustering wih som: U* c," in *Proceedings of the 5th Workshop on Self-Organizing Maps*, vol. 2, pp. 75–82, 2005.

[16] R. Kahn, C. Der, and G. Bokoch, "The ras superfamily of gtp-binding proteins: guidelines on nomenclature.," *The FASEB journal*, vol. 6, no. 8, pp. 2512–2513, 1992.

[17] M. Kennedy, H. Beale, H. Carlisle, and L. Washburn, "Achieving signalling specificity: the ras superfamily.," *Nature Reviews Neuroscience*, vol. 6, pp. 423 – 434, 2005.

[18] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, 2002.

[19] A. M. Rojas, G. Fuentes, A. Rausell, and A. Valencia, "The ras protein superfamily: evolutionary tree and role of conserved amino acids," *The Journal of cell biology*, vol. 196, no. 2, pp. 189–201, 2012.