

PROTEIN STRUCTURE ANALYSIS WITH SOM

BY

SEONJOO LIM

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

UNIVERSITY OF RHODE ISLAND

2015

MASTER OF YOUR SCIENCE THESIS
OF
SEONJOO LIM

APPROVED:

Thesis Committee: Joan Peckham
Bethany Jenkins

Major Professor Lutz Hamel

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2015

ABSTRACT

The exponential growth of proteome databases has increased the demand for methodologies that can reveal the structural relationships between proteins. In general, large protein families need to be approached on several different levels in order to be fully understood. In such families, key characteristics and relationships are hidden under their sophisticated structures. While similarities in the primary sequences of two proteins give basic clues about their relationship, three-dimensional structural information provides crucial details needed for determining protein functionality.

As such, powerful and efficient computational analytic methods are becoming all the more essential. In the case of proteins, functionalities are most closely related with their three-dimensional structures. Thus, analysis based on the three-dimensional structure is absolutely necessary. The functions of proteins, particularly the functions of specific functional sites, are determined primarily by structural features. Thus, it can be said that structural similarities often point to functional similarities as well.

This analysis, based on the functional site, suggests a unique way of constructing a structural comparison model using SOM, an unsupervised machine learning algorithm. The experiment was performed with two popular protein families. Structural alignment of protein structure was performed prior to the analysis, in hopes of minimizing the error in the three-dimensional structures of the proteins. The SOM technique was then applied to the aligned structures. The results obtained with the SOM algorithm highlight the similarity and dissimilarity of the proteins. Finally, by analyzing clusters in a SOM grid, the structure-function relationship between proteins could be identified.

ACKNOWLEDGMENTS

First and foremost, I would like to express my appreciation to my thesis advisor Dr. Hamel, who has supported me and advised me throughout my research. Without his guidance, I would never have been able to finish my thesis.

I would also like to thank the members of my committee, Dr. Peckham and Dr. Jenkins, who were willing to participate in my defense.

A special thanks to my family—my husband Kyukwang, and my son Andrew—for their love and encouragement. I would also like to thank my parents for their love and prayers.

LIST OF TABLES

TABLE	PAGE
Table 1. Hierarchy of Ras superfamily and the list of proteins used for SOM analysis	12
Table 2. Hierarchy of Ste Kinase Family and the Binding Sites	13

LIST OF FIGURES

FIGURE	PAGE
Figure 1. Local Structure Alignment Tool and Visualization of Aligned Structures .	15
Figure 2. Globally aligned and superposed structure of 121P and 1A2B using FATCAT	16
Figure 3. Alignments of p-loop motif of 121P and 1A2B with (a) global alignment, (b) local alignment.....	16
Figure 4. Preprocessing the protein structural data.....	18
Figure 5. Feature vector construction, unfolded xyz coordinates	19
Figure 6. SOM result with Local Alignment (25 X 20 SOM with 500 iterations)	22
Figure 7. SOM result with Global Alignment (25 X 20 SOM with 200 iterations) ...	23
Figure 8. 25 x20 SOM of STE kinase family with local alignment (50 iterations)	24
Figure 9. 25 x20 SOM of STE kinase families with global alignment (50 iterations)	25
Figure 10. Cluster Dendrograms with results of (a) Local Alignment and (b) Global Alignment.....	26
Figure 11. Cluster Dendrograms with results of (a) Local Alignment and (b) Global Alignment.....	27

CHAPTER 1

INTRODUCTION

The accelerating growth of proteomic data demands effective analytical methods for revealing the relationship between proteins. Proteome science offers various approaches for comparing structural characteristics of proteins and continuously attempts to divulge the correlation between protein structure and function. However, there are many difficulties in identifying such relationships due to their structural complexity. Structural analysis is carried out on several different levels. Comparing amino acid sequences, the primary structure of proteins, for instance, is primitive yet important. There exist several sequence similarity search tools such as BLAST [1] that help find regions with similar sequences and optimal sequence alignments. However, only comparing the primary sequences or secondary structures--the alpha-helices and the beta-sheets-- of these proteins is not sufficient for uncovering the finer structural characteristics. These structural characteristics, including adopting a particular fold or conformation, can lead to a deeper understanding of the functional relationship between proteins [2]. Thus, since protein function is significantly related to its specific three-dimensional structure, a structure-based approach is crucial for identifying the relationship between proteins. The most common method for 3D protein structure comparison is global Root Mean Square Deviation (RMSD) that represents the average distance between the two equivalent atoms for the all pairs in global structure [25]. By focusing on the functional core, not

comparing global structure, it is able to show the meaningful structural functional relationship.

A protein family is a group of proteins with common sequence features and similar biological functions. A large protein family often has a hierarchical relationship and can be arranged in a tree representing their evolutionary origin and their subfamilies (e.g. the Ras superfamily is divided into five major subfamilies) [3]. Proteins generally interact with their substrates at a particular site called the active site. The functional site is considered a decisive factor for discerning which kinds of molecules they will interact with. Ultimately, we expect that a structural comparison of the functional sites will allow us to classify the protein family based on the structure-function relationship of the proteins.

One of the most well-known proteomic structural databases is the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [4]. PDB can be found on the web site <http://www.rcsb.org>, which contains information about the 3D structures of large biological molecules. PDB provides information on over 100,000 protein structures and seems to be expanding. With such a rapidly growing proteomic database, more effective analytical methods are becoming increasingly necessary for identifying the relationships between proteins.

The primary advantage of using SOM in this research is the ability to represent the similarities between the protein structures

There was an initial experiment on functional center-based analysis of protein structure using Self-Organizing Maps (SOM) [5]. This novel method recognized a functionally important local structure, the functional center, and extracted out the

surrounding structure within a certain radius. After performing structural alignment on the selected functional local structures, SOM was finally applied to these aligned structures. However, primitive local structural alignment techniques which had been performed manually with DS Viewer [6], were big hurdles in performing a fast and accurate analysis. Converting three-dimensional structural coordinates into linear vectors in order to construct feature vectors for SOM was also very difficult. In this paper, a new local structural alignment tool was used to improve the effectiveness of the research. In addition, straightforward feature vector constructions for SOM introduced here made the complex steps remarkably simple.

SOM is one of the artificial neural network algorithms, with an unsupervised learning aspect. Unsupervised learning trains the data without pre-defined categories whereas supervised learning has specified classes. SOM technique is often used as an analysis algorithm because it has many capabilities that other structural classification tools such as SCOP [7] and CATH [8] do not have. The greatest advantage of using SOM is its great ability to reduce dimensionality. In addition, SOM can process multiple objects at the same time and has the benefits of having graphical representations and easy interpretation. With Popsom [9], a new SOM package, a map can be constructed, as well as evaluated on its reliability, by computing the convergence rate of the map. The map can be trained until it has converged well, and this converged map can later be a criterion for selecting models that enhance the accuracy of the analysis of this research.

The objective of this research is to elucidate structural-functional relationships by classifying proteins from families into subfamilies using their structural features,

given 3D coordinate information on the proteins via unsupervised machine learning. In this paper, a unique structure-based approach is suggested, focusing on the structure of the functional site via the SOM algorithm with automated structural alignment techniques.

CHAPTER 2

BACKGROUND

2.1 Self-Organizing Maps

The Self-Organizing Map (SOM) [10], introduced by Kohonen, is one of the most prominent artificial neural network algorithms with aspects of unsupervised learning. The main goal of unsupervised learning is to discover hidden patterns underlying data without explicit target definition. SOM is used in a wide variety of fields such as market analysis, image processing, and bioinformatics, fields that typically require finding clusters that group data by similarity. The main idea of the SOM technique is to project multi-dimensional data into a low-dimensional map, where the map represents the similarity or dissimilarity of the input. For each observation, a corresponding neuron is calculated in the SOM and a simple topological map shows the nice low-dimensional representation of the input data. By competitive learning, the SOM algorithm finds the best matching neuron and updates the winning node and its neighborhood neurons.

Training a map is similar to regression process. Let x be an n -dimensional input vector. At each iteration, vector x is compared with all the m_i , the reference models, which have the same dimensionality as the input vector and are randomly initialized at the beginning. Then, the best matching unit or winning node using Euclidean distance between vector x and reference model m_i , that is the minimal $\|x - m_i\|$, is computed,

$$c = \underset{i}{\operatorname{argmin}} \{ \|x - m_i\| \} \quad (1)$$

where c is the index of the winning reference model. The winning reference model is the reference model with the shortest distance to the input vector x .

Next, the following formula shows the adjustment of the weights of all the reference models m_i ,

$$m_i(t+1) = m_i(t) + h_{ci}(t) [x(t) - m_i(t)] \quad (2)$$

where $t = 0, 1, 2, \dots$ is the step index. Here $h_{ci}(t)$ is the neighborhood function defined as follows,

$$h_{ci} = \begin{cases} 0 & \text{if } |c-i| > \beta, \\ \alpha & \text{if } |c-i| \leq \beta \end{cases} \quad (3)$$

where α is the learning rate and β is the neighborhood radius. The neighborhood function selects the reference models that need to be updated and only selects nodes that are within the neighborhood β . The neighborhood function gets increasingly smaller over time (that is, both α and β are functions of time t) and the adjusting steps are repeated consistently over the specified iteration.

The greatest advantage of SOM is data visualization. The low-dimensional SOM result can be interpreted intuitively. In addition, SOM achieves dimension reduction of data by projecting high-dimensional input data onto a two-dimensional grid that represents the essential clusters underlying input data with minimal loss of information. The gradient colors of grid units in map show the relative distances between reference vectors. Lighter colors represent greater similarity or closeness, while darker colors represent greater dissimilarity or distance.

2.2 Structural-Functional Relationship of Protein Family

A protein family is typically defined by similarities in the sequences of amino acids or similarities in their biological functions. Members of the same protein family are evolutionary-related so that they share a common ancestor and can thus often be arranged in a hierarchical system. For the most part, protein families can be divided into subfamilies and sometimes into even smaller families. For instance, the Ras superfamily is divided into 5 major subfamilies: Rho, Ras, Rab, Ran, and Arf. These divisions are made according to the structural and functional similarities, with each subfamily involved with a specific function [11].

Some computational methods for protein family classification are sequence-based, which finds the relationship among proteins based on similarity in amino acid sequence profiles [12]. However, it is well known that similarities in sequence do not indicate structural similarity [13]. Therefore, searching for sequential similarity alone is insufficient for determining other important functional properties which are more related to the three-dimensional structure.

The classification of protein families based on structural similarity is a major issue in computational biology. Comparing the 3D structure of proteins requires more intensive computation than sequential comparison. In general, the 3D structure of functional sites in a protein is highly conserved during evolution and is more related to the function of proteins. Comparing the structure of specific functional sites, such as the active site or the binding pocket, for example, helps to identify functional properties, since most proteins interact with other molecules and function by binding

onto these sites. As the name suggests, a binding site is shaped so that other molecules or proteins can recognize it.

Thus, the structural similarity of proteins is a good measure for the classification of proteins. Furthermore, we believe that it is highly useful for predicting the functionalities and classification of more-newly discovered protein structures.

2.3 Protein Data Bank

There are a number of biological data repositories. The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [4] is one of the most widely used databases and provides vast information on the structure of proteins and other macromolecules. The PDB archive stores general structural descriptions, including primary and secondary structures of proteins, as well as more detailed descriptions, including atomic coordinates. The RCSB Protein Data Bank (<http://www.rcsb.org>) offers a variety of methods such as advanced search options for PDB entry and other useful tools for exploring and visualizing proteins. Several protein comparison tools can be used to analyze sequential and structural relationships based on the representative domains on the website. Such structural information is collected via X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy and is used to dictate relative locations of atoms and identify the coordinates of atoms in the molecule. The coordinate files stored in the PDB archive can be viewed using visualizing tools such as Jmol. These files are downloadable from the server in a variety of types. Jmol[14] is an interactive 3D viewer for molecular

structures and can read over 60 file formats including PDB, CIF, SDF, MOL, and PyMOL. Jmol provides a variety of options for presenting protein structure. A typical PDB formatted file consists of several sections. The title section has a summary of the protein, the summary section goes over primary and secondary structure, the connectivity section describes the bonds and links between sheets and helices, and the coordinate section lists atoms along with 3D coordinates of the atoms in the protein.

CHAPTER 3

METHODOLOGY

3.1 Functional Site Based Analysis

A protein is a large and complex molecule composed of amino acid sequences that fold up into a unique three-dimensional structure. It is believed that this unique three-dimensional structure determines its biological properties and thus that protein function can be identified by detecting local structural similarities [15]. In this way, there are a number of methods dedicated to predicting protein function by means of analyzing similarities in sequence or structure. Most proteins are composed of several hundreds of amino acid sequences and the functional site of a protein can be defined as a common local structure that defines the functionality of a set of proteins. Some computational methods have been developed based on the fact that only a few key amino acids of the functional site in the protein are involved in interacting with other molecules. For instance, the property that proteins bind to other molecules to work as a molecular switch gives rise to the fact that binding sites that interact with other molecules are deeply related to protein functionality. An approach to classify protein kinase based on the binding pockets is a good example [16]. Thus, recognizing a functionally important local structure such as binding sites and functional motifs of a protein is essential in structure based analysis of proteins, and this aspect of protein behavior is also applied to the core of the approach in this paper.

3.1.1 Functional Site of 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 [11]. This superfamily is divided into five major families based on their structural and functional similarities: Rho, Ras, Rab, Ran, and Arf. Rho, Ras, and Rab are the most closely related among the five [17]. The protein members of the Ras superfamily have 40% - 85% of high primary sequence identity, while each subfamily has individual functions and different targets [18]. 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 [19].

There is a particular motif in the proteins of the Ras superfamily that determines the features of each subfamily. Each subfamily either 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 [11]. 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. A comparative analysis based on functional sites begins with finding the p-loop motif and comparing its three-dimensional shape. Table 1 shows the hierarchical relationship of the Ras superfamily and the list of PDB IDs chosen for analysis in this research project.

Table 1: Hierarchy of Ras superfamily and the list of proteins used for SOM analysis

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

3.1.2 Binding Site of Protein Kinase Family

Protein kinases catalyze proteins by attaching phosphate groups to them. For example, protein kinase helps transfer ATP to proteins so that they can be phosphorylated. The sterile (STE) group, which is one of ten human kinase families, including protein kinases, is involved with mitogen-activated protein (MAP) kinases. Three main families in the STE group operate on each other sequentially: STE 20 activates STE11, STE11 activates STE 7, and STE7 directly acts on MAPKs.

STE 20, the largest of the three STE families, can be further divided into the p21-activated kinase (PAK) group and the germinal center kinase (GCK) group. These two groups are involved with interactions dealing with various signaling and regulatory proteins of the cytoskeleton [20].

Table 2 organizes the subfamilies and their members and indicates the binding sites for each member. Not all of the coordinate information on the proteins of this family is available in PDB, so only the proteins with discoverable coordinate data were selected and used in the analysis.

Table 2: Hierarchy of Ste Kinase Family and the 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

3.2 Structural Alignment of Proteins

Protein structure alignment is crucial to computational biology. In particular, the comparison of protein structures is imperative because structural similarities often imply evolutionary relationships or common functional characteristics. Proteins are comprised of amino acids chains, which fold into unique three-dimensional shapes that dictate functionality. In general, structural alignment refers to the three-dimensional structural alignment between two or more proteins without taking into consideration sequence arrangement. Structure alignment is necessary to perform a precise comparison, even before comparing protein structures. Each protein structure has a different size and different coordinates. Protein structures pulled out from the Protein Data Bank (PDB) need to be aligned using a structural alignment tool since the structures may have similar shapes but different 3D orientations. In other words, the

three-dimensional coordinates stored in the PDB file for each individual protein only have the relative location of the atoms in the whole structure. Structural alignment is performed in two ways: locally and globally. By performing structural alignment both locally and globally, it is possible to discover any differences that may come up in the results. Alignment is performed based on the backbone structure of the protein: the skeletal structure composed of α -carbons for each residue.

3.2.1 Local Structural Alignment

The main purpose of local structural alignment is to minimize error by aligning smaller, selected regions without taking into consideration the rest of the structure, before the proteins are compared. In this paper, the local structure states the functional site to be observed and the local structural alignment performed in a pairwise manner, based on the one of the protein structures selected for the analysis.

Protein Local Alignment Tool (PLAT) [21] is a newly developed, web-based local structure alignment tool that performs pairwise alignment. PLAT provides simple but convenient ways to align local structures and makes the process of selecting specific residues to be aligned much easier. The protein data, more specifically the PDB ID and the chain type, is pulled straight from the PDB, and the local region to be aligned is selected as well. Aligned structures can then be viewed in jmol and saved as a .pdb file. Figure 1 includes a screenshot of plat and an example of aligned structure viewed using jmol. The regions shaded in yellow indicate the local structures chosen to be aligned. In order to perform an alignment, the number of the residues chosen should be the same.

After performing a local alignment, plat shows the new origin of the coordinate system and the rotation matrix. In this case, the P-loop structure of every protein is aligned based on the structure of 121P.

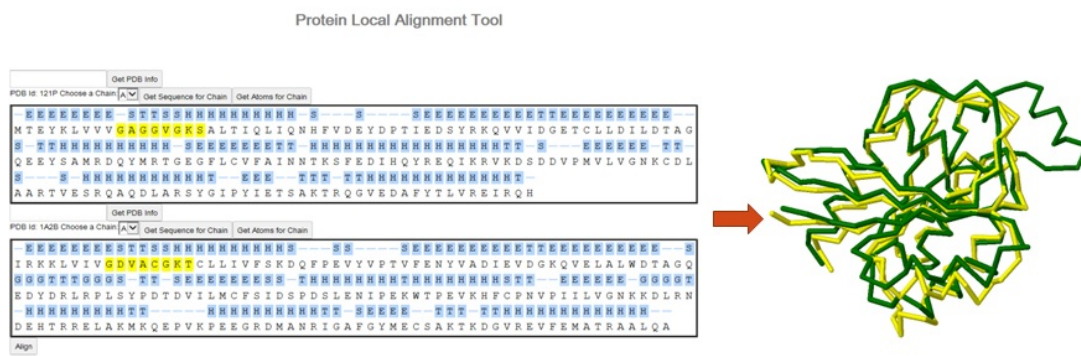


Figure 1: Local Structure Alignment Tool and Visualization of Aligned Structures

3.2.2 Global Structural Alignment

The main purpose of global structural alignment is to find the overall optimized 3D structural alignment. The result will suggest a method for comparing proteins. A superposition of two or more structures is computed by looking at the number of matched α -carbons and the minimal root-mean-square deviation (RMSD). RMSD is widely used to indicate the distance between atoms in superposed structures when comparing the structures of biomolecules. FATCAT (Flexible structure AlignmentT by Chaining Aligned fragment pairs with Twists) [22] provides flexible pairwise 3D structure alignment functions. Comparison starts by searching for AFPs (aligned fragment pairs) between two protein structures (in PDB format), and the algorithm finds optimal alignment by detecting hinges and twisting them during the process of connecting the AFPs. Figure 2 displays the global alignment of 121P and

1A2B, which is the best 3D superposition of matched α -carbons, produced with FATCAT.

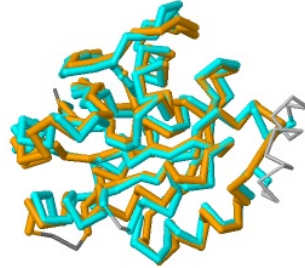


Figure 2: Globally aligned and superposed structures of 121P(light brown) and 1A2B(light blue) using FATCAT

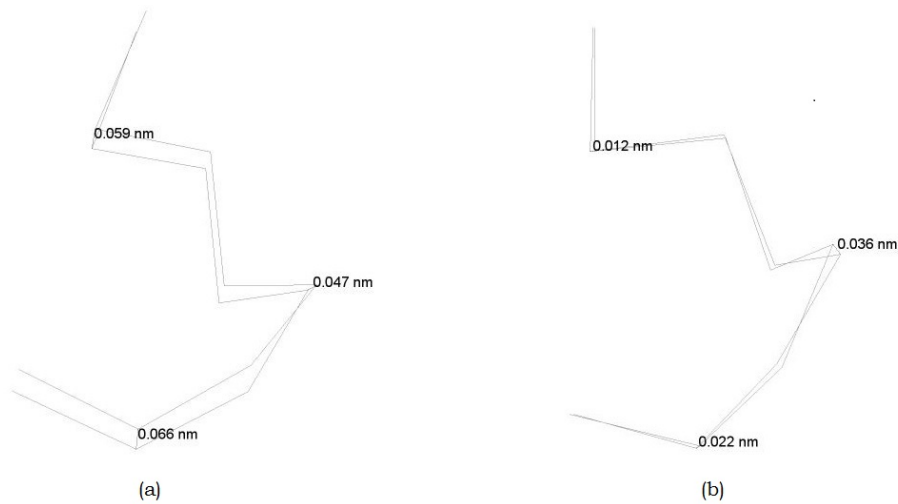


Figure 3: Alignments of p-loop motif of 121P and 1A2B with (a) global alignment, (b) local alignment.

Figure 3 shows the two different alignments of the p-loop motif of 121P and 1A2B using the two alignment techniques: global structural alignment and local

structural alignment, respectively. The six corners and two end points, for a total of eight points, represent the α -carbons of the p-loop structure. jmol is used to visualize the aligned xyz coordinates. The numbers indicate the distance in Å between the two corresponding α -carbons, and we can note in (b) that the local alignment technique tends to align the structure more precisely.

3.3 Preprocessing the Protein Structure Information and Feature Vector Construction

An innovative method is needed to describe the 3D structure of proteins, especially when the structural data is complex. The major steps for preprocessing protein data are summarized in Figure 4. First, the protein structures for proteins under investigation are pulled from the Protein Data Bank (PDB). Proteins are then aligned using either a global or local structural alignment tool. However, protein structures, even after local or global alignment, often still contain irrelevant data pertaining to the functional site. In order to achieve functional site based analysis, the functional sites must be filtered out completely. In order to filter out the functional sites, key structural information must be 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 must be simplified by collecting only the α -carbons in these functional sites. This process provides information on the backbone structure of the functional site only 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 mapped into a linear vector. This vector is called the feature vector of the functional site.

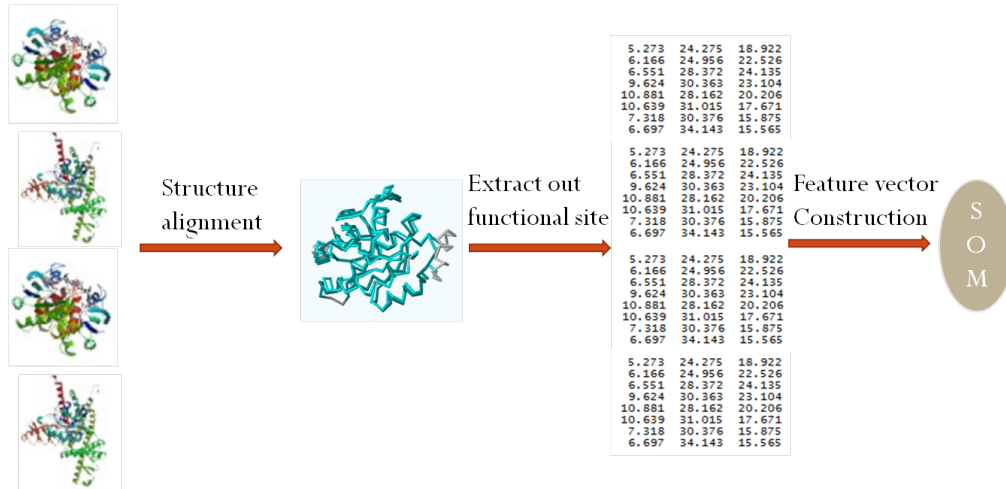


Figure 4: Preprocessing the protein structural data

Figure 5 represents a part of the actual feature vector constructed by the method described above. There are sixteen structures shown among the twenty-six protein structures of the Ras superfamily under investigation. Each structure has two labels, family name and PDB ID. In addition, there are three sets of attributes representing the first three residues (GXX) of the eight residues making up the p-loop motif. Each set shows the x,y, and z coordinates of these residues. As previously mentioned, the p-loop motif has the pattern GXXXXGK[S/T]. For example, G(x,y,z), grouping the attributes X1, X2, and X3, denote the x,y, and z coordinates, respectively, for the first residue G of the p-loop. The eight coordinate sets of the eight residues are unfolded and arranged in the same order as the consensus of the p-loop; thus, there are a total of twenty-four (eight sets of xyz) attributes used to represent the p-loop motif.

	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	2fo1	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

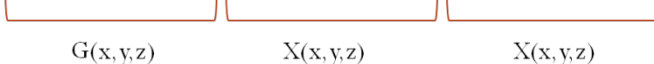


Figure 5: Feature vector construction, unfolded xyz coordinates

CHAPTER 4

Structural Analysis using SOM

4.1 SOM Analysis of the Ras Superfamily

The primary advantage of using Self-Organizing Map (SOM) is the ability to train models in which the categories are not defined. SOM groups together similarities in the data and creates grid maps representing these similarities. Specifically, the geometric similarity of two proteins can be described as the distance between their corresponding atoms [23]. Due to the property of structure and function relationship, proteins are classified into families by structural similarities in their functional sites. The Ras superfamily is a large superfamily consisting of structurally distinguishable families. One way to examine the structural-functional relationship of such proteins is to observe the clustering of the Ras superfamily through the SOM algorithm. All pairwise structural alignments using local and global techniques are performed based on the structure of 121P.

4.1.1 SOM with Local Structural Alignment

Local structural alignment focuses on more specific regions without taking into consideration any peripheral structures. 121P is selected as the base structure, and an alignment with each protein based on the p-loop structure is thus performed in pairwise manner. As a result, the aligned structure of each protein is preserved, while the coordinate data of the p-loop structure is taken out of the aligned structure. The

feature vector is composed of the three-dimensional coordinate data on the eight α -carbons in the eight residues of the p-loop motif. Figure 6 shows the SOM result obtained by the local alignment technique. The SOM result was generated with a size of 25 X 20 with 500 iterations. SOM generates a different map at every execution. The number of iterations was increased by 100 so that the map converged, avoiding overfitting and demonstrating a correct model. It can be noted that most proteins were clustered appropriately, with only a few proteins not. Starbursts [24] in the map make it easier to recognize each cluster easily. Most of the Ras family can be found in the upper left corner of the map, while most of the Ran family can be found in the bottom left corner. Similarly, most of the Arf family can be found in the bottom right corner of the map, while most of the Rho family can be found in the upper right corner. One thing to note is that the Rab family tended to disperse more so than the other families. It is also important to note that the Rho, Ras, and Rab families tended to be closer or more mixed with each other because they were more closely related among the five subfamilies.

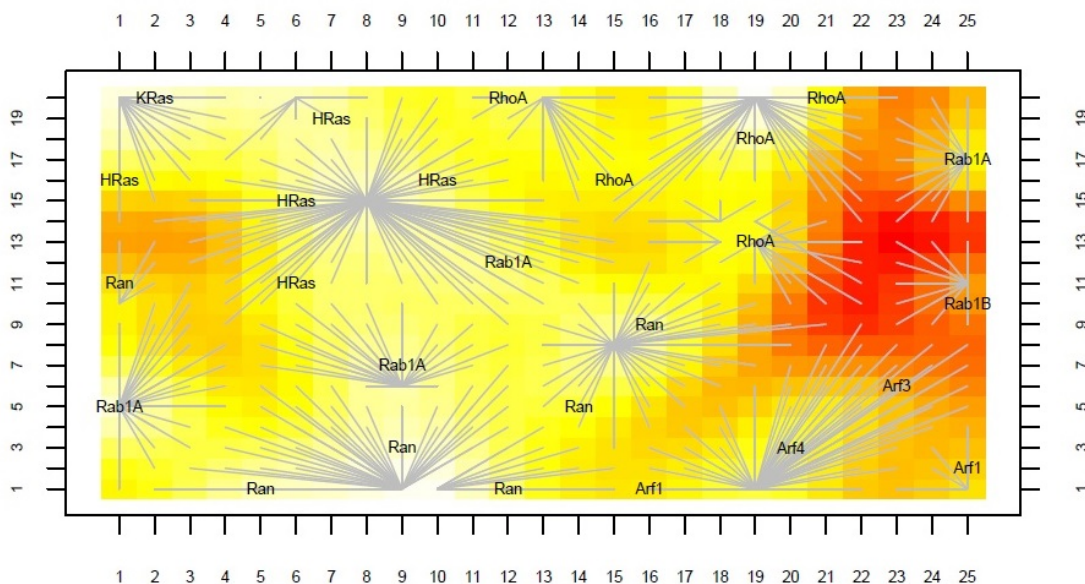


Figure 6: SOM result with local alignment (25 x 20 SOM with 500 iterations)

4.1.2 SOM with Global Alignment with Flexibility

Like local alignment, global alignment is performed in a pairwise manner, based on the structure of 121P. Unlike local alignment, however, it is executed over the entire structure, not just the p-loop structure. FATCAT is a flexible structure alignment tool allowing twists around hinges. If only a query structure is provided, FATCAT will search mainly for similar structures. On the other hand, if both a query structure and target structure are provided, it will find the structural superposition by comparing their global structures. Three-dimensional coordinate information on the functional sites is then extracted from these globally superposed structures.

Figure 7 represents the 25 x 20 SOM result for the Ras superfamily using global alignment. The size of the map was matched to the size of the SOM map created following local alignment, and the map was trained until it converged (200

iterations). Most centers of crowded starbursts display clusters for each family. The most distinct cluster is that of the Ras family in the center of the map, above the clusters where the Arf and Ran families are located. The two clusters for the Rab family are appeared but cannot be seen in the SOM of Figure 5. Very few structures, Arf3 and are mis-clustered overall.

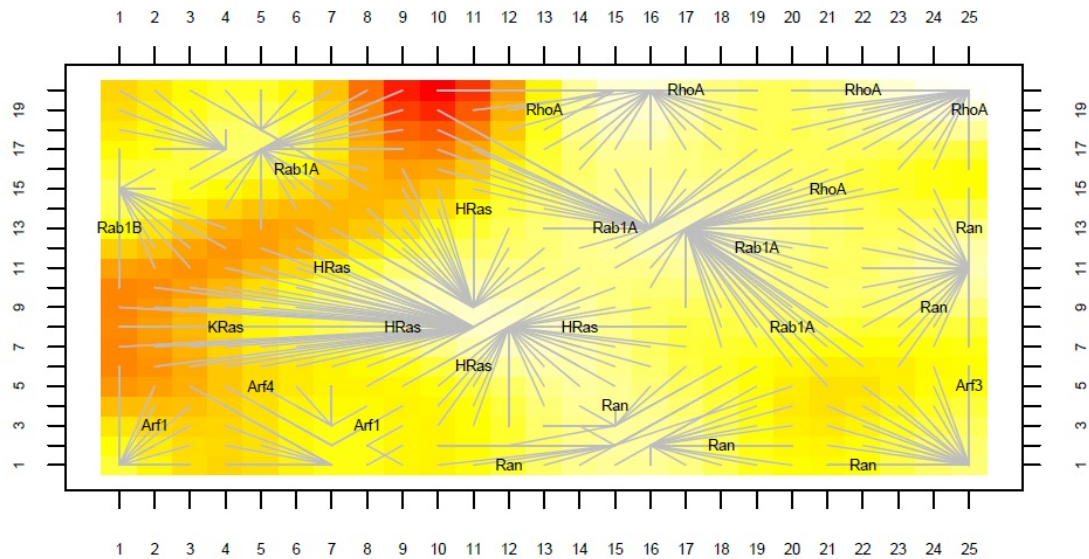


Figure 7: SOM result with global alignment (25 x 20 SOM with 200 iterations)

Although the SOM following global alignment looks a bit more organized than the one following local alignment, it is not completely obvious as to which map and thus which alignment technique, represents clustering better. Because of this, another clustering method is adopted to see the difference even better.

4.2 SOM result of Protein Kinase Family

In order to validate the assumptions of functional site based analysis, another protein group is adopted. STE group is one of the protein kinase families, and it contains three homologs of yeast, Sterile 7, Sterile 11, and Sterile 20. ATP binding regions of STE family are selected as functional sites of this group, and local and global structural alignments are performed prior to the application of SOM.

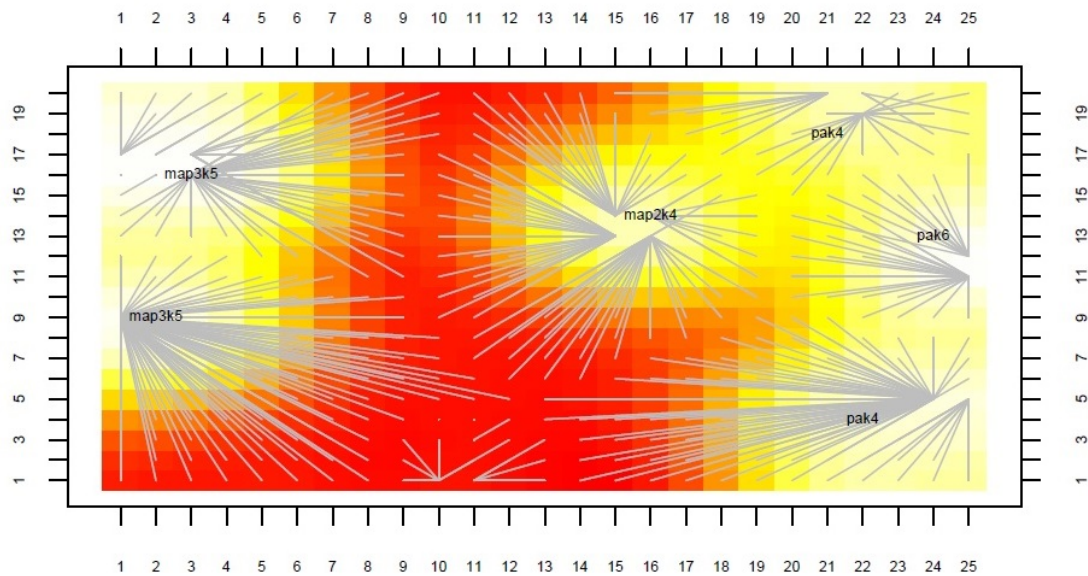


Figure 8: 25x20 SOM of STE kinase family with local alignment (50 iterations)

Figure 8 shows the 25 x 20 SOM result with 50 iterations of training. The map exhibits three distinctive clusters, one on the upper left corner (STE 11), one on the right side (STE 20), and one in the center (STE 7). The classifications are distinguishable by the nodes' different shades of color and the distances between the clusters.

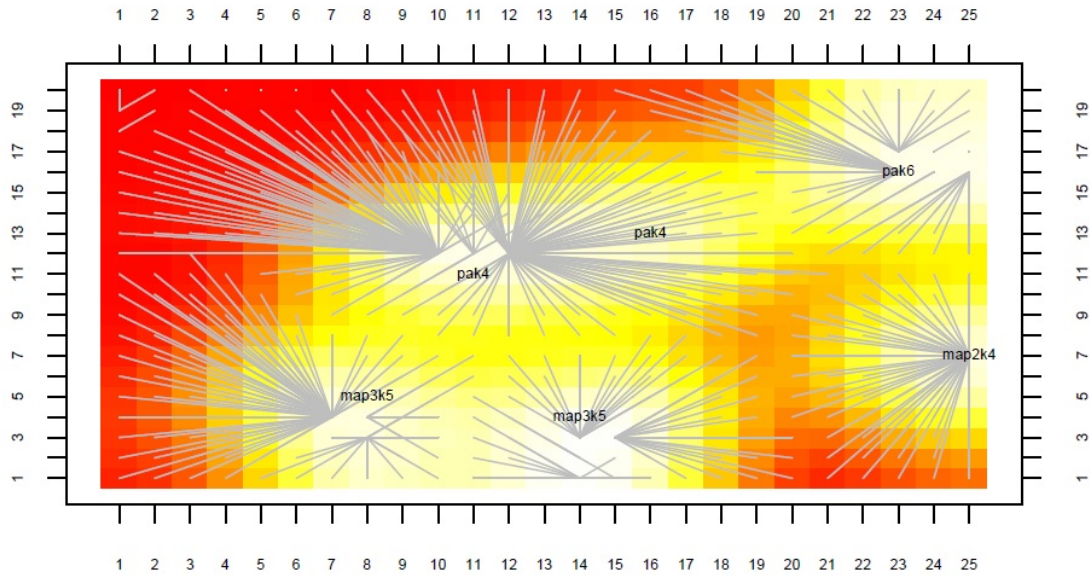


Figure 9: 25x20 SOM of STE kinase family with global alignment (50 iterations)

In Figure 9, the SOM result of STE family is obtained following the global alignment technique. STE 7 is on the right-hand side of the map, the cluster for STE 11 is on the bottom part, and the cluster for STE 20 is on the center of the upper right corner. Like local alignment technique, three visible clusters for the subfamilies are appeared according to the structural similarities among them.

4.3 Comparing Hierarchical Clustering with the SOM results

Yet another clustering method can be adopted to validate the approach suggested in this research project. A hierarchical clustering technique, using dendrograms, can also help create visualizations of the many hierarchical relationships of protein families. Difference or distance in data can be demonstrated in one of several different ways. One is by calculating the distance matrix between the rows or

columns of the data matrix. The distance matrix generated from the original data matrix can then be visualized using cluster dendrograms. Dendrograms are treelike graphs that arrange the clustering of hierarchical structures between data based on their distances to each other. Dendrograms help identify relationships between data via graphical representations. The data matrix is the feature vector for SOM, which consists of the functional sites' three-dimensional coordinates collected after local or global alignment with dimensions of (24 x the number of structures). Hierarchical clustering technique is applied to the distance matrix calculated from the feature vector. In other words, the smaller the distance from the joint in the tree, the more similar they are to each other, and the greater the distance, the more different they are. The hierarchical clustering technique is a simple but effective way to view the similarities or differences of protein structures, given that it is possible to compute the distance matrix. In addition, this clustering result can be a suitable way to compare the SOM results obtained previously.

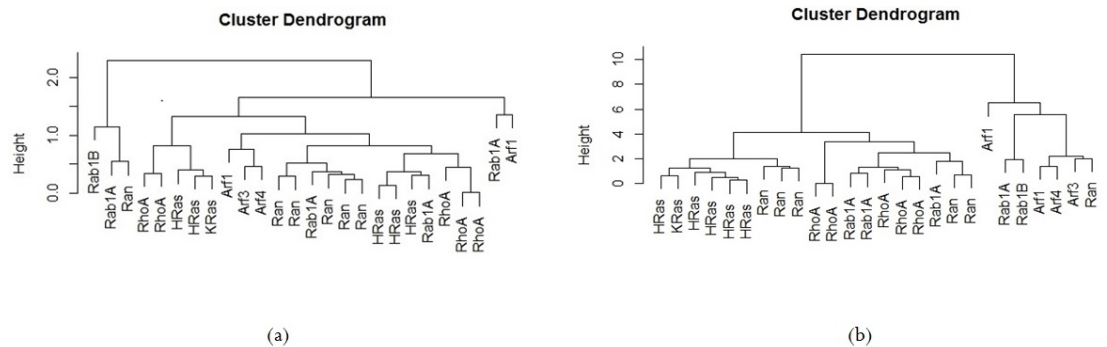


Figure 10: Cluster Dendrograms with results of (a) Local Alignment and (b) Global Alignment of Ras Superfamily

Figure 10(a) is the cluster dendrogram generated from the distance matrix of the feature vector with the coordinate result of local structural alignment, and (b) is the cluster dendrogram generated from the distance matrix of the feature vector with the coordinate result of global structural alignment. Leaves on the tree represent protein structures, with each structure labeled with its subfamily name. Most proteins in the same subfamily are clustered together, with a few minor exceptions. On the left side of the trees, the height value is shown. The height represents the difference between clusters; thus (b) shows stronger clustering than (a). Also, (b) has more homogenous clusters, especially for the Ras, Rho and Arf families. Therefore, it is possible to conclude that the results of global alignment are clearly more useful.

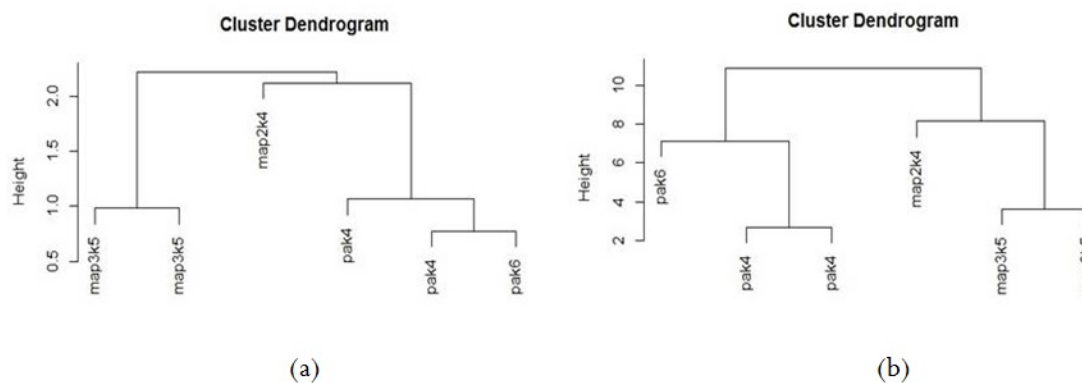


Figure 11: Cluster Dendrograms with results of (a) Local Alignment and (b) Global Alignment of STE Group

Figure 11 displays the two cluster dendrograms of the STE kinase group, (a) with the local alignment technique and (b) with the global alignment technique. In this case, the height value of (b) is bigger than (a), so (b) shows stronger clusters.

Both SOM and hierarchical clustering trees allow for easy visualization and interpretation. Dendrograms are relatively easy to read and interpret until the size of the clustering tree gets much bigger and more complicated. SOM maps, on the other hand, although harder to understand at first, are more useful than dendrograms when the number of observations gets much larger.

CHAPTER 5

CONCLUSION

5.1 Conclusions

We have developed a unique method for comparing proteins and for discovering similarities and differences between functional sites via an unsupervised machine learning technique using SOM. SOM has the superior ability to recognize patterns in data. It maps structural patterns in protein families into low-dimensional grid maps by grouping proteins with similar structural patterns closer together. It is difficult to understand the relationships embedded in high-dimensional data simply by inspection. SOM helps to identify such relationships, especially among complex protein structures, through visualizations, which minimize the loss of information.

The nature of protein conformation indicates that structure and function are deeply related. The function of a protein is determined primarily by its tertiary structure, and then, although to a lesser extent, by its primary sequence. In this way, the functional core of the protein plays a critical role in classifying proteins into their respective subfamilies. The study of structural analysis based on functional sites of proteins began by merely identifying functionally important local structures. SOM expanded this study by investigating and comparing the three-dimensional shape of these functionally important local structures. Prior to the construction of SOM models, structural alignments were used solely to minimize errors existing in the coordinate system between protein structures. PLAT, a newly developed web-based protein local

alignment tool, allows users to now select specific residues and align structures focused on these residues. Unlike local alignment, global alignment demonstrates that other domain structures affect the alignment of functional sites. Thus, it is remarkable that the small distortions in the functional sites extracted from globally aligned structures contributed to better clustering results than local alignment structures did. The convergence rate of SOM made certain the reliability of SOM results. In a functional site-based analysis, similarities between proteins are found by using relatively small local structures and excluding all other unrelated structures. SOM successfully identified the clusters of subfamilies of two protein groups, the Ras superfamily and the STE kinase family, proving the structure-function relationship of proteins and the effectiveness of the functional site based approach.

The most notable improvements from preliminary research are, by far, the automated local structure extraction technique and the structural alignment technique (e.g., the backbone of the p-loop motif). This paper also introduced SOM's simple but effective feature vector construction component by unfolding the coordinate data on protein structure.

5.2 Future Work

Although the analysis was conducted in regards to two large protein families, only a limited number of proteins from each family were chosen. In order to consolidate the conclusions reached in this research project, more protein structures or protein groups should be added. If not, more domain structures can be added (e.g., the whole G-domain structure of the Ras superfamily can be used) so that the analysis is

not just restricted to one functional site. Study of these strong predictive structural features will provide guidance in classification of newly discovered protein structures. Both of these improvements enable broad understanding on the classification of protein structures.

LIST OF REFERENCES

“Self Organizing Maps” Aug. 2014[online] Available:

http://en.wikipedia.org/wiki/Self-organizing_map

“Protein family” Nov. 2014[online] Available:

http://en.wikipedia.org/wiki/Protein_family

“Protein Function” Feb. 2015[online] Available:

<http://www.nature.com/scitable/topicpage/protein-function-14123348>

“Introduction to protein classification at the EBI” Aug. 2014[online] Available:

<http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi>

“Understanding PDB Data : Looking at Structures” Nov. 2014[online] Available:

http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/intro.html

Nejc C, et al. “Protein-Protein Binding Site Prediction by Local Structural Alignment”

J. Chem, Inf. Model, vol. 50, pp:1906-1913, 2010.

“Jmol, An open-source Java viewer for chemical structures in 3D” Dec 2014[online]

Available:

<http://jmol.sourceforge.net/>

“Kinase Group STE” Feb. 2015[online] Available:

http://kinase.com/wiki/index.php/Kinase_Group_STE

“The STE Group” Feb. 2015[online] Available:

<http://www.compbio.dundee.ac.uk/kinomer/families/STE.html>

“Chapter 7 Hierarchical cluster analysis” Feb. 2015 [online] Available:
www.econ.upf.edu/~michael/stanford/maeb7.pdf

“What is Protein Binding?” Dec 2014[online] Available:
www.wisegeek.org/what-is-protien-binding.htm

“Structural alignment” Feb 2015 [online] Available:
http://en.wikipedia.org/wiki/Structural_alignment

“Root-mean-square-deviation of atomic positions” Mar 2015 [online] Available:
https://en.wikipedia.org/wiki/Root-mean-square_deviation_of_atomic_positions

“Site” Mar 2015 [online] Available:
<http://proteopedia.org/wiki/index.php/Site>

BIBLIOGRAPHY

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., “Basic local alignment search tool.” *Journal of Molecular Biology*, vol. 215, pp:403-410, 1990.

[2] Berg, J.M., Tymoczko, J.L., and Stryer, L. “Biochemistry” 5th ed. New York: W H Freeman, Chapter 3, Protein Structure and Function. Available from:
<http://www.ncbi.nlm.nih.gov/books/NBK21177/>

[3] Rojas, A.M., Fuentes, G., Rausell, A., and Valencia, A., “The Ras protein superfamily: Evolutionary tree and role of conserved amino acids.” *The Journal of Cell Biology*, vol. 196, pp:189-201, 2012.

[4] Rose P.W., et al., “The RCSB Protein Data Bank: new resources for research and education.” *Nucleic Acids Research*, vol. 41, pp:475-482, 2013.

[5] Hamel, L., Sun, G., and Zhang, J., “Toward Protein Structure Analysis with Self-Organizing Maps” *Computational Intelligence in Bioinformatics and Computational Biology*, 2005.

[6] Pazel, D.P., “DS-viewer—an interactive graphical data structure presentation facility.” *IBM Systems Journal*, vol. 28, Issue 2, pp:307-323, 1989.

[7] Murzin A.G., Brenner S.E., Hubbard T.J.P., and Chothia C. “**SCOP**: a structural classification of proteins database for the investigation of sequences and structures.” *Journal of Molecular Biology*, vol. 247, pp:536-540, 1995.

[8] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. “CATH- a hierarchic classification of protein domain structures” *Structure*, vol. 5, pp:1093-1109, 1997.

- [9] Hamel, L., Ott, B., and Breard, G. “popsom: Self-Organizing Maps With Population Based Convergence Criterion” Available from: <http://cran.r-project.org/web/packages/popsom/index.html>
- [10] Kohonen, T., “Self-organizing maps” 3rd ed. Berlin, New York, Springer, 2001.
- [11] Wennerberg, K., Rossman, K.L., and Der, C.J. “The Ras superfamily at a glance” *Journal of Cell Science*, vol. 118, pp: 843-846, 2005.
- [12] Xu, Y., Xu, D., and Liang, J. “Computational Methods for Protein Structure Prediction and Modeling” 2007.
- [13] Hegyi, H., and Gerstein, M. “The relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome”, *Journal of Molecular Biology*, vol. 288, pp: 147-164, 1999
- [14] “Jmol: an open source Java viewer for chemical structures in 3D.” Available from: <http://www.jmol.org/>
- [15] Najmanovich, R.J., Torrance, J.W., and Thornton J.M., “Prediction of Protein Function from Structure: Insights from Methods for the Detection of Local Structural Similarities” *Bio Techniques*, vol. 38, no. 6, pp: 847-851, 2005 mm
- [16] Kuhn, D., Wescamp, N., Hullermeier, E., and Klebe, G. “Functional Classification of Protein Kinase Binding Sites Using Cavbase”, *ChemMedChem*, vol. 2, pp:1432-1447, 2007.
- [17] Kahn, R.A., Der, C.J. and Bokoch, G.M. “The Ras superfamily of GTP-binding proteins: guidelines on nomenclature.” *The FASEB Journal*, vol.6, no.8, pp: 2512-2513, 1992.

[18] Kennedy, M.B., Beale, H.C., Carlisle, H.J., and Washburn, L.R. "Achieving signalling specificity: the Ras superfamily" *Nature Reviews Neuroscience*, vol. 6, pp: 423-434, 2005

[19] Alberts B, et al. "Molecular biology of the cell", 4th ed. New York; Garland Science, 2002.

[20] Dan, I., Norinobu M., Watanabe N.M., and Kusumi, A. "The Ste20 group kinases as regulators of MAP kinase cascades." *Trends in cell biology*, vol. 11, pp: 220-230, 2001.

[21] Jaegle, S. "Protein Local Alignment Tool" Available from:
<http://plat.cs.uri.edu/plat/>

[22] Ye, Y., Godzik, A., "Flexible structure alignment by chaining aligned fragment pairs allowing twists", *Bioinformatics* Vol. 19, Suppl. 2, pp:246-255, 2003

[23] Lisa H. et al, "Protein Structure Comparison by Alignment of Distance Matrices" *Journal of Molecular Biology*, vol. 233, pp: 123-138, 1993

[24] Hamel, L., Ott, H.B., "A Population Based Convergence Criterion for Self-Organizing Maps"

[25] Maiorov VN., Crippen GM., "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins", *Journal of Molecular Biology*, vol. 235, pp: 625-634, 1994