Toward Protein Structure Analysis with Self-Organizing Maps

Lutz Hamel
Department of Computer Science
University of Rhode Island
Kingston, RI 02881
hamel@cs.uri.edu

Gongqin Sun
Department of Cell and Molecular Biology
University of Rhode Island
Kingston, RI 02881
gsun@uri.edu

Jing Zhang
Department of Computer Science
University of Rhode Island
Kingston, RI 02881
zhangji@cs.uri.edu

Abstract - **Establishing structure-function relationships on the proteomic scale is a unique challenge faced by bioinformatics and molecular biosciences. Large protein families represent natural libraries of analogues of a given catalytic or protein function, thus making them ideal targets for the investigation of structure-function relationships in proteins. To this end, we have developed a new technique for analyzing large amounts of detailed molecular structure information focusing on the functional centers of homologous proteins. Our approach uses unsupervised machine learning, in particular, self-organizing maps. The information captured by a self-organizing map and stored in its reference models highlights the essential structure of the proteins under investigation and can be effectively used to study detailed structural differences and similarities among homologous proteins. Our preliminary results obtained with a prototype based on these techniques demonstrate that we can classify proteins and identify common and unique structures within a family and, more importantly, identify common and unique structural features of different conformations of the same protein. The approach developed here outperforms many of today's structure analysis tools. These tools are usually either limited by the number of proteins they can process at the same time or they are limited by the structural resolution they can accommodate, that is, many of the structural analysis tools that can handle multiple proteins at the same time limit themselves to secondary structure analysis and therefore miss fine structural nuances within proteins. It is worthwhile noting that the ability of our approach to analyze different conformations of the same protein is beyond the capabilities of multiple residue sequence alignment techniques.**

I. INTRODUCTION

The eukaryotic proteome contains a number of large protein families, including the protein kinases, GTPases, and G-protein coupled receptors, vital to the proper development and functioning of the cells in an organism. For example, the protein kinase family contains over 500 proteins that are involved in various specific cellular signal pathways controlling virtually all aspects of cellular function [1]. Disruption of their function or regulation by gene amplification or mutations often leads to a number of serious diseases, making many of the protein families important targets for drug discovery.

Large protein families represent natural libraries of analogues of a given catalytic or protein function ideal for the investigation of structure-function relationships in proteins. In particular, proteins that belong to large families tend to perform similar functions with varying specificities. They usually have similar overall structures with minor variations in certain substructures. In particular, functional centers enabling the protein functions are highly conserved within protein families. While these conserved structures are responsible for functions that are common to all the proteins within a family, the variable substructures tend to be responsible for protein specific functions. It is the combination of these common and unique functions that gives each member a unique functional identity, allowing each to perform specific roles in the cell, and to respond to different regulatory signals.

In order to understand the structural basis of functional similarities and specificities of proteins, it is essential that we analyze the structural information in detail and correlate the structural patterns to the functional patterns. It is also important to distinguish between different functional conformations of the same protein. Here, the function of the protein or enzyme is regulated by covalent or non-covalent modifications that result in conformational changes. Many proteins have been crystallized in multiple functional states. The detailed analysis of the structural information of such multiple conformations is beyond the capabilities of tools which are mostly concerned with structure alignment based on the secondary structure of proteins, e.g., VAST [2, 3], DALI [4, 5]. It is also beyond structural comparison tools such as K2 [6] due to the fact that these tools can only perform pair wise comparisons and do not perform any kind of feature selection.

In this paper, we develop an approach to protein structure analysis based on unsupervised learning that goes beyond looking at secondary structure. It distinguishes itself from structural comparison tools such as K2 by the fact that it can process detailed three dimensional structure information of more than two proteins at a time. Just as in structural alignment tools we only consider the structure of the protein proper; ligands and other complexes are ignored. Theoretically there is no limit as to how many proteins we can process at a time; it is purely a function of available computing power. Another essential difference between tools such as K2 and our approach is that we compute the relative importance of differences and similarities between proteins whereas tools such as K2 leave this interpretation up to the user.

To summarize, our unsupervised machine learning approach to the structural analysis of proteins based on self-organizing maps is driven by the following observations:

i) The poignant structural differences or similarities between proteins in a protein family are function specific, that is, the specificity of a protein function is supported by local structural variations around a particular functional center [7].

ii) Machine learning allows us to discern structural patterns by considering many proteins at the same time [8].

iii) Machine learning constructs patterns by considering highly predictive or the most relevant structures. Part of the machine learning process is the differentiation of relevant versus non-relevant features [9-11].

iv) A fairly large number of the protein kinase structures have already been deposited in PDB [12] (108 protein structures representing 46 unique protein kinases, ca. December 2004) and more are added continuously. This is also true for other large protein families such as the GTPases [13].

v) The detailed structure analysis envisioned here is beyond the capabilities of most other structural comparison tools.

The remainder of the paper is structured as follows: Section II briefly introduces self-organizing maps. We explain functional center-based protein structure analysis in Section III. Our first set of experiments with GTPases using this technique is discussed in Section IV and the second set of experiments with human protein kinases is discussed in Section V. In Section VI we highlight some related work and we conclude the paper with final remarks and notes on further research in Section VII.

II. Self-Organizing Maps

Self-organizing maps [14] were introduced by Kohonen in 1982 and can be viewed as tools to visualize structure in high-dimensional data [15]. Self-organizing maps are considered members of the class of unsupervised machine learning algorithms, since they do not require a predefined concept but will learn the structure of a target domain without supervision [16].

Typically, a self-organizing map consists of a rectangular grid of processing units. Multidimensional observations are represented as feature vectors. Each processing unit in the self-organizing map also consists of a feature vector called a reference vector or reference model. The goal of the map is to assign values to the reference models on the map in such a way that all observations can be represented on the map with the smallest possible error. However, the map is constructed under constraints similar to regression surfaces in multiple-regression analysis in the sense that the reference models cannot take on arbitrary values but are subject to a smoothing function called the neighborhood function. During training the values of the reference models on the map become ordered so that similar reference models are close to each other on the map and dissimilar ones are further apart from each other.

The training of the map is carried out by a sequential regression process, where $t = 1,2,...$ is the step index. For each observation $\mathbf{x}(t)$, we first identify the index c of some reference model which represents the best match in terms of Euclidean distance by the condition,

$$c = \arg\min_{i} \left\| \mathbf{x}(t) - \mathbf{m}_i(t) \right\|, \forall i . \qquad (1)$$

Here, the index i ranges over all reference models on the map. The construction $\|\boldsymbol{x} - \boldsymbol{y}\|$ represents the Euclidean distance between feature vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Next, all reference models on the map are updated with the following regression rule where model index c is the reference model index as computed in (1),

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}[\mathbf{x}(t) - \mathbf{m}_i(t)], \forall i . \qquad (2)$$

Here h_{ci} is the neighborhood function that is defined as follows,

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

$|c - i|$ represents the distance between the best matching reference model c and some other reference model i on the map, β is the neighborhood distance and α is the learning rate. It is customary to express α and β also as functions of time. This regression is usually repeated over the available observations many times during the training phase of the map.

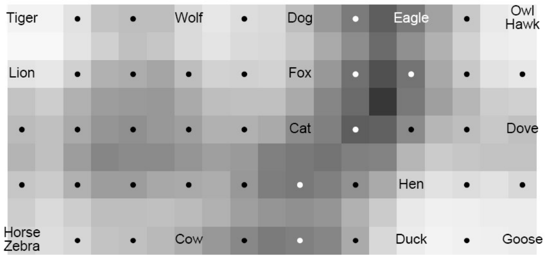

Fig. 1: Mapping animals on to a self-organizing map.

An advantage of self-organizing maps is that they have an appealing visual representation. Fig. 1 shows animals mapped onto a self-organizing map. Each animal is described by a set of 13 features [14] such as how many legs, does it possess feathers, does it hunt, *etc*. Each square in the map represents a reference model. The shading of the map represents the level of quantization or mapping error for the map: light shading represents a small quantization error; dark shading represents a large quantization error. Contiguous areas of low quantization error represent clusters of similar entities.

For example, in Fig. 1 we find two major clusters: mammals and birds. Within these major clusters we can find areas of small quantization error representing sub-clusters such as large predatory mammals in the top left corner of the map and domesticated birds in the bottom right corner. The same reasoning applies to the maps computed in this paper where the structure of proteins is described by appropriate feature vectors and the resulting maps display clusters of similar protein structures.

III. Functional Center-Based Analysis

In the approach discussed here we assume that the functional center is the most conserved and stable structure across individual proteins within a family. It is the functional center that is essential for the core functions of a protein with peripheral structures playing important roles in assisting and differentiating the functional center. Comparing proteins within a protein family based on the molecular structure surrounding a chosen functional center provides a detailed view of the structure-function relationship of a given functional site on a protein by protein basis.

In order to maximize the possibility of extracting interesting structural patterns around the functional center we use local protein alignment techniques. Local alignment techniques tend to minimize the local alignment error compared to global alignment techniques. Fig. 2 shows the alignment of the catalytic loops for the protein kinases 1FPU and 1PHK with a) a global technique [17] and b) a local technique [18]. As can be seen from Fig. 2, the alignment error of the global alignment technique can be substantial due to the global optimization criterion. Global alignment errors are often of the same order of magnitude of the typical resolution in the crystallographic process and therefore there exists the distinct possibility that this global alignment error obliterates important structural patterns surrounding a functional center.

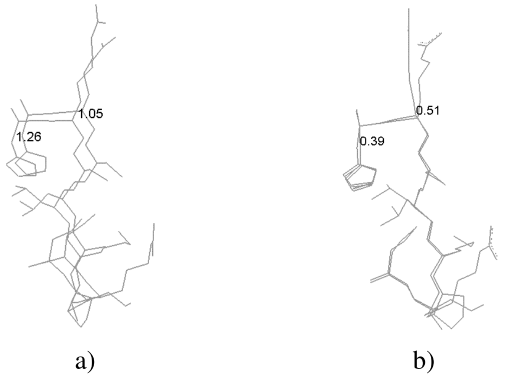

Fig. 2: Global versus local structural alignment of proteins: a) global alignment with FAST, b) local alignment with DS Viewer. The numbers indicate the distance in Å between the residues of the aligned proteins.

In our current prototype active site structures are extracted with a filter that uses the coordinates of a given residue as the functional center and the size of the analysis radius. The resulting protein fragments of the functional sites are then read into a tool such as DS ViewerPro [18] for local alignment. Fig. 3 illustrates this process for two proteins. The functional centers of both proteins are extracted - black circles in the original protein structures a) and b) - and then locally aligned in c).

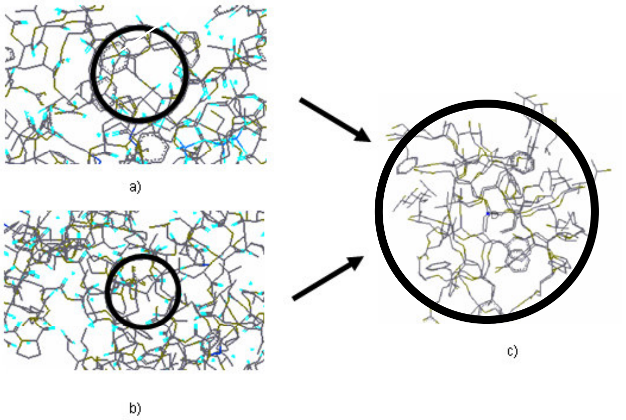

Fig.3: Alignment of active site structures in proteins; a) active site of cAMP-dependent pro.ein-kinase (1ATP), b) active site of glycogen synthase kinase-3β (1GNG), c) the extracted and locally aligned structures surrounding the active sites are shown.

The extracted and locally aligned protein structures are then encoded in such a way that they become amenable to mechanical analysis. This encoding encompasses a number of steps and is conceptually similar to the process in [19]. Fig. 4 illustrates this process. First we represent protein structures by the α-carbons of their amino acid residues. We then normalize the positions of these α-carbons according to a grid structure with a user defined resolution (typically chosen to be close to the resolution of the crystallographic process). We call the resulting structures our *normalized protein models*. Finally, we use the grid structure as a way to partition the space holding the protein. Each subspace in the grid is assigned either a 1 or 0 depending on whether it holds a normalized α-carbon atom or not, respectively. Finally, we unfold the 3D subspace structure into a linear feature vector where each element of the feature vector describes the state of exactly one subspace. This gives rise to a feature vector which holds a 0 or a 1 at each subspace location depending whether the corresponding subspace holds an α-carbon atom or not.

We can view our feature vector computation as a transformation from 3D protein structure space into a k-dimensional feature space where k is the number of subunits of the original 3D protein space. In this k-dimensional feature space each protein is represented by a k-dimensional bit vector and proteins appear as points in this high-dimensional feature space. Moreover, proteins with similar structure will appear close together in this feature space, whereas, proteins with

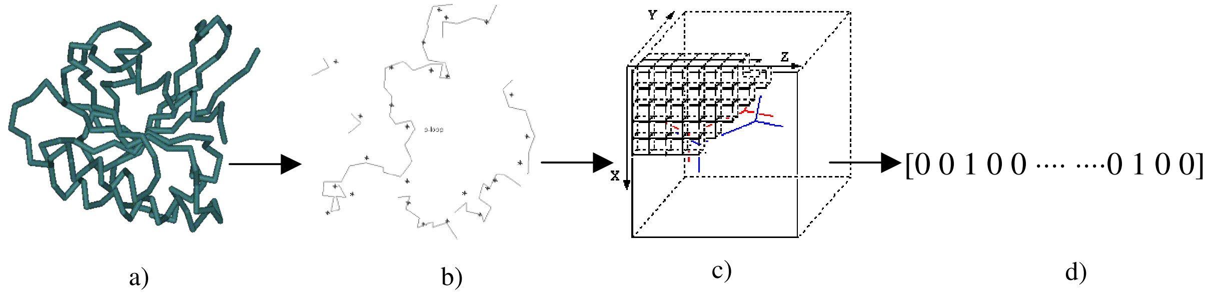

Fig. 4: Protein feature vector construction: a) the 3D structure of a protein without side-chains; b) the normalized structure of the functional center of the protein, the crosses pinpoint the normalized locations of the α-carbons representing our normalized model; c) encoding the normalized model by using cubic subunits; if there is a normalized α-carbon atom in a cubic subunit then the subunit is assigned a 1, otherwise it is assigned a 0; d) the 3D structure of the cubic subunits is unfolded giving rise to a one dimensional feature vector describing the structure of the protein; each position in the feature vector describes the state of a single subunit of the original 3D structure.

dissimilar structure will appear further apart. The self-organizing map algorithm investigates this space for protein similarity/dissimilarity and the structure of the feature space can then be visualized with the typical 2D SOM visualization.

The workflow of our approach is summarized in Fig. 5. We start with a set of PDB files describing the 3D structures of the proteins under investigation. We decide on a functional center *F* and an analysis horizon *r*. We then extract the relevant structures and perform a local alignment. Finally, we compute the normalized models and the corresponding feature vectors which are then submitted to the self-organizing map for analysis. Once the analysis is completed the characteristic 2D visualization for the self-organizing map can be obtained.

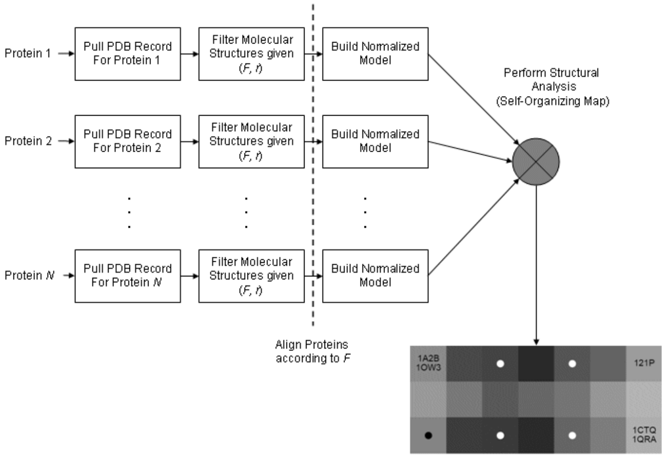
Protein 1
Pull PDB Record For Protein 1
Filter Molecular Structures given (F, r)
Build Normalized Model
Protein 2
Pull PDB Record For Protein 2
Filter Molecular Structures given (F, r)
Build Normalized Model
Protein N
Pull PDB Record For Protein N
Filter Molecular Structures given (F, r)
Build Normalized Model
Perform Structural Analysis (Self-Organizing Map)
Align Proteins according to F
1A2B 1OW3
121P
1CTQ 1QRA

Fig. 5: Summary of Workflow.

IV. Structural Classification Of Proteins

The small GTPases include two large subfamilies, the Rho GTPases and the Ras GTPases [13]. We used two Rho GTPases, 1A2B and 1OW3, and three Ras GTPases, 121P, 1CTQ, and 1QRA for this experiment. One of the highly preserved Glycine residues on the respective p-loops was defined as the reaction center *F* and the analysis horizon *r* was set to be 10Å.

The question we investigated in this experiment was: *Can our technique structurally distinguish between Rho and Ras GTPases?*

Fig. 6 shows that our technique can structurally distinguish between the Rho and Ras GTPases: Rho GTPases appear on the left side of the self-organizing map and Ras GTPases appear on the right side of the map. The fact that the proteins 1A2B and 1OW3 are mapped to one square and the proteins 1CTQ and 1QRA are mapped to another square, respectively, means that each pair shares substantial structure.

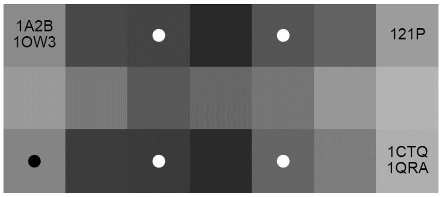

Fig. 6: Self-organizing map showing the structure analysis of small GTPases: Rho GTPases appear on the left side of the map; Ras GTPases appear on the right.

It is comforting that this agrees with the phylogenetic tree for these proteins as in Fig. 7 obtained with ClustalW [20]. This allows us to conclude that structure can be used to classify the proteins. It is perhaps also surprising that our small analysis horizon of 10Å suffices to identify the characteristic structural features of each subfamily reinforcing the power of functional center-based analysis. If we were to identify the functional properties of these GTPases subfamilies we could directly relate them to the structural patterns identified here.

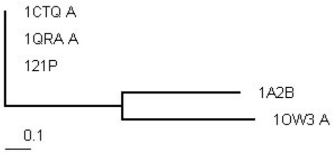

Fig. 7: A phylogenetic tree of the five GTPases.

V. Analyzing Protein Tyrosine Kinase Regulation

Protein tyrosine kinases are a family of important enzymes in cellular regulation [1]. Their activities are often under the control of multiple activation and inactivation mechanisms. It is difficult to elucidate the structural basis of their activation and inactivation. Even in cases where the structures of both the activated and inactivated kinases are available, it is still difficult to determine what conformational changes are responsible for the activation or inactivation. Since the active and inactive protein tyrosine kinases are different conformations of the same protein with the same primary sequence, traditional sequence alignment is completely useless for this analysis (as we will demonstrate). We determined whether our technique can distinguish active versus inactive kinases and identify the conformational features that make a kinase active or inactive. For this purpose we chose two families (Table 1). Csk and Src/Lck represent two distinct protein tyrosine kinase families with different regulatory mechanisms that lead to activation and inactivation. The catalytic domain of Csk is activated by the presence of the regulatory domains, 1K9A_A is the structure of the full length and active Csk, and 1BYG is the structure of only the catalytic domain, and represents an inactive Csk structure. Src/Lck is inactivated by phosphorylation on Tyr527. 3LCK is the structure of unphosphorylated, and thus active Lck, while 2Src is the structure of the Tyr527 phosphorylated and inactive Src. Not only is it important to identify the conformational features that make a kinase active or inactive, it is also interesting to determine if different mechanisms of activation/inactivation lead to the same conformational changes in different kinase families or not.

We break our analysis into three parts. First, we use our SOM-based technique. Second, we use sequence-based techniques. Finally, we use structure comparison and alignment tools for the analysis.

TABLE I
ACTIVE AND INACTIVE HUMAN KINASES

Kinase Family	*Active*	*Inactive*
Csk	1K9A_A*	1BYG
Src/Lck	3LCK	2SRC

*here the A indicates that we used the first chain from the PDB file for our analysis.

A. Self-Organizing Maps

The first question: *Can we structurally distinguish between active and inactive kinases?* In order to answer this question we constructed two maps according to our methodology. As the functional center (*F*) for the proteins we picked the highly conserved Aspartic acid residue in the catalytic loops and we chose an analysis horizon (*r*) of 8Å. The training set for each map contained the kinases 3LCK and 2SRC as "prototypes" of active and inactive kinases, respectively. Then we added the active kinase 1K9A_A to one training set and the inactive kinase 1BYG to the other. Fig. 7 shows the resulting self-organizing maps. Map a) shows the active kinases clustered at the bottom right corner of the map. This means that the self-organizing map recognized that the 1K9A_A kinase was structurally more similar to the active kinase 3LCK than to the inactive kinase 2SRC. The converse is true for map b). Here we see the inactive kinase 1BYG appears in the same area of low quantization error (light grey area) as the inactive kinase 2SRC. This means that in this case the self-organizing map recognized that 1BYG and 2SRC structurally more similar than 1BYG and 3LCK. It is intriguing that the active kinases form a "tighter" cluster in the sense that both active kinases are mapped to the same reference model whereas the inactive kinases merely appear in the same low quantization error region of the map. Perhaps regulated structures of active kinases are structurally much more uniform than the structures of inactive kinases, consequently the active kinases cluster much more tightly than the inactive kinases. These results demonstrate that out SOM-based methodology can group protein tyrosine kinases according to their activation state.

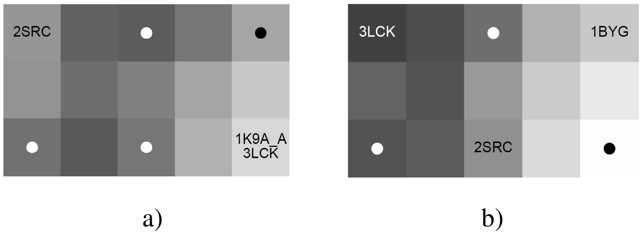

Fig. 7: Self-organizing maps of active and inactive kinases: a) the kinases 1K9A_A and 3LCK are recognized as active kinases and clustered together; b) the kinases 2SRC and 1BYG are recognized as inactive kinases and clustered together.

We investigated next: *Can we detect the regulated sub-structures?* The balls in Fig. 8 identify the normalized α-carbon atoms which represent the predictive features of the active kinases 3LCK and 1K9A_A as computed by the self-organizing map. The predictive features in this case are the structural features that uniquely identify active kinases. There exists no analogous structure to the identified structure in Fig. 8 within our analysis horizon for the 2SRC protein. Consequently we have to assume that this represents the regulated substructure of the kinases. The fact that the self-organizing map identified the regulated structures as predictive features for active kinases underscores the power of this approach. This kind of automatic feature extraction is not available in other structural analysis tools.

From the above we can conclude that the shape of the regulated sub-structures is not depended on the precise activation mechanism.

Fig. 9 shows the structures of the two inactive kinases 1BYG and 2SRC. The balls indicate the predictive features of 1BYG. Two observations are interesting: 1) The predictive features of inactive kinases are different and distinct from the predictive features of the active kinases; there does not exist a structure in the inactive kinases which is analogous to the regulated structure in the active kinases. 2) The predictive features of inactive kinases are not as uniform as the predictive features of active kinases corroborating our findings with the self-organizing maps in Fig. 7.

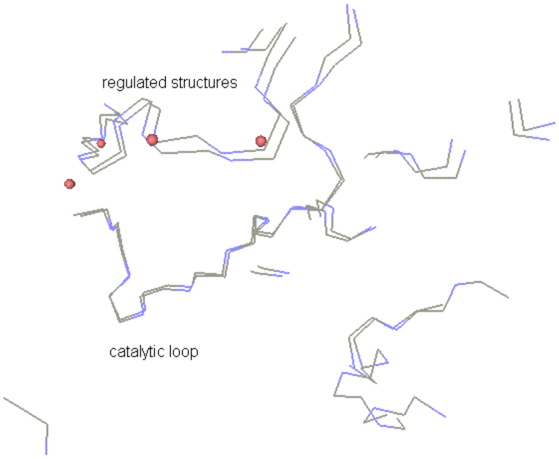

Fig. 8: The predictive features of active kinases. The balls indicate the α-carbon atoms computed by the self-organizing map as predictive features. Note that all the features lie on the regulated structures of the active kinases.

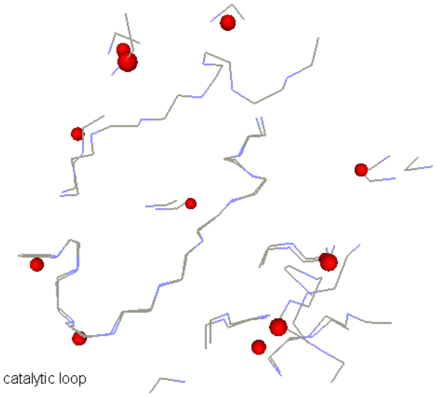

Fig. 9: Comparing the structures of the inactive kinases 1BYG and 2SRC with the predictive features of 1BYG shown.

B. *Residue Sequences*

It is instructive to look at residue sequence based tools and show that an analysis of regulated structures in proteins is not possible with these tools. Fig. 10 shows the phylogenetic tree based on the residue sequences of the active and inactive proteins from Table 1 using ClustalW [20]. There is virtually no structure in this tree; all proteins appear to be identical. The structure that does appear can be considered noise.

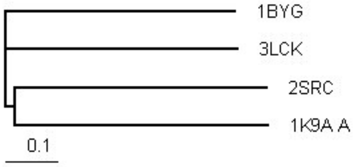

Fig. 10: Phylogenetic tree based on the protein residue sequences.

Examining the multiple residue alignments of our proteins confirms the findings of the phylogenetic tree: the residue sequences line up almost perfectly with many substantially conserved sections. Fig. 11 shows the residue sequence alignments for all four proteins.

```
             10        20        30        40        50        60
     ....*....|....*....|....*....|....*....|....*....|....*....|
3LCK KPWWEDEWEVPRETLKLVERLGAgqFGEVWMGYYNGHTKVAVKSLKQGSMSPDAFLAEAN
2SRC qglakDAWEIPRESLRLEVKLGQgcfGEVWMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQ
1K9A DEFYRSGWALNMKELKLLQTIGKgefgDVMLGDY-RGNKVAVKCik-nDATAQAFLAEAS
1BYG defyrsGWALNMKELKLLQTIGKgeFGDVMLGDYR-GNKVAVKCIKnd-atAQAFLaeas
             70        80        90       100       110       120
     ....*....|....*....|....*....|....*....|....*....|....*....|
3LCK LMKQLQHQRLVRLYAVVTQ--E--PIYIITEYMENGSLVDFLKTPSGIKLTINKLLDMAA
2SRC VMKKLRHEKLVQLYAVVSE--E--PIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAA
1K9A VMTQLRHSNLVQLLGVIVEekG--GLYIVTEYMAKGSLVDYLRSrgrsVLGGDCLLKFSL
1BYG vMTQLRHSNLVQLLGVIVE--EkgGLYIVTEYMAKGSLVDYLRSRgrsVLGGDCLLKFSL
            130       140       150       160       170       180
     ....*....|....*....|....*....|....*....|....*....|....*....|
3LCK QIAEGMAFIEERNYIHRDLRAANILVSDTLSCKIADFGLARLIednextaregaKFPIKW
2SRC QIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARliedneytarqgakFPIKW
1K9A DVCEAMEYLEGNNFVHRDLAARNVLVSEDNVAKVSDFGLTKEAsstq----dtgKLPVKW
1BYG DVCEAMEYLEGNNFVHRDLAARNVLVSEDNVAKVSDFGltkeass----tqdtgkLPVKW
            190       200       210       220       230       240
     ....*....|....*....|....*....|....*....|....*....|....*....|
3LCK TAPEAINYGTFTIKSDVWSFGILLTEIVTHGRIPYPGMTNPEVIQNLERGYRMVRPDNCP
2SRC TAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVERGYRMPCPPECP
1K9A TAPEALREKKFSTKSDVWSFGILLWEIYSFGRVPYPRIPLKDVVPRVEKGYKMDAPDGCP
1BYG TAPEALREKKFSTKSDVWSFGILLWEIYSFGRVPYPRIPLKDVVPRVEKGYKMDAPDGCP
            250       260       270
     ....*....|....*....|....*....|....*
3LCK EELYQLMRLCWKERPEDRPTFDYLRSVLEDFFTAT
2SRC ESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSt
1K9A PAVYDVMKNCWHLDAATRPTFLQLREQLEHIRTHE
1BYG PAVYEVMKNCWHLDAAMRPSFLQLREQLEHIKTHE
```

Fig. 11: Multiple residue sequence alignment of the proteins under investigation.

As expected, multiple sequence alignment techniques cannot provide any details on the structure of regulated proteins.

C. *Structure Comparison and Alignment Tools*

Structure servers such as K2 [6] and DALI [4, 5] simply return structurally aligned proteins. It is up to the user to interpret the results. No mechanical support is provided for these interpretations. The limitation to pair wise structural alignments makes studies such as the one undertaken here very cumbersome: instead of constructing two maps as we have done above one would have to construct four pair wise alignments, extract the essential structural features from these alignments, and then perform an overall comparison of the essential structures over the four pair wise alignments by hand. Tools such as VAST [2, 3] use protein structure as query parameters in order to find the structural neighbors of the proteins under investigation. That VAST is not sensitive to regulated conformational changes in proteins is witnessed by the fact that a query given an active conformation of a protein returns the inactive conformation of this protein as a structural neighbor. This makes VAST unsuitable for the kind of studies we envision here.

VI. RELATED WORK

It is clear from the above discussion that our approach is closely related to structure comparison and alignment tools such as K2, VAST, *etc* [3, 5, 6, 17, 21]. A relatively new algorithm for 3D protein structure alignment is FAST [22]. It utilizes a directionality-based scoring scheme to compare the intra-molecular residue-residue relationships in two structures. Another approach is to model the folding of proteins given an amino acid sequence [23]. The protein folding simulation program Wurst [24] is a protein threading program with an emphasis on high quality sequence to structure alignments. It takes submitted sequences and aligns them to a large number of templates with a conventional dynamic programming algorithm. The 3-D structures of submitted sequences are deduced from a log-odds probability of sequence to structure fragment compatibility, obtained from a Bayesian classification procedure. In both cases overall statements about structural similarity between proteins can be made but both techniques have limited facilities for providing in-depth analyses on function-specific structures.

We can also consider purely visual interpretations and comparisons of protein structures [18, 25] which allow the user to examine the molecular structures in much more detail. However, our innate ability to see patterns in protein structures is easily overwhelmed by the vast amount of structural information available for typical proteins.

Unsupervised machine learning techniques, particularly the self-organizing map technique, have been widely used in the evaluation of biological data. To evaluate protein secondary structures, Unneberg and co-workers trained a SOM with a set of protein circular dichroism data and the SOM was able to classify secondary structures of a group of proteins [26]. In their attempt to locate HIV protease cleavage sites in proteins, Yang and Chou partitioned a set of protein sequences using SOM and applied conventional homology alignment to each cluster to determine the conserved local motif (biological pattern) for the cluster. These local motifs were then regarded as rules for prediction and classification. They found that the rules derived from this method are much more robust than those derived from the decision tree method [27]. Andrade and co-workers have applied SOM to classify sequences within a protein family into subgroups that generally correspond to biological subcategories. Combining maps generated at different levels of resolution, they captured the structure of relations in protein families that could not otherwise be represented in a single map. The underlying representation of maps enabled them to retrieve characteristic sequence patterns for individual subgroups of sequences. Such patterns tended to correspond to functionally important regions. Their modified SOM algorithm included a convergence test that dynamically controls the learning parameters to adapt them to the learning set instead of being fixed and externally optimized by trial and error [28]. Kohonen has applied SOM's to protein sequence data by considering a sequence distance measure based on phylogenetic distances between the sequences [29, 30]. The advantage of this method is that the sequences under consideration are not constrained by the "equal length" requirement imposed by the SOM algorithm. However, these approaches did not take protein structure into consideration.

VII. Conclusions And Further Research

Here we have introduced a novel technique for the analysis of protein structure based on self-organizing maps. Our technique goes beyond the capabilities of similar existing structure comparison tools. Preliminary results demonstrate that our technique can be successfully applied to protein families in order to classify the proteins within these families by their local structure around a functional center. We have also demonstrated that we can analyze conformational changes due to different functional states of the same proteins. The information captured by the self-organizing map and stored in its reference models highlights the essential or predictive structures of the proteins under consideration and can be effectively used to study detailed structural differences between the proteins further, promising answers to interesting and difficult biochemical and biological questions. The automatic nature of our approach due to the machine learning techniques will enable us to scale this to a high-throughput structure analysis in the future allowing us to study large subsets, if not whole families of proteins. It is hoped that studying larger number of proteins will solidify the initial findings reported here. In order to accomplish this effectively we are currently investigating a more compact feature vector representation and automatic local protein alignment techniques. Our feature vector computation will need to be more sophisticated. In our current version small conformational deviations between proteins can lead to large differences in the associated normalized models inducing artificially inflated dissimilarities between the molecules. Algorithms and techniques from computer vision and computational geometry seem promising for improving our feature vector computation.

We have also conducted some initial experiments that go beyond the α-carbon representation and integrate amino acid side chain structures into our normalized model representation of proteins. The results of these experiments look promising but the visual presentation of these normalized and reference models represents a challenge and will be addressed as part of our research program. Another research aspect is the extension of the analysis horizon from a local neighborhood around the functional center to the inclusion of whole protein domain structures and perhaps complete proteins.

Acknowledgements

This research was supported in part by NIH Grant Number P20 RR016457 from the BRIN/INBRE Program of the National Center for Research Resources.

References

[1] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The Protein Kinase Complement of the Human Genome," *Science*, vol. 298, pp. 1912-1934, 2002.

[2] T. Madej, J.-F. Gibrat, and S. H. Bryant, "Threading a database of protein cores," *Protein Struct. Funct. Genet.*, vol. 23, pp. 356-369, 1995.

[3] NCBI, "VAST - Vector Alignment Search Tool," 2002.

[4] L. Holm and C. Sander, "Touring protein fold space with Dali/FSSP," *Nucl. Acids Res.*, vol. 26, pp. 316-319, 1998.

[5] EMBL-EBI, "Dali Structure Server," 2000.

[6] J. Szustakowski and Z. Weng, "Protein structure alignment using evolutionary computing," in *Evolutionary Computation in Bioinformatics*, G. Fogel and D. Corne, Eds.: Morgan Kaufman, 2002.

[7] S. R. Hubbard and J. H. Till, "Protein tyrosine kinase structure and function," *Annu Rev Biochem*, vol. 69, pp. 373-398, 2000.

[8] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. Cambridge, Mass.: MIT Press, 2001.

[9] L. Hamel, "A Brief Tutorial on Database Queries, Data Mining and OLAP," in *The Encyclopedia of Data Warehousing and Mining*, J. Wang, Ed.: Idea Group Publishers, in press.

[10] P. Baldi and S. Brunak, *Bioinformatics : the machine learning approach*, 2nd ed. Cambridge, Mass.: MIT Press, 2001.

[11] T. M. Mitchell, *Machine learning*. New York: McGraw-Hill, 1997.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne., "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.

[13] D. D. Leipe, Y. I. Wolf, E. V. Koonin, and L. Aravind, "Classification and evolution of P-loop GTPases and related ATPases," *Journal of Molecular Biology*, vol. 317, pp. 41-72, 2002.

[14] T. Kohonen, *Self-organizing maps*, 3rd ed. Berlin ; New York: Springer, 2001.

[15] L. Hamel, J. Bang, B. Ioerger, and N. Dholakia, "Market Segmentation for CRM using Multiple, Hierarchically Constructed Self-Organizing Maps," *Decision Support Systems*, submitted.

[16] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*: Morgan Kaufman, 1999.

[17] J. Zhu and Z. Weng, "FAST: A Novel Protein Structure Alignment Algorithm," *Proteins: Structure, Function and Bioinformatics*, submitted.

[18] Accelrys, "DS Viewer," 2004.

[19] M. d. Rinaldis, G. Ausiello, G. Cesareni, and M. Helmer-Citterich, "Three-dimensional Profiles: A New Tool to Identify Protein Surface Similarities," *J. Mol. Biol.*, vol. 284, pp. 1211-1221, 1998.

[20] D. Higgins, J. Thompson, and T. Gibson, "CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice.," *Nucleic Acids Res.*, vol. 22, pp. 4673-4680, 1994.

[21] T. J. P. Hubbard, A. G. Murzin, S. E. Brenner, and C. Chothia, "Scop: a structural classification of proteins database," *Nucleic Acids Res.*, vol. 25, pp. 236- 239, 1997.

[22] J. Zhu and Z. Weng, "FAST: A Novel Protein Structure Alignment Algorithm," *Proteins: Structure, Function and Bioinformatics*, vol. 3, pp. 618-627, 2005.

[23] W. F. v. Gunsteren, T. Huber, and A. E. Torda, "Biomolecular Modelling: Overview of Types of Methods to Search and Sample Conformational Space," presented at 1st European Conference on Computational Chemistry, 1995.

[24] A. E. Torda, J. B. Procter, and T. Huber, "Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices," *Nucleic Acids Research*, vol. 32, pp. W532-W535, 2004.

[25] R. Sayle and E. J. Milner-White, "RasMol: Biomolecular graphics for all," *Trends in Biochemical Sciences*, vol. 20, pp. 374, 1995.

[26] P. Unneberg, J. J. Merelo, P. Chacon, and F. Moran, "SOMCD: Method for evaluating protein secondary structure from UV circular dichroism spectra," *Proteins-Structure Function and Genetics*, vol. 4, pp. 460-470, 2001.

[27] Z. R. Yang and K. C. Chou, "Mining biological data using self-organizing map," *Journal of Chemical Information and Computer Sciences*, vol. 6, pp. 1748-1753, 2003.

[28] M. A. Andrade, G. Casari, C. Sander, and A. Valencia, "Classification of protein families and detection of the determinant residues with an improved self-organizing map," *Biological Cybernetics*, vol. 6, pp. 441-450, 1997.

[29] P. Somervuo and T. Kohonen, "Clustering and Visualization of Large Protein Sequence Databases by Means of an Extension on the Self-Organizing Map " in *Proceedings of the Third International Conference on Discovery Science* Springer-Verlag, 2000 pp. 76-85

[30] T. Kohonen and P. Somervuo, "How to make large self-organizing maps for nonvectorial data " *Neural Netw.* , vol. 15 pp. 945-952 2002