ABSTRACT

We have developed a Graphical User Interface (GUI) implemented by homology modeling software to enhance the performance of protein modeling as well as to increase user-friendliness and reduce the requirement of prior user knowledge of computers. In addition, by making the process easier we hope to promote the use of homology modeling as an effective technique. The interface was created entirely using PERL and CGI computer languages and comprises approximately 600 lines of code with future improvements expected. Homology modeling entails the determination of a protein's three-dimensional structure from that of its homolog. Homology constraints are used to conform highly homologous regions while energy functions delimit those regions that are not as homologous (Li et. al, 1997). The GUI primarily utilizes DYANA to calculate three-dimensional protein structures from homology constraints created using PDBSTAT.

INTRODUCTION

The advent of genome-wide sequencing projects has triggered the realization of the immediate need for methods to decipher the enormous amounts of genetic information that result from these projects. Complete sequencing of entire genomes has been performed for many organisms; most notably the Human Genome Project is nearing completion. Entire genomes have been completed for 27 organisms including such model organisms as *Caenorhabditis elegans* and *Drosophilia melanogaster* leading to the availability of much needed genetic information. The data that result from these projects have the potential to help us discover and understand almost every gene in the organism of study. Unfortunately, just knowing the sequence will not provide us with these answers. Instead, we may increase the value of this information by devising methods and strategies to delineate and decode this information. In doing so, we may discover gene functionality and hence new pharmaceutical targets. The products of these genes are widely recognized as the next generation of therapeutics and targets for the development of pharmaceuticals (Montelione and Anderson, 1998). The future of these projects is therefore seen in extending the analysis of the genome to probe into the meaning of the organism's sequence and how we may use this information to further benefit society. Due to the length of such immense projects and the rapid growth of technology, sequencing tools and techniques are improving alongside the project's progression. For this reason, sequencing projects are completed earlier than expected. They present a formidable challenge in their analysis. For example, the Human Genome Project began in October 1990 and was expected to last fifteen years. Instead, because of vast technological success in the field, the project is now foreseen to be complete by 2003; two years in advance (http://www.ornl.gov/hgmis/faq/faqs1.html).

In the Protein NMR Lab we find simple and effective solutions for these challenges by concentrating our work in structure-based functional genomics. This field involves the application of protein structure determination and/or modeling to identify potential biochemical functions of novel genes. In other words, we establish a functional

relationship through the analysis and grouping of structurally similar proteins. In addition, we attempt to effectively speed up the modeling of a protein in order to increase the capability of our software to handle large amounts of data in shorter periods of time. The value of this advantage is realized when one considers the information base forming from the projects mentioned above. In essence, we hope to develop a method for high-throughput structural analysis on a genomic scale.

The field of bioinformatics has equipped us with methods to perform these tasks, hence becoming a major focus and player/component of our lab. Bioinformatics involves the computational analysis of gene sequences and protein structures on a large scale. It is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research (http://www.whatis.com). Major sub-fields of bioinformatics that interest our group include structural and functional genomics. Structural genomics is the systematic analysis of the three-dimensional structures of all the proteins within a particular genome. Functional genomics may be described as the analysis of all functions of all proteins within a genome.

The advantages of this field have already been applied in drug research as the following article comments about a drug company Vertex: "Its researchers use 3-D protein structure to help accelerate the developmental process…" (http://www.doubletwist.com/news/features.jsp). More structurally oriented companies such as Structural GenomiX, Inc. (SGX) are making similar efforts. In reference to the company's CEO, an article states the following:

> He describes the mission of SGX as high-throughput structure determination —essentially taking an industrial approach to the laborious process of structure elucidation, building a database of high-resolution structure, and selling access to pharmaceutical and biotech companies (http://www.doubletwist.com/news/features.jsp).

The market of high-throughput structural determination has already been tapped and competition is growing. Our lab comes in with similar goals of determining large amounts of accurate structures while not concerning ourselves with the commercial value of our results. Our efforts are thus seen as a "public protein initiative" rather than attempting to commercially exploit the process.

In order for researchers to increase the "value" of genomic data there must be some way to infer function from sequence. Modern day technology has not provided a workable way to reliably predict function from sequences. In the future, this may no longer pose a problem as research is currently being performed in the hopes of developing such techniques. Recognizing these facts and shortcomings of modern technology, a method must be devised to get from knowing an organism's sequence to predicting its function. The relationship between a protein's sequence, structure, and function has been the subject of much study, as well as much argument for quite some time now. It is known that a protein's sequence determines its' structure which in turn has a profound influence on the function of that protein, as shown in figure 1.

Figure 1: Flow of genetic information

From the figure, it can be seen that protein structure is somewhat of an intermediate between a gene's sequence and its function. Through research, it has been found that if two or more proteins have similarities in sequence and a common ancestor they will most likely have similar structures. Therefore, in attempting to establish functional homologies, employing sequential similarities as in indicator would not be the best approach. Rather, to detect functional relationships between genes, structural similarities prove to be more of an accurate indication. One advantage of this technique is that structural similarities may provide functional homologies not readily seen in analyzing sequence homology. The following group provides support for this fact by stating, "because evolution tends to conserve function and function depends more directly on structure than on sequence, structure is more conserved in evolution than sequence" (Sanchez and Sali, 1998). Although this is true, we would be mistaken in stating that similarities in structure *always* lead to similarities in function. Rather, structural homology *often* implies functional homology.

As an example of the disadvantages of sequential genomics in establishing functional homologies, consider the three genes below in Figure 2. The portions of the

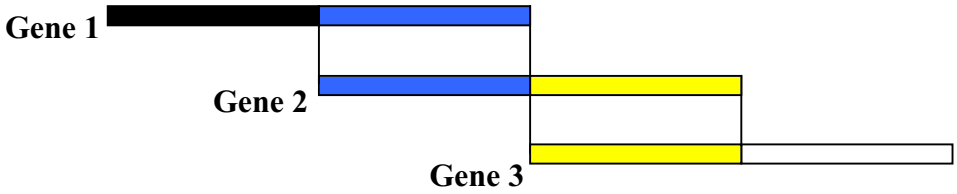

Figure 2: Cartoon representation of hypothetical 3-gene family

gene that are colored similarly are homologous to each other. The region in black seen only on gene 1 represents the location of residues coding for a hypothetical function X. Furthermore, it may be inferred from the figure that gene 1 is 50% homologous to gene 2 and gene 2 is 50% homologous to gene 3. In one situation, we may want to establish homology between the 3 genes. Since we know that gene 1 is homologous to gene 2 which is in turn homologous to gene 3, we may place all three genes in the same homologous family when in fact gene 1 and gene 3 do not have homologous regions. We would therefore be mistaken in attributing any homology between genes 1 and 3. In another situation with the genes 1 and 2, assume we wish to determine the function of gene 2 and realize that the 50% sequence homology between genes 1 and 2 is significant. We may be tempted to attribute function X to gene 2 since gene 1 has this function and

they are significantly homologous. However it can be seen that gene 2 may not, in fact, have this function since it is not homologous to the functional domain of gene 1 where X is coded for. These two situations provide examples of where sequential homology may not provide the best path to establishing functional similarities. In addition, "[another] advantage of 3D modeling over sequence matching is that some binding and active sites cannot possibly be found by searching for local sequence patterns" (Sanchez and Sali, 1998). Certain structural characteristics that would implicate a potential biochemical function may include such factors as protein fold class, locations and clustering of conserved residues not adjacent in sequence, and surface electrostatic field distributions (Montelione and Anderson, 1998). It can then be hypothesized that an alternative approach to functional genomics may be through structural genomics.

Within a genome, a certain minimum number of protein structures must be experimentally determined in order to be able to infer common functions between proteins. These protein structures for which the sequence is known can be referred to as "template structures" since their structure is the basis of conformation. "Query protein" is the term used to describe proteins being modeled. Once these structures are determined, they may then be analyzed for homology and families of structures may be formed. Li et al. provide support by stating the following:

> Because the relative positions of certain structurally and/or functionally crucial atoms should be similar among a family of homologous proteins, the three-dimensional structure of a protein can often be modeled reliably based on the known structures of homologous proteins (Li et. al, 1997).

To be members of the same family would indicate significant homology between two or more proteins. Then for each family, if a member already has a known function we may be able to attribute that function to the rest of the group after further structural analysis. As stated above, structural homology does not always lead to functional homology. Yet, when modeling we realize that we are calculating *approximate* three-dimensional structures. From these minimally accurate structures we are still able to determine such functionally relevant factors as surface electrostatics. If we know the function of the protein we are utilizing as a template, we can compare these factors and consider whether they compare favorably. For example, when considering electrostatic field potentials we may discover that a protein is mostly positive on one side while it is mostly negative on the other. In addition, we may know that the template is a surface protein sharing that characteristic. Then it would be reasonable to attribute similar functions to the two proteins. So, from the precise, experimental determination of a number of proteins, we may model many times that number of proteins assuming they belong to the same family. As a hypothetical situation consider the following. We have 10 proteins each of which has had their structure experimentally determined. In addition, we know that each of the 10 proteins has significant homology to a family of 10 other proteins. Using homology modeling, we may then theoretically predict the three-dimensional structure of 100 proteins (10 proteins with known structure multiplied by 10 proteins with significant homology to those with known structure). This is significant considering the fact that only 10 experimental structures were required. The advantages of this method are

therefore realized once a small percentage of protein structures are determined. Therefore, in establishing homology between genes, we institute a path for modeling protein structures assuming that one gene has a previously determined structure. This point is further sustained by Gerstein: "…if sequences can be related to specific functions and pathways, one can see whether homologous sequences in two organisms truly have the same role (orthlog vs. paralog) and whether particular pathways are present or absent in different organisms" (Gerstein, 1998).

In modeling a protein, regions of the query protein can be subdivided into two categories: those regions, which are highly homologous to the template structure and those that are not. These regions can be differentiated from the alignment of the template and query proteins. This alignment will ultimately determine the viability of the predicted models. Based on the chosen protocol, a certain amount of gaps and insertions will be present in the final alignment. These gaps and insertions should not interrupt non-flexible regions such as secondary structure elements and core atoms. The researcher may modify poor alignments by adjusting the placement of these interruptions; "…for sequences with sequence identity <40%, large errors in the alignment can sometimes be prevented by examining and editing the alignment manually" (Sali, 1995). The characterization of sequence domains from the alignment is important as the two regions are modeled differently. Li et al. explains the point further:

> Regions of the unknown protein structure that are highly homologous to the known template structure are constrained by 'homology distance constraints,' whereas the conformations of nonhomologous regions of the unknown protein are defined only by the potential energy function (Li et. al, 1997).

In other words, regions that are similar in both proteins can be conformed solely from constraints while dissimilar regions must make use of an atom's natural tendencies to form the correct structure. Due to the lack of atomic data we have for the dissimilar regions from the template structure, we must allow it to conform as it would in nature.

First, one must realize that highly homologous regions will most likely share certain characteristics such as protein folds. Sahasrabudhe and colleagues provide support for this point by stating that:

> For structural and other physiochemical reasons, the relative positions of key atoms or groups of atoms in a family of homologous proteins with similar biochemical functions must remain similar in order to maintain proper functioning of these proteins" (Sahasrabudhe et. al, 1998).

For this reason, we may model these regions of the query protein by constraining residues in the query protein to positions of the corresponding residues in the template structure. Only heavy atoms are considered when defining homologous atoms as can be seen from Figure 3 (Li et. al, 1997).

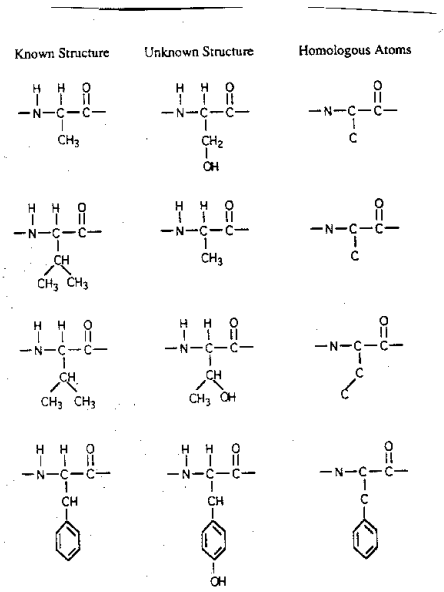

Fig. 8. Representative examples of approach used to define homologous atoms between aligned residues.

Figure 3: Visual explanation of homologous atoms

Measuring and applying constraints is possible because we know the three-dimensional coordinates of the template protein. From these spatial coordinates we can survey atomic distances for each residue. By applying a small percentage of these atomic distances, referred to as "homology constraints" (Li et. al, 1997), to the homologous regions of the query protein we have defined a range of space for those residues. Also, it should be noted that the constraints applied are not strict distances to which residues must adhere. Instead a range of space is created thereby introducing a small amount of freedom so that structures are not forced into unfavorable conformations. A "forced" structure can be described as a conformation that is not energetically favorable and would most likely not arise in nature. This freedom is accounted for by adding and subtracting 10% from the measured constraint forming upper and lower constraints, respectively. Therefore, based on the fact that highly homologous primary structures will most likely have similar folds, we can define "valid space" for residues in the query protein from the three-dimensional coordinates of the template structure.

The regions of the query protein we are left with are those that are not highly homologous. Because of this lack of homology, we may not assume that these query regions will have folds similar to corresponding regions in the template structure. Therefore, we must devise a method to accurately conform these regions while being impartial in the process so as not to force structures. This is possible by recognizing that in nature, molecules tend to exist in their most energetically favorable conformations. Having minimal energy thereby stabilizes the molecule that attains these conformations. Realizing this, we have developed a method for allowing proteins to explore all possible conformations thereby giving the molecule a chance to find its most energetically favorable structure. This process is described as "simulated annealing" (Li et. al, 1997).

Each protein molecule can be represented by a potential function, which represents its energy at differing conformations. A simple example of a potential function would be that of a Lennard-Jones Potential (Figure 4).

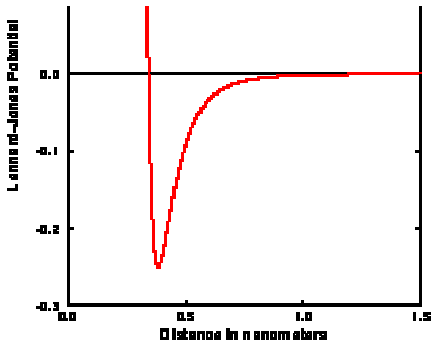

Figure 4: Lennard-Jones Potential Function for two atoms

This figure (http://polymer.bu.edu/Wasser/robert/work/node8.html) represents the potential used to describe the interaction of two non-bonded atoms. From the graph it can be seen that the atoms start to mildly attract each other at a certain distance due to induced dipole-dipole interactions. The atoms then continue to move closer and eventually reach an ideal distance from each other and hence form a bond at a minimum energy level in the graph. Further proximity causes the atomic distance to exceed an understood threshold and the repulsion of the atoms' electrons overcomes the bond's attractive strength. The energy of the bond then exponentially increases breaking the bond due to instability.

The vertical axis of a graph in a potential or energy function is always energy while the horizontal axis can be most anything. The energy function takes into account a number of atomic and structural properties in order to find the most energetically favorable conformation and to avoid "forced" structures. In the technique of simulated annealing, a natural system is emulated and through the use and manipulation of the specific protein's energy functions, the protein is allowed to conform to its most energetically favorable form.

MATERIALS AND METHODS

Figure 5 (http://biotech.icmb.utexas.edu/pages/bioinfo.html) presents an overview of the process of homology modeling indicating its attempt to mimic the process protein's naturally go through.

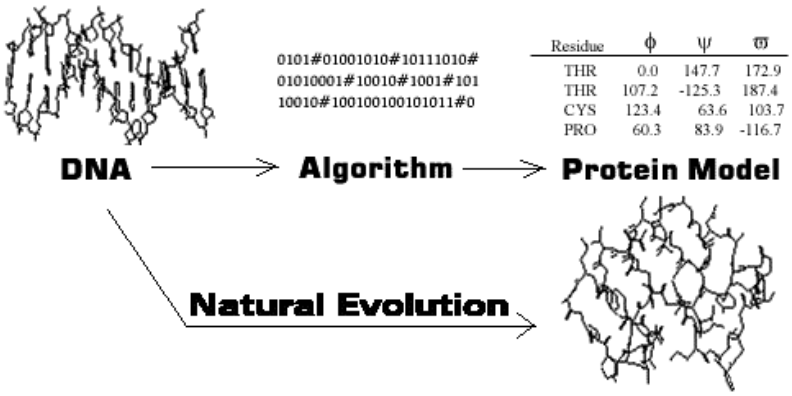

Figure 5: Overview of Homology Modeling

The process of homology modeling is accomplished in the lab through the effective use and manipulation of various network configurations. All the steps are computer-based and run on a local network shared between three labs. Specifically the software used for homology modeling includes PDBSTAT, CONGEN, DYANA, and some 3-D viewing software. To run these programs remotely, a Graphical User Interface (GUI) was created in order to allow access to the software from the World Wide Web as well as to increase the efficiency and productivity of the software's use. The GUI provides a simple and efficient way to practice the method of homology modeling described in this paper. One of the main goals in creating the GUI was to conceal the details of the software in order to allow lab members less skilled with computers to be able to perform the process. In addition, the GUI amplifies the effectiveness of the process by combining many underlying steps into a simple click of the mouse. For example, the first submission of data through the interface runs approximately 25 commands as shown below.

pdb.step1
read coor pdb /nmrusr/people/reza/public_html/cgi-bin/1COF.pdb
get
drosocof.dst

10000

change
drosocof.dst
/nmrusr/people/reza/public_html/cgi-bin/alinea.new.for2
drosocof.cng
exit

pdb.step2
read seq /nmrusr/people/reza/public_html/cgi-bin/DROSOCOF/drosocof.seq

rea cons congen drosocof.cng
wri
drosocof.dya
cons
diana
exit

Figure 5: PDBSTAT commands

Each line shown above represents a different command excluding those reading pdb.step1 and pdb.step2. These commands are run through PDBSTAT to create homology constraints from three-dimensional coordinates. The files containing these commands are also copied to the user's directory allowing those who are familiar with PDBSTAT to view the steps taken. Particular commands will automatically be changed according to the protein being modeled. For example, longer residue chains need more constraints so the command line that reads 10000 would be increased. Advanced users will also be able to customize these commands to their own liking.

The GUI was created on the Internet using Perl (Practical Extraction and Report Language) and CGI (Common Gateway Interface) scripts. A CGI is a way for a web server to communicate user-entered information to local software. For example, when a user purchases an item over the Internet they must fill out a form requiring them to enter such fields as their name, address, and credit card authorization number. This information is then passed through a CGI to software at the site where the website is maintained. The software then processes the data and sends results back to the user, such as a confirmation in this case. Perl is a programming language that has many advantages in performing text manipulation. This is important when considering that the information these programs will be receiving, such as three-dimensional coordinates and alignments, will be variant. Therefore, being able to effectively process this data is key to the GUI's performance. Perl provides quick and easy implementation of such tasks and has thus become the primary language used in establishing the GUI.

To begin the process, the researcher must provide two things: 1) the PDB coordinates of the template structure and 2) an alignment of the two proteins. As can be seen in Figure A.1, there are entry fields for each parameter. The alignment of the two genes may be entered manually in the text box or as a file from the user's home directory. Also, the pdb coordinates of the template structure may be entered in more than one way. The user may provide a file containing the pertinent information from their home directory. Alternatively, the user may simply enter a PDB identification code and our software will retrieve the coordinates from the PDB website (http://www.rcsb.org/pdb).

In dealing with user files, one must realize that our web server does not possess these files. This ownership is obviously necessary as all software is run off our server. Therefore, it was necessary to implement an uploading scheme to transfer files from the user to our server. Uploading is needed for the alignment file and coordinate file only when the user chooses to provide us with files rather than manual entering the alignment or pdb identification code. To perform this task, we utilized an existing uploading CGI script created by Todd Sambar (http://www.sambar.com) modified by Dr. Daniel Monleon to be compatible with our network configuration.

Once these files are attained by either method, PDBSTAT (developed by Dr. Roberto Tejero) evaluates various macromolecular statistics from the given pdb coordinates. For homology modeling, PDBSTAT takes the user-entered 3-D coordinates of the template structure and generates homology constraints. The amount of constraints generated is user-dependent and ultimately determines how many will be applied to the

query protein. In addition, the user has the choice to enter the amount of constraints to be measured. If the user decides not to do so, the amount will default to 10,000. These constraints are then formatted for the next step in the process and split into upper and lower constraints. These boundaries are calculated by respectively adding and subtracting 10 percent of the original homology constraint. In general, it has been found that approximately 16 constraints per atom is sufficient (Li et. al, 1997), though users may choose to utilize a higher constraint density.

Files created by PDBSTAT require inspection for errors and a method of division for certain procedures. These tasks were performed using Perl. A filtering procedure was embedded into the program as well as a procedure to divide constraint files. These methods involved advanced manipulation of text strings. Unnecessary lines of the file are recognized and discarded while protecting wanted information. The challenge in dividing and filtering procedures came in the fact that unnecessary and necessary data lines were extremely similar. Text patterns were established so that unessential data could be distinguished and discarded. All software used in our method of homology modeling is extremely sensitive to format. For this reason, the filtering and dividing of data was vital to be able to create input for the following steps.

Once PDBSTAT has completed these calculations, target functions are generated and energy minimization techniques are applied through the use of DYANA and CONGEN. Both DYANA and CONGEN utilize molecular dynamics to create target functions for the query protein while focusing on separate aspects. Specifically, CONGEN concentrates on energy functions while DYANA concentrates more on distance geometry. DYANA has fewer degrees of freedom and as such has a simpler and less accurate energy function calculation. Dihedral angles provide the sole degree of freedom in DYANA. The energy function primarily includes terms regarding dihedral angles, hydrogen bonding, and van der Waal forces. In addition, the starting point from which the query protein is modeled is a random structure. This is advantageous because for each conformation the query protein is able to explore all conformations and "choose" the most energetically favorable one. When using DYANA, much time is saved and because of this, DYANA is seen as an optimization of the process. Guntert et al. present some advantages of DYANA:

> Torsion angle dynamics can be more efficient than molecular dynamics in Cartesian coordinate space because of the reduced number of degrees of freedom and the concomitant absence of high frequency bond and angle vibrations, which allows for the use of longer time-steps and/or higher temperatures in the structure's calculation (Guntert et al., 1997).

Nevertheless, there are drawbacks such as a pre-requisite of increased user knowledge of the protein and increased refinement before achieving a final structure. These disadvantages are mostly due to the fact that terms are excluded from the function. Although DYANA calculations take approximately one minute as opposed to two hours for CONGEN, the tradeoff comes when one realizes that energetically unfavorable results may occur. CONGEN, on the other hand, introduces many more degrees of freedom

requiring its energy function to contain all terms from DYANA in addition to many other terms. The starting point for CONGEN is a fully extended conformation. Starting from a fully extended conformation allows the protein the same advantages of exploration as in DYANA. The loosening of constraints with CONGEN calculations has led to the observation of "improved convergence" and "lower energy structures" (Tejero et al., 1996). When using the interface, the decision to utilize DYANA or CONGEN is made by the user. CONGEN is not available through the GUI as of yet leaving users the only option of DYANA. Although we plan to offer CONGEN in the future, DYANA is recommended. This is due to the fact that when a user is processing data, all information is susceptible to Internet traffic. For this reason, we plan to offer a procedure of emailing results for those users who choose CONGEN. CONGEN calculations take approximately two hours as opposed to one-minute calculations by DYANA. Waiting for results from CONGEN through the Internet is not a viable option.

The next step in the process is the actual calculation of the models. Multiple models are created to insure that the protein has examined every conformational setting. The starting point for each model is different thereby releasing any bias in the conforming methods of the software. CreateProc is a standard control file for homology modeling calculations and is utilized next. The tasks of this program are split between multiple processors thereby expediting the calculation. The results of this program are three-dimensional structure files, which are copied to a file in the user's directory. These structures may then be viewed either one of two ways. First, the user may opt to utilize a graphics program on their local system. Secondly, we plan to offer a plug-in through the Internet that will allow users to view their protein directly through the interface. This will be slightly delayed, as all information must pass through the Internet to the user. The user is recommended to utilize local graphics tools to avoid these delays. Specifically, the plug-in present on our server will be RasMol software. RasMol provides many advantages in 3-D viewing such as quick and smooth rotational viewing advantages as well as detailed viewing of important secondary structure motifs. Once the structures are calculated, the results are than manually checked for user-entered errors and modified accordingly. It has been found in practice that a number of attempts are required before attaining the correct structure. Usually each attempt increases in its precision due to the manual refinement of user-entered data.

The success of the process relies on well-configured networks, which connect all users within the lab. The lab is very much dependent on these networks simplifying many tasks in the lab. There are two network configurations in the lab performing separate tasks. The main setup we have may be represented as a client – server model (Figure 6). The client server model represents a number of computers, or clients, running applications off a central location, or server. In implementing such a configuration, installation of software is simplified as this task is only performed on the server. All clients may run all programs in the lab off the server without having these programs installed on their hard disk. In addition, the server provides more memory, faster processors, and a multitasking operating system (OS). The multitasking OS provides advantages such as allowing many users to run the same application at the same time. In

addition, users may run processes overnight without actually being logged into the system.

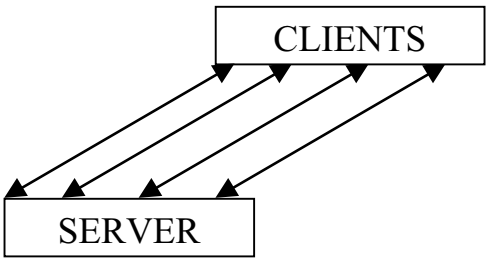

Figure 6: Client-Server Model

The second network setup we have in the lab runs on a master – slave model (Figure 7).

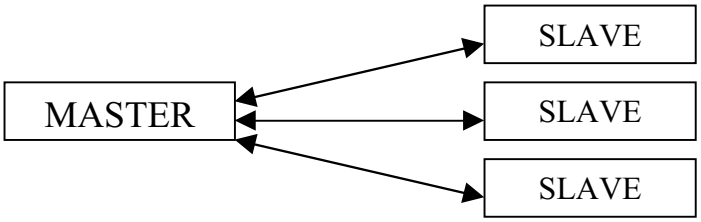

Figure 7: Master – Slave Model

The computers are setup as a queuing system where one computer acts as a master of the rest of the computers on the network, which are referred to as slaves. Specifically, the network consists of 15 pairs of Pentium III processors each running on LINUX operating systems. Installed on this network is a web server from which the GUI for homology modeling is run. Jobs that are sent to the master are separated into smaller tasks and individually assigned to the slaves. The slaves then perform their assigned work and send the results back to the master who compiles all the results and presents them to the user. In essence, we are using multiple processors as opposed to one to perform large tasks. These network setups greatly assist lab members in performing their work quickly and effectively by providing much needed acceleration to large-scale calculations.

RESULTS

The creation of the CABM Homology Modeling Interface has increased the effectiveness and productivity of protein modeling in the lab. Overall, approximately six hundred lines of code were written in PERL and CGI scripts to accomplish the GUI's creation. With these scripts, prior computer knowledge is no longer a requirement for running the software. In addition, multiple steps have been combined to perform the same tasks while greatly decreasing the amount of work required by the user. We estimate that time spent by the user has decreased approximately 40%. Overall time saving was not an issue as total structure calculation times are protein-specific. The utilization and effectiveness of our homology modeling software are also estimated to significantly increase. By implementing these timesaving techniques we have increased user-friendliness and decreased the amount of unnecessary work performed by the user.

Without the interface, all commands including those of PDBSTAT, DYANA, CONGEN, and CreateProc would each have to be typed in manually. We have thus replaced manual typing labor with simple clicks of the mouse in the hopes of expediting and easing the modeling of a protein by a researcher. In regards to Internet delays, a possible solution would be to email results to users once structures are calculated. This solution is more relevant to calculations done using CONGEN, as time becomes a major issue in its use as opposed to DYANA. Finally, all files used in the process are copied to the user's directory hopefully providing a fuller view of the process to the user. By providing this information, we present the user with more control over their calculation.

To investigate the usefulness of the interface, a sample calculation was done with a protein that had been previously modeled. The twinstar gene (*tsr*) is found in *Drosophilia melanogaster*. This gene encodes for a 17-kD homologue of a cofilin/ADF protein called twinstar. Cofilin/ADF proteins are a family of small actin severing proteins (Gunsalus et. al, 1995). Members of this family of proteins are also found in *Saccharomyces cerevisiae* and *Caenorbhabditus elegans* (Gunsalus et al., 1995). Mutations in this gene result in defective centrosome migration and cytokinesis (Gunsalus et al., 1995). This suggests a mitotic/meiotic role for the protein encoded by *tsr*. This protein was modeled using DYANA.

The starting point for these calculations was acquiring the pdb coordinates of a template structure and aligning the two sequences. The template structure was an actin-binding cofilin found in yeast (Gunsalus et al., 1995). The alignment between the 147-residue twinstar protein and 143-residue Yeast cofilin protein resulted in exact sequence homology of 51/147 residues or 35%. This percentage represents an approximate lower limit homology to obtain reasonable structures. Modeling may be done with lower sequence homology but the calculated structures may not be reliable. By using the GUI we were able to get from having these files to the point of calculating structures in less than one minute. The user simply enters the names of the alignment and pdb files into labeled text boxes. Submitting the data then processes the data. In effect, the user is writing two filenames and clicking one button to run PDBSTAT. The actual calculation of structures took approximately five minutes. Once the structures were calculated, the ten best were chosen and superposed as can be seen in Figure (8a). The ten best were chosen from a group of forty calculated structures.
The default CreateProc command calculates structures using 10 high-speed processors with each processor calculating 4 structures. "Best structures" are those with minimal target functions and are chosen by the software. These structures were viewed using MolMol. In addition, for better representation of secondary structure elements, RasMol was used to create Figure (8b). As can be seen from Figure (8a), much of the protein was highly similar in most regions. The variable regions are represented by the "frayed" loops. These regions were those that were not highly homologous and were defined solely by potential energy functions. Important secondary structure elements were not affected by this variability and because of this these models may still be deemed "good" models. The consistency of the other regions provides proof for this point showing that from many different, random starting points the homology of the two proteins was significant enough to produce extremely homologous models with minimal energy

conformations. Recall that when using DYANA, the starting point for each calculation is random.

Figure 8a: Superposition of ten best models of *tsr* shown by MolMol

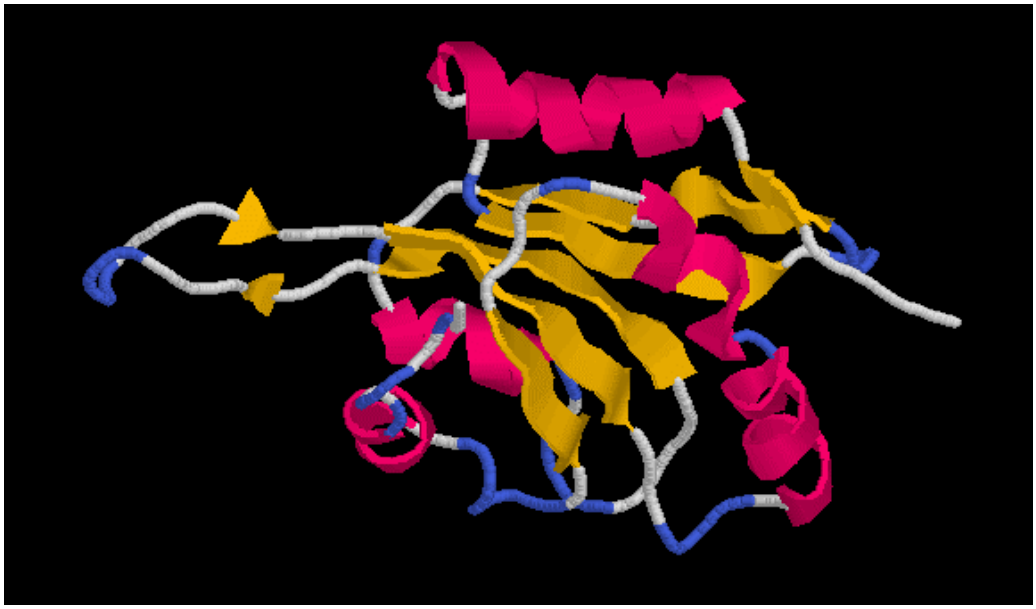

Figure 8b: Cartoon representation of best model of *tsr* shown by RasMol

DISCUSSION

The completion of the GUI will involve the addition of more advanced options for users, allowing use of CONGEN and possible emailing of results, as well as a 3-D viewing application to be run as a plug-in through the Internet. At this time, the interface sets default parameters for most commands during software execution. We hope to expand user capabilities by introducing the advanced user with the ability to change these parameters. Advanced users should have a general knowledge of the underlying software so that their calculations will compute correctly. If a user would only like to change some fields, the unchanged fields would default to previous values. In addition, with some manipulation of software we will be able to allow users to utilize CONGEN for their energy functions. Because of the length of this process, we would email final structures to the user. This would require users to enter their email address as a parameter when beginning the modeling process. Finally, we would like to provide users with the ability to view their structures immediately after they are calculated. The PERL and CGI scripts that run the GUI currently copy all files created during the process to the user's home directory. The user may then follow a hyperlink to a Rasmol viewing of their structure. To be able to view this link, the user must have Rasmol as a local application on their machine in addition to certain pre-set Netscape preferences. Instructions on how to accomplish these procedures will be provided to the user. By providing a plug-in, we will be allowing those who do not possess this software on their local system to view their structures. In addition, for those who do own such software, we will be simplifying the viewing process by providing structure viewing alongside its calculation. The only drawback foreseen in viewing structures through a plug-in would be Internet delays. Because all data traffic will be flowing through multiple servers, time may become an obstacle in receiving data. This point is true for utilizing the GUI as well. However, it should be noted that Internet time transfer rates are not always delayed and when they are, delay times may still be very brief.

Protein modeling provides many advantages for the field of structural genomics. By being able to group genes into homologous families and experimentally determine a minimum amount of structures we are able to infer structural similarities between homologous proteins. This accomplishment then sets the stage for structure-function analysis. Functional discovery becomes more evident due to the fact that structure and function have co-evolved strongly. Recognizing genomic sequencing projects as the first step is important, as they have provided the ability to establish homology between gene sequences. However, it is important to realize that structural similarities do not always lead to identical functions. Rather, being able to form families of homologous proteins will allow us a more in-depth analysis of structural motifs which may or may not implicate a specific function.

The success of the GUI lies in its ability to make easier the process of homology modeling. By concentrating our efforts on this task, we believe more people will access and employ our software for their research thereby promoting the use of homology modeling. In modeling the twinstar protein of *Drosophilia melanogaster* we showed the advantages of using the GUI as opposed to manually running homology modeling

software. The interface we have created eliminates the requirement of prior software knowledge thereby increasing the range of users who wish to model proteins. User-friendliness has thus been increased which will permit faster and more frequent use of homology modeling as a tool for structural genomics. Protein modeling provides many biological advantages as well. In simplifying the structure of a protein we are able to more easily establish distant homologues. Efficiency is also important as the on going sequencing of entire genomes will continue to provide us with large amounts of data that serve as the first step in a long process. The interface has facilitated these advantages by simplifying the process of homology modeling. This has been achieved by concealing the details of the process from the user so as to allow those researchers who are less proficient with computers the ability to perform complex tasks with advanced software.

Acknowledgments

I gratefully acknowledge Dr. Gaetano T. Montelione for allowing me to pursue such work. I would also like to thank Dr. Daniel Monleon for greatly aiding me in my research. In addition, I appreciate help given to me by Dr. Kris Gunsalus, Dr. Hunter Moseley, Dr. Roberto Tejero, Gurmukh Sahota, and Aneerban Bhattacharya.

WORKS CITED

Gerstein M. Dec 1998. Patterns of protein fold usage in 8 microbial genomes: A comprehensive structural census. *PROTEINS: Structure, Function, and Genetics 33(4)*: 518-534.

Gunsalus KC, Bonaccorsi S, Williams E, Verni F, Gatti M, Goldberg ML. Dec 1995. Mutations in twinstar, a Drosophila gene encoding a cofilin/ADF homologue, result in defects in centrosome migration and cytokinesis. *J Cell Biol. 131(5)*: 1243-59.

Guntert P, Mumenthaler M, Wutrich K. 1997. Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. *J. Mol. Biol. 273*:283-298.

Li H, Tejero R, Monleon D, Bassolino-Klimas D, Abate-Shen C, Bruccoleri RE, Montelione GT. 1997. Homology modeling using simulated annealing of restrained molecular dynamics and conformational search calculations with CONGEN: Application in predicting the three-dimensional structure of murine homeodomain Msx-1. *Protein Sci. 6*:956-970.

Montelione GT, Anderson S. Jan 1999. Structural Genomics: keystone for a Human Proteome Project. *Nature Structural Biology 6(1)*:11-12.

Sahasrabudhe PV, Tejero R, Kitao S, Furuichi Y, Montelione GT. 1998. Homology modeling of an RNP Domain From a Human RNA-Binding Protein: Homology-Constrained Energy Optimization Provides a Criterion for Distinguishing Potential Sequence Alignments. *PROTEINS: Structure, Function, and Genetics 33*:558-566.

Sali A. 1995. Modeling mutations and homologous proteins. *Current Opinion in Biotechnology 6*:437-451.

Sanchez R, Sali A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. 95*:13597-13602.

Tejero R, Bassolino-Klimas D, Bruccoleri RE, Montelione GT. 1996. Simulated Annealing with restrained molecular dynamics using CONGEN: Energy refinement of the NMR solution structures of epidermal and type-α transforming growth factors. *Protein Science 5:578*-592.

High Throughput Homology Modeling for Structural Genomics

Reza Y. Akhtar
Principal Investigator: Dr. Gaetano T. Montelione
Protein NMR Lab
Center for Advanced Biotechnology and Medicine