Pacific Symposium on Biocomputing 4:354-367 (1999)

PROTALIGN: A 3-DIMENSIONAL PROTEIN ALIGNMENT ASSESSMENT TOOL

DOANNA MEADS, MARC D. HANSEN, ALEX PANG

Computer Science Department
University of California
Santa Cruz, CA 95064
(doanna,mhansen,pang)@cse.ucsc.edu

Abstract

Protein fold recognition (sometimes called threading) is the prediction of a protein's 3-dimensional shape based on its similarity to a protein of known structure. Fold predictions are *low resolution*; that is, no effort is made to rotate the protein's component amino acid side chains into their correct spatial orientations. The goal is simply to recognize the protein family member that most closely resembles the target sequence of unknown structure and to create a sensible alignment of the target to the known structure (i.e., a structure-sequence alignment). To facilitate this type of structure prediction, we have designed a low resolution molecular graphics tool. **ProtAlign** introduces the ability to interact with and edit alignments directly in the 3-dimensional structure as well as in the usual 2-dimensional layout. It also contains several functions and features to help the user assess areas within the alignment. **ProtAlign** implements an open pipe architecture to allow other programs to access its molecular graphics capabilities. In addition, it is capable of "driving" other programs. Because amino acid side chain orientation is not relevant in fold recognition, we represent amino acid residues as abstract shapes or glyphs much like Lego (tm) blocks and we borrow techniques from comparative flow visualization using streamlines to provide clean depictions of the entire protein model. By creating a low resolution representation of protein structure, we are able to at least double the amount of information on the screen. At the same time, we create a view that is not as busy as the corresponding representations using traditional high resolution visualization methods which show detailed atomic structure. This eliminates distracting and possibly misleading visual clutter resulting from the mapping of protein alignment information onto a high resolution display of the known structure. This molecular graphics program is implemented in OpenGL to facilitate porting to other platforms.

Pacific Symposium on Biocomputing 4:354-367 (1999)

1 Introduction

Proteins are responsible for such diverse tasks as facilitating chemical reactions and transporting molecules. By studying protein structure, we gain insight into how proteins function, and how their properties can be modulated, either in a directed manner as in protein engineering, or in an unwanted way as is the case in genetic disease.

As the genome sequencing projects proceed, scientists have gained access to tremendous amounts of biological information. Due to the difficulties inherent in understanding large quantities of data, information visualization techniques have become an attractive option for the field of bioinformatics [1,2]. Using information visualization, researchers can see experimental results more clearly than by simply viewing raw numbers. For example, a protein sequence alignment may obtain a reasonable numerical score, but visual inspection of the structural model might reveal incongruencies with the physical demands placed on protein structures, such as the need for an intact structural core. In developing and using tools for biological visualization, we have observed that it is difficult to incorporate 3-dimensional data into visual displays for the purpose of analyzing the validity of individual amino acid placements. This problem arises because of the normal visual clutter which ensues when large amounts of atomic data are displayed at high resolution (see Color Plate 1). Another problem is that while there exist several tools for displaying 2-dimensional biosequence alignments (see Figure 1), the tools for viewing the corresponding 3D comparisons either show too much information or not enough [3]. Furthermore, while there are tools that allow one to fine tune and edit an alignment in 2-dimensions, virtually no tools exist to support alignment editing directly in the 3-dimensional structure.

To address these concerns, we describe the **ProtAlign** system. In particular, we describe its:

1. 3-dimensional editing capabilities. While analyzing the structure of a protein, the user now has the ability to directly manipulate and edit the position of the residues. This feature saves the user a context switch in going from the 3D representation to 2D then back to 3D, and allows them to focus more on the problem at hand. Traditional 2-dimensional editing is still supported. Editing in either 2-dimension or 3-dimension will result in the corresponding changes in the 3-dimensional or 2-dimensional displays respectively.
2. Open pipe architecture to facilitate integration with other applications. We demonstrate this ability by integrating **ProtAlign** with the DINAMO[4] alignment editing and scoring program. We see this architecture

as a means for extending the capabilities of **ProtAlign**.

3. Lego-like glyphs used to represent amino acids (see Color Plate 3). The design of these glyphs takes into account the overall size and residue type of the amino acid. Furthermore, the pairing of these glyphs quickly gives the user an impression of goodness of fit. For example, fitting a round peg into a square hole indicates a poor fit.
4. Comparative visualization techniques to highlight the quality of an alignment. In particular, we draw from and adapt techniques used in comparing vector field data from aerodynamics [5] to bear upon the problem of showing how well a structure-sequence fits together.

In order to facilitate the discussion of our visualization techniques in the context of the protein folding problem, the next section provides a brief overview of fold recognition for predicting the 3D shape of proteins. This is followed by a description of methods for assessing protein sequence alignments. Next we preface a more detailed description of our visualization techniques with a discussion of previous work in this area. We follow this with a description of our open pipe architecture and our editing capabilities. We conclude by summarizing our results and outlining plans for future research.

2 Background

2.1 Protein Structure Prediction

Knowing a protein's structure gives some insight into how the protein works. This insight can be used to guide biological experiments (such as site-directed mutagenesis) to verify the details of functionality and to help discover the genetic basis for inherited diseases. The ability to deduce a protein's structure from its amino acid sequence alone would simplify protein engineering (the modification of an existing protein's residue sequence for the purpose of creating a change in the protein's stability or function) and protein design (the creation of an entirely new protein).

Proteins with similar amino acid sequences will likely possess similar structures and function [6,7]. This makes it possible to predict the overall shape, or fold, of a protein when its amino acid sequence is similar to that of another protein whose structure is already known. An alignment is made between a known structure and a target sequence (see Figure 1). Using the alignment, the target is "threaded" through the structure [8] creating a structural model in which the aligned portions of the target sequence backbone are placed in the same orientations as the corresponding backbone segments of the known structure. In this way, the overall shape of the protein is predicted. But when the similarities between the target sequence and the protein with known structure are small, structural modeling is difficult. In these cases, the alignments

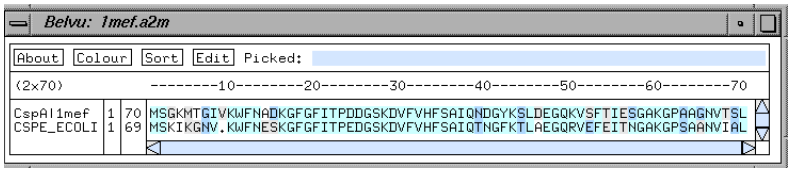

Figure 1: Example protein sequence alignment shown in belvu.

and the corresponding structural models must be studied closely in order to ascertain that they do not violate the accepted heuristics of protein folding.

2.2 Analysis of Alignments and Structural Models

There are many methods for quantifying the similarity of individual amino acids. Some methods compare the amino acid sizes, possible charges, bonding patterns, and other chemical properties [9]. We have several integrated scoring methods available to help assess an alignment. The BLOSUM 62 [10] amino acid substitution matrix can be used as an indicator of alignment quality independent of structural information. This matrix contains a measure of the likelihood of finding a particular amino acid substitution in nature.

In addition to using amino acid similarity measures, when building a structural model of a protein, it is important to analyze the validity of the alignment in the context of the structure's 3-dimensional environment (using criteria such as the preference for an intact core and preferences of the individual amino acid for certain environments and neighboring amino acids). The alignment can be scored using the environmental data as determined by the program *Environments* [11]. This information can be either visualized or, in the future, sonified (e.g. with PROMUSE [12]).

3 Previous Work

Most molecular graphics programs are designed to allow scientists to study a single structure in detail. An example of such a program is RasMol [13] (see Color Plate 1). RasMol allows you to display a molecule in many different modes (backbone, wireframe, ball and stick, etc.). However, RasMol is strictly for molecular visualization, and will neither read nor analyze alignment files.

Of those programs that allow the scientist to use 3-dimensional structural information to analyze alignments, the majority focus on homology modeling rather than threading and therefore display either not enough or too much atomic detail at the level of individual amino acids. One example of a homology modeling package is the Swiss-Model [14] web server, and its associated

visualization tool, Swiss-PDB Viewer [14]. Swiss-PDB Viewer allows the user to thread the target sequence through one or more structures and highlight problem areas. Several other homology modeling visualization systems exist, including the Molecular Applications Group's LOOK, a stand-alone molecular modeling program, and Molecular Simulations Inc.'s HOMOLOGY, an adjunct to the company's molecular graphics package Insight II.

Apart from the alignment evaluation programs based on homology modeling, there are a few notable products designed specifically for analyzing the results of protein threading. One example of such a tool is ANALYST [15], which was developed to visualize the output of the THREADER [16] program. Two other programs useful in analyzing structure-sequence alignments are DINAMO [4] and CINEMA [17]. CINEMA is currently limited to showing only a backbone view of the protein, without any detail at the amino acid level. DINAMO uses Chime [18], a web browser plug-in for viewing molecules. Because Chime is based on RasMol, it is limited to high resolution display.

DINAMO [a] allows multiple sequence alignments, where the first sequence is considered the guide sequence. This tool has an editor which maps colors, as determined by the assessment plug-ins, onto the 1-letter amino acid codes in the 2D alignment and the 3D display.

4 Structure-Sequence Data

In order to display correct structural representations of proteins, we parse PDB [19] files. There are many formats for storing biosequence alignments. Our program reads a format known as the FASTA [20] format. An example of a protein sequence alignment is given in Figure 1.

5 Structure-Sequence Visualizations

The analyses of fold recognition structural models do not involve amino acid rotational angles. In fact, similarity of amino acid angles between the known structure and the target may give the deceptive impression that the region of the model under inspection is superior. As a result, displaying this data can detract from the rest of the picture. One of the tenets of information visualization is to maximize the ratio of information to "ink" [21]. Clearly, in the case of protein fold recognition, showing detailed amino acid structure violates this precept. **ProtAlign** aims to give as much information as necessary to the scientist while eliminating those elements that are unnecessary, detracting, and potentially misleading.

[a] DINAMO [4] is available on the world wide web at http://tito.ucsc.edu/dinamo/

5.1 Visualizing the Amino Acids

To prevent discarding all structural information, we have designed glyphs shaped like children's building blocks to represent the amino acids. The dimensions of the block reflect the overall structure of the amino acid and the shape of the pegs reflects the residue type [3] (see Color Plate 3). Square pegs are used to represent hydrophobic amino acids, conical pegs for charged acidic amino acids, trapezoidal pegs for charged basic amino acids, and cylindrical pegs for polar amino acids. Illustrating our representation, Phenylalanine, an amino acid with a seven carbon side chain, is depicted by a block with seven square pegs. The pegs are arranged to roughly mirror the structural features of phenylalanine's actual chemical framework (see Color Plate 3 and Figure 2B). By varying the layout and shape of the building blocks, we can show why one amino acid might not be a good substitution for another, despite possible similarities in overall shape. Consider the ball and stick depictions of histidine and phenylalanine (see Figures 2A and 2B). Note that for clarity, only the amino acid side chains are drawn. Someone without a background in chemistry might think that the two amino acids are similar enough to be acceptable substitutions for each other. However, as shown in Figure 2C, an alignment containing a substitution of histidine for phenylalanine in our program would give visual cues to the user regarding the poor plausibility of this match. Phenylalanine's hydrophobic nature is indicated by its square shape; similarly, because histidine is a polar molecule it is represented by a cylindrical shape. The "goodness of substitution" between the two residues can be mapped to the color of the two blocks. The color is decided by the current scoring mode chosen by the user. In our case we scored using BLOSUM 62 [10], and red indicates the poor match. In this manner, our glyph depictions convey information on similarity in amino acid structure and properties in a way that is more easily accessible. Further, the compact glyphs present residue information without appearing as busy as a display that contains every atom in the protein structure and the target sequence. This is demonstrated by comparing Color Plate 1 with Color Plate 2. The latter contains twice the information (structure and the target) as the former.

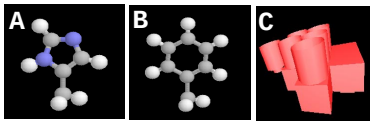

Figure 2: Ball and stick representation of histidine (A) and phenylalanine (B) side chains.[b] Compare to aligned histidine and phenylalanine glyphs (C).

[b] Pictures A and B were created using RasMol.

5.2 Visualizing the Protein Main Chain

Color Plate 2 shows the residue glyphs superimposed on a simple wireframe (backbone) depiction of the protein main chain. The backbone and glyph color indicates the current scoring of the protein (in our case BLOSUM 62).

In addition to amino acid glyphs, we use a protein structure depiction borrowed from comparative streamline visualization. A representation of the protein is created whereby the target and structure proteins are represented as individual "streamlines" following the general shape of the known protein structure. Correspondence between residues in the target and the structure are indicated by line segments connecting the streamlines, similar to rungs on a ladder. The color of each rung is mapped to values in the current scoring scheme, reflecting the suitability of the amino acid substitution at that position in the alignment (see Color Plate 4). Similar to the streamline mode, the ribbon mode enables the user to view the alignment quality using a filled ribbon rather than a wireframe. The strand mode is much like the ribbon mode, except it can be seen through (see Color Plates 5 and 6). There are other modes of drawing the protein. For example, the cartoon mode allows the user to render the main backbone chain using the familiar Richardson's ribbon format[22]. The invisible mode allows the user to select unimportant portions of the protein and make them invisible. This allows closer inspection of the more important areas of the protein alignment.

ProtAlign presents a cleaner picture of the overall structure of the protein. Structural motifs are easier to detect. As evidence of this, compare Color Plate 1 with Color Plates 2, 4, 5, and 6. All show the same protein in the same orientation, but the beta barrel structure is completely obscured in Plate 1.

5.3 Assessing the Alignment

If desired, the user can choose from several local scoring functions to help assess an alignment: BLOSUM 62, Environment, Sonify, or none. When visual cues and scoring are used, color is mapped as follows: red is bad, orange is very poor, yellow is poor, green is considered good, while blue is perfect.

Choosing BLOSUM scores and colors the alignment using the BLOSUM 62 matrix. Environment scoring uses output from the program *Environments*. Visual coloring shows how likely an amino acid substitution is given the environment of the amino acid, the secondary structure, amino acid exposure and overall goodness of fit. Sonification scoring uses the output of *Environments*, but will generate both aural and visual cues to help the user access areas of the alignment[12]. When DINAMO is used with **ProtAlign**, all of the DINAMO scoring plug-in functions can be used to color our 3D alignment.

It is possible to label the amino acid positions with their 1-letter amino

acid codes. As shown in Color Plate 4, our program gives a true 3D analogue of the traditional 2D alignments such as the one in Figure 1.

In streamline mode, streamline rungs indicate the correspondence of amino acid positions in the alignment created versus positions in an ideal (i.e. reference) alignment. The angles of the rungs, as determined by *MeasureShift* [23], reflect the quality of the alignment under assessment. Perpendicular rungs indicate that the amino acids are well aligned, while slanted rungs would indicate that the amino acids were misaligned. Figure 3 depicts a 2D alignment using angled line segments between amino acids to indicate problem areas.

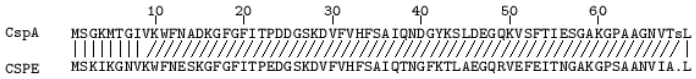

Figure 3: Two Dimensional Alignment With Angled Lines Indicating Mismatched Regions [d]

All of the depictions of the protein backbone (backbone, ribbon, etc.) are capable of low resolution assessment of the alignment. When scoring is used, (red hot for trouble areas, cool blue for great areas) coloring reflects the overall quality of the alignment. When no scoring is enabled, alignment coloration simply reflects the direction of the protein from *'N'* terminus to the *'C'* terminus. This allows scientists to make their own decisions without outside scoring influence.

It is possible to get additional information about individual positions in the alignment. When the user selects an amino acid pair, the position in the alignment and the numerical score of the position are displayed. It is also possible to select portions of the alignment by secondary structure as indicated in PDB file. This allows the user to select specific structural areas to more closely examine without extra clutter. In Color Plate 4, the beta barrel is labeled with the 1-letter residue codes.

6 Pipe Architecture

6.1 Communication mechanisms

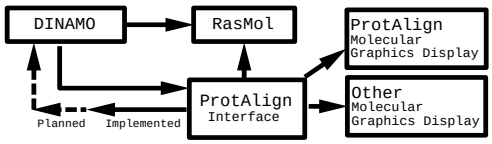

Figure 4: Communication pathways through pipe

[d] This figure generated by the program shift_figure. [23]

ProtAlign implements an open pipe communication design. While the tool is running, it can receive commands from standard in and write commands to standard out. Because of the open pipe communication, drawing commands can be received both externally and from the GUI (see Figure 4). This allows us to make use of the editing and scoring capabilities from a precursor program DINAMO [4]. It also allows our program to drive other graphical visualization tools such as RasMol. **ProtAlign** has all the capabilities of DINAMO plus many visualization capabilities not afforded by RasMol alone.

6.2 DINAMO driving ProtAlign

Because of the piping mechanism, it is possible to make use of the DINAMO 2D alignment editor while accessing all of the scoring features of both DINAMO and **ProtAlign**. In fact, any alignment editor that can give commands through a pipe can access our comparative visualization display. In the past, DINAMO was only capable of using the high resolution drawing program, RasMol, to visualize the alignment. Now, it is possible to drive the low resolution representation of **ProtAlign** from DINAMO.

6.3 ProtAlign driving other programs

When high resolution is desired, the pipe allows **ProtAlign** to use RasMol as visual output. Though the display program is RasMol, the images can be assessed using our scoring functions. Aside from RasMol, ProtAlign can also be used in conjunction with PROMUSE where the user can determine how good the alignment is by listening to aural cues. Finally, it will be possible for DINAMO to drive **ProtAlign** which in turn drives RasMol.

7 Alignment Editing Capabilities

7.1 3-Dimensional editing of the alignment

ProtAlign introduces the ability to directly edit the alignment on the 3D display. It is possible to select a section of the target sequence and drag it along the structural model (see Color Plate 4). This feature is more intuitive than changing a 2D alignment and looking back to the image to see the results. This is especially helpful in closing small gaps in helices and adding gaps in loop regions.

7.2 2-Dimensional editing of the alignment

Because of the command pipe architecture, it is possible to make use of the DINAMO 2D alignment editor. All changes to the alignment are reflected in our 3D graphical display. Currently, **ProtAlign** can be updated from DINAMO. However, the converse is not currently true. That is, when the 3D alignment is

edited directly from within **ProtAlign**, DINAMO's 2D alignment is not correctly updated. We anticipate that this problem can be resolved shortly, and relies primarily on DINAMO to correctly listen (rather than just talk) through the pipe. To address this problem, we currently have a basic 2D alignment viewer and editor to allow the the user to interactively edit the alignment by changing the 2D alignment or the 3D alignment within **ProtAlign**.

8 Conclusions and Future Work

This project builds upon our earlier work [3]. With **ProtAlign** we introduce 3D interactive editing capabilities. We implement a pipe mechanism that allows other programs to interface with our molecular graphics capabilities, and allows **ProtAlign** to communicate with other packages. We offer new methods for viewing and assessing structure-sequence alignments. Building-block glyphs display amino acid structural information in a way that is both compact and accessible to chemists and non-chemists. The streamline representation permits the display of high level structural motifs along with both directional information and alignment quality data. Visual and aural cues make it possible to easily identify problems with the alignment.

As short term goals, we are working with the DINAMO developers to allow DINAMO to listen for **ProtAlign** inputs. This will give full communication between DINAMO and **ProtAlign**. With full communication, **ProtAlign** will have full access to DINAMO's 2D alignment editing tool and scoring plug-ins. In the interim, **ProtAlign** uses its own simple 2D alignment representation.

We are currently researching the value of mapping alignment quality to other display options such as the use of texture mapped images, shininess, opacity, emissivity, building block size, and strand width, thickness, or smoothness.

We are also interested in viewing structure-structure alignments (coordinate files for two protein structures that have been superimposed in three dimensions). Again, our streamline methods could be used to indicate where two protein are most similar in their structures.

As a long term goal, we will be extending our tool for use in high resolution homology modeling. This will require more detailed depiction of amino acids, and would entail implementing the following features:

1. Display of ϕ and ψ backbone angle rotations with the alignment.
2. Estimation of the angles for insertions, deletions, and mutations. These would be generated using molecular dynamics.
3. Improved navigational and interrogation aid for working within the complex 3D structure.

4. The ability to save the coordinate files for the predicted structure in standard PDB format.

Check the following URL for updated information on this work:
www.cse.ucsc.edu/research/avis/bio.html.

Acknowledgements

We would like to thank Leslie Grate, Albion Baucom, Jesse Bentz, Lydia Gregoret, and Srividya Ananthanarayanan for their help in the development of the precursors to this project. We would also like to thank the Santa Cruz Laboratory for Visualization and Graphics (SLVG) for the research environment. Marc Hansen and Doanna Meads are supported by GAANN fellowships. This projected is supported by DARPA grant N66001-97-8900, ONR grant N00014-96-1-0949, NSF grant IRI-9423881, and NASA grant NCC2-5281.

References

1. E. H. Chi, P. Barry, E. Shoop, J. Carlis, E. Retzel, and J. Riedl. Visualization of biological sequence similarity search results. In *Proceedings of Visualization 95*, pages 44–51, 1995.
2. E. H. Chi, J. Riedl, E. Shoop, J. V. Carlis, E. Retzel, and P. Barry. Flexible information visualization of multivariate data from biological sequence similarity searches. In *Proceedings of Visualization 96*, pages 133–140,477, 1996.
3. Marc Hansen, Doanna Meads, and Alex Pang. Comparative visualization of protein structure–sequence alignments. In *IEEE Information Visualization*, 1998. To appear.
4. Marc Hansen, Jesse Bentz, Albion Baucom, and Lydia Gregoret. DINAMO: a coupled sequence alignment editor/molecular graphics tool for interactive homology modeling of proteins. In *Pacific Symposium on Biocomputing*, volume 3, pages 106–117, 1998.
5. S. K. Lodha, Alex Pang, Robert E. Sheehan, and Craig M. Wittenbrink. UFLOW: Visualizing uncertainty in fluid flow. In *Proceedings of Visualization 96*, pages 249–254, October 1996.
6. C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991.
7. Jürgen Bajorath, Ronald Stenkamp, and Alejandro Aruffo. Knowledge-based model building of proteins: Concepts and examples. *Protein Science*, 2:1798–1810, 1993.

8. David Jones and Janet Thornton. Protein fold recognition. *Journal of Computer-Aided Molecular Design*, 7:439–456, 1993.
9. Markéta J. Zvelebil, Geoffrey J. Barton, William R. Taylor, and Michael J. E. Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology*, 195:957–961, 1987.
10. Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89:10915–10919, November 1992.
11. James U. Bowie, Roland Luthy, and David Eisenberg. A method to identify protein sequences that fold into a known three–dimensional structure. *Science*, 253:164–170, 1991.
12. Marc Hansen and Erik Charp. Multi–modal visualization of local environment data for protein structural alignments. Technical Report UCSC–CRL–98–08, UCSC Computer Science Department, 1998.
13. Roger Sayle and E.J. Milner-White. RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences*, 20:374–376, 1995.
14. N. Guex and M. C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis*, 18:2714–2723, 1997.
15. R. T. Miller, David T. Jones, and Janet M. Thornton. Protein fold recognition by sequence threading: tools and assessment techniques. *Fasb J*, 10(1):171–178, 1996.
16. David Jones, W. Taylor, and Janet Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.
17. T. K. Attwood, A. W. R. Payne, A. D. Michie, and D. J. Parry-Smith. A Colour INteractive Editor for Multiple Alignments - CINEMA. *EMBnet.news*, 3(3), 1997.
18. Chime home page. http://www.mdli.com/chemscape/chime, 1997.
19. F.C. Bernstein, T.F. Koetzle, Jr. G.J.B Williams, et al. The protein data bank: A computer based archival file for macromolecular structures. *Journal of Molecular Biology*, 112:535–542, 1977.
20. W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.
21. Edward R. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, 1983.
22. J. S. Richardson. Schematic drawings of protein structures. *Methods of Enzymology*, 115:358–380, 1985.
23. Melissa Cline and Kevin Karplus. On alignment shift and its measures. Tech Report UCSC–CRL–97–27, UCSC Computer Science Dept., 1998.