Comparative Visualization of Protein Structure-Sequence Alignments

Marc Hansen, Doanna Meads, Alex Pang
Computer Science Department, UCSC
(mhansen,doanna,pang)@cse.ucsc.edu

Abstract

Protein fold recognition (threading) involves the prediction of a protein's three-dimensional shape based on its similarity to a protein whose structure is known. Fold predictions are low resolution; no effort is made to rotate the protein's component amino acid side chains into their correct spatial orientations. Rather, the goal is 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 structure (i.e., a structure-sequence alignment). To complement this structure prediction method we have implemented a low resolution molecular graphics tool. Since amino acid side chain orientation is not relevant in fold recognition, amino acid residues are represented by abstract shapes or glyphs much like Lego (tm) blocks. We also borrow techniques from comparative streamline visualization to provide clean depictions of the entire protein structure model. By creating a low resolution representation of protein structure, we are able to approximately double the amount of information on the screen. This implementation also possesses the advantage of eliminating distracting and possibly misleading visual clutter resulting from the mapping of protein alignment information onto a high resolution display of a known structure.

Key Words and Phrases: proteins, structure, alignment, fold recognition, threading, similarity, glyphs, streamlines, amino acids.

1 INTRODUCTION

Proteins are the "machinery" of the cell, and are responsible for such diverse tasks as facilitating chemical reactions and transporting molecules. It is the protein's central role in the workings of the cell that makes it an important target for investigation. 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 human and other 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, the field of bioinformatics is becoming an attractive target for the application of visualization techniques. [8] [9] Using information visualization techniques, researchers can often see the results of their experimental methods more clearly than by simply looking at raw numbers. For example, a protein sequence alignment (see figure 4) may obtain a reasonable numerical score, but visual inspection of the structural model might reveal incongruencies with 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 two problems: **(1)** It is difficult to incorporate 3D data into visual displays for the purpose of analyzing the validity of individual amino acid placements. This difficulty arises as a result of the visual clutter which normally ensues when large amounts of atomic data are displayed at high resolution (see figure 1). **(2)** While there exist several tools for displaying 2D bio-sequence alignments (see figure 4), the tools for viewing the corresponding 3D comparisons are limited to showing too much information or not enough.

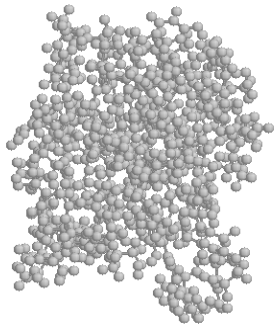

Figure 1: RasMol's [22] ball and stick depiction of cold-shock protein 1mef's chain B

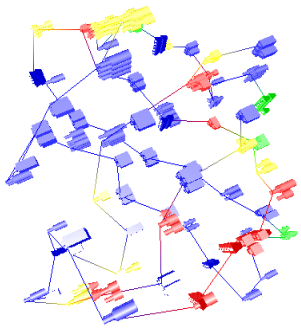

Figure 2: Glyph depiction of cold-shock protein 1mef's chain B along with target sequence

To address the first problem, we have developed a low resolution representation of protein molecular structure, using building block glyphs to represent amino acids (see figure 6). By focusing on the major structural aspects of amino acids, we are able to display approximately twice as much information as is commonly shown in high resolution protein depictions, while making the overall representation cleaner (compare figures 1 and 2). We have borrowed techniques used in UFLOW [18] comparative streamline visualization to remedy the second problem. The target and structure proteins are represented as individual "streamlines". Correspondences between residues in the target and the structure are indicated by line segments connecting the streamlines, much like rungs on a ladder (see figure 8).

In order to facilitate a discussion of our visualization techniques in the context of the protein folding problem, the following section provides a brief overview of protein structure and introduces a specific technique, threading (also called fold recognition), for predicting the 3D shape of proteins. A description of methods for assessing protein sequence alignments follows. We preface a more detailed description of our visualization techniques with a discussion of previous work in this area. Finally, we conclude by summarizing our results and outlining plans for future research.

2 BACKGROUND

2.1 Proteins

Proteins are naturally occurring polymers usually consisting of hundreds of amino acids. As shown in figure 3, each amino acid contains a side chain, or R group, which makes that amino acid unique. A full listing of these side chains is given in figure 6. In a protein, the amino acids (also called residues) are linked together

somewhat like beads on a string. Proteins have direction: along the main chain, or backbone, amino acids are linked from their end C' carbon to the next residue's beginning Nitrogen (N) atom. The chemical and physical properties of the amino acids which comprise a protein determine how the protein will fold upon itself.

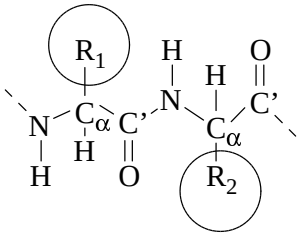

Figure 3: Protein backbone segment[1]

2.2 Protein Structure Prediction

In the late 1950's, Christian Anfinsen [3] discovered that after the protein ribonuclease was unfolded (denatured) by chemical means, the denaturants could be removed, allowing the protein to refold and slowly regain its original catalytic ability. Since a protein's function is dependent on its shape, this pioneering work showed that the information for determining the three dimensional structure of the protein is contained in its one dimensional string of amino acids.

Since this discovery, scientists have grappled with the protein folding problem, which can be succinctly stated as: Given a protein's amino acid sequence, what will its three-dimensional shape be? As an indication of the importance of this question, protein folding is often called the second half of the genetic code.

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. In addition, structural information can be used to develop new pharmaceutical products to interact with the protein and modulate its function. Drug design is an extremely lengthy and costly process, and any tool which could help bring a product to market more quickly would be a boon, not only to the pharmaceutical industry, but to the entire population. The ability to deduce a protein's structure from its amino acid sequence alone would also 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, much as an architect would design a new building).

It is believed that proteins fold so as to minimize the energy level of the molecule as a whole. Unfortunately, there currently exists no method that uses positional atomic energy levels to accurately predict a protein's structure. The number of possible orientations each amino acid in the sequence could take is so great that it would be impractical for a computer to sample the entire conformational space and pick the fold that results in the lowest energy level. For example, if we were to allow each amino acid only seven accessible rotational states, a small protein with only 50 residues would have 7^{50} ($\approx 10^{42}$) possible conformations for the backbone alone. [15] Groups have labored trying to predict the structures of small loop regions using sequence information alone. Even for chains of only 2 to 5 amino acids long, however, the success rate is under 65%. [21]

[1]This figure was adapted from Branden & Tooze. [7]

Fortunately, it has been shown that proteins with similar amino acid sequences will likely possess similar structures and function. [5] This fact makes it possible to predict the overall shape, or fold, of a protein if its amino acid sequence is similar to that of another protein whose structure is already known.

2.3 Methods for Building Structure Models

The process of fold recognition begins with the acquisition of a target protein whose sequence is known, but whose structure is not. This target is used to query a database of proteins whose structures are known. If a match is found, the two sequences can be aligned such that similar amino acids are placed in the same columns (see figure 4).

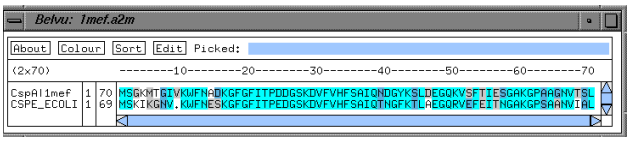

Figure 4: Example protein sequence alignment shown in belvu [1]

Once the alignment between the target and the structure has been created, the target can be "threaded" through the structure. [16] This involves the creation of 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.

Homology modeling (sometimes called comparative modeling) is a structure prediction method similar to threading, but in which the aim is to create a high resolution estimate of the protein's structure. This prediction method attempts to go beyond the threading method and also predict the rotational positions of the target's amino acid side chains. [5] This is usually accomplished using energy minimization algorithms. Typically, a higher degree of similarity between the protein sequences in the alignment is needed in order to carry out homology modeling.

In general, structural modeling is more difficult when the similarities between the target sequence and the protein with the known structure are small. Especially in these cases, the alignments and the corresponding structural models must be studied closely in order to ascertain that they do not violate the accepted heuristics of protein folding.

2.4 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. [24] The measures based on evolutionary mutation rates do not specifically take structural information into consideration. Nonetheless, they provide a useful initial analysis of the quality of a sequence alignment.

We use the BLOSUM 62 [14] amino acid substitution matrix as an indicator of alignment quality independent of structural information. Rather than taking into consideration an individual residue's environment in relation to the structure of the entire protein, this matrix contains a measure of the likelihood of finding a particular amino acid substitution in nature. The BLOSUM 62 matrix was derived by aligning segments of proteins that were 62% identical, and transforming the ratios of observed versus expected amino acids into whole numbers. In the BLOSUM 62 matrix, a zero indicates that the amino acid substitution frequencies observed are those expected due to random chance. Positive numbers indicate that the

frequencies observed were higher than expected, while the converse is true for negative numbers.

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 three-dimensional environment. There are several important guidelines used to evaluate such a model. Bajorath et al. [5] provides an excellent review of typical assessment methods.

1. The protein's inner core tends to be more evolutionarily conserved than the outer loops. Since the core regions are responsible for the general structure, the protein's stability is more sensitive to mutations in this area. As a corollary, if there are insertions or deletions in the target protein relative to the structure, they should preferably occur in the variable loop regions.
2. Hydrophobic, i.e., water avoiding, amino acids are more energetically favorable when they reside in the interior of the protein, away from the solvent. Conversely, hydrophilic, or water preferring, amino acids tend to occur on the exterior regions of the protein in contact with the solvent.
3. The patterns in the local residue conformations of a protein are known as secondary structure. A protein's secondary structure can be predicted using only its sequence of amino acids. Due to the relatively high accuracy of secondary structure predictions, (approximately 65-72% [20]), they are useful as another test in determining the quality of a structural model: If the predicted secondary structure regions for the target sequence alone match the actual secondary structure regions of the sequence with known structure, this will lend credence both to the alignment between the two, as well as the model.
4. Another important criterion is packing, i.e., the density of amino acids in the protein. Differences in the volumes of target versus structure amino acids can lead to over- or under-packing. [10]
5. Each amino acid has preferences both for which other amino acids it is more likely to contact, and for which environments are more likely to be found surrounding it. Preference profiles are useful in judging whether an alignment makes sense structurally. [11]
6. Finally, structural positions crucial to the shape of the protein tend to be conserved. For example, if a protein contains a tight hairpin turn, it may be the case that only a very small amino acid such as glycine will fit there. In such cases it is important to verify that the target protein's amino acid at this location fits the constraints imposed by the existing structure.

3 PREVIOUS WORK

Most molecular graphics programs are designed to allow scientists to study one structure in detail. An example of such a program is RasMol [22] (see figures 1 and 7). 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 therefore neither read nor analyze alignment files.

Of those programs which do allow the scientist to use three-dimensional structural information to analyze alignments, the majority focus on the problem of 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 [12] web server, and its associated visualization tool, Swiss-PDB Viewer. [12] Swiss-PDB Viewer allows the user to thread the target sequence through one or more structures and highlight problem areas. Swiss-PDB Viewer also allows the protein to be displayed in traditional modes such as backbone, ribbon and wireframe. 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, [19] which was developed to visualize the output of the THREADER [15] program. Two other programs useful in analyzing structure-sequence alignments are DINAMO [13] and CINEMA. [4] DINAMO uses Chime, [2] a web browser plug-in for viewing molecules. Chime is based on RasMol, and is therefore limited to displaying an existing structure and using display options such as color to indicate alignment quality. CINEMA is currently limited to showing only a backbone view of the protein, without any detail at the amino acid level.

Most of the programs described previously have the advantage of allowing the user to interactively edit a sequence alignment and view the resulting analyses in relation to the protein structure. None of them, however, use structurally based glyphs to represent the amino acids, nor do they offer a streamline representation of an alignment. They therefore suffer from the resulting problems of either showing too little or too much information as a consequence of the complexity inherent in most protein structures.

4 STRUCTURE-SEQUENCE DATA

Detailed protein structural information is most commonly found in a format established by the Brookhaven Protein Data Bank (PDB). [6] This file format contains x, y, and z coordinates for atoms contained in the protein (see columns 6-8 of figure 5). In order to display correct structural representations of proteins, we wrote a parser for PDB files.

ATOM	1	N	MET	1	-3.860	-7.574	-8.551	1.00	0.00	N
ATOM	2	CA	MET	1	-4.253	-7.422	-9.942	1.00	0.00	C
ATOM	3	C	MET	1	-5.676	-6.870	-10.087	1.00	0.00	C
ATOM	4	O	MET	1	-5.879	-5.794	-10.641	1.00	0.00	O
ATOM	5	CB	MET	1	-4.166	-8.805	-10.639	1.00	0.00	C
ATOM	6	CG	MET	1	-2.824	-9.532	-10.458	1.00	0.00	C
ATOM	7	SD	MET	1	-2.821	-11.078	-11.398	1.00	0.00	S
ATOM	8	CE	MET	1	-2.281	-10.472	-13.016	1.00	0.00	C
ATOM	9	H	MET	1	-3.989	-6.698	-8.000	1.00	0.00	H
ATOM	10	2H	MET	1	-4.425	-8.334	-8.131	1.00	0.00	H
ATOM	11	3H	MET	1	-2.879	-7.900	-8.478	1.00	0.00	H
ATOM	12	HA	MET	1	-3.572	-6.748	-10.441	1.00	0.00	H
ATOM	13	1HB	MET	1	-4.943	-9.460	-10.265	1.00	0.00	H
ATOM	14	2HB	MET	1	-4.331	-8.667	-11.698	1.00	0.00	H
ATOM	15	1HG	MET	1	-2.007	-8.909	-10.795	1.00	0.00	H
ATOM	16	2HG	MET	1	-2.682	-9.786	-9.415	1.00	0.00	H
ATOM	17	1HE	MET	1	-2.537	-9.429	-13.121	1.00	0.00	H
ATOM	18	2HE	MET	1	-1.214	-10.602	-13.121	1.00	0.00	H
ATOM	19	3HE	MET	1	-2.800	-11.023	-13.787	1.00	0.00	H
ATOM	20	N	SER	2	-6.634	-7.641	-9.557	1.00	0.00	N
	...									
ATOM	1022	2HD1	LEU	70	2.560	7.981	-7.732	1.00	0.00	H
ATOM	1023	3HD1	LEU	70	1.220	8.379	-8.796	1.00	0.00	H
ATOM	1024	1HD2	LEU	70	0.921	6.241	-6.388	1.00	0.00	H
ATOM	1025	2HD2	LEU	70	-0.706	6.843	-6.123	1.00	0.00	H
ATOM	1026	3HD2	LEU	70	-0.092	6.653	-7.766	1.00	0.00	H
TER	1027		LEU	70						

Figure 5: Example of the protein data bank (PDB) file format

There are many formats for storing biosequence alignments. Our program reads a format known as A2M, for "align-to-model". An example of a protein sequence alignment is given in figure 4.

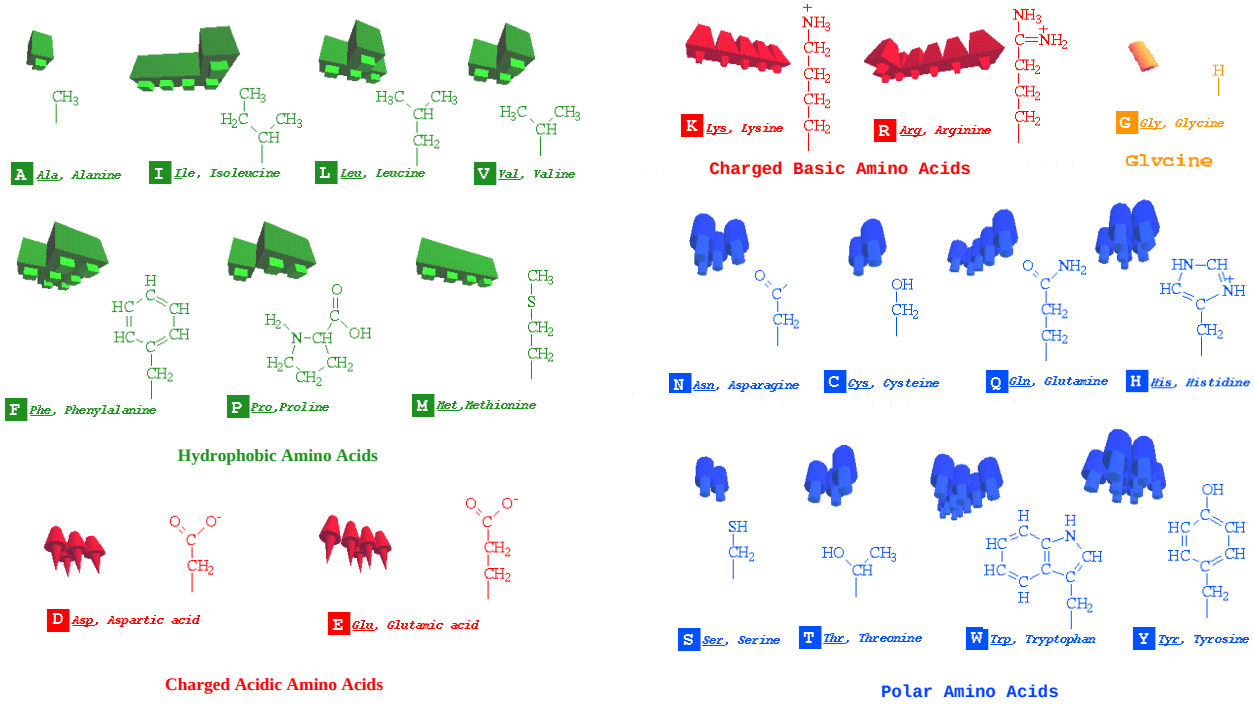

Figure 6: Amino acid building blocks

5 STRUCTURE-SEQUENCE VISUALIZATIONS

The analyses of fold recognition structural models do not involve amino acid rotational angles. As a result, displaying this data can detract from the rest of the picture. Further, if the amino acid angles are similar, this similarity may give the deceptive impression that the region of the model under inspection is superior. One of the tenets of information visualization is to maximize the ratio of information to "ink". [23] Clearly, in the case of protein fold recognition, showing detailed amino acid structure violates this precept. Our program aims to give as much information as necessary to the scientist while eliminating elements that are unnecessary, detracting, and possibly misleading.

5.1 Visualizing the Amino Acids

We have implemented glyphs to represent amino acids in order to prevent discarding all of their structural information. Our glyphs are shaped like children's building blocks, with the dimensions of the block reflecting the overall shape of the amino acid, and the shape of the pegs reflecting the residue type (see figure 6). In all cases (except glycine) we omit information for hydrogens. Glycine's side chain consists solely of one hydrogen; therefore, we depict it using a smaller peg to distinguish it from the other larger side chains. The majority of PDB files do not include positional information for all of the hydrogens, and for our purposes the potential distraction imposed by these atoms is not offset by the limited additional structural information they would provide.

As an example of one of our building block representations: phenylalanine, an amino acid with a side chain containing seven carbon atoms, is depicted by a block consisting of seven square pegs roughly arranged to mirror the structural features of phenylalanine's actual chemical framework (see figure 6 and 7B). By varying the layout and shape of our building blocks, we are able to show why one amino acid might not be a good substitution for another, despite possible similarities in overall shape. For example, based on their ball and stick chemical pictures the amino acids histidine and phenylalanine appear similar (see figures 7A and 7B). (Note: 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. For the non-chemist, it is difficult to determine an amino acid's properties (hydrophobicity, polarity, etc.) from a traditional chemical drawing of the amino acid.

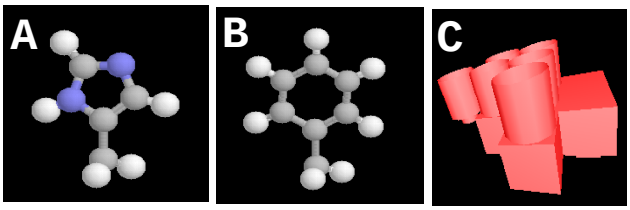

Figure 7: Ball and stick representation of histidine (A) and phenylalanine (B) side chains.[2] Compare to aligned phenylalanine and histidine glyphs (C).

As shown in figure 7C, an alignment containing a substitution of histidine for phenylalanine in our program would give several visual cues to the user regarding the plausibility of this match. The similarity between the two residues according to the BLOSUM 62

[2]Pictures A and B were created using RasMol. [22]

matrix is mapped to the color of the two blocks. In this case red indicates a poor match (in fact, the actual score is -2). Phenylalanine's hydrophobic nature is indicated by its square shape; similarly, the fact that histidine is a polar molecule is represented by its cylindrical structure. In this manner, our glyph depictions convey information on similarity in amino acid structure and properties in a way that is more easily accessible to those possessing a limited familiarity with chemistry. Further, the compact glyphs present this information without appearing as busy as a display of every atom in the protein structure and the target sequence would be.

5.2 Visualizing the Protein Main Chain

Figure 2 shows the residue glyphs superimposed on a simple wireframe depiction of the protein main chain. In this example the color of the backbone indicates the direction of the protein, and is interpolated along the length of the protein from Red (at the beginning N terminus) to Orange, Yellow, Green, and finally Blue (at the ending C' terminus).

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 two streamlines follow the general shape of the protein. Similar to rungs on a ladder, line segments connect the two streamlines at amino acid positions to indicate the suitability of the amino acid substitution occurring at that position (see figure 8). The color of the rungs is mapped to values in the BLOSUM 62 amino acid substitution matrix, and reflects the similarity of the amino acids matched at that position in the alignment.

Similar to the streamline mode, the ribbon mode enables the user to view the alignment quality using a filled ribbon rather than a wireframe one (see figure 9).

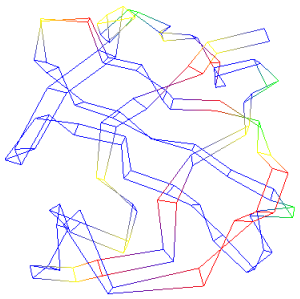

Figure 8: Streamline style display of alignment

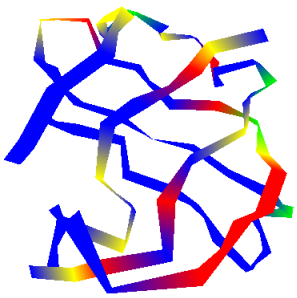

Figure 9: Ribbon style display of alignment

Both the streamline and ribbon representations have the benefit of presenting a clean picture of the overall structure of the protein. Overall structural motifs are easier to detect in this mode. As evidence of this, compare figure 1 with figures 2, 8, and 9. All show the same protein in the same orientation, but the barrel structure is completely obscured in the first figure.

6 CONCLUSIONS AND FUTURE WORK

This project is a continuation of earlier work developed to assist our computational biology group during the protein structure prediction contest CASP (Critical Assessment of Techniques for Protein Structure Prediction). [17] Our group used Leslie Grate's SAE, (a prototype tool not intended for release), which combined an alignment editor with RasMol. After the contest, we decided that it would be useful both to create an alignment assessment tool for release and to develop additional ways to view structure-sequence alignments in order to take advantage of available 3D information. To those ends, we previously developed DINAMO, and now present our current efforts.

With this tool we offer two new methods for viewing structure-sequence alignments: **(1)** Building block glyphs display amino acid structural information in a way that is both compact and accessible to non-chemists; **(2)** Streamline representation permits the display of high level structural motifs along with both directional information and alignment quality data.

We have three immediate goals to be completed in the next several months. Our first goal is to label the amino acid positions with either their 1 or 3 letter amino acid codes or their sequence numbers. This goal will make our program the first to give a true 3D analogue of the traditional alignments such as the one in figure 4.

Our second immediate goal is to add more options for drawing the molecules. An example of such an option would be a cartoon representation of secondary structure in which corkscrews or cylinders would stand for alpha helices, and directed arrow strands would represent beta sheets.

Our third immediate goal is to read alignment reference files. This enhancement would allow us to use streamline rungs to 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 would then reflect the quality of the alignment under assessment. Vertical rungs would indicate that the amino acids were well aligned, while slanted rungs would indicate that the amino acids were misaligned. See figure 10 for an alignment depiction that uses angled line segments between amino acids to indicate problem areas.

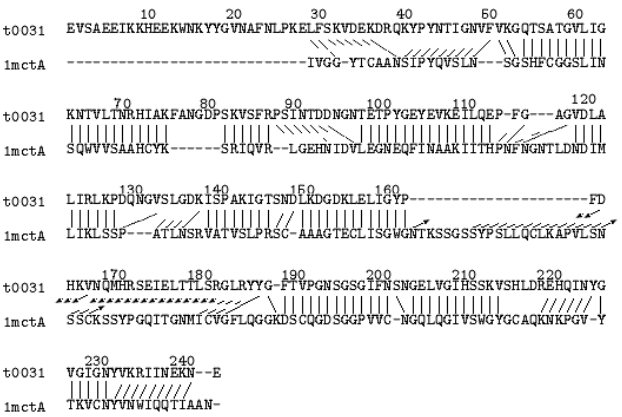

Figure 10: Example of an alignment with angled lines to indicate mismatch regions [3]

Our last short term goal is to modify DINAMO to allow the use of our molecular graphics program as its display tool. This would allow us to avail ourselves of existing display options and alignment assessment algorithms already in place in DINAMO, (such as coloring by predicted versus actual secondary structure). In addition, it would allow the user to edit and save alignment files interactively, while immediately seeing the results.

As longer term goals, we are 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,

It is somewhat cumbersome to rotate the entire molecule when the interest may lie in a small stretch of the protein. It may be useful to provide a pop-up window for viewing a single amino acid

[3] This figure duplicated with author's permission. [17]

substitution pair (as depicted in figure 7C) independently of the rest of the protein.

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

1. Allow the display of ϕ and ψ backbone angles in addition to the alignment.
2. Estimate the amino acid angles for insertions, deletions, and mutations. These would be generated using molecular dynamics.
3. 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 for creating the Sam Alignment Editor (SAE), a prototype tool for assessing alignments in their 3D context. Much of the work undertaken here is a continuation of that project. Albion Baucom helped create a prior program for visualizing proteins using graphics primitives as glyphs, and along with Jesse Bentz and Lydia Gregoret helped developed DINAMO, the immediate precursor to this project. Bruce Meads took part in discussions on breaking a protein down to the atomic level. Srividya Ananthanarayanan helped write the code for parsing and storing data from PDB files. Dawn Davis reviewed the manuscript and provided editorial assistance. The following faculty have been the driving force in protein folding and bioinformatics at UCSC: Tony Fink, Lydia Gregoret, David Haussler, Richard Hughey, Kevin Karplus, and Todd Wipke. Marc Hansen and Doanna Meads are supported by GAANN fellowships. This project is supported by NASA grant NCC2-5207, NSF grant IRI-9423881, DARPA grant N66001-97-8900, and ONR grant N00014-92-J-1807.

References

[1] Belvu home page. http://www.sanger.ac.uk/~esr/Belvu.html.

[2] Chime home page. http://www.mdli.com/chemscape/chime, 1997.

[3] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[4] 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.

[5] Jürgen Bajorath, Ronald Stenkamp, and Alejandro Aruffo. Knowledge-based model building of proteins: Concepts and examples. *Protein Science*, 2:1798–1810, 1993.

[6] 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.

[7] Carl Branden and John Tooze. *Introduction to Protein Structure*. Garland Publishing, 1991.

[8] 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.

[9] 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.

[10] Lydia M. Gregoret and Fred E. Cohen. Novel method for the rapid evaluation of packing in protein structures. *Journal of Molecular Biology*, 211:959–974, 1990.

[11] M. Gribskov, A.D. McLaschlan, and D.E. Eisenberg. Profile analysis: detection of distantly related proteins. In *Proceedings of National Academy of Science*, volume 84, pages 4355–4358, 1987.

[12] N. Guex and M. C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. *Electrophoresis*, 18:2714–2723, 1997.

[13] 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.

[14] 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.

[15] David Jones, W. Taylor, and Janet Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[16] David Jones and Janet Thornton. Protein fold recognition. *Journal of Computer-Aided Molecular Design*, 7:439–456, 1993.

[17] Kevin Karplus, Kimmen Sjölander, Christian Barrett, Melissa Cline, and David Haussler. Predicting protein structure using hidden markov models. *Proteins: Structure, Function, and Genetics*, Supplement 1(1):134–139, 1997.

[18] 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.

[19] 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.

[20] Burkhard Rost and Sean O'Donoghue. Sisyphus and protein structure prediction. *CABIOS*, 13:345–356, 1997.

[21] Stephen D. Rufino, Luis E. Donate, Luis E. Donate, Luc H. J. Cananrd, and Tom L. Blundell. Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modelling. *Journal of Molecular Biology*, 267(2):352–367, March 1997.

[22] Roger Sayle and E.J. Milner-White. RasMol: Biomolecular graphics for all. *Trends in Biochemical Sciences*, 20:374–376, 1995.

[23] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[24] 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.