**DINAMO: interactive protein alignment and model building**

Jesse Bentz¹, Albion Baucom², Marc Hansen² and Lydia M. Gregoret³

Departments of ¹Biology, ²Computer Science and ³Chemistry & Biochemistry, University of California, Santa Cruz, CA 95064, USA

Received on September 17, 1998; revised on December 18, 1998; accepted on January 2, 1999

**Abstract**

**Motivation:** To facilitate the process of structure prediction by both comparative modeling and fold recognition, we describe DINAMO, an interactive protein alignment building and model evaluation tool that dynamically couples a multiple sequence alignment editor to a molecular graphics display. DINAMO allows the user to optimize the alignment and model to satisfy the known heuristics of protein structure by means of a set of analysis tools. The analysis tools return information to both the alignment editor and graphics model in the form of visual cues (color, shape), allowing for rapid evaluation. Several analysis tools may be employed, including residue conservation, residue properties (charge, hydrophobicity, volume), residue environmental preference, and secondary structure propensity.

**Results:** We demonstrate DINAMO by building a model for submission in the 3rd annual Critical Assessment of Techniques for Protein Structure Prediction (CASP3) contest.

**Availability:** DINAMO is freely available as a local application or Web-based Java applet at http://tito.ucsc.edu/dinamo

**Contact:** gregoret@hydrogen.ucsc.edu

**Introduction**

A three-dimensional (3D) model can be invaluable in guiding experiments to investigate protein function (Klein et al., 1997; Quondam et al., 1997; Strahs and Weinstein, 1997; Moro et al., 1998; Navaratnam et al., 1998). However, the availability of sequence information generally precedes a high-resolution crystal or NMR structure by several years, should the structure ever be solved. In such cases, building an approximate model of the protein under investigation can provide a rational starting point for experimental design, particularly for site-directed mutagenesis experiments aimed at elucidating the role of specific amino acid residues. Insights into the relative positions of key functional or structural residues can save time and energy in discerning the molecular details of protein function.

Two excellent methods for building an approximate structure are comparative modeling and fold recognition. Both methods involve using the template structure as a scaffold to position the residues of the target sequence in space. In the case of comparative modeling, the structural template is provided by a protein sharing high sequence homology with the target sequence (Bajorath et al., 1993). When two sequences share significant sequence homology, their structures tend to be highly similar. Specifically, when two proteins share at least 50% residue identity, the relative mean square deviation (rmsd) of their Cα coordinates is predicted to be only 1 Å (Chothia and Lesk, 1986). Thus, if the novel protein sequence is at least 25–30% identical with a protein of known structure, an accurate model of the novel protein can be built.

Fold recognition methods can serve as an alternative basis for structure approximation when no sequence homologs of known structure are available. Fold recognition depends on the observation that sequences sharing no sequence homology can have similar folds. Indeed, distinct fold families are emerging which are independent of sequence (Chothia, 1992; Orengo et al., 1994; Russell et al., 1997). For example, triosephosphate isomerase (TIM) and tryptophan synthase are both members of the ‘TIM barrel’ fold family, yet are functionally unrelated, non-homologous proteins (Orengo et al., 1994). Fold recognition methods score alignments of the target protein against a library of fold models, and select the model with the highest score as a template. Many scoring methods exist (Bowie et al., 1990, 1991; Flockner et al., 1995; Fischer and Eisenberg, 1996; Rost et al., 1997), but in general the score is based on a structure-based alignment between the two sequences, and is determined by the probability of the target sequence adopting the template fold.

To provide an environment in which to build sequence alignments and assess the plausibility of resulting models, we have developed DINAMO, an interactive sequence alignment and model building tool (Hansen et al., 1998). DINAMO provides alignment and model quality assessment via a number of analysis tools. The analysis tools return information in the form of visual cues which alert the user as to the feasibility of each residue’s position in the alignment and the model. The user may then evaluate and modify the alignment/model based on his or her judgement. As such,
DINAMO serves as a workbench to organize and visualize available information, but does not automate the modeling process.

Table 1. Internet locations

<table>
<thead>
<tr>
<th>Service</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DINAMO</td>
<td><a href="http://tito.ucsc.edu/dinamo">http://tito.ucsc.edu/dinamo</a></td>
</tr>
<tr>
<td>CHIME</td>
<td><a href="http://www.mdli.com/">http://www.mdli.com/</a></td>
</tr>
<tr>
<td>BLAST</td>
<td><a href="http://www.ncbi.nlm.nih.gov/BLAST">http://www.ncbi.nlm.nih.gov/BLAST</a></td>
</tr>
<tr>
<td>THREADER</td>
<td><a href="http://globin.bio.warwick.ac.uk/~jones/threader.html">http://globin.bio.warwick.ac.uk/~jones/threader.html</a></td>
</tr>
<tr>
<td>DALI</td>
<td><a href="http://www2.ebi.ac.uk/dali/">http://www2.ebi.ac.uk/dali/</a></td>
</tr>
<tr>
<td>Protein Data</td>
<td><a href="http://www.pdb.bnl.gov/">http://www.pdb.bnl.gov/</a></td>
</tr>
</tbody>
</table>

DINAMO may operate as either a Web-based applet in conjunction with the Internet browser plug-in Chime (Table 1), or as a browser-independent local application that utilizes Rasmd (Sayle and Milner-White, 1995) as the molecular graphics display. We chose Java as the programming language because of its cross-platform compatibility and widespread acceptance. DINAMO is simple to use and does not require the acquisition of complicated and expensive software or hardware. Our intended user is the biologist, chemist or computer scientist who would benefit from insight into the structure of their protein of interest, yet may not have access to high-end molecular graphics software. DINAMO may also be suitable for student use in biochemistry courses.

DINAMO differs from other comparative modeling tools (Sali et al., 1995; Miller et al., 1996; Guex and Peitsch, 1997; Sanchez and Sali, 1997) in that it is simple, Web based and completely interactive. As a result, the user may employ his or her scientific judgement and experience throughout the process by manually refining the alignment and hence the model. However, prediction accuracy and the level of detail may be improved by using DINAMO in conjunction with external tools and services, such as CLUSTALW (Thompson et al., 1994) for generating the initial alignment and SWISS-MODEL (Guex and Peitsch, 1997) for generating atomic coordinates. Alignments based on the superposition of 3D structures can be obtained from the FSSP (Fold classification based on Structure-Structure alignment of Proteins) database (Holm and Sander, 1998). Format converters to facilitate input and output between DINAMO and these applications have been developed.

System and methods

DINAMO consists of five major components: (i) a sequence alignment editor; (ii) a molecular graphics display; (iii) analysis tools to evaluate alignment and model quality; (iv) an analysis tool manager which controls how analysis tool information is visualized in the editor and the graphics; and (v) an alignment router which sends alignment information to active analysis tools. A diagram of DINAMO’s architecture is shown in Figure 1. In summary, as the user adjusts the

Fig. 1. DINAMO architecture. Alignment information is sent to active analysis tools from the editor via the alignment router. Analysis tools return information to both the editor and the graphics, where it is displayed by color or shape.
sequence alignment, the editor sends information to active analysis tools. The analysis tools return alignment assessment information to both the alignment editor and graphics display.

**Sequence alignment editor**

The sequence alignment editor may act as an independent alignment editor or in conjunction with the analysis tools and/or molecular graphics display. When used as a 3D modeling tool, the editor allows the user to align a target sequence to a template sequence (i.e. the sequence for which a structure has been experimentally determined). The editor also implements a set of radio buttons that allow selection of the sequence that is currently being modeled against the template sequence. Thus, although only one 3D model is displayed at a time, the user can switch back and forth between various target sequences to evaluate better a model based on a multiple alignment.

Sequences are loaded into DINAMO by pasting them into a window activated by the ‘Load Alignment’ menu option. The sequences can be listed sequentially, separated by new lines, and must be in FASTA format: >sequence1[CR]sequence of protein #1 in one letter codes (where [CR] indicates carriage return). There is no upper limit to the number of sequences that can be loaded into the editor.

Once loaded, sequences are represented in the alignment editor by a string of residue tiles. Each tile has the capacity to convey at least three pieces of information returned by the analysis tools through the use of color, tile shape and font color. As the user adjusts the alignment by clicking on the tiles and sliding them, alignment information is sent to the analysis tools, then back to the editor and the graphics. Thus,
Fig. 3. The editor containing sequences 1ndh, 1fdr, 2cnd and t0062. The static Hydrophobicity analysis tool is mapped to tile color and dynamic Charge analysis tool is mapped to tile shape.

Analysis tool response to the state of the alignment is immediate and allows for real-time alignment and model evaluation.

DINAMO was written in Java to allow it to run as a Web-based applet or a local application. A major limitation inherent to Java security is that files cannot be read from or written to a local file system. Although these read/write limitations do not apply to the local application version, when run as a Web-based applet, alignments cannot be saved to the disk. They can, however, be displayed into a local text editor or e-mailed to the user in either FASTA format or in a printable format in which sequences are aligned and numbered. Similarly, structural coordinates cannot be read from a local disk, but may be pasted into an input window or retrieved directly from the Protein Data Bank (Table 1) by DINAMO.

Molecular graphics

The molecular graphics portion of DINAMO allows the visualization of a proposed model in the context of the 3D Cα backbone of the template structure as a ribbon diagram. By viewing the results of the analysis tools relative to the sequence-structure model, the user may ‘tune’ the alignment to best fit the template structure by, for example, moving regions of low sequence identity to loop regions or insertions/deletions.

Analysis tool information is mapped to the 3D model by color. When the sequence being modeled is moved relative to the template in the editor, the molecular graphics are updated to reflect the new positions of the amino acids. In this way, the user may view the results of changes in the alignment relative to the template, and thus determine the plausibility of the new alignment in the context of the model.

Insertions or deletions in the target sequence can be displayed in the molecular graphics using specific display attributes to reveal their position. Residues of interest may also be labeled in both the editor and the graphics to orient the user to specific locations in the model and alignment.

Because DINAMO does not calculate or optimize the conformation of the amino acid side chains, only the α-carbon backbone is displayed. The β-carbons of the template structure, however, may be optionally viewed to determine the orientation of the side chains. This is particularly useful for distinguishing exterior and interior residues in β-sheet regions.

Analysis tools

As the alignment is moved, the alignment information is sent to each of the active analysis tools via the Alignment Router class. Each analysis tool processes the information and returns
the results via a callback function in both the editor and graphics. Each analysis tool operates within its own Java thread, enabling several tools to work efficiently in parallel.

There are two types of analysis tools in DINAMO: dynamic and static. Dynamic analysis tools are alignment dependent and are updated as the alignment is adjusted, while static analysis tools are not. Hydrophobicity is an example of a static analysis tool: each residue tile in alignment is colored according to hydrophobicity, independent of its position in the alignment. Alternatively, Blosum62 is an example of a dynamic analysis tool. When mapped to tile color, each residue tile in the target sequence(s) is colored according to its degree of conservation to the corresponding residue in the template sequence as determined by the Blosum62 matrix (Henikoff and Henikoff, 1992). As the user adjusts the alignment, the color of each tile in the template and target sequence(s), as well as residue color in the model, are dynamically updated to reflect the state of the new alignment of the target against the template sequence. The classification of static versus dynamic only applies in relationship to the editor, as all analysis tools are dynamic with respect to the graphics. This is due to the fact that as the alignment is moved, the alignment will change with respect to the template sequence and thus residue position along the model will change. Available analysis tools include Blosum62, Hydrophobicity, Charge, Volume, Helix Propensity and Sheet Propensity.

Fig. 4. Molecular graphics. This figure shows t0062 threaded on the 1ndh structure with the dynamic Hydrophobicity analysis tool activated. Deletions are represented as strands. The Pie Chart tool is also shown displaying the results of dynamic Hydrophobicity analysis of t0062 versus 1ndh.
The analysis tool architecture is designed to be plug-in based so that the addition of future tools is a simple process. Each tool is a Java class that extends either dinamo.plugins.Static.java or dinamo.plugins.Dynamic.java, depending on whether the tool is dynamic or static. An analysis tool Java class file consists of the name of the tool, and the names of the categories it returns. Examples of dynamic return categories include conserved, non-conserved, etc. Examples of static return categories include insertions, hydrophobic and hydrophilic. The file also includes a data matrix that assigns each amino acid (static) or amino acid pair (dynamic) to the appropriate return category. To create a new analysis tool, the developer would use an existing analysis tool Java class file as a template, and replace the appropriate variables.

The analysis tools may be used singly or in combination by mapping the information to different display options. This allows the alignment to be evaluated simultaneously by multiple criteria. For example, tile color might reflect residue charge, while tile shape might reflect residue volume. By using the tools in combination, new patterns may emerge that may have been undetectable using only one type of analysis.

The Pie Chart tool displays analysis tool information by percentage categories and an overall score (for dynamic analysis tools only). For example, if the results of Blosum62 were activated in the Pie Chart, the Blosum62 categories (indel, not conserved, conserved, highly conserved, exactly conserved) would be displayed by percentage. This allows global alignment quality to be monitored while refining a local area of the alignment.

Implementation

We have used DINAMO to build a model for flavin reductase (CASP3 target ID: t0062) for submission to the 3rd annual Critical Assessment of Techniques for Protein Structure Prediction (CASP3) contest. Here we describe our model building process in order to illustrate how DINAMO might be applied to a real-world problem. (An extensive, on-line tutorial is also available on the DINAMO Web page.)

Establish template sequence/structure

The template sequence/structure was established in the following way. First, a BLAST (Table 1) search was performed with the target sequence to determine whether sequence homologs exist. The target sequence was simultaneously submitted to THREADER (Jones et al., 1992), a fold recognition program (Table 1), to determine putative template folds. Both BLAST and THREADER returned Protein Data Bank file 2cnd as the highest scoring hit. 2cnd, NADH-dependent nitrate reductase from *Escherichia coli*, was subsequently submitted to the FSSP database via the DALI server (Holm and Sander, 1998) (Table 1) to find structurally related homologs. This query returned a multiple structure-based sequence alignment of proteins structurally homologous to 2cnd. As 1ndh, pig liver cytochrome b₅ reductase, is the representative protein of this fold family, it was selected as the template structure. 1ndh and t0062 were then aligned by CLUSTALW (Thompson et al., 1994) to generate a complete, gapped alignment. The sequence identity of t0062 to 1ndh and 2cnd is ~17–18%. Therefore, although this is formally a case of comparative modeling, this very low level of sequence identity approaches modeling by fold recognition. The CLUSTALW-based alignment of 1ndh and t0062 was then manually merged with the FSSP alignment of 1ndh versus 2cnd versus 1fdr (flavodoxin reductase from *E.coli*). The merging process did not involve any alignment editing—the two sets of alignments were simply combined to provide the multiple alignment that was loaded into DINAMO and used as a starting point for further optimization. This alignment is shown in Figure 2, with the initial alignment of t0062 to 1ndh indicated by t0062 (INIT.).

Optimize sequence alignment/model

Initial refinement of the t0062 alignment was accomplished by using the Hydrophobicity, Charge and Volume analysis tools in both the static and dynamic format. Mapping these analysis tools to different display options allowed the information returned by each to be visualized and assessed simultaneously. Figure 3 shows the static version of Hydrophobicity mapped to tile color, and the dynamic version of Charge mapped to tile shape.

Based on these three properties, the target sequence was manually aligned to best fit the template sequence and thus structure. In this case, the structural homologs to 2cnd were also used as subtemplates and aided in achieving an overall sequence-based structural alignment. To facilitate alignment, the Pie Chart tool, displayed in Figure 4, was also employed to quantify alignment quality.

The plausibility of the resulting model was addressed in parallel with the sequence alignment optimization. Specifically, the Hydrophobicity analysis tool was activated in the model (Figure 4) and the likelihood of the residue position with respect to hydrophobicity was examined. The following criteria were used as guidelines in this process:

1. Maximization of hydrophobic core.
2. Maximization of exposed hydrophilic residues.
3. Maximization of hydrophobic residue–residue contacts (this was best achieved by examining side-chain orientation by viewing β-carbons in the model to note in which direction side chains pointed).

The Charge analysis tool was next activated in the model and a similar set of criteria were used:

1. Minimization of buried charge residues.
These parameters were observed as general guidelines in assessing model probability. However, these parameters are not absolute and corresponding residue positions in the original structure served as a reference point.

Generally speaking, a good alignment and resulting model will have a minimum number of insertions and deletions (indels) within elements of secondary structure. By viewing the model, we made an effort to restrict indels to loop regions. The final alignment submitted to CASP3 is shown in Figure 2, and is indicated by t0062 (FINAL).

Note added in proof
As of the time of publication, the three-dimensional structure of target 0062 has not been released. However, results are available for the other three targets for which we made predictions using DINAMO (0055, 0074, 0085). Our predictions for targets 0055 and 0085 scored highly.

Conclusions
As with any comparative modeling tool, the accuracy of a model built in DINAMO is inherently limited by the quality of the sequence alignment. The template structure simply provides a scaffold for the target sequence, and thus target residue position in the model ultimately depends upon the sequence alignment. Because no energy minimization, spatial calculations or side-chain positioning are performed in DINAMO, the model is solely a reflection of the alignment. Detailed amino acid side-chain orientation is difficult to predict (Vasquez, 1996), particularly for surface residues, and in most cases is not necessary. Therefore, DINAMO makes no presumptions and serves simply as a workbench to visualize available information and make reason-based calculations. As such, the model should be treated as a useful approximation of structure, not an actual structure. Models built using DINAMO can be extremely beneficial in rationalizing experimental design, but do not substitute for an actual crystal structure, not an actual structure. Models built using DINAMO are inherently limited by the quality of the sequence alignment. The template structure simply serves as a reference point. Not absolute and corresponding residue positions in the original structure served as a reference point. Generally speaking, a good alignment and resulting model will have a minimum number of insertions and deletions (indels) within elements of secondary structure. By viewing the model, we made an effort to restrict indels to loop regions. The final alignment submitted to CASP3 is shown in Figure 2, and is indicated by t0062 (FINAL).

Acknowledgements
We would like to thank Leslie Grate for providing the inspiration for DINAMO and for allowing us to use his in-house prototype program, SAE. This work was supported in part by NIH grant GM52885-R29 and a Junior Faculty Research award from the American Cancer Society to L.M.G.

References


